**Universidad de Valladolid**

**Facultad de Ciencias**

Máster en matemáticas

TRABAJO DE FIN DE MÁSTER:

# ANÁLISIS CLUSTER Y CLASIFICACIÓN BASADOS EN MODELOS

AUTORA:
Carla Perucha Jurjo

TUTORES:
Carlos Matrán Bea, Luis Ángel García Escudero

Julio 2023

# Índice

# 1. Introducción

En el contexto actual, donde la disponibilidad de datos masivos y la complejidad de las aplicaciones en diversos campos han aumentado exponencialmente, el Análisis Cluster y la Clasificación se han convertido en herramientas fundamentales para descubrir estructuras subyacentes y organizar grandes conjuntos de datos de manera significativa. Estas técnicas multivariantes se encargan del estudio de métodos y algoritmos ideados para agrupar objetos en conjuntos de acuerdo con sus características internas, a la par que asignar nuevos individuos a estos grupos en base a características compartidas.

Una gran cantidad de los enfoques del Aprendizaje Automático se sustentan en técnicas heurísticas para particionar los datos, pudiendo dar la sensación de que las similitudes intrínsecas entre individuos no se basan en un modelo probabilístico subyacente. Sin embargo, las agrupaciones que surgen en un conjunto de datos dependen fuertemente del método de agrupamiento elegido, siendo su actuación muy dependiente de la morfología de las observaciones. El Análisis Cluster y la Clasificación basados en modelos incorporan modelos estadísticos de los datos haciendo posible modelar distribuciones subyacentes complejas y capturar diferentes formas de agrupamiento.

En este documento presentaremos los modelos finitos de mezclas, incidiendo especialmente en la familia de modelos en los que las componentes de mixtura consideradas son gaussianas. Por su naturaleza, estos modelos admiten un estudio matemático que permite demostrar propiedades como la existencia o la consistencia de soluciones, realizar comparaciones entre ellos, resaltar sus bondades y prever su comportamiento y sus inconvenientes o sus limitaciones.

En las últimas dos décadas, se han producido progresos significativos en el problema de ajustar modelos de mezclas a conjuntos de datos. Muchos de los avances más destacados en este campo se han producido hace apenas 20 años, en particular la aplicación del algoritmo de Expectation-Maximization (EM) para el cálculo del estimador máximo verosímil revolucionó el modelado a través de modelos de mezclas y catapultó su uso en diversas aplicaciones.

En este trabajo hemos procurado introducir el modelo de mezclas junto con sus más recientes avances y retos metodológicos respecto a la selección de modelos, estimación del número de componentes e inicialización del algoritmo EM, planteados en [2]. Además, hemos completado esta exposición con el estudio de algunos de los resultados matemáticos más interesantes ya que a menudo no encontramos resultados teóricos en los manuales sobre el modelo de mezclas.

Por último, destacar que en las matemáticas, en especial en el Análisis Cluster, y en general en Clasificación, una visualización por medio de ejemplos gráficos siempre resulta de gran ayuda para conseguir comprender mejor qué está sucediendo. Para ello, hemos empleando el lenguaje de programación R, ampliamente utilizado en Estadística, donde existen múltiples paquetes que implementan algoritmos de clustering y funciones para visualizar sus resultados. En este documento se emplean los siguientes:

- **rgl:** Proporciona funciones para visualizaciones interactivas en 3D, ya sea a la hora de proporcionar gráficos (plot3d()) como funciones para contruir representaciones de objetos geométricos

tridimensionales (ellipse3d()). Los resultados se visualizan a través de una ventana denominada RGL device 5 que permite efectuar rotaciones para poder explorar las diferentes perspectivas de los gráficos generados.

- **mclust:** Permite el ajuste de modelos de mezclas gaussianos por medio del algoritmo EM para el Análisis Cluster basado en modelos, clasificación y estimación de funciones de densidad, incluyendo la regularización Bayesiana, inferencia basada en resampling y reducción de dimensiones para una mejor visualización.

- **tclust:** Incluye funciones de recorte propias del Análisis Cluster robusto. La función tclust() busca $k$ (o menos) clusters con diferentes matrices de covarianza en un conjunto de datos dado. Se pueden especificar dos parámetros muy interesantes: $restr.fact$, que restringe el ratio entre los autovalores de las matrices de covarianza y $alpha$, que permite un porcentaje de observaciones a recortar o eliminar a la hora de ajustar el modelo con el fin de hacer la estimación más robusta.

- **mixtools:** Permite el análisis de los modelos finitos de mixtura para varias familias de densidades paramétricas, entre ellas la gaussiana. Por medio de funciones como normalmixEM() o mvnormalmixEM(), se puede estudiar el comportamiento del algoritmo EM al ajustar el MLE (como por ejemplo el valor de la función de log-verosimilitud en cada iteración, número de iteraciones...)

- **MASS:** Recoge numerosos conjuntos de datos y las distribuciones de probabilidad más utilizadas en Estadística. En concreto, con la función mvrnorm() se pueden generar muestras aleatorias de una normal multivariante especificando el vector de medias y la matriz de covarianzas.

- **Rmixmod:** Interfaz del software "MIXMOD" para clasificación supervisada, semi-supervisada y no supervisada por medio del modelo de mixturas. Contiene la función mixmodCluster() que ofrece la posibilidad de hallar el iterante inicial del EM en el con el algoritmo smallEM. También está presente mixmodGaussianModel(), donde podemos introducir una lista de modelos según su clasificación VSO para el ajuste de los datos con el fin de compararlos entre sí.

# 2. Presentación del análisis clúster basado en modelos

Entre las numerosas técnicas que existen para abordar el problema de discriminar individuos a partir de la medición de algunas de sus características, el modelo finito de mixturas resulta una herramienta muy poderosa para identificar y modelar estas subpoblaciones.

En esta sección presentaremos unos primeros ejemplos donde se verá patente el interés del método y daremos unas primeras definiciones relativas al problema de ajuste de un modelo de mixturas. Por otro lado, expondremos algunas de las situaciones en las que podría ser interesante utilizar esta metodología. Como ya hemos comentado, los ejemplos expuestos a lo largo de este trabajo serán procesados con el lenguaje de programación $R$, de uso extendido en el ámbito de la Estadística.

## 2.1. Planteamiento intuitivo del modelo finito de mixturas

Supongamos que contamos con un conjunto de datos constituido por 600 observaciones de individuos a los que hemos medido valores en las variables $V1$ y $V2$. Mostramos una representación bivariante de los datos a continuación:

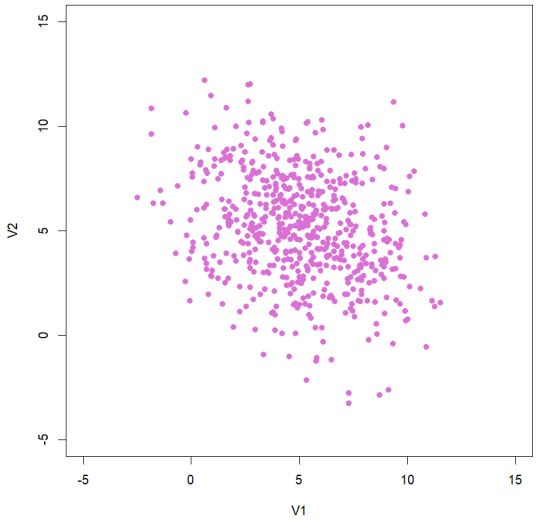



Figura 1: Representación gráfica de las observaciones en el plano para el primer conjunto de datos

Las observaciones de este conjunto de datos están distribuidas en el plano formando una figura elipsoidal. Además, existe una mayor densidad de puntos a medida que nos acercamos al centro de la nube de puntos. Construimos dos histogramas (uno para cada variable), donde dividimos el rango de valores que toman las observaciones en intervalos y mostramos la frecuencia con la que los datos caen en cada intervalo.

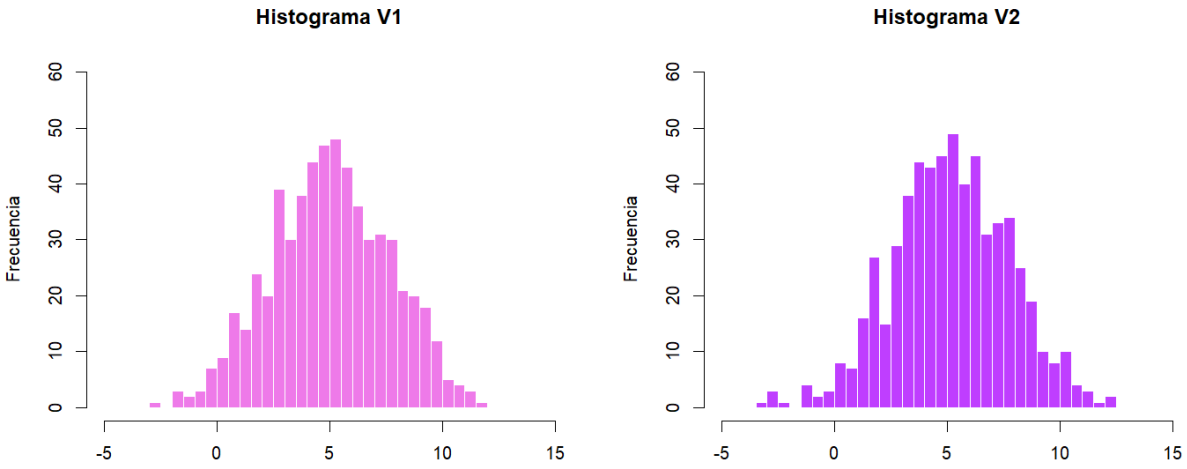



Figura 2: Histogramas para ambas variables

Observamos que ambos histogramas tienen una forma bastante simétrica y acampanada que, junto con la información que recogíamos en el primer gráfico, nos hace pensar que tal vez una distribución normal bivariante con los parámetros adecuados sea capaz de describir la variabilidad de las observaciones. Esto puede ser interesante ya que la distribución gaussiana es ampliamente utilizada en Estadística, por lo que el modelo resultaría simple y familiar, a la par que nos permitiría utilizar propiedades conocidas de esta distribución relativas a la inferencia y a la predicción.

Un posible ajuste de los datos mediante una distribución normal bivariante podría ser mediante una función de densidad normal con vector de medias $\mu = (5{,}03 \quad 5{,}15)$ y matriz de covarianzas $\Sigma = \begin{pmatrix} 6{,}93 & -1{,}87 \\ -1{,}87 & 6{,}78 \end{pmatrix}$ estimadas por el método de máxima verosimilitud. En el gráfico siguiente, representamos el conjunto de datos señalando con una "X" el centro de la distribución y con una línea verde la elipse de confianza del 95 % asociada a $\Sigma$ :

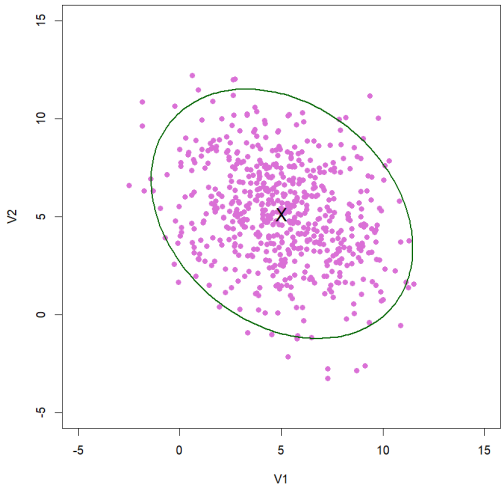



Figura 3: Ajuste por medio de una distribución Gaussiana

Consideremos ahora otro conjunto de datos constituido por otras 600 observaciones bivariantes. Si nos fijamos en su representación en el plano como puntos de $\mathbb{R}^2$, observamos que los individuos toman valores bastante diferentes en las variables V1 y V2. Además, al dibujar la función de densidad de los datos para cada variable vemos que claramente son bimodales (observamos dos cumbres, dos modas).

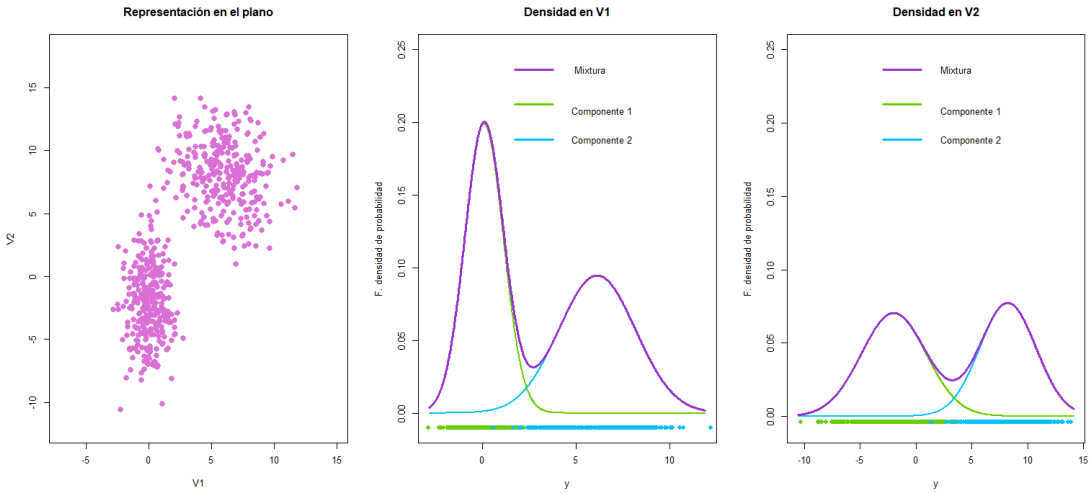



Figura 4: Representación bivariante de las observaciones del segundo conjunto de datos

En el gráfico, se han dibujado por separado las dos densidades normales junto con la densidad de probabilidad total, que es una combinación lineal de dos densidades normales ponderadas. Los puntos que aparecen en la parte inferior del gráfico son una muestra de esta distribución y están coloreados en función de qué componente los ha generado. Observamos además que la función de densidad es más baja entre las dos jorobas, por lo que existe algo de separación entre las dos componentes en cada una de las dos variables. Sin embargo, esta separación no es total y existe incertidumbre a la hora de saber qué puntos provienen de cada una de las componentes.

Parece razonable tratar de extender la idea anterior de resumir la distribución y variabilidad de los datos al caso en el que contamos con $G$ distribuciones en lugar de solo una. Este procedimiento se denomina modelo de mixturas y trata de representar cada subgrupo o componente del conjunto de datos mediante una distribución de probabilidad, en el ejemplo que estamos planteando esta sería la distribución Gaussiana.

Si consideramos el siguiente conjunto de datos, por la colocación que tienen los puntos en el plano, parece coherente pensar que existen tres subploblaciones, por lo que tratamos de describir a los individuos por medio de tres distribuciones Gaussianas. El objetivo será elegir los parámetros de las funciones de densidad y las proporciones en las que aparecen con el fin de lograr la mejor representación de los datos.

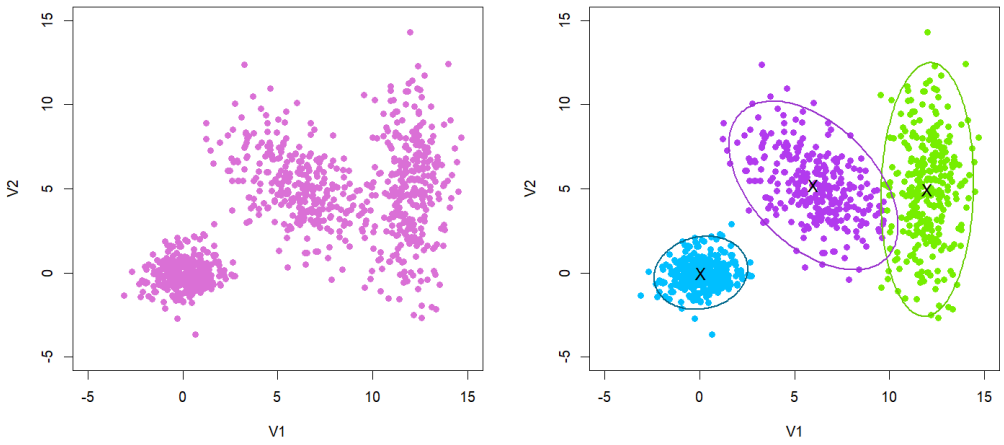



Figura 5: Tres componentes para ajustar los datos

En las situaciones que hemos expuesto, las agrupaciones podrían parecer triviales y factibles de detectar con casi cualquier procedimiento de Análisis cluster ya que existen diferencias bastante evidentes entre los individuos de distintas poblaciones. Sin embargo, podemos encontrarnos con configuraciones mucho más complejas donde no es evidente una separación y donde las observaciones se presentan como una mezcla indistinguible sin el conocimiento de los grupos. Un ejemplo de esto

podría ser el siguiente conjunto con 600 observaciones donde, por sus características, el modelo de mixturas puede desentrañar su estructura latente.

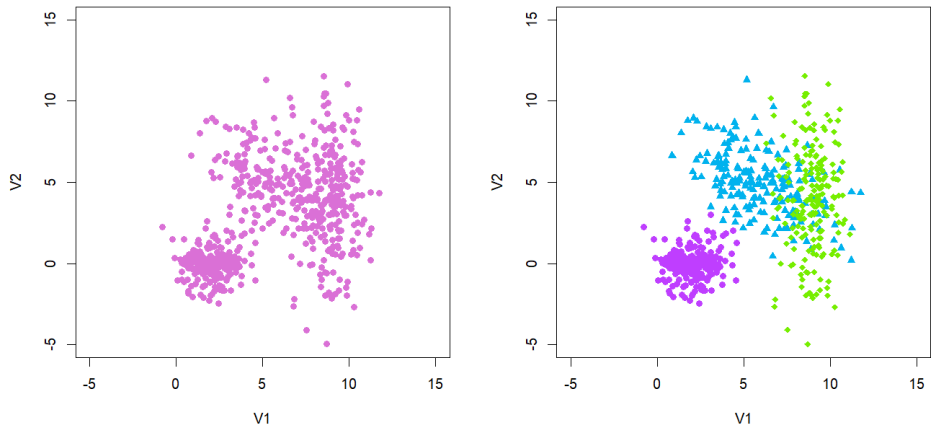



Figura 6: Un conjunto de datos donde los grupos no son tan triviales

## 2.2. Formulación del modelo finito de mixturas

El objetivo del modelo de mezclas es estimar los parámetros de cada componente y las proporciones con las que se presentan para explicar de la mejor manera posible la variablidad de los datos. Concretamos un poco más a qué nos referimos con todo esto en la siguiente definición:

**Definición 2.1.** *Sea $A = \{x_1, .., x_n\}$ un conjunto con $n$ observaciones multivariantes pertenecientes a $\mathbb{R}^d$, de tal manera que $x_i = (x_{i,1}, .., x_{i,d})$. Sea $G$ un entero positivo y $\pi_j$ pesos tales que $\pi_j \geq 0$ para $j = 1, .., G$ con $\sum_{j=1}^{G} \pi_j = 1$. Sea $\mathscr{F} = \{f(\cdot;\theta)/\theta \in \Theta\}$ una familia de funciones de densidad con vector de parámetros $\theta$ en el correspondiente espacio de parámetros $\Theta$. Un modelo finito de mezclas es un modelo estadístico que representa la función de densidad de una observación $x_i$, $f(x_i)$, como una combinación lineal ponderada de $G$ funciones de densidad $f_j(\cdot;\theta_j) \in \mathscr{F}$:*

$$f(x_i) = \sum_{j=1}^{G} \pi_j f(x_i;\theta_j) \tag{1}$$

*Las funciones de densidad $f(\cdot;\theta_j) = f_j$ se denominan componentes de la mixtura y los valores $\pi_j$ proporciones de mezcla o pesos.*

**Observación 2.2.** *Dado que las funciones $f_1(\cdot;\theta_1), .., f_G(\cdot;\theta_G)$ son densidades, es claro que la expresión (1) define una función de densidad.*

El vector que recoge los parámetros desconocidos a estimar es $\Psi = (\pi_1, .., \pi_{G-1}, \theta_1, .., \theta_G)$, donde hemos omitido $\pi_G$ por ser redundante ya que las proporciones $\pi_j$ suman uno. Denotamos por $\Theta_c$

el espacio paramétrico del vector $\Psi$. Los parámetros $\theta_1, .., \theta_G$ son distintos entre sí para considerar diferentes elementos de la familia de densidades.

En esta formulación del problema de mezclas, el número de componentes $G$ se considera fijo. Sin embargo, en múltiples ocasiones prácticas el valor de $G$ deberá ser inferido a través de los datos disponibles. A menudo, las funciones de densidad consideradas son las gaussianas y por lo tanto $f_j$ es la función de densidad normal multivariante $\phi_j$, parametrizada por su vector de medias $\mu_j$ y su matriz de covarianzas $\Sigma_j$. Es decir, en este caso $\theta_j = (\mu_j, \Sigma_j)$ y $\Theta = \{(\mu, \Sigma) : \mu \in \mathbb{R}^p, \Sigma \in M_{p\times p}\}$, donde $M_{p\times p}$ denota el subespacio de las matrices cuadradas definidas positivas de dimensión $p$.

Una forma de visualizar conceptualmente el modelo de mezclas es considerar que el vector $x_i$ ha sido generado por la componente $j$-ésima con probabilidad $\pi_j$, y que la función de densidad de $x_i$ sabiendo que proviene de la componente $j$ es $f_j$.

De este modo, si $\{x_1, .., x_n\}$ es una realización de $n$ vectores aleatorios $p-$dimensionales $X_1, .., X_n$ definidos en el mismo espacio probabilístico, consideramos $Z_i$ una variable aleatoria categórica que toma los valores $1, .., G$ con probabilidades $\pi_1, .., \pi_G$ respectivamente. Suponemos que la densidad condicional de $X_i$ dado que $Z_i = j$ es $f_j(x_i)$ . En este contexto, la variable $Z_i$ se puede ver como la etiqueta del vector $X_i$ que nos informa de qué componente lo generó. Es conveniente que trabajemos con un vector $G$-dimensional $Z_i$ de etiquetas donde sus componentes $Z_{ij}$ son cero o uno dependiendo de si el origen de la observación $x_i$ en la mixtura proviene de la $j$-ésima componente.

Resulta entonces que $Z_i$ tiene una distribución multinomial que consiste en elegir entre $G$ posibles categorías con probabilidades $\pi_1, .., \pi_G$. Es decir

$$P(Z_i = z_i) = \pi_1^{z_{i1}} \pi_2^{z_{i2}} ... \pi_G^{z_{iG}}$$

Es decir, el modelo de mezclas tiene una relación directa con aquellas situaciones en las que las observaciones pueden ser identificadas o provienen de una población con $G$ grupos que se presentan en proporciones $\pi_1, .., \pi_G$. Por ejemplo, en el contexto del Análisis Cluster, las observaciones se agrupan en $G$ clusters asignando cada individuo a la componente que lo originó.

A pesar de que en un sentido físico haya ocasiones en las que no será de todo apropiado considerar que existen etiquetas desconocidas asociadas a las observaciones, este enfóque es muy útil ya que nos permite utilizar el algoritmo EM, una herramienta muy potente para computar el estimador de máxima verosimilitud (MLE) de la que hablaremos más adelante.

### 2.2.1. Estimación de los parámetros del modelo

A lo largo de los años, se han considerado multitud de enfoques para abordar el problema de estimar la densidad de la mixtura, entre ellos métodos gráficos, el método de momentos, procedimientos Bayesianos... La gran variedad de técnicas consideradas responde al hecho de que no existen expresiones cerradas de los parámetros a estimar, por lo que habitualmente deben ser calculados iterativamente. Sin embargo, como ya mencionábamos anteriormente, el algoritmo EM permite el cálculo directo del MLE, por lo que el método de máxima verosimilitud es el procedimiento que detallaremos en este documento.

El principio de máxima verosimilitud tiene como objetivo encontrar los valores de los parámetros que maximizan la probabilidad de aparición de los datos observados. Este procedimiento se utiliza ampliamente en estadística no solo por su capacidad para encontrar los valores de los parámetros que maximizan la probabilidad de observar los datos registrados, sino que además proporciona estimaciones consistentes y eficientes a medida que aumenta el tamaño de la muestra bajo ciertas restricciones. Es decir, a medida que se tienen más datos, la estimación dada por máxima verosimilitud se acerca cada vez más a los valores reales del parámetro y además tiene menor varianza que otras estimaciones consistentes.

La función de verosimilitud en el problema de ajustar un modelo de mixturas al conjunto de observaciones $A$ por medio de $G$ funciones pertenecientes a la familia $\mathscr{F}$ tiene la siguiente expresión

$$\mathscr{L}(\Psi, A) = \prod_{i=1}^{n} \left( \sum_{j=1}^{G} \pi_j f(x_i; \theta_j) \right), \ f \in \mathscr{F}, \Psi = (\pi_1, .., \pi_G, \theta_1, .., \theta_G)$$

A menudo, dado que la función logaritmo es creciente, se considera el problema equivalente de maximizar el logaritmo de la función de verosimilitud, esto es,

$$L(\Psi, A) = \log(\mathscr{L}(\Psi, A)) = \sum_{i=1}^{n} \left( \log \left( \sum_{j=1}^{G} \pi_j f(x_i, \theta_j) \right) \right)$$

Este enfoque es utilizado a menudo ya que el logaritmo tiene la propiedad de convertir el producto de probabilidades en una suma de logaritmos. Esto evita problemas numéricos asociados con la multiplicación de muchos valores pequeños y facilita el cálculo de la función objetivo. Además, las derivadas parciales de la función objetivo se simplifican en comparación con las de la función de verosimilitud, lo cual resulta verdaderamente interesante ya que estos cálculos suelen ser necesarios en numerosos algoritmos de optimización.

