FACULTADE DE MATEMÁTICAS

Traballo Fin de Grao

Técnicas de formación de grupos: Métodos de particionamiento

Iria Lago Portela

2019/2020

UNIVERSIDADE DE SANTIAGO DE COMPOSTELA

GRAO DE MATEMÁTICAS

Traballo Fin de Grao

Técnicas de formación de grupos: Métodos de particionamiento

Iria Lago Portela

07/09/2020

UNIVERSIDADE DE SANTIAGO DE COMPOSTELA

Trabajo propuesto

Área de Coñecemento: Estatística e Investigación Operativa
Título: Técnicas de formación de grupos: métodos de particionamiento
Breve descrición do contido
Dado un conjunto de observaciones, las técnicas de formación de grupos tienen por objetivo reconocer patrones o estructuras dentro de una población general. Entre las técnicas de formación de grupos distinguimos los métodos jerárquicos y los métodos de particionamiento. En este trabajo la alumna deberá hacer una revisión exhaustiva de los métodos de particionamiento existentes en la literatura y centrarse en el método de $k-$medias. Además de la revisión teórica deberá aplicar el método a un conjunto de datos reales y a datos simulados que permitan ilustrar su aplicabilidad y también identificar situaciones en las que los resultados no son los deseables.
Recomendacións
Outras observacións

Índice general

Resumen

El análisis cluster es un conjunto de técnicas de análisis multivariante que permiten clasificar conjuntos de datos en grupos, de forma que los individuos dentro de cada grupo presenten cierto grado de homogeneidad respecto de las variables observables. En este trabajo hemos revisado de forma teórica los métodos de $K-$medias y de mixturas finitas para la agrupación de un conjunto de datos, y hemos realizado un análisis comparativo de ambos métodos. En el primer capítulo se ofrece una introducción al análisis cluster y, en particular, a los métodos de particionamiento. En el segundo y tercer capítulos se hizo una revisión teórica de los algoritmo de $K-$medias y mixturas finitas, respectivamente, y se ilustran los algoritmos mediante un ejemplo con datos reales. En el cuarto, y último capítulo, hemos realizado varias simulaciones que nos permitieron comparar los algoritmos de $K-$medias y mixturas finitas y hemos podido ver en qué circunstancias es más adecuado aplicar cada uno de los métodos.

Abstract

Cluster analysis is a set of multivariate analysis techniques that allow data sets to be classified into groups, so that the individuals within each group present a certain degree of homogeneity with respect to the observable variables. In this work we have theoretically reviewed the $K-$means and finite mixture methods for grouping a data set, and we have performed a comparative analysis of both methods. The first chapter provides an introduction to cluster analysis and, in particular, to partitioning methods. In the second and third chapters it was made a theoretical review of the algorithms of $K-$ means and finite mixtures, and the algorithms are illustrated by an example with real data. In the fourth, and last chapter, we have carried out several simulations that allowed us to compare the algorithms of $K-$means and finite mixtures and we have been able to see in which circumstances it is more appropriate to apply each of the methods.

Capítulo 1

Formación de grupos en análisis multivariante

El análisis cluster es un conjunto de técnicas de análisis multivariante que permiten clasificar conjuntos de datos en un número determinado de grupos o clusters. El objetivo es encontrar grupos cuyos datos sean los más similares entre sí, mientras que los datos entre grupos difieran lo máximo posible. Esta rama de la Estadística tiene aplicaciones en minería de datos, reconocimiento de patrones, aprendizaje automático, análisis de imágenes, etc.

1.1. Problemas del análisis cluster

Antes de comenzar con la teoría relativa al análisis cluster explicaremos los dos problemas fundamentales que surgen a la hora de desarrollar un método cluster efectivo.

El primero de ellos surge cuando se intenta encontrar grupos no superpuestos en los datos. Una forma lógica de afrontar este problema sería formar cada grupo de datos posible, evaluar cada grupo y luego seleccionar la mejor partición de los datos. Sin embargo, este método no es práctico y es costoso computacionalmente, puesto que el número de formas en que se pueden dividir n elementos en K grupos no superpuestos viene dado por el número de Stirling del segundo tipo (Weisstein, 2003, [33]):

$$\frac{1}{n!}\sum_{i=1}^{n}(-1)^{n-i}\binom{n}{i}n^K. \tag{1.1}$$

Para valores de n entre 20 y 30 se ha usado la programación dinámica para encontrar particiones óptimas. Sin embargo, a medida que n aumenta, es prácticamente imposible la evaluación de una función criterio para cada una de las posibles particiones. Por lo tanto,

se han diseñado algoritmos heurísticos, es decir, aquellos que buscan una solución dentro de un subconjunto del total de particiones.

El segundo problema es que la mayoría de los algoritmos de agrupamiento proporcionan particiones de los datos aunque no exista una estructura cluster presente en los datos.

Teniendo en cuenta estos antecedentes, es lógico pensar que no existe un único método de agrupación que sea válido en todas las aplicaciones.

1.2. Estructura del análisis cluster

El análisis cluster puede organizarse como una secuencia de siete pasos fundamentales que representan las decisiones o pasos críticos tomados en el proceso de agrupamiento. En casos particulares puede ser necesaria una variación de esta secuencia de pasos.

Antes de explicar la secuencia de pasos, es importante conocer las diferencias entre análisis cluster y método cluster. Un método cluster se refiere al medio por el cual se forman los grupos, mientras que el análisis cluster se refiere a toda la secuencia de pasos que representan un análisis completo. Es decir, un método cluster representa un único paso dentro del análisis cluster.

Los pasos esenciales del análisis cluster son:

1. Selección del conjunto de datos. Estos datos deben ser escogidos de tal forma que sean representativos de la estructura de grupos que se cree que está presente en los datos.

2. Selección de variables. Estas variables representan las medidas que consideraremos en cada dato. Las variables deben contener información suficiente para permitir la correcta agrupación de los datos.

3. Debe tomarse una decisión acerca de la estandarización de cada una de las variables usadas en el análisis cluster. La estandarización no siempre es necesaria. En caso de serlo, se debe seleccionar un método adecuado de estandarización.

4. Debe seleccionarse una medida de similitud o disimilitud. Esta medida refleja el grado de cercanía o separación entre los datos que serán agrupados. Explicaremos este tipo de medidas de forma detallada en la sección 1.4.

5. Selección de un método adecuado para el tipo de agrupación que se espera que esté presente en los datos. Esta decisión es importante porque los diferentes métodos cluster tienden a encontrar diferentes tipos de estructuras de agrupación.

6. Se debe determinar el número de clusters en la solución. Esta decisión es uno de los grandes problemas del análisis cluster, puesto que no existe información a priori en cuanto al número esperado de grupos en los datos.

7. El último paso consta de tres componentes: interpretación, testeo y replicación. Se comienza con la interpretación de los resultados dentro del contexto del problema. A continuación, se usan test para determinar si existe una estructura de cluster significativa en los datos. Finalmente, la replicación se usa para determinar si la estructura de cluster obtenida se puede replicar en una segunda muestra.

En las siguientes secciones explicaremos con detalle algunos de estos pasos. Para ello será útil fijar la notación que emplearemos.

Dada una población de n individuos para los cuales observamos p variables, la forma general de representación de este conjunto de datos es mediante la matriz de datos multivariante. Esta matriz, que denotamos por X, tiene dimensión $n \times p$, y viene dada por:

$$X = \begin{pmatrix} x_{11} & x_{12} & \dots & x_{1p} \\ x_{21} & \dots & \dots & \dots \\ \vdots & \vdots & \vdots & \vdots \\ x_{n1} & \dots & \dots & x_{np} \end{pmatrix}, \tag{1.2}$$

donde cada fila $i \in \{1, \dots, n\}$ representa un individuo y cada columna $j \in \{1, \dots, p\}$ una variable. Por lo tanto, x_{ij} representa el valor de la variable $j-$ésima para el individuo $i-$ésimo.

Las variables de la matriz X pueden ser categóricas, continuas o una mezcla de ambas. En este trabajo consideraremos únicamente variables continuas.

1.3. Estandarización de las variables

El siguiente paso consiste en la estandarización de las variables, aunque este paso puede no ser necesario. El proceso de estandarización es un paso lógico si se cree que los grupos existen en el espacio de la variable transformada, puesto que si los grupos existen en el espacio de la variable original, entonces la estandarización puede distorsionar o esconder los grupos presentes en los datos.

La forma más conocida de estandarización de las variables es el método $z-$score, pero existen muchos otros métodos. En el contexto del análisis cluster, Milligan y Cooper (1988, [26]) investigaron ocho métodos diferentes de estandarización de variables continuas bajo diferentes condiciones de error y concluyeron que la estandarización por división por el rango es el método más efectivo, más incluso que el método $z-$score.

Explicaremos ambos tipos de estandarizaciones. La primera forma de estandarización de la que hablaremos es la fórmula $z-$score, usada para transformar variables con distribución normal a estándar. Para el caso univariante, si $X_j \in N(\mu, \sigma^2)$, entonces

$$Z_1 = \frac{X_j - \bar{X}_j}{s} \in N(0,1) \tag{1.3}$$

siendo X_j la columna $j-$ésima de la matriz original de datos multivariantes, $\bar{X}_j$ su media muestral y s su cuasi-desviación típica.

Algunos autores de advierten que la estandarización $z-$score puede no funcionar correctamente si existen diferencias sustanciales entre las desviaciones dentro del cluster. Nótese que la estandarización Z_1 debe ser aplicada de forma global y no dentro de grupos individuales, puesto que puede llevar a una solución del problema cluster incorrecta. Las siguientes estandarizaciones usan el rango de la variable como divisor. Vienen dadas por:

$$Z_2 = \frac{X_j}{max(X_j) - min(X_j)}, \tag{1.4}$$

$$Z_3 = \frac{X_j - min(X_j)}{max(X_j) - min(X_j)}. \tag{1.5}$$

1.4. Selección de las medidas de similitud o disimilitud

Un elemento importante a la hora de identificar grupos dentro de un conjunto de datos es la distancia entre individuos, es decir, si los individuos están próximos o lejanos unos de otros. Muchas técnicas de formación de grupos comienzan con una matriz donde sus elementos reflejan una medida cuantitativa de la cercanía. Nosotros consideraremos dos formas de medir la cercanía: la semejanza y la disimilitud.

Definición 1.1. La semejanza entre dos objetos es una medida numérica del grado en el que dos objetos son parecidos. Por lo tanto, la semejanza será alta para pares de objetos que sean más parecidos. La semejanza es una medida no negativa y puede ser escalada para tomar valores en el intervalo $[0,1]$. Denotaremos a esta medida por s_{ij}, $i \in \{1,\ldots,n\}$, $j \in \{1,\ldots,p\}$.

Definición 1.2. La disimilitud, al contrario que la semejanza, es una medida numérica del grado en el que dos objetos son diferentes. Por tanto, la disimilitud es baja para los pares de objetos que sean más parecidos. Al igual que con la semejanza, la disimilitud es una medida no negativa y puede ser escalada para tomar valores en el intervalo $[0,1]$. Denotaremos a esta medida por δ_{ij}, $i \in \{1,\ldots,n\}$, $j \in \{1,\ldots,p\}$.

Diremos que dos individuos están próximos cuando su disimilitud sea pequeña, o bien cuando su semejanza sea grande. La disimilitud se denominará medida de distancia cuando también verifique la desigualdad triangular:

$$\delta_{ij} + \delta_{im} \geq \delta_{jm} \tag{1.6}$$

De forma rigurosa podemos definir el concepto de distancia.

Definición 1.3. Dado un conjunto X, una distancia sobre X es una aplicación $d : X \times X \to \mathbb{R}$ que a cada par de puntos $x, y \in X$ le asocia un número real $d(x, y)$, que verifica las siguiente propiedades:

1. $d(x, y) \geq 0$ para todo $x, y \in X$ y $d(x, y) = 0$ si y sólo si $x = y$
2. $d(x, y) = d(y, x)$ para todo $x, y \in X$
3. $d(x, y) \leq d(x, z) + d(z, y)$ para todo $x, y, z \in X$ (Desigualdad triangular)

En la figura 1.1 se puede observar una representación gráfica de la desigualdad triangular.

