MEZCLAS FINITAS DE DISTRIBUCIONES NORMALES: UNA ALTERNATIVA PARA CLASIFICAR.

HÉCTOR JAVIER MOYANO NIÑO

UNIVERSIDAD INDUSTRIAL DE SANTANDER
FACULTAD DE CIENCIAS
ESCUELA DE MATEMÁTICAS
BUCARAMANGA
2007

MEZCLAS FINITAS DE DISTRIBUCIONES NORMALES: UNA ALTERNATIVA PARA CLASIFICAR.

HÉCTOR JAVIER MOYANO NIÑO

Monografía presentada como
requisito para optar al título
de *Licenciado en Matemáticas*

Henry Lamos Díaz
Director

UNIVERSIDAD INDUSTRIAL DE SANTANDER
FACULTAD DE CIENCIAS
ESCUELA DE MATEMÁTICAS
BUCARAMANGA
2007

Nota de aceptación

4.5 (Cuatro Cinco)

Henry Lamos Díaz
Director

Gabriel Yañez Canal
Calificador

Tulia Esther Rivera Floréz
Calificador

Bucaramanga, Agosto de 2007

Dedicatoria.

A la princesa por haberle regalado al guardián el idilio vivido;
este reposará por siempre en mi recuerdo.

Agradecimientos

Doy mi más profundo agradecimiento:

A Dios, por darme día a día la fuerza suficiente para seguir luchando por mis sueños.

A mis padres, Margarita Niño y Esteban A. Moyano; a mis hermanos Jhon Jairo, Luisa Fernanda y Andrés Felipe por su apoyo incondicional, compresión y aliento en los momentos difíciles.

A mi director de monografía Ph.D Henry Lamos Díaz por su invaluable colaboración y compromiso.

A mi Abuela, mi tío jaime, mis otros tíos y tías, mis primos y primas, y demás familiares que me han colaborado a través de los años.

A la señora Sandra, por las palabras justas en el momento preciso.

A mis amigos Isnardo, Carlos, Pacho, Guti, Hernán, Jose Luis, Arturo; a mis amigas Angelica, Juanita, Luisa, Diana, Maria, Betty, Andrea, y demás compañeros de carrera, por los momentos vividos.

A las Instituciones Ased, Grupo Gauss, CPE-UIS y La Presentación, por haberme dado la oportunidad de empezar a forjar mi vida laboral.

A todos mis profesores, de colegio y de universidad, por que de cada uno he tomado las herramientas necesarias que me han permitido construir mi proyecto de vida.

A todas aquellas personas que en su momento me han tendido la mano o me han dado una voz de aliento para seguir adelante.

TITTLE: FINITE MIXTURE OF NORMAL DISTRIBUTIONS: AN ALTERNATIVE TO CLASSIFY.[1].

AUTHOR: HÉCTOR JAVIER MOYANO NIÑO[2].

PALABRAS CLAVES:
Classification, Homogeneity, Likelihood, Mixture, Normal, Algorithm, Expectation, Maximization.

DESCRIPTION:

The classification is one of the most interest problems in the science, since all phenomenons should be ordered to be understood. Classify consists on dividing a heterogeneous population in homogeneous groups. The multivariate analysis data offers traditionally techniques directed to the supervised classification and no supervised one. In the some way other techniques have been built to classify such as the mixtures of distributions.

In this monograph the finite mixtures of normal distributions are presented, it consists in supposing the sample to classify divided in G groups or components, where to each one is assigned a certain number of elements, a function of normal distribution and a pondered weigh. Then, the classification problem through a finite mixture of normal distributions is solved when all the data you is used in the estimate of the parameters for each group. The parameters are estimated by the method of Maximum Likelihood.

The classification through mixtures of distributions offers techniques for such as purpose. Here is studied using the algorithm EM for mixtures. This is a recurrent method of optimization used to estimate the parameters of each group in which the observed variables are enlarged (Y), introducing not unites observed (Z) that has the function of indicating that component of the mixture comes each fact. The algorithm EM this made up of two alternate steps that involve an Expectation and maximization. It begins of a previous estimation of the parameters, then it finds the Expectation of the Log-Likelihood function L(Y, Z) conditioned to the parameters and the distribution of (Z) and it concludes with the maximization of this Expectation to find the new parameters. If the values of the found parameters converge to a fixed value, it stops the calculation and found parameters are shown.

---

[1]Monograph.

[2]DEPARTMENT OF SCIENCES, DEGREE IN MATHEMATICS. Supervisor: Henry Lamos Díaz.

TITULO: MEZCLAS FINITAS DE DISTRIBUCIONES NORMALES: UNA ALTERNATIVA PARA CLASIFICAR[1].

AUTOR: HÉCTOR JAVIER MOYANO NIÑO[2].

PALABRAS CLAVES:
Clasificación, Homogeneidad, verosimilitud, Mezcla, Normal, Algoritmo, Esperanza, Maximización.

DESCRIPCIÓN:

La clasificación es uno de los problemas de mayor interés en la ciencia ya que todo fenómeno debe ser ordenado para ser entendido. Clasificar consiste en dividir una población heterogénea en grupos homogéneos. El análisis de datos al nivel multivariado ofrece tradicionalmente técnicas dirigidas a la clasificación supervisada y no supervisada. Paralelamente se han construido otras técnicas para clasificar como lo son las mezclas de distribuciones.

En esta monografía se presentan las mezclas finitas de distribuciones normales, que consiste en suponer que la muestra a clasificar esta dividida en G grupos o componentes, donde a cada uno se le asigna un número determinado de elementos, una función de distribución normal y un peso de ponderación. Luego el problema de la clasificación a través de una mezcla finita de distribuciones normales queda resuelto al utilizar todos los datos muéstrales en la estimación de los parámetros para cada grupo. Los parámetros son estimados por el método de máxima verosimilitud.

La clasificación mediante mezclas de distribuciones ofrece técnicas para tal propósito. Aquí se estudia el algoritmo EM para mezclas. Este es un método de optimización iterativo usado para estimar los parámetros de cada grupo en el cual se amplían las variables observadas (Y), introduciendo unas no observadas (Z), que tiene la función de indicar de que componente de la mezcla proviene cada dato. El algoritmo EM esta compuesto de dos pasos alternados que involucran una esperanza y una maximización. Parte inicialmente de una estimación previa de los parámetros, luego halla la esperanza de la función de soporte L(Y,Z) condicionada a los parámetros y a la distribución de (Z) y finaliza con la maximización de esta esperanza para encontrar los nuevos parámetros. Si los valores de los parámetros hallados convergen a un valor fijo se detiene el cálculo y se muestran los parámetros hallados.

---

[1]Monografía.

[2]FACULTAD DE CIENCIAS, LICENCIATURA EN MATEMÁTICAS. Director Henry Lamos Díaz.

Índice general

Prólogo

Esta monografía aparte de ser un requisito para optar al titulo de Licenciado en Matemáticas, es el resultado de la revisión y el posterior estudio bibliográfico de las técnicas existentes para la clasificación de individuos en grupos más homogéneos. Aunque existe una buena cantidad de trabajos realizados entorno a esta temática, son muy escasos los escritos en español donde se muestren la clasificación de individuos que proceden de la mezcla finita de distribuciones normales multivariantes. Este tipo de clasificación ha venido tomando gran importancia en diferentes campos de investigación, ya que a partir de una serie de datos generados por una mezcla de G distribuciones normales y el uso de algunos métodos de agrupamiento, podemos particionar una muestra heterogénea en grupos mas homogéneos.

El presente trabajo esta compuesto de cuatro capítulos y cuatro apéndices.

El capítulo 1 presenta algunos conceptos de algebra matricial que en la mayoría de los cursos, o no son presentados, o se dejan como temas de profundización para el estudiante. Además en este capitulo se exponen conceptos básicos para el análisis estadístico multivariante, así como el estudio de la distribución normal para dimensión $p > 1$. Paralelamente se muestran en el **Apéndice A** ciertas distribuciones especiales que están estrechamente relacionadas con la distribución normal multivariante.

En el capítulo 2 se hace la revisión de algunos tópicos de Inferencia Multivatiante en distribuciones normales y la presentación de los métodos para realizar este proceso. Debido a la orientación de este trabajo, se han excluido de este capítulo la mayoría de los contrastes sobre la distribución normal multivariante para ser anexada mas adelante en el **Apéndice B**. En este apartado se pueden apreciar algunos ejemplos de estos contrastes, cosa que pocos libros realizan cuando exponen esta temática.

Ya con las herramientas dadas en los capítulos 1 y 2, podemos estudiar las mezclas de

distribuciones normales p-variantes expuestas en el capitulo 3. Aquí se realiza un completo estudio, partiendo de sus primeras apariciones históricas hasta la definición formal que hoy conocemos, así como el hallazgo de sus principales parámetros y la estimación de aquellos que hacen parte de sus componentes. Además se muestra como se hace la clasificación de los datos asumiendo una mezcla finita por medio del **algoritmo EM**. Adicionalmente el **Apéndice C** trata en forma detallada el método de las G medias (o k-medias) para mezclas.

Para el capitulo 4 se presenta una aplicación para el software comercial $MATLAB$, el cual efectúa el calculo de los parámetros componentes de mezcla, así como la clasificación de los datos en grupos altamente homogéneos. El **Apéndice D** muestra una tabla con datos mezclados, usada en un ejemplo ilustrativo del capitulo 4.

La evolución de los ordenadores y la facilidad que estos han mostrado para el procesamiento de grandes bases de datos, han permitido que el estudio de las técnicas multivariantes tengan grandes avances en los últimos años. Es por esto que la mayoría de los ejemplos mostrados en este trabajo han sido desarrollados con algunas aplicaciones del software comercial $MATLAB$.

Espero que esta monografía pueda ser útil a distintos tipos de audiencias, especialmente a aquellas personas que al igual que yo sean amantes de la estadística.

Héctor Javier Moyano Niño.

Capítulo 1

PRELIMINARES

1.1. Introducción

Aquí revisamos inicialmente ciertos conceptos del algebra matricial que por lo general no son tenidos en cuenta en un curso básico de algebra. Estos conceptos son: *la descomposición espectral de una matriz cuadrada*, necesaria en la selección del modelo estadístico que mas se ajusta a los datos muestrales; *las formas cuadráticas*, que son campos escalares donde la imagen esta formada por el producto combinado de un vector y una matriz; *la partición de una matriz* en componentes mas pequeñas denominadas *submatrices*, necesarias para el agrupamiento de datos muestrales en componentes mas sencillas y *las derivadas matriciales* con algunas propiedades básicas, indispensables en la demostración de ciertas propiedades que se verán mas adelante.
Luego se dan los conceptos básicos usados en el *análisis de datos multivariantes*, que no son otra cosa que la generalización (para el caso n-dimensional) de los ya estudiados en estadística unidimensional. Estos conceptos son: *variable aleatoria (vectorial)*, *distribuciones conjuntas*, *distribuciones marginales*, *distribuciones condicionadas* y *los parámetros básicos de una distribución (vector de medias y matriz de covarianzas)*. Además se hace una breve descripción de las esperanzas y varianzas condicionadas.
Finalizamos este capitulo con el estudio de *la distribución normal multivariante* donde describimos en forma detallada sus principales propiedades y hacemos un breve acercamiento geométrico para interpretación de los datos que presentan esta distribución. La distribución normal multivariante es el componente base en el estudio de las mezclas finitas de distribuciones normales.

1.2. Conceptos de álgebra matricial

1.2.1. Descomposición espectral

Sea $\mathbf{A}$ una matriz cuadrada y simétrica de orden p. Por algebra matricial se sabe que para esta clase de matrices los valores propios son números reales y los vectores propios son ortogonales. Luego la matriz $\mathbf{A}$ puede ser escrita como

$$\mathbf{A} = \mathbf{UDU}', \tag{1.1}$$

donde $\mathbf{D}$ es una matriz diagonal formada por los valores propios de $\mathbf{A}$ y $\mathbf{U}$ es una matriz ortogonal cuyas columnas son los vectores propios unitarios asociados con los elementos de la diagonal de la matriz $\mathbf{D}$. Esta propiedad se conoce con el nombre de **la descomposición espectral**. Llamando $\lambda_1, \ldots, \lambda_p$ a los valores propios de la matriz $\mathbf{A}$ y $\mathbf{u}_1, \ldots, \mathbf{u}_p$ a sus respectivos vectores propios, la descomposición dada en (1.1) puede escribirse:

$$\mathbf{A} = \sum_{i=1}^{p} \lambda_i \mathbf{u}_i \mathbf{u}_i', \tag{1.2}$$

que descompone la matriz $\mathbf{A}$ como la suma de p matrices de rango uno, $\mathbf{u}_i\mathbf{u}_i'$, con coeficientes λ_i.
Si la matriz $\mathbf{A}$ tiene rango r la descomposición espectral (1.2) indica que puede expresarse como suma de r matrices de rango unidad. La importancia de esta descomposición es que si algunos valores propios son muy pequeños, podemos reconstruir aproximadamente $\mathbf{A}$ utilizando los restantes valores y vectores propios.
Observemos que la descomposición espectral de $\mathbf{A}^{-1}$ es

$$\mathbf{A}^{-1} = \sum_{i=1}^{p} \lambda_i^{-1} \mathbf{u}_i \mathbf{u}_i',$$

ya que $\mathbf{A}^{-1}$ tiene los mismos vectores propios de $\mathbf{A}$ y valores propios λ_i^{-1}.

Ejemplo 1.1. Encontremos la descomposición espectral para la matriz

$$\mathbf{A} = \begin{pmatrix} 3 & 10 & 52 \\ 10 & 75 & 61 \\ 52 & 61 & 7 \end{pmatrix}$$

con el apoyo de la función de referencia *eig* de MATLAB 7.1.

```
>> A=[3 10 52; 10 75 61; 52 61 7] %.......Introducimos la matriz.
>> A =
     3    10    52
    10    75    61
    52    61     7
>> % Hallamos la matriz diagonal de valores propios D y la ortogonal
>> % de vectores propios U de la matriz A con la función (eig).
>> [U, D] = eig(A)
U =
   -0.5891    0.7481    0.3054
   -0.2996   -0.5532    0.7773
    0.7505    0.3664    0.5500
D =
  -58.1635         0         0
         0   21.0737         0
         0         0  122.0898
>> % Luego la descomposición espectral de la matriz A esta dada
>> % por el producto de matrices U*D*U'.
>> U*D*U'
ans =
    3.0000   10.0000   52.0000
   10.0000   75.0000   61.0000
   52.0000   61.0000    7.0000
```

1.2.2. Formas cuadráticas

Sea $\mathbf{A}$ una matriz simétrica de tamaño $(p \times p)$ y $\mathbf{x}$ un vector de tamaño $(p \times 1)$, la función

$$Q(\mathbf{x}) = \mathbf{x}'\mathbf{A}\mathbf{x}, \tag{1.3}$$

se llama una **forma cuadrática** de $\mathbf{x}$. $Q(\mathbf{x})$ es un escalar y puede ser expresado alternativamente por la ecuación

$$Q(\mathbf{x}) = \sum_{i=1}^{p}\sum_{j=1}^{p} a_{ij}x_i x_j = \sum_{i=1}^{p} a_{ii}x_i^2 + 2\sum_{i=1}^{p}\sum_{j=i+1}^{p} a_{ij}x_i x_j,$$

con a_{ij} elemento de la matriz $\mathbf{A}$, x_i y x_j elementos del vector $\mathbf{x}$.
Si $Q(\mathbf{x}) > 0$ para todo $\mathbf{x} \neq 0$, se dice que $\mathbf{A}$ es *definida positiva*. Si $Q(\mathbf{x}) \geq 0$ para todo

$\mathbf{x} \neq 0$, $\mathbf{A}$ se llama *semidefinida positiva*. Si $\mathbf{A}$ es definida positiva se nota $\mathbf{A} > 0$ y si es semidefinida positiva se nota $\mathbf{A} \geq 0$.
Algunas propiedades de las formas cuadráticas son:

1. Si $\mathbf{A} > 0$, entonces todos sus valores propios $\lambda_1, \ldots, \lambda_p$ son positivos. Si $\mathbf{A} \geq 0$, entonces $\lambda_i \geq 0$ para $i = 1, \ldots, p$ y $\lambda_i = 0$ para algún i.
2. Si $\mathbf{A} > 0$, entonces $\mathbf{A}$ es no singular y en consecuencia $|\mathbf{A}| > 0$.
3. Si $\mathbf{A} > 0$, entonces $\mathbf{A}^{-1} > 0$.
4. Si $\mathbf{A} > 0$ y $\mathbf{C}$ es una matriz no singular $(p \times p)$, entonces $\mathbf{C}'\mathbf{A}\mathbf{C} > 0$.

1.2.3. Partición de una matriz

A veces resulta mas cómodo expresar una matriz en forma de *submatrices* , es decir que los elementos que la conformen sean matices de tamaño más pequeño (sea por filas, columnas o ambos) que la original. En general sea $\mathbf{A}$ una matriz de tamaño $(n \times p)$, la matriz $\mathbf{A}$ se puede escribir así:

$$\mathbf{A} = \begin{pmatrix} \mathbf{A}_{11} & \ldots & \mathbf{A}_{1j} & \ldots & \mathbf{A}_{1p} \\ \vdots & \ddots & \vdots & \ddots & \vdots \\ \mathbf{A}_{i1} & \ldots & \mathbf{A}_{ij} & \ldots & \mathbf{A}_{ip} \\ \vdots & \ddots & \vdots & \ddots & \vdots \\ \mathbf{A}_{n1} & \ldots & \mathbf{A}_{nj} & \ldots & \mathbf{A}_{np} \end{pmatrix}, \tag{1.4}$$

donde la *submatriz* $\mathbf{A}_{ij}$ es de tamaño $(n_i \times p_j)$, con $\sum_{i=1}^{n} n_i = n$ y $\sum_{j=1}^{p} p_j = p$.
La suma y el producto entre este tipo de matrices se conforma de manera semejante a como se describen en las operaciones con matrices habituales. De esta forma, si las matrices $\mathbf{A}$ y $\mathbf{B}$ se particionan similarmente entonces

$$\mathbf{A} + \mathbf{B} = \begin{pmatrix} \mathbf{A}_{11} + \mathbf{B}_{11} & \ldots & \mathbf{A}_{1j} + \mathbf{B}_{1j} & \ldots & \mathbf{A}_{1p} + \mathbf{B}_{1p} \\ \vdots & \ddots & \vdots & \ddots & \vdots \\ \mathbf{A}_{i1} + \mathbf{B}_{i1} & \ldots & \mathbf{A}_{ij} + \mathbf{B}_{ij} & \ldots & \mathbf{A}_{ip} + \mathbf{B}_{ip} \\ \vdots & \ddots & \vdots & \ddots & \vdots \\ \mathbf{A}_{n1} + \mathbf{B}_{n1} & \ldots & \mathbf{A}_{nj} + \mathbf{B}_{nj} & \ldots & \mathbf{A}_{np} + \mathbf{B}_{np} \end{pmatrix}. \tag{1.5}$$

Si las matrices $\mathbf{A}$ y $\mathbf{B}$, son de tamaño $(m \times n)$ y $(n \times p)$, respectivamente, y se particionan adecuadamente para el producto, se tiene que

$$\mathbf{AB} = \begin{pmatrix} \sum_{k=1}^{n} \mathbf{A}_{1k}\mathbf{B}_{k1} & \cdots & \sum_{k=1}^{n} \mathbf{A}_{1k}\mathbf{B}_{kj} & \cdots & \sum_{k=1}^{n} \mathbf{A}_{1k}\mathbf{B}_{kp} \\ \vdots & \ddots & \vdots & \ddots & \vdots \\ \sum_{k=1}^{n} \mathbf{A}_{ik}\mathbf{B}_{k1} & \cdots & \sum_{k=1}^{n} \mathbf{A}_{ik}\mathbf{B}_{kj} & \cdots & \sum_{k=1}^{n} \mathbf{A}_{ik}\mathbf{B}_{kp} \\ \vdots & \ddots & \vdots & \ddots & \vdots \\ \sum_{k=1}^{n} \mathbf{A}_{mk}\mathbf{B}_{k1} & \cdots & \sum_{k=1}^{n} \mathbf{A}_{mk}\mathbf{B}_{kj} & \cdots & \sum_{k=1}^{n} \mathbf{A}_{mk}\mathbf{B}_{kp} \end{pmatrix}. \tag{1.6}$$

Para una matriz $\mathbf{A}$ particionada en la siguiente forma

$$\mathbf{A} = \begin{pmatrix} \mathbf{A}_{11} & \mathbf{A}_{12} \\ \mathbf{A}_{21} & \mathbf{A}_{22} \end{pmatrix}, \tag{1.7}$$

donde $\mathbf{A}_{11}$ y $\mathbf{A}_{22}$ son matrices cuadradas *no singulares* , la inversa de $\mathbf{A}$ se calcula mediante

$$\mathbf{A}^{-1} = \begin{bmatrix} \mathbf{B}^{-1} & -\mathbf{B}^{-1}\mathbf{A}_{12}\mathbf{A}_{22}^{-1} \\ -\mathbf{A}_{22}^{-1}\mathbf{A}_{21}\mathbf{B}^{-1} & \mathbf{A}_{22}^{-1} + \mathbf{A}_{22}^{-1}\mathbf{A}_{21}\mathbf{B}^{-1}\mathbf{A}_{12}\mathbf{A}_{22}^{-1} \end{bmatrix}, \tag{1.8}$$

donde

$$\mathbf{B} = (\mathbf{A}_{11} - \mathbf{A}_{12}\mathbf{A}_{22}^{-1}\mathbf{A}_{21}).$$

El determinante de la matriz $\mathbf{A}$ se puede calcular basándonos en su partición, en tanto las submatrices $\mathbf{A}_{ii}$, con $i = 1, 2$, sean no singulares. Es decir,

$$|\mathbf{A}| = |\mathbf{A}_{11}| \cdot |\mathbf{A}_{22} - \mathbf{A}_{21}\mathbf{A}_{11}^{-1}\mathbf{A}_{12}|, \tag{1.9}$$

o de igual manera

$$|\mathbf{A}| = |\mathbf{A}_{22}| \cdot |\mathbf{A}_{11} - \mathbf{A}_{12}\mathbf{A}_{22}^{-1}\mathbf{A}_{21}| = |\mathbf{A}_{22}| \cdot |\mathbf{B}|. \tag{1.10}$$

Ejemplo 1.2. La matriz

$$\mathbf{A} = \begin{pmatrix} 6 & 11 & 12 & 10 & 1 \\ 6 & 7 & 10 & 6 & 2 \\ 21 & 11 & 4 & 6 & 3 \\ 0 & 4 & 12 & 10 & 4 \\ 1 & 2 & 3 & 4 & 5 \end{pmatrix},$$

puede escribirse como una matriz 2×2 particionada:

$$\mathbf{A} = \begin{pmatrix} \mathbf{A}_{11} & \mathbf{A}_{12} \\ \mathbf{A}_{21} & \mathbf{A}_{22} \end{pmatrix} = \left(\begin{array}{ccc:cc} 6 & 11 & 12 & 10 & 1 \\ 6 & 7 & 10 & 6 & 2 \\ 21 & 11 & 4 & 6 & 3 \\ \hdots & \hdots & \hdots & \hdots & \hdots \\ 0 & 4 & 12 & 10 & 4 \\ 1 & 2 & 3 & 5 & 4 \end{array}\right),$$

donde

$$\mathbf{A}_{11} = \begin{pmatrix} 6 & 11 & 12 \\ 6 & 7 & 10 \\ 21 & 11 & 4 \end{pmatrix}, \quad \mathbf{A}_{12} = \begin{pmatrix} 10 & 1 \\ 6 & 2 \\ 6 & 3 \end{pmatrix}, \quad \mathbf{A}_{21} = \begin{pmatrix} 0 & 4 & 12 \\ 1 & 2 & 3 \end{pmatrix}, \quad \mathbf{A}_{22} = \begin{pmatrix} 10 & 4 \\ 5 & 4 \end{pmatrix}.$$

Podemos obtener la inversa y el determinate de esta matriz particionada según las ecuaciones (1.8) y (1.10). Con la ayuda de MATLAB 7.1 encontramos los siguientes resultados.

```
>> % Introduzcamos la matriz A.
>> A=[6 11 12 10 1; 6 7 10 6 2; 21 11 4 6 3; 0 4 12 10 4; 1 2 3 4 5];
>> % La instrucción Aij = A(k:l,m:n) muestra las submatrices de A con filas
>> % de la (k) a la (l) y  columnas de la (m) a la (n), donde i,j=1,2 ;
>> % k,m =1,4 y l,n=3,5.
>> A11=A(1:3,1:3); A12=A(1:3,4:5);
>> A21=A(4:5,1:3); A22=A(4:5,4:5);
>> % Hallemos la matriz B de la ecuación (1.8) para luego encontrar
>> % la inversa de A que llamaremos AI.
>> B= A11 -(A12*inv(A22)*A21)
B =
    6.8824    7.3529   -1.5882
    6.1176    4.6471    2.5882
   20.8235    8.5294   -2.8824
>> % Llamando AI11, AI12, AI21, AI22 a las submatrices componentes de la
>> % matriz inversa dada en (1.8), se tiene que:
>> AI11=inv(B)
AI11 =
```

```
   -0.1006    0.0217    0.0749
    0.2028    0.0375   -0.0781
   -0.1264    0.2677   -0.0369
>> AI12= (-inv(B))*(A12)*(inv(A22))
AI12 =
    0.0824   -0.0994
   -0.2574    0.1972
    0.0173   -0.0736
>> AI21 = (-inv(A22))*(A21)*(inv(B))
AI21 =
    0.0951   -0.3887    0.0884
   -0.0612    0.1309   -0.0324
>> AI22 = inv(A22)+(inv(A22)*(A21)*(inv(B))*(A12)*(inv(A22)))
AI22 =
    0.2231   -0.0951
   -0.1024    0.2612
>> % Luego la inversa de A se obtiene así:
>> AI= [AI11 AI12; AI21 AI22]
AI =
   -0.1006    0.0217    0.0749    0.0824   -0.0994
    0.2028    0.0375   -0.0781   -0.2574    0.1972
   -0.1264    0.2677   -0.0369    0.0173   -0.0736
    0.0951   -0.3887    0.0884    0.2231   -0.0951
   -0.0612    0.1309   -0.0324   -0.1024    0.2612
>> % Nótese que es el mismo resultado obtenido al usar la función inv(A).
>> inv(A)
ans =
   -0.1006    0.0217    0.0749    0.0824   -0.0994
    0.2028    0.0375   -0.0781   -0.2574    0.1972
   -0.1264    0.2677   -0.0369    0.0173   -0.0736
    0.0951   -0.3887    0.0884    0.2231   -0.0951
   -0.0612    0.1309   -0.0324   -0.1024    0.2612
>> % El determinante se halla con la ecuación (1.10)
>> dete_A = (det(A22))*(det(B))
dete_A =
  1.1990e+004
```

```
>> % Obtenemos el mismo resultado con la función det(A).
>> det(A)
ans =
       11990
```

1.2.4. Derivadas matriciales

Definición 1.1. Sea f una función que asigna a un vector $\mathbf{x} \in \mathbb{R}^p$ un número real, esquemáticamente esto es:

$$f : \mathbb{R}^p \longrightarrow \mathbb{R}$$
$$\mathbf{x} = \begin{pmatrix} x_1 \\ \vdots \\ x_p \end{pmatrix} \longmapsto f(\mathbf{x}).$$

Se define la derivada de $f(\mathbf{x})$ con respecto al vector $\mathbf{x}$ de tamaño $(p \times 1)$ como

$$\frac{\partial f(\mathbf{x})}{\partial \mathbf{x}} = \begin{bmatrix} \frac{\partial f}{\partial x_1} \\ \vdots \\ \frac{\partial f}{\partial x_p} \end{bmatrix} = \left[\frac{\partial f(\mathbf{x})}{\partial x_i} \right]. \tag{1.11}$$

Los siguientes resultados son consecuencia de la definición 1.1.

1. Si $f(\mathbf{x}) = \mathbf{a}'\mathbf{x} = a_1x_1 + \cdots + a_px_p$ donde $\mathbf{a}$ y $\mathbf{x}$ son vectores de $\mathbb{R}^p$, la derivada de la función f respecto al vector $\mathbf{x}$, de acuerdo con la ecuación 1.11, está dada por:

$$\frac{\partial f(\mathbf{x})}{\partial \mathbf{x}} = \frac{\partial}{\partial \mathbf{x}}(\mathbf{a}'\mathbf{x}) = \frac{\partial}{\partial \mathbf{x}}(\mathbf{x}'\mathbf{a}) = \begin{bmatrix} \frac{\partial f}{\partial x_1} \\ \vdots \\ \frac{\partial f}{\partial x_p} \end{bmatrix} = \begin{pmatrix} a_1 \\ \vdots \\ a_p \end{pmatrix} = \mathbf{a}. \tag{1.12}$$

2. Sea $Q(\mathbf{x}) = \mathbf{x}'\mathbf{A}\mathbf{x}$, donde $\mathbf{A}$ es cuadrada y simétrica. La derivada de la función Q respecto al vector $\mathbf{x}$ es:

$$\begin{aligned} \frac{\partial Q(\mathbf{x})}{\partial \mathbf{x}} &= \frac{\partial}{\partial \mathbf{x}}(\mathbf{x}'\mathbf{A}\mathbf{x}) \\ &= \frac{\partial}{\partial \mathbf{x}} \left(\sum_{i=1}^{p} a_{ii}x_i^2 + 2 \sum_{i=1}^{p} \sum_{j=i+1}^{p} a_{ij}x_ix_j \right), \end{aligned}$$

luego tendremos que:

$$\frac{\partial}{\partial x_1}(\mathbf{x}'\mathbf{A}\mathbf{x}) = 2a_{11}x_1 + \cdots + 2a_{1p}x_p = 2\mathbf{a_1}'\mathbf{x}$$

$$\vdots$$

$$\frac{\partial}{\partial x_p}(\mathbf{x}'\mathbf{A}\mathbf{x}) = 2a_{p1}x_1 + \cdots + 2a_{pp}x_p = 2\mathbf{a_p}'\mathbf{x}$$

donde $\mathbf{a_i}'$ es la i-esima fila de la matriz. Por lo tanto

$$\frac{\partial}{\partial \mathbf{x}}(\mathbf{x}'\mathbf{A}\mathbf{x}) = \begin{bmatrix} 2\mathbf{a_1}'\mathbf{x} \\ \vdots \\ 2\mathbf{a_p}'\mathbf{x} \end{bmatrix} = 2\mathbf{A}\mathbf{x}. \tag{1.13}$$

Ejemplo 1.3. Calculemos la derivada con respecto al vector $\mathbf{x} = (x_1, x_2)'$ de las siguientes funciones:

1. $f_1(\mathbf{x}) = 2x_1 + 3x_2$.
$$\frac{\partial f_1(\mathbf{x})}{\partial \mathbf{x}} = \begin{bmatrix} \frac{\partial f_1(\mathbf{x})}{\partial x_1} \\ \frac{\partial f_1(\mathbf{x})}{\partial x_2} \end{bmatrix} = \begin{bmatrix} 2 \\ 3 \end{bmatrix}.$$

2. $f_2(\mathbf{x}) = 4x_1^2 + 3x_1x_2$.
$$\frac{\partial f_2(\mathbf{x})}{\partial \mathbf{x}} = \begin{bmatrix} \frac{\partial f_2(\mathbf{x})}{\partial x_1} \\ \frac{\partial f_2(\mathbf{x})}{\partial x_2} \end{bmatrix} = \begin{bmatrix} 8x_1 + 3x_2 \\ 3x_1 \end{bmatrix}.$$

3. $f_3(\mathbf{x}) = 3x_1^4x_2^3 + 2x_1^2x_2^2 - 7x_1x_2^3 + 6$.
$$\frac{\partial f_2(\mathbf{x})}{\partial \mathbf{x}} = \begin{bmatrix} \frac{\partial f_3(\mathbf{x})}{\partial x_1} \\ \frac{\partial f_3(\mathbf{x})}{\partial x_2} \end{bmatrix} = \begin{bmatrix} 12x_1^3x_2^3 + 4x_1x_2^2 - 7x_2^3 \\ 9x_1^4x_2^2 + 4x_1^2x_2 - 21x_1x_2^2 \end{bmatrix} = \begin{bmatrix} x_2^2(12x_1^3x_2 + 4x_1 - 7x_2) \\ x_1x_2(9x_1^3x_2 + 4x_1 - 21x_2) \end{bmatrix}.$$

Definición 1.2. Sea f una función que asigna a una matriz $\mathbf{Y} \in (\mathbb{R}^n \times \mathbb{R}^p)$ un número real, esquemáticamente esto es:

$$f: \quad (\mathbb{R}^n \times \mathbb{R}^p) \quad \longrightarrow \quad \mathbb{R}$$

$$\mathbf{Y} = \begin{pmatrix} y_{11} & \cdots & y_{1p} \\ \vdots & \ddots & \vdots \\ y_{n1} & \cdots & y_{np} \end{pmatrix} \longmapsto \quad f(\mathbf{Y}).$$

Se define la derivada de $f(\mathbf{Y})$ con respecto a la matriz $\mathbf{Y}$ de tamaño $(n \times p)$ como

$$\frac{\partial f(\mathbf{Y})}{\partial \mathbf{Y}} = \frac{\partial f}{\partial \mathbf{Y}} \begin{pmatrix} y_{11} & \cdots & y_{1p} \\ \vdots & \ddots & \vdots \\ y_{n1} & \cdots & y_{np} \end{pmatrix} = \begin{pmatrix} \frac{\partial f}{y_{11}} & \cdots & \frac{\partial f}{y_{1p}} \\ \vdots & \ddots & \vdots \\ \frac{\partial f}{y_{n1}} & \cdots & \frac{\partial f}{y_{np}} \end{pmatrix} = \left(\frac{\partial f(\mathbf{Y})}{\partial y_{ij}} \right). \tag{1.14}$$

Los siguientes resultados son consecuencia de la definición 1.2.

1. Si $f(\mathbf{Y}) = \mathbf{a}'\mathbf{Y}\mathbf{b}$, donde $\mathbf{a} \in \mathbb{R}^n, \mathbf{Y} \in (\mathbb{R}^n \times \mathbb{R}^p)$ y $\mathbf{b} \in \mathbb{R}^p$, entonces la derivada de la función f respecto a la matriz $\mathbf{Y}$ está dada por:

$$\frac{\partial}{\partial \mathbf{Y}}(\mathbf{a}'\mathbf{Y}\mathbf{b}) = \mathbf{b}\mathbf{a}' \tag{1.15}$$

2. Sea $f(\mathbf{Y}) = \mathbf{a}'\mathbf{Y}'\mathbf{Y}\mathbf{b}$, entonces la derivada de f viene dada por:

$$\frac{\partial}{\partial \mathbf{Y}}(\mathbf{a}'\mathbf{Y}'\mathbf{Y}\mathbf{b}) = (\mathbf{a}\mathbf{b}' + \mathbf{b}\mathbf{a}')\mathbf{Y}' \tag{1.16}$$

Definición 1.3. Dado un vector $\mathbf{y}$ cuyos componentes son funciones f_i de un vector de variables $\mathbf{x}' = (x_1, \ldots, x_p)$, definimos la derivada de $\mathbf{y}$ respecto a $\mathbf{x}$ como la matriz cuyas columnas son las derivadas de las componentes f_i respecto a $\mathbf{x}$. Es decir, si

$$\mathbf{y}' = (f_1(\mathbf{x}), \ldots, f_p(\mathbf{x})),$$

entonces

$$\frac{\partial \mathbf{y}}{\partial \mathbf{x}} = \left[\frac{\partial f_1}{\partial \mathbf{x}}, \ldots, \frac{\partial f_p}{\partial \mathbf{x}} \right] = \begin{bmatrix} \frac{\partial f_1}{\partial x_1} & \cdots & \frac{\partial f_p}{\partial x_1} \\ \vdots & \ddots & \vdots \\ \frac{\partial f_1}{\partial x_p} & \cdots & \frac{\partial f_p}{\partial x_p} \end{bmatrix},$$

a esta matriz de derivadas se le denomina *matriz jacobiana*.

Ejemplo 1.4. Calculemos la derivada con respecto al vector $\mathbf{x} = (x_1, x_2)'$ de las siguientes funciones vectoriales, construidas con las funciones $f_i(\mathbf{x})$ del ejemplo (1.3):

1. $h_1(\mathbf{x}) = (f_1(\mathbf{x}), f_2(\mathbf{x}), f_3(\mathbf{x}))'$.

$$\frac{\partial h_1(\mathbf{x})}{\partial \mathbf{x}} = \begin{bmatrix} \frac{\partial f_1(\mathbf{x})}{\partial x_1} & \frac{\partial f_2(\mathbf{x})}{\partial x_1} & \frac{\partial f_3(\mathbf{x})}{\partial x_1} \\ \frac{\partial f_1(\mathbf{x})}{\partial x_2} & \frac{\partial f_2(\mathbf{x})}{\partial x_2} & \frac{\partial f_3(\mathbf{x})}{\partial x_2} \end{bmatrix} = \begin{bmatrix} 2 & 8x_1 + 3x_2 & x_2^2(12x_1^3x_2 + 4x_1 - 7x_2) \\ 3 & 3x_1 & x_1x_2(9x_1^3x_2 + 4x_1 - 21x_2) \end{bmatrix}.$$

2. $h_2(\mathbf{x}) = (g_1(\mathbf{x}), g_2(\mathbf{x}))'$, donde $g_1(\mathbf{x}) = 2f_1(\mathbf{x}) + 5f_2(\mathbf{x})$ y $g_2(\mathbf{x}) = -6f_3(\mathbf{x})$. Puesto que

$$\begin{aligned} g_1(\mathbf{x}) &= 2f_1(\mathbf{x}) + 5f_2(\mathbf{x}) \\ &= 2(2x_1 + 3x_2) + 5(4x_1^2 + 3x_1x_2) \\ &= 20x_1^2 + 4x_1 + 15x_1x_2 + 3x_2 \end{aligned}$$

entonces

$$\frac{\partial g_1(\mathbf{x})}{\partial \mathbf{x}} = \begin{bmatrix} \frac{\partial g_1(\mathbf{x})}{\partial x_1} \\ \frac{\partial g_1(\mathbf{x})}{\partial x_2} \end{bmatrix} = \begin{bmatrix} 4(10x_1 + 1) + 15x_2 \\ 3(5x_1 + 1) \end{bmatrix}.$$

De igual forma se tiene que

$$g_2(\mathbf{x}) = -6f_3(\mathbf{x}) = -6(3x_1^4x_2^3 + 2x_1^2x_2^2 - 7x_1x_2^3 + 6)$$

entonces

$$\frac{\partial g_2(\mathbf{x})}{\partial \mathbf{x}} = \begin{bmatrix} \frac{\partial g_2(\mathbf{x})}{\partial x_1} \\ \frac{\partial g_2(\mathbf{x})}{\partial x_2} \end{bmatrix} = -6 \begin{bmatrix} x_2^2(12x_1^3x_2 + 4x_1 - 7x_2) \\ x_1x_2(9x_1^3x_2 + 4x_1 - 21x_2) \end{bmatrix}.$$

Luego

$$\frac{\partial h_2(\mathbf{x})}{\partial \mathbf{x}} = \begin{bmatrix} \frac{\partial g_1(\mathbf{x})}{\partial x_1} & \frac{\partial g_2(\mathbf{x})}{\partial x_1} \\ \frac{\partial g_1(\mathbf{x})}{\partial x_2} & \frac{\partial g_2(\mathbf{x})}{\partial x_2} \end{bmatrix} = \begin{bmatrix} 4(10x_1 + 1) + 15x_2 & -6x_2^2(12x_1^3x_2 + 4x_1 - 7x_2) \\ 3(5x_1 + 1) & -6x_1x_2(9x_1^3x_2 + 4x_1 - 21x_2) \end{bmatrix}.$$

Teorema 1.1. *Si* $\mathbf{y} = \mathbf{Ax}$*, donde* $\mathbf{A}$ *es una matriz cualquiera, entonces*

$$\frac{\partial \mathbf{y}}{\partial \mathbf{x}} = \frac{\partial}{\partial \mathbf{x}}(\mathbf{Ax}) = \mathbf{A}'.$$

Demostración. Para deducir este resultado de la definición anterior, escribamos la matriz $\mathbf{A}$ como:

$$\mathbf{A} = \begin{bmatrix} \mathbf{a_1}' \\ \vdots \\ \mathbf{a_p}' \end{bmatrix},$$

donde cada $\mathbf{a_i}'$ es una fila de la matriz. Entonces:

$$\mathbf{y} = \mathbf{Ax} = \begin{bmatrix} \mathbf{a_1}'\mathbf{x} \\ \vdots \\ \mathbf{a_p}'\mathbf{x} \end{bmatrix} = \begin{bmatrix} f_1(\mathbf{x}) \\ \vdots \\ f_p(\mathbf{x}) \end{bmatrix},$$

con lo que:

$$\frac{\partial f_i(\mathbf{x})}{\partial \mathbf{x}} = \frac{\partial}{\partial \mathbf{x}}(\mathbf{a_i}'\mathbf{x}) = \mathbf{a_i},$$

por lo tanto, según lo anterior:

$$\frac{\partial \mathbf{y}}{\partial \mathbf{x}} = [\mathbf{a_1}, \ldots, \mathbf{a_p}] = \mathbf{A}'.$$

■

Otras propiedades.

Sea $\mathbf{X}$ una matriz cuadrada y no singular. De las anteriores definiciones se deduce las siguientes propiedades:

$a) \frac{\partial |\mathbf{X}|}{\partial \mathbf{X}} = |\mathbf{X}|(\mathbf{X}')^{-1}$

$b) \frac{\partial \ln |\mathbf{X}|}{\partial \mathbf{X}} = \left(\frac{1}{|\mathbf{X}|}\right) \frac{\partial |\mathbf{X}|}{\partial \mathbf{X}} = (\mathbf{X}')^{-1}$

$c) \frac{\partial tra(\mathbf{X})}{\partial \mathbf{X}} = \mathbf{I}$

$d) \frac{\partial tra(\mathbf{XB})}{\partial \mathbf{X}} = \mathbf{B}'$

$e) \frac{\partial tra(\mathbf{BXC})}{\partial \mathbf{X}} = \mathbf{B}'\mathbf{C}'$

$f) \frac{\partial tra(\mathbf{X}'\mathbf{AXB})}{\partial \mathbf{X}} = \mathbf{BX}'\mathbf{A} + \mathbf{B}'\mathbf{X}'\mathbf{A}'$

$g) \frac{\partial tra(\mathbf{BX}^{-1})}{\partial \mathbf{X}} = -(\mathbf{X}^{-1}\mathbf{BX}^{-1})$

además, si $\mathbf{X}$ es simétrica se tiene:

$$\frac{\partial |\mathbf{X}|}{\partial \mathbf{X}} = |\mathbf{X}|(2\mathbf{X}^{-1} - diag(\mathbf{X}^{-1}))$$
$$\frac{\partial \ln |\mathbf{X}|}{\partial \mathbf{X}} = \left(\frac{1}{|\mathbf{X}|}\right) \frac{\partial |\mathbf{X}|}{\partial \mathbf{X}} = 2\mathbf{X}^{-1} - diag(\mathbf{X}^{-1})$$
$$\frac{\partial tra(\mathbf{XB})}{\partial \mathbf{X}} = \mathbf{B} + \mathbf{B}' - diag(\mathbf{B}).$$

1.3. Conceptos básicos del análisis de datos multivariantes

El análisis multivariante de datos trata la asociación de conjuntos de medidas sobre un numero de individuos u objetos. El conjunto de individuos junto con sus variables pueden disponerse una matriz $\mathbb{X}$ como se define a continuación:

Definición 1.4. Se define **la matriz de datos multivariantes** como el arreglo

$$\mathbb{X} = \begin{pmatrix} x_{11} & x_{12} & \dots & x_{1p} \\ x_{21} & x_{22} & \dots & x_{2p} \\ \vdots & \vdots & \ddots & \vdots \\ x_{n1} & x_{n2} & \dots & x_{np} \end{pmatrix}, \tag{1.17}$$

donde x_{ij} es la observación de la $j-esima$ variable en el $i-esimo$ individuo.

La matriz $\mathbb{X}$ muestra los valores observados de p-variables numéricas en un conjunto de n-elementos. Cada una de estas p-variables se denomina una variable **escalar o univariada** y el conjunto de las p-variables forman una variable **vectorial o multivariante**.
La matriz $\mathbb{X}$ puede definirse como el arreglo de vectores fila o vectores columna. Así $\mathbf{x_i'} = (x_{i1}, \dots, x_{ip})$ es el vector fila que contiene las observaciones de todas las variables del $i-esimo$ individuo y $\mathbf{X_j} = (x_{1j}, \dots, x_{nj})'$ es vector de la columna que contiene las observaciones de la $j-esima$ variable. Luego podemos escribir la matriz de datos así:

$$\mathbb{X} = \begin{pmatrix} \mathbf{x_1'} \\ \mathbf{x_2'} \\ \vdots \\ \mathbf{x_n'} \end{pmatrix} = \begin{pmatrix} \mathbf{X_1} & \mathbf{X_2} & \dots & \mathbf{X_p} \end{pmatrix}.$$

Definimos además *la media muestral* de la $j-esima$ variable por

$$\overline{\mathbf{x}}_\mathbf{j} = \frac{1}{n}\sum_{i=1}^{n} x_{ij}, \quad con \quad j = 1, \dots, p, \tag{1.18}$$

luego *el vector de medias muestral* esta dado por:

$$\overline{\mathbf{x}} = \begin{pmatrix} \overline{\mathbf{x}}_\mathbf{1} \\ \vdots \\ \overline{\mathbf{x}}_\mathbf{p} \end{pmatrix} \tag{1.19}$$

La matriz de varianzas y covarianzas muestral esta dada por

$$\mathbf{S} = \begin{pmatrix} s_{11} & s_{12} & \dots & s_{1p} \\ s_{21} & s_{22} & \dots & s_{2p} \\ \vdots & \vdots & \ddots & \vdots \\ s_{p1} & s_{p2} & \dots & s_{pp} \end{pmatrix}, \tag{1.20}$$

donde

$$s_{jk} = \frac{1}{n}\sum_{i=1}^{n}(x_{ij} - \overline{\mathbf{x}}_\mathbf{j})(x_{ik} - \overline{\mathbf{x}}_\mathbf{k}); \quad j,k = 1,\ldots,p, \tag{1.21}$$

es la covarianza muestral entre la variable j y la variable k. Nótese que si $j = k$, se obtiene la varianza muestral asociada a la variable j-esima. La matriz $\mathbf{S}$ es simétrica; es decir $s_{jk} = s_{kj}$, para todas las entradas $j,k = 1,\ldots,p$.
La matriz de correlación la definimos así:

$$\mathbf{R} = \begin{pmatrix} 1 & r_{12} & \ldots & r_{1p} \\ r_{21} & 1 & \ldots & r_{2p} \\ \vdots & \vdots & \ddots & \vdots \\ r_{p1} & r_{p2} & \ldots & 1 \end{pmatrix}, \tag{1.22}$$

con

$$r_{jk} = \frac{s_{jk}}{\sqrt{s_{jj}s_{jk}}}, -1 \leq r_{jk} \leq 1, \tag{1.23}$$

de igual forma, esta matriz es simétrica.
Una característica muy importante de un conjunto de datos es su *homogeneidad*; si las desviaciones, $(x_{ij} - \overline{x}_j)^2$, son muy distintas es porque hay datos que se alejan mucho de la media y por lo tanto tenemos alta *heterogeneidad*.

Ejemplo 1.5. Los siguientes datos hacen referencia a los kilómetros promedio por galón $\mathbf{X}_1$ (en Km), la velocidad maxima $\mathbf{X_2}$ (en Km/h), el peso X_3 (en toneladas), el volumen $\mathbf{X}_4$ (en m^3) y potencia del motor $\mathbf{X}_5$ (en hp) de 18 diversas clases de camionetas. Los datos se representan en el cuadro 1.1.

Clase	$\mathbf{X_1}$	$\mathbf{X_2}$	$\mathbf{X_3}$	$\mathbf{X_4}$	$\mathbf{X_5}$	Clase	$\mathbf{X_1}$	$\mathbf{X_2}$	$\mathbf{X_3}$	$\mathbf{X_4}$	$\mathbf{X_5}$
1	23	190	2.41	14.37	212	10	25	160	2.75	14.18	126
2	32	185	2.51	10.37	150	11	23	220	2.95	18.80	260
3	23	200	2.96	16.85	215	12	23	220	3.06	18.80	260
4	35	185	2.82	15.26	185	13	28	180	1.51	10.56	75
5	23	185	2.58	12.87	140	14	28	180	1.55	10.56	105
6	24	190	2.59	13.37	140	15	33	200	1.22	11.01	141
7	23	185	2.58	13.17	140	16	40	181	1.35	9.25	81
8	24	190	3.03	15.60	212	17	33	205	1.71	11.22	125
9	25	160	2.37	12.13	99	18	35	155	1.20	9.67	67

Cuadro 1.1: Datos experimentales para el Ejemplo 1.5

Encontremos el vector de medias $\overline{x}$, la matriz de covarianza muestral $\mathbf{S}$ y la matriz de correlación $\mathbf{R}$ usando las funciones de referencia *Statistics* y *Correlation* de MATLAB 7.1. Previamente se importa la tabla de datos de *Excel* y se archiva como un documento $*.mat$, luego ejecutamos la siguiente serie de comandos

```
>> Data=load('ejemplo15.mat') %..... Carga los datos.
Data =
    X: [18x5 double] %...Matriz de datos X de dimensión 18x5.
>> X1=X(:,1); %.........Guarda la primera columna en la variable X1.
>> X2=X(:,2); %                 .
>> X3=X(:,3); %                 .
>> X4=X(:,4); %                 .
>> X5=X(:,5); %.........Guarda la quinta columna en la variable X5.
>> X1bar=mean(X1) %......Encuentra la media de X1.
X1bar =
   27.7778
>> % Hay que tener cuidado con MATLAB, pues este arroja siempre vectores
>> % filas y nosotros asumimos los vectores como vectores columnas.
>> Tra_Xbar=mean(X) %.....Encuentra el vector de medias transpuesto.
Tra_Xbar =
   27.7778  187.2778    2.2861   13.2267  151.8333
>> Xbar =(Tra_Xbar)' %....Este es el vector de medias.
Xbar =
    27.7778
   187.2778
     2.2861
    13.2267
   151.83
>> s11= var(X1) %........Encuentra la varianza de X1.
s11 =
   29.8301
>> S=cov(X) %............Encuentra la matriz de covarianzas muestral.
S =
  1.0e+003 *
    0.0298   -0.0243   -0.0025   -0.0104   -0.1732
   -0.0243    0.3234    0.0041    0.0319    0.8103
   -0.0025    0.0041    0.0004    0.0017    0.0306
```

```
   -0.0104    0.0319    0.0017    0.0087    0.1614
   -0.1732    0.8103    0.0306    0.1614    3.5741
>> R=corrcoef(X) %.......Encuentra la matriz correlación.
R =
    1.0000   -0.2479   -0.6845   -0.6481   -0.5305
   -0.2479    1.0000    0.3384    0.6009    0.7537
   -0.6845    0.3384    1.0000    0.8427    0.7678
   -0.6481    0.6009    0.8427    1.0000    0.9147
   -0.5305    0.7537    0.7678    0.9147    1.0000
```

Se observa que la correlación mas fuerte aparece entre volumen X_4 y la potencia del motor X_5 y es de 0.9147. La velocidad máxima X_2 y el peso X_3 tiene una baja relación lineal con valor de 0.3384. Nótese además que existe una relación negativa entre la primera variable (Km por galón) y el resto de las variables; esto nos sugiere realizar un estudio mas preciso con alguna técnica de visualización de datos multivariantes.

1.3.1. Variables aleatorias vectoriales

Una *variable aleatoria p-dimensional*, es un vector en el que cada una de sus componentes es una variable aleatoria. Así;

$$\mathbf{X} = (X_1, \ldots, X_p)' \tag{1.24}$$

es un vector aleatorio, con X_i variable aleatoria para cada $i = 1, \ldots, p$. De acuerdo con esta definición, los vectores aleatorios pueden estar conformados por variables aleatorias de tipo discreto, continuo ó ambos. Para simplificar aquí la exposición del tema, y salvo indicación en otro sentido, supondremos que la variable aleatoria es continua.

1.3.2. Distribuciones conjuntas

Similar al caso unidimensional, se define la *función de distribución conjunta* para el vector $\mathbf{X}$ mediante:

$$F(x_1, \ldots, x_p) = P(X_1 \leq x_1, \ldots, X_p \leq x_p) \ . \tag{1.25}$$

Si el vector aleatorio $\mathbf{X}$ es continuo y F es absolutamente continua, entonces la *función de densidad conjunta* es:

$$\frac{\partial^p F(x_1, x_2, \ldots, x_p)}{\partial x_1 \partial x_2 \ldots \partial x_p} = f(x_1, x_2, \ldots, x_p) \ . \tag{1.26}$$

Si las p-variables aleatorias que conforman el vector $\mathbf{X}$ son variables aleatorias independientes, entonces

$$F(x_1, x_2, \ldots, x_p) = F_1(x_1)F_2(x_2)\ldots F_p(x_p)$$

y

$$f(x_1, x_2, \ldots, x_p) = f_1(x_1)f_2(x_2), \ldots f_p(x_p)\ .$$

1.3.3. Distribuciones marginales

Dada una variable aleatoria p-dimensional $\mathbf{X}$ con función de distribución $F(x_1, x_2, \ldots, x_p)$, se define la *función de distribución marginal* para algún subconjunto de variables $X_1, \ldots, X_r$ con $(r \leq p)$ como :

$$\begin{aligned} P(X_1 \leq x_1, \ldots, X_p \leq x_p) &= P(X_1 \leq x_1, \ldots, X_r \leq x_r, X_{r+1} \leq \infty, \ldots, X_p \leq \infty) \\ &= F(x_1, \ldots, x_r, \infty, \ldots, \infty) \\ &= F(x_1, \ldots, x_r). \end{aligned} \tag{1.27}$$

La función de densidad marginal de $X_1, \ldots, X_p$ es

$$f(x_1, \ldots, x_r) = \int_{-\infty}^{\infty}\int_{-\infty}^{\infty}\cdots\int_{-\infty}^{\infty} f(x_1, \ldots, x_p)dx_{r+1}, \ldots, dx_p\ . \tag{1.28}$$

1.3.4. Distribuciones condicionales

Definimos la función de distribución condicional de un subconjunto de variables aleatorias $X_1, \ldots, X_r$ dadas las variables $X_{r+1} = x_{r+1}, \ldots, X_p = x_p$, como:

$$F(x_1, \ldots, x_r \mid x_{r+1}, \ldots, x_p) = \frac{F(x_1, \ldots, x_p)}{F(x_{r+1}, \ldots, x_p)}\ . \tag{1.29}$$

La función de densidad condicional está definida en forma semejante a como se muestra en seguida

$$f(x_1, \ldots, x_r \mid x_{r+1}, \ldots, x_p) = \frac{f(x_1, \ldots, x_p)}{f(x_{r+1}, \ldots, x_p)}\ . \tag{1.30}$$

1.3.5. Algunos parámetros asociados a las variables vectoriales

Dado un vector aleatorio $\mathbf{X}$, *el valor esperado* o *vector de medias*, notado por $\mathrm{E}(\mathbf{X})$, es el vector cuyas componentes son las esperanzas, o medias de las componentes de la variable

aleatoria, así;

$$\mu = \mathrm{E}(\mathbf{X}) = \begin{pmatrix} \mathrm{E}(X_1) \\ \vdots \\ \mathrm{E}(X_p) \end{pmatrix} = \begin{pmatrix} \mu_1 \\ \vdots \\ \mu_p \end{pmatrix} \tag{1.31}$$

Además definimos la matriz de *varianzas y covarianzas* de $\mathbf{X}$, la cual notaremos por $\mathbf{V}$, mediante:

$$\begin{aligned}
\mathbf{V} &= \mathrm{E}\{(\mathbf{X}-\mu)(\mathbf{X}-\mu)'\} \\
&= \mathrm{E}\left(\begin{bmatrix} X_1-\mu_1 \\ \vdots \\ X_p-\mu_p \end{bmatrix} [X_1-\mu_1 \ \ldots \ X_p-\mu_p] \right) \\
&= \mathrm{E} \begin{pmatrix} (X_1-\mu_1)^2 & (X_1-\mu_1)(X_2-\mu_2) & \ldots & (X_1-\mu_1)(X_p-\mu_p) \\ (X_1-\mu_1)(X_2-\mu_2) & (X_2-\mu_2)^2 & \ldots & (X_2-\mu_2)(X_p-\mu_p) \\ \vdots & \vdots & \ddots & \vdots \\ (X_p-\mu_p)(X_1-\mu_1) & (X_p-\mu_p)(X_2-\mu_2) & \ldots & (X_p-\mu_p)^2 \end{pmatrix} \\
&= \begin{pmatrix} \mathrm{E}\{(X_1-\mu_1)^2\} & \mathrm{E}\{(X_1-\mu_1)(X_2-\mu_2)\} & \ldots & \mathrm{E}\{(X_1-\mu_1)(X_p-\mu_p)\} \\ \mathrm{E}\{(X_1-\mu_1)(X_2-\mu_2)\} & \mathrm{E}\{(X_2-\mu_2)^2\} & \ldots & \mathrm{E}\{(X_2-\mu_2)(X_p-\mu_p)\} \\ \vdots & \vdots & \ddots & \vdots \\ \mathrm{E}\{(X_p-\mu_p)(X_1-\mu_1)\} & \mathrm{E}\{(X_p-\mu_p)(X_2-\mu_2)\} & \ldots & \mathrm{E}\{(X_p-\mu_p)^2\} \end{pmatrix} \\
&= \begin{pmatrix} \sigma_{11} & \ldots & \sigma_{1p} \\ \vdots & \ddots & \vdots \\ \sigma_{p1} & \ldots & \sigma_{pp} \end{pmatrix}
\end{aligned} \tag{1.32}$$

De estos dos parámetros podemos destacar las siguientes propiedades:

1. La matriz $\mathbf{V}$ es simétrica; es decir $\mathbf{V}' = \mathbf{V}$ puesto que $\sigma_{ij} = \sigma_{ji}$.

2. Los elementos de la diagonal de $\mathbf{V}$ corresponden a las varianzas de las respectivas variables ($\sigma_{ii} \equiv Var(X_i) \equiv \sigma_i^2$). Para $i \neq j$ decimos que σ_{ij} es la covarianza entre la variable X_i y la variableX_j.