Para maximizar la expresión anterior, buscamos las raices de la ecuación

$$\frac{\partial L(\Psi, A)}{\partial \Psi} = \frac{\partial \left( \sum_{i=1}^{n} \left( \log \left( \sum_{j=1}^{G} \pi_j f(x_i, \theta_j) \right) \right) \right)}{\partial \Psi} = 0 \tag{2}$$

Una cuestión fundamental a tener en cuenta es que rara vez la función de verosimilitud es convexa, por lo que es frecuente que tenga varios máximos locales. A la hora de elegir la raíz de (2) para estimar el vector de parámetros, idealmente querríamos escoger aquella en la que se alcanza el máximo global en las situaciones en las que la función de verosimilitud está acotada en el espacio paramétrico. Sin embargo, esto no siempre ocurre: si consideramos por ejemplo el modelo de mixturas gaussianas con con varianzas diferentes en el caso univariante y matrices de covarianza distintas en el caso multivariante, esta función no está acotada, ya que si alguna de las matrices de covarianzas $\Sigma_j$ verifica que $|\Sigma_j| \to 0$ resulta que $\phi(\cdot; x_i, \Sigma_j) \to \infty$ para cualquier $x_i$ y la función de verosimilitud no estaría acotada.

Debemos tener cuidado por lo tanto con aquellos máximos locales que surgen como consecuencia de elegir parámetros que provocan que el determinante de la matriz de covarianza de algunos grupos sea muy pequeño (aunque no nulo). Estas agrupaciones contienen pocas observaciones o su disposición es tal que están prácticamente contenidas en un subespacio de dimensión menor.

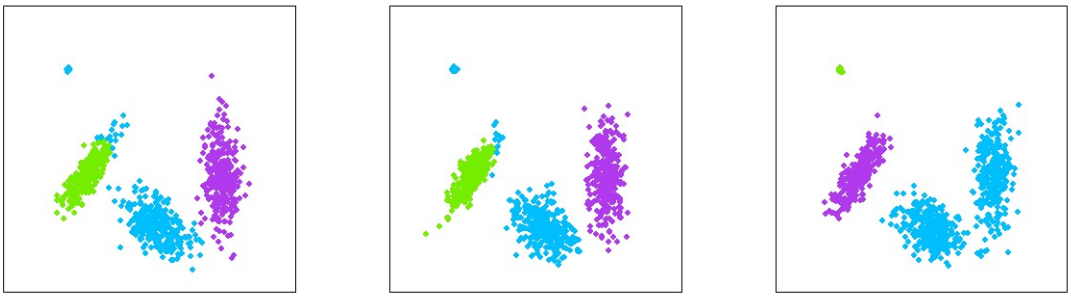


Figura 7: Diferentes soluciones al considerar distintas muestras de una misma distribución

En esta imagen observamos la gran variabilidad que se produce en el ajuste del modelo de mezclas gaussiano a la distribución a medida que generamos diferentes muestras. Esto se debe a que arriba a la izquierda encontramos un pequeño grupo de observaciones alejadas del resto pero muy cercanas entre sí, de tal manera que la componente que las describiría tendría el determinante de la matriz muy proximo a cero. En estas situaciones, el método no resulta robusto y pequeñas fluctuaciones en los datos generados provoca un cambio drástico en los parámetros elegidos, por lo que evidentemente no podremos asegurar la existencia de estimadores consistentes.

Vemos por lo tanto que si no consideramos ninguna restricción adicional, podemos caer en soluciones engañosas o espurias dado que existen múltiples caminos en el espacio paramétrico a través de los cuales la función de verosimilitud no está acotada. En estos casos, el estimador máximo verosímil de $\Psi$ no existe o al menos no como aquel vector de parámetros que maximiza globalmente la verosimilitud.

Sin embargo, en algunas casos existe el MLE de $\Psi$ pero como un maximizador local. Imponiendo algunas condiciones de regularidad, podremos asegurar que existe una sucesión de raíces de la ecuación de verosimilitud con propiedades de consistencia y eficiencia asintótica. Más adelante, veremos que bajo ciertas restricciones sobre las matrices de covarianzas y la distribución de probabilidad subyacente, está asegurada al existencia de soluciones al problema concreto de ajuste del modelo de mezclas gaussiano al igual que la consistencia de estas.

### 2.2.2. El modelo de mezclas en el Análisis Cluster

En ocasiones, la razón por la que buscamos ajustar un modelo de mezclas es obtener un buen modelo que capture la distribución de nuestro conjunto de datos. Una vez conseguido, podría ser interesante observar si las componentes utilizadas pueden ser identificadas con grupos existentes en el conjunto o si los clusters que surgen como consecuencia del modelado pueden revelar la existencia

de subpoblaciones que no habían sido definidas o reconocidas hasta ahora.

No obstante, en muchos otros casos en los que modelamos un conjunto de datos por medio del modelo de mezclas, la búsqueda de agrupamientos o clusters en un conjunto de datos suele ser el principal objetivo. En estas condiciones, se supone que cada observación es una realización de una de las $G$ componentes de la mixtura, donde cada componente corresponde con uno de los grupos desconocidos. Una vez se han encontrado los parámetros para ajustar el modelo, se puede realizar una partición del conjunto de datos en términos de las probabilidades a posteriori: Asignamos cada individuo a la componente a la que tiene la probabilidad posterior estimada más alta de pertenecer.

Consideramos de nuevo $\{x_1, .., x_n\}$ una realización de $X_1, .., X_n$ y $Z_i$ el vector $G-$dimensional de etiquetas donde sus componentes $Z_{ij}$ son cero o uno dependiendo de si el origen de la observación $x_i$ en la mixtura proviene de la $j$-ésima componente. Una vez ajustado un modelo de mezclas, estamos interesados en inferir el valor de los $z_{ij}$ basándonos en la realización $x_i$, $1 \leq i \leq n$. Una vez obtenida la estimación del vector de parámetros desconocidos, $\hat{\Psi}$, podemos definir un agrupamiento para las observaciones en términos las probabilidades a posteriori de pertenecer a cada componente que hemos ajustado.

Para cada $x_i$, denotamos $\tau_j(x_i; \hat{\Psi})$ la probabilidad a posteriori estimada de que $x_i$ provenga de la $j$-ésima componente, $1 \leq j \leq G$. Es decir

$$\tau_j(x_i; \hat{\Psi}) = \frac{\hat{\pi}_j f(x_i; \hat{\theta}_j)}{\sum_{j=1}^{G} \hat{\pi}_j f(x_i; \hat{\theta}_j)}$$

De esta forma, estimamos cada etiqueta como

$$\hat{z}_{ij} = 1 \text{ si } j = \arg\max_h \tau_h(x_i; \hat{\Psi})$$

$$\hat{z}_{ij} = 0 \text{ en otro caso}$$

para $1 \leq j \leq G$, $1 \leq i \leq n$.

Esta regla no genera una única partición del espacio ya que pueden existir varias componentes que maximicen $\tau_h(x_i; \hat{\Psi})$. En este caso, se puede asignar aleatoriamente ese individuo a cualquiera de las componentes que maximizan la probabilidad a posteriori o definir una partición del espacio en grupos $C_1, .., C_G$ que será única:

$$C_1 = \{x \in \mathbb{R}^d / \arg\max_h \tau_h(x_i; \hat{\Psi}) = 1\}$$

$$C_j = \{x \in \mathbb{R}^d / \arg\max_h \tau_h(x_i; \hat{\Psi}) = j\} - C_{j-1}, \quad j = 1, .., G$$

Se define la incertidumbre en la clasificación de $x_i$ como

$$u(x_i) = 1 - \max_{j \in 1,..,G} \tau_j(x_i; \hat{\Psi})$$

Dado que $\tau_j(x_i; \hat{\Psi})$ es una probabilidad, la incertidumbre para un punto $x_i$ será mayor cuanto más parecidos sean entre sí los valores $\tau_1(x_i; \hat{\Psi}), .., \tau_G(x_i; \hat{\Psi})$, alcanzando su máximo cuando sean

iguales, en cuyo caso $u(x_i) = \frac{1}{G}$. En cambio, será más baja cuando uno de los valores $\tau_j(x_i; \hat{\Psi})$ sea próximo a uno, es decir, cuando la probabilidad de que $x_i$ provenga de la componente $j$ considerando el vector de parámetros $\hat{\Psi}$ sea alta.

**Observación 2.3.** *En el contexto de la clasificación, la partición planteada tiene sentido si asumimos que el coste de clasificar erróneamente una observación es el mismo para todos los grupos. Si en algún caso concreto esto no fuese así, cambiaríamos la regla para minimizar los costes.*

## 2.3. Aplicaciones del modelo finito de mixturas

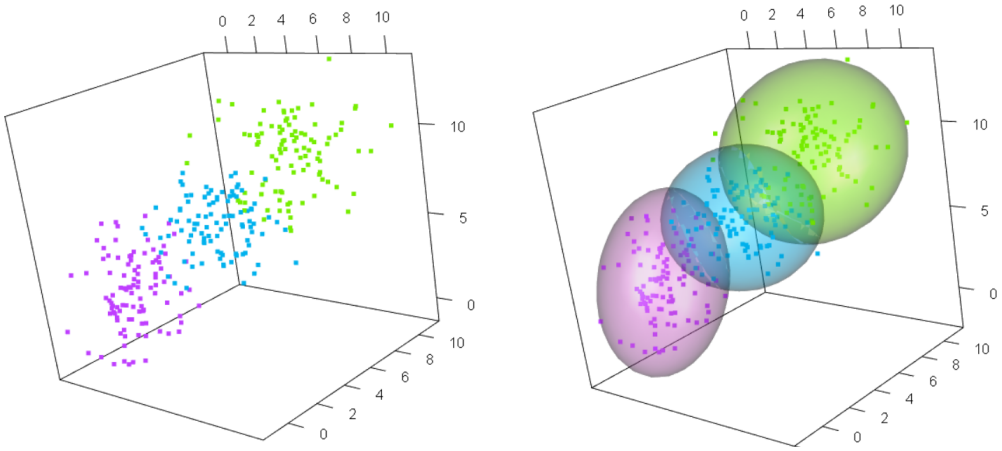
Recogemos en este apartado algunas situaciones en las que resulta verdaderamente interesante la utilización del modelo de mezclas a la hora de ajustar o describir un conjunto de observaciones, particularizando en el modelo de mixturas gaussianas. Por otro lado, detallamos algunos de los escenarios en los que la actuación del modelo de mixturas es insuficiente o inadecuada por la naturaleza intrínseca de los datos.

### 2.3.1. Conjuntos de datos apropiados para el modelo de mezclas

Nos limitamos en este apartado al caso en el que las componentes de mixtura son funciones de densidad normales, una situación muy frecuente en la práctica.

- **Conjuntos elipsoidales**

  Una variable aleatoria con distribución normal multivariante toma valores cercanos a la media con mayor probabilidad y tiene una función de densidad elipsoidal. El modelo de mezclas gaussiano tiende por lo tanto a favorecer disposiciones elipsoidales de los datos, con una mayor concentración de puntos entorno al centro de la distribución. Es por ello que configuraciones de puntos donde encontramos agrupaciones con forma de elipsoide, bien alargados o bien más esféricos, pueden ser modelados fácilmente por un modelo de mixturas gaussianas.




(a) (b)

Figura 8: Conjuntos elipsoidales en el espacio

- **Distribuciones multimodales**

  Las densidades multimodales son aquellas que presentan múltiples modas o picos distintos en su distribución. Estas distribuciones representan situaciones en las que los datos pueden provenir de diferentes grupos o subpoblaciones dentro de la muestra, ya que cada moda puede estar asociada a una agrupación. El modelo de mezclas es especialmente útil para abordar este tipo de situaciones ya que las componentes del modelo se encargarán de ajustar cada una de las modas.

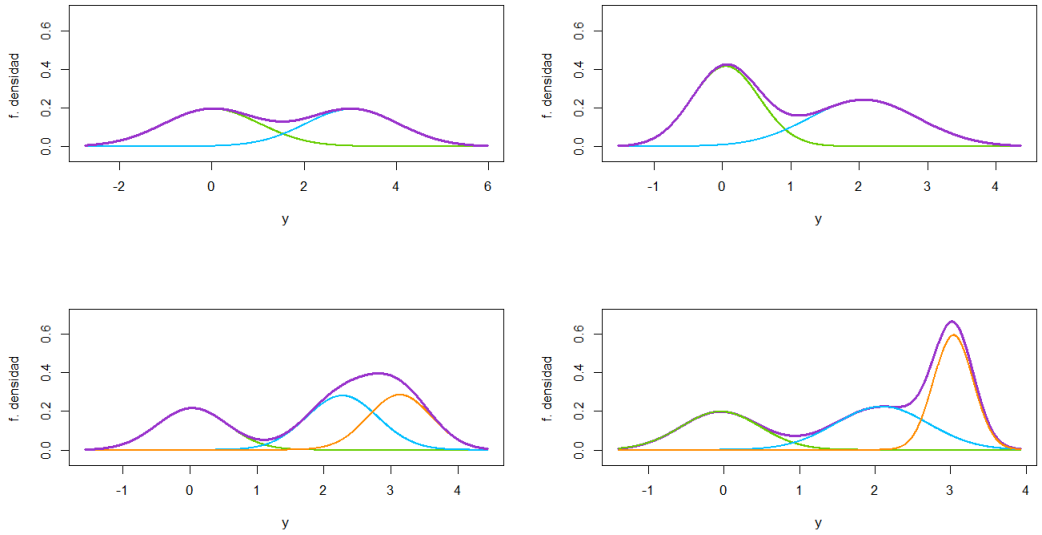



Figura 9: Diferentes ajustes de densidades multimodales

En la figura, observamos varios ejemplos de conjuntos de datos unidimensionales con distribuciones multimodales. En el primer caso, se ven claramente dos protuberancias en la función de densidad, con un valle entre medias. Además, parece que el tamaño de las mismas es idéntico, por lo que si utilizásemos un modelo de mezclas gaussiano, sería adecuado ajustar un modelo de dos mezclas con la misma varianza. En el segundo caso, uno de los picos es más alto que el otro, debido a varianzas diferentes en las subpoblaciones. A continuación, en el tercer caso, tenemos un conjunto de puntos que ha sido generado por tres distribuciones normales, a pesar de que en la función de densidad resultante al combinarse pudiera parecer que solo existen dos modos. Para finalizar, observamos una función de densidad con tres modos de diferentes tamaños, por lo que parecería razonable utilizar una mixtura con tres componentes de distinta varianza.

### 2.3.2. Tratamiento de datos desequilibrados

Dado que para ajustar el modelo de mezclas tenemos que estimar las proporciones de las componentes, este procedimiento puede resultar interesante en conjuntos de datos con clases desequilibradas. Puede ser útil a la hora de compensar la falta de observaciones de un grupo en el contexto de la clasificación o bien para capturar la asimetría de una función de densidad, modelándola a través de varias componentes con diferentes pesos.

Consideremos el siguiente conjunto de datos desequilibrados cuya separación es relativamente buena en el que contamos con 8 grupos de observaciones bidimensionales, tres de los cuales tienen 2000 individuos mientras que los cinco restantes tienen 100. La flexibilidad a la hora de ajustar los

pesos $\pi_1, .., \pi_8$ consigue evitar la situación desfavorable en la que el método de mezclas gaussiano trate de fragmentar uno de los grupos más densos. Observamos como la partición que se genera es bastante buena.

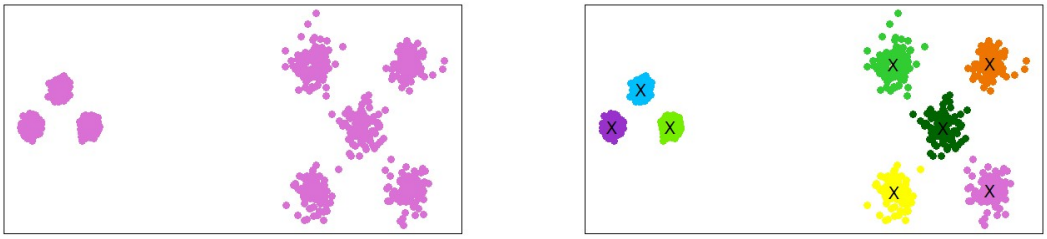


Figura 10: Conjunto de datos desequilibrado

Por otro lado, si tenemos interés en modelar un conjunto de datos cuya distribución tiene asimetría, es posible capturar esta concentración de la masa de probabilidad hacia alguna dirección por medio del modelo de mixturas.

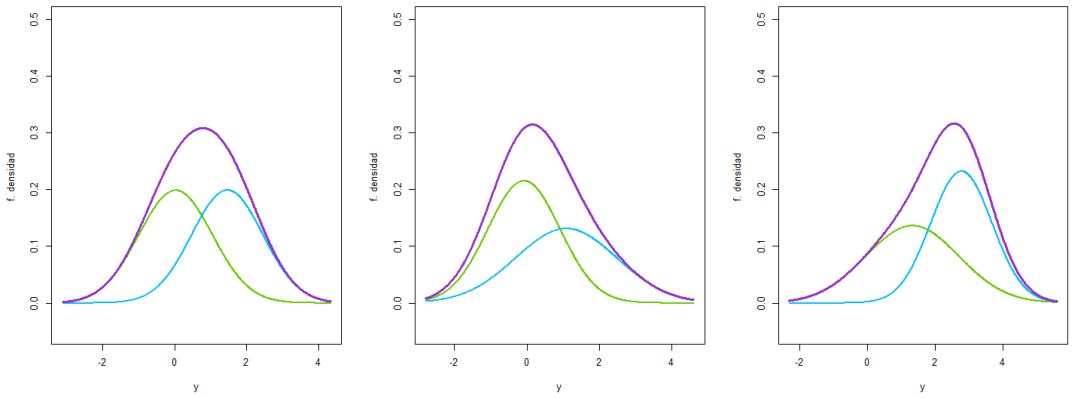



Figura 11: Modelado de funciones de densidad variando su sesgo

En la imagen podemos observar cómo utilizando dos componentes en la mezcla y otorgando mayor peso a una de ellas conseguimos capturar la asimetría de las funciones de densidad consideradas, algo que no hubiera sido posible utilizando una sola densidad gaussiana debido a su carácter simétrico.

### 2.3.3. Detección de valores atípicos

De manera general en el Análisis Estadístico, un outlier se define como un elemento cuyas características no siguen el mismo patrón que la mayoría del resto de observaciones. Estos puntos, en el

contexto del modelo de mixturas, pueden ser observaciones atípicas que están alejadas de los grupos definidos por las componentes o que están dispuestas entremedias de ellos, en particular cuando existe una buena separación entre clusters. Suponen un problema ya que al distorsionar la distribución real de los datos, influyen en la estimación de parámetros. Además, los modelos estadísticos pueden ser sensibles a los valores extremos, por lo que incluso unos pocos outliers pueden tener un impacto desproporcionadamente grande en los resultados del modelo de mezclas. Es por lo tanto importante identificar y tratar adecuadamente las observaciones atípicas para obtener resultados más precisos, robustos y confiables.

El modelo de mezclas, debido a la flexibilidad que tiene a la hora de elegir las proporciones de las componentes, es una herramienta útil en algunas ocasiones para detectar estas observaciones atípicas. Consideramos el siguiente conjunto de datos donde se tienen dos grupos, de 60 y 120 observaciones cada uno, elipsoidales y fácilmente distinguibles. Por otro lado, generamos 20 observaciones con distribución normal pero con una matriz de covarianza de determinante mayor. Algunos de estos outliers son fácilmente identificables, mientras que aquellos que se encuentran en las vecindades de los grupos resultan más difíciles de detectar. Si ajustamos tres componentes gaussianas, el modelo de mixturas prácticamente es capaz de recuperar las dos agrupaciones iniciales con bastante éxito, asignando a la tercera componente aquellos puntos que son considerados outlier.

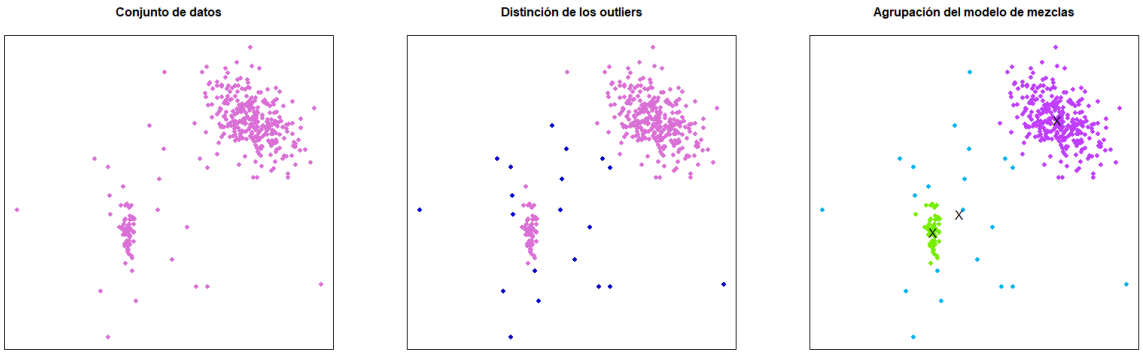



Figura 12: Detección de observaciones atípicas utilizando el modelo de mezclas.

En ocasiones, el método no tiene capacidad suficiente para tratar con éxito este tipo de observaciones. Existen multitud de enfoques para abordar esta situación, pero nosotros presentaremos dos de ellos en este apartado.

- **Componente uniforme para los valores atípicos**

  Una forma de enfrentarnos a los outliers o al ruido en el modelo de mixturas consiste en añadir una componente adicional al modelo para representar la aportación de estas observaciones atípicas. Por ejemplo, se puede elegir una componente uniforme para tratar de reflejar esa dispersión equitativa de los valores atípicos en la región del espacio donde se sitúan los grupos,

de tal manera que se localicen tanto en las afueras de los clusters como entremedias de ellos. De este modo, el modelo que habíamos definido antes tendría ahora la siguiente forma:

$$f(x_i) = \frac{\pi_0}{V} + \sum_{j=1}^{G} \pi_j f(x_i; \theta_j),\ 1 \leq i \leq n$$

donde $V$ es el volumen aproximado de la región que ocupan los datos y $\pi_0$ es la proporción esperada de valores atípicos. Podemos encontrar más información sobre este planteamiento en [2] y sobre la sensibilidad del algoritmo EM dependiendo de la inicialización en el caso de existir outliers.

- **Procedimientos de recorte**

Podemos encontrar estas ideas en artículos como [4], donde se presenta una técnica que consiste en eliminar parte de los datos de manera no arbitraria con el fin de hacer más robusto el procedimiento de mixturas, en concreto el caso en el que las funciones de densidad son gaussianas. Para ello, se consideran ciertas restricciones para asegurar la robustez de los parámetros estimados y se construye la función de verosimilitud recortada

$$\sum_{i=1}^{n} z(x_i) \log\Big(\sum_{j=1}^{G} \pi_j \phi(x_i; \mu_j, \Sigma_j)\Big)$$

Donde $z(\cdot)$ es una función de recorte que vale 0 si la observación ha sido eliminada y 1 en el caso contrario. Previamente, se elige un número $\alpha \in (0, 1)$ denominado nivel de recorte que será la proporción de datos que se desea recortar de la muestra $\{x_1, .., x_n\}$. Esa proporción de datos será la que asumimos como observaciones espurias que no queremos incluir cuando estimemos los parámetros del modelo de mixturas. Por lo tanto, se tiene que $\sum_{i=1}^{n} z(x_i) = [n(1-\alpha)]$, donde $[\cdot]$ denota el redondeo hacia arriba.

Este artículo proporciona además una versión robusta del algoritmo EM que incluye un paso de recorte adicional. El criterio que se utiliza para decidir qué observaciones parecen ser espurias es calcular $d_i = \text{máx}\{\pi_1\phi_1(y_i), .., \pi_G\phi_G(y_i)\}$ y y seleccionar la proporción $\alpha$ de puntos que toman los valores más pequeños de $d_i$.

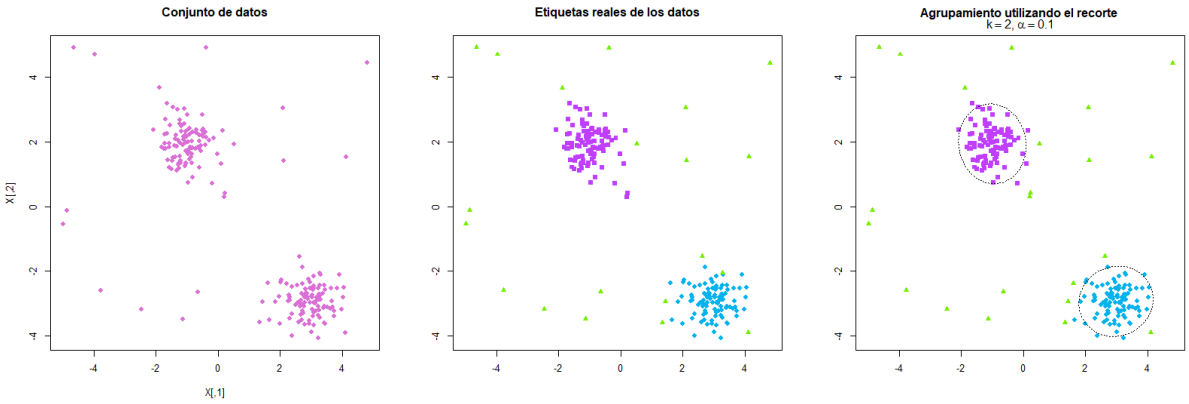



Figura 13: Recorte de nivel $\alpha = 0{,}2$

### 2.3.4. Superposición entre los grupos

La existencia de superposición entre los grupos en un conjunto de datos supone un desafío para el modelo de mezclas ya que puede tener dificultades para distinguir las subpoblaciones. En la siguiente figura observamos la actuación del modelo de mezclas gaussiano al tratar de distinguir los 15 grupos de la figura, cada uno con 300 observaciones. Dado que los clusters son elipsoidales y no existe una superposición muy notoria, con pocos puntos entre los clusters, la partición resulta bastante satisfactoria.

