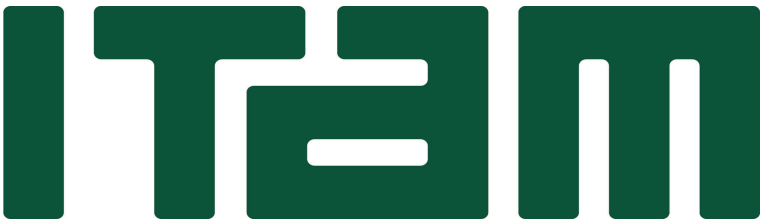
Instituto Tecnológico Autónomo de México



Clasificación no supervisada utilizando un modelo bayesiano de mezclas

TESIS

QUE PARA OBTENER EL TÍTULO DE

Licenciada en Actuaría

PRESENTA

Montserrat Vizcayno García

ASESOR

Dr. Juan Carlos Martínez Ovando

MÉXICO, D.F. 2019

"Con fundamento en los artículos 21 y 27 de la Ley Federal de Derecho de Autor y como titular de los derechos moral y patrimonial de la obra titulada "**Clasificación no supervisada utilizando un modelo bayesiano de mezclas**", otorgo de manera gratuita y permanente al Instituto Tecnológico Autónomo de México y a la biblioteca Raúl Baillères Jr., autorización para que fijen la obra en cualquier medio, incluido el electrónico, y la divulguen entre sus usuarios, profesores, estudiantes o terceras personas, sin que pueda percibir por tal divulgación una prestación"

Montserrat Vizcayno García

_____

Fecha

_____

Firma

Índice general

Índice de figuras

Índice de figuras

Índice de tablas

Introducción

Objetivo de la Tesis

La motivación para realizar este trabajo, surge de la necesidad de segmentar a los clientes de una casa de empeño, con base en información de las operaciones que han realizado anteriormente con la empresa, con el propósito de que dichas agrupaciones proporcionen información sobre el nivel de riesgo de los clientes para ofrecerles distintos tipos de productos, además de ayudar en la gestión de la cartera de empeño de la empresa.

Con base en esta motivación, el objetivo de este trabajo de tesis es desarrollar un algoritmo de clasificación para agrupar observaciones. En este caso, dado que la información no cuenta con una segmentación predefinida, se busca clasificar a los clientes en terminos de la similaridad de sus propiedades o

características, buscando que el grado de asociación entre las observaciones de un grupo sea máximo; a este tipo de segmentación se le conoce como clasificación no supervisada.

Tomando en cuenta el objetivo planteado y dadas las características de la información disponible, se definió realizar una clasificación no supervisada mediante un modelo bayesiano de mezclas, y con base en éste construir un algoritmo capaz de mostrar como se segmentan los clientes K grupos distintos.

El algoritomo, mediante iteraciones de un modelo Gibbs Sampler, estima las distribuciones predictivas que determinan las probabilidades de pertenencia de un cliente a cada uno de los K grupos, para después generar una segmentación medianite su asignación al grupo cuya pertenencia es mas probable.

Estructura de la Tesis

La estructura de este trabajo de tesis está compuesta de seis capítulos. El primer capítulo da una breve introducción a los modelos de clasificación no supervisada y sus distintos enfoques. Más adelante en el segundo capítulo, se exponen las bases y los principios del paradigma bayesiano de inferencia. En

el tercer capítulo, describe el modelo de clasificación empleando una mezcla de modelos para datos mixtos (discretos y continuos). Una vez expuesta la metodología, en el capítulo cuatro se procede a desarrollar un algoritmo para enfrentar al modelo con un conjunto de datos. El quinto capítulo presenta los resultados de la aplicación práctica sobre una base de datos de una empresa de empeño, cuyo objetivo es segmentar los clientes en distintos grupos con base en la información proporcionada por cinco variables previamente seleccionadas en un análisis exploratorio. Por último en el sexto capítulo, se describen las conclusiones acerca del ejercicio práctico y los resultados del modelo. En este capítulo se discute la viabilidad de utilizar un modelo de clasificación no supervisada de esta índole.

Capítulo 1

Preliminares

La segmentación de un conjunto de datos consiste en descubrir patrones comunes subyacentes, no directamente observables, que con base en un criterio, se definen similitudes dentro de un grupo de datos y diferencias entre los diferentes grupos de datos. Basandose en estos patrones identificados, se definen J grupos, y se crean reglas de asociación con las cuales la observación x_i es asignada a un grupo $C_j \in \{1, 2, ..., J\}$. Dichas reglas de asignación pueden plantearse bajo dos enfoques propuestos para resolver el problema que aborda la clasificación no supervisada; el primero y más conocido es mediante argumentos geométricos o distancias (hay distintitos tipos de distancias), mientras que la segunda, en la que se basa este trabajo, utiliza distribuciones

o modelos de probabilidad para determinar dicha similitud. Ambos enfoques buscan lograr que los elementos asignados a un grupo posean características similares, y sean diferentes con respecto a los objetos en los otros grupos; es decir, que sean homogéneos dentro del grupo y heterogéneos entre sí. En las siguientes secciones, se explica de forma mas detallada en que consisten estos enfoques.

1.1. Segmentación basada en distancias

En la escuela tradicional, la asignación de las observaciones x_i a una clase C_j se define con base en distancias.

$$x_i \in C_{j^*} \Leftrightarrow j^* = \text{argmin}_j \, d(x_i, C_j).$$

Dicha asignación depende directamente de la cercanía, en términos de distancia, de una observación a un punto dentro del grupo.

Por ejemplo, para segmentar un