ESCUELA POLITÉCNICA NACIONAL

FACULTAD DE CIENCIAS

Aplicación De Mezclas Finitas Para Segmentación De Clientes De Una Entidad Bancaria (CRM).

TRABAJO DE TITULACIÓN PREVIO A LA OBTENCIÓN DEL TÍTULO DE INGENIERA MATEMÁTICA

PAULINA ALEXANDRA DÍAZ ORDÓÑEZ
paulipollet@yahoo.es

DIRECTOR: MÉNTHOR OSWALDO URVINA MAYORGA
menthor.urvina@epn.edu.ec

Quito, Diciembre 2018

DECLARACIÓN

Yo, Paulina Alexandra Díaz Ordóñez, declaro bajo juramento que el trabajo aquí descrito es de mi autoría; que no ha sido previamente presentada para ningún grado o calificación profesional; y que he consultado las referencias bibliográficas que se incluyen en este documento.

La Escuela Politécnica Nacional puede hacer uso de los derechos correspondientes a este trabajo, según lo establecido por la Ley de Propiedad Intelectual, por su Reglamento y por la normatividad institucional vigente.

_____

PAULINA ALEXANDRA DÍAZ ORDÓÑEZ

CERTIFICACIÓN

Certifico que el presente trabajo fue desarrollado por PAULINA ALEXANDRA DÍAZ ORDÓÑEZ, bajo mi supervisión.

MÉNTHOR OSWALDO URVINA
DIRECTOR

AGRADECIMIENTO

Mis sinceros agradecimientos:

A Dios, por estar presente en cada momento de mi vida y darme la fuerza para alcanzar mis metas.

A mis padres Paco Díaz y Olga Ordóñez, a mis hermanos Carlos, Gaby y Francisco y mi abuelita Laura Flor por su apoyo incondicional.

A mi director Ménthor Urvina por su invaluable ayuda en este trabajo.

A mi querida Politécnica Nacional en donde no solo obtuve conocimientos técnicos sino tuve la suerte de contar con grandes docentes.

A mis amigos Diana y Raúl por su constante apoyo y aliento en cada uno de mis proyectos de vida.

A mis pequeñas Doménica, Abigail y Paula por alegrar cada día de mi vida y llenarme de paz y amor.

A mi esposo Daniel por ser un compañero de vida y por todo el apoyo brindando.

DEDICATORIA

A mis padres por su ejemplo y constante apoyo impulsándome a ser mejor cada día y a mi familia mi esposo y mis tres pequeñas que son el motor y fuerza motivadora en mi vida.

Paulina.

ÍNDICE DE CONTENIDO

LISTA DE FIGURAS

LISTA DE TABLAS

LISTA DE ANEXOS

RESUMEN

Este trabajo busca dar una alternativa de clasificación en conjunto de variables categóricas y continuas, dónde las variables continuas están modeladas dentro de las componentes de la mezcla por distribuciones gaussianas y las variables categóricas dentro de las componentes por distribuciones multinomiales independientes. El método estadístico que se utiliza eses modelos de mezclas finitas. Esta metodología se fundamenta en la estimación de probabilidades condicionales, lo que permite, analizar variables medidas en diferentes métricas. Para la estimación de las distribuciones mixtas se utiliza el algoritmo de Máxima Expectación (EM) que está compuesto por dos pasos alternados iterativamente que involucran una esperanza y una maximización na vez el algoritmo EM alcanza la convergencia, se tiene a los individuos clasificados en grupos homogéneos sobre los cuales se puede focalizar campañas en diferentes ámbitos por parte de una entidad bancaria cuyo objetivo es ser más eficiente y asertivos en la relación con el cliente CRM. Para identificar el número adecuado de grupos se consideran los criterios de información bayesiana BIC y de Akaike AIC; finalmente para desarrollar las estrategias de negocio se utiliza un score comportamental basado en la información a priori proporcionada por el conocimiento experto del área comercial de la entidad con lo cual se afinan las estrategias a proponerse para cada uno de los grupos.

Palabras clave: Clasificación, CRM, AIC, BIC, Mezclas Gaussianas, Esperanza, Maximización, Convergencia

ABSTRACT

This paper seeks to provide an alternative for classifying categorical and continuous variables, where the continuous variables are modeled within the components of the mixture by Gaussian distributions and the categorical variables within the components by independent multinomial distributions. The statistical method used is by means of a statistical method using is finite mixture models. This methodology is based on the estimation of conditional probabilities, which allows analyzing variables measured in different metrics. For the estimation of the mixed distributions, the Maximun Expectation (MS) is used. This algorithm that is composed of two iterative steps that involve a hope and a maximization. Once the algorithm reaches convergence, individuals are classified in homogeneous groups on which campaigns can be focused in different areas by a bank whose objective is to be more efficient and assertive in the relationship with the CRM client, to identify the appropriate number of groups are considered the criteria Bayesian Information Criteria BIC and Akaike AIC finally to develop the businesss strategies a behavioral score is used based on the a priori information provided by the expert knowledge of the commercial area with which the strategies to be proposed for each of the groups are refined.

Keywords: Classification, CRM, AIC, BIC, Gaussian Mixes, Hope, Maximization, Convergence

CAPÍTULO I

1. Introducción

Como resultado de la automatización de procesos de carga de información de datos estructurados, semiestructurados y no estructurados de diferentes fuentes; para el tratamiento de información y el desarrollo de tecnología en el almacenamiento de datos, el crecimiento de estos se ha incrementado exponencialmente. Toda esta información requiere ser analizada y explorada para extraer conocimiento, patrones, comportamientos y comprensión de fenómenos en cualquier entorno (médico, ambiental, económico, etc).

Existe muy poca literatura y software sobre la resolución de estos problemas que son importantes para la clasificación donde los componentes con variables no numéricas.

Por tanto, es relevante la metodología que permita la clasificación de variables no solo cuantitativas sino cualitativas, que permitan segmentar la población.

La clasificación de variables permite a la entidad toma de decisiones importantes para definir estrategias de marketing diferenciado, atención al cliente, gestiones de cobranza y seguimiento de los clientes mediante indicadores basados en la información que proporciona cada segmento.

Este tipo de clasificación ha ido evolucionando en diferentes campos de la investigación, ya que a partir de una serie de datos generados por una mezcla de G distribuciones normales y el uso de algunos métodos de agrupamiento, se puede particionar una muestra heterogénea en grupos más homogéneos.

En un problema de reconocimiento de patrones el objetivo es obtener una función de decisión que permita la clasificación entre diferentes clases. Un modelo de mezclas gaussianas es un modelo de distribución de probabilidad de mezclas finitas que ha resultado exitoso en el reconocimiento de patrones [1]. Ante esta problemática para la clasificación se incluyen esquemas generativos, donde el

clasificador aprende las densidades por clase, y los esquemas discriminativos, los cuales se enfocan en el aprendizaje de los límites de las clases.

La estimación de máxima verosimilitud (MLE) es el método más utilizando para el diseño de clasificadores, esta estimación ofrece un método simple de estimación de las densidades de probabilidad por clase y consecuentemente representar la distribución de probabilidad de clases relacionadas. MLE aproxima el problema del diseño del clasificador desde una perspectiva indirecta, no optimiza directamente el desempeño de los datos a ser clasificados.

La clasificación es uno de los fenómenos de mayor interés en la ciencia ya que este debe ser ordenado para ser entendido. A nivel multivariado se tienen técnicas dirigidas a la clasificación supervisada y no supervisada de manera paralela se ha desarrollado técnicas de mezclas de distribuciones.

En este trabajo se presentan las mezclas finitas de distribuciones normales, las mismas que suponen que la muestra a ser clasificada está dividida en G grupos o componentes, a cada uno se le asigna un número determinado de elementos, una función de distribución normal y un peso de ponderación. Para resolver el problema a través de mezclas finitas de distribuciones normales se utiliza todos los datos muestrales en la estimación de los parámetros para cada grupo, los mismos que son estimados por el método de máxima verosimilitud.

Como método para las mezclas se estudia el algoritmo iterativo EM para estimar los parámetros de cada grupo en el cual se amplían las variables observadas (Y), introduciendo las no observadas (Z) cuya función es indicar la componente de la mezcla de la que proviene cada uno.

El algoritmo EM está compuesto de dos pasos alternados que consisten en una esperanza y una maximización. Se inicia con una estimación previa de los parámetros, luego se halla la esperanza de la función de soporte $L(Y, Z)$ condicionada a los parámetros y a la distribución de (Z) y finaliza con la maximización de esta esperanza para encontrar los parámetros. Se indican los valores de los parámetros cuando los valores de los parámetros convergen a un valor fijo.

El score comportamental recoge la información a priori del conocimiento de negocio así como el comportamiento del cliente en la entidad basado en su información transaccional.

1.1. Reseña Histórica

En la actualidad es ampliamente útil para el reconocimiento de patrones el uso de modelos de mezclas finitas en particular las basadas en núcleos gaussianos como una potente herramienta probabilística para modelar datos en una o más dimensiones.

Los modelos de mezclas finitas permiten representar las observaciones de forma normal, asumiendo que han sido generadas por una fuente perteneciente a un conjunto de posibles fuentes aleatorias, aunque se desconoce inicialmente la fuente que generó cada uno de los datos.

Las distribuciones mixtas son utilizadas para la modelización de datos heterogéneos en situaciones experimentales, los mismos que pueden interpretarse como procedentes de dos o más subpoblaciones (componentes), al obtener estas componentes se tiene la estimación de los parámetros de la mixtura. Esta estimación se inicia con Pearson (1894) posteriormente utilizan esta aproximación Charlier (1906), Charlier y Wicksell (1924), Cohen (1967), y Than y Chang (1972). Más tarde utilizan la estimación de máxima verosimilitud Rao (1948) y Hasselblad (1966, 1969).

Aplicaciones prácticas y un análisis estadístico detallado de las mixturas finitas que abarcaron diferentes métodos de estimación fueron presentados por Everitt y Hand (1981), y Titterington (1985) Descripciones más generales fueron publicadas por McLachlan y Basford (1988), McLachlan y Jones (1988), McLachlan y Krishnan (1997), y McLachlan y Peel (2000). Algunas aplicaciones en un contexto médico fueron presentadas por Delmar (2005) en bioinformática y genetica, Schlattmann (2009) y Frühwirth-Schnatter (2010). El trabajo reciente de Mengerse et al. (2011) muestra la relevancia de los modelos mixtos considerando un esquema bayesiano.

1.2. Conceptos Estadísticos

1.2.1. Descomposición Espectral

En el álgebra matricial se conoce que para una matriz cuadrada y simétrica real, de orden p los valores propios son números reales y los vectores propios son ortogonales, entonces para esta matriz A puede escribirse como:

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

donde D es la matriz diagonal formada por los valores propios de A y U es una matriz ortogonal cuyas columnas son los vectores propios asociados con los elementos de la diagonal de la matriz D. Esta propiedad es la que se denomina descomposición espectral.

Sean $\lambda 1, \lambda 2, \lambda 3, \ldots \ldots \lambda p$, los valores propios de la matriz A y $\mu 1, \mu 2, \mu 3, \ldots \mu p$ sus correspondientes vectores propios, la descomposición de (1.1) puede notarse como:

$$A = \sum_{i=1}^{p} \lambda_i \mu_i \mu_i' \tag{1.2}$$

que descompone la matriz A como la suma de p matrices de rango uno, $u_i u_{i'}$ con coeficientes λi.

La descomposición espectral de A^{-1} es

$$A^{-1} = \sum_{i=1}^{p} \lambda_i^{-1} u_i u_i'$$

ya que A'^1 tiene los mismos vectores propios de A y valores propios λi^{-1}

1.2.2. Formas Cuadráticas

Sea A una matriz simétrica de tamaño (pxp) y v un vector de tamaño $(px1)$, la función [2]

$$Q(x) = x'Ax$$

se denomina forma cuadrática de x.

$Q(x)$ es un escalar que se puede expresar mediante la ecuación

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

siendo a_{ij} elementos de A y xi, xj elementos de x.

Si $Q(x) > 0$ para todo $x \neq 0$, se dice que A es definida positiva y se denota $A > 0$

Si $Q(x) \geq 0$ para todo $x \neq 0$, se dice que A es semidefinida positiva y se denota $A \geq 0$

Para las funciones cuadráticas existen algunas propiedades [2]:

1. Si A> 0 todos sus valores propios λ_1, λ_2, λ_3,……. λ_p son mayores a cero. Si $A \geq 0$,entonces

   $\lambda_i > 0$ para i=1,2,....,p y $\lambda_i = 0$ para algún i.

2. Si A> 0, A es no singular entonces $|A| > 0$
3. Si A > 0, entonces $A^{-1} > 0$.
4. Si A> 0 y C es un matriz (rxr), C'AC > 0.

1.2.3. Partición de una Matriz

En ocasiones es más práctico escribir una matriz en forma de submatrices, es decir que los elementos que la conforman sean matrices de menor dimensión que la original. Una matriz se puede expresar por:

$$A = \begin{pmatrix} A_{11} & A_{1j} & A_{1p} \\ \vdots & \vdots & \vdots \\ A_{i1} & A_{ij} & A_{ip} \\ & \vdots & \\ & A_{nj} & \end{pmatrix} y\ B = \begin{pmatrix} B_{11} & B_{1j} & B_{1p} \\ \vdots & \vdots & \vdots \\ B_{i1} & B_{ij} & B_{ip} \\ & \vdots & \\ & B_{nj} & \end{pmatrix}$$

Donde la submatriz A_{ij} es de dimensión $(n_i \times p_j)$ con $\sum_{i=1}^{n} n_i = n$ y $\sum_{j=1}^{p} p_j = p$.

Las operaciones suma y productos para este tipo de matrices se conforman de manera similar que con las matrices habituales, de esta forma si las matrices A y B se particionan similarmente entonces.

$$A+B=\begin{pmatrix} A_{11}+B_{11} \dots A_{1j}+B_{1j} \dots & A_{1p}+B_{1p} \\ A_{i1}+B_{i1} \dots A_{ij}+B_{ij} \dots & A_{ip}+B_{ip} \\ A_{n1}+B_{n1} \dots A_{nj}+B_{nj} \dots & A_{np}+B_{np} \end{pmatrix}$$

A manera de ejemplo se particiona A en la siguiente forma:

$$A=\begin{pmatrix} A_{11} & A_{12} \\ A_{21} & A_{22} \end{pmatrix}$$

Donde A_{11} y A_{22} son matrices cuadradas no singulares, la inversa de A se obtiene:

$$A^{-1}=\begin{pmatrix} B^{-1} & -B^{-1}A_{12}A_{22}^{-1} \\ -A_{22}^{-1}A_{21}B^{-1} & A_{22}^{-1}+A_{22}^{-1}A_{21}B^{-1}A_{12}A_{22}^{-1} \end{pmatrix}$$

El determinante de la matriz A puede calcularse en base a su partición, para lo cual las submatrices sean no singulares.

dónde

$$B=(A_{11}-A_{12}A_{22}^{-1}A_{21})$$

1.2.4. Derivadas Matriciales

Definición 1.1 Sea f una función que asigna a un vector $x \in \mathbb{R}^p$ un número real, esquemáticamente esto es [2]:

$$f: \mathbb{R}^p \longrightarrow \mathbb{R}$$

$$x=\begin{pmatrix} x_1 \\ \vdots \\ x_p \end{pmatrix} \rightarrow f(x).$$

Se define la derivada de $f(x)$ con respecto al vector x de tamaño $(px1)$ como

$$\frac{\partial f(x)}{\partial (x)}=\begin{bmatrix} \frac{\partial f}{\partial x_1} \\ \vdots \\ \frac{\partial f}{\partial x_p} \end{bmatrix}=\left[\frac{\partial f(x)}{\partial x_i}\right] \qquad \text{Definición (1.1)}$$

Los siguientes resultados son consecuencia de la definición 1.1.