3. Toda matriz de varianzas y covarianzas es *definida no negativa (semidefinida positiva)*. Es decir dado un vector cualquiera $\mathbf{y}$ de dimension $(p \times 1)$ se cumple que: $\mathbf{y}'\mathbf{V}\mathbf{y} \geq 0$.

4. Sean a, b constantes y X_i, X_j variables aleatorias para $i, j \in \{1, \ldots, p\}$, luego:

   - $\mathrm{E}(aX_i) = a\mathrm{E}(X_i) = a\mu_i$.
   - $Var(aX_i) = \mathrm{E}\{(aX_i - a\mu_i)^2\} = a^2 Var(X_i) = a^2\sigma_{ii} = a^2\sigma_i^2$.
   - $Cov(aX_i, bX_j) = \mathrm{E}\{(aX_i - a\mu_i)(bX_j - b\mu_j)\} = ab\mathrm{E}\{(X_i - \mu_j)(bX_i - b\mu_j)\}$. $= ab\ Cov(X_i, X_j) = ab\ \sigma_{ij}$.
   - $\mathrm{E}(aX_i + bX_j) = a\mathrm{E}(X_i) + b\mathrm{E}(X_j) = a\mu_i + b\mu_j$.
   - $Var(aX_i + bX_j) = a^2 Var(X_i) + b^2 Var(X_j) + 2abCov(X_i, X_j)$ $= a^2\sigma_{ii} + b^2\sigma_{jj} + 2ab\sigma_{ij}$.

5. Sea $\mathbf{c} = (c_1, \ldots, c_p)'$ un vector de constantes y $\mathbf{X} = (X_1, \ldots, X_p)'$ una variable aleatoria p-dimensional, luego la combinación lineal $\mathbf{c'X} = c_1X_1 + \cdots + c_pX_p$ tiene:

   - $MEDIA = \mathrm{E}(\mathbf{c'X}) = \mathbf{c'}\mu$.
   - $VARIANZA = Var(\mathbf{c'X}) = \mathbf{c'V1}$.

   donde $\mu = \mathrm{E}(\mathbf{X}), \mathbf{V} = Cov(\mathbf{X})$ y $\mathbf{1} = (1, \ldots, 1)'$.

6. Si $\mu = \mathrm{E}(\mathbf{X})$ y $\mathbf{V} = Cov(\mathbf{X})$ entonces:

$$\mathrm{E}(\mathbf{A'X} + \mathbf{b}) = \mathbf{A'}\mu + \mathbf{b} \quad y \quad Cov(\mathbf{A'X} + \mathbf{b}) = \mathbf{A'VA},$$

   con $\mathbf{A}$ matriz de constantes de tamaño $(p \times q)$ y $\mathbf{b}$ vector de constantes de tamaño $(q \times 1)$.

1.3.6. Esperanza y varianza condicionada

- La *esperanza* de un vector aleatorio $\mathbf{X}$ condicionado a otro vector $\mathbf{Y}$ se define como la esperanza de la distribución de $\mathbf{X}$ condicionada a un valor (y) de $\mathbf{Y}$, $(\mathbf{Y} = y)$, la cual viene dada por:

$$\mu_{\mathbf{X}|\mathbf{Y}} = \mathrm{E}(\mathbf{X}|\mathbf{Y}) = \int \mathbf{X} f(\mathbf{X}|\mathbf{Y}) d\mathbf{X}. \tag{1.33}$$

  En general, esta expresión será una función de $\mathbf{Y}$. Cuando $\mathbf{Y}$ es un valor fijo, $(\mathbf{Y} = y)$, la esperanza condicionada sera una constante. Si $\mathbf{Y}$ es una variable aleatoria, la esperanza condicionada sera también una variable aleatoria.
  La esperanza de una variable aleatoria puede obtenerse promediando, de forma ponderada, las esperanzas condicionadas por sus probabilidades de aparición, es decir, la

esperanza media condicional es la esperanza marginal o incondicional que podemos escribir:

$$\mathrm{E}(\mathbf{X}) = \mathrm{E}[\mathrm{E}(\mathbf{X}|\mathbf{Y})] \tag{1.34}$$

- La *varianza* de $\mathbf{X}$ condicionado a $\mathbf{Y}$ se define como la varianza de la distribución de $\mathbf{X}$ condicionado a un valor (y) de $\mathbf{Y}$, $(\mathbf{Y} = y)$, es decir

$$Var(\mathbf{X}|\mathbf{Y}) = \mathbf{V}_{\mathbf{X}|\mathbf{Y}} = \mathrm{E}\{[(\mathbf{X} - \mu_{\mathbf{X}|\mathbf{Y}})|\mathbf{Y}][(\mathbf{X} - \mu_{\mathbf{X}|\mathbf{Y}})|\mathbf{Y}]'\}, \tag{1.35}$$

luego esta matriz tendrá las propiedades ya estudiadas de la matriz de covarianzas. La varianza de una variable aleatoria puede descomponerse en dos fuentes de variación así:

$$Var(\mathbf{X}) = \mathrm{E}[Var(\mathbf{X}|\mathbf{Y})] + Var[\mathrm{E}(\mathbf{X}|\mathbf{Y})], \tag{1.36}$$

esta expresión se conoce *la descomposición de la varianza* y en la primera componente muestra el promedio de las varianzas de las distribuciones condicionadas, $Var(\mathbf{X}|\mathbf{Y})$, que en general son distintas, influyendo así en la variabilidad total. En el segundo componente hay también variabilidad ya que las medias de las distribuciones condicionadas pueden ser distintas y este segundo termino recoge la diferencia entre medias condicionadas, $\mathrm{E}(\mathbf{X}|\mathbf{Y})$, y la media total $\mu_{\mathbf{X}}$, mediante el termino $Var[\mathrm{E}(\mathbf{X}|\mathbf{Y})]$.
Si $\mathbf{X}, \mathbf{Y}$ son independientes, entonces todas las medias condicionadas serán igual a $\mu_{\mathbf{X}}$ y el termino $Var(\mathbf{X}|\mathbf{Y}) = 0$

1.4. La distribución normal multivariante

Generalizando la distribución normal escalar notada por $N(\mu, \sigma)$, una variable aleatoria vectorial $\mathbf{X}$ sigue una *distribución normal p-dimensional* si su función de densidad viene dada por

$$f(\mathbf{X}) = \frac{1}{(2\pi)^{\frac{p}{2}}|\mathbf{V}|^{\frac{1}{2}}} \exp\left\{-\frac{1}{2}(\mathbf{X} - \mu)'\mathbf{V}^{-1}(\mathbf{X} - \mu)\right\}, \tag{1.37}$$

donde $\mu = (\mu_1, \ldots, \mu_p)'$, $\mathbf{V}$ es una matriz simétrica definida positiva de tamaño $(p \times p)$ denominada la matriz de covarianzas y $|\mathbf{V}|$ denota el determinante de $\mathbf{V}$.
La figura 1.1 muestra la distribución normal bivariate con vector de medias $\mu = \mathbf{0}$ y matriz de covarianzas $\begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix}$.

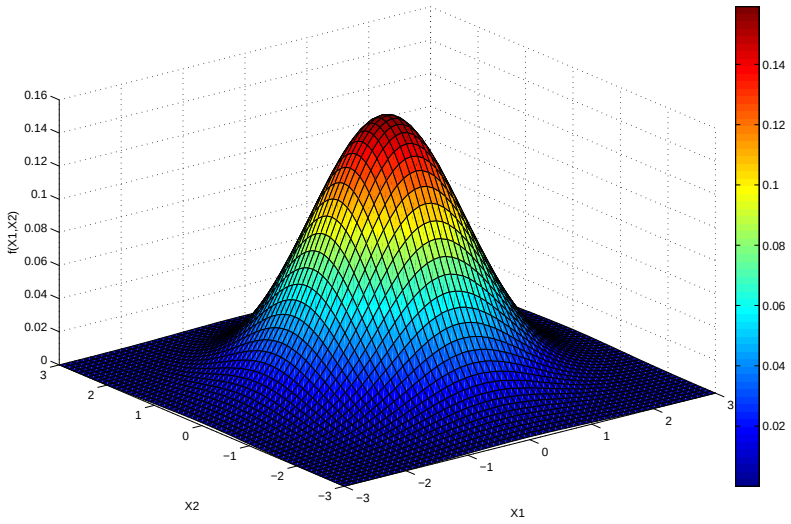


Figura 1.1: Distribución normal bivariate.

De igual forma, decimos que un vector p-dimensional $\mathbf{X}$ tiene una *distribución normal p-variante* con un vector de medias μ y una matriz de covarianzas $\mathbf{V}$, si y solo si, *la función generadora de momentos*[1] es

$$\mathbf{M_X}(\mathbf{t}) = \exp\left\{\mu'\mathbf{t} + \frac{\mathbf{t'Vt}}{2}\right\} \quad con \quad \mathbf{t}' = (t_1, \ldots, t_p)'. \tag{1.38}$$

Las principales propiedades de la distribución normal multivariante son:

1. **Simetría:** La distribución normal p-variante es simétrica alrededor de μ.
2. **Punto máximo:** La distribución normal p-variante tiene un único máximo en μ.
3. **Determinación:** Si $\mathbf{X}_{p\times 1}$ es un vector aleatorio con distribución normal multivariante, su media es μ y su matriz de varianzas y covarianzas es $\mathbf{V}$, lo cual notaremos $\mathbf{X} \sim N_p(\mu, \mathbf{V})$.
4. **Linealidad:** Si $\mathbf{X}$ es un vector aleatorio p-dimensional distribuido normalmente, con vector de medias μ y matriz de covarianzas $\mathbf{V}$, entonces el vector $\mathbf{Y} = \mathbf{AX} + \mathbf{b}$ con

[1] La función generadora de momentos (fgm) puede consultarse en el apéndice B de [2].

$\mathbf{A}_{q\times p}$ y con $\mathbf{b}_{q\times 1}$ tiene distribución normal q-variante con vector de media $\mathbf{A}\mu+\mathbf{b}$ y matriz de varianzas y covarianzas $\mathbf{AVA'}$, es decir, si

$$\mathbf{X}\sim N_p(\mu,\mathbf{V})\ entonces\ \mathbf{Y}=(\mathbf{AX}+b)\sim N_q(\mathbf{A}\mu+\mathbf{b},\mathbf{AVA'}) \tag{1.39}$$

5. **Marginales:** Considérese el vector $\mathbf{X}$ particionado como $\mathbf{X}=(X_{(1)},X_{(2)})$ con $X_{(1)}=(X_1,\ldots,X_{p_1})$ y $X_{(2)}=(X_{p_1+1},\ldots,X_p)$ y además $\mathbf{V}$ particionada:

$$\mathbf{V}=\begin{pmatrix}V_{11} & V_{12}\\ V_{21} & V_{22}\end{pmatrix}, \tag{1.40}$$

donde V_{11} es la submatriz superior izquierda de $\mathbf{V}$ de tamaño $(p_1\times p_1)$.
Si $\mathbf{X}$ tiene una distribución normal con media μ, matriz de varianzas y covarianzas $\mathbf{V}$ (definida positiva) y $V_{12}=V_{21}'=0$ entonces los vectores $X_{(1)}$ y $X_{(2)}$ son independientes y normalmente distribuidos con vectores de medias $\mu_{(1)}$, $\mu_{(2)}$ y matrices de varianzas y covarianzas V_{11} y V_{22}.

6. **Independencia:** La matriz de varianzas y covarianzas de un vector $\mathbf{X}_{p\times 1}$ con distribución normal p-variante es diagonal si y solo si las componentes de $\mathbf{X}$ son variables aleatorias normales e independientes.

7. **Estandarización:** Sea $\mathbf{X}$ un vector aleatorio p-dimensional distribuido normalmente con vector de medias μ y matriz de varianzas y covarianzas $\mathbf{V}$. Si $\mathbf{V}$ es una matriz no singular entonces:

$$\mathbf{Z}=\mathbf{V}^{-1/2}(\mathbf{X}-\mu), \tag{1.41}$$

tiene distribución normal p-variante con vector de medias cero y matriz de varianzas y covarianzas I_p, donde $\mathbf{V}^{-1/2}=(\mathbf{V}^{-1})^{1/2}$. [2] Es decir si

$$\mathbf{X}\sim N_p(\mu,\mathbf{V})\ entonces\ \mathbf{Z}=\mathbf{V}^{-1/2}(\mathbf{X}-\mu)\sim N_p(0,I) \tag{1.42}$$

8. **Distribución condicional:** Consideremos la misma partición del vector aleatorio $\mathbf{X}$ dada en el numeral 5 con $X_{(1)}$ y $X_{(2)}$ de tamaños $(P_1\times 1)$ y $(P_2\times 1)$ respectivamente.

[2] En semejanza con los numeros reales, para las matrices no definidas positivas existe una única matriz que corresponde a su *raíz cuadrada*; es decir que para la matriz $\mathbf{V}\geq 0$ existe una única matriz $B\geq 0$ tal que: $B^2=\mathbf{V}$. Esta es $B=\mathbf{V}^{-1/2}$. Además,

$$(\mathbf{V}^{-1})^{1/2}=(\mathbf{V}^{1/2})^{-1}=\mathbf{V}^{-1/2}.$$

La función de densidad condicional de $X_{(1)}$ dado $X_{(2)} = x_{(2)}$ es :

$$g(x_{(1)}|x_{(2)}) = \frac{f(x_{(1)}, x_{(2)})}{h(x_{(2)})}, \tag{1.43}$$

donde h es la función de densidad marginal para $x_{(2)}$, es decir

$$h(x_{(2)}) = \frac{1}{(2\pi)^{P_2/2}|V_{22}|^{1/2}} \exp\left\{-\frac{1}{2}(x_{(2)} - \mu_{(2)})' V_{22}^{-1}(x_{(2)} - \mu_{(2)})\right\} \tag{1.44}$$

y la función $f(x_{(1)}, x_{(2)})$ es normal multivariante.
Reemplazando por las funciones indicadas y después de hacer las operaciones y simplificaciones pertinentes, se llega al siguiente resultado

$$\begin{aligned} g(x_{(1)}|x_{(2)}) = & \frac{1}{(2\pi)^{P_2/2}|V_{11} - V_{12}V_{21}^{-1}V_{12}'|^{1/2}} \\ & \exp\left\{-\frac{1}{2}[\ x_{(1)} - (\mu_{(1)} + V_{12}V_{12}^{-1}(x_{(2)} - \mu_{(2)}))]'(V_{11} - V_{12}V_{21}^{-1}V_{12}')^{-1}\right. \\ & \left.[\ x_{(1)} - (\mu_{(1)} + V_{12}V_{12}^{-1}(x_{(2)} - \mu_{(2)}))]\right\}. \end{aligned} \tag{1.45}$$

La función $g(x_{(1)}|x_{(2)})$ es la función de densidad normal p_1-variante con vector de medias

$$\mu_{X_{(1)}|X_{(2)}} = \mu_{(1)} + V_{12}V_{22}^{-1}(x_{(2)} - \mu_{(2)}). \tag{1.46}$$

y matriz de varianzas y covarianzas

$$\mathbf{V}_{X_{(1)}|X_{(2)}} = V_{11\cdot 2} = V_{11} - V_{12}V_{22}^{-1}V_{12}'. \tag{1.47}$$

La matriz $\beta \equiv V_{12}V_{22}^{-1}$ es la matriz de los coeficientes de *regresión* de $X_{(1)}$ sobre $X_{(2)}$. El vector $\mu_{X_{(1)}|X_{(2)}} = V_{11} - \beta(x_{(2)} - \mu_{(2)})$ se llama con frecuencia *función de regresión* de $X_{(1)}$ sobre $X_{(2)}$.

9. **Vector de residuos:** Los subvectores $X_{(2)}$ y $X_{(1)}^* = X_{(1)} - V_{12}V_{22}^{-1}(X_{(2)} - \mu_{(2)})$ son independientes y normalmente distribuidos con medias

$$\mu_{(2)} \quad y \quad \mu_{(1)}^* = \mu_{(1)} - V_{12}V_{22}^{-1}\mu_{(2)},$$

y matices de varianzas y covarianzas (definidas positivas)

$$V_{11} \quad y \quad V_{11\cdot 2} = V_{11} - V_{12}V_{22}^{-1}V_{21},$$

respectivamente.
La independencia entre los subvectores $X_{(2)}$ y $X^*_{(1)}$ se garantiza demostrando que:

$$V^*_{21} = \mathrm{E}[(x_{(2)} - \mu_{(2)})x^*_{(1)}] = 0.$$

Para terminar con el paralelo con la regresión lineal, el vector

$$\mathbf{E}_{(1\cdot 2)} = X_{(1)} - \mu_{X_{(1)}|X_{(2)}} = X_{(1)} - [\mu_{(1)} + V_{12}V_{22}^{-1}(x_{(2)} - \mu_{(2)})], \tag{1.48}$$

es el **vector de residuales** entre $X_{(1)}$ y los valores predichos por la regresión de $X_{(1)}$ sobre $X_{(2)}$. De lo anterior se establece que bajo el criterio de normalidad, los residuales y las variables regresoras (fijas) son independientes.

10. **Combinación lineal de multinormales:** Sean $X_1, \ldots, X_n$ vectores aleatorios independientes de tamaño $(P \times 1)$ con distribución $\mathbf{X} \sim N_p(\mu, \mathbf{V})$. Entonces, la distribución lineal $L_1 = c_1X_1 + \ldots + c_nX_n$ se distribuye $(\sum_{i=1}^n c_i\mu_i, (\sum_{i=1}^n c_i^2)\mathbf{V})$.Ademas, L_1 y $L_2 = d_1X_1 + \ldots + d_nX_n$ tiene distribución normal conjunta, con vector de medias

$$\begin{pmatrix} \sum_{i=1}^n c_i\mu_i \\ \sum_{i=1}^n d_i\mu_i \end{pmatrix},$$

y matriz de covarianzas

$$\begin{pmatrix} \sum_{i=1}^n c_i^2\mathbf{V} & (d'c)\mathbf{V} \\ (d'c)\mathbf{V} & \sum_{i=1}^n d_i^2\mathbf{V} \end{pmatrix}.$$

Las dos combinaciones lineales L_1 y L_2 son independientes si $(d'c) = \sum_{i=1}^n c_id_i = 0$.

Ejemplo 1.6. En el Toolbox Statistic de MATLAB 7.1 encontramos los comandos *mvnrnd* y *mvnpdf*; este primero nos permite generar datos aleatorios para simular una distribución normal multivariate dado el vector de medias μ y la matriz de covarianzas $\mathbf{V}$. El segundo comando nos permite encontrar los valores de la densidad normal multivariante (fdp) para los datos suministrados, el vector de medias μ y la matriz de covarianzas $\mathbf{V}$. Veamos para el caso de la estandarización de la distribución normal bivariante, como con la ayuda de estos comandos podemos esquematizar la (fdp) y la proyección de los datos sobre el plano. En el *Command Window* de MATLAB 7.1 escribimos:

```
>> Tra_mu = [0 0];%..............El vector de medias transpuesto.
>> mu = (Tra_mu)';%..............El vector de medias.
>> sigma = [1 0;0 1];%...........La matriz de covarianzas.
```

```
>> r = mvnrnd(Tra_mu,sigma,1000);%Genera 1000 datos aleatorios.
>> z = mvnpdf(r,Tra_mu,sigma);%...Valores de la (fdp) para los datos.
>> x = r(:,1);%...................Guarda la primera columna de r en x.
>> y = r(:,2);%...................Guarda la segunda columna de r en y.
>> plot3(x,y,z,'.')%..............Gráfica en 3D para los datos.
>> grid on %......................Grilla para la gráfica.
>> plot(x,y,'+') % Proyección de datos sobre el plano XY.
>> grid on
```

Estas gráficas las podemos apreciar en las figuras 1.2 y 1.3, respectivamente.

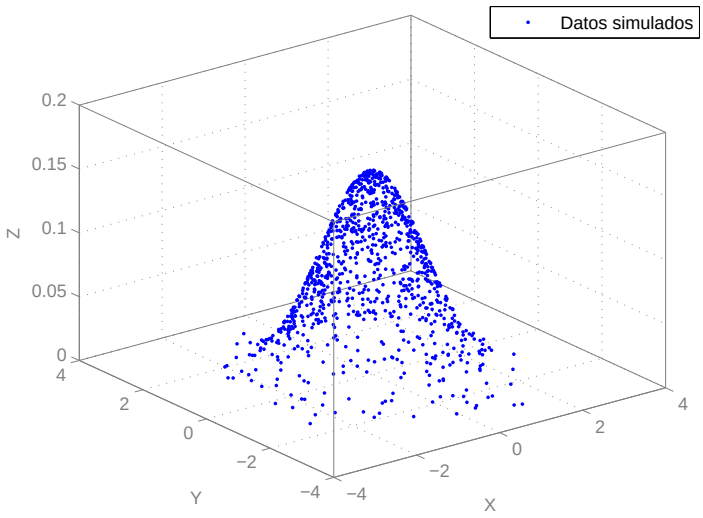


Figura 1.2: Simulación de la distribución normal bivariante estandarizada.

1.4.1. Una visión geométrica de la distribución normal multivariante.

El exponente $(\mathbf{X}-\mu)'\mathbf{V}^{-1}(\mathbf{X}-\mu)$ de la función de densidad normal multivariate dada por la ecuación (1.37), corresponde a la ecuación de un elipsoide en el espacio $p-$*dimensinal*

cuando este es igual a una constante $\mathcal{C}$; este elipsoide se obtiene al cortar con un hiperplano, paralelo al definido por las p variables que forman la variable vectorial $\mathbf{X}$, la distribución normal p variante.

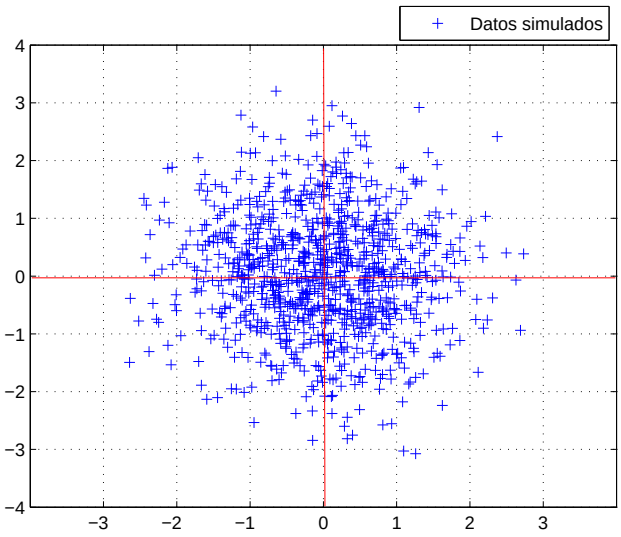


Figura 1.3: Proyección sobre el plano **XY** de los datos simulados en el **Ejemplo 1.6**.

Luego la ecuación

$$(\mathbf{X}-\mu)'\mathbf{V}^{-1}(\mathbf{X}-\mu)=\mathcal{C} \tag{1.49}$$

define las curvas de nivel (elipsoides) y estas a su vez una medida de la distancia de un punto al centro de la distribución. Esta medida se denomina *distancia de Mahalanobis*[3] y la representamos por:

$$D_i^2=(\mathbf{x_i}-\mu)'\mathbf{V}^{-1}(\mathbf{x_i}-\mu), \tag{1.50}$$

[3]La distancia de Mahalanobis se distribuye como una χ^2 con p grados de libertad; esto se demuestra realizando la transformación lineal o *estandarización* del vector $\mathbf{X}$ planteada en (1.41) y tomando $\mathbf{V}^{-1}=\mathbf{V}^{-1/2}\mathbf{V}^{-1/2}$ con lo que se obtiene que

$$D^2=\mathbf{Z}'\mathbf{Z}=\sum Z_i^2$$

donde cada Z_i es $N(0,1)$. Por lo tanto $D^2\sim\chi^2$.

donde $\mathbf{x_i} = (x_{i1}, \ldots, x_{ip})'$ representa un individuo particular, seleccionado aleatoriamente de una población con centro en $\mu = (\mu_1, \ldots, \mu_p)'$ y $\mathbf{V}$ como matriz de covarianzas. La familia de elipsoides concéntricos, generados al variar $\mathcal{C}$, tienen su centro común en μ. El eje principal de cada elipsoide esta en la linea que pasa a través de los puntos mas distantes de la elipse; es decir, es el segmento principal de la elipse (o diámetro) el cual pasa por μ y tiene sus extremos en la superficie de un elipsoide, éste tiene las coordenadas que maximizan el cuadrado de la mitad de su longitud.
Así,

$$\|(\mathbf{x} - \mu)\|^2 = (\mathbf{x} - \mu)'(\mathbf{x} - \mu) \tag{1.51}$$

es la distancia entre $\mathbf{x}$ y μ, que debe maximizarse bajo la restricción:

$$\mathcal{C} = (\mathbf{x} - \mu)'\mathbf{V}^{-1}(\mathbf{x} - \mu) \tag{1.52}$$

la cual dice que el punto $\mathbf{x}$ pertenece al elipsoide. En la figura 1.4 se muestra las curvas de nivel para una distribución normal bivariate donde las dos variables aleatorias X_1 y X_2 son independientes.

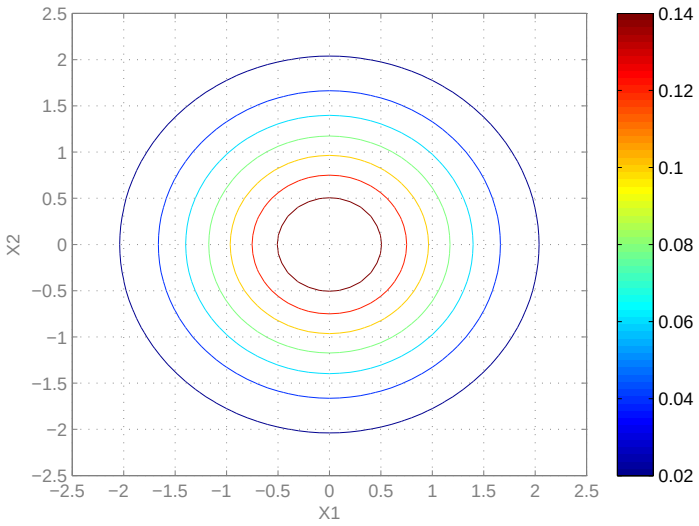


Figura 1.4: Curvas de nivel para una normal bivariate con media $\mu = \mathbf{0}$ y matriz de covarianzas $\mathbf{V} = \mathbf{I}_{(2\times2)}$.

Capítulo 2

INFERENCIA MULTIVARIANTE

2.1. Introducción

Tradicionalmente, los problemas de inferencia estadística se dividen en *problemas de estimación* y problemas de *contraste de hipótesis*, aunque en realidad todo esta englobado en los llamados *problemas de decisión*. La diferencia principal entre las dos clases de problemas es que en los de estimación debemos determinar el valor de un vector de parámetros θ en un conjunto $\Omega \subset \mathbb{R}^n$ de posibles alternativas, mientras que en los de contrastes debemos decidir si aceptamos o rechazamos un vector de valores especifico del parámetro llamado $\widehat{\theta}$.

En este capitulo se presentan algunos tópicos respecto a estimación, distribución, propiedades y contrastes de hipótesis de los parámetros y sus estimadores. Adicionalmente en la sección 2.4 estudiaremos la estimación y los contrastes Bayesianos que nos complementará el estudio de la inferencia multivariante. Finalizamos con la presentación de dos medidas de contraste usadas en la selección del modelo que mejor se ajusta a una muestra poblacional. Estas medidas son los criterios **AIC** y **BIC**.

Suponemos que el lector ya esta familiarizado con los conceptos básicos de inferencia al nivel univariado, pues esto facilitará la plena comprensión del capitulo. Al finalizar la inferencia multivariante estaremos en la capacidad de estudiar las mezclas finitas de las distribuciones normales.

2.2. Estimación multivariante: el método de la maxima verosimilitud

Ronald Fisher (1890-1962), matemático británico cuyas teorías estadísticas hicieron mucho más precisos los experimentos científicos, publicó a principios del siglo $\overline{\text{XX}}$ dos artículos en estadística donde propuso un método general de estimación llamado **el método de la maxima verosimilitud**, el cual toma como estimador de los parámetros aquel valor que maximiza la probabilidad de que un modelo estadístico genere los datos muestrales.
Así la característica esencial del método de maxima verosimilitud es examinar los valores de la muestra y escoger como estimado de los parámetros desconocidos ha aquellos para los cuales es maxima la probabilidad (densidad) de los datos muestrales.
Supongamos que se dispone de una muestra aleatoria simple de n elementos de una variable aleatoria p -dimensional $\mathbf{X}$ con función de densidad $f(\mathbf{X}|\theta)$, donde $\theta = (\theta_1, \ldots, \theta_r)'$ es un vector de parámetros que supondremos tiene dimensión $r < np$. Si llamamos
$\mathbb{X} = (\mathbf{x}_1, \ldots, \mathbf{x}_n)'$ a los datos muestrales donde $\mathbf{x}_i = (x_{i1}, \ldots, x_{ip})'$ representa un individuo particular, entonces **la función de densidad conjunta** de la muestra será:

$$f(\mathbb{X}|\theta) = \prod_{i=1}^{n} f(\mathbf{x}_i|\theta), \tag{2.1}$$

debido a la independencia de las observaciones. Además si el parámetro θ es conocido, la función $f(\mathbb{X}|\theta)$ determina la probabilidad de aparición de la muestra.
Por lo general en los problemas de estimación se disponen de muestras con parámetro θ desconocido, quedando así nuestro modelo estadístico definido en (2.1) incompleto; en estos casos es de utilidad el concepto de **modelo saturado**. Un modelo se denomina saturado cuando utiliza tantos parámetros como observaciones hemos efectuado y por tanto se ajusta perfectamente a los datos. Es así como debemos considerar en la expresión (2.1) a θ como una variable y particularizar esta función para los datos observados, obteniendo así una función que llamaremos *función de verosimilitud* la cual notaremos $l(\theta|\mathbb{X})$, o $l(\theta)$.

Definición 2.1. Sea $\mathbb{X} = (\mathbf{x}_1, \ldots, \mathbf{x}_n)'$ los datos muestrales de una población con vector de parámetros $\theta = (\theta_1, \ldots, \theta_r)'$, con $r < np$. **La función de verosimilitud** esta definida por:

$$l(\theta|\mathbb{X}) \equiv l(\theta) = \prod_{i=1}^{n} f(\mathbf{x}_i|\theta), \tag{2.2}$$

donde $\mathbb{X}$ es fijo y θ variable.

El estimador de maxima verosimilitud $\widehat{\theta}$ de θ, o estimador MV, es el valor de θ que maximiza la probabilidad de aparición de los valores muestrales observados y que obtenemos al calcular el valor máximo de la función $l(\theta)$. Asumamos que $l(\theta)$ es diferenciable, entonces:

$$\begin{aligned}\frac{\partial l(\theta)}{\partial \theta_1} &= 0\\ &\vdots\\ \frac{\partial l(\theta)}{\partial \theta_r} &= 0.\end{aligned}$$

En la practica suele ser mas cómodo obtener el máximo del logaritmo de la función de verosimilitud:

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

que llamaremos *función soporte*. Como el logaritmo es una función monótona, ambas funciones tienen el mismo máximo, pero trabajar con la función soporte es mas sencillo y hace mas cómodo el obtener el máximo. Formalmente el soporte es:

Definición 2.2. Sea $l(\theta)$ la función de verosimilitud de los datos muestrales $\mathbb{X} = (\mathbf{x}_1, \ldots, \mathbf{x}_n)'$, con $\theta = (\theta_1, \ldots, \theta_r)'$. Definimos **la función soporte** como:

$$L(\theta) = \ln l(\theta) = \ln\left[\prod_{i=1}^{n} f(\mathbf{x}_i|\theta)\right] = \sum_{i=1}^{n} \ln\big[f(\mathbf{x}_i|\theta)\big]. \tag{2.3}$$

Definición 2.3. Sea $f(\mathbf{X}|\theta)$ un modelo estadístico con $\theta = (\theta_1, \ldots, \theta_r)'$. **La función** "**score**"se define como:

$$Z(\theta) = \frac{\partial}{\partial \theta} \ln f(\mathbf{X}|\theta) = \begin{bmatrix} \frac{\partial \ln f}{\partial \theta_1} \\ \vdots \\ \frac{\partial \ln f}{\partial \theta_r} \end{bmatrix} = \left[\frac{\partial \ln f(\mathbf{X}|\theta)}{\partial \theta_i}\right]. \tag{2.4}$$

Para la función de densidad conjunta $f(\mathbb{X}|\theta)$ de una muestra, dada en (2.1), tenemos que la función "**score**"se puede expresar de la siguiente forma:

$$\begin{aligned} Z(\theta) &= \frac{\partial}{\partial \theta} \ln f(\mathbb{X}|\theta)\\ Z(\theta) &= \frac{\partial}{\partial \theta} \ln \left\{\prod_{i=1}^{n} f(\mathbf{x}_i|\theta)\right\} \\ Z(\theta) &= \frac{\partial}{\partial \theta} \ln l(\theta) \end{aligned} \tag{2.5}$$

$$Z(\theta) = \frac{\partial}{\partial \theta} L(\theta) \tag{2.6}$$

Luego para obtener el estimador máximo verosímil $\widehat{\theta}$ de θ maximizamos la función de soporte $L(\theta)$, que equivale (bajo condiciones de regularidad) a igualar la función "**score**"para una muestra de n elementos a cero, con lo que tendremos un sistema de ecuaciones homogéneo. Simplificando la expresión dada en (2.5), podemos obtener $\widehat{\theta}$ resolviendo el sistema de ecuaciones:

$$z(\theta) = \frac{\partial}{\partial \theta} L(\theta) = \sum_{i=1}^{n} \frac{\partial}{\partial \theta} \ln f(\mathbf{x}_i|\theta) = 0 \tag{2.7}$$

Además el doble de la función soporte cambiada de signo proporciona un método general para juzgar el ajuste de un modelo a los datos, esta nueva función se denomina *función desviación* y la definimos así:

Definición 2.4. Sea $L(\theta)$ la función de soporte de los datos muestrales $\mathbb{X} = (\mathbf{x}_1, \ldots, \mathbf{x}_n)'$, con $\theta = (\theta_1, \ldots, \theta_r)'$. Definimos **la función de desviación** como:

$$\mathcal{D}(\theta) = -2L(\theta), \tag{2.8}$$

donde $\mathcal{D}(\theta)$ mide la discrepancia entre el modelo y los datos.

Cuanto mayor sea el soporte, $L(\theta)$, mayor es la concordancia entre el valor del parámetro y los datos, y menor es la desviación $\mathcal{D}(\theta)$. Luego maximizar la verosimilitud equivale a minimizar la desviación.
Para distribuciones cuyo rango de posibles valores es conocido *a priori* y para las cuales no hay una ninguna dependencia de sus parámetros, bajo condiciones muy generales respecto al modelo de distribución de probabilidad, se tiene que el método de maxima verosimilitud (MV) proporcionan estimadores que son:

- Asintóticamente centrados.
- Con distribución asintóticamente normal.
- Asintóticamente de varianza minima (eficientes).
- Si existe un estadístico suficiente para el parámetro, el estimador MV es suficiente.
- *Invariantes* en el siguiente sentido: Si $\widehat{\theta}$ es el estimador MV de θ y $g(\theta)$ es una función cualquiera del vector de parámetros, entonces en condiciones bastantes generales , $\widehat{g(\theta)}$ es el estimador MV de $g(\theta)$.

Las demostraciones de estas propiedades pueden consultarse en [12].

2.2.1. Estimación de parámetros en la distribución normal multivariante

Partiendo de una muestra aleatoria de una población normal p-variante se obtiene los estimadores de μ y $\mathbf{V}$ por el método de *máxima verosimilitud (MV)*. Es decir se buscan los valores de μ y de $\mathbf{V}$ que maximizan la probabilidad de que la muestra aleatoria proceda de esta población.
Sea $\mathbb{X} = (\mathbf{x}_1, \ldots, \mathbf{x}_n)'$ una muestra aleatoria simple donde $\mathbf{x}_i \sim N_p(\mu, \mathbf{V})$. La *función de verosimilitud* es :

$$l(\mu, \mathbf{V}|\mathbb{X}) = \prod_{i=1}^{n} |\mathbf{V}|^{-1/2}(2\pi)^{-p/2} \exp\left\{-\frac{1}{2}(\mathbf{x}_i - \mu)'\mathbf{V}^{-1}(\mathbf{x}_i - \mu)\right\}$$

$$l(\mu, \mathbf{V}) = |\mathbf{V}|^{-n/2}(2\pi)^{-np/2} \exp\left\{-\frac{1}{2}\sum_{i=1}^{n}(\mathbf{x}_i - \mu)'\mathbf{V}^{-1}(\mathbf{x}_i - \mu)\right\} \tag{2.9}$$

y la *función de soporte* será:

$$L(\mu, \mathbf{V}|\mathbb{X}) = \ln l(\mu, \mathbf{V})$$

$$L(\mu, \mathbf{V}) = -\frac{1}{2}pn\ln(2\pi) - \frac{1}{2}n\ln|\mathbf{V}| - \frac{1}{2}\sum_{i=1}^{n}(\mathbf{x}_i - \mu)'\mathbf{V}^{-1}(\mathbf{x}_i - \mu). \tag{2.10}$$

Esta ultima función nos indica el apoyo o el soporte que reciben los posibles valores de los parámetros dados los datos. Cuanto mayor sea esta función para unos valores de los parámetros, mayor sera la concordancia entre estos parámetros y los datos.
Expresemos esta función en una forma mas conveniente. Tomando el vector de medias muestral definido en la ecuación (1.19), el cual esta notado por

$$\overline{\mathbf{x}} = \begin{pmatrix} \overline{\mathbf{x}}_\mathbf{1} \\ \vdots \\ \overline{\mathbf{x}}_\mathbf{p} \end{pmatrix} \quad donde \quad \overline{\mathbf{x}}_\mathbf{j} = \frac{1}{n}\sum_{i=1}^{n} x_{ij}, \quad con \quad j = 1, \ldots, p,$$

podemos desarrollar la forma cuadrática contenida en la ecuación (2.10) al introducir $(-\overline{\mathbf{x}} + \overline{\mathbf{x}})$, obteniendo así:

$$\begin{aligned}
&(\mathbf{x}_i - \mu)'\mathbf{V}^{-1}(\mathbf{x}_i - \mu) = \\
&= [(\mathbf{x}_i - \overline{\mathbf{x}}) + (\overline{\mathbf{x}} - \mu)]'\mathbf{V}^{-1}[(\mathbf{x}_i - \overline{\mathbf{x}}) + (\overline{\mathbf{x}} - \mu)] \\
&= (\mathbf{x}_i - \overline{\mathbf{x}})'\mathbf{V}^{-1}[(\mathbf{x}_i - \overline{\mathbf{x}}) + (\overline{\mathbf{x}} - \mu)] + (\overline{\mathbf{x}} - \mu)'\mathbf{V}^{-1}[(\mathbf{x}_i - \overline{\mathbf{x}}) + (\overline{\mathbf{x}} - \mu)]
\end{aligned}$$

$$= (\mathbf{x}_i - \overline{\mathbf{x}})'\mathbf{V}^{-1}(\mathbf{x}_i - \overline{\mathbf{x}}) + \underbrace{(\mathbf{x}_i - \overline{\mathbf{x}})'\mathbf{V}^{-1}(\overline{\mathbf{x}} - \mu)}_{(a)} + \underbrace{(\overline{\mathbf{x}} - \mu)'\mathbf{V}^{-1}(\mathbf{x}_i - \overline{\mathbf{x}})}_{(b)} + (\overline{\mathbf{x}} - \mu)'\mathbf{V}^{-1}(\overline{\mathbf{x}} - \mu)$$

Puesto que (a) y (b) son iguales [1] se concluye que:

$$(\mathbf{x}_i - \mu)'\mathbf{V}^{-1}(\overline{\mathbf{x}} - \mu) = (\mathbf{x}_i - \overline{\mathbf{x}})'\mathbf{V}^{-1}(\mathbf{x}_i - \overline{\mathbf{x}}) + (\overline{\mathbf{x}} - \mu)'\mathbf{V}^{-1}(\overline{\mathbf{x}} - \mu) + 2(\overline{\mathbf{x}} - \mu)'\mathbf{V}^{-1}(\mathbf{x}_i - \overline{\mathbf{x}}). \tag{2.11}$$

Ahora si sumamos sobre el subíndice i, el ultimo termino de la identidad anterior se anula[2], de donde resulta la expresión

$$\sum_{i=1}^{n}(\mathbf{x}_i - \mu)'\mathbf{V}^{-1}(\mathbf{x}_i - \mu) = \sum_{i=1}^{n}\underbrace{(\mathbf{x}_i - \overline{\mathbf{x}})'\mathbf{V}^{-1}(\mathbf{x}_i - \overline{\mathbf{x}})}_{(c)} + n(\overline{\mathbf{x}} - \mu)'\mathbf{V}^{-1}(\overline{\mathbf{x}} - \mu).$$

La forma cuadrática dada en (c) es un escalar, por lo tanto sera igual a su traza; además por conceptos básicos de matrices se sabe que $tra(AB) = tra(BA)$, luego se tiene que:

$$(\mathbf{x}_i - \overline{\mathbf{x}})'\mathbf{V}^{-1}(\mathbf{x}_i - \overline{\mathbf{x}}) = tra\left[\mathbf{V}^{-1}(\mathbf{x}_i - \overline{\mathbf{x}})(\mathbf{x}_i - \overline{\mathbf{x}})'\right], \tag{2.12}$$

en consecuencia decimos

$$\begin{aligned}\sum_{i=1}^{n}(\mathbf{x}_i - \mu)'\mathbf{V}^{-1}(\mathbf{x}_i - \mu) &= \sum_{i=1}^{n} tra\left[\mathbf{V}^{-1}(\mathbf{x}_i - \overline{\mathbf{x}})(\mathbf{x}_i - \overline{\mathbf{x}})'\right] + n(\overline{\mathbf{x}} - \mu)'\mathbf{V}^{-1}(\overline{\mathbf{x}} - \mu) \\ &= tra\left[\sum_{i=1}^{n}\mathbf{V}^{-1}(\mathbf{x}_i - \overline{\mathbf{x}})(\mathbf{x}_i - \overline{\mathbf{x}})'\right] + tra\left[n(\overline{\mathbf{x}} - \mu)'\mathbf{V}^{-1}(\overline{\mathbf{x}} - \mu)\right]\end{aligned}$$

[1] Esto es fácil de demostrar ya que tanto (a) como (b) son matrices de dimensiones (1×1), luego serán iguales a sus traspuestas. Además la matriz $\mathbf{V}^{-1}$ es simétrica, siendo así $\mathbf{V}^{-1}$ igual a su traspuesta. Entonces:

$$\begin{aligned}(\mathbf{x}_i - \overline{\mathbf{x}})'\mathbf{V}^{-1}(\overline{\mathbf{x}} - \mu) &= \{[(\mathbf{x}_i - \overline{\mathbf{x}})'\mathbf{V}^{-1}](\overline{\mathbf{x}} - \mu)\}' = \{[\mathbf{V}^{-1}(\mathbf{x}_i - \overline{\mathbf{x}})]'(\overline{\mathbf{x}} - \mu)\}' \\ &= (\overline{\mathbf{x}} - \mu)'\{[\mathbf{V}^{-1}(\mathbf{x}_i - \overline{\mathbf{x}})]'\}' = (\overline{\mathbf{x}} - \mu)'[\mathbf{V}^{-1}(\mathbf{x}_i - \overline{\mathbf{x}})] = (\overline{\mathbf{x}} - \mu)'\mathbf{V}^{-1}(\mathbf{x}_i - \overline{\mathbf{x}}). \blacksquare\end{aligned}$$

[2] Demostremos que $\sum_{i=1}^{n}(\overline{\mathbf{x}} - \mu)'\mathbf{V}^{-1}(\mathbf{x}_i - \overline{\mathbf{x}}) = 0$. Llamando $\mathbf{a} = (\overline{\mathbf{x}} - \mu)'\mathbf{V}^{-1}$, que es un vector de dimension $(1 \times p)$ con valores constantes e independientes de la i-esima observación y recordando que $\overline{\mathbf{x}} = \frac{1}{n}\sum_{i=1}^{n}\mathbf{x}_i$, se tiene:

$$\sum_{i=1}^{n}(\overline{\mathbf{x}} - \mu)'\mathbf{V}^{-1}(\mathbf{x}_i - \overline{\mathbf{x}}) = \sum_{i=1}^{n}\mathbf{a}(\mathbf{x}_i - \overline{\mathbf{x}}) = \sum_{i=1}^{n}\mathbf{a}\mathbf{x}_i - \sum_{i=1}^{n}\mathbf{a}\overline{\mathbf{x}} = \mathbf{a}\sum_{i=1}^{n}\mathbf{x}_i - n(\mathbf{a}\overline{\mathbf{x}}) = n(\mathbf{a}\overline{\mathbf{x}}) - n(\mathbf{a}\overline{\mathbf{x}}) = 0. \blacksquare$$

$$= tra\left[\mathbf{V}^{-1}\sum_{i=1}^{n}(\mathbf{x}_i-\overline{\mathbf{x}})(\mathbf{x}_i-\overline{\mathbf{x}})'\right] + tra\left[n(\overline{\mathbf{x}}-\mu)'\mathbf{V}^{-1}(\overline{\mathbf{x}}-\mu)\right],$$

llamando

$$\mathbf{S} = \frac{1}{n}\sum_{i=1}^{n}(\mathbf{x}_i-\overline{\mathbf{x}})(\mathbf{x}_i-\overline{\mathbf{x}})' \tag{2.13}$$

a la *matriz de covarianza muestral*, definida anteriormente en (1.21), tenemos:

$$\sum_{i=1}^{n}(\mathbf{x}_i-\mu)'\mathbf{V}^{-1}(\mathbf{x}_i-\mu) = n\ tra[\mathbf{V^{-1}S}] + tra[\ n(\overline{\mathbf{x}}-\mu)'\mathbf{V}^{-1}(\overline{\mathbf{x}}-\mu)]. \tag{2.14}$$

Sustituyendo en la función de soporte (2.10):

$$\begin{aligned} L(\mu,\mathbf{V}) &= -\frac{1}{2}pn\ln(2\pi) - \frac{1}{2}n\ln|\mathbf{V}| - \frac{1}{2}\Big[\ n\ tra[\mathbf{V^{-1}S}] + tra[\ n\ (\overline{\mathbf{x}}-\mu)'\mathbf{V}^{-1}(\overline{\mathbf{x}}-\mu)]\Big] \\ &= -\frac{1}{2}pn\ln(2\pi) - \frac{1}{2}n\ln|\mathbf{V}| - \frac{n}{2}tra\left[\mathbf{V^{-1}S}\right] - \frac{n}{2}tra\left[(\overline{\mathbf{x}}-\mu)'\mathbf{V}^{-1}(\overline{\mathbf{x}}-\mu)\right]. \end{aligned} \tag{2.15}$$

Ahora, para encontrar los estimadores de maxima verosimilitud se debe resolver la ecuación dada en (2.7), la cual nos sugiere el siguiente sistema de ecuaciones homogéneo:

$$\begin{aligned} \frac{\partial L(\mu,\mathbf{V})}{\partial\mu} &= 0 \\ \frac{\partial L(\mu,\mathbf{V})}{\partial\mathbf{V}} &= 0. \end{aligned}$$

En la primera ecuación, tomando los mismos argumentos que se utilizaron para hallar (2.12), se tiene:

$$\begin{aligned} \frac{\partial}{\partial\mu}L(\mu,\mathbf{V}) &= \frac{\partial}{\partial\mu}\left\{-\frac{1}{2}pn\ln(2\pi) - \frac{1}{2}n\ln|\mathbf{V}| - \frac{n}{2}tra\left[\mathbf{V^{-1}S}\right] - \frac{n}{2}tra\left[(\overline{\mathbf{x}}-\mu)'\mathbf{V}^{-1}(\overline{\mathbf{x}}-\mu)\right]\right\} \\ &= -\frac{n}{2}\left\{\frac{\partial}{\partial\mu}\left[(\overline{\mathbf{x}}-\mu)'\mathbf{V}^{-1}(\overline{\mathbf{x}}-\mu)\right]\right\} = 0. \end{aligned}$$

Aplicando el resultado dado en la ecuación (1.13) decimos que:

$$\frac{\partial}{\partial\mu}L(\mu,\mathbf{V}) = -\frac{n}{2}\left\{(2)(\mathbf{V}^{-1})(\overline{\mathbf{x}}-\mu)(-1)\right\} = 0$$

$$= n\mathbf{V}^{-1}(\overline{\mathbf{x}} - \mu) = 0,$$

entonces :

$$\begin{aligned} \frac{n(\overline{\mathbf{x}} - \mu)}{\mathbf{V}} &= 0 \\ n(\overline{\mathbf{x}} - \mu) &= 0 \\ \overline{\mathbf{x}} - \mu &= 0 \\ \widehat{\mu} &= \overline{\mathbf{x}}. \end{aligned} \tag{2.16}$$

Luego el estimador de maxima verosimilitud de μ es $\widehat{\mu} = \overline{\mathbf{x}}$.
Similarmente en la segunda ecuación del sistema homogéneo se tiene:

$$\begin{aligned} \frac{\partial}{\partial \mathbf{V}} L(\mu, \mathbf{V}) &= \frac{\partial}{\partial \mathbf{V}} \left\{ -\frac{1}{2} pn \ln(2\pi) - \frac{1}{2} n \ln|\mathbf{V}| - \frac{n}{2} tra\left[\mathbf{V^{-1}S}\right] - \frac{n}{2} tra\left[(\overline{\mathbf{x}} - \mu)'\mathbf{V}^{-1}(\overline{\mathbf{x}} - \mu)\right] \right\} \\ &= -\frac{n}{2}\left\{ \frac{\partial}{\partial \mathbf{V}} \ln|\mathbf{V}| \right\} - \frac{n}{2}\left\{ \frac{\partial}{\partial \mathbf{V}} tra\left[\mathbf{V^{-1}S}\right] \right\} - \frac{n}{2}\left\{ \frac{\partial}{\partial \mathbf{V}} tra\left[(\overline{\mathbf{x}} - \mu)'\mathbf{V}^{-1}(\overline{\mathbf{x}} - \mu)\right] \right\} \\ &= -\frac{n}{2}\left\{ \frac{\partial}{\partial \mathbf{V}} \ln|\mathbf{V}| + \frac{\partial}{\partial \mathbf{V}} tra\left[\mathbf{V^{-1}S}\right] + \frac{\partial}{\partial \mathbf{V}} tra\left[\mathbf{V}^{-1}(\overline{\mathbf{x}} - \mu)(\overline{\mathbf{x}} - \mu)'\right] \right\} = 0. \end{aligned}$$

Según las propiedades b) y g) de las derivadas matriciales vistas en la sección 1.2.4 y puesto que $\mathbf{V}$ es simétrica, se tiene:

$$\begin{aligned} -\frac{n}{2}\left\{ (\mathbf{V}')^{-1} - \left[\mathbf{V^{-1}SV^{-1}}\right] - \left[\mathbf{V^{-1}}(\overline{\mathbf{x}} - \mu)(\overline{\mathbf{x}} - \mu)'\mathbf{V^{-1}}\right] \right\} &= 0 \\ (\mathbf{V})^{-1} - \left[\mathbf{V^{-1}SV^{-1}}\right] - \left[\mathbf{V^{-1}}(\overline{\mathbf{x}} - \mu)(\overline{\mathbf{x}} - \mu)'\mathbf{V^{-1}}\right] &= 0. \end{aligned}$$

Multiplicando por la matriz $\mathbf{V}$, primero a derecha y después a izquierda, obtenemos:

$$\begin{aligned} (\mathbf{V} \times \mathbf{V^{-1}}) - (\mathbf{V} \times \mathbf{V^{-1}})\mathbf{SV^{-1}} - (\mathbf{V} \times \mathbf{V^{-1}})(\overline{\mathbf{x}} - \mu)(\overline{\mathbf{x}} - \mu)'\mathbf{V^{-1}} &= 0 \\ \mathbf{I} - \mathbf{SV^{-1}} - (\overline{\mathbf{x}} - \mu)(\overline{\mathbf{x}} - \mu)'\mathbf{V^{-1}} &= 0 \\ (\mathbf{I} \times \mathbf{V}) - \mathbf{S}(\mathbf{V^{-1}} \times \mathbf{V}) - (\overline{\mathbf{x}} - \mu)(\overline{\mathbf{x}} - \mu)'(\mathbf{V^{-1}} \times \mathbf{V}) &= 0 \\ \mathbf{V} - \mathbf{S} - (\overline{\mathbf{x}} - \mu)(\overline{\mathbf{x}} - \mu)' &= 0. \end{aligned}$$

Luego reemplazando el resultado obtenido en (2.16) tenemos:

$$\begin{aligned} \mathbf{V} - \mathbf{S} &= 0 \\ \widehat{\mathbf{V}} &= \mathbf{S} \end{aligned} \tag{2.17}$$

Entonces los estimadores MV de μ y $\mathbf{V}$ son $\overline{\mathbf{x}}$ y $\mathbf{S}$ respectivamente.
En resumen los estimadores de maxima verosimilitud para μ y $\mathbf{V}$ son:

$$\widehat{\mu} \equiv \overline{\mathbf{x}} = \begin{pmatrix} \overline{\mathbf{x}}_1 \\ \vdots \\ \overline{\mathbf{x}}_p \end{pmatrix}, \quad y \quad \widehat{\mathbf{V}} \equiv \mathbf{S} = \left(\frac{1}{n} \sum_{k=1}^{n} (x_{ik} - \overline{\mathbf{x}}_i)(x_{jk} - \overline{\mathbf{x}}_j) \right), \quad i, j = 1, \ldots, p, \quad (2.18)$$

los cuales tienen las siguientes propiedades

1. $\overline{\mathbf{x}}$ y $n\mathbf{S}$ tienen distribución $N(\mu, \frac{1}{n}\mathbf{V})$ y $\mathcal{W}_p(n-1, \mathbf{V})$ [3] respectivamente.
2. $\overline{\mathbf{x}}$ y $\mathbf{S}$ son independientes.
3. $\overline{\mathbf{x}}$ y $\mathbf{S}$ son estimadores insesgados de μ y de $\mathbf{V}$.
4. $\overline{\mathbf{x}}$ y $\mathbf{S}$ son consistentes.
5. $\overline{\mathbf{x}}$ y $\mathbf{S}$ son estadísticas suficientes.