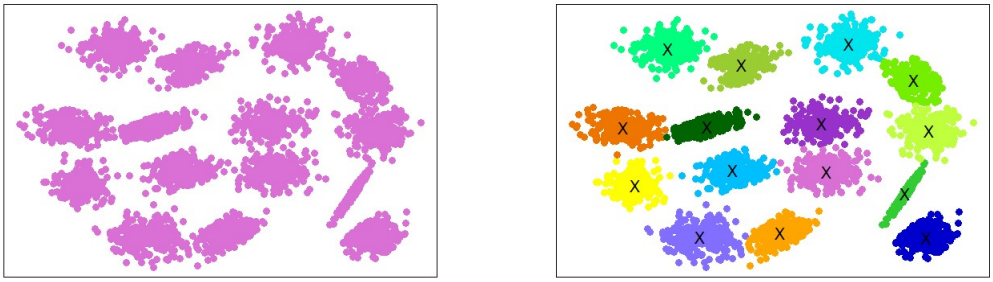


Figura 14: Grupos bien separados

Sin embargo, al considerar un conjunto de datos con mayor grado de superposición, la estimación de parámetros se vuelve más sensible a pequeñas variaciones y además el método fracasa a la hora de detectar los grupos. Observamos el siguiente ejemplo donde se ha realizado un ajuste por medio de un modelo de mezclas gaussiano a un conjunto de datos con tres subpoblaciones. En la figura de la derecha vemos una comparación gráfica entre los parámetros reales utilizados para generar los datos y los parámetros estimados por el modelo de mezclas gaussiano.

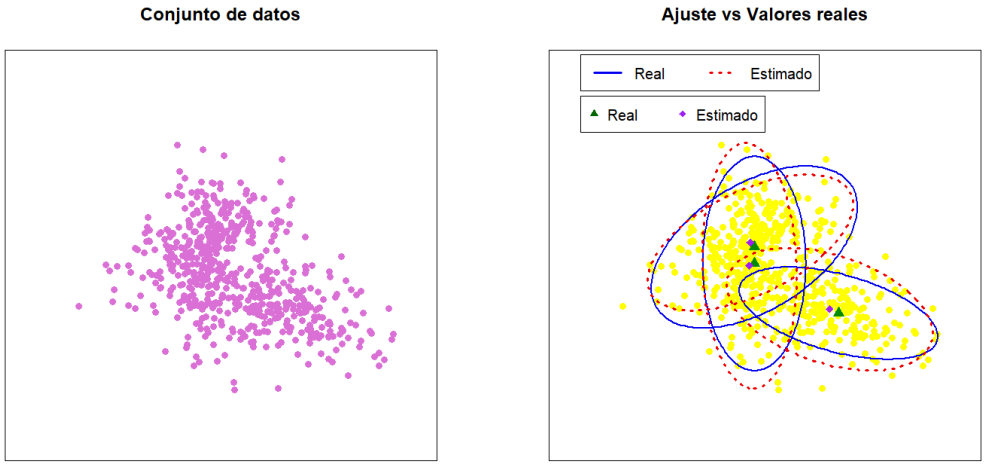



Figura 15: Elección de parámetros con superposición entre tres grupos

A la hora de elegir los vectores de medias, la actuación del modelo es bastante buena, mientras que el ajuste de las matrices de covarianza no resulta del todo exacto. Aun así, el método consigue capturar relativamente bien la disposición y orientación de las componentes.

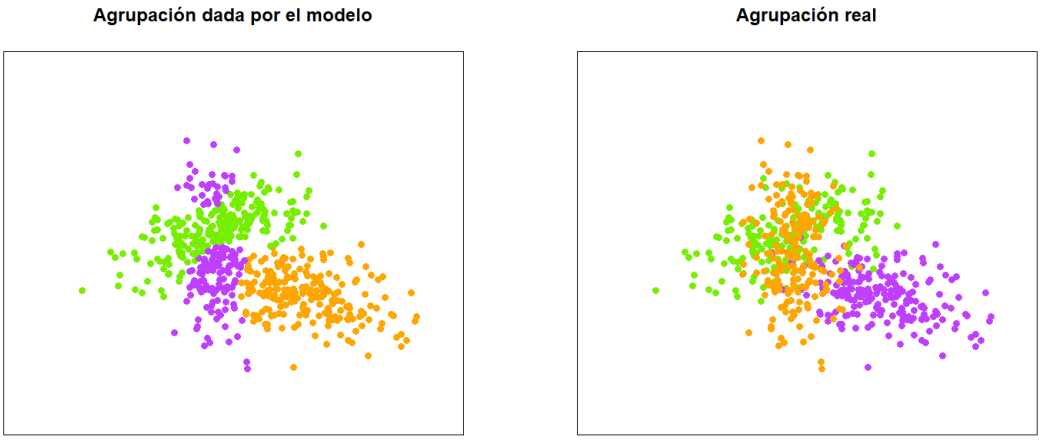



Figura 16: Diferencia entre la partición generada por el modelo y la agrupación real

Sin embargo, la presencia de puntos con probabilidades a posteriori parecidas de pertenecer a varios grupos produce una gran incertidumbre a la hora de asignar las observaciones a un cluster o a otro. Esto genera una mayor variabilidad en los resultados al variar la muestra y dificulta la interpretación de los grupos obtenidos, ya que en la agrupación real los individuos están entremezclados en ciertas zonas de su representación en el plano y el modelo en cambio realiza una separción total de los grupos de manera que no existe solapamiento.

Lo mismo ocurre al considerar conjuntos incluso con mayor grado de superposición. En el gráfico siguiente vemos señalados con una equis los vectores de medias estimados por el modelo de mezclas gaussiano y con círculos rojos los verdaderos centros. De nuevo, la elección de los centros parece razonable, mientras que la actuación a la hora de crear los clusters de nuevo es bastante deficiente, ya que las fronteras son variables y poligonales.

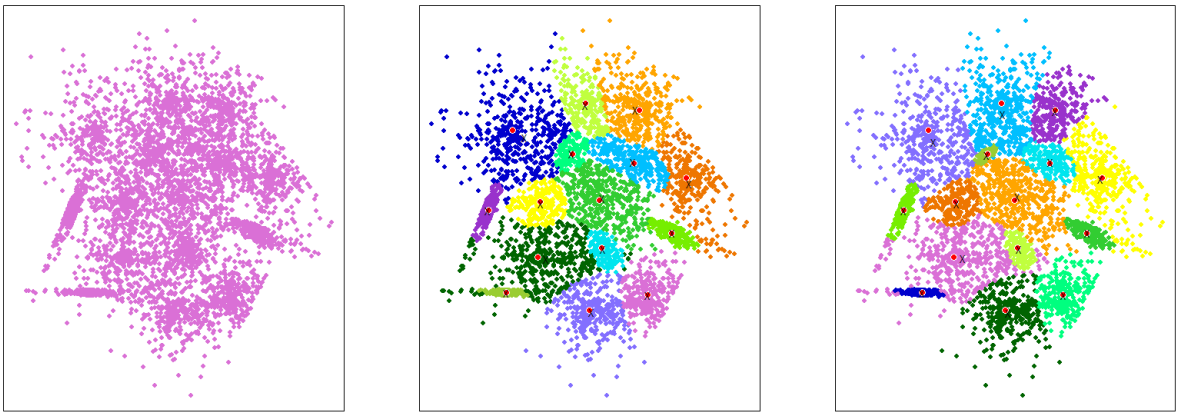


Figura 17: Variabilidad y rigidez en las fronteras en el caso de superposición entre grupos

Es importante tener en cuenta que la creación de estas particiones tan fluctuantes en presencia de superposición entre grupos no siempre implica un fallo del modelo de mezclas. En algunos casos, puede ser un reflejo de la naturaleza de las observaciones: si existe un solapamiento real entre los grupos es posible que simplemente sus individuos sean difíciles de distinguir.

# 3. Estudio teórico del modelo de mixturas

Hemos planteado la idea de configurar una descripción de un conjunto de $n$ puntos de la mejor manera posible utilizando $G$ funciones de densidad pertenecientes a una familia de distribuciones $\{f(\cdot;\theta), \theta \in \Theta\}$. En esta sección, buscamos enfocar este problema desde una perspectiva más general, a través de un estudio teórico que incluya tanto conjuntos de datos como distribuciones de probabilidad. Para ello, generalizaremos de manera natural la expresión de la función de verosimilitud utilizando la esperanza matemática.

A continuación, particularizando en la familia de distribuciones Gaussianas, hablaremos de algunas propiedades matemáticas de este problema como la existencia, la identificabilidad o la consitencia. Esta familia de distribuciones es especialmente interesante ya que encontramos de manera frecuente el problema en el que los individuos provienen de diferentes poblaciones normales. Tanto para hablar de la existencia de un vector de parámetros que maximice la expresión de la verosimilitud como para hablar de la consistencia del vector de estimadores muestrales es necesario imponer ciertas restricciones en el problema, algunas de ellas especialmente confeccionadas para las funciones de densidad normales, siendo esta otra razón por la que nos enfocamos en el caso de la familia de distribuciones Gaussianas.

A lo largo de esta sección trataremos de ilustrar gráficamente estos resultados teóricos a fin de facilitar la visualización y la comprensión de los resultados planteados.

## 3.1. Planteamiento del modelo general de mixturas

Inicialmente habíamos introducido el ajuste un modelo de $G$ funciones de densidad a un conjunto de $n$ observaciones $\{x_1, .., x_n\}$. Supongamos que en lugar de tener un conjunto con $n$ datos contamos con una distribución de probabilidad: $X$ será una variable aleatoria definida en un espacio probabilístico $(\Omega, \sigma, \mathbb{P})$ y $P$ la ley de probabilidad que induce en $\mathbb{R}$. Podemos reescribir la función de verosimilitud para esta probabilidad P en lugar de para $n$ observaciones de $X$ por medio de la esperanza matemática, que no es sino la generalización de la media a través del concepto de integral respecto de una probabilidad:

$$\mathbb{E}(X) = \int_{\mathbb{R}} x dP(x) \tag{3}$$

Para variables aleatorias multidimensionales, su esperanza o valor esperado se define componente a componente, esto es, dado un vector aleatorio $X = (X_1, .., X_p) : \Omega \longrightarrow \mathbb{R}^p$, definimos su esperanza como

$$\mathbb{E}[(X_1, .., X_p)] = [\mathbb{E}(X_1), .., \mathbb{E}(X_p)]$$

De este modo, para dicho vector aleatorio $X$, la función de log-verosimilitud que tratábamos de maximizar en la sección anterior se reescribiría como

$$L(\Psi, P) = \mathbb{E}_P \left( log \left( \sum_{j=1}^{G} \pi_j f(x; \theta_j) \right) \right), \ f \in \mathscr{F} = \{ f(\cdot; \theta) : \theta \in \Theta \}$$

Este planeamiento del problema mediante distribuciones de probabilidad generaliza el caso de tener un conjunto con $n$ puntos. En el caso de tener $x_1, .., x_n$ elementos de $\mathbb{R}^d$, podemos considerar $P_n$, la probabilidad muestral que asigna probabilidad $\frac{1}{n}$ a cada $x_i$ y utilizar que la esperanza no es sino el promedio de los $n$ puntos. Estamos en disposición de plantear el problema de ajuste de un modelo de mezclas a la distribución P:

**Definición 3.1.** *Sea $X : (\Omega, \sigma, \mathbb{P}) \longrightarrow \mathbb{R}^d$ un vector aleatorio y sea $P$ la probabilidad inducida por $X$ en $\mathbb{R}^d$. Sea $G$ un entero positivo y $\pi_1, .., \pi_{G-1} \geq 0$ pesos tales que $\sum_{j=1}^{G} \pi_j = 1$. Sea $\mathscr{F} = \{ f(\cdot; \theta) / \theta \in \Theta \}$ una familia de funciones de densidad donde $\Theta$ denota el espacio paramétrico y $\Psi = (\pi_1, .., \pi_G, \theta_1, .., \theta_G)$ el vector que recoge los parámetros a estimar. El problema de ajustar a la variable $X$ un modelo de $G$ distribuciones respecto a la familia $\mathscr{F}$ consiste en encontrar $\Psi^*$ que maximice la expresión*

$$L(\Psi, P) = \mathbb{E}_P \left( log \left( \sum_{j=1}^{G} \pi_j f(x; \theta_j) \right) \right) \tag{4}$$

## 3.2. Propiedades matemáticas

Abordaremos en esta sección el estudio teórico del modelo de mixturas para el caso particular en el que la familia de distribuciones sea la gaussiana: $f_j$ será la función de densidad normal multivariante $\phi_j$, parametrizada por su vector de medias $\mu_j$ y su matriz de covarianzas $\Sigma_j$, con la forma

$$\phi(x_i; \mu_j, \Sigma_j) = \frac{1}{(2\pi)^{\frac{n}{2}} |\Sigma_j|^{\frac{1}{2}}} \exp \{ -\frac{1}{2} (x_i - \mu_j)^T \Sigma_j^{-1} (x_i - \mu_j) \}$$

Es decir, consideraremos el problema en el que $\mathscr{F} = \{ \phi(\cdot; \mu, \sigma) : (\mu, \sigma) \in \Theta \}$, donde $\Theta = \mathbb{R}^p \times M_{p \times p}$. En primer lugar, plantearemos las hipótesis necesarias para poder garantizar la existencia de una solución al problema de maximización de la log-verosimilitud. Después, definiremos el concepto de identificabilidad de un modelo y demostraremos que los modelos de mezclas gaussianos poseen esta propiedad. Por último, estudiaremos la consistencia del método, que resulta interesante ya que a menudo el problema aparece en un ámbito estadístico en el que contamos con conjuntos de datos que podemos interpretar como una muestra.

### 3.2.1. Existencia de solución

A continuación, abordaremos el problema de existencia de un vector de parámetros que minimicen la expresión (4). La existencia de una solución no está garantizada sin imponer ciertas restricciones sobre la medida de probabilidad $P$ y la relación entre los autovalores de las matrices de covarianzas. La función de verosimilitud no está acotada en el caso en el que nos propusiesemos ajustar un modelo de $G$ mixturas a un conjunto de $G$ puntos o a un conjunto de datos en el que su disposición se aproxime demasiado a un subespacio de dimensión menor, ya que en estos casos el determinante de la

matriz de covarianzas tendería a cero y, en consecuencia, $\phi_j$ tendería a infinito para esa componente en concreto. Si excluimos el caso en el que manejemos una probabilidad concentrada en $G$ puntos e impedimos que las proporciones entre los tamaños de los autovalores de las matrices de convarianza sean arbitrarias, seremos capaces de asegurar la existencia y consistencia de soluciones.

Los resultados que encontramos en esta sección son una adaptación de los que aparecen en los artículos de García Escudero et al. (2008, 2015) [6], [5], donde se trabaja con un problema más general denominado ajuste de mixturas Gaussianas recortado. En él, se presenta una técnica que consiste en eliminar una proporción $\alpha$ de los datos de manera no arbitraria con el fin de hacer más robusto el procedimiento. En esta versión del problema, una de las restricciones que expondremos a continuación para garantizar la existencia y consistencia de un vector de parámetros que maximice la verosimilitud dejaría de ser necesaria, por lo que cuando lleguemos a ese punto resaltaremos este hecho.

Dada la matriz de covarianzas $\Sigma_j$, correspondiente a la $j-$ésima componente de la mixtura, dado que es simétrica y definida positiva, podemos encontrar matrices $E_j$ y $D_j$ de tal manera que

$$\Sigma_j = E_j D_j E_j^T$$

Donde $D_j$ es la matriz diagonal que recoge los autovalores $\lambda_{j,1}, .., \lambda_{j,G}$ de $\Sigma_j$ y $E_j$ es la matriz de cambio de base cuyas columnas son los autovectores unitarios $v_{j,1}, .., v_{j_G}$ asociados a los autovalores anteriores (al ser los autovectores ortonormales, $E_j^T$ coincide con $E_j^{-1}$). Esta descomposición nos resultará útil a la hora de probar ciertos resultados posteriores y nos ayudará a entender una de las condiciones que imponemos sobre los autovalores de las matrices de covarianzas.
Dicho esto, presentamos a continuación las restricciones necesarias para poder garantizar la existencia y la consistencia:

1. **(PR)** Restricciones sobre la distribución P:
   - La distribución P no debe estár concentrada en $G$ puntos.
   - P debe de tener momento de segundo orden finito, esto es: $\mathbb{E}_P(\|\cdot\|^2) < \infty$. En el caso de estar trabajando con recortes, esta hipótesis no sería necesaria.
2. **(ER)** Restricciones sobre el ratio entre autovalores: Debe cumplirse, para una constante $c \geq 1$, que
   $$M_n/m_n \leq c$$
   Donde
   $$M_n = \max_{1\leq j\leq G} \max_{1\leq l\leq p} \lambda_{jl}$$
   $$m_n = \min_{1\leq j\leq G} \min_{1\leq l\leq p} \lambda_{jl}$$
   Es decir, queremos tener controlado el ratio entre el autovalor más grande y el más pequeño entre todas las matrices de covarianza. En el ámbito de la búsqueda de agrupaciones o de la clasificación de individuos, esta restricción controla simultaneamente las diferencias entre grupos y cuánto distan estos de ser esféricos.

Existen otras maneras de poner un control para evitar soluciones espurias. De hecho, las restricciones consideradas son una adaptación de las que se contemplan en Hathaway (1983,1986) [7], que, en el caso de $G$ componentes univariantes propone controlar el cociente entre las varianzas de las diferentes componentes, probando que el estimador máximo verosímil bajo estas restricciones es consistente. Otra posible perspectiva es la de utilizar una función que penalice valores pequeños de las componentes de las matrices de covarianza. Nosotros consideraremos en este documento la restricción **(ER)** descrita anteriormente, presente en [6].

Nuestro principal objetivo será demostrar el siguiente resultado de existencia bajo las restricciones descritas.

**Teorema 3.2.** *Si se verifican las restricciones* **(PR)** *y* **(ER)**, *existe* $\Psi \in \Theta_c$ *en el cual se alcanza el máximo de*

$$L(\Psi, P) = \mathbb{E}_P \left( log \left( \sum_{j=1}^{G} \pi_j \phi(x; \mu_j, \Sigma_j) \right) \right)$$

Para probar este resultado, consideramos $\{\Psi_n\}_{n=1}^{\infty} = \{(\pi_1^n, .., \pi_G^n, \mu_1^n, .., \mu_G^n, \Sigma_1^n, ..\Sigma_G^n)\}_{n=1}^{\infty}$ una sucesión del espacio paramétrico $\Theta_c$ de tal manera que

$$\lim_{n\to\infty} L(\Psi_n, P) = \sup(L(\Psi, P)) = K > -\infty$$

Esto siempre es posible porque, dado que P tiene momento de orden dos finito, tomando el vector de parámetros $\tilde{\Psi}$ donde elegimos $\pi_1 = 1, \pi_j = 0$ si $j \neq 1$, $\mu_1 = 0$, $\Sigma = I$ y el resto de valores de forma arbitraria, se tiene que

$$\sup(L(\Psi, P)) \geq L(\tilde{\Psi}, P) > -\infty$$

Nuestro objetivo es extraer una subsucesión convergente de $\{\Psi_n\}_{n=1}^{\infty}$. Para ello, pensaremos cómo podemos hacerlo para los pesos $\pi_j$, los vectores de medias $\mu_j$ y para las matrices de covarianza $\Sigma_j$ respectivamente.

- Para los pesos $\pi_j$:

  En este caso resulta sencillo extraer una subsucesión convergente ya que $[0,1]^G$ es compacto: Podemos encontrar una subsucesión de $\{\Psi_n\}_{n=1}^{\infty}$ que denotaremos como la original de tal manera que $\pi_j^n \to \pi_j \in [0,1]$

- Para las matrices de covarianzas:

  Dada la restricción que hemos impuesto sobre los autovalores, tenemos tres posibilidades:

  1. $\Sigma_j^n \to \Sigma_j$ para $1 \leq j \leq G$
  2. $M_n \to \infty$, por lo que consecuentemente $m_n \to \infty$ para que se verifique **(ER)**.
  3. $m_n \to 0$, con lo que $M_n \to 0$ para que se cumpla **(ER)**.

  Vamos a ver en primer lugar que, con las condiciones que hemos impuesto, solo puede darse una de las convergencias.

**Lema 3.3.** *Dada la subsucesión* $\{\Psi_n\}_{n=1}^{\infty}$ *con* $\lim_{n\to\infty} L(\Psi_n, P) > -\infty$ *y dadas las restricciones* **(ER)** *y* **(PR)**, *solo es posible que se de la primera de las convergencias que hemos enunciado.*

**Demostración**

Vamos a suponer que se cumple alguna de las otras dos convergencias que hemos contemplado y, en ambos casos, llegaremos a un absurdo.

Consideremos en primer lugar que $M_n \to \infty$ y $m_n \to \infty$. Denotaremos por $\phi_j^n = \phi(\cdot; \mu_j^n, \Sigma_j^n)$. Se tiene que

$$L(\Psi_n, P) = \mathbb{E}_P\left(\log\left(\sum_{j=1}^{G} \pi_j^n \phi_j^n\right)\right) \leq \mathbb{E}_P\left(\log\left(G \cdot \max_{1\leq j\leq G} \pi_j^n \phi_j^n\right)\right) =$$

$$= \mathbb{E}_P\left(\log(G) + \max_{1\leq j\leq G} \log(\pi_j^n \phi_j^n)\right)$$

Siendo esta última igualdad consecuencia del carácter creciente de la función logaritmo. Si sustituímos $\phi_j^n$ en la expresión tendríamos

$$\mathbb{E}_P\left(\log(G) + \max_{1\leq j\leq G}\left(\log(\pi_j^n) - \frac{p}{2}\log(2\pi) - \frac{1}{2}\log(|\Sigma_j^n|) - \frac{1}{2}(x_i - \mu_j^n)^T(\Sigma_j^n)^{-1}(x_i - \mu_j^n)\right)\right) \leq$$

$$\leq \mathbb{E}_P\left(\log(G) + \max_{1\leq j\leq G}\left(-\frac{1}{2}\log(|\Sigma_j^n|) - \frac{1}{2}(x_i - \mu_j^n)^T(\Sigma_j^n)^{-1}(x_i - \mu_j^n)\right)\right)$$

Ya que $\log(2\pi) > 0$ y $\log(\pi_j^n) < 0$ por ser los $\pi_j^n$ pesos. Denotamos por $\lambda_{jl}^n$ el $l-$ésimo autovalor de $\Sigma_j^n$ y por $v_{jl}^n$ el autovector unitario asociado. Sabemos que el determinante de una matriz es el producto de sus autovalores, por lo que $|\Sigma_j^n| = \prod_{l=1}^{p} \lambda_{jl}^n$. Si sustituimos esto en la expresión anterior junto con la descomposición anteriormente mencionada de $\Sigma_j^n$ en una matriz diagonal, tendríamos

$$\mathbb{E}_P\left(\log(G) + \max_{1\leq j\leq G}\left(-\frac{1}{2}\sum_{l=1}^{p}\log(\lambda_{jl}^n) - \frac{1}{2}(x_i - \mu_j^n)^T((E_j^n)^{-1}D_j^n E_j^n)^{-1}(x_i - \mu_j^n)\right)\right) =$$

$$= \mathbb{E}_P\left(\log(G) + \max_{1\leq j\leq G}\left(-\frac{1}{2}\sum_{l=1}^{p}\left(\log(\lambda_{jl}^n) - \frac{1}{2\lambda_{jl}^n}(x - \mu_j^n)^T v_{jl}^n (v_{jl}^n)^T (x_i - \mu_j^n)\right)\right)\right) \leq$$

$$\leq \mathbb{E}_P\left(\log(G) + \max_{1\leq j\leq G}\left(-\frac{p}{2}\log(m_n) - \frac{1}{2M_n}\|x - \mu_j^n\|^2\right)\right)$$

Donde la última desigualdad es consecuencia de acotar cada autovalor $\lambda_{jl}^n$ por $M_n$ o $m_n$ respectivamente. Si en esta expresión hacemos tender $M_n$ y $m_n$ a infinito, resulta que $-\frac{p}{2}\log(m_n) \to$

$-\infty$, con lo que, dado que el resto de sumandos están acotados, $\lim_{n\to\infty} L(\Psi_n, P) = -\infty$ y llegaríamos a un absurdo. Descartamos por lo tanto que pueda darse esa convergencia.

Consideramos ahora el caso en el que $m_n \to 0$ y consecuentemente $M_n \to 0$. Para estudiar qué ocurre en este caso, necesitamos probar un resultado previo.

**Lema 3.4.** *En las condiciones del problema anterior y habiendo establecido las restricciones* **(PR)** *y* **(ER)**, *existe* $h > 0$ *tal que* $\mathbb{E}_P\left(\text{máx}_{1\leq j\leq G}\|x-\mu_j^n\|^2\right) \geq h > 0$.