1. Si $f(x) = a´x = a_1x_1 + \cdots + a_px_p$ donde a y x son vectores de $\mathbb{R}^p$, la derivada de la función f respecto al vector x, de acuerdo a la ecuación (1.1), está dada por:

$$\frac{\partial f(x)}{\partial (x)} = \frac{\partial}{\partial x}(a´x) = \frac{\partial}{\partial x}(x´a) = \begin{bmatrix} \frac{\partial f}{\partial x_1} \\ \vdots \\ \frac{\partial f}{\partial x_p} \end{bmatrix} = \begin{pmatrix} a_1 \\ \vdots \\ a_p \end{pmatrix} = a$$

2. Sea $Q(x) = x´Ax$, donde A es cuadrada y simétrica. La derivada de la función Q respecto al vector x es:

$$\frac{\partial Q(x)}{\partial (x)} = \frac{\partial}{\partial x}(x\acute{A}x)$$

$$= \frac{\partial}{\partial x}\left(\sum_{i=1}^{p} a_{ii}x_i^2 + 2\sum_{i=1}^{p}\sum_{j=i+1}^{p} a_{ij}x_ix_j\right),$$

Luego se tiene que:

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

$$\vdots$$

$$\frac{\partial}{\partial x_p}(x´Ax) = 2a_{p1}x_1 + \cdots + 2a_{pp}x_p = 2a_p´x$$

Donde $a_i´$ es la i −esima fila de la matriz. Por lo tanto

$$\frac{\partial}{\partial x}(x´Ax) = \begin{bmatrix} 2a_1´x \\ \vdots \\ 2a_p´x \end{bmatrix} = 2Ax.$$

Definición 1.2 Sea f una función que asigna a una matriz $Y \in (\mathbb{R}^n \times \mathbb{R}^p)$ un número real, esquemáticamente esto es [2]:

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

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

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

$$\frac{\partial f(Y)}{\partial Y} = \frac{\partial f}{\partial Y}\begin{pmatrix} y_{11} & \cdots & y_{1p} \\ \vdots & \ddots & \vdots \\ y_{n1} & \cdots & y_{np} \end{pmatrix} = \begin{pmatrix} \frac{\partial f}{y_{11}} & \cdots & \frac{\partial f}{y_{1p}} \\ \vdots & \ddots & \vdots \\ \frac{\partial f}{y_{n1}} & \cdots & \frac{\partial f}{y_{np}} \end{pmatrix} = \frac{\partial f(Y)}{\partial y_{ij}}. \qquad (1.2)$$

Los siguientes resultados son consecuencia de la definición 1.2.

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

$$\frac{\partial}{\partial Y}(a'Yb) = ba' \qquad (1.3)$$

Sea $f(Y) = a'Y'Yb$, entonces la derivada de f viene dada por:

$$\frac{\partial}{\partial Y}(a'Y'Yb) = (ab' + ba')Y' \qquad (1.4)$$

Definición 1.3. Dado un vector y cuyos componentes son funciones f_i de un vector de variables $x' = (x_1, \dots, x_p)$, se define la derivada de y respecto a x como la matriz cuyas columnas son las derivadas de las componentes f_i respecto a x [2]. Es decir, si

$$y' = (f_1(x), \dots, f_p(x)),$$

Entonces

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

A esta matriz de derivadas se le denomina $matriz\ jacobiana$.

CAPÍTULO II

2. Modelos de Mezclas Finitas

En el contexto estadístico se utiliza un modelo de mezclas como un modelo probabilístico para representar la presencia de subpoblaciones dentro de una misma población. Se puede definir también como una distribución de mezcla que representa la distribución de probabilidad de alguna observación en la población general. Los modelos de mezcla son usados para crear inferencias, aproximaciones y predicciones acerca de las propiedades de las subpoblaciones a partir de las observaciones o datos adquiridos de la población estudiada.

2.1. Modelos de Mezclas Gaussianas

El objetivo del modelo de mezcla Gaussiana es encontrar una aproximación o estimación a partir de sus componentes, encontrando un acomodamiento de los datos que contienen las componentes como se muestra en el la siguiente figura:

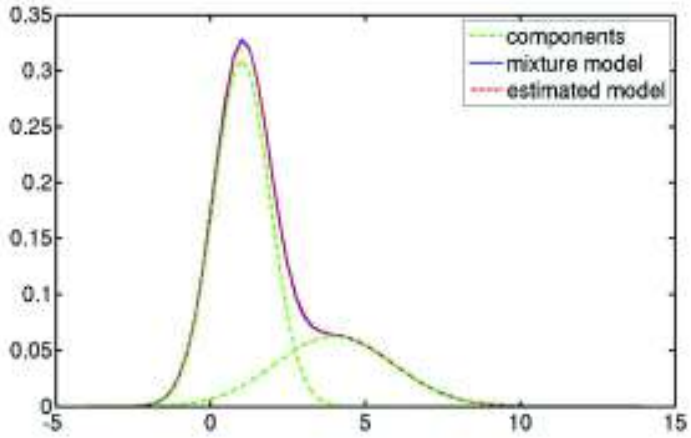


Figura 1.1 – Modelo de mezclas gaussianas

No es suficiente explicar las distribuciones de unos datos mediante una única distribución estadística, si los datos pueden agruparse en subpoblaciones o asociados a distintos procesos generadores. Es necesaria la utilización de una composición de distribuciones, las mismas que se suelen describir mediante

modelos de mixturas, los cuales se definen por los parámetros de cada componente y las proporciones en las que cada una de ellas contribuye a la distribución general. El conjunto de parámetros que definen a estos modelos pueden ser estimados mediante muchas técnicas como: métodos gráficos, método de momentos máxima verosimilitud, aproximaciones bayesianas, entre otras.

El algoritmo de Esperanza-Maximización (EM) es una herramienta iterativa para la estimación de máxima verosimilitud de las distribuciones mixtas. El principio de esta herramienta es introducir una variable indicadora multinomial que identifica la pertenencia a un clúster de cada observación del conjunto de datos.

Los modelos de mezcla finita pertenecen a los modelos de clase latente, son conocidos por su flexibilidad e importante ventajas. En estos modelos se parte de una variable aleatoria que se asume extraída de una población que es una mezcla aditiva de subpoblaciones distintas en proporciones $\pi_1, \dots, \pi_c$ donde:

$$\sum_{j=1}^{c} \pi_j = 1, \qquad \pi_j \geq 0, \quad j = 1, \dots, C.$$

La función de probabilidad de la mezcla viene dada por:

$$f(y_i/\theta) = \sum_{j=1}^{C} \pi_j f(y_i/\theta_j) \qquad i = 1, \dots, n\,,$$

siendo $f(y_i/\theta_j)$ la función de probabilidad de cada subpoblación $j(j = 1, \dots, C)$. Estas distribuciones podrían pertenecer a familias paramétricas distintas, aunque normalmente se supone que pertenecen a la misma (por ejemplo, Poisson, Binomial negativa, etc) [3]. Por lo general, los π_j son parámetros desconocidos que deben ser estimados conjuntamente con el resto de parámetros del modelo, θ, además $\pi_C = 1 - \sum_{j=1}^{C-1} \pi_j$. Estos π_j podrían ser parametrizados como función de un conjunto de variables explicativas [3].

La mezcla finita genera una representación natural e intuitivamente atractiva de la heterogeneidad en un número finito, usualmente pequeño de clases latentes, cada una se la considera como un tipo o grupo.

El análisis de clases latentes (ACL) fue reportado por primera vez por Lazarsfeld (1950) como herramienta para construir una tipología en el análisis de un conjunto de variables dicotómicas. Leo Goodman (1974) logró que los modelos de clases latentes pudieran aplicarse en una mayor diversidad de estudios, desarrollando un algoritmo para obtener estimaciones de máxima verosimilitud. Él propuso la extensión del modelo para variables manifiestas politómicas y realizó importantes mejoras para la identificación de los modelos [4].

Diversos investigadores (Agresti, 2002; Bartholomew, 2002; Hagenaars, 1990; McCutcheon, 1987; Vermunt, 2003) resaltaron las bondades los modelos de clases latentes.

- Reducen la complejidad de los datos identificando un número pequeño de variables (clases latentes)
- Explican relaciones “verdaderas” entre variables observadas, ya que al incorporarlas controlan fuentes de error como casos ausentes, correlaciones entre las observaciones, etc.
- Permiten estimar la probabilidad que tiene cada uno de los participantes de pertenecer a una de las clases latentes.
- Analizan datos categóricos en las escalas que fueron medidos, sin requerir transformaciones para lograr normalidad multivariada.

El ACL, se base en un modelo probabilístico, utilizado cuando se tienen variables nominales, ordinales, continuas o conteos, a diferencia de otros modelos estadísticos incorpora variables discretas no observadas para explicar la relación entre variables observadas, sin basarse en supuestos tradicionales del modelado (distribución normal, relaciones lineales y homogeneidad de varianzas)

La probabilidad de las clases latentes $\pi(X_t)$ describe la distribución de los niveles detectados en una variable no observada, a través de las cuales las variables observadas son independientes. Dos aspectos importantes en las probabilidades de las clases latentes son: el número de clases y el tamaño de las clases.

Los parámetros de los modelos de clases latentes representan las probabilidades de que un individuo obtenga un valor determinado en una variable, dada su pertenencia a una clase latente, en cada clase, las probabilidades condicionales de

las variables observadas deben sumar 1, por lo tanto cada observación tiene una probabilidad específica de estar en un nivel de la variable observada [4].

Los parámetros de modelos de clases latentes se estiman por el método de máxima verosimilitud (ML), es decir, que la solución en valores de parámetros que maximizan la función de probabilidad y su logaritmo natural, para aproximarse a los estimadores más verosímiles se utiliza el algoritmo de maximización de la esperanza (EM), este se inicializa con valores arbitrarios de parámetros.

El planteamiento del modelo de mezclas gaussianas es el siguiente:

Sea $\boldsymbol{Y} = [Y_1, Y_2, \dots, Y_D]^T$ una variable aleatoria real D-dimensional. Se dice que la distribución de $\boldsymbol{Y}$ sigue una distribución mezcla finita si su función densidad de probabilidad (fdp) se puede escribir como una combinación lineal de fdp´s elementales

$$p(y/\theta) = \sum_{i=1}^{I} \alpha_i p(y/C = i, \beta_i), \qquad i \in \{1, \dots, I\} \qquad (2.1)$$

Donde I representa el número de distribuciones elementales (componentes) de la mezcla, $C = 1,2, \dots, I, y\ \theta$ representa el conjunto de parámetros

$$\theta = \{\alpha_1, \dots, \alpha_I, \beta_1, \dots, \beta_I\}$$

Siendo $\beta = \{\beta_1, \dots, \beta_I\}$ el conjunto de parámetros asociados a cada distribución de la mezcla y $\alpha = \{\alpha_1, \dots, \alpha_I\}$ la probabilidad o peso de cada distribución de la mezcla.

Cuando las distribuciones que componen la mezcla son Gaussianas, la función densidad de probabilidad (2.1) se conoce como mezcla gaussiana. Por tanto una mezcla Gaussiana es una distribución probabilística cuya fdp es una combinación lineal de distribuciones Gaussianas

$$p(y/\theta) = \sum_{i=1}^{I} \alpha_i N(y/C = i, \beta_i)$$

Siendo

$$N(y/C = i, \beta_i) = \frac{1}{(2\pi)^{\frac{D}{2}} |\Sigma_i|^{\frac{1}{2}}} \exp\left(-\frac{1}{2}(y - \mu_i)^T \sum_i^{-1} (y - \mu_i)\right)$$

Y los parámetros de cada Gaussiana

$$\beta_i = \{\mu_i, \Sigma_i\}$$

Donde $\mu_i \in \mathbb{R}^D$ y $\Sigma_i \in \mathbb{R}^{DxD}$ son la media y la matriz de covarianza de la componente i-ésima [1], [2], [3], [4].

Los parámetros α y $\sum_i$ presentan una serie de restricciones a considerarse. Las probabilidades de las mezclas, deben verificar:

$$\alpha_i \geq 0, \quad i \in \{1, \dots, I\}, \qquad \sum_{i=1}^{I} \alpha_i = 1$$

Y las matrices de covarianzas deben cumplir las siguientes restricciones

$$\sum_i = \sum_i^T$$

$$\sigma_{jj}^2 \geq 0, \qquad j \in \{1, \dots, D\}$$

Las matrices de covarianzas deben ser simétricas, además en los elementos de la diagonal se encuentran las varianzas y deben ser no negativas.

2.2. Función de Densidad de Probabilidad de Mezclas Gaussianas

En análisis multivariado la mayor parte de los procedimientos existentes de deducción han sido desarrollados bajo la suposición de normalidad y en modelos lineales el vector de error es frecuentemente normalmente distribuido. Si no existe conocimiento a priori de una función de densidad de probabilidad del fenómeno en estudio, se puede usar solo un modelo general y la distribución Gaussiana es una buena opción.

Para un modelo de reconocimiento de patrones lo que se busca es una función de decisión con la cual realizar la clasificación entre las diferentes clases, una manera de hacerlo es representar cada clase como una función de densidad de probabilidad (PDF) [1].

Aunque la aplicación de las PDFs Gaussianas multivariantes ha sido exitosa para representar características y discriminar diferentes clases en muchos problemas, la suposición de una sola componente conducen a requerimientos estrictos para las características de un fenómeno: una sola clase básica la cual varía ligeramente alrededor de la media. La suposición de unimodalidad para características

distribuidas multimodalmente podría causar un error en la estimación de la PDF y, como consecuencia, en la discriminación entre clases. Para una variable cuyos valores son generados por fuentes independientes que ocurren de forma aleatoria en lugar de una sola fuente, se puede usar un modelo de mezclas finitas para aproximar las (PDFs), si la forma gaussiana es suficiente para la fuentes individuales entonces se puede usar un modelo de mezclas gaussianas.

$$p(x/C_i) = \sum_{k=1}^{M} \omega_{ik} * p[x_t|\mu_{ik}, \sum_{ik}] \qquad 2.1$$

donde ω_{ik} para $k = 1,2, \dots, M$ son los pesos de mezclas que representan la probabilidad de cada distribución Gaussiana, M son las distribuciones Gaussianas en total y debe cumplir $\sum_{k=1}^{M} \omega_{ik} = 1$. La función densidad de probabilidad $p[x_t|\mu_{ik}, \sum_{ik}]$ es la distribución Gaussiana D-dimensional definida en (2.1).