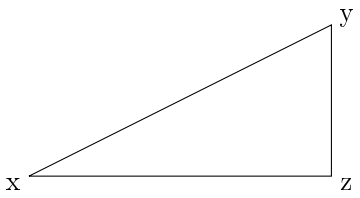

Figura 1.1: Representación geométrica de la desigualdad triangular

En el análisis cluster será de gran importancia este concepto para calcular las distancias entre individuos y entre grupos. Por ello, introduciremos varias distancias que nos serán de utilidad en la práctica. En primer lugar definiremos la distancia Euclidiana, que viene dada por:

$$d_{ij} = \left[\sum_{k=1}^{p} (x_{ik} - x_{jk})^2 \right]^{1/2}, \tag{1.7}$$

donde x_{ik} y x_{jk} son, respectivamente, los valores de la $k-$ésima variable de la observación $p-$dimensional para los individuos i y j. Esta medida de distancia tiene la propiedad de que d_{ij} puede ser interpretada como la distancia física entre dos puntos $p-$dimensionales $x'_i = (x_{i1}, \ldots, x_{ip})$ y $x'_j = (x_{j1}, \ldots, x_{jp})$ en el espacio Euclidiano. En la figura 1.2 se puede observar la representación gráfica de la distancia euclidiana entre dos puntos.

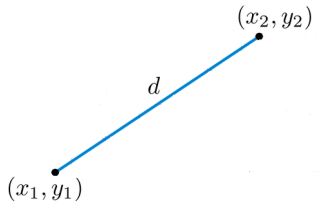

Figura 1.2: Representación de la distancia euclidiana

La siguiente distancia se denomina distancia del taxi o distancia de Manhattan, y viene dada por

$$d_{ij} = \sum_{k=1}^{p} w_k |x_{ik} - x_{jk}|, \tag{1.8}$$

donde w_k es el peso de la variable $k-$ésima. Otra distancia bien conocida es la distancia de Minkowski, definida como

$$d_{ij} = \left(\sum_{k=1}^{p} w'_k |x_{ik} - x_{jk}|^r \right)^{1/r}, \ r \geq 1, \tag{1.9}$$

donde w'_k es el peso de la variable $k-$ésima. Tanto la distancia Euclidiana como la distancia del taxi son casos particulares de la distancia de Minkowski para los valores $r = 2$ y $r = 1$ respectivamente. En la figura 1.3 se puede observar la representación gráfica de la distancia de Minkowski para distintos valores de r.

Por último explicaremos la distancia de Mahalanobis, utilizada para determinar la similitud entre variables aleatorias multidimensionales. Esta distancia se define como

$$d_{ij} = (x_i - x_j)' \Sigma^{-1} (x_i - x_j), \tag{1.10}$$

siendo Σ^{-1} la inversa de la matriz de varianzas-covarianzas.

Se han realizado estudios comparando el uso de algunas de estas distancias. Sin embargo, el problema de estos estudios es que las medidas de semejanza o disimilitud óptimas son muy dependientes de la naturaleza de los datos generados.

1.5. Métodos clúster

Dentro del análisis cluster, las técnicas de agrupamiento se dividen en dos categorías: jerárquicas o de particionamiento. A continuación explicaremos ambas técnicas, centrándonos en estas últimas.

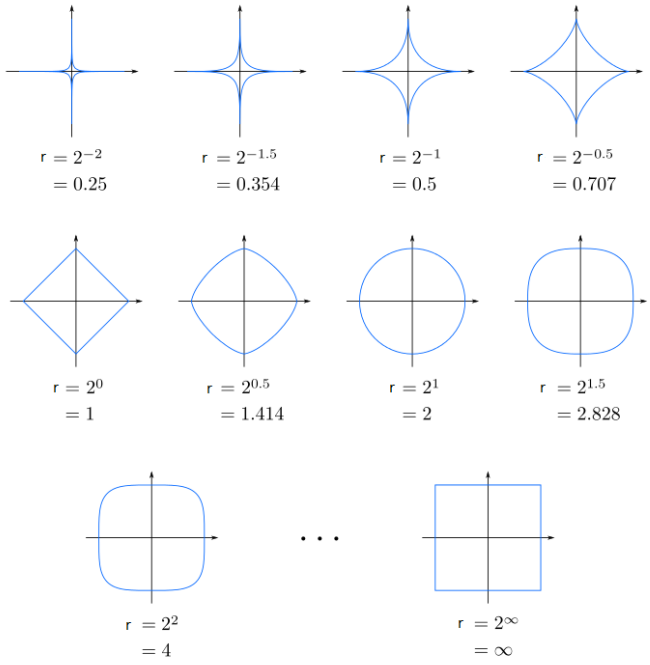

Figura 1.3: Representación de las distancias de Minkowski

1.5.1. Técnicas jerárquicas

En una clasificación jerárquica, los datos no se dividen en un número particular de clases o grupos en un único paso, sino que la clasificación consiste en una serie de particiones, que pueden ejecutarse desde un solo clúster que contiene todos los individuos, hasta n grupos cada uno conteniendo un individuo. Las técnicas de agrupamiento jerárquico pueden subdividirse en métodos aglomerativos, que proceden de una serie de fusiones sucesivas de los n individuos en grupos y métodos divisivos, que separan a los n individuos sucesivamente en agrupaciones más finas.

Cuando en los métodos jerárquicos se realizan fusiones o divisiones, éstas son irreversibles. Esto es, cuando en un algoritmo aglomerativo se juntan dos individuos, no pueden posteriormente ser separados; y cuando en un algoritmo divisivo dos individuos se separan, no pueden volver a ser juntados.

Las clasificaciones jerárquicas producidas tanto por métodos aglomerativos o divisivos pueden ser representados mediante diagramas bi-dimensionales, conocido como dendogramas, que ilustra las fusiones o divisiones realizadas en cada etapa del análisis. Un ejemplo de este diagrama es dado en la figura 1.4.

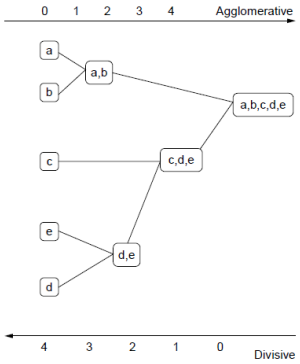

Figura 1.4: Ejemplo de dendograma

1.5.2. Técnicas de particionamiento

Las técnicas de particionamiento son una clase de técnicas de formación de grupos que producen una partición de los individuos en un número específico de grupos, mediante la minimización o maximización de algún criterio numérico.

A continuación introduciremos la formulación matemática del problema.

Dada una población de n individuos para los cuales observamos p variables, representados mediante la matriz de datos multivariante (1.2), nuestro objetivo es particionar el conjunto de datos X en K grupos, siendo K fijado de antemano. Diremos que $P = \{C_1, \ldots, C_K\}$ es una partición de X en K grupos si verifica:

1. $C_k \neq \emptyset$ para todo $k \in \{1, \ldots, K\}$
2. $C_k \cap C_{k'} = \emptyset$ para todo $k, k' \in \{1, \ldots, K\}$
3. $\cup_k C_k = X$.

Denotaremos al conjunto de las particiones de X en K grupos mediante $\mathcal{P}(X)$.

Por lo tanto, las técnicas de particionamiento consisten en encontrar la partición de X en K grupos que maximice una función criterio prefijada J que mide la calidad de la partición P.

Este problema puede ser formulado mediante el siguiente modelo de optimización:

$$\begin{array}{ll} \texttt{maximizar} & J(P) \\ \texttt{sujeto a} & P \in \mathcal{P}(X) \end{array} \tag{1.11}$$

En el caso particular de que el número de objetos sea finito y el número de clusters sea K, el conjunto factible del problema de optimización (1.11) se puede expresar del siguiente

modo:

$$\sum_{k=1}^{K} w_{ik} = 1;\ i = 1, \ldots, n \tag{1.12}$$

$$\sum_{i=1}^{n} w_{ik} \geq 1;\ k = 1, \ldots, K \tag{1.13}$$

$$w_{ik} \in \{0, 1\};\ i = 1, \ldots, n;\ k = 1, \ldots, K; \tag{1.14}$$

donde la variable de decisión w_{ik} tomará el valor 1 si el objeto i es asignado al cluster k y 0 en caso contrario. La condición (1.12) indica que todos los objetos deben ser asignados a un único cluster. La condición (1.13) impone que no existan clusters vacíos y (1.14) establece la naturaleza binaria en las variables.

En principio, una partición óptima, basada en la función criterio J, podría encontrarse enumerando todas las posibles particiones. Sin embargo, ya hemos visto que es un problema computacionalmente costoso, puesto que el número de formas de particionar n elementos en K grupos viene dado por (1.11), por lo que conviene utilizar algoritmos numéricos.

Capítulo 2

$K-$medias

En este capítulo presentaremos el algoritmo de $K-$medias, propuesto por Lloyd (1957) [21] y Forgy (1965) [15] de forma independiente. Este algoritmo presenta distintas formulaciones, entre ellas destacan las propuestas por MacQueen (1967) [22] y Hartigang-Wong (1979) [19]. A pesar de ser uno de los métodos de particionamiento más antiguos, sigue siendo muy utilizado, puesto que es un algoritmo fácil de implementar, computacionalmente eficiente y necesita de poco espacio de almacenamiento. Además, en este capítulo ilustraremos el funcionamiento del algoritmo mediante dos ejemplos con datos reales.

2.1. Formulación del método de K-medias

El algoritmo de $K-$medias está diseñado para particionar n objetos en K grupos, $(C_1, \ldots, C_k)$, donde C_k es el conjunto de n_k objetos en el grupo $k-$ésimo. Dado el conjunto de datos X, definido en (1.2), el algoritmo de K-medias construye las particiones de modo que cada individuo queda asignado al grupo cuyo centroide es el más próximo en distancia euclidiana. El centroide de un grupo C_k es un punto de un espacio p-dimensional que se construye promediando los valores en cada variable sobre los objetos del grupo. Por ejemplo, el valor del centroide para la variable $j-$ésima en el grupo C_k es:

$$\bar{x}_j^{(k)} = \frac{1}{n_k} \sum_{i \in C_k} x_{ij}, \tag{2.1}$$

donde n_k es el número de objetos en el grupo C_k. El vector de centroides para el grupo C_k estará dado por:

$$\bar{x}^{(k)} = (\bar{x_1}^{(k)}, \ldots, \bar{x_p}^{(k)})'. \tag{2.2}$$

A pesar de que el algoritmo de $K-$medias posee distintas formulaciones, una de las más típicas consiste en el siguiente proceso iterativo:

(1) Se seleccionan K puntos iniciales, que denominaremos semillas, y que denotaremos mediante $(s_1^{(k)}, \ldots, s_p^{(k)}) \in \mathbb{R}^p$, para $1 \leq k \leq K$.

(2) Se calcula la distancia euclidiana al cuadrado, $d^2(i,k)$, entre el objeto $i-$ésimo y el $k-$ésimo vector semilla:

$$d^2(i,k) = \sum_{j=1}^{p} (x_{ij} - s_j^{(k)})^2. \tag{2.3}$$

Los individuos son asignados al grupo donde (2.3) es mínimo.

(3) Después de la asignación inicial, se obtienen los centroides de cada grupo mediante (2.2).

(4) Posteriormente, los individuos son comparados con cada centroide usando (2.3), de modo que cada individuo será reasignado al grupo cuyo centroide esté más próximo.

(5) Se calculan los centroides de los nuevos grupos.

(6) Se repiten los pasos 4 y 5 hasta que ningún individuo sea reasignado a otro grupo.

2.2. Criterio global para escoger la partición

Cuando intentamos encontrar una "buena" partición de un objeto mediante un método iterativo como el descrito, es interesante notar que también intentamos minimizar el criterio de suma de errores al cuadrado (SSE):

$$SSE = \sum_{j=1}^{p} \sum_{k=1}^{K} \sum_{i \in C_k} (x_{ij} - \bar{x}_j^{(k)})^2 \tag{2.4}$$

El algoritmo de $K-$medias puede reformularse considerando el vector de centroides para el grupo C_k, dado por (2.2), y una matriz adicional, la matriz de miembros $M = \{m_{ik}\}_{n \times K}$ donde m_{ik} vale 1 si el individuo i pertenece al cluster C_k y vale 0 en caso contrario.