La demostración de este Lema está recogida en el Apéndice. De este modo, se tiene

$$\mathbb{E}_P\left(\log(G) + \max_{1\leq j\leq G}\left(-\frac{p}{2}\log(m_n) - \frac{1}{2M_n}\|x-\mu_j^n\|^2\right)\right) =$$
$$= \log(G) + \max_{1\leq j\leq G}(-\frac{p}{2}\log(m_n)) - \frac{1}{2M_n}\mathbb{E}_P\left(\max_{1\leq j\leq G}\|x-\mu_j^n\|^2\right) \leq$$
$$\leq \log(G) + \max_{1\leq j\leq G}(-\frac{p}{2}\log(m_n)) - \frac{1}{2M_n}h$$

En este caso, cuando $m_n$ y $M_n$ tienden a 0, la expresión anterior tiende a $-\infty$ por órdenes de infinitud, por lo que de nuevo llegaríamos a un absurdo y solo sería posible la primera de las convergencias mencionadas. □

Tenemos por lo tanto que existen $\Sigma_1, .., \Sigma_G \in M_{p\times p}$ tales que $\Sigma_j^n \to \Sigma_j$ para $j = 1, .., G$.

- Para los vectores de medias:

Podemos afirmar que se cumple una de estas dos opciones:

1. Para todo $j = 1, .., G$, resulta que $\|\mu_j^n\| \to \infty$
2. Con un reordenamiento adecuado de los índices, existe $k$ tal que
   - Para $1 \leq j \leq k$, $\mu_j^n \to \mu_j \in \mathbb{R}^p$
   - Para $k < j \leq G$, $\min_{j>k}\|\mu_j^n\| \to \infty$

Veamos que la primera opción no puede darse ya que llegaríamos a un absurdo. Habíamos establecido anteriormente la siguiente desigualdad:

$$L(\Psi_n, P) \leq \mathbb{E}_P\left(\log(G) + \max_{1\leq j\leq G}\left(-\frac{p}{2}\log(m_n) - \frac{1}{2M_n}\|x_i-\mu_j^n\|^2\right)\right)$$

Hemos demostrado que necesariamente las matrices de covarianza convergen, por lo que $\exists M, m \in \mathbb{R}$ tales que $M_n \to M, m_n \to m$. Esto provoca que todos los términos estén acotados salvo $\|x-\mu_j^n\|^2$, que tendería a infinito ya que para todo $j = 1, .., G$ se tiene que $\|\mu_j^n\| \to \infty$. De este modo, estaríamos afirmando que $L(\Psi_n, P) \leq -\infty$, con lo que descartamos esta primera opción.

Nuestro siguiente objetivo es ver que la convergencia se da para todos los vectores de medias:

**Lema 3.5.** *Para* $\{\Psi_n\}_{n=1}^{\infty}$ *con* $\lim_{n\to\infty} L(\Psi_n, P) > -\infty$, *dadas las restricciones* ***(ER)*** *y* ***(PR)***, *si se tiene que cada* $\pi_j$ *es estrictamente positivo, resulta que* $k = G$.

**Demostración**

En primer lugar, vamos a probar que

$$\mathbb{E}_P\left(\log\left(\sum_{j=1}^{G}\pi_j^n\phi_j^n\right)\right)-\mathbb{E}_P\left(\log\left(\sum_{j=1}^{k}\pi_j^n\phi_j^n\right)\right)\to 0$$

Utilizaremos el Teorema de la Convergencia Dominada de Lebesgue para demostrarlo.

Veamos que la sucesión de variables aleatorias $X_n=\log\left(\sum_{j=1}^{G}\pi_j^n\phi_j^n\right)-\log\left(\sum_{j=1}^{k}\pi_j^n\phi_j^n\right)\to 0$ $P-c.s.$ Se tienen las siguientes desigualdades:

$$0\leq X_n=\log\left(1+\frac{\sum_{j>k}^{G}\pi_j^n\phi_j^n}{\sum_{j=1}^{k}\pi_j^n\phi_j^n}\right)\leq\log\left(1+\frac{\sum_{j>k}^{G}\pi_j^n\phi_j^n}{\pi_1^n\phi_1^n}\right)$$

En la expresión anterior podemos dividir cada término del sumatorio entre $\pi_1^n\phi_1^n$. Esos cocientes pueden reescribirse como

$$\frac{\pi_j^n\phi_j^n}{\pi_j^n\phi_1}=\frac{\pi_j^n}{\pi_1^n}\left(\frac{|\Sigma_1^n|}{|\Sigma_j^n|}\right)^{1/2}\exp\{\frac{-1}{2}(x-\mu_j^n)^T(\Sigma_j^n)^{-1}(x-\mu_j^n)+\frac{1}{2}(x-\mu_1^n)^T(\Sigma_1^n)^{-1}(x-\mu_j^n)\}$$

Dado que $\left(\frac{|\Sigma_1^n|}{|\Sigma_j^n|}\right)^{1/2}=\frac{\prod_{l=1}^{G}\lambda_{l1}^n}{\prod_{l=1}^{G}\lambda_{lj}^n}$, podemos acotar cada autovalor del numerador por $M_n$ y cada autovalor del denominador por $m_n$. Si además utilizamos la descomposición en autovectores de las matrices de covarianzas, podemos afirmar que

$$\frac{\pi_j^n\phi_j^n}{\pi_j^n\phi_1}\leq\frac{\pi_j^n}{\pi_1^n}\left(\frac{M_n}{m_n}\right)^{p/2}\exp\{\frac{1}{2M_n}\|x-\mu_1^n\|^2-\frac{1}{2m_n}\|x-\mu_j^n\|^2\}$$

Por lo tanto, volviendo a la expresión con la que trabajábamos anteriormente

$$\log\left(1+\frac{\sum_{j>k}^{G}\pi_j^n\phi_j^n}{\pi_1^n\phi_1^n}\right)\leq\log\left(1+\sum_{j>k}^{G}\frac{\pi_j^n}{\pi_1^n}\left(\frac{M_n}{m_n}\right)^{p/2}\exp\{\frac{1}{2M_n}\|x-\mu_1^n\|^2-\frac{1}{2m_n}\|x-\mu_j^n\|^2\}\right)\leq$$

$$\leq\log\left(1+\sum_{j>k}^{G}\frac{\pi_j^n}{\pi_1^n}\left(\frac{M_n}{m_n}\right)^{p/2}\exp\{\frac{1}{2M_n}\|x-\mu_1^n\|^2+\frac{1}{2m_n}\|x-\mu_1^n\|^2-\frac{1}{2m_n}\|\mu_1^n-\mu_j^n\|^2\}\right)$$

Donde en el último paso hemos aplicado la desigualdad triangular. Dado que $1/M_n\leq 1/m_n$ y la función exponencial es creciente, podemos acotar las líneas anteriores por

$$\log\left(1+\exp\{\frac{-1}{2m_n}\min_{j>k}\|\mu_1^n-\mu_j^n\|^2\}\sum_{j>k}^{G}\frac{\pi_j^n}{\pi_1^n}c^{p/2}\exp\{\frac{1}{m_n}\|\mu_1^n-\mu_j^n\|^2\}\right)$$

Como los pesos $\pi_1^n, .., \pi_G^n$, los autovalores $M_n, m_n$ y el vector $\mu_1^n$ convergen, los términos están acotados y para un $x$ fijo la expresión anterior tiende a cero, ya que $\min_{j>k}\|x-\mu_j^n\|^2 \to \infty$.

A continuación, veamos que la sucesión de variables aleatorias $X_n$ está acotada por una variable $Y$ con $\mathbb{E}(Y) < \infty$. Para ello, utilizaremos una desigualdad elemental:

$$Si\ x \geq 0, resulta\ que \log(1+ae^x) \leq x + \log(1+a)$$

Esto se da trivialmente, ya que cuando $x \geq 0$, $e^x \geq 1$ y por lo tanto

$$\log(1+ae^x) = log\left((1/e^x + a)\, e^x\right) \leq log((1+a)e^x) = x + \log(1+a)$$

Como $X_n \geq 0$, podemos aplicárselo a la expresión anterior tomando

$$a = \exp\{\frac{-1}{2m_n} \min_{j>k}\|\mu_1^n - \mu_j^n\|^2\} \sum_{j>k}^{G} \frac{\pi_j^n}{\pi_1^n} c^{p/2}$$

De esta manera, se tendría

$$0 \leq X_n \leq \frac{\|x-\mu_1^n\|^2}{m_n} + \log\left(1 + \exp\{\frac{-1}{2m_n} \min_{j>k}\|\mu_1^n - \mu_j^n\|^2\} \sum_{j>k}^{G} \frac{\pi_j^n}{\pi_1^n} c^{p/2}\right)$$

La exponencial considerada en esta expresión tiende a 0, por lo que podríamos escribir, para $k_1, k_2$ constantes adecuadas, que

$$0 \leq X_n \leq k_1 + k_2\|x\|^2 = Y$$

Por la restricción **(PR)**, podemos afirmar que $\mathbb{E}(Y) < \infty$ por tener $X$ momento de segundo orden finito. En estas condiciones, aplicando el Teorema de la Convergencia Dominada de Lebesgue, se tiene que $\mathbb{E}(X_n) \to 0$. Si $\Psi^*$ es el límite de la subsucesión $\{(\pi_1^n, .., \pi_k^n, \mu_1^n, .., \mu_k^n, \Sigma_1^n, .., \Sigma_k^n)\}_{n=1}^{\infty}$, resulta que $\lim_{n\to\infty} \sup L(\Psi_n, P) = L(\Psi^*, P)$.

Como $\sum_{j=1}^{k} \pi_j < 1$, definimos $\{\tilde{\Psi}_n\}_{n=1}^{\infty} = \{(\tilde{\pi}_1^n, .., \tilde{\pi}_G^n, \tilde{\mu}_1^n, .., \tilde{\mu}_G^n, \tilde{\Sigma}_1^n, .., \tilde{\Sigma}_G^n)\}_{n=1}^{\infty}$ donde

$$\tilde{\pi}_j^n = \frac{\pi_j^n}{\sum_{j=1}^{k} \pi_j^n}\ para\ 1 \leq j \leq k, \quad \tilde{\pi}_{k+1}^n = .. = \tilde{\pi}_G^n = 0$$

$$\tilde{\mu}_j^n = \mu_j^n\ y\ \tilde{\Sigma}_j^n = \Sigma_j^n\ si\ 1 \leq j \leq k,\ elegidas\ arbitrariamente\ en\ \Theta\ si\ j > k$$

De esta manera, resulta

$$\lim_{n\to\infty} \sup L(\tilde{\Psi}_n, P) < \lim_{n\to\infty} \sup L(\Psi_n, P) = K$$

Esto llevaría a una contradicción con la optimalidad de $\{\Psi_n\}_{n=1}^{\infty}$, por lo que concluiríamos que necesariamente $k = G$.

□

Una vez comentados estos resultados, estamos en disposición de probar el Teorema 3.2 de existencia.

**Demostración**

Teniendo en cuenta los lemas anteriores, se debe dar una de las siguientes posibilidades:

1. Si $\pi_j^n \to \pi_j > 0$ para $1 \leq j \leq G$, la elección de $\Psi$ es trivial.
2. Si resulta que $\pi_j^n \to \pi_j > 0$ para $1 \leq j \leq k$ y $\pi_j^n = 0$ si $j > k$, podemos definir los pesos como $\pi_j = \lim_{n\to\infty} \pi_j^n$ para $1 \leq j \leq k$ y $\pi_{k+1} = .. = \pi_G = 0$. Para los vectores de medias y matrices de covarianza, tomamos $\mu_j = \lim_{n\to\infty} \mu_j^n$ y $\Sigma_j = \lim_{n\to\infty} \Sigma_j^n$ si $1 \leq j \leq k$ y arbitrarios (cumpliendo la restricción **(ER)**) si $j > k$.

□

### 3.2.2. Identificabilidad

La identificabilidad de un modelo es una propiedad matemática que se refiere a la capacidad de recuperar los parámetros reales del modelo a partir de los datos observados. Es necesaria para poder asegurar la existencia de estimadores consistentes en cualquier modelo estadístico, ya que sin ella es posible que existan varias soluciones para el problema de estimación de parámetros que maximicen (4). Además, supone un problema en el cálculo práctico de los estimadores ya que los algoritmos numéricos corren el riesgo de encontrar solo parte de estas soluciones o peor aún, el investigador que ajusta el modelo podría ni siquiera ser consciente de que la solución encontrada computacionalmente es solo una de muchas posibilidades.

En artículos como el de Teicher (1963) [11] se expone una condición necesaria y suficiente para que una clase de mixturas finitas sea identificable y se deduce la identificabilidad de todas las mixturas finitas gaussianas univariantes. Nuestro objetivo será demostrar este mismo resultado para el caso multivariante, para lo cual daremos antes unas definiciones más concretas sobre qué es la identificabilidad en modelos finitos de mixtura. Encontramos estos resultados en [13], donde no solo se comenta el caso del modelo de mezclas gaussiano sino que se contemplan varias familias de densidades más.

Decimos que una familia de funciones de densidad paramétricas es identificable cuando valores distintos de los parámetros determinan elementos distintos de la familia. Sea $\mathscr{H}$ una familia de funciones de densidad paramétricas. Cuando ajustamos una mixtura de G componetnes con elementos de $\mathscr{H}$ a una variable $X$ escribimos la función de densidad de $f$ como una combinación lineal convexa de las funciones de densidad $f(\cdot;\theta_j)$ de la clase $\mathscr{H}$:

$$f(x) = \sum_{j=1}^{G} \pi_j f(x;\theta_j)$$

Si $F(\cdot)$ y $F(\cdot;\theta_j)$ son las funciones de distribución asociadas a $f(\cdot)$ y $f(\cdot;\theta_j)$ respectivamente para $1 \leq j \leq G$, la línea anterior se puede escribir equivalentemente como

$$F(x) = \sum_{j=1}^{G} \pi_j F(x; \theta_j)$$

Trabajaremos con esta expresión análoga para aprovechar las propiedades de las funciones de distribución a la hora de dar definiciones o planear resultados.

Consideremos $\mathscr{F} = \{F(\cdot;\theta)/\theta \in \Theta\}$ una familia de funciones de distribución parmétricas y definamos $\mathscr{M}$ la familia de todas las mixturas finitas de la clase $\mathscr{F}$, que coincide con su envoltura convexa:

$$\mathscr{M} = \{F(\cdot;\Psi) = \sum_{j=1}^{G} \pi_j F(\cdot;\theta_j) : \ \pi_j \geq 0, \ \sum_{j=1}^{G} \pi_j = 1, \ \theta_j \in \Theta, \ G \in \mathbb{N}, \ 1 \leq j \leq G\}$$

Esta familia es la que consideramos cuando tratamos de ajustar un modelo de mezclas a un conjunto de datos o a una distribución de probabilidad. En este caso, decimos que $F(\cdot;\Psi)$ es identificable si es invariante bajo las $G!$ permutaciones de las etiquetas de $\Psi$:

**Definición 3.6.** *Sea $\mathscr{M}$ la familia paramétrica de distribuciones propia del modelo de mezclas. Sean $F(\cdot;\Psi) = \sum_{j=1}^{G} \pi_j F(\cdot;\theta_j)$ y $F'(\cdot;\Psi') = \sum_{j=1}^{H} \pi'_j F(\cdot;\theta'_j)$ dos elementos de $\mathscr{M}$. Decimos que $\mathscr{M}$ es identificable para $\Psi$ si $F(\cdot;\Psi) \equiv F'(\cdot;\Psi')$ si y solo si $G = H$ y podemos permutar las etiquetas de las componentes de mixtura de tal manera que $\pi_j = \pi'_j$ y $F(\cdot;\theta'_j) = F(\cdot;\theta_j)$ para $1 \leq j \leq G$.*

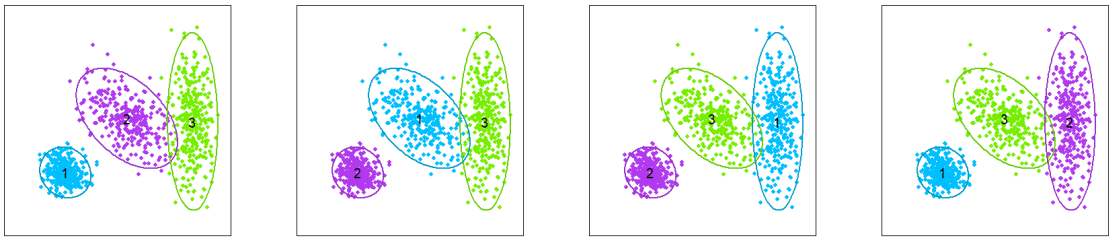



Figura 18: Diferentes permutaciones de un modelo identificable

Para evitar la falta de identificabilidad de $\Psi$ debido al intercambio de etiquetas entre las componentes, se suelen imponer restricciones sobre $\Psi$ como $\pi_1 \leq \pi_2 \leq ... \leq \pi_G$.

Se tiene la siguiente caracterización sobre las familias de funciones de distribución paramétricas identificables.

**Teorema 3.7.** *Una condición necesaria y suficiente para que la clase $\mathscr{M}$ sea identificable es que $\mathscr{F}$ sea un conjunto linealmente independiente sobre el cuerpo de los números reales.*

**Demostración**

Denotamos por $F_j = F(\cdot;\theta_j)$ y por $\langle A\rangle$ el conjunto de las combinaciones lineales finitas de elementos de $A$ sobre el cuerpo de los reales

$$\langle A\rangle = \Big\{\sum_{i=1}^{k}\alpha_i a_i : k\in\mathbb{N}, a_i\in A, \alpha_i\in\mathbb{R} \;\; para \;\; 1\leq i\leq k\Big\}$$

Supongamos que $\mathscr{F}$ no es linealmente independiente, esto es, existen $\alpha_1,..,\alpha_k\in\mathbb{R}$ no todos nulos tales que $\sum_{j=1}^{k}\alpha_j F_j = 0$. Estos $\alpha_j$ pueden ser estrictamente negativos, cero o estrictamente positivos. Si hay $M\in\{1,..,k\}$ coeficientes estrictamente negativos, reordenemos los elementos de tal manera que coloquemos al principio: $a_j<0$ si $1\leq j\leq M$. Así, se tendría

$$\sum_{j=1}^{M}|\alpha_j|F_j = \sum_{j=M+1}^{k}|\alpha_j|F_j$$

Dado que los elementos de $\mathscr{F}$ son funciones de distribución, sabemos que $\lim_{x\to\infty}F(x)=1$, por lo que haciendo tender $x$ a infinito en ambos lados de la igualdad, llegamos a que

$$\sum_{j=1}^{M}|\alpha_j| = \sum_{j=M+1}^{k}|\alpha_j| = \beta > 0$$

ya que al menos uno de los $\alpha_j$ es no nulo. Si tomamos $\gamma_j = \frac{|\alpha_j|}{\beta}$, entonces $\sum_{j=1}^{M}|\gamma|_j F_j = \sum_{j=M+1}^{k}|\gamma|_j F_j$ serían dos representaciones distintas de la misma mixtura finita, por lo que $\mathscr{M}$ no sería identificable.

Por otro lado, supongamos que $\mathscr{F}$ es linealmente independiente. Eso quiere decir que es una base de $\langle\mathscr{F}\rangle$, por lo que dos representaciones distintas de la misma mixtura supondrían una contradicción de las propiedades de representación única que tienen las bases, teniendo en cuenta que $\mathscr{M}$ sería un subespacio de $\langle\mathscr{F}\rangle$.

□

**Teorema 3.8.** *La familia $\mathscr{F}$ de funciones de distribución gaussianas $n-$variantes generan mixturas finitas identificables.*

Supongamos que $\mathscr{F}$ no es identificable y lleguemos a una contradicción. Por el teorema visto anteriormente, es equivalente suponer que la clase de funciones de distribución normales multivariantes de dimensión $n$ es un conjunto linealmente dependiente sobre $\mathbb{R}$. Esto implica que existen $\alpha_1,..,\alpha_k$ no todos nulos tal que $\sum_{j=1}^{k}\alpha_j F_j = 0$. Reescribimos como hacíamos en el resultado anterior, colocando en los primeros índices los $M$ elementos cuyo coeficiente es estrictamente negativo para tener

$$\sum_{j=1}^{M}\alpha_j F_j = \sum_{j=M+1}^{k}\alpha_j F_j \tag{5}$$

Es decir, en los dos sumatorios considerados, los coeficientes son no negativos.

Utilizaremos dos resultados relativos a la función característica de una variable aletoria y la relación entre la función de distribución de una variable y su función generadora de momentos. El

Teorema de Unicidad de la función característica asegura que distribuciones de probabilidad distintas tienen diferentes funciones características (ver Teorema 26.2 en Billingsley [1]). Como consecuencia, teniendo en cuenta que la integral de una función respecto de una mixtura de probabilidades es igual a la mixtura de las integrales, dadas $F_1, F_2, ...$ distribuciones de probabilidad con funciones características $\varphi_0, \varphi_1, ...$ y $\beta_i \geq 0$ con $\sum_{i=1}^{r} \beta_i = 1$, la mixtura $U = \sum_{i=1}^{r} \beta_i F_i$ es una distribución de probabilidad con función característica $\varphi = \sum_{i=1}^{r} \beta_i \varphi_i$.

Volviendo a nuestra expresión, consideremos $\sum_{j=1}^{M} \alpha_j = \gamma > 0$. Si dividimos entre $\gamma$ a ambos lados de la expresión (5), obtendríamos

$$\sum_{j=1}^{M} \frac{\alpha_j}{\gamma} F_j = \sum_{j=M+1}^{k} \frac{\alpha_j}{\gamma} F_j$$

Es claro que los coeficientes de la expresión de la izquierda son pesos (mayores o iguales que cero cuya suma es uno), por lo que $U = \sum_{j=1}^{M} \frac{\alpha_j}{\gamma} F_j$ sería una distribución de probabilidad. Además, como las $F_j$ son funciones de distribución, se tiene que

$$\lim_{x \to \infty} \sum_{j=M+1}^{k} \frac{\alpha_j}{\gamma} F_j(x) = \sum_{j=M+1}^{k} \frac{\alpha_j}{\gamma} = \sum_{j=1}^{M} \frac{\alpha_j}{\gamma} = \lim_{x \to \infty} \sum_{j=1}^{M} \frac{\alpha_j}{\gamma} F_j(x)$$

Pero por cómo hemos definido $\gamma$, $\sum_{j=1}^{M} \frac{\alpha_j}{\gamma} = 1$, por lo que los coeficientes $\frac{\alpha_{M+1}}{\gamma}, .., \frac{\alpha_k}{\gamma}$ son pesos y se tiene que $V = \sum_{j=M+1}^{k} \frac{\alpha_j}{\gamma} F_j$ es una distribución de probabilidad también.

Dado que $U$ y $V$ son distribuciones de probabilidad iguales, necesariamente sus funciones características también lo son. Si $\varphi_j$ denota la función característica de $F_j$, se tendría entonces

$$\sum_{j=1}^{M} \frac{\alpha_j}{\gamma} \varphi_j = \sum_{j=M+1}^{k} \frac{\alpha_j}{\gamma} \varphi_j \Rightarrow \sum_{j=1}^{M} \alpha_j \varphi_j = \sum_{j=M+1}^{k} \alpha_j \varphi_j \Rightarrow \sum_{j=1}^{k} \alpha_j \varphi_j = 0$$

Denotando por $\mu_j$ y $\Sigma_j$ el vector de medias y la matriz de coviarianzas de $F_j$ respectivamente, dado que $F_j$ es la función de distribución de una normal $N(\mu_j, \Sigma_j)$, su función característica es $\varphi_j = \exp\{\frac{1}{2} x^T \Sigma_j x + x^T \mu_j\}$. Por lo tanto

$$\sum_{j=1}^{k} \alpha_j \exp\{\frac{1}{2} t^T \Sigma_j t + t^T \mu_j\} = 0$$

Si tomamos $t = cu$ donde $c$ es un escalar y $u$ es un vector, obtenemos

$$\sum_{j=1}^{k} \alpha_j \exp\{\frac{c^2}{2} u^T \Sigma_j u + u^T \mu_j\} = 0 \tag{6}$$

Si todas las matrices de covarianza son iguales, los vectores de medias deben ser todos diferentes entre sí para que exista diferencia de parámetros entre los índices. Por lo tanto, salvo para un número finito de hiperplanos, los pares de números reales $(u^T \Sigma_j u, u^T \mu_j)$ son distintos para cada $j \in \{1, .., k\}$. De otro modo, supongamos que $\Sigma_1, .., \Sigma_N$ son las únicas matrices de covarianza distintas entre

$\Sigma_1, .., \Sigma_k$. Entonces, para un vector $u$ que no pertenezca a un número finito de cónicas, los números reales $u^t \Sigma_j u$ son distintos para $1 \leq j \leq N$. Para los índices $N+1, .., k$, los $\mu_j$ son diferentes entre sí, por lo que para los vectores que no pertenezcan a un número finito de hiperplanos y cónicas, los pares $(u^T \Sigma_j u, u^T \mu_j)$ son distintos dos a dos, $1 \leq j \leq k$. Si elegimos un $u$ con estas características, lo que estaríamos diciendo en (6) es que la clase de las mixturas finitas de normales univariantes no es identificable, lo cual contradice los resultados demostrados por Teicher [11].

## 3.3. Consistencia

Una vez planteado el problema de existencia, buscamos evaluar si el modelo finito de mixturas produce resultados estables y consistentes a medida que se aumenta el tamaño de la muestra.