2.3. Formulación Paramétrica de un Modelo de Mezclas

Las distribuciones mixtas son utilizadas para la modelización de datos heterogéneos en algunas situaciones experimentales, estas pueden interpretarse como procedentes de dos o más subpoblaciones, a las que se las denomina componentes. Para obtenerlas es necesario la estimación de parámetros de mixtura.

La estimación se remonta a Pearson (1894) [5], quien inicio el trabajo con la mixtura de dos componentes con varianzas iguales utilizando el método de momentos, hasta el trabajo más reciente de Mergense en el 2011, muestra la relevancia de los modelos mixtos considerando un esquema bayesiano [5].

El desarrollo conceptual del algoritmo EM, que se describió anteriormente, necesita de una formulación paramétrica para la representación del modelo, considerando la notación de McLachlan y Peel (2000) [5].

Para denotar la muestra aleatoria de tamaño n, se considerará $Y = (Y_1, Y_2, \dots, Y_n)$.

Donde Y_j es un vector aleatorio q-dimensional con función de densidad de probabilidad $f(y_j)$ en $\mathbb{R}^q$. Así, $y = (y_1, y_2, \dots, y_n)$ representan una muestra observada o realización de Y, donde y_j constituye un valor observado aleatorio Y_j.

Definición 2.1.- La distribución de una variable aleatoria Y_j cuya función de densidad se escribe [5]

$$f(y_j|\Psi) = \sum_{i=1}^{g} \pi_i f_i(y_j|\theta_i), \quad y_j \in \mathbb{R}^q,$$

Se denomina distribución de mixtura finita de g componentes, con un vector de parámetros del modelo.

$$\Psi = (\pi_1, \dots, \pi_g, \theta_1, \dots, \theta_g).$$

Así, $f_i(y_j|\theta_i)_{i=1,\dots,g}$ denotan las $densidades\ de\ las\ componentes$ de la mixtura con parámetros θ_i, y $\pi_1, \dots, \pi_g$, las $proporciones\ o\ pesos$. Mediante la notación $f_i(.|\theta_i)$, se asume que estas densidades pueden pertenecer a diferentes familias paramétricas, agrupándose cada uno de sus vectores paramétricos en θ_i.

Las proporciones de la mixtura representan las probabilidades de que la realización y_j de la variable aleatoria haya sido generada por las g diferentes densidades y, como probabilidades que son, están sujetas a las restricciones:

$$0 \leq \pi_i \leq 1 \quad i = 1, \dots, g$$

$$\text{Y} \qquad \sum_{i=1}^{g} \pi_i = 1,$$

Por lo que uno de los pesos resulta redundante. En la figura 2.1 se muestra un ejemplo de mixturas de distribuciones gaussianas con dos componentes y diferentes parametrizaciones, creadas a partir de muestras sintéticas generadas mediante una misma semilla aleatoria.

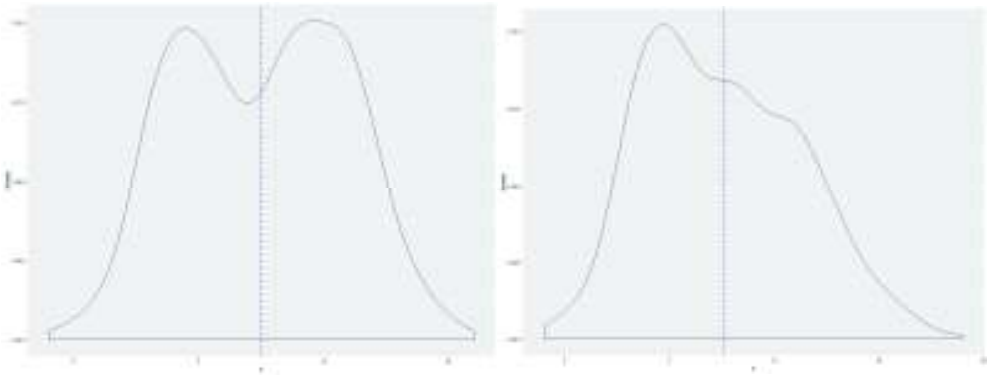


Figura 2.1 – Mixturas de distribuciones gaussianas con dos componentes

Para el ejemplo las densidades de las componentes son del tipo gaussiana univariantes heterocedásticas por lo que podrían representarse mediante $f(y_j|\theta_i)$, con $\theta_i = (\mu_i, \sigma_i)$, o bien $\emptyset(y_j|\mu_i, \sigma_i)$. Cuya función de densidad se describe a continuación:

Defínción 2.2: Se dice que una variable aleatoria Y sigue una distribución normal o gaussiana si su función de densidad puede describirse como [5]

$$f(y|\mu, \sigma^2) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{(y-\mu)2}{2\sigma^2}\right) \qquad -\infty < y < \infty$$

$con\ -\infty < \mu < \infty,\ \sigma^2 > 0 \quad y \quad \theta = (\mu, \sigma^2)$ los parámetros de la distribución.

2.3.1. Vector de Medias

Las mezclas de distribución es una técnica de modelado estadístico con la que se obtiene una estimación de la función de densidad de probabilidad de datos en una muestra aleatoria, la cual se representa como una suma ponderada finita de las componentes de densidad multivariables.

Entre las características principales de las mezclas de distribución se observa que no requieren un parámetro de suavizado de la función. La utilidad de esta técnica radica en el análisis y modelado de conglomerados. Las muestras de distribución multivariables se expresan así:

$$f(x) = \sum_{g=1}^{G} p_g f_g(\boldsymbol{x}, \boldsymbol{\theta})$$

Las mezclas de distribución sobre una muestra poblacional X compuesta de observaciones d-dimensionales, contiene G componentes $f_g(\boldsymbol{x}, \boldsymbol{\theta})$ referentes a la densidad de distribución multivariada seleccionada.

Cada componente de densidad f_g intenta describir el comportamiento de un grupo dentro de la población en el cual los datos relativos a dicho grupo poseen características similares, establecidas en el vector estimador θ correspondiente a

los parámetros de cada distribución. Un estimador es un parámetro que define el comportamiento de los datos dentro de un grupo, y por tanto, su forma.

Para una distribución de probabilidad normal los estimadores son: el vector de medias μ y la matriz de covarianza V, que definen el punto central de la distribución y la forma como se concentran los datos alrededor de dichos puntos respectivamente. La cantidad p_g, llamada proporción de la mezcla o coeficiente de mezclado, brinda información sobre la importancia del grupo dentro de la mezcla.

Las condiciones que deben cumplir los coeficientes de mezclado están dadas por:

$$\sum_{g=1}^{G} p_g = 1 \qquad p_g > 0$$

El objetivo de la mezcla de distribuciones es identificar una cantidad desconocida de grupos en las cuales se agrupen los datos de una muestra, es decir, se busca la homogeneidad dentro de una muestra heterogénea.

2.3.2. Matriz de Varianzas y Covarianzas

La variabilidad de los datos y la información relativa a las relaciones lineales entre las variables se resumen en la matriz de varianzas y covarianzas. Esta matriz es cuadrada y simétrica de orden k, donde los términos diagonales son las varianzas y los no diagonales, las covarianzas entre las variables. Llamando S a esta matriz, se tiene que, por definición:

$$S = \begin{bmatrix} s_1^2 & s_{12} & \dots & s_{1k} \\ \vdots & \vdots & \ddots & \vdots \\ s_{k1} & s_{k2} & \dots & s_k^2 \end{bmatrix}$$

Esta matriz puede calcularse como:

$$S = \frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})(x_i - \bar{x})'.$$

La comprobación es inmediata como:

$$\begin{bmatrix} x_{i1} - \bar{x}_1 \\ \vdots \\ x_{ik} - \bar{x}_k \end{bmatrix} [x_{i1} - \bar{x}_1 \dots \bar{x}_{ik} - \bar{x}_k] = \begin{bmatrix} (x_{i1} - \bar{x}_1)^2 & \dots & (x_{i1} - \bar{x}_1)(x_{k1} - \bar{x}_k) \\ \vdots & \ddots & \vdots \\ (x_{ik} - \bar{x}_k)(x_{i1} - \bar{x}_1) & \dots & (x_{ik} - \bar{x}_k)^2 \end{bmatrix}$$

Al sumar para todos los elementos y dividir por n se obtienen las varianzas y covarianzas entre las variables. Otra manera de calcular S es a partir de la matriz de datos cuadrados X, que se obtiene restando a cada dato su media. Esta matriz puede calcularse mediante:

$$\tilde{X} = X - 1\bar{x}',$$

Y al sustituir el vector de medias por su expresión se tiene:

$$\tilde{X} = X - \frac{1}{n}11'X = PX,$$

Donde la matriz cuadrada P está definida por

$$P = I - \frac{1}{n}11'$$

Y es simétrica e idempotente por tanto la matriz S puede escribirse:

$$S = \frac{1}{n}\,\tilde{X}'\tilde{X} = \frac{1}{n}\,X'PX.$$

Nota: La matriz de covarianzas es semidefinida positiva, es decir, si y es cualquier vector $y'S'Z \geq 0$. Lo que implica que los autovalores de esta matriz λ_i son no negativos, es decir, $Sv_i = \lambda_i V_i$, por tanto $\lambda_i \geq 0$.

2.4.Estimación de Máxima Verosimilitud

Existen varios métodos de estimación como el método de los momentos, procedimientos gráficos, bayesiana, mínimo cuadrática, mínimo X^2 o máxima verosimilitud. Tan y Chang (1972), realizaron una comparación entre el método de momentos y el de máxima verosimilitud, demostrando que el segundo es mejor, a esta misma conclusión llegaron Holgerson y Jorner (1978) [5].

La evaluación por el método de máxima verosimilitud procura encontrar los valores más probables de los parámetros de la distribución para un conjunto de datos. Maximizando el valor de lo que se conoce como la “función de verosimilitud”, la cual

se basa en la función de la densidad de la probabilidad (fdp) para una distribución dada.

Definición 2.3: Sea $y = (y_1, y_2, \dots, y_n)$ observaciones independientes de una variable aleatoria Y con función de densidad $f(y/\theta)$, donde θ es el vector de parámetros n desconocidos que se quiere estimar; la función de densidad conjunta de y se escribe como [5]:

$$f(y|\theta) = \prod_{j=1}^{n} f\left(y_j|\theta\right) = L(\theta|y)$$

donde $L(\theta|y)$ representa la función de verosimilitud y se considera una función de θ.

Definición 2.4: Un estimador de máxima verosimilitud $\theta y \theta$ es un valor que maximiza $L(\theta/y)$, es decir,

$$\hat{\theta} = \arg\max_{\theta \in \theta} L(\theta).$$

Donde θ representa el espacio paramétrico.

Bajo esta definición se tiene la posibilidad de tener más de un estimador de máxima verosimilitud, como puede ocurrir en aproximaciones experimentales donde se detectan múltiples máximos.

Definición 2.5: Las ecuaciones normales o de verosimilitud vienen dadas por

$$S(y|\theta) = \frac{\partial \ell(\theta|y)}{\partial \theta_j} = 0, \qquad j = 1, \dots, k.$$

En el supuesto de que $\theta = (\theta_1, \theta_2, \dots, \theta_k)$ sea un parámetro k dimensional.

Definición 2.6: Sea $I(\theta|y) = \frac{\partial^2}{\partial \theta \partial \theta^T} \ell(\theta|y)$ las derivadas segundas parciales negativas de la función de log-verosimilitud con respecto a θ, donde el θ^T denota el vector transpuesto de θ. Entonces, $I(\theta|y)$ se denota la matriz de información observada. La matriz de información esperada viene dada, bajo condiciones de regularidad, por [5]

$$\chi(\theta|y) = E[S(y|\theta)S^T(y|\theta)] = E[I(\theta|y)].$$

La solución es compleja y requiere procedimientos iterativos para su resolución, los métodos más utilizados son el método de Newton – Rapson y scoring de Fisher. Otra técnica estándar para el cálculo de los EMV es el algoritmo EM, la idea de esta técnica es resolver problemas de datos incompletos de cierta complejidad abordando de forma repetida una situación de datos completos de fácil resolución.

Se asigna a cada dato observado una etiqueta que indica su pertenencia a una u otra subpoblación en la muestra de partida, asumiendo que estas existen. Es así que la asignación de un conjunto de datos a una y otra componente de la mixtura conduce de forma natural al agrupamiento de estos o a la formación de clústers.

El valor de estas etiquetas o indicadores de pertenencia a un grupo es, a priori, desconocido. Bajo una distribución mixta, la estimación de su valor puede considerarse como un problema de datos faltantes, y el algoritmo EM puede utilizarse. A diferencia del método de scoring de Fisher, el EM no requiere del cálculo de una matriz Hessiana en cada iteración, esta matriz puede ser aproximada numéricamente mediante simulación empleando métodos de Montecarlo, aunque el costo computacional puede ser alto si es necesario numerosas iteraciones para su obtención.

Los estimadores de máxima verosimilitud deben cumplir las siguientes propiedades:

INVARIANZA:

Si θMV es el estimador máximo verosímil de θ, entonces $h(\theta$MV) es estimador máximo verosímil de $h(\theta)$.

CONSISTENCIA:

Bajo condiciones generales θMV es un estimador consistente de θ, es decir, que mientras el tamaño de la muestra aumenta, las estimaciones convergen a los valores correctos.

$$lim_{n\to\infty} E[\theta_{nMV}] = \theta.$$

INSESGADEZ ASINTÓTICA

Un estimador $\hat{\theta}$ es insesgado para estimar el parámetro θ si

$$E\left[\hat{\theta}\right]= \theta,\ \theta \in \Omega$$

NORMALIDAD ASINTÓTICA

Bajo ciertas condiciones generales donde:

$$\sqrt{n}(\hat{\theta}_{MV} - \theta) \underset{\sim}{\overset{A}{}} N\left(0, \sqrt{i(\theta)^{-1}}\right)$$

$$i(\theta) = E\left[\left(\frac{\partial}{\partial\theta}\ln f(X;\theta)\right)^2\right]$$

es la cantidad de información de Fisher correspondiente a una observación.

La cantidad de información de Fisher correspondiente a n observaciones es

$$I(\theta) = E\left[\left(\frac{\partial}{\partial\theta}\ln f(X_1,\dots,X_n;\theta)\right)^2\right] \underset{=}{\overset{m.a.s.}{}} n * E\left[\left(\frac{\partial}{\partial\theta}\ln f(X;\theta)\right)^2\right]$$

$$= \mathrm{n.\,i(\theta)}$$

Se tiene que

$$\mathrm{I(\theta)} = \mathrm{E}\left[\left(\frac{\partial}{\partial\theta}\ln f(X_1,\dots,X_n;\theta)\right)^2\right] = -E\left[\frac{\partial^2}{\partial\theta^2}\ln f(X_1,\dots,X_n;\theta)\right]$$

La varianza asintótica de $\ominus\ MV$ es

$$Var\left[\hat{\theta}_{MV}\right] \underset{=}{\overset{A}{}} \frac{1}{n * i(\theta)} = \frac{1}{I(\theta)} = -\frac{1}{E\left[\frac{\partial}{\partial\theta}\ln f(X_1,\dots,X_n;\theta)\right]}$$

$$\approx -\frac{1}{\frac{\partial^2}{\partial\theta^2}\ell(\theta)|_{\theta=\hat{\theta}_{MV}}}$$

2.4.1. Verosimilitud, Soporte y Score para Mezclas

Definición 2.7: Sea $X_1 \dots X_n$ una muestra de datos independientes que pueden estratificarse en G grupos, de forma que existe n_1 observaciones del grupo 1, n_2 observaciones del grupo 2,…n_G del grupo G. El vector $X_i = \left(X_{i1}, \dots, X_{ip}\right)'$ representa un individuo particular para $i = i, \dots n$ [2].