La matriz suma de cuadrados y productos cruzados intragrupos W se expresa como:

$$W = \sum_{k=1}^{K} \sum_{i \in C_k} (x_i - \bar{x}^{(k)})(x_i - \bar{x}^{(k)})', \tag{2.5}$$

siendo x_i la fila $i-$ésima de la matriz X y $\bar{x}^{(k)}$ el vector de medias dentro del grupo C_k. También se puede expresar como $W = (X - M\bar{x}^{(k)})'(X - M\bar{x}^{(k)})$, donde X es la matriz de datos. De este modo podemos escribir (2.4) como la traza de la matriz W,

$$SSE = tr(W) = \sum_{k=1}^{K} tr(W_k) \tag{2.6}$$

donde W_k es la matriz de suma de cuadrados y productos cruzados para el $k-$ésimo cluster y está definida por

$$W_k = \frac{1}{2n_k} \sum_{i \in C_k} \sum_{i^* \in C_k} (x_i - x_{i^*})(x_i - x_{i^*})' \; \forall i, i^* = 1, \ldots, n_k \tag{2.7}$$

Por lo tanto, minimizar $tr(W)$ es equivalente a minimizar SSE.

Aunque $tr(W)$ es la función de W más popular para minimizar, existen otras alternativas que surgen al considerar la siguiente relación:

$$T = W + B \tag{2.8}$$

donde T es la matriz total de suma de cuadrados y productos cruzados, que viene dada por

$$T = \sum_{k=1}^{K} \sum_{l=1}^{n_k} (x_{kl} - \bar{x})(x_{kl} - \bar{x})' \tag{2.9}$$

siendo x_{kl} el vector de observaciones del objeto $l-$ésimo en el grupo k y $\bar{x}$ es el vector de medias dado por

$$\bar{x}_j = \frac{1}{n} \sum_{i=1}^{n} x_{ij} \;\; \forall j \in \{1, \ldots, p\}. \tag{2.10}$$

Por otro lado, B es la matriz de suma de cuadrados y productos cruzados entre grupos, que viene dada por:

$$B = \sum_{k=1}^{K} n_k (\bar{x}^{(k)} - \bar{x})(\bar{x}^{(k)} - \bar{x})', \tag{2.11}$$

donde $\bar{x}_k$ es el vector de medias dentro del grupo k.

2.3. Óptimo local

Uno de los principales problemas del algoritmo de $K-$medias es que cuando intentamos minimizar (2.4), el algoritmo no necesariamente proporciona un óptimo global, y dependiendo de la partición inicial que se tome, el algoritmo proporciona un óptimo local que no puede verificarse si coincide con el óptimo global. Gersho y Gray (1992, [16]), proponen que después de aplicar el algoritmo de las $K-$medias se realice una inspección final entre todos los puntos y centroides. Si existe un objeto dentro de un cluster C_k y otro cluster C_{k*} cumpliendo

$$\frac{n_k}{n_k - 1} d^2(i, k) > \frac{n_{k*}}{n_{k*} + 1} d^2(i, k*) \tag{2.12}$$

entonces el elemento $i-$ésimo se reasigne del cluster C_k al cluster C_{k*} y así el valor de SSE se verá reducido Späth (1980, [29]). Si realizamos de manera reiterada este último paso,

hasta no obtener ningún cambio, se garantiza que hemos finalizado, obteniendo un óptimo global. Para evitar el óptimo local, también se sugirió aplicar el método de $K-$medias repetidamente modificando los valores de la partición inicial y aceptando aquel valor que diera mejores resultados en términos de la suma de errores al cuadrado. Sin embargo, Steinley (2003, [31]) demostró que el número de óptimos locales en conjuntos de datos con tamaño moderado puede ser del orden de 1000, por lo que estudios con un número pequeño de particiones iniciales pueden dar lugar a error y no conducir a un óptimo global. No obstante, se ha demostrado que el algoritmo de las $K-$medias usualmente exhibe buenas propiedades de reconstrucción de los clusters, Dimitriadou et al. (2002,[11]).

2.4. Métodos de inicialización

Dado que puede haber numerosos óptimos locales para un mismo conjunto de datos, la elección de los valores iniciales en el algoritmo de $K-$medias es crucial, y se han propuesto varias alternativas en un intento de evitar soluciones localmente óptimas. MacQueen (1967, [22]) propuso escoger K puntos del conjunto de datos como semillas; sin embargo, este procedimiento sufre la influencia del orden inicial de los datos. Otra posibilidad consiste en escoger como semillas K puntos del conjunto de datos de forma aleatoria (McRae, 1971). Milligan (1980, [24]) propuso comenzar el método de $K-$medias utilizando los resultados obtenidos en un procedimiento jerárquico aglomerativo, idea que tuvo bastante apoyo en la literatura.

2.5. Determinación del número de grupos

En secciones anteriores hemos explicado cómo obtener una partición de un conjunto de n individuos en un número específico de grupos, K, mediante el método de $K-$medias. Hemos asumido que el valor de K era conocido de antemano, sin embargo, éste raramente es el caso. En esta sección explicaremos varios métodos para averiguar el número de grupos dentro de un conjunto de datos.

Una evaluación informal consistiría en estudiar gráficamente cómo varía el criterio de agrupamiento utilizado en función de distintos valores de K; un aplanamiento de la curva indica los valores correctos de K. En la figura 2.1 está representada gráficamente la variación de la suma de errores al cuadrado intragrupos con respecto al número de grupos para el conjunto de datos "diabetes", para el cual aplicaremos posteriormente el método de $K-$medias en la sección 2.6.2. Puede observarse un aplanamiento de la curva para los valores de K a partir de 4, luego a la vista de este gráfico se tomarían $K = 4$ grupos.

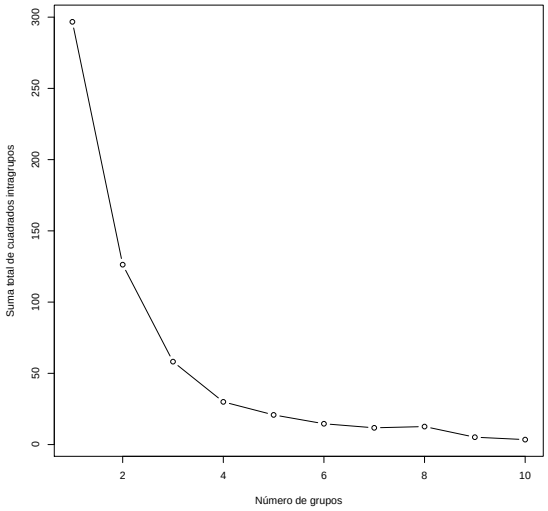

Figura 2.1: Representación gráfica de la variación de la suma de errores al cuadrado intragrupos con respecto al número de grupos para el conjunto de datos "diabetes"

Algunos métodos formales para decidir los valores apropiados para K se han denominado reglas de detención, puesto que consisten en detener el proceso de combinación de los grupos en el valor seleccionado de K. Es conveniente clasificar las reglas de detención en globales o locales. Las reglas globales evalúan una medida, $J(K)$, de la bondad de la partición en K grupos, generalmente basada en la variabilidad intra y entre grupos, e identificando el valor de K para el cual $J(K)$ es óptimo. La desventaja de muchas de las reglas de detención globales es que no poseen una definición natural de $J(1)$, por lo tanto, no aportan información de si los datos deberían o no particionarse. Las reglas de detención locales examinan si un par de grupos debería o no combinarse o si un único grupo debería subdividirse. A diferencia de las reglas globales, éstas están basadas en una parte del conjunto de datos, excepto cuando la comparación se hace para los valores de $K \in \{1, 2\}$. Una desventaja de este tipo de reglas es que generalmente necesitan especificar el nivel de significación.

El estudio más detallado de la comparativa de reglas de detención fue llevado a cabo por Milligan y Cooper (1985, [25]), en el cual investigaron el grado en el que 30 reglas de detención podían detectar el número correcto de grupos en conjuntos de datos simulados, construidos de forma que tuvieran una estructura de grupos clara.

Explicaremos cinco reglas de las que aparecen en el estudio de Milligan y Cooper (1985, [25]). La primera de ellas es el índice de Calinski y Harabasz (1974, [5]), que denotaremos por G_1, para evaluar una partición de un conjunto de individuos descritos por variables

cuantitativas. Está definido por:

$$G_1(K) = \frac{\frac{tr(B)}{K-1}}{\frac{tr(W)}{n-K}}, \tag{2.13}$$

donde W y B fueron definidos en 2.5 y 2.11 respectivamente. El valor que maximiza este índice es tomado como el número correcto de grupos.

La siguiente regla se denomina Gamma (Baker y Hubert, 1975, [1]), y la denotaremos por G_2. El índice viene dado por:

$$G_2(K) = \frac{S_+ - S_-}{S_+ + S_-} \tag{2.14}$$

donde S_+ (resp. S_-) denota el número de comparaciones consistentes que involucran distancias intra y entre grupos (resp. el número de comparaciones inconsistentes). El valor que maximiza este índice se toma como el número de particiones del conjunto de datos.

La siguiente regla, L_1, se trata de una regla local y fue propuesta por Duda y Hart (1973, [12]). Esta regla decide si un grupo debe ser o no dividido en dos subgrupos, y está basada en la comparación de las sumas de las distancias al cuadrado intra grupos ($tr(W)$) con la suma de las distancias al cuadrado entre grupos ($tr(B)$) cuando el grupo fue optimamente dividido en dos. Si el grupo contiene n individuos descritos mediante p variables cuantitativas, la hipótesis de que el grupo es homogéneo, y por tanto no debe ser subdividido, se rechaza si

$$\frac{tr(W)}{tr(B)} < 1 - \frac{2}{\pi p} - z\left[2\frac{1-\frac{8}{\pi^2 p}}{np}\right]^{1/2} \tag{2.15}$$

donde z es una desviación normal estándar especificando el nivel de significación del test. Se toma como número de grupos el mínimo valor de n tal que el índice es mayor que un valor crítico.

La siguiente regla local, L_2, fue propuesta por Beale (1969, [4]). Beale propuso un test para decidir si un grupo debería ser subdividido, involucrando la comparación de

$$F = \frac{\frac{tr(W)-tr(B)}{tr(B)}}{(\frac{n-1}{n-2})2^{2/p} - 1} \tag{2.16}$$

(donde $tr(W)$, $tr(B)$, n y p fueron definidos en L_1) con una distribución $F_{p,(n-2)p}$. Rechazamos la hipótesis de que el grupo debe ser subdividido para valores significativamente grandes de F. El número de grupos es el mínimo valor de n tal que el índice es menor que un valor crítico.

Los autores Tibshirani, Walther y Hastie (2001, [32]) propusieron el estadístico Gap para determinar K, definido como

$$Gap_n(K) = E_n^*\{log(W_k)\} - log(W_k), \tag{2.17}$$

donde W_k viene dado por 2.7 y E_n^* denota la esperanza bajo una muestra de tamaño n de la distribución de referencia. Se tomará el valor de K que maximice $Gap_n(K)$.

Es recomendable no depender unicamente de una regla de detención, sino sintetizar los resultados de varias de ellas, para no llegar a error. En la sección 4.1. realizaremos una comparación de las reglas de detención explicadas anteriormente usando datos simulados.

2.6. Ilustración del método de K-medias

En esta sección ilustraremos el algoritmo de $K-$medias mediante dos ejemplos. Para el primero de ellos usaremos la base de datos "xclara" del paquete "cluster" de R, y será un ejemplo donde el algoritmo de $K-$medias aporte buenos resultados. Para el segundo ejemplo utilizaremos la base de datos "diabetes" del paquete "mclust" de R, y será un ejemplo donde el algoritmo no aporte buenos resultados.

2.6.1. Base de datos "xclara" del paquete "cluster" de R

En esta sección ilustraremos el algoritmo de $K-$medias mediante un ejemplo con la base de datos "xclara" del paquete "cluster" de R. Esta base de datos contiene datos de $n = 3000$ individuos para los cuales observamos $p = 2$ variables V_1 y V_2. En primer lugar seleccionaremos el número de grupos usando las técnicas explicadas en la sección 2.5. del segundo capítulo. Después aplicaremos el método de $K-$medias para el número de grupos obtenidos y compararemos los grupos obtenidos con los grupos reales.

En la figura 2.2 aparece representado gráficamente el conjunto de datos que analizaremos en este ejemplo, indicando el grupo real al que pertenecen.