El marco de trabajo característico de la Estadística Matemática es el escenario en el que tenemos un conjunto de $n$ individuos $\{x_1, .., x_n\}$ que toman medidas en $p$ variables y se considera que estas observaciones se obtienen como muestra en un modelo probabilístico donde el mecanismo generador de los datos $P$ es desconocido. A la hora de ajustar un modelo de mixturas a una variable aleatoria $X$ con distribución $P$, los vectores de medias y las matrices de covarianza que hallamos son concretos y no dependen de la aleatoriedad. Sin embargo, estimar estos mismos parámetros para una muestra dada depende del $\omega$ escogido, por lo que obtenemos valores distintos al generar muestras nuevas.

Buscamos por lo tanto estudiar la consistencia estadística del modelo, es decir, analizar la capacidad del método para obtener resultados consistentes y estables a medida que tamaño de la muestra aumenta indefinidamente. En términos prácticos, nos preguntamos si los vectores de medias y matrices de covarianza que hallamos para una muestra se acercan a aquellos de la distribución de probabilidad teórica.

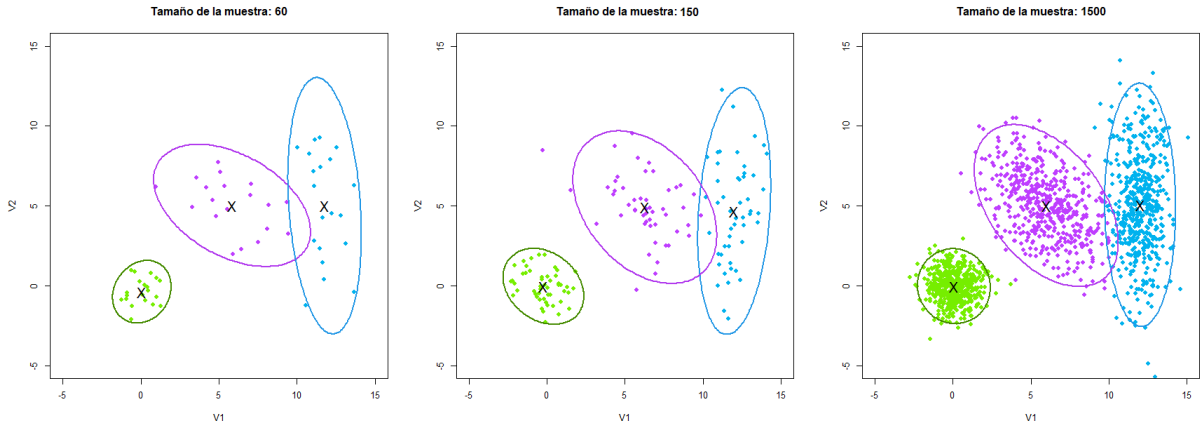



Figura 19: Vectores de medias y matrices de covarianza según aumenta el tamaño de la muestra

En esta figura podemos ver por ejemplo el resultado de estimar los vectores de medias y las ma-

trices de covarianza en el modelo de mixturas cuando generamos una muestra de tamaño 60, 150 y 1500 obsevaciones respectivamente de una combinación balanceada de tres normales multivariantes $X_1 \sim N((0,0), \left(\begin{smallmatrix} 1 & 0 \\ 0 & 1 \end{smallmatrix}\right))$, $X_2 \sim N((6,5), \left(\begin{smallmatrix} 4 & -2 \\ -2 & 4 \end{smallmatrix}\right))$ y $X_3 \sim N((12,5), \left(\begin{smallmatrix} 1 & 0 \\ 0 & 10 \end{smallmatrix}\right))$.

Podemos observar a grandes rasgos que, a medida que aumenta el tamaño de la muestra considerada, se estabilizan los valores de $\mu$ y $\Sigma$ para cada distribución normal. Este es el resultado que probaremos de manera teórica en esta subsección, cuya demostración será muy similar al que llevábamos a cabo en el apartado de existencia.

Por último, decir que las propiedades de consistencia permiten extraer información sobre la capacidad del método para identificar patrones subyacentes en los datos y generar agrupamientos coherentes y significativos.

**Definición 3.9.** *Sea $X_0$ un vector aleatorio definido en el espacio probabilístico $(\Omega, \sigma, \mathbb{P})$ con llegada en $\mathbb{R}^d$. Sean $X_1, ..., X_n, ...$ vectores aleatorios i.i.d. definidas en el mismo espacio con $\mathscr{L}(X) = \mathscr{L}(X_i) = P$. Se define $P_n^w$ la distribución de probabilidad empírica como aquella función de masa de probabilidad discreta que otorga masa $\frac{1}{n}$ a cada $X_i(\omega)$, $i = 1, .., n$.*

Es claro que, dado que $P_n^\omega$ depende del $\omega \in \Omega$ elegido, también lo harán los parámetros que estamos estimando. Denotamos por $\{\Psi_n^\omega\}_{n=1}^\infty = \{(\pi_1^{n,\omega}, .., \pi_G^{n,\omega}, \mu_1^{n,\omega}, .., \mu_G^{n,\omega}, \Sigma_1^{n,\omega}, ..\Sigma_G^{n,\omega})\}_{n=1}^\infty$ la sucesión de estimadores empíricos donde cada $\Psi_n^\omega$ es el vector de parámetros que maximiza

$$L(\Psi, P, \omega) = \mathbb{E}_{P_n}\left(log\left(\sum_{j=1}^{G} \pi_j^{n,w} \phi_j^{n,w}\right)\right) = \sum_{i=1}^{n}\left(\log\left(\sum_{j=1}^{G} \pi_j^{n,w} \phi(X_i(w); \mu_j^{n,w}, \Sigma_j^{n,w})\right)\right) \tag{7}$$

donde el tamaño de la muestra es $n$.

La propiedad de identificabilidad de la familia de mixturas considerada anteriormente es necesaria si queremos asegurar la consistencia del problema de máxima verosimilitud, como veremos en el siguinte lema.

**Lema 3.10.** *Consideramos $\mathscr{M}'$ una familia identificable de mixturas finitas definida como*

$$\mathscr{M}' = \{f(\cdot; \Psi) = \sum_{j=1}^{G} \pi_j f(\cdot; \theta_j) : \ \pi_j \geq 0, \ \sum_{j=1}^{G} \pi_j = 1, \theta_j \in \Theta, \ G \in \mathbb{N}, \ 1 \leq j \leq G\}$$

*Sea $X$ una variable aleatoria definida en un espacio probabilístico con densidad $f(\cdot; \Psi_0) \in \mathscr{M}'$. Denotemos por $\mathbb{E}_{\Psi_0}(Y)$ el valor esperado de una variable $Y$ cuando el valor verdadero del parámetro es $\Psi_0$. Entonces, se tiene que*

$$\mathbb{E}_{\Psi_0} \log f(X, \Psi) < \mathbb{E}_{\Psi_0} \log f(X, \Psi_0)$$

*Para cualquier $\Psi \neq \Psi_0$ en el espacio paramétrico.*

**Demostración**

Dado que $\mathscr{M}$ es identificable, para vectores de parámetros distintos $\Psi_0, \Psi_1$, $f(x;\Psi) \neq f(x;\Psi_0)$ para al menos un valor de $x$. Como las funciones de $\mathscr{F}$ son continuas, podemos asegurar que existe un entorno de $x$ donde se tiene bien $f(x;\Psi_1) > f(x;\Psi_0)$ o bien $f(x;\Psi_1) < f(x;\Psi)$. Esto quiere decir que el conjunto $\{x \in \mathbb{R}^p : f(x;\Psi) \neq f(x;\Psi_0)\}$ tiene medida de Lebesgue estrictamente positiva. Si escribimos $\log(f(X;\Psi)) - \log(f(X;\Psi_0)) = \log\Big(\frac{f(X;\Psi)}{f(X;\Psi_0)}\Big)$, como la función logaritmo es estrictamente cóncava, la desigualdad de Jensen estricta establece que

$$\mathbb{E}_{\Psi_0} \log\Big(\frac{f(X;\Psi)}{f(X;\Psi_0)}\Big) < \log \mathbb{E}_{\Psi_0}\Big(\frac{f(X;\Psi)}{f(X;\Psi_0)}\Big) = \int \frac{f(X;\Psi)}{f(X;\Psi_0)} dP(\Psi_0) =$$

$$= \int \frac{f(X;\Psi)}{f(X;\Psi_0)} f(X;\Psi_0)dx = \int f(X;\Psi)dx = \log(1) = 0$$

Donde la última igualdad es consecuencia de que $f(X;\Psi)$ sea una función de densidad. □

**Teorema 3.11.** *En el marco descrito anteriormente, sea $\{\Psi_n^\omega\}_{n=1}^\infty$ una sucesión de estimadores empíricos donde n denota el tamaño de la muestra. Suponemos que $X_0$ es un vector aleatorio definido en el espacio probabilístico $(\Omega, \sigma, \mathbb{P})$ con llegada a $\mathbb{R}^d$ siendo $\Psi_0$ el único vector de parámetros que maximiza la verosimilitud (4). Entonces, se tiene que $\Psi_n^\omega \longrightarrow \Psi_0$ para $\omega \in \Omega_0 \subset \Omega$ con $\mathbb{P}(\Omega_0) = 1$.*

**Observación 3.12.** *Se puede calcular la sucesión de estimadores empíricos $\{\Psi_n^w\}_{n=1}^\infty$ (cuya existencia está garantizada) resolviendo el problema de maximización de la función de verosimilitud correspondiente para cada n, a menudo mediante procedimientos iterativos. Cabe destacar que cada uno de los $\Psi_n^w$ puede no ser único.*

Para demostrar este teorema, probaremos en primer lugar que existe un conjunto compacto K en el espacio paramétrico tal que para un $n$ suficientemente avanzado, $\Psi_n^w \in K$ casi seguro. Sean $X_1, ..., X_n, ...$ vectores aleatorios i.i.d. definidos en el mismo espacio con $\mathscr{L}(X) = \mathscr{L}(X_i) = P$, y sea $P_n^w$ la distribución de probabilidad empírica para $X_1(\omega), ..., X_n(\omega)$.

De nuevo, impondremos las restricciones **(PR)** y **(ER)** sobre P que enunciábamos en el apartado anterior. Se verificará una versión de los dos lemas que utilizábamos previamente:

**Lema 3.13.** *Dada la sucesión $\{\Psi_n^w\}_{n=1}^\infty$ y dadas las restricciones* **(ER)** *y* **(PR)**, *no puede darse ni $m_n \to 0$ ni $M_n \to \infty$, donde $m_n$ y $M_n$ denotaban el autovalor más pequeño y el autovalor más grande de las matrices de covarianza respectivamente.*

Este lema se prueba con un razonamiento similar al que llevábamos a cabo en el lema (3.3). En este caso, también necesitamos una cota del tipo

$$\mathbb{E}_{P_n}\left(\max_{1\leq j\leq G} \|x - \mu_j^{n,w}\|^2\right) \geq h'$$

Para un $h' > 0$. Este resultado auxiliar está disponible en el Apéndice en (5.6) y utiliza el argumento de que la clase de las bolas en $\mathbb{R}^p$ es una clase de Glivenko-Cantelli. Por otro lado, también se verifica el siguiente resultado:

**Lema 3.14.** *Dada la sucesión $\{\Psi_n^w\}_{n=1}^\infty$ y dadas las restricciones* **(ER)** *y* **(PR)**, *es posible elegir los vectores de medias empíricos $\mu_1^{n,w}, .., \mu_G^{n,w}$ de tal manera que sus normas estén uniformemente acotadas con probabilidad 1.*

De nuevo, la demostración de este lema es muy similar a la que exponíamos anteriormente en (3.5). Es decir, dado que para los pesos $\pi_j^{n,w}$ era sencillo encontrar un compacto que los contuviese, podemos afirmar que existen $n_0$ y $K$ tal que $\Psi_n^w \in K$ si $n \geq n_0$.

A continuación, ahora que sabemos que la sucesión $\{\Psi_n^w\}_{n=1}^{\infty}$ es uniformemente ajustada, buscamos probar la convergencia casi seguro de $\Psi_n^w$ a $\Psi_0$. Para ello, recurriremos a varios resultados presentes en el libro Vaart and Wellner [12], donde se utiliza una extensión del Teorema de Glivenko-Cantelli para clases de conjuntos y funciones más generales. En el Apéndice se recoge más información sobre los conceptos necesarios para plantear estos resultados, como las clases de Glivenko-Cantelli o las de Vapnik-Chervonenkis.

Consideremos, dado un compacto $K$ en el espacio paramétrico, la clase de funciones

$$\mathscr{H} = \Big\{ \log\Big( \sum_{j=1}^{G} \pi_j \phi_j \Big) : \theta \in K \Big\}$$

El Teorema 3.2.3 presente en Vaart and Wellner [12] nos permite asegurar que se da la convergencia de $\Psi_n^w$ a $\Psi_0$ con probabilidad 1 si demostramos que la clase $\mathscr{H}$ es de Glivenko Cantelli. Es decir, buscamos demostrar la siguiente proposición.

**Proposición 3.15.** *La clase $\mathscr{H}$ definida anteriormente satisface*

$$\|P_n^w - P\|_{\mathscr{H}} = \sup_{h \in \mathscr{H}} |P_n^w(h) - P(h)| = \sup_{h \in \mathscr{H}} \Big| \int h dP_n^w - \int h dP \Big| \longrightarrow 0 \quad c.s.$$

**Demostración**
Para probar esto, definiremos la clase

$$\mathscr{G} = \Big\{ I_B(\cdot) \log\Big( \sum_{j=1}^{G} \pi_j \phi_j \Big) : \theta \in K \Big\}$$

Donde $B$ es un conjunto compacto fijo. Demostraremos que esta clase de funciones es de Glivenko-Cantelli, lo que nos ayudará a concluir que $\mathscr{H}$ también lo es. Se tiene que $\Phi = \{\phi : \mu \in \mathbb{R}^p, \Sigma \in M_{p\times p}\}$ es un espacio vectorial de funciones medibles de dimensión finita. Por el lema 2.6.15 de [12] , sabemos que su dimensión de Vapnik-Chervonenkis es finita y por lo tanto es una clase $VC$.

Dado que $\pi_1, .., \pi_G$ son pesos, $\{\sum_{j=1}^{G} \pi_j \phi_j : \theta \in K\}$ es la envoltura convexa de $\Phi$, por lo que es una clase envolvente (o clase H) de Vapnik Chervonenkis. Si aplicamos el Teorema 2.10.20 presente en [12] con $\phi(x) = I_B(x)\log(x)$, se tiene que $\mathscr{G}$ satisface la condición de entropía uniforme. Como está acotada uniformemente, $\mathscr{G}$ es una clase de Glivenko-Cantelli.

A continuación, veremos que existen $a$ y $b$ para acotar el tamaño de cada elemento $h$ de $\mathscr{H}$ :

$$|h(x)| \leq a\|x\|^2 + b$$

- Cota inferior:

Dado que $K$ es un compacto y las matrices son definidas positivas, existen constantes positivas $m$ y $M$ que acotan superior e inferiormente todos los autovalores de las $G$ matrices de covarianza. Resulta entonces:

$$\sum_{j=1}^{G} \pi_j \phi_j \geq \sum_{j=1}^{G} \pi_j \frac{1}{(2\pi M)^{-p/2}} \exp\{\frac{-1}{2m}\|x-\mu_j\|^2\}$$

Por la identidad del paralelogramo, se tiene que $\|x-\mu\|^2 \leq 2(\|x\|^2 + \|\mu\|^2)$. Por lo tanto, resulta que

$$\sum_{j=1}^{G} \pi_j \frac{1}{(2\pi M)^{-p/2}} \exp\{\frac{-1}{2m}\|x-\mu_j\|^2\} \geq \exp\{\frac{-1}{m}\|x\|^2\} \sum_{j=1}^{G} \pi_j \frac{1}{(2\pi M)^{-p/2}} \exp\{\frac{-1}{m}\|\mu_j\|^2\}$$

Habíamos demostrado anteriormente que $\text{máx}_{1\leq j\leq G}\|\mu_j\| < \infty$, así que existen $a'$ y $b'$ tales que

$$\log\left(\sum_{j=1}^{G} \pi_j \phi_j\right) \geq a'\|x\|^2 + b'$$

- Cota superior:

Sabiendo que existen $m, M > 0$, escribimos

$$\sum_{j=1}^{G} \pi_j \phi_j \leq \sum_{j=1}^{G} \pi_j \frac{1}{(2\pi m)^{-p/2}} \exp\{\frac{-1}{2M}\|x-\mu_j\|^2\} \leq \sum_{j=1}^{G} \pi_j \frac{1}{(2\pi m)^{-p/2}}$$

Donde la última desigualdad se tiene ya que el exponente de la función exponencial es negativo y por lo tanto, toda ella es un factor menor que uno. Llegamos a que existen $a''$ y $b''$ con

$$\log\left(\sum_{j=1}^{G} \pi_j \phi_j\right) \leq a''\|x\|^2 + b''$$

De esta manera, tomando adecuadamente $a$ y $b$ teniendo en cuenta las dos desigualdades, llegamos a que $|h(x)| \leq a\|x\|^2 + b \ \forall h \in \mathscr{H}$.

Sabiendo esto, para cada $h \in \mathscr{H}$ y para $B$ un compacto prefijado de $\mathbb{R}^p$, se tiene

$$|\mathbb{E}_{P_n^w}[h(\cdot)] - \mathbb{E}_P[h(\cdot)]| = |\mathbb{E}_{P_n^w}[h(\cdot)I_B(\cdot)] + \mathbb{E}_{P_n^w}[h(\cdot)I_{B^c}(\cdot)] - \mathbb{E}_P[h(\cdot)I_B(\cdot)] - \mathbb{E}_P[h(\cdot)I_{B^c}(\cdot)]| \leq$$

$$\leq |\mathbb{E}_{P_n^w}[h(\cdot)I_B(\cdot)] - \mathbb{E}_P[h(\cdot)I_B(\cdot)]| +$$

$$+|\mathbb{E}_{P_n^w}[(a\|x\|^2+b\|x\|+c)I_{B^c}(\cdot)]-\mathbb{E}_P[(a\|x\|^2+b\|x\|+c)I_B(\cdot)]|$$

La expresión anterior tiende a 0 cuando $n$ tiende a infinito teniendo en cuenta que $h(\cdot)I_B(\cdot)\in\mathscr{G}$, que es una clase de Glivenko-Cantelli. Esto provoca que

$$\|P_n^w-P\|_{\mathscr{H}}=\sup_{h\in\mathscr{H}}|\int hdP_n^w-\int hdP|=\sup_{h\in\mathscr{H}}|\mathbb{E}_{P_n^w}(h(\cdot))-\mathbb{E}_P[h(\cdot)]|\longrightarrow 0 \quad c.s.$$

Y que por lo tanto $\mathscr{H}$ sea una clase de Glivenko-Cantelli.

□

Una vez probados estos lemas, procedemos a demostrar el Teorema 3.11 de Consistencia que enunciábamos anteriormente.

**Demostración**

Teniendo en cuenta que $\Psi_0$ es único y que los lemas anteriores que aseguran que $\{\Psi_n^w\}_{n=1}^{\infty}$ es uniformemente ajustada y $\mathscr{H}$ es una clase de Glivenko-Cantelli, podemos afirmar que $\Psi_n^w\to\Psi_0$ con probabilidad uno utilizando el Corolario 3.2.3 presente en [12]. □

# 4. Aspectos Metodológicos

Tras presentar el modelo de mixturas y haber estudiado sus principales propiedades matemáticas, abordamos en esta sección el cálculo práctico de estimaciones de los parámetros del modelo, limitándonos al modelo de mezclas gaussiano.

En primer lugar, hablaremos sobre las restricciones geométricas que se pueden introducir en el modelo para aliviar el número de parámetros a estimar, tras lo cual introduciremos algunos criterios útiles para inferir las características geométricas de los grupos. Otra cuestión importante es determinar el número de componentes del modelo cuando esta información no es conocida a priori y debemos inferirla de los datos, al igual que considerar el modelado de datos con agrupaciones no gaussianas fusionando clusters creados a partir del modelo de mezclas gaussiano. Recientemente, se han propuesto multitud de criterios para todas estas cuestiones, muchos de los cuales tienen resultados alentadores en estudios empíricos para estudiar su actuación.

Por otro lado, presentaremos el algoritmo EM incidiendo en el impacto que tuvo en el problema de ajuste del modelo de mezclas y detallaremos los pasos del mismo. Comentaremos someramente sus propiedades de convergencia bajo ciertas restricciones e introduciremos algunos procedimientos relativos a la incialización del algoritmo.

## 4.1. Restricciones geométricas en el modelo de mezclas

Volvamos a la situación en la que consideramos un conjunto de $n$ observaciones $p$-variantes $\{x_1, .., x_n\}$ y buscamos expresar su función de densidad a través de una combinación convexa de $G$ funciones de densidad gaussianas:

$$f(x_i) = \sum_{j=1}^{G} \pi_j \phi(x_i; \mu_j, \Sigma_j) = \sum_{j=1}^{G} \pi_j \phi_j(x_i),\ 1 \leq i \leq n \tag{8}$$

Analicemos el número de parámetros que debemos estimar en este modelo:

1. En primer lugar, será necesario estimar $(G-1)$ de los pesos $\pi_j$ ya que $\pi_G$ se puede obtener como $1 - \sum_{j=1}^{G-1} \pi_j$.

2. En el caso de los vectores de medias, será necesario estimar $G$ vectores $p-$dimensionales, por lo que tendríamos que calcular $Gp$ parámetros.

3. Para finalizar, dado que contamos con $G$ matrices de covarianza (simétricas, solo es necesario obtener el triángulo superior de la matriz para reconstruirla) que son de tamaño $p \times p$, en este caso acumulamos $\frac{Gp(p+1)}{2}$ parámetros a estimar.

En total, el número de parámetros para el modelo planteado en (8) sería $(G-1) + Gp + \frac{Gp(p+1)}{2}$. En un caso muy sencillo en el que contásemos con datos bidimensionales y buscásemos ajustar un modelo con dos componentes, tendríamos que calcular 11 parámetros, algo que sin algoritmos y computadoras que realicen los cálculos resultaría bastante tedioso. Pero no solo eso: este número puede crecer rápidamente y ser bastante elevado en el momento en el que aumentemos el número de variables $p$, considerando conjuntos de datos en dimensiones mayores. Esto hace patente la necesidad

de un algoritmo que sea capaz de computar de manera eficiente estimaciones de los parámetros y de mitigar la carga de parámetros cuando los grupos compartan alguna característica común en su estructura y podamos simplificar las hipótesis del problema.

Todos los modelos tienen $Gp$ parámetros para los vectores de medias y $(G-1)$ para los pesos de las componentes. Sin embargo, con el fin de aliviar el problema, es común utilizar versiones más parsimoniosas del modelo que supongan propiedades de las matrices de covarianza, como por ejemplo igualdad de matrices de covarianza entre componentes, esfericidad o igualdad de volúmenes. Una forma de recoger estas condiciones es considerar la descomposición VSO (Volume-Shape-Orientation) de las matrices de covarianza, dada de la siguiente manera:

$$\Sigma_j = V_j E_j D_j E_j^T \tag{9}$$

En esta expresión, $V_j$ es una constante de proporcionalidad, $D_j = \{\lambda_{j1}, .., \lambda_{jp}\}$ es la matriz que tiene en su diagonal valores proporcionales a los autovalores de $\Sigma$ ordenados de mayor a menor de tal manera que su determinante sea 1, y $E_j$, $E_j^T$ son las matrices de cambio de base formadas por los autovectores unitarios asociados a los autovalores. Podemos asociar cada elemento en la descomposición con una propiedad geométrica:

- La matriz de autovectores $E_j$ determina la orientación en $\mathbb{R}^p$ de la componente.
- La matriz diagonal $D_j$ determina la forma de la componente: Si resulta que el autovalor $\lambda_{j1}$ es mucho mayor que $\lambda_{j2}$, nos dice que la componente $j-$ésima de la mezcla está concentrada prácticamente en un subespacio de dimensión uno. En el caso opuesto, si todos los valores $\lambda_{ji}$ son aproximadamente iguales, esto nos dice que la mixtura $j-$ésima es prácticamente esférica.
- La constante de proporcionalidad $V_j$ determina el volumen que ocupa la componente en $\mathbb{R}^d$.

A la hora de ajustar un modelo de mixturas, es frecuente que las componentes compartan alguna de estas características geométricas. Estas consideraciones llevan a analizar submodelos en los que alguna de las variables volumen, forma u orientación puede suponerse constante entre las componentes. Además, puede considerarse que la matriz de covarianzas es la identidad para cada $j$ (es decir, las componentes son esféricas ) o que es diagonal (las componentes están alineadas con los ejes). Esto permite hacer una división de dos modelos en el caso univariante y de 14 en el caso multivariante:

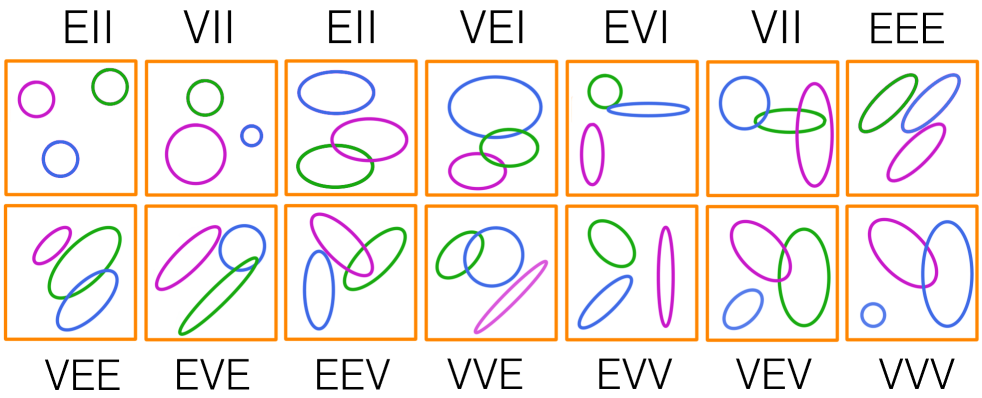



Figura 20: Ejemplo sobre el aspecto que tendrían tres componentes gaussianas en el caso bivariante para cada una de las 14 categorias VSO.