Se define la función de verosimilitud para la mezcla, $I(\theta)$ como [2]:

$$l(\theta) = \prod_{i=1}^{n} G(x_i|\theta) = \prod_{i=1}^{n}\left(\sum_{g=1}^{G} \pi_g f_g(x_i)\right), \qquad (2.1)$$

donde $\theta = (\theta_1, \dots, \theta_n)$ es el vector de parámetros para los G grupos, $G(x_i/\theta)$ es valor de la mezcla de densidades para el i-ésimo individuo dado el vector de parámetros θ, $\theta\pi_g$ es el peso de la mezcla para el g-ésimo grupo y $fg(x_i)$ es el valor de la densidad en el g-ésimo grupo para el i-ésimo individuo.

Se observa que $l(\theta)$ se puede escribir como la suma de G_n términos, que corresponden a todas las posibles clasificaciones de las n observaciones entre los G grupos como se muestra:

$$l(\theta) = \prod_{i=1}^{n}\left(\sum_{g=1}^{G} \pi_g f_g(x_i)\right) = \overbrace{\underbrace{\left[\sum_{g=1}^{G} \pi_g f_g(x_1)\right]}_{G-sumando}\underbrace{\left[\sum_{g=1}^{G} \pi_g f_g(x_2)\right]}_{G-sumandos} \dots \underbrace{\left[\sum_{g=1}^{G} \pi_g f_g(x_n)\right]}_{G-sumandos}}^{n-datos}.$$

Definición 2.8: Sea $l(\theta)$ la función de verosimilitud de una mezcla de G densidades, con $\theta = (\theta_1, \dots, \theta_G)'$. La función soporte para la mezcla se define como [2]:

$$L(\theta) = \sum_{i=1}^{n} \ln(G(x_i)|\theta) = \sum_{i=1}^{n} \ln\left(\sum_{i=1}^{n} \pi_g f_g\,(x_i)\right). \qquad (2.2)$$

Esta definición es la particularización de la definición soporte dada en la ecuación (2.1), puesto que:

$$L(\theta) = \ln[l(\theta)] = \ln\left[\prod_{i=1}^{n}\left(\sum_{g=1}^{G}\pi_g f_g(x_i)\right)\right] = \sum_{i=1}^{n}\ln\left(\sum_{g=1}^{G}\pi_g f_g\,(x_i)\right)$$

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

$$Z(\theta) = \sum_{i=1}^{n}\frac{\frac{\partial}{\partial\theta}[G(x_i|\theta)]}{G(x_i|\theta)} = \sum_{i=1}^{n}\frac{\frac{\partial}{\partial\theta}[\sum_{g=1}^{G}\pi_g f_g(x_i)]}{\sum_{g=1}^{G}\pi_g f_g(x_i)}. \qquad (2.3)$$

De igual forma que se planteó la función de soporte para la mezcla, esta definición es consecuencia de las ecuaciones (2.2) y (2.3), ya que:

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

$$= \frac{\partial}{\partial\theta}\left\{\sum_{i=1}^{n}\ln\left(\sum_{g=1}^{G}\pi_g f_g(x_i)\right)\right\}$$

$$= \sum_{i=1}^{n}\left\{\frac{\partial}{\partial\theta}\ln\left(\sum_{g=1}^{G}\pi_g f_g(x_i)\right)\right\}$$

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

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

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

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

Donde la componente vectorial

$$\sum_{i=1}^{n} \frac{\frac{\partial}{\partial \theta_g}[\sum_{g=1}^{G} \pi_g f_g(x_i)]}{\sum_{g=1}^{G} \pi_g f_g(x_i)}, \qquad (2.4)$$

Para $g = 1, \ldots, G$ es nuevamente un vector con dimensión determinada por θ_g. Recuérdese que para este caso $\theta_g = (\pi_g, \mu_g, V_g)$.

2.4.2. Ecuaciones MV para Mezclas de Densidades Normales.

Se asume $f_g(x)$ como una densidad normal p-variante con vector de medias μ_g y matriz de varianza V_g, por tanto $\theta = (\pi_1, \ldots., \pi_G;\ \mu_1, \ldots\ \mu_G;\ V_v, .. V_G)$, el objetivo es maximizar la función soporte lo que equivale a encontrar la solución del sistema homogéneo que produce esta función para mezclas.

Se asume que el orden de las G distribuciones estará determinado por $\pi_1 \geq \pi_2 \geq \cdots \geq \pi_G$, asi como que hay mínimo p-observaciones en cada distribución buscando un máximo local que genere un estimador consistente de los parámetros.

La función de score genera el siguiente sistema homogéneo para el g-ésimo grupo

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

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

La primera ecuación dada por (2.4), se tiene $\sum_{g=1}^{G} \pi_g = 1$; lo que lleva a maximizar $L(\theta)$ incluyendo en ella el multiplicador de Lagrange, la función que se debe maximizar es:

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

Derivado respecto a las probabilidades:

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

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

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

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

Multiplicando por π_g, con $\pi_g \neq 0$, se obtiene:

$$\frac{\partial L_\lambda(\theta)}{\partial \pi_g}=\sum_{i=1}^{n}\frac{\pi_g f_g(x_i)}{\sum_{g=1}^{G}\pi_g f_g(x_i)}-\pi_g\lambda=0, \qquad (2.5)$$

luego

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

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

Donde se detecta π_{ig} a:

$$\pi_{ig}=\frac{\pi_g f_g(x_i)}{\sum_{g=1}^{G}\pi_g f_g(x_i)}, \qquad (2.6)$$

El cual se traduce como la probabilidad $a\ posteriori$ de que la observación i se haya generado por la población g. Es claro que para cada dato se cumplirá: $\sum_{g=1}^{G}\pi_g=1$ .Así el valor de λ se obtiene, sumando (2.5) para todos los grupos, de la siguiente forma:

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

$$\sum_{g=1}^{G}\pi_g\lambda=\sum_{g=1}^{G}\sum_{i=1}^{n}\pi_{ig}$$

$$\lambda \sum_{g=1}^{G} \pi_g = \sum_{i=1}^{n} \left(\sum_{g=1}^{G} \pi_{ig} \right)$$

$$\lambda = \sum_{i=1}^{n} 1$$

$$\lambda = n.$$

Sustituyendo este valor en (2.5) se tiene la probabilidad $a\ priori$ de pertenecer al g −ésimo grupo así:

$$\hat{\pi}_g = \frac{1}{n} \sum_{i=1}^{n} \pi_{ig}$$

Para la segunda ecuación dada en (2.6) no hay restricciones por lo que se tiene

$$\frac{\partial L(\theta)}{\partial \mu_g} = \sum_{i=1}^{n} \frac{\frac{\partial}{\partial \mu_g} \left[\sum_{g=1}^{G} \pi_g f_g(x_i) \right]}{\sum_{g=1}^{G} \pi_g f_g(x_i)} = 0$$

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

$$= \sum_{i=1}^{n} \frac{\pi_g \left[\frac{\partial}{\partial \mu_g} f_g(x_i) \right]}{\sum_{g=1}^{G} \pi_g f_g(x_i)} = 0.$$

Puesto que:

$$f_g(x_i) = |V_g|^{-\frac{1}{2}} (2\pi)^{-\frac{p}{2}} exp\left\{ -\tfrac{1}{2} (x_i - \mu_g)' V_g^{-1} (x_i - \mu_g) \right\}, \tag{2.7}$$

Se tiene, según el resultado dado en (2.6), que:

$$\frac{\partial}{\partial \mu_g} \left[f_g(x_i) \right] = \frac{\partial}{\partial \mu_g} \left\{ |V_g|^{-\frac{1}{2}} (2\pi)^{-\frac{p}{2}} exp\left\{ -\frac{1}{2} (x_i - \mu_g)' V_g^{-1} (x_i - \mu_g) \right\} \right\}$$

$$= |V_g|^{-\frac{1}{2}} (2\pi)^{-\frac{p}{2}} \left\{ \frac{\partial}{\partial \mu_g} exp \underbrace{\left\{ -\frac{1}{2} (x_i - \mu_g)' V_g^{-1} (x_i - \mu_g) \right\}}_{U} \right\}$$

$$= \underbrace{\left|V_g\right|^{-\frac{1}{2}}(2\pi)^{-\frac{p}{2}}exp\{U\}}_{f_g(x_i)}\left\{\frac{\partial U}{\partial \mu_g}\right\}$$

$$= f_g(x_i)\left\{-\frac{1}{2}\left[(2)V_g^{-1}(x_i-\mu_g)\right](-1)\right\}$$

$$= f_g(x_i)\left[V_g^{-1}(x_i-\mu_g)\right]. \qquad (2.8)$$

Reemplazando (2.8) en (2.7) y sustituyendo el resultado dado en (2.6) se tiene:

$$\sum_{i=1}^{n}\frac{\pi_g\left[f_g(x_i)\right]V_g^{-1}(x_i-\mu_g)}{\sum_{g=1}^{G}\pi_g f_g(x_i)} = 0$$

$$\sum_{i=1}^{n}\frac{\pi_g f_g(x_i)}{\sum_{g=1}^{G}\pi_g f_g(x_i)}V_g^{-1}(x_i-\mu_g) = 0$$

$$\sum_{i=1}^{n}\pi_{ig}\,V_g^{-1}(x_i-\mu_g) = 0$$

$$\sum_{i=1}^{n}\frac{\pi_{ig}(x_i-\mu_g)}{V_g} = 0$$

$$\sum_{i=1}^{n}\pi_{ig}(x_i-\mu_g) = 0$$

$$\sum_{i=1}^{n}\pi_{ig}x_i - \sum_{i=1}^{n}\pi_{ig}\mu_g = 0,$$

Entonces:

$$\sum_{i=1}^{n}\pi_{ig}\mu_g = \sum_{i=1}^{n}\pi_{ig}x_i$$

$$\hat{\mu}_g = \frac{1}{\sum_{i=1}^{n}\pi_{ig}}\sum_{i=1}^{n}\pi_{ig}x_i$$

Para la tercera ecuación, (2.6) se basa en (2.7) y (2.8)

$$\frac{\partial L(\theta)}{\partial V_g} = \sum_{i=1}^{n} \frac{\frac{\partial}{\partial V_g}\left[\sum_{g=1}^{G} \pi_g f_g(x_i)\right]}{\sum_{g=1}^{G} \pi_g f_g(x_i)} = 0$$

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

$$= \sum_{i=1}^{n} \frac{\pi_g \left[\frac{\partial}{\partial V_g} f_g(x_i)\right]}{\sum_{g=1}^{G} \pi_g f_g(x_i)} = 0.$$

puesto que

$$f_g(x_i) = |V_g|^{-\frac{1}{2}} (2\pi)^{-\frac{p}{2}} exp\left\{-\frac{1}{2}(x_i - \mu_g)' V_g^{-1}(x_i - \mu_g)\right\},$$

se tiene

$$\frac{\partial}{\partial V_g}\left[f_g(x_i)\right] = \frac{\partial}{\partial V_g}\left\{|V_g|^{-\frac{1}{2}} (2\pi)^{-\frac{p}{2}} exp\left\{-\frac{1}{2}(x_i - \mu_g)' V_g^{-1}(x_i - \mu_g)\right\}\right\}$$

$$= (2\pi)^{-\frac{p}{2}} \frac{\partial}{\partial V_g}\left\{\overbrace{|V_g|^{-\frac{1}{2}}}^{U(V_g)} e \underbrace{xp\left\{-\frac{1}{2}(x_i - \mu_g)' V_g^{-1}(x_i - \mu_g)\right\}}_{W(V_g)}\right\}$$

$$= (2\pi)^{-\frac{p}{2}} \frac{\partial}{\partial V_g}\left\{\left[U(V_g)\right]\left[W(V_g)\right]\right\}$$

$$= (2\pi)^{-\frac{p}{2}} \left\{\left[U(V_g)\right]\frac{\partial[W(V_g)]}{\partial V_g} + \left[W(V_g)\right]\frac{\partial[U(V_g)]}{\partial V_g}\right\} \qquad (2.9)$$

Se encuentran las derivadas planteadas en (2.9)

Al considerar la propiedad de las derivadas parciales matriciales se tiene.

Según la propiedad a) de las derivadas matriciales vistas en la sección (2.6) se tiene:

$$\frac{\partial}{\partial V_g}\left[U(V_g)\right] = \frac{\partial\left[|V_g|^{-\frac{1}{2}}\right]}{\partial V_g}$$

$$= -\frac{1}{2}|V_g|^{-\frac{3}{2}}\left[|V_g|(V'_g)^{-1}\right]$$

$$= -\frac{1}{2}|V_g|^{-\frac{1}{2}}(V_g)^{-1}$$

$$= -\frac{1}{2}\left[\frac{|V_g|^{-\frac{1}{2}}}{V_g}\right]\left[\frac{V_g}{V_g}\right]$$

$$= -\frac{1}{2}\left[\frac{|V_g|^{-\frac{1}{2}}V_g}{V_g}\right]$$

Según la definición 2.7 se tiene que:

$$\frac{\partial}{\partial V_g}[W(V_g)] = \frac{\partial}{\partial V_g}\left[exp\left\{-\frac{1}{2}(x_i-\mu_g)'V_g^{-1}(x_i-\mu_g)\right\}\right]$$

$$= exp\left\{-\frac{1}{2}(x_i-\mu_g)'V_g^{-1}(x_i-\mu_g)\right\}\frac{\partial}{\partial V_g}\left\{-\frac{1}{2}(x_i-\mu_g)'V_g^{-1}(x_i-\mu_g)\right\}$$

$$= exp\left\{-\frac{1}{2}(x_i-\mu_g)'V_g^{-1}(x_i-\mu_g)\right\}\left\{-\frac{1}{2}(x_i-\mu_g)(x_i-\mu_g)'(-1)(V_g^{-2})\right\}$$

$$= exp\left\{-\frac{1}{2}(x_i-\mu_g)'V_g^{-1}(x_i-\mu_g)\right\}\left\{\frac{(x_i-\mu_g)(x_i-\mu_g)'}{2V_g^2}\right\}$$