2.3. Contrastes multivariantes: el método de la razón de verosimilitudes

En el análisis de datos multivariantes, los contrastes de hipótesis son mas complejos de realizar que en los casos univariados; así por ejemplo tomemos la distribución normal p -variante, la cual tiene p -medias, p-varianzas y $\binom{p}{2}$ covarianzas. Obsérvese además que si se realizan pruebas de hipótesis en forma separada para el numero total de parámetros, es necesario formular $\frac{1}{2}p(p+3)$ hipótesis, lo cual hace engorroso estas pruebas cuando el tamaño de p es grande. Existen además otra serie de inconvenientes al trabajar con contrastes univariados para los datos multivariantes como son los casos de el incremento en la tasa de error tipo I y la no consideración de las posibles correlaciones existentes entre variables. Es así como en esta sección se ve la necesidad de estudiar pruebas multivariantes para el contraste de hipótesis de datos multivariantes.
Con frecuencia se desea comprobar si una muestra dada puede provenir de una distribución con ciertos parámetros conocidos y en otros casos nos interesa comprobar si varias muestras multivariantes provienen o no de la misma población (muchas veces basando esta inferencia

[3] La distribución Wishart $\mathcal{W}_p(m, \mathbf{V})$, donde m son los grados de libertad y $\mathbf{V}$ es la matriz de covarianzas, se sintetiza en el **Apéndice A**.

bajo la hipótesis de normalidad), por lo que se hace necesario realizar un contraste para ver si nuestra hipótesis no es rechazada por los datos observados. Para realizar contrastes de parámetros vectoriales aplicamos la teoría de contraste de verosimilitudes, que proporciona pruebas estadísticas con propiedades óptimas para tamaños muestrales grandes.
Supongamos que la función de densidad de $(X_1, \ldots, X_p)'$ es $f(\mathbf{X}|\theta)$ donde $\mathbf{X} \in \mathbb{R}^p$ y θ es un parámetro vectorial que toma valores en la region paramétrica Ω, con $\Omega \subset \mathbb{R}^p$. Sean $\Omega_0 \subset \Omega$ y $\Omega_1 = \Omega - \Omega_0$ sub-regiones paramétricas. Planteamos el contraste de hipótesis:

$$H_0 : \theta \in \Omega_0 \quad \text{vs} \quad H_1 : \theta \in \Omega_1 \ , \tag{2.19}$$

el cual establece en la hipótesis nula H_0 que θ está contenida en una región Ω_0 del espacio paramétrico, frente a una hipótesis alternativa H_1 que supone que θ no esta restringida a la region Ω_0.
Ahora derivemos la estadística de prueba pertinente. Este problema se conoce como *el problema de Hotelling*. El problema consiste en contrastar la hipótesis nula H_0 contra la hipótesis alternativa H_1, comparando las respectivas probabilidades de obtener los datos: situación en la cual requeriremos el vector de parámetros, que por lo general es desconocido.
El método de la razón de verosimilitudes resuelve este problema tomando el valor que hace mas probable obtener la muestra observada y que es compatible con la hipótesis. Observemos que si H_0 y H_1 son ambas hipótesis simples, Ω_0 y Ω_1 tienen cada uno un solo elemento, entonces la razón de las verosimilitudes se obtiene encontrando el valor del cociente entre las verosimilitudes $l(H_0)$ y $l(H_1)$ supuesto θ_0 y θ_1 respectivamente. En el caso que Ω_0 y Ω_1 tengan muchos valores, compararemos entre ellos las dos cantidades $\max l(H_0) \equiv \max l_0$ y $\max l(H_1) \equiv \max l_1$ donde $\max l_0$ es el valor máximo de la función de verosimilitud para todos los valores θ en Ω_0, y $\max l_1$ es el valor máximo de la función de verosimilitud para todos los valores de θ en Ω. Este ultimo valor debería calcularse estrictamente sobre el espacio Ω_1, pero es mas simple hacerlo sobre todo el espacio, ya que en general se obtiene el mismo resultado.
La definición de *la razón de verosimilitudes* (λ) resume las ideas anteriormente expuestas.

Definición 2.5. Sea Ω_0 y Ω_1 sub-regiones paramétricas del espacio $\Omega \subset \mathbb{R}^n$. *La razón de verosimilitudes* es el estadístico:

$$\lambda = \frac{\max l_0}{\max l_1} \quad \text{que satisface} \quad 0 \leq \lambda \leq 1, \tag{2.20}$$

donde $\max l_0$ y $\max l_1$ son los valores máximos de la función de verosimilitud para los valores de θ en Ω_0 y Ω respectivamente.

Nótese que el denominador de esta razón es el máximo sobre todo el espacio paramétrico Ω y además como el conjunto de los parámetros restringido por la hipótesis nula Ω_0 esta contenido en el espacio de parámetros completo Ω, el numerador es menor que el denominador. Valores de λ cercanos a 1 provocan decisiones a favor de H_0, en tanto valores cercanos a 0 sugieren el rechazo de H_0. Luego *la región critica o región de rechazo* de H_0 vendrá definida por:

$$\lambda \leq k, \quad donde \quad 0 < k < 1. \tag{2.21}$$

Esta desigualdad genera un **test de razón de verosimilitud** de la hipótesis nula H_0 contra la hipótesis alternativa H_1, donde el valor de k es determinado por el nivel de significancia α escogido para el test; luego se escoge k de manera que el tamaño de la region critica sea α. Para hallar k es necesario conocer la distribución de λ cuando H_0 es cierta, lo que suele ser difícil en la practica.
El test basado en λ tiene muchas aplicaciones en el análisis multivariate, pero en la mayoría de casos la distribución de los datos es desconocida. Sin embargo cuando el tamaño muestral es grande, existe un importante resultado atribuido a Wilks[4], el cual garantiza que la distribución del doble de la diferencia de las funciones de soporte entre la alternativa y la nula (cuando H_0 es cierta), o lo que es lo mismo, -2 veces el logaritmo de λ, es una distribución χ^2. Este resultado se da formalmente en el siguiente teorema, el cual se encuentra en [4].

Teorema 2.1. *Sea*

$$\delta_L \equiv 2[L(H_1) - L(H_0)] = -2\ln(\lambda) = \mathcal{D}(H_0) - \mathcal{D}(H_1),$$

donde $L(H_i) = \ln \mathrm{l}(H_i) = \ln \mathrm{l}_i$, *con* $i = 0, 1$. *Bajo ciertas condiciones de regularidad se verifica que:*

$$\delta_L \quad se\ distribuye\ asintóticamente \quad \chi^2_{s-t}, \tag{2.22}$$

donde $t = \dim(\Omega_0) < s = \dim(\Omega)$.

Es frecuente que la dimension de Ω sea p y la dimension de Ω_0 sea $p-r$, siendo r el numero de restricciones lineales sobre el vector de parámetros. Entonces el numero de grados de libertad de la diferencia de soporte δ_L es r.

[4]Samuel S. Wilks (1906-1964): Estadístico estadounidense. Construyó generalizaciones multivariantes para el análisis de varianza y el coeficiente de correlación multiple. Fue uno de los fundadores del Institute of Mathematical Statistics (1935) y editor de la revista Annals of Mathematical Statistics durante once años.

2.3.1. Comparación de medias: el análisis de varianza multivariante.

Supongamos que hemos observado una muestra de datos independientes, con tamaño n provenientes de una población normal p-variante, que puede estratificarse en G clases o grupos de manera que existe n_1 observaciones del grupo 1, n_2 observaciones del grupo $2, \ldots, n_G$ del grupo G. Luego podemos decir que tenemos G matrices con las características dadas en el cuadro (2.1).

Matriz	Orden	Medias	Covarianzas	Distribución
$\mathbb{X}_1$	$n_1 \times p$	$\overline{\mathbf{x}}_1$	$\mathbf{S}_1$	$N_p(\mu_1, \mathbf{V})$
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
$\mathbb{X}_g$	$n_g \times p$	$\overline{\mathbf{x}}_g$	$\mathbf{S}_g$	$N_p(\mu_g, \mathbf{V})$
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
$\mathbb{X}_G$	$n_G \times p$	$\overline{\mathbf{x}}_G$	$\mathbf{S}_G$	$N_p(\mu_G, \mathbf{V})$

Cuadro 2.1: Datos muestrales estratificados en G grupos.

Nuestro problema esta en contrastar las hipótesis:

$$H_0 : \mu_1 = \ldots = \mu_G = \mu,\ \mathbf{V}_i = \mathbf{V} \quad para \quad i = 1, \ldots, G,$$
$$\text{vs}$$
$$H_1 : no\ todas\ las\ \mu_i\ son\ iguales;\ \mathbf{V}_i = \mathbf{V} \quad para \quad i = 1, \ldots, G. \tag{2.23}$$

Según los posibles contrastes a realizar en la distribución normal multivariante, presentados en este documento en el **Apéndice B**, se sabe que la función de verosimilitud bajo H_0 de una muestra normal homogénea alcanza su máximo en $\widehat{\mu} = \overline{\mathbf{x}}$ y $\widehat{\mathbf{V}} = \mathbf{S}$ y que su función de soporte estará dada de forma similar a la ecuación (B.11) por:

$$L(H_0) = -\frac{n}{2} \ln |\mathbf{S}| - \frac{np}{2}.$$

Bajo H_1, los n vectores de observaciones se subdividen en n_1 del grupo 1, n_2 del grupo2, así sucesivamente hasta el grupo G que tiene n_G observaciones. La función de verosimilitud bajo H_1 es:

$$l(\mu_1, \mu_2, \ldots, \mu_p; \mathbf{V}|\mathbb{X}) = |\mathbf{V}|^{-n/2}(2\pi)^{-np/2} \exp\left\{ -\frac{1}{2} \sum_{g=1}^{G} \sum_{h=1}^{n_g} (x_{hg} - \mu_g)' \mathbf{V}^{-1} (x_{hg} - \mu_g) \right\}, \tag{2.24}$$

donde x_{hg} es el elemento h del grupo g y μ_g es el vector de medias de dicho grupo. La maximización de esta función en el espacio paramétrico definido por H_1 se realiza de forma similar a la estudiada en la sección 2.2.1. La estimación de la media de cada grupo será la media muestral, $\widehat{\mu}_g = \overline{\mathbf{x}}_g$, y la estimación de la matriz de covarianzas común se obtiene utilizando el hecho que:

$$\begin{aligned}\sum_{g=1}^{G}\sum_{h=1}^{n_g}(x_{hg}-\overline{\mathbf{x}}_g)'\mathbf{V}^{-1}(x_{hg}-\overline{\mathbf{x}}_g) &= tra\left(\sum_{g=1}^{G}\sum_{h=1}^{n_g}(x_{hg}-\overline{\mathbf{x}}_g)'\mathbf{V}^{-1}(x_{hg}-\overline{\mathbf{x}}_g)\right)\\ &= \sum_{g=1}^{G}\sum_{h=1}^{n_g}tra\left(\mathbf{V}^{-1}(x_{hg}-\overline{\mathbf{x}}_g)(x_{hg}-\overline{\mathbf{x}}_g)'\right)\\ &= tra\big(\mathbf{V}^{-1}\mathbf{W}\big) \qquad (2.25)\end{aligned}$$

donde

$$\mathbf{W} = \sum_{g=1}^{G}\sum_{h=1}^{n_g}(x_{hg}-\overline{\mathbf{x}}_g)(x_{hg}-\overline{\mathbf{x}}_g)'. \qquad (2.26)$$

La matriz $\mathbf{W}$ se llama *matriz suma de cuadrados dentro de los grupos*, la cual verifica bajo nuestra hipótesis H_0 que: $\mathbf{W} \sim \mathcal{W}_p(n-G, \mathbf{V})$. Sustituyendo (2.25) en la función de verosimilitud (2.24) y tomando logaritmos se obtiene:

$$L(\mathbf{V}|\mathbb{X}) = -\frac{n}{2}\ln|\mathbf{V}| - \frac{n}{2}tra\left(\mathbf{V}^{-1}\frac{\mathbf{W}}{n}\right), \qquad (2.27)$$

y según los resultados obtenidos en la sección 2.2.1, la varianza común a los grupos cuando estos tienen distinta media se estima por:

$$\boxed{\widehat{\mathbf{V}} = \mathbf{S}_w = \frac{1}{n}\mathbf{W}.}$$

Luego la función de soporte para la hipótesis alternativa viene dada por:

$$L(H_1) = -\frac{n}{2}\ln|\mathbf{S}_w| - \frac{np}{2}, \qquad (2.28)$$

con la cual se establece que el doble de la diferencia de los soportes es:

$$\delta_L = 2[L(H_1) - L(H_0)] = n\ln\left[\frac{|\mathbf{S}|}{|\mathbf{S}_w|}\right]. \qquad (2.29)$$

Rechazaremos H_0 cuando esta diferencia sea grande; es decir cuando la variabilidad suponiendo H_0 cierta, medida por $|\mathbf{S}|$, sea mucho mayor que la variabilidad cuando permitimos que

las medias de los grupos sea distinta, medida por $|\mathbf{S}_w|$.
La distribución de δ_L es, asintóticamente, una χ^2_g donde los grados de libertad, g, se obtienen por la diferencia de las dimensiones de los espacios paramétricos. Dicha diferencia es:

$$g = dim(\Omega) - dim(\Omega_0) = \left[p + \frac{p(p+1)}{2}\right] - \left[Gp + \frac{p(p+1)}{2}\right] = p(G-1).$$

La aproximación a la distribución χ^2_g del cociente de verosimilitudes puede mejorarse para tamaños pequeños. En [1] se muestra que el estadístico

$$\delta_L = m \ln \frac{|\mathbf{S}|}{|\mathbf{S}_w|}, \tag{2.30}$$

donde

$$m = (n-1) - (p+G)/2,$$

sigue asintóticamente una distribución χ^2_g donde $g = p(G-1)$ y la aproximación es mejor que tomando $m = n$ en muestras pequeñas.

El análisis de la varianza multivariante.

Sea $\mathbf{T}$ *la variabilidad total* de los datos dada por:

$$\mathbf{T} = \sum_{i=1}^{n} (x_i - \overline{\mathbf{x}})(x_i - \overline{\mathbf{x}})' \quad donde \quad \overline{\mathbf{x}} = \frac{1}{n}\sum_{g=1}^{G} n_g \overline{\mathbf{x}}_g, \tag{2.31}$$

la cual mide las desviaciones respecto a la media común y que bajo nuestra hipótesis H_0 verifica que $\mathbf{T} \sim \mathcal{W}_p(n-1, \mathbf{V})$. Vamos a descomponer la matriz $\mathbf{T}$ como la suma de dos matrices. La primera es nuestra ya conocida $\mathbf{W}$, la matriz de suma de cuadrados dentro de los grupos, que mide las desviaciones respecto a las medias de cada grupo. La segunda medirá la variabilidad explicada por las diferencias entre las medias y la llamaremos $\mathbf{B}$, la cual se obtendrá sumando y restando las medias de cada grupo en la expresión de $\mathbf{T}$, es decir:

$$\mathbf{T} = \sum_{g=1}^{G}\sum_{h=1}^{n_g} (x_{gh} - \overline{\mathbf{x}}_g + \overline{\mathbf{x}}_g - \overline{\mathbf{x}})(x_{gh} - \overline{\mathbf{x}}_g + \overline{\mathbf{x}}_g - \overline{\mathbf{x}})'.$$

Desarrollando se comprueba que el doble producto se anula y resulta:

$$\boxed{\mathbf{T} = \mathbf{B} + \mathbf{W}.} \tag{2.32}$$

donde $\mathbf{B}$ es:

$$\mathbf{B} = \sum_{g=1}^{G} n_g(\overline{\mathbf{x}}_g - \overline{\mathbf{x}})(\overline{\mathbf{x}}_i - \overline{\mathbf{x}})', \tag{2.33}$$

y se llama *la matriz de sumas de cuadrados entre grupos*, en la cual bajo la hipótesis nula se verifica que $\mathbf{B} \sim \mathcal{W}_p(G-1, \mathbf{V})$. Esta descomposición también puede expresarse como:

$$\boxed{Variabilidad\ total\ (\mathbf{T}) = Variabilidad\ explicada\ (\mathbf{B})\ +\ Variabilidad\ residual\ (\mathbf{W})} \tag{2.34}$$

que es la descomposición habitual del análisis de la varianza.
Luego el contraste planteado en (2.23) lo podemos hacer comparando el tamaño de las matrices $\mathbf{T}$ y $\mathbf{W}$. La medida de tamaño adecuada es el determinante, con lo que concluimos que el contraste debe basarse en el cociente

$$\frac{|\mathbf{W}|}{|\mathbf{T}|} = \frac{|\mathbf{W}|}{|\mathbf{B}+\mathbf{W}|} \equiv \Lambda$$

Esta estadística se conoce como *la distribución Λ de Wilks*, la cual se muestra en forma detallada en [4]. Entonces nuestro estadístico para el contraste planteado en (2.23) es:

$$\Lambda = \frac{|\mathbf{W}|}{|\mathbf{B}+\mathbf{W}|} \sim \Lambda_{(p,n-G.G-1)}. \tag{2.35}$$

La region critica a un nivel de significación de α es $\Lambda < \Lambda_{(\alpha,p,n-G.G-1)}$. La distribución exacta de Λ ha sido obtenida en para algunos casos especiales que se resumen en la siguiente tabla dada en [2].

No. Variables	**No. Grupos**	**Transformación**	**Distribución F**
$p=1$	$G \geq 2$	$\left(\frac{1-\Lambda}{\Lambda}\right)\left(\frac{n-G}{G-1}\right)$	$F_{(G-1,n-G)}$
$p=2$	$G \geq 2$	$\left(\frac{1-\Lambda^{1/2}}{\Lambda^{1/2}}\right)\left(\frac{n-G-1}{G-1}\right)$	$F_{(2(G-1),2(n-G-1))}$
$p \geq 1$	$G=2$	$\left(\frac{1-\Lambda}{\Lambda}\right)\left(\frac{n-p-1}{p}\right)$	$F_{(p,n-p-1)}$
$p \geq 1$	$G=2$	$\left(\frac{1-\Lambda^{1/2}}{\Lambda^{1/2}}\right)\left(\frac{n-p-2}{p}\right)$	$F_{(2p,2(n-p-2))}$

Para muestras de tamaño grande se tiene **la estadística de Bartlett**

$$\mathcal{V} = \left(n-1-\frac{(p+G)}{2}\right)\ln\Lambda = -\left(n-1-\frac{(p+G)}{2}\right)\ln\frac{|\mathbf{W}|}{|\mathbf{B}+\mathbf{W}|}, \tag{2.36}$$

la cual tiende aproximadamente a una $\chi^2_{(p(G-1))}$. Se rechaza H_0 para valores $\mathcal{V}$ mayores que $\chi^2_{(\alpha,p(G-1))}$.

2.4. Estimación y Contrastes Bayesianos.

2.4.1. Estimación.

Hasta ahora hemos supuesto que los parámetros que se quieren estimar están conformados por constantes desconocidas; en el enfoque bayesiano los parámetros son variables aleatorias que tienen *distribuciones a priori* $p(\theta)$, llamadas también *previas o neutras*, que suele reflejar la fortaleza de las creencias de uno sobre los posibles valores que se pueden asumir, además la inferencia respecto a los posibles valores de los parámetros se realiza sobre la ecuación (2.15) para obtener su distribución condicionada a la información disponible. Una vez obtenida la distribución de probabilidad del parámetro, los problemas habituales de inferencia quedan resueltos con la *distribución a posteriori o posterior*, de manera simple. Si llamamos $\mathbb{X}$ a la matriz de datos, con distribución conjunta $f(\mathbb{X}|\theta)$ que nos proporciona las probabilidades de los valores muestrales conocido el vector de parámetros, la distribución *a posteriori* $p(\theta|\mathbb{X})$ será:

$$p(\theta|\mathbb{X}) = \frac{f(\mathbb{X}|\theta)p(\theta)}{\int f(\mathbb{X}|\theta)p(\theta)d(\theta)}, \tag{2.37}$$

donde el denominador de esta expresión es la distribución marginal de los datos.
En la practica el calculo de la *a posteriori* se simplifica observando que el denominador no depende de θ y este actúa únicamente como una constante de normalización para que la integral de $p(\theta|\mathbb{X})$ sea la unidad. Por lo tanto podemos calcular la distribución *a posteriori* escribiendo:

$$p(\theta|\mathbb{X}) = k\ l(\theta|\mathbb{X})p(\theta), \tag{2.38}$$

ya que $\mathbb{X}$ es constante y al considerar $f(\mathbb{X}|\theta)$ como una función de θ que se convierte en la función de verosimilitud $l(\theta|\mathbb{X})$. Multiplicando para cada valor de θ las ordenadas de $l(\theta|\mathbb{X})$ y $p(\theta)$ resulta la distribución *a posteriori*.
Si no disponemos de información *a priori*, o se desea que los datos hablen por si solos, se debe establecer una *distribución a priori no informativa* o *distribución de referencia*. Intuitivamente una distribución a priori no informativa, para un vector de parámetros de localización es aquella que es localmente uniforme sobre la zona relevante del espacio paramétrico, luego escribiremos $p(\theta) = c$.
Jeffreys (1961) ha estudiado el problema de establecer distribuciones de referencia con propiedades razonables. Para distribuciones normales, la distribución de referencia para un vector de parámetros podemos tomarla como localmente uniforme y suponer en la

zona relevante para la inferencia que $p(\theta) = c$. Para matrices de covarianza Jeffreys, por consideraciones de invarianza ante transformaciones, propuso tomar la distribución de referencia proporcional al determinante de la matriz de covarianzas elevado al exponente $-(p+1)/2$, donde p es la dimensión de la matriz.
En [1] se encuentra en forma mas detallada la estimación bayesiana, además de citar varias referencias bibliográficas para la posterior consulta.

2.4.2. Contrastes.

En el enfoque bayesiano la hipótesis nula no se acepta o se rechaza como en el enfoque clásico, lo que hacemos es determinar su probabilidad *a posteriori* dados los datos. Supongamos el contraste general; dado un parámetro vectorial p-dimensional θ, que toma valores en Ω, deseamos contrastar las hipótesis:

$$H_0 : \theta \in \Omega_0, \quad \text{vs} \quad H_1 : \theta \in \Omega - \Omega_0. \tag{2.39}$$

Supongamos que existen probabilidades *a priori* para cada una de las dos hipótesis. Estas probabilidades quedan automáticamente determinadas si establecemos una distribución *a priori* sobre θ, ya que entonces:

$$p_0 = P(H_0) = P(\theta \in \Omega_0) = \int_{\Omega_0} p(\theta) d\theta, \tag{2.40}$$

$$p_1 = P(H_1) = P(\theta \in \Omega - \Omega_0) = \int_{\Omega - \Omega_0} p(\theta) d\theta. \tag{2.41}$$

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

$$P(H_i|\mathbb{X}) = \frac{P(\mathbb{X}|H_i)P(H_i)}{P(\mathbb{X})} \; i = 0, 1 \tag{2.42}$$

y de aquí se obtiene el resultado fundamental:

$$\frac{P(H_0|\mathbb{X})}{P(H_1|\mathbb{X})} = \frac{f(\mathbb{X}|H_0)}{f(\mathbb{X}|H_1)} \cdot \frac{P(H_0)}{P(H_1)} \tag{2.43}$$

que puede expresarse como

$$\boxed{\textit{razón de posterioris} = \textit{razón de verosimilitudes} \times \textit{razon de prioris.}}$$

Al cociente entre las verosimilitudes se le denomina factor de Bayes ($\mathfrak{B}$) y si las probabilidades *a priori* de ambas hipótesis son las mismas este factor determina las probabilidades

a posteriori de las hipótesis. Expresando las probabilidades a posteriori en términos del parámetro, se obtiene

$$P(H_i|\mathbb{X}) = P(\theta \in \Omega_i|\mathbb{X}) = \int_{\Omega_i} p(\theta|\mathbb{X})d\theta \quad i = 1, 0 \tag{2.44}$$

donde $\Omega_1 = \Omega - \Omega_0$ y $p(\theta|\mathbb{X})$ es la distribución a posteriori para el vector de parámetros de interés. Por lo tanto.

$$P(H_i|\mathbb{X}) = \frac{1}{f(\mathbb{X})} \int_{\Omega_i} f(\mathbb{X}|\theta)p(\theta)d(\theta) \quad i = 0, 1 \tag{2.45}$$

Sustituyendo en (2.43) se obtiene que el factor de Bayes de la primera hipótesis respecto a la segunda ,$\mathfrak{B}_{01}$, es

$$\mathfrak{B}_{01} = \frac{p_1}{p_0} \frac{\int_{\Omega_0} f(\mathbb{X}|\theta)p(\theta)d(\theta)}{\int_{\Omega-\Omega_0} f(\mathbb{X}|\theta)p(\theta)d(\theta)}.$$

Harold Jeffreys[5] ha dado una escala de evidencia para el factor de Bayes que se representa en la siguiente tabla dada en [2].

$\log_{10}\mathfrak{B}_{01}$	$\mathfrak{B}_{01}$	$P(H_0) \quad para \quad (p_0/p_1 = 1)$	Interpretación
0	1	0.5	indecisión
-1	10^{-1}	0.1	débil rechazo de H_0
-2	10^{-2}	0.01	rechazo de H_0
-3	10^{-3}	0.001	rechazo sin duda de H_0

Comparación entre los contrastes bayesianos y los clásicos.

Si suponemos que, *a priori*, las probabilidades de ambas hipótesis son las mismas, el factor de Bayes es comparable con la razón del contraste de las verosimilitudes, pero existe una diferencia fundamental: en el contraste de verosimilitudes se toma el máximo de la verosimilitud mientras que en el enfoque bayesiano se toma el promedio sobre la region relevante, promediando con la distribución *a priori*. Por lo tanto el contraste tiene en cuenta al calcular la integral el tamaño del espacio definido por Ω_0 y por Ω_1. Por

[5]Harold Jeffreys (1891-1989) fue un físico y estadístico británico. Desarrollo el enfoque bayesiano para la inferencia estadística en su celebre libro *Theory of Probability*, donde resolvió problemas de inferencia multivariante desde el punto de vista bayesiano.

ejemplo supongamos que θ es un parámetro escalar $0 \leq \theta \leq 1$ y que contrastamos las hipótesis:

$$H_0 : \theta = \theta_0, \quad \text{vs} \quad H_1 : \theta \neq \theta_0, \tag{2.46}$$

para que ambas probabilidades sean las mismas, supongamos que fijamos $p(\theta = \theta_0) = 1/2$ y que $p(\theta) = 1/2 \quad \text{si} \quad \theta \neq \theta_0$. Luego el factor de Bayes compara $f(\mathbb{X}|\theta_0)$ con el valor promedio de la verosimilitud cuando $\theta \neq \theta_0$, mientras que el contrastes de verosimilitudes compara $f(\mathbb{X}|\theta_0)$ con el valor máximo de la verosimilitud. Si el valor $\theta = \theta_0$ no es exactamente cierto, sino aproximadamente cierto, y el tamaño de la muestra es muy grande $(n \to \infty)$, decimos que el enfoque clásico rechaza H_0 en la practica, mientras que con en el enfoque bayesiano cuando $(n \to \infty)$ es mas difícil rechazar H_0 en la practica. Esto es consecuencia de que en el enfoque bayesiano tiene en cuenta la verosimilitud de H_0 y de H_1, mientras que el enfoque clásico mira sólo H_0.
Es importante señalar que esta contradicción desaparece en el momento que reformulamos el problema como uno de estimación. Entonces ambos métodos coincidirán con muestras grandes en la estimación del parámetro.

2.5. Selección de modelos.

El método de maxima verosimilitud supone que la forma del modelo estadístico es conocida y solo falta estimar los parámetros. Cuando no es así debe aplicarse con cuidado. Es por esto que si tenemos varios candidatos a ser *el modelo estadístico* que mejor se ajusta a nuestros datos, escogemos el que posea mayor soporte para los datos. Además si añadimos más términos o más variables a un modelo, el soporte mejorará y si la muestra es grande será difícil distinguir mediante la razón de verosimilitudes entre una mejora *real* y una aportación trivial.
El modelo perfecto no existe, puesto que todos constituyen simplificaciones de la realidad y siempre son preferibles modelos con menos variables, puesto que además de ser más sencillos, son más estables y menos sometidos a sesgo. Es por esto que selección de modelos estadísticos tiene como objetivo último el comparar varios modelos alternativos con el objeto de elegir aquél más adecuado. La solución habitual para escoger entre varios modelos es hacer un contraste de hipótesis utilizando el contraste de verosimilitudes mediante el teorema 2.1 y tomar como nuestro modelo aquel que maximice δ_L, es decir al que haga máximo el valor:

$$\delta_L = 2[L(M_r) - L(M_s)] = \mathcal{D}(M_s) - \mathcal{D}(M_r),$$

donde $L(M_r)$ y $L(M_s)$ son los soportes de los modelos M_r y M_s, respectivamente, al sustituir en cada una de estas funciones el parámetro θ por su estimador MV y $\mathcal{D}(M_s) = -2L(M_s)$, $\mathcal{D}(M_r) = -2L(M_r)$ las desviaciones. Debido a que el modelo con mas parámetros es el que maximiza δ_L, se han propuesto medidas de contraste entre modelos que penalizan en alguna medida que éstos tengan muchos parámetros. Las más conocidas son **el criterio de información de Akaike (AIC)** y **el criterio de información bayesiano (BIC)**.

2.5.1. El criterio de información de Akaike (AIC).

El criterio de información de Akaike **(AIC)** fue desarrollado por Hirotsugu Akaike en (1974), el cual es una medida de la bondad de ajuste de un modelo de estimación estadística. El **AIC** es una manera operacional de acortar distancia entre la complejidad de un modelo estimado y un buen modelo ajustado a los datos. El criterio definido por Akaike se basa en **la medida de información de Kullback-Leibler (KL)** dado en (1951), el cual permite interpretar la distancia entre las dos distribuciones a partir del soporte de un modelo.
Si asumimos que nuestros datos siguen una función de densidad verdadera $f(\mathbf{y})$ y nuestro modelo estadístico aproximado de $f(\mathbf{y})$ es $g(\mathbf{y})$, la medida de información de Kullback-Leibler **(KL)** sera un patrón para medir la similitud entre el modelo estadístico y la verdadera distribución. La medida **KL** viene dada por:

$$\begin{aligned}\mathbf{KL}[g(\mathbf{y})|f(\mathbf{y})] &= \mathrm{E}_{\mathbf{y}}\left\{\ln\frac{g(\mathbf{y})}{f(\mathbf{y})}\right\}\\ &= \mathrm{E}_{\mathbf{y}}[\ln g(\mathbf{y})] - \mathrm{E}_{\mathbf{y}}[\ln f(\mathbf{y})]\\ &= \int_{-\infty}^{\infty}\ln[g(\mathbf{y})]g(\mathbf{y})dy - \int_{-\infty}^{\infty}\ln[f(\mathbf{y})]f(\mathbf{y})dy. \qquad (2.47)\end{aligned}$$

Las propiedades de la medida de información **(KL)** son:

(i) $\mathbf{KL}[g(\mathbf{y})|f(\mathbf{y})] \geq 0$

(ii) $\mathbf{KL}[g(\mathbf{y})|f(\mathbf{y})] = 0 \Longleftrightarrow g(\mathbf{y}) = f(\mathbf{y})$

Luego nosotros diremos que $f \longrightarrow g$ cuando $\mathbf{KL}(g|f) \longrightarrow 0$. Para estimar la medida de información $\mathbf{KL}(g|f)$ dada en (2.47) debemos observar que en la expresión:

$$\mathbf{KL}(g|f) = \mathrm{E}_{\mathbf{y}}[\ln g(\mathbf{y})] - \mathrm{E}_{\mathbf{y}}[\ln f(\mathbf{y})],$$

sólo el segundo término es importante en la evaluación del modelo estadístico $f(\mathbf{y})$, ya que el objetivo final es minimizar la distancia entre $g(\mathbf{y})$ y $f(\mathbf{y})$ y por tanto hacer el primer termino lo mas pequeño posible.
Entonces la medida de información **KL** para la densidad $f(\mathbf{y})$ puede ser aproximada en base al soporte dado en la ecuación (2.3) por la siguiente expresión:

$$\widehat{\mathbf{KL}}(f) = \widehat{\mathrm{E}}_{\mathbf{y}}[\ln f(\mathbf{y})] \approx \frac{1}{n}\sum_{i=1}^{n} \ln[f(y_i|\widehat{\theta})] = \frac{1}{n} L(\widehat{\theta}) \tag{2.48}$$

Por consiguiente $\widehat{\mathbf{KL}}(f)$ puede reemplazar la medida de información **KL** como un criterio para evaluar modelos.
Akaike tomó la medida de información **KL** e hizo las siguientes consideraciones:

(a) El soporte puede usarse para estimar los valores de parámetros. Sin embargo, no puede usarse para comparar modelos diferentes sin algunas correcciones.

(b) Usar $\frac{1}{n}L(\widehat{\theta})$ como una estimación de $\mathrm{E}_{\mathbf{y}}[\ln f(\mathbf{y})]$ produce un prejuicio. La razón para que ocurra tal prejuicio es que los mismos datos son usados para estimar los parámetros y calcular el soporte.

Luego hizo la siguiente estimación imparcial de $\mathrm{E}_{\mathbf{y}}[\ln f(\mathbf{y})]$, el cual sera nuestro factor de corrección en la **Definición (2.4**):

$$C = \mathrm{E}\left\{ \mathrm{E}_{\mathbf{y}}[\ln f(\mathbf{y})] - \frac{1}{n}\sum_{i=1}^{n} \ln[f(y_i|\widehat{\theta})] \right\} \approx \frac{p}{n} \tag{2.49}$$

donde p son la cantidad de parámetros de $f(\mathbf{y})$. Entonces Akaike tomo como criterio de información de la densidad $f(\mathbf{y})$ el producto de los n datos y la función de desviación dada en (2.8) con el soporte igual a la diferencia entre (2.48) y (2.49), es decir:

$$\begin{aligned} \mathbf{AIC} &= n\mathcal{D}(\widehat{\theta}) \\ &= -2n\left[\frac{1}{n}L(\widehat{\theta}) - \frac{p}{n}\right] \\ &= -2[L(\widehat{\theta}) - p] \\ &= -2L(\widehat{\theta}) + 2p \\ &= \mathcal{D}(\widehat{\theta}) + 2p \end{aligned} \tag{2.50}$$

Llamado M_i, con $i = 1, \ldots, N$, a los posibles modelos para nuestros datos, el proceso de seleccionar el mejor modelo se realiza tomando aquel para el cual sea mínimo el criterio **AIC**, es decir:

$$\mathbf{AIC}[f(\mathbf{y})] = \min_{i=1,\ldots,N}[\mathcal{D}(M_i) + 2p_i] \tag{2.51}$$

2.5.2. El criterio de información bayesiano (BIC).

Desde el punto de vista bayesiano, el problema de la selección de modelos es abordado con los mismos principios usados para el contraste de hipótesis. Las estructuras bayesianas usan en las pruebas de hipótesis el factor de Bayes para cuantificar la evidencia de un modelo supuesto contra otro. Luego se hace necesario calcular las probabilidades *a posteriori* de cada modelo para hacer los contrastes.
Gideon Schwarz en (1978) derivó un criterio de información muestral con principios bayesianos, el cual estaba fundamentado en su criterio de información estadístico **SIC** (*Schwarz Information Criterion*). El nuevo criterio lo denomino *Bayesian Information Criterion* **(BIC)**, que no es otra cosa que el criterio **SIC** bajo argumentos bayesianos.
El **BIC** parte de una aproximación a la probabilidad *a posteriori* de un modelo para una muestra grande; esta aproximación es dada por el valor del soporte en su máximo afectado por el número de parámetros mas adecuado en el modelo, o para el caso de clasificación de datos, la elección mas adecuada del número de grupos. Luego el **BIC** es la desviación del modelo aplicada en esta aproximación.
Al minimizar este criterio para todos los posibles modelos se están haciendo comparaciones con diferentes cantidades de parámetros y diferentes cantidades de grupos. En general el valor mas pequeño del **BIC**, es la más fuerte evidencia para ese modelo y ese número de clusters.
Sean $\mathbb{X} = \{x_1, \ldots, x_n\}$ los datos muestrales y $M_1, \ldots, M_m$ los modelos candidatos para los datos observados. Si consideramos los modelos como posibles hipótesis (bajo el enfoque bayesiano), calculamos las probabilidades *a posteriori* para cada modelo así:

$$P(M_j|\mathbb{X}, \theta) = \frac{l(\theta|M_j)}{l(\theta)} P(M_j) \quad j = 1, \ldots, m, \tag{2.52}$$

donde $P(M_j)$ es la probabilidad *a priori* del modelo j y $l(\theta|M_j)$ es igual a la integral respecto a los parámetros θ_j del producto entre la verosimilitud para el modelo j y su probabilidad *a priori*. Esta resultado se conoce comúnmente como *la verosimilitud marginal* de los datos para el modelo j, pues no depende del valor de los parámetros, y viene dada

por:

$$l(\theta|M_j) = \int l_j(\theta_j)p(\theta_j|M_j)d\theta_j.$$

donde $p(\theta_j|M_j)$ es la probabilidad *a priori* para los parámetros del modelo j.
Además $l(\theta)$ es la media ponderada de las verosimilitudes marginales, siendo los coeficientes de ponderación $P(M_j)$, es decir:

$$l(\theta) = \sum_{j=1}^{m} l(\theta_j|M_j)P(M_j)$$

Una manera clásica para escoger un modelo es seleccionar este aumentando al máximo la verosimilitud marginal $l(\theta|M_j)$. Una aproximación asintótica de $l(\theta|M_j)$, válido bajo las condiciones de regularidad, es la propuesta por Schwarz en (1978). Schwarz consideró a la distribución *a posteriori* del vector de parámetros para el modelo j como asintóticamente normal multivariante, es decir:

$$p(\theta_j|\mathbb{X}, M_j) = (2\pi)^{-p_j/2}|\mathbf{S_j}|^{-1/2}\exp\left\{-\frac{1}{2}(\theta_j - \widehat{\theta}_j)'\mathbf{S_j}^{-1}(\theta_j - \widehat{\theta}_j)\right\} \tag{2.53}$$

donde p_j es la dimension del vector de parámetros del modelo M_j, $\widehat{\theta}_j$ es el estimador MV de θ_j y $\mathbf{S_j}$ la matriz de covarianzas de este estimador. Además por el teorema de Bayes tenemos:

$$p(\theta_j|\mathbb{X}, M_j) = \frac{l_j(\theta_j)p(\theta_j|M_j)}{l(\theta|M_j)}$$

entonces,

$$l(\theta|M_j) = \frac{l_j(\theta_j)p(\theta_j|M_j)}{p(\theta_j|\mathbb{X}, M_j)}. \tag{2.54}$$

Tomando logaritmos, particularizando esta expresión para $\widehat{\theta}_j$ y reemplazando la ecuación (2.53) en (2.54) se tiene:

$$\begin{aligned}
\ln\Big[l(\widehat{\theta}_j|M_j)\Big] &= \ln\left[\frac{l_j(\widehat{\theta}_j)p(\widehat{\theta}_j|M_j)}{p(\widehat{\theta}_j|\mathbb{X}, M_j)}\right] \\
L(\widehat{\theta}_j|M_j) &= \ln\Big[l_j(\widehat{\theta}_j)p(\widehat{\theta}_j|M_j)\Big] - \ln\Big[p(\widehat{\theta}_j|\mathbb{X}, M_j)\Big] \\
&= \ln\Big[l_j(\widehat{\theta}_j)\Big] + \ln\Big[p(\widehat{\theta}_j|M_j)\Big] - \ln\Big[p(\widehat{\theta}_j|\mathbb{X}, M_j)\Big] \\
&= L_j(\widehat{\theta}_j) + \ln\Big[p(\widehat{\theta}_j|M_j)\Big] - \left[-\frac{p_j}{2}\ln(2\pi) - \frac{1}{2}\ln|\mathbf{S_j}|\right].
\end{aligned} \tag{2.55}$$

La matriz $\mathbf{S_j}$ tiene habitualmente términos de la forma a/n, luego podemos redefinirla así:

$$\mathbf{S_j} \equiv \frac{1}{n}\mathbf{R_j},$$

entonces $|\mathbf{S_j}| = n^{-p_j}|\mathbf{R_j}|$. Sustituyendo este resultado en (2.55) se tiene:

$$\begin{aligned} L(\widehat{\theta}_j|M_j) &= L_j(\widehat{\theta}_j) + \ln\left[p(\widehat{\theta}_j|M_j)\right] - \left[-\frac{p_j}{2}\ln(2\pi) - \frac{1}{2}\ln\left[n^{-p_j}|\mathbf{R_j}|\right]\right] \\ &= \underbrace{L_j(\widehat{\theta}_j)}_{Primer\ termino} + \underbrace{\ln\left[p(\widehat{\theta}_j|M_j)\right]}_{Segundo\ termino} + \underbrace{\frac{p_j}{2}\ln(2\pi)}_{Tercer\ termino} - \underbrace{\frac{p_j}{2}\ln(n)}_{Cuarto\ termino} + \underbrace{\frac{1}{2}\ln|\mathbf{R_j}|}_{Quinto\ termino}. \end{aligned} \tag{2.56}$$

Para n suficientemente grande, el soporte $L(\widehat{\theta}_j|M_j)$ dado en (2.56) se mantiene casi constante en el segundo termino en relación con la verosimilitud. El tercer termino es de orden constante y el quinto termino por construcción es acotado. Luego para n grande podemos escribir:

$$L(\widehat{\theta}_j|M_j) \simeq L_j(\widehat{\theta}_j) - \frac{p_j}{2}\ln(n). \tag{2.57}$$

Según lo expuesto en la sección 2.2, el método de maxima verosimilitud busca maximizar el soporte de los datos para encontrar los estimadores MV. El criterio **BIC** garantiza que para muestras grandes el soporte a maximizar es asintóticamente aproximado a $L(\widehat{\theta}_j|M_j)$, el cual viene dado (2.57). Por la definición dada en la ecuación (2.8), definimos el criterio **BIC** así:

$$\begin{aligned} \mathbf{BIC} &= -2L(\widehat{\theta}_j|M_j) \\ &= -2[L_j(\widehat{\theta}_j) - \frac{p_j}{2}\ln(n)] \\ &= -2L_j(\widehat{\theta}_j) + p_j\ln(n) \\ &= \mathcal{D}(M_j) + p_j\ln(n). \end{aligned} \tag{2.58}$$

Llamado M_j, con $j = 1, \ldots, m$, a los posibles modelos para nuestros datos, el proceso de seleccionar el mejor modelo se realiza tomando aquel para el cual sea mínimo el criterio **BIC**, es decir:

$$\min_{j=1,\ldots,m} \mathbf{BIC}[M_j] = \min_{j=1,\ldots,m} [\mathcal{D}(M_j) + p_j\ln(n)]. \tag{2.59}$$

Capítulo 3

MEZCLAS DE DISTRIBUCIONES NORMALES P-VARIANTES

3.1. Introducción

En el desarrollo de la estadística moderna las mezclas de distribuciones probabilísticas han sido usadas para modelar problemas de clasificación de datos. Una mezcla de distribuciones esta compuesta de densidades estadísticas provenientes de poblaciones heterogéneas, para las cuales asumimos una función de densidad respectiva; por lo tanto una mezcla finita de distribuciones tendrá un número finito de distribuciones componentes.

Varios modelos de mezclas finitas han venido surgiendo para resolver problemas en diferentes campos de investigación. Por ejemplo, en *Mercadeo* se analiza el gasto realizado en distintos productos en una muestra de consumidores donde es de esperar que hayan grupos (de consumidores) con patrones de gasto distinto; esta clasificación de consumidores se realiza por medio de un modelo de mezclas de distribuciones. En el campo de la *Salud* se observa como en un centro asistencial los pacientes presentan diversos valores en las medidas diagnosticas, lo cual es un indicador de la presencia o ausencia de una enfermedad especifica en cada uno de ellos; luego es de interés en este campo conocer la distribución de las medidas diagnosticas que determinan una enfermedad especifica para luego establecer el tratamiento a seguir en cada grupo de pacientes. De igual forma en estudios de *piscicultura* se analiza una nueva especie de pez tomando las medidas longitudinales en la anatomía externa para así determinar grupos que diferencien sus etapas de crecimiento; un modelo de mezclas de distribuciones nos permite realizar esta clasificación.

Así podemos encontrar ejemplos importantes en Psicología, Biología, Economía y en diferentes campos de ciencia y tecnología. Es importante resaltar que los modelos mas interesantes que han surgido, utilizan la distribución normal (univariada o multivariada) como componente de la mezcla.

3.2. Distribuciones normales mezcladas

3.2.1. Reseña histórica

Los procedimientos de modelación mediante mezclas de distribuciones tuvieron su origen en el trabajo de Francis Galton (1822-1921) sobre mezclas de variables normales y de Karl Pearson (1857-1936) que utilizó por primera vez el método de los momentos para estimarlas.

A principios de 1890, el profesor W. R. Weldon (1860-1906) le consultó al estadístico Karl Pearson acerca de un conjunto de medidas de la razón (frente/longitud de cuerpo) tomadas a una muestra de 1000 cangrejos. Un estudio gráfico de datos mostró que estos estaban sesgados a la derecha. Weldon sugirió que la razón para esta asimetría podría ser que la muestra tenía representantes de dos tipos de cangrejo, pero cuando los datos fueron recolectados no habían sido diferenciados como tal. Esto indujo a Pearson a proponer que la distribución de las medidas podían ser modeladas por la suma de los productos entre la proporción del tipo de cangrejo y su distribución normal, con los dos pesos dados a la proporción de cangrejos de cada tipo. Esto parece ser la primera aplicación de lo que ahora comúnmente se conoce con el terminó de mezcla finita de distribuciones.

En términos matemáticos, la distribución sugerida por Pearson para las medidas en los cangrejos era de la forma

$$f(\mathsf{x}) = p\ N(\mu_1, \sigma_1) + (1-p)\ N(\mu_2, \sigma_2), \tag{3.1}$$

donde p es la proporción del tipo de cangrejo para la cual la razón (frente/longitud del cuerpo) tiene media μ_1 y desviación estándar σ_1, y $(1-p)$ es la proporción del tipo de cangrejo para que los valores correspondientes sean μ_2 y σ_2. En la ecuación (3.1), la función $N(\mu_i, \sigma_i)$, para $i = 1, 2$, es la distribución normal univariate estudiada en un curso básico de estadística.

Para poder ajustar los datos al modelo planteado en (3.1), había que estimar los cinco parámetros $(p, \mu_1, \sigma_1, \mu_2, \sigma_2)$ con los datos suministrados por la muestra de 1000 cangrejos. Pearson en un articulo de 1894 titulado *"Contributions to the mathematical theory*

of evolution" publicado en *Philosophical Transactions A, 185, 71-110*, dio a conocer un método (basado en *el método de los momentos*) que requirió la solución de un polinomio de grado nueve; una tarea computacionalmente exigente en ese momento histórico, lo cual llenó de mas motivos a Pearson para hacerla publica. Así Pearson manejó la tarea heroica de encontrar una solución al ajuste de los datos a dos componentes de distribución normal.

3.2.2. Función de densidad para una mezcla

Definición 3.1. Sea $\mathbf{X}$ una variable aleatoria vectorial definida en una población que contiene G grupos homogéneos. Si su distribución puede ser representada por la función de densidad (o función de probabilidad en el caso discreto) de la forma:

$$\mathbf{G}(\mathbf{X}) = G_1(\mathbf{X}) + \cdots + G_G(\mathbf{X}) = \sum_{g=1}^{G} G_g(\mathbf{X}) \tag{3.2}$$

$$\textit{donde} \quad G_g(\mathbf{X}) = \pi_g f_g(\mathbf{X}), \quad \textit{con} \quad 0 \leq \pi_g \leq 1 \quad \textit{y} \quad \sum_{g=1}^{G} \pi_g = 1 \quad \textit{para} \quad g \in \{1, \ldots, G\} \ ,$$

diremos que $\mathbf{X}$ tiene una **mezcla finita de distribuciones** y que $\mathbf{G}(\cdot)$ es *la función de densidad de la mezcla o mezcla de densidades* de toda la población. Además $G_1(\cdot), \ldots, G_G(\cdot)$ son llamadas *componentes de la mezcla*, los parámetros $\pi_1, \ldots, \pi_G$ son llamados *pesos de las componentes de la mezcla o proporciones de las componentes de la mezcla* y $f_1(\cdot), \ldots, f_G(\cdot)$ *densidades de las componentes de la mezcla.*

De esta definición podemos decir que la mezcla de densidades de una población, que ha sido subdividida de forma finita, viene dada por:

$$\mathbf{G}(\mathbf{X}) = \sum_{g=1}^{G} \pi_g f_g(\mathbf{X}) \tag{3.3}$$

Nótese que el observar un elemento al azar en una población de este estilo puede plantearse en dos etapas : En la primera seleccionamos el estrato o grupo al azar mediante una variable escalar auxiliar i que toma valores de 1 *a* G, cada uno con probabilidades $\pi_1, \ldots, \pi_G$; en la segunda seleccionamos un elemento del estrato $f_g(\mathbf{X})$ con $g = 1, \ldots, G$. Luego la probabilidad de que el elemento seleccionado tome valor x en un evento A de nuestro espacio muestral vendrá dada por la ecuación:

$$P(x \in A) = \sum_{g=1}^{G} P(x \in A |\ i = g) \cdot P(i = g),$$

entonces las densidades componentes serán distribuciones condicionadas al estrato elegido, lo cual podemos escribir como $f_g(\mathbf{X}) = f(\mathbf{X}|\ i = g)$.

No es necesario que las densidades pertenezcan a la misma familia de distribuciones, pero en nuestro caso centraremos la atención en que $f_1(\cdot), \ldots, f_G(\cdot)$ sean G funciones con distribuciones normales p-variantes, por lo tanto lo que aquí se estudia son combinaciones de distribuciones normales.

Así la mezcla de densidades de la variable aleatoria unidimensional $\mathbf{x}$ vendrá dada por:

$$\mathbf{G}(\mathbf{x}) = \pi_1 f_1(\mathbf{x}; \mu_1, \sigma_1) + \ldots + \pi_G f_G(\mathbf{x}; \mu_G, \sigma_G) \tag{3.4}$$

donde $f_i(\mathbf{x}; \mu_i, \sigma_i)$, para $i = 1, \ldots, G$, denota la densidad normal $N(\mu_i, \sigma_i)$.
En la figura 3.1 observamos un ejemplo de una mezcla normal unidimensional, compuesta por cuatro distribuciones normales con pesos: $\pi_1 = 0.3; \pi_2 = 0.25; \pi_3 = 0.1$ *y* $\pi_4 = 0.35$.

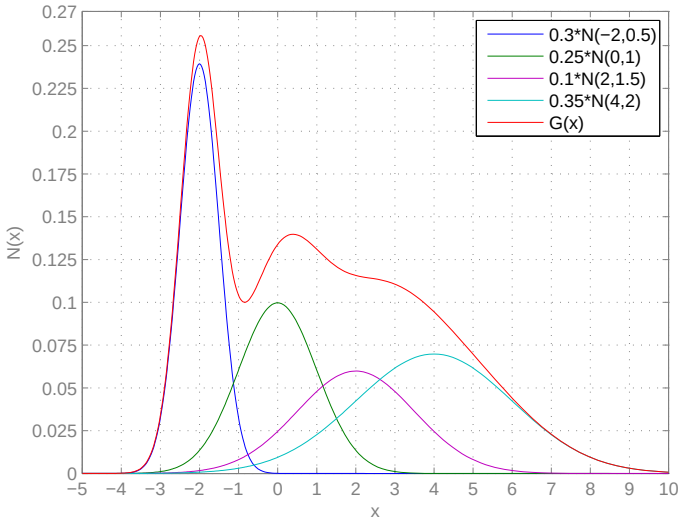


Figura 3.1: Mezcla de 4 distribuciones normales unidimensionales con medias en $-2, 0, 2, 4$, desviaciones $0.5; 1; 1.5; 2$ y pesos $0.3; 0.25; 0.1; 0.35$ respectivamente.

Para problemas de mayor dimensión ($p > 1$), asumimos en la mezcla la variable aleatoria vectorial $\mathbf{X} = (X_1, \ldots, X_p)'$ con densidades componentes $N_p(\mu_g, \mathbf{V}_g)$ que son distribuciones normales multivariantes con vector de medias μ_g y matrices de varianzas y covarianzas $\mathbf{V}_g$, donde $g = 1, \ldots, G$. En el caso multivariante, las mezclas de distribuciones normales se ven representadas en una amplia gama de distribuciones. En la figura 3.2 se presenta un ejemplo de estas.

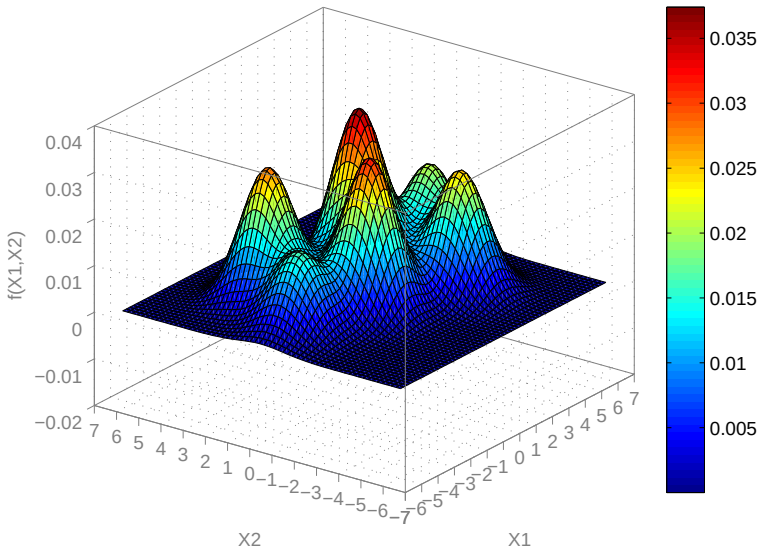


Figura 3.2: Mezcla de seis distribuciones normales bivariantes con matrices de covarianzas $\mathbf{I}$, medias en $(-4,0);(-2.5,2.5);(-1,-1);(1,1);(2.5,-2.5);(4,0)$ y pesos $0.1;0.18;0.2;0.23;0.16;0.13$ respectivamente.

3.3. Parámetros en la mezcla de densidades normales

Los parámetros de la mezcla de densidades normales p-variantes vienen dados por la pareja $(\mu(\mathbf{G}), \mathbf{V}(\mathbf{G}))$, donde llamaremos $\mu(\mathbf{G})$ al vector de medias de la mezcla y $\mathbf{V}(\mathbf{G})$ a la matriz de varianzas y covarianzas de la distribución mezclada; estos se obtienen fácilmente conocidas los vectores de medias μ_g y las matrices de varianzas $\mathbf{V}_g$, con $g = 1, \ldots, G$, de la densidades componentes de las mezcla.

3.3.1. Vector de medias

Teorema 3.1. *La media de una distribución mezclada con densidades componentes normales viene dada por*

$$\mu(\mathbf{G}) = \sum_{g=1}^{G} \pi_g \mu_g \tag{3.5}$$

donde π_g y μ_g representan el peso y la media, respectivamente, de la componente g-esima para $g \in \{1, \ldots, G\}$.