Cada uno de los 14 modelos se describe a partir de tres letras, que se refieren al Volumen, Forma y Orientación. El volumen puede ser igual entre componentes (E) o variable (V). Respecto a la forma, de manera similar, la letra E denota que las matrices $D_j$ son iguales para todas las componentes, la letra V si no se impone ningua condición al respecto y la letra "I" en el caso de que los clusters sean esféricos. Para terminar, la letra "E" impone que las orientaciones inducidas por las matrices de autovectores $E_j$ son iguales para todos los grupos, "V" si se impone ninguna condición sobre la orientación e "I" si las componentes son esféricas, en cuyo caso $E_j = I_p$, $1 \leq j \leq G$.

En ocasiones estos modelos en los que hemos impuesto ciertas condiciones sobre $\Sigma$ tienen muchos menos parámetros que el modelo planteado en (8) y ajustan los datos prácticamente igual de bien. Cuando esto sucede, obtendremos estimaciones más precisas, con un grado mayor de interpretabilidad. En cuanto consideramos conjuntos de datos con dimensión relativamente grande, el potencial ganado es inmenso. Sin embargo, estos modelos deben ser solo utilizados cuando la estructura intrínseca de los datos lo permite, ya que un modelo más parsimonioso no tiene por qué ajustar los datos mejor en un caso general sino que además podría hacer que perdiésemos la capacidad de capturar la disposición y la variabilidad de las observaciones. Veremos a continuación un criterio para escoger el tipo de modelo de acuerdo con la descomposición VSO que más parece ajustarse a nuestros datos.

## 4.2. Elección del Número de Componentes y el modelo de Clustering

Una vez descritos los 14 diferentes modelos en función de sus características geométricas, la siguiente pregunta que nos hacemos es cómo detectar en cada caso concreto cuál sería el mejor modelo a utilizar en función de las características de los datos con los que trabajamos. Una cuestión íntimamente relacionada con esto es determinar el número de componentes del modelo, ya que el número de clusters viene determinado por la naturaleza de los datos y las relaciones que albergan los individuos entre sí. Estas dos elecciones constituyen dos problemas imposibles de resolver

si consideramos un conjunto de datos arbitrario, por lo que no existe un método general para hacerlo.

En el contexto de los modelos de mixtura, estas dos preguntas pueden reducirse a una única cuestión si consideramos que cada combinación de un número de componentes y una elección del VSO corresponde con un modelo Estadístico diferente. En esa situación, el problema se restringe a realizar una comparación entre un conjunto de posibles modelos para los datos. Usualmente, debemos llegar a un compromiso entre el modelo VSO escogido y el número de componentes, ya que el un modelo con menos parámetros a menudo necesitará de un mayor número de clusters para poder describir la variabilidad de los datos.

Un enfoque posible para seleccionar un modelo de clustering es el Criterio de Información Bayesiano (BIC). La idea básica es que, dados $M_1, .., M_K$ modelos estadísticos con probabilidades a priori $p(M_k)$, $1 \leq k \leq K$, el teorema de Bayes asegura que la probabilidad a posteriori de un modelo $M_k$ dado un conjunto de datos $D$ es proporcional a la probabilidad de los datos dado el modelo $M_k$ multiplicado por $p(M_k)$:

$$p(M_k|D) = \frac{p(D, M_k)}{p(D)} = \frac{p(D|M_k)}{p(D)} P(M_k)$$

Cuando los modelos tienen parámetros desconocidos, por la Ley de Probabilidad Total, $p(D|M_k)$ se obtiene integrando respecto de los parámetros:

$$p(D|M_k) = \int p(D|\theta_{M_k}, M_k) p(\theta_{M_k}|M_k) d\theta_{M_k}$$

Donde $\theta_{M_k}$ es el vector de parámetros del modelo $M_k$ y $p(\theta_{M_k}|M_k)$ es la distribución a priori de $\theta_{M_k}$. La cantidad $p(D|M_k)$ se denomina verosimilitud marginal del modelo $M_k$. Dado que buscamos escoger el modelo que es más probable a posteriori y usualmente se presupone que las probabilidades a priori $p(M_k)$ son iguales para todos los modelos, escogeremos aquel modelo con mayor verosimilitud marginal.

Para comparar dos modelos $M_1$ y $M_2$ el factor de Bayes se define como el ratio de las verosimilitudes marginales:

$$B_{12} = \frac{p(D|M_1)}{p(D|M_2)}$$

De este modo, diríamos que bajo este criterio el modelo $M_1$ es más favorable que el $M_2$ si $B_{12}$ es mayor que uno y viceversa. Para tener evidencia de que un modelo es mucho mejor que otro, establecemos umbrales más altos que uno (por ejemplo, $B_{12} > 100$). Con el fin de utilizar una escala más sencilla, se suele considerar $2\log(B_{12})$ (ahora por ejemplo el umbral para tener evidencias estadísticas fuertes de que un modelo es mejor que otro podría ser 10 en lugar de 100). Es un criterio que en el modelo de mixturas resulta bastante adecuado ya que permite considerar más de dos modelos y estos no tienen por qué estar necesariamente anidados.

Para evaluar la integral que hemos descrito anteriormente, si el modelo es regular, podemos aproximarla por el criterio BIC:

$$2\log p(D|M_k) \cong 2\log p(D|\hat{\theta}_{M_k}, M_k) - \beta_{M_k}\log(n) = BIC_{M_k}$$

En la expresión, $\beta_{M_k}$ denota el número de parámetros independientes a estimar en el modelo $M_k$. Como podemos ver, el BIC penaliza los modelos con un mayor número de parámetros, con el fin de evitar el sobreajuste y buscar modelos más parsimoniosos e interpretables. Los modelos de mezclas no satisfacen por lo general las condiciones que permiten asegurar que la verosimilitud marginal pueda ser aproximada por el BIC (Sin ir más lejos, la verosimilitud no está acotada en el caso del modelo de mixturas gaussiano). Sin embargo, considerando restricciones como las que hemos planteado anteriormente sobre la probabilidad $P$ y el ratio entre autovalores de las matrices de covarianzas, nos permiten afirmar que la aproximación es válida.

En el caso de considerar modelos de mezclas gaussianos multivariantes, los modelos que buscamos comparar corresponden con diferente número de componentes y distintos conjuntos de restricciones sobre las matrices de covarianza. Si fijamos $G_{max}$ el número máximo de componentes razonable, consideraríamos $14 \times G_{max}$ modelos diferentes.

El criterio BIC está confeccionado para elegir no tanto el número de clusters del conjunto de datos sino el número de componentes en un modelo de mezclas. Esta diferencia parece sutil pero es importante, ya que en ocasiones podemos encontrar agrupaciones en los datos que no pueden describirse adecuadamente mediante una sola componente gaussiana, mientras que sí conseguimos representarla por medio de varias componentes. En el caso en el que estemos interesados en estimar el número de clusters en lugar de encontrar el mejor modelo de mixturas, podemos utilizar el criterio Integrated Completed Likelihood (ICL) en lugar del BIC. Este criterio está basado en la verosimilitud conjunta de los datos observados $x_i$ y las etiquetas desconocidas $z_{ij}$, de modo que quedaría de la siguiente manera

$$ICL = BIC - \sum_{i=1}^{n}\sum_{j=1}^{G_{M_k}} \hat{z}_{ij}\log(\hat{z}_{ij})$$

El último término es la entropía esperada de la clasificación por medio del modelo $M_k$, por lo que el ICL penaliza la incertidumbre en la clasificación. En consecuencia, el ICL se decanta más por modelos que producen separaciones más claras entre los grupos que el BIC, por lo que en la práctica el ICL suele elegir un número menor o igual de clusters que el BIC.

A la hora de considerar conjuntos formados por agrupaciones claramente gaussianas y bien separadas, los criterios BIC e ICL coinciden. Sin embargo, presentamos a continuación dos conjuntos de datos en los que encontramos distribuciones de puntos que no provienen en un principio de una mixtura gaussiana.

En el primer conjunto, ambos criterios seleccionan tanto modelos VSO como número de componentes distintas: según el BIC, el mejor modelo sería "EEV" con 7 componentes, mientras que para el caso del ICL, el mejor ajuste sería "VVI" con 5 clusters. El modelo EEV considera componentes de igual forma y volumen pero con orientaciones variables, argumento que tiene sentido si desmenuzamos el conjunto de datos en clusters más pequeños y construimos cada agrupación con varios, ajustando su orientación para cubrir bien cada componente. Por otro lado, el modelo VVI explica los

datos por medio de elipsoides de volumen y forma variable pero alineados con los ejes de abscisas y ordenadas. Vemos que en esta ocasión ambos modelos consiguen recuperar las agrupaciones iniciales salvo por el conjunto de dos observaciones apartadas del resto, si bien la partición creada por el modelo elegido con el criterio ICL concuerda más con la impresión visual.

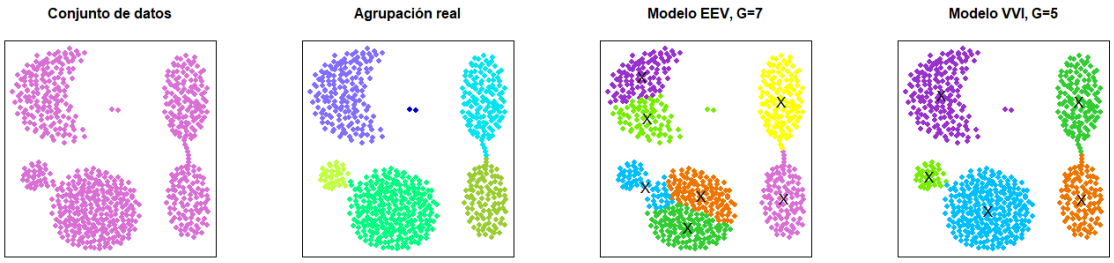



Figura 21: Primera comparación criterios BIC e ICL

Por otro lado, observamos el desempeño de los dos criterios con otro conjunto de datos en el que de nuevo las componentes no parecen gaussianas, esta vez sin incluir la agrupación real de los datos.

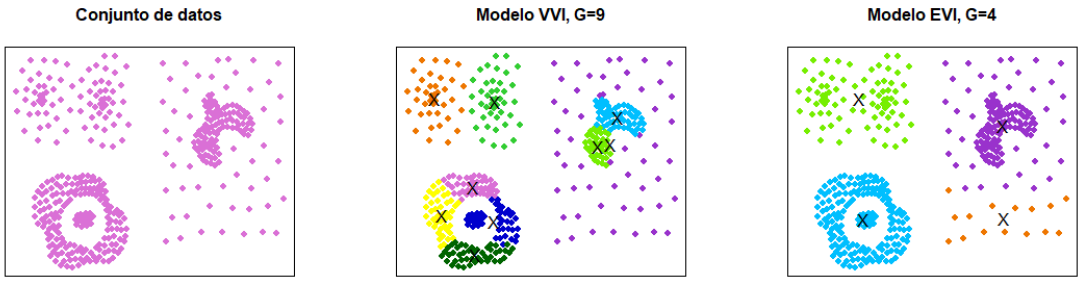



Figura 22: Segunda comparación criterios BIC e ICL

En este caso, el criterio ICL determina que el modelo óptimo es EVI con 4 grupos de igual volumen, forma variable y alineados con los ejes, cuya actuación no es la más idónea para representar todas las particularidades del conjunto de datos ya que no es capaz de distinguir las dos componentes de la esquina izquierda inferior ni separar la nube de puntos dispersos del conjunto con forma de herradura de la derecha. El BIC en este caso es más prudente y opta por explicar los conjuntos con formas más disparatadas por medio de 9 pequeñas componentes gaussianas de volumen y forma

variables alineadas con los ejes, logrando una partición razonable si realizamos una reconstrucción de los grupos salvo por la pequeña agrupación abajo a la izquierda, que incluye como parte de otro cluster.

### 4.2.1. Componentes no gaussianas y fusión de clusters

La presencia de agrupaciones cuya distribución no es gaussiana provoca que criterios como el BIC escojan usualmente un número mayor de componentes gaussianas para reconstruir un cluster de estas características. Esto produce una buena estimación de la densidad de mixtura pero una sobreestimación del número de clusters. Por otro lado, el ICL tiende a seleccionar tantas componentes en la mezcla como clusters hay en el conjunto de datos pero, al tratar de representar un grupo cuya distribución no es gaussiana por medio de una componente que sí lo es, el ajuste suele ser mucho más pobre y los datos no se ven bien representados.

Una solución a este dilema consiste en mantener la densidad de mixtura estimada por el mejor modelo según el BIC para ajustar correctamente los datos pero fusionar algunas de las componentes más cercanas para terminar teniendo una partición con el número de clusters correcto. De este modo, es necesario desarrollar técnicas para decidir cuándo fusionar o juntar dos o más componentes y cuáles. De manera general, un método para fusionar las componentes sigue los siguientes pasos:

1. Comenzar considerando tantos clusters como componentes hemos utilizado en el modelo de mixturas.
2. Encontrar el par de clusters más prometedor para ser unidos.
3. Aplicar un criterio de parada para determinar si fusionar estos dos clusters o si considerar la partición actual como la final.
4. Si decidimos unirlos, volver al paso dos.

Es decir, se necesita un criterio para seleccionar los dos clusters más prometedores para ser fusionados y otro para establecer la parada. Expondremos en este apartado el método de la entropía de clasificación como ejemplo de criterio para unir grupos.

Recordemos que la incertidumbre en la clasificación de un punto se definía como

$$u(x_i) = 1 - \max_{j \in 1,..,G} \tau_j(x_i; \hat{\Psi})$$

Denotemos por $\hat{z}_{ij}^G$ la probabilidad a posteriori de que la observación $i$-ésima provenga de la componente $j-$ésima utilizando un modelo con $G$ componentes. Esta clasificación tiene entropía

$$Ent(G) = -\sum_{i=1}^{n} \sum_{j=1}^{G} \hat{z}_{ij}^G \log(\hat{z}_{ij}^G) \geq 0$$

La entropía es mínima cuando no existe incertidumbre respecto a la clasificación de la observación $x_i$, es decir, cuando cada $\hat{z}_{ij}^G$ bien es 1 o bien es 0. Por el contrario, la entropía es máxima cuando la incertidumbre es máxima, es decir, cuando $\hat{z}_{ij}^G = \frac{1}{G}$ (cuando existe la misma probabilidad a posteriori

de pertencer a cualquiera de los $G$ grupos). Resulta que $Entr(G) \geq Entr(G-1)$, ya que al añadir grupos se tienen más posibilidades a la hora de asignar un individuo a un cluster y por lo tanto mayor incertidumbre para cada observación.

La idea fundamental de este procedimiento es escoger el par de grupos que minimizan el incremento de entropía al ser fusionados. Sean $C_k$ y $C_h$ dos clusters de la partición en $G$ grupos $\{C_1, .., C_G\}$ y $h, k$ sus respectivos índices. Si unimos $C_k$ y $C_h$, dado que son disjuntos, el nuevo cluster $C_h \cup C_k$ con índice $\{h, k\}$ tiene la siguiente probabilidad a posteriori:

$$\hat{z}^G_{i,\{h,k\}} = \hat{z}^G_{ik} + \hat{z}^G_{ih}$$

Mientras que el resto de valores $\hat{z}^G_{ij}$ se mantienen iguales, con $j \neq k, h$. La entropía resultante será

$$-\Big(\sum_{i=1}^{n}(\hat{z}^G_{ih} + \hat{z}^G_{ik})\log(\hat{z}^G_{ih} + \hat{z}^G_{ik}) + \sum_{j\neq h,k}\hat{z}^G_{ij}\log(\hat{z}^G_{ij})\Big)$$

Si restamos ambas dos entropías, buscamos dos clusters $C_h$ y $C_k$ que maximicen la diferencia

$$-\sum_{i=1}^{n}(\hat{z}^G_{ih} + \hat{z}^G_{ik})\log(\hat{z}^G_{ih} + \hat{z}^G_{ik}) + \sum_{i=1}^{n}\hat{z}^G_{ik}\log(\hat{z}^G_{ik}) + \hat{z}^G_{ih}\log(\hat{z}^G_{ih})$$

Una vez seleccionado el par de clusters, se renombra su unión y se acualizan las probabilidades a posteriori $\hat{z}^{G-1}_{ij}$ (Ahora tenemos $G-1$ grupos). El objetivo es continuar uniendo grupos hasta que se provoca un incremento grande en la entropía. Usualmente, se puede utilizar el criterio del codo en un gráfico que compare el número de componentes con la entropía para decidir en qué momento parar.

Expondremos el funcionamiento del método de Entropía de Clasificación con el siguiente conjunto de datos. En la siguiente figura, observamos su representación bidimensional y su partición en clusters a medida que variamos el número de grupos.

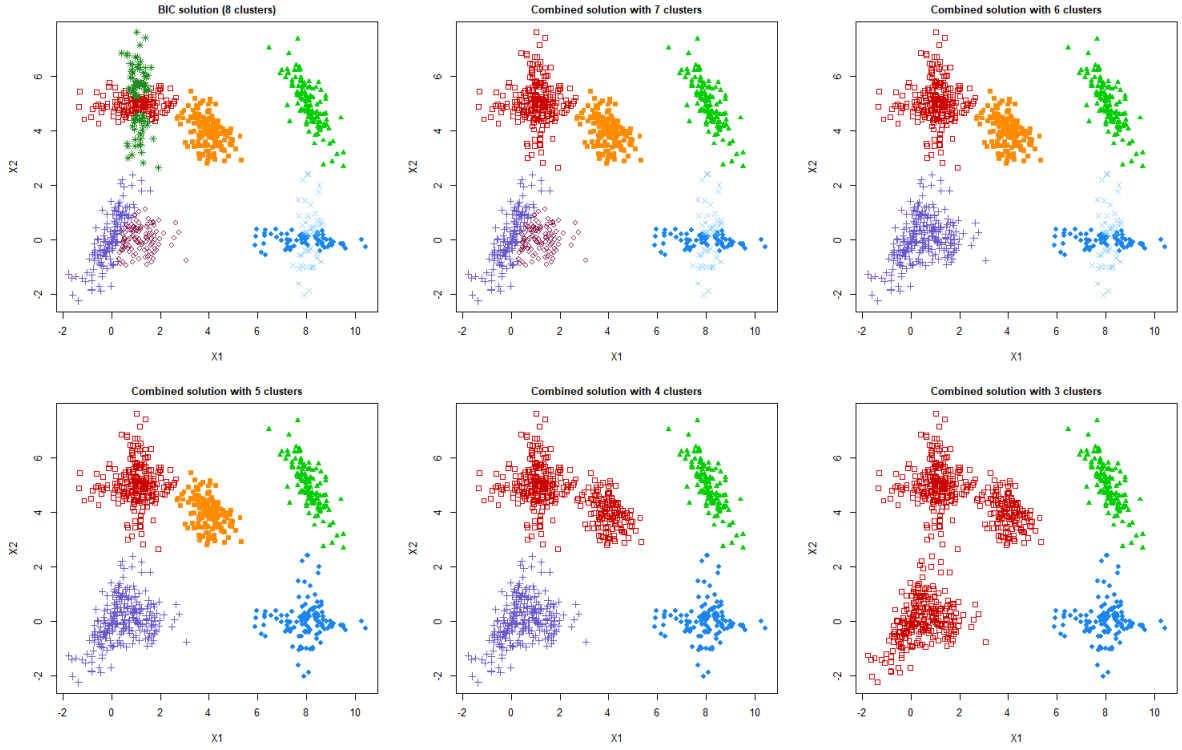



Figura 23: Agrupaciones variando $G$

Observamos cómo este conjunto de datos tiene 5 grupos más o menos separados, alguno de apariencia elipsoidal mientras que otros son claramente no gaussianos. Si construimos un gráfico que muestre la entropía de clasificación según el número de componentes consideradas, observamos como su variación es pequeña hasta que consideramos 5 clusters. A partir de ese momento, se dispara, por lo que el método de la Entropía de Clasificación sugeriría que existen 5 grupos en este conjunto de datos.

Como comentábamos, el procedimiento parte del modelo escogido por el BIC, que a menudo contiene una partición más exhaustiva en componentes elípticas de los clusters no gaussianos con el fin de representarlos mejor. Es decir, en este caso el criterio BIC determinó que el número adecuado de grupos es 8. Por otro lado, es interesante estudiar cuántas componentes preveíamos esperar en el modelo con mejor ICL, por lo que producimos dos gráficos que muestran el BIC e ICL respectivamente para los 14 modelos según la clasificación VSO variando el número de componentes.

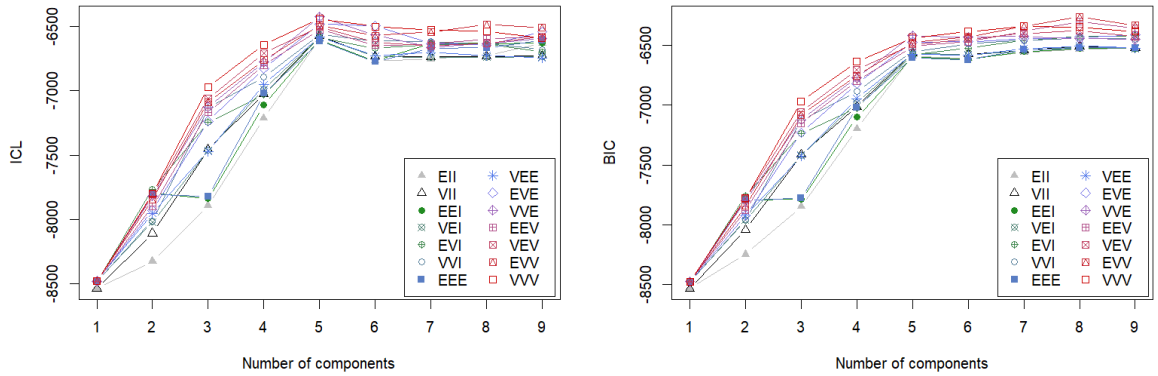



Figura 24: Selección de modelos utilizando BIC e ICL

Observamos cómo el criterio ICL consigue recuperar el número de clusters que visualmente se aprecian en el conjunto de datos, es decir, cinco.

## 4.3. El algoritmo EM

Hace unas décadas, a pesar de contar con ordenadores que permitían realizar cálculos con mucha mayor rapidez, existía cierta reticencia a utilizar los modelos de mezclas en el caso multidimensional, posiblemente por la falta de información sobre algunas cuestiones abiertas que aún no habían sido estudiadas en profundidad, como la presencia de múltiples máximos en la función de verosimilitud y la no acotación de esta en el caso de utilizar componentes gaussianas con distintas matrices de covarianza.

En el momento en el que estas dificultades se comprendieron satisfactoriamente y se consensuaron algunas soluciones, se incrementó el uso de estos modelos en la práctica. En los sesenta, el ajuste de modelos de mezclas por máxima verosimilitud había sido estudiado en numerosos artículos, pero fue la publicación de Dempster, Laird y Rubin en 1977 sobre el algoritmo EM lo que puso el foco de interés en este método para modelar datos heterogéneos.

El algoritmo EM es una herramienta de uso extendido para el cálculo del estimador máximo verosimil especialmente provechosa en problemas con datos incompletos donde métodos como el de Newton-Raphson pueden resultar más complicados. Este procedimiento iterativo puede utilizarse no solo en situaciones en las que hay datos faltantes, distribuciones truncadas o agrupaciones en las observaciones, sino en circunstancias donde el carácter incompleto de los datos no es tan evidente, como puede ser el modelo de mixtura, los modelos log-lineales, convoluciones... Abordaremos su estudio en el problema de la estimación de parámetros en el modelo de mezclas y observaremos cómo se ve simplificado por el EM, además de proporcionar en algunos casos expresiones cerradas de los estimadores.

### 4.3.1. Formulación del algoritmo EM

Consideramos el caso general del modelo de mixturas en el que la familia de densidades de probabilidad es arbitraria. Más tarde, particularizaremos estos resultados para el caso del modelo de mezclas gaussiano. Es decir, nos encontramos en el caso en el que buscamos estimar el vector de parámetros $\Psi = (\pi_1, .., \pi_G, \theta_1, .., \theta_G)$ dada una muestra aleatoria $A = \{x_1, .., x_n\}$ a la que hemos ajustado un modelo de mixturas

$$f(x_i; \Psi) = \sum_{j=1}^{G} \pi_j f(x_i; \theta_j),\ f \in \mathscr{F} = \{f(\cdot; \theta) : \theta \in \Theta\},\ 1 \leq i \leq n$$

Su función de log-verosimilitud correspondiente se escribe como

$$L(\Psi, A) = \sum_{i=1}^{n} \log\Big(\sum_{j=1}^{G} \pi_j f(x_i; \theta_j)\Big)$$

El estimador máximo versolímil de $\Psi$ se calcula resolviendo la ecuación de verosimilitud

$$\frac{\partial L(\Psi, A)}{\partial \Psi} = 0$$

A menudo en la práctica, la función de log-verosimilitud no puede ser maximizada de forma analítica. En muchos casos, es posible calcular de forma iterativa el MLE de $\Psi$ utilizando el procedimiento de Newton-Raphson o alguno similar cuando el número de parámetros en el modelo no es demasiado grande.