Luego reemplazando en (2.8) se tiene:

$$\frac{\partial}{\partial V_g}[f_g(x_i)] = (2\pi)^{-\frac{p}{2}}\left\{[U(V_g)]\frac{\partial[W(V_g)]}{\partial V_g} + [W(V_g)]\frac{\partial[U(V_g)]}{\partial V_g}\right\}$$

$$= (2\pi)^{-\frac{p}{2}}\left\{\left[|V_g|^{-\frac{1}{2}}\right] * \left[exp\left\{-\frac{1}{2}(x_i-\mu_g)'V_g^{-1}(x_i-\mu_g)\right\}\left\{\frac{(x_i-\mu_g)(x_i-\mu_g)'}{2V_g^2}\right\}\right]\right.$$

$$\left. + \left[exp\left\{-\frac{1}{2}(x_i-\mu_g)'V_g^{-1}(x_i-\mu_g)\right\}\right] * \left[\left(-\frac{1}{2}\right)\frac{|V_g|^{-\frac{1}{2}}V_g}{V_g^2}\right]\right\}$$

$$\left\{\underbrace{|V_g|^{-\frac{1}{2}}(2\pi)^{-\frac{p}{2}}exp\left\{-\frac{1}{2}(x_i-\mu_g)'V_g^{-1}(x_i-\mu_g)\right\}}_{f_g(x_i)}\left[\frac{(x_i-\mu_g)(x_i-\mu_g)'}{2V_g^2}\right]\right\}$$

$$+\left\{\underbrace{|V_g|^{-\frac{1}{2}}(2\pi)^{-\frac{p}{2}}exp\left\{-\frac{1}{2}(x_i-\mu_g)'V_g^{-1}(x_i-\mu_g)\right\}}_{f_g(x_i)}\left[\frac{V_g}{2V_g^2}\right]\right\}$$

$$= f_g(x_i)\left\{\frac{(x_i-\mu_g)(x_i-\mu_g)'}{2V_g^2}\right\} + f_g(x_i)\left\{-\frac{V_g}{2V_g^2}\right\}$$

Entonces:

$$\frac{\partial}{\partial V_g}\left[f_g(x_i)\right] = f_g(x_i)\left\{\frac{(x_i-\mu_g)(x_i-\mu_g)'}{2V_g^2}\right\} + f_g(x_i)\left\{-\frac{V_g}{2V_g^2}\right\}$$

$$\frac{f_g(x_i)}{2V_g^2}\left\{(x_i-\mu_g)(x_i-\mu_g)' - V_g\right\} \tag{2.9}$$

Reemplazando (2.9) en (2.7) y sustituyendo los resultados dados en (2.8) se tiene:

$$\sum_{i=1}^{n}\frac{\pi_g\left[\frac{f_g(x_i)}{2V_g^2}\left\{(x_i-\hat{\mu}_g)(x_i-\hat{\mu}_g)' - V_g\right\}\right]}{\sum_{g=1}^{G}\pi_g f_g(x_i)} = 0$$

$$\sum_{i=1}^{n}\frac{\pi_g f_g(x_i)}{\sum_{g=1}^{G}\pi_g f_g(x_i)}\left[\frac{(x_i-\hat{\mu}_g)(x_i-\hat{\mu}_g)' - V_g}{2V_g^2}\right] = 0$$

$$\sum_{i=1}^{n}\pi_{ig}\left[\frac{(x_i-\hat{\mu}_g)(x_i-\hat{\mu}_g)' - V_g}{2V_g^2}\right] = 0$$

$$\sum_{i=1}^{n}\pi_{ig}\left[(x_i-\hat{\mu}_g)(x_i-\hat{\mu}_g)' - V_g\right] = 0$$

$$\sum_{i=1}^{n}\pi_{ig}\,(x_i-\hat{\mu}_g)(x_i-\hat{\mu}_g)' - \sum_{i=1}^{n}\pi_{ig}\,V_g = 0,$$

Entonces:

$$\sum_{i=1}^{n}\pi_{ig}\,V_g = \sum_{i=1}^{n}\pi_{ig}\,(x_i-\hat{\mu}_g)(x_i-\hat{\mu}_g)'$$

$$\hat{V}_g = \frac{1}{\sum_{i=1}^{n} \pi_{ig}} \sum_{i=1}^{n} \pi_{ig} \left(x_i - \hat{\mu}_g\right)\left(x_i - \hat{\mu}_g\right)'$$

Para la solución de las ecuaciones planteadas (2.6), (2.8), (2.9) y obtener los estimadores es necesario las probabilidades π_{ig}, para lo que es necesario los parámetros del modelo. Las mezclas de distribuciones probabilísticas son utilizadas para modelar problemas de clasificación de datos, aquí es donde se utilizan algoritmos de clasificación con los cuales se van obteniendo elementos suficientes para la estimación de los parámetros [2].

2.5. Clasificación de Datos Asumiendo una Mezcla Finita

En la clasificación de datos cuyo origen son distribuciones mezcladas se consideran los resultados del Análisis de conglomerados y del Análisis discriminante; para variables categóricas Análisis de correspondencias múltiples, el primero tiene como objetivo agrupar elementos en grupos homogéneos basando en función de las similitudes entre ellos, es decir, lo hace desde un punto de vista descriptivo. El segundo es una herramienta que asigna o clasifica nuevos elementos en grupos previamente reconocidos; conociendo algunas características de un individuo y partiendo del hecho que pertenece a uno o varios grupos definidos a priori, se asigna al individuo a uno de estos grupos basados en la información disponible.

Asumiendo que los datos generados a partir de una mezcla de G distribuciones desconocidas se presenta a continuación un método para dividir la muestra en grupos más homogéneos, mediante el algoritmo EM para mezclas, con el que se obtiene la estimación de los parámetros de las componentes de la mezcla en forma iterativa, con lo cual se realizará la clasificación de los individuos en los grupos por su probabilidades de pertenencia.

CAPÍTULO III

3. Algoritmo de Máxima Expectación

3.1. Generalidades

El algoritmo EM fue descrito y analizado por Dempster, Laird y Rubin en (1977), en el documento titulado Maximum Likelihood Incomplete Data via the EM Algorithm aunque el método ya había sido usado mucho antes, para la obtención de un máximo de la función de verosimilitud cuando el cálculo no es posible. Este algoritmo es un método de optimización iterativo utilizado para la estimación de parámetros desconocidos en la función de máxima verosimilitud de un conjunto de datos muestrales. Posteriormente el algoritmo simplifica la estimación ampliando el conjunto de datos, ingresando los datos no observados de los datos muestrales, es decir, de los datos observados [2].

Su nombre proviene de dos pasos que realiza una esperanza y una maximización, su popularidad en el uso de investigaciones se debe a que permite estimar datos faltantes en diferentes problemas multivariantes donde los algoritmos como el método iterativo de Newton resultan complicado. Su éxito se basa en la simplicidad, estabilidad y sus propiedades de convergencia.

Los pasos del algoritmo son:

- **Paso Esperanza (E):** consiste en calcular la esperanza de la función de verosimilitud condicionada de los datos faltantes a través de los datos observados y de la distribución de los datos ausentes.

- **Paso Maximización (M):** se maximiza la esperanza encontrada en el paso anterior respecto a los parámetros.

Los parámetros encontrados en el paso de maximización se toman como valores iniciales en el paso de esperanza, repitiéndose alternadamente estos dos pasos hasta que los valores de los parámetros no cambien.

A pesar de las bondades del algoritmo como lo son la simplicidad y la convergencia garantizada una falencia que se discute es su lentitud en la convergencia comparado con algoritmos como Newton Raphson cuya convergencia es rápida. Computacionalmente se busca un equilibrio entre la ganancia de reducción de tiempo de computo con el aumento en el tiempo de procesamiento con un acelerador para este algoritmo.

Se han desarrollado modificaciones y extensiones del algoritmo como la de Redner y Homer (1984) que lo usan de manera conjunta con el algoritmo de Newton sumando a las buenas propiedades del algoritmo la rápida convergencia del método de Newton; otro híbrido es la combinación con Gauss-Newton introducido por Aitkin (1996).

3.2. Fundamentos

Para tener conocimiento del funcionamiento del algoritmo se supone una muestra aleatoria de tamaño n=20 de una variable aleatoria vectorial $x = (x_a, x_b, x_c, x_d, x_e)$, donde para algunos de los n individuos no se tiene valores en algunas variables, como en el ejemplo.

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

Tabla 3.1 – Ejemplo de 20 individuos con datos ausentes de las variables $(x_a, x_b, x_c, x_d, x_e)$

Un algoritmo que permita llegar a la solución puede ser el siguiente:

- Se estiman parámetros del modelo estadístico con los datos observados.
- Se inicializa con los parámetros encontrados para estimar los datos faltantes.
- Con los datos observados y los estimados se realiza la estimación de máxima verosimilitud de los parámetros.
- Con esta estimación se vuelve a estimar los datos iterando hasta obtener la convergencia en los parámetros.

El algoritmo analiza la aparición de los datos ausentes desde la unificación de dos enfoques:

Observaciones con datos faltantes: Si X es el vector con n elementos de la muestra, se puede particionar como $X = (Y, Z)$, siendo $Y = (y_1, \dots, y_{ny})$ los individuos con datos completos y $Z = (z_1, \dots, z_{nz})$ los individuos con datos faltantes, X por tanto es una matriz de tamaño $(n * p)$ y $n_y + n_z = n$.

Si la muestra es de tamaño $n = 20$, $n_y = 7$, $n_z = 13$ y $p = 5$ la partición sería:

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

Variables con datos faltantes: si X es el vector de variables, se lo puede particionar como $x = (Y; Z)$, donde $Y = (y_1, \dots, y_{py})$ son las variables con datos completos y $Z = (z_1, \dots, z_{pz})$ las variables con datos incompletos, por tanto X es una matriz de tamaño $(n * p)$ y $p_y + p_z = p$.

Si la muestra de tamaño $n = 20$ con $p = 5$, $p_y = 2$ y $p_z = 3$ esta partición será:

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

Al unificar en X, matriz de datos de la muestra los criterios expuestos, esta se puede particionar en $Y = (y_1, \dots, y_{ny})$, considerada la matriz de datos observados, donde y_i es un vector $(p_y x_1)$ y $Z = ((z_1, \dots, z_{nz})$ considerada matriz de datos faltantes, donde z_i es un vector $(p_z x_1)$. En los criterios se tiene n_y individuos con datos completos y n_z con datos faltantes cuya dimensión $p_y = p_z = p$, en el segundo criterio se tiene p_y variables con datos completos y p_z variables con datos incompletos con dimensión $n_y = n_z = n$.

Si la muestra es de tamaño $n = 20$, $n_y = 7$, $n_z = 13$ y $p = 5$, $p_y = 2$, $p_z = 3$ la unificación será:

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

El algoritmo toma la matriz $X = (Y, Z)$ como la partición de los datos muestrales ya sea de individuos o variables o los dos.

Dada la relación en las distribuciones de probabilidad condicionada se tiene que la densidad conjunta de (Y, Z)con los parámetros θ se describe como:

$$f(Y, Z|\theta) = f(Z|Y, \theta)f(Y|\theta),$$

entonces:

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

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

$$l(\theta|Y) = \frac{l(\theta|Y,Z)}{l(\theta,Y|Z)} \quad (3.1)$$

Donde $l(\theta|Y,Z)$ es la verosimilitud para toda la muestra y $l(\theta,Y|Z)$ es la verosimilitud de los datos ausentes conocidos los datos observados. Operando en la ecuación (3.1) se tiene:

$$\ln l(\theta|Y) = \ln l(\theta|Y,Z) - \ln l(\theta,Y|Z)$$

$$L(\theta|Y) = L(\theta|Y,Z) - \ln l(\theta,Y|Z)$$

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

El objetivo del algoritmo es la estimación de los parámetros desconocidos de la muestra poblacional. Lo más sencillo es la maximización de $L(\theta|Y,Z)$ que es la función de soporte de la muestra completa que la maximización de los datos observados $L(\theta|Y)$, por esta razón el algoritmo EM usa $L(\theta|Y,Z)$ como función de verosimilitud para el estimado MV de θ.

Con todas estas consideraciones la estructura del algoritmo es:

a. Partir de un estimador inicial $\hat{\theta}^{[o]}$ teniendo en cuenta un margen de error para los valores de los parámetros estimados.
b. Iniciar un contador $k = 0$.
c. Hacer $k = k + 1$
d. Paso E:

Con la estimación de θ, calcular la esperanza de $L(\theta|Y,Z)$ respecto a los valores ausentes Z y los datos observados. Que nos da una nueva verosimilitud

$$L*(\theta|Y) = E_{Z|\hat{\theta}^{[k-1]}}[L(\theta|Y,Z)]$$

4. Paso M:

   Maximizar $L*(\theta|Y)$ respecto al vector de variables θ y se tiene el nuevo vector de parámetros

$$\hat{\theta}^{[k]} = máx[L*(\theta|Y)].$$

e. Se calcula $\left\|\hat{\theta}^{[k]} - \hat{\theta}^{[k-1]}\right\|$ evaluando si es suficientemente pequeña, menor a la tolerancia que se establezca, si lo es, se tiene el estimador de máxima verosimilitud caso contrario se vuelve a iterar en el paso $k = k + 1$ iterando hasta lograr la convergencia.

3.3 Definición del Algoritmo

El propósito de los modelos mixtos es proporcionar la partición de los datos en g grupos, siendo este número de grupos ya preestablecido. La i-ésima proporción de la mixtura $(\pi_i, i = 1, \dots, g)$ se puede interpretar como la probabilidad a priori de que una observación pertenezca a un grupo [2]:

$$P(Z_{ij} = 1) = \pi_i \qquad i = 1, \dots, g$$

El procedimiento de agrupamiento busca asociar cada variable $z_1, \dots, z_n$ con los datos observados $y_1, \dots, y_n$; cuando el modelo de mixtura ha sido ajustado y estimado su parámetro, las observaciones tienen un agrupamiento en probabilidades posteriores de pertenecían a uno u otro grupo.

$$\hat{\tau}_{ij} = P\{Z_{ij} = 1 | Y_j = y_j\} = \frac{\pi_i f_i(y_i|\widehat{\theta}_i)}{\sum_{\ell=1}^{g} \pi_\ell f_\ell(y_\ell|\widehat{\theta}_\ell)} \qquad \ell = 1, .., g \quad j = 1, \dots, n$$

La asignación de las observaciones a uno u otro grupo se decide mediante la mayor probabilidad

$$\hat{z}_{ij} = \begin{Bmatrix} 1 & si\ i = \arg_\ell max\ \hat{\tau}_{ij} \\ 0 & c.c \end{Bmatrix} \qquad i = 1, \dots, g \quad j = 1, \dots, n$$

3.4. Aplicación del Algoritmo en Mezclas de Distribución

Se realiza una descripción rápida de las variables y conjunto de datos que se van a utilizar, se recodifican las variables categóricas con valores enteros, posteriormente se determina la semilla y el número de grupos que se busca obtener, después se utiliza la librería “fcp”, el paquete “flexmix” y el modelo lcmixed del programa R, se aplica a los resultados obtenidos un score creado en base a

criterio experto de las diferentes áreas de la institución y finalmente se caracteriza los resultados estadísticamente proponiendo una matriz para la toma de decisiones.

3.4.1. Descripción de los Datos

En el trabajo que se está presentando se considera la información de los clientes de una institución bancaria que corresponden a datos sociodemográficos, si bien el objetivo es clasificar variables cualitativas también se pueden incluir variables continuas.

Sobre las variables continuas se verifica su distribución para la aplicación de la metodología, estas distribuciones no presentan normalidad sin embargo se considera el teorema de límite central (TLC) y se realizan simulaciones con lo que se comprueba que dado el número de observaciones es suficiente para asumir normalidad de la distribución de los datos.