Demostración. Introduzcamos la variable escalar auxiliar de clasificación que nombramos anteriormente, i, la cual toma valores de 1 a G con probabilidades $\pi_1, \ldots, \pi_G$ y tomemos la esperanza de la variable aleatoria $\mathbf{X}$ condicionada a la probabilidad de aparición en i, es decir:

$$\mathrm{E}(\mathbf{X}|i = g) = \mu_g \equiv \mathrm{E}_{\mathbf{X}|i}(\mathbf{X}).$$

Aplicando la propiedad de la esperanza condicionada dada en la ecuación (1.34), se tiene que:

$$\mu(\mathbf{G}) = \mathrm{E}(\mathbf{X}) = \mathrm{E}_i[\mathrm{E}_{\mathbf{X}|i}(\mathbf{X})] = \mathrm{E}_i[\mu_g].$$

Ahora, como el promedio de las esperanzas condicionadas para los i grupos es el promedio ponderado de todas las medias y $\sum_{g=1}^{G} \pi_g = 1$, se concluye que:

$$\mu(\mathbf{G}) = \mathrm{E}_i[\mu_g] = \frac{\sum_{g=1}^{G} \pi_g \mu_g}{\sum_{g=1}^{G} \pi_g} = \sum_{g=1}^{G} \pi_g \mu_g$$

■

Luego $\mu(\mathbf{G})$ es el vector de medias p-dimensional de la distribución mezclada, donde una componente de este vector $\mu(G)_j$, con $j = 1, \ldots, p$, es la media ponderada de la mezcla en la j-esima variable; esta se obtiene sumando, para los G-grupos, los productos entre las proporciones de mezcla π_g (con $g = 1, \ldots, G$) y las j-esimas componentes del vector de medias de cada grupo μ_{jg}, es decir:

$$\mu(\mathbf{G}) = \begin{pmatrix} \mu(G)_1 \\ \vdots \\ \mu(G)_p \end{pmatrix} = \sum_{g=1}^{G} \pi_g \begin{pmatrix} \mu_{1g} \\ \vdots \\ \mu_{pg} \end{pmatrix} = \sum_{g=1}^{G} \begin{pmatrix} \pi_g \ \mu_{1g} \\ \vdots \\ \pi_g \ \mu_{pg} \end{pmatrix} = \begin{pmatrix} \sum_{g=1}^{G} \pi_g \ \mu_{1g} \\ \vdots \\ \sum_{g=1}^{G} \pi_g \ \mu_{pg} \end{pmatrix} \tag{3.6}$$

3.3.2. Matriz de varianzas y covarianzas

Teorema 3.2. *La matriz de varianzas y covarianzas de la distribución mezclada con densidades componentes normales viene dada por*

$$\mathbf{V}(\mathbf{G}) = \sum_{g=1}^{G} \pi_g \mathbf{V}_g + \sum_{g=1}^{G} \pi_g (\mu_g - \mu(\mathbf{G}))(\mu_g - \mu(\mathbf{G}))' \tag{3.7}$$

Demostración. Por definición de la matriz de varianzas y covarianzas e introduciendo el termino $(-\mu_g + \mu_g)$ se tiene:

$$\mathbf{V}(\mathbf{G}) = \mathrm{E}\{[\mathbf{X} - \mu(\mathbf{G})][\mathbf{X} - \mu(\mathbf{G})]'\} = \mathrm{E}\{[(\mathbf{X} - \mu_g) + (\mu_g - \mu(\mathbf{G}))][(\mathbf{X} - \mu_g) + (\mu_g - \mu(\mathbf{G}))]'\},$$

la cual, según la propiedad de la esperanza condicionada dada en la ecuación (1.34), puede ser escrita como:

$$\mathrm{E}\{[(\mathbf{X}-\mu_g)+(\mu_g-\mu(\mathbf{G}))][(\mathbf{X}-\mu_g)+(\mu_g-\mu(\mathbf{G}))]'\} = \mathrm{E}_i\{\underbrace{\mathrm{E}_{\mathbf{X}|i}\{[(\mathbf{X}-\mu_g)+(\mu_g-\mu(\mathbf{G}))][(\mathbf{X}-\mu_g)+(\mu_g-\mu(\mathbf{G}))]'\}}_{(a)}\}, \quad (3.8)$$

donde i es la variable escalar auxiliar introducida en el proceso de clasificación, asumiendo valores de 1 a G.
Resolviendo la matriz planteada por la esperanza condicionada (a) se tiene:

$$\mathrm{E}_{\mathbf{X}|i}\{[(\mathbf{X}-\mu_g)+(\mu_g-\mu(\mathbf{G}))][(\mathbf{X}-\mu_g)+(\mu_g-\mu(\mathbf{G}))]'\} =$$

$$\begin{pmatrix} \mathrm{E}_{\mathbf{X}|i}[(X_1-\mu_{1g})+(\mu_{1g}-\mu(G)_1)]^2 & \cdots & \mathrm{E}_{\mathbf{X}|i}[(X_1-\mu_{1g})+(\mu_{1g}-\mu(G)_1)][(X_p-\mu_{pg})+(\mu_{pg}-\mu(G)_p)] \\ \vdots & \ddots & \vdots \\ \mathrm{E}_{\mathbf{X}|i}[(X_p-\mu_{pg})+(\mu_{pg}-\mu(G)_p)][(X_1-\mu_{1g})+(\mu_{1g}-\mu(G)_1)] & \cdots & \mathrm{E}_{\mathbf{X}|i}[(X_p-\mu_{pg})+(\mu_{pg}-\mu(G)_p)]^2 \end{pmatrix},$$

luego en la posición (k,l) de esta matriz, con $k,l \in \{1,\ldots,p\}$, está el elemento

$$\begin{aligned}\mathrm{E}_{\mathbf{X}|i}\{[(X_k-\mu_{kg})+(\mu_{kg}-\mu(G)_k)][(X_l-\mu_{lg})+(\mu_{lg}-\mu(G)_l)]\} = \\ = \mathrm{E}_{\mathbf{X}|i}\{[(X_k-\mu_{kg})(X_l-\mu_{lg})]\} + \overbrace{\mathrm{E}_{\mathbf{X}|i}\{[(X_k-\mu_{kg})(\mu_{lg}-\mu(G)_l)]\}}^{(b)} \\ + \underbrace{\mathrm{E}_{\mathbf{X}|i}\{[(X_l-\mu_{lg})(\mu_{kg}-\mu(G)_k)]\}}_{(c)} + \mathrm{E}_{\mathbf{X}|i}\{[(\mu_{kg}-\mu(G)_k)(\mu_{lg}-\mu(G)_l)]\} \quad (3.9)\end{aligned}$$

donde $\mathrm{E}_{\mathbf{X}|i}\{[(\mu_{kg}-\mu(G)_k)(\mu_{lg}-\mu(G)_l)]\} = (\mu_{kg}-\mu(G)_k)(\mu_{lg}-\mu(G)_l)$, ya que ambos términos son constantes. Además (b) y (c) son iguales a cero, pues en (b) se tiene:

$$\begin{aligned}\mathrm{E}_{\mathbf{X}|i}\{[(X_k-\mu_{kg})(\mu_{lg}-\mu(G)_1)]\} &= \mu_{lg}.\mathrm{E}_{\mathbf{X}|i}\{X_k\} - \mu(G)_1.\mathrm{E}_{\mathbf{X}|i}\{X_k\} - \mu_{kg}.\mu_{lg} + \mu_{kg}.\mu(G)_1 \\ &= \mu_{lg}.\mu_{kg} - \mu(G)_1.\mu_{kg} - \mu_{kg}.\mu_{lg} + \mu_{kg}.\mu(G)_1 \\ &= 0.\end{aligned}$$

De igual forma se demuestra que (c) es cero.

$$\mathrm{E}_{\mathbf{X}|i}\{[(X_l-\mu_{lg})(\mu_{kg}-\mu(G)_k)]\} = \mu_{kg}.\mathrm{E}_{\mathbf{X}|i}\{X_l\} - \mu(G)_k.\mathrm{E}_{\mathbf{X}|i}\{X_l\} - \mu_{lg}.\mu_{kg} + \mu_{lg}.\mu(G)_k$$

$$= \mu_{kg}.\mu_{lg} - \mu(G)_k.\mu_{lg} - \mu_{lg}.\mu_{kg} + \mu_{lg}.\mu(G)_k = 0.$$

Retornando a (3.9), se observa que un elemento (k, l) de la matriz planteada en (a) es de la forma:

$$\mathrm{E}_{\mathbf{X}|i}\{[(X_k - \mu_{kg}) + (\mu_{kg} - \mu(G)_k)][(X_l - \mu_{lg}) + (\mu_{lg} - \mu(G)_1)]\} =$$
$$\mathrm{E}_{\mathbf{X}|i}\{[(X_k - \mu_{kg})(X_l - \mu_{lg})]\} + (\mu_{kg} - \mu(G)_k)(\mu_{lg} - \mu(G)_1),$$

lo que sugiere descomponer la matriz (a) como la suma de otras dos, donde el primer miembro es precisamente la matriz de varianzas y covarianzas en el g-esimo grupo, es decir:

$$\mathrm{E}_{\mathbf{X}|i}\{[(\mathbf{X} - \mu_g) + (\mu_g - \mu(\mathbf{G}))][(\mathbf{X} - \mu_g) + (\mu_g - \mu(\mathbf{G}))]'\} =$$
$$\mathrm{E}_{\mathbf{X}|i}\{(\mathbf{X} - \mu_g)(\mathbf{X} - \mu_g)'\} + (\mu_g - \mu(\mathbf{G}))(\mu_g - \mu(\mathbf{G}))' = \mathbf{V}_g + (\mu_g - \mu(\mathbf{G}))(\mu_g - \mu(\mathbf{G}))'.$$

Reemplazando este resultado en (3.8) tenemos que:

$$\mathrm{E}\{[(\mathbf{X} - \mu_g) + (\mu_g - \mu(\mathbf{G}))][(\mathbf{X} - \mu_g) + (\mu_g - \mu(\mathbf{G}))]'\} =$$
$$\mathrm{E}_i\{\mathrm{E}_{\mathbf{X}|i}\{[(\mathbf{X} - \mu_g) + (\mu_g - \mu(\mathbf{G}))][(\mathbf{X} - \mu_g) + (\mu_g - \mu(\mathbf{G}))]'\}\} =$$
$$\mathrm{E}_i\{\mathbf{V}_g + (\mu_g - \mu(\mathbf{G}))(\mu_g - \mu(\mathbf{G}))'\}.$$

Puesto que $\mathrm{E}_i(\cdot)$ es el promedio ponderado para los pesos de la mezcla asignados a cada grupo y $\sum_{g=1}^{G} \pi_g = 1$ se tiene que:

$$\begin{aligned}
\mathbf{V}(\mathbf{G}) &= \mathrm{E}\{[(\mathbf{X} - \mu_g) + (\mu_g - \mu(\mathbf{G}))][(\mathbf{X} - \mu_g) + (\mu_g - \mu(\mathbf{G}))]'\} \\
&= \mathrm{E}_i\{\mathbf{V}_g + (\mu_g - \mu(\mathbf{G}))(\mu_g - \mu(\mathbf{G}))'\} \\
&= \sum_{g=1}^{G} \pi_g\{\mathbf{V}_g + (\mu_g - \mu(\mathbf{G}))(\mu_g - \mu(\mathbf{G}))'\} / \sum_{g=1}^{G} \pi_g \\
&= \sum_{g=1}^{G} \pi_g\{\mathbf{V}_g + (\mu_g - \mu(\mathbf{G}))(\mu_g - \mu(\mathbf{G}))'\} \\
&= \sum_{g=1}^{G} \pi_g \mathbf{V}_g + \sum_{g=1}^{G} \pi_g(\mu_g - \mu(\mathbf{G}))(\mu_g - \mu(\mathbf{G}))'
\end{aligned}$$

■

Luego la variabilidad total, que es la matriz de varianzas y covarianzas $\mathbf{V}(\mathbf{G})$, se descompone en una variabilidad explicada, $\sum_{g=1}^{G} \pi_g(\mu_g - \mu(\mathbf{G}))(\mu_g - \mu(\mathbf{G}))'$ que tiene en cuenta las diferencias entre medias de las densidades de las componentes μ_g y el vector de medias $\mu(\mathbf{G})$, y una variabilidad no explicada, $\sum_{g=1}^{G} \pi_g \mathbf{V}_g$, que es la variabilidad con respecto a las componentes.

3.4. Estimación MV de parámetros en las componentes de la mezcla

Como se puede apreciar en la sección anterior, los parámetros de una mezcla de densidades normales quedan totalmente determinados si conocemos los parámetros de cada una de las componentes que forman la mezcla; esto es para nuestro caso conocer la terna $(\pi_g, \mu_g, \mathbf{V}_g)$, que llamaremos θ_g, ya que esta formada por los parámetros del grupo g para $g = 1, \ldots, G$. Difícilmente en la practica contamos con esta información, por lo que se hace necesario estimar los valores del peso π_g, el vector de medias μ_g y la matriz de varianzas y covarianzas $\mathbf{V}_g$, para $g = 1, \ldots, G$. Estimaremos estos parámetros por el método de maxima verosimilitud (MV). Veamos ahora algunos conceptos básicos.

3.4.1. Verosimilitud, soporte y score para mezclas

Definición 3.2. Sea $\mathbf{x}_1, \ldots, \mathbf{x}_n$ una muestra de datos independientes que pueden estratificarse en G grupos, de manera que existe n_1 observaciones del grupo 1, n_2 observaciones del grupo $2, \ldots, n_G$ del grupo G. El vector $\mathbf{x}_i = (x_{i1}, \ldots, x_{ip})'$ representa un individuo particular para $i = 1, \ldots, n$.
Definimos **la función de verosimilitud para la mezcla**, $l\big(\theta\big)$, como:

$$l\big(\theta\big) = \prod_{i=1}^{n} \mathbf{G}(\mathbf{x}_i|\theta\big) = \prod_{i=1}^{n} \left(\sum_{g=1}^{G} \pi_g f_g(\mathbf{x}_i) \right), \tag{3.10}$$

donde $\theta = (\theta_1, \ldots, \theta_G)$ es el vector de parámetros para los G grupos, $\mathbf{G}(\mathbf{x}_i|\theta\big)$ es valor de la mezcla de densidades para el i-esimo individuo dado el vector de parámetros θ, π_g es el peso de la mezcla para el g-esimo grupo y $f_g(\mathbf{x}_i)$ es el valor de la densidad en el g-esimo grupo para el i-esimo individuo.

Nótese que $l\big(\theta\big)$ puede escribirse como la suma de G^n términos, que corresponden a todas

las posibles clasificaciones de las n observaciones entre los G grupos, esto es:

$$l(\theta) = \prod_{i=1}^{n}\left(\sum_{g=1}^{G}\pi_g f_g(\mathbf{x}_i)\right) = \overbrace{\underbrace{\left[\sum_{g=1}^{G}\pi_g f_g(\mathbf{x}_1)\right]}_{G-sumandos}\underbrace{\left[\sum_{g=1}^{G}\pi_g f_g(\mathbf{x}_2)\right]}_{G-sumandos}\ldots\underbrace{\left[\sum_{g=1}^{G}\pi_g f_g(\mathbf{x}_n)\right]}_{G-sumandos}}^{n-datos}.$$

Definición 3.3. Sea $l(\theta)$ la función de verosimilitud de una mezcla de G densidades, con $\theta = (\theta_1, \ldots, \theta_G)'$. **La función soporte para la mezcla** se define como:

$$L(\theta) = \sum_{i=1}^{n}\ln\Big(\mathbf{G}(\mathbf{x}_i|\theta)\Big) = \sum_{i=1}^{n}\ln\left(\sum_{g=1}^{G}\pi_g f_g(\mathbf{x}_i)\right). \tag{3.11}$$

Esta definición es la particularización de la definición de función soporte dada en la ecuación (2.3), puesto que:

$$L(\theta) = \ln\big[l(\theta)\big] = \ln\left[\prod_{i=1}^{n}\left(\sum_{g=1}^{G}\pi_g f_g(\mathbf{x}_i)\right)\right] = \sum_{i=1}^{n}\ln\left(\sum_{g=1}^{G}\pi_g f_g(\mathbf{x}_i)\right).$$

Definición 3.4. Sea $L(\theta)$ la función soporte de una mezcla de G densidades, con $\theta = (\theta_1, \ldots, \theta_G)'$. **La función score para la mezcla** se define como:

$$Z(\theta) = \sum_{i=1}^{n}\frac{\frac{\partial}{\partial\theta}\big[\mathbf{G}\big(\mathbf{x}_i|\theta\big)\big]}{\mathbf{G}\big(\mathbf{x}_i|\theta\big)} = \sum_{i=1}^{n}\frac{\frac{\partial}{\partial\theta}\Big[\sum_{g=1}^{G}\pi_g f_g(\mathbf{x}_i)\Big]}{\sum_{g=1}^{G}\pi_g f_g(\mathbf{x}_i)}. \tag{3.12}$$

De igual forma que se planteó la función de soporte para la mezcla, esta definición es consecuencia de las ecuaciones (2.4) y (3.11), ya que:

$$\begin{aligned}
Z(\theta) &= \frac{\partial}{\partial\theta}L(\theta)\\
&= \frac{\partial}{\partial\theta}\left\{\sum_{i=1}^{n}\ln\left(\sum_{g=1}^{G}\pi_g f_g(\mathbf{x}_i)\right)\right\}\\
&= \sum_{i=1}^{n}\left\{\frac{\partial}{\partial\theta}\ln\left(\sum_{g=1}^{G}\pi_g f_g(\mathbf{x}_i)\right)\right\}\\
&= \sum_{i=1}^{n}\left\{\left(\frac{1}{\sum_{g=1}^{G}\pi_g f_g(\mathbf{x}_i)}\right)\frac{\partial}{\partial\theta}\Big[\sum_{g=1}^{G}\pi_g f_g(\mathbf{x}_i)\Big]\right\}
\end{aligned}$$

$$= \sum_{i=1}^{n} \frac{\frac{\partial}{\partial\theta}\left[\sum_{g=1}^{G} \pi_g f_g(\mathbf{x}_i)\right]}{\sum_{g=1}^{G} \pi_g f_g(\mathbf{x}_i)}.$$

Obsérvese que la función score es un campo vectorial; luego para una mezcla de G densidades tenemos:

$$Z(\theta) = \begin{bmatrix} \sum_{i=1}^{n} \frac{\frac{\partial}{\partial\theta_1}\left[\sum_{g=1}^{G} \pi_g f_g(\mathbf{x_i})\right]}{\sum_{g=1}^{G} \pi_g f_g(\mathbf{x_i})} \\ \vdots \\ \sum_{i=1}^{n} \frac{\frac{\partial}{\partial\theta_G}\left[\sum_{g=1}^{G} \pi_g f_g(\mathbf{x_i})\right]}{\sum_{g=1}^{G} \pi_g f_g(\mathbf{x_i})} \end{bmatrix},$$

donde la componente vectorial

$$\sum_{i=1}^{n} \frac{\frac{\partial}{\partial\theta_g}\left[\sum_{g=1}^{G} \pi_g f_g(\mathbf{x_i})\right]}{\sum_{g=1}^{G} \pi_g f_g(\mathbf{x_i})}, \tag{3.13}$$

para $g = 1, \ldots, G$ es nuevamente un vector con dimension determinada por θ_g. Recuérdese que para nosotros $\theta_g = (\pi_g, \mu_g, \mathbf{V}_g)$.

3.4.2. Ecuaciones MV para mezclas de densidades normales

En nuestro caso, asumimos cada $f_g(\mathbf{X})$ como una densidad normal p-variante con vector de medias μ_g y matriz de varianzas $\mathbf{V}_g$; de esta forma $\theta = (\pi_1, \ldots, \pi_G; \mu_1, \ldots, \mu_G; \mathbf{V}_1, \ldots, \mathbf{V}_G)$. Luego nuestra tarea ahora es maximizar la función soporte dada en (3.11), que equivale a encontrar solución al sistema homogéneo que se produce en la función score para mezclas dado en (3.12).

Para evitar inconvenientes en la maximización supondremos que el orden de las G distribuciones estará determinado por $\pi_1 \geq \pi_2 \geq \ldots \geq \pi_G$, pues el orden $1, \ldots, G$ dado anteriormente es arbitrario. También supondremos que como mínimo hay p-observaciones en cada distribución y trataremos de encontrar un máximo local que proporcione un estimador consistente de los parámetros. Luego la función score para mezclas nos proporciona el siguiente sistema homogéneo para el g-esimo grupo:

$$\frac{\partial L(\theta)}{\partial \pi_g} = 0 \tag{3.14}$$

$$\frac{\partial L(\theta)}{\partial \mu_g} = 0 \tag{3.15}$$

$$\frac{\partial L(\theta)}{\partial \mathbf{V}_g} = 0 \tag{3.16}$$

Resolvamos este sistema:

1. En la primera ecuación dada por (3.14), debemos tener en cuenta que $\sum_{g=1}^{G} \pi_g = 1$; luego maximizar $L(\theta)$ bajo esta restricción nos obliga a introducir en ella un multiplicador de Lagrange. Entonces la función a maximizar será:

$$L_\lambda(\theta) = \sum_{i=1}^{n} \ln\left(\sum_{g=1}^{G} \pi_g f_g(\mathbf{x_i})\right) - \lambda\left(\sum_{g=1}^{G} \pi_g - 1\right). \tag{3.17}$$

Derivando respecto a las probabilidades:

$$\frac{\partial L_\lambda(\theta)}{\partial \pi_g} = \frac{\partial}{\partial \pi_g}\left[\sum_{i=1}^{n} \ln\left(\sum_{g=1}^{G} \pi_g f_g(\mathbf{x_i})\right)\right] - \frac{\partial}{\partial \pi_g}\left[\lambda\left(\sum_{g=1}^{G} \pi_g - 1\right)\right] = 0$$

$$\Longrightarrow \sum_{i=1}^{n}\left[\frac{\partial}{\partial \pi_g} \ln\left(\sum_{g=1}^{G} \pi_g f_g(\mathbf{x_i})\right)\right] - \lambda\left[\frac{\partial}{\partial \pi_g}\left(\sum_{g=1}^{G} \pi_g\right)\right] = 0$$

$$\Longrightarrow \sum_{i=1}^{n}\left[\frac{\partial}{\partial \pi_g}\left(\sum_{g=1}^{G} \pi_g f_g(\mathbf{x_i})\right) \Big/ \sum_{g=1}^{G} \pi_g f_g(\mathbf{x_i})\right] - \lambda = 0$$

$$\Longrightarrow \sum_{i=1}^{n}\left[f_g(\mathbf{x_i}) \Big/ \sum_{g=1}^{G} \pi_g f_g(\mathbf{x_i})\right] - \lambda = 0.$$

Multiplicando por π_g, con $\pi_g \neq 0$, tenemos:

$$\frac{\partial L_\lambda(\theta)}{\partial \pi_g} = \sum_{i=1}^{n} \frac{\pi_g f_g(\mathbf{x_i})}{\sum_{g=1}^{G} \pi_g f_g(\mathbf{x_i})} - \pi_g \lambda = 0,$$

luego

$$\pi_g \lambda = \sum_{i=1}^{n} \frac{\pi_g f_g(\mathbf{x_i})}{\sum_{g=1}^{G} \pi_g f_g(\mathbf{x_i})}$$

$$\pi_g \lambda = \sum_{i=1}^{n} \pi_{ig} \tag{3.18}$$

donde hemos llamado π_{ig} a:

$$\boxed{\pi_{ig} = \frac{\pi_g f_g(\mathbf{x_i})}{\sum_{g=1}^{G} \pi_g f_g(\mathbf{x_i})},} \tag{3.19}$$

el cual se traduce como la probabilidad *a posteriori* de que la observación i halla sido generada por la población g. Es claro que para cada dato se cumplirá: $\sum_{g=1}^{G} \pi_{ig} = 1$. Así el valor de λ se obtiene, sumando en (3.18) para todos los grupos, de la siguiente forma:

$$\begin{aligned}
\pi_g \lambda &= \sum_{i=1}^{n} \pi_{ig} \\
\sum_{g=1}^{G} \pi_g \lambda &= \sum_{g=1}^{G} \sum_{i=1}^{n} \pi_{ig} \\
\lambda \sum_{g=1}^{G} \pi_g &= \sum_{i=1}^{n} \left(\sum_{g=1}^{G} \pi_{ig} \right) \\
\lambda &= \sum_{i=1}^{n} 1 \\
\lambda &= n.
\end{aligned} \tag{3.20}$$

Sustituyendo este valor en (3.18) se tiene la probabilidad *a priori* de pertenecer al g-esimo grupo así:

$$\boxed{\widehat{\pi}_g = \frac{1}{n} \sum_{i=1}^{n} \pi_{ig}.} \tag{3.21}$$

2. Para la ecuación dada en (3.15) no hay la clase de restricciones que se plantearon en el anterior numeral, luego según la **definición 3.4** y la ecuación (3.13) tenemos:

$$\begin{aligned}
\frac{\partial L(\theta)}{\partial \mu_g} &= \sum_{i=1}^{n} \frac{\frac{\partial}{\partial \mu_g} \left[\sum_{g=1}^{G} \pi_g f_g(\mathbf{x}_i) \right]}{\sum_{g=1}^{G} \pi_g f_g(\mathbf{x}_i)} = 0 \\
&= \sum_{i=1}^{n} \frac{\frac{\partial}{\partial \mu_g} \left[\pi_g f_g(\mathbf{x}_i) \right]}{\sum_{g=1}^{G} \pi_g f_g(\mathbf{x}_i)} = 0 \\
&= \sum_{i=1}^{n} \frac{\pi_g \left[\frac{\partial}{\partial \mu_g} f_g(\mathbf{x}_i) \right]}{\sum_{g=1}^{G} \pi_g f_g(\mathbf{x}_i)} = 0.
\end{aligned} \tag{3.22}$$

Puesto que:

$$f_g(\mathbf{x}_i) = |\mathbf{V}_g|^{-1/2}(2\pi)^{-p/2}\exp\left\{-\frac{1}{2}(\mathbf{x}_i-\mu_g)'\mathbf{V}_g^{-1}(\mathbf{x}_i-\mu_g)\right\},$$

se tiene, según el resultado dado en (1.13), que:

$$\begin{aligned}
\frac{\partial}{\partial\mu_g}\big[f_g(\mathbf{x}_i)\big] &= \frac{\partial}{\partial\mu_g}\left\{|\mathbf{V}_g|^{-1/2}(2\pi)^{-p/2}\exp\left\{-\frac{1}{2}(\mathbf{x}_i-\mu_g)'\mathbf{V}_g^{-1}(\mathbf{x}_i-\mu_g)\right\}\right\}\\
&= |\mathbf{V}_g|^{-1/2}(2\pi)^{-p/2}\left\{\frac{\partial}{\partial\mu_g}\exp\underbrace{\left\{-\frac{1}{2}(\mathbf{x}_i-\mu_g)'\mathbf{V}_g^{-1}(\mathbf{x}_i-\mu_g)\right\}}_{U}\right\}\\
&= \underbrace{|\mathbf{V}_g|^{-1/2}(2\pi)^{-p/2}\exp\{U\}}_{f_g(\mathbf{x}_i)}\left\{\frac{\partial U}{\partial\mu_g}\right\}\\
&= f_g(\mathbf{x}_i)\left\{-\frac{1}{2}\left[(2)\mathbf{V}_g^{-1}(\mathbf{x}_i-\mu_g)\right](-1)\right\}\\
&= f_g(\mathbf{x}_i)\left[\mathbf{V}_g^{-1}(\mathbf{x}_i-\mu_g)\right]. \qquad (3.23)
\end{aligned}$$

Reemplazando (3.23) en (3.22) y sustituyendo el resultado dado en (3.19) tenemos:

$$\begin{aligned}
\sum_{i=1}^{n}\frac{\pi_g\Big[f_g(\mathbf{x}_i)\mathbf{V}_g^{-1}(\mathbf{x}_i-\mu_g)\Big]}{\sum_{g=1}^{G}\pi_g f_g(\mathbf{x}_i)} &= 0\\
\sum_{i=1}^{n}\frac{\pi_g f_g(\mathbf{x}_i)}{\sum_{g=1}^{G}\pi_g f_g(\mathbf{x}_i)}\mathbf{V}_g^{-1}(\mathbf{x}_i-\mu_g) &= 0\\
\sum_{i=1}^{n}\pi_{ig}\mathbf{V}_g^{-1}(\mathbf{x}_i-\mu_g) &= 0\\
\sum_{i=1}^{n}\frac{\pi_{ig}(\mathbf{x}_i-\mu_g)}{\mathbf{V}_g} &= 0\\
\sum_{i=1}^{n}\pi_{ig}(\mathbf{x}_i-\mu_g) &= 0\\
\sum_{i=1}^{n}\pi_{ig}\mathbf{x}_i - \sum_{i=1}^{n}\pi_{ig}\mu_g &= 0,
\end{aligned}$$

entonces:

$$\sum_{i=1}^{n} \pi_{ig}\mu_g = \sum_{i=1}^{n} \pi_{ig}\mathbf{x}_i$$

$$\boxed{\widehat{\mu}_g = \frac{1}{\sum_{i=1}^{n} \pi_{ig}} \sum_{i=1}^{n} \pi_{ig}\mathbf{x}_i} \tag{3.24}$$

3. La tercera ecuación del sistema homogéneo, es decir la ecuación (3.16), se plantea según las ecuaciones (3.12) y (3.13) como:

$$\begin{aligned}
\frac{\partial L(\theta)}{\partial \mathbf{V}_g} &= \sum_{i=1}^{n} \frac{\frac{\partial}{\partial \mathbf{V}_g}\left[\sum_{g=1}^{G} \pi_g f_g(\mathbf{x}_i)\right]}{\sum_{g=1}^{G} \pi_g f_g(\mathbf{x}_i)} = 0 \\
&= \sum_{i=1}^{n} \frac{\frac{\partial}{\partial \mathbf{V}_g}\left[\pi_g f_g(\mathbf{x}_i)\right]}{\sum_{g=1}^{G} \pi_g f_g(\mathbf{x}_i)} = 0 \\
&= \sum_{i=1}^{n} \frac{\pi_g\left[\frac{\partial}{\partial \mathbf{V}_g} f_g(\mathbf{x}_i)\right]}{\sum_{g=1}^{G} \pi_g f_g(\mathbf{x}_i)} = 0.
\end{aligned} \tag{3.25}$$

Puesto que

$$f_g(\mathbf{x}_i) = |\mathbf{V}_g|^{-1/2}(2\pi)^{-p/2} \exp\left\{-\frac{1}{2}(\mathbf{x}_i - \mu_g)'\mathbf{V}_g^{-1}(\mathbf{x}_i - \mu_g)\right\},$$

se tiene:

$$\begin{aligned}
\frac{\partial}{\partial \mathbf{V}_g}\big[f_g(\mathbf{x}_i)\big] &= \frac{\partial}{\partial \mathbf{V}_g}\left\{|\mathbf{V}_g|^{-1/2}(2\pi)^{-p/2} \exp\left\{-\frac{1}{2}(\mathbf{x}_i - \mu_g)'\mathbf{V}_g^{-1}(\mathbf{x}_i - \mu_g)\right\}\right\} \\
&= (2\pi)^{-p/2}\frac{\partial}{\partial \mathbf{V}_g}\left\{\overbrace{|\mathbf{V}_g|^{-1/2}}^{U(\mathbf{V}_g)} \underbrace{\exp\left\{-\frac{1}{2}(\mathbf{x}_i - \mu_g)'\mathbf{V}_g^{-1}(\mathbf{x}_i - \mu_g)\right\}}_{W(\mathbf{V}_g)}\right\} \\
&= (2\pi)^{-p/2}\frac{\partial}{\partial \mathbf{V}_g}\left\{\big[U(\mathbf{V}_g)\big]\big[W(\mathbf{V}_g)\big]\right\} \\
&= (2\pi)^{-p/2}\left\{\big[U(\mathbf{V}_g)\big]\frac{\partial\big[W(\mathbf{V}_g)\big]}{\partial \mathbf{V}_g} + \big[W(\mathbf{V}_g)\big]\frac{\partial\big[U(\mathbf{V}_g)\big]}{\partial \mathbf{V}_g}\right\}
\end{aligned} \tag{3.26}$$

Encontremos las derivadas planteadas en 3.26.

- Según la propiedad a) de las derivadas matriciales vistas en la sección 1.3.4, tenemos que:

$$\begin{aligned}\frac{\partial}{\partial \mathbf{V}_g}\big[U(\mathbf{V}_g)\big] &= \frac{\partial\big[|\mathbf{V}_g|^{-1/2}\big]}{\partial \mathbf{V}_g} \\ &= -\frac{1}{2}|\mathbf{V}_g|^{-3/2}\big[|\mathbf{V}_g|(\mathbf{V}_g')^{-1}\big] \\ &= -\frac{1}{2}|\mathbf{V}_g|^{-1/2}(\mathbf{V}_g)^{-1} \\ &= -\frac{1}{2}\left[\frac{|\mathbf{V}_g|^{-1/2}}{\mathbf{V}_g}\right]\left[\frac{\mathbf{V}_g}{\mathbf{V}_g}\right] \\ &= -\frac{1}{2}\left[\frac{|\mathbf{V}_g|^{-1/2}\mathbf{V}_g}{\mathbf{V}_g^2}\right]\end{aligned}$$

- Según la propiedad (1.15) se tiene que:

$$\begin{aligned}\frac{\partial}{\partial \mathbf{V}_g}\big[W(\mathbf{V}_g)\big] &= \frac{\partial}{\partial \mathbf{V}_g}\left[\exp\left\{-\frac{1}{2}(\mathbf{x}_i-\mu_g)'\mathbf{V}_g^{-1}(\mathbf{x}_i-\mu_g)\right\}\right] \\ &= \exp\left\{-\frac{1}{2}(\mathbf{x}_i-\mu_g)'\mathbf{V}_g^{-1}(\mathbf{x}_i-\mu_g)\right\}\frac{\partial}{\partial \mathbf{V}_g}\left\{-\frac{1}{2}(\mathbf{x}_i-\mu_g)'\mathbf{V}_g^{-1}(\mathbf{x}_i-\mu_g)\right\} \\ &= \exp\left\{-\frac{1}{2}(\mathbf{x}_i-\mu_g)'\mathbf{V}_g^{-1}(\mathbf{x}_i-\mu_g)\right\}\left\{-\frac{1}{2}(\mathbf{x}_i-\mu_g)(\mathbf{x}_i-\mu_g)'(-1)(\mathbf{V}_g^{-2})\right\} \\ &= \exp\left\{-\frac{1}{2}(\mathbf{x}_i-\mu_g)'\mathbf{V}_g^{-1}(\mathbf{x}_i-\mu_g)\right\}\left\{\frac{(\mathbf{x}_i-\mu_g)(\mathbf{x}_i-\mu_g)'}{2\mathbf{V}_g^2}\right\}.\end{aligned}$$

Luego reemplazando en (3.26) tenemos:

$$\begin{aligned}\frac{\partial}{\partial \mathbf{V}_g}\big[f_g(\mathbf{x}_i)\big] &= (2\pi)^{-p/2}\left\{[U(\mathbf{V}_g)]\frac{\partial\big[W(\mathbf{V}_g)\big]}{\partial \mathbf{V}_g} + \big[W(\mathbf{V}_g)\big]\frac{\partial\big[U(\mathbf{V}_g)\big]}{\partial \mathbf{V}_g}\right\} \\ &= (2\pi)^{-p/2}\left\{\left[|\mathbf{V}_g|^{-1/2}\right]\times\left[\exp\left\{-\frac{1}{2}(\mathbf{x}_i-\mu_g)'\mathbf{V}_g^{-1}(\mathbf{x}_i-\mu_g)\right\}\left\{\frac{(\mathbf{x}_i-\mu_g)(\mathbf{x}_i-\mu_g)'}{2\mathbf{V}_g^2}\right\}\right]\right. \\ &\qquad \left.+\left[\exp\left\{-\frac{1}{2}(\mathbf{x}_i-\mu_g)'\mathbf{V}_g^{-1}(\mathbf{x}_i-\mu_g)\right\}\right]\times\left[\left(-\frac{1}{2}\right)\frac{|\mathbf{V}_g|^{-1/2}\mathbf{V}_g}{\mathbf{V}_g^2}\right]\right\} \\ &= \left\{\underbrace{|\mathbf{V}_g|^{-1/2}(2\pi)^{-p/2}\exp\left\{-\frac{1}{2}(\mathbf{x}_i-\mu_g)'\mathbf{V}_g^{-1}(\mathbf{x}_i-\mu_g)\right\}}_{f_g(\mathbf{x}_i)}\left[\frac{(\mathbf{x}_i-\mu_g)(\mathbf{x}_i-\mu_g)'}{2\mathbf{V}_g^2}\right]\right\}\end{aligned}$$

$$+\left\{\underbrace{|\mathbf{V}_g|^{-1/2}(2\pi)^{-p/2}\exp\left\{-\frac{1}{2}(\mathbf{x}_i-\mu_g)'\mathbf{V}_g^{-1}(\mathbf{x}_i-\mu_g)\right\}}_{f_g(\mathbf{x}_i)}\left[-\frac{\mathbf{V}_g}{2\mathbf{V}_g^2}\right]\right\}$$

$$= f_g(\mathbf{x}_i)\left\{\frac{(\mathbf{x}_i-\mu_g)(\mathbf{x}_i-\mu_g)'}{2\mathbf{V}_g^2}\right\} + f_g(\mathbf{x}_i)\left\{-\frac{\mathbf{V}_g}{2\mathbf{V}_g^2}\right\}.$$

Entonces:

$$\frac{\partial}{\partial \mathbf{V}_g}\big[f_g(\mathbf{x}_i)\big] = f_g(\mathbf{x}_i)\left\{\frac{(\mathbf{x}_i-\mu_g)(\mathbf{x}_i-\mu_g)'}{2\mathbf{V}_g^2}\right\} + f_g(\mathbf{x}_i)\left\{-\frac{\mathbf{V}_g}{2\mathbf{V}_g^2}\right\}$$

$$= \frac{f_g(\mathbf{x}_i)}{2\mathbf{V}_g^2}\left\{(\mathbf{x}_i-\mu_g)(\mathbf{x}_i-\mu_g)'-\mathbf{V}_g\right\} \tag{3.27}$$

Reemplazando (3.27) en (3.25) y sustituyendo los resultados dados en (3.19) y (3.24) tenemos:

$$\sum_{i=1}^{n}\frac{\pi_g\left[\frac{f_g(\mathbf{x}_i)}{2\mathbf{V}_g^2}\left\{(\mathbf{x}_i-\widehat{\mu}_g)(\mathbf{x}_i-\widehat{\mu}_g)'-\mathbf{V}_g\right\}\right]}{\sum_{g=1}^{G}\pi_g f_g(\mathbf{x}_i)} = 0$$

$$\sum_{i=1}^{n}\frac{\pi_g f_g(\mathbf{x}_i)}{\sum_{g=1}^{G}\pi_g f_g(\mathbf{x}_i)}\left[\frac{(\mathbf{x}_i-\widehat{\mu}_g)(\mathbf{x}_i-\widehat{\mu}_g)'-\mathbf{V}_g}{2\mathbf{V}_g^2}\right] = 0$$

$$\sum_{i=1}^{n}\pi_{ig}\left[\frac{(\mathbf{x}_i-\widehat{\mu}_g)(\mathbf{x}_i-\widehat{\mu}_g)'-\mathbf{V}_g}{2\mathbf{V}_g^2}\right] = 0$$

$$\sum_{i=1}^{n}\pi_{ig}\left[(\mathbf{x}_i-\widehat{\mu}_g)(\mathbf{x}_i-\widehat{\mu}_g)'-\mathbf{V}_g\right] = 0$$

$$\sum_{i=1}^{n}\pi_{ig}(\mathbf{x}_i-\widehat{\mu}_g)(\mathbf{x}_i-\widehat{\mu}_g)' - \sum_{i=1}^{n}\pi_{ig}\mathbf{V}_g = 0,$$

entonces:

$$\sum_{i=1}^{n}\pi_{ig}\mathbf{V}_g = \sum_{i=1}^{n}\pi_{ig}(\mathbf{x}_i-\widehat{\mu}_g)(\mathbf{x}_i-\widehat{\mu}_g)'$$

$$\boxed{\widehat{\mathbf{V}}_g = \frac{1}{\sum_{i=1}^{n}\pi_{ig}}\sum_{i=1}^{n}\pi_{ig}(\mathbf{x}_i-\widehat{\mu}_g)(\mathbf{x}_i-\widehat{\mu}_g)'} \tag{3.28}$$

Para resolver las ecuaciones (3.21), (3.24) y (3.28) y así obtener los estimadores se necesitan las probabilidades π_{ig} dadas con (3.19), pero para esto se necesitan los parámetros del modelo. Entonces tenemos el conflicto de estimar parámetros los cuales necesitamos conocer previamente para estimarlos.
Como se menciono al iniciar este capitulo, las mezclas de distribuciones probabilísticas son usadas para modelar problemas de clasificación de datos y es precisamente allí donde a medida que utilicemos un algoritmo de clasificación, vamos obteniendo elementos suficientes para resolver el problema de estimación de los parámetros.

3.5. Clasificación de datos asumiendo una mezcla finita.

Para clasificar datos que proviene de distribuciones mezcladas debemos considerar los resultados obtenidos en el *Análisis de conglomerados* y en *el Análisis discriminante*. En [3] se estudia a fondo estas temáticas. El análisis de conglomerados (clusters) tiene por objeto agrupar elementos en grupos homogéneos en función de las similitudes o similaridades entre ellos; es decir analiza la homogeneidad de la muestra y encuentra grupos desde un punto de vista descriptivo. El análisis discriminante es una herramienta que permite asignar o clasificar nuevos individuos dentro de grupos previamente reconocidos o definidos; es así como conocidas algunas características (variables) de un individuo y partiendo del hecho de que pertenece a uno o varios grupos (población) definidos de antemano , se debe asignar tal individuo en alguno de estos, con base en la información que de el se dispone. La técnica del análisis discriminante suministra los requerimientos y criterios para tomar la decisión.
En la siguiente sección supondremos que los datos se han generado a partir de una mezcla de G distribuciones desconocidas y se presenta un método para dividir la muestra en grupos mas homogéneos. Este es **el algoritmo EM para mezclas**, el cual nos permite hacer la estimación de los parámetros de las componentes de la mezcla en forma iterada. Conocidos los parámetros se realiza la clasificación de los individuos en los grupos por sus probabilidades de pertenencia.
Existen otros métodos de clasificación como lo es **el método de las G-medias** o k-medias que se estudia en el análisis de conglomerados y el cual se presenta aquí en el **Apéndice C** observándose así que el criterio estudiado inicialmente es óptimo para una determinada configuración de los datos y que las hipótesis que se hacen respecto a las componentes de la mezcla implicaran distintos criterios que podremos maximizar de forma similar al criterio de las k-medias. Un buen método es **el muestreo de Gibbs o Gibbs sampling**

para mezclas, de enfoque netamente bayesiano y el cual se ha implantado en algunos programas computacionales, partiendo del supuesto de datos provenientes de una mezcla de G poblaciones normales. Este procedimiento hace parte de la gran familia de los métodos de Monte Carlo con cadenas de Markov, o métodos $(MC)^2$ los cuales buscan generar muestras de la distribución *a posteriori* establecida para la mezcla. En [1] se describe como trabaja este método.
Otro es **el método de la minima Kurtosis**, el cual esta clasificado como un método de proyección de datos; este busca e investiga direcciones de proyección de los datos donde puedan aparecer los distintos grupos para así después buscar los grupos sobre estas direcciones univariadas. En [1] se expone este método para la clasificación de datos mediante mezclas y se da una dirección en internet donde se encuentra un algoritmo en $MATLAB$ para aplicar estos procedimientos.
En [1] también se menciona **el método SAR**, el cual es un método para encontrar grupos heterogéneos en distribuciones normales mezcladas.

3.6. El algoritmo EM

3.6.1. Generalidades del algoritmo

El algoritmo **EM** fue descrito y analizado por Dempster, Laird y Rubin en (1977), en el documento titulado *Maximum Likelihood from Incomplete Data via the EM Algorithm* publicado en *Journal of the Royal Statistical Society. Series B, 39:1, pp. 1-38*, aunque el método se había usado mucho antes por otros investigadores como Hartley (1958). Este algoritmo es un método de optimización iterativo usado para estimar parámetros desconocidos en la función de maxima verosimilitud de un conjunto de datos muestrales (estimación que puede ser un problema intratable analíticamente). Luego lo que realiza el algoritmo **EM** es simplificar la estimación ampliando el conjunto de datos con los que se trabaja; es así como introduce unos *datos no observados* o *datos ausentes* independientes de los datos muestrales que denominaremos *datos observados*.
Su nombre de algoritmo **EM** se debe a que esta compuesto de dos pasos alternados que involucran una esperanza y una maximización. Además este método se ha convertido en una herramienta ampliamente usada por investigadores en las diferentes áreas, pues permite estimar datos faltantes o ausentes en diversos problemas multivariantes dónde los algoritmos basados en los métodos iterativos de Newton pueden resultar más complicados. El éxito del algoritmo **EM** radica en su simplicidad, su estabilidad y sus propiedades de

convergencia. Como se mencionó anteriormente el algoritmo **EM** implica dos pasos: el paso de la **E**speranza y el paso de la **M**aximización. Suponiendo una muestra aleatoria para la cual se establece un modelo estadístico, los pasos del algoritmo a nivel general son:

- **Paso E:** Partiendo de unos valores iniciales de los parámetros, este paso consiste en calcular para los datos ausentes la esperanza de la función de verosimilitud condicionada a los valores de los datos observados y a la distribución de los datos ausentes.
- **Paso M:** En el se maximiza la esperanza encontrada en el **paso E** respecto a los parámetros.

Después los parámetros encontrados en el **paso M** se toman como valores iniciales en el **paso E**, estos dos pasos se repiten alternadamente hasta que el valor de los parámetros no cambie. Nótese que en el **paso E** se hace necesario conocer la distribución de los datos ausentes.
Aunque el algoritmo **EM** es simple y su convergencia se garantiza teóricamente, es muy criticado a causa de que precisamente su convergencia por lo general es sumamente lenta. De hecho, la convergencia del algoritmo **EM** puede ser muy lenta cuando otros métodos como Newton-Raphson convergen muy rápidamente. Este problema ha llevado a investigadores a pensar en diferentes formas para aumentar la velocidad o acelerar la convergencia del algoritmo y así reducir el número de iteraciones en los pasos **E** y **M**. En investigación computacional un acelerador bueno tiene que equilibrar las ganancias entre la reducción del tiempo computo y el aumento en el tiempo de procesamiento. En [6] se estudian los aceleradores mas importantes para este algoritmo, además de mostrar una nueva propuesta para acelerar el algoritmo **EM**.

3.6.2. Fundamentos del algoritmo

Para comprender como funciona el algoritmo supongamos que tenemos una muestra de tamaño $n = 20$ de una variable aleatoria vectorial $\mathbf{X} = (\mathbf{X_a}, \mathbf{X_b}, \mathbf{X_c}, \mathbf{X_d}, \mathbf{X_e})$, donde algunos de los n individuos observados carecen de valores en algunas variables. Estos se pueden observar en el cuadro 3.1.
Este problema de los datos faltantes intuitivamente se podría resolver estableciendo un algoritmo que funcione así:

1. Con los individuos que poseen datos observados de forma completa se estima los parámetros del modelo estadístico y a estos parámetros los llamamos $\theta^{[0]}$. Estos

Individuo	$\mathbf{X_a}$	$\mathbf{X_b}$	$\mathbf{X_c}$	$\mathbf{X_d}$	$\mathbf{X_e}$	Individuo	$\mathbf{X_a}$	$\mathbf{X_b}$	$\mathbf{X_c}$	$\mathbf{X_d}$	$\mathbf{X_e}$
1	a_1	b_1	c_1	d_1	e_1	11	a_{11}	b_{11}	c_{11}	d_{11}	e_{11}
2	a_2	b_2	c_2	d_2	e_2	12	a_{12}	b_{12}	c_{12}	·	e_{12}
3	·	b_3	c_3	·	e_3	13	·	b_{13}	c_{13}	d_{13}	e_{13}
4	a_4	b_4	c_4	d_4	e_4	14	a_{14}	b_{14}	·	d_{14}	e_{14}
5	a_5	b_5	·	d_5	e_5	15	a_{15}	b_{15}	·	·	e_{15}
6	a_6	b_6	c_6	·	e_6	16	a_{16}	b_{16}	c_{16}	d_{16}	e_{16}
7	·	b_7	·	d_7	e_7	17	a_{17}	b_{17}	c_{17}	d_{17}	e_{17}
8	a_8	b_8	c_8	d_8	e_8	18	a_{18}	b_{18}	c_{18}	·	e_{18}
9	a_9	b_9	c_9	·	e_9	19	·	b_{19}	c_{19}	d_{19}	e_{19}
10	a_{10}	b_{10}	·	d_{10}	e_{10}	20	a_{20}	b_{20}	·	d_{20}	e_{20}

Cuadro 3.1: Ejemplo de 20 individuos con datos ausentes en las variables $\mathbf{X_a}, \mathbf{X_c}$ y $\mathbf{X_d}$

parámetros también puede ser tomados de forma arbitraria, según las consideraciones del caso.

2. Los datos faltantes o no observados se estiman a partir de los datos conocidos y el valor inicial $\theta^{[0]}$.

3. Con estos datos estimados y los datos observados en forma completa se realiza una nueva estimación via máxima verosimilitud del parámetro θ y la llamamos $\theta^{[1]}_{MV}$.

4. Esta estimación se utiliza para re-estimar los datos y se continua iterando hasta la convergencia en los parámetros.

El anterior procedimiento intuitivo es bueno en muchos casos, pero no tiene en cuenta como se utilizan los datos ausentes para estimar los parámetros a partir de la verosimilitud; es decir no tiene en cuenta como se distribuyen las datos faltantes. Por esta y otras razones el algoritmo **EM** es un procedimiento óptimo y eficiente para problemas con datos ausentes. En el algoritmo **EM** analiza la aparición de los datos ausentes desde la unificación de dos enfoques:

(a) Individuos con datos faltantes: Si llamamos X al vector con los n elementos de la muestra, podemos particionarlo así: $\mathrm{X} = (\mathrm{Y}, \mathrm{Z})'$, donde $\mathrm{Y} = (\mathrm{y}_1, \ldots, \mathrm{y}_{n_y})'$ son los individuos con datos completos y $\mathrm{Z} = (\mathrm{z}_1, \ldots, \mathrm{z}_{n_z})'$ son los individuos con datos ausentes. Note además que y'_i,z'_i son vectores de tamaño $(p \times 1)$, con lo que X es una matriz de tamaño $(n \times p)$ y $n_y + n_z = n$.

Para la muestra dada en el cuadro 3.1, con $n = 20, n_y = 7, n_z = 13$ y $p = 5$, esta partición será:

$$\mathbb{X} = \begin{pmatrix} a_1 & b_1 & c_1 & d_1 & e_1 \\ a_2 & b_2 & c_2 & d_2 & e_2 \\ \cdot & b_3 & c_3 & \cdot & e_3 \\ a_4 & b_4 & c_4 & d_4 & e_4 \\ a_5 & b_5 & \cdot & d_5 & e_5 \\ a_6 & b_6 & c_6 & \cdot & e_6 \\ \cdot & b_7 & \cdot & d_7 & e_7 \\ a_8 & b_8 & c_8 & d_8 & e_8 \\ a_9 & b_9 & c_9 & \cdot & e_9 \\ a_{10} & b_{10} & \cdot & d_{10} & e_{10} \\ a_{11} & b_{11} & c_{11} & d_{11} & e_{11} \\ a_{12} & b_{12} & c_{12} & \cdot & e_{12} \\ \cdot & b_{13} & c_{13} & d_{13} & e_{13} \\ a_{14} & b_{14} & \cdot & d_{14} & e_{14} \\ a_{15} & b_{15} & \cdot & \cdot & e_{15} \\ a_{16} & b_{16} & c_{16} & d_{16} & e_{16} \\ a_{17} & b_{17} & c_{17} & d_{17} & e_{17} \\ a_{18} & b_{18} & c_{18} & \cdot & e_{18} \\ \cdot & b_{19} & c_{19} & d_{19} & e_{19} \\ a_{20} & b_{20} & \cdot & d_{20} & e_{20} \end{pmatrix} \Longrightarrow \mathrm{X} = \begin{pmatrix} a_1 & b_1 & c_1 & d_1 & e_1 \\ a_2 & b_2 & c_2 & d_2 & e_2 \\ a_4 & b_4 & c_4 & d_4 & e_4 \\ a_8 & b_8 & c_8 & d_8 & e_8 \\ a_{11} & b_{11} & c_{11} & d_{11} & e_{11} \\ a_{16} & b_{16} & c_{16} & d_{16} & e_{16} \\ a_{17} & b_{17} & c_{17} & d_{17} & e_{17} \\ \cdots & \cdots & \cdots & \cdots & \cdots \\ \cdot & b_3 & c_3 & \cdot & e_3 \\ a_5 & b_5 & \cdot & d_5 & e_5 \\ a_6 & b_6 & c_6 & \cdot & e_6 \\ \cdot & b_7 & \cdot & d_7 & e_7 \\ a_9 & b_9 & c_9 & \cdot & e_9 \\ a_{10} & b_{10} & \cdot & d_{10} & e_{10} \\ a_{12} & b_{12} & c_{12} & \cdot & e_{12} \\ \cdot & b_{13} & c_{13} & d_{13} & e_{13} \\ a_{14} & b_{14} & \cdot & d_{14} & e_{14} \\ a_{15} & b_{15} & \cdot & \cdot & e_{15} \\ a_{18} & b_{18} & c_{18} & \cdot & e_{18} \\ \cdot & b_{19} & c_{19} & d_{19} & e_{19} \\ a_{20} & b_{20} & \cdot & d_{20} & e_{20} \end{pmatrix} \begin{matrix} \\ \longrightarrow \mathrm{Y} \\ \\ \\ \\ \\ \\ \\ \\ \\ \\ \\ \\ \\ \\ \\ \\ \longrightarrow \mathrm{Z} \end{matrix}$$

(b) Variables con datos faltantes: Si llamamos X al vector de variables, podemos particionar este vector así: $\mathrm{X} = (\mathrm{Y}, \mathrm{Z})$, donde $\mathrm{Y} = (\mathrm{y}_1, \ldots, \mathrm{y}_{p_y})$ son las variables con datos completos y $\mathrm{Z} = (\mathrm{z}_1, \ldots, \mathrm{z}_{p_z})$ son las variables con datos incompletos. Note además que y_i,z_i son vectores de tamaño $(n \times 1)$, con lo que X es una matriz de tamaño $(n \times p)$ y $p_y + p_z = p$.