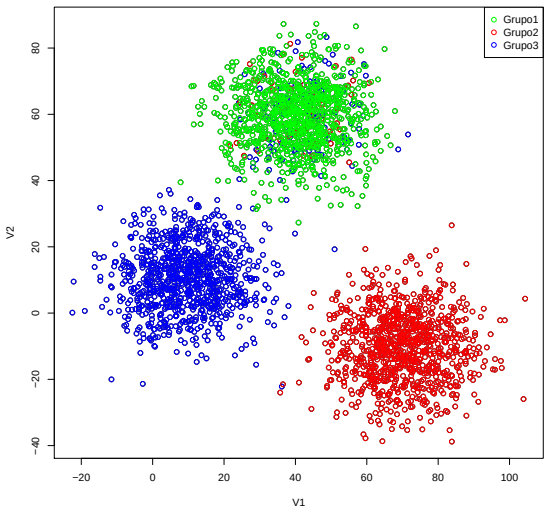

Figura 2.2: Representación gráfica de los grupos reales

Para determinar el número de clusters utilizaremos la librería "NbClust", donde implementamos el índice de Calinski y Harabasz (2.13), la regla Gamma (2.14), la regla de Duda y Hart (2.15), la regla de Beale (2.16) y el estadístico Gap (2.17). Los resultados obtenidos se recogen en el cuadro 2.1.

Regla	Número de grupos
Calinski y Harabasz	3
Gamma	3
Duda y Hart	2
Beale	2
Gap	3

Cuadro 2.1: Resumen del número de grupos obtenidos con las distintas reglas

Debido a problemas computacionales, el resultado $K = 3$ de la regla gamma se ha obtenido utilizando una submuestra aleatoria de $n = 300$ individuos.

Dado que el número de grupos $K = 3$ es el más seleccionado, aplicaremos el método de $K-$medias para este valor.

Además, aplicando el método de aplanamiento para el cálculo del número de grupos, obtenemos un valor de $K = 3$ grupos, como muestra la figura 2.3.

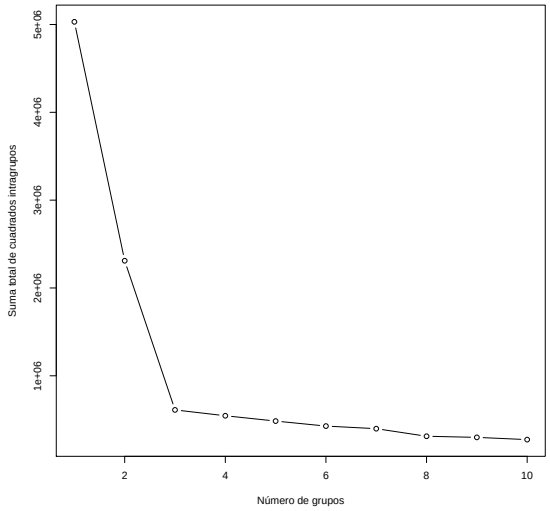

Figura 2.3: Representación gráfica de la variación de la suma de errores al cuadrado intragrupos con respecto al número de grupos para el conjunto de datos "xclara"

Comenzaremos entonces con el primer paso del algoritmo de $K-$medias, la inicialización de los centroides. Como ya hemos explicado en la sección 2.4, existen diversos métodos de inicialización de los datos. En este caso tomaremos tres puntos del conjunto de datos de forma aleatoria como semillas.

Continuamos con el segundo paso del algoritmo. Se calculan las distancias de los datos de las muestras a cada una de las semillas, asignando cada dato a la semilla donde esta distancia sea menor. El siguiente paso consiste en obtener los centroides de los grupos iniciales mediante (2.2). En la figura 2.4 aparece representado gráficamente, a la izquierda, los grupos inciales con las semillas, y a la derecha los mismos grupos con sus respectivos centroides.

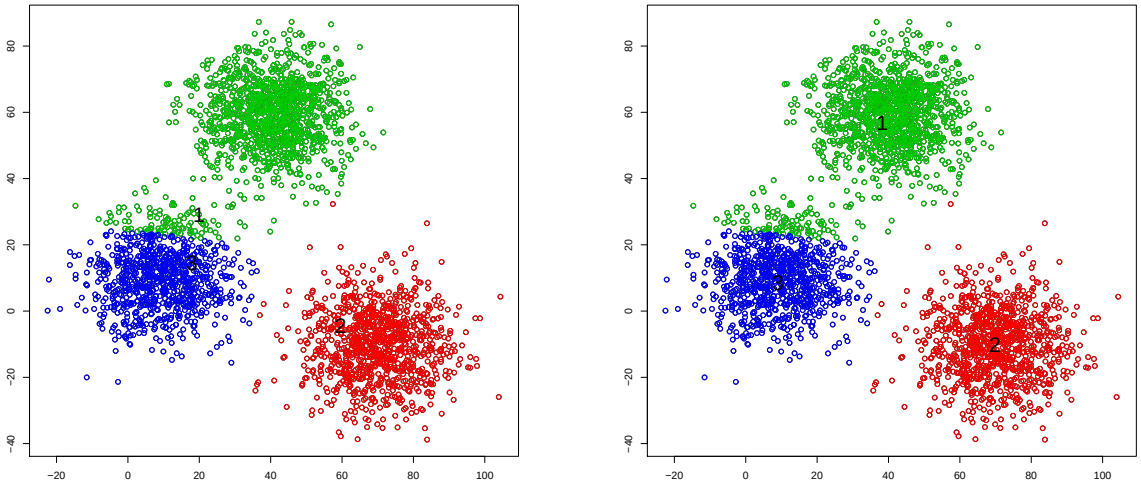

Figura 2.4: Representación gráfica de los centroides iniciales y la asignación inicial (gráfico de la izquierda), junto con el cálculo de centroides en los grupos creados (gráfico de la derecha)

Calcularemos de nuevo las distancias de los datos a los nuevos centroides, reasignando cada individuo al centroide más cercano, y el cálculo de los centroides asociados a esos nuevos grupos. Los nuevos grupos con sus respectivos centroides se muestran en 2.5.

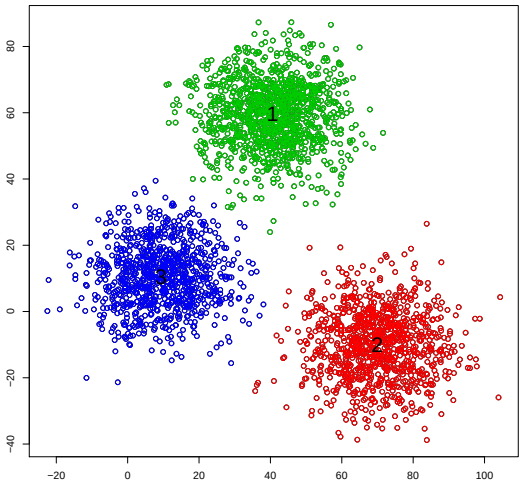

Figura 2.5: Representación gráfica de los nuevos grupos con sus respectivos centroides

Repetimos de nuevo el proceso del cálculo de las distancias de los datos al centroide y reasignamos los datos. Dado que no coinciden con los grupos calculados anteriormente, se realiza una iteración más del algoritmo.

De nuevo, reasignamos los datos a los nuevos centroides, y dado que ningún dato se reasigna, el algoritmo de $K-$medias termina.

En la figura 2.6 aparecen, a la izquierda los grupos reales presentes en los datos, mientras que a la derecha aparecen los grupos obtenidos mediante el algoritmo de $K-$medias.

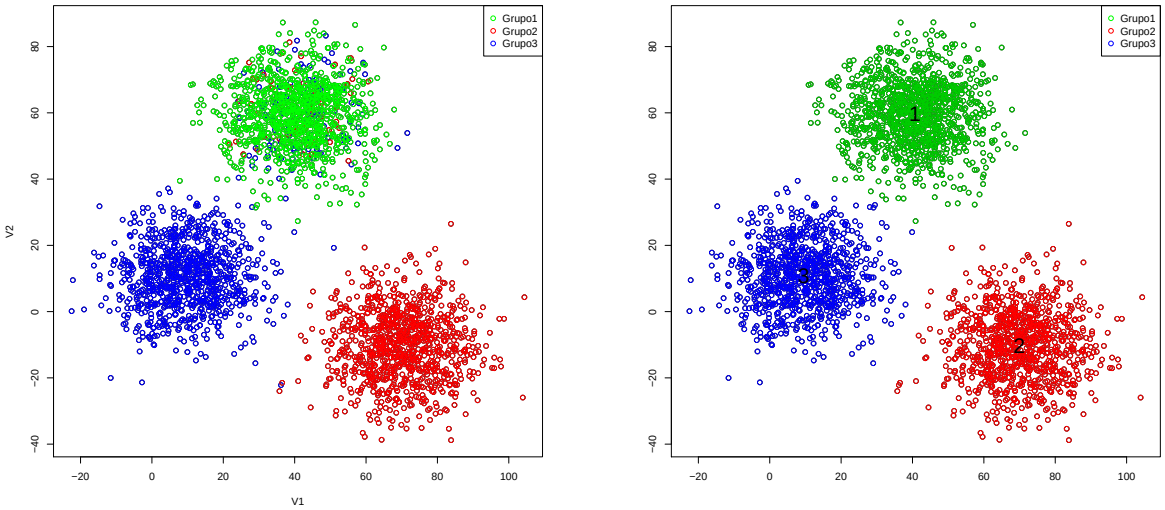

Figura 2.6: Representación gráfica de las diferencias entre los grupos reales y los grupos obtenidos mediante el método de $K-$medias para $K = 3$

En el cuadro 2.2 aparece la relación entre los grupos reales y los obtenidos mediante el algoritmo de $K-$medias.

		Grupo asignado		
		1	2	3
Grupo real	1	997	0	3
	2	50	950	0
	3	102	2	896

Cuadro 2.2: Relación entre los grupos reales y los grupos obtenidos mediante $K-$medias para $K = 3$

Aunque algunos individuos de los grupos 2 y 3 están asignados al grupo 1, se observa

que existen tres grupos bien definidos. Por lo tanto, el algoritmo de $K-$medias ofrece buenos resultados para este ejemplo. Además, no se percibe que pueda haber un método mejor de clasificación para este caso, pues los puntos rojos y azules mal clasificados como verdes están muy mezclados con el grueso de los puntos verdes.

2.6.2. Base de datos "diabetes" del paquete "mclust" de R

En esta sección veremos otro ejemplo del algoritmo de $K-$medias mediante un ejemplo con datos reales del paquete "mclust" de R. Esta base de datos contiene datos de $n = 145$ individuos para los cuales observamos $p = 2$ variables: el nivel de insulina (`insulin`) y de glucosa plasmática en estado estacionario (`sspg`). Estos individuos se encuentran en tres grupos según el nivel de diabetes: "Normal", "Chemical" y "Overt". En primer lugar seleccionaremos el número de grupos usando las técnicas explicadas en la sección 2.5. del segundo capítulo. Después aplicaremos el método de $K-$medias para el número de grupos obtenidos y compararemos los grupos resultantes con los grupos reales. Por último veremos si seleccionando otro número de grupos, se obtienen mejores resultados.

En la figura 2.7 aparece representado gráficamente el conjunto de datos que analizaremos en este ejemplo, indicando el grupo real al que pertenecen.

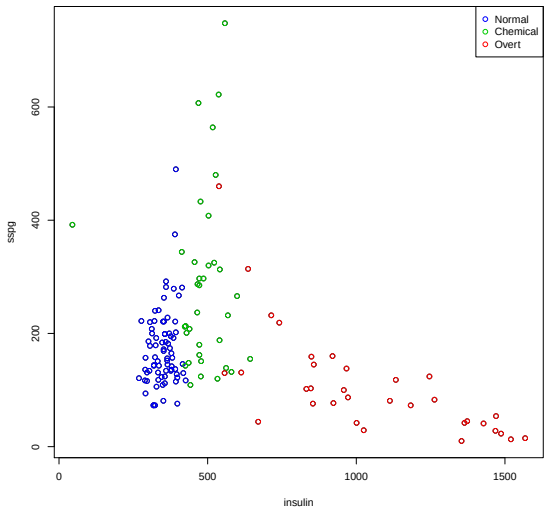

Figura 2.7: Representación gráfica de los grupos reales presentes en los datos

Para determinar el número de clusters utilizaremos la librería "NbClust", donde implementamos el índice de Calinski y Harabasz (2.13), la regla Gamma (2.14), la regla de Duda y Hart (2.15), la regla de Beale (2.16) y el estadístico Gap (2.17). Los resultados obtenidos

se recogen en el cuadro 2.3.

Regla	Número de grupos
Calinski y Harabasz	10
Gamma	2
Duda y Hart	2
Beale	2
Gap	2

Cuadro 2.3: Resumen del número de grupos obtenidos con las distintas reglas

Dado que el número de grupos $K = 2$ es el más seleccionado, aplicaremos el método de $K-$medias para este valor. En la figura 2.8 observamos las diferencias entre los grupos reales y los obtenidos mediante el algoritmo de $K-$medias para $K = 2$ grupos.