Definición TLC Sea $X_1, X_2, \dots, X_n$ un conjunto de variables aleatorias, independientes e idénticamente distribuidas con media μ y varianza $0 < \sigma^2 < 0\infty$. Sea

$$S_n = X_1, X_2, \dots, X_n$$

Entonces

$$\lim_{n\to\infty} P_r\left(\frac{S_n - n\mu}{\sigma\sqrt{n}} \leq z\right) = \Phi(z)$$

La base contiene información transaccional, sociodemográfica, comportamiento en los canales y variables de relación entre el cliente y la institución el período de análisis abarca información de 3 años lo que permite la creación de un score comportamental que no se ve afectado por temas de estacionalidad.

Las variables a considerarse son:

VARIABLE	DESCRIPCION	TIPO DE VARIABLE
Ingresos	ingresos que percibe el cliente	continua
tipo empresa	clasificación de la empresa donde presta servicios	categórica
Ocupación	ocupación registrada en las fuentes de datos	categórica
Genero	género de la persona	categórica
Cedulado	tipo de cedula que posee el cliente	categórica
Nacionalidad	nacionalidad del cliente	categórica
estado civil	estado civil del cliente	categórica
nivel estudios	nivel de estudios del cliente	categórica
Profesión	profesión registrada en la cédula	categórica
lugar de nacimiento	ciudad de nacimiento	categórica
División	sectorización da nivel parroquial	categórica
Zona	identifica zona rural o urbana	categórica
actividad económica	actividad a la que se dedica el cliente	categórica
barrio	barrio de residencia del cliente	categórica
Antigüedad	antigüedad en la institución	continua
tipo vivienda	tipo de vivienda que posee el cliente	categórica
numero cargas	número de cargas familiares	continua
Estado	si el cliente a la fecha de análisis está activo para la institución	categórica
Edad	edad del cliente a la fecha del análisis	continua

Tabla 3.2 – Descripción de datos sociodemográficos

Se observa que hay 19 variables de las cuales 4 son continuas y 15 discretas, esta información es tomada de las bases de datos de la institución, al ser información que no es tan variable en el tiempo asegura que las características generales de los grupos encontrados se mantengan en el tiempo. Otra característica importante sobre los datos es que son el resultado de una campaña de actualización de datos realizada por la institución.

3.4.2. Resultados de los Grupos Obtenidos

Mediante el programa R, se utilizan las librerías fpc y flexmix en las cuales ya se encuentra implementado el algoritmo EM así como la metodología de mezclas finitas para la clasificación de variables categóricas.

Se parte de un valor semilla que inicializara el algoritmo y se elige el número de grupos que se busca obtener esta elección se basa en el manejo de grupos que se desea obtener con los resultados se analiza los estadísticos y el tamaño de los grupos.

Número de Grupos	Población	%
1	6.264	8%
2	20.643	25%
3	15.911	19%
4	13.228	16%
5	26.159	32%
TOTAL	**82.205**	**100%**

Tabla 3.3 – Distribución del número de grupos

Se tiene una distribución adecuada de los clientes en los diferentes grupos basado en el conocimiento empírico y de negocio, así como en los criterios AIC y BIC, el número de grupos es adecuado ya que no existen concentraciones de población y se adapta a la gestión comercial que la entidad puede manejar.

3.4.3. Estadísticos de Validación

Elegir el número de componentes para el modelo de mixtura no es trivial, depende de varios factores entre los que destacan la distribución de los datos modelizados y la forma de las componentes. Una alternativa son los criterios de información, para esta aplicación se utilizan dos criterios para llegar a un número adecuado de grupos.

El primer criterio que se evalúa es el de Akaike (AIC) desarrollado en 1974, el cual es una medida de la bondad de ajuste de un modelo de estimación estadística. El AIC es una manera operacional de acortar distancia entre la complejidad de un modelo estimado y un buen modelo ajustado a los datos y que se expresa como [5]:

$$AIC = -2LogL(\Psi) + 2k$$

donde $L(\Psi)$ es la función de verosimilitud del modelo, y k el número de parámetros independientes de la mixtura según el número de componentes propuesto (3_{g-1}). Al seleccionar entre los distintos modelos cada uno definido por diferentes componentes se elige aquel que presente menor AIC, este criterio penaliza el sobreajuste en modelos grandes mediante k.

El criterio AIC se basa en la medida de información de Kullback-Leibler (KL), que permite interpretar la distancia entre las dos distribuciones.

El segundo criterio a considerarse es el bayesiano (BIC), expresado mediante [5]:

$$BIC = -2logL(\Psi) + klog(n)$$

es similar a AIC excepto por la penalización, ahora incluye el número de observaciones independientes de la muestra univariante (n). Este parámetro es menos susceptible de sobreestimar el número de componentes.

El BIC parte de una aproximación a la probabilidad a posteriori de un modelo para una muestra grande, el BIC es la desviación del modelo, aplicada en esta aproximación.

Al minimizar este criterio para todos los posibles modelos se están haciendo comparaciones con diferentes cantidades de parámetros y diferentes cantidades de grupos.

Número Grupos	K	AIC	BIC
4	4	6.511.043	6.724.840
5	5	6.353.747	6.620.995
6	6	6.376.925	6.597.625
8	7	6.685.617	7.059.767
10	8	6.393.656	6.641.258

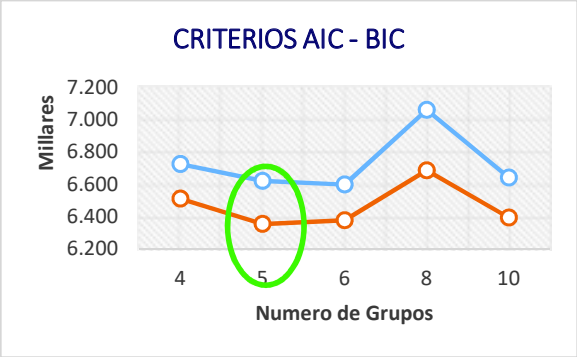
Tabla 3.4 – Descripción de AIC y BIC

El proceso de seleccionar el mejor modelo se realiza tomando el valor mínimo de los dos criterios.

Basados en los criterios AIC y BIC el número adecuado de grupos es 5, al observar la distribución del tamaño de grupos se evidencia que es el número idóneo ya que no existen concentraciones en los grupos estos datos están en los anexos.

No existe un criterio mejor para todos los modelos, el objetivo es obtener el modelo que presente un asunto adecuado y que sea el más parsimonioso.

Como se observa en la gráfica los valores mínimos de AIC y BIC son los que se observan cuando se tienen 5 grupos (clústers).



Gráfica 3.4 – Valores de AIC y BIC

3.4.4. Descripción de los Grupos Obtenidos

Para la descripción de los grupos obtenidos se realizan cuadros con los valores promedios y la distribución de la población de las diferentes variables, como se muestra en la tabla de los anexos.

GRUPO 1

Corresponde al 8% de la población con un ingreso promedio de 2,749, predomina el género masculino, casados con estudios superiores, con vivienda propia hipotecada, activos, en promedio de 42 años con una carga familiar y antigüedad en la institución de 10 años, predominan como profesión empleado privado en sociedades y compañías.

GRUPO 2

Corresponde al 32% de la población con un ingreso promedio de 416, predomina el género femenino, viudas con educación elemental, con vivienda propia no hipotecada, activo, en promedio de 48 años con dos cargas familiares y antigüedad en la institución de 10 años, predomina como profesión estudiantes y quehaceres domésticos, se destacan tipos de empresas unipersonales y de pequeña industria, este grupo tiene el 16% de concentración de la población en la zona rural.

GRUPO 3

Corresponden al 19% de la población con un ingreso promedio de 668, predomina el género femenino, casadas con estudios secundarios y superiores, vive con familiares, activos, en promedio de 33 años sin cargas con antigüedad en la

institución de 6 años, predomina como profesión estudiantes en la empresa privada y organismos del estado.

GRUPO 4

Corresponden al 16% de la población con un ingreso promedio de 1,071, predomina el género femenino, con estado civil unión de hecho y estudios superiores, con vivienda propia, activo en promedio de 46 años con una carga y de 10 años de antigüedad en la institución, predomina como profesión empleado privado y público en organismos del estado, en este grupo se encuentran el mayor porcentaje de extranjeros.

GRUPO 5

Corresponden al 25% de la población con un ingreso promedio de 407, predomina el género masculino, solteros con estudios primarios, inactivos de 26 años sin cargar familiares y antigüedad en la institución de 4 años, predomina como profesión estudiantes en la empresa privada , esta población tiene el 13% de concentración en la zona rural.

3.4.5. Composición Score

Como complemento para los grupos obtenidos se genera un score basado en información transaccional de los clientes lo que permite generar una matriz de estrategia en función de sus características grupales (clústers) así como de su comportamiento en la institución.

Es score está compuesto por 43 variables agrupadas en 4 grandes grupos en base a su procedencia y tipo de información que aportan.

1. Variables Transaccionales
2. Variables de Productos
3. Variables de Canales
4. Frecuencia

A cada grupo de variables se les asigna un peso y a cada variable un porcentaje de importancia basado en el criterio comercial como se muestra en el anexo.

El aporte del conocimiento de cada área de la institución facilita la sociabilización de la metodología y los resultados ya que al realizar la descripción de los grupos se observa la coherencia entre las diferentes variables, es decir, si la edad en el grupo es alta la relación con las variables de canales se inclinan hacia medios no tecnológicos.

3.4.6. Matriz de Estrategias

Con los grupos demográficos y el score se genera una matriz de estrategia que permite tomar decisiones focalizadas en cada uno de los perfiles.

GRUPOS	Menor a 87	De 88 a 125	De 126 a 317	Mayor a 317
G1	1.299	1.367	1.810	1.788
G3	3.093	3.747	3.730	5.341
G4	2.979	2.290	3.064	4.895
G5	5.227	4.543	5.201	5.672
G2	5.224	10.940	7.080	2.915

GRUPOS	Menora 87	De 88 a 125	De 126 a 317	Mayor a 317
G1	MANTENER		MANTENER SELECTIVAMENTE	DESARROLLAR
G3				
G4				
G5	RETIRARSE	MANTENER AL MINIMO	DEFENDER	
G2				

ESTRATEGIA	#	%
RETIRARSE	10.451	13%
MANTENER	14.775	18%
MANTENER SELECTIVAMENTE	8.604	10%
DESARROLLAR	12.024	15%
DEFENDER	20.868	25%
MANTENER AL MINIMO	15.483	19%

CAPÍTULO IV

4. Conclusiones y Recomendaciones

4.1. Conclusiones

En este trabajo se presenta la aplicación de la estadística, utilizando el método de mezclas finitas para la clasificación de variables categóricas y continuas, permitiendo estimar la probabilidad que tiene cada uno de los clientes de pertenecer a uno u otro grupo, el objetivo central es la clasificación de una población de una entidad bancaria aplicando el algoritmo EM en el software estadístico R, que nos permite tener grupos homogéneos de clientes en los cuales se pueden observar patrones de comportamiento y características diferentes en cada uno de ellos con lo cual se pueden generar estrategias diferenciadas y ser más efectivos en las diferentes campañas que proponga la entidad así como en la toma de decisiones sobre los diferentes grupos.

Cada cliente corresponde a un y solo uno de los grupos encontrados, siendo su probabilidad de pertenencia similar con los individuos de su mismo grupo y diferente a la probabilidad de los individuos que pertenecen a los otros grupos.

Los datos provienen de la información de la entidad bancaria, los mismos que mediante el uso del método de clasificación se encontraron 5 grupos con información sociodemográfica.

El primer grupo representa el 8% de la población donde los ingresos son altos hombres de edad madura con vivienda propia y estudios superiores.

El segundo grupo representa el 16% de la población mujeres con ingresos medios de estudios superiores con edad madura

El tercer grupo representa el 19% de la población con ingresos bajos mujeres solteras de edad y estudios medios.

El cuarto grupo corresponde al 32% de la población con ingresos bajos casados, educación media de edad madura.

El quinto grupo representa el 25% de la población, hombres solteros con ingresos bajos.

Al describir el perfil se ve que existen diferencias entre los grupos, como complemento se incluye el score que recoge la información de comportamiento del cliente en la institución con lo cual se genera la matriz de estrategia basadas en los grupos demográficos y el score comportamental.

El algoritmo a pesar no tener una convergencia rápida es muy eficiente al momento de clasificar a los individuos. Como alcance a este algoritmo y para mejorar su velocidad de convergencia se debería utilizar el algoritmo en sus versiones mixtas.

Una ventaja en la construcción del modelo es la determinación de grupos de características bien definidos que permite optimizar la clasificación de los datos, el software para la implementación de la solución propuesta a la clasificación de variables categóricas es de libre acceso lo que permite establecer de manera fácil y sencilla la implementación.

Una ventaja que presentan los modelos de mixturas finitas frente a otros modelos estadísticos es que no se basan en las hipótesis de modelamiento como son que las distribuciones sean normales, y que no se necesitan conocer los parámetros a priori ya que el modelo se basa en las probabilidades condicionales de pertenecer a una clase dado que ya pertenece a un grupo; estas probabilidades iniciales las re estima hasta obtener la convergencia.

El algoritmo EM sirve para un número arbitrario de dimensiones lo que permite su aplicación a datos de diversas fuentes y que no conlleve a realizar transformación para obtener normalidad

Una limitación que se puede presentar en el modelo propuesto es la capacidad de fuerza comercial de la entidad financiera para ejecutar las estrategias por lo que el número de grupos no debe ser muy amplio; ante esta limitante se utiliza criterios teóricos y comerciales por medio del score comportamental construido a partir del

trabajo en conjunto con varias áreas de la institución basado en sus conocimientos a priori.

Se debe tener en consideración que el algoritmo EM es altamente dependiente de los parámetros iniciales ya que estos influyen sobre la velocidad de convergencia y la capacidad para alcanzar el máximo global.

4.2. Recomendaciones

Se recomienda que la base de datos sea de calidad, es decir que se mantengan campañas de actualización de la información del cliente, para una mejor clasificación de los mismos, ya que como parte del tratamiento de datos se utilizaron datos por default para aquellos clientes que no contaban con la información al momento de realizar el modelo.

Se recomienda que para la aplicación de esta metodología en bases de volúmenes más grandes, se implemente la versión del algoritmo EM con métodos de convergencia más rápida como Newton-Raphson, buscando equilibrar las ganancias de reducción computacional y el aumento en tiempo de procesamiento.

Se recomienda hacer seguimiento de la evolución de cada grupo mediante indicadores para determinar la estabilidad de los mismos y tener el control del crecimiento de cada grupo que no afecte la gestión comercial.

Para la utilización del algoritmo EM se recomienda testear varias formas para determinar los valores iniciales que ayuden a la rápida convergencia al máximo global. Se deberá elegir como estrategia de inicialización del algoritmo aquella que muestre mejores resultados en velocidad y convergencia.

REFERENCIAS

[1] J. P. C. González, *METODOLOGÍA DE ENTRENAMIENTO DE MODELOS DE MEZCLAS GAUSSIANAS EMPLEANDO CRITERIOS DE GRNA MARGEN PARA LA DETECCIÓN DE PATOLOGÍAS EN BIOSEÑALES,* Manizales, 2010.

[2] H. J. M. Niño, *MEZCLAS FINITAS DE DISTRIBUCIONES NORMALES UNA ALTERNATIVA PARA CLASIFICAR,* Bucaramanga, 2007.