Consideremos las etiquetas $z_{ij}$ no observadas que toman el valor 1 si la observación $x_i$ proviene de la componente $j$-ésima y 0 en otro caso ($1 \leq i \leq n, 1 \leq j \leq G$), y denotamos $x = (x_1, .., x_n)$. Los vectores $z_i = (z_{i1}, .., z_{iG})$ se consideraban una realización de $Z_1, .., Z_n$ vectores aleatorios independientes igualmente distribuidos definidos en un mismo espacio probabilístico que siguen una distribución multinomial, donde se elige una de $G$ categorías con probabilidades $\pi_1, .., \pi_G$. Asumimos que la densidad de una observación $x_i$ dado $z_i$ es

$$\prod_{j=1}^{G} f(x_i; \theta_j)^{z_{ij}}$$

Es decir, si $x_i$ proviene de la componente $j$, todos los factores distintos de $j$ serán 1 y solo quedará $f(x_i; \theta_j)$. La log-verosimilitud de los datos completos es por lo tanto

$$L_c(\Psi, A) = \sum_{j=1}^{G} \sum_{i=1}^{n} z_{ij}(\log \pi_j + \log f(x_i; \theta_j))$$

El algoritmo EM cuenta con dos pasos que describimos a continuación:

- **E Step (Expectation Step):**

  Sea $\Psi^{(0)}$ un valor inicial para $\Psi$. En este paso, se procede al cálculo de

  $$Q(\Psi; \Psi^{(0)}) = \mathbb{E}_{\Psi^{(0)}}\big(L_c(A, \Psi)|x\big)$$

Es decir, calculamos el valor esperado de la verosimilitud completa condicionado a x, utilizando el ajuste $\Psi^{(0)}$ para $\Psi$. Como hemos visto, la log-verosimilitud de los datos completos es una función lineal en $z_{ij}$, por lo que basta con calcular la esperanza condicional de $Z_{ij}$ dados los datos observados $X$. Por la definición de esperanza condicional, podemos reescribir

$$\mathbb{E}_{\Psi^{(0)}}\big(Z_{ij}|X=x\big) = 1 \cdot P(Z_{ij}=1|X=x) + 0 \cdot P(Z_{ij}=0|X=x) = \frac{P(Z_{ij}=1, X=x)}{P(X=x)}$$

Por el Teorema de Bayes, la última expresión puede reescribirse como

$$\frac{P(Z_{ij}=1, X=x)}{P(X=x)} = \frac{P(X=x|Z_{ij}=1)P(Z_{ij}=1)}{P(X=x)} = \frac{\pi_j f(x_i;\theta_j)}{f(x_i;\Psi^{(0)})} = \tau_j(x_i;\Psi^{(0)})$$

Para $1 \leq i \leq n, 1 \leq j \leq G$. La cantidad $\tau_j(x_i;\Psi^{(0)})$ es la probabilidad a posteriori de que la observación $i-$ésima pertenezca a la $j-$ésima componente. Sustituyendo esto en la esperanza condicional de la log-verosimilitud, se tiene

$$Q(\Psi;\Psi^{(0)}) = \sum_{j=1}^{G}\sum_{i=1}^{n} \tau_j(x_i;\Psi^{(0)})(\log \pi_j + \log f(x_i;\theta_j))$$

- **M Step (Maximization Step):**

El primer paso consiste en elegir $\Psi^{(1)}$ el valor que maximiza $Q(\Psi;\Psi^{(0)})$ respecto de $\Psi$. Si las $z_{ij}$ fueran datos observados, el MLE de los pesos se calcularía como

$$\hat{\pi}_j = \sum_{i=1}^{n} \frac{z_{ij}}{n}$$

Por lo tanto, en este primer paso $\pi_j^{(1)} = \sum_{i=1}^{n} \frac{\tau_j(x_i;\Psi^{(0)})}{n}$. Es decir, al formar la estimación del peso de la componente $j-$ésima en la primera iteración, $\pi_j^{(1)}$, se tiene en cuenta la contribución de cada observación $x_i$ por medio de su probabilidad posterior (evaluada en ese momento) de pertenecer a la j-ésima componente del modelo de mezclas. Esta estimación por medio del EM tiene por lo tanto una interpretación bastante intuitiva.

A la hora de estimar los parámetros $\theta_j^{(1)}$ correspondientes, debemos buscar una raíz de la ecuación

$$\sum_{j=1}^{G}\sum_{i=1}^{n} \tau_j(x_i;\Psi^{(0)}) \frac{\partial(\log f(x_i;\theta_j))}{\partial\theta} = 0$$

Donde $\theta$ denota $(\theta_1, .., \theta_G)$. La razón por la que el algoritmo EM resulta interesante en este contexto se debe al hecho de que a menudo la solución de esta ecuación tiene una forma cerrada, como veremos más adelante en el caso del modelo de mezclas gaussiano.

Hemos descrito los dos pasos del algoritmo EM en la primera iteración. De manera general, se tendría:

1. Inicializamos el vector de parámetros $\Psi^{(0)}$.
2. En la iteración $k \geq 1$, se efectúan
   - E-Step:
     Se calcula
     $$Q(\Psi; \Psi^{(k-1)}) = \mathbb{E}_{\Psi^{(k-1)}}\left(L_c(A, \Psi)|x\right) = \sum_{j=1}^{G}\sum_{i=1}^{n} \tau_j(x_i; \Psi^{(k-1)})(\log \pi_j + \log f(x_i; \theta_j^{(k-1)}))$$
   - M-Step:
     Se busca el vector de parámetros $\Psi^{(k)}$ tal que
     $$\Psi^{(k)} = \arg\max_{\Psi} Q(\Psi; \Psi^{(k-1)})$$
   - Se calcula $L(\Psi^{(k)}) - L(\Psi^{(k-1)})$
3. Cuando la diferencia calculada anteriormente es suficientemente pequeña o no ha habido un cambio significativo en los parámetros, se detiene el proceso. En caso contrario, se actualiza el valor del vector de parámetros $\Psi^{(k-1)} := \Psi^{(k)}$ y se repite el segundo paso.

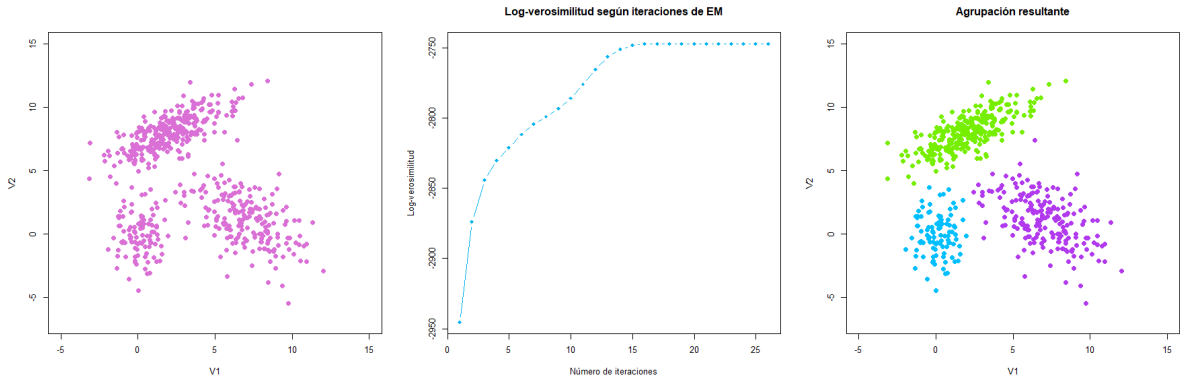



Figura 25: Ejemplo de actuación del algoritmo EM

En esta imagen observamos la actuación del algoritmo EM a la hora de ajustar tres componentes de mixtura al conjunto de datos de la izquierda. En el gráfico central, observamos los valores que toma la función de log-verosimilitud en cada una de las iteraciones. Observamos cómo a partir de la iteración 15, los cambios en la log-verosimilitud son muy pequeños. Una vez superada la tolerancia, el algoritmo se detiene y ofrece la agrupación final en tres componentes que vemos a la derecha.

Por otro lado, en aquellas situaciones en las que existe un nivel alto de superposición entre los grupos, la convergencia del algoritmo es mucho más lenta y necesita un número mayor de iteraciones. En este caso, vemos como la actuación no resulta del todo satisfactoria a la hora de hallar las tres componentes del conjunto de datos y observamos que fueron necesarias más de 300 iteraciones para la convergencia del algoritmo.

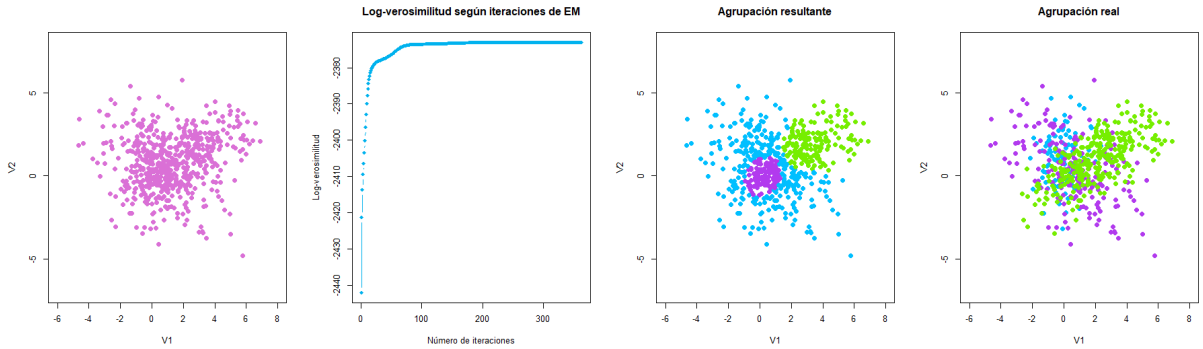



Figura 26: Ejemplo de actuación del algoritmo EM

### 4.3.2. El algoritmo EM en el modelo de mezclas gaussiano

Si consideramos el caso particular en el que las componentes de la distribución son gaussianas, existe una expresión cerrada para el cálculo de los parámetros por medio del algoritmo EM. Recordemos que en este caso trabajamos con $\Psi = (\pi_1, .., \pi_G, \mu_1, .., \mu_G, \Sigma_1, .., \Sigma_G)$ el vector de parámetros y $\phi(\cdot; \mu_j, \Sigma_j)$ la $j-$ésima componente de la mixtura. Consideramos en primer lugar el caso heterocedástico (no hay restricciones sobre las matrices de covarianza): Si nos encontramos en la iteración $k-$ésima en la etapa de maximización, los vectores de medias y matrices de covarianzas se actualizan del siguiente modo:

$$\mu_j^{(k)} = \sum_{i=1}^{n} \frac{z_{ij}^{(k-1)} x_i}{\sum_{i=1}^{n} z_{ij}^{(k)}}, \ 1 \leq j \leq G$$

$$\Sigma_j^{(k)} = \sum_{i=1}^{n} \frac{z_{ij}^{(k-1)} (x_i - \mu_j^{(k)})(x_i - \mu_j^{(k)})^T}{\sum_{i=1}^{n} z_{ij}^{(k-1)}}, \ 1 \leq j \leq G$$

En la práctica, cuando la naturaleza de los datos lo permite, se trabaja bajo la hipótesis de que las componentes tienen matriz de covarianza común $\Sigma$ desconocida. En este caso de homocedasticidad, se sustituye la actualización correspondiente de las $\Sigma_j^{(k)}$ por

$$\Sigma^{(k)} = \sum_{j=1}^{G} \sum_{i=1}^{n} \frac{z_{ij}^{(k-1)} (x_i - \mu_j^{(k)})(x_i - \mu_j^{(k)})^T}{n}$$

En este caso, el MLE existe como el maximizador global de la verosimilitud y además es fuertemente consistente.

En el caso heteroscedástico, la maximización de la verosimilitud sin ninguna restricción podría llevarnos a una solución degenerada. Una vez seleccionamos un modelo, podemos plantear condiciones sobre el ratio entre los autovalores de las matrices de covarianza en cada iteración

$$\frac{M^{(k)}}{m^{(k)}} \geq c,\ c \geq 1$$

$M^{(k)}$ *mayor autovalor de las matrices de covarianza en la iteración k-ésima*

$m^{(k)}$ *menor autovalor de las matrices de covarianza en la iteración k-ésima*

De la misma forma, cuando buscamos elegir un modelo por medio del BIC, podríamos descartar aquellos modelos cuyo ratio entre autovalores de las matrices finales de covarianza están por debajo del umbral $c$. En este caso, la cuestión se reduce a cómo elegir este parámetro para asegurar que la restricción que hemos hecho del espacio paramétrico contienen el verdadero valor del vector de parámetros $\Psi$, o para garantizar que no descartamos modelos que podrían ajustar mejor los datos pero que no superan esta restricción. La elección de $c$ dependerá de las particularidades del conjunto de datos en cada caso.

### 4.3.3. Convergencia del Algoritmo EM

Nuestro objetivo es recoger brévemente los resultados más importantes disponibles sobre la teoría básica relativa a la convergencia del algoritmo EM. Para ello, utilizaremos como guía el libro de McLachlan [10] donde se encuentran las pruebas de los teoremas que enunciaremos.

En primer lugar, demostraremos que la función de log-verosimilitud $L(\Psi)$ que buscamos maximizar es no decreciente al actualizar los iterantes.

**Proposición 4.1.** ***(Motononía del EM)*** *Consideremos $x = (x_1, .., x_n)$ una muestra aleatoria, $\Psi = (\pi_1, .., \pi_G, \theta_1, .., \theta_G)$ el vector de parámetros a estimar y $\mathscr{L}(\Psi, A)$ la función de verosimilitud*

$$\mathscr{L}(\Psi, x) = \prod_{i=1}^{n}\Big(\sum_{j=1}^{G} \pi_j f(x_i; \theta_j)\Big) = \prod_{i=1}^{n} f(x_i, \Psi)$$

*Consideramos $z = (z_1, .., z_n)$ los datos no observados donde $z_i = (z_{i1}, .., z_{iG})$ es el vector de etiquetas de la observación $i-$ésima. Sea $\{\Psi^{(k)}\}_{k\geq 0}$ una sucesión de iterantes obtenidos por medio del algoritmo EM. Entonces, para todo $k \geq 0$, se tiene que*

$$L(\Psi^{(k+1)}) \geq L(\Psi^{(k)})$$

**Demostración**

Recordemos que la función de verosimilitud completa tenía la expresión

$$\mathscr{L}_c(\Psi, x, z) = \prod_{i=1}^{n}\prod_{i=1}^{n}\big(\pi_j f(x_i; \theta_j)\big)^{z_{ij}} = \prod_{i=1}^{n} f_c(\Psi, x_i, z_i)$$

Sea $k(z|x,\Psi)$ la densidad condicional de $Z$ dado $X=x$, es decir

$$k(z|x,\Psi)=\frac{\prod_{i=1}^{n} f_c(\Psi,x_i,z_i)}{\prod_{i=1}^{n} f(x_i,\Psi)}$$

La log-verosimilitud se escribe entonces

$$L(\Psi,x)=\log(\mathscr{L}(\Psi,x))=\log\big(\prod_{i=1}^{n} f(x_i,\Psi)\big)=\log\big(\prod_{i=1}^{n} f_c(\Psi,x_i,z_i)\big)-\log k(z|x,\Psi)$$

Si tomamos esperanzas condicionales respecto de $Z$ dado $X=x$ a ambos lados y utilizando el ajuste $\Psi^{(k)}$ para $\Psi$, se tiene que

$$L(\Psi,x)=\mathbb{E}_{\Psi^{(k)}}\Big(\log\big(\prod_{i=1}^{n} f(x_i,\Psi)\big)\Big)-\mathbb{E}_{\Psi^{(k)}}\Big(\log k(z|x,\Psi)\Big)=Q(\Psi;\Psi^{(k)})-H(\Psi;\Psi^{(k)})$$

De esta manera, se tiene que

$$L(\Psi^{(k+1)},x)-L(\Psi^{(k)},x)=\Big(Q(\Psi^{(k+1)};\Psi^{(k)})-Q(\Psi^{(k)};\Psi^{(k)})\Big)-\Big(H(\Psi^{(k+1)};\Psi^{(k)})-H(\Psi^{(k)};\Psi^{(k)})\Big) \tag{10}$$

El iterante $\Psi^{(k+1)}$ se elige de tal manera que maximice $Q(\Psi;\Psi^{(k)})$, por lo que

$$Q(\Psi^{(k+1)};\Psi^{(k)})\geq Q(\Psi^{(k)};\Psi^{(k)})$$

Y con esto, la resta del primer paréntesis es no negativa. Si logramos ver que

$$H(\Psi^{(k+1)};\Psi^{(k)})-H(\Psi^{(k)};\Psi^{(k)})\leq 0$$

se tendría que la expresión (10) es no negativa y habríamos acabado. Para un $\Psi$ cualquiera

$$H(\Psi;\Psi^{(k)})-H(\Psi^{(k)};\Psi^{(k)})=\mathbb{E}_{\Psi^{(k)}}\Big(\log k(z|x,\Psi)-\log k(z|x,\Psi^{(k)})|x\Big)=$$

$$=\mathbb{E}_{\Psi^{(k)}}\Big(\log\Big(\frac{k(z|x,\Psi)}{k(z|x,\Psi^{(k)})}\Big)\Big|x\Big)\leq\log\Big(\mathbb{E}_{\Psi^{(k)}}\Big(\frac{k(z|x,\Psi)}{k(z|x,\Psi^{(k)})}\Big|x\Big)\Big)$$

Donde la última desigualdad es consecuencia de la desigualdad de Jensen ya que la función logaritmo es cóncava. Esta última expresión, recordando la definición de esperanza condicional respecto de $Z$ dado $x$ y utilizando el ajuste $\Psi^{(k)}$ para $\Psi$, puede reescribirse como

$$\log\Big(\mathbb{E}_{\Psi^{(k)}}\Big(\frac{k(z|x,\Psi)}{k(z|x,\Psi^{(k)})}\Big|x\Big)\Big)=\int\frac{k(z|x,\Psi)}{k(z|x,\Psi^{(k)})}k(z|x,\Psi^{(k)})dx=\log(1)=0$$

Por lo tanto, hemos probado que la verosimilitud no decrece al iterar el algoritmo EM. □

Una consecuencia de esto es que si $\hat{\Psi}$ es el MLE que maximiza $\mathscr{L}(\Psi,x)$, necesariamente

$$Q(\hat{\Psi};\hat{\Psi})\geq Q(\Psi;\hat{\Psi})$$

Para cualquier $\Psi$ del espacio paramétrico. Si no fuese así, llegaríamos a una contradicción ya que $\hat{\Psi}$ no maximizaría $\mathscr{L}(\Psi,x)$.

Bajo condiciones no demasiado restrictivas, podemos asegurar la convergencia de los iterantes hacia puntos estacionarios, es decir, donde el gradiente de la función de verosimilitud se anula.

**Teorema 4.2.** *Supongamos que la función $Q(\Psi;\Phi)$ es continua en ambas variables. Entonces todos los puntos límites de cualquier sucesión de iterantes $\{\Psi^{(k)}\}_{k\geq 0}$ del algoritmo EM son puntos estacionarios de $L(\Psi,x)$ y $L(\Psi^{(k)})$ converge monótonamente hacia un valor $L^* = L(\Psi^*,x)$ para un punto estacionario $\Psi^*$*

Esto provoca que en numerosas situaciones prácticas podamos encontrar un máximo local $L^* = L(\Psi^*,x)$ por medio del algoritmo EM. En general, si $L(\Psi,x)$ tiene varios puntos estacionarios, la convergencia de los iterantes hacia cualquier punto que anula el gradiente de la función de verosimilitud (máximos locales o globales, puntos de silla) depende de la elección del iterante inicial $\Psi^{(0)}$.

En el caso de considerar el problema de mezclas gaussiano, este teorema nos garantiza que imponiendo una condición sobre el ratio de los autovalores de las matrices de covarianza como la ya planteada, podemos asegurar que el algoritmo convergerá hacia un punto estacionario.

Por último, si la función de verosimilitud es unimodal con un solo punto estacionario en el interior del espacio paramétrico y se cumplen ciertas condiciones de regularidad, tenemos el siguiente resultado.

**Corolario 4.3.** *Supongamos que $L(\Psi,x)$ es unimodal con $\Psi^*$ es su único punto estacionario y $\frac{\partial Q(\Psi;\Phi)}{\partial \Psi}$ es continua en ambas variables. Entonces cualquier sucesión $\{\Psi^{(k)}\}_{k\geq 0}$ de iterantes del algoritmo EM converge al único maximizador $\Psi^*$ de $L(\Psi;x)$, es decir, su MLE.*

### 4.3.4. Inicialización del Algoritmo EM

Dado que usualmente existen máximos locales en la función de verosimilitud, a menudo el vector de parámetros obtenido por medio del algoritmo EM depende del valor inicial $\Psi^{(0)}$ elegido. Es por ello que existen diferentes técnicas de inicialización más o menos sofisticadas con el fin de seleccionar un vector de parámetros inciales que permitan un buen punto de partida a la hora de comenzar el algoritmo.

En la figura siguiente, mostramos la actuación del algoritmo EM cuando la inicialización se realiza por medio de una partición aleatoria en dos grupos equilibrados. En este caso, observamos que el algoritmo tiene una convergencia rápida ya que a partir de la octava iteración se producen cambios muy pequeños en la log-verosimilitud. En la figura de la derecha podemos observar la partición resultante efectuada por el EM, donde las líneas muestran los valores que han tomado los dos vectores de medias en cada iteración. Dado que la partición inicial en grupos es aleatoria, los iterantes iniciales están muy próximos entre sí, prácticamente en medio de la nube de puntos. Los triángulos azul y verde son las estimaciones finales de los centros de las distribuciones.

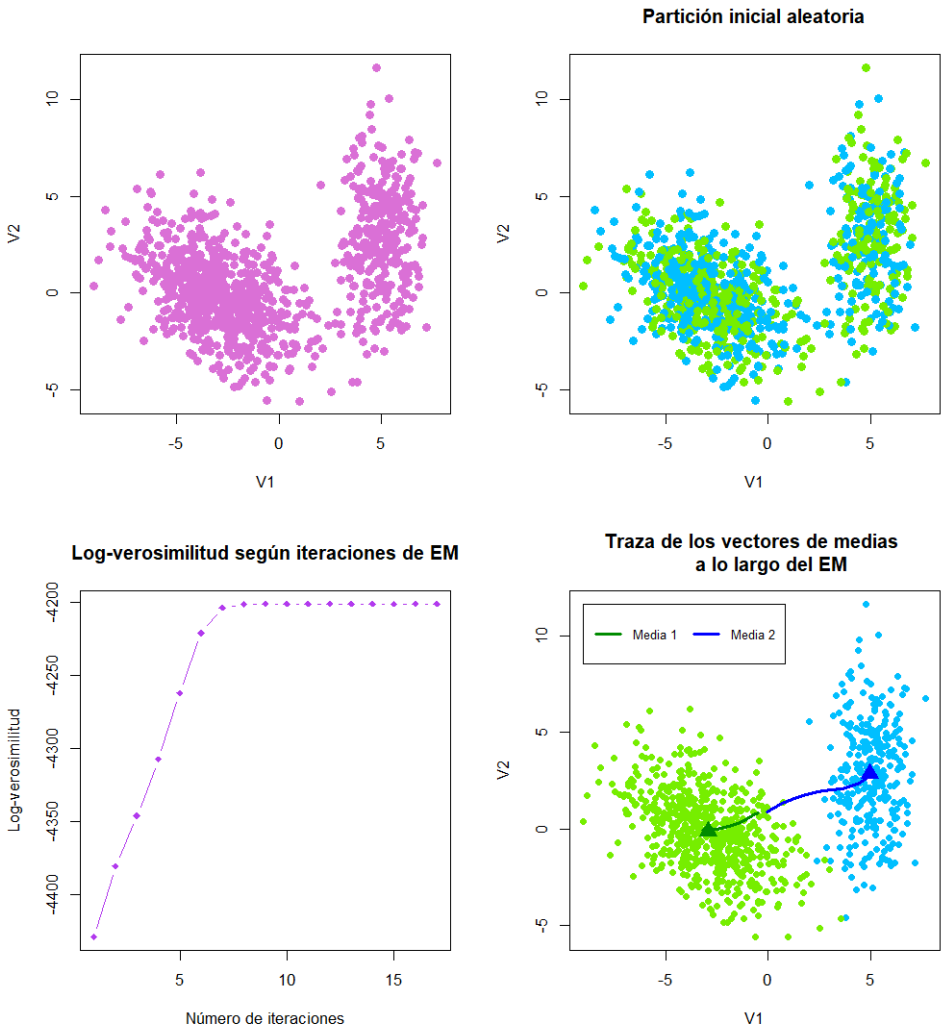



Figura 27: Partición aleatoria para inicializar el EM

Otros posibles métodos para inicializar el procedimiento EM son el clústering jerárquico basado en modelos y el algoritmo smallEM. El primero es un procedimiento aglomerativo que en cada iteración fusiona el par de grupos que supone un incremento mayor en la función de verosimilitud. El procedimiento comienza tratando cada observación como un cluster y termina cuando se tiene una partición del conjunto de datos en $G$ grupos. El clustering jerárquico basado en modelos consigue una partición en clusters válida para los 14 modelos según VSO (en el caso multivariante) y para cualquier número de grupos. Este hecho lo convierte en un procedimiento atractivo ya que con frecuencia el Análisis Cluster basado en modelos necesita ajustar un número muy grande de modelos

diferentes para elegir el más adecuado, por lo que conseguir una partición inicial para todos ellos con una sola ejecución del procedimiento a menudo es computacionalmente eficiente.