Para la muestra dada en el cuadro 3.1, con $p = 5, p_y = 2, p_z = 3$ y $n = 20$, esta

partición será:

$$\mathbb{X} = \begin{pmatrix} a_1 & b_1 & c_1 & d_1 & e_1 \\ a_2 & b_2 & c_2 & d_2 & e_2 \\ \cdot & b_3 & c_3 & \cdot & e_3 \\ a_4 & b_4 & c_4 & d_4 & e_4 \\ a_5 & b_5 & \cdot & d_5 & e_5 \\ a_6 & b_6 & c_6 & \cdot & e_6 \\ \cdot & b_7 & \cdot & d_7 & e_7 \\ a_8 & b_8 & c_8 & d_8 & e_8 \\ a_9 & b_9 & c_9 & \cdot & e_9 \\ a_{10} & b_{10} & \cdot & d_{10} & e_{10} \\ a_{11} & b_{11} & c_{11} & d_{11} & e_{11} \\ a_{12} & b_{12} & c_{12} & \cdot & e_{12} \\ \cdot & b_{13} & c_{13} & d_{13} & e_{13} \\ a_{14} & b_{14} & \cdot & d_{14} & e_{14} \\ a_{15} & b_{15} & \cdot & \cdot & e_{15} \\ a_{16} & b_{16} & c_{16} & d_{16} & e_{16} \\ a_{17} & b_{17} & c_{17} & d_{17} & e_{17} \\ a_{18} & b_{18} & c_{18} & \cdot & e_{18} \\ \cdot & b_{19} & c_{19} & d_{19} & e_{19} \\ a_{20} & b_{20} & \cdot & d_{20} & e_{20} \end{pmatrix} \Longrightarrow \mathrm{X} = \left(\begin{array}{ccc:cc:c} b_1 & e_1 & a_1 & c_1 & d_1 \\ b_2 & e_2 & a_2 & c_2 & d_2 \\ b_3 & e_3 & \cdot & c_3 & \cdot \\ b_4 & e_4 & a_4 & c_4 & d_4 \\ b_5 & e_5 & a_5 & \cdot & d_5 \\ b_6 & e_6 & a_6 & c_6 & \cdot \\ b_7 & e_7 & \cdot & \cdot & d_7 \\ b_8 & e_8 & a_8 & c_8 & d_8 \\ b_9 & e_9 & a_9 & c_9 & \cdot \\ b_{10} & e_{10} & a_{10} & \cdot & d_{10} \\ b_{11} & e_{11} & a_{11} & c_{11} & d_{11} \\ b_{12} & e_{12} & a_{12} & c_{12} & \cdot \\ b_{13} & e_{13} & \cdot & c_{13} & d_{13} \\ b_{14} & e_{14} & a_{14} & \cdot & d_{14} \\ b_{15} & e_{15} & a_{15} & \cdot & \cdot \\ b_{16} & e_{16} & a_{16} & c_{16} & d_{16} \\ b_{17} & e_{17} & a_{17} & c_{17} & d_{17} \\ b_{18} & e_{18} & a_{18} & c_{18} & \cdot \\ b_{19} & e_{19} & \cdot & c_{19} & d_{19} \\ b_{20} & e_{20} & a_{20} & \cdot & d_{20} \end{array}\right) \begin{matrix} \\ \downarrow \\ \mathrm{Y} \end{matrix} \quad \begin{matrix} \\ \downarrow \\ \mathrm{Z} \end{matrix}$$

Luego los criterios **(a)** y **(b)** se unifican al llamar a X como ***la matriz de datos de la muestra***, los cuales se pueden particionar en $\mathrm{Y} = (\mathrm{y}_1, \ldots, \mathrm{y}_{n_y})$, que denominamos ***matriz de datos observados***, donde y_i es un vector de tamaño $(p_y \times 1)$ y $\mathrm{Z} = (\mathrm{z}_1, \ldots, \mathrm{z}_{n_z})$ que denominaremos ***matriz de datos ausentes*** donde z_i es un vector de tamaño $(p_z \times 1)$. Esta formulación cubre los dos casos anteriormente nombrados, ya que para **(a)** tenemos n_y individuos con datos completos y n_z datos ausentes, cada uno de estos con dimension $p_y = p_z = p$. Para el enfoque dado en **(b)** tendremos p_y variables con datos completos y

p_z variables con datos ausentes, cada una de estas con dimension $n_y = n_z = n$.
Para la muestra dada en el cuadro 3.1, con $n = 20, n_y = 7, n_z = 13$ y $p = 5, p_y = 2, p_z = 3$, esta unificación será:

$$\mathbb{X} = \begin{pmatrix} a_1 & b_1 & c_1 & d_1 & e_1 \\ a_2 & b_2 & c_2 & d_2 & e_2 \\ \cdot & b_3 & c_3 & \cdot & e_3 \\ a_4 & b_4 & c_4 & d_4 & e_4 \\ a_5 & b_5 & \cdot & d_5 & e_5 \\ a_6 & b_6 & c_6 & \cdot & e_6 \\ \cdot & b_7 & \cdot & d_7 & e_7 \\ a_8 & b_8 & c_8 & d_8 & e_8 \\ a_9 & b_9 & c_9 & \cdot & e_9 \\ a_{10} & b_{10} & \cdot & d_{10} & e_{10} \\ a_{11} & b_{11} & c_{11} & d_{11} & e_{11} \\ a_{12} & b_{12} & c_{12} & \cdot & e_{12} \\ \cdot & b_{13} & c_{13} & d_{13} & e_{13} \\ a_{14} & b_{14} & \cdot & d_{14} & e_{14} \\ a_{15} & b_{15} & \cdot & \cdot & e_{15} \\ a_{16} & b_{16} & c_{16} & d_{16} & e_{16} \\ a_{17} & b_{17} & c_{17} & d_{17} & e_{17} \\ a_{18} & b_{18} & c_{18} & \cdot & e_{18} \\ \cdot & b_{19} & c_{19} & d_{19} & e_{19} \\ a_{20} & b_{20} & \cdot & d_{20} & e_{20} \end{pmatrix} \Longrightarrow \mathrm{X} = \begin{pmatrix} b_1 & e_1 & \vdots & a_1 & c_1 & d_1 \\ b_2 & e_2 & \vdots & a_2 & c_2 & d_2 \\ b_4 & e_4 & \vdots & a_4 & c_4 & d_4 \\ b_8 & e_8 & \vdots & a_8 & c_8 & d_8 \\ b_{11} & e_{11} & \vdots & a_{11} & c_{11} & d_{11} \\ b_{16} & e_{16} & \vdots & a_{16} & c_{16} & d_{16} \\ b_{17} & e_{17} & \vdots & a_{17} & c_{17} & d_{17} \\ \cdots & \cdots & \cdots & \cdots & \cdots & \cdots \\ b_3 & e_3 & \vdots & \cdot & c_3 & \cdot \\ b_5 & e_5 & \vdots & a_5 & \cdot & d_5 \\ b_6 & e_6 & \vdots & a_6 & c_6 & \cdot \\ b_7 & e_7 & \vdots & \cdot & \cdot & d_7 \\ b_9 & e_9 & \vdots & a_9 & c_9 & \cdot \\ b_{10} & e_{10} & \vdots & a_{10} & \cdot & d_{10} \\ b_{12} & e_{12} & \vdots & a_{12} & c_{12} & \cdot \\ b_{13} & e_{13} & \vdots & \cdot & c_{13} & d_{13} \\ b_{14} & e_{14} & \vdots & a_{14} & \cdot & d_{14} \\ b_{15} & e_{15} & \vdots & a_{15} & \cdot & \cdot \\ b_{18} & e_{18} & \vdots & a_{18} & c_{18} & \cdot \\ b_{19} & e_{19} & \vdots & \cdot & c_{19} & d_{19} \\ b_{20} & e_{20} & \vdots & a_{20} & \cdot & d_{20} \end{pmatrix} \begin{matrix} \\ \\ \\ \\ \longrightarrow \\ \\ \\ \\ \\ \\ \\ \\ \\ \\ \\ \\ \\ \\ \\ \\ \longrightarrow \\ \\ \\ \\ \end{matrix} \begin{matrix} \\ \\ \\ \\ \mathrm{Y} \\ \\ \\ \\ \\ \\ \\ \\ \\ \\ \\ \\ \\ \\ \\ \\ \mathrm{Z} \\ \\ \\ \\ \end{matrix}$$

$$\begin{matrix} \downarrow & & \downarrow \\ \mathrm{Y} & & \mathrm{Z} \end{matrix}$$

Entonces el algoritmo **EM** toma a $\mathrm{X} = (\mathrm{Y}, \mathrm{Z})$ como la partición de los datos muestrales, ya sea como vectores de individuos, como vectores de variables o ambos.
Aprovechando la relación existente en las distribuciones de probabilidad condicionadas, podemos asegurar que la función de densidad conjunta de (Y, Z), bajo los parámetros θ

es:

$$f(\mathrm{Y},\mathrm{Z}|\theta) = f(\mathrm{Z}|\mathrm{Y},\theta)f(\mathrm{Y}|\theta),$$

entonces:

$$f(\mathrm{Y}|\theta) = \frac{f(\mathrm{Y},\mathrm{Z}|\theta)}{f(\mathrm{Z}|\mathrm{Y},\theta)}.$$

Luego podemos decir que la función de verosimilitud para los datos observados Y será:

$$l(\theta|\mathrm{Y}) = \frac{l(\theta|\mathrm{Y},\mathrm{Z})}{l(\theta,\mathrm{Y}|\mathrm{Z})} \tag{3.29}$$

donde $l(\theta|\mathrm{Y},\mathrm{Z})$ es la verosimilitud para toda la muestra y $l(\theta,\mathrm{Y}|\mathrm{Z})$ es la verosimilitud de los datos ausentes conocidos los datos observados. Operando en la ecuación (3.29) se tiene:

$$\begin{aligned} \ln l(\theta|\mathrm{Y}) &= \ln l(\theta|\mathrm{Y},\mathrm{Z}) - \ln l(\theta,\mathrm{Y}|\mathrm{Z}) \\ L(\theta|\mathrm{Y}) &= L(\theta|\mathrm{Y},\mathrm{Z}) - \ln l(\theta,\mathrm{Y}|\mathrm{Z}), \end{aligned} \tag{3.30}$$

donde $L(\theta|\mathrm{Y})$ es el soporte de los datos observados, $L(\theta|\mathrm{Y},\mathrm{Z})$ es el soporte para toda la muestra y $\ln l(\theta,\mathrm{Y}|\mathrm{Z})$ es la mejor densidad de los datos ausentes conocidos los datos observados y los parámetros.

Como se mencionó en las generalidades del algoritmo **EM**, el objetivo real de este es estimar los parámetros desconocidos de una muestra poblacional. Entonces si asumimos que los datos ausentes fueron introducidos para simplificar la estimación de los parámetros desconocidos, nuestro interés radica es en maximizar la función de soporte $L(\theta|\mathrm{Y})$ dada en (3.30). Para maximizar esta función, necesitamos que en la diferencia allí planteada el soporte de la muestra completa $L(\theta|\mathrm{Y},\mathrm{Z})$ sea máximo. En la practica, la maximización de $L(\theta|\mathrm{Y},\mathrm{Z})$ es mas fácil de realizar que la maximización de los datos observados $L(\theta|\mathrm{Y})$. Es esta la razón por la cual el algoritmo **EM** usa el soporte $L(\theta|\mathrm{Y},\mathrm{Z})$ como función de verosimilitud en la búsqueda del estimador MV de θ.

Así podemos ya dar en forma completa la estructura funcional de el algoritmo **EM**:

- Partir de un estimador inicial $\widehat{\theta}^{[0]}$ y estipular un margen de error o tolerancia para los valores de los parámetros estimados.
- Iniciar un contador con $k = 0$.
- Hacer $k = k + 1$.
- **Paso E:**
  Usar la estimación actual de θ, es decir $\widehat{\theta}^{[k-1]}$, para calcular la esperanza de la

verosimilitud $L(\theta|\mathrm{Y},\mathrm{Z})$ con respecto a la distribución de los valores ausentes Z dados $\widehat{\theta}^{[k-1]}$ y los datos observados Y. Esto nos dará una nueva verosimilitud que denominaremos $L^*(\theta|\mathrm{Y})$, es decir:

$$L^*(\theta|\mathrm{Y}) = \mathrm{E}_{Z|\widehat{\theta}^{[k-1]}}\Big[L(\theta|\mathrm{Y},\mathrm{Z})\Big]. \tag{3.31}$$

- **Paso M:**
  Maximizar $L^*(\theta|\mathrm{Y})$ con respecto al vector de variables θ. Llamaremos a este nuevo vector de parámetros $\widehat{\theta}^{[k]}$, es decir:

$$\widehat{\theta}^{[k]} = \underset{\theta}{\max}\Big[L^*(\theta|\mathrm{Y})\Big]. \tag{3.32}$$

- Se calcula $\|\widehat{\theta}^{[k]} - \widehat{\theta}^{[k-1]}\|$ y se evalúa si es lo suficiente pequeña; es decir comparar si es menor que la tolerancia establecida anteriormente.
  Si lo es, nuestro estimador máximo verosímil de los parámetros es $\widehat{\theta}^{[k]}$. Si la diferencia no es suficientemente pequeña se retorna al paso $k = k + 1$, incrementando así este contador. Seguimos con el **paso E** y repetimos el proceso hasta lograr la convergencia.

En Dempster, Laird y Rubin (1977), se demuestra que el algoritmo converge y en [1], apéndice 11.1 se demuestra que valor estimado con el algoritmo es realmente el estimador MV para la muestra.

3.6.3. El algoritmo EM en mezclas

En un proceso de clasificación de n individuos donde suponemos que estos provienen de una mezcla de G distribuciones normales, donde G se a determinado de antemano, es de nuestro interés conocer el vector de parámetros $\theta = (\theta_1, \ldots, \theta_G)'$ para así realizar la posterior asignación de estos (los individuos) en los grupos determinados; esta clasificación se hace según la probabilidad de pertenencia de cada individuo al grupo.
Por medio del algoritmo **EM** podemos estimar el vector de parámetros θ. Para esto introduzcamos un conjunto de variables no observadas $\mathbf{Z} = (\mathbf{z_1}, \ldots, \mathbf{z_n})'$ donde
$\mathbf{z_i} = (z_{i1}, \ldots, z_{iG})$, con $i = 1, \ldots, n$, tiene como función indicar de que componente de la mezcla proviene cada observación. Luego para cada arreglo en la variable $\mathbf{z_i}$ se tiene:

$$z_{ig} = \begin{cases} 1, & \text{si } i \text{ proviene de la población } g; \\ 0, & \text{en otro caso.} \end{cases} \tag{3.33}$$

Por lo tanto cada individuo i tiene G posibilidades:

$$\text{Pertenecer al grupo } 1 \Longrightarrow \mathbf{z_i} = (1, 0, 0, \ldots, 0)'_{G\times 1}$$

$$\text{Pertenecer al grupo } 2 \Longrightarrow \mathbf{z_i} = (0, 1, 0, \ldots, 0)'_{G\times 1}$$

$$\vdots$$

$$\text{Pertenecer al grupo } G \Longrightarrow \mathbf{z_i} = (0, 0, 0, \ldots, 1)'_{G\times 1}$$

Esta ampliación en los datos muestrales nos abliga a renombrarlos asi: $\mathbf{X} = (\mathbf{Y}, \mathbf{Z})$ donde $\mathbf{Y}$ representa los datos observados y $\mathbf{Z}$ la procedencia de estos (datos no observados). Ilustremos esta notación en el siguiente ejemplo.

Ejemplo 3.1. En el ejemplo 2.10 de [3] se clasifican doce objetos en tres grupos asi:

Objeto	$\mathbf{X_1}$	$\mathbf{X_2}$	**Grupo**
O_1	1	2	**I**
O_2	1	3	**I**
O_3	2	3	**I**
O_4	2	5	**II**
O_5	0	6	**II**
O_6	3	1	**I**
O_7	4	5	**II**
O_8	6	5	**III**
O_9	7	5	**III**
O_{10}	7	2	**III**
O_{11}	8	3	**III**
O_{12}	9	1	**III**

Cuadro 3.2: Clasificación de 12 individuos en 3 grupos. La distribución se realizó por el método de las G-medias.

Las variables no observadas para este ejemplo son:

$$\begin{array}{lll} \mathbf{z_1} = (1,0,0) & \mathbf{z_5} = (0,1,0) & \mathbf{z_9} = (0,0,1) \\ \mathbf{z_2} = (1,0,0) & \mathbf{z_6} = (1,0,0) & \mathbf{z_{10}} = (0,0,1) \\ \mathbf{z_3} = (1,0,0) & \mathbf{z_7} = (0,1,0) & \mathbf{z_{11}} = (0,0,1) \\ \mathbf{z_4} = (0,1,0) & \mathbf{z_8} = (0,0,1) & \mathbf{z_{12}} = (0,0,1) \end{array}$$

Luego la matriz de datos observados junto con su procedencia se representan así:

$$\mathbf{X} = \left(\begin{array}{cc:ccc} 1 & 2 & 1 & 0 & 0 \\ 1 & 3 & 1 & 0 & 0 \\ 2 & 3 & 1 & 0 & 0 \\ 2 & 5 & 0 & 1 & 0 \\ 0 & 6 & 0 & 1 & 0 \\ 3 & 1 & 1 & 0 & 0 \\ 4 & 5 & 0 & 1 & 0 \\ 6 & 5 & 0 & 0 & 1 \\ 7 & 5 & 0 & 0 & 1 \\ 7 & 2 & 0 & 0 & 1 \\ 8 & 3 & 0 & 0 & 1 \\ 9 & 1 & 0 & 0 & 1 \end{array}\right).$$

Bajo las anteriores consideraciones y por la definición de distribución condicionada, decimos que el valor en la mezcla de densidades para el i-esimo individuo viene dada por:

$$\mathbf{G}(\mathbf{x_i}) = \mathbf{G}(\mathbf{y_i}, \mathbf{z_i}) = \mathbf{G}(\mathbf{y_i}|\mathbf{z_i})\mathbf{G}(\mathbf{z_i}), \tag{3.34}$$

donde $\mathbf{G}(\mathbf{y_i}|\mathbf{z_i})$ es la función de densidad en la mezcla del vector de variables $\mathbf{y_i}$ dado $\mathbf{z_i}$, la cual se define como:

$$\mathbf{G}(\mathbf{y_i}|\mathbf{z_i}) = \prod_{g=1}^{G} f_g(\mathbf{x_i})^{z_{ig}}. \tag{3.35}$$

La definición dada en (3.35) es consistente, pues en el vector $\mathbf{z_i}$ solo una de sus componentes z_{ig} es distinta de cero y es precisamente esa componente la que define cual es la función de densidad de la observación $\mathbf{x_i}$.
En la ecuación (3.34), $\mathbf{G}(\mathbf{z_i})$ define la función de probabilidad de la variable $\mathbf{z_i}$, la cual esta dada por:

$$p(\mathbf{z_i}) = p(\mathbf{z_1}, \ldots, \mathbf{z_n}) = \mathbf{G}(\mathbf{z_i}) = \prod_{g=1}^{G} (\pi_g)^{z_{ig}}. \tag{3.36}$$

Luego, reemplazando (3.35) y (3.36) en (3.34), podemos escribir la mezcla de densidades como:

$$\mathbf{G}(\mathbf{y_i}, \mathbf{z_i}) = \left[\prod_{g=1}^{G} f_g(\mathbf{x_i})^{z_{ig}}\right]\left[\prod_{g=1}^{G} (\pi_g)^{z_{ig}}\right]$$

$$= \prod_{g=1}^{G} f_g(\mathbf{x_i})^{z_{ig}} (\pi_g)^{z_{ig}}$$
$$= \prod_{g=1}^{G} [\pi_g f_g(\mathbf{x_i})]^{z_{ig}}. \tag{3.37}$$

La función de verosimilitud para la mezcla, definida en (3.10), vendrá dada por:

$$l\big(\theta\big|\mathbf{Y}, \mathbf{Z}\big) = \prod_{i=1}^{n} \mathbf{G}(\mathbf{y_i}, \mathbf{z_i}|\theta\big)$$
$$= \prod_{i=1}^{n} \prod_{g=1}^{G} [\pi_g f_g(\mathbf{x_i})]^{z_{ig}}. \tag{3.38}$$

Por lo tanto la función de soporte para la mezcla, definida en (3.11), será:

$$L\big(\theta\big|\mathbf{Y}, \mathbf{Z}\big) = \ln\big\{ l\big(\theta\big|\mathbf{y_i}, \mathbf{z_i}\big)\big\}$$
$$= \ln\left\{ \prod_{i=1}^{n} \prod_{g=1}^{G} [\pi_g f_g(\mathbf{x_i})]^{z_{ig}} \right\}$$
$$= \sum_{i=1}^{n} \ln\left\{ \prod_{g=1}^{G} [\pi_g f_g(\mathbf{x_i})]^{z_{ig}} \right\}$$
$$= \sum_{i=1}^{n} \sum_{g=1}^{G} \ln[\pi_g f_g(\mathbf{x_i})]^{z_{ig}}$$
$$= \sum_{i=1}^{n} \sum_{g=1}^{G} z_{ig} \ln[\pi_g f_g(\mathbf{x_i})]$$
$$= \sum_{i=1}^{n} \sum_{g=1}^{G} z_{ig} \ln[\pi_g] + \sum_{i=1}^{n} \sum_{g=1}^{G} z_{ig} \ln[f_g(\mathbf{x_i})]. \tag{3.39}$$

La función de soporte hallada en (3.39) es precisamente la implementada por el algoritmo **EM** para realizar sus operaciones internas.
Ahora analicemos como funciona el algoritmo **EM** para mezclas paso a paso.

- **Partir de un estimador inicial $\widehat{\theta}^{[0]} = (\theta_1^{[0]}, \ldots, \theta_G^{[0]})$ y determinar una tolerancia tol.**
  Este estimador inicial es establecido por el investigador usando algún método de

clasificación multivariante, o por información *a priori* de la muestra, o arbitrariamente si lo permite la investigación. Para nuestro caso $\theta_g^{[0]} = (\widehat{\pi}_g^{[0]}, \widehat{\mu}_g^{[0]}, \widehat{\mathbf{V}}_g^{[0]})$, con $g = 1, \ldots, G$.
La tolerancia *tol* se stablece lo suficientemente pequeña para garantizar la precisión de la estimación.

- **Iniciar un contador con $k = 0$.**
- **Hacer $k = k + 1$.**
- **Paso E:**

$$\begin{aligned} L^*\big(\theta|\mathbf{Y}\big) &= \mathrm{E}_{Z|\widehat{\theta}^{[k-1]}}\Big\{L\big(\theta|\mathbf{Y},\mathbf{Z}\big)\Big\} \\ &= \mathrm{E}_{Z|\widehat{\theta}^{[k-1]}}\left\{\sum_{i=1}^{n}\sum_{g=1}^{G} z_{ig}\ln[\pi_g] + \sum_{i=1}^{n}\sum_{g=1}^{G} z_{ig}\ln\big[f_g(\mathbf{x_i})\big]\right\} \\ &= \sum_{i=1}^{n}\sum_{g=1}^{G}\mathrm{E}\Big\{z_{ig}|\mathbf{Y},\widehat{\theta}^{[k-1]}\Big\}\ln[\pi_g] + \sum_{i=1}^{n}\sum_{g=1}^{G}\mathrm{E}\Big\{z_{ig}|\mathbf{Y},\widehat{\theta}^{[k-1]}\Big\}\ln\big[f_g\big(\mathbf{x_i}\big)\big]. \end{aligned} \tag{3.40}$$

Notese ademas que, para nuestro caso, la funcion $f_g\big(\mathbf{x_i}\big)$ implementada en este paso y los siguientes es

$$\boxed{f_g\big(\mathbf{x_i}\big) = N_p\left(\mathbf{x_i}; \widehat{\mu}_g^{[k-1]}, \widehat{\mathbf{V}}_g^{[k-1]}\right).}$$

Puesto que z_{ig} es una variable binaria, su esperanza coincide con la probabilidad de que tome el valor de uno; esto sucede cuando la observación $\mathbf{x_i}$ proviene de la población g, que es precisamente la probabilidad π_{ig} dada en (3.19). Es decir:

$$\mathrm{E}\Big\{z_{ig}|\mathbf{Y},\widehat{\theta}^{[k-1]}\Big\} = p\Big(z_{ig}=1|\mathbf{Y},\widehat{\theta}^{[k-1]}\Big) = p\Big(z_{ig}=1|\mathbf{y_i},\widehat{\theta}^{[k-1]}\Big) = \widehat{\pi}_{ig}^{[k]}$$

Para la k-esima iteración se tiene por (3.19) que:

$$\boxed{\widehat{\pi}_{ig}^{[k]} = \frac{\widehat{\pi}_g^{[k-1]} f_g(\mathbf{x_i})}{\sum_{g=1}^{G}\widehat{\pi}_g^{[k-1]} f_g(\mathbf{x_i})}.} \tag{3.41}$$

Reemplazando (3.41) en (3.40) obtenemos:

$$L^*\big(\theta|\mathbf{Y}\big) = \sum_{i=1}^{n}\sum_{g=1}^{G}\widehat{\pi}_{ig}^{[k]}\ln[\pi_g] + \sum_{i=1}^{n}\sum_{g=1}^{G}\widehat{\pi}_{ig}^{[k]}\ln\big[f_g\big(\mathbf{x_i}\big)\big]. \tag{3.42}$$

- **Paso M:**

$$\begin{aligned}\widehat{\theta}^{[k]} &= \underset{\theta}{\text{máx}}\Big\{L^*(\theta|\mathrm{Y})\Big\} \\ &= \underset{\theta}{\text{máx}}\left\{\sum_{i=1}^{n}\sum_{g=1}^{G}\widehat{\pi}_{ig}^{[k]}\ln[\pi_g] + \sum_{i=1}^{n}\sum_{g=1}^{G}\widehat{\pi}_{ig}^{[k]}\ln\big[f_g\big(\mathbf{x_i}\big)\big]\right\} \end{aligned} \tag{3.43}$$

En este paso se maximiza la verosimilitud hallada en (3.42) respecto al vector de parámetros θ, el cual podemos reescribir así: $\theta = \big(\pi_1,\ldots,\pi_G;\mu_1,\ldots,\mu_G;\mathbf{V}_1,\ldots,\mathbf{V}_G\big)$. El objetivo de esta maximización es encontrar los parámetros componentes del vector de parámetros θ, para ello derivamos $L^*\big(\theta|\mathbf{Y}\big)$ respecto a π_g, μ_g, y $\mathbf{V}_g$, con $g = 1,\ldots,G$. Hallemos la función score para (3.42):

$$\begin{aligned} Z^*\big(\theta|\mathbf{Y}\big) &= \frac{\partial}{\partial\theta}\Big\{L^*\big(\theta|\mathbf{Y}\big)\Big\} \\ &= \frac{\partial}{\partial\theta}\left\{\sum_{i=1}^{n}\sum_{g=1}^{G}\widehat{\pi}_{ig}^{[k]}\ln[\pi_g] + \sum_{i=1}^{n}\sum_{g=1}^{G}\widehat{\pi}_{ig}^{[k]}\ln\big[f_g\big(\mathbf{x_i}\big)\big]\right\} \\ &= \sum_{i=1}^{n}\frac{\partial}{\partial\theta}\left\{\sum_{g=1}^{G}\widehat{\pi}_{ig}^{[k]}\ln[\pi_g]\right\} + \sum_{i=1}^{n}\frac{\partial}{\partial\theta}\left\{\sum_{g=1}^{G}\widehat{\pi}_{ig}^{[k]}\ln\big[f_g\big(\mathbf{x_i}\big)\big]\right\} \end{aligned} \tag{3.44}$$

Encontremos los parámetros π_g, μ_g, y $\mathbf{V}_g$ para $g = 1,\ldots,G$.

1. Observemos que π_g aparece solo en el primer termino de las sumas tanto de (3.43) como (3.44), luego podemos omitir las segundas partes. Recordemos que estos parámetros están sujetos a la restricción $\sum_{g=1}^{G}\pi_{ig} = 1$, por lo que introducimos un multiplicador de lagrange. La función score bajo las consideraciones dichas anteriormente e igualada a cero nos permitirá encontrar el parámetro, esto es:

$$\begin{aligned} Z_\lambda^*\big(\pi_g|\mathbf{Y}\big) &= \sum_{i=1}^{n}\frac{\partial}{\partial\pi_g}\left\{\sum_{g=1}^{G}\widehat{\pi}_{ig}^{[k]}\ln[\pi_g]\right\} - \frac{\partial}{\partial\pi_g}\left\{\lambda\left(\sum_{g=1}^{G}\pi_{ig} - 1\right)\right\} = 0 \\ &= \sum_{i=1}^{n}\widehat{\pi}_{ig}^{[k]}\left\{\frac{\partial\ln[\pi_g]}{\partial\pi_g}\right\} - \lambda\left\{\frac{\partial}{\partial\pi_g}\left(\sum_{g=1}^{G}\pi_{ig}\right)\right\} = 0 \\ &= \sum_{i=1}^{n}\frac{\widehat{\pi}_{ig}^{[k]}}{\pi_g} - \lambda = 0, \end{aligned}$$

entonces:

$$\lambda\pi_g = \sum_{i=1}^{n} \widehat{\pi}_{ig}^{[k]} \ ,$$

que es precisamente la igualdad (3.18) con la cual dedujimos que $\lambda = n$ y posteriormente la ecuación (3.21) que nos da el estimador MV de π_g. Luego:

$$\boxed{\widehat{\pi}_{g}^{[k]} = \frac{1}{n}\sum_{i=1}^{n} \widehat{\pi}_{ig}^{[k]}.} \tag{3.45}$$

2. Para hallar μ_g nótese que la las ecuaciones (3.43) y (3.44) no contiene el termino en sus primeros sumandos, luego los podemos omitir para así igualar la función score (3.44) a cero y encontrar el valor de μ_g. Esto es:

$$\begin{aligned} Z^*\big(\mu_g|\mathbf{Y}\big) &= \sum_{i=1}^{n} \frac{\partial}{\partial\mu_g}\left\{\sum_{g=1}^{G} \widehat{\pi}_{ig}^{[k]} \ln\big[f_g\big(\mathbf{x_i}\big)\big]\right\} = 0 \\ &= \sum_{i=1}^{n}\left\{\widehat{\pi}_{ig}^{[k]} \frac{\partial}{\partial\mu_g}\Big(\ln\big[f_g\big(\mathbf{x_i}\big)\big]\Big)\right\} = 0 \\ &= \sum_{i=1}^{n}\left\{\widehat{\pi}_{ig}^{[k]}\left(\frac{\frac{\partial}{\partial\mu_g}\big[f_g\big(\mathbf{x_i}\big)\big]}{f_g\big(\mathbf{x_i}\big)}\right)\right\} = 0, \end{aligned}$$

por el resultado dado en (3.23) tenemos:

$$\begin{aligned} Z^*\big(\mu_g|\mathbf{Y}\big) &= \sum_{i=1}^{n}\left\{\widehat{\pi}_{ig}^{[k]}\left(\frac{f_g\big(\mathbf{x_i}\big)\mathbf{V}_g^{-1}(\mathbf{x_i}-\mu_g)}{f_g\big(\mathbf{x_i}\big)}\right)\right\} = 0 \\ &= \sum_{i=1}^{n}\Big\{\widehat{\pi}_{ig}^{[k]}\mathbf{V}_g^{-1}(\mathbf{x_i}-\mu_g)\Big\} = 0. \end{aligned} \tag{3.46}$$

Según el análisis realizado para encontrar el estimador MV de μ_g, después de la ecuación (3.23), la igualdad dada en (3.46) nos conduce a su estimador MV dado en (3.24). Luego:

$$\boxed{\widehat{\mu}_{g}^{[k]} = \frac{1}{\sum_{i=1}^{n}\widehat{\pi}_{ig}^{[k]}}\sum_{i=1}^{n} \widehat{\pi}_{ig}^{[k]}(\mathbf{x_i}).} \tag{3.47}$$

3. Por la misma razón del numeral (2), igualamos la función score a cero para poder encontrar el estimador MV de $\mathbf{V}_g$. Esto es:

$$\begin{aligned} Z^*\big(\mathbf{V}_g|\mathbf{Y}\big) &= \sum_{i=1}^{n} \frac{\partial}{\partial \mathbf{V}_g}\left\{\sum_{g=1}^{G} \widehat{\pi}_{ig}^{[k]} \ln\big[f_g\big(\mathbf{x_i}\big)\big]\right\} = 0 \\ &= \sum_{i=1}^{n}\left\{\widehat{\pi}_{ig}^{[k]} \frac{\partial}{\partial \mathbf{V}_g}\Big(\ln\big[f_g\big(\mathbf{x_i}\big)\big]\Big)\right\} = 0 \\ &= \sum_{i=1}^{n}\left\{\widehat{\pi}_{ig}^{[k]}\left(\frac{\frac{\partial}{\partial \mathbf{V}_g}\big[f_g\big(\mathbf{x_i}\big)\big]}{f_g\big(\mathbf{x_i}\big)}\right)\right\} = 0, \end{aligned}$$

por el resultado dado en (3.27) tenemos:

$$\begin{aligned} Z^*\big(\mu_g|\mathbf{Y}\big) &= \sum_{i=1}^{n}\left\{\widehat{\pi}_{ig}^{[k]}\left(\frac{\left[\frac{f_g(\mathbf{x_i})}{2\mathbf{V}_g^2}\big\{(\mathbf{x_i}-\widehat{\mu}_g^{[k]})(\mathbf{x_i}-\widehat{\mu}_g^{[k]})' - \mathbf{V}_g\big\}\right]}{f_g\big(\mathbf{x_i}\big)}\right)\right\} = 0 \\ &= \sum_{i=1}^{n}\left\{\widehat{\pi}_{ig}^{[k]}\left[\frac{(\mathbf{x_i}-\widehat{\mu}_g^{[k]})(\mathbf{x_i}-\widehat{\mu}_g^{[k]})' - \mathbf{V}_g}{2\mathbf{V}_g^2}\right]\right\} = 0 \\ &= \sum_{i=1}^{n}\left\{\widehat{\pi}_{ig}^{[k]}\Big[(\mathbf{x_i}-\widehat{\mu}_g^{[k]})(\mathbf{x_i}-\widehat{\mu}_g^{[k]})' - \mathbf{V}_g\Big]\right\} = 0. \end{aligned} \tag{3.48}$$

De igual forma, la ecuación (3.48) nos conduce al estimador MV de $\mathbf{V}_g$, el cual esta dado por:

$$\boxed{\widehat{\mathbf{V}}_g^{[k]} = \frac{1}{\sum_{i=1}^{n}\widehat{\pi}_{ig}^{[k]}}\sum_{i=1}^{n}\widehat{\pi}_{ig}^{[k]}(\mathbf{x_i}-\widehat{\mu}_g^{[k]})(\mathbf{x_i}-\widehat{\mu}_g^{[k]})'.} \tag{3.49}$$

En resumen, en el **paso M** hallamos:

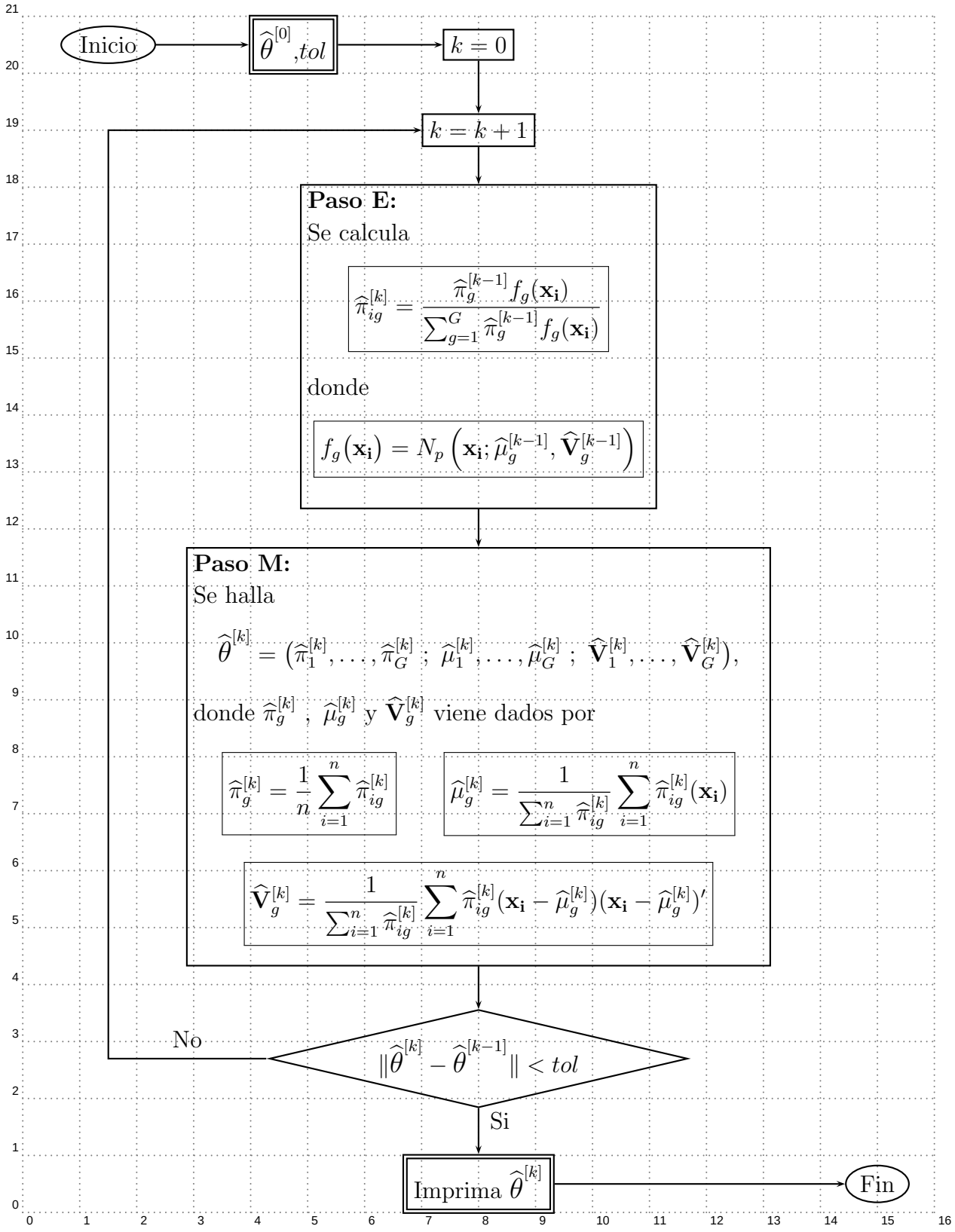
$$\widehat{\theta}^{[k]} = \big(\widehat{\pi}_1^{[k]},\ldots,\widehat{\pi}_G^{[k]}\ ;\ \widehat{\mu}_1^{[k]},\ldots,\widehat{\mu}_G^{[k]}\ ;\ \widehat{\mathbf{V}}_1^{[k]},\ldots,\widehat{\mathbf{V}}_G^{[k]}\big),$$

donde $\widehat{\pi}_g^{[k]}$, $\widehat{\mu}_g^{[k]}$ y $\widehat{\mathbf{V}}_g^{[k]}$ viene dados por (3.45), (3.47) y (3.49) respectivamente.

- **Se evalua** $\|\widehat{\theta}^{[k]} - \widehat{\theta}^{[k-1]}\| < tol.$

Si la comparacion es cierta, nuestro estimador máximo verosímil de los parámetros es $\widehat{\theta}^{[k]}$. Si la diferencia no es suficientemente pequeña se retorna a $\mathbf{k} = \mathbf{k} + \mathbf{1}$ y se incrementa este contador, seguimos con el **paso E** y repetimos el proceso hasta lograr la convergencia.

Veamos este algoritmo en el siguiente diagrama de flujo.


Inicio
$\widehat{\theta}^{[0]}$,tol
$k = 0$
$k = k + 1$
Paso E:
Se calcula
$\widehat{\pi}_{ig}^{[k]} = \dfrac{\widehat{\pi}_g^{[k-1]} f_g(\mathbf{x_i})}{\sum_{g=1}^{G} \widehat{\pi}_g^{[k-1]} f_g(\mathbf{x_i})}$
donde
$f_g(\mathbf{x_i}) = N_p\left(\mathbf{x_i}; \widehat{\mu}_g^{[k-1]}, \widehat{\mathbf{V}}_g^{[k-1]}\right)$
Paso M:
Se halla
$\widehat{\theta}^{[k]} = (\widehat{\pi}_1^{[k]}, \ldots, \widehat{\pi}_G^{[k]}\ ;\ \widehat{\mu}_1^{[k]}, \ldots, \widehat{\mu}_G^{[k]}\ ;\ \widehat{\mathbf{V}}_1^{[k]}, \ldots, \widehat{\mathbf{V}}_G^{[k]})$,
donde $\widehat{\pi}_g^{[k]}$, $\widehat{\mu}_g^{[k]}$ y $\widehat{\mathbf{V}}_g^{[k]}$ viene dados por
$\widehat{\pi}_g^{[k]} = \dfrac{1}{n} \sum_{i=1}^{n} \widehat{\pi}_{ig}^{[k]}$
$\widehat{\mu}_g^{[k]} = \dfrac{1}{\sum_{i=1}^{n} \widehat{\pi}_{ig}^{[k]}} \sum_{i=1}^{n} \widehat{\pi}_{ig}^{[k]}(\mathbf{x_i})$
$\widehat{\mathbf{V}}_g^{[k]} = \dfrac{1}{\sum_{i=1}^{n} \widehat{\pi}_{ig}^{[k]}} \sum_{i=1}^{n} \widehat{\pi}_{ig}^{[k]}(\mathbf{x_i} - \widehat{\mu}_g^{[k]})(\mathbf{x_i} - \widehat{\mu}_g^{[k]})'$
No
$\|\widehat{\theta}^{[k]} - \widehat{\theta}^{[k-1]}\| < tol$
Si
Imprima $\widehat{\theta}^{[k]}$
Fin

Capítulo 4

SOFTWARE PARA MEZCLAS NORMALES MULTIVARIANTES

4.1. Introducción

Para facilitar el estudio de las mezclas finitas de distribuciones normales multivariantes, se han venido desarrollando estos últimos años una serie de softwares que permiten no solo realizar la estimación de los parámetros de densidades componentes y la clasificación de los datos en los clusters así determinados, si no la asignación de nuevos datos a través de algoritmos basados en el análisis discriminante. Dentro de estos softwares se destacan el **MCLUST**, que es el paquete informático mas difundido para realizar esta clase de análisis. El **MCLUST** fue desarrollado por Fraley y Raftery (1999) para el software comercial **S-PLUS**. Una página web donde se encuentra disponible este software es: *http://www.stat.washington.edu/mclust*. Otros softwares para el ajuste de datos a las mezclas de normales multivariantes con G componentes son el **EMMIX** desarrollado por McLachlan, Peel, Basford y Adams (1999) y el **AutoClass** de Cheeseman y Stutz (1996), cuya ultima version se encuentra en la página web *http://ic.arc.nasa.gov/ic/projects/bayes-group/autoclass/autoclass-c-program.html*.
En este capitulo se presenta un software o aplicación compatible con el software comercial **MATLAB**; este es el **MBC Toolbox**. El principal objetivo aquí es mostrar la necesidad del uso del ordenador en el proceso de clasificación de los datos mezclados provenientes de poblaciones normales multivariantes, luego se aclará que la escogencia de este software es solo por la facilidad que le brinda al usuario en su manipulación y la posterior presentación

de resultados del proceso de clasificación.

4.2. MBC Toolbox: Una aplicación en $MATLAB$ para modelos basados en agrupamientos.

4.2.1. Generalidades

El **MBC**[1] es una aplicación en $MATLAB$ para la resolución de problemas de mezclas en el análisis de agrupamiento de datos, basado en las mezclas finitas de distribuciones normales multivariantes:

$$\mathbf{G}(\mathbf{X}) = \sum_{g=1}^{G} \pi_g f_g(\mathbf{X}).$$

El **MBC** fue desarrollado en Enero de 2003 por Wendy L. Martinez y Angel R. Martinez, integrantes de la Oficina de Investigación del Centro de Guerra Naval de los Estados Unidos, División Dahlgren. El nombre original de esta aplicación es *"Model-Based Clustering Toolbox"* y solo se ha hecho publica la version 1.0.
La idea general en el **MBC** es generar estimaciones de los parámetros π_g, μ_g y $\mathbf{V}_g$, con $g = 1, \ldots, G$, basados en la implementación del **algoritmo EM para mezclas** y en ciertas restricciones impuestas a las matrices de covarianzas $\mathbf{V}_g$. La mejor estimación y el mejor modelo (es decir en número de componentes, parámetros estimados y en la forma de las matrices de covarianza) es escogido según el modelo que posea mayor valor del Criterio de Información Bayesiano (**BIC**).
Como ya habíamos visto en la sección 1.4.1, la forma cuadrática $(\mathbf{x_i} - \mu)'\mathbf{V}^{-1}(\mathbf{x_i} - \mu)$ (contenida en el exponente de la distribución normal multivariante y que equivale a la *la distancia de Mahalanobis* entre el individuo $\mathbf{x_i}$ y el vector de medias μ) determina la forma en que se distribuyen los datos; la matriz de covarianzas $\mathbf{V}$ proporciona las características geométricas importantes de la distribución $f_g(\mathbf{X})$. Luego las matrices $\mathbf{V}_g$ son la clave del algoritmo utilizado por la aplicación **MBC**.

4.2.2. Condiciones sobre las matrices de covarianzas

Banfield y Raftery en el articulo de [1993] titulado *"Model-based Gaussian and non-Gaussian clustering"*, publicado en *Biometrics 49:803-821*, desarrollaron un modelo basa-

[1]Esta aplicación es para datos multivariantes ($p \geq 2$). Para una mejor comprensión de esta véase [7].

do en agrupamientos por la parametrización de la matriz de covarianzas en términos de la descomposición espectral[2], como sigue:

$$\mathbf{V}_g = \lambda_g \mathbf{U}_g \mathbf{D}_g \mathbf{U}'_g, \tag{4.1}$$

donde $\lambda_g \mathbf{D}_g$ es una matriz diagonal formada por los valores propios de $\mathbf{V}_g$ (siendo el escalar λ_g el mayor valor propio de la matriz) y $\mathbf{U}_g$ es una matriz ortogonal cuyas columnas son los vectores propios unitarios asociados con los elementos de la diagonal de la matriz $\mathbf{D}_g$. Usando esta descomposición se pueden establecer las siguientes propiedades geométricas de la distribución:

- **La orientación** determinada por los vectores propios dados en la matriz $\mathbf{U}_g$.
- **La forma** de la distribución determinada por la matriz $\mathbf{D}_g$.
- **El volumen** ocupado en el espacio por el grupo g determinado por el escalar λ_g.

En virtud a estas propiedades Banfield y Raftery establecieron 6 modelos parametrizados para el agrupamiento de datos. Estos se especifican a detalle en el cuadro 4.1.

$\mathbf{V}_g$	Distribución	Volumen	Forma	Orientación
$\lambda\mathbf{I}$	Esférica	Igual	Igual	NA
$\lambda_g\mathbf{I}$	Esférica	Variable	Igual	NA
$\lambda\mathbf{U}\mathbf{D}\mathbf{U}'$	Elipsoidal	Igual	Igual	Igual
$\lambda_g\mathbf{U}_g\mathbf{D}_g\mathbf{U}'_g$	Elipsoidal	Variable	Variable	Variable
$\lambda\mathbf{U}_g\mathbf{D}\mathbf{U}'_g$	Elipsoidal	Igual	Igual	Variable
$\lambda_g\mathbf{U}_g\mathbf{D}\mathbf{U}'_g$	Elipsoidal	Variable	Igual	Variable

Cuadro 4.1: Parametrización para modelos basados en agrupamientos

El **MBC** aplica la parametrización de Banfield y Raftery usando solo los cuatro primeros modelos del cuadro 4.1. Estos se describen en el cuadro 4.2

4.2.3. Funcionamiento del MBC

Como se mencionó anteriormente, la aplicación **MBC** utiliza para su funcionamiento **el algoritmo EM para mezclas**. Recordemos que **el algoritmo EM** requiere de una

[2]Véase la sección 1.2.1.

suposición inicial de los parámetros de las densidades componentes y los pesos de mezcla así como el conocimiento del número de componentes en la mezcla; la misma información se necesita para el **MBC**.

♯ del modelo (M)	Covarianza	Modelo	Descripción
1	Esférica e igual	$\widehat{\mathbf{V}}_g = \sigma^2\mathbf{I}$	Las matrices de covarianzas son diagonales. Los elementos de la diagonal tiene el mismo valor. Las matrices de covarianzas son iguales.
2	Esférica y no igual	$\widehat{\mathbf{V}}_g = \sigma_g^2\mathbf{I}$	Las matrices de covarianzas son diagonales. Se permite que las matrices de covarianzas varíen entre grupos. La matriz de covarianzas en cada grupo tiene el mismo valor en sus elementos diagonales.
3	Elipsoidal e igual	$\widehat{\mathbf{V}}_g = \mathbf{V}$	Las matrices de covarianzas pueden tener elementos no ceros fuera de su diagonal principal. La matriz de covarianzas son iguales.
4	Elipsoidal y no igual	$\widehat{\mathbf{V}}_g = \mathbf{V}_g$	Las matrices de covarianzas se hallan según la ecuación (3.49). Las matrices de covarianzas pueden tener elementos no ceros fuera de su diagonal principal. Las matrices de covarianzas varían de grupo en grupo.

Cuadro 4.2: Descripción de los cuatro modelos usados en el **MBC**

En esta aplicación se inicia **el algoritmo EM** con una estimación obtenida por métodos de agrupamientos aglomerativos jerárquicos. En aquellos métodos cada dato puntual inicia como un cluster (grupo), luego los dos más cercanos (en lo referente a la distancia euclídea)

se unen para formar un nuevo cluster, se continúa así el algoritmo hasta agrupar todos los datos en un solo cluster. En la aplicación **MBC** se usa una metodología similar dónde en este caso los clusters (grupos) se unen de tal manera que la función de verosimilitud aumente al máximo; es decir se mide la cercanía de los clusters tomando *las distancias de Mahalanobis* entre los vectores de medias de cada grupo y el vector de medias total, luego los dos mas cercanos forman el nuevo cluster. Se continúa así el algoritmo hasta agrupar todos los datos en un solo cluster. Esta forma de agrupar crea una jerarquía que se observa en gráficos denominados *dendogramas*. [3]

Fraley en el articulo de [1998] titulado *"Algorithms for model-based Gaussian hierarchical clustering "*, publicado en *SIAM Journal on Scientific Computing, 20:270-281*, describe algoritmos aglomerativos jerárquicos para todos los modelos basados en agrupamientos descritos en el cuadro (4.2). En el **MBC** se usa ***el modelo número* 4** del cuadro (4.2), que ***no tiene restricciones*** con su matriz de covarianzas general, para dar inicio a los agrupamientos aglomerativos jerárquicos en todos los modelos; es decir en cada paso se calcula la matriz de covarianzas general con los representantes (vectores de medias) de cada grupo.

Hay que tener especial cuidado con la ecuación que actualiza la matriz de covarianzas en **el algoritmo EM para mezclas**, pues si el modelo ***tiene restricciones*** para $\mathbf{V}_g$, la ecuación (3.49) debe modificarse para el buen funcionamiento del algoritmo; es decir para los modelos 1, 2 y 3 se debe cambiar la ecuación que actualiza $\mathbf{V}_g$ en el **algoritmo EM**. Las ecuaciones de actualización para estos modelos se encuentran en la publicación de Celeux y Govaert [1995] titulada *"Gaussian Parsimonious clustering models."*, publicado en *Pattern Recognition, 28:781-793*.

Lo último que se necesita para garantizar el buen funcionamiento del **MBC** es alguna manera de determinar la mejor estimación (en cuanto al número de componentes y al modelo de la matriz de covarianzas) en el ajuste de los datos. Como previamente se mencionó, la opción del mejor modelo se realiza via **BIC**[4], dado por:

$$\mathbf{BIC} = 2L_M(\theta|\mathbf{X}) - n(p, G)\ln(n), \tag{4.2}$$

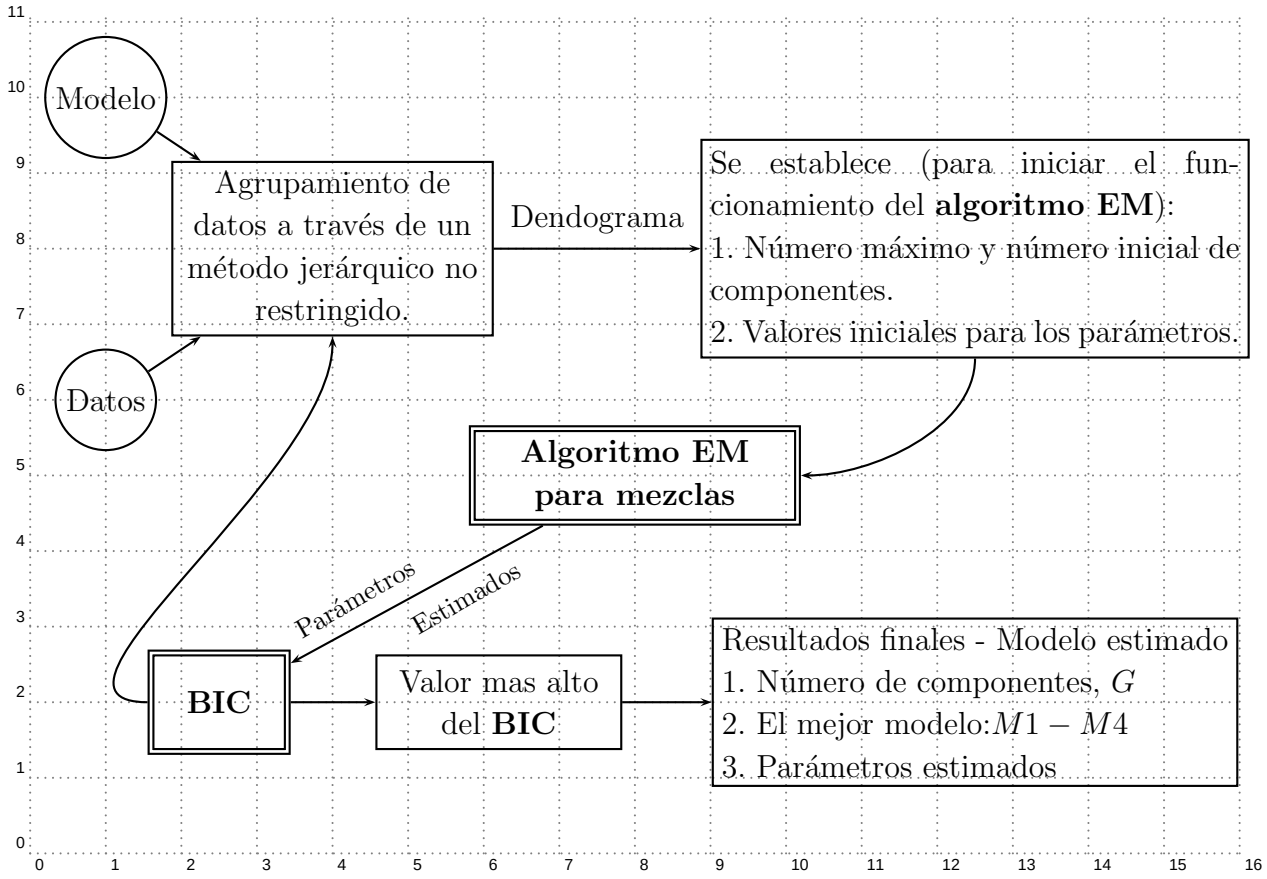
donde $n(p, G)$ es el numero de parámetros y $L_M(\theta|\mathbf{X})$ es el soporte para el modelo M. El modelo y la estimación final corresponderán al valor más alto del **BIC**.

Todos estos elementos se reúnen en el siguiente algoritmo que describe los pasos usados en el **MBC**.

[3] En [1] se hace una pequeña descripción de estos gráficos.

[4] Véase la sección 2.5.2.

4.2.4. Algoritmo implementado en el MBC



1. Aplíquese un procedimiento aglomerativo jerárquico, ***sin restricciones*** en la matriz de covarianzas, para encontrar un agrupamiento inicial de los datos en el número deseado de clusters.

2. Escoja un modelo: $M = 1, 2, 3, 4$. (vea cuadro 4.2).

3. Escoja el número de grupos o densidades componentes, G.

4. Encuentre la partición dada por el procedimiento aglomerativo del **paso 1** para el valor dado G.

5. Usando esta partición, encuentre los pesos de la mezcla, los vectores de medias y la matriz de covarianzas para cada grupo. Esta partición se basa en la forma del modelo escogido en el **paso 2**.

6. Usando el G escogido **(paso 3)** y los valores iniciales **(paso 5)**, aplique **el algoritmo EM para mezclas** y obtenga las estimaciones finales.

7. Calcule el valor del **BIC** para este valor de G y M.

8. Vaya al **paso 3** para escoger otro valor de G.

9. Vaya al **paso 2** para escoger otro modelo M.

4.3. Ejemplos de aplicación

4.3.1. Mezcla de 3 distribuciones normales bivariantes

Para ilustrar el procedimiento del **MBC**, consideramos una muestra con $n = 30$ generada por una simulación bivariante de un modelo de mezcla normal con parámetros

$$\mu_1 = \begin{pmatrix} -3 \\ -3 \end{pmatrix}, \quad \mu_2 = \begin{pmatrix} 0 \\ 3 \end{pmatrix}, \quad \mu_3 = \begin{pmatrix} 4 \\ 0 \end{pmatrix}; \quad \mathbf{V} = \mathbf{V}_1 = \mathbf{V}_2 = \mathbf{V}_3 = \begin{pmatrix} 1 & -0.5 \\ -0.5 & 1 \end{pmatrix},$$

y pesos de mezcla $\pi_1 = 0.3, \pi_2 = 0.2$ y $\pi_3 = 0.5$.
La simulación de los datos se genera tecleando en el *Command Window* de $MATLAB$ la instrucción "genmix ",[5] la cual despliega la ventana mostrada en la figura 4.1. Seguimos los pasos allí indicados, introduciendo así los parámetros y generando en $MATLAB$ los datos que guardaremos en el archivo ejemplo1_mbc.mat. Estos datos obtenidos se observan en el cuadro 4.3.
La ventana mostrada en 4.1 nos brinda la opción de ver los datos en una *plotmatrix* (matriz gráfica), que hace un *scatterplot* (gráfico de dispersion) sobre las variables para observar la variabilidad de los datos. De igual forma se puede observar en el plano $\mathbb{R}^2$ la distribución de los datos. La plotmatrix se observa en la figura 4.2 y el plano en $\mathbb{R}^2$ se observa en la figura 4.3.
Ahora agruparemos estos datos ignorando los parámetros que los generaron para así ajustarlos a una mezcla de tres componentes normales. Para ello seguiremos el procedimiento utilizado en el **MBC** y descrito anteriormente en el diagrama de flujo o en la sección 3.6.4.

[5]Previamente se debe instalar la aplicación *Model-Based Clustering Toolbox Version 1.0*. El proceso de instalación puede consultarse en [7].

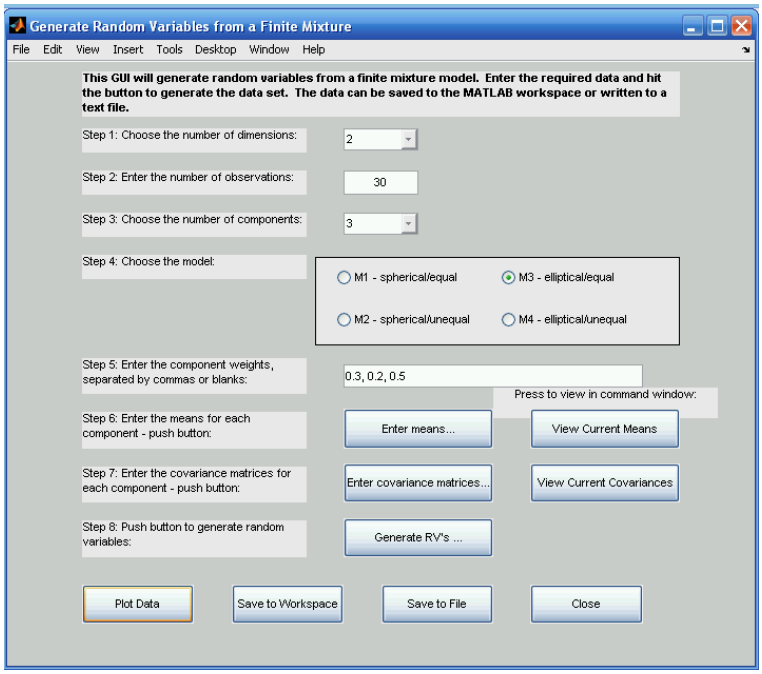


Figura 4.1: Esta ventana muestra el GUI que se invoca con genmix. Al lado izquierdo de la ventana se muestran los pasos que deben seguirse para generar los datos.