[3] V. J. Cano Fernández, ANÁLISIS DEL NÚMERO DE TIPOS DE VINO CONSUMIDOS EN TENERIFE, TENERIFE, 2001.

[4] M. C. Lucía, ANÁLISIS DE CLASES LATENTES UNA TÉCNICA PARA DETECTAR HETEROGENEIDAD EN POBLACIONES, MÉXICO, 2009.

[5] Á. G. Losada, *MODELOS DE MIXTURAS FINITAS PARA LA CARACTERIZACIÓN Y MEJORA DE LAS REDES DE MONITORIZACIÓN DE LA CALIDAD DEL AIRE,* Granada, 2014.

BIBLIOGRAFÍA

1. Biernacki, C., Celeux, G. and Govaert, G. (2000). Assessing a mixture model for clustering with the integrated completed likelihood. IEEE transactions on pattern analysis and machine intelligence.

2. D. Titterington, A. Smith, and U. Makov, (1985) Statistical Analysis of Finite Mixture Distributions. Chichester, U.K.: John Wiley & Sons.

3. Hartigan, J. A.; Wong, M. A.; Algorithm AS 136: A K-means Clustering Algorithm, Journal of the Royal Statistical Society Series C, vol. 28.

4. Mario A.T. Figueiredo, (2002) Unsupervised Learning of Finite Mixture Models, IEEE Transactions on Pattern Analysis and Machine Intelligence.

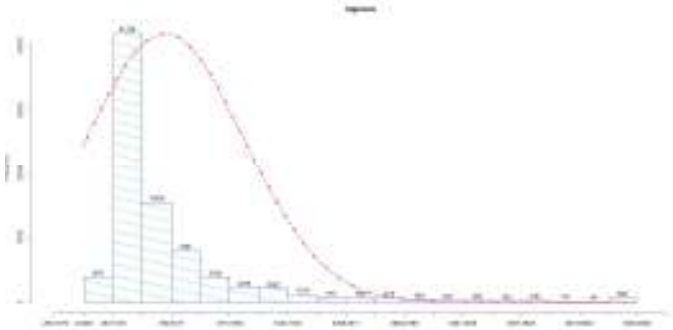
5. R. Torres, R. Salas, H. Allende, and C. Moraga. (2002) *Estimador robusto en modelos de mezcla de expertos locales*. CLATSEV.

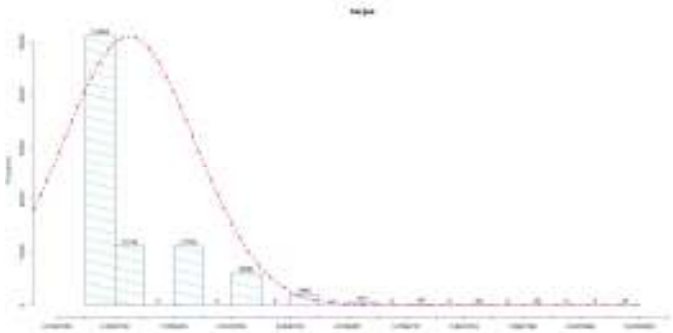
6. Teukolsky, SA; Vetterling, WT; Flannery, (2007). Gaussian Mixture Models and K-means Clustering, Numerical Recipes: The Art of Scientific Computing. New York Cambridge University Press.

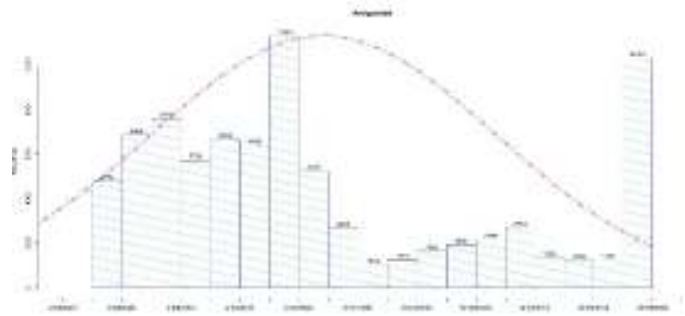
7. V. Melnykov and I. Melnykov, (2011), Initializing the EM algorithm in Gaussian mixture models with an unknown number of components, Computational Statistics & Data Analysis.

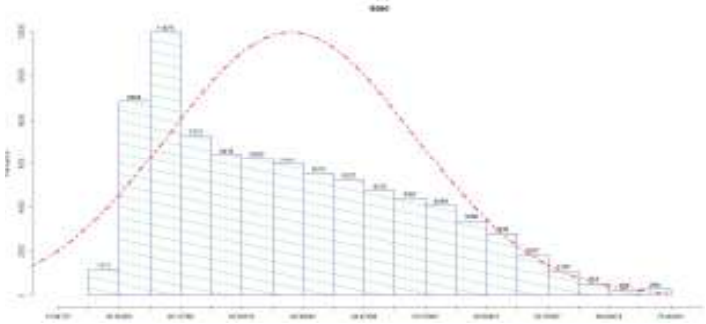
ANEXOS

A. Distribuciones de las Variables Cuantitativas

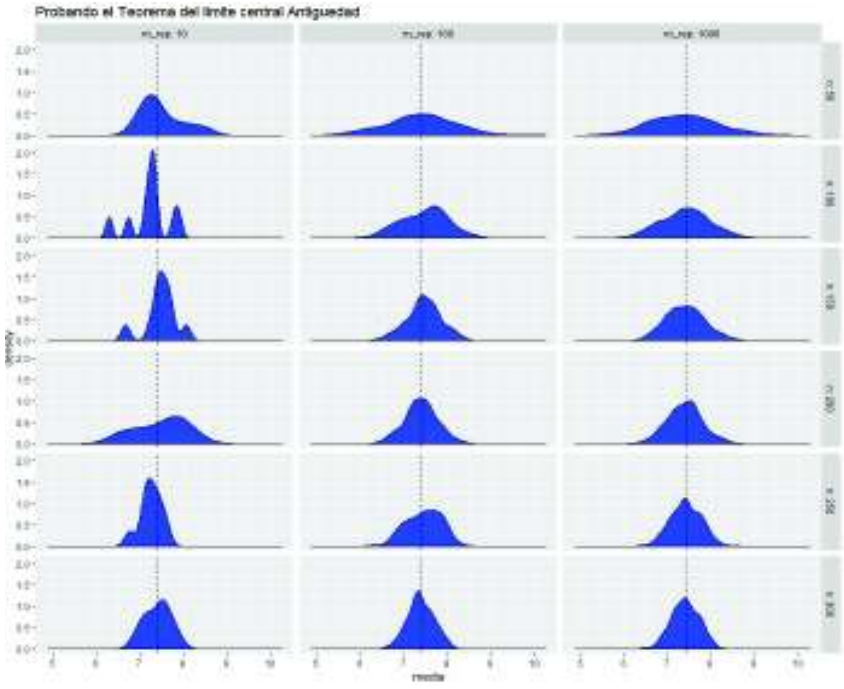


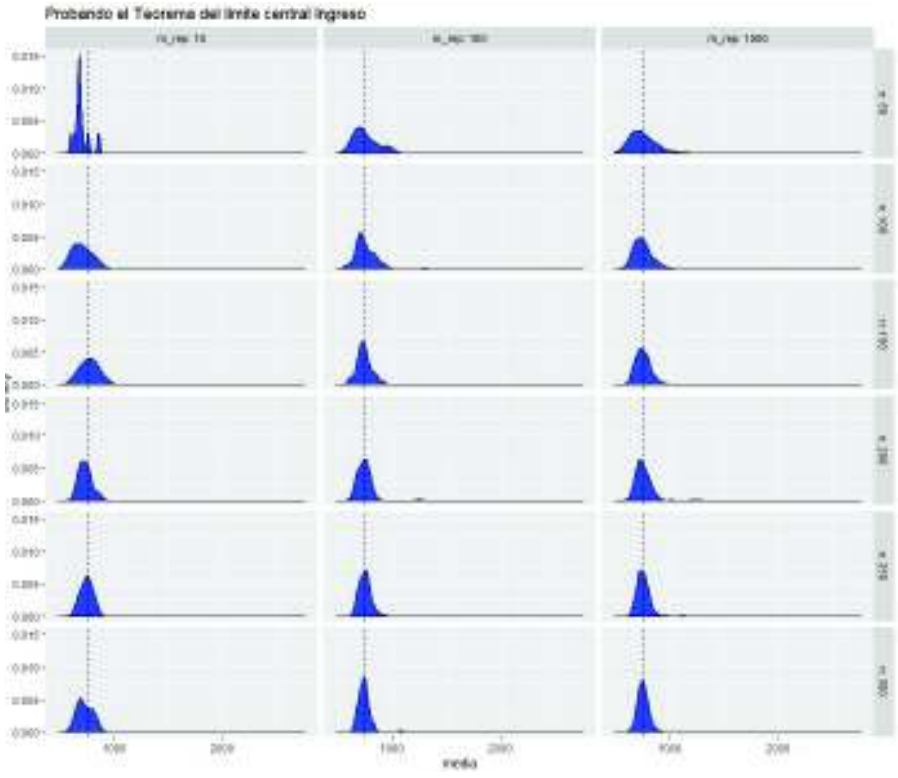


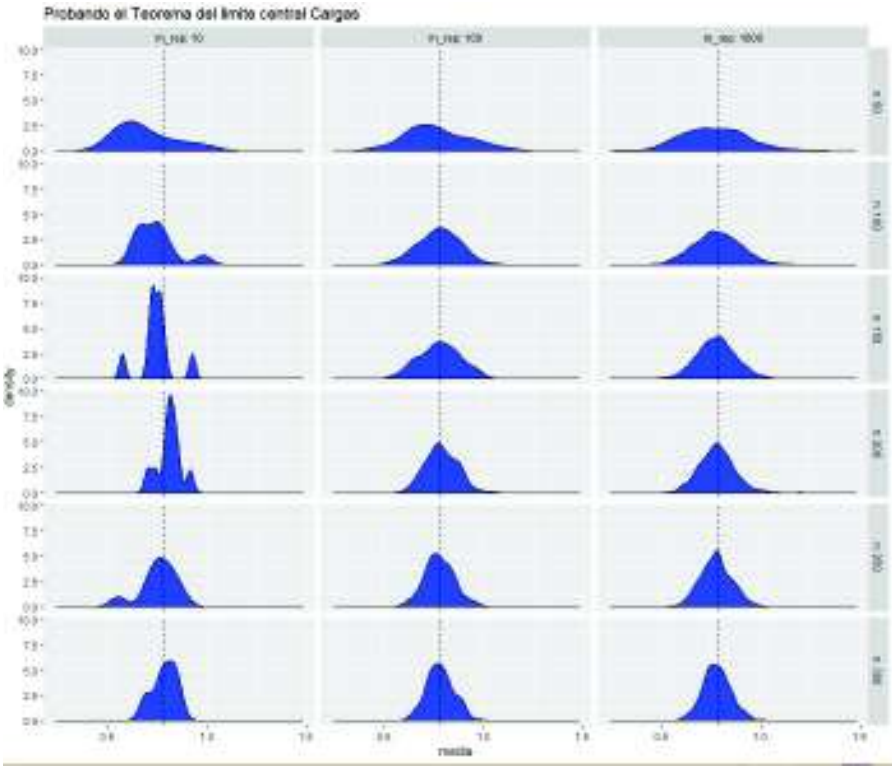


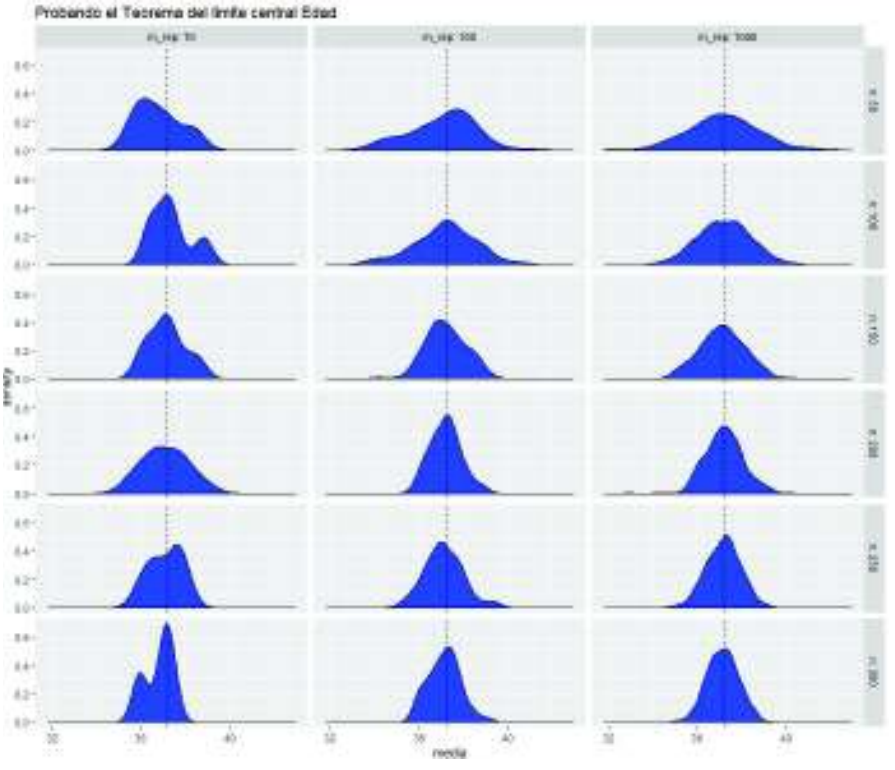


B. Simulaciones de Teorema de Limite Central






Probando el Teorema del límite central Caigao
densidad
media

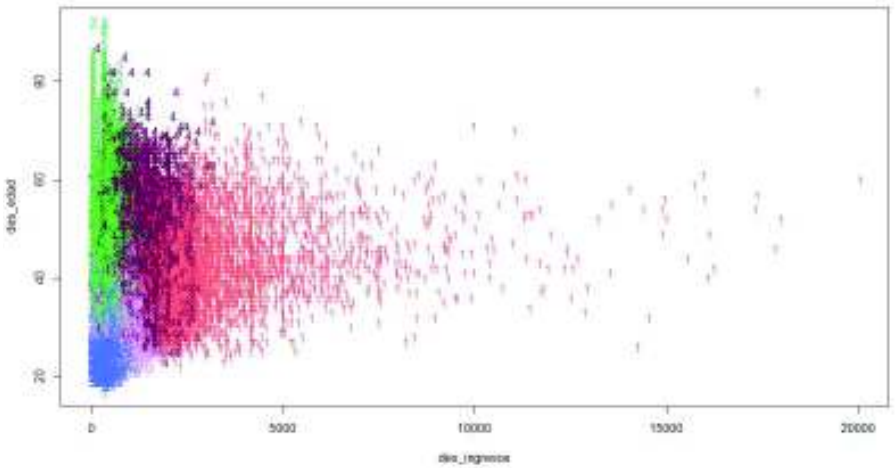

Probando el Teorema del límite central Edad
density
media

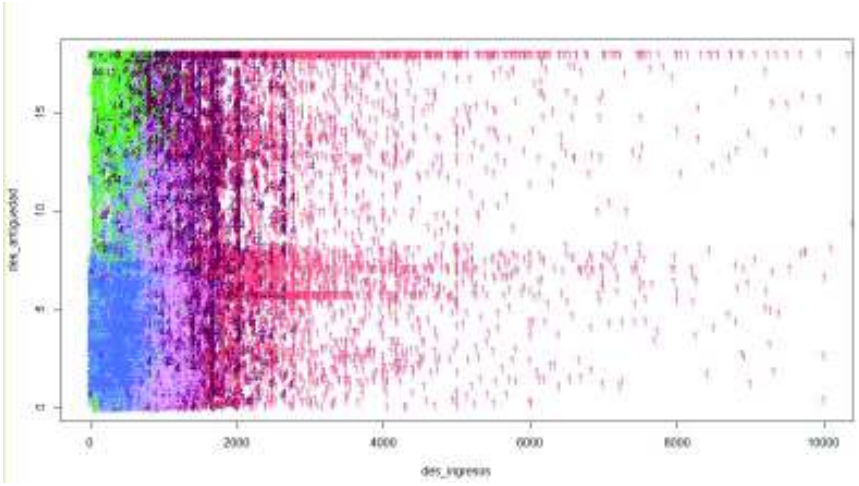
C. Variables para el score comportamental