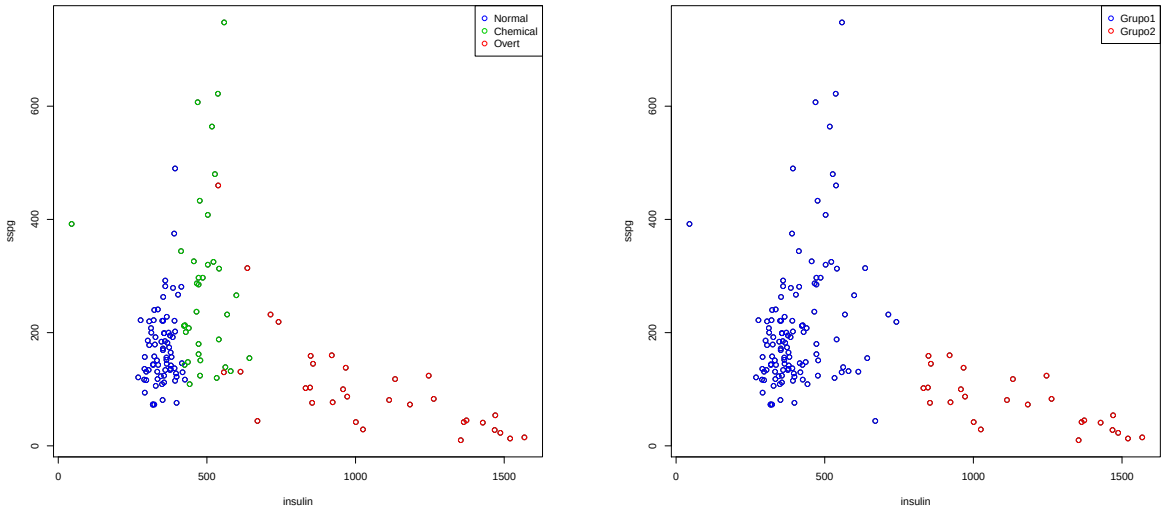

Figura 2.8: Representación gráfica de las diferencias entre los grupos reales y los grupos obtenidos con el método de $K-$medias para $K = 2$

En el cuadro 2.4 aparece la relación entre los grupos reales y los obtenidos mediante el algoritmo de $K-$medias. Observamos que el primer grupo contiene los datos de los grupos "Normal" y "Chemical", mientras que los datos de "Overt" se reparten en 7 datos para el primer grupo y 26 datos para el segundo grupo.

	1	2
Normal	76	0
Chemical	36	0
Overt	7	26

Cuadro 2.4: Relación entre los grupos reales y los grupos obtenidos mediante $K-$medias para $K = 2$

Dado que en realidad existen tres grupos en el conjunto de datos, dados por "Normal", "Chemical" y "Overt", aplicaremos también el algoritmo de $K-$medias para tres grupos. En la figura 2.9 se muestran las diferencias entre los grupos obtenidos y los grupos reales.

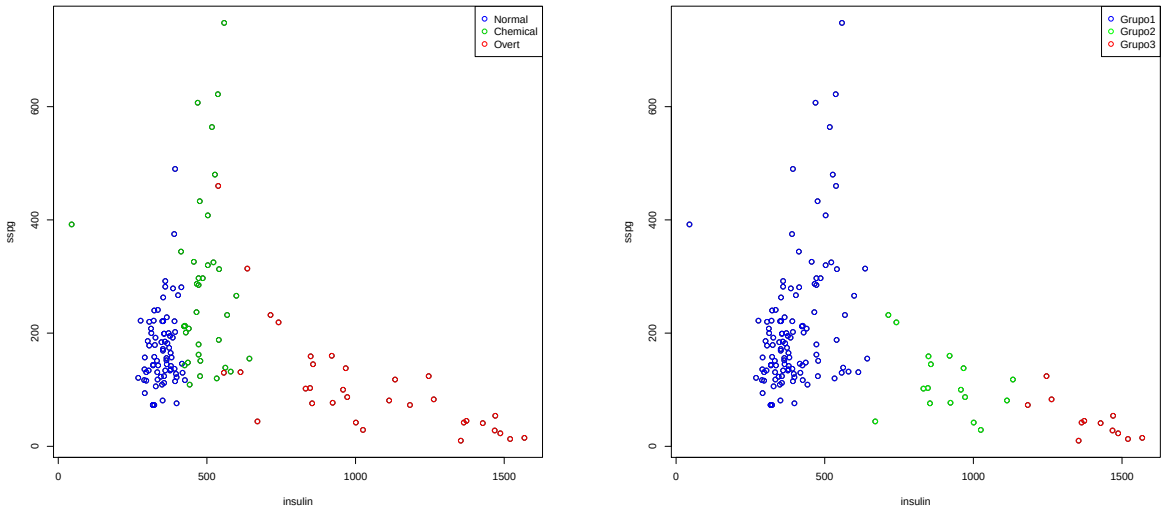

Figura 2.9: Representación gráfica de las diferencias entre los grupos reales y los grupos obtenidos mediante el método de $K-$medias para $K = 3$

En el cuadro 2.5 aparece la relación entre los grupos reales y los grupos obtenidos mediante el algoritmo de $K-$medias.

	1	2	3
Normal	76	0	0
Chemical	36	0	0
Overt	4	17	12

Cuadro 2.5: Relación entre los grupos reales y los grupos obtenidos mediante $K-$medias para $K = 3$

Observamos que el primer grupo obtenido con el algoritmo de $K-$medias contiene los datos de los grupos "Normal" y "Chemical", y además, los grupos 2 y 3 separan los datos del grupo "Overt". Por lo tanto, en este caso, el algoritmo de $K-$medias no es capaz de clasificar correctamente el conjunto de datos en los grupos reales. Además, hemos visto al inicio de este ejemplo que los métodos de selección del número de grupos tampoco dan buenos resultados.

Este mal funcionamiento del método de $K-$medias, que es relativamente frecuente en situaciones como ésta, donde los grupos no se disponen de manera circular (o esférica) entorno a su media, justifica la aplicación de procedimientos algo más complejos del análisis cluster, como el modelo de mixturas finitas que veremos en el capítulo 3.

Capítulo 3

Modelo de mixturas finitas

En este capítulo introduciremos el método de mixturas finitas para la formación de grupos, desarrollado por [9]. Se trata de un modelo estadístico formal, que supone que la población consta de una serie de subpoblaciones (los clusters) donde en cada una de las subpoblaciones las variables tienen una función de densidad multivariante diferente, que resulta en lo que se conoce como una densidad de mixtura finita. Al usar densidades de mixtura finita, el problema de agrupamiento se convierte en un problema de estimación de los parámetros bajo el modelo asumido y cálculo posterior con esos parámetros estimados de los miembros de cada cluster. Además, el problema de determinar el número de grupos se reduce a un problema de selección de un modelo, donde existen procedimientos objetivos.

3.1. Formulación del modelo de mixturas finitas

Dada X una población de n individuos para los cuales observamos p variables continuas, y que representamos mediante la matriz de datos multivariante (1.2), consideraremos las filas de dicha matriz, es decir, las observaciones independientes $x_1, \ldots, x_n$. Supongamos que la población, es una mixtura de un número finito, digamos K, de sub-poblaciones $G_1, \ldots, G_K$, en las proporciones $\pi_1, \pi_2, \ldots, \pi_K$, respectivamente, donde:

$$\begin{array}{l} \sum_{i=1}^{K} \pi_i = 1 \\ \pi_i \geq 0 \;\; i = 1, \ldots, K. \end{array} \tag{3.1}$$

La función de densidad de una observación x en X puede ser representada en forma de mixtura finita,

$$f(x) = \sum_{i=1}^{K} \pi_i g_i(x; \theta_i), \tag{3.2}$$

donde $g_i(x; \theta)$ es la función de densidad en G_i, π_i son las proporciones de la mixtura y θ_i es el vector de parámetros desconocidos asociado a cada función de densidad g_C.

3.2. Modelo de mixturas finitas de distribuciones normales

El modelo de distribución más empleado, y el único que consideraremos, es el modelo de distribución Gaussiana. En este caso el parámetro θ_i contiene los elementos del vector de medias μ_i y los elementos de la matriz de covarianzas Σ_i para $i = 1, \ldots, K$. Además, las funciones de densidad para una variable aleatoria x, de dimensión p, tendrán la siguiente forma

$$f(x; \pi, \mu, \Sigma) = \sum_{i=1}^{K} \pi_i g_i(x; \mu_i, \Sigma_i) \tag{3.3}$$

donde $\pi = (\pi_1, \ldots, \pi_{k-1})$, las $K-1$ proporciones de mixtura independientes, son tales que

$$0 < \pi_i < 1, \;\; \pi_K = 1 - \sum_{i=1}^{K-1} \pi_i. \tag{3.4}$$

Cada función de densidad g_i viene dada por

$$g_i(x; \mu_i, \Sigma_i) = (2\pi)^{-p/2} |\Sigma_i|^{-1/2} exp\{-\frac{1}{2}(x - \mu_i)' \Sigma_i^{-1} (x - \mu_i)\}. \tag{3.5}$$

Los conjuntos de datos generados por mixturas de densidades normales multivariantes se caracterizan por tener grupos centrados en las medias μ_k, con mayor densidad en puntos cercanos a la media.

Una vez formulado el modelo, el problema de particionamiento del conjunto de datos en K grupos se reduce a estimar los parámetros bajo dicho modelo.

3.3. Estimación de máxima verosimilitud en mixturas de distribuciones normales

Explicaremos unicamente el método de estimación de máxima verosimilitud (EMV) para estimar los parámetros de una mixtura de distribuciones normales. Es uno de los métodos más utilizados, puesto que bajo condiciones muy generales obtiene estimaciones con propiedades deseables como la consistencia y la distribución asintótica normal.

Dado el conjunto de datos X con observaciones multivariantes e independientes $x_1, \ldots, x_n$, la función de máxima probabilidad para los parámetros viene dada por:

$$\mathcal{L} = \prod_{i=1}^{n} f(x_i; \pi, \mu, \Sigma) = \prod_{i=1}^{n} \sum_{k=1}^{K} \pi_k g_k(x_i; \Sigma_k, \mu_k) \tag{3.6}$$

donde $g_k(x_i; \Sigma_k, \mu_k)$ representa la función de densidad bajo el modelo en el grupo k, evaluada en la observación x_i. Para maximizar esta expresión podríamos diferenciarla e igualarla

a cero, pero antes conviene tomar el logaritmo:

$$\begin{aligned} L = & \sum_{i=1}^{n} log f(x_i; \pi, \Sigma, \mu) \\ = & \sum_{i=1}^{n} log(\sum_{k=1}^{K} \pi_k g_k(x_i; \Sigma_k, \mu_k)) \end{aligned} \tag{3.7}$$

Esta formulación es mucho más sencilla y dado que el logarítmo es una transformación monótona, L tomará su máximo en los mismos parámetros que lo hacía $\mathcal{L}$.