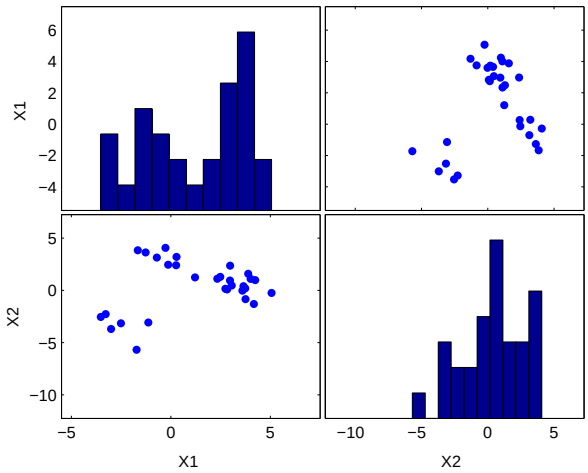


Figura 4.2: Plotmatrix generada por la orden PlotData en el GUI mostrado la figura 4.1 para 30 datos normalmente distribuidos con los parámetros estipulados.

Dato	X1	X2	Dato	X1	X2
1	−2.5145	−3.1532	16	2.3360	1.1061
2	−3.0050	−3.6969	17	2.9710	0.9473
3	−3.2762	−2.2726	18	4.2431	0.9853
4	−1.7235	−5.6860	19	2.7434	0.1539
5	−1.1366	−3.0742	20	3.6528	0.3995
6	−3.5226	−2.5491	21	3.0586	0.4593
7	0.2617	2.4005	22	2.8254	0.0848
8	1.2134	1.2384	23	2.9789	2.3607
9	−0.2747	4.0663	24	3.5983	−0.0223
10	−0.1331	2.4499	25	4.1737	−1.3075
11	−1.2705	3.6255	26	3.8839	1.5910
12	−1.6636	3.8311	27	5.0641	−0.2501
13	−0.7036	3.1358	28	3.7546	−0.8464
14	0.2809	3.2030	29	2.4825	1.2960
15	3.7360	0.1938	30	4.0097	1.0948

Cuadro 4.3: Los 30 datos generados por en el GUI mostrado la figura 4.1. Estos datos simulan una mezcla de tres componentes normales bivariantes.

Lo primero es realizar el agrupamiento aglomerativo jerárquico. Este se lleva acabo digitando en el *Command Window* de $MATLAB$ la función "agmbclust " , la cual produce el siguiente resultado:

```
>> % Iniciamos importando los datos
>> data = load('ejemplo1_mbc.mat')
data =
    Xa: [30x2 double]
>> % Se hace el agrupamiento aglomerativo jerárquico.
>> Z=agmbclust(Xa)
Merging clusters ... step 1
Merging clusters ... step 2
          .
          .
          .
Merging clusters ... step 26
Merging clusters ... step 27
```

```
Z =
   19.0000   22.0000   54.4988
   15.0000   20.0000   54.5028
   16.0000   29.0000   54.5075
   18.0000   30.0000   54.5129
    3.0000    6.0000   54.5240
   24.0000   32.0000   54.5360
    7.0000   10.0000   54.5488
   11.0000   12.0000   54.5648
   17.0000   21.0000   54.5847
   25.0000   28.0000   54.6160
   26.0000   34.0000   54.6551
    1.0000    2.0000   54.6982
   31.0000   39.0000   54.7621
   14.0000   37.0000   54.8335
    9.0000   13.0000   54.9171
    8.0000   33.0000   55.0690
   38.0000   45.0000   55.2410
   36.0000   43.0000   55.4262
   27.0000   40.0000   55.6350
   35.0000   42.0000   55.8763
   23.0000   41.0000   56.1613
    4.0000    5.0000   56.6725
   44.0000   47.0000   57.3248
   46.0000   48.0000   58.1152
   51.0000   54.0000   59.2000
   50.0000   52.0000   60.5356
   49.0000   55.0000   62.4980
   53.0000   57.0000   76.5650
   56.0000   58.0000  119.3946
```

El "agmbclust" genera la matriz de rendimiento Z, la cual contiene en el primera columna los clusters que se unen con los clusters de la segunda columna a una distancia dada en la tercera columna; al unirse dos clusters, forman uno nuevo y "agmbclust" lo etiqueta con un número superior a n. La matriz Z permite gráficar el dendograma dado en la figura 4.4. Ahora, si determinamos que los clusters son $G = 3$, en el dendograma se sugiere la clasificación de los individuos según lo mostrado en la figura 4.5.

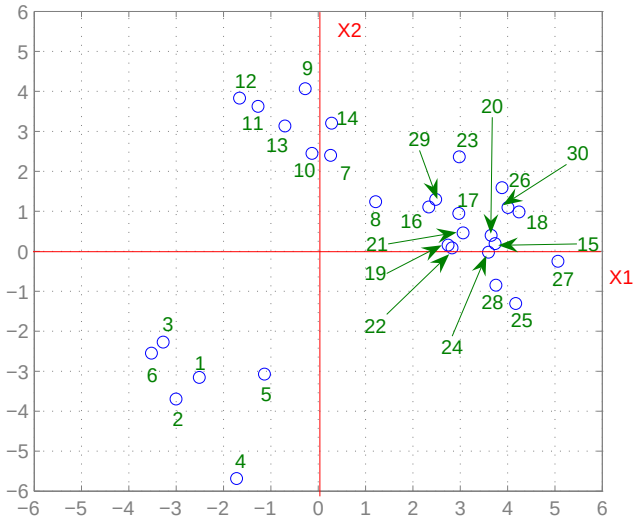


Figura 4.3: Ubicación en $\mathbb{R}^2$ de los 30 datos generados aleatoriamente.

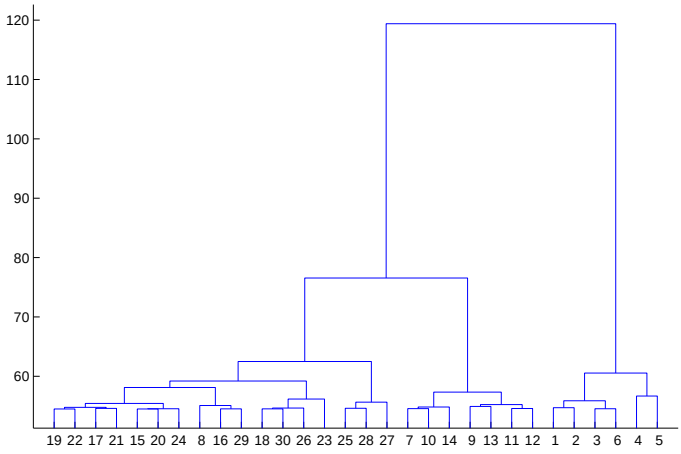


Figura 4.4: Este es el dendograma para los 30 datos del agrupamiento aglomerativo. El eje vertical representa la distancia a la cual se unen los clusters. Esta distancia se estipula en la matriz Z. El dendograma se genera con la instrucción dendrogram(Z).

La matriz del rendimiento Z también puede usarse en la función de $MATLAB$ llamada "cluster". Esta se encuentra disponible en la Caja de herramientas Estadísticas y devuelve un vector T de n etiquetas que indican el número del cluster para cada observación. Esta función contiene algunas opciones de entrada como:

```
>>T = cluster(Z,'maxclust',k)
```

donde k es el número máximo de clusters a formar para el árbol jerárquico en Z. Para nuestro ejemplo obtenemos el siguiente resultado

```
>> T=cluster(Z,'maxclust',3);
>> Trans_T=T'
Trans_T =
  Columns 1 through 12
     3     3     3     3     3     3     1     2     1     1     1     1
  Columns 13 through 24
     1     1     2     2     2     2     2     2     2     2     2     2
  Columns 25 through 30
     2     2     2     2     2     2
```

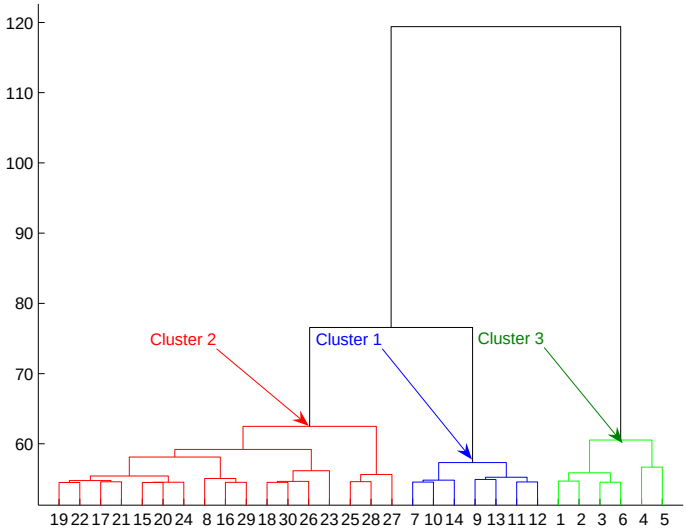


Figura 4.5: Clasificación aglomerativa para los 30 datos con $G = 3$

El **MBC** trae incorporado *los gráficos de rectángulo* desarrollados por Samuel S. Wills en 1998, los cuales se basan en los despliegues de mapa de árbol de Jhnoson Y Shneiderman de 1991. Este método de *los gráficos de rectángulo* trabaja con el rendimiento del agrupamiento jerárquico (Z) y despliega una serie de rectángulos anidados donde el rectángulo padre (o raíz del árbol) se da por la área entera del despliegue. los rectángulos pequeños (sub-rectángulo) se obtiene por subdivisiones recursivas del rectángulo padre dónde el tamaño (Area) de cada sub-rectángulo es proporcional al tamaño del nodo (valor en la tercera columna de la matriz Z). El gráfico de rectángulo de Wills divide el rectángulo padre a lo largo del lado más largo y siguen dividiéndose hasta alcanzar un nodo o hasta que la distancia de corte sea alcanzada. El método del gráfico de rectángulo esta incluido en esta aplicación como la función "rectplot ". La sintaxis básica para esta es

```
>> rectplot(Z,nc,T)
```

Esta función usa la matriz Z que se usa para crear dendograma. La variable de entrada nc representa el número de clusters o rectángulos para incluir en el gráfico. El tercer argumento T es opcional que es un vector n-dimensional que contiene las etiquetas de la clase para las observaciones conocidas. Para la muestra de 30 datos generada anteriormente obtiene los gráficos de rectángulos dados en las figuras 4.6 y 4.7.

```
>> rectplot(Z,3)
```

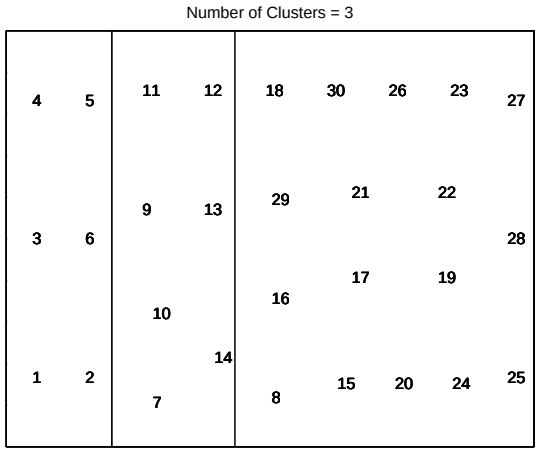


Figura 4.6: Gráfico de rectángulo para $nc = 3$

```
>> rectplot(Z,30)
```

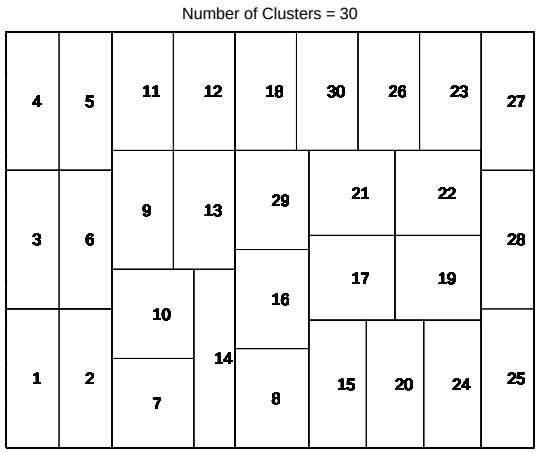


Figura 4.7: Gráfico de rectángulo para $nc = 30$

Continuando con el algoritmo del **MBC**, el paso a seguir es determinar $\pi_g, \mu_g, \mathbf{V}_g$ para el agrupamiento jerárquico hallado anteriormente. Esto se realiza ejecutando en $MATLAB$ las siguientes sentencias:

```
>> %Inicialmente se importan los datos
>> data = load('ejemplo1_mbc.mat')
data =
    Xa: [30x2 double]
>> % Creamos las matrices C1, C2 y C3 que contendrán los elementos de los
clusters encontrados en el agrupamiento aglomerativo.
>> C1=[Xa(7,:);Xa(9,:);Xa(10,:);Xa(11,:);Xa(12,:);Xa(13,:);Xa(14,:)]
C1 =
    0.2617    2.4005
   -0.2747    4.0663
   -0.1331    2.4499
   -1.2705    3.6255
   -1.6636    3.8311
   -0.7036    3.1358
    0.2809    3.2030
```

```
>> C2=[Xa(8,:);Xa(15,:);Xa(16,:);Xa(17,:);Xa(18,:);Xa(19,:);Xa(20,:);
Xa(21,:);Xa(22,:);Xa(23,:);Xa(24,:);Xa(25,:);Xa(26,:);Xa(27,:);Xa(28,:);
Xa(29,:);Xa(30,:)]
C2 =
    1.2134    1.2384
    3.7360    0.1938
    2.3360    1.1061
    2.9710    0.9473
    4.2431    0.9853
    2.7434    0.1539
    3.6528    0.3995
    3.0586    0.4593
    2.8254    0.0848
    2.9789    2.3607
    3.5983   -0.0223
    4.1737   -1.3075
    3.8839    1.5910
    5.0641   -0.2501
    3.7546   -0.8464
    2.4825    1.2960
    4.0097    1.0948
>> C3=[Xa(1,:);Xa(2,:);Xa(3,:);Xa(4,:);Xa(5,:);Xa(6,:)]
C3 =
   -2.5145   -3.1532
   -3.0050   -3.6969
   -3.2762   -2.2726
   -1.7235   -5.6860
   -1.1366   -3.0742
   -3.5226   -2.5491
>> %Encontremos los vectores de medias muestrales para cada cluster
>> Tra_C1bar=mean(C1); C1bar =(Tra_C1bar)'
C1bar =
   -0.5004
    3.2446
>> Tra_C2bar=mean(C2); C3bar =(Tra_C2bar)'
C2bar =
```

```
    3.3368
    0.5579
>> Tra_C3bar=mean(C3); C1bar =(Tra_C3bar)'
C3bar =
   -2.5297
   -3.4054
>> % Guardemos estos vectores en una matriz cuyas columnas representen los
vectores de medias de cada cluster.
>> Cbar=[Tra_C1bar;Tra_C2bar;Tra_C3bar]
Cbar =
   -0.5004    3.2446
    3.3368    0.5579
   -2.5297   -3.4054
>> Tra_Cbar=Cbar'
Tra_Cbar =
   -0.5004    3.3368   -2.5297
    3.2446    0.5579   -3.4054
>> % Ahora hallemos las matrices de covarianza muestral para cada cluster.
>> SC1=Cov(C1)
Warning: Could not find an exact (case-sensitive) match for 'Cov'.
C:\Archivos de programa\MATLAB71\toolbox\matlab\datafun\cov.m is a
case-insensitive match and will be used instead. You can improve the
performance of your code by using exact name matches and we
therefore recommend that you update your usage accordingly.
Alternatively, you can disable this warning using
warning('off','MATLAB:dispatcher:InexactMatch').
SC1 =
    0.5607   -0.2893
   -0.2893    0.4203
>> SC2=Cov(C2)
SC2 =
    0.8095   -0.3425
   -0.3425    0.8278
>> SC3=Cov(C3)
SC3 =
    0.8720   -0.5861
```

```
   -0.5861    1.4952
>> % Guardamos estos datos en una hipermatriz, es decir una matriz de mas de
dos dimensiones.
>> SC(:,:,1)=SC1;
>> SC(:,:,2)=SC2;
>> SC(:,:,3)=SC3;
>> SC
SC(:,:,1) =
    0.5607   -0.2893
   -0.2893    0.4203
SC(:,:,2) =
    0.8095   -0.3425
   -0.3425    0.8278
SC(:,:,3) =
    0.8720   -0.5861
   -0.5861    1.4952
>> % Finalmente asignemos los pesos de la mezcla. Es muy sano suponer
inicialmente que los pesos en cada componente son iguales, claro si
no hay lugar a contradicciones.
>> Pi=[0.3333 0.3333 0.3334]
Pi =
    0.3333    0.3333    0.3334
```

Nótese que en calculo de los parámetros para los clusters formados en el agrupamiento aglomerativo, al hallar las matrices de covarianza se genero el mensaje de advertencia:

```
Warning: Could not find an exact (case-sensitive) match for 'Cov'.
C:\Archivos de programa\MATLAB71\toolbox\matlab\datafun\cov.m is a
case-insensitive match and will be used instead. You can improve the
performance of your code by using exact name matches and we
therefore recommend that you update your usage accordingly.
Alternatively, you can disable this warning using
warning('off','MATLAB:dispatcher:InexactMatch').
```

que nos previene de la posible inexactitud en el calculo de las matrices de covarianza debido al origen de los datos. En la figura 4.8 se visualizan en el plano $\mathbb{R}^2$ los clusters aglomerativos con sus medias muestrales respectivas.

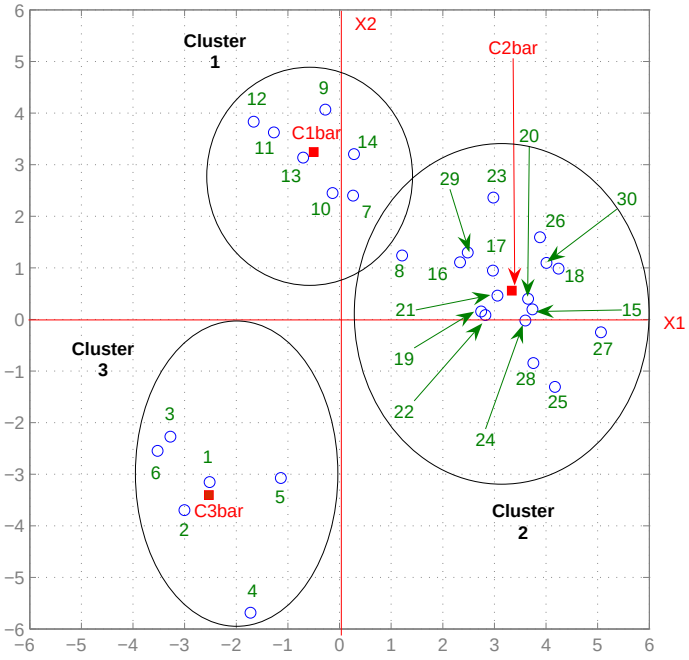


Figura 4.8: Clusters formados en el agrupamiento aglomerativo con sus respectivos vectores de medias muestrales.

El **Algoritmo EM para mezclas** aplicado en el **MBC**, para cada uno de los modelos dados en el cuadro 4.2, se obtiene en forma detallada gracias a la función "mbcfinmix " que se encuentra en esta aplicación. La sintaxis básica para esta función es

```
>> [wts,mus,vars] = mbcfinmix(data,muin,varin,wtsin,model);
```

La función "mbcfinmix " devuelve estimaciones del modelo con los parámetros: pesos, medias, y covarianzas. El argumento wts es un vector que contiene los G pesos, uno para cada término. La variable mus es una matriz $d \times G$: dónde cada columna corresponde a una media de la componente. Recuérdese que G es el número de componentes en la mezcla, y d es la dimensionalidad de los datos. La variable vars es un arreglo 3-D dónde cada página (es decir, tercera dimensión) corresponde a una matriz de covarianzas. Así, las dimensiones de vars son $d \times d \times G$. Los argumentos de entrada muin, varin, y wtsin son similares en forma y contienen los valores iniciales para estos parámetros. En el caso del agrupamiento aglomerativo, éstos se obtienen de las particiones (clusters) allí encontrados. En general, estos valores iniciales pueden ser de cualquier fuente razonable. La entrada de la variable data es una matriz de $n \times d$ que contiene las observaciones, y model es una

variable numérica que indica uno de los 4 modelos básicos (véase el cuadro 4.2). Para nuestro ejemplo de los 30 datos bidimensionales, esta función se ejecuta en el *Command Window* de $MATLAB$ de la siguiente manera:

```
>> % Previamente se cargan los datos y se toman los arreglos Tra_Xabar,
SC, y Pi hallados anteriormente para el agrupamiento aglomerativo.
>> % Hallemos los parámetros de las componentes de la mezcla para G=3
y el modelo 1.
>> [wst,mus,vars]=mbcfinmix(Xa,Tra_Cbar,SC,Pi,1)
Warning: Could not find an exact (case-sensitive) match for
'mbcfinmix'. C:\Archivos de
programa\MATLAB71\toolbox\mbctool\mbcfinmix.M is a case- insensitive
match and will be used instead. You can improve the performance of
your code by using exact name matches and we therefore recommend
that you update your usage accordingly. Alternatively, you can
disable this warning using
warning('off','MATLAB:dispatcher:InexactMatch').

wst =
    0.2000    0.2373    0.5627
mus =
   -2.5297   -0.4715    3.3515
   -3.4054    3.2112    0.5532
vars(:,:,1) =
    0.7360         0
         0    0.7360
vars(:,:,2) =
    0.7360         0
         0    0.7360
vars(:,:,3) =
    0.7360         0
         0    0.7360
>> % De igual forma se hallan los parámetros de las componentes de la mezcla
para G=3 y el modelo 2.
>> [wst,mus,vars]=mbcfinmix(Xa,Tra_Cbar,SC,Pi,2)
wst =
    0.2000    0.2334    0.5666
```

```
mus =
   -2.5297   -0.4994    3.3371
   -3.4054    3.2433    0.5579
vars(:,:,1) =
    0.9863         0
         0    0.9863
vars(:,:,2) =
    0.4226         0
         0    0.4226
vars(:,:,3) =
    0.7706         0
         0    0.7706
>> % Los parámetros componentes de la mezcla para G=3 y el modelo 3 son:
>> [wst,mus,vars]=mbcfinmix(Xa,Tra_Cbar,SC,Pi,3)
wst =
    0.2000    0.2561    0.5439
mus =
   -2.5297   -0.3467    3.4250
   -3.4054    3.0657    0.5297
vars(:,:,1) =
    0.6459   -0.3767
   -0.3767    0.8480
vars(:,:,2) =
    0.6459   -0.3767
   -0.3767    0.8480
vars(:,:,3) =
    0.6459   -0.3767
   -0.3767    0.8480
>> % Los parámetros componentes de la mezcla para G=3 y el modelo 4 son:
>> [wst,mus,vars]=mbcfinmix(Xa,Tra_Cbar,SC,Pi,4)
wst =
    0.2000    0.2643    0.5357
mus =
   -2.5297   -0.2946    3.4574
   -3.4054    3.0083    0.5190
vars(:,:,1) =
```

```
    0.7266   -0.4884
   -0.4884    1.2460
vars(:,:,2) =
    0.7487   -0.5858
   -0.5858    0.7385
vars(:,:,3) =
    0.5375   -0.2546
   -0.2546    0.7962
```

El **MBC** proporciona una función llamada "mbclust " que lleva a cabo el procedimiento completo para la clasificación mediante mezclas finitas, incluso el proceso de la inicialización, el **EM**, y la selección del mejor modelo. La sintaxis básica para esta función es:

```
>>[bics,bestmodel,allmodels,Z,clabs]=mbclust(data,maxclus);
```

Como se mencionó anteriormente, la entrada de la variable data es una matriz de $n \times d$ que contiene las observaciones. La variable maxclus es el número aceptable máximo de clusters o densidades componente en la mezcla (es decir, el máximo permitido del valor G). La variable de rendimiento bics es una matriz que contiene todos los valores **BIC** para cada modelo y número de clusters. La variable bics contiene 4 filas y columnas igual al valor maxclus dónde cada fila corresponde a un modelo y cada columna corresponde al número de condiciones o clusters.
La variable bestmodel es una estructura de $MATLAB$ que contiene los parámetros para el mejor modelo, indicado por el valor de **BIC** más alto. La estructura tiene los siguientes campos:

```
bestmodel.pies bestmodel.mus bestmodel.vars
```

La variable allmodels es una estructura de $MATLAB$ que contiene la información sobre todos los modelos. Cada registro (hay 4) de allmodels contiene la información para uno de los modelos. El campo clus es otra estructura dónde cada registro (hay maxclus de ellos) contiene las estimaciones del parámetro para el modelo. Finalmente, la estructura clus contiene 3 campos: pies, mus, vars. Por ejemplo,

```
allmodels(2).clus(5).pies
```

tiene los pesos para modelo 2 con 5 clusters. La estructura de clus es realmente un sub-estructura (o campo) bajo la estructura principal llamada allmodels.
La variable Z es la misma matriz descrita en el agrupamiento aglomerativo. La variable de rendimiento clabs contiene las etiquetas de las n observaciones, como las dadas por bestmodel.
Para nuestro ejemplo de los 30 datos bidimensionales, esta función se ejecuta en el *Command Window* de $MATLAB$ de la siguiente manera:

```
>> [bics,bestmodel,allmodels,Z,clabs]=mbclust(Xa,4)
Warning: Could not find an exact (case-sensitive) match for
'mbclust'. C:\Archivos de programa\MATLAB71\toolbox\mbctool\mbclust.M
is a case-insensitive match and will be used instead. You can
improve the performance of your code by using exact name matches and
we therefore recommend that you update your usage accordingly.
Alternatively, you can disable this warning using
warning('off','MATLAB:dispatcher:InexactMatch'). Getting the
agglomerative model based clustering structure

Merging clusters ... step 1
Merging clusters ... step 2
Merging clusters ... step 3
Merging clusters ... step 4
Merging clusters ... step 5
         .
         .
         .
Merging clusters ... step 27
Merging clusters ... step 28

Getting the finite mixture estimate for model 1, 2 clusters.
Getting the finite mixture estimate for model 2, 2 clusters.
Getting the finite mixture estimate for model 3, 2 clusters.
Getting the finite mixture estimate for model 4, 2 clusters.
Getting the finite mixture estimate for model 1, 3 clusters.
Getting the finite mixture estimate for model 2, 3 clusters.
Getting the finite mixture estimate for model 3, 3 clusters.
Getting the finite mixture estimate for model 4, 3 clusters.
```

```
Getting the finite mixture estimate for model 1, 4 clusters.
Getting the finite mixture estimate for model 2, 4 clusters.
Getting the finite mixture estimate for model 3, 4 clusters.
Getting the finite mixture estimate for model 4, 4 clusters.

Maximum BIC is -239.3577. Model number 3. Number of clusters is 3

bics =
 -289.4720 -276.7183 -240.9345 -248.7121
 -289.4720 -275.5715 -245.5116 -255.6220
 -293.9860 -247.9193 -239.3577 -247.1562
 -293.9860 -252.9455 -256.4803       NaN
bestmodel =
    pies: [0.2561 0.5439 0.2000]
     mus: [2x3 double]
    vars: [2x2x3 double]
allmodels = 1x4 struct array with fields:
    clus
Z =
   19.0000   22.0000   54.4988
   15.0000   20.0000   54.5028
   16.0000   29.0000   54.5075
   18.0000   30.0000   54.5129
    3.0000    6.0000   54.5240
              .
              .
              .
   51.0000   54.0000   59.2000
   50.0000   52.0000   60.5356
   49.0000   55.0000   62.4980
   53.0000   57.0000   76.5650
   56.0000   58.0000  119.3946
clabs =

  Columns 1 through 12
     3     3     3     3     3     3     1     1     1     1     1     1
```

```
  Columns 13 through 24
     1     1     2     2     2     2     2     2     2     2     2     2
  Columns 25 through 30
     2     2     2     2     2     2
```

Luego la función "mbclust" sugiere, según el máximo valor del BIC, que los datos se clasifiquen en tres grupos de distribuciones elipsoidales diferentes (modelo 3). Los parámetros para el mejor modelo se obtienen así:

```
>> % Los mejores pesos de mezcla.
>> bestmodel.pies
ans =
    0.2561    0.5439    0.2000
>> % Los mejores vectores de medias para cada componente.
>> bestmodel.mus
ans =
   -0.3467    3.4250   -2.5297
    3.0657    0.5297   -3.4054
>> % Las mejores matrices de covarianza para cada componente.
>> bestmodel.vars
ans(:,:,1) =
    0.6459   -0.3767
   -0.3767    0.8480
ans(:,:,2) =
    0.6459   -0.3767
   -0.3767    0.8480
ans(:,:,3) =
    0.6459   -0.3767
   -0.3767    0.8480
```

Podemos también visualizar la conclusión del modelo escogido (El mejor en cuanto a cluster y cantidad de componentes) con la función "plotbic" que gráfica todos los valores **BIC** para todos los modelos bajo las consideraciones de las matrices de covarianzas. La función "plotbic" usa la variable de rendimiento bics de mbclust, y la sintaxis básica es

```
>> plotbic(bics)
```

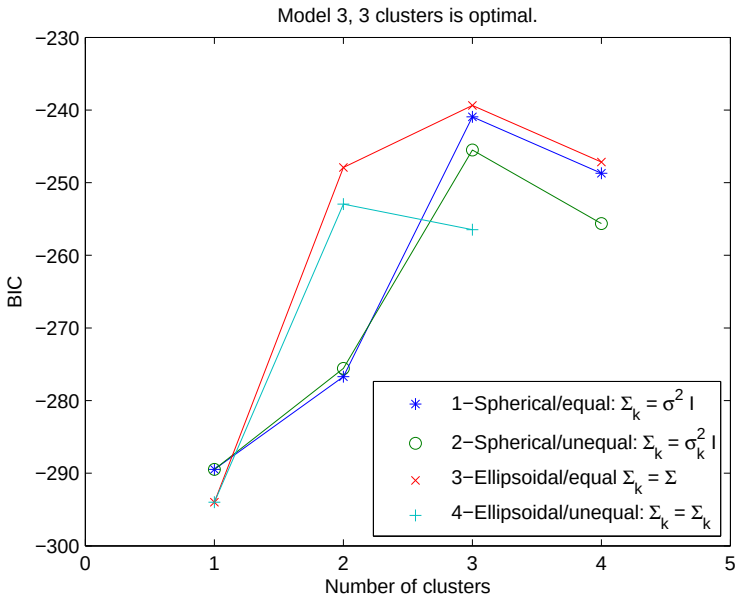


Figura 4.9: Gráfico de los valores **BIC** para cada modelo, según la cantidad de clusters escogidos. El máximo se da para 3 clusters con el modelo 3.

Usando los resultados de la función mbclust aplicados al conjunto de datos ejemplo1_mbc.mat, obtenemos el gráfico de los valores **BIC** en la Figura 4.9.
En la figura 4.10 podemos observar la clasificación dada por mbclust en el plano $\mathbb{R}^2$. Nótese que el dato 8 por métodos aglomerativos jerárquicos pertenecía inicialmente al grupo 2. Con el mbclust encontramos que este provenía del grupo 1.
El **MBC** ofrece la función reclus como una manera de extender las ideas del método del rectángulo para desplegar las configuraciones de otros métodos de clasificación (no solo los aglomerativos). Como en el gráfico de rectángulo, ReClus usa el área entera del despliegue como el rectángulo padre, luego este se divide en rectángulos mas pequeños dónde el área es proporcional al número de observaciones que pertenecen a ese cluster. Para poder usar esta herramienta se hace necesario encontrar primero las etiquetas del cluster para cada dato, como las dadas por la función cluster en el agrupamiento aglomerativo y guardadas en el vector T o como las que nos proporciona la salida clabs de la función mbclust. Estas etiquetas se obtienen para cualquier método con las siguientes instrucciones en *MATLAB*:

```
>>[clabs,errdata] = ...
mixclass(data,bestmodel.pies,bestmodel.mus,bestmodel.vars);
```

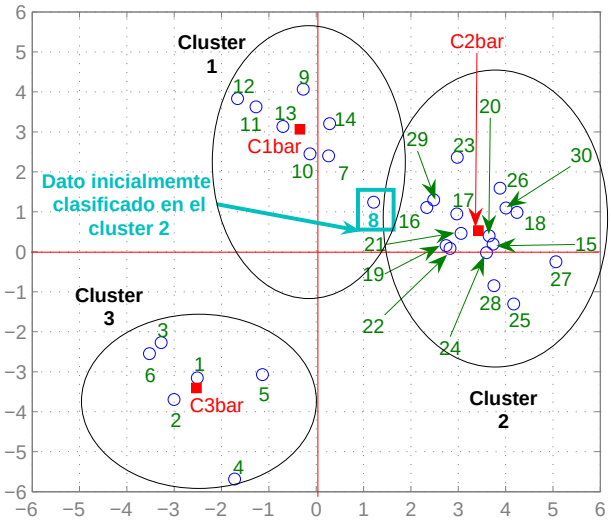


Figura 4.10: Clusters formados en el agrupamiento por el mbclust con sus respectivos vectores de medias muestrales. Nótese que el dato 8 se clasificó inicialmente por métodos jerárquicos en el cluster 2 y ahora se encuentra en el cluster 1.

La función mixclass produce la salida clabs que son las etiquetas de las cuales se ha venido hablando y errdata que nos proporciona el error de clasificación cometido por el método. Obsérvese que las entradas son los datos (data) y los parámetros obtenidos al final del proceso, que para el caso de la clasificación obtenida con mbclust son: bestmodel.pies,bestmodel.mus,bestmodel.vars. Luego, para los resultados obtenidos con mbclust, la función mixclass arroja como salida el vector clabs, que es el mismo obtenido con mbclust y el vector errdata con los errores en la clasificación del mbclust. En nuestro ejemplo con los 30 datos bivariantes, la función mixclass produce:

```
>> [clabs,errXa]=mixclass(Xa,bestmodel.pies,bestmodel.mus,bestmodel.vars)
clabs =
  Columns 1 through 12
     3     3     3     3     3     3     1     1     1     1     1     1
  Columns 13 through 24
     1     1     2     2     2     2     2     2     2     2     2     2
  Columns 25 through 30
     2     2     2     2     2     2
errXa =
  Columns 1 through 7
         0         0         0         0    0.0000         0    0.0013
```

```
Columns 8 through 14
  0.3213    0.0000    0.0001    0.0000    0.0000    0.0000    0.0010
Columns 15 through 21
  0.0000    0.0040    0.0001    0.0000    0.0003    0.0000    0.0001
Columns 22 through 28
  0.0002    0.0002    0.0000    0.0000    0.0000    0.0000    0.0000
Columns 29 through 30
  0.0020    0.0000
```

ReClus(reclus) tiene varias opciones para presentar sus graficos. La mas sencilla muestra los datos según la clasificación hecha por mbclust. La sintaxis para esto es: reclus(clabs). Otra alternativa es, conocida las verdaderas etiquetas de clase o el origen de los datos, mostrar un gráfico que compare esta clasificación con la dada en el vector clabs. Luego así tendremos un cuadro visual de que tan precisa esta la clasificación según la verdadera información. Si no tenemos las verdaderas etiquetas de la clase para los datos, podemos comparar las etiquetas asignadas en clasificación obtenida por los métodos aglomerativos y archivadas anteriormente en el vector T con las etiquetas clabs halladas con mbclust. Esto se hace con la instrucción:

```
>> reclus(clabs,T);
```

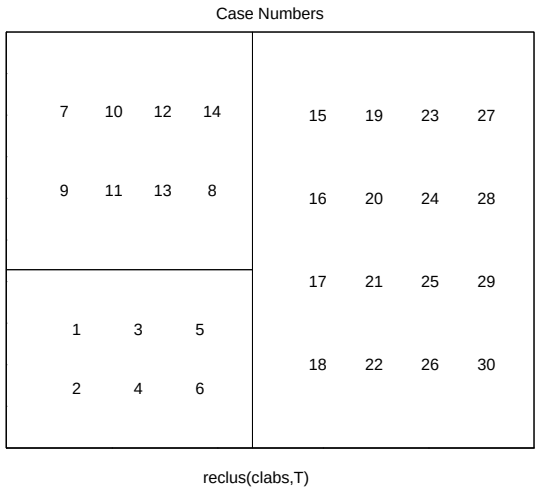


Figura 4.11: Este gráfico muestra los elementos muestrales según la clasificación hecha por mbclust comparados respecto a la clasificación dada por la función cluster.

Para nuestro ejemplo con los 30 datos bivariantes, el gráfico de rectángulo generado por reclus(clabs,T) se observa en la figura 4.11. Además podemos observar la proximidad de las dos clasificaciones con el comando:

```
>> reclus(clabs,T,errdata);
```

Este diseño se muestra en la figura 4.12. Alternativamente podríamos estar interesados en ver qué observaciones tienen una probabilidad alta de pertenecer al cluster. Podemos llamar a reclus con un umbral como sigue:

```
>> reclus(clabs,T,errdata,.999);
```

Observaciones que tienen una probabilidad posterior o superior que el umbral se muestra en negro y escritas en negrita. Observemos este gráfico en la figura 4.13. Para finalizar este ejemplo veamos en la figura 4.14 la distribución obtenida como mezcla de las tres distribuciones bivariantes con parámetros:

$$\mu_1 = \begin{pmatrix} -3 \\ -3 \end{pmatrix}, \quad \mu_2 = \begin{pmatrix} 0 \\ 3 \end{pmatrix}, \quad \mu_3 = \begin{pmatrix} 4 \\ 0 \end{pmatrix}; \quad \mathbf{V} = \mathbf{V}_1 = \mathbf{V}_2 = \mathbf{V}_3 = \begin{pmatrix} 1 & -0.5 \\ -0.5 & 1 \end{pmatrix},$$

y pesos de mezcla $\pi_1 = 0.3, \pi_2 = 0.2$ y $\pi_3 = 0.5$.
Además en la figura 4.15 se aprecia las curvas de nivel para esta distribución.

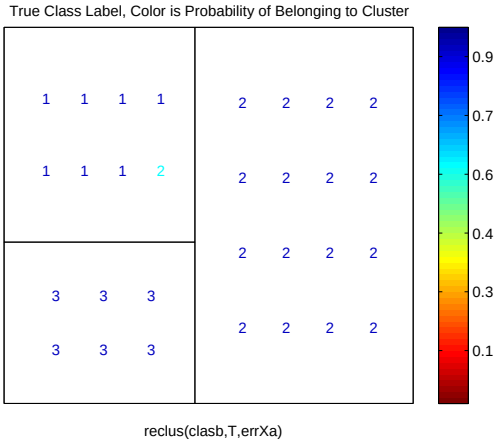


Figura 4.12: En este gráfico se compara la asignación hecha por la calificación aglomerativa con la hallada en el mbclust. La barra de color indica la probabilidad de que la observación pertenezca al cluster.

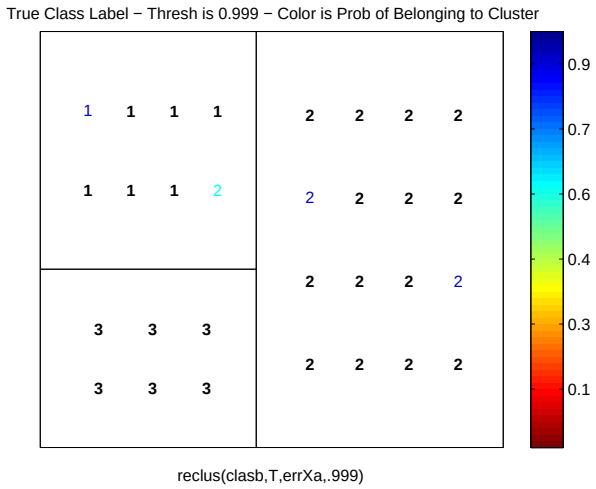


Figura 4.13: Este gráfico del reclus muestra las observaciones con una probabilidad mayor que 0.999 en negrilla.

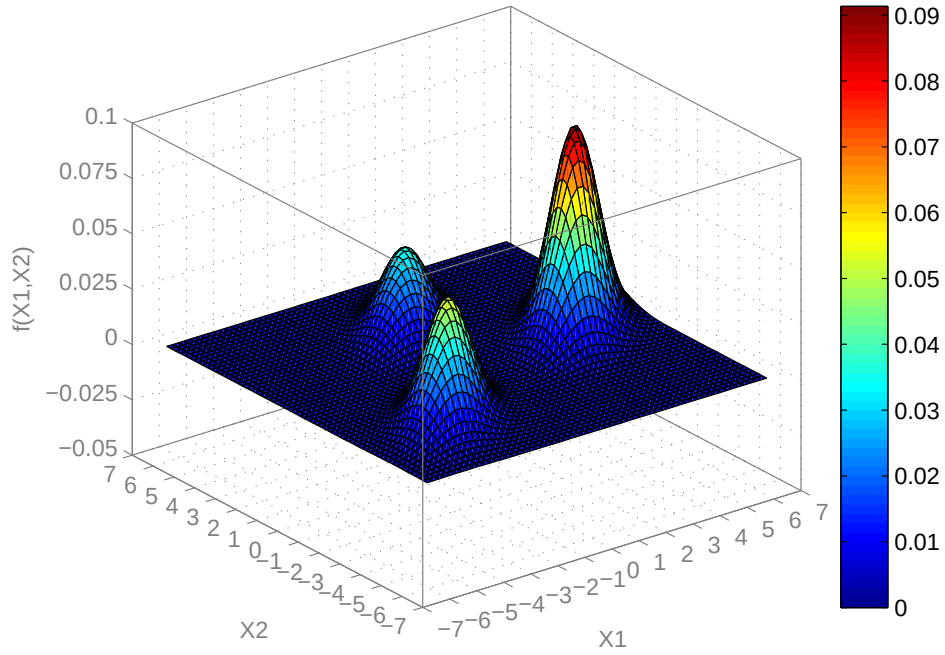


Figura 4.14: Distribución de la mezcla: $\mathbf{G(X_1, X_2)} = 0.3N_2(\mu_1, \mathbf{V}) + 0.2N_2(\mu_2, \mathbf{V}) + 0.5N_2(\mu_3, \mathbf{V})$

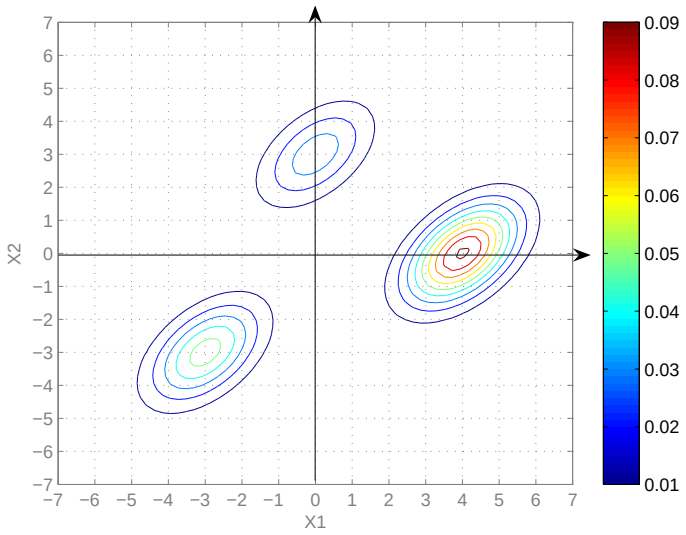


Figura 4.15: En las curvas de nivel para la distribución de la mezcla $\mathbf{G}(\mathbf{X_1}, \mathbf{X_2})$, se observan como los datos en cada componente (cluster) se distribuyen en forma elipsoidal. Además las tres componentes tienen igual orientación.

4.3.2. Mezcla de 5 distribuciones normales con dimension $p = 4$.

De forma similar al ejemplo anterior, consideramos una muestra con $n = 100$ generada por "genmix " para un modelo de mezcla normal multivariante con vectores de medias:

$$\mu_1 = \begin{pmatrix} 0 \\ 0 \\ 0 \\ 5 \end{pmatrix}, \quad \mu_2 = \begin{pmatrix} 0 \\ 0 \\ 5 \\ 0 \end{pmatrix}, \quad \mu_3 = \begin{pmatrix} 0 \\ 0 \\ 0 \\ 0 \end{pmatrix}, \quad \mu_4 = \begin{pmatrix} 0 \\ 5 \\ 0 \\ 0 \end{pmatrix}, \quad \mu_5 = \begin{pmatrix} 5 \\ 0 \\ 0 \\ 0 \end{pmatrix},$$

matrices de covarianzas:

$$\mathbf{V}_1 = 2\mathbf{I}, \quad \mathbf{V}_2 = 3\mathbf{I}, \quad \mathbf{V}_3 = 5\mathbf{I}, \quad \mathbf{V}_4 = \mathbf{I}, \quad \mathbf{V}_3 = 0.7\mathbf{I},$$

y pesos de mezcla:

$$\pi_1 = 0.1, \quad \pi_2 = 0.15, \quad \pi_3 = 0.2, \quad \pi_4 = 0.25, \quad \pi_5 = 0.3.$$

La función genmix despliega el **GUI** que ya se ha mostrado anteriormente en la figura 4.1. Siguiendo los pasos allí indicados se da entrada a los anteriores parámetros, generando así en $MATLAB$ la matriz de datos (Xa) que se guardan en el archivo mezcla_P4G5.mat.

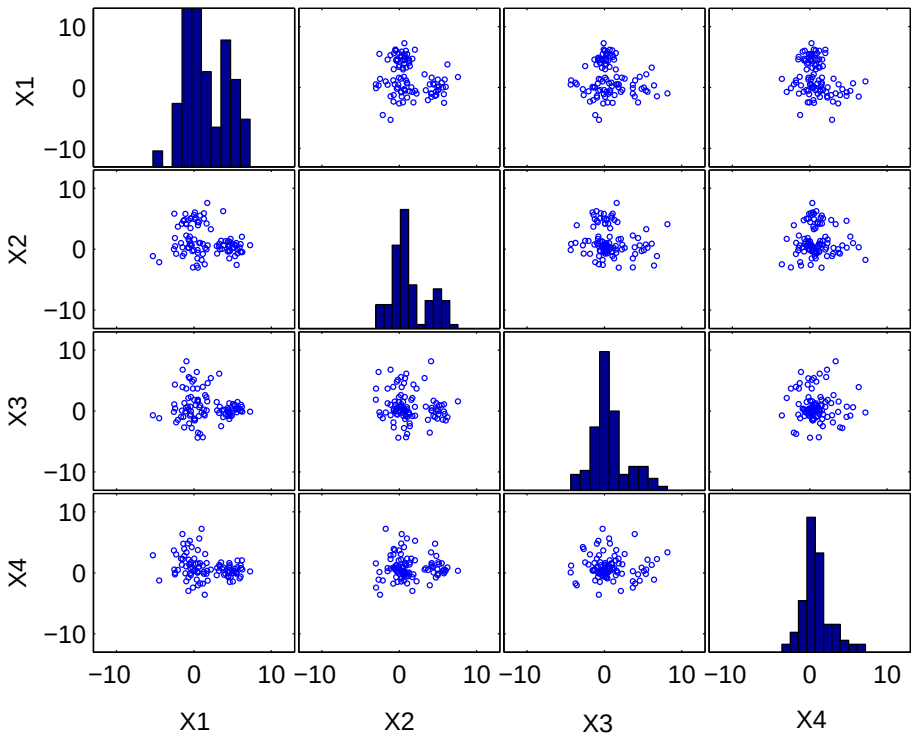


Figura 4.16: Plotmatrix generada por el botón PlotData en el GUI mostrado la figura 4.1 para 100 datos multivariantes con distribución normal.

Por su extension, estos datos se muestran en el **apéndice D**. Con el botón Plot Data obtenemos la matriz de dispersion dada en la figura 3.18. En la figura 4.16 se observan relaciones de dependencia no lineales y homocedasticas (varianza constante) entre las variables, reflejándose esto en los histogramas de la diagonal principal de la matriz, donde se observa simetría entre las barras.
Realicemos un agrupamiento aglomerativo jerárquico acorde al procedimiento utilizado por el **MBC**; esto se realiza con el fin de tener un punto de comparación para la clasificación final. Este agrupamiento se lleva acabo digitando en el *Command Window* de $MATLAB$ la función "agmbclust" , la cual produce la matriz de rendimiento Z, que contiene en cada fila los cluster que se van uniendo y a la distancia que lo hacen. La matriz Z nos permite gráficar el dendograma dado en la figura 4.17. El procedimiento realizado en $MATLAB$ se describe a continuación.

```
>> % Iniciamos importando los datos
```

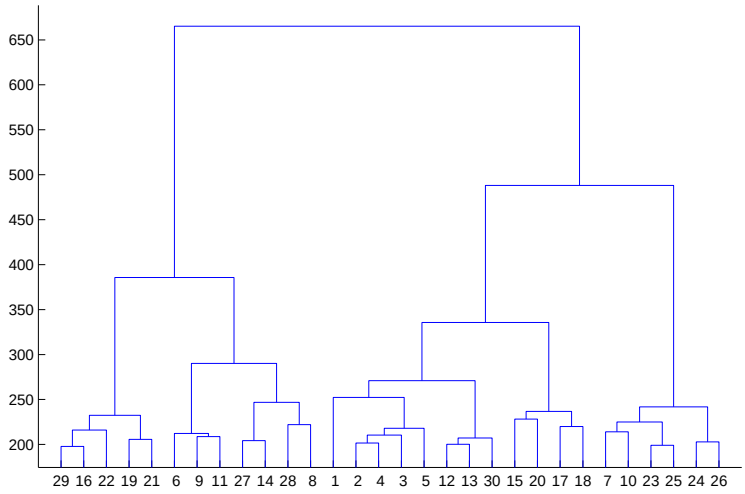


Figura 4.17: Este es el dendograma para los 100 datos del agrupamiento aglomerativo. El eje vertical representa la distancia a la cual se unen los clusters. Note que algunos de los números en las etiquetas de cada rama no corresponden al número verdadero de la observación; estas etiquetas aquí mostradas podrían incluir algunas de las observaciones reales.

```
>> data = load('mezcla_P4G5.mat');
>> % Se hace el agrupamiento aglomerativo jerárquico.
>> Z=agmbclust(Xa);
>> % Se genera el dendograma.
>> dendogram(Z);
```

Una cosa para notar con respecto a la función dendrogram es que despliega como valor predeterminado máximo 30 nodos. Así, los nodos de la rama desplegada podrían corresponder a observaciones múltiples. Por ejemplo en Figura 4.17 se muestran sólo 30 nodos, cuando se tienen en realidad 100 observaciones. Para encontrar las observaciones contenidas en cada nodo de las ramas, procedemos en el *Command Window* de *MATLAB* como sigue:

```
>> % Se corta el dendograma a una altura de 150 para formar los 30 clusters,
que serán cada uno de los nodos contenidos en este.
>> [H,E] = dendrogram(Z,'colorthreshold',150);
>> % H es el vector de rendimiento para las alturas del dendograma.
>> % E es el vector con los nombres de cada uno de los 30 clusters .
```

```
>> % Ahora con la función find encontramos los elementos de cada nodo.
Los elementos de cada nodo los guardamos en vectores Ei, donde i es
la etiqueta mostrada en el dendograma.
>> TE1=find(E==1);E1=TE1'
E1 =
     1    16    19    42
>> TE2=find(E==2);E2=TE2'
E2 =
     2     5     7
>> TE3=find(E==3);E3=TE3'
E3 =
     3
>> TE4=find(E==4);E4=TE4'
E4 =
     4
>> TE5=find(E==5);E5=TE5'
E5 =
    33
>> TE6=find(E==6);E6=TE6'
E6 =
     6     8    10
>> TE7=find(E==7);E7=TE7'
E7 =
    34    46    75    80    93    96    98
>> TE8=find(E==8);E8=TE8'
E8 =
    39
>> TE9=find(E==9);E9=TE9'
E9 =
     9    36
>> TE10=find(E==10);E10=TE10'
E10 =
    40    71    74    81
>> TE11=find(E==11);E11=TE11'
E11 =
    11    26    37
```

```
>> TE12=find(E==12);E12=TE12'
E12 =
    12    14
>> TE13=find(E==13);E13=TE13'
E13 = =
    13    47
>> TE14=find(E==14);E14=TE14'
E14 =
    41
>> TE15=find(E==15);E15=TE15'
E15 =
    15    21    23    24    25
>> TE16=find(E==16);E16=TE16'
E16 =
    43    51    60    64
>> TE17=find(E==17);E17=TE17'
E17 =
    17
>> TE18=find(E==18);E18=TE18'
E18 =
    18    35
>> TE19=find(E==19);E19=TE19'
E19 =
    45    57
>> TE20=find(E==20);E20=TE20'
E20 =
    20    22    38
>> TE21=find(E==21);E21=TE21'
E21 =
    48    49    50    52    54    56
>> TE22=find(E==22);E22=TE22'
E22 =
    55    59    62    63    65
>> TE23=find(E==23);E23=TE23'
E23 =
    67    76    82    84    86    87    90    91    95    97
```

```
>> TE24=find(E==24);E24=TE24'
E24 =
    68    69    73    83    88    89    92    94
>> TE25=find(E==25);E25=TE25'
E25 =
    70    72    78    85    99   100
>> TE26=find(E==26);E26=TE26'
E26 =
    77    79
>> TE27=find(E==27);E27=TE27'
E27 =
    27    31    32
>> TE28=find(E==28);E28=TE28'
E28 =
    28
>> TE29=find(E==29);E29=TE29'
E29 =
    29    53    58    61    66
>> TE30=find(E==30);E30=TE30'
E30 =
    30    44
```

Luego esto quiere decir que, por ejemplo, la etiqueta marcada en el dendograma con el número 21, contiene los datos mostrados en el **apéndice D** con etiquetas 48, 49, 50, 52, 54, 56.
Ahora, si determinamos que los clusters son $G = 5$, el dendograma nos sugiere la clasificación mostrada en la figura 4.18. Para determinar los datos reales que pertenecen a cada grupo procedemos, como en el ejemplo anterior, a utilizar la función cluster. Esto se visualiza digitando los comandos en $MATLAB$ que se describen a continuación.

```
>> % Se agrupan los datos en 5 clusters.
>> T = cluster(Z,'maxclust',5);
>> % Ahora con la función find encontramos los elementos de cada cluster.
Los elementos del cluster se guardan en vectores Ti, donde i es la
etiqueta mostrada en el dendograma.
>> TC1=find(T==1); C1=TC1'
```

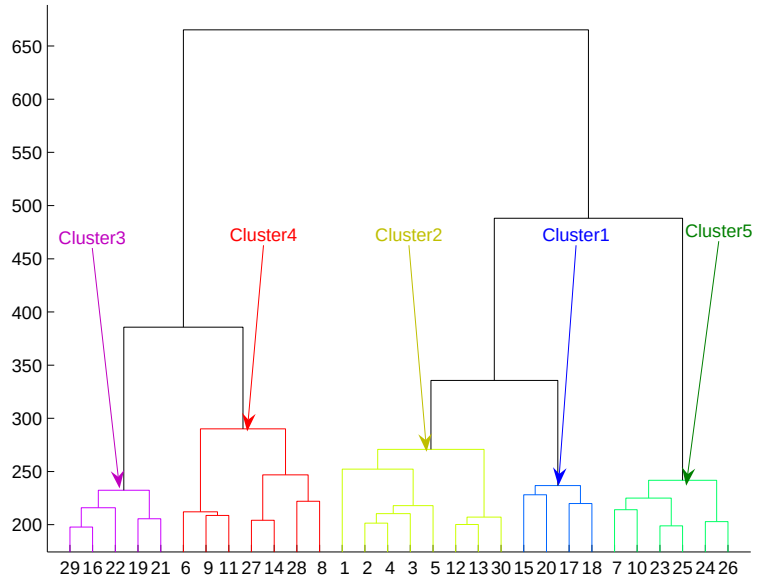


Figura 4.18: Clasificación aglomerativa jerárquica para los 100 datos con $G = 5$.

```
C1 =
    15    17    18    20    21    22    23    24    25    35    38
>> TC2=find(T==2); C2=TC2'
C2 =
  Columns 1 through 12
     1     2     3     4     5     7    12    13    14    16    19    30
  Columns 13 through 16
    33    42    44    47
>> TC3=find(T==3); C3=TC3'
C3 =
  Columns 1 through 12
    29    43    45    48    49    50    51    52    53    54    55    56
  Columns 13 through 22
    57    58    59    60    61    62    63    64    65    66
>> TC4=find(T==4); C4=TC4'
C4 =
  Columns 1 through 12
```

```
     6     8     9    10    11    26    27    28    31    32    36    37
  Columns 13 through 14
    39    41
>> TC5=find(T==5); C5=TC5'
C5 =
  Columns 1 through 12
    34    40    46    67    68    69    70    71    72    73    74    75
  Columns 13 through 24
    76    77    78    79    80    81    82    83    84    85    86    87
  Columns 25 through 36
    88    89    90    91    92    93    94    95    96    97    98    99
  Column 37
   100
```

Debido a lo tedioso que se convierte el encontrar los elementos que pertenecen a cada uno de los 30 nodos de las ramas del dendograma, cuando la muestra es muy grande, se generan los gráficos de rectángulo por medio de la función rectplot. Para este ejemplo obtenemos los gráficos de rectángulo dados en las figuras 4.19 y 4.20.

```
>> rectplot(Z,5)
```

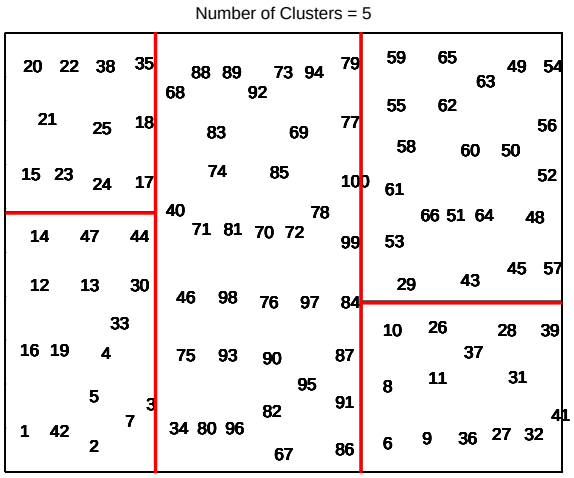


Figura 4.19: Gráfico de rectángulo para $G = 5$

```
>> rectplot(Z,100)
```

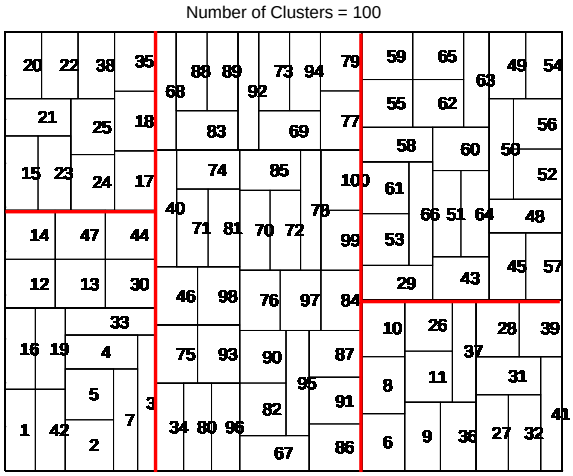


Figura 4.20: Se genera el gráfico de rectángulo para $G = 100$, donde el area de cada rectángulo es proporcional a la distancia de unión con el cluster mas cercano.