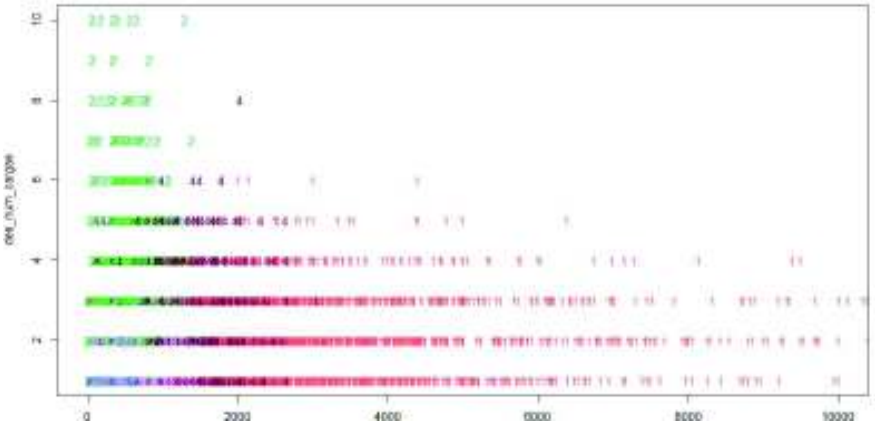
		VARIABLES		PESOS	PUNTAJES
TRANSACCIONAL	550	valor 1	monto de transacciones de depósito	2%	11
		count 1	número de transacciones de depósito	2%	11
		valor 2	monto transacciones de retiro	2%	11
		count 2	número transacciones de retiro	20%	110
		valor 3	monto de transacciones de tarjeta	2%	11
		count 3	número de transacciones de tarjeta	2%	11
		valor 4	monto de transacciones de crédito	2%	11
		count 4	número de transacciones de crédito	2%	11
		valor 5	monto de transacciones de débito	3%	17
		count 5	número de transacciones de débito	25%	138
		i1	número de transacciones de crédito/número total trans.	3%	17
		i2	monto de transacciones de crédito/monto total trans.	3%	17
		i3	número de transacciones de débito/número total trans.	10%	55
		i4	monto de transacciones de débito/monto total trans.	3%	17
		i5	número de transacciones de depósito/número total trans.	3%	17
		i6	monto de transacciones de depósito/monto total trans.	3%	17
		i7	número transacciones de retiro/número total trans.	10%	55
		i8	monto transacciones de retiro/monto total trans.	3%	17
PRODUCTOS	350	Depósitos a Plazo	saldo corte depósitos a plazo	2%	7
		Ahorros	saldo promedio ahorros	2%	7
		Corrientes	saldo promeido corrientes	2%	7
		Visa	saldo corte visa	2%	7
		Consumo	saldo corte consumo	30%	105
		Vivienda	saldo corte vivienda	10%	35
		Patrimonio	total de patrimonio del cliente	2%	7
		Pasivos	total de pasivos del cliente	2%	7
		p1	total patrimonio (ahorros, depósitos,etc)	2%	7
		p2	total pasivos (hipotecario, consumo,tarjetas,etc)	3%	11
		p3	ahorros/total patrimonio + total pasivos	3%	11
		p4	depósitos/total patrimonio + total pasivos	3%	11
		p5	vivienda/total patrimonio + total pasivos	10%	35
		p6	consumo/total patrimonio + total pasivos	24%	84
		p7	tarjeta/total patrimonio+ total pasivos	3%	11
CANAL	75	ATM_M	monto de transacciones por atm	10%	8
		ATM_N	número de transacciones por atm	30%	23
		OFICINAS_M	monto de transacciones por oficina	10%	8
		OFICINAS_N	número de transacciones por oficina	10%	8
		INTERNET_M	monto de transacciones por internet	10%	8
		INTERNET_N	número de transacciones por internet	10%	8
		TARJETA_M	monto de transacciones por tarjeta	5%	4
		TARJETA_N	número de transacciones por tarjeta	5%	4
		SIN_CANAL_M	monto de transacciones sin canal	5%	4
		SIN_CANAL_N	número de transacciones sin canal	5%	4
RECENCY	25	Recency	Fecha de la última transacción	100%	25
TOTAL	**1000**				**1000**

D. Distribución del tamaño de grupos

E. Dispersión de las variables numéricas por grupos.







F. Descripción de variables de los grupos

GRUPO	1	4	3	2	5
% POBLACION	8%	16%	19%	32%	25%
INGRESO PROMEDIO	2.749	1.071	668	416	407
% Hombres	71%	40%	46%	57%	91%
% Mujeres	29%	60%	54%	43%	9%
% CASADO	71%	61%	36%	61%	14%
% DIVORCIADO	9%	14%	6%	12%	1%
% SOLTERO	20%	23%	58%	24%	85%
% UNION DE HECHO	0%	0%	0%	0%	0%
% VIUDO	0%	1%	0%	2%	0%
% BACHILLERATO	15%	10%	34%	32%	43%
% BASICA	1%	0%	2%	18%	19%
% ELEMENTAL	0%	0%	0%	0%	0%
% ESPECIAL	0%	0%	0%	1%	0%
% INICIAL	0%	0%	0%	0%	0%
% NINGUNA	0%	0%	0%	0%	0%
% PRIMARIA	0%	0%	1%	14%	12%
% SECUNDARIA	13%	8%	25%	26%	21%
% SUPERIOR	70%	81%	37%	8%	4%
% ALQUILER	27%	27%	30%	29%	12%
% ANTICRESIS	0%	0%	0%	0%	0%
% NO DEFINIDO	0%	0%	1%	0%	42%
% PRESTADA	0%	0%	0%	0%	0%
% PROPIA HIPOTECADA	3%	2%	0%	1%	0%
% PROPIA NO HIPOTECADA	36%	37%	10%	39%	3%
% VIVE CON FAMILIARES	34%	34%	58%	31%	42%
% Activo	66%	66%	71%	56%	32%
% Inactivo	34%	34%	29%	44%	68%
Edad Promedio	42	46	33	48	26
Cargas Promedio	1	1	0	2	0
Antigüedad Promedio	10	11	6	10	4

G. Código R

Para la obtención de los resultados se utiliza el software libre R, en el cual se realizan los siguientes pasos:

- Carga de la fuente de información
- Consolidación de los datos
- Depuración de datos faltantes
- Análisis univariado de cada variable detectanto: valores faltantes, atípicos, etc.
- Corrida de varias iteraciones en las que se obtienen los criterios AIC y BIC con los cuales se define el número de grupos adecuado para la clasificación de los clientes.
- Finalmente se obtiene la base con el grupo asignado a nivel de individuo que se une al valor del score comportamental para dar como resultado la matriz de estrategia con la cual se realizan las diferentes toma de decisiones y focalización de campañas que requiera hacer la institución.

```
# Carga de fuentes de información
                setwd("C:/PROYECTOS/SEGMENTACION")
                load('Objetos_R/Civiles/Base_Demo.RData')
                base <-
                subset(Base_Demo,des_tipo_cliente=='DEPENDIENTE')
                base <- subset(base,!des_nombre_empresa=='ENTIDAD
                FINANCIERA')
                base_lc <-
                subset(base,select=c('des_tipo_cliente','des_ingresos','des_ti
                poemp',
                'des_ocupacion2','des_genero','des_cedulado','des_nacionali
                dad',
                'des_estado_civil3','des_nivel_estudios','des_profesion3',
                'des_lugar_nac','des_division_p','des_zona','des_act_eco',
                'des_barrio','des_antiguedad','des_tipo_vivienda',
                'des_num_cargas','des_estado','des_edad','R_Edad','R_Sala
                rio'))
#%%          Depuración faltantes de datos en cada variable          %%
```

```
base_lc <-
transform(base_lc,des_ingresos=ifelse(is.na(des_ingresos)==TRUE,0,des_ingresos))

base_lc <-
transform(base_lc,des_tipoemp=ifelse(is.na(des_tipoemp)==TRUE,'NO DISPONIBLE',des_tipoemp))


base_lc <-
transform(base_lc,des_nombre_empresa=ifelse(is.na(des_nombre_empresa)==TRUE,'NO DISPONIBLE',des_nombre_empresa))

base_lc <-
transform(base_lc,des_ocupacion2=ifelse(is.na(des_ocupacion2)==TRUE,'NO DISPONIBLE',des_ocupacion2))

base_lc <-
transform(base_lc,des_genero=ifelse(is.na(des_genero)==TRUE,'NO DIPSONIBLE',des_genero))

base_lc <-
transform(base_lc,des_cedulado=ifelse(is.na(des_cedulado)==TRUE,'NO DISPONIBLE',des_cedulado))

base_lc <-
transform(base_lc,des_nacionalidad=ifelse(is.na(des_nacionalidad)==TRUE,'NO DISPONIBLE',des_nacionalidad))

base_lc <-
transform(base_lc,des_estado_civil3=ifelse(is.na(des_estado_civil3)==TRUE,'NO DISPONIBLE',des_estado_civil3))

base_lc <-
transform(base_lc,des_nivel_estudios=ifelse(is.na(des_nivel_estudios)==TRUE,'NO DISPONIBLE',des_nivel_estudios))

base_lc <-
transform(base_lc,des_profesion3=ifelse(is.na(des_profesion3)==TRUE,'NO DISPONIBLE',des_profesion3))

base_lc <-
transform(base_lc,des_lugar_nac=ifelse(is.na(des_lugar_nac)==TRUE,'NO DISPONIBLE',des_lugar_nac))

base_lc <-
transform(base_lc,des_division_p=ifelse(is.na(des_division_p)==TRUE,'NO DISPONIBLE',des_division_p))

base_lc <-
transform(base_lc,des_zona=ifelse(is.na(des_zona)==TRUE,'NO DISPONIBLE',des_zona))
```

```
base_lc <-
transform(base_lc,des_act_eco=ifelse(is.na(des_act_eco)==
TRUE,'NO DISPONIBLE',des_act_eco))

base_lc <-
transform(base_lc,des_barrio=ifelse(is.na(des_barrio)==TRU
E,'NO DISPONIBLE',des_barrio))

base_lc <-
transform(base_lc,des_antiguedad=ifelse(is.na(des_antigued
ad)==TRUE,0,des_antiguedad))

base_lc <-
transform(base_lc,des_tipo_vivienda=ifelse(is.na(des_tipo_vi
vienda)==TRUE,'NO DISPONIBLE',des_tipo_vivienda))

base_lc <-
transform(base_lc,des_num_cargas=ifelse(is.na(des_num_c
argas)==TRUE,0,des_num_cargas))

base_lc <-
transform(base_lc,des_estado=ifelse(is.na(des_estado)==TR
UE,'NO DISPONIBLE',des_estado))

base_lc <-
transform(base_lc,des_edad=ifelse(is.na(des_edad)==TRUE
,0,des_edad))

base_lc <-
transform(base_lc,R_Edad=ifelse(is.na(R_Edad)==TRUE,'N
O DISPONIBLE',R_Edad))

base_lc <-
transform(base_lc,R_Salario=ifelse(is.na(R_Salario)==TRUE
,'NO DISPONIBLE',R_Salario))

#%%%%%          Modelo Clúster          %%%%#

library(fpc)

base_lc <- base_lc[,c(2:20)]

cc=discrete.recode(base_lc,xvarsorted=FALSE,continuous=c
      (1,15,17,19),discrete=c(2:14,16,18))

summary(cc)

str(cc)

#numgrupos <-
flexmixedruns(cc$data,continuous=4,discrete=15,simruns=5,
n.cluster=2:10,allout=FALSE)

#%%   Iteraciones para detector el número de grupos   %%#

set.seed(123123)
```

```
fcc=flexmix(cc$data~1,k=10,
model=lcmixed(continuous=4,discrete=15,ppdim=c(25,117,2,
16,40,5,9,1242,520,814,4,300,2638,8,3),diagonal=TRUE,pre
d.ordinal=TRUE,printlik=TRUE),
control=list(iter.max=10, verbose=3))

summary(fcc)


set.seed(123123)

fcc1=flexmix(cc$data~1,k=8,
model=lcmixed(continuous=4,discrete=15,ppdim=c(25,117,2,
16,40,5,9,1242,520,814,4,300,2638,8,3),diagonal=TRUE,pre
d.ordinal=TRUE,printlik=TRUE), control=list(iter.max=8,
verbose=3))

summary(fcc1)


set.seed(123123)

fcc2=flexmix(cc$data~1,k=4,
model=lcmixed(continuous=4,discrete=15,ppdim=c(25,117,2,
16,40,5,9,1242,520,814,4,300,2638,8,3),diagonal=TRUE,pre
d.ordinal=TRUE,printlik=TRUE), control=list(iter.max=4,
verbose=3))

summary(fcc2)


set.seed(123123)

fcc3=flexmix(cc$data~1,k=6,
model=lcmixed(continuous=4,discrete=15,ppdim=c(25,117,2,
16,40,5,9,1242,520,814,4,300,2638,8,3),diagonal=TRUE,pre
d.ordinal=TRUE,printlik=TRUE), control=list(iter.max=6,
verbose=3))

summary(fcc3)

print(fcc)

#%%                    base final                    %%#

b=data.frame(fcc@cluster)

A=cbind(base_lc,b)

base_grupos <- A

base_grupos                                          <-
transform(base_grupos,des_cedula=row.names(base_lc))
```

H. Simulación Teorema Límite Central

```
# TEOREMA CENTRAL DE LIMITE CENTRAL

simularTLC <- function(n_muestras, m_repeticiones, poblacion){
  df <- data.frame()
  for(i in n_muestras) {
    col <- c()
    for(j in m_repeticiones) {
      repeticion <- 1:j
      contadorDeMuestra <- j
      medias <- c()
      while(contadorDeMuestra > 0) {
        bucket <- sample(poblacion, i, replace = TRUE)
        xbar <- mean(bucket)
        medias <- c(medias, xbar)
        contadorDeMuestra <- contadorDeMuestra - 1
      }
      sbar <- sd(medias)
      col <- cbind(repeticion, medias, sbar, i, j)
      df <- rbind(df, col)
    }
  }
  names(df) <- c("repeticion", "media", "desviaicon", "n", "m_rep")
  return(df)
}
```