Las ecuaciones de máxima verosimilitud se obtienen igualando a cero las primeras derivadas parciales de (3.7) con respecto a π_k, los elementos de cada matriz Σ_k, y a aquellos de cada vector μ_k. Estas operaciones son más sencillas si dejamos que los elementos independientes de Σ_k^{-1} en vez de Σ_k sean los parámetros desconocidos. Nótese que, por simetría, solo la mitad de las celdas fuera de la diagonal de Σ_k^{-1} son independientes y que $\sum_{k=1}^{K} \pi_k = 1$. Obtenemos así las siguientes ecuaciones:

$$\begin{aligned} & \tfrac{\partial L}{\partial \pi_k} = \sum_{i=1}^{n} \tfrac{1}{f(x_i;\pi,\Sigma,\mu)} [g_k(x_i; \Sigma_k, \mu_k) - g_K(x_i; \Sigma_K, \mu_K)] = 0 \\ & k = 1, 2, \ldots, K-1 \end{aligned} \tag{3.8}$$

$$\begin{aligned} & \tfrac{\partial L}{\partial \mu_k l} = \sum_{i=1}^{n} \tfrac{\pi_k g_k(x_i;\Sigma_k,\mu_k)}{f(x_i;\pi,\Sigma,\mu)} \sum_{j=1}^{p} \sigma_k^{lj}(x_{ij} - \mu_{kj}) = 0 \\ & k = 1, 2, \ldots, K \\ & l = 1, 2, \ldots, p \end{aligned} \tag{3.9}$$

$$\begin{aligned} & \tfrac{\partial L}{\partial \sigma_k^{lj}} = \sum_{i=1}^{n} \tfrac{\pi_k g_k(x_i;\Sigma_k,\mu_k)}{f(x_i;\pi,\Sigma,\mu)} (1 - \delta_{ij}/2)[\sigma_{lj}^k - (x_{il} - \mu_{kl})(x_{ij} - \mu_{kj})] = 0 \\ & k = 1, 2, \ldots, K \\ & l, j = 1, 2, \ldots, p \end{aligned} \tag{3.10}$$

donde x_{ij} con $j = 1, \ldots, p$ son los elementos del vector x_i, μ_{kj} con $j = 1, \ldots, p$ son los elementos del vector μ_k, y σ_{ij}^k y σ_k^{ij}, con $i, j = 1, \ldots, p$, son los elementos de Σ_k y Σ_k^{-1} respectivamente. Además, δ_{ij} es la delta de Kronecker, dada por:

$$\delta_{ij} = \begin{cases} 1, & \text{si i=j,} \\ 0, & \text{si } i \neq j. \end{cases} \tag{3.11}$$

Como se supone que las componentes existen en una proporción fija en la mixtura, podemos hablar de la probabilidad de que un individuo particular de la muestra pertenezca a una de estas componentes. Si $P(s|x_i)$ es la probabilidad de que la observación x_i pertenezca a la componente s, entonces tenemos que

$$P(s|x_i) = \frac{\pi_s g_s(x_i; \Sigma_s, \mu_s)}{f(x_i; \pi, \Sigma, \mu)}. \tag{3.12}$$

Aplicando (3.12) a las soluciones de las ecuaciones (3.8), (3.9) y (3.10) en términos de π_k, μ_k y Σ_k, podemos escribir las estimaciones de los parámetros de la siguiente forma:

$$\hat{\pi}_k = \frac{1}{n}\sum_{i=1}^{n}\hat{P}(k|x_i),\ \ k = 1,\ldots,K-1 \tag{3.13}$$

$$\hat{\mu}_k = \frac{1}{n\hat{\pi}_k}\sum_{i=1}^{n}\hat{P}(k|x_i)x_i,\ \ k = 1,\ldots,K \tag{3.14}$$

$$\hat{\Sigma}_k = \frac{1}{n\hat{\pi}_k}\sum_{i=1}^{n}\hat{P}(k|x_i)(x_i-\hat{\mu}_k)(x_i-\hat{\mu}_k)',\ \ k = 1,\ldots,K. \tag{3.15}$$

Las ecuaciones (3.13), (3.14) y (3.15) no dan las estimaciones de los parámetros de forma explícita, sino que es necesario usar algún procedimiento iterativo. Explicaremos uno de los algoritmos más utilizados: el algoritmo EM.

Algoritmo EM

El algoritmo de esperanza-maximización (Dempster, Laird, y Rubin 1977, [10]; McLachlan y Krishnan 1997, [23]), también conocido como algoritmo EM, es un procedimiento iterativo para la estimación de máxima verosimilitud en problemas en los que los datos consisten en n observaciones multivariantes $y_i = (x_i, z_i)$, que indican a qué grupo pertenece cada individuo. Para el modelo de mixturas, los datos no observables vienen dados por $z_i = (z_{i1},\ldots,z_{iK})$, donde

$$z_{ik} = \begin{cases} 1, & \text{si } x_i \text{ pertenece al grupo } k, \\ 0, & \text{en otro caso.} \end{cases} \tag{3.16}$$

Si los datos z_i son independientes e idénticamente distribuidos de acuerdo con una distribución multinomial de un sorteo de K categorías con probabilidades $\pi_1,\ldots,\pi_K$, y la función de densidad de una observación x_i dado z_i viene dada por

$$\prod_{k=1}^{K} g_k(x_i;\theta_k)^{z_{ik}}, \tag{3.17}$$

entonces la probabilidad logarítmica de los datos completos es:

$$l(\theta_k,\pi_k,z_{ik};x) = \sum_{i=1}^{n}\sum_{k=1}^{K} z_{ik}log(\pi_k g_k(x_i;\theta_k)) \tag{3.18}$$

El algoritmo EM alterna dos etapas, una etapa "E" y una etapa "M". En la etapa "E" se calcula la esperanza condicional de la probabilidad logarítmica de los datos completos

dados los datos observados y las estimaciones de los parámetros actuales. Para el modelo de mixturas esta etapa viene dada por:

$$\hat{z}_{ik} = \frac{\hat{\pi}_k g_k(x_i; \hat{\theta}_k)}{\sum_{j=1}^{K} \hat{\pi}_j g_j(x_i; \hat{\theta}_j)}, \tag{3.19}$$

que consiste en calcular el valor esperado de las variables no observables z_{ik}.

En la etapa "M" se determinan los parámetros que maximizan la probabilidad logarítmica esperada del paso "E". Para el modelo de mixturas, en esta etapa se maximiza (3.18) en términos de π_k y θ_k, con z_{ik} fijado para los valores calculados en la etapa "E", $\hat{z}_{ik}$. En el caso particular de mixturas de distribuciones normales multivariantes, los estimadores obtenidos son

$$\hat{\pi}_k = \frac{n_k}{n}, \tag{3.20}$$

$$\hat{\mu}_k = \frac{\sum_{i=1}^{n} \hat{z}_{ik} x_i}{n_k}, \tag{3.21}$$

siendo $n_k = \sum_{i=1}^{n} \hat{z}_{ik}$. La estimación de la matriz de covarianzas $\hat{\Sigma}_k$ depende de su parametrización.

Este algoritmo presenta ciertas limitaciones. En primer lugar, se trata de un algoritmo que puede tener convergencia lenta. Sin embargo, cuando los datos se ajustan bien al modelo y las iteraciones comienzan en un valor adecuado, da lugar a buenos resultados. La siguiente limitación es que, para mixturas de distribuciones normales multivariantes, el algoritmo falla si la matriz de covarianzas posee una o más componentes singulares o próximas a la singularidad. Además, el algoritmo puede fallar o dar resultados incorrectos si algunos grupos poseen pocas observaciones o si las observaciones que contienen se concentran próximas a un subespacio lineal con dimensión menor que el conjunto de datos.

3.4. Características geométricas de los grupos

Las características geométricas como la forma, volumen y orientación de los grupos pueden ser determinadas mediante la matriz de varianzas-covarianzas Σ_k. Cuando esta matriz toma la forma $\Sigma_k = \lambda I$, los clusters poseen forma esférica y son del mismo tamaño. Si $\Sigma_k = \Sigma$ es constante, todos los grupos tienen la misma geometría, pero no necesariamente es esférica. En el caso de que Σ_k no tenga restricciones, los clusters pueden tener distinta geometría.

Banfield y Raftery (1993, [2]) propusieron un procedimiento general para restricciones de grupos en mixturas finitas de distribuciones normales mediante la parametrización de la matriz de varianzas-covarianzas Σ_k usando la siguiente descomposición de autovalores

$$\Sigma_k = \lambda_k D_k A_k D_k', \tag{3.22}$$

donde A_k es una matriz diagonal cuyos elementos son proporcionales a los autovalores, λ_k es la constante asociada a dicha proporcionalidad y D_k es una matriz ortogonal de autovectores. Cuando los parámetros independientes A_k, λ_k y D_k son fijados, los grupos comparten ciertas características geométricas; A_k posee la forma de la $k-$ésima componente de la mixtura, D_k la orientación y λ_k su volumen.

Dependiendo de la descomposición de la matriz Σ_k surgieron distintos modelos, cuyas características pueden observarse en el cuadro 3.1.

Modelo	Σ_k	Distribución	Volumen	Forma	Orientación
EII	λI	Esférico	Igual	Igual	
VII	$\lambda_k I$	Esférico	Variable	Igual	
EEI	λA	Diagonal	Igual	Igual	Ejes coordenados
VEI	$\lambda_k A$	Diagonal	Variable	Igual	Ejes coordenados
EVI	λA_k	Diagonal	Igual	Variable	Ejes coordenados
VVI	$\lambda_k A_k$	Diagonal	Variable	Variable	Ejes coordenados
EEE	$\lambda DAD'$	Elipsoidal	Igual	Igual	Igual
EVE	$\lambda DA_kD'$	Elipsoidal	Igual	Variable	Igual
VEE	$\lambda_k DAD'$	Elipsoidal	Variable	Igual	Igual
VVE	$\lambda_k DA_kD'$	Elipsoidal	Variable	Variable	Igual

Cuadro 3.1: Características geométricas de los grupos

La distribución "esférica" significa que las variables están incorreladas y tienen la misma varianza, "diagonal" significa que están incorreladas pero pueden tener distinta varianza y "elipsoidal" significa que pueden tener cualquier tanto correlación como varianzas distintas. El resto de columnas del cuadro se refieren a si los grupos se suponen iguales en volumen (iguales varianzas en el caso "esférico") o en forma.

Estos modelos serán importantes a la hora de implementar el algoritmo del modelo de mixturas en R.

3.5. Ilustración del método de mixturas finitas para el caso real

En esta sección ilustraremos el método de mixturas finitas de distribuciones normales en datos reales. Utilizaremos los mismos datos que los considerados en la sección 2.6.2. Al igual que en el ejemplo de $K-$medias comenzaremos considerando que el número de grupos es $K = 2$.

Para aplicar el método de mixturas utilizaremos la función "Mclust" del paquete "MClust" de R. En esta función tendremos que especificar el conjunto de datos, el número de grupos y el modelo que usaremos, que será alguno de los expuestos en el cuadro 3.1. Para elegir el modelo que mejor se ajuste a nuestros datos utilizaremos la función "mclustModel". En este caso, el modelo que mejor se ajusta a los datos de diabetes es "VVE", es decir, un modelo con grupos con distribución elipsoidal, volumen y forma variables y misma orientación.

Aplicando el método de mixturas a nuestros conjunto de datos obtuvimos los grupos representados en la figura 3.1.

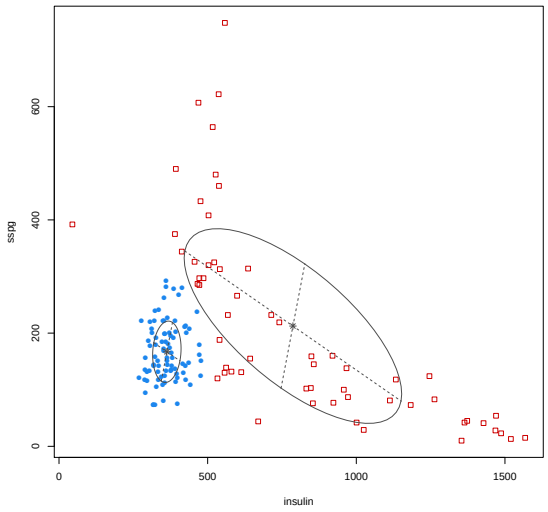

Figura 3.1: Representación gráfica de los grupos obtenidos y sus distribuciones

Además, en la figura 3.2 se pueden observar las diferencias entre los grupos reales y los grupos obtenidos mediantes mixturas.

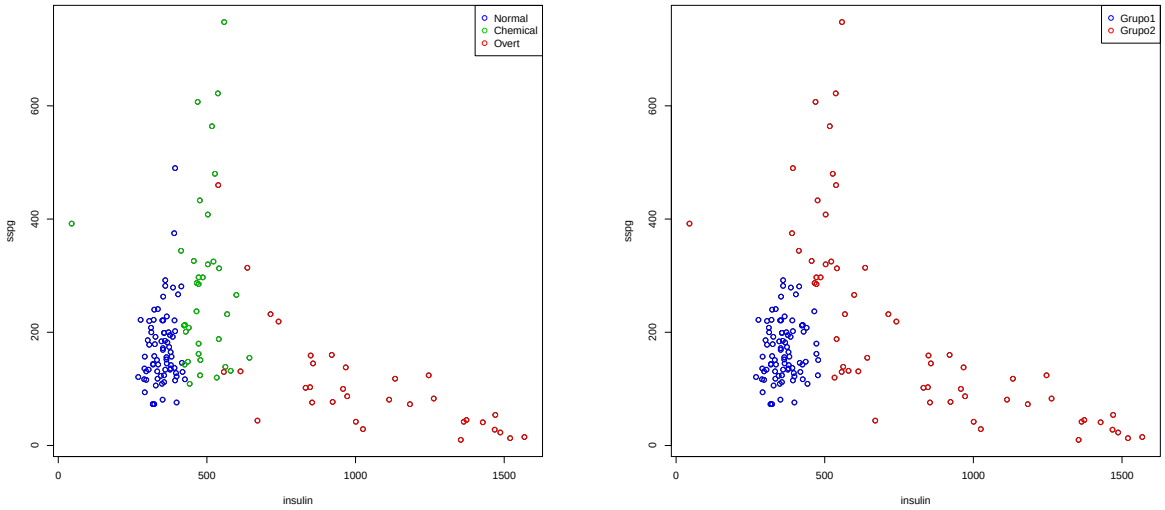

Figura 3.2: Representación gráfica de las diferencias entre los grupos reales y los grupos obtenidos con el método mixturas finitas de distribuciones normales con el modelo "VVE" para $K = 2$

En el cuadro 3.2 se puede observar que el grupo "Normal" pertenece mayoritariamente al primer grupo, "Chemical" tiene datos en ambos grupos y "Overt" pertenece al segundo grupo.