Con la función mbclust que lleva a cabo el procedimiento completo para la clasificación mediante mezclas finitas, incluyendo el agrupamiento aglomerativo jerárquico hecho anteriormente, **el algoritmo EM**, y la selección del mejor modelo. Para nuestro ejemplo con 100 datos multivariantes, esta función se ejecuta en el *Command Window* de *MATLAB* de la siguiente manera:

```
>> [bics,bestmodel,allmodels,Z,clabs]=mbclust(Xa,8)
Warning: Could not find an exact (case-sensitive) match for
'mbclust'. C:\Archivos de programa\MATLAB71\toolbox\mbctool\mbclust.M
is a case-insensitive match and will be used instead. You can
improve the performance of your code by using exact name matches and
we therefore recommend that you update your usage accordingly.
Alternatively, you can disable this warning using
warning('off','MATLAB:dispatcher:InexactMatch').
Getting the agglomerative model based clustering structure
Merging clusters ... step 1
Merging clusters ... step 2
Merging clusters ... step 3
```

```
Merging clusters ... step 4
Merging clusters ... step 5
Merging clusters ... step 6
              .
              .
              .
Merging clusters ... step 93
Merging clusters ... step 94
Merging clusters ... step 95
Merging clusters ... step 96
Merging clusters ... step 97
Merging clusters ... step 98
Getting the finite mixture estimate for model 1, 2 clusters.
Getting the finite mixture estimate for model 2, 2 clusters.
Getting the finite mixture estimate for model 3, 2 clusters.
Getting the finite mixture estimate for model 4, 2 clusters.
Getting the finite mixture estimate for model 1, 3 clusters.
Getting the finite mixture estimate for model 2, 3 clusters.
Getting the finite mixture estimate for model 3, 3 clusters.
Getting the finite mixture estimate for model 4, 3 clusters.
                              .
                              .
                              .
Getting the finite mixture estimate for model 1, 8 clusters.
Getting the finite mixture estimate for model 2, 8 clusters.
Getting the finite mixture estimate for model 3, 8 clusters.
Getting the finite mixture estimate for model 4, 8 clusters.

Maximum BIC is -1688.8346. Model number 2. Number of clusters is 4

bics =
  1.0e+003 *
  Columns 1 through 7
   -1.8532   -1.8416   -1.8066   -1.7582   -1.7525   -1.7532   -1.7586
   -1.8532   -1.7227   -1.6889   -1.6888   -1.7044   -1.7228   -1.7462
   -1.8611   -1.8350   -1.8057   -1.7720   -1.7766   -1.7825   -1.7911
```

```
   -1.8611   -1.8365   -1.7818   -1.8024   -1.8352   -1.8757   -1.9031
  Column 8
   -1.7739
   -1.7675
       NaN
       NaN
bestmodel =
    pies: [0.1987 0.2756 0.3456 0.1802]
     mus: [4x4 double]
    vars: [4x4x4 double]
allmodels = 1x4 struct array with fields:
    clus

Z =
   82.0000   90.0000  173.7790
   55.0000   59.0000  173.7958
   95.0000  101.0000  173.8168
   99.0000  100.0000  173.8456
   70.0000   72.0000  173.8783
               .
               .
               .
  181.0000  192.0000  290.1217
  190.0000  194.0000  335.6823
  189.0000  195.0000  385.7570
  191.0000  196.0000  488.0677
  197.0000  198.0000  665.1850

clabs =
  Columns 1 through 12
     2     2     2     2     2     2     2     2     2     2     2     4
  Columns 13 through 24
     4     4     4     2     4     4     4     4     4     4     4     4
  Columns 25 through 36
     4     2     2     2     1     2     2     2     2     3     4     2
  Columns 37 through 48
```

```
   2     4     2     2     2     2     1     4     2     4     4     1
Columns 49 through 60
   1     1     1     1     1     1     1     1     1     1     1     1
Columns 61 through 72
   1     1     1     1     1     1     3     3     3     3     3     3
Columns 73 through 84
   3     3     3     3     3     3     3     3     3     3     3     3
Columns 85 through 96
   3     3     3     3     3     3     3     3     3     3     3     3
Columns 97 through 100
   3     3     3     3
```

Podemos visualizar la conclusión del modelo escogido (El mejor en cuanto a cluster y cantidad de componentes) con la función "plotbic" que gráfica los valores **BIC** para todos los modelos bajo las consideraciones de las matrices de covarianzas.

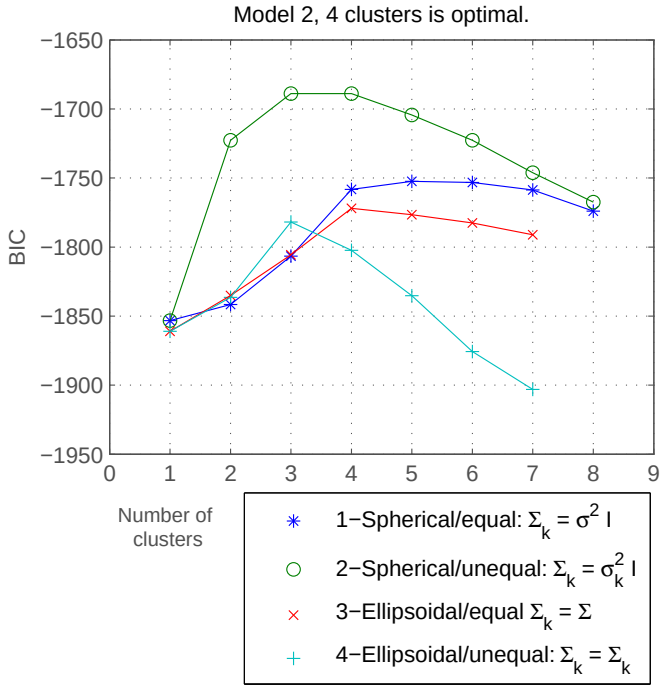


Figura 4.21: Gráfico de los valores **BIC** para cada modelo, según la cantidad de clusters escogidos. El máximo se da para 4 cluster con el modelo 2.

Luego la figura 4.21 sugiere, según el máximo valor del BIC, que los datos se clasifiquen

en cuatro grupos de distribuciones esféricas diferentes (modelo 2). Los parámetros para el mejor modelo se obtienen así:

```
>> % Los mejores pesos de mezcla.
>> bestmodel.pies
ans =
    0.1987    0.2756    0.3456    0.1802
>> % Los mejores vectores de medias para cada componente.
>> bestmodel.mus
ans =
   -0.0557   -0.3884    4.9045    0.1877
    5.0658    0.7416    0.2010    0.2872
    0.0932   -0.6852    0.1548    4.6008
    0.6744    2.1160    0.2710    0.2643
>> % Las mejores matrices de covarianza para cada componente.
>> bestmodel.vars
ans(:,:,1) =
    0.9336         0         0         0
         0    0.9336         0         0
         0         0    0.9336         0
         0         0         0    0.9336
ans(:,:,2) =
    5.1284         0         0         0
         0    5.1284         0         0
         0         0    5.1284         0
         0         0         0    5.1284
ans(:,:,3) =
    0.7328         0         0         0
         0    0.7328         0         0
         0         0    0.7328         0
         0         0         0    0.7328
ans(:,:,4) =
    2.7548         0         0         0
         0    2.7548         0         0
         0         0    2.7548         0
         0         0         0    2.7548
```

Se esperaba que la función mbclust clasificara al conjunto de datos en 5 clusters esféricos no iguales (modelo2), pues los datos fueron creados artificialmente con la función genmix bajo estas condiciones. La posible **"mala clasificación "** puede ser debida a multiples razones, como lo es el mismo origen de los datos, la evidente densidad que existe entre ellos (reflejada en la figura 4.16), o la imprecisión de calculo originada por el algoritmo de la función mbclust (esto se advierte en un mensaje tan pronto la función esta en ejecución). Luego se hace necesario realizar previamente otros análisis de correspondencia de datos para observar relaciones entre datos o variables. De igual forma el investigador es el que conoce sus datos y sabe como actuar ante estos resultados.
Para finalizar observemos el agrupamiento realizado por mbclust en un gráfico de rectángulo generado por la función reclus. Para esto encontremos primero las etiquetas del cluster para cada dato y los errores de clasificación; ello se obtiene con la función mixclass como sigue:

```
>> [clabs,errdata] = mixclass(Xa,bestmodel.pies,bestmodel.mus,bestmodel.vars)
clabs =
  Columns 1 through 12
     2     2     2     2     2     2     2     2     2     2     2     4
  Columns 13 through 24
     4     4     4     2     4     4     4     4     4     4     4     4
  Columns 25 through 36
     4     2     2     2     1     2     2     2     2     3     4     2
  Columns 37 through 48
     2     4     2     2     2     2     1     4     2     4     4     1
  Columns 49 through 60
     1     1     1     1     1     1     1     1     1     1     1     1
  Columns 61 through 72
     1     1     1     1     1     1     3     3     3     3     3     3
  Columns 73 through 84
     3     3     3     3     3     3     3     3     3     3     3     3
  Columns 85 through 96
     3     3     3     3     3     3     3     3     3     3     3     3
  Columns 97 through 100
     3     3     3     3

errdata =
  Columns 1 through 7
```

```
  0.0973    0.0075    0.0738    0.0001    0.0027    0.0000    0.0002
Columns 8 through 14
  0.0000    0.0024    0.0002    0.0024    0.0682    0.0251    0.0045
Columns 15 through 21
  0.0118    0.2791    0.0566    0.0131    0.3819    0.0683    0.0097
Columns 22 through 28
  0.0741    0.0205    0.0233    0.0137    0.0016    0.0000    0.0154
Columns 29 through 35
  0.2241    0.3257    0.0000    0.0000    0.0050    0.0822    0.0300
Columns 36 through 42
  0.0002    0.0663    0.3366    0.0008    0.0788    0.0049    0.4777
Columns 43 through 49
  0.0782    0.0563    0.1893    0.2342    0.0148    0.0356    0.0094
Columns 50 through 56
  0.0902    0.0161    0.0119    0.0313    0.0131    0.0127    0.0104
Columns 57 through 63
  0.0725    0.4880    0.0145    0.0071    0.0687    0.0261    0.0174
Columns 64 through 70
  0.0083    0.0165    0.0835    0.0007    0.0011    0.0327    0.0083
Columns 71 through 77
  0.0038    0.0034    0.0023    0.0258    0.0262    0.0019    0.0507
Columns 78 through 84
  0.0091    0.0020    0.0089    0.0022    0.0005    0.0008    0.0052
Columns 85 through 91
  0.0026    0.0008    0.0031    0.0010    0.0013    0.0006    0.0007
Columns 92 through 98
  0.0121    0.0043    0.0126    0.0007    0.0319    0.0017    0.1770
Columns 99 through 100
  0.0026    0.0017
```

En la gráfica 4.22 se visualizan los datos clasificados por mbclust en sus 4 clusters. Esta asignación puede ser comparada con la obtenida con métodos aglomerativos jerárquicos para así detectar los datos que cambiaron de grupo; esto se aprecia en la figura 4.23. Puesto que esta escala de colores sobre los datos en ocasiones es difícil de distinguir, generamos los gráficos de las figuras 4.24 y 4.25 donde se destacan en negrilla los datos con una probabilidad alta de pertenecía al cluster (0.999 y 0.99 respectivamente).

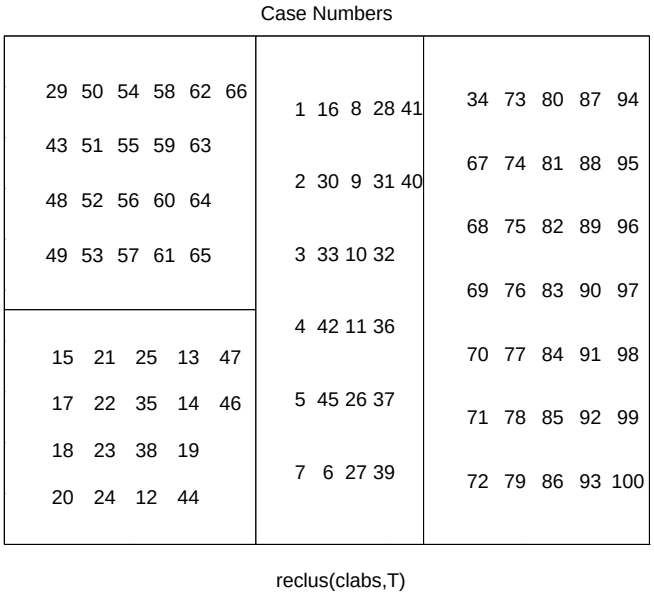


Figura 4.22: Este gráfico muestra la clasificación dada por mbclust usando los resultados del mejor modelo.

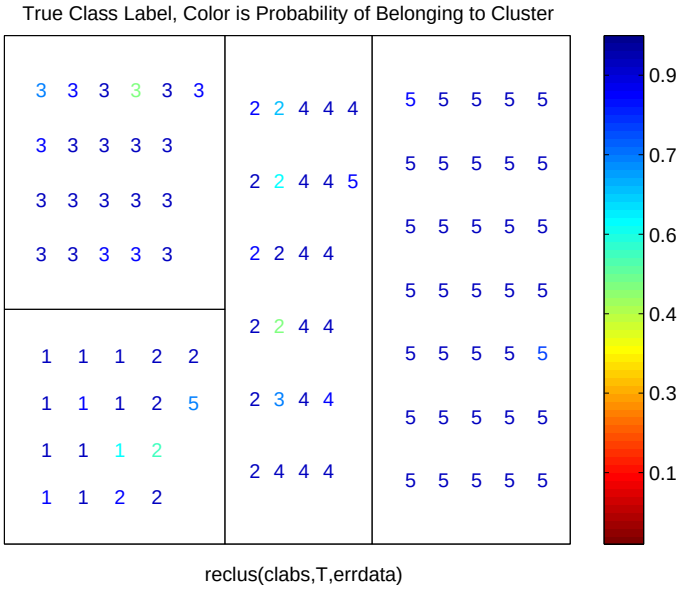


Figura 4.23: En este gráfico se compara la asignación hecha por la calificación aglomerativa con la hallada en el mbclust. La barra de color indica la probabilidad de que la observación pertenezca al cluster.

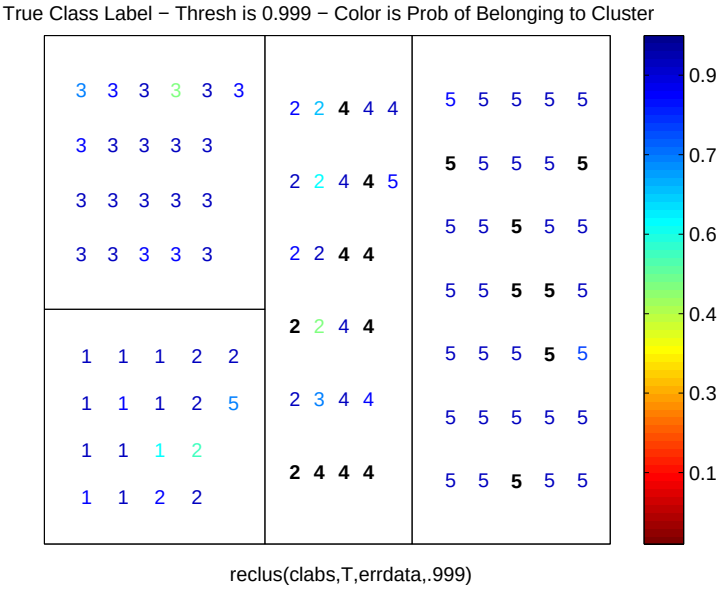


Figura 4.24: Este gráfico del reclus muestra las observaciones con una probabilidad de pertenecía mayor que 0.999 en negrilla.

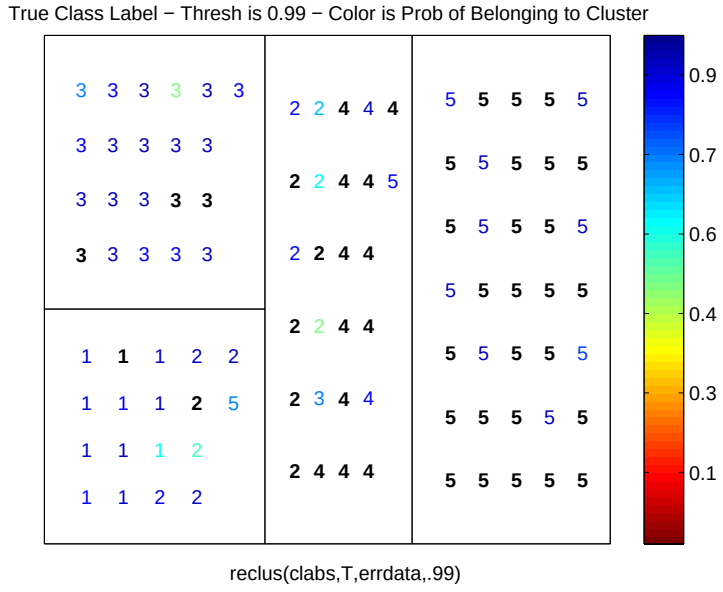


Figura 4.25: Este gráfico del reclus muestra las observaciones con una probabilidad de pertenecía mayor que 0.99 en negrilla.

Apéndice A

Distribuciones multivariantes especiales

Veamos dos distribuciones multivariantes con sus respectivas propiedades asociadas a la distribución normal multivariante; estas son la *distribución de Wishart* y *la distribución T^2 de Hotelling* ó *la distribución de Hotelling*. La primera está asociada con la distribución muestral de matrices de covarianzas para variables normales multivariantes. La otra es ampliamente usada en la inferencia multivariante cuando desconocemos la variabilidad de la muestra. Estas son de gran apoyo en la conformación de grupos en el análisis que se realiza a las mezclas de las distribuciones normales.

A.1. La distribución Wishart

La distribución de Wishart[1] es la que sigue una matriz aleatoria simétrica definida positiva y la cual generaliza la distribución χ^2 de Pearson. Esta distribución juega un papel importante en la inferencia multivariante; se utiliza para representar la distribución muestral de las matrices de covarianza en muestras de variables aleatorias multivariantes, calculada a partir de la matriz de datos donde las filas son observaciones normales multivariantes.
Si $\mathbf{x}_1, \ldots, \mathbf{x}_m$ son observaciones aleatorias independientes de tamaño $(p \times 1)$, con $m > p$, normalmente distribuidos con media cero; es decir si $\mathbf{x}_i \sim N_p(0, \mathbf{V})$, $i\epsilon\{1, \ldots, m\}$ y

[1]John Wishart (1898-1956): Estadístico Británico que en 1928 que encontró la distribución que lleva su nombre. Fue profesor de la universidad de Cambridge en Estadística y Agricultura.

$\mathrm{X} = (\mathbf{x}_1, \ldots, \mathbf{x}_m)$, entonces

$$\mathbf{W} = \mathrm{XX}' = \sum_{i=1}^{m} \mathbf{x}_i \mathbf{x}_i', \tag{A.1}$$

es una matriz de tamaño $(p \times p)$, simétrica y definida positiva, y la llamamos *matriz de Wishart*, la cual sigue *la distribución de Wishart* con m grados de libertad y matriz de covarianzas $\mathbf{V}$ y se denotará por $\mathbf{W} \sim \mathcal{W}_p(m, \mathbf{V})$. Cuando $\mathbf{V} = \mathbf{I}_p$ la distribución esta en su forma estándar.
La función de distribución de Wishart esta dada por:

$$f(\mathbf{W}) = c|\mathbf{V}|^{-m/2}|\mathbf{W}|^{(m-p-1)/2} \exp\left\{-\frac{1}{2} tra[\mathbf{V}^{-1}\mathbf{W}]\right\}, \tag{A.2}$$

donde c es una constante para que la función integre a uno.

A.1.1. Propiedades de la distribución Wishart

La distribución de Wishart tiene las siguientes propiedades:

1. La esperanza de la distribución es:

   $$\mathrm{E}[\mathbf{W}] = m\mathbf{V},$$

   lo que implica que $\mathbf{W}/m$ tiene esperanza $\mathbf{V}$.

2. Si $\mathbf{W_1}, \mathbf{W_2}$ son independientes Wishart $\mathcal{W}_p(m, \mathbf{V}), \mathcal{W}_p(n, \mathbf{V})$, entonces

   $$\mathbf{W_1} + \mathbf{W_2} \sim \mathcal{W}_p(m + n, \mathbf{V}).$$

3. Sea la partición de $\mathbf{V}$ y $\mathbf{W}$ en q-filas y $(p-q)$-columnas así:

   $$\mathbf{V} = \begin{pmatrix} V_{11} & V_{12} \\ V_{21} & V_{22} \end{pmatrix}, \qquad \mathbf{W} = \begin{pmatrix} W_{11} & W_{12} \\ W_{21} & W_{22} \end{pmatrix},$$

   Si $\mathbf{W}$ se distribuye como $\mathcal{W}_p(m, \mathbf{V})$, entonces W_{ii} se distribuye como $\mathcal{W}_\alpha(m, V_{ii})$ para $i = 1, 2$ y $\alpha = q, p-q$. Esta propiedad se pueda hacer extensiva para cualquier partición adecuada de las matrices $\mathbf{V}$ y $\mathbf{W}$.

4. Si $\mathbf{W}$ se distribuye como $\mathcal{W}_p(m, \mathbf{V})$ y $\mathbf{T}$ es una matriz de tamaño $(h \times p)$ de constantes, entonces $\mathbf{T'WT} \sim \mathcal{W}_p(m, \mathbf{T'WT})$. En particular si $\mathbf{t}$ es un vector, entonces

   $$\frac{\mathbf{t'Wt}}{\mathbf{t'Vt}}, \quad es \quad \chi^2_m.$$

A.2. La distribución de Hotelling

La distribución Hotelling[2] o distribución $\mathbf{T}^2$ de Hotelling es una generalización de la distribución t-Student.

Definición A.1. Si $\mathbf{x}$ es un vector aleatorio $N_p(0, \mathbf{I})$, $\mathbf{W}$ es Wishart $\mathcal{W}_p(m, \mathbf{I})$ y además $\mathbf{x}, \mathbf{W}$ son independientes, entonces

$$\mathbf{T}^2 = m\mathbf{x}'\mathbf{W}^{-1}\mathbf{x}$$

sigue la distribución $\mathbf{T}^2$ de Hotelling que se indica por $\mathbf{T}^2_{(p,m)}$.

A.2.1. Propiedades de la distribución Hotelling

La distribución Hotelling tiene las siguientes propiedades:

1. Si $\mathbf{x}$ es $N_p(\mu, \mathbf{V})$ independiente de $\mathbf{M}$ que es $\mathcal{W}_p(m, \mathbf{V})$, entonces

$$\mathbf{T}^2_0 = m(\mathbf{x} - \mu)'\mathbf{M}^{-1}(\mathbf{x} - \mu) \sim \mathbf{T}^2_{(p,m)}.$$

2. $\mathbf{T}^2$ esta directamente relacionada con la distribución F de Fisher-Snedecor

$$\mathbf{T}^2_{(p,m)} \equiv \frac{mp}{m - p + 1} F_{(p,m-p+1)}.$$

3. Si $\overline{\mathbf{x}}$ y $\mathbf{S}$ son el vector de medias y la matriz de covarianzas de la matriz $\mathbb{X}_{n\times p}$ con filas independientes $N_p(\mu, \mathbf{V})$, entonces

$$(n-1)(\overline{\mathbf{x}} - \mu)'\mathbf{S}^{-1}(\overline{\mathbf{x}} - \mu) \sim \mathbf{T}^2_{(p,n-1)}$$

y por lo tanto

$$\frac{n-p}{p}(\overline{\mathbf{x}} - \mu)'\mathbf{S}^{-1}(\overline{\mathbf{x}} - \mu) \sim F_{(p,n-p)}$$

[2]Harold Hotelling (1895-1973): Estadístico americano fundador del Statistical Research Group en la Universidad de columbia en Nueva York y del Departamento de Estadística en Chapel Hill, uno de los centros lideres de investigación estadística moderna. Es el creador de los componentes principales, el análisis de correlaciones canónicas y los nuevos métodos de inferencia multivariante.

4. Si $\overline{\mathbf{x}}, \mathbf{S}_1, \overline{\mathbf{y}}, \mathbf{S}_2$ son el vector de medias y la matriz de covarianzas de las matrices $\mathbb{X}_{n_1 \times p}, \mathbb{Y}_{n_2 \times p}$, respectivamente con filas independientes $N_p(\mu, \mathbf{V})$ y consideramos la estimación conjunta centrada de $\mathbf{V}$

$$\widetilde{\mathbf{S}} = (n_1\mathbf{S}_1 + n_2\mathbf{S}_2)/(n_1 + n_2 - 2)$$

entonces

$$\mathbf{T}_0^2 = \frac{n_1 n_2}{n_1 + n_2}(\overline{\mathbf{x}} - \overline{\mathbf{y}})'\mathbf{S}^{-1}(\overline{\mathbf{x}} - \overline{\mathbf{y}}) \sim \mathbf{T}^2_{(p,n_1+n_2-2)}$$

y por lo tanto

$$\frac{(n_1 + n_2 - 1 - p)}{(n_1 + n_2 - 2)p}\mathbf{T}^2 \sim F_{(p,n_1+n_2-1-p)}$$

Apéndice B

Contrastes en la distribución normal multivariante

B.1. Contrastes sobre la media

En este apartado se desarrollan procedimientos de inferencia estadística sobre el vector de medias para poblaciones con distribución normal multivariante, cuando se conoce y cuando no se conoce la matriz de covarianzas; en cada uno de estos analizamos los casos para una o mas poblaciones y se dan algunos ejemplos ilustrativos acerca de estos contrastes.

B.1.1. Matriz de covarianzas conocida

- **Una población**
  Sea $\mathbf{x}_1, \ldots, \mathbf{x}_n$, con $n > p$, una muestra aleatoria de una población $N_p(\mu, \mathbf{V})$ donde $\mathbf{V}$ es conocida. Queremos realizar la prueba de hipótesis:

$$H_0 : \mu = \mu_0,\ \mathbf{V} = \widetilde{\mathbf{V}} \qquad \text{vs} \qquad H_1 : \mu \neq \mu_0,\ \mathbf{V} = \widetilde{\mathbf{V}}. \tag{B.1}$$

  Aquí μ_0 es un vector especifico y $\widetilde{\mathbf{V}}$ es la matriz de covarianzas conocida. Construimos el contraste de razón de verosimilitudes calculando el máximo de la función de verosimilitud bajo H_0 y H_1, donde encontramos que la distribución de λ es una χ^2 con p grados de libertad. Luego el estadístico de contraste es:

$$\lambda = \underbrace{n(\overline{\mathbf{x}} - \mu_0)'\widetilde{\mathbf{V}}^{-1}(\overline{\mathbf{x}} - \mu_0)}_{\chi_0^2} \sim \chi_p^2. \tag{B.2}$$

Para verificar la hipótesis nula H_0 se usa como *region critica* el conjunto de puntos tales que:

$$\chi_0^2 \geq \chi^2_{(\alpha,p)}, \tag{B.3}$$

donde $\chi^2_{(\alpha,p)}$ es el número tal que

$$P(\chi_p^2 > \chi^2_{(\alpha,p)}) = \alpha.$$

Además para una media muestral $\overline{\mathbf{x}}$, la desigualdad

$$n(\overline{\mathbf{x}} - \mu^*)'\widetilde{\mathbf{V}}^{-1}(\overline{\mathbf{x}} - \mu^*) \leq \chi^2_{(\alpha,p)} \tag{B.4}$$

se satisface con una probabilidad $(1-\alpha)$ para una muestra de tamaño n, extraída de una población $N_p(\mu, \mathbf{V})$. El conjunto de valores μ^* que satisfacen la ultima desigualdad forman una *region de confianza* para μ, con un coeficiente de confiabilidad de $(1-\alpha)$. Esta ultima expresión representa el interior y la superficie de un elipsoide con centro en $\mu = \overline{\mathbf{x}}$, cuya forma y tamaño dependen de $\widetilde{\mathbf{V}}$ y $\chi^2_{(\alpha,p)}$.

Ejemplo B.1. Los 30 datos artificiales dados en el cuadro B.1 han sido generados en MATLAB 7.1 y hacen referencia a una muestra de niños de mas o menos dos años de edad de una región de gran altitud en Asia. Llamaremos X_1 a la altura, X_2 a la medida de la circunferencia del tórax y X_3 a la circunferencia a la mitad del antebrazo. Todas las mediciones son en centímetros[1]. Asumamos que la muestra es generada por una población normal 3-variante donde

$$\widetilde{\mathbf{V}} = \begin{pmatrix} 30 & 6 & 3 \\ 6 & 2 & 1 \\ 3 & 1 & 2 \end{pmatrix}.$$

Por estudios realizados en años anteriores se sabe que en esta region de Asia los promedios de X_1, X_2 y X_3 son 90, 58 y 16 centímetros respectivamente. Probemos la hipótesis que los niños de esta generación conservan las misma medidas en X_1, X_2 y X_3 que los de anteriores generaciones, a un nivel de significación de 0.05. Es decir

$$H_0 : \mu = (90, 58, 16)',\ \mathbf{V} = \widetilde{\mathbf{V}} \qquad \text{vs} \qquad H_1 : \mu \neq (90, 58, 16)',\ \mathbf{V} = \widetilde{\mathbf{V}}.$$

Con la ayuda de MATLAB 7.1 generemos el estadístico de contraste apropiado y luego verifiquemos si se cumple o no la hipótesis nula. Previamente se importa la tabla de datos del *Bloc de notas* y se archiva como un documento $*.mat$, luego ejecutamos la siguiente serie de comandos:

[1]Este ejemplo esta basado en el ejercicio número 1 propuesto en la página 434 de [5]

Individuo	X_1	X_2	X_3	Individuo	X_1	X_2	X_3
1	90.0	58.3	16.3	16	88.7	57.9	17.8
2	88.3	57.7	16.7	17	90.6	58.1	17.5
3	96.0	58.6	14.9	18	91.7	57.4	14.9
4	79.7	55.4	14.2	19	97.9	58.7	14.8
5	92.3	58.9	16.1	20	88.1	57.3	15.5
6	94.9	58.1	14.6	21	93.4	57.6	15.4
7	94.0	59.5	15.1	22	94.4	57.9	14.9
8	93.2	59.1	17.7	23	95.1	60.3	17.8
9	90.2	57.3	15.7	24	84.6	57.0	17.3
10	93.7	58.5	15.5	25	91.2	57.1	14.9
11	93.1	57.6	16.8	26	91.3	58.2	15.1
12	88.6	55.7	15.2	27	84.5	55.9	14.6
13	87.9	58.5	15.0	28	85.9	56.0	15.8
14	88.4	57.2	17.2	29	95.9	58.9	15.4
15	81.9	56.7	15.7	30	89.3	58.7	14.9

Cuadro B.1: Datos artificiales para el Ejemplo B.1

```
>>% Primero introduzcamos los datos y la matriz de covarianzas.
>> data = load('ejemploB1.mat')
data =
    X: [30x3 double]
>> V=[30 6 3;6 2 1;3 1 2];
>>% Ahora introduzcamos la media muestral y la media a contrastar.
>> Tra_Xbar=mean(X); Xbar = (Tra_Xbar)'
Xbar =
   90.4933
   57.8033
   15.7767
>> Tra_mu_0 = [90 58 16]; mu_0 =(Tra_mu_0)'
mu_0 =
    90
    58
    16
>>% Construyamos el estadístico de prueba.
```

```
>> n=30;p=3;
>> DM=(Xbar-mu_0)
DM =
    0.4933
   -0.1967
   -0.2233
>> IV=inv(V)
IV =
    0.0833   -0.2500   -0.0000
   -0.2500    1.4167   -0.3333
   -0.0000   -0.3333    0.6667
>> lamda=n*(DM)'*(IV)*(DM)
lamda =
  3.8267
```

Para $\alpha = 0.05$, $\chi^2_{(0.05,3)} = 7.815$, luego se acepta la hipótesis $H_0 : \mu = (90, 58, 16)'$, pues $\lambda = 3.8267 < 7.815$. [2]

- **Dos poblaciones**
  Supongamos ahora que tenemos dos matrices de datos muestrales independientes, $\mathbb{X}_{n_1 \times p}$ y $\mathbb{Y}_{n_1 \times p}$, provenientes de distribuciones $N_p(\mu_1, \mathbf{V})$, $N_p(\mu_2, \mathbf{V})$ respectivamente. Consideremos el problema de contrastar las hipótesis

$$H_0 : \mu_1 = \mu_2,\ \mathbf{V} = \widetilde{\mathbf{V}} \qquad \text{vs} \qquad H_1 : \mu_1 \neq \mu_2,\ \mathbf{V} = \widetilde{\mathbf{V}},$$

que equivale a la prueba:

$$H_0 : \mu_1 - \mu_2 = 0,\ \mathbf{V} = \widetilde{\mathbf{V}} \qquad \text{vs} \qquad H_1 : \mu_1 - \mu_2 \neq 0,\ \mathbf{V} = \widetilde{\mathbf{V}} \tag{B.5}$$

Las medias muestrales son $\overline{\mathbf{x}} = \frac{1}{n_1}\sum_{i=i}^{n_1} x_i$, $\overline{\mathbf{y}} = \frac{1}{n_2}\sum_{j=i}^{n_2} y_j$, donde x_i y y_j son los datos muestrales de cada una de las matrices anteriormente mencionadas.
Como $(\overline{\mathbf{x}} - \overline{\mathbf{y}})$ se distribuye $N_p(\mu_1 - \mu_2, (n_1^{-1} + n_2^{-1})\mathbf{V})$, la construcción del contraste de razón de verosimilitudes es similar al caso de una población; luego el estadístico de contraste es

$$\lambda = \underbrace{\frac{n_1 n_2}{n_1 + n_2}(\overline{\mathbf{x}} - \overline{\mathbf{y}})'\widetilde{\mathbf{V}}^{-1}(\overline{\mathbf{x}} - \overline{\mathbf{y}})}_{\chi^2_0} \sim \chi^2_p. \tag{B.6}$$

[2] La tabla de valores de $\chi^2_{(\alpha,p)}$ se puede consultar en [2] o [8].

De forma análoga, *la region critica* para contrastar la hipótesis (B.5) es determinada por los puntos que satisfacen la desigualdad $\chi_0^2 \geq \chi^2_{(\alpha,p)}$. La *region de confianza* para estimar la diferencia entre los dos vectores de medias poblacionales es

$$\frac{n_1 n_2}{n_1+n_2}\Big((\overline{\mathbf{x}}-\overline{\mathbf{y}})-(\mu_1-\mu_2)\Big)'\widetilde{\mathbf{V}}^{-1}\Big((\overline{\mathbf{x}}-\overline{\mathbf{y}})-(\mu_1-\mu_2)\Big) \leq \chi_p^2. \tag{B.7}$$

Ejemplo B.2. Un grupo de investigadores ambientales desea comprobar si dos plantas de tratamiento de aguas residuales, ubicadas en ciudades diferentes, presentan comportamientos similares en la depuración del agua. Sospechan que ambas ciudades tiene un alto nivel de contaminación en sus aguas causada por los desperdicios industriales. Para ello realizan la medición de cuatro variables durante 45 dias en las dos plantas. Las variables medidas son:

- X_1 : Demanda Biológica de Oxígeno empleado por los microorganismos a lo largo de un periodo de cinco días para descomponer la materia orgánica de las aguas residuales, (DBO_5), en la corriente de entrada, dada en miligramos por litro (mg/L).
- X_2 : Temperatura dada en grados Fahrenheit (°F).
- X_3 : Solidos en suspensión dados en miligramos por litro (mg/L).
- X_4 : La DBO_5 en la corriente de salida dada en miligramos por litro (mg/L).

Los datos se observan en el cuadro B.2. Si suponemos que las dos muestras tomadas en las plantas de tratamiento tienen un comportamiento normal multivariate, con matriz de covarianzas

$$\widetilde{\mathbf{V}} = \begin{pmatrix} 1086 & -18 & 19 & 105 \\ -18 & 23 & 62 & -9 \\ 19 & 62 & 20938 & -22 \\ 105 & -9 & -22 & 44 \end{pmatrix},$$

probemos la hipótesis nula $\mu_1 - \mu_2 = 0$, frente a la hipótesis alternativa $\mu_1 - \mu_2 \neq 0$ a un nivel de significación de 0.05.
Con la ayuda de MATLAB 7.1 generemos el estadístico de contraste apropiado y luego verifiquemos si se cumple o no la hipótesis nula. Previamente se importa las tablas de datos de *Excel* y se archiva como un documento $*.mat$, luego ejecutamos la siguiente serie de comandos:

\	Planta A				Planta B				\	Planta A				Planta B			
Dia	$\mathbf{X_1}$	$\mathbf{X_2}$	$\mathbf{X_3}$	$\mathbf{X_4}$	$\mathbf{X_1}$	$\mathbf{X_2}$	$\mathbf{X_3}$	$\mathbf{X_4}$	**Dia**	$\mathbf{X_1}$	$\mathbf{X_2}$	$\mathbf{X_3}$	$\mathbf{X_4}$	$\mathbf{X_1}$	$\mathbf{X_2}$	$\mathbf{X_3}$	$\mathbf{X_4}$
1	219	52	1380	23	177	48	1170	17	**24**	199	50	1330	8	161	52	1240	19
2	219	47	1410	23	160	47	1190	18	**25**	209	46	1360	16	181	54	1110	11
3	127	48	1410	8	183	50	1230	17	**26**	185	46	1310	15	156	54	1120	15
4	195	45	1380	26	144	53	1070	25	**27**	203	47	1170	17	156	57	1180	15
5	219	45	1340	20	159	53	1090	18	**28**	225	50	1360	28	139	56	1180	6
6	192	46	1380	19	159	50	1130	18	**29**	225	52	1190	23	167	54	1160	17
7	239	50	1320	24	118	52	1130	8	**30**	225	51	1290	23	155	56	1220	13
8	177	47	1330	18	159	47	1110	19	**31**	119	50	1290	10	172	54	1210	11
9	177	50	1350	18	147	52	1080	14	**32**	174	46	1220	16	212	55	1120	15
10	166	48	1350	6	154	54	1110	16	**33**	190	52	1280	18	253	58	1060	13
11	190	50	1310	16	179	56	1020	14	**34**	238	50	1290	22	253	53	1140	13
12	171	48	1360	19	222	47	1040	15	**35**	225	50	1240	24	134	55	1140	6
13	201	50	1310	13	222	48	1180	15	**36**	257	54	1250	24	177	53	1100	14
14	188	53	1300	19	162	50	1180	10	**37**	257	50	1030	24	192	52	1180	13
15	174	60	1230	19	196	49	1150	22	**38**	121	51	1030	10	179	55	1250	22
16	174	50	1330	19	212	53	1220	25	**39**	163	50	1170	14	163	56	1170	13
17	136	50	1330	7	205	56	1210	18	**40**	144	50	1300	15	216	58	1120	10
18	159	42	1470	14	192	55	1210	15	**41**	172	49	1250	10	216	56	1140	10
19	216	44	1400	18	170	53	1310	9	**42**	226	48	1190	18	137	56	1140	8
20	204	51	1400	13	170	52	1330	9	**43**	184	48	1260	11	185	55	1140	18
21	250	47	1310	21	106	54	1330	9	**44**	184	51	1280	11	156	58	1210	13
22	233	54	1370	22	140	53	1020	27	**45**	184	49	1280	8	160	58	1250	8
23	233	53	1330	22	143	50	1310	25									

Fuente: Datos tomados de la página 646 de [13].

Cuadro B.2: Datos para el Ejemplo 2.2

```
>> data1 = load('ejemplo_plantaT_CA.mat')
data1 =
    X: [45x4 double]
>> data2 = load('ejemplo_plantaT_CB.mat')
data2 =  Y: [45x4 double]
>> % Encontremos los vectores de medias.
>> Tra_Xbar=mean(X); Xbar =(Tra_Xbar)'
Xbar =
  1.0e+003 *
    0.1948
    0.0493
    1.2993
    0.0172
>> Tra_Ybar=mean(Y); Ybar =(Tra_Ybar)'
Ybar =
  1.0e+003 *
    0.1733
    0.0533
    1.1644
    0.0148
>> % Introduzcamos la matriz de covarianza conocida y hallemos su inversa.
>> V=[1086 -18 19 105; -18 23 62 -9;19 62 20938 -22; 105 -9 -22 44];
>> IV=inv(V)
IV =
    0.0012   -0.0002   -0.0000   -0.0029
   -0.0002    0.0476   -0.0001    0.0101
   -0.0000   -0.0001    0.0000    0.0000
   -0.0029    0.0101    0.0000    0.0317
>> % Construyamos el estadístico de prueba apropiado.
>> nx=45;ny=45;p=4;
>> dm_xy=(Xbar-Ybar)
dm_xy =
   21.5333
   -3.9333
  134.8889
    2.3556
```

```
>> a=(nx*ny)/(nx+ny)
a =
   22.5000
>> lamda_xy=(a)*(dm_xy)'*(IV)*(dm_xy)
lamda_xy =
   45.3607
```

Para $\alpha = 0.05$, $\chi^2_{(0.05,4)} = 9.488$, luego se rechaza la hipótesis $H_0 : \mu_1 - \mu_2 = 0$, pues $\lambda = 45.3607 > 9.488$. [3]

B.1.2. Matriz de covarianzas desconocida

En la mayoría de las situaciones practicas, rara vez se conocemos la matriz de covarianzas. Desarrollamos ahora una prueba de hipótesis para un vector de medias μ de una población normal multivariante con matriz de varianzas y covarianzas desconocida.
En una población normal univariada, el problema de verificar si la media es igual a cierto valor especifico, cuando se desconoce la varianza, se realiza mediante la estadística *t-Student* con $n-1$ grados de libertad. Para el campo multivariate se tiene una expresión análoga a esta, conocida como la estadística T^2 de Hotelling, presentada aquí anteriormente en el **Apéndice A**.

- **Una población**
  Sea $\mathbf{x}_1, \ldots, \mathbf{x}_n$, con $n > p$, una muestra aleatoria de una población $N_p(\mu, \mathbf{V})$ donde $\mathbf{V}$ es desconocida. Con base en esta muestra queremos realizar la prueba de hipótesis:

$$H_0 : \mu = \mu_0,\ \mathbf{V} = cualquiera \qquad \text{vs} \qquad H_1 : \mu \neq \mu_0,\ \mathbf{V} = cualquiera. \tag{B.8}$$

  Construyamos el contraste de razón de verosimilitudes calculando la función de verosimilitud bajo H_0 y H_1 según la función de soporte dada en (2.15), para lo cual debemos encontrar los estimadores MV de μ y $\mathbf{V}$ en H_0 y H_1.
  Bajo H_0 el estimador de μ es μ_0. Ahora operando en forma adecuada y eliminando la constante aditiva dada en (2.15), ya que esta desaparece al derivar, podemos escribir el soporte como

$$L(\mu, \mathbf{V}) = -\frac{n}{2} \ln |\mathbf{V}| - \frac{n}{2} tra \left[\mathbf{V^{-1}} \mathbf{S}_0\right] \tag{B.9}$$

[3] La tabla de valores de $\chi^2_{(\alpha,p)}$ se puede consultar en [2] o [8].

donde

$$\mathbf{S}_0 = \frac{1}{n}\sum_{i=1}^{n}(x_i - \mu_0)(x_i - \mu_0)'.$$

Derivemos la ecuación (B.9) respecto a $\mathbf{V}$ e igualando a cero para encontrar su estimador MV. Aquí se aplica nuevamente las propiedades a) y g) de la sección 1.2.4 y el hecho que $\mathbf{V}$ es simétrica. Luego tenemos:

$$\begin{aligned}\frac{\partial}{\partial \mathbf{V}}L(\mu, \mathbf{V}) &= -\frac{n}{2}\left\{\frac{\partial}{\partial \mathbf{V}}\ln|\mathbf{V}| + \frac{\partial}{\partial \mathbf{V}}tra\left[\mathbf{S}_0\mathbf{V}^{-\mathbf{1}}\right]\right\} = 0 \\ &= -\frac{n}{2}\left\{(\mathbf{V}')^{-1} - [\mathbf{V}^{-1}\mathbf{S}_0\mathbf{V}^{-1}]\right\} = 0 \\ &= (\mathbf{V})^{-1} - \mathbf{V}^{-1}\mathbf{S}_0\mathbf{V}^{-1} = 0.\end{aligned}$$

Multiplicando por la matriz $\mathbf{V}$, primero a derecha y después a izquierda obtenemos:

$$\begin{aligned}\mathbf{V}\times\mathbf{V}^{-1} - (\mathbf{V}\times\mathbf{V}^{-1})\mathbf{S}_0\mathbf{V}^{-1} &= 0 \\ \mathbf{I} - \mathbf{S}_0\mathbf{V}^{-1} &= 0 \\ \mathbf{I}\times\mathbf{V} - \mathbf{S}_0(\mathbf{V}^{-1}\times\mathbf{V}) &= 0 \\ \mathbf{V} - \mathbf{S}_0 &= 0 \\ \mathbf{V} &= \mathbf{S}_0.\end{aligned}$$

Luego el estimador MV bajo H_0 de $\mathbf{V}$ es $\mathbf{S}_0$. Sustituyendo este resultado en (B.9) se tiene que el soporte para H_0 es:

$$L(H_0) = -\frac{n}{2}\ln|\mathbf{S_0}| - \frac{np}{2}. \tag{B.10}$$

Bajo H_1 los estimadores MV son $\overline{\mathbf{x}}$ y $\mathbf{S}$, luego sustituyendo en (2.15) tenemos que el soporte para H_1 es:

$$L(H_1) = -\frac{n}{2}\ln|\mathbf{S}| - \frac{np}{2}. \tag{B.11}$$

Si asumimos una muestra de tamaño grande, por el **Teorema 2.1** decimos que la diferencia de soportes es:

$$\begin{aligned}\delta_L &= 2[L(H_1) - L(H_0)] \\ &= 2\left[-\frac{n}{2}\ln|\mathbf{S}| - \frac{np}{2} + \frac{n}{p}\ln|\mathbf{S}_0| + \frac{np}{2}\right] \\ &= n\left[\ln|\mathbf{S}| - \ln|\mathbf{S}_0|\right]\end{aligned}$$

$$= n \ln \frac{|\mathbf{S}_0|}{|\mathbf{S}|}. \tag{B.12}$$

El resultado obtenido en (B.12) tiene una distribución χ^2 con grados de libertad igual a la diferencia de las dimensiones del espacio en que se mueven los parámetros bajo ambas hipótesis.
En [2] se demuestra que

$$\frac{|\mathbf{S}_0|}{|\mathbf{S}|} = 1 + \frac{\mathbf{T}_0^2}{n-1},$$

luego podemos escribir la ecuación (B.12) como

$$\delta_L = n \ln \left(1 + \frac{\mathbf{T}^2}{n-1}\right), \tag{B.13}$$

donde el estadístico $\mathbf{T}_0^2$ es:

$$\mathbf{T}_0^2 = (n-1)(\overline{\mathbf{x}} - \mu_0)' \mathbf{S}^{-1} (\overline{\mathbf{x}} - \mu_0), \tag{B.14}$$

el cual sigue la distribución $\mathbf{T}^2$ de Hotelling, con p y $n-1$ grados de libertad, lo cual se nota $\mathbf{T}_0^2 \sim \mathbf{T}_{(p,n-1)}^2$, según lo estudiado en el **Apéndice A**, sección **A.2.1**.
Puesto que la diferencia de soportes es una función monótona de $\mathbf{T^2}$, podemos utilizar el estadístico (B.14) en lugar de la razón de verosimilitudes.
Así como se muestra que la estadística univariada t-Student es un caso especial de la distribución F a través de la relación $t_{(n)}^2 = F_{(1,n)}$, la distribución de la estadística $\mathbf{T}^2$ de Hotelling se relaciona con la F a través de

$$\frac{\mathbf{T}^2}{n-1} \left(\frac{n-p}{p}\right) \sim F_{(p,n-p)}. \tag{B.15}$$

Esta ultima relación se puede observar mas a detalle en [2].
En resumen, sustituyendo (B.14) en (B.15), el estadístico de contraste es

$$\underbrace{\left(\frac{n-p}{p}\right) (\overline{\mathbf{x}} - \mu_0)' \mathbf{S}^{-1} (\overline{\mathbf{x}} - \mu_0)}_{F_0} \sim F_{(p,n-p)}. \tag{B.16}$$

Rechazamos H_0 para valores significativamente grandes del estadístico; es decir $F_0 \geq F_{(\alpha,p,n-p)}$, con α como nivel de significación.

Ejemplo B.3. En una tertulia organizada en el mes de diciembre del 2006, un profesor de economía realiza una pequeña charla sobre el desempleo en Colombia, en

medio de esta arroja la siguiente frase: "Podemos asegurar que en los dos últimos años en promedio la tasa de ocupación en Colombia es del 53.0 %, la tasa de desempleo del 11.0 % y la tasa de subempleo del 34.0 %". Un estudiante, un poco escéptico ante estas afirmaciones, decide consultar por Internet las últimas estadísticas del DANE y encuentra la información suministrada en el cuadro B.3.
Luego decide realizar la siguiente prueba de hipótesis:

$$H_0 : \mu = (53.0, 11.0, 34.0) \qquad \text{vs} \qquad H_1 : \mu \neq (53.0, 11.0, 34.0),$$

con $\mathbf{V}$ desconocida para ambos casos. Para ello genera el estadístico de prueba apropiado con la ayuda de MATLAB 7.1. Sus resultados fueron:

```
>> % Introducimos nuestra base de datos
>> data = load('ejemploB3.mat')
data =
    X: [25x3 double]
>> X1=X(:,1); %tasa de ocupación
>> X2=X(:,2); %tasa de desempleo.
>> X3=X(:,3); %tasa de subempleo.
>> % Hallamos la media y la matriz de covarianzas muestrales.
>> Tra_Xbar = mean(X);Xbar =(Tra_Xbar)'
Xbar =
   52.3320
   11.9760
   32.3360
>> S= cov(X)
S =
    2.4289   -1.1792   -1.2604
   -1.1792    1.1252   -0.4999
   -1.2604   -0.4999    6.3491
>> IS= inv(S)
IS =
   1.4846    1.7479    0.4324
    1.7479    2.9788    0.5815
    0.4324    0.5815    0.2891
>> % Construyamos el estadístico de prueba.
>> n=25; p=3;
```

Mes y Año	Tasa de ocupación	Tasa de desempleo	Tasa de subempleo
$oct/04$	53.8	12.4	31.8
$nov/04$	53.9	11.7	31.9
$dic/04$	52.4	12.1	30.6
$ene/05$	51.5	13.2	28.2
$feb/05$	51.6	14.0	30.4
$mar/05$	51.6	13.1	29.5
$abr/05$	52.1	12.0	32.0
$may/05$	52.3	12.5	33.9
$jun/05$	52.1	11.4	31.5
$jul/05$	53.6	11.8	33.9
$ago/05$	52.7	11.3	32.5
$sep/05$	53.3	11.2	32.6
$oct/05$	55.0	10.0	32.6
$nov/05$	54.4	10.2	31.1
$dic/05$	54.5	10.4	31.2
$ene/06$	51.3	13.4	28.4
$feb/06$	51.7	13.2	29.9
$mar/06$	53.4	11.3	30.0
$abr/06$	51.4	12.1	31.9
$may/06$	52.1	11.8	34.3
$jun/06$	54.2	10.5	$34,4$
$jul/06$	50.1	12.6	34.1
$ago/06$	49.7	12.9	37.7
$sep/06$	49.0	12.9	36.4
$oct/06$	50.6	11.4	37.6

Fuente: DANE-ECH

Cuadro B.3: Datos para el Ejemplo B.3

```
>> Tra_mu_0 = [53.0 11.0 34.0]; mu_0 =(Tra_mu_0)'
mu_0 =
    53
    11
    34
>> DM=(Xbar-mu_0)
DM =
   -0.6680
    0.9760
   -1.6640
>> b=(n-p)/p
b =
    7.3333
>> F_0 = b*(DM)'*(IS)*(DM)

F_0 =
    8.0201
```

Para $\alpha = 0.01$, $F_{(0.01,3,22)} = 4.82$, luego rechaza la hipótesis dada por el profesor, pues $F_0 = 8.0201 > 4.82$. [4]

- **Dos poblaciones**

  Supongamos ahora que tenemos dos matrices de datos muestrales independientes, $\mathbb{X}_{n_1\times p}$ y $\mathbb{Y}_{n_1\times p}$, provenientes de distribuciones $N_p(\mu_1, \mathbf{V})$, $N_p(\mu_2, \mathbf{V})$ respectivamente. Consideremos el problema de contrastar las hipótesis

$$H_0 : \mu_1 - \mu_2 = 0,\ \mathbf{V} = cualquiera \quad \text{vs} \quad H_1 : \mu_1 - \mu_2 \neq 0,\ \mathbf{V} = cualquiera. \tag{B.17}$$

  Si $\overline{\mathbf{x}}, \overline{\mathbf{y}}, \mathbf{S}_1, \mathbf{S}_2$ son los vectores de medias y las matrices de covarianzas de los datos muestrales $\mathbb{X}_{n_1\times p}$, $\mathbb{Y}_{n_2\times p}$ respectivamente, consideramos la estimación conjunta centrada de $\mathbf{V}$

$$\widetilde{\mathbf{S}} = (n_1\mathbf{S}_1 + n_2\mathbf{S}_2)/(n_1 + n_2 - 2)$$

  Según las propiedades estudiadas en la sección **A.2.1**, el estadístico

$$\mathbf{T}_0^2 = \frac{n_1 n_2}{n_1 + n_2}(\overline{\mathbf{x}} - \overline{\mathbf{y}})'\widetilde{\mathbf{S}}^{-1}(\overline{\mathbf{x}} - \overline{\mathbf{y}}) \tag{B.18}$$

[4] La tabla de valores de $F_{(\alpha,p,n-p)}$ se puede consultar en [2] o [8].

sigue la distribución $\mathbf{T}^2$ de Hotelling con p y $n_1 + n_2 - 2$ grados de libertad, lo cual se nota $\mathbf{T}_0^2 \sim \mathbf{T}^2_{(p,n_1+n_2-2)}$. Análogo al caso anterior y a la relación dada en (B.15), el estadístico de contraste es

$$\underbrace{\left[\frac{n_1+n_2-1-p}{(n_1+n_2-2)p}\right]\frac{n_1 n_2}{n_1+n_2}(\overline{\mathbf{x}}-\overline{\mathbf{y}})'\widetilde{\mathbf{S}}^{-1}(\overline{\mathbf{x}}-\overline{\mathbf{y}})}_{F_0} \sim F_{(p,n_1+n_2-1-p)} \tag{B.19}$$

De igual forma, la region critica esta determinada por los puntos tales que $F_0 \geq F_{(\alpha,p,n_1+n_2-1-p)}$ con α como nivel de significación.