Por otro lado, el algoritmo smallEM se implementa de forma separada para cada modelo según la clasificación VSO y el número de componentes. Dado $A = \{x_1, .., x_n\}$, se deben de seguir $M$ veces los siguientes pasos:

- Se genera un valor inicial para vector de parámetros $\Psi$ aleatoriamente.
- Ese vector se considera $\Psi^{(0)}$, el primer iterante. Se ejecuta el algoritmo EM en una versión más corta, deteniendo el procedimiento en la primera iteración $s$ en la que

$$\frac{L(\hat{\Psi}^{(s)}, A) - L(\hat{\Psi}^{(s-1)}, A)}{L(\hat{\Psi}^{(s)}, A) - L(\hat{\Psi}^{(1)}, A)} < \epsilon$$

  Para un $\epsilon$ prefijado.

- Se guarda el valor de la función de log-verosimilitud en el punto de parada del smallEM.

Entre los $M$ puntos de parada considerados, escogemos aquel con mayor valor de la log-verosimilitud. Ese será nuestro iterante inicial para el algoritmo EM.

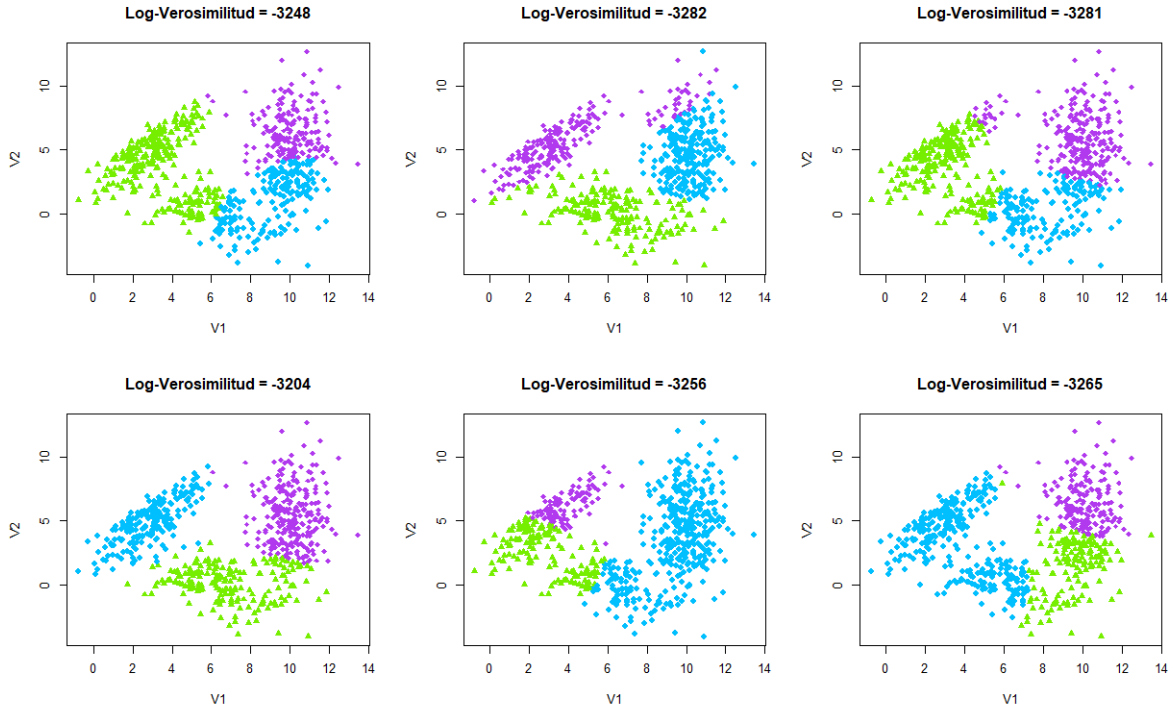



Figura 28: Inicializaciones propuestas por el EM

En la figura, se recogen seis particiones iniciales propuestas por el algoritmo smallEM, de las cuales la que está situada abajo a la izquierda es aquella con mayor log-verosimilitud. Entre estas opciones, esta será la elegida como punto de partida para desarrollar el algoritmo EM. Podemos observar que con este procedimiento de inicialización, que solo ha necesitado de 5 iteraciones, hemos conseguido un buen esbozo de la división en tres componentes del conjunto de datos, por lo que la actuación del EM será en muchas ocasiones más rápida que si consideráramos una partición aleatoria inicial de los individuos.

La agrupación tras aplicar el algoritmo EM es la siguiente, donde la función de log-verosimilitud toma finalmente el valor de -2943.

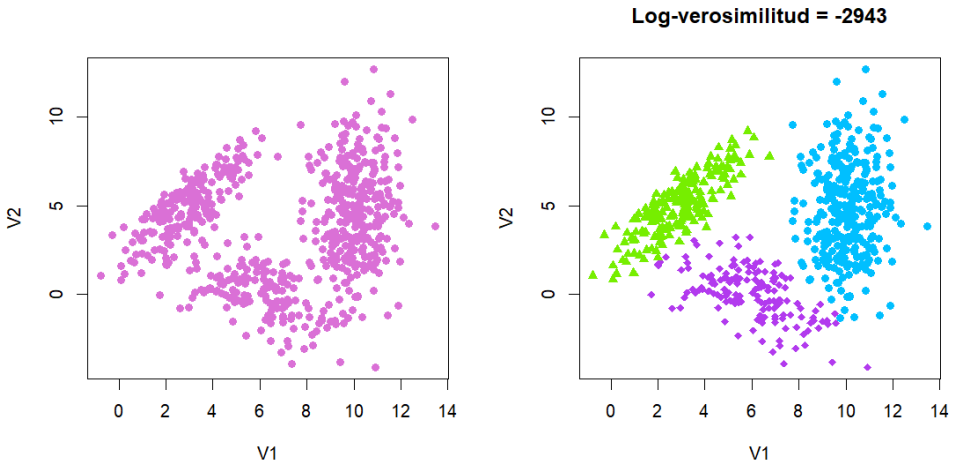



Figura 29: Inicializaciones propuestas por el EM

# 5. Apéndice

Esta sección tiene la finalidad de reunir algunas definiciones o demostraciones de carácter secundario necesarias para el desarrollo y mejor entendimiento del resto del documento.

## 5.1. Clases de Glivenko Cantelli y Vapnik-Chervonenkis

El problema característico de la Estadística consiste en considerar una muestra que ha sido obtenida por medio de un experimento con cierto grado de incertidumbre y estudiar las propiedades de la ley $P$ que gobierna los datos, que es desconocida, extrayendo información de dicha muestra. La idea clave de la Estadística Matemática se basa en tratar la muestra no como un conjunto de datos abstractos sino asociarle una distribución muestral en la que consideramos que los resultados son equiprobables. Para ello, recurrimos a todas las herramientas típicas de la Teoría de la Probabilidad en su versión muestral que tendrán un carácter aleatorio y dependerán de la muestra, que podrá ser más o menos complicada.

Consideremos $X_1, .., X_n, ...$ v.a.i.i.d. definidas en un mismo espacio probabilístico con función de distribución común $F(x)$. Si $I_A$ denota la función indicadora de $A$, se define la función de distribución empírica como

$$F_n(x) = \frac{1}{n} \sum_{i=1}^{n} I_{[X_i, \infty)}(x)$$

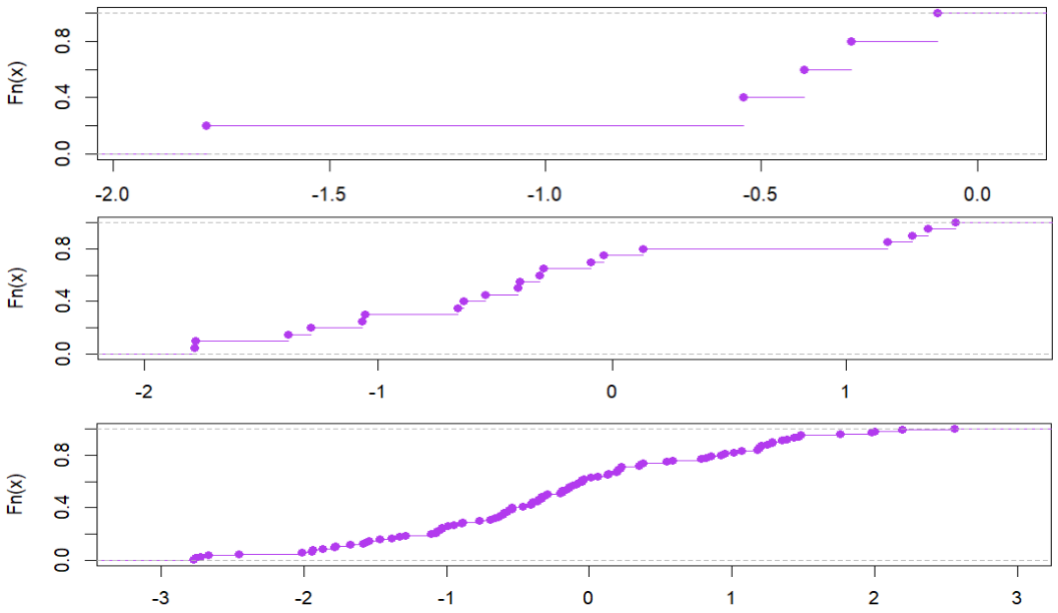



Figura 30: Función de distribución empírica a medida que aumenta la muestra

Dada una muestra de tamaño $n$, la probabilidad muestral $P_n$ es la función de masa de probabilidad discreta que otorga masa $\frac{1}{n}$ a cada $X_i$, $i = 1, .., n$.

Necesitamos tener garantías de que podemos aproximar la medida de probabilidad teórica por la muestral de manera uniforme y que el $n$ a partir del cual se da una aproximación de cierto nivel no depende del $x$ escogido. Resultados como el Teorema de Glivenko-Cantelli (ver Teorema en 20.6 [1]) aseguran la convergencia en el sentido uniforme de la función de distribución empírica a la función de distribución teórica, teniendo como consecuencia

$$\sup_{x\in\mathbb{R}}|P_n((-\infty, x]) - P((-\infty, x])| \longrightarrow 0 \quad c.s.$$

Esto nos dice que para un $n$ suficientemente avanzado podemos aproximar la medida teórica $P$ por la medida empírica $P_n$. Además, este índice resulta común para todos los $x$ $P - c.s.$

Una forma alternativa de expresar este resultado, si escribimos $\mathscr{A} = \{(-\infty, t]/t \in \mathbb{R}\}$, sería

$$\sup_{A\in\mathscr{A}}|P_n(A) - P(A)| \longrightarrow 0 \quad c.s.$$

Para la clase de conjuntos $\mathscr{A}$ en $\mathbb{R}$ es trivial que se cumple esta convergencia uniforme, pero estaríamos interesados ahora en estudiar si esta propiedad se verifica en familias de conjuntos más interesantes, como pueden ser el producto de semi intervalos $\prod_{i=1}^{p}(-\infty, x_i]$ en $\mathbb{R}^p$, semiespacios, bolas, elipsoides...

Por otro lado, dada una colección $\mathscr{F}$ de funciones medibles en el espacio considerado con llegada a $\mathbb{R}$, la medida de probabilidad empírica $P_n$ induce una aplicación de $\mathscr{F}$ a $\mathbb{R}$ de la forma

$$f \to P_n f = \int f dP_n$$

Si denotamos $\|Q\|_F = \sup\{|Qf|/f \in \mathscr{F}\}$, la versión uniforme de la Ley de los Grandes Números para el conjunto de funciones $\mathscr{G} = \{I_{[x,\infty)} : x \in \mathbb{R}\}$ se escribiría como $\|P_n - P\|_{\mathscr{G}} \to 0$ $c.s.$, acorde con la idea de que la Esperanza Matemática no es más que la extensión natural de la probabilidad por linealidad y continuidad monótona. Surge entonces también la cuestión de determinar si se verifica esta propiedad en el caso de considerar clases más generales de funciones como aquellas que son indicador de conjuntos más "complicados", la clase de funciones continuas y diferenciables, la clase de funciones monótonas...

**Definición 5.1.** *Sean $X_1, .., X_n, ...$ v.a.i.i.d con distribución de probabilidad común $P$. Sea $P_n$ la medida empírica. Una clase de funciones $\mathscr{F}$ que satisface $\|P - P_n\|_{\mathscr{F}} \to 0$ c.s se denominará clase de Glivenko-Cantelli (clase GC).*

**Observación 5.2.** *Al igual que planteamos el problema de encontrar aquellas clases de funciones para las que se verifica una versión uniforme de la Ley Fuerte de los Grandes Numeros, podríamos plantearnos para qué clases se da el Teorema Central del Límite uniformemente. Estas clases se denominan de Donsker y quedan fuera de los objetivos de este documento, por lo que no hablaremos de ellas aquí. Las clases de Donsker y de Glivenko Cantelli están relacionadas ya que si una clase es de Donsker, es también GC.*

Las clases Glivenko-Cantelli son cerradas bajo transformaciones lineales y afines, combinaciones lineales finitas y bajo límites puntuales. Esto permite extender esta propiedad a clases más complicadas que pueden expresarse como combinación lineal, transformación afín o límite de clases más simples que sí poseen la propiedad de ser GC.

Introducimos a continuación las clases de Vapnik-Chervonenkis, cuya relación con las clases de Glivenko-Cantelli nos facilitará el estudio de las mismas.

**Definición 5.3.** *Sea $A$ un conjunto y $\mathscr{C}$ una clase de conjuntos. Decimos que la clase $\mathscr{C}$ rompe o fragmenta el conjunto $A$ si para cada subconjunto $B$ de $A$ existe $C \in \mathscr{C}$ tal que $C \cap A = B$. Definimos la dimensión $VC$ de una clase de conjuntos, $V(\mathscr{C})$, como el cardinal máximo que puede tener el conjunto $A$ para que la clase $\mathscr{C}$ pueda romperla, es decir, si consigue fragmentar los $2^n$ subconjuntos de $A = \{x_1, .., x_n\}$. Una clase con una dimensión VC finita se denomina clase Vapnik-Chervonenkis o clase VC.*

Por ejemplo, podríamos considerar $\mathscr{L}$ la clase de los intervalos en la recta real y $\mathscr{F}$ la clase de los conjuntos finitos en la recta real.

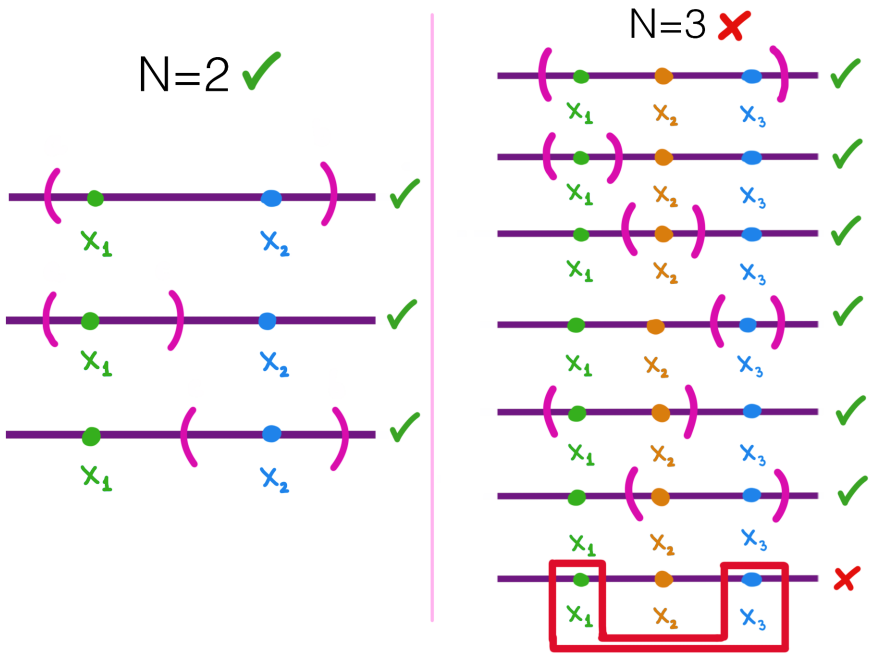



Figura 31: $\mathscr{L}$ fragmenta conjuntos de cardinal 2 pero no de cardinal 3

1. Dado un conjunto $A = \{x_1, x_2\}$, $\mathscr{L}$ fragmenta $A$: Esto significa que para cualquier conjunto finito de 2 puntos, existe al menos un intervalo que puede "rodear" estos dos puntos o separarlos. Sin embargo, para cualquier conjunto de 3 puntos, $A = \{x_1, x_2, x_3\}$, no existe un intervalo

que pueda separarlos de todas las formas posibles. Por lo tanto, la dimensión VC de $\mathscr{L}$ es 2.

2. La dimensión VC de la clase $\mathscr{F}$ es infinita, ya que dado cualquier conjunto finito de puntos $A$ en $\mathbb{R}$, sus propios subconjuntos ya pertenecen a $\mathscr{F}$. Dado que podemos aumentar el tamaño de $A$ arbitrariamente, la dimensión VC de $\mathscr{F}$ es infinita.

En este documento utilizaremos el siguiente resultado que permite relacionar las clases VC con las clases de Glivenko-Cantelli.

**Teorema 5.4.** *Una clase $\mathscr{C}$ es de Glivenko-Cantelli si y solo si es una clase Vapnik-Chervonenkis*

La idea intuitiva que subyace bajo este resultado es que algunas clases de conjuntos son "demasiado grandes" para que se pueda verificar el Teorema de Glivenko-Cantelli.

Por ejemplo, si consideramos la clase de las bolas en $\mathbb{R}^p$, que denotamos por $\mathscr{B}$, es claro que su dimensión VC es finita ya que a partir de un cardinal del conjunto $A$, una bola no es capaz de separar todos los subconjuntos de $A$. Esto provoca que $\mathscr{B}$ sea una clase de Glivenko Cantelli, y que por lo tanto

$$\sup_{B\in\mathscr{B}} |P_n(B) - P(B)| \longrightarrow 0 \quad c.s.$$

## 5.2. Algunos resultados auxiliares

Recordemos que las restricciones que establecíamos al plantear la existencia y consistencia de soluciones en el modelo de mezclas eran las siguientes:

1. **(PR)** Restricciones sobre la distribución P:
   - La distribución P no debe estár concentrada en $G$ puntos.
   - P debe de tener momento de segundo orden finito, esto es: $\mathbb{E}_P(\|\cdot\|^2) < \infty$. En el caso de estar trabajando con recortes, esta hipótesis no sería necesaria.
2. **(ER)** Restricciones sobre el ratio entre autovalores: Debe cumplirse, para una constante $c \geq 1$, que
   $$M_n/m_n \leq c$$
   Donde
   $$M_n = \max_{1\leq j\leq G} \max_{1\leq l\leq p} \lambda_{jl}$$
   $$m_n = \min_{1\leq j\leq G} \min_{1\leq l\leq p} \lambda_{jl}$$

**Lema 5.5.** *Sea $\{\Psi_n\}_{n=1}^{\infty}$ una sucesión de estimadores tal que $\lim_{n\to\infty} L(\Psi_n, P) > -\infty$. Dadas las restricciones* **(PR)** *y* **(ER)**, *existe $h > 0$ tal que $\mathbb{E}_P\left(\max_{1\leq j\leq G}\|x - \mu_j^n\|^2\right) \geq h > 0$.*

**Demostración**

Dado que $P$ no está concentrada en $G$ puntos, existen $y_1, .., y_{G+1} \in \mathbb{R}^p$ tales que $P(B(y_j, \epsilon)) > \delta_\epsilon > 0$, donde $B(y, \epsilon)$ denota la bola de centro $y$ con radio $\epsilon$.

Consideramos ahora $\epsilon_0 < \text{mín}_{1\leq j<k\leq G+1} \frac{\|y_j - y_k\|}{2}$ y las bolas $B(y_j, \epsilon_0)$ con $j = 1, .., G+1$. Observamos que, dado que son bolas abiertas, son disjuntas dos a dos. Dado que tenemos $G$ vectores $\mu_j^n$ y $G+1$ puntos $y_j$, sabemos que existe un índice, vamos a decir que es el $G+1$ tras reordenarlos, de tal manera que la bola $B(y_{G+1}, \epsilon_0)$ no contiene a ninguno de los $\mu_j^n$. Denotamos como $B_{G+1}$ esta bola.
De esta manera, podemos escribir

$$\mathbb{E}_P(\max_{1\leq j\leq G}\|x - \mu_j^n\|^2) \geq \mathbb{E}_P(\min_{1\leq j\leq G}\|x - \mu_j^n\|^2) =$$

$$= \mathbb{E}_P(\min_{1\leq j\leq G}\|x - \mu_j^n\|^2 \cdot I_{B_{G+1}}) + \mathbb{E}_P(\min_{1\leq j\leq G}\|x - \mu_j^n\|^2 \cdot I_{B_{G+1}^c}) \geq \mathbb{E}_P(\min_{1\leq j\leq G}\|x - \mu_j^n\|^2 \cdot I_{B_{G+1}})$$

La última desigualdad es consecuencia de que $\text{mín}_{1\leq j\leq G}\|X - \mu_j^n\|^2$ sea una variable positiva. Dado que $x \in B_{G+1}$ y $\mu_j^n$ no, se tiene que

$$\mathbb{E}_P(\min_{1\leq j\leq G}\|x - \mu_j^n\|^2 \cdot I_{B_{G+1}}) \geq \mathbb{E}_P\left(\left(\min_{1\leq j<k\leq G+1}\frac{\|y_j - y_k\|}{2}\right)^2 \cdot I_{B_{G+1}}\right) \geq \epsilon_0^2 \cdot P(B_{G+1}) > 0$$

□

**Lema 5.6.** *Sea $\{P_n\}_{n=1}^\infty$ una sucesión de distribuciones de probabilidad empíricas que satisfacen* **(ER)**. *Sea $\{\Psi_n\}_{n=1}^\infty$ una sucesión de estimadores muestrales. Si $P$ satisface* **(PR)**, *entonces existe $h' > 0$ tal que $\mathbb{E}_{P_n}\left(\text{máx}_{1\leq j\leq G}\|x - \mu_j^n\|^2\right) \geq h'$.*

**Demostración**

*Hemos comentado que $\sup_{B\in\mathscr{B}}|P_n(B) - P(B)| \longrightarrow 0$ c.s. por ser $\mathscr{B}$ de Glivenko-Cantelli. La expresión anterior se puede reescribir como*

$$\sup_{B\in\mathscr{B}}|P_n(B) - P(B)| = \sup_{B\in\mathscr{B}}|\int I_B dP_n - \int I_B dP| = \sup_{B\in\mathscr{B}}|\mathbb{E}_{P_n}(I_B) - \mathbb{E}_P(I_B)| \longrightarrow 0 \quad c.s.$$

*Esto nos permite afirmar que, de manera uniforme en $\mathscr{B}$, para un $n$ elevado $\mathbb{E}_{P_n}(I_B)$ dista muy poco de $\mathbb{E}_P(I_B)$. Por lo tanto, tomando un índice suficientemente avanzado podríamos sustituir una esperanza por la otra y conseguiríamos demostrar $\mathbb{E}_{P_n}\left(\text{máx}_{1\leq j\leq G}\|x - \mu_j^n\|^2\right) \geq h'$ siguiendo unos pasos análogos a los que utilizábamos en el lema anterior para probar $\mathbb{E}_P\left(\text{máx}_{1\leq j\leq G}\|x - \mu_j^n\|^2\right) \geq h$.*
□

# Referencias


[1] Billingsley, P. (2013). *Convergence of probability measures.* John Wiley and Sons.

[2] Bouveyron, C., Celeux, G., Murphy, T. B., and Raftery, A. E. (2019). *Model-based clustering and classification for data science: with applications in R (Vol. 50).* Cambridge University Press.

[3] Fritz, H., García-Escudero, L. A., Mayo-Iscar, A. (2013). *Computational Statistics and Data Analysis, 61, 124-136.* Computational Statistics and Data Analysis, 61, 124-136.

[4] García-Escudero, L. A., Gordaliza, A., and Mayo-Íscar, A. (2014). *A constrained robust proposal for mixture modeling avoiding spurious solutions.*. Advances in Data Analysis and Classification, 8(1), 27-43.

[5] García-Escudero, L. A., Gordaliza, A., Matrán, C., and Mayo-Iscar, A. (2008). *A general trimming approach to robust cluster analysis.* 1324-1345.

[6] García-Escudero, L. A., Gordaliza, A., Matrán, C., and Mayo-Iscar, A. (2015). *Avoiding spurious local maximizers in mixture modeling.* Statistics and Computing, 25, 619-633.

[7] Hathaway, R. J. (1985). *A constrained formulation of maximum-likelihood estimation for normal mixture distributions.* . The Annals of Statistics, 13(2), 795-800.

[8] McLachlan, G.J and Peel, D.(1999) *Computing Issues for the EM Algorithm in Mixture Models.* Interface'99, Schaumbury Illinois, Virginia: Interface Foundation of North America.

[9] McLachlan, G. J., Lee, S. X., and Rathnayake, S. I. (2019). *Finite mixture models.* Annual review of statistics and its application, 6, 355-378.

[10] McLachlan, G. J., and Krishnan, T. (2007). *The EM algorithm and extensions.* John Wiley and Sons.

[11] Teicher, H. (1963): *Identifiability of Finite Mixtures.* Mathematical Statistics, 34,4,1265–1269

[12] Van der Vaart, A., and Wellner, J. A. 1996. *Weak convergence and empirical processes.* Springer New York.

[13] Yakowitz, S. J., and Spragins, J. D. (1968). *On the identifiability of finite mixtures.* The Annals of Mathematical Statistics, 39(1), 209-214.