	1	2
Normal	74	2
Chemical	12	24
Overt	0	33

Cuadro 3.2: Relación entre los grupos reales y los grupos obtenidos mediante el método de mixturas finitas con el modelo "VVE"

Dado que realmente los datos poseen tres grupos: "Normal", "Chemical" y "Overt", consideraremos el modelo de mixturas con $K = 3$. El modelo que mejor se ajusta a los datos es "VVE", al igual que en el caso anterior. En la figura 3.3 aparecen representados los grupos.

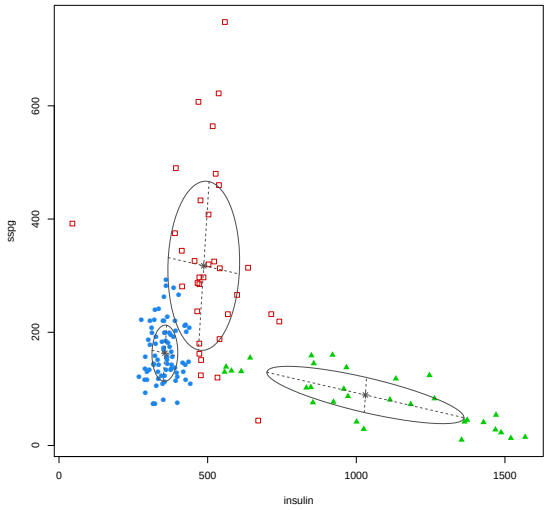

Figura 3.3: Representación gráfica de los grupos con sus distribuciones

En la figura 3.4 se pueden observar las diferencias entre los grupos reales y los grupos obtenidos mediante el método de mixturas.

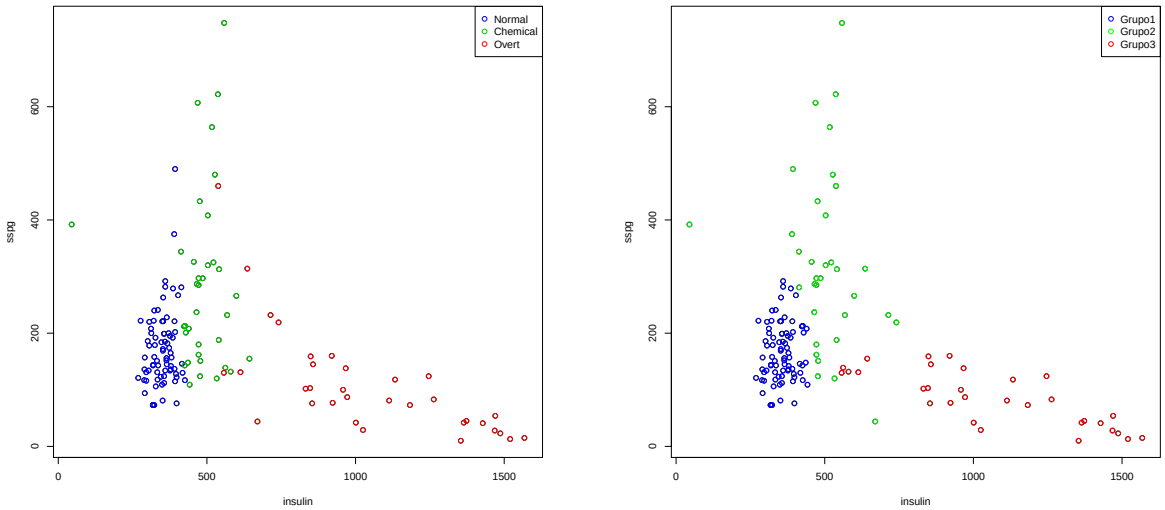

Figura 3.4: Representación gráfica de las diferencias entre los grupos reales y los grupos obtenidos con el método de mixturas finitas de distribuciones normales con el modelo "VVE" para $K = 3$

En el cuadro 3.3 se observan tres grupos diferenciados, el grupo "Normal" está asociado al grupo 1, "Chemical" al grupo 2, y "Overt" al grupo 3.

	1	2	3
Normal	73	3	0
Chemical	7	26	3
Overt	0	5	28

Cuadro 3.3: Relación entre los grupos reales y los grupos obtenidos mediante el método de mixturas finitas con el modelo "VVE"

De los modelos explicados anteriormente éste claramente es el que mejores resultados ofrece.

Capítulo 4

Simulaciones

En este capítulo realizaremos un análisis crítico de los métodos vistos en los anteriores capítulos. En primer lugar realizaremos diferentes simulaciones de distintas poblaciones de individuos para conocer la bondad de los distintos métodos de selección de grupos. A continuación, compararemos el algoritmo de $K-$medias y el algoritmo de mixturas finitas mediante dos simulaciones de individuos de distintas poblaciones. Por último, veremos la importancia de la elección del modelo correcto para el método de mixturas finitas.

4.1. Comparación de los métodos para la selección de grupos

En esta sección realizaremos una comparación de los métodos para determinar el número de grupos explicados en la sección 2.5.

En primer lugar comenzaremos simulando 100 repeticiones de $n = 40$ datos de dos poblaciones con distribuciones normales de vectores de medias $\mu_1 = (8, 10)$ y $\mu_2 = (11, 12)$, y matrices de varianzas-covarianzas $\Sigma_1 = \begin{pmatrix} 4 & 2 \\ 2 & 4 \end{pmatrix}$ y $\Sigma_2 = \begin{pmatrix} 6 & 3 \\ 3 & 6 \end{pmatrix}$.

Los resultados se muestran en el cuadro 4.1.

		Número de grupos						
	$n = 40$	2	3	4	5	6	7	8
	Calinski y Harabasz	100						
	Gamma							100
Método	Duda y Hart	100						
	Beale	100						
	Gap	100						

Cuadro 4.1: Resultados del número de grupos obtenidos para cada uno de los métodos simulando 40 datos de dos poblaciones

Como podemos observar, todos los métodos obtienen un resultado de $K = 2$ grupos, mientras que el método Gamma obtiene $K = 8$ grupos.

Además simularemos 100 repeticiones de $n = 100$ datos de las poblaciones anteriores. Los resultados se muestran en el cuadro 4.2.

		Número de grupos						
	$n = 100$	2	3	4	5	6	7	8
	Calinski y Harabasz							100
	Gamma							100
Método	Duda y Hart	100						
	Beale	100						
	Gap	100						

Cuadro 4.2: Resultados del número de grupos obtenidos para cada uno de los métodos simulando 100 datos de dos poblaciones

Los resultados son idénticos a la tabla anterior, sin embargo, para el índice de Calinski y Harabasz el resultado es $K = 8$ grupos.

A continuación simularemos 100 repeticiones de $n = 60$ datos de tres poblaciones con distribuciones normales de vectores de medias $\mu_1 = (8, 10)$, $\mu_2 = (11, 16)$ y $\mu_3 = (16, 9)$, y matrices de varianzas-covarianzas $\Sigma_1 = \begin{pmatrix} 4 & 2 \\ 2 & 4 \end{pmatrix}$, $\Sigma_2 = \begin{pmatrix} 6 & 3 \\ 3 & 6 \end{pmatrix}$ y $\Sigma_3 = \begin{pmatrix} 4 & 1 \\ 1 & 4 \end{pmatrix}$.

Los resultados de la selección del número de grupos se muestra en el cuadro 4.3.

		Número de grupos						
	$n = 60$	2	3	4	5	6	7	8
	Calinski y Harabasz		100					
	Gamma		99	1				
Método	Duda y Hart	1	99					
	Beale	100						
	Gap		100					

Cuadro 4.3: Resultados del número de grupos obtenidos para cada uno de los métodos simulando 60 datos de tres poblaciones

Como podemos observar, todos los métodos llegan al resultado correcto de $K = 3$ grupos, a excepción del método de Beale, que toma $K = 2$ grupos. Además, simularemos 100 repeticiones de $n = 150$ individuos de las mismas poblaciones. Los resultados se muestran en el cuadro 4.4.

		Número de grupos						
	n=150	2	3	4	5	6	7	8
	Calinski y Harabasz		100					
	Gamma			100				
Método	Duda y Hart		100					
	Beale	1	99					
	Gap	1	99					

Cuadro 4.4: Resultados del número de grupos obtenidos para cada uno de los métodos simulando 150 datos de tres poblaciones

Ahora simularemos 100 repeticiones de 80 datos de cuatro poblaciones con distribuciones normales de vectores de medias $\mu_1 = (8, 10)$, $\mu_2 = (11, 16)$, $\mu_3 = (16, 9)$ y $\mu_4 = (6, 14)$, y matrices de varianzas-covarianzas $\Sigma_1 = \begin{pmatrix} 4 & 2 \\ 2 & 4 \end{pmatrix}$, $\Sigma_2 = \begin{pmatrix} 6 & 3 \\ 3 & 6 \end{pmatrix}$, $\Sigma_3 = \begin{pmatrix} 4 & 1 \\ 1 & 4 \end{pmatrix}$ y $\Sigma_4 = \begin{pmatrix} 6 & 2 \\ 2 & 6 \end{pmatrix}$. Los resultados obtenidos se muestran en el cuadro 4.5.

		Número de grupos						
	$n = 80$	2	3	4	5	6	7	8
	Calinski y Harabasz			100				
	Gamma							100
Método	Duda y Hart	1	99					
	Beale	1	99					
	Gap	100						

Cuadro 4.5: Resultados del número de grupos obtenidos para cada uno de los métodos simulando 80 datos de cuatro poblaciones

Como podemos observar, tan sólo el índice de Calinski y Harabasz llega al resultado correcto de $K = 4$ grupos.

Por último, simularemos 100 repeticiones de $n = 200$ datos de las poblaciones anteriores. Los resultados se muestran en el cuadro 4.6.

		Número de grupos						
	$n = 200$	2	3	4	5	6	7	8
	Calinski y Harabasz			50	50			
	Gamma						50	50
Método	Duda y Hart	100						
	Beale	100						
	Gap	100						

Cuadro 4.6: Resultados del número de grupos obtenidos para cada uno de los métodos simulando 200 datos de cuatro poblaciones

En el cuadro anterior podemos ver que tan solo el índice de Calinski y Harabasz llega al valor correcto de $K = 4$ grupos la mitad de las veces.

4.2. Comparación de K-medias y Mixturas

En esta sección realizaremos una comparación del algoritmo de $K-$medias con el algoritmo de mixturas finitas. En primer lugar veremos un ejemplo para el cual el algoritmo de $K-$medias funciona peor que el algoritmo de Mixturas finitas. Para ello simularemos 1000 repeticiones de $n = 40$ datos de dos poblaciones de individuos con distribuciones normales de vectores de medias $\mu_1 = (10, 10)$ y $\mu_2 = (15, 10)$, y matrices de varianzas-covarianzas $\Sigma_1 = \Sigma_2 = \begin{pmatrix} 1 & 0 \\ 0 & 16 \end{pmatrix}$.

En la figura 4.1 se muestran las curvas de nivel de las distribuciones de las poblaciones. Para este ejemplo las poblaciones parece que poseen distribuciones elípticas, diagonales y con el mismo volumen, forma y orientación.