Ejemplo B.4. Se desea comparar el promedio del rendimiento atlético en pruebas de velocidad de los continentes americano y asiático en relación a los récords nacionales en las pruebas de $100m$, $200m$ y $400m$. Para ello se ha tomado dos muestras de los países de cada continente con los respectivos tiempos en cada prueba.
El cuadro B.4 suministra los datos con las variables:
X_1 : Record en los $100m$ dado en segundos.
X_2 : Record en los $200m$ dado en segundos.
X_3 : Record en los $400m$ dado en segundos.
Con la ayuda de MATLAB 7.1 generemos el estadístico de contraste apropiado y luego verifiquemos si se cumple o no la hipótesis nula. Previamente se importa las tablas de datos del *Bloc de notas* y se archiva como un documento $*.mat$, luego ejecutamos la siguiente serie de comandos:

```
>> data1 = load('ejemploB4america.mat')
data1 =
    X: [11x3 double]
>> data2 = load('ejemploB4asia.mat')
data2 =
    Y: [14x3 double]
>> % Encontremos los vectores de las medias.
>> Tra_Xbar=mean(X); Xbar = (Tra_Xbar)'
Xbar =
   10.3855
   20.8464
   46.3418
>> Tra_Ybar=mean(Y);Ybar = (Tra_Ybar)'
Ybar =
```

America Y Asia							
País	$\mathbf{X_1}$	$\mathbf{X_2}$	$\mathbf{X_3}$	**País**	$\mathbf{X_1}$	$\mathbf{X_2}$	$\mathbf{X_3}$
Argentina	10.39	20.81	46.84	China	10.51	21.04	47.30
Bermuda	10.28	20.58	45.91	India	10.60	21.42	45.73
Brazil	10.22	20.43	45.21	Indonesia	10.59	21.49	47.80
Canada	10.17	20.22	45.68	Israel	10.71	21.00	47.80
Chile	10.34	20.80	46.20	Japan	10.34	20.81	45.86
Colombia	10.43	21.05	46.10	Korea	10.34	20.89	46.90
Costa Rica	10.94	21.90	48.66	P Korea	10.91	21.94	47.30
Dom Rep	10.14	20.65	46.80	Malaysia	10.40	20.92	46.30
Guatemala	10.98	21.82	48.40	Philippines	10.78	21.64	46.24
Mexico	10.42	21.30	46.10	Singapore	10.38	21.28	47.40
USA	9.93	19.75	43.86	Taiwan	10.59	21.29	46.80
				Thailand	10.39	21.09	47.91
				Turkey	10.71	21.43	47.60
				USSR	10.07	20.00	44.60

Cuadro B.4: Datos para el Ejemplo B.4

```
   10.5229
   21.1600
   46.8243
>> % Encontremos las matrices de covarianzas muestrales.
>> SX=cov(X)
SX =
    0.1019    0.1969    0.3892
    0.1969    0.4186    0.7908
    0.3892    0.7908    1.8270
>> SY=cov(Y)
SY =
    0.0482    0.0891    0.1033
    0.0891    0.2125    0.2413
    0.1033    0.2413    0.9358
>> % Hallemos la estimación centrada de la matriz de covarianzas común.
>> nx=11; % datos de X.
>> ny=14; % datos de Y.
```

```
>> a=1/(nx+ny-2)
a =
    0.0435
>> Smono=a*(nx*SX+ny*SY)
Smono =
    0.0781    0.1484    0.2490
    0.1484    0.3296    0.5251
    0.2490    0.5251    1.4434
>> ISmono = inv(Smono)
ISmono =
   91.4467  -38.1856   -1.8847
  -38.1856   23.1627   -1.8386
   -1.8847   -1.8386    1.6868
>> % Hallemos el estadístico T^2 dado en la ecuación B.18.
>> b=(nx*ny)/(nx+ny)
b =
    6.1600
>> T2=b*(Xbar-Ybar)'*ISmono*(Xbar-Ybar)
T2 =
    1.8485
>> % Luego nuestro estadístico de prueba sera:
>> p=3;
>> c=(nx+ny-1-p)/((nx+ny-2)*p)
c =
    0.3043
>> F_0=c*T2
F_0 =
    0.5626
```

Para $\alpha = 0.01$, $F_{(0.01,3,21)} = 4.87$, luego aceptamos la hipótesis de que en promedio el rendimiento atlético en pruebas de velocidad de los continentes americano y asiático es igual, pues $F_0 = 0.5626 < 4.87$. [5]

[5]La tabla de valores de $F_{(\alpha,p,n_1+n_2-1-p)}$ se puede consultar en [2] o [8].

B.2. Contrastes sobre la matriz de covarianzas

Anteriormente se definió la matriz de covarianzas junto con algunas de sus propiedades, se hizo su estimación, se determinó su distribución bajo el supuesto de normalidad y se empleo en la inferencia sobre el vector de medias. Ahora presentamos la distribución de la matriz de covarianzas y la inferencia sobre esta para una o varias poblaciones, basados en el contraste de la razón de verosimilitudes el cual aplicamos para hacer contrastes de vectores de medias.

- **Una población**
  Veamos tres contrastes sobre la matriz de covarianzas de una población normal.

  1. **Contraste de un valor particular:** Supongamos una muestra aleatoria de n observaciones vectoriales $\mathbf{x}_1, \ldots, \mathbf{x}_n$ de una población $N_p(\mu, \mathbf{V})$, se desea contrastar las hipótesis:

  $$H_0 : \mathbf{V} = \widetilde{\mathbf{V}}, \quad \mu \equiv cualquiera, \quad \text{vs} \quad H_1 : \mathbf{V} \quad y \quad \mu \equiv cualquiera. \tag{B.20}$$

  La razón de maxima verosimilitud suministra la estadística de prueba para el anterior contraste. Si asumimos una muestra grande, hallamos el doble de la diferencia de soportes para las dos hipótesis H_0 y H_1. Los estimadores MV para los parámetros de la distribución normal multivariante restringidos a H_0 son $\overline{\mathbf{x}}$ y $\widetilde{\mathbf{V}}$ respectivamente, mientras que los estimadores en todo el espacio de parámetros son $\overline{\mathbf{x}}$ y $\mathbf{S}$. En [1] se muestra que el doble de la diferencia de soportes reduce el contraste a:

  $$\delta_L = \underbrace{n \ \ln \left[\frac{|\widetilde{\mathbf{V}}|}{|\mathbf{S}|} \right] + n \ tra(\widetilde{\mathbf{V}}^{-1}\mathbf{S}) - np}_{\chi_0^2}, \tag{B.21}$$

  el cual tiene una distribución χ^2 con $p(p+1)/2$ grados de libertad, valor que corresponde al numero de términos distintos de $\mathbf{V}$. La region critica está determinada por $\chi_0^2 > \chi^2_{(\alpha, p(p+1)/2)}$ con α como el nivel de significancia.

  2. **Contraste de independencia:** Otro contraste de interés (consecuencia del anterior) es el de independencia, donde suponemos que la matriz $\widetilde{\mathbf{V}}$ es diagonal. Deseamos realizar la prueba de hipótesis:

  $$H_0 : \mathbf{V} = Diagonal, \ \mu \equiv cualquiera, \quad \text{vs} \quad H_1 : \mathbf{V} \quad y \quad \mu \equiv cualquiera. \tag{B.22}$$

En [1] y [2] se muestra que el doble de la diferencia de los soportes reduce el contraste a:

$$\delta_L = -n \ln |\mathbf{R}|, \tag{B.23}$$

siendo $\mathbf{R}$ la matriz de correlaciones. El estadístico δ_L es asintóticamente χ^2 con grados de libertad igual a

$$p(p+1)/2 - p = p(p-1)/2 \equiv q.$$

Se rechaza H_0 cuando $\delta_L > \chi^2_{(\alpha,q)}$. Si las variables son independientes, tendremos que $\mathbf{R} \approx \mathbf{I}$, luego $\delta_L \approx 0$ y es probable que $\chi^2_{(\alpha,q)} = -n \ln |\mathbf{R}|$ no sea significativo.

3. **Contraste de Esfericidad:** La hipótesis respecto a la independencia y a la homocedasticidad de las variables, asumida en la mayoría de los modelos de regresión lineal y en el análisis de varianza clásico, se expresa como
$H_0 : \mathbf{V} = \sigma^2\mathbf{I}$, donde σ^2 es la varianza común y desconocida. Este contraste recibe el nombre de *esfericidad*, ya que la distribución de las variables tienen curvas de nivel que son esferas. Las hipótesis a contrastar son:

$$H_0 : \mathbf{V} = \sigma^2\mathbf{I},\ \mu \equiv cualquiera, \quad \text{vs} \quad H_1 : \mathbf{V} \quad y \quad \mu \equiv cualquiera. \tag{B.24}$$

Si asumimos una muestra grande, hallamos el doble de la diferencia de soportes para las dos hipótesis H_0 y H_1. Los estimadores MV para los parámetros de la distribución normal multivariante restringidos a H_0 son $\overline{\mathbf{x}}$ y $\widehat{\sigma}^2 = tra(\mathbf{S}/p)$, mientras que los estimadores en todo el espacio de parámetros son $\overline{\mathbf{x}}$ y $\mathbf{S}$.
En [1] se muestra que el doble de la diferencia de soportes reduce el contraste a:

$$\delta_L = np \ln(\widehat{\sigma}^2) - n \ln |\mathbf{S}|, \tag{B.25}$$

que se distribuye asintóticamente como una χ^2 con $\frac{(p+2)(p+1)}{2}$ grados de libertad. La region critica está determinada por $\delta_L > \chi^2_{(\alpha,(p+2)(p+1)/2)}$ con α como el nivel de significancia.

Ejemplo B.5. Una empresa que fabrica pizzas no perecederas esta realizando un estudio sobre la relación existente entre algunas sustancias nutritivas de su producto. Para ello toma de ciertos supermercados a los cuales provee una muestra aleatoria de 40 pizzas congeladas. Cada una de las pizzas se convirtió en puré y se mezcló a conciencia, después de lo cual se tomo una muestra de cada mezcla para el análisis de las sustancias nutritivas. Las variables medidas en cada mezcla fueron:

- X_1 : Cantidad de proteína por cada 100 Gramos.

- X_2 : Cantidad de grasa por cada 100 Gramos.
- X_3 : Cantidad de ceniza por cada 100 Gramos.
- X_4 : Cantidad de sodio por cada 100 Gramos.
- X_5 : Cantidad de carbohidratos por cada 100 Gramos.

Esta empresa quiere determinar la relación que existe en estas variables y para ello decide realizar inicialmente un contraste de esfericidad; es decir desea contrastar las hipótesis:

$$H_0 : \mathbf{V} = \sigma^2\mathbf{I},\ \mu \equiv cualquiera, \quad \text{vs} \quad H_1 : \mathbf{V} \quad y \quad \mu \equiv cualquiera.$$

Los datos se observan en el cuadro B.5.
Con la ayuda de MATLAB 7.1 se genera el estadístico de contraste apropiado y luego se verifica si se cumple o no la hipótesis nula. Previamente se importa la tabla de datos del *Bloc de notas* y se archiva como un documento $*.mat$, luego ejecutamos la siguiente serie de comandos:

```
>> % Introducimos nuestra base de datos
>> data = load('ejemploB5.mat')
data =
    X: [40x5 double]
>> % Hallamos la matriz de covarianza muestral.
>> S=cov(X)
S =

   56.7312   13.7484    0.0134    0.0050   14.9465
   13.7484   96.4544    1.8195   -0.2458    3.1393
    0.0134    1.8195    1.7404    0.0458   -1.8008
    0.0050   -0.2458    0.0458    0.3087   -0.2741
   14.9465    3.1393   -1.8008   -0.2741   63.3105
>> % Hallemos la traza y el determinate de S.
>> Traza_S=trace(S)
Traza_S =
  218.5451
>> det_S=det(S)
det_S =
  1.5826e+005
```

Obs.	X_1	X_2	X_3	X_4	X_5	Obs.	X_1	X_2	X_3	X_4	X_5
1	8.578	10.954	2.990	1.317	23.059	21	11.824	25.184	4.746	1.727	14.030
2	15.090	16.103	1.444	1.135	16.308	22	21.659	36.645	4.444	1.961	19.317
3	9.186	25.668	5.079	1.583	0.265	23	3.365	14.019	3.266	0.305	0.002
4	13.299	12.482	1.171	1.666	1.354	24	14.003	13.900	4.665	0.383	0.017
5	4.401	38.441	3.156	1.278	11.283	25	19.922	17.785	1.321	1.338	0.489
6	21.722	17.547	4.998	1.544	19.679	26	8.449	27.767	1.060	0.759	15.464
7	7.855	34.740	2.492	0.883	0.038	27	21.254	43.075	1.452	0.966	22.271
8	23.515	25.600	3.364	1.216	19.043	28	29.585	36.636	4.790	0.350	22.676
9	14.200	34.509	3.833	0.004	18.964	29	27.085	37.041	3.331	1.580	0.951
10	13.532	25.092	5.476	1.027	0.827	30	24.135	32.383	4.771	0.426	12.602
11	24.866	40.636	2.743	0.206	14.114	31	4.988	12.754	1.441	0.314	19.180
12	11.817	15.946	4.084	0.815	19.497	32	15.622	32.556	4.927	0.815	12.102
13	21.544	18.102	1.868	0.105	20.055	33	17.076	15.965	4.097	1.883	11.775
14	13.824	18.745	3.105	0.299	0.506	34	13.359	33.389	2.801	0.768	14.490
15	26.323	16.088	1.220	0.622	16.663	35	13.304	31.556	4.110	0.337	16.919
16	10.989	18.000	1.316	1.793	23.563	36	9.076	15.253	3.768	0.645	19.254
17	25.555	18.335	3.298	1.468	18.435	37	22.784	39.038	1.070	0.821	21.657
18	28.493	16.105	3.010	0.799	24.774	38	16.738	41.470	3.024	1.011	12.598
19	4.270	28.496	3.297	0.338	15.727	39	17.885	17.528	1.689	1.049	19.815
20	24.486	31.169	3.182	1.282	11.216	40	29.313	16.682	2.515	0.032	13.109

Fuente: Datos tomados de la página 331 de [5].

Cuadro B.5: Datos para el Ejemplo B.5

```
>> % Luego el estimador de sigma^2 es:
>> p=5;n=40;
>> sigma2=Traza_S/p
sigma2 =
   43.7090
>> % Construyamos el estadístico de prueba.
>> delta_L1=(n*p*log(sigma2))-(n*log(det_S))
delta_L1 =
  276.6305
```

Puesto que $\frac{(p+2)(p+1)}{2} = 21$, con $\alpha = 0.05$ se tiene que: $\chi^2_{(0.05,21)} = 41.401$. Luego se

rechaza que las variables tengan la misma varianza y estén incorreladas, es decir la hipótesis $H_0 : \mathbf{V} = \sigma^2\mathbf{I},\ \mu \equiv cualquiera$, pues

$$\delta_{L1} = 276.6305 > 41.401.^{6}$$

Ahora probemos la hipótesis de independencia:

$$H_0 : \mathbf{V} = Diagonal,\ \mu \equiv cualquiera, \quad \text{vs} \quad H_1 : \mathbf{V} \quad y \quad \mu \equiv cualquiera.$$

En MATLAB 7.1 generemos el estadístico de contraste apropiado y luego verifiquemos si se cumple o no la hipótesis nula.

```
>> R=corrcoef(S)
R =
    1.0000    0.0490   -0.2313   -0.2001    0.1892
    0.0490    1.0000    0.5309   -0.5262   -0.2472
   -0.2313    0.5309    1.0000    0.1393   -0.8582
   -0.2001   -0.5262    0.1393    1.0000   -0.6008
    0.1892   -0.2472   -0.8582   -0.6008    1.0000
>> det_R=det(R)
det_R =
  5.9706e-018
>> delta_L2=-n*log(det_R)
delta_L2 =
  1.5864e+003
```

Puesto que $\frac{p(p-1)}{2} = 10$, con $\alpha = 0.05$ se tiene que: $\chi^2_{(0.05,10)} = 25.188$. Luego se rechaza sin duda la hipótesis de independencia, pues $\delta_{L2} = 1586.4 > 25.188$.[7]

- **Varias poblaciones**

  En esta parte tratamos de contrastar la hipótesis sobre la igualdad de las matrices de covarianzas asociadas a varias poblaciones normales multivariantes, mediante la información contenida en una muestra aleatoria de cada una de ellas.
  Sean $x_{1_g}, \ldots, x_{n_g}$, con $g = 1, \ldots, G$, una muestra aleatoria de una población $N_p(\mu_g, \mathbf{V}_g)$; es decir se dispone de G-muestras aleatorias independientes de poblaciones normales multivariantes. Las hipótesis a contrastar son:

$$H_0 : \mathbf{V}_1 = \ldots = \mathbf{V}_G = \mathbf{V}, \quad \mu_g \equiv cualquiera,$$

[6] La tabla de valores de $\chi^2_{(\alpha,p)}$ se puede consultar en [2] o [8].
[7] La tabla de valores de $\chi^2_{(\alpha,p)}$ se puede consultar en [2] o [8].

vs

$$H_1 : \textit{no todas las } \mathbf{V}_g \textit{ son iguales}, \quad \mu_g \equiv \textit{cualquiera}. \tag{B.26}$$

De los datos muestrales se obtienen las matrices

$$\mathbf{T}_g = \sum_{\alpha=1}^{n_g} (x_{\alpha_g} - \overline{\mathbf{x}}_g)(x_{\alpha_g} - \overline{\mathbf{x}}_g)'$$
$$\mathbf{T} = \sum_{\alpha=1}^{G} T_g,$$

donde $\sum_{\alpha=1}^{G} n_g = N$, con $g = 1, \ldots, G$. Con estas matrices $\mathbf{T}_g$ y $\mathbf{T}$ estimamos $\mathbf{V}_G$ y $\mathbf{V}$, en el espacio de parámetros general y en el espacio de parámetros reducido por H_0, respectivamente. Entonces

$$\widehat{\mathbf{V}}_g = \frac{1}{n_g}\mathbf{T}_g \quad y \quad \widehat{\mathbf{V}} = \frac{1}{N}\mathbf{T}.$$

Si consideramos a $\upsilon_g = (n_g - 1)$ y $\upsilon = \sum_{g=1}^{G} \upsilon_g = (N - G)$, obtenemos los estimadores insesgados de $\mathbf{V}_g$ y $\mathbf{V}$, los cuales son $\mathbf{S}_g$ y $\mathbf{S}_p$, respectivamente; es decir

$$\mathbf{S}_g = \frac{1}{\upsilon_g}\mathbf{T}_g \quad y \quad \mathbf{S}_p = \frac{1}{\upsilon}\mathbf{T} = \frac{1}{\upsilon}\sum_{g=1}^{G} \upsilon_g \mathbf{S}_g. \tag{B.27}$$

En [2] se muestra que la razón de verosimilitud para verificar (B.26) es:

$$\lambda = \left[\frac{\prod_{g=1}^{G} |\mathbf{T}_g|^{\frac{n_g}{2}}}{|\mathbf{T}|^{\frac{N}{2}}}\right] \left[\frac{n^{\frac{p.N}{2}}}{\prod_{g=1}^{G} n_g^{\frac{p.n_g}{2}}}\right]. \tag{B.28}$$

Se rechaza H_0 para valores pequeños de λ a un nivel de significación α; es decir,se rechaza H_0 para valores $\lambda \leq \lambda(\alpha)$. De la misma forma en [2] se muestra que si introducimos la cantidad ρ dada por

$$\rho = 1 - \frac{2p^2 + 3p - 1}{6(p+1)(G-1)} \left(\sum_{g=1}^{G} \frac{1}{\upsilon_g} - \frac{1}{\upsilon}\right), \tag{B.29}$$

entonces el contraste se reduce a

$$\varphi = -2\rho \ln(\lambda_n), \tag{B.30}$$

que se distribuye asintóticamente como una χ^2 con $p(p+1)(G-1)/2$ grados de libertad (el subíndice n resalta la distribución asintótica).

Apéndice C

El método de las G-medias

C.1. Criterios para la agrupación de datos en G-medias

En los métodos de partición estudiados en el análisis de cluster se persigue minimizar la varianza dentro de los grupos para así detectar las diferencias entre ellos. Si tomamos como referencia el análisis de varianza visto aquí en la sección 2.3.1, decimos que para una muestra de tamaño n de una variable p-dimensional que puede estratificarse en G clases o grupos, la variación total de los datos esta dada por:

$$\mathbf{T} = \mathbf{B} + \mathbf{W},$$

que es la ecuación dada en (2.32), donde $\mathbf{T}$ es la variación total de los datos, $\mathbf{B}$ es la matriz de variabilidad explicada entre grupos y $\mathbf{W}$ la matriz variabilidad residual dentro de los grupos.
Si se asumen G grupos, entonces se tiene que $\mathbf{W} = \sum_{g=1}^{G} \mathbf{W}_g$, donde $\mathbf{W}_g$ mide la variabilidad dentro del grupo g, la cual llamaremos *matriz suma de cuadrados en el grupo g*.
En cualquier conjunto de datos $\mathbf{T}$ es fijo , luego el criterio para la formación de clusters recae sobre $\mathbf{B}$ ó $\mathbf{W}$. Algunos criterios para agrupar son los siguientes:

- **La traza de W**. Se trata de minimizar la traza de la matriz combinada de sumas de cuadrados y productos cruzados. Por la identidad (2.32), minimizar $tra(\mathbf{W})$ equivale a maximizar $tra(\mathbf{B})$, es decir:

$$\text{mín}\, tra(\mathbf{W}) \equiv \text{máx}\, tra(\mathbf{B}).$$

- **Determinante de W**. La minimización del determinate de $\mathbf{W}$ es un criterio para la partición de grupos. Minimizar $|\mathbf{W}|$ equivale a maximizar $|\mathbf{T}|/|\mathbf{W}|$.
- **Traza de $\mathbf{W}^{-1}\mathbf{B}$**. La maximización de $(\mathbf{W}^{-1}\mathbf{B})$ [1] puede ser expresada en términos de los valores propios $\lambda_1, \ldots, \lambda_p$ asociados con la matriz $\mathbf{W}^{-1}\mathbf{B}$, porque

$$tra(\mathbf{W}^{-1}\mathbf{B}) = \sum_{i=1}^{p} \lambda_i.$$

C.2. Fundamentos del método G-medias

El procedimiento de G-medias consiste en particionar un conjunto de n individuos en G grupos con el siguiente criterio: primero se escoge los *centroides* [2] de los grupos que minimicen la distancia de cada individuo a ellos, luego se asigna cada individuo al grupo cuyo centroide este mas cercano a dicho centroide. Se asume que entre los individuos se puede establecer una distancia euclídea.
El *algoritmo de G-medias* requiere cuatro etapas:

1. Seleccionar G puntos como centroides de los grupos iniciales. Esto puede hacerse:
   - Asignando aleatoriamente los objetos a los grupos y tomando los centroides de los grupos así formados.
   - Tomando como centroides los G puntos mas lejanos entre sí.
   - Construyendo unos grupos iniciales con información *a priori* y calculando sus centroides, o bien seleccionando sus centroides *a priori*.

[1] La matriz $\mathbf{W}^{-1}\mathbf{B}$ a sido llamada por Calyampudi R. Rao (1920) matriz de distancias de Mahalanobis generalizada, ya que su traza es la suma de las distancias de Mahalanobis entre la media del grupo y la media total. En efecto tenemos que:

$$tra(\mathbf{W}^{-1}\mathbf{B}) = \sum_{g=1}^{G} (\overline{\mathbf{x}}_g - \overline{\mathbf{x}}_T)' (\frac{\mathbf{W}}{n_g})^{-1} (\overline{\mathbf{x}}_g - \overline{\mathbf{x}}_T)$$

[2] Llamamos centroide de los datos al vector de promedios o de medias , formado por las p-medias muestrales; es decir

$$\overline{\mathbf{x}} = \frac{1}{n}\mathbf{1}'\mathbb{X} = (\overline{\mathbf{x}}_1, \ldots, \overline{\mathbf{x}}_p)'.$$

2. Calcular las distancias euclídeas de cada elemento a los centroides de los G grupos y asignar cada elemento al grupo cuyo centroide esté más próximo. La asignación se realiza secuencialmente y al introducir un nuevo elemento en un grupo se recalculan las coordenadas del nuevo centroide del grupo.

3. Definir un criterio de optimalidad para la agrupación de los datos y luego comprobar si reasignado alguno de los elementos mejora el criterio.

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

C.3. Implementación del método G-medias

El algoritmo G-medias utiliza como criterio de homogeneidad o de optimalidad el hacer minima *la suma de cuadrados dentro de los grupos* ($SCDG$) para todas las variables, la cual viene dada por:

$$SCDG = \sum_{g=1}^{G} \sum_{j=1}^{p} \sum_{i=1}^{n_g} (x_{ijg} - \overline{\mathrm{x}}_{jg})^2, \tag{C.1}$$

donde x_{ijg} es el valor de la variable j en el elemento i del grupo g y $\overline{\mathrm{x}}_{jg}$ la media de esta variable en el grupo. Hay que tener mucho cuidado de no confundir este valor con $\mathbf{W}$, la *matriz suma de cuadrados dentro de los grupos*, ya que en $SCDG$ operamos con los *escalares* de las variables en los grupos, mientras que en $\mathbf{W}$ operamos con los *elementos (vectores p-dimensionales)* de los grupos.
El criterio de minimizar ($SCDG$) es equivalente al criterio de minimizar *la suma ponderada de las varianzas de las variables de los grupos*, lo cual que puede escribirse así:

$$\begin{aligned}
\text{mín}\, SCDG &= \text{mín} \sum_{g=1}^{G} \sum_{j=1}^{p} \sum_{i=1}^{n_g} (x_{ijg} - \overline{\mathrm{x}}_{jg})^2 \\
&= \text{mín} \sum_{g=1}^{G} \sum_{j=1}^{p} \left[\sum_{i=1}^{n_g} (x_{ijg} - \overline{\mathrm{x}}_{jg})^2 \right] \\
&= \text{mín} \sum_{g=1}^{G} \sum_{j=1}^{p} \left[n_g \cdot \frac{1}{n_g} \sum_{i=1}^{n_g} (x_{ijg} - \overline{\mathrm{x}}_{jg})^2 \right] \\
&= \text{mín} \sum_{g=1}^{G} \sum_{j=1}^{p} n_g s_{jg}^2 \ ,
\end{aligned} \tag{C.2}$$

donde n_g es el numero de elementos en el grupo g y s_{jg}^2 es la varianza de la variable j en dicho grupo.
Las varianzas de las variables en los grupos son claramente una medida de heterogeneidad de la clasificación y al minimizarlas obtendremos grupos mas homogéneos. Un criterio alternativo de la homogeneidad seria minimizar las distancias al cuadrado entre los individuos del grupo y su centro. Si medimos las distancias con las normas euclidianas, este criterio se escribe:

$$\text{mín} \sum_{g=1}^{G} \sum_{i=1}^{n_g} (\mathbf{x}_{ig} - \overline{\mathbf{x}}_g)'(\mathbf{x}_{ig} - \overline{\mathbf{x}}_g) = \text{mín} \sum_{g=1}^{G} \sum_{i=1}^{n_g} d^2(i,g) \tag{C.3}$$

donde $d^2(i,g)$ es el cuadrado de la distancia euclídea entre el elemento i del grupo g y la media de su grupo. Probemos que el criterio de minimizar $d^2(i,g)$ es equivalente al criterio de minimizar $(SCDG)$.
Como un escalar es igual a su traza, podemos escribir la distancia al cuadrado entre los individuos del grupo y su centro así:

$$\begin{aligned} \text{mín} \sum_{g=1}^{G} \sum_{i=1}^{n_g} d^2(i,g) &= \text{mín} \sum_{g=1}^{G} \sum_{i=1}^{n_g} tra\big[d^2(i,g)\big] \\ &= \text{mín } tra\left[\sum_{g=1}^{G} \sum_{i=1}^{n_g} (\mathbf{x}_{ig} - \overline{\mathbf{x}}_g)'(\mathbf{x}_{ig} - \overline{\mathbf{x}}_g)\right]. \end{aligned}$$

Por propiedades de la traza se tiene que:

$$\text{mín } tra\left[\sum_{g=1}^{G} \sum_{i=1}^{n_g} (\mathbf{x}_{ig} - \overline{\mathbf{x}}_g)'(\mathbf{x}_{ig} - \overline{\mathbf{x}}_g)\right] = \text{mín } tra\left[\sum_{g=1}^{G} \sum_{i=1}^{n_g} (\mathbf{x}_{ig} - \overline{\mathbf{x}}_g)(\mathbf{x}_{ig} - \overline{\mathbf{x}}_g)'\right],$$

ya que los productos $(\mathbf{x}_{ig} - \overline{\mathbf{x}}_g)'(\mathbf{x}_{ig} - \overline{\mathbf{x}}_g)$, $(\mathbf{x}_{ig} - \overline{\mathbf{x}}_g)(\mathbf{x}_{ig} - \overline{\mathbf{x}}_g)'$ están bien definidos. Luego podemos decir que:

$$\text{mín} \sum_{g=1}^{G} \sum_{i=1}^{n_g} d^2(i,g) = \text{mín } tra\left[\sum_{g=1}^{G} \sum_{i=1}^{n_g} (\mathbf{x}_{ig} - \overline{\mathbf{x}}_g)(\mathbf{x}_{ig} - \overline{\mathbf{x}}_g)'\right] \tag{C.4}$$

$$= \text{mín } \sum_{g=1}^{G} \sum_{i=1}^{n_g} tra\big[(\mathbf{x}_{ig} - \overline{\mathbf{x}}_g)(\mathbf{x}_{ig} - \overline{\mathbf{x}}_g)'\big]. \tag{C.5}$$

Introduciendo el subíndice j para identificar las variables encontramos que la traza de la matriz $(\mathbf{x}_{ig} - \overline{\mathbf{x}}_g)(\mathbf{x}_{ig} - \overline{\mathbf{x}}_g)'$ es

$$tra\big[(\mathbf{x}_{ig} - \overline{\mathbf{x}}_g)(\mathbf{x}_{ig} - \overline{\mathbf{x}}_g)'\big] = \sum_{j=1}^{p} (x_{ijg} - \overline{\mathbf{x}}_{jg})^2$$

entonces:

$$\begin{aligned}\text{mín}\sum_{g=1}^{G}\sum_{i=1}^{n_g} d^2(i,g) &= \text{mín}\sum_{g=1}^{G}\sum_{j=1}^{p}\sum_{i=1}^{n_g}(x_{ijg}-\bar{\mathsf{x}}_{jg})^2 \\ &= \text{mín}\, SCDG\end{aligned}$$

que era precisamente lo que queríamos probar.
Observemos que en (C.4) tenemos es precisamente a $\mathbf{W}$, *la matriz de suma de cuadrados entre grupos*; por lo tanto minimizar la suma de cuadrados entre los grupos equivale a minimizar la traza de la matriz de suma de cuadrados entre grupos

$$\text{mín}\, SCDG \equiv \text{mín}\, tra\big(\mathbf{W}\big).$$

Este nuevo criterio se denomina el *criterio de la traza* de $\mathbf{W}$, ya mencionado anteriormente y el cual fue propuesto por Ward (1963).
En resumen, tenemos las siguientes equivalencias para los criterios de homogeneidad

$$\boxed{\text{mín}\, SCDG \equiv \text{mín}\sum_{g=1}^{G}\sum_{j=1}^{p} n_g s_{jg}^2 \equiv \text{mín}\, d^2(i,g) \equiv \text{mín}\, tra\big(\mathbf{W}\big).}$$

El algoritmo de G-medias busca la partición optima con la restricción de que en cada iteración solo se permite mover un elemento de un grupo a otro. Con el criterio de la traza el algoritmo funciona como sigue:

1. Partir de una asignación inicial.
2. Comparar si moviendo algún elemento se reduce $tra(\mathbf{W})$.
3. Si es posible reducir $tra(\mathbf{W})$ moviendo un elemento hacerlo, recalcular las medias de los grupos afectados por el cambio y volver a 2. Si no es posible reducir $tra(\mathbf{W})$, terminar.

En consecuencia, el resultado del algoritmo puede depender de la asignación inicial y del orden de los elementos. Conviene siempre repetir el algoritmo con distintos valores iniciales y permutando los elementos de la muestra. El efecto del orden suele ser pequeño, pero conviene asegurarse en cada caso.
Una propiedad importante del criterio de la traza es que al minimizar la distancia euclídea se producen grupos aproximadamente esféricos, lo cual se verá mas adelante.

C.4. Criterios para agrupar en una mezcla.

Para obtener criterios que se puedan aplicar en la conformación de clusters, tomando datos que provengan de la mezcla de G distribuciones normales, debemos retomar el problema planteado en el análisis discriminante en el cual se estudia la discriminación G poblaciones normales multivariantes $N(\mu_g, \mathbf{V}_g)$, cuando disponemos de una muestra de entrenamiento donde se conoce la procedencia de las observaciones; luego aquí no se tiene en cuenta en el proceso de clasificación los pesos π_g para cada grupo.
Sea n_g el numero de elementos de la muestra que proviene de la población g, donde $g = 1, \ldots, G$ y $\sum_{g=1}^{G} n_g = n$. Aplicando los resultados del **Apéndice B**, la sección B.1, y tomando la ecuación de soporte (2.15) con matriz de covarianzas $\mathbf{V}_g$ conocida, diremos que el soporte para el g-esimo grupo, eliminando las constantes, es:

$$L(\mu_g, \mathbf{V}_g) = -\frac{n_g}{2} \ \ln|\mathbf{V}_g| - \frac{n_g}{2} \ tra\left[\mathbf{V}_g^{-1}\mathbf{S}(\mu_g)\right]$$

donde $\mathbf{S}(\mu_g) = \frac{1}{n_g}\sum_{i=1}^{G}(x_i - \mu_g)(x_i - \mu_g)'$. Entonces el soporte para toda la muestra, sumando los soportes de cada población será:

$$\begin{aligned} L(\mu_1, \ldots, \mu_G; \mathbf{V}_1, \ldots, \mathbf{V}_G) &= \sum_{g=1}^{G} L(\mu_g, \mathbf{V}_g) \\ &= \sum_{g=1}^{G}\left\{-\frac{n_g}{2} \ \ln|\mathbf{V}_g| - \frac{n_g}{2} \ tra\left[\mathbf{V}_g^{-1}\mathbf{S}(\mu_g)\right]\right\} \\ &= -\frac{1}{2}\sum_{g=1}^{G} n_g \ \ln|\mathbf{V}_g| - \frac{1}{2}\sum_{g=1}^{G} n_g \ tra\left[\mathbf{V}_g^{-1}\mathbf{S}(\mu_g)\right] \end{aligned} \tag{C.6}$$

Puesto que la estimación de cada vector de medias (μ_g) es la media muestral ($\overline{\mathbf{x}}_g$), la función de soporte concentrada en estos parámetros será:

$$L(\mu_1, \ldots, \mu_G; \mathbf{V}_1, \ldots, \mathbf{V}_G) = -\frac{1}{2}\sum_{g=1}^{G} n_g \ \ln|\mathbf{V}_g| - \frac{1}{2}\sum_{g=1}^{G} n_g \ tra\left[\mathbf{V}_g^{-1}\mathbf{S}_g\right] \tag{C.7}$$

donde

$$\mathbf{S}_g = \frac{1}{n_g}\sum_{i=1}^{n_g}(x_i - \overline{\mathbf{x}}_g)(x_i - \overline{\mathbf{x}}_g)'. \tag{C.8}$$

Supongamos que admitimos la hipótesis $\mathbf{V}_g = \sigma^2\mathbf{I}$ (la cual mencionamos en los contrastes de esfericidad de la sección 2.3.2), esta nos indica que las variables están incorreladas y

tienen la misma varianza entre sí y en todos los grupos. Entonces la función de soporte se reduce a:

$$\begin{aligned} L(\mu_1,\ldots,\mu_G;\mathbf{V}_1,\ldots,\mathbf{V}_G) &= -\frac{1}{2}\left[\sum_{g=1}^{G} n_g\ \ln|\sigma^2\mathbf{I}|\right] - \frac{1}{2}\sum_{g=1}^{G} n_g\ tra\big[(\sigma^2\mathbf{I})^{-1}\mathbf{S}_g\big] \\ &= -\frac{1}{2}\left[\sum_{g=1}^{G} n_g\ \ln|\sigma^2| + \sum_{g=1}^{G} n_g\ \ln|\mathbf{I}|\right] - \frac{1}{2\sigma^2}\sum_{g=1}^{G} n_g\ tra\big[\mathbf{I}\cdot\mathbf{S}_g\big] \\ &= -\frac{1}{2}\Big[n\ln(\sigma^2) + n\ln(p)\Big] - \frac{1}{2\sigma^2}\sum_{g=1}^{G} n_g\ tra\big[\mathbf{S}_g\big] \\ &= -\frac{n}{2}\ln(\sigma^2 p) - \frac{1}{2\sigma^2}\ tra\left(\sum_{g=1}^{G} n_g\ \mathbf{S}_g\right), \end{aligned} \tag{C.9}$$

obsérvese que

$$\begin{aligned} \sum_{g=1}^{G} n_g\ \mathbf{S}_g &= \sum_{g=1}^{G} n_g\ \left[\frac{1}{n_g}\sum_{i=1}^{G}(x_i-\overline{\mathbf{x}}_g)(x_i-\overline{\mathbf{x}}_g)'\right] \\ &= \sum_{g=1}^{G}\sum_{i=1}^{n_g}(x_i-\overline{\mathbf{x}}_g)(x_i-\overline{\mathbf{x}}_g)' \\ &= \mathbf{W}. \end{aligned}$$

Luego maximizar la verosimilitud, para nosotros maximizar el soporte dado en la ecuación (C.9), equivale a:

$$\boxed{\text{mín}\ \ tra(\mathbf{W})}$$

que es el *criterio de la traza*. Nótese además que en la ecuación dada en (C.9) el criterio de la traza consiste en minimizar la suma ponderada de las varianzas estimadas de cada grupo. Este criterio se obtuvo por otros métodos en el algoritmo de las G-medias y tiene la ventaja de ser simple y fácil de calcular pero no es invariante ante transformaciones lineales y no tiene en cuenta las correlaciones.

Si admitimos la hipótesis $\mathbf{V}_g = \mathbf{V}$, la verosimilitud es equivalente a la del problema de discriminación clásica y viene dado por:

$$\begin{aligned} L(\mu_1,\ldots,\mu_G;\mathbf{V}_1,\ldots,\mathbf{V}_G) &= -\frac{1}{2}\sum_{g=1}^{G} n_g\ \ln|\mathbf{V}| - \frac{1}{2}\sum_{g=1}^{G} n_g\ tra\left[\mathbf{V}^{-1}\mathbf{S}_g\right] \\ &= -\frac{1}{2}\sum_{g=1}^{G} n_g\ \ln|\mathbf{V}| - \frac{1}{2}\sum_{g=1}^{G}\ tra\left[n_g\mathbf{V}^{-1}\mathbf{S}_g\right] \end{aligned}$$

$$= -\frac{1}{2}\sum_{g=1}^{G} n_g \ \ln|\mathbf{V}| - \frac{1}{2}\ tra\left(\sum_{g=1}^{G} n_g \mathbf{V}^{-1}\mathbf{S}_g\right)$$
$$= -\frac{n}{2}\ln|\mathbf{V}| - \frac{1}{2}\ tra\left(\mathbf{V}^{-1}\sum_{g=1}^{G} n_g\ \mathbf{S}_g\right), \qquad \text{(C.10)}$$

además, puesto que la estimación MV de $\mathbf{V}$ es:

$$\widehat{\mathbf{V}} = \frac{1}{n}\sum_{g=1}^{G} n_g \mathbf{S}_g = \frac{1}{n}\mathbf{W} \equiv \mathbf{S}_W,$$

la función de soporte es ahora,

$$L(\mu_1, \ldots, \mu_G; \mathbf{V}_1, \ldots, \mathbf{V}_G) = -\frac{n}{2}\ln|\mathbf{V}| - \frac{1}{2}\ tra\left(\mathbf{V}^{-1}\cdot\frac{1}{n}\mathbf{W}\right)$$
$$= \frac{n}{2}\ln|\mathbf{V}^{-1}| - \frac{1}{2}\ tra\left(\frac{\mathbf{W}}{n}\cdot\mathbf{V}^{-1}\right). \qquad \text{(C.11)}$$

Maximizar la verosimilitud equivale a

$$\boxed{\text{mín}\,|\mathbf{W}|}$$

que es el *criterio del determinante* propuesto por Friedman y Rubin en 1967, en el articulo publicado en *Journal of the American Statistical Association, 62, 1159-1178* y titulado: *On some Invariant Criteria for Grouping Data*. Este criterio es invariante a transformaciones lineales y tiende a identificar grupos elípticos.
En el caso general que las poblaciones tiene distinta matriz de varianzas y covarianzas, la estimación máximo verosímil de $\mathbf{V}_g$ es $\mathbf{S}_g$ y el máximo de la función de verosimilitud es:

$$L(\mu_1, \ldots, \mu_G; \mathbf{V}_1, \ldots, \mathbf{V}_G) = -\frac{1}{2}\sum_{g=1}^{G} n_g \ \ln|\mathbf{S}_g| - \frac{np}{2}, \qquad \text{(C.12)}$$

y maximizar la verosimilitud equivale a:

$$\boxed{\text{mín}\sum_{g=1}^{G} n_g \ \ln|\mathbf{S}_g|.} \qquad \text{(C.13)}$$

En otros términos cada grupo debe tener *volumen* mínimo. Suponemos que cada grupo tiene $n_g > p+1$, de manera que $\mathbf{S}_g$ sea no singular, lo que exige que $n > G(p+1)$.

Un criterio adicional propuesto por Friedman y Rubin (1967) es partir de la descomposición del análisis de la varianza multivariante y maximizar el tamaño de la *distancia de Mahalanobis generalizada* dada por $\mathbf{W}^{-1}\mathbf{B}$. De nuevo el tamaño de esta matriz puede medirse por la traza o el determinante, pero este criterio no ha dado buenos resultados en el análisis de clusters.
Cualquiera de estos criterios puede implementarse en un algoritmo similar al G-medias dado en la sección C.2. El criterio del determinante es fácil de identificar y como se puede se comprueba en [1], apéndice 15.1, tiende a producir grupos elípticos mientras que la traza grupos esféricos.
El criterio (C.13) tiene el inconveniente de que es necesario imponer restricciones fuertes sobre el numero de observaciones en cada grupo para que las matrices no sean singulares, ya que si el numero de grupos es grande, el número de parámetros a estimar es alto. En la practica, parece mejor permitir algunos rasgos comunes en las matrices de covarianzas y esto no es fácil de imponer con este criterio. Además si un grupo tiene pocas observaciones y $\mathbf{S}_g$ es casi singular, este grupo tendrá un peso desproporcionado en el criterio y el algoritmo tiende a caer en este tipo de soluciones. Por esta razón, este criterio aunque de interés teórico no se utiliza en la practica.

C.5. Determinación del numero de grupos.

Generalmente el numero de grupos es desconocido y se estima con los datos aplicando el algoritmo para distintos valores de G para luego seleccionar el resultado mas adecuado. Comparar la soluciones obtenidas no es simple, porque cualquiera de los criterios disminuirá si aumentamos el numero de grupos. En efecto según el análisis de varianza multivariante, la variabilidad total puede descomponerse como :

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

Intuitivamente, el objetivo de la division de los grupos es conseguir que $\mathbf{B}$, la variabilidad entre los grupos, sea lo mas grande posible, mientras que $\mathbf{W}$, la variabilidad dentro el grupo, sea lo mas pequeña posible. Dada una división cualquiera en grupos, si elegimos uno cualquiera de ellos podemos aplicarle de nuevo esta descomposición, con lo que reduciremos de nuevo la variabilidad descomponiendo más este grupo. Por lo tanto no podemos utilizar ningún criterio basado en el tamaño de $\mathbf{W}$ para compara soluciones con grupos distintos, ya que siempre podemos disminuir $\mathbf{W}$ haciendo más grupos.

Para seleccionar el numero de grupos en el análisis de conglomerados estudiado en [1] se realiza un *test* F de reducción de variabilidad, el cual compara la disminución de variabilidad al aumentar un grupo con la varianza promedio. Para mezclas se realiza un test F aproximado (*que llamaremos* H), en el cual calculamos la reducción proporcional de variabilidad que se obtiene aumentando un grupo adicional. El test es :

$$H = \frac{tra(\mathbf{W}_G) - tra(\mathbf{W}_{G+1})}{tra(\mathbf{W}_{G+1})/(n-G-1)}, \tag{C.14}$$

y bajo el supuesto que G grupos sean suficientes, el valor H puede compararse con una *distribución* f con p y $p(n-G-1)$ grados de libertad. Una regla empírica que da resultados razonables, sugerida por Hartigan en 1975 en el articulo: *A k-means Clustering Algorithm* publicado en *Applied statistics ,28, 100-108*, e implantada por algunos programas informáticos, es introducir un grupo más si este cociente H es mayor que 10.
Un criterio adicional que suele funcionar mejor que el anterior, es el propuesto por Calinski y Harabasz (1974). Este criterio parte de la descomposición de la variabilidad total dada en (2.32) y selecciona el valor de G maximizando:

$$CH = \text{máx}\left\{\frac{tra(\mathbf{B})/(G-1)}{tra(\mathbf{W})/(n-G)}\right\}. \tag{C.15}$$

Además de estos dos criterios existen otras alternativas, como utilizar *los criterios de selección de modelos* estudiados aquí en la sección 2.5.

C.6. Resumen:

- Los criterios de optimidad mas implementados en el algoritmo de G-medias para mezclas de distribuciones normales p-variantes $N_p(\mu_g, \mathbf{V}_g)$, son:

  1. El criterio de la mínima traza, mín $tra(\mathbf{W})$, que se obtiene al maximizar la función de soporte dada en la ecuación (C.6), tomando $\widehat{\mu}_g = \overline{\mathbf{x}}_g$ y $\widehat{\mathbf{V}}_g = \sigma^2\mathbf{I}$.
  2. El criterio del mínimo determinante, mín $|\mathbf{W}|$, que se obtiene al maximizar la función de soporte dada en la ecuación (C.6), tomando $\widehat{\mu}_g = \overline{\mathbf{x}}_g$ y $\widehat{\mathbf{V}}_g = \mathbf{V}$.

- Una vez establecido el criterio de optimidad, el algoritmo para las mezclas consiste en:

  1. Partir de una asignación inicial de G grupos con distribuciones normales p-variantes. Calcular $\overline{\mathbf{x}}_g$, la matriz $\mathbf{W} = \sum_{g=1}^{G} \mathbf{W}_g$ (con $g = 1, \ldots, G$) y el valor del criterio a evaluar.

2. Comparar si moviendo algún elemento se reduce el criterio, $tra(\mathbf{W})$ o $|\mathbf{W}|$. Para esto se deben calcular la distancia euclídea de cada elemento a cada media grupal $\overline{\mathbf{x}}_g$, con $g = 1, \ldots, G$. Luego se asigna de forma secuencial cada elemento a la media grupal mas cercana; si en esta nueva asignación se introduce un nuevo elemento al grupo, se recalcula $\overline{\mathbf{x}}_g$, $\mathbf{W}_g$ y por consiguiente $\mathbf{W}$.
3. Si es posible reducir el criterio moviendo un elemento de un grupo al otro, hacerlo. luego calculase las medias de os grupos afectados y el valor del criterio implementado. Vuelva a 2) y repita el proceso.
4. Si no es posible reducir el criterio terminar.

- La determinación del numero de grupos se realizan según los test establecidos en las ecuaciones (C.14) y (C.15).

Apéndice D

Datos mezclados

A continuación se dan los datos hallados para el ejemplo de la mezcla de 5 distribuciones normales multivariantes con dimension $p = 4$.

Datos	X1	X2	X3	X4
001	0.5134	-0.5875	1.5392	3.7239
002	-0.8019	0.2476	0.7079	4.7811
003	-1.4767	0.3245	3.9199	6.3396
004	0.9859	-1.7549	-0.2267	7.1926
005	0.6845	0.9900	0.6074	5.6068
006	-0.2740	0.6037	-2.7815	4.2069
007	-0.5347	2.0574	-0.7722	5.2536
008	-1.2535	-0.7215	-2.6706	3.9089
009	-2.6024	-0.0094	-0.1527	3.6658
010	-2.3026	-0.7432	-1.8613	3.0094
011	-1.6600	1.0150	-0.9512	2.3043
012	-0.1131	-3.0099	3.6849	-2.4332
013	1.1642	-0.7786	3.9691	-1.7850
014	0.3571	-2.6811	6.4071	-1.1144
015	-0.0141	-0.1659	5.1204	0.2958
016	0.0347	1.5722	1.8240	2.3293
017	-0.9673	4.1043	8.1651	3.3537
018	3.2668	0.9003	6.1329	1.2840
019	-0.3810	0.7111	2.3244	1.4064

020	-2.4498	1.8231	4.3503	-0.2474
021	-0.5245	0.7427	5.3597	-0.1730
022	-0.9865	2.2432	3.6739	-1.3859
023	-0.2104	-0.3223	4.8158	0.8543
024	-0.6758	0.2264	5.5868	2.1435
025	-1.4623	-1.1390	6.7900	2.2447
026	-0.6221	1.3491	-2.0440	1.3462
027	0.4855	-0.1306	-4.3634	0.0399
028	1.4104	-2.4791	-0.7098	-3.6010
029	-1.2266	4.7946	1.3156	2.7699
030	0.5135	-3.0250	1.8537	1.5285
031	0.7945	1.0219	-3.7451	-1.7457
032	1.1656	0.8747	-4.2984	1.1873
033	-1.3774	4.6354	-0.9764	4.7729
034	3.0093	-0.7230	0.1006	0.7924
035	2.1800	3.2828	5.4037	0.5181
036	-5.3172	-1.1234	-0.6929	2.8800
037	-2.4425	0.4687	0.4194	-0.0302
038	-0.7283	1.6878	2.1191	-2.9814
039	-4.4995	-2.1202	-1.1755	-1.2440
040	3.5054	1.3711	-2.4946	1.6896
041	0.5218	3.9366	-3.5605	-2.0390
042	1.4454	0.1985	2.6270	3.0672
043	-2.5254	5.8039	1.0848	0.5491
044	0.4406	-1.5104	3.6794	0.2657
045	3.7943	6.2315	-1.0157	0.8602
046	1.6235	-0.0376	2.2557	-0.1569
047	1.7721	0.6074	4.5826	-1.2931
048	0.4693	4.4374	1.7026	0.0596
049	1.2997	5.8781	-0.4690	0.1491
050	1.6348	4.1854	0.0946	1.5959
051	-0.7028	4.7416	0.2871	-0.7773
052	0.8073	5.4933	0.9194	1.5503
053	-1.0275	4.1973	0.5101	1.0550
054	1.2945	4.9917	0.2454	-0.1667
055	0.0149	5.6276	-1.4005	0.3145

056	0.2187	5.1544	0.9696	1.4196
057	1.7132	7.5807	1.5937	0.3273
058	-2.0788	3.6938	-1.4379	0.4757
059	0.1129	6.0235	-1.5342	0.3988
060	-1.0865	5.7778	-0.0747	-0.0728
061	-1.5583	4.1661	0.0815	1.3148
062	0.6374	4.4133	-0.8432	0.9783
063	-0.4046	5.0657	-0.5646	1.7221
064	-0.4033	4.9877	-0.0282	-0.4123
065	0.0841	4.9230	-1.2437	0.5651
066	-0.4353	3.4414	0.7330	0.7399
067	5.1842	0.2511	0.7818	0.3146
068	6.0984	-0.4886	1.0279	0.7612
069	5.5264	-2.5717	-0.2087	0.1291
070	4.0730	1.2977	-0.5920	-0.1692
071	4.6260	-0.3408	-0.4968	1.2455
072	4.3926	1.1948	-0.2195	-0.5201
073	5.2962	-1.1322	1.0398	0.6773
074	4.5760	0.7563	-1.2959	1.6137
075	3.2399	0.4532	-0.3236	0.3314
076	4.4439	-0.3891	0.2302	-0.7207
077	6.2132	2.0334	0.6912	2.0347
078	4.7241	1.6905	-0.8193	-0.7032
079	7.2605	0.6671	-0.0873	0.2346
080	3.6321	0.0259	0.1070	0.6864
081	4.5513	0.4524	0.0523	1.0273
082	5.4578	0.5722	0.3109	-0.0532
083	6.2488	-0.4937	-0.0870	0.5399
084	4.6191	-0.2184	-0.5830	-1.4820
085	4.5847	1.2693	-0.3236	0.0499
086	6.0335	0.8427	0.0135	-0.6360
087	5.0342	0.2539	1.1456	-1.4147
088	5.6262	-0.6836	0.3488	0.9234
089	6.0297	-0.4110	0.0575	1.2236
090	5.2551	0.7258	0.2461	0.1976
091	5.9017	0.3019	0.3954	-0.9184

092	5.6402	-0.0672	1.4915	2.0207
093	3.8960	0.6269	0.2180	-0.3364
094	4.5740	-1.4993	1.2640	0.7648
095	5.4644	1.0151	0.2652	-0.1138
096	3.3621	-0.0506	0.6722	1.0995
097	4.3638	-0.3284	0.4860	0.2697
098	2.9553	0.5100	1.4881	-0.3987
099	4.4483	0.5385	-0.7848	0.0637
100	4.9040	0.8530	-0.7669	-0.0880

Bibliografía

[1] PEÑA, Daniel. *Análisis de Datos Multivariantes*. Primera edición en español. Editorial McGrawHill, España, 2002.

[2] DIAZ, Luis G. *Estadística Multivariada: Inferencia y metodos*. Primera edición. Facultad de ciencias. Universidad Nacional de Colombia, Bogota, 2002.

[3] CASTRILLÓN, Oscar Y. *Análisis de Cluster y Análisis Discriminante*, Monografia. Universidad industrial de santander, 2002.

[4] CUADRAS, Carles M. *Análisis Multivariante*. Primera edición, version PDF. Universidad de Barcelona, España, 2004.

[5] DALLAS, Johnson. *Métodos Multivariados Aplicados al Análisis de Datos* . Primera edición. International Thomson editores, México, 2000.

[6] CIFUENTES A, Julio César. *A New Accelerator for the EM Algorithm*. Version preliminar en PDF. Universidad Icesi. Colombia.

[7] MARTINEZ Wendy L. y MARTINEZ Angel R. *Model-Based Clustering Toolbox for MATLAB*. Documento en PDF. Oficina de investigaciones del Centro de Guerra Naval de los Estados unidos, 2004.

[8] FREUND, Jhon E. *Estadistica Matematica con Aplicaciones*. Sexta edición. Editorial Prentice Hall, Mexico, 2000.

[9] FRALEY, Chris. y RAFTERY Adrian E. *MCLUST: Software for model-Based cluster Analysis*. Journal of American Statistical Association, 16, 297-307.

[10] FRALEY, Chris. y RAFTERY Adrian E. *Model-Based clustering, Discriminant Analysis, and Density Estimation*. Journal of American Statistical Association: Jun 2002;97,485; ABI/INFORM Global pg 611-631.

[11] MCLACHLAN G.J, PEEL D, BASFORD K.E y ADAMS P. *The EMMIX Software for the Fitting of Mixtures of Normal and t-Components*. Documento en PDF. Universidad de Queensland, St. Lucia, AUSTRALIA, 1999.

[12] CASELLA G. y BERGER L. *Statistical Inference*. Editorial Thomson.

[13] BOX, George P. *Estadistica para inverstigadores*. Editorial Reverte S.A.

[14] CHATFIELD, Chris. *Problem Solving A Statistician's Guide*. Editorial Chapman & Hall. 1995.

[15] COOLEY, William W. y LOHNES Paul R. *Multivariate Data Analysis*. Editorial Wiley & Sonc Inc. 1971.