UNIVERSIDAD NACIONAL AUTÓNOMA
DE MÉXICO

FACULTAD DE CIENCIAS

Un Enfoque Bayesiano al Análisis de Grupos en Datos Bivariados

T E S I S

QUE PARA OBTENER EL TÍTULO DE:

Actuario

PRESENTA:

Jesús Alberto Galis García

DIRECTOR DE TESIS

Dra. Ruth Selene Fuentes García

Ciudad Universitaria, CD. MX., 2018

Hoja de Datos del Jurado

1. Datos del alumno

Apellido paterno	Galis
Apellido materno	García
Nombre(s)	Jesús Alberto
Teléfono	55 29 61 22 38
Universidad Nacional Autónoma de México	
Facultad de Ciencias	
Carrera	Actuaría
Número de cuenta	411004024

2. Datos del tutor

Grado	Dra
Nombre(s)	Ruth Selene
Apellido paterno	Fuentes
Apellido materno	García

3. Datos del sinodal 1

Grado	Dr
Nombre(s)	Carlos
Apellido paterno	Díaz
Apellido materno	Ávalos

4. Datos del sinodal 2

Grado	Dr
Nombre(s)	Ricardo
Apellido paterno	Ramírez
Apellido materno	Aldana

5. Datos del sinodal 3

Grado	Mat
Nombre(s)	Margarita Elvira
Apellido paterno	Chávez
Apellido materno	Cano

6. Datos del sinodal 4

Grado	Act
Nombre(s)	Edna Gabriela
Apellido paterno	López
Apellido materno	Estrada

7. Datos del trabajo escrito

Título	Un Enfoque Bayesiano al Análisis de Grupos en Datos Bivariados
Número de Páginas	82
Año	2018

Agradecimientos

A mis padres Alberto y Doli por todo el apoyo, la paciencia y el impulso para conseguir mis metas.

A mi hermana Ana por enseñarme de honestidad y de pasión por hacer las cosas.

A mi novia Agélica por estar siempre ahí, por su cariño, la motivación y todas las otras cosas por agradecer y que no terminaría de escribir.

A mis abuelos por su ejemplo, sus enseñanzas y su amor incondicional, a pesar de la edad y a pesar de las distancias.

A mis amigos por compartir los momentos y las experiencias, los buenos y no tan buenos ratos.

A mi tutora Ruth por la supervisión y por lograr que se terminara este trabajo.

Finalmente a mi universidad la UNAM, por las oportunidades, las vivencias, las enseñanzas y por hacerlas posibles.

Índice general

Prólogo

En los últimos años los campos de la estadística y el análisis de datos han enfrentado retos cada vez más complejos derivados de la necesidad de problemas encontrados en áreas científicas e industriales. La basta cantidad de datos acumulados en todo el mundo y la creciente necesidad de análisis cada vez más precisos han propiciado el desarrollo de nuevas técnicas estadísticas para afrontar problemas cada vez más diversos.

En este documento se hace referencia al problema de encontrar grupos en datos no etiquetados, más aún, cuando no se sabe tampoco el número de grupos en el que se desean agrupar los datos, por ejemplo, si se tienen un conjunto de puntos en un plano que representan a un conjunto de datos bivariados, se busca "colorearlos" o "etiquetarlos" de alguna manera que todos los elementos que comparten el color o la etiqueta, compartan también rasgos comunes, en este caso al tratarse de puntos en un plano, se busca que tengan localizaciones similares en el mismo.

Para tratar el problema se hace referencia a la solución propuesta por Fuentes y Walker en [11] para casos unidimensionales y se trata de extender el modelo a datos de dimensiones superiores, generando, a partir de variables latentes, elipsoides en lugar de intervalos.

En el capítulo 1 se describe el problema con mayor detalle, se dan ejemplos y se describen métodos existentes para este problema, asimismo se aplican a conjuntos de datos a fin de ilustrar su funcionamiento, ventajas y desventajas.

El capítulo 2 presenta información previa, herramientas matemáticas y estadísticas de las que se hace uso en el resto del documento para dar el sustento matemático al enfoque presentado en el capítulo 3, en donde se desarrolla la teoría del modelo en el caso univariado, se presentan los supuestos para el caso multivariado y se definen los procesos del modelo. El desarrollo teórico que acompaña al modelo, las distribuciones que apoyan la convergencia, la selección de parámetros y las reglas de clasificación se encuentran en esta sección.

En el capítulo 4 se define el algoritmo formalmente, se enuncian los problemas con la implementación de dicho algoritmo y se propone un proceso semejante. Se dan supuestos adicionales a fin de facilitar la programación en un software estadístico conocido y así visualizar los resultados en conjuntos de datos bivariados estudiados apliamente, así como los datos que se analizaron con los métodos expuestos en el capítulo 1. A grandes rasgos, se toma lo que se propone en el capítulo previo y se busca llevarlo lo más cercano posible a un lenguaje de programación, a fin de que una computadora realice las iteraciones necesarias para obtener resultados.

más adelante se comparan los resultados con los de métodos existentes antes descritos y se dan las conclusiones a las que se llega tras el análisis de los resultados.

Los análisis y algunos ejemplos numéricos se implementaron en la herramienta R, un software estadístico libre para manejo de datos disponible en la siguiente dirección URL: `https://www.r-project.org/`.

Capítulo 1

Descripción del Problema

1.1. Análisis de Conglomerados: una revisión a la literatura

El análisis de conglomerados o *cluster analysis* es el problema estadístico que consiste en dividir muestras en grupos de acuerdo a sus características. Éste es considerado también como el arte de encontrar grupos en los datos. En el contexto de reconocimiento de patrones se refiere al Análisis de Conglomerados como aprendizaje no supervisado o *unsupervised learning* contrario al aprendizaje supervisado, que comprende las técnicas del análisis discriminante en los que se cuenta con un *training set* o muestra de entrenamiento, es decir, las observaciones de la muestra incluyen una variable o etiqueta que indica el grupo al cual pertenece dicho dato, cosa que no sucede en los problemas de aprendizaje no supervisado.

Esta distinción entre el Análisis de Conglomerados y el análisis discriminante se puede traducir a que el primero no involucra observaciones conocidas que describan los grupos y no necesariamente información previa del número de grupos. Esta diferencia en ocasiones es poco clara cuando en los datos obtenidos se tiene solo una pequeña porción de la información etiquetada y hay cabida para grupos en la parte de la población que no pertenece a la porción etiquetada del conjunto muestral.

El Análisis de Conglomerados está fuertemente ligado a los métodos de representación gráfica de la información. El objetivo del Análisis de Conglomerados es la división de datos en un número determinado de grupos, y los métodos gráficos hacen valiosas aportaciones a cualquier propuesta de partición. Se dice que es, generalmente, imposible e inapropiado hacer inferencia estadística formal sobre el comportamiento de los grupos, por lo que casi siempre es necesaria alguna ilustración que muestre cuán cercanos están los datos, o las similitudes entre datos.

Existe una distinción entre la idea de encontrar grupos "naturales" y la de dividir el conjunto muestral en un número κ de grupos por conveniencia. En [15] se le llamó a este último problema "disección" y a pesar de que las técnicas en ambas variantes están fuertemente relacionadas, nosotros nos enfocaremos en el primero. Esto es, el que no busca dividir la muestra en una cantidad κ de grupos dada, sino que se busca la partición más natural de los datos cuando no se tiene una cantidad de grupos previa. A pesar de centrarnos en ese problema, también es posible aplicar las técnicas desarrolladas en este documento a problemas de disección.

Krzanowski y Marriot dividen en [16] a la gran mayoría de los métodos propuestos para resolver el problema en cuatro grupos:

Los métodos jerárquicos que dependen de la medida de similitud o disimilitud de los datos, los métodos con base en una partición óptima y también es común ajustar modelos de mezclas de distribuciones, en donde se supone una distribución previa y se supone que los datos provienen de variables aleatorias con esas mezclas por distribución.

Finalmente están los métodos no paramétricos, que debido al avance computacional iniciado

en los años 80's, han tenido un auge en los modelos. También deben su popularidad a que en ocasiones las suposiciones acerca de la distribución parecen ambiciosas y los datos se no ajustan a las distribuciones supuestas. Estos métodos no serán profundizados en este documento por no figurar ni dentro de los objetivos, ni dentro del marco conceptual necesario, y nos limitamos a esta breve mención.

1.2. Métodos jerárquicos

Los métodos jerárquicos pueden ser clasificados en dos grupos, según la dirección que toma la jerarquía con respecto al número de grupos. Al primer método donde se parte de un único y gran grupo, y a partir de éste se van desprendiendo más conforme la jerarquía se incrementa se le denomina divisivo, mientras que al segundo método donde se parte de un número grande de grupos para ir refinando la partición conforme la jerarquía sube es llamado método aglomerativo.

Métodos jerárquicos divisivos: son aquellos que inician con un solo grupo de la totalidad del conjunto de entrenamiento y en cada paso se hacen divisiones entre los grupos mientras se baja la jerarquía.

Estos métodos son computacionalmente más costosos que los métodos aglomerativos, por lo que su área de aplicación e implementación es mucho más limitada que los anteriores.

Un ejemplo de estos métodos se encuentra en la función *diana()* en la paquetería *cluster* para R, que implementa el algoritmo descrito en el capítulo 6 de Kaufman [14]. Éste a su vez es de lo poco, en comparación con los otros casos, que existe implementado para lo métodos divisivos dado que la mayoría de las aplicaciones en software están dirigidas a los métodos aglomerativos.

Métodos jerárquicos aglomerativos: son aquellos tratamientos que se hacen a una muestra no supervisada para conglomerar partiendo de una cantidad de grupos grande, cantidad igual al tamaño de la muestra, y a partir de ahí se reduce el número de grupos según algún criterio.

Los métodos jerárquicos se basan en una matriz llamada matriz de distancias, semejanzas o similaridades. Esta matriz tiene en cada una de sus entradas i, j la distancia entre el i-ésimo y el j-ésimo elemento de la muestra, la matriz tiene dimensión $n \times n$ donde n es el tamaño de la muestra y la distancia es alguna previamente definida. Una de las más usadas es la distancia euclidiana, o en su forma más general la distancia Minkowsky.

La principal herramienta de visualización de los métodos jerárquicos es el dendrograma, que ha sido una de las principales herramientas en el Análisis de Conglomerados desde los años 50s y que se refiere a un diagrama de árbol que en el fondo muestra los nodos de cada uno de los datos, mientras que conforme se avanza se pierde detalle hasta llegar a un único grupo que contiene a todos los datos.

El dendrograma se apoya en una función de disimilitud, que mide qué tan distintos son los puntos o las observaciones, por ejemplo, si cada observación se representa por un vector en $\mathbb{R}^n$ una función de disimilitud entre dos puntos que es ampliamente utilizada es la distancia euclidiana, aunque existen diversas alternativas.

Para ilustrar los métodos jerárquicos se considera la base `mtcars` [13] de Henderson y Velleman y el código en R que se muestra en el Apéndice, en el que primero se calculan las disimilitudes entre observaciones con la función `dist` que por default calcula las distancias euclidianas, luego genera el modelo de dendrograma para posteriormente concentrar los datos, en este caso, en tres grupos. Finalmente se grafica el dendrograma y se señalan dichos grupos, ver figura 1.1.

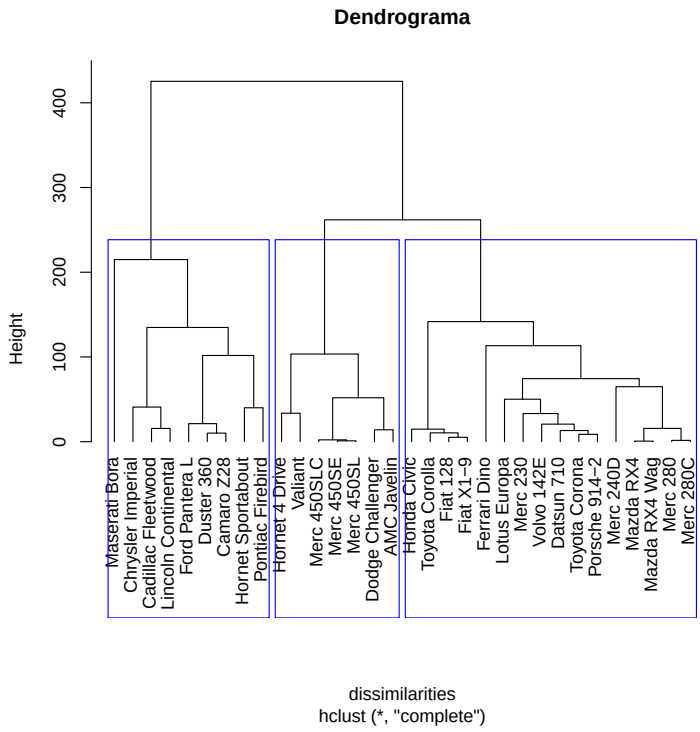


Figura 1.1: Ilustración del dendrograma con una partición de 3 grupos.

Para hacerlo más claro, considérense los datos que corresponden únicamente a las variables Desplazamiento y Cociente del eje trasero para tener una muestra de dos variables.

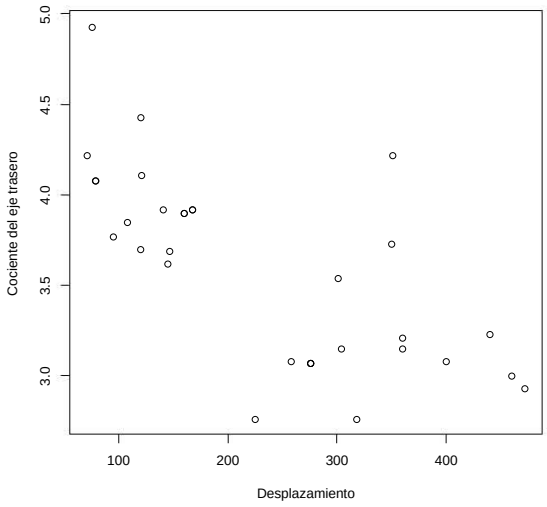
Al aplicar el mismo proceso que al conjunto de datos completo se obtiene un dendrograma, éste separara las observaciones en tres grupos. Los datos, el dendrograma y los grupos se muestran en la figura 1.2

1.3. Métodos de partición

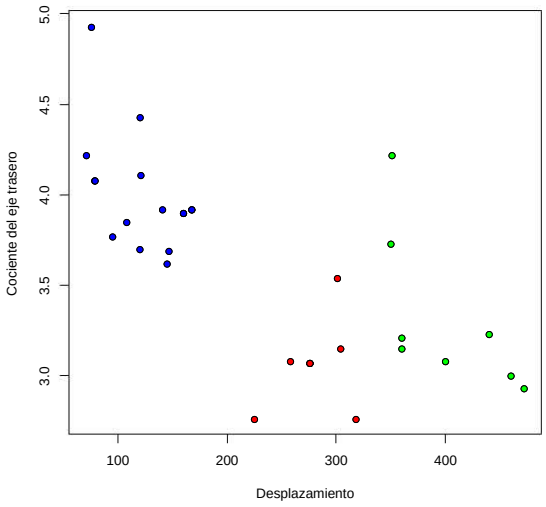
Los métodos de partición hacen uso de algún criterio de agrupamiento, y a partir de éste se busca la partición que optimice dicho criterio o que no se pueda mejorar cambiando parámetros del modelo.

Un ejemplo es el método κ-centroides, que dada una cantidad de grupos κ, supone $\{\mu_1, \mu_2, ..., \mu_\kappa\}$ medias iniciales, calcula distancias con respecto a los centros y agrupa las observaciones relacionando cada una con la media μ_j que minimiza la distancia a la observación. Así se forman κ conglomerados y posteriormente las actualiza de manera iterativa hasta satisfacer el criterio previamente definido.

Para ilustrar estos métodos se implementó un código en R con el mismo conjunto de datos bivariados obtenido de `mtcars`, se utilizó la función `kmeans` que ejecuta el método y cuyos parámetros son un conjunto de datos, los centros iniciales de los conjuntos o en su defecto la cantidad de grupos que se tienen y un número máximo de iteraciones para lograr la convergencia. El código a detalle se puede ver en la parte correspondiente del Apéndice. Los resultados del método tras una, dos y tres iteraciones se muestran en las gráficas de la figura 1.3, donde



Gráfica de los datos considerando dos variables



Gráfica de los tres grupos generados por el segundo dendrograma.

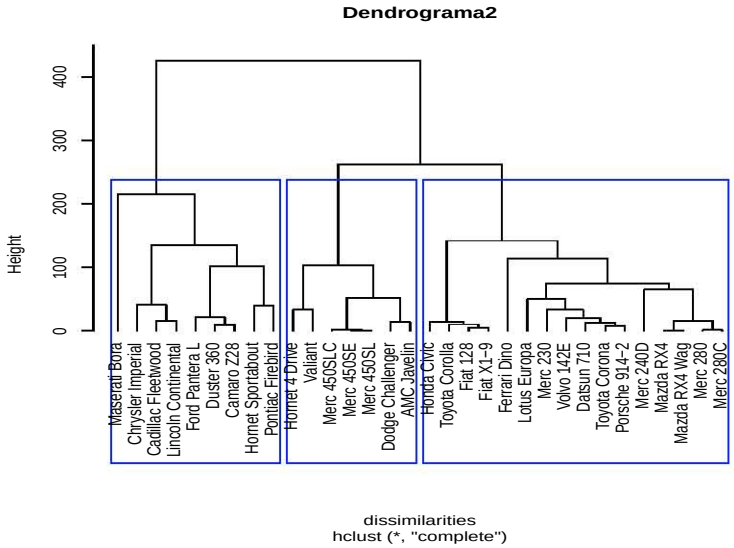
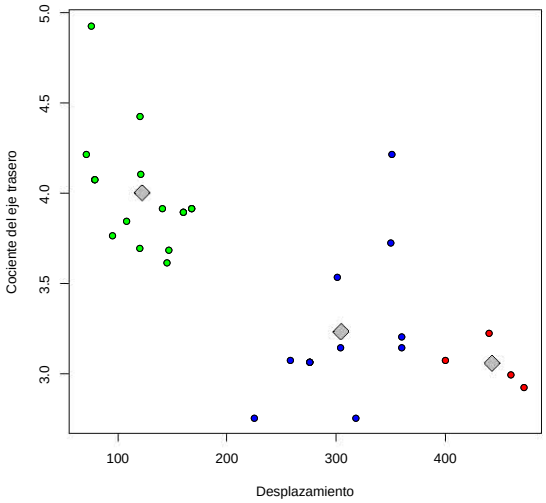


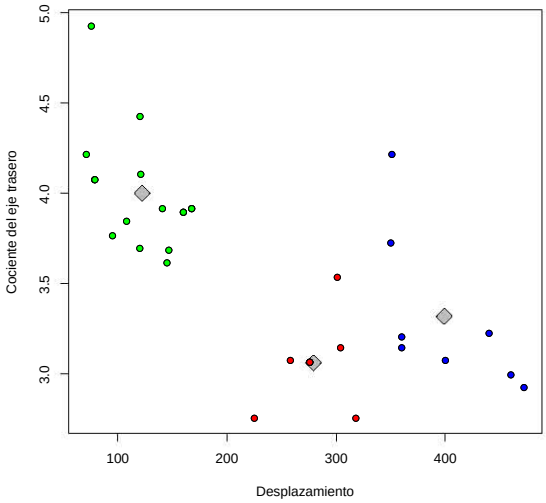
Ilustración del segundo dendrograma con una partición de 3 grupos.

Figura 1.2

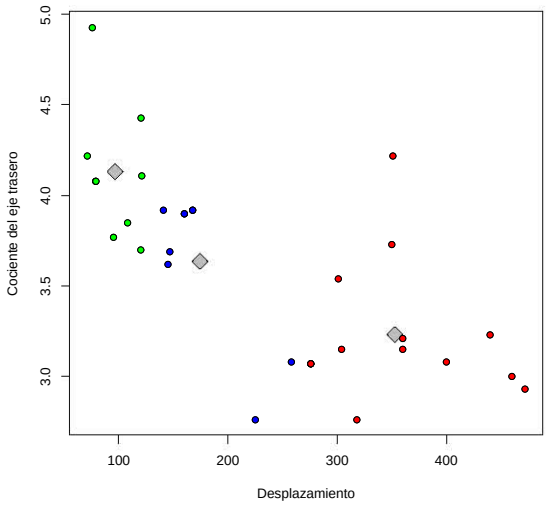
también se muestran los centroides finales como rombos de color gris.



Gráfica de los datos tras la primera iteración del método



Gráfica de los datos tras la segunda iteración del método



Gráfica de los datos tras la tercera iteración del método

Figura 1.3

1.4. Métodos basados en modelos

Los métodos que se basan en modelos buscan explicar el comportamiento de los datos a través de supuestos distribucionales y la estimación de parámetros adecuados para describir el comportamiento de los mismos.

Debido a la naturaleza multimodal del problema del análisis de conglomerados, los supuestos distribucionales de modelos de mezclas son ampliamente aceptados.

Cuando los modelos de mezclas asocian un componente a cada grupo en la muestra y se concentran en distribuciones de mezclas finitas, producen ciertos problemas pues las distribuciones

de las poblaciones de interés no siempre se ajustan a las elegidas en la mezcla.

Entre los modelos más comunes está la mezcla de distribuciones normales, que tiene la bondad de que facilita notablemente los cálculos, sin embargo, a pesar de usarse el ajuste normal, no se cuenta con un método de estimación estándar definido para los parámetros en dicho modelo, y en la práctica hay diversidad de alternativas.

Usualmente los modelos de este tipo consideran la probabilidad de que cierta observación x pertenezca al grupo j en términos de una variable indicadora z, es decir, z se define como una variable aleatoria tal que $z = j$ si y solo si $x \in C_j$, y el conjunto de pesos $w = (w_1, w_2...)$ está definido por $w_j = \mathbb{P}[z = j] = \mathbb{P}\left[x \in C_j\right]$, por lo que se satisface la restricción $\sum_{j=1}^{\infty} w_j = 1$, y se considera la densidad conjunta

$$f(x|\mu,\sigma) = \sum_{j=1}^{\infty} w_j \frac{1}{\sqrt{2\pi\sigma_j^2}} exp\Big\{\frac{-\left(x-\mu_j\right)^2}{2\sigma_j^2}\Big\},$$

donde $\mu = (\mu_1, ...)$ se refiere al vector formado por las medias de cada grupo y $\sigma = (\sigma_1, ...,)$ al vector de varianzas.

En el caso multivariado el razonamiento es el mismo, sea z una variable latente que indica el grupo al que pertenece el vector aleatorio $x \in \mathbb{R}^p$ bajo el supuesto de una probabilidad que obedece a la función de densidad de una mezcla de funciones normales multivariadas

$$f(x|\mu,\Sigma) = \sum_{j=1}^{\infty} w_j \frac{|\Sigma_j|^{-1/2}}{(2\pi)^{p/2}} exp[-\frac{1}{2}\left(x-\mu_j\right)^T \Sigma_j^{-1}\left(x-\mu_j\right)],$$

donde nuevamente $\mu = (\mu_1, ..., \mu_J)$ se refiere a las medias de cada grupo y $\sigma = (\sigma_1, ..., \sigma_J)$ a las varianzas.

Una vez que, que se tienen estimaciones para los parámetros se utiliza el modelo junto con estas estimaciones para calcular las probabilidades de pertenencia de cada observación a los distintos grupos que plantee el modelo. Así, asignar los datos a los grupos se hace con base en la probabilidad de pertenencia que nos devuelve el modelo evaluado en los parámetros estimados proporcionando al proceso de clasificación una medida de incertidumbre.

Por ejemplo, si se plantea una muestra de distribuciones normales, clasificar a la observación $x \in \mathbb{R}$ puede reducirse a que $x \in C_j$ si y solo si

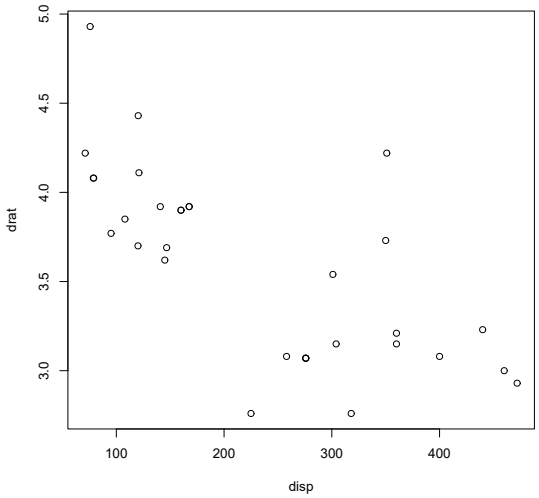
$$\widehat{w_j} \frac{1}{\sqrt{2\pi\widehat{\sigma_j^2}}} exp\Big\{\frac{-\left(x-\widehat{\mu_j}\right)^2}{2\widehat{\sigma_j^2}}\Big\} > \widehat{w_k} \frac{1}{\sqrt{2\pi\widehat{\sigma_k^2}}} exp\Big\{\frac{-\left(x-\widehat{\mu_k}\right)^2}{2\widehat{\sigma_k^2}}\Big\}$$

para cualquier componente k de la mezcla del modelo tal que $k \neq j$, y para agrupar los datos basta con calcular las densidades de probabilidad que genera el modelo.

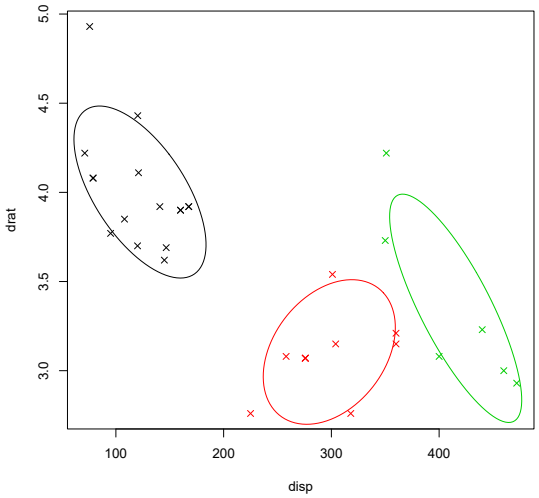
El algoritmo EM, por Expectation-Maximization, es un ejemplo de estos métodos. El modelo consiste en alternar entre dos pasos hasta lograr una convergencia. El primer paso es el de la Esperanza, en donde se estima la distribución de los grupos dado un modelo con parámetros determinados. Posteriormente, en el paso de Maximización, se toman en cuenta los resultados del paso anterior y se estiman nuevos parámetros al modelo, de manera que maximicen la función de log-verosimilitud esperada. Propiedades en el diseño del proceso garantizan su convergencia, y su condición de paro generalmente se da en términos del cambio en los parámetros o en los grupos encontrados en cada iteración.

A continuación se muestra el código para ejecutar el proceso para los mismos datos, inicialmente fijando un número de grupos igual a tres y posteriormente dejando que el modelo decida

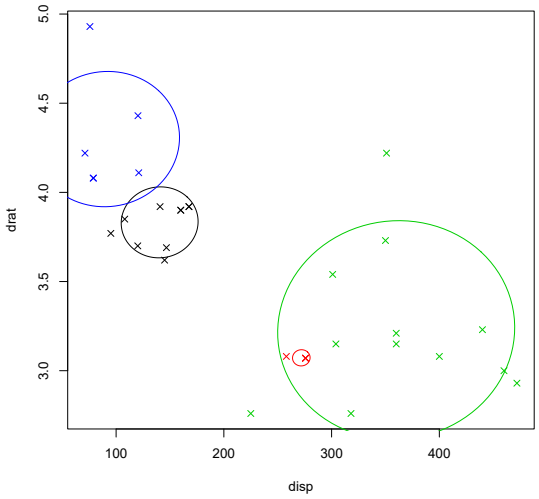
el número de clusters según la evaluación del Criterio de Información Bayesiano. En la figura 1.4 se muestran los datos y los resultados de cada ejemplo, se grafican también elipses con un nivel de confianza de 0.7.



Gráfica de los datos



Gráfica de los datos tras ejecutar EM con $n = 3$



Gráfica de los datos tras ejecutar EM con n que maximiza el BIC

Figura 1.4

Capítulo 2

Preliminares

Como herramientas preliminares se enuncian resultados sobre Teoría matricial, Geometría, Probabilidad, Estadística, Análisis Multivariado y el Muestreo de Gibbs.

Se considera que es un marco suficientemente amplio para el desarrollo de la teoría en el resto del documento, pues en el desarrollo del algoritmo se trabaja de manera general con datos multivariados, donde el álgebra matricial aparece; los principios de geometría se aplican para trabajar con elipses y elipsoides, su definición y propiedades; los fundamentos de teoría de probabilidad y estadística son necesarios pues se trabaja con funciones de densidad de probabilidades, probabilidad condicional, familias paramétricas y variables aleatorias normales, así como muestras aleatorias y funciones de verosimilitud, probabilidades posteriores y conceptos de estadística bayesiana; el desarrollo hace uso de herramientas del análisis multivariado debido a que el proceso corresponde a una generalización de una propuesta para datos univariados, y el manejo de datos de dimensiones múltiples se ilustra en dicha sección; finalmente aparece una breve descripción de la técnica llamada muestreo de Gibbs, misma que juega un rol central en todo el proceso a través de diversos usos que se hace de la misma durante el desarrollo del algoritmo.

2.1. Álgebra Matricial

Se considerará a $\mathbb{R}^{p\times q}$ el espacio de todas las matrices con entradas en los reales, de p renglones y q columnas, con las definiciones de suma, producto escalar y producto matricial usuales.

Definición 2.1.1. Se define, dada una matriz A en $\mathbb{R}^{p\times q}$, a la matriz traspuesta de A, denotada por A^T como la matriz en $\mathbb{R}^{q\times p}$ tal que $A_{ij} = (A^T)_{ji}$.

Teorema 2.1.1. *Sea A en $\mathbb{R}^{p\times q}$, entonces su matriz traspuesta A^T satisface que:*
$\left(A^T\right)^T = A, \qquad (A+B)^T = A^T + B^T \qquad y \qquad (AB)^T = B^T A^T.$

Definición 2.1.2. Se dice que A en $\mathbb{R}^{p\times p}$ es invertible si existe A^{-1} en $\mathbb{R}^{p\times p}$ tal que $AA^{-1} = I_p$ donde I_p es la matriz identidad. La matriz A^{-1} se denomina la matriz inversa de A.

Definición 2.1.3. Sea A en $\mathbb{R}^{p\times p}$, se defne al determinante de A, denotado como $|A|$ a

$$|A| = \sum_{i=1}^{p} (-1)^{i+j} A_{ij} \cdot |\widetilde{A_{ij}}|$$

donde $\widetilde{A_{ij}} \in \mathbb{R}^{(p-1)\times(p-1)}$ se refiere a la matriz obtenida de A al suprimir el renglón i y la columna j.

Teorema 2.1.2. *Sean A,B en $\mathbb{R}^{p\times p}$, el determinante satisface que*
$|A^T| = |A|$ *y que* $|AB| = |A|\cdot|B|$.

Definición 2.1.4. Se dice que $A \in \mathbb{R}^{p\times p}$ es singular si $|A| = 0$.

Teorema 2.1.3. *Las siguientes enunciados son equivalentes:*

- *A no tiene inversa.*
- *$|A| = 0$, es decir, A es singular.*
- *La igualdad $Ax = 0$ tiene infinitas soluciones x.*

Definición 2.1.5. Sea A en $\mathbb{R}^{p\times p}$ es una matriz simétrica si $A = A^T$, y se denotará por $\mathbb{R}^{p\times p}_{sym}$ al conjunto de matrices reales simétricas, de p columnas y p renglones.

Definición 2.1.6. Sea A en $\mathbb{R}^{p\times p}_{sym}$, se dice que A es definida positiva si para cualquier vector $x \neq \vec{0}$ en $\mathbb{R}^p$, se tiene que $x^TAx > 0$ donde

$$x = \begin{bmatrix} x_1 \\ x_2 \\ \vdots \\ x_p \end{bmatrix} \qquad y \qquad x^T = [x_1, x_2, \cdots, x_p]$$

y se denota a $\mathbb{R}^{p\times p}_{sym,+}$ como el conjunto de matrices reales, de p columnas y p renglones, simétricas y definidas positivas.

Teorema 2.1.4. *Si $A \in \mathbb{R}^{p\times p}_{sym,+}$, entonces A^{-1}, el inverso multiplicativo de A, también está en $\mathbb{R}^{p\times p}_{sym,+}$.*

Definición 2.1.7 (Traza)**.** Sea $A \in \mathbb{R}^{p\times p}$, se define la traza de la matriz A o $tr(A)$ como $\sum_{i=1}^{p} a_{ii}$.

Teorema 2.1.5. *La traza satisface las siguientes condiciones:*
- $tr(A+B) = tr(A) + tr(B)$
- $tr(AB) = tr(BA)$.

2.2. Consideraciones Geométricas

2.2.1. Elipses en $\mathbb{R}^2$

Sección cónica es el nombre que recibe la figura que se obtiene al cortar con un plano diferentes partes de un cono doble con una inclinación determinada. En esta sección se tratará únicamente el caso de las elipses, pero se sugiere consultar [22], [17] o [5] si se desea profundizar en el tema.

Definición 2.2.1. Una elipse es el lugar geométrico de los puntos P del plano tales que la suma de sus distancias a dos puntos fijos F_1 y F_2 llamados focos es una constante denotada con $2a$, es decir,

$$d(P,F_1) + d(P,F_2) = 2a$$

donde $d(P,F)$ denota la distancia del punto P a F.

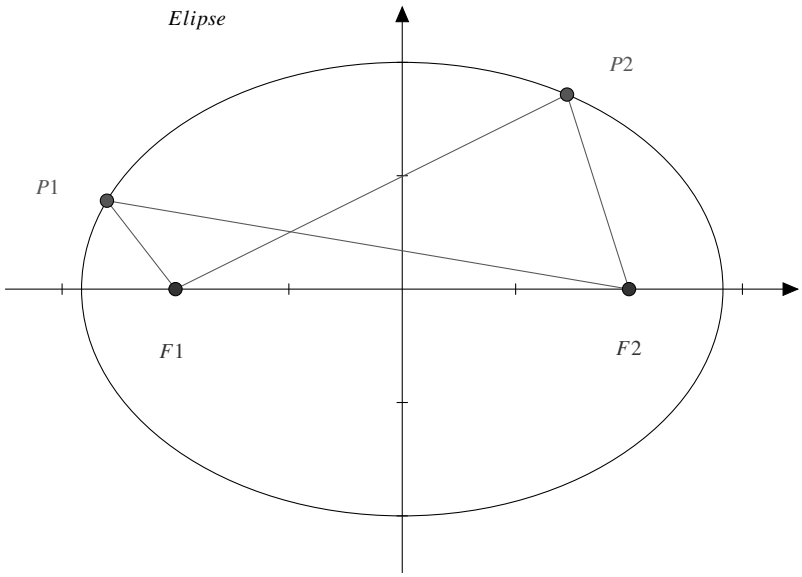


Figura 2.1: Elipse con focos en F_1 y F_2 y centro en el origen.

Formalmente se puede reescribir la definición de una elipse como el conjunto

$$\mathscr{E} := \left\{ P \in \mathbb{R}^2 : d(P,F_1) + d(P,F_2) = 2a \right\}.$$

En la figura 2.1 se muestra una elipse con focos sobre el eje horizontal y dos puntos P_1 y P_2 que pertenecen a ésta. La definición 2.2.1 establece que la suma de la longitud de los segmentos F_1P_1 y P_1F_2 es igual a la suma de la longitud de los segmentos $\overline{F_1P_2}$ y $\overline{P_2F_2}$.

Usualmente las elipses y las cónicas en general se representan con base en su ecuación canónica. En el caso de las elipses con centro en el origen, ésta tiene una forma $\frac{x^2}{a^2} + \frac{y^2}{b^2} = 1$, donde los parámetros a y b dependen de los focos y la distancia entre ellos y la elipse. Cuando la elipse se encuentra centrada en un punto (h,k), la ecuación adquiere la forma $\frac{(x-h)^2}{a^2} + \frac{(y-k)^2}{b^2} = 1$. Cuando los ejes de la elipse no son paralelos a los ejes del plano, es decir, la elipse está inclinada, la ecuación tiene también un término mixto mxy. Esta ecuación se puede desaarrollar en forma de polinomio como se muestra en el siguiente teorema.

Teorema 2.2.1. *Cualquier sección cónica se puede representar como el conjunto de puntos que satisface una ecuación de la forma*

$$Ax^2 + Bxy + Cy^2 + Dx + Ey + F = 0,$$

donde A,B,C,D,E y F son números reales y al menos uno de A,B y C es distinto de cero. Inversamente cualquier conjunto de parejas ordenada (x,y) que satisface la ecuación anterior es una sección cónica, más aún, si $B^2 - 4AC < 0$, entonces la ecuación corresponde a una elipse [5].

Teorema 2.2.2. *Sean $F_1, F_2 \in \mathbb{R}^2$, $a \in \mathbb{R}^+$ y $\mathscr{E}$ la elipse generada por éstos parámetros, entonces se puede reescribir a $\mathscr{E}$ como el conjunto*

$$\mathscr{E} = \left\{ X \in \mathbb{R}^2 : (X-O)^T M (X-O) = 1 \right\}$$

donde M es una matriz simétrica y definida positiva, y $O \in \mathbb{R}^2$ denota al centro del elipsoide.

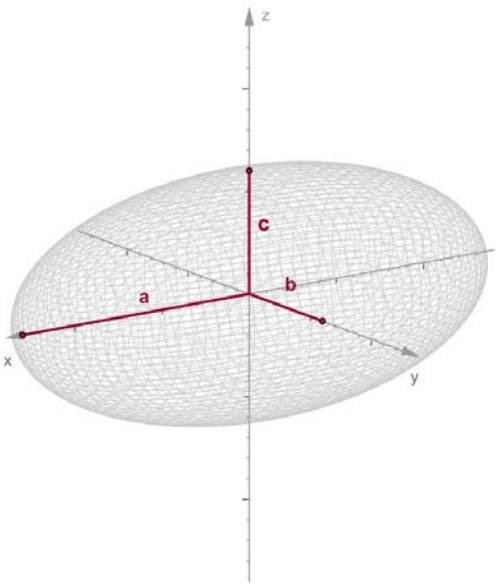


Figura 2.2: Elipsoide con centro en el origen y ejes paralelos a los ejes coordenados.

2.2.2. Elipsoides en $\mathbb{R}^n$

Para generalizar la idea de una elipse a dimensiones superiores se utilizará una definición de elipsoide distinta a la anterior pero que es equivalente y que además resulta conveniente en notación.

Definición 2.2.2. Sea M una matriz simétrica definida positiva de $p \times p$ y μ un vector en $\mathbb{R}^p$ se define al elipsoide generado por M con centro en μ como el conjunto de puntos

$$\mathscr{E} := \left\{ x \in \mathbb{R}^p : (x-\mu)^T M (x-\mu) = 1 \right\}, \text{ o}$$

$$\mathscr{E} := \left\{ x \in \mathbb{R}^p : \|M^{1/2}(x-\mu)\| = 1 \right\}.$$

Es importante hacer notar que cuando $\mu = \vec{0}$, la ecuación del elipsoide tiene una forma $x^T M x = 1$ y si además M es la matriz identidad el elipsoide coincide con la circunferencia unitaria, por lo que, por definición, la esfera unitaria se considera una elipse.

A continuación, se define la función gamma que aparece en la expresión para el volumen de un elipsoide $\mathscr{E}$, cabe señalar que esta función también se menciona repetidamente en las secciones de Probabilidad e Inferencia Estadística de éste capítulo.

Definición 2.2.3. Sea $\alpha > 0$, se define a la función gamma de α como

$$\Gamma(\alpha) := \int_0^\infty t^{\alpha-1} e^{-t} dt$$

.

Teorema 2.2.3 (Volumen de un elipsoide)**.** *Sean M una matriz definida positiva, $\mu \in \mathbb{R}^p$ y $\mathscr{E}_{(\mu,M)}$ la elipse generada por estos dos elementos, $\mathscr{E}_{(\mu,M)} := \left\{ x \in \mathbb{R}^p : \|M^{1/2}(x-\mu)\| = 1 \right\}$. Entonces el volumen de $\mathscr{E}_{(\mu,M)}$ está en términos del determinante de M y la dimensión p.*

$$Vol\left(\mathscr{E}_{(\mu,M)}\right) = |M|^{-1/2} \cdot \frac{\pi^{p/2}}{\Gamma\left(\frac{p}{2}+1\right)}$$

La prueba del teorema anterior se deriva de que el cambio en el volumen de una figura es al ser sometida a una transformación lineal es proporcional al determinante de la matriz asociada a la tranformación y a un escalar que está en función de la dimensión. Tomando en cuenta que los elipsoides son transformaciones lineales de esferas unitarias $\mathscr{E}_{(\vec{0},I_p,1)}$ con volumen conocido [9, p. 302]

$$Vol\left(\mathscr{E}_{(\vec{0},I_p)}\right) = \frac{\pi^{p/2}}{\Gamma\left(\frac{p}{2}+1\right)}$$

y que las traslaciones no afectan al volumen de la figura, se puede deducir la formula anterior para el volumen del elipsoide.

2.3. Probabilidad

A continuación se definen conceptos importantes y se enuncian resultados relevantes dentro de la teoría de probabilidad y que, en su mayoría, se encuentran en las aplicaciones más adelante.

Definición 2.3.1 (Espacio muestral)**.** El espacio muestral denotado por Ω es la colección o la totalidad de los posibles resultados de un experimento aleatorio.

¡

Definición 2.3.2 (Eventos)**.** Un evento se define como un subconjunto del espacio muestral Ω, es decir, un evento es una colección de posibles resultados de un experimento.

Definición 2.3.3. Sea $\mathscr{F}$ una colección de subconjuntos de Ω, se dice que $\mathscr{F}$ es una σ-álgebra si satisface las siguientes propiedades:

i $\emptyset \in \mathscr{F}$.

ii Si $A \in \mathscr{F}$, entonces $A^c \in \mathscr{F}$.

iii Si $A_1, A_2, \ldots \in \mathscr{F}$, entonces $\bigcup_{i=1}^{\infty} A_i \in \mathscr{F}$

Definición 2.3.4. Dados un espacio muestral Ω y $\mathscr{F}$ una σ-álgebra asociada, una función de probabilidad $\mathbb{P}$ es una función con dominio $\mathscr{F}$que satisface las siguientes propiedades:

i $\mathbb{P}(A) \geq 0$ para todo $A \in \mathscr{F}$

ii $\mathbb{P}(\Omega) = 1$

iii Si $A_1, A_2, \ldots \in \mathscr{F}$ es una secuencia de eventos mutuamente excluyentes, entonces $\mathbb{P}(\bigcup_{i=1}^{\infty} A_i) = \sum_{i=1}^{\infty} \mathbb{P}(A_i)$

Definición 2.3.5. Sean Ω, $\mathscr{F}$ y $\mathbb{P}$ un espacio muestral, una σ-álgebra asociada y una función de probabilidad con dominio en $\mathscr{F}$ respectivamente, entonces se dice que $(\Omega, \mathscr{F}, \mathbb{P})$ es un espacio de probabilidad.

Definición 2.3.6 (Probabilidad condicional)**.** Sea $(\Omega, \mathscr{F}, \mathbb{P})$ un espacio de probabilidad, A y B eventos en $\mathscr{F}$ y $\mathbb{P}(B) > 0$, entonces la probabilidad condicional de A dado B, denotado como $\mathbb{P}(A|B)$ es:

$$\mathbb{P}(A|B) = \frac{\mathbb{P}(A \cap B)}{\mathbb{P}(B)}.$$

Teorema 2.3.1 (Regla de Bayes)**.** *Para un espacio de probabilidad* $(\Omega, \mathscr{F}, \mathbb{P})$*, sea* $A_1, A_2, \ldots$ *una particion del espacio muestral en* $\mathscr{F}$*, y sea B cualquier conjunto en* $\mathscr{F}$*, entonces para cada* $i = 1, 2, \ldots$*, tal que* $\mathbb{P}(A_i) > 0$ *se tiene que*

$$\mathbb{P}(A_i|B) = \frac{\mathbb{P}(B|A_i)\,\mathbb{P}(A_i)}{\sum_{j=1}^{\infty} \mathbb{P}\left(B|A_j\right)\mathbb{P}\left(A_j\right)}$$

Definición 2.3.7 (Independencia)**.** Sea $(\Omega, \mathscr{F}, \mathbb{P})$ un espacio de probabilidad, A y B eventos en $\mathscr{F}$, se dice que A y B son estadísticamente independientes si

$$\mathbb{P}(A \cap B) = \mathbb{P}(A)\,\mathbb{P}(B).$$

Definición 2.3.8. Sea $(\Omega, \mathscr{F}, \mathbb{P})$ un espacio de probabilidad, una variable aleatoria es una función del espacio muestral Ω a los números reales. La función $X(\cdot)$ debe satisfacer que cada conjunto $\{\omega : X(\omega) \leq r\}$ está en $\mathscr{F}$ para cada $r \in \mathbb{R}$.

Definición 2.3.9 (Función de distribución)**.** La función de distribución de una variable aleatoria denotada por $F_X(\cdot)$ está defninida como la función con dominio en la recta real que va al intervalo $[0,1]$ y que satisface $F_X(x) = \mathbb{P}[X \leq x] = \mathbb{P}[\{\omega : X(\omega) \leq x\}]$ para cada $x \in \mathbb{R}$.

Teorema 2.3.2. *La función* $F(x)$ *es una función de distribución si y solo si satisface las siguientes tres propiedades:*

i $\lim_{x \to -\infty} F(x) = 0$, *y* $\lim_{x \to \infty} F(x) = 1$.

ii $F(x)$ *es una función no decreciente de x.*

iii $F(x)$ *es una función continua por la derecha, es decir,* $\lim_{x \downarrow x_0} F(x) = x_0$

Definición 2.3.10 (Variables aleatorias discretas)**.** Una variable aleatoria X está definida como una variable discreta si el rango de X es a lo más numerable. Si una variable aleatoria es discreta entonces se dice que su función de distribución es discreta.

Definición 2.3.11. Sea X una variable aleatoria discreta, se define a su función de masa de probabilidad denotada por $f_X(\cdot)$ como

$$f_X(x) = \mathbb{P}[X = x]$$

Definición 2.3.12 (Variables aleatorias continuas)**.** Una variable aleatoria X es llamada una variable continua si existe una función $f_X(\cdot)$ tal que la función de distribución $F_X(\cdot)$ se puede escribir como $F_X(x) = \int_{-\infty}^{x} f_x(u)du$ para cualquier x en la recta real. Se dice entonces que la función de distribución $F_X(\cdot)$ es absolutamente continua.

Definición 2.3.13. Sea X una variable aleatoria continua, a la función $f_X(\cdot)$ que aparece en $F_X(x) = \int_{-\infty}^{x} f_x(u)du$ se le denomina función de densidad de probabilidad de la variable aleatoria X.

Definición 2.3.14 (Media)**.** El valor esperado o media de la variable aleatoria X denotada por $\mathbb{E}[X]$ se define como

$$\mathbb{E}[X] = \begin{cases} \int_{-\infty}^{\infty} x f_X(x)dx & \text{si } X \text{ es continua} \\ \sum_{x \in \chi} x f_X(x) & \text{si } X \text{ es discreta} \end{cases}$$

siempre que la integral o la suma exista, en caso contrario se dice que la media de la variable aleatoria no existe.

Teorema 2.3.3. *Sean X y Y variables aleatorias, a, b y c números reales, entonces*

$$\mathbb{E}[aX+bY+c] = a\mathbb{E}[X]+b\mathbb{E}[Y]+c.$$

Definición 2.3.15. Sea X una variable aleatoria y sea $\mu_X = \mathbb{E}[X]$. La varianza de X, denotada por σ_X^2 o var$[X]$, se define como

$$\sigma_X^2 = \mathbb{E}[(X-\mu_X)^2].$$

Definición 2.3.16 (Distribución normal)**.** Una variable aleatoria continua se define como distribuida normal o $X \sim N(\mu, \sigma^2)$ si su función de densidad está dada por

$$f_X(x) = \frac{1}{\sqrt{2\pi\sigma^2}} \cdot e^{\frac{-(x-\mu)^2}{2\sigma^2}},$$

donde los parámetros μ y σ^2 satisfacen $-\infty < \mu < \infty$ y $\sigma^2 > 0$.

Teorema 2.3.4. *Si X una variable aleatoria* $X \sim N(\mu, \sigma^2)$*, entonces* $\mathbb{E}[X] = \mu$ *y var*$[X] = \sigma^2$.

Teorema 2.3.5. *Sea X una variable aleatoria normal* $X \sim N(\mu, \sigma^2)$*, si se define a Y como* $Y := aX + b$*, donde a y b números en la recta real, entonces Y también tiene una distribución normal como sigue,* $Y \sim N(a\mu + b, a^2\sigma^2)$.

Definición 2.3.17 (Distribución Gamma)**.** Si una variable aleatoria X tiene una densidad dada por

$$f_X(x|r,\lambda) = \frac{\lambda^r}{\Gamma(r)} x^{r-1} e^{-\lambda x} \mathbb{1}_{[0,\infty)}$$

donde $r > 0$ y $\lambda > 0$, entonces se dice que X tiene una distribución gamma o $X \sim Gamma(r,\lambda)$, donde $\Gamma(\cdot)$ se refiere a la función definida en 2.2.3.

Teorema 2.3.6. *Si X es una variable aleatoria con distribución gamma con parámetros r y* λ*, entonces*

$$\mathbb{E}[X] = \frac{r}{\lambda} \;\; y \;\; Var[X] = \frac{r}{\lambda^2}.$$

Definición 2.3.18. Sea X una variable aleatoria, se dice que X tiene una distribución ji-cuadrada con k grados de libertad o $X \sim \chi_k^2$ si $X \sim Gamma(1/2, k/2)$.

Definición 2.3.19 (Covarianza)**.** Sean X y Y variables aleatorias, la covarianza de X y Y, denotada como Cov$[X,Y]$ está definida como

$$Cov[X,Y] = \mathbb{E}[(X-\mathbb{E}[X])(Y-\mathbb{E}[Y])].$$

Definición 2.3.20 (Correlación)**.** Sean X y Y variables aleatorias, la correlación de X y Y, denotada como $\rho[X,Y]$ está definida como

$$\rho[X,Y] = \frac{Cov[X,Y]}{\sqrt{\sigma_X^2 \sigma_Y^2}}$$

siempre que la covarianza, y las varianzas existan.

Definición 2.3.21. Un vector aleatorio p-dimensional es una función de un espacio muestral a $\mathbb{R}^p$, el espacio euclideano p-dimensional. Es una generalización del concepto de variable aleatoria.

Definición 2.3.22. Sea $X = (X_1, X_2, ..., X_p)$ un vector aleatorio, la función de distribución conjunta del vector es una función que va de $\mathbb{R}^p$ al intervalo $[0,1]$ definida como

$$F_{X_1,X_2,...,X_p}(x_1, x_2, ..., x_p) = \mathbb{P}[X_1 \leq x_1, X_2 \leq x_2, ..., X_p \leq x_p].$$

Definición 2.3.23. Se dice que el vector aleatorio $X = (X_1, X_2, ..., X_p)$ es absolutamente continuo si existe $f_{X_1,X_2,...,X_p(\cdot)}$ función de $\mathbb{R}^p$ a la recta real y que satisface que

$$F_{X_1,X_2,...,X_p}(x_1, x_2, ..., x_p) = \int_{-\infty}^{x_1}\int_{-\infty}^{x_2}\cdots\int_{-\infty}^{x_p} f_{X_1,X_2,...,X_p}(x_1, x_2, ..., x_p)dx_1dx_2...dx_p,$$

a esta función $f_{X_1,X_2,...,X_p(\cdot)}$ se le denomina función de densidad de probabilidad del vector aleatorio.

Definición 2.3.24. Se dice que se tiene un vector aleatorio $X = (X_1, X_2, ..., X_p)$ si los puntos $(x_1, x_2, ..., x_p)$ donde tiene valores en el espacio p-dimensional es a lo más numerable.

Definición 2.3.25. Si $X = (X_1, X_2, ..., X_p)$ es un vector aleatorio discreto, se define a la función de masa de probabilidad como

$$f_{X_1,X_2,...,X_p}(x_1, x_2, ..., x_p) = \mathbb{P}[X_1 = x_1, X_2 = x_2, \cdots, X_p = x_p].$$

Definición 2.3.26. Sea $X = (X_1, X_2, ..., X_p)$ un vector aleatorio, se define a la distribución marginal de X_i como

$$f_{X_i}(x) = \begin{cases} \int_{-\infty}^{\infty}\cdots\int_{-\infty}^{\infty} f_{X_1,...,X_p}(x_1, ..., x_{i-1}, x, x_{i+1}, ..., x_p)dx_1..., dx_{i-1}dx_{i+1}...dx_p & \text{si X es continua} \\ \sum_{x_1\in\chi}\cdots\sum_{x_p\in\chi} f_{X_1,...,X_p}(x_1, ..., x_{i-1}, x, x_{i+1}, ..., x_p) & \text{si X es discreta.} \end{cases}$$

Definición 2.3.27. Sean X un vector aleatorio en $\mathbb{R}^p$, $\mu \in \mathbb{R}^p$ y $\Lambda \in \mathbb{R}^{p\times p}_{sym,+}$, decimos que tiene una distribución uniforme en el elipsoide $\mathscr{E}_{(\mu,M)} = \left\{x \in \mathbb{R}^p : \|M^{1/2}(x-\mu)\| = 1\right\}$ si su densidad está dada por

$$f(x|\mu,\Lambda) = \begin{cases} \frac{1}{Vol(\mathscr{E}_{(\mu,M)})} & \text{si } x \in \mathscr{E}_{(\mu,M)} \\ 0 & \text{c.o.c.} \end{cases}$$

Definición 2.3.28. Sean Y_i $i = 1, ..., n$, variables aleatorias con κ posibles resultados $s_1, ..., s_\kappa$ tales que $\mathbb{P}[Y_i = s_j] = p_j$ y $\sum_{j=1}^{\kappa} p_j = 1$. Sea $X_j = \sum_{i=1}^{n} \mathbb{I}_{(Y_i=s_j)}$, $j = 1, ..., \kappa$ el vector κ-dimensional que cuenta las observaciones para cada posible salida, entonces se dice que el vector $X = (X_1, ..., X_\kappa)$ tiene una distribución multinomial, y su función de másа de probabilidad es

$$f_{X_1,...,X_\kappa}(x_1, ..., x_\kappa) = \frac{n!}{\prod_{j=1}^{\kappa} x_j}\prod_{j=1}^{\kappa} p_j.$$

Definición 2.3.29. Si $X = (X_1, X_2)$ es un vector aleatorio en $\mathbb{R}^2$, la probabilidad condicional de X_1 dada X_2 está definida como

$$f_{X_1}(x_1|X_2 = x_2) = \frac{f_{X_1,X_2}(x_1, x_2)}{f_{X_2}(x_2)}$$

donde $f_{X_2}(x_2) > 0$

Definición 2.3.30. Sea $X = (X_1, X_2, ..., X_p)$ un vector aleatorio, se dice que tiene una distriución Dirichlet o que $X \sim \text{Dir}(\alpha)$, con $\alpha = (\alpha_1, \alpha_2, ..., \alpha_p) \in \mathbb{R}^{p,+}$ un vector de reales positivos si

$$f_{X_1,X_2,...,X_p}(x_1, x_2, ..., x_p) = \frac{1}{B(\alpha)} \prod_{i=1}^{p} x_i^{\alpha_i - 1}$$

donde $B(\alpha) = \frac{\prod_{i=1}^{p} \Gamma(\alpha_i)}{\Gamma(\sum_{i=1}^{p} \alpha_i)}$.

Teorema 2.3.7. *Si $X \in \mathbb{R}^p$ un vector aleatorio que tiene una distribución Dirichlet con parámetro $\alpha \in \mathbb{R}^{p,+}$, entonces*

$$\mathbb{E}[X_i] = \frac{\alpha_i}{\sum_{j=1}^{p} \alpha_j}$$

y

$$Var[X_i] = \frac{\alpha_i \left(\sum_{j=1}^{p} \alpha_j - \alpha_i \right)}{\left(\sum_{j=1}^{p} \alpha_j \right)^2 \left(\sum_{j=1}^{p} \alpha_j + 1 \right)}.$$

2.4. Inferencia Estadística

2.4.1. Muestras Aleatorias

En muchas ocasiones la información obtenida de un experimento consiste en diversas observaciones de una variable de interés. Por ello la siguiente definición establece matemáticamente qué se entiende por una muestra aleatoria, este concepto es la base del enfoque con que generalmente se tratan estas situaciones .

Definición 2.4.1 (Muestra aleatoria)**.** Las variables aleatorias $X_1, X_2, ..., X_n$ son llamadas una muestra aleatoria de tamaño n de la distribución $f(\cdot)$ si $X_1, X_2, ..., X_n$ son mutuamente independientes, y la función de densidad marginal de cada X_i es la función $f(\cdot)$.

Definición 2.4.2 (Media muestral)**.** Sea $X_1, X_2, ..., X_n$ una muestra aleatoria, se define a la media muestral denotada por $\bar{X}$ como

$$\bar{X} = \frac{1}{n} \sum_{i=1}^{n} X_i.$$

Definición 2.4.3 (Varianza muestral)**.** Sea $X_1, X_2, ..., X_n$ una muestra aleatoria, se define a la varianza muestral denotada por S^2 como

$$S^2 = \frac{1}{n-1} \sum_{i=1}^{n} (X_i - \bar{X})^2.$$

Teorema 2.4.1. *Sea $X_1, X_2, ..., X_n$ una muestra aleatoria de una distribución con media μ y varianza $\sigma^2 < \infty$, entonces*

$$\mathbb{E}[\bar{X}] = \mu, \;\; Var[\bar{X}] = \frac{\sigma^2}{n} \;\; y \;\; \mathbb{E}[S^2] = \sigma^2$$

Teorema 2.4.2. *Sea $X_1, X_2, ..., X_n$ una muestra aleatoria con distribución $N(\mu, \sigma^2)$, entonces*

i $\bar{X}$ y S^2 son variables aleatorias independientes,

ii $\bar{X}$ *tiene una distribución* $N\left(\mu, \sigma^2/n\right)$,

iii $(n-1)S^2/\sigma^2$ *tiene una distribución ji-cuadrada con* $n-1$ *grados de libertad.*

Teorema 2.4.3. *Si* $Z \sim N(0,1)$*, entonces* $Z^2 \sim \chi_1^2$ *además, si* $X_1, X_2, ..., X_n$ *son variables aleatorias independientes, tales que* $X_i \sim \chi_{k_i}^2$ *y* $Y := \sum_{i=1}^n X_i$*, entonces* $Y \sim \chi_{k_1+k_2+...+k_n}^2$.

2.4.2. Estimación

Si se obtienen datos a través de una muestra aleatoria, y estos se rigen por una distribución con densidad $f(x|\theta)$, donde el vector θ representa a todos los parámetros de la distribución, entonces el comportamiento de la variable a observar está completamente determinado por el parámetro θ. Es por eso que se busca, a partir de la muestra, obtener un buen estimador para θ. Esto se hace mediante diversos métodos de estimación, de los cuales dos de los más comunes se enunciaran en esta sección, el estimador máximo verosímil y el estimador bayesiano.

Definición 2.4.4. Un estimador es cualquier función $W(X_1, ..., X_n)$ de la muestra.

Máxima Verosimilitud

Definición 2.4.5. Sea $X_1, X_2, ..., X_n$ una muestra aleatoria de la distribución $f(x|\theta)$, definimos a la función de verosimilitud $L(\cdot|x_1, ..., x_n)$ como la función de densidad o masa conjunta de la muestra, vista en términos del parámetro θ, es decir,

$$L(\theta|x_1, ..., x_n) := \prod_{i=1}^n f(x_i|\theta).$$

Definición 2.4.6. Sea $x = (x_1, ..., x_n)$ un punto muestral, sea además $\hat{\theta}(x)$ el valor de θ que maximiza a la función de verosimilitud $L(\theta|x)$, entonces decimos que $\hat{\theta}(x)$ es un estimador máximo verosímil de θ.

Teorema 2.4.4. *Sea* $X_1, X_2, ..., X_n$ *una muestra aleatoria de la distribución normal de parámetros media* μ *y varianza* σ^2*. Los estimadores de máxima verosimilitud respectivamente son*

$$\hat{\mu} = \bar{X} \;\; y \;\; \hat{\sigma^2} = \frac{1}{n}\sum_{i=1}^n (x_i - \bar{X}).$$

Estimación Bayesiana

Existen claras diferencias entre los enfoques clásico y bayesiano de la estadística. En el enfoque clásico el parámetro θ es pensado desconocido, pero fijo, mientras que en el enfoque bayesiano el parámetro θ es considerado como una cantidad cuyos cambios pueden ser descritos mediante una variable aleatoria con distribución inicial o previa, llamada distribución *a piori*. Posteriormente se toma en cuenta la información que aporta la muestra y a partir de ésta y de la distribución previa, se obtiene, mediante el teorema 2.3.1 (Regla de Bayes), una función de distribución de probabilidad posterior.

Si se considera a $\pi(\theta)$ como la función de densidad que describe el comportamiento de la distribución de probabilidad inicial o a piori del parámetro θ y a $f(x_1, ..., x_n|\theta)$ como la función de densidad o masa de la muestra, entonces aplicando la regla de Bayes se tiene que la distribución posterior de θ dada la muestra está dada por

$$\pi(\theta|x_1, ..., x_n) = f(x_1, ..., x_n|\theta)\,\pi(\theta)\,/m(x_1, ..., x_n)$$

donde m es la distribución marginal de $(X_1,...X_n)$, es decir,

$$m(x_1,...,x_n) := \int f(x_1,...,x_n|\theta)\,\pi(\theta)\,d\theta.$$

Debido a que la función $m(\cdot)$ es una constante con respecto a θ, se suele escribir que

$$\pi(\theta|x_1,...,x_n) \propto f(x_1,...,x_n|\theta)\,\pi(\theta),$$

además si $X_1,...,X_n$ es una muestra aleatoria, de acuerdo a 2.4.5, la función $f(x_1,...,x_n|\theta)$ se puede ver como la función de verosimilitud $L(\cdot|x_1,...,x_n)$, por lo tanto se puede escribir a la distribución posterior de θ como

$$\pi(\theta|x_1,...,x_n) \propto L(\theta|x_1,...,x_n)\,\pi(\theta)$$

que para tener la igualdad sólo hace falta multiplicarla por una constante que haga que la integral sea uno.

Teorema 2.4.5. *Sea $X_1,...,X_n$ una muestra aleatoria de la distribución $N(\mu,\sigma^2)$, si se supone σ^2 conocida, y una distribución a priori $\mu \sim N(\mu_0,\tau^2)$, con $\sigma^2,\tau^2 > 0$ y $\mu_0 \in \mathbb{R}$, entonces $\mu|x_1,...,x_n$, se distribuye normal con media $\frac{n\tau^2\bar{x}+\sigma^2\mu}{n\tau^2+\sigma^2}$ y varianza $\frac{\sigma^2\tau^2}{n\tau^2+\sigma^2}$.*

Definición 2.4.7. Sea F la clase de funciones de masa o de densidad de probabilidad $f(x|\theta)$, indexadas por θ. Una clase Π de distribuciones *a priori* es una familia conjudada de F si la distribución posterior está en la clase Π para toda $f \in F$, todas las previas en Π y todos los puntos muestrales $(x_1,...,x_n)$.

En el teorema anterior, decimos que la normal es su propia conjugada, pues si una muestra aleatoria se distribuye normal y la distribución *a priori* de la media es normal, entonces la distribución posterior de la media también es normal.

Definición 2.4.8. Sea X una variable aleatoria normal con media μ y varianza σ^2, definimos al parámetro precisión de X como $\lambda = 1/\sigma^2$.

Teorema 2.4.6. *Sea $X_1,...,X_n$ una muestra aleatoria distribuida normal con media μ y precisión λ tal que $\mu|\lambda$ tiene una distribución a priori condicional normal con media μ_0 y varianza $(\kappa\lambda)^{-1}$, y que λ se distribuye gamma con parámetros α y β. Entonces las distribuciones posteriores de λ y μ quedan*

$$\mu|\lambda \sim N\left(\mu_n, \frac{\lambda}{n+\kappa}\right) \qquad \lambda \sim Gamma(\alpha_n,\beta_n).$$

Donde

$$\begin{aligned}
\mu_n &= \frac{n\bar{x}+\kappa\mu_0}{n+\kappa} \\
\alpha_n &= \alpha + n/2 \\
\beta_n &= \beta + \frac{(n-1)S^2}{2} + \frac{n\kappa(\bar{x}-\mu_0)^2}{2(n+\kappa)}.
\end{aligned}$$

2.5. Estadística Multivariada

Definición 2.5.1 (Normal multivariada)**.** Se dice que $X = [X_1, ..., X_n]^T$ un vector aleatorio tiene una distribución normal con parámetros μ y Σ, denotado como $X \sim N_p(\mu, \Sigma)$, si su función de densidad está dada por

$$f_X(x|\mu,\Sigma) = \frac{|\Sigma|^{-1/2}}{(2\pi)^{p/2}} \cdot exp[-\frac{1}{2}(x-\mu)^T \Sigma^{-1}(x-\mu)].$$

Definición 2.5.2. Se define a una matriz aleatoria Z como una matriz donde cada elemento Z_{ij} es una variable aleatoria.

Definición 2.5.3. El valor esperado o media de una matriz aleatoria en $\mathbb{R}^{p\times q}$ se define como la matriz de valores esperados de cada una de sus entradas, es decir,

$$\mathbb{E}[A]_{ij} = \mathbb{E}[A_{ij}], \qquad i = 1, ..., p, \quad j = 1, ..., q.$$

Definición 2.5.4. Si X un vector aleatorio en $\mathbb{R}^p$, definimos a $\mathbb{E}[(X - \mathbb{E}[X])(X - \mathbb{E}[X])^T]$ como la matriz de covarianzas de X.

Teorema 2.5.1. *Si X un vector aleatorio tal que $X \sim N_p(\mu, \Sigma)$, entonces la media y la matriz de covarianzas de X son respectivamente μ y Σ. En ocasiones se denotara a la densidad de este vector aleatorio normal como $N_p(x|\mu, \Sigma)$.*

Definición 2.5.5. Si $X \sim N_p(\mu, \Sigma)$ se define a su matriz de precisión como $\Lambda = \Sigma^{-1}$.

Teorema 2.5.2. *Si $X \sim N_p(\mu, \Sigma)$, $a \in \mathbb{R}^q$, $C \in \mathbb{R}^{q\times p}$ no singular y*

$$Y := CX + B$$

entonces $Y \sim N_q\left(C\mu + a, C\Sigma C^T\right)$.

Definición 2.5.6. La función *Gamma* multivariada se define como

$$\Gamma_p(t) = \pi^{p(p-1)/4} \prod_{i=1}^{p} \Gamma\left(t - \frac{1}{2}(i-1)\right).$$

Definición 2.5.7 (Distribución Wishart)**.** Sea A una matriz aleatoria en $\mathbb{R}^{p\times p}$, se dice que A se distribuye Wishart con grados de libertad $v > p-1$ y matriz de dispersión $T \in \mathbb{R}^{p\times p}_{sym,+}$, denotado por $A \sim W(v, T)$, si su función de densidad es

$$f(A|T, v) = \frac{|A|^{\frac{1}{2}(v-p-1)} e^{-\frac{1}{2}tr\left(T^{-1}A\right)}}{2^{\frac{vp}{2}} |T|^{\frac{v}{2}} \Gamma_p(v/2)}.$$

Definición 2.5.8. Si $X_1, ..., X_n$ es una muestra de vectores aleatorios, se dice que $\underline{X} = [X_1, X_2, ..., X_n]^T \in \mathbb{R}^{n\times p}$ es una matriz de datos.
Decimos que una matriz de datos $\underline{X} = [X_1, X_2, ..., X_n]^T \in \mathbb{R}^{n\times p}$ se distribuye $\underline{N}_{n\times p}(\mu, \Sigma)$ si $X_i \sim N_p(\mu, \Sigma)$ para $i = 1, 2, ..., n$.

Teorema 2.5.3. *Sea $\underline{X}$ una matriz de datos, si $\underline{X} \sim \underline{N}_{n\times p}(0, \Sigma)$, entonces $\underline{Y} = \underline{X}^T \underline{X}$ tiene una distribución Wishart con n grados de libertad y matriz de dispersión Σ.*

Teorema 2.5.4 (Estimación por Máxima Verosimilitud)**.** *Si $X_1, ..., X_n$ es una muestra de vectores aleatorios cuya distribución es $N_p(\mu, \Sigma)$ entonces los estimadores máximo verosímiles para μ y Σ respectivamente son $\hat{\mu} = \bar{X} = 1/n \sum_{i=1}^{n} X_i$ y $\hat{\Sigma} = 1/n \sum_{i=1}^{n} (X_i - \bar{X})(X_i - \bar{X})^T$.*

Los siguientes teoremas que se refieren a las estimaciones bayesiana de parámetros para las distribuciones normal y Dirichlet, se presentarán con su demostración, pues el modelo enunciado en el documento utiliza fundamentalmente estos resultados.

Lema 2.5.5.

$$\sum_{i=1}^{n} (X_i - \mu)^T \Lambda (X_i - \mu) = \sum_{i=1}^{n} (X_i - \bar{X})^T \Lambda (X_i - \bar{X}) + n(\bar{X} - \mu)$$

Demostración.

$$\begin{aligned}
\sum_{i=1}^{n} (X_i - \mu)^T \Lambda (X_i - \mu) &= \sum_{i=1}^{n} ([X_i - \bar{X}] + [\bar{X} - \mu])^T \Lambda ([X_i - \bar{X}] + [\bar{X} - \mu]) \\
&= \sum_{i=1}^{n} (X_i - \bar{X})^T \Lambda (X_i - \bar{X}) - \sum_{i=1}^{n} (X_i - \bar{X})^T \Lambda (\bar{X} - \mu) \\
&\quad - \sum_{i=1}^{n} (\bar{X} - \mu)^T \Lambda (X_i - \bar{X}) + \sum_{i=1}^{n} (\bar{X} - \mu)^T \Lambda (\bar{X} - \mu) \\
&= \sum_{i=1}^{n} (X_i - \bar{X})^T \Lambda (X_i - \bar{X}) + n(\bar{X} - \mu)^T \Lambda (\bar{X} - \mu)
\end{aligned}$$

□

Teorema 2.5.6. *Sea $\underline{X}$ una matriz de datos, tal que $\underline{X} \sim \underline{N}_{n \times p}(\mu, (\Lambda)^{-1})$. Si $\mu \sim N_p\left(\mu_0, (\kappa\Lambda)^{-1}\right)$ y $\Lambda \sim W(v, M)$, entonces las distribuciones posteriores son:*

$$\begin{aligned}
\Lambda \mid v, M, \underline{X} &\sim Wishart(v + n, M_n) \\
\mu \mid \mu_0, \kappa, \Lambda, \underline{X} &\sim Normal\left(\mu_n, (k\Lambda)^{-1}\right)
\end{aligned}$$

donde $\mu_n = \frac{n\bar{X} + \kappa\mu_0}{n + \kappa}$ y $M_n = \left(\sum_{i=1}^{n} (X_i - \bar{X})(X_i - \bar{X})^T + (\frac{n\kappa}{n+\kappa})(\bar{X} - \mu_0)(\bar{X} - \mu_0)^T + M^{-1}\right)^{-1}$.

Demostración del Teorema 2.5.6. Considérese a la función de verosimilitud, que por el lema anterior se puede reescribir de la siguiente manera:

$$\begin{aligned}
L(\mu, \Lambda) &= \prod_{i=1}^{n} |\Lambda|^{1/2} (2\pi)^{-p/2} e^{\{-(X_i - \mu)^T \Lambda (X_i - \mu)/2\}} \\
&= |\Lambda|^{n/2} (2\pi)^{-np/2} e^{\{-\frac{1}{2} \sum_{i=1}^{n} (X_i - \mu)^T \Lambda (X_i - \mu)\}} \\
&= |\Lambda|^{n/2} (2\pi)^{-np/2} e^{\{-\frac{1}{2} \sum_{i=1}^{n} (X_i - \bar{X})^T \Lambda (X_i - \bar{X}) - \frac{n}{2} (\bar{X} - \mu)^T \Lambda (\bar{X} - \mu)\}}
\end{aligned}$$

De ahí que la función de densidad conjunta de μ y Λ pueda escribirse de la siguiente manera:

$$\begin{aligned}
p(\mu,\Lambda|\mu_0,\kappa,v,M,\underline{X}) &\propto L(\mu,\Lambda)\cdot p(\mu,\Lambda|\mu_0,\kappa,v,M)\\
&= L(\mu,\Lambda)\cdot p(\mu|\Lambda,\mu_0,\kappa)\cdot p(\Lambda \mid v,M)\\
&\propto |\Lambda|^{\frac{v+n-p}{2}}\cdot exp\{-\frac{1}{2}\sum_{i=1}^{n}(X_i-\bar{X})^T\Lambda(X_i-\bar{X})-\\
&\quad \frac{n}{2}(\bar{X}-\mu)^T\Lambda(\bar{X}-\mu)-\frac{\kappa}{2}(\mu-\mu_0)^T\Lambda(\mu-\mu_0)\}exp\{-\frac{1}{2}tr(M^{-1}\Lambda)\}\\
&\propto |\Lambda|^{\frac{v+n-p}{2}}\cdot exp\{-\frac{1}{2}\sum_{i=1}^{n}(X_i-\bar{X})^T\Lambda(X_i-\bar{X})-\frac{n+\kappa}{2}\mu^T\Lambda\mu+\\
&\quad \mu^T\Lambda(n\bar{X}+\kappa\mu_0)-\frac{n}{2}\bar{X}^T\Lambda\bar{X}-\frac{\kappa}{2}\mu_0^T\Lambda\mu_0\}\cdot exp\{-\frac{1}{2}tr(M^{-1}\Lambda)\}
\end{aligned}$$

Si ahora $\mu_n := \frac{n\bar{X}+\kappa\mu_0}{n+\kappa}$, entonces

$$\begin{aligned}
p(\mu,\Lambda \mid \mu_0,\kappa,v,M,\underline{X}) &\propto |\Lambda|^{\frac{v+n-p}{2}}\cdot exp\{-\frac{1}{2}\sum_{i=1}^{n}(X_i-\bar{X})^T\Lambda(X_i-\bar{X})-\frac{n+\kappa}{2}\mu^T\Lambda\mu+\\
&\quad \mu^T(n+\kappa)\Lambda\mu_n-\mu_n^T(n+\kappa)\Lambda\mu_n+\mu_n^T(n+\kappa)\Lambda\mu_n-\\
&\quad \frac{n}{2}\bar{X}^T\Lambda\bar{X}-\frac{\kappa}{2}\mu_0^T\Lambda\mu_0\}\cdot exp\{-\frac{1}{2}tr(M^{-1}\Lambda)\}\\
&\propto |\Lambda|^{\frac{v+n-p}{2}}\cdot exp\{-\frac{1}{2}\sum_{i=1}^{n}(X_i-\bar{X})^T\Lambda(X_i-\bar{X})-\frac{1}{2}(\mu-\mu_n)^T(n+\kappa)\Lambda(\mu-\mu_n)+\\
&\quad \mu_n^T(n+\kappa)\Lambda\mu_n-\frac{n}{2}\bar{X}^T\Lambda\bar{X}-\frac{\kappa}{2}\mu_0^T\Lambda\mu_0\}\cdot exp\{-\frac{1}{2}tr(M^{-1}\Lambda)\}
\end{aligned}$$

Ahora desarrollando $\frac{n}{2}\bar{X}^T\Lambda\bar{X}+\frac{\kappa}{2}\mu_0^T\Lambda\mu_0-\mu_n^T(n+\kappa)\Lambda\mu_n$

$$\begin{aligned}
\frac{n}{2}\bar{X}^T\Lambda\bar{X}+\frac{\kappa}{2}\mu_0^T\Lambda\mu_0-\mu_n^T(n+\kappa)\Lambda\mu_n &= \frac{n}{2}\bar{X}^T\Lambda\bar{X}+\frac{\kappa}{2}\mu_0^T\Lambda\mu_0-(\frac{n\bar{X}+\kappa\mu_0}{n+\kappa})^T(n+\kappa)\Lambda(\frac{n\bar{X}+\kappa\mu_0}{n+\kappa})\\
&= \frac{n}{2}\bar{X}^T\Lambda\bar{X}+\frac{\kappa}{2}\mu_0^T\Lambda\mu_0-[\frac{(n\bar{X}+\kappa\mu_0)^T\Lambda(n\bar{X}+\kappa\mu_0)}{n+\kappa}]\\
&= \frac{1}{2}\cdot\{\frac{n(n+\kappa)\bar{X}^T\Lambda\bar{X}+\kappa^2(n+\kappa)\mu_0^T\Lambda\mu_0}{n+\kappa}\\
&\quad -[\frac{(n\bar{X}+\kappa\mu_0)^T\Lambda(n\bar{X}+\kappa\mu_0)}{n+\kappa}]\}\\
&= \frac{1}{2}\cdot\{\frac{n^2\bar{X}^T\Lambda\bar{X}+n\kappa\bar{X}^T\Lambda\bar{X}+\kappa n\mu_0^T\Lambda\mu_0+\kappa^2\mu^T\Lambda\mu}{n+\kappa}\\
&\quad -\frac{n^2\bar{X}^T\Lambda\bar{X}-2n\kappa\bar{X}^T\Lambda\mu+\kappa^2\mu^T\Lambda\mu}{n+\kappa}\}\\
&= \frac{1}{2}\left(\frac{n\kappa}{n+\kappa}\right)\left(\bar{X}^T\Lambda\bar{X}-\bar{X}^T\Lambda\mu_0+\mu_0^T\Lambda\mu_0\right)\\
&= \frac{1}{2}\left(\frac{n\kappa}{n+\kappa}\right)(\bar{X}-\mu_0)^T\Lambda(\bar{X}-\mu_0)
\end{aligned}$$

Y sustituyendo en $p(\mu,\Lambda \mid \mu,\kappa,v,M,\underline{X})$ se tiene que

$$p(\mu,\Lambda \mid \mu,\kappa,v,M,\underline{X}) \propto |\Lambda|^{\frac{v+n-p}{2}} \cdot exp\{-\frac{1}{2}\sum_{i=1}^{n}(X_i-\bar{X})^T\Lambda(X_i-\bar{X}) - \frac{1}{2}(\mu-\mu_n)^T(n+\kappa)\Lambda(\mu-\mu_n)+$$
$$-\frac{1}{2}(\frac{n\kappa}{n+\kappa})(\bar{X}-\mu)^T\Lambda(\bar{X}-\mu)\} \cdot exp\{-\frac{1}{2}tr(M^{-1}\Lambda)\}$$
$$\propto |\Lambda|^{\frac{v+n-p}{2}} \cdot exp\{-\frac{1}{2}(\mu-\mu_n)^T(n+\kappa)\Lambda(\mu-\mu_n)\}\cdot$$
$$exp\{-\frac{1}{2}[\sum_{i=1}^{n}(X_i-\bar{X})^T\Lambda(X_i-\bar{X}) + (\frac{n\kappa}{n+\kappa})(\bar{X}-\mu_0)^T\Lambda(\bar{X}-\mu) + tr(M^{-1}\Lambda)]\}$$
$$\propto |\Lambda|^{\frac{v+n-p}{2}} \cdot exp\{-\frac{1}{2}(\mu-\mu_n)^T(n+\kappa)\Lambda(\mu-\mu_n)\}\cdot$$
$$exp\left\{-\frac{1}{2}\left[\sum_{i=1}^{n} tr\left((X_i-\bar{X})(X_i-\bar{X})^T\Lambda\right) + (\frac{n\kappa}{n+\kappa})tr\left((\bar{X}-\mu_0)(\bar{X}-\mu_0)^T\Lambda\right) + tr\left(M^{-1}\Lambda\right)\right]\right\}$$
$$\propto |\Lambda|^{\frac{v+n-p}{2}} \cdot exp\{-\frac{1}{2}(\mu-\mu_n)^T(n+\kappa)\Lambda(\mu-\mu_n)\}\cdot$$
$$exp\left\{-\frac{1}{2}\left[tr\left(\left(\sum_{i=1}^{n}(X_i-\bar{X})(X_i-\bar{X})^T + (\frac{n\kappa}{n+\kappa})(\bar{X}-\mu_0)(\bar{X}-\mu_0)^T + M^{-1}\right)\Lambda\right)\right]\right\}$$

Ahora de igual manera que con μ_n, se define a M_n como:
$M_n := \left(\sum_{i=1}^{n}(X_i-\bar{X})(X_i-\bar{X})^T + (\frac{n\kappa}{n+\kappa})(\bar{X}-\mu_0)(\bar{X}-\mu_0)^T + M^{-1}\right)^{-1}$,
y finalmente

$$p(\mu,\Lambda \mid \mu,\kappa,v,M,\underline{X}) \propto |\Lambda|^{\frac{1}{2}} \cdot exp\{-\frac{1}{2}(\mu-\mu_n)^T(n+\kappa)\Lambda(\mu-\mu_n)\}\cdot$$
$$|\Lambda|^{\frac{v+n-p-1}{2}} exp\left\{-\frac{1}{2}\, tr\left(M_n^T\Lambda\right)\right\}$$
$$\propto N\left(\mu \mid \mu_n\, , ((n+\kappa)\Lambda)^{-1}\right) \cdot W\left(\Lambda \mid v+n\, , M_n\right)$$

Por lo tanto, las distribuciones posteriores quedan *Normal* y *Wishart* nuevamente.

$$\Lambda \mid v\, , M\, , \underline{X} \sim Wishart\left(v+n, M_n\right)$$
$$\mu \mid \mu\, , \kappa\, , \Lambda\, , \underline{X} \sim Normal\left(\mu_n, (k\Lambda)^{-1}\right))$$

□

Teorema 2.5.7. *Sea $X = (X_1,...,X_\kappa)$ un vector aleatorio que se distribuye multinomial de parámetro $p = (p_1,...,p_\kappa)$, y sea la distribución a priori de p Dirichlet con parámetro $\alpha = (\alpha_1,...,\alpha_p)$. Entonces la función de densidad posterior de p es Dirichlet de parámetro $\alpha_n = (\alpha_1+x_1,...,\alpha_p+x_\kappa)$.*

Demostración. Se sabe de p que

$$f(p \mid \alpha) = \frac{\Gamma(\sum_{i=1}^{K}\alpha_i)}{\prod_{i=1}^{K}\Gamma(\alpha_i)} \cdot \prod_{i=1}^{K} p_i^{\alpha_i-1}$$
$$L(p) = f(x_1,...,x_k|p_1,...,p_k)$$
$$= \frac{\Gamma(n+1)}{\prod_{i=1}^{K}\Gamma(x_i+1)} \prod_{i=1}^{K} p^{x_i} \cdot \mathbb{1}_{(\sum_{i=1}^{K} x_i = n)}$$

Entonces

$$\begin{aligned} f(p|\alpha\ ,X) &\propto f(p|\alpha)\cdot L(p) \\ &\propto \left(\frac{\Gamma(\sum_{i=1}^{\kappa}\alpha_i)}{\prod_{i=1}^{\kappa}\Gamma(\alpha_i)}\cdot\prod_{i=1}^{\kappa}p_i^{\alpha_i-1}\right)\left(\frac{\Gamma(n+1)}{\prod_{i=1}^{\kappa}\Gamma(x_i+1)}\prod_{i=1}^{\kappa}p^{x_i}\right) \\ &\propto \prod_{i=1}^{\kappa}p_i^{\alpha_i-1}\cdot\prod_{i=1}^{\kappa}p^{x_i} \\ &\propto \prod_{i=1}^{\kappa}p_i^{\alpha_i+x_i-1} \end{aligned}$$

Que es el núcleo de la densidad de una variable aleatoria con distribución Dirichlet de parámetros $\alpha_1+x_1,\alpha_2+x_2,...,\alpha_\kappa+x_\kappa$. □

2.6. Criterios de Información para Evaluación de Modelos

A fin de evaluar y comparar el comportamiento de los modelos respecto a los datos se consideran diversos criterios. El más comunmente utilizado es el Criterio de Información de Akaike (AIC) , que fue propuesto por Hirotugu Akaike, estableciendo una medida con base en maximizar la entropía esperada del modelo, es decir, maximiza la información esperada. A partir de este criterio se formuló el Criterio de Bayesiano de Información (BIC) que, como su nombre lo dice, es una extensión bayesiana del criterio anterior.

Las expresiones dadas para estos criterios son como siguen:

$$AIC = 2\kappa - 2ln\,(L)\,, BIC = ln(n)\kappa - 2ln(L),$$

donde L se refiere a la verosimilitud de los datos bajo el modelo a evaluar, κ el número de parámetros a ser estimados y n el tamaño de la muestra.

Estos métodos de evaluación y comparación de modelos han ganado popularidad recientemente debido a que pueden ser utilizados en situaciones de selección de modelos cada vez mas complicadas. Son frecuentes en modelos lineales generalizados, series de tiempo, estadística espacial, redes bayesianas, entre otras aplicaciones.

2.7. Cadenas de Markov y Muestreo de Gibbs

En esta sección se introducen los conceptos de proceso estocástico,cadenas de Markov y distribución estacionaria para terminar enunciando los principios en los que se sustenta la teoría del muestreo de Gibbs.

2.7.1. Cadenas de Markov a tiempo finito y con espacio de estados no numerable

Definición 2.7.1. Un proceso estocástico $X = \{X(t) : t \in T\}$ es una colección de variables aleatorias, es decir, para cada t en el conjunto de índices T, $X(t)$ es una variable aleatoria. Es común que se interprete a t como el tiempo y a $X(t)$ como el estado del proceso al tiempo t. Si el conjunto de índices T es un conjunto numerable, entonces se dice que X es un proceso estocástico a tiempo discreto, si T es no numerable, se dice que es es un proceso estocástico a tiempo continuo.

Definición 2.7.2. Sea $X = \{X(i) : i = 1,2,...\}$ un proceso estocástico. Se dice que es una cadena de Markov si

$$\mathbb{P}[X_{n+1} \in A_{n+1} | X_1 \in A_1, ..., X_{n-1} \in A_{n-1}, X_n \in A_n] = \mathbb{P}[X_{n+1} \in A_{n+1} | X_n \in A_n].$$

Es decir, que la probabilidad de X_{n+1} depende únicamente del estado del proceso al tiempo n. A esta propiedad se le llama propiedad de Markov para procesos estocásticos.

Definición 2.7.3. Sea $\{X_n : n = 0,1,2,...\}$una cadena de Markov, se define a la distribución estacionaria π como la función que satisface

$$\pi_j = \sum_{i \in S} \pi_i p_{ij}, \qquad \text{Si el soporte } S \text{ de las variables aleatorias } X_n \text{ es igual a } \{0,1,2,...\}$$

$$\pi(A) = \int_S \pi(dx) P(x,A) \qquad \text{Si el soporte } S \text{ de las variables aleatorias } X_n \text{ es un conjunto no numerable,}$$

donde p_{ij} es la probabilidad de pasar del estado i al estado j en S y $P(x,A)$ es la probabilidad de llegar a un estado en el subconjunto medible de S, A dado que se inicia en el punto $x \in S$.

Definición 2.7.4. Se dice que una cadena de Markov es irreducible si para cualquier subconjunto medible $A \subset S$ y medida mayor o igual a cero, se tiene que $\mathbb{P}_x[\tau_A < \infty] > 0$ para cualquier $x \in S$, donde $\tau_A = inf\{n \geq 0 : X_n \in A\}$ y $\mathbb{P}_x[B] = \mathbb{P}[B|X_0 = x]$, es decir si la cadena tiene probabilidad positiva de tomar cualquier valor dentro de cualquier subconjutno medible A de S.

Teorema 2.7.1. *Si una cadena de markov a tiempo discreto $\{X_n : n \geq 1\}$ es irreducible, aperiodica, tiene una distribución estacionaria $\pi(\cdot)$ y $P^n(x,A)$ es la probabilidad de llegar a un estado en el subconjunto medible de S, A dado que se inicia en el punto $x \in S$ en n pasos, entonces*

$$\lim_{n \to \infty} \sup |P^n(x,A) - \pi(A)| = 0,$$

es decir, la cadena de Markov converge a su distribución estacionaria.

2.7.2. Muestreo de Gibbs

Supóngase una función de densidad conjunta de la forma $f\left(x, y_1, ..., y_q\right)$, y se busca generar una muestra aleatoria de la variable X. En muchos casos las integrales para obtener $f(x)$ son muy difíciles de resolver incluso a través de métodos numéricos y es en estas situaciones donde el muestreo de Gibbs es útil.

Se asume que es posible sin mayor complicación generar variables aleatorias de las densidades

$$f\left(x|y_1, ..., y_q\right) \text{ y } f\left(y_i|x, y_1, ..., y_{i-1}, y_{i+1}, ..., y_q\right).$$

Se inicia el proceso fijando valores para $\left(Y_1, ..., Y_q\right) = \left(y_1, ..., y_q\right)$. A partir de estos valores y usando que podemos generar x de $f\left(x|y_1, ..., y_q\right)$ se obtiene $x^{(1)}$, luego se descarta el valor de Y_1 y se genera otro a partir de el resto de las variables Y_i y del nuevo valor de X, luego el proceso se repite para cada una de las Y_i hasta sustituir la totalidad por nuevos valores, al concluir el proceso se dice que ha pasado la primera iteración.

Para el resto de las iteraciones el proceso es el mismo, pero en lugar de usar valores iniciales para las variables, se usan los resultantes de la iteración anterior. El proceso se repite una cantidad de iteraciones grande, y si lo es suficientemente, de la mano de las hipótesis del teorema 2.7.1 se puede garantizar la convergencia de X a una variable aleatoria generada por $f(x)$.

El proceso completo del muestreo de Gibbs se puede escibir como se muestra en el Algoritmo 1 usando la notación estándar:

Algoritmo 1 Muestreo de Gibbs

Inicializar $Y_2^{(0)} = y_2^*,\ ...,\ Y_q^{(0)} = y_q^*,\ X^{(0)} = x^*$
Para iteración $l \leftarrow 1, 2, ..., L$ **hacer**
    Generar $Y_1^{(l)} \sim p\left(y_1 | Y_2^{(l-1)}, ...,\ Y_q^{(l-1)},\ X^{(l-1)}\right)$
    Generar $Y_2^{(l)} \sim p\left(y_2 | Y_3^{(l-1)}, ...,\ Y_q^{(l-1)},\ X^{(l-1)}\right)$
    $\vdots$
    Generar $Y_q^{(l)} \sim p\left(y_q | Y_2^{(l)} =, ...,\ Y_{q-1}^{(l)},\ X^{(l-1)}\right)$
    Generar $X^{(l)} \sim p\left(x | Y^{(l)},\ Y_2^{(l)} =, ...,\ Y_q^{(l)}\right)$
Fin Para
Devuelve $X^{(L)}$

Como ilustración del funcionamiento del muestreo de Gibbs tómese el ejemplo que aparece en [3, p. 542] en donde se buscan simular dos variables aleatorias normales correlacionadas X y Y con medias μ_1 y μ_2, varianzas σ_1^2 y σ_2^2, y una correlación ρ de las que, hipotéticamente, se supone que no se puede generar una observación marginal de X, sin embargo se pueden generar realizaciones de cada una de las variables condicionadas a la otra. Así se seguirían los siguientes pasos si se busca generar un valor que se distribuya como X a través del muestreo de Gibbs:

1. Fijar $X^{(0)} = x^*$
2. Determinar un número de iteraciones a realizar, p. ejem. 4
3. Inicializar $l = 1$
4. Generar $Y^{(l)} \sim N\left(\mu_2 + \frac{\sigma_2}{\sigma_1}\rho\left(x^{(l-1)} - \mu_1\right), \left(1-\rho^2\right)\sigma_2^2\right)$
5. Generar $x^{(l)} \sim N\left(\mu_1 + \frac{\sigma_1}{\sigma_2}\rho\left(y^{(l)} - \mu_2\right), \left(1-\rho^2\right)\sigma_1^2\right)$
6. Si $l < 4$ redefinir a l como $l+1$ y regresar al paso 4, en caso contrario continuar al paso 7
7. Devuelve $X_1 = x_1^{(4)}$

En la figura 2.3 se muestra una realización del ejemplo mencionado, con un total de cuatro iteraciones, con un valor inicial igual a x_0, seguido por los valores y_1, x_1, y_2, x_2, y_3, x_3, y_4 y que genera un valor final $X = x_4$, donde la elipse dibujada representa un nivel de confianza para las variables conjuntas e ilustra la correlación entre ambas tomando como base una distribución normal multivariada con parámetros $\mu_1, \mu_2, \sigma_1^2, \sigma_2^2$ y ρ.

Para profundizar en el muestreo de Gibbs, sus aplicaciones y su sustento teórico se recomienda consultar en [8], [19], [4] y [3].

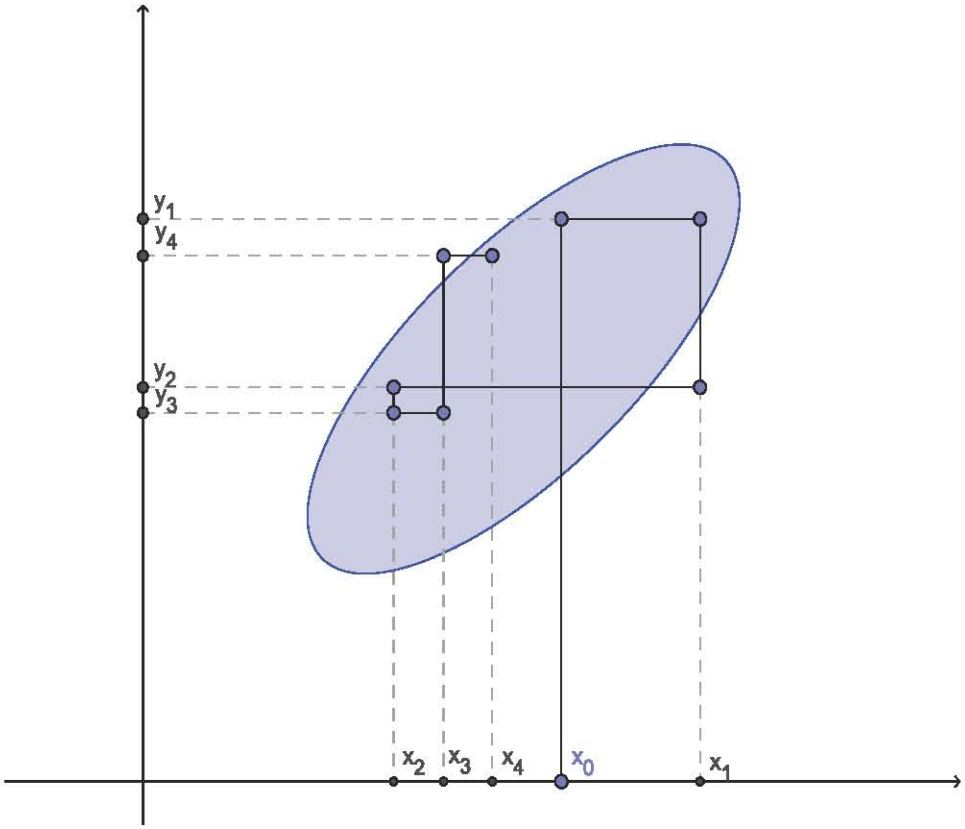


Figura 2.3: Ilustración del muestreo de Gibbs

Capítulo 3

Descripción del Modelo

El modelo se basa en el que fue desarrollado por Fuentes y Walker en [11], donde se presenta el caso univariado del razonamiento. A continuación se enuncia y desarrolla brevemente con variables unidimensionales a fin de introducir el panorama al momento del paso a dimensiones superiores.

3.1. El caso unidimensional

Sea y una observación por clasificar dentro de algún *cluster*. Se define la regla para determinar el *cluster* de la siguiente manera: Si dada la sucesión $\left(\theta_j, u_j\right)_{j=1}^{\infty}$, donde $u_j > 0$ y $\theta_j \in \mathbb{R}$, la observación y satisface que $|y - \theta_j| < \sqrt{u_j}$, entonces la observación y se asigna al *cluster* j. La sucesión $\left(\theta_j, u_j\right)_{j=1}^{\infty}$ tendrá asociada una distribución a priori de modo que garantice las siguientes dos condiciones, esto dirigido a enfatizar el objetivo principal que es agrupar los datos:

- Los intervalos definidos como $A_j := \left(\theta_j - \sqrt{u_j}, \theta_j + \sqrt{u_j}\right)$ no se traslapan.
- La observación y debe estar dentro de alguno de los intervalos A_j.

El punto de partida para la presentación de una distribución a priori para $\left(\theta_j, u_j\right)_{j=1}^{\infty}$ será la siguiente afirmación, que se basa en la densidad de probabilidad de y dados los parámetros θ, U y W:

$$\begin{aligned} p\left(y \mid w, u, \theta\right) &= \sum_{j=1}^{\infty} w_j \cdot Unif\left(\theta_j - \sqrt{u_j}, \theta_j + \sqrt{u_j}\right) \\ &= \sum_{j=1}^{\infty} w_j \cdot Unif\left(A_j\right) \end{aligned}$$

Donde $w = \left(w_j\right)_{j=1}^{\infty}$, $u = \left(u_j\right)_{j=1}^{\infty}$, $\theta = \left(\theta_j\right)_{j=1}^{\infty}$ y $Unif\left(\theta_j - \sqrt{u_j}, \theta_j + \sqrt{u_j}\right)$ denota a la densidad de una variable aleatoria continua uniformemente distribuida en el intervalo $\left(\theta_j - \sqrt{u_j}, \theta_j + \sqrt{u_j}\right)$.

Este punto debido a que la probabilidad de que la observación y sea clasificada en el j-ésimo cluster, es decir, que $y \in C_j$ dados los parámetros w, u y θ, puede ser escrita como:

$$\begin{aligned}
\mathbb{P}\left[y \in C_j | w,u,\theta\right] &= \int_{A_j} p\left(y \in C_j \mid w,u,\theta\right) dy \\
&= \int_{A_j} \sum_{k=1}^{\infty} w_k \cdot Unif\left(A_k\right) dy \\
&= \sum_{k=1}^{\infty} \int_{A_j} w_k \cdot Unif\left(A_k\right) dy \\
&= \int_{A_j} w_j \cdot Unif\left(A_j\right) dy \\
&= w_j \int_{A_j} Unif\left(A_j\right) dy \\
&= w_j
\end{aligned}$$

Que es una propiedad deseable en el modelo dado que coincide con los modelos usuales de mezclas que consideran a w como los pesos asociados a cada cluster.

Previo a pensar en una función de distribución conjunta para u y θ, primero consideraremos una función de distribución para u_j condicionada a una θ_j dada por

$$f\left(u \mid \theta\right) \propto \left\{ \prod_{j=1}^{\infty} Gamma\left(u \mid 3/2, \lambda_j/2\right) \right\} \cdot \mathbb{1}\left(u \ \in \ D\right)$$

con $Gamma\left(u \mid 3/2, \lambda_j\right)$ la densidad de una variable aleatoria gamma con parámetros $3/2$ y $\lambda_j/2$, y D el conjunto de u's que satisfacen las restricciones previamente descritas para los intervalos A_j, es decir, $D := \{(u_1, u_2, ...) : A_i \bigcap A_j = \emptyset \wedge \exists\ k \in \mathbb{N},\ \text{tal que}\ \ y \in A_k\}$ y A_j como está definido anteriormente.

Antes de continuar se demostrará un lema que liga la función anterior con la mezcla de distribuciones normales.

Lema 3.1.1. *Sean Y y U variables aleatorias tales que $Y \mid (U = u) \sim Unif\left(\theta - \sqrt{u}, \theta + \sqrt{u}\right)$ y $U|\lambda \sim Gamma\left(3/2, \lambda/2\right)$, entonces la distribución marginal de Y está dada por $f\left(y \mid \theta, \lambda\right) = \sqrt{\frac{\lambda}{2\pi}} \exp\left\{-\frac{\lambda^2 (y-\theta)^2}{2}\right\}$, es decir,*

$$y \mid \theta, \lambda \sim Normal(\theta, 1/\lambda)$$

Demostración.

$$\begin{aligned}
f\left(y \mid \theta, \lambda\right) &= \int_{\mathbb{R}^+} f\left(y \mid u, \theta, \lambda\right) \cdot f\left(u \mid \lambda\right) du \\
&= \int_{\mathbb{R}^+} \frac{1\{y \in (\theta - \sqrt{u}, \theta - \sqrt{u})\}}{2\sqrt{u}} \cdot \frac{\left(\lambda/2\right)^{3/2}}{\Gamma(3/2)} \sqrt{u}\, e^{-\lambda u/2} du
\end{aligned}$$

Pero

$$\begin{aligned}
y \in \left(\theta - \sqrt{u}, \theta - \sqrt{u}\right) &\Leftrightarrow \theta - \sqrt{u} < y < \theta + \sqrt{u} \\
&\Leftrightarrow -\sqrt{u} < y - \theta < \sqrt{u} \\
&\Leftrightarrow |y - \theta| < \sqrt{u} \\
&\Leftrightarrow (y - \theta)^2 < u \\
\therefore 1_{\{y \in (\theta - \sqrt{u}, \theta - \sqrt{u})\}} &= 1_{\{u > (y-\theta)^2\}}
\end{aligned}$$

Por lo que $f(y \mid \theta, \lambda)$ se puede escribir como:

$$\begin{aligned} f(y \mid \theta, \lambda) &= \int_{(y-\theta)^2}^{\infty} \frac{1}{2\sqrt{u}} \cdot \frac{(\lambda/2)^{3/2}}{\Gamma(3/2)} \sqrt{u}\, e^{-\lambda u/2} du \\ &= \frac{\lambda^{1/2}}{2^{3/2}\,\Gamma(3/2)} \int_{(y-\theta)^2}^{\infty} (\lambda/2) e^{-(\lambda/2)u} du \\ &= \sqrt{\frac{\lambda}{2\pi}} e^{-\frac{\lambda(y-\theta)^2}{2}} \end{aligned}$$

Que es la densidad de una variable aleatoria con distribución normal.

□

Ahora considérese nuevamente a la densidad de u

$$f(u \mid \theta) \propto \left\{ \prod_{j=1}^{\infty} Gamma\left(u \mid 3/2, \lambda_j/2\right) \right\} \cdot \mathbb{1}(u \in D),$$

para concretar la relación entre este modelo y una mezcla de distribuciones normales se tomará a $\hat{f}(u \mid \theta)$ como la función de densidad de U omitiendo las restricciones definidas en D, es decir,

$$\hat{f}(u \mid \theta) := \prod_{j=1}^{\infty} Gamma\left(u \mid 3/2, \lambda_j/2\right).$$

Partiendo ahora de $p(y \mid w, \theta, \lambda)$ se tiene

$$p(y \mid w, \theta, \lambda) = \int_{\mathbb{R}} \int_{\mathbb{R}} \cdots p(y \mid w, u_1, u_2, \ldots, \theta, \lambda) \cdot f(u_1, u_2, \cdots \mid \lambda)\, du_1 du_2 \ldots$$

luego intercambiando $f(u \mid \theta)$ por $\hat{f}(u \mid \theta)$, y haciendo uso del lema anterior, se puede hacer el desarrollo siguiente:

$$\begin{aligned} p(y \mid w, \theta, \lambda) &= \int_{\mathbb{R}} \int_{\mathbb{R}} \cdots p(y \mid w, u_1, u_2, \ldots, \theta, \lambda) \cdot \hat{f}(u_1, u_2, \cdots \mid \lambda)\, du_1 du_2 \ldots \\ &= \int_{\mathbb{R}} \int_{\mathbb{R}} \cdots \left[\sum_{j=1}^{\infty} w_j Unif\left(y \mid \theta_j, \lambda_j, u_j\right) \right] \cdot \left[\prod_{i=1}^{\infty} Gamma\left(u_i \mid 3/2, \lambda_j/2\right) \right] du_1 du_2 \ldots \\ &= \int_{\mathbb{R}} \int_{\mathbb{R}} \cdots \left[\sum_{j=1}^{\infty} w_j Unif\left(y \mid \theta_j, \lambda_j, u_j\right) \cdot \prod_{i=1}^{\infty} Gamma\left(u_i \mid 3/2, \lambda_j/2\right) \right] du_1 du_2 \ldots \\ &= \sum_{j=1}^{\infty} \left\{ \int_{\mathbb{R}} \int_{\mathbb{R}} \cdots \left[w_j Unif\left(y \mid \theta_j, \lambda_j, u_j\right) \cdot \prod_{i=1}^{\infty} Gamma\left(u_i \mid 3/2, \lambda_j/2\right) \right] du_1 du_2 \ldots \right\} \\ &= \sum_{j=1}^{\infty} \left\{ \left(\int_{\mathbb{R}} Gamma(u_1 \mid 3/2, \lambda_1/2)\, du_1 \right) \left(\int_{\mathbb{R}} Gamma(u_2 \mid 3/2, \lambda_2/2)\, du_2 \right) \cdots \right. \\ &\quad \left. \left(w_j \int_{\mathbb{R}} Unif\left(y \mid \lambda_j, u_j\right) \cdot Gamma\left(u_j \mid 3/2, \lambda_j/2\right) du_j \right) \cdots \right\} \\ &= \sum_{j=1}^{\infty} \left\{ (1) \cdot (1) \cdots \left(w_j \cdot Normal\left(y \mid \theta_j, \lambda_j\right) \right) \cdots \right\} \\ &= \sum_{j=1}^{\infty} w_j \cdot Normal\left(y \mid \theta_j, \lambda_j\right) \end{aligned}$$

y así se demuestra que el modelo es similar al que se tiene con una mezcla de normales, de manera que coinciden cuando se omiten algunas restricciones.

Como última observación se explicará el procedimiento en el caso general de la muestra $y_1, y_2, ..., y_n$, en lugar de considerar a una sola observación y. El poceso es bastante similar, salvo que ahora se considerará una sucesión $\left(u_{ij}\right)_{j=1}^{\infty}$, es decir, $y_i \in C_j$ si y solo si $|y_i - \theta_j| < \sqrt{u_{ij}}$ con $\left(u_{ij}\right)_{j=1}^{\infty}$ satisfaciendo todas las restricciones que en el caso anterior se habían definido para $\left(u_j\right)_{j=1}^{\infty}$. Así ahora la distribución para $u = \left(\left(u_{1j}, u_{2j}, \cdots, u_{nj}\right)\right)_{j=1}^{\infty}$ está dada por

$$f\left(u \mid \theta, \lambda\right) \propto \left\{ \prod_{i=1}^{n} \prod_{j=1}^{\infty} Gamma\left(u_{ij} \mid 3/2, \lambda_j/2\right) \right\} \cdot \mathbb{1}\left(u \in \tilde{\mathrm{D}}\right).$$

Donde $\tilde{\mathrm{D}}$ denota al conjunto de u's que satisfacen las restricciones antes enunciadas.

Esto hace posible que al final del proceso se tenga para cada j una submuestra $y_{j1}, y_{j2}, ..., y_{jn_j}$ de tamaño n_j entre cero y n que facilita el uso de los supuestos distribucionales, en especial permitirá usar herramienta bayesiana para actualizar los parámetros θ_j, y λ_j; una vez que estos estén actualizados se origina un proceso iterativo realizando las mismas acciones con nuevos parámetros y que se detendrá una vez que se considere que el número de iteraciones ha sido suficiente para perder la dependencia a los θ_j, y λ_j iniciales, siguiendo la estructura de un muestreo de Gibss.

Para mayor detalle sobre este algoritmo consultar [11], donde además se implementa para conjuntos de datos simulados, así como para conjuntos de datos reales.

3.2. Supuestos del modelo multivariado

Ahora se desarrollará un razonamiento para dimensiones superiores utilizando la mayor parte de la estructura de la sección anterior.

Se supone que el conjunto a clasificar está formado por observaciones que tienen la forma de un punto $y \in \mathbb{R}^p$, cuyos valores pueden ser vistos como resultados de experimentos normales con medias y varianzas a estimar.

El enfoque con el que se trabaja el problema en este documento, y que permite cierta bondad en los resultados matemáticos, es el de una mezcla normal, es decir, que la densidad de la muestra es de la forma

$$f\left(y_i \mid w_1, \theta_1, \Lambda_1, w_2, \theta_2, \Lambda_2, ...\right) = \sum_{j=1}^{\infty} w_j \cdot N\left(y_i \mid \theta_j, \Lambda_j\right)$$

donde μ_j y Σ_j se refieren a la media y a la varianza de cada una de las normales que aparecen en la mezcla, y $\sum_{j=1}^{\infty} w_j = 1$ son los pesos de cada componente de la mezcla. A esta distribución se llegará mediante los desarrollos de la siguiente sección y a partir de los supuestos enunciados a continuación:

I. Se asume que el número de *clusters* está acotado con base en el argumento de que no es posible obtener información de más de *n clusters* si el conjunto de muestra sólo cuenta con n observaciones, es decir, que no se puede saber nada sobre grupos que no están representados por nunguna observación. Nótese que a pesar de que la densidad de mezcla de normales cubre el caso de un número infinito de *clusters*, también cubre el caso en que hay un número finito de estos considerando una $J \in \mathbb{N}$ tal que $w_j = 0$ si $j > J$.

II. Se considera a la función de densidad del vector aleatorio Y dados los parámetros θ, Λ, U y w como:

$$\begin{aligned} p(Y \mid w,U,\Lambda,\theta) &= \sum_{j=1}^{\infty} w_j \cdot Unif\left(\mathscr{E}_j\right) \\ &= \sum_{j=1}^{\infty} w_j \cdot Unif\left(\mathscr{E}_j\right) \end{aligned}$$

Donde $\theta = \left(\theta_j\right)_{j=1}^{\infty}$, $\Lambda = \left(\Lambda_j\right)_{j=1}^{\infty}$, $U = \left(U_j\right)_{j=1}^{\infty}$, $W = \left(w_j\right)_{j=1}^{\infty}$ y $Unif\left(\mathscr{E}_j\right)$ se refiere a la densidad de un vector aleatorio uniforme dentro del elipsoide $\mathscr{E}_j := \{y \in \mathbb{R}^p : \|\Lambda_j^{1/2}(y-\mu_j)\|^2 < u_j\}$.

III. La distribución de U está definida por la densidad

$$f(u \mid \theta,\lambda) \propto \left\{ \prod_{i=1}^{n} \prod_{j=1}^{\infty} \chi^2_{p+2}\left(u_{ij}\right) \right\} \cdot \mathbb{1}\left(u \in \tilde{\mathrm{D}}\right)$$

con $\tilde{\mathrm{D}}$ el conjunto de los posibles resultados para U que satisface que los elipsoides $\mathscr{E}_j$ no se traslapan y que la observación y necesariamente pertenece a alguno de éstos, donde $\chi^2_{p+2}(u)$ denota la densidad de una variable aleatoria ji cuadrada de $p+2$ grados de libertad y valuada en el punto $u \in \mathbb{R}^+$. De manera similar a la sección anterior, este supuesto será relajado tratando de mantener la esencia del procedimiento, suprimiendo de la densidad a la función indicadora en el conjunto $\tilde{\mathrm{D}}$, es decir, se continuará con una función

$$\hat{f}(u \mid \theta) := \prod_{j=1}^{\infty} \chi^2_{p+2}\left(u_{ij}\right).$$

El algoritmo en este documento es para ajustar un modelo de mezclas a datos multivariados, pero se ilustrarán e implementarán sólo casos de mezclas normales bivariadas.

Este enfoque tiene como fin agrupar en cada iteración una muestra de n observaciones en grupos que se tienen una distribución de probabilidad normal por sí mismos, una especie de submuestras $\{C_1, C_2, ..., C_\kappa\}$ que siguen una distribución normal con medias $\{\mu_1, \mu_2, ..., \mu_\kappa\}$ y varianzas $\{\Sigma_1, \Sigma_2, ..., \Sigma_\kappa\}$.

Es necesario echar mano de este razonamiento debido a que a pesar de que la densidad mezcla de normales tiene un modelo amigable en los cálculos y con una forma explícita, la estimación de parámetros es un problema que carece de una solución definitiva y que varía según el enfoque que se de al problema.

3.3. Clasificación

Ya con punto de partida se define, tal como en [11], una regla para clasificar la observación y. Sea z el número que indica el cluster al que pertenece la observación. Ésta se clasifica en el grupo C_j cuando la distancia de Mahalanobis entre y y θ_j es menor a una u_j dada, es decir, si $||\Sigma_j^{-1/2}\left(y-\theta_j\right)||^2 < u_j$, o bien $||\Lambda_j^{1/2}\left(y-\theta_j\right)||^2 < u_j$ con parámetros $\left(\Lambda_j, \theta_j, u_u\right)_{j=1}^{\infty}$ fijos, donde Σ_j son las matrices varianzas de las normales involucradas en la mezcla, Λ_j las matrices de precisión que serán usadas por comodidad en los cálculos, θ_j las medias y u_j números reales positivos que se calcularán más adelante. La idea es asignar una distribución a priori a cada

uno de los parámetros $\left(\Lambda_j, \theta_j, u_u\right)_{j=1}^{\infty}$, de manera que se asegure que la unión de los conjuntos $\mathscr{E}_j = \{y \in \mathbb{R}^p : \|\Lambda_j^{1/2}(y-\theta_j)\|^2 < u_j\}$ contiene a y, y que son disjuntos, esto para evitar que y pertenezca a dos o más grupos.

Definición 3.3.1. Sean $\theta \in \mathbb{R}^p$, $u \in \mathbb{R}^+$ y $\Lambda \in \mathbb{R}_{sym,+}^{p\times p}$ se denotará como $\mathscr{E}_{(\theta,\Lambda,u)}$ al elipsoide generado por

$$\mathscr{E}_{(\theta,\Lambda,u)} = \{x \in \mathbb{R}^p : \|\Lambda^{1/2}(x-\theta)\|^2 < u\}$$

esto quiere decir que, siguiendo la definición de $\mathscr{E}_{(\theta,\Lambda)}$ en el teorema 2.2.3, $\mathscr{E}_{(\theta,\Lambda,u)} = \mathscr{E}_{\left(\theta,\left(\frac{1}{u}\Lambda\right)\right)}$.

Así se inicia con la probabilidad condicional de

$$p(y|\theta,\Lambda,u,w) = \sum_{j=1}^{\infty} w_j \cdot Unif(y|\mathscr{E}_{(\theta_j,\Lambda_j,u_j)})$$

Donde $w = \left(w_j\right)_{j=1}^{\infty}$, $u = \left(u_j\right)_{j=1}^{\infty}$, $\theta = \left(\theta_j\right)_{j=1}^{\infty}$, $\Lambda = \left(\Lambda_j\right)_{j=1}^{\infty}$ y $Unif(y|\mathscr{E}_{(\theta_j,\Sigma_j,u_j)})$ se refiere a la densidad de un vector aleatorio uniforme en el conjunto elipsoidal generado por $\mathscr{E}_{(\theta_j,\Sigma_j,u_j)} = \{x \in \mathbb{R}^p : \|\Lambda^{1/2}\left(x-\theta_j\right)\|^2 < u_j\}$.

Ahora, como se hizo en el caso unidimensional, se requiere verificar que la probabilidad de que una observación y esté en el *cluster* j coincide con el peso w_j.

$$\begin{aligned}
\mathbb{P}\left[y \in C_j|w,u,\theta,\Lambda\right] &= \mathbb{P}\left[y \in \mathscr{E}_{(\theta_j,\Sigma_j,u_j)}|w,u,\theta,\Lambda\right] \\
&= \int_{\mathscr{E}_{(\theta_j,\Sigma_j,u_j)}} p(y|\theta,\Lambda,u,w)\,dy \\
&= \int_{\mathscr{E}_{(\theta_j,\Sigma_j,u_j)}} \sum_{l=1}^{\infty} w_l \cdot Unif(y|\mathscr{E}_{(\theta_l,\Lambda_l,u_l)})dy \\
&= \sum_{l=1}^{\infty} w_l \int_{\mathscr{E}_{(\theta_j,\Sigma_j,u_j)}} Unif(y|\mathscr{E}_{(\theta_l,\Lambda_l,u_l)})dy
\end{aligned}$$

Nótese que la densidad $Unif(y|\mathscr{E}_{(\theta_l,\Lambda_l,u_l)})$ es igual a cero para todo $y \notin \mathscr{E}_{(\theta_l,\Lambda_l,u_l)}$, anteriormente también se mencionó que se consideraba como hipótesis que los elipsoides $\mathscr{E}_{(\theta_l,\Lambda_l,u_l)}$ y $\mathscr{E}_{(\theta_j,\Lambda_j,u_j)}$ son disjuntos cuando $l \neq j$, de ahí que si $y_o \in \mathscr{E}_{(\theta_j,\Lambda_j,u_j)}$ y $l \neq j$, entonces $Un(y_o|\mathscr{E}_{(\theta_l,\Lambda_l,u_l)}) = 0$

Así siguiendo con el cálculo de la probabilidad de que y pertenezca al j-ésimo *cluster*

$$\begin{aligned}
\mathbb{P}\left[y \in C_j|w,u,\theta,\Lambda\right] &= \sum_{l=1}^{\infty} w_l \int_{\mathscr{E}_{(\theta_j,\Sigma_j,u_j)}} Unif(y|\mathscr{E}_{(\theta_l,\Lambda_l,u_l)})dy \\
&= \left[\sum_{l\neq j} w_l \int_{\mathscr{E}_{(\theta_j,\Sigma_j,u_j)}} Unif(y|\mathscr{E}_{(\theta_l,\Lambda_l,u_l)})dy\right] + \left[w_j \int_{\mathscr{E}_{(\theta_j,\Sigma_j,u_j)}} Unif(y|\mathscr{E}_{(\theta_j,\Lambda_j,u_j)})dy\right] \\
&= \left[\sum_{l\neq j} w_l \int_{\mathscr{E}_{(\theta_j,\Sigma_j,u_j)}} 0\,dy\right] + \left[w_j \int_{\mathscr{E}_{(\theta_j,\Sigma_j,u_j)}} Unif(y|\mathscr{E}_{(\theta_j,\Lambda_j,u_j)})dy\right] \\
&= \left[\sum_{l\neq j}^{\infty} w_l \cdot 0\right] + \left[w_j \cdot 1\right] \\
\mathbb{P}\left[y \in C_j|w,u,\theta,\Lambda\right] &= w_j
\end{aligned}$$

Por lo que el enfoque coincide con la distribución a priori del modelo de mezclas enunciado en [11].

Ya que se ha explicado el cómo clasificar a una sola variable o una muestra aleatoria y de longitud uno, se explica brevemente el procedimiento para una muestra o conjunto de observaciones $y_1, y_2, ..., y_n$, se define a la sucesión $z_1, z_2, \ldots, z_n$ de manera que

$$z_i = \begin{cases} j & \text{si } y_i \in C_j \text{ para algún } j \\ 0 & \text{si } y_i \notin C_j \text{ para todo } j. \end{cases}$$

Para el caso de n observaciones, el modelo de clasificación, dadas las sucesiones $\left(\theta_j\right)_{j=1}^{\infty}$, $\left(\Lambda_j\right)_{j=1}^{\infty}$, $\left(\left(u_{ij}\right)_{i=1}^{n}\right)_{j=1}^{\infty}$, $\left(w_j\right)_{j=1}^{\infty}$, está dado por

$$y_i \in C_j \Longleftrightarrow \|\Lambda_j^{1/2}\left(y-\theta_j\right)\|^2 < u_{ij}$$

Se puede observar que esto coincide con la propiedad vista para el caso de una muestra compuesta por una sola observación y, por lo que análogamente se infiere que

$$\mathbb{P}\left[y_i \in C_j | w, u, \theta, \Lambda\right] = w_j.$$

3.4. Sobre las distribuciones

Ahora se verán las distribuciones a priori de u, w, μ y Λ.

Primero se propone una distribución para $u = (u)_{i=1}^{\infty}$ condicionada a los demás parámetros$\{\theta, \Lambda, w\}$. Y para ello se hara uso del siguiente resultado.

Teorema 3.4.1. *Sea Λ una matriz simétrica, definida positiva en $\mathbb{R}^{p\times p}$, θ un vector fijo en $\mathbb{R}^p$ y U una variable aleatoria con distribución ji-cuadrada de $p+2$ grados de libertad, es decir, U tiene una función de densidad de la forma*

$$f_U(u) = \frac{1}{2^{\frac{p+2}{2}}\Gamma\left(\frac{p+2}{2}\right)} u^{\frac{p}{2}} e^{-\frac{u}{2}}.$$

Sea X un vector aleatorio tal que, condicionado a un valor fijo de U, $X|U$ se distribuye uniformemente en el elipsoide $\mathscr{E}_{(\theta,\Sigma,u)} = \{y \in \mathbb{R}^p : \|\Lambda(y-\theta)\|^2 < u\}$, entonces X tiene una distribución marginal normal con media θ y matriz de precisión Λ.

Demostración. Considérese la función de densidad marginal de X.

$$\begin{aligned} f_X(x) &= \int_{-\infty}^{\infty} f_{X|U}(x|u) \cdot f_U(u) du \\ &= \int_{-\infty}^{\infty} \left(\frac{\mathbb{1}_{\{\|\Lambda^{1/2}(y-\theta)\|^2<u\}}^{(x)}}{Vol\left(\mathscr{E}_{(\theta,\Sigma,u)}\right)} \right) \cdot \left(\frac{\mathbb{1}_{(0,\infty)}^{(u)}}{2^{\frac{p+2}{2}}\Gamma\left(\frac{p+2}{2}\right)} u^{\frac{p}{2}} e^{-\frac{u}{2}} \right) du \\ &= \int_{\|\Lambda^{1/2}(y-\theta)\|^2}^{\infty} \left(\frac{1}{Vol\left(\mathscr{E}_{(\theta,\Sigma,u)}\right)} \right) \cdot \left(\frac{u^{\frac{p}{2}} e^{-\frac{u}{2}}}{2^{\frac{p+2}{2}}\Gamma\left(\frac{p+2}{2}\right)} \right) du \end{aligned}$$

Donde $Vol\left(\mathscr{E}_{(\theta,\Sigma,u)}\right)$ denota al volumen del elipsoide. Pero se sabe que el volumen de un elipsoide de esas características se calcula como sigue.

$$\begin{aligned} Vol\left(\mathscr{E}_{(\theta,\Sigma,u)}\right) &= Vol\left(\{y\in\mathbb{R}^p : \|\Lambda^{1/2}(y-\theta)\|^2 < u\}\right) \\ &= Vol\left(\{y\in\mathbb{R}^p : \|\left[(1/u)\Lambda\right]^{1/2}(y-\theta)\|^2 < 1\}\right) \\ &= |(1/u)\Lambda|^{-1/2}\cdot\frac{\pi^{p/2}}{\Gamma\left(\frac{p}{2}+1\right)} \\ &= u^{p/2}\cdot|\Lambda|^{-1/2}\cdot\frac{\pi^{p/2}}{\Gamma\left(\frac{p}{2}+1\right)} \end{aligned}$$

Y sustituyendo en la fórmula anterior se obtiene.

$$\begin{aligned} f_X(x) &= \int_{\|\Lambda^{1/2}(y-\theta)\|^2}^{\infty}\left(\frac{1}{Vol\left(\mathscr{E}_{(\theta,\Sigma,u)}\right)}\right)\cdot\left(\frac{u^{\frac{p}{2}}e^{-\frac{u}{2}}}{2^{\frac{p+2}{2}}\Gamma\left(\frac{p+2}{2}\right)}\right)du \\ &= \int_{\|\Lambda^{1/2}(y-\theta)\|^2}^{\infty}\left(\frac{1}{u^{p/2}|\Lambda|^{-1/2}\cdot\frac{\pi^{p/2}}{\Gamma\left(\frac{p}{2}+1\right)}}\right)\cdot\left(\frac{u^{\frac{p}{2}}e^{-\frac{u}{2}}}{2^{\frac{p+2}{2}}\Gamma\left(\frac{p+2}{2}\right)}\right)du \\ &= |\Lambda|^{1/2}\cdot(2\pi)^{-p/2}\cdot\int_{\|\Lambda^{1/2}(y-\theta)\|^2}^{\infty}\frac{e^{-\frac{u}{2}}}{2}du \end{aligned}$$

Y se resuelve

$$\int_{\|\Lambda^{1/2}(y-\theta)\|^2}^{\infty}\frac{e^{-\frac{u}{2}}}{2}du = e^{-\frac{\|\Lambda^{1/2}(y-\theta)\|^2}{2}}$$

Por lo que finalmente se puede concluir que

$$f_X(x) = |\Lambda|^{1/2}\cdot(2\pi)^{-p/2}\cdot e^{-\frac{\|\Lambda^{1/2}(y-\theta)\|^2}{2}}$$

Que no es más que la definición de la densidad de un vector aleatorio con distribución normal multivariada. □

Se hará uso de el teorema anterior para definir una distribución a priori para U. Suponemos a U como una variable aleatoria con distribución ji-cuadrada con $p+2$ grados de libertad pero con las restricciones que se mencionaron anteriormente para justificar que las observaciones pertenecen unicamente a un grupo y que todas las observaciones están clasificadas. Esta distribución casi coincide con el modelo de mezcla normal con el que se inició, salvo por las restricciones de que los elipsoides no se traslapan y de que cada observación se encuentra incluida en un elipsoide. Sea $D\subseteq\mathbb{R}$ el conjunto de los valores de U que satisfacen las condiciones que se mencionan anteriormente, entonces U tendrá una densidad de la forma

$$f_U(u) \propto \frac{1}{2^{\frac{p+2}{2}}\Gamma\left(\frac{p+2}{2}\right)}u^{\frac{p}{2}}e^{-\frac{u}{2}}\mathbb{1}\left(u\ \in\ D\right)$$

Así se puede expresar a la densidad de $u=((u_{ij})_{j=1}^{\infty})_{i=1}^{n}$ como una función proporcional a

$$p(u|\Lambda,\mu) \propto \prod_{i=1}^{n}\prod_{j=1}^{\infty}\frac{1}{2^{\frac{p+2}{2}}\Gamma\left(\frac{p+2}{2}\right)}(u_{ij})^{\frac{p}{2}}e^{-\frac{u_{ij}}{2}}\mathbb{1}_{\{u\ \in\ \mathbb{D}\}}$$

Nótese que el modelo sin las restricciones que define D es exactamente un modelo de mezcla de normales. Se sujetó a estas condiciones para completar un modelo de clasificación y posteriormente hacer la inferencia.

Respecto a las distribuciones a priori de los parámetros Λ_j y θ_j, se supondrá que el modelo bayesiano Normal-Wishart con parámetros Λ los describe, pero será necesario hacer una aclaración para la distribución de θ_j.

La distribución de probabilidad de y_i dados los parámetros θ, Λ, u y w obedece a la función

$$p(y_i|\theta,\Lambda,u,w) = \sum_{j=1}^{\infty} w_j \cdot Unif(y_i|\mathscr{E}_{(\theta_j,\Lambda_j,u_j)}),$$

mas aún, si también se condiciona sobre el valor de $z_i = j$ se tendrá que

$$p\left(y_i|z_i = j,\theta_j,\Lambda_j,u_{ij}\right) = Unif(y_i|\mathscr{E}_{(\theta_j,\Lambda_j,u_j)}),$$

es decir, y_i se distribuye uniformemente sobre el elipsoide $\mathscr{E}_{(\theta_j,\Lambda_j,u_j)}$ y bajo el supuesto de que U se distribuye ji-cuadrada como se aclara previamente, se tiene que la distribución marginal de $y_i|z_i = j,\theta_j,\Lambda_j$ es normal de media θ_j y matriz de precisión Λ_j.

Siguiendo el modelo bayesiano Normal-Wishart se supone que θ_j sigue una distribución inicial, condicionada al parámetro adicional κ y a Λ_j, normal con media μ_j y matriz de precisión $\left(\kappa_j\Lambda_j\right)$. Mientras que Λ_j sigue una distribución Wishart de parámetros v y T. Recordando el principal criterio de clasificación para colocar a y_i en C_j se tiene que

$$y_i \in C_j \iff ||\Lambda_j^{1/2}\left(y_i - \theta_j\right)||^2 \leq u_{ij},$$

que sigue siendo parte del objetivo al momento de actualizar θ_j, por lo que si se fija cada una de las observaciones y_i que pertenecen a c_j y se busca hacer variar a θ_j a fin de actualizar su valor utilizando la distribución posterior que se enuncia en el teorema 2.5.6 esta condición dice que cada una de las observaciones y_i deben estar dentro de su respectivo elipsoide $\mathscr{E}_{\left(\mu_j,\Lambda_j,u_{ij}\right)}$ y cambia a decir que la actualización del parámetro θ_j debe estar en la región donde se traslapan los elipsoides $\mathscr{E}_{\left(y_i,\Lambda_j,u_{ij}\right)}$.

Así las distribuciones posteriores siguiendo de nuevo el teorema 2.5.6 y la observación anterior quedan como:

$$\begin{aligned}\Lambda_j \mid v_j, T_j, C_j &\sim Wishart\left(v_j + n_j, T_j^*\right)\\ \theta_j \mid \mu_j, \kappa_j, \Lambda_j, C_j &\sim Normal\left(\mu_j^*, (\kappa_j\Lambda_j)^{-1}\right))\mathbb{1}_{\bigcap_{y_i\in C_j}\mathscr{E}_{y_i,\Lambda_j,u_{ij}}}\end{aligned}$$

donde $\mu_j^* = \frac{n_j\bar{Y}_j+\kappa_j\mu_j}{n_j+\kappa_j}$, $T_j^* = \left(\sum_{i=1}^{n_j}(y_i-\bar{Y}_j)(y_i-\bar{Y}_j)^T + (\frac{n_j\kappa_j}{n_j+\kappa_j})(\bar{Y}_j-\mu_j)(\bar{Y}_j-\mu_j)^T + T_j^{-1}\right)^{-1}$, n_j el número de observaciones asignadas a C_j, $\bar{Y}_j = \frac{1}{n_j}\sum_{\substack{i=1\\ y_i\in C_j}}^{n_j} y_i$ y $\bigcap_{y_i\in C_j}\mathscr{E}_{\left(y_i,\Lambda_j,u_{ij}\right)}$ se refiere a la región en que se traslapan los elipsoides $\mathscr{E}_{\left(y_i,\Lambda_j,u_{ij}\right)}$.

La distribución del parámetro de pesos en la mezcla w se mantiene bajo el supuesto del modelo bayesiano de distribución Dirichlet-multinomial, y sin restricciones se sigue tal y como se enuncia en el teorema 2.3.7.

Capítulo 4

Implementación

4.1. Problemáticas en la Implementación

Se puede resumir del capítulo anterior que un proceso iterativo como el que se describe:

- que parte de parámetros iniciales $\left(\theta_j, T_j, k_j, \alpha_j\right)_{j=1}^{\infty}$ para generar $\left(w_j, \mu_j, \Lambda_j\right)_{j=1}^{\infty}$,
- que utiliza variables latentes U cuya probabilidad está definida por la distribución descrita antes para obtener elipsoides que agrupan a las observaciones de una muestra $Y_1, \ldots, Y_n$ en submuestras,
- y que dadas esas submuestras actualiza los parámetros iniciales

asemeja a un método de estimación de parámetros para datos que siguen una distribución de probabilidad descrita por una mezcla de normales.

Usando un formato cercano al pseudocódigo, el algoritmo se puede condensar como en Algoritmo 2.

Si el algoritmo anterior satisface las condiciones para ser un muestreo de Gibbs, entonces teóricamente se puede llegar a una estimación coherente de los parámetros θ, Λ, w y z, así como el número de componentes de la mezcla, sin embargo, hay algunos pasos en el proceso que presentan complicaciones cuando se busca plasmar en código que una computadora pueda leer.

Un problema encontrado es el hecho de no poder generar un número infinito de parámetros, mismo que se resuelve si asumimos que n es una cota coherente para el número de componentes de la mezcla y por lo tanto de los parámetros a estimar, el principal problema surge al tratar de generar las variables U, pues se busca que siga una distribución aleatoria acotada de manera que ninguna de las elipses que son generadas por estas variables choquen, éste representa el principal obstáculo y en este documento nos limitamos a evitarlo proponiendo una definición alternativa del algoritmo, una adaptación de los recursos disponibles buscando la mayor semejanza con el algoritmo 2.

Vale la pena mencionar que a pesar de lograr un parecido consistente entre los algoritmos, el resultado obtenido de la aplicación del algoritmo planteado en el código no logra la convergencia de los parámetros. Es por ello que para aprovechar la información obtenida del proceso se consideraron varios criterios para comparar las iteraciones, los parámetros, cómo ajustan a los datos y la cantidad de información estimada, es decir, el número de parámetros estimados.

Se emplearon el Criterio de Información de Akaike, el Criterio de Información Bayesiano y la función de Verosimilitud con el objetivo de comparar entre iteraciones y considerar el conjunto de parámetros que mejor ajustó a los datos de acuerdo a cada criterio.

Algoritmo 2 Definición del proceso planteado

1: Se establecen parámetros iniciales $\left(\alpha_j^0,\ \kappa_j^0,\ \mu_j^0,\ v_j^0,\ T_j^0\right)_{j=1}^{\infty}$
2: Se generan parametros $\Lambda_j^0 \sim Wishart\left(v_j^0, T_j^0\right)$ y $\theta_j^0 \sim Normal\left(\mu_j^0, \Lambda_j^0\right)$
3: *Proceso Iterativo*:
4: **Para** $l \leftarrow 0$ **hasta** L grande que asegura convergencia **hacer**
5: Generar variables aleatorias U^l de acuerdo a 3.4
6: A partir de U^l se calculan las variables $\left(Z_i^l\right)_{i=1}^n$, de acuerdo a la regla $Z_i^l \iff ||(\Lambda_j^l)_j^{1/2}\left(y_i - \mu_j^l\right)||^2 < U_{ji}^l$.
7: Se obtienen las submuestras $\left(D_j^l\right)_{j=1}^J$.
8: **Subproceso** ACTUALIZACIÓN DE PARÁMETROS NORMAL-WISHART (Para cada j se calculan los parámetros bayesianos posteriores tomando como muestra a los registros en C_j)
9: Se actualiza el parámetro μ para θ, según el modelo Normal-Wishart
10: Se actualizan los parámetros T y v para Λ, según el modelo Normal-Wishart
11: Se actualiza el parámetro κ para θ, según el modelo Normal-Wishart
12: Se actualiza el parámetro α para w, según el modelo Dirichlet-Multinomial
13: **Fin Subproceso**
14: **Subproceso** GENERAR w, Λ, θ (Se generan nuevos parámetros para el modelo de mezclas)
15: Se generan $\Lambda_j \sim Wishart\,(T_j, v_j)$, $\theta_j \sim Normal\left(mu_j, (\kappa\Lambda)^{-1}\right)$ y $w \sim Dirichlet\,(\alpha_1, ..., \alpha_J)$
16: **Fin Subproceso**
17: **Fin Para**
18: **Devuelve** θ, Λ, w y z.

4.2. Algoritmo, definición

Para la adaptación del algoritmo se iniciará por definir los parámetros iniciales, mismos que seguirán las distribuciones más naturales para variables normales pues se supone que los datos provienen de varias muestras con distribución normal.

Para las actualizaciones de los parámetros Λ y μ será utilizado el supuesto bayesiano de una distribución Normal-Wishart, mientras que para el parámetro w el modelo Multinomial-Dirichlet será el supuesto distribucional.

Por lo tanto:

$$\begin{aligned} w &\sim Dirichlet(\alpha_1, ..., \alpha_\kappa) \\ \Lambda &\sim Wishart(v, T) \\ \mu|\Lambda &\sim Normal(\theta, (k\Lambda)^{-1}) \end{aligned}$$

En la siguiente sección se explicarán los detalles y las justificaciones de cada uno de los puntos en el Algoritmo 3. Además de enunciar también en pseudocódigo aquellos puntos que por sí mismos representan procesos lo suficientemente complejos a fin de aclararlos por completo.

Algoritmo 3 Definición completa del proceso implementado

Se fija J grande en sustitución de ∞
Se determinan J parámetros iniciales $\alpha = \alpha_1, ..., \alpha_J$, $\kappa = \kappa_1, ...\kappa_J$, $\mu = \mu_1, ..., \mu_J$, $v = v_1, ..., v_J$ y $T = T_1, ..., T_j$
Se generan parámetros $w_1, ..., w_J|\alpha$; $\Lambda_1, ..., \Lambda_J|v, T$ y $\theta_1, ..., \theta_J|\mu, \Lambda$
Proceso Iterativo:
Para $l \leftarrow 1$ **hasta** L grande que asegura convergencia **hacer**
    Genera variables aleatorias U de acuerdo a 3.4
    A partir de U y $(Z_i)_{i=1}^n$ se obtienen $(C_j)_{j=1}^J$ y $m_1, ..., m_J$
    Subproceso ACTUALIZACIÓN DE PARÁMETROS NORMAL-WISHART (Para cada j se calculan los parámetros bayesianos posteriores tomando como muestra a los registros en C_j)
        Se actualiza el parámetro μ para θ, según el modelo Normal-Wishart
        Se actualizan los parámetros T y v para Λ, según el modelo Normal-Wishart
        Se actualiza el parámetro κ para θ, según el modelo Normal-Wishart
        Se actualiza el parámetro α para w, según el modelo Dirichlet-Multinomial
    Fin Subproceso
    Subproceso GENERAR w, Λ, θ (Se generan nuevos parámetros para el modelo de mezclas)
        Se generan $\Lambda_j \sim Wishart\,(T_j, v_j)$ para $j = 1, ..., J$
        Se generan $\theta_j \sim Normal\left(\mu_j, (\kappa\Lambda)^{-1}\right)$ para $j = 1, ..., J$
        Se generan $w \sim Dirichlet\,(\alpha_1, ..., \alpha_J)$ para $j = 1, ..., J$
    Fin Subproceso
Fin Para
Devuelve θ, Λ, w y z.

1. Se toma un número natural $J \leq n$ como el máximo número de grupos de los que se puede obtener información, esto apoyado en la idea de que el número de grupos debe ser menor al número de observaciones ya que no es posible hacer inferencia sobre algún grupo que no está representado por alguna observación. Así la mezcla queda caracterizada por la densidad $f\,(x|w, \mu, \Lambda) = \sum_{j=1}^J w \cdot N\left(x|\mu_j, \Lambda_j\right)$

2. Determinar los parámetros α, κ, μ, v y T iniciales para generar los parámetros de la mezcla. En este ejercicio se tomaron de la siguiente manera:

$$\begin{aligned}
\alpha_j &= 1 && \text{para } j = 1, ..., J \\
\kappa_j &= 1 && \text{para } j = 1, ..., J \\
\mu_j &= \frac{1}{n}\sum_{i=1}^n x_i && \text{para } j = 1, ..., J \\
v_j &= 2 && \text{para } j = 1, ..., J \\
T_j &= 1/n \sum_{i=1}^n (X_i - \bar{X})\,(X_i - \bar{X})^T && \text{para } j = 1, ..., J.
\end{aligned}$$

Los parámetros α_j se escogieron de manera que inicialmente los grupos tengan pesos equilibrados entre ellos, κ_j a fin de asignar más peso en la actualización de parámetros a los datos agrupados que a los parámetros previos y v_j para evitar variaciones muy grandes en las matrices de precisión en cada iteración.

3. A partir de los parámetros anteriores se generan w, Λ y θ de acuerdo a lo siguiente

$$\begin{aligned} w &\sim Dirichlet\,(\alpha_1, ..., \alpha_J) \\ \Lambda_j &\sim Wishart\left(v_j, T_j\right) && \text{para } j = 1, ..., J \\ \theta_j &\sim Normal\left(\mu_j, (\kappa\Lambda)^{-1}\right) && \text{para } j = 1, ..., J \end{aligned}$$

4. En el paso 3 se inicia el proceso iterativo, es decir, es la parte del algoritmo que se repetirá un número de iteraciones suficientemente grande como para asumir que se presenta una convergencia.

5. El proceso iterativo del paso anterior se logra a través de la variable l, que inicia en 1 y avanza de uno en uno hasta llegar a una cota.

6. Se generan las variables U. En el capítulo 3 se establecía que estas variables debían satisfacer dos condiciones que se englobaban en el conjunto $\mathbb{D}$. La primera se refiere al hecho de que los elipsoides $\mathscr{E}_{j_1}$ y $\mathscr{E}_{j_2}$ son mutuamente excluyentes cuando $j_1 \neq j_2$, y la segunda se establece que para todo $i = 1, ..., n$ la observación x_i pertenece a alguno de los grupos C_j, es decir, $|\Lambda_j\left(\theta_j - x_i\right)| < u_{ij}$ para algún $j \in \{1, 2, ..., J\}$.

   El proceso para generar variables aleatorias U bajo estas condiciones es complicado y enfrenta su principal obstáculo en la parte donde se buscaría una cota para U que baste para decir que el elipsoide correspondiente y el resto no se traslapan. Así al tratar de encontrar los puntos en que dos elipsoides se tocan y, más aún, hacer variar U hasta encontrar el mínimo se prefirió, para fines de este trabajo, generar U de manera similar, pero mucho más simplificada.

   El proceso que se siguió en este documento consiste únicamente en considerar los centros θ que no se encuentren dentro del elipsoide del 99 % de confianza de alguno de los anteriores, como se muestra el pseudocódigo, en el paso 2, justo después de iniciar las variables. La cota para cada U_j será la mínima de las U que hace que se toque al más próximo de los centros de acuerdo a la distancia de Mahalanobis. Se asigna después a d_j la distancia de Mahalanobis entre x_i y θ_j con la matriz Λ_j. A continuación se genera U_i siguiendo una distribución ji-cuadrada en este caso con 4 grados de libertad pero estos depende de la dimensión de las variables, en el intervalo $\left(0, b_j\right)$. Ahora para asignar un x_i al grupo j se verifican dos condiciones, la primera es que $d_j > e_j$, es decir que x_i está dentro del elipsoide $\mathscr{E}_{\theta_j, \Lambda_j, e_j}$ y que j sea el índice que maximiza $w_j \cdot e^{-d_j/2}$, en caso de satisfacer ambas condiciones se fijan $U_i = e_c$ y $z_i = c$, donde Z_i se refiere al grupo donde se clasificó a la observación x_i.

   En caso de que ningún j cumpla las dos condiciones necesarias para clasificar a la observación i, ésta se quedará sin grupo durante la iteración. El detalle de este paso puede verse en el algoritmo 4.

7. Ya que se han generado las variables U y las etiquetas Z, se agrupa cada grupo o submuestra en C_j. Es decir, C_j será el conjunto de observaciones clasificadas en el grupo j.

   Por otro lado m_j es la variable que mide el tamaño de la submuestra C_j.

8. Se inicia el subproceso de, a partir de las submuestras, actualizar los parámetros κ, μ, v, T y α.

Algoritmo 4 Generar variables aleatorias U

Se fijan $U_i = (0,...,0)$, $z_i = 0$ y $b_j = \infty$ para $i = 1,...,n$ y $j = 1,...,J$
Se determina si dentro de la elipse del 99 % de confianza de μ_j hay un μ_k con $k \in \{j, j+1,...,J\}$, en caso de que sí, se omite en el proceso para dar prioridad a μ_j.
Para $j \leftarrow 1$ **hasta** J **hacer**
    Se determina la cota superior de u_{ij} con $i \in \{1,...,n\}$ como $b_j = min_{k \neq j} |\Lambda_j (\mu_j - \mu_k)|$
Fin Para
Para $i \leftarrow 1$ **hasta** n **hacer**
    Se definen $d_j = 0$ y $e_j = 0$ para $j = 1,...,J$, y $c = 0$
    Para $j \leftarrow 1$ **hasta** J **hacer**
        Se cambia el valor de d_j a $|\Lambda_j (\mu_j - x_i)|$
        De manera similar se genera e_j a partir de una densidad $\sim \chi_4 \cdot \mathbb{1}_{(0,b_j)}$
    Fin Para
    Se fija c como la j en que se maximiza $w_j \cdot e^{-d_j/2}$
    Si $d_c - e_c < 0$ **Entonces**
        Se reescribe a $U_i = e_c$ y $z_i = c$
    De otro modo
        Se reescriben las variables $U_i = NA$ y $z_i = 0$
    Fin Si
Fin Para
Devuelve

9. Se actualiza para el modelo Normal-Wishart el parámetro μ de la siguiente manera:

$$\mu_j = \begin{cases} \frac{\kappa_j \mu_j + m_j \frac{1}{m_j} \sum_{x_i \in C_j} x_i}{\kappa_j + m_j} & \text{Si } m_j \neq 0 \\ \frac{1}{n - \sum_{j=1}^{J} m_j} \sum_{x_i \notin C_j} x_i & \text{Si } m_j = 0. \end{cases}$$

   Donde $1/(n - \sum_{j=1}^{J} m_j) \cdot \sum_{x_i \notin C_j} x_i$ se refiere a la media aritmética de las observaciones que no tienen un grupo asignado, es decir, que las medias de los grupos que no consiguieron observaciones se actualizarán a la media de los datos que no tienen grupo. Esta estimación se verá afectada en caso de que dichos puntos fueran muy distantes, sin embargo, esta aproximación coincide con la idea de hacer un nuevo proceso para estos datos sin etiquetar, es decir, se toman los datos sin clasificar y se tratan de manera similar a como se trataron los datos de la muestra completa cuando no se tenían grupos identificados.

10. Se actualizan para el modelo Normal-Wishart los parámetros T y v de la siguiente manera:

$$T_j = \begin{cases} \left(T^{-1} + (m-1)\widehat{\Sigma}_j + \frac{m_j \cdot \kappa_j}{m_j + \kappa_j} \left(\mu_j - \bar{X}_j\right) \left(\mu_j - \bar{X}_j\right)^T \right) & \text{Si } m_j \neq 0 \\ \widehat{\Sigma}_* & \text{Si } m_j = 0. \end{cases}$$

   donde $\overline{X}_j = \frac{1}{m_j} \sum_{x_i \in C_j} x_i$ la media muestral de $j-$ésimo grupo, $\widehat{\Sigma}_j = \frac{1}{n-1} \sum_{x_i \in C_j} \left(x_i - \overline{X}_j\right) \left(x_i - \overline{X}_j\right)^T$, la varianza muestral de C_j y $\widehat{\Sigma}_*$ la covarianza muestral de los datos sin agrupar.

   El parámetro v se actualiza de manera que si $m_j \neq 0$, entonces $V_j = v_j + m_j$ y si $m_j = 2$, entonces $V_j = 2$.

11. Se actualiza para el modelo Normal-Wishart el parámetro κ de la siguiente manera:

$$\kappa_j = \begin{cases} 2m_j & \text{Si } m_j \neq 0 \\ 1 & \text{Si } m_j = 0. \end{cases}$$

12. En este paso se actualiza el parámetro α para el modelo Dirichlet-Multinomial con el que se generan los pesos w_j asignados a cada componente de la mezcla, así la actualización es tal que $\alpha_j \leftarrow \alpha_j + m_j$.

13. Se termina el subproceso de cálculo de parámetros bayesianos.

14. Ahora a través de los modelos bayesianos y con las cantidades previas calculadas, se generan nuevos parámetros para para el modelo de mezclas.

15. El primero en actualizarse es $\Lambda_j, j = 1, ..., J$, que consiste unicamente en generar $\Lambda_j \sim Wishart\left(v_j, T_j\right)$. Vale la pena aclarar que el valor previo de Λ_j se utilizará en el siguiente paso para determinar elipses relacionadas a la actualización del parámetro θ_j.

16. Para implementar la actualización de los parámetros θ_j en el modelo de mezclas a través del supuesto bayesiano que se hizo al inicio, es importante tomar en cuenta las distribuciones en la última sección del capítulo anterior, donde se establece que una vez que se ha actualizado el parámetro Λ_j, μ_j sigue una distribución normal acotada en una región definida por la intersección de varias elipses. Para esto se sigue un proceso que se detalla a continuación. Es importante hacer notar que las elipses $\mathscr{E}_{\left(y_i, \Lambda_j, u_{ij}\right)}$ se generan a partir de las matrices Λ_j previas y no las que se acaban de actualizar que solo modifican a la precisión de la distribución normal θ_j.

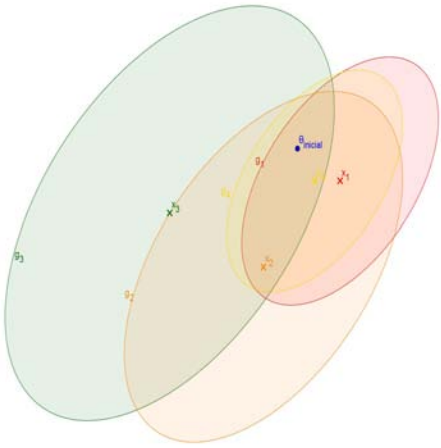
    El algoritmo consiste en diferenciar los j que tienen observaciones asignadas y los que no, pues el parámetro de estos últimos queda actualizado como la media muestral de las observaciones que no tienen ningún grupo asignado. Para el resto se generan las elipses $\mathscr{E}_{\left(y_i, \Lambda_j^*, u_{ij}\right)}$ para cada y_i en C_j. Se busca que θ_j tenga una distribución normal bivariada acotada en la región donde se traslapan las elipses anteriores. Así se inicia con θ_j como valor inicial pues este punto satisface encontrarse dentro de la región deseada. Posteriormente se sigue un proceso iterativo, en donde se toma el valor de θ_j y se toma la recta vertical con valor x igual a la constante abscisa de θ_j, sobre esta recta se genera un nuevo valor para la ordenada de θ_j acotada en la región anterior con la distribución posterior normal que sigue θ_j condicional a que el la abscisa está fija, después de manera análoga, se toma la recta que cumple que y es igual al valor de las ordenadas de θ_j, donde también se genera la variable posterior normal dado que el valor de la ordenada está fijo. Finalmente se repite este mismo proceso tomando cada vez los nuevos valores que se obtienen para θ_j, donde tras un número suficiente de iteraciones se pierde la dependencia al valor inicial siguiendo el argumento del muestreo de Gibbs ya antes descrito.

17. Se generan variables Dirichlet a partir de los nuevos parámetros $\alpha_1, ..., \alpha_J$.

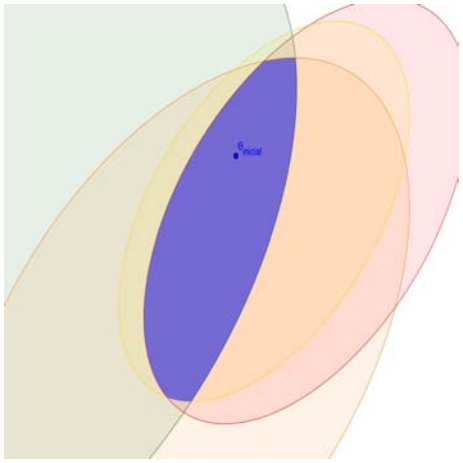
18. Se concluye la parte del proceso que corresponde a la actualización de parámetros para la mezcla.

19. Si se considera que han sido suficientes iteraciones se avanza al paso 20, de no ser así, se inicia de nuevo el proceso desde el paso 5.

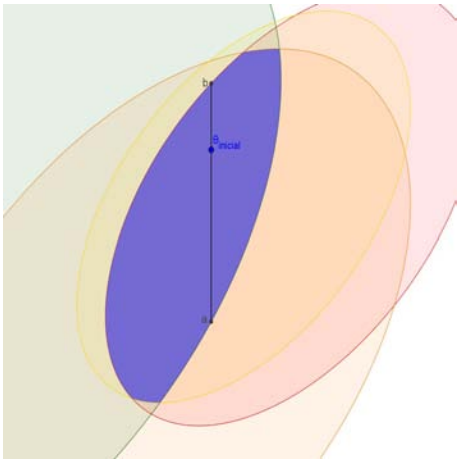
20. Se devuelven los últimos parámetros θ, Λ, w y z, que son las estimaciones del algoritmo al modelo.



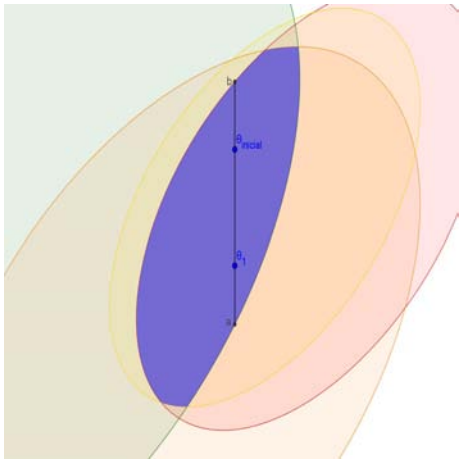
Se generan elipses como se describe en el paso 3.



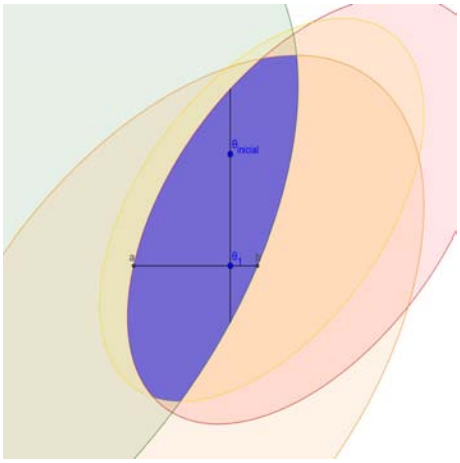
Región donde θ_j tiene probabilidad positiva.



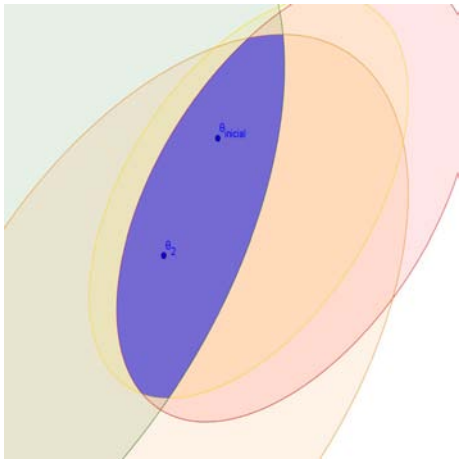
Como se detalla en 5, 6 y 7, se obtiene la recta y el intervalo.



Se genera un valor para θ_{2_j} condicional a θ_{1_j}.



Nuevamente se obtienen a y b en la recta para describir el intervalo.



Esta vez se genera θ_{1_j} dado θ_{2_j} y se inicia la siguiente iteración.

Figura 4.1: Proceso mediante el cual se actualizan los valores para el parámetro θ, Algoritmo 5.

Algoritmo 5 Generar variables aleatorias posteriores θ, figura 4.1

Para $j \leftarrow 1$ **hasta** J **hacer**
  Si $m_j \neq 0$, es decir, C_j no es vacío **Entonces**
    Se generan $\mathscr{E}_{(y_i,\Lambda_j^*,u_{ij})}$, donde Λ_j^* es el parámetro antes de la actualización del paso anterior
    Para $l \leftarrow 1$ **hasta** L grande que asegura convergencia **hacer**
      Se considera la recta $x = \theta_{j_1}$, la constante abscisa de θ_j
      $a \leftarrow \text{máx}_{y_i \in C_j}\{\text{mín}\left(\mathscr{E}_{(y_i,\Lambda_j^*,u_{ij})} \cap \{(x,y) : x = \theta_{1_j}\}\right)\}$
      $b \leftarrow \text{mín}_{y_i \in C_j}\{\text{máx}\left(\mathscr{E}_{(y_i,\Lambda_j^*,u_{ij})} \cap \{(x,y) : x = \theta_{1_j}\}\right)\}$
      Se genera θ_{2_j} con la distribución normal bivariada de θ_j condicionada a que $\theta_{1_j} = \theta_{1_j}$ y acotada dentro del intervalo (a,b)
      Se considera ahora la recta $y = \theta_{j_2}$, la constante ordenada de θ_j
      $a \leftarrow \text{máx}_{y_i \in C_j}\{\text{mín}\left(\mathscr{E}_{(y_i,\Lambda_j^*,u_{ij})} \cap \{(x,y) : y = \theta_{2_j}\}\right)\}$
      $b \leftarrow \text{mín}_{y_i \in C_j}\{\text{máx}\left(\mathscr{E}_{(y_i,\Lambda_j^*,u_{ij})} \cap \{(x,y) : y = \theta_{2_j}\}\right)\}$
      Se genera θ_{1_j} con la distribución normal bivariada de θ_j condicionada a que $\theta_{2_j} = \theta_{2_j}$ y acotada dentro del intervalo (a,b)
      Así la actualización de θ_j queda como los últimos valores.
    Fin Para
  De otro modo
    Si, en caso contrario, $m_j = 0$, se genera a $\theta_j \sim N\left(\mu_j, (\kappa_j\Lambda_j)^{-1}\right)$
  Fin Si
Fin Para
Devuelve $\theta_1, \theta_2, ..., \theta_J$

4.3. El Proceso como Muestreo de Gibbs

En el fondo la este procedimiento no es más que un ejemplo de un muestreo de Gibbs y ahora se explica ese razonamiento.

De acuerdo al capítulo II un muestreo de Gibbs es un método para generar variable aleatoria cuya distribución marginal con respecto a otras variables es difícil de obtener, pero se puede generar fácilmente condicionando con respecto a otras variables.

La idea central es considerar a los parámetros posteriores $\left(w_j, \mu_j, \Lambda_j\right)_{j=1}^{\infty}$ como esas cantidades a generar, es decir, que dados los supuestos distribucionales bayesianos se busca simular una observación de esos parámetros dada la muestra C. Se sabe que dado que no se conoce z por ser un problema de aprendizaje no supervisado, por ello generar esta observación no es sencillo y usualmente se usan algoritmos de reducción de dimensión que llevan a cabo la tarea.

Sea $z^l = \left(z_1^l, z_2^l, ..., z_n^l\right)$ un vector de etiquetas donde $z_i^l \in \{1,2,...\}$ y que se refiera al número de *cluster* al que pertenece el i-ésimo dato, en la iteración l del algoritmo. Se hace énfasis en que z_i^l no es el identificador definitivo del *cluster* de x_i sino el que en dicha iteración atrapó al dato.

Nótese que si se fijan las etiquetas z de los datos, el problema de generar parámetros posteriores se reduce a considerar para cada j la submuestra de la muestra total que tiene etiquetas j y generar $\left(w_j, \mu_j, \Lambda_j\right)$. Es decir, tomar $C_j = \{x_i \in C : z_i = j\}$ y generar parámetros con base en esa muestra. Entonces el problema completo se divide en varios problemas pequeños de solución bien conocida.

Así se puede, a partir de cada z^l, generar parámetros posteriores $(w_{l+1}, \mu_{l+1}, \Lambda_{l+1})$. De estos parámetros posteriores se pueden generar las u^l's que se mencionan en el desarrollo del

algoritmo y a partir de ellas etiquetas z^{l+1}, a partir de las cuales podemos repetir el proceso.

Esta manera de entender el algoritmo a través de un muestreo de Gibbs garantiza, en medida de lo posible, una convergencia por parte de (w_l, μ_l, Λ_l) a una observación (w, μ, Λ) de la distribución posterior de los parámetros dado el conjunto de datos observados.

4.4. Pruebas Numéricas

A continuación se ilustra el funcionamiento del algoritmo aplicándolo a distintos conjuntos de datos, el primero un conjunto sintético que proviene de una mezcla de normales, y tres conjuntos conocidos y ampliamente estudiados, el conjunto de datos propuesto por Enrique H. Ruspini en [23] para estudiar modelos de análisis de conglomerados y los datos en la base `mtcars`, una base de datos extraída de la revista Motor Trend US de 1974 en el que Henderson y Velleman cuantifican 10 aspectos en el diseño y rendimiento de 32 vehículos modelos 1973 y 1974 [13]. Todas las bases de datos empleadas se muestran en el Apéndice 2.

Para cada uno de los conjuntos de datos se implementó el código en el Apéndice 1, con cambios en las líneas entre la 26 y la 29. En cada caso se asigna a `Data` el conjunto de datos a analizar y se comentan las líneas que corresponden a los de más conjuntos.

4.4.1. Conjunto de Datos Generados Normales

El primer conjunto de datos que se utiliza para las pruebas es un conjunto generado en R, donde cada observación es generada de la mezcla de tres distribuciones normales, es decir, se generan variables normales de tres conjuntos de parámetros distintos y se asumen como una sola muestra, en este caso, se generaron de tres distintos parámetros μ_1, μ_2 y μ_3, pero todas las variables tienen varianza 1.

Para generar esta muestra se implementó el siguiente código:

```
set.seed(5)
M    <-matrix(c(1,0,0,1),2,2)
datos<-data.frame(rmvnorm(25,c(0,0),M))
datos<-rbind(datos,data.frame(rmvnorm(25,c(10,10),M)))
datos<-rbind(datos,data.frame(rmvnorm(30,c(0,20),M)))
Data <-as.matrix(datos)
```

Dicho código devuelve el conjunto de datos que se muestra en la figura 4.2, donde se puede ver que la normalidad de los datos y las matrices de varianzas angostas hacen visible la separación de los grupos en los datos.

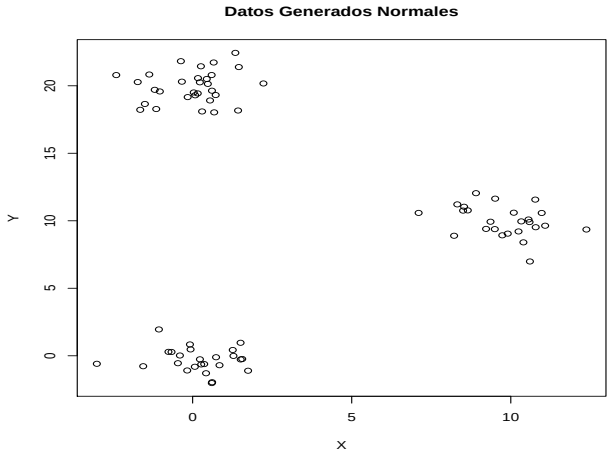


Figura 4.2: Gráfica de los datos normales

Iteración	AIC	BIC	LKH
7,139	676.68	719.56	-320.34
14,719	679.17	722.04	-321.58
7,140	679.82	722.69	-321.91
7,380	679.88	722.76	-321.94
9,261	679.96	722.84	-321.98

Tabla 4.1: Iteraciones mejor evaluadas

j	θ_1^j	θ_2^j
1	0.0185	0.3592
2	10.0218	9.3104
3	0.0673	20.1459

Λ_{11}^j	Λ_{12}^j	Λ_{21}^j	Λ_{22}^j
0.3827	-0.1226	-0.1226	0.7746
0.9159	0.0949	0.0949	1.3386
0.8390	0.3323	0.3323	0.8072

Tabla 4.2: Parámetros de la iteración número $7,139$.

Finalmente, *Data* es una muestra de longitud 80 formada por tres submuestras de variables con distribución normal, de varianza $\Sigma_j = I_2, j = 1,2,3$ y con medias $(0,0),(10,10)$ y $(0,20)$.

El análisis, después de un total de $20,000$ iteraciones considerando se concluye en que los tres conjuntos de parámetros mejor evaluados, de acuerdo a los cálculos de los criterios de Logverosimilitud, Akaike y Bayesiano (LKH, AIC y BIC respectivamente), son los que se muestran en la tabla 4.1.

El resultado obtenido por los parámetros generados en la iteración número $7,139$ está entre los mejor calificados según los tres criterios con esos datos, esta iteración devuelve un total de 3 clusters, en los que agrupa la totalidad de la muestra salvo 13 observaciones.

Las medias y varianzas obtenidas se muestran en la tabla 4.2, estos generan la clasificación que se muestra en la gráfica 4.3a y las distribuciones que se muestran en la gráfica 4.3b. Se puede observar que esta estimación ajusta bien a los datos y atina al número de grupos.

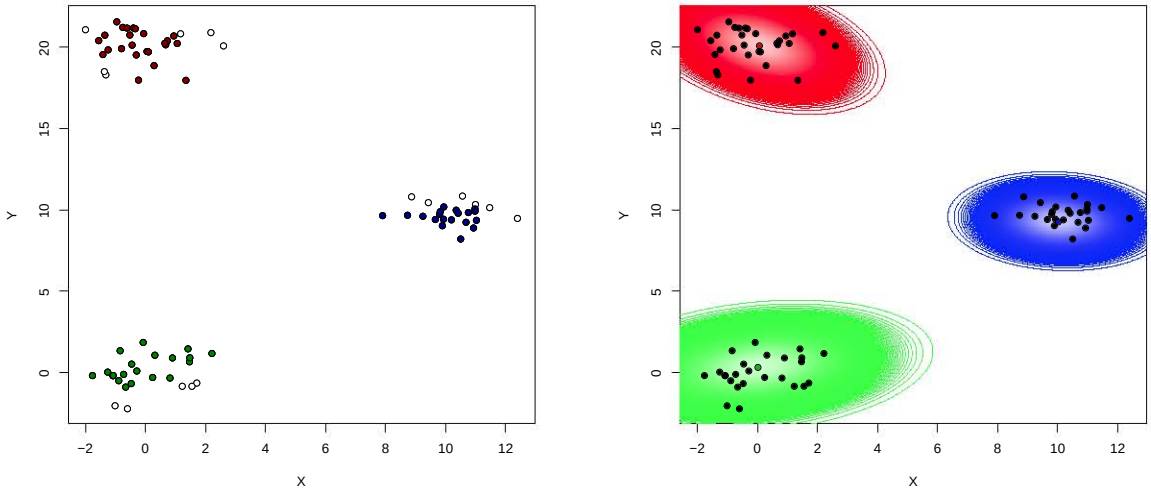
Se muestran ahora los parámetros obtenidos a partir de la iteración $14,719$, que también fueron un conjunto de parámetros con buen desempeño en los criterios. Las medias y varianzas se muestran en la tabla 4.3, los grupos obtenidos a partir de ellos en 4.4a y la gráfica con la representación de las densidades en 4.4b.

j	θ_1^j	θ_2^j
1	0.49	0.24
2	10.17	9.73
3	0.05	20.41

Λ_{11}^j	Λ_{12}^j	Λ_{21}^j	Λ_{22}^j
1.0376	-0.2882	-0.2882	1.0396
0.6303	0.2326	0.2326	1.4823
0.5511	0.1394	0.1394	1.0194

Tabla 4.3: Parámetros de la iteración número $14,719$.

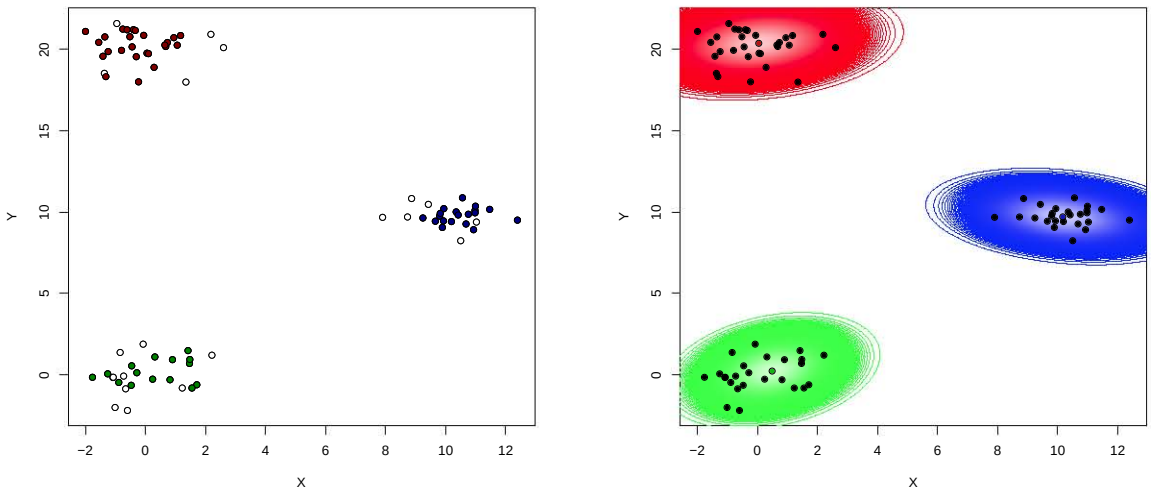
La tercera iteración con mejor desempeño en los criterios es $7,140$, cuyos parámetros están descritos en la tabla 4.4, los grupos que estos formaron se muestran en la gráfica 4.5a y el comportamiento de su densidad en la gráfica 4.5b.



(a) Clasificación de los datos

(b) Elipses de acuerdo a los parámetros.

Figura 4.3: Gráficas generadas en la iteración $7,139$.



(a) Clasificación de los datos

(b) Elipses de acuerdo a los parámetros.

Figura 4.4: Gráficas generadas en la iteración $14,719$.

j	θ_1^j	θ_2^j
1	-0.51	0.11
2	10.30	9.67
3	-0.00	20.47

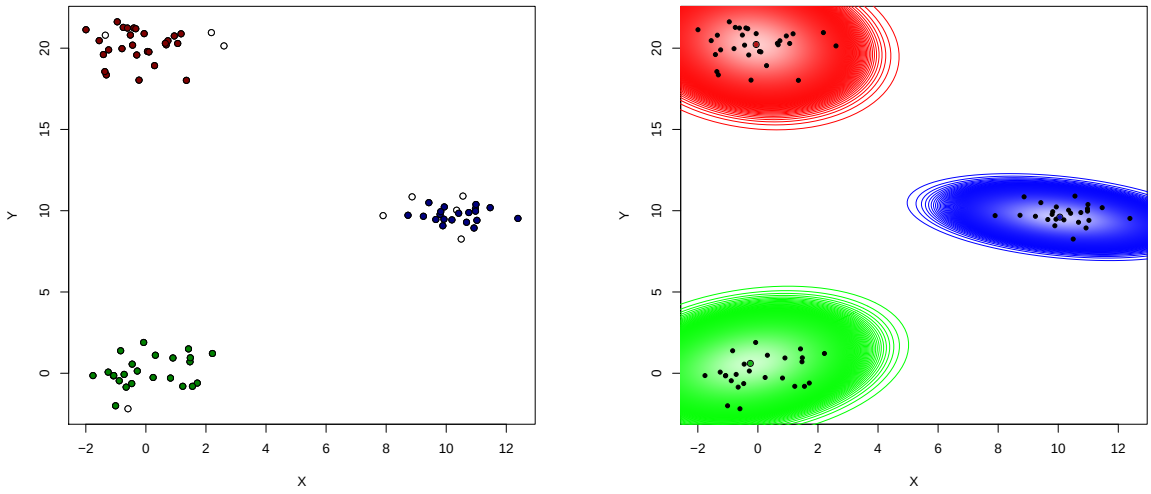
Λ_{11}^j	Λ_{12}^j	Λ_{21}^j	Λ_{22}^j
0.3972	-0.1063	-0.1063	0.3111
0.8995	-0.0254	-0.0254	2.1442
0.4762	-0.1890	-0.1890	0.8102

Tabla 4.5: Parámetros de la iteración número $7,380$.

j	θ_1^j	θ_2^j
1	-0.25	0.61
2	10.05	9.61
3	-0.06	20.23

Λ_{11}^j	Λ_{12}^j	Λ_{21}^j	Λ_{22}^j
0.4716	-0.1255	-0.1255	0.5806
0.5361	0.3358	0.3358	1.9354
0.5562	0.0715	0.0715	0.4567

Tabla 4.4: Parámetros de la iteración número $7,140$.



(a) Clasificación de los datos

(b) Elipses de acuerdo a los parámetros.

Figura 4.5: Gráficas generadas en la iteración $7,140$.

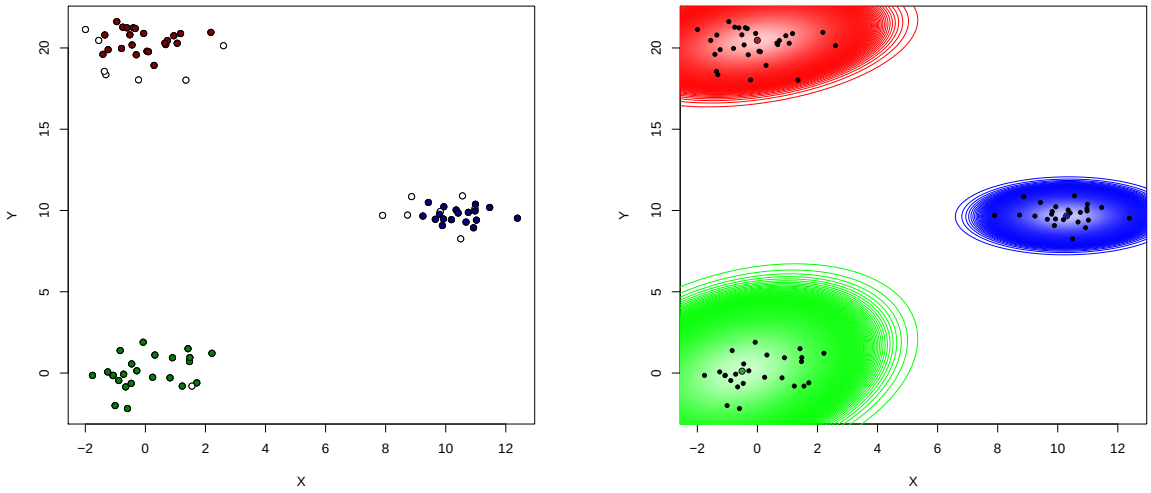
La iteración $7,380$ devuelve el siguiente conjunto de parámetros mejor evaluados, explícitamente mostrados en la tabla 4.5, cuyos agrupamientos se muestran en la gráfica 4.6a y con una densidad que se comporta como se muestra en la gráfica 4.6b.

Finalmente los parámetros de la última iteración a mostrarse son los que se formaron a partir de la repetición $9,261$ del proceso. Igualmente estos se muestran en la tabla 4.6, los grupos que generan en la gráfica 4.7a y el comportamiento de la probabilidad dados estos parámetros es como se muestra en la gráfica 4.7b.

j	θ_1^j	θ_2^j
1	0.08	0.14
2	10.30	9.95
3	-0.34	19.87

Λ_{11}^j	Λ_{12}^j	Λ_{21}^j	Λ_{22}^j
0.3915	-0.1062	-0.1062	0.4584
0.4260	-0.1279	-0.1279	1.4338
1.1459	-0.0387	-0.0387	0.6770

Tabla 4.6: Parámetros de la iteración número $9,261$.



(a) Clasificación de los datos

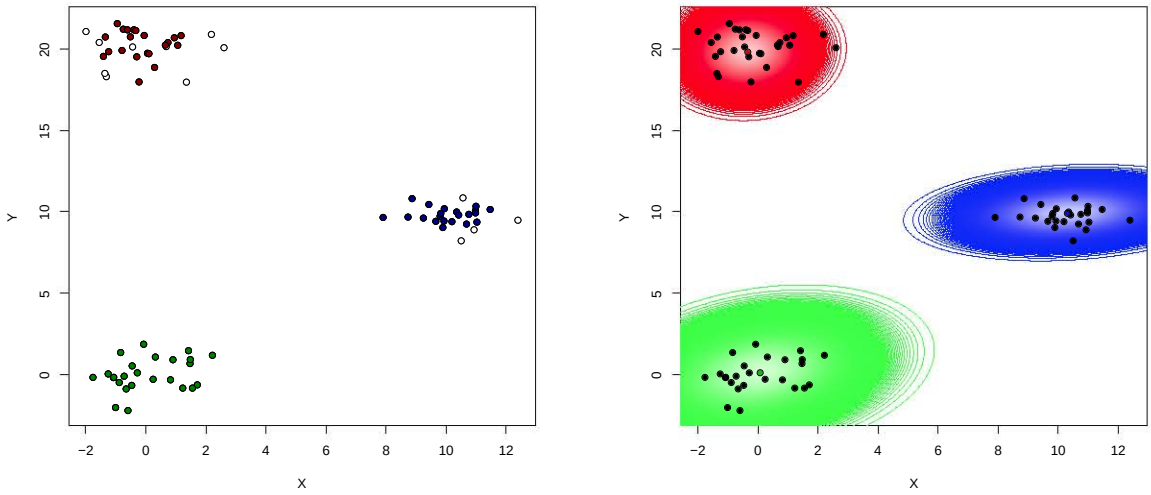
(b) Elipses de acuerdo a los parámetros.

Figura 4.6: Gráficas generadas en la iteración $7,380$.

A partir de los resultados anteriores se puede observar que el proceso estimó de manera adecuada tanto el número de grupos, como los parámetros de las mezclas con datos que siguieron una distribución de probabilidad de una mezcla gaussiana.

4.4.2. Conjunto de Datos Ruspini

El conjunto de datos Ruspini consiste en 75 puntos en $\mathbb{R}^2$ que son populares para ilustrar técnicas de análisis de conglomerados. Tienen su origen en [23] por Enrique Ruspini.



(a) Clasificación de los datos

(b) Elipses de acuerdo a los parámetros.

Figura 4.7: Gráficas generadas en la iteración $9,261$.

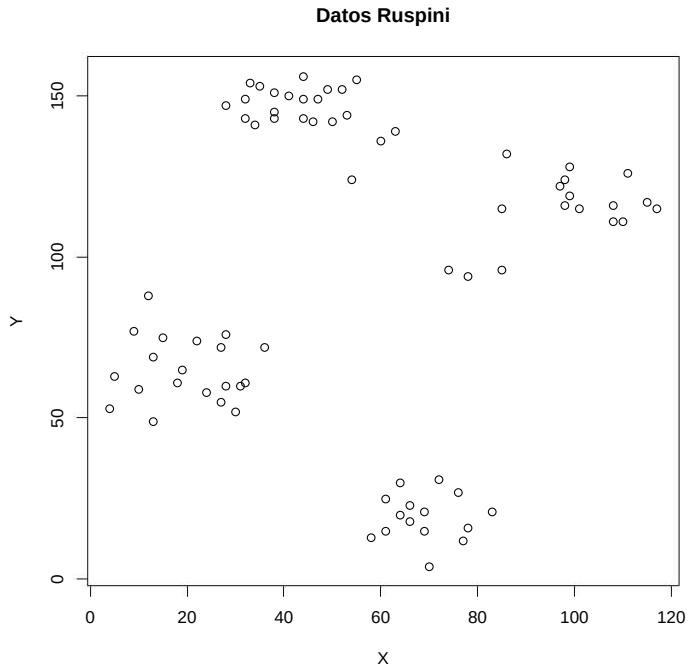


Figura 4.8: Gráfica del conjunto de datos Ruspini

Después de un total similar de $20,000$ repeticiones del proceso, los resultados muestran que los conjuntos de parámetros mejor evaluados de acuerdo a los mismos criterios son los que se enuncian en la tabla 4.7, y se detallan a continuación.

Iteración	AIC	BIC	LKH
12,201	1,382.93	1,438.55	-667.46
19,472	1,375.08	1,430.70	-663.54
5,883	1,377.96	1,447.48	-658.98
525	1,371.82	1,427.44	-661.91
18,763	1,379.14	1,434.76	-665.57

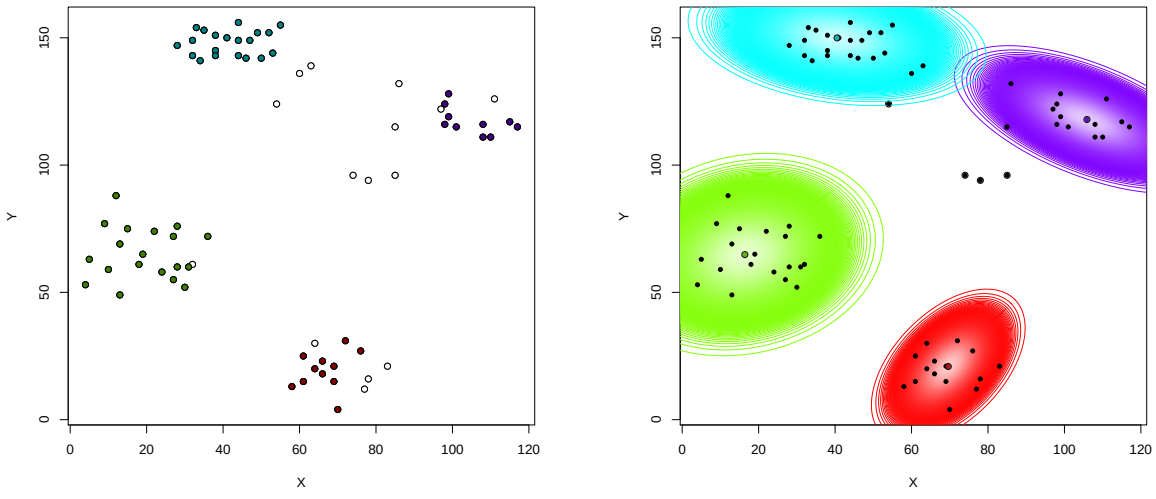
Tabla 4.7: Iteraciones mejor evaluadas

Los parámetros resultado de la iteración número $12,201$ forman uno de los conjuntos mejor evaluados que, bajo el supuesto distribucional de probabilidad de una mezcla gaussiana, devolvió un modelo de cuatro grupos con medias y varianzas obtenidas se muestran en la tabla 4.8, estos clasifican a los datos como se muestra en la gráfica 4.9a y los componentes de la mezcla se muestran en la figura 4.9b.

j	θ_1^j	θ_2^j
1	16.39	64.84
2	40.53	149.96
3	105.88	117.91
4	69.61	20.89

Λ_{11}^j	Λ_{12}^j	Λ_{21}^j	Λ_{22}^j
0.0096	-0.0013	-0.0014	0.0080
0.0087	0.0032	0.0032	0.0186
0.0157	0.0111	0.0111	0.0225
0.0407	-0.0134	-0.0134	0.0178

Tabla 4.8: Parámetros de la iteración número $12,201$.



(a) Clasificación de los datos

(b) Elipses de acuerdo a los parámetros.

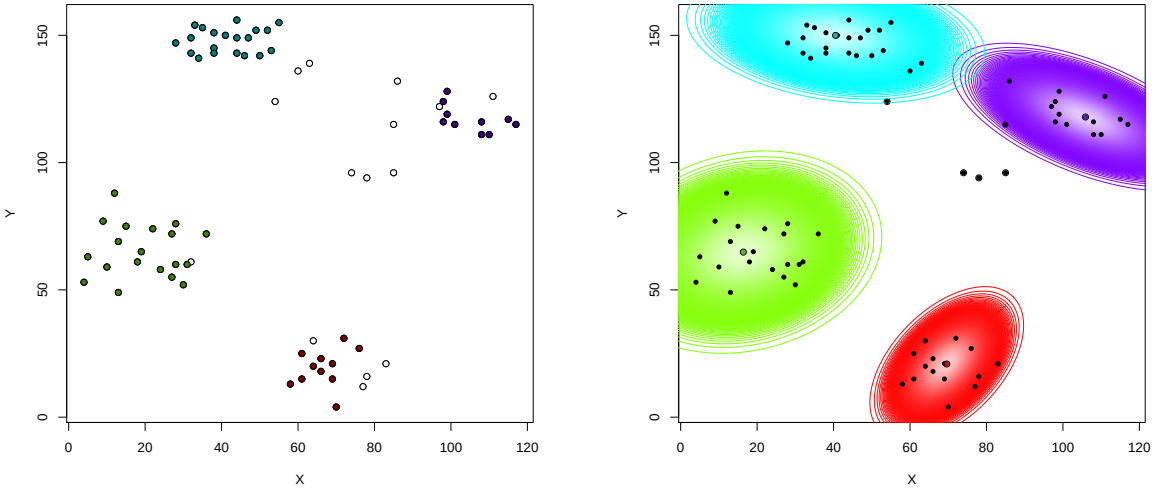
Figura 4.9: Gráficas generadas en la iteración $12,201$.

Otro conjunto de parámetros bien evaluados es el formado en la repetición número $19,472$, dichos parámetros se enlistan en 4.9, y clasifican a los datos como en la gráfica 4.10a. La mezcla considera los componentes como se muestra en la gráfica 4.10b.

j	θ_1^j	θ_2^j
1	22.43	61.66
2	39.80	146.86
3	101.67	120.47
4	66.86	26.65

Λ_{11}^j	Λ_{12}^j	Λ_{21}^j	Λ_{22}^j
0.0133	-0.0009	-0.0009	0.0095
0.0041	0.0012	0.0012	0.0126
0.0027	-0.0025	-0.0025	0.0081
0.0125	0.0031	0.0031	0.0127

Tabla 4.9: Parámetros de la iteración número 19,472.



(a) Clasificación de los datos

(b) Elipses de acuerdo a los parámetros.

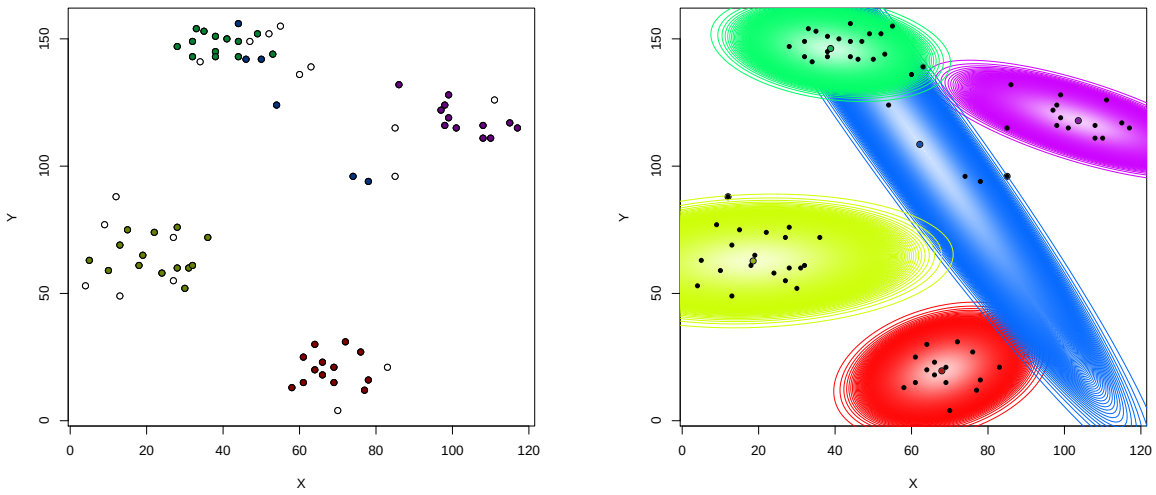
Figura 4.10: Gráficas generadas en la iteración 19,472.

Otra combinación de parámetros con buen desempeño según los criterios fue la obtenida a partir de la repetición 5,883 del proceso, ésta considera un total de 5 grupos, que contrasta con la mayoría para el conjunto de datos de Ruspini, sin embargo, se puede ver que el grupo central surge de observaciones que no siempre han sido considerados por los conjuntos de parámetros anteriores, mismos que no parecen ajustar de manera normal a alguno de los otros grupos formados. Las medias y varianzas obtenidas se muestran en la tabla 4.10, dichos parámetros surgen de la clasificación que se muestra en la gráfica 4.11a y esta considera que la mezcla normal está formada por las variables normales que se muestran en la figura 4.11b.

j	θ_1^j	θ_2^j
1	18.56	62.76
2	38.84	146.21
3	62.16	108.56
4	103.62	117.86
5	67.93	19.61

Λ_{11}^j	Λ_{12}^j	Λ_{21}^j	Λ_{22}^j
0.0045	-0.0009	-0.0009	0.01804699
0.0135	0.0054	0.0054	0.03063918
0.0421	0.0198	0.0198	0.01003433
0.0163	0.0196	0.0196	0.04534126
0.0170	-0.0053	-0.0053	0.01862544

Tabla 4.10: Parámetros de la iteración número 5,883.



(a) Clasificación de los datos

(b) Elipses de acuerdo a los parámetros.

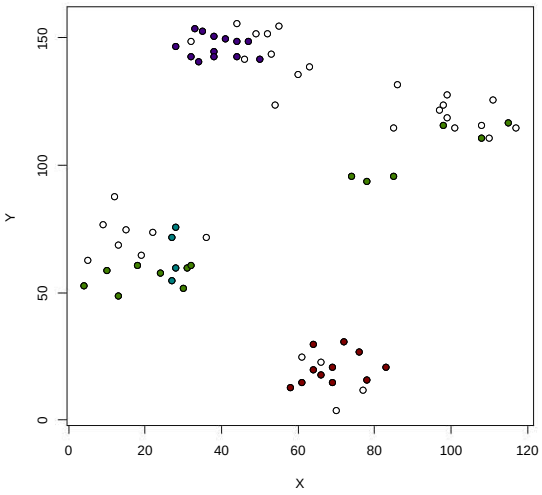
Figura 4.11: Gráficas generadas en la iteración $5,883$.

La iteración 525 también está en el grupo de las mejor calificadas, el conjunto de parámetros correspondiente, que se muestra en 4.11, difiere de los anteriores por la presencia de un componente en la mezcla que aporta una media en el punto $(-24.10, 17.28)$ y que agrupa un porcentaje importante de puntos como se puede ver en 4.9a, y que debido a su matriz de varianzas devuelve el componente verde de la gráfica, 4.9b, mismo que difiere con el comportamiento de los datos. En este ejemplo particular se aprecia la debilidad del modelo cuando los datos no siguen las distribuciones normales de los supuestos iniciales.

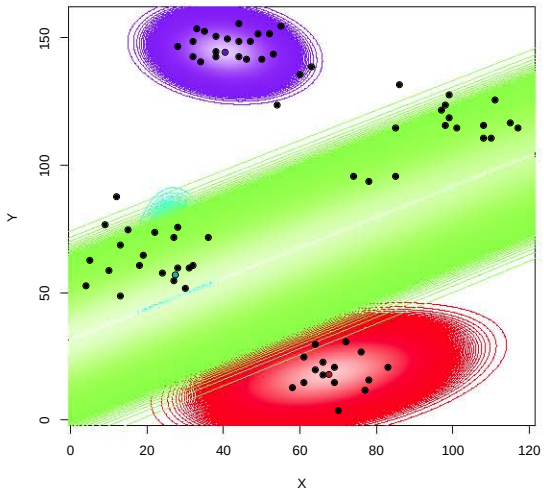
j	θ_1^j	θ_2^j
1	-24.10	17.28
2	27.38	57.28
3	40.40	144.73
4	67.55	18.19

Λ_{11}^j	Λ_{12}^j	Λ_{21}^j	Λ_{22}^j
0.0025	-0.0041	-0.0041	0.0068
0.1040	0.0018	0.0018	0.0107
0.0196	0.0041	0.0041	0.0293
0.0069	-0.0047	-0.0047	0.0191

Tabla 4.11: Parámetros de la iteración número 525.



(a) Clasificación de los datos



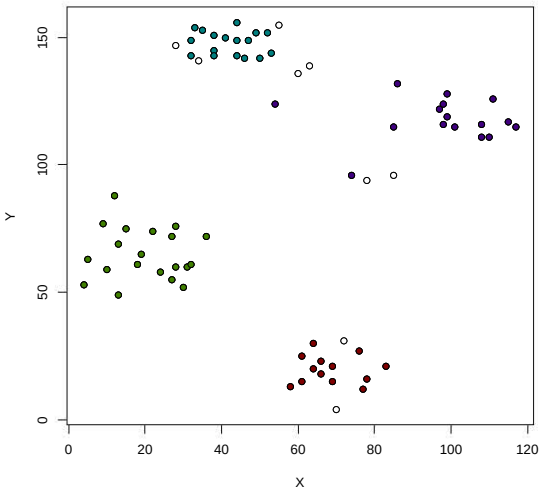
(b) Elipses de acuerdo a los parámetros.

Figura 4.12: Gráficas generadas en la iteración 525.

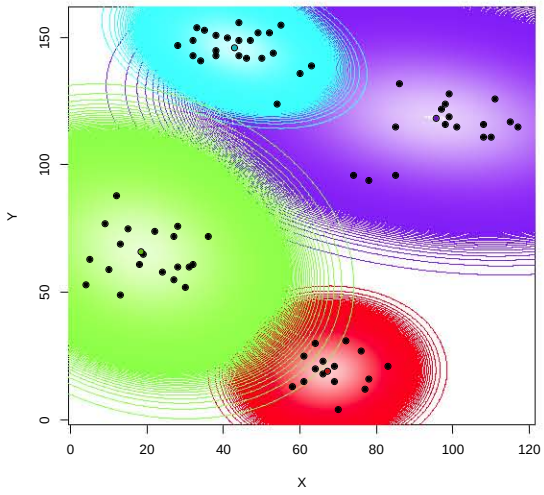
j	θ_1^j	θ_2^j
1	18.40	65.99
2	42.85	146.10
3	95.62	118.40
3	67.18	19.05

Λ_{11}^j	Λ_{12}^j	Λ_{21}^j	Λ_{22}^j
0.0042	0.0007	0.0007	0.0030
0.0115	0.0042	0.0042	0.0137
0.0017	0.0005	0.0005	0.0034
0.0126	-0.0001	-0.0001	0.0111

Tabla 4.12: Parámetros de la iteración número 18,763.



(a) Clasificación de los datos



(b) Elipses de acuerdo a los parámetros.

Figura 4.13: Gráficas generadas en la iteración 18,763.

Finalmente la última ilustración del desempeño del proceso con los datos de Ruspini es la correspondiente a la iteración número 18,763. En esta repetición los parámetros listados en la

tabla 4.8, misma que presenta uno de los totales de observaciones sin clasificar más bajos con 9 como se puede ver en la gráfica 4.9a. La mezcla que produce dicha partición se muestra en 4.9b, cuyos componentes presentan varianzas amplias, lo que ayuda a explicar el bajo nivel de datos sin clasificar.

Los datos Ruspini no siguen una distribución normal de probabilidad, y ponen a prueba la flexibilidad del modelo para enfrentar datos de esta naturaleza, sin embargo, no reflejan la complejidad que generalmente presentan los datos reales. A pesar de presentar resultados menos contundentes que el conjunto de datos anterior, éstos mantienen la coherencia respecto a la impresión gráfica que dan los datos y el desempeño del modelo parece satisfactorio.

4.4.3. Conjunto de Datos mtcars

Los datos en el conjunto `mtcars` se refieren a información de diversos modelos de automóviles entre 1973 y 1974. El conjunto incluye mediciones de millas recorridas por galón de gasolina, el número de cilindros, desplazamiento, caballos de fuerza, relación de eje trasero, peso en toneladas, tiempo en recorrer un cuarto de milla, entre otras; sin embargo, como en el capítulo 1 de este documento, el modelo se probó considerando únicamente las dos variables desplazamiento y cociente del eje trasero, que devuelven una gráfica como la que se muestra en 4.14.

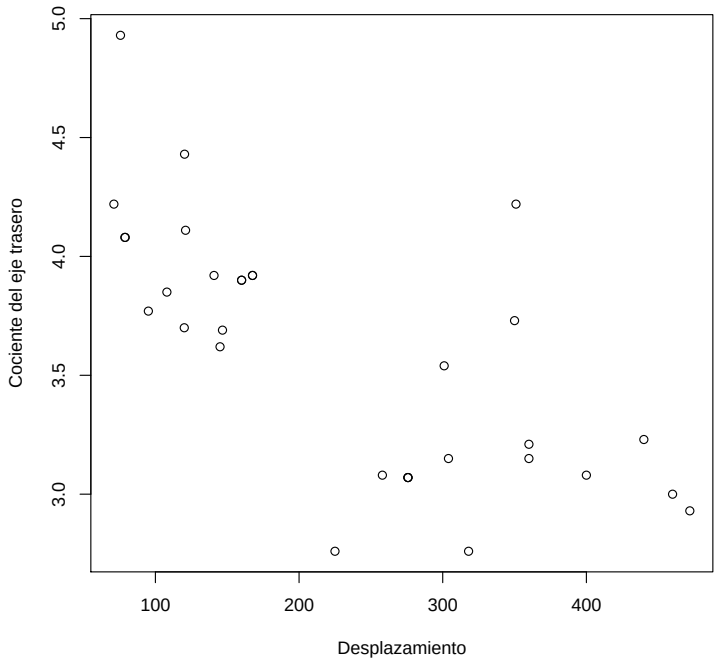


Figura 4.14: Gráfica del conjunto de datos mtcars

Siguiendo el procedimiento que se siguió con los dos conjuntos de datos anteriores, se ejecutaron $20,000$ iteraciones del proceso, que arrojaron los resultados que se muestran en la tabla 4.13, y se detallan a continuación a través de las mismas tablas y gráficas.

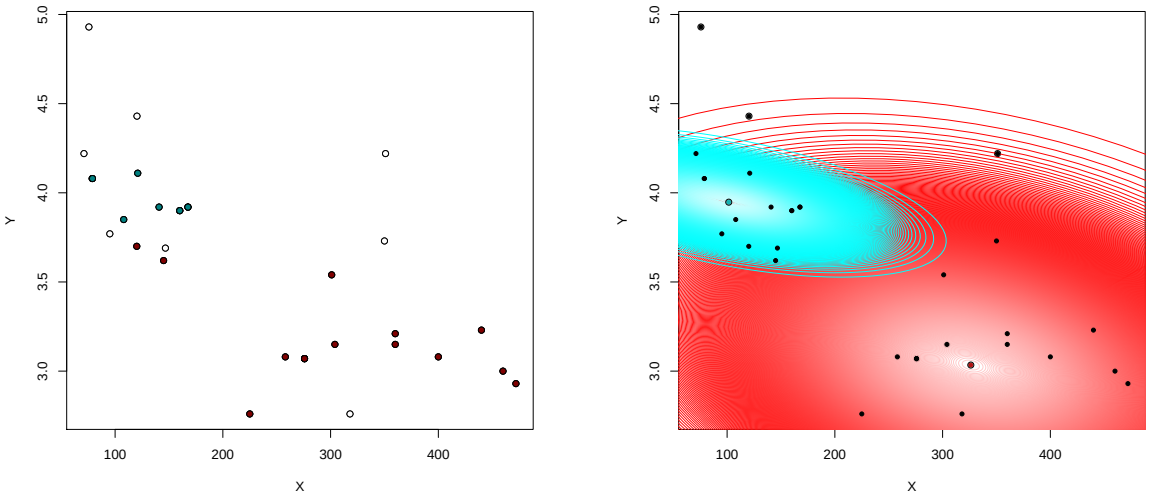
Iteración	AIC	BIC	LKH
16,825	430.57	448.16	-203.28
5,886	432.95	450.54	-204.48
13,154	431.32	448.91	-203.66
17,901	434.97	443.77	-211.49
149	443.04	469.42	-203.52

Tabla 4.13: Iteraciones mejor evaluadas

La iteración número $16,825$ devuelve las medias y varianzas listadas en la tabla 4.14, estos parámetros generan la clasificación que se muestra en la gráfica 4.15a y la mezcla gaussiana considera densidades como se muestran en la figura 4.15b. Este conjunto de parámetros agrupa a los datos en dos grupos, el primero agrupando los datos de la parte superior izquierda de la gráfica y la mayoría del resto de los datos en otro grupo con una varianza amplia que trata de ajustar a la dispersión de los datos.

j	θ_1^j	θ_2^j	Λ_{11}^j	Λ_{12}^j	Λ_{21}^j	Λ_{22}^j
1	101.52	3.95	0.00041	0.10290	0.10290	95.38839
2	326.18	3.03	0.00007	0.00565	0.00565	5.96041

Tabla 4.14: Parámetros de la iteración número $16,825$.



(a) Clasificación de los datos

(b) Elipses de acuerdo a los parámetros.

Figura 4.15: Gráficas generadas en la iteración $16,825$.

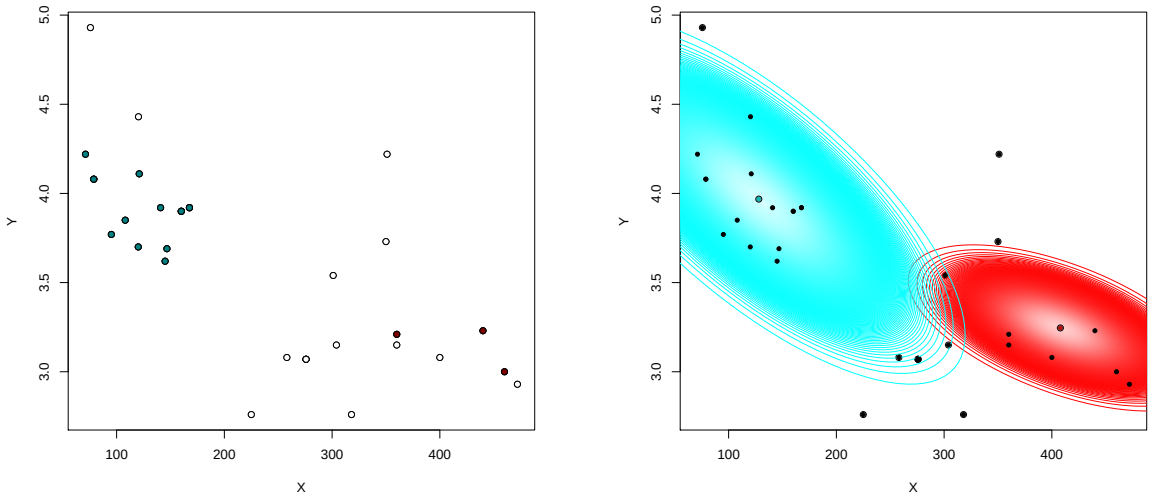
Otra iteración de los datos que generó buen desempeño en la evaluación de los criterios fue la de número $5,886$, misma que estima una mezcla gaussiana de dos componentes, cada uno con parámetros como se muestran en la tabla 4.15, dicha mezcla se obtuvo de la partición que se muestra en la gráfica 4.16a. La figura 4.16b muestra el comportamiento estimado de los componentes de la mezcla, donde se observa que el grupo de la parte inferior derecha de la

fracción del plano dibujada mantiene una varianza menor comparada con la que presentó un grupo similar la iteración anterior.

j	θ_1^j	θ_2^j
1	128.01	3.97
2	407.92	3.25

Λ_{11}^j	Λ_{12}^j	Λ_{21}^j	Λ_{22}^j
0.00073	0.09947	0.09947	24.95711
0.00092	0.15955	0.15955	84.50723

Tabla 4.15: Parámetros de la iteración número $5,886$.



(a) Clasificación de los datos

(b) Elipses de acuerdo a los parámetros.

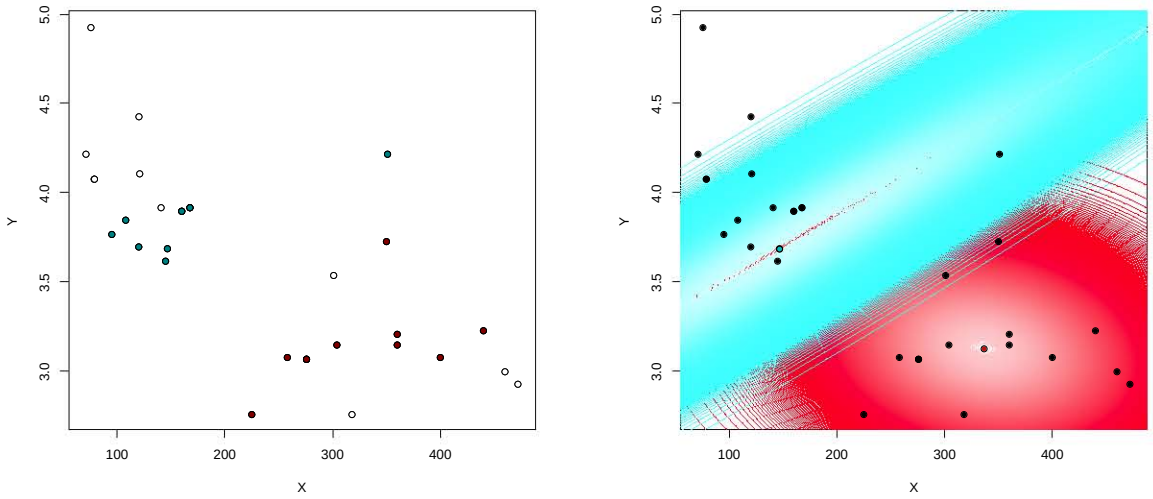
Figura 4.16: Gráficas generadas en la iteración $5,886$.

Además de las anteriores, la repetición $13,154$ también se desempeña bien respecto a los criterios, sin embargo, los parámetros que se muestran en la tabla 4.16 y que cuyo comportamiento se ilustra en la gráfica 4.17b ajusta a los datos de manera deficiente, esto principalmente debido a que considera varianzas significativamente más amplias y por lo tanto, una media dispersa en el grupo de color azul de la misma gráfica. La agrupación que se generó en dicha iteración se puede observar en la figura 4.17a.

j	θ_1^j	θ_2^j
1	146.88	3.69
3	336.65	3.13

Λ_{11}^j	Λ_{12}^j	Λ_{21}^j	Λ_{22}^j
0.00028	-0.07241	-0.07241	20.54812
0.00014	0.00572	0.00572	9.18323

Tabla 4.16: Parámetros de la iteración número $13,154$.



(a) Clasificación de los datos

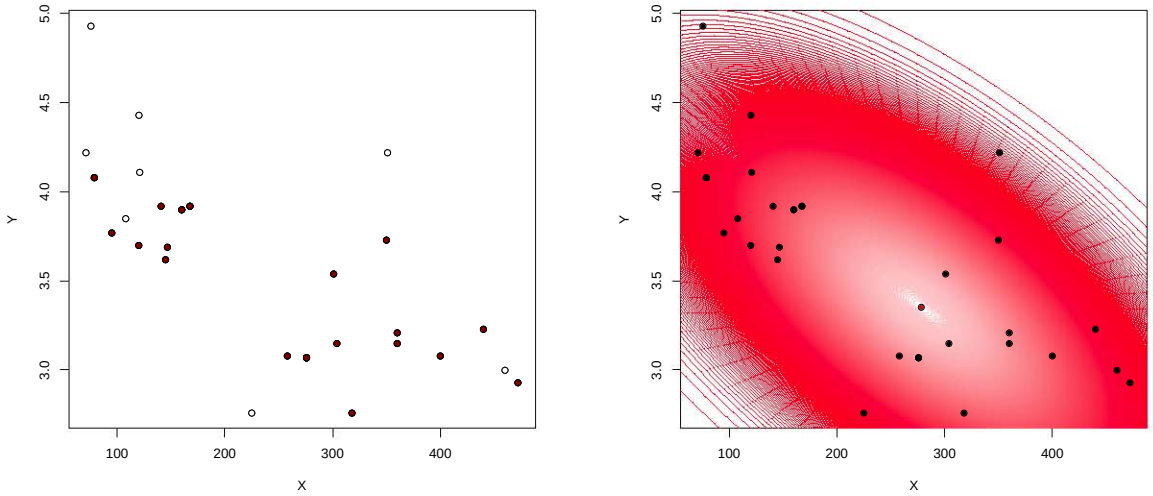
(b) Elipses de acuerdo a los parámetros.

Figura 4.17: Gráficas generadas en la iteración 13,154.

j	θ_1^j	θ_2^j
1	278.3834	3.35

Λ_{11}^j	Λ_{12}^j	Λ_{21}^j	Λ_{22}^j
0.00015	0.01836	0.01836	5.57384

Tabla 4.17: Parámetros de la iteración número 17,901.



(a) Clasificación de los datos

(b) Elipses de acuerdo a los parámetros.

Figura 4.18: Gráficas generadas en la iteración 17,901.

Los criterios también destacan el conjunto de parámetros correspondiente a la repetición 17,901 del proceso. Ésta devuelve un solo grupo en los datos con la media y la matriz de varianzas que aparecen en la tabla 4.17. Como también puede observarse en la figura 4.18a, estos parámetros generan una clasificación que agrupa a la mayoría de los datos y trata de ajustar

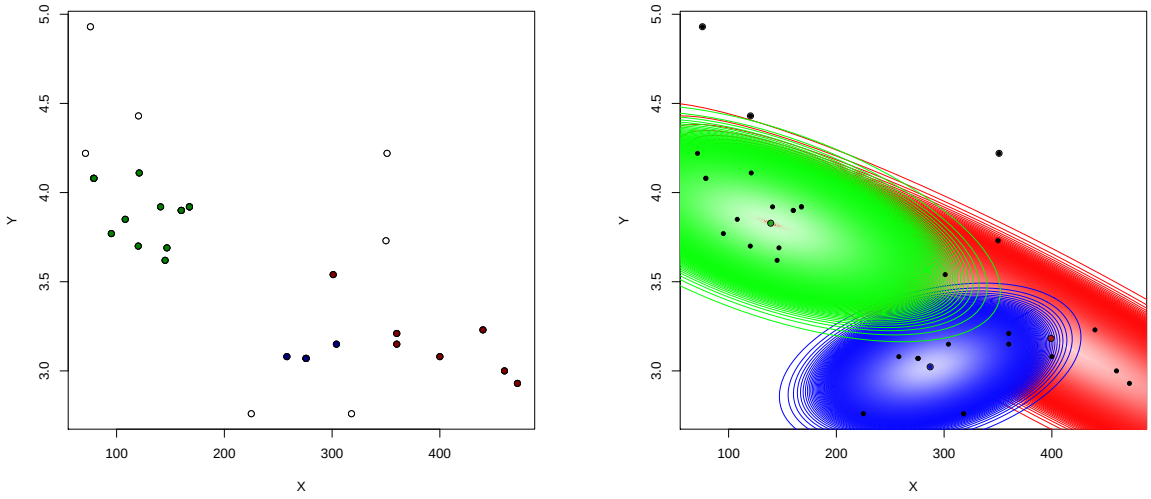
una densidad normal como se puede ver en la gráfica 4.18b, nótese que el comportamiento de los datos difícilmente coincide con una distribución de probabilidad normal.

Finalmente, los criterios consideran el modelo obtenido de la repetición 149 del algoritmo, donde la mezcla está formada por los tres componentes cuyas medias y varianzas están listadas en la tabla 4.18, la clasificación que devolvió a estos parámetros está en la figura 4.19a y no considera las observaciones en la parte superior de la gráfica ni las dos de la parte inferior por lo que las tres medias se concentran en la parte central. El comportamiento de la probabilidad de acuerdo a los parámetros anteriores se puede ver en la gráfica 4.19b donde también se puede ver que los datos, nuevamente, con dificultad seguirían dicho comportamiento.

j	θ_1^j	θ_2^j
1	139.08	3.83
2	287.06	3.02
3	399.22	3.18

Λ_{11}^j	Λ_{12}^j	Λ_{21}^j	Λ_{22}^j
2 0.00040	0.07417	0.07417	41.53008
5 0.00072	-0.07604	-0.07604	63.95920
4 0.00054	0.14869	0.14869	47.81917

Tabla 4.18: Parámetros de la iteración número 149.



(a) Clasificación de los datos

(b) Elipses de acuerdo a los parámetros.

Figura 4.19: Gráficas generadas en la iteración 149.

El proceso como puede observarse con este último conjunto de datos `mtcars`, es sensible a la falta de normalidad en los datos, el modelo no presentó homogeneidad respecto al comportamiento de los parámetros en los componentes de las mezclas ni al número de grupos en los datos, se forzaron probabilidades normales con medias altas para tratar de compensar esa misma falta de normalidad y como resultado no se pueden hacer conclusiones concretas.

Capítulo 5

Conclusiones

Sobre los Resultados

A partir de los tres conjuntos de datos usados en el algoritmo, y de los resultados obtenidos se pueden concluir algunas afirmaciones.

- El modelo es sensible a la normalidad de las muestras.
- Los datos atípicos no parecen influir demasiado.
- El modelo tiene un desempeño de regular a bueno identificando el número de grupos.
- El modelo utiliza la libertad que tiene sobre la varianza para ajustar los datos.

Las pruebas numéricas del algoritmo del modelo se realizaron en datos de mayor a menor ajuste normal. De esa misma manera se comportan los resultados que con el primer conjunto de datos, los cinco mejor resultados obtenidos coinciden en el número de grupos y en la similitud de sus parámetros. Con el conjunto siguiente de datos los resultados no siempre muestran el mismo número de grupos, principalmente por las observaciones centrales que son difíciles de agrupar, en la mayoría de los resultados se asume que los datos siguen una mezcla de variables normales con cuatro componentes. El último conjunto de datos `mtcars` devuelve resultados mucho más diversos, esto debido al tamaño menor de la muestra y a la ausencia de normalidad en los datos, los resultados nos muestran posibilidad de uno, dos y hasta tres componentes en la función mezclada de densidad de probabilidades, cada una con parámetros muy distintos a las demás, de ahí se puede decir que la normalidad de los datos o un tamaño de muestra mayor son factores que influyen en el desempeño del algoritmo.

El número de grupos en los datos muestra uniformidad en el primer conjunto, donde todos los resultados que se consideraron muestran a tres componentes, en el segundo conjunto de datos todos los resultados mostrados consideran el número de grupos como cuatro con excepción de uno que toma como base una mezcla de cinco componentes, para el conjunto de datos `mtcars`, como se describió en el párrafo previo, los resultados muestran conclusiones más diversas y es complicado llegar a una conclusión adecuada.

A partir del análisis hecho del conjunto de datos obtenido de la base `mtcars`, se puede ver que cuando el algoritmo no logra conciliar los datos con una mezcla paramétrica adecuada, este hace uso de la libertad que tiene sobre la generación de parámetros y extiende las varianzas a fin de agrupar los más datos posibles, a pesar de que eso no siempre concluya en un mejoramiento de la evaluación de los criterios de Akaike y Bayesiano, o de su función de verosimilitud.

Principales Problemas

El modelo parece funcionar bien a pesar de enfrentar problemáticas y de diferir en la idea inicial del algoritmo, principalmente en el paso de qué parámetros utilizar.

Al iniciar el presente trabajo se creía que se podía garantizar la convergencia del modelo hasta que las diferencias en los parámetros fueran irrelevantes, dicha convergencia no se logró y el resultado consistía en conjuntos de parámetros ajustados a los datos donde algunos de estos conjuntos ajustaban adecuadamente a los datos, otro enfoque era tomar en cuenta la moda de las clasificaciones, es decir, considerar a las etiquetas del total de observaciones por iteración como una variable y tomar la moda, dicho de otro modo, tomar la agrupación que más repeticiones tuviera por iteración y a partir de esta obtener parámetros, no obstante, con 20,000 iteraciones no se lograron repeticiones en los agrupamientos de los datos mostrados. Partiendo de esto se decidió considerar criterios para comparar entre conjuntos de parámetros, criterios que dependieran de los datos, de los parámetros y de el supuesto distribucional del que se había partido, así se consideraron los criterios de Akaike, Bayesiano y la función de verosimilitud para seleccionar los conjuntos de parámetros que mejor ajustaran a los datos y se desarrolló el proceso tal como está definido el el capítulo 4.

Otro problema importante del enfoque, o específicamente de la implementación como está definida, es el tiempo en realizar la totalidad de las iteraciones que implica horas o días para ejecutar la totalidad de las 20 mil iteraciones que se consideran suficientes para los fines de este documento.

Discusión

Son diversos los elementos que se pueden modificar sobre el algoritmo presentado y aplicarse en muchas partes del mismo.

Los siguientes son algunos elementos que se considera que se pueden alterar, modificar o tomar en cuenta para alguna posible generalización.

Criterios de Selección

En el presente documento se consideran como criterios de selección de un conjunto de parámetros sobre otros al desempeño de estos en criterios como el Criterio de Información de Akaike, el Criterio de Información Bayesiano o la evaluación de la función de verosimilitud, sin embargo, posiblemente habrá diferentes criterios de decisión de diferente naturaleza, se podrían considerar el número de observaciones que no fueron tomadas en cuenta para generar los parámetros o algún factor de ajuste distinto.

Supuestos Distribucionales

El modelo descrito en este documento supone que los datos pueden ser vistos como realizaciones de una variable que sigue una función de densidad de probabilidad de la forma de una mezcla de variables normales. Este supuesto podría relajarse a tomar otra densidad en lugar de la distribución gaussiana a reserva del cálculo de parámetros posteriores que para dicha distribución son ampliamente manejados en la literatura y que con otras familias paramétricas podrían presentar dificultades o la necesidad de recurrir a soluciones numéricas. Recordando que las variables normales en el documento se construyen a partir de otras con distribución Ji cuadrada y uniformes en el elipsoide que devuelve la anterior. Tomando esa información en consideración se podría pensar en alterar los grados de libertad de la distribución de la variable

Ji cuadrada y tratar de ver si esto modifica los parámetros posteriores. Es importante mencionar que esta función carecerá de forma explicita. Se podría pensar que no tenga efecto en la distribución del parámetro media, sin embargo, el comportamiento de la probabilidad del parámetro varianza seguramente sufriría diferencias respecto a la del parámetro posterior de la normal, esto a reserva de hacer los cálculos necesarios.

Omitir las Correlaciones entre las Dimensiones

Otra posibilidad es la de omitir las correlaciones entre las dimensiones y efectuar un proceso univariado como el descrito en [11] para cada dimensión, alterando las cotas que se manejan de manera que en lugar de los elipsoides que se tienen en el modelo actual se tengan rectángulos a fin de agrupar a cada uno de los datos. Las diferencias de este modelo únicamente se dan en el paso de agrupar a los datos en alguno de los clusters, de manera que los datos una vez etiquetados generan submuestras normales cada una con sus parámetros.

Apéndice A

Código en R

A.1. Métodos Comunes

A.1.1. Métodos Jerárquicos

```
dissimilarities <-dist(mtcars)
Dendrogramm     <-hclust(dissimilarities)
grupos          <-cutree(Dendrogramm,k=3)
plot(Dendrogramm,hang=-1)
rect.hclust(Dendrogramm, k=3, border="blue")
```

A.1.2. Métodos de Partición

```
fit <- kmeans(data, 3 , 1)
plot(data,xlab="Desplazamiento",ylab="Cociente del eje trasero")
points(data[which(fit$cluster==1),1],data[which(fit$cluster==1),2],bg="blue",pch=21)
points(data[which(fit$cluster==2),1],data[which(fit$cluster==2),2],bg="red",pch=21)
points(data[which(fit$cluster==3),1],data[which(fit$cluster==3),2],bg="green",pch=21)
fit <- kmeans(data, 3 , 2)
plot(data,xlab="Desplazamiento",ylab="Cociente del eje trasero")
points(data[which(fit$cluster==1),1],data[which(fit$cluster==1),2],bg="blue",pch=21)
points(data[which(fit$cluster==2),1],data[which(fit$cluster==2),2],bg="red",pch=21)
points(data[which(fit$cluster==3),1],data[which(fit$cluster==3),2],bg="green",pch=21)
fit <- kmeans(data, 3 , 3)
plot(data,xlab="Desplazamiento",ylab="Cociente del eje trasero")
points(data[which(fit$cluster==1),1],data[which(fit$cluster==1),2],bg="blue",pch=21)
points(data[which(fit$cluster==2),1],data[which(fit$cluster==2),2],bg="red",pch=21)
points(data[which(fit$cluster==3),1],data[which(fit$cluster==3),2],bg="green",pch=21)
```

A.1.3. Métodos basados en Modelos

```
library(mclust)
library(ellipse)
E.M<-Mclust(Data,3)
plot(Data,type="n")
for(i in 1:3){
points((E.M$parameters)$mean[,i],pch=19,col=i)
points(Data[which(E.M$classification==i),],pch=4,col=i)
lines(ellipse(((E.M$parameters)$variance)$sigma[,,i],
centre=(E.M$parameters)$mean[,i],
level=.7),col=i)
}
E.M<-Mclust(Data)
plot(Data,type="n")
for(i in 1:E.M$G){
points((E.M$parameters)$mean[,i],pch=19,col=i)
```

```
points(Data[which(E.M$classification==i),],pch=4,col=i)
lines(ellipse(((E.M$parameters)$variance)$sigma[,,i],
centre=(E.M$parameters)$mean[,i],
level=.7),col=i)
}
```

A.2. Implementación

A.2.1. Algoritmo

```
##################################### Paqueteria #####################################
install.packages(c("plyr","conics","mvtnorm","msm","stats",
                   "matrixcalc","truncdist","gtools","cluster",
                   "RConics","ellipse"))
library(plyr)
library(conics)
library(mvtnorm)
library(msm)
library(stats)
library(matrixcalc)
library(truncdist)
library(gtools)
library(cluster)
library(RConics)
library(ellipse)

ellipse <- ellipse::ellipse
join <- plyr::join
tcrossprod <-base::tcrossprod

##################### Inicializacion y definicion de variables ######################

M<-matrix(c(1,0,0,1),2,2)
set.seed(5)

#datos<-data.frame(rmvnorm(25,c(0,0),M))
#datos<-rbind(datos,data.frame(rmvnorm(25,c(10,10),M)))
#datos<-rbind(datos,data.frame(rmvnorm(30,c(0,20),M)))
#datosnorm<-as.matrix(datos)


#Data<-as.matrix(datosnorm)
#Data<-data.frame(ruspini)
#Data<-data.frame(faithful)
#Data<-as.matrix(mtcars[,c(3,5)])

###################### Generar los parametros Wishart iniciales ######################

n<-length(Data[,1])
J<-5
k<-rep(1,J)
# Generate the
Tw<-list()
Tw<-rep(list(matrix(solve(cov(Data)),2,2)),J)
mu<-matrix(rep(colMeans(Data),J),J,2,byrow=T)
v<-rep(2,J)
alpha <- rep(1,J)

########################## Generar los parametros normales ##########################

lambda <-list()
```

```
theta <- matrix(rep(numeric(),2*J),J,2);
w<-rdirichlet(1,alpha)
for(j in 1:J){lambda[[j]]<-matrix(rWishart(1,v[j],Tw[[j]]),2,2)
              theta[j,]<-rmvnorm(1,mu[j,],solve(k[j]*lambda[[j]]))}

########################### Parametros usasdos para graficar ##########################

cols<-c("royalblue4","darkred","black","mediumpurple4","skyblue3",
        "aquamarine4","bisque4","darkolivegreen4","goldenrod4",
        "plum4","seagreen4")[1:J]
innercolors<-c("royalblue1","brown1","gray75","mediumpurple1","skyblue1",
               "aquamarine1","bisque1","darkolivegreen1","goldenrod1",
               "plum1","seagreen1")[1:J]
ellipoints<-1000

################################### Otros Parametros ##################################

m<-rep(0,J);L<-20000
results<-data.frame(matrix(rep(0,n*L),L,n))
InfoCriteria<-data.frame(matrix(rep(Inf,3*L),L,3))
colnames(InfoCriteria)<-c("Akaike","Bayesian","Loglikelihood")
InfoCriteria$Loglikelihood<-rep(-Inf,L)

plot(Data,main=paste("Datos"));
#setwd("Directory")

hist_theta<-list()
hist_lambda<-list()

#######################################################################################
################################ Definicion de Funciones ##############################
#######################################################################################

#################### Funciones para actualizar parametros Wishart ###################

update_mu <- function(Data,mu,means,m,kw,z,variable)
{
  for(j in 1:length(mu[,1]))
  {
    if(m[j]!=0 & variable1[j]==T)
           {mu[j,]<-(kw[j]*mu[j,]+m[j]*means[j,])/(kw[j]+m[j])}
    else if(length(which(z==0))>1)
           {mu[j,]<-colMeans(Data[which(z==0),])}
    else{mu[j,]<-colMeans(Data)}
  }
  return(mu)
}

update_alpha <- function(alpha,m)
{
  alpha<-alpha+m
  alpha[which(m==0)]<-0
  return(alpha)
}

update_Tw<-function(Data,Tw,mu,kw,D,m,z,variable)
{
  for(j in 1:length(Tw))
  {
    if(m[j]>2 & variable[j]==T)
    {
           Tw[[j]]<-update_Tw_j(Tw[[j]],mu[j,],kw[j],D[[j]],m[j])
    }
```

```
    else
    {
      if(length(which(z==0))<=2)
        {Tw[[j]]<-matrix(c(1,0,0,1),2,2)}
      else
        {Tw[[j]]<-length(which(z==0))*
                  matrix(solve(cov(Data[which(z==0),])),2,2)}

    }
  }
  return(Tw)
}

update_Tw_j<-function(Tw,mu,kw,D,m)
{
  ifelse(m<=2,A<-matrix(rep(0,4),2,2),A<-(m-1)*cov(D))
  return(solve(solve(Tw)+A+
           ((m*kw)/(m+kw))*tcrossprod(mu-c(mean(D[,1]),mean(D[,2])))))
}

update_v <- function(v,m,Mpr)
{
  res<-pmax(2,Mpr+m)
  res[which(m==0)]<-2
  #res<-v+m
  #res[which(m==0)]<-rep(2,length(which(m==0)))
  #res[which(m-Mpr< 0)]<-res[which(m-Mpr< 0)]+1
  #res[which(m-Mpr>=0)]<-res[which(m-Mpr>=0)]-1
  #res[which(res<2)]<-rep(2,length(which(res<2)))
  return(res)
}

################### Funciones para actualizar parametros Normales ###################

update_lambda <-function(Tw,v,variable)
{
  for(j in 1:length(lambda))
  {
    lambda[[j]]<-matrix(rWishart(1,v[j],Tw[[j]]),2,2)
  }
  return(lambda)
}

projective_matrix_ellipse<-function(O,B,u)
{
  O<-matrix(O,1,2);B<-matrix(B,2,2)
  return(conicMatrix(c(B[1,1],2*B[2,1],B[2,2],-2*(B[1,1]*O[1]+B[2,1]*O[2]),
              -2*(B[2,1]*O[1]+B[2,2]*O[2]),(O%*%B%*%t(O))-u)))
}

plot_ellipse<-function(O,A,u)
{
  conicPlot(projective_matrix_ellipse(O,A,u))
}

limits<-function(DD,theta,lambda,m,u,vert)
{
  a<--Inf;b<-Inf
  for(i in 1:m)
  {
    A<-matrix(projective_matrix_ellipse(c(DD[i,1],DD[i,2]),lambda,u[i]),3,3)
    if(vert==F)
    {P<-sort(intersectConicLine(A,c(0,1,-theta[2]))[1,])}
```

```
    else
    {P<-sort(intersectConicLine(A,c(1,0,-theta[1]))[2,])}
    a<-max(a,P[1]);b<-min(b,P[2])
  }
  return(list(a,b))
}


update_theta<-function(DD,m,mu,k,theta,lambda,prevLambda,u,iter,variable)
{
  for(j in 1:length(lambda))
  {
    if(m[j]!=0 & variable[j]==T)
    {
      for(l in 1:iter)
      {
        uu<-u[which(z==j)]
        a<-limits(D[[j]],theta[j,],prevLambda[[j]],m[j],uu,F)[[1]];
        b<-limits(D[[j]],theta[j,],prevLambda[[j]],m[j],uu,F)[[2]];
        theta[j,][1]<-rtnorm(1,(mu[j,][1]-
                                  (k[j]*lambda[[j]][1,2]/k[j]*lambda[[j]][2,2])*
                                  (theta[j,][2]-mu[j,][2])),
                               1/(k[j]*lambda[[j]][2,2]),
                               a,b)
        a<-limits(D[[j]],theta[j,],prevLambda[[j]],m[j],uu,T)[[1]];
        b<-limits(D[[j]],theta[j,],prevLambda[[j]],m[j],uu,T)[[2]];
        theta[j,][2]<-rtnorm(1,(mu[j,][2]-
                                  (k[j]*lambda[[j]][1,2]/k[j]*lambda[[j]][1,1])*
                                  (theta[j,][1]-mu[j,][1])),
                               1/(k[j]*lambda[[j]][1,1]),
                               a,b)
      }
    }
    else
        {
            theta[j,]<-rmvnorm(1,mu[j,],solve(lambda[[j]]))
        }
    }
    return(theta)
}

######################### Funciones para generar variables U ########################

generate_U <- function (data,theta,lambda,w,variable)
{
  N<-length(Data[,1]);u<-rep(0,N);z<-rep(0,N);bounds<-rep(Inf,J)
  Da<-replicate(length(theta[,1]),data.frame())
  for(j in 1:J)
  {
    for(k in 1:J)
    {
      if(j!=k & variable[k]==T)
        {
          bounds[j]<-min(c(bounds[j],
                           sqrt(mahalanobis(theta[j,],theta[k,],lambda[[j]]))))
        }
    }
  }
  for(i in 1:N)
  {
    aux<-generate_U_i(Data[i,],theta,lambda,i,w,bounds,variable)
    z[i]<-aux[[1]];u[i]<-aux[[2]]
    if(z[i]!=0){Da[[z[i]]]<-rbind(Da[[z[i]]],Data[i,])}
```

```
  }
  return(list(z,u,Da))
}

generate_U_i <- function(x,theta,lambda,i,w,bounds,variable)
{
  J<-length(theta[,1]);chi<-rep(0,J);
  distances<-rep(0,J);cluster<-0;
  for(j in 1:J)
  {
    #print(variable[j]==T)
    if(variable[j]==T)
    {
      chi[j]<-rtrunc(1,spec="gamma",shape=2,scale=2,a=0,b=bounds[j])
      distances[j]<-mahalanobis(Data[i,],theta[j,],lambda[[j]],inverted=T)
    }
    else
    {
      chi[j]<-NA;
      distances[j]<-Inf
    }
  }
  cluster<-which.min(distances)
  if((distances-chi)[cluster]>0){return(list(0,NA))}
  return(list(cluster,chi[cluster]))
}

##########################################################################################
######################################## Main ############################################
##########################################################################################

for(l in 1:20000)
{
  repeat
  {
    variable1<-rep(TRUE,J)
    for(j in 2:J)
    {
      for(jj in 1:(j-1))
      {
        if(mahalanobis(theta[j,],
                       theta[jj,],
                       lambda[[jj]],
                       inverted=T)<qchisq(.6,4))
        {
          if(dist(rbind(colMeans(Data),theta[j]))<
             dist(rbind(colMeans(Data),theta[jj])))
            {variable1[j]<-F}
          else{variable1[jj]<-F}
        }
      }
    }

    aux<-generate_U(Data,theta,lambda,w,variable1)
    z<-aux[[1]]
    u<-aux[[2]]
    D<-aux[[3]]

    means<-matrix(rep(NA,2*J),J,2);prevM<-m;m<-rep(0,J);
    for(j in 1:J)
       {
          if(length(D[[j]])!=0)
             {
```

```
            means[j,]<-colMeans(D[[j]]);
            m[j]<-length(D[[j]][,1])
            }
     }

  if(sum(m)/n>=.5)
  {
    break
  }
  else
  {
    for(j in 1:J)
    {
      if(variable1[j]==F | m[j]==0)
      {
        if(length(which(z==0))<=1)
          {theta[j,]<-rmvnorm(1,colMeans(Data),cov(Data))}
        else
          {theta[j,]<-rmvnorm(1,colMeans(Data[which(z==0),]),cov(Data))}
      }
    }
  }
}

mh<-matrix(rep(0,n*J),n,J)

for(i in 1:n)
{
  for(j in 1:J)
  {
    mh[i,j]<-(det(lambda[[j]])^(1/2))*
              w[j]*
              exp(-(mahalanobis(Data[i,],theta[j,],lambda[[j]],inverted=T))/2)
  }
}

## Calcular criterios

lkh<-(2*pi)^(-n)*prod(rowSums(mh))

InfoCriteria$Loglikelihood[l]<-(-n*log(2*pi))+sum(log(rowSums(mh)))
InfoCriteria$Akaike[l]<-(-2*InfoCriteria$Loglikelihood[l])+
                          2*length(which(m!=0))*6
InfoCriteria$Bayesian[l]<-(-2*InfoCriteria$Loglikelihood[l])+
                            log(n)*length(which(m!=0))*6
print(paste("Iteracion numero ",l,
            "AIC: ", InfoCriteria$Akaike[l],
            ", BIC: ",InfoCriteria$Bayesian[l],
            ", LLd: ",InfoCriteria$Loglikelihood[l]))

## Actualizar a los parametros posteriores

mu    <-update_mu(Data,mu,means,m,k,z,variable1)
Tw    <-update_Tw(Data,Tw,mu,k,D,m,z,variable1)
v     <-update_v(v,m,prevM)
alpha <-update_alpha(alpha,m)

prevLambda<-lambda

hist_theta[[l]]<-as.data.frame(theta)
hist_lambda[[l]]<-data.frame(lamd_11=numeric(0),
                             lamd_22=numeric(0),
                             lamd_12=numeric(0))
```

```
  for(j in 1:J)
  {
    hist_lambda[[l]][j,1]<-lambda[[j]][1,1]
    hist_lambda[[l]][j,2]<-lambda[[j]][2,2]
    hist_lambda[[l]][j,3]<-lambda[[j]][1,2]
  }

  lambda <-update_lambda(Tw,v,variable1)
  theta  <-update_theta(D,m,mu,k,theta,lambda,prevLambda,u,25,variable1)
  w      <-rdirichlet(1,alpha)
  alpha  <-m+1;k<-m/2
  k[which(m==0)]<-1
  results[l,]<-z
}
```

A.2.2. Pruebas Numéricas

```
##############################################################################
######################################## Final ###############################
##############################################################################


row.names(head(InfoCriteria[order(InfoCriteria$Akaike),],10))
row.names(head(InfoCriteria[order(InfoCriteria$Bayesian),],5))
row.names(head(InfoCriteria[order(-InfoCriteria$Loglikelihood),],5))

AkaikeResults<-data.frame(iHCL=row.names(head(
InfoCriteria[order(InfoCriteria$Akaike),],10)),
AkaikeValue=head(
InfoCriteria[order(InfoCriteria$Akaike),1],10))
AkaikeResults<-cbind(AkaikeResults,zeros=rowSums(results[AkaikeResults$i,]==0))

BayesianResults<-data.frame(i=row.names(head(
InfoCriteria[order(InfoCriteria$Bayesian),],10)),
BayesianValue=head(InfoCriteria[order(
InfoCriteria$Bayesian),2],10))
BayesianResults<-cbind(BayesianResults,
zeros=rowSums(results[BayesianResults$i,]==0))

LoglikelihoodResults<-data.frame(i=row.names(head(
InfoCriteria[order(-InfoCriteria$Loglikelihood),],10)),
LoglikelihoodValue=head(
InfoCriteria[order(-InfoCriteria$Loglikelihood),3],10))
LoglikelihoodResults<-cbind(LoglikelihoodResults,
zeros=rowSums(results[LoglikelihoodResults$i,]==0))
AkaikeResults
BayesianResults
LoglikelihoodResults

#######################
#######################
####### Testing #######

#selected<-c(7139,14719,7140,7138,9261)
#selected<-c(12201,19472,5883,525,18763)
#selected<-c(16825,5886,13154,17901,149)
InfoCriteria[selected,]

for(s in 1:5){
test<-selected[s]
InfoCriteria[test,]

```

```
z<-as.numeric(results[test,])
z_ord<-z
for(i in 1:n) ## j variable represents the column
{
z_ord[i]<-ifelse(z[i]==0,0,
ifelse(i==1,1,
ifelse(match(z[i],z)==i,
max(z_ord[1:i-1])+1,
z_ord[match(z[i],z)])))
}

def_J<-max(z_ord)
def_theta<-data.frame(theta_X=rep(0,def_J),theta_Y=rep(0,def_J))
def_lambda<-list()

pdf(paste("mtcars_resultado1_",test,".pdf",sep=""))
plot(Data,xlab="X",ylab="Y")
points(Data,bg="white",pch=21);
for(j in 1:def_J){
clr=hsv(j/def_J, 1 ,.5)
points(Data[which(z_ord==j),1],Data[which(z_ord==j),2],bg=clr,pch=21)
}
dev.off()

as.numeric(row.names(hist_theta[[test]][unique(as.numeric(results[test,])),]))
for(j in 1:def_J)
{
def_theta[j,]   <- as.matrix(hist_theta[[test]][unique(z[which(z!=0)])[j],])
def_lambda[[j]] <- matrix(c(hist_lambda[[test]][unique(z[which(z!=0)])[j],1],
hist_lambda[[test]][unique(z[which(z!=0)])[j],3],
hist_lambda[[test]][unique(z[which(z!=0)])[j],3],
hist_lambda[[test]][unique(z[which(z!=0)])[j],2]),2,2)
}

pdf(paste("mtcars_resultado2_",test,".pdf",sep=""))
plot(Data,xlab="X",ylab="Y")
for(j in def_J:1)
{
for(t in 1:480)
{
lines(ellipse(solve(def_lambda[[j]]),
centre=c(def_theta[j,1],
def_theta[j,2]),
level=(t/480)),
col=hsv(j/def_J,
(t/600) +.2 ,1))
}
}

points(Data,pch=20,cex=1)

for(j in 1:def_J){
clr=hsv(j/def_J, .9 ,.7)
points(def_theta[j,],col=clr,pch=20,cex=1)
points(def_theta[j,],pch=21,cex=1)
}
dev.off()
}

print(InfoCriteria[selected,])
for(s in 1:5){
print(paste("Los valores de Theta con l=",selected[s]))
print(hist_theta[[selected[s]]][unique(as.numeric(results[selected[s],])),])
```

```
109 print(paste("Los valores de Lambda con l=",selected[s]))
110 print(hist_lambda[[selected[s]]][unique(as.numeric(results[selected[s],
111 ])),c(1,3,3,2)])
112 }
```

.

Apéndice B

Conjuntos de datos

A continuación se presentan los conjuntos de datos utilizados en el documento.

B.1. Datos mtcars

	mpg	cyl	disp	hp	drat	wt	qsec	vs	am	gear	carb
Mazda RX4	21.00	6.00	160.00	110.00	3.90	2.62	16.46	0.00	1.00	4.00	4.00
Mazda RX4 Wag	21.00	6.00	160.00	110.00	3.90	2.88	17.02	0.00	1.00	4.00	4.00
Datsun 710	22.80	4.00	108.00	93.00	3.85	2.32	18.61	1.00	1.00	4.00	1.00
Hornet 4 Drive	21.40	6.00	258.00	110.00	3.08	3.21	19.44	1.00	0.00	3.00	1.00
Hornet Sportabout	18.70	8.00	360.00	175.00	3.15	3.44	17.02	0.00	0.00	3.00	2.00
Valiant	18.10	6.00	225.00	105.00	2.76	3.46	20.22	1.00	0.00	3.00	1.00
Duster 360	14.30	8.00	360.00	245.00	3.21	3.57	15.84	0.00	0.00	3.00	4.00
Merc 240D	24.40	4.00	146.70	62.00	3.69	3.19	20.00	1.00	0.00	4.00	2.00
Merc 230	22.80	4.00	140.80	95.00	3.92	3.15	22.90	1.00	0.00	4.00	2.00
Merc 280	19.20	6.00	167.60	123.00	3.92	3.44	18.30	1.00	0.00	4.00	4.00
Merc 280C	17.80	6.00	167.60	123.00	3.92	3.44	18.90	1.00	0.00	4.00	4.00
Merc 450SE	16.40	8.00	275.80	180.00	3.07	4.07	17.40	0.00	0.00	3.00	3.00
Merc 450SL	17.30	8.00	275.80	180.00	3.07	3.73	17.60	0.00	0.00	3.00	3.00
Merc 450SLC	15.20	8.00	275.80	180.00	3.07	3.78	18.00	0.00	0.00	3.00	3.00
Cadillac Fleetwood	10.40	8.00	472.00	205.00	2.93	5.25	17.98	0.00	0.00	3.00	4.00
Lincoln Continental	10.40	8.00	460.00	215.00	3.00	5.42	17.82	0.00	0.00	3.00	4.00
Chrysler Imperial	14.70	8.00	440.00	230.00	3.23	5.34	17.42	0.00	0.00	3.00	4.00
Fiat 128	32.40	4.00	78.70	66.00	4.08	2.20	19.47	1.00	1.00	4.00	1.00
Honda Civic	30.40	4.00	75.70	52.00	4.93	1.61	18.52	1.00	1.00	4.00	2.00
Toyota Corolla	33.90	4.00	71.10	65.00	4.22	1.83	19.90	1.00	1.00	4.00	1.00
Toyota Corona	21.50	4.00	120.10	97.00	3.70	2.46	20.01	1.00	0.00	3.00	1.00
Dodge Challenger	15.50	8.00	318.00	150.00	2.76	3.52	16.87	0.00	0.00	3.00	2.00
AMC Javelin	15.20	8.00	304.00	150.00	3.15	3.44	17.30	0.00	0.00	3.00	2.00
Camaro Z28	13.30	8.00	350.00	245.00	3.73	3.84	15.41	0.00	0.00	3.00	4.00
Pontiac Firebird	19.20	8.00	400.00	175.00	3.08	3.85	17.05	0.00	0.00	3.00	2.00
Fiat X1-9	27.30	4.00	79.00	66.00	4.08	1.94	18.90	1.00	1.00	4.00	1.00

B.2. Datos Generados Normales

	mpg	cyl	disp	hp	drat	wt	qsec	vs	am	gear	carb
Porsche 914-2	26.00	4.00	120.30	91.00	4.43	2.14	16.70	0.00	1.00	5.00	2.00
Lotus Europa	30.40	4.00	95.10	113.00	3.77	1.51	16.90	1.00	1.00	5.00	2.00
Ford Pantera L	15.80	8.00	351.00	264.00	4.22	3.17	14.50	0.00	1.00	5.00	4.00
Ferrari Dino	19.70	6.00	145.00	175.00	3.62	2.77	15.50	0.00	1.00	5.00	6.00
Maserati Bora	15.00	8.00	301.00	335.00	3.54	3.57	14.60	0.00	1.00	5.00	8.00
Volvo 142E	21.40	4.00	121.00	109.00	4.11	2.78	18.60	1.00	1.00	4.00	2.00

Datos Generados Normales								
i	X	Y	*i*	X	Y	*i*	X	Y
1	-0.84	1.38	2	-1.26	0.07	3	1.71	-0.60
4	-0.47	-0.64	5	-0.29	0.14	6	1.23	-0.80
7	-1.08	-0.16	8	-1.07	-0.14	9	-0.60	-2.18
10	0.24	-0.26	11	0.90	0.94	12	1.47	0.71
13	0.82	-0.29	14	1.42	1.50	15	-0.66	-0.85
16	0.32	1.11	17	2.22	1.22	18	1.48	0.95
19	-1.01	-2.00	20	-1.76	-0.14	21	1.55	-0.80
22	-0.07	1.90	23	-0.46	0.56	24	-0.89	-0.46
25	-0.72	-0.07	26	11.46	10.19	27	11.02	9.41
28	9.89	9.08	29	10.75	9.89	30	9.94	10.23
31	8.86	10.85	32	9.42	10.50	33	9.24	9.66
34	7.90	9.70	35	8.73	9.72	36	9.80	9.77
37	10.35	10.03	38	10.41	9.84	39	10.97	10.12
40	10.19	9.44	41	10.50	8.26	42	10.98	9.98
43	10.68	9.29	44	12.39	9.53	45	9.92	9.48
46	10.93	8.94	47	10.56	10.90	48	10.99	10.38
49	9.65	9.46	50	9.82	9.94	51	-2.00	21.14
52	0.68	20.21	53	-0.06	20.89	54	-0.23	18.03
55	-0.75	21.28	56	-0.95	21.62	57	2.60	20.14
58	-1.35	20.80	59	-1.55	20.46	60	0.05	19.80
61	1.17	20.88	62	-1.32	18.36	63	1.06	20.29
64	-0.40	21.24	65	-1.37	18.56	66	1.35	18.02
67	-1.24	19.90	68	0.73	20.46	69	0.29	18.93
70	0.65	20.30	71	-0.80	19.97	72	2.18	20.96
73	-0.31	19.58	74	0.10	19.77	75	-1.42	19.61
76	0.95	20.75	77	-0.52	20.81	78	-0.61	21.24
79	-0.34	21.20	80	-0.44	20.19			

B.3. Datos Ruspini

Datos Ruspini					
x	y	x	y	x	y
4	53	41	150	98	124
5	63	38	145	99	119
10	59	38	143	99	128
9	77	32	143	10 1	115
13	49	34	141	108	111
13	69	44	156	110	111
12	88	44	149	108	116
15	75	44	143	111	126
18	61	46	142	115	117
19	65	47	149	117	115
22	74	49	152	70	4
27	72	50	142	77	12
28	76	53	144	83	21
24	58	52	152	61	15
27	55	55	155	69	15
28	60	54	124	78	16
30	52	60	136	66	18
31	60	63	139	58	13
32	61	86	132	64	20
36	72	85	115	69	21
28	147	85	96	66	23
32	149	78	94	61	25
35	153	74	96	76	27
33	154	97	122	72	31
38	151	98	116	64	30

Bibliografía

[1] H. Akaike. *Information Theory and an Extension of the Maximum Likelihood Principle*. Second International Symposium on Information Theory. 1963.

[2] T.W. Anderson. *An Introduction to Multivariate Statistical Analysis*. Wiley Series in Probability and Statistics. Wiley, 2003.

[3] Christopher M. Bishop. *Pattern Recognition and Machine Learning (Information Science and Statistics)*. Springer-Verlag New York, Inc., Secaucus, NJ, USA, 2006.

[4] W. M. Bolstad.

[5] D. A. Brannan, M. F. Esplen, and J. J. Gray. *Geometry*. Cambridge University Press, Cambridge, 2 edition, 12 2011.

[6] G. Casella and R. Berger.

[7] G. Casella and E. I. George. Explaining the gibbs sampler. *The American Statistician*, 46(3):167–174, 1992.

[8] G. Casella and E. I. George. Explaining the gibbs sampler. *The American Statistician*, 46:167–174, 1992.

[9] R. Courant. *Differential and Integral Calculus*, volume 2. Blackie & Son, 1961.

[10] B. Everitt and D. J. Hand. *Finite Mixture Distributions*. Monographs on applied probability and statistics. Chapman and Hall, 1981.

[11] R. Fuentes-García and S. G. Walker. A new approach to classification. *Journal of Applied Statistics*, 37(1):137–146, 2010.

[12] T. Hastie, R. Tibshirani, and J. Friedman. *The Elements of Statistical Learning*. Springer Series in Statistics. Springer New York Inc., New York, NY, USA, 2001.

[13] H. V. Henderson and P. F. Velleman. Building multiple regression models interactively. *Biometrics*, 37:391–411, 1981.

[14] L. Kaufman and P. Rousseeuw. *Finding groups in data : an introduction to cluster analysis*. Wiley series in probability and mathematical statistics. Wiley, New York, 1990. A Wiley-Interscience publication.

[15] H. A. L. Kiers. Multivariate analysis, part 2: Classification, covariance structure, and repeated measurements, by w.j. krzanowski and f.h.c. marriott. *Journal of Classification*, 15(2), 1998.

[16] W. J. Krzanowski, editor. *Principles of Multivariate Analysis: A User's Perspective*. Oxford University Press, Inc., New York, NY, USA, 1988.

[17] C. H. Lehmann. *Analytic geometry*. J. Wiley & sons, inc., 1942.

[18] B. G. Lindsay. *Mixture Models: Theory, Geometry, and Applications*. Conference Board of the Mathematical Sciences: NSF-CBMS regional conference series in probability and statistics. Institute of Mathematical Statistics, 1995.

[19] S. M. Lynch. *Introduction to Applied Bayesian Statistics and Estimation for Social Scientists*. Springer, 2007.

[20] J. M. Marin, K. L. Mengersen, and C. Robert. Bayesian modelling and inference on mixtures of distributions. 2005.

[21] A. M. F. Mood, F. A. Graybill, and D. C. Boes. *Introduction to the Theory of Statistics*. International Student edition. McGraw-Hill, 1974.

[22] A. I. Ramírez Galarza. *Geometría analítica : una introducción a la geometría.* Temas de matemáticas. 3ª edición, 2013.

[23] E. H. Ruspini. Numerical methods for fuzzy clustering. *Inf. Sci.*