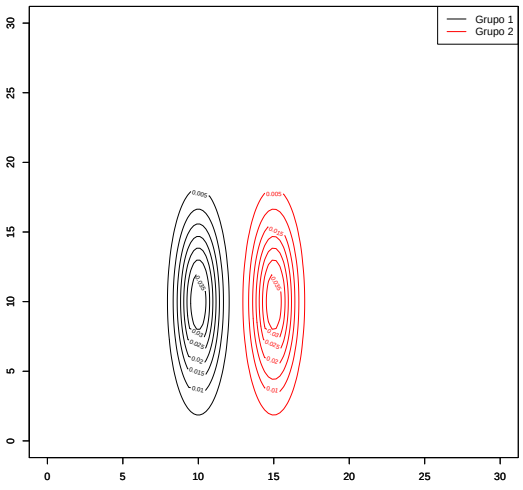

Figura 4.1: Representación gráfica de las curvas de nivel de las distribuciones de los dos grupos

En primer lugar aplicaremos el algoritmo de $K-$medias para este conjunto de datos. Dado que sabemos de antemano el número de grupos, seleccionaremos $K = 2$ grupos.

El cuadro 4.7 muestra los resultados de la agrupación de los datos mediante el método de $K-$medias. El porcentaje de acierto es del 63'9425 %.

		Grupo asignado	
		1	2
Grupo real	1	12727	7273
	2	7150	12850

Cuadro 4.7: Resultados de la asignación de los grupos por el método de $K-$medias para $K = 2$

Para el método de mixturas usaremos dos modelos distintos. El primero de ellos será "EEE", es decir, aquel con distribución elipsoidal y volumen, forma y orientación iguales. Los resultados para este modelo aparecen en el cuadro 4.8. El porcentaje de acierto es del 89'6725 %.

		Grupo asignado	
		1	2
Grupo	1	17977	2023
real	2	2108	17892

Cuadro 4.8: Resultados de la asignación de grupos por el método de las mixturas con el modelo "EEE"

El segundo será el modelo "EEI", con distribución diagonal y volumen y forma iguales. Los resultados para este modelo se muestran en el cuadro 4.9. El porcentaje de acierto es del 90'0375 %.

		Grupo asignado	
		1	2
Grupo	1	18074	1926
real	2	2059	17941

Cuadro 4.9: Resultados de la asignación de grupos por el método de las mixturas con el modelo "EEI"

Por lo tanto, en este caso el método de mixturas agrupa considerablemente mejor el conjunto de datos.

A continuación veremos un ejemplo en el cual el método de $K-$medias obtiene mejores resultados que el método de Mixturas finitas. Para ello simularemos 1000 repeticiones de 40 datos de dos poblaciones de individuos con distribuciones normales de vectores de medias $\mu_1 = (10, 10)$ y $\mu_2 = (17, 27)$, y matrices de varianzas-covarianzas $\Sigma_1 = \Sigma_2 = \begin{pmatrix} 16 & 0 \\ 0 & 16 \end{pmatrix}$.

Para este ejemplo las curvas de nivel de las distribuciones de los grupos poseen forma esférica e idéntico tamaño, como podemos observar en la figura 4.2.

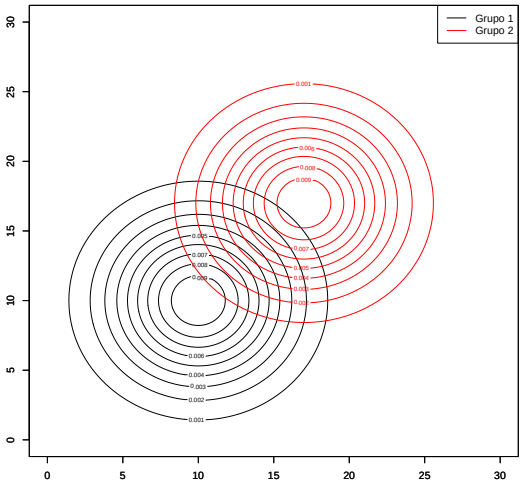

Figura 4.2: Representación gráfica de las curvas de nivel de las distribuciones de los grupos

Los resultados para el método de $K-$medias con $K = 2$ se muestran en la tabla 4.10. En este caso el porcentaje de acierto es del 88'265 %.

		Grupo asignado	
		1	2
Grupo	1	17710	2290
real	2	2404	17596

Cuadro 4.10: Resultados de la asignación de grupos por el método de $K-$medias para $K = 2$

Para el método de mixturas seleccionamos el modelo "EII", es decir, distribución esférica y forma y volumen iguales. En este caso el porcentaje de acierto es menor que para el método de $K-$medias, con un valor de 74'875 %. Los resultados se muestran en el cuadro 4.11.

		Grupo asignado	
		1	2
Grupo	1	15184	4816
real	2	5234	14766

Cuadro 4.11: Resultados de la asignación de grupos por el método de las mixturas con el modelo "EII"

Por tanto, cuando las poblaciones tienen distribución esférica y tamaños similares, el método de $K-$medias tiende a agrupar mejor los datos que con el método de las mixturas. Sin embargo, cuando estas condiciones no se cumplen, es más recomendable emplear el método de las mixturas con un modelo que se adapte a las circunstancias del problema en cuestión.

4.3. Comparación de distintos modelos en el método de mixturas finitas

En esta sección veremos un ejemplo en el cual se muestra la importancia de la elección correcta del modelo para el método de mixturas finitas. Para ello simularemos 800 repeticiones de 40 datos de dos poblaciones de individuos con distribuciones normales de vectores de medias $\mu_1 = (10, 10)$ y $\mu_2 = (15, 10)$, y matrices de varianzas-covarianzas $\Sigma_1 = \Sigma_2 = \begin{pmatrix} 1 & 2 \\ 2 & 16 \end{pmatrix}$.

En la figura 4.3 se muestran las curvas de nivel de las distribuciones de los grupos. Para este ejemplo los grupos poseen distribuciones elípticas, no diagonales y con el mismo volumen, forma y orientación.

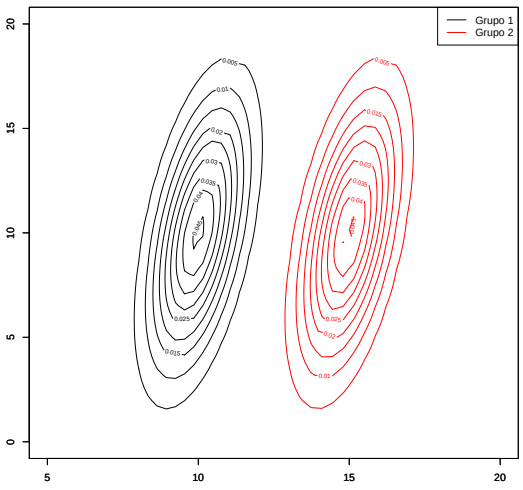

Figura 4.3: Representación gráfica de las curvas de nivel de las distribuciones de los grupos

Los resultados obtenidos para los distintos modelos del método de mixturas se muestran en el cuadro 4.12.

Modelo	Porcentaje de acierto
VII	61'84717 %
EEI	91'8201 %
VEI	88'71841 %
EVI	91'79904 %
VVI	89'1787 %
EEE	92'41877 %
EVE	91'94043 %
VEE	89'64501 %
VVE	89'62696 %

Cuadro 4.12: Resultados de la asignación de los grupos por el método de las mixturas con distintos modelos

Podemos observar que el modelo "VII" posee el peor porcentaje de acierto, con un 61'84717 %. Esto se debe a que es el único que considera que las poblaciones siguen distribuciones esféricas. Además, los modelos "EEI", "EVI", "EEE" y "EVE" poseen un porcentaje de acierto mayor al 90 % debido a que en todos ellos el volumen no es variable.

El modelo con mejores resultados es el "EEE", pues supone una distribución elipsoidal, no esférica ni diagonal, con igual volumen, forma y orientación, que son precisamente las condiciones en las que se generaron las muestras simuladas.

Bibliografía

[1] Baker, F.B., y Hubert, L.J., Measuring the Power of Hierarchical Cluster Analysis, *Journal of the American Statistical Association*, Vol. 70, pp. 31-38, (1975)

[2] Banfield, J.D., y Raftery, A.E., Model-Based Gaussian and Non-Gaussian Clustering, *Biometrics*, Vol. 49, pp. 803-821, (1993)

[3] V. Barnett y T.Lewis, Outliers in statistical data, *International Journal of Forecasting*, (1994)

[4] Beale, Euclidean cluster analysis, *Bulletin of the International Statistical Institute*, (1969)

[5] Calinski, T. and Harabasz, J., A dendrite method for cluster analysis, *Communications in Statistics*, (1974)

[6] Cerioli, A., A new method for detecting influential observations in nonhierarchical cluster analysis, *Springer*, (1998)

[7] Cheng, R., y Milligan, G. W., $K-$means clustering methods with influence detection, *Educational and Psychological Measurement*, Vol. 56, pp. 833-838, (1996a)

[8] Cheng, R., y Milligan, G. W., Measuring the influence of individual data points in cluster analysis, *Journal of Classification*, Vol. 13, pp. 315-335, (1996b)

[9] Chris Fraley y Adrian E. Raftery, Model-Based Clustering, Discriminant Analysis, and Density Estimation, *Journal of the American Statistical Association*, Vol. 97, pp. 611-631, (2002)

[10] Dempster, A.P., Laird, N.M., & Rubin, D.B., Maximum Likelihood from Incomplete Data via the EM Algorithm, *Journal of the Royal Statistical Society*, Vol. 39, pp. 1-38, (1977)

[11] Dimitriadou, E., Dolnicar, S., & Weingessel, A., An examination of indexes for determining the number of clusters in binary data sets, *Psychometrika*, Vol. 67, pp. 137-160, (2002)

[12] Duda, R.O., y Hart, P.E., Pattern classification and scene analysis, *Wiley*, (1973)

[13] Everitt, B.S. y Hand. D.J., Finite Mixture Distributions, *Chapman and Hall*, (1981)

[14] Everitt, B.S. et al., Cluster Analysis 5th Edition, *Wiley*, (2011)

[15] Forgy, E.W., Cluster analysis of multivariate data: efficiency vs interpretability of classifications, *Biometrics*, Vol. 21, pp. 768-780, (1965)

[16] Gersho, A., y Gray, R.M., Vector quantization and signal compression, *Boston: Kluwer Academic*, (1992)

[17] Goodman, L.A. and Kruskal, W.H., Measures of association for cross-clasifications, *Journal of the American Statistical Association*, Vol. 49, pp. 732-764, (1954)

[18] Gordon A.D. , Classification, 2nd edition, *Chapman and Hall- CRC.*, (1999)

[19] Hartigang, J.A. y Wong, M.A. , A K-means clustering algorithm, *Applied Statistics*, Vol. 28, pp. 100-108, (1979)

[20] Kaufman, L., y Rousseeuw, P. Finding groups in data: An introduction to cluster analysis, *New York: Wiley*, (1990)

[21] Lloyd, S.P., Least squares quantization in PCM, *IEEE Transactions on Information Theory*, Vol. 28, pp. 129-137, (1982)

[22] MacQueen, J., Some methods for classification and analysis of multivariate observations, *Proceedings of the Fifth Berkeley Symposium on Mathematical Statistics and Probability*, Vol. 1, pp. 281-297, (1967)

[23] McLachlan, G., y Krishnan, T., The EM algorithm and extensions, *Wiley series in probability and statistics*, (1997)

[24] Milligan, G.W., An examination of the effect of six types of error perturbation on fifteen clustering algorithms, *Psychometrika*, Vol. 45, pp. 325-342, (1980)

[25] Milligan, G.W. y Cooper, M.C., An examination of procedures for determining the number of clusters in a data set, *Psychometrika*, Vol. 50, pp. 159-179, (1985)

[26] Milligan, G.W. y Cooper, M.C., A study of standardization of variables in cluster analysis, *Journal of Classification*, Vol. 5, pp. 181-204, (1988)

[27] Milligan, G.W. and Richard Gheng, Measuring the ingluence of individual data points in a cluster analysis, *Journal of Classification*, Vol. 13, pp. 315-335, (1996)

[28] Milligan, G.W., Clustering validation: Results and implications for applied analyses, *Clustering and Classification*, (1996)

[29] Späth, H., Cluster analysis algorithms for data reduction and classification of objects, *Wiley*, (1980)

[30] Steinley D., K-means clustering: A half-century synthesis, *The British Psychological Society*, (2006)

[31] Steinley D., Initializing K-means Batch Clustering: A Critical Evaluation of Several Techniques, *Journal of Classification*, Vol. 24, pp. 99-121, (2007)

[32] Tibshirani, R., Walther, G., y Hastie, T., Estimating the number of clusters in a data set via the gap statistic, *Journal of the Royal Statistical Society*, Vol. 63, pp. 411-423, (2001)

[33] Weisstein, E.W., CRC concise encyclopedia of mathematics, *Chapman & Hall/CRC.*, (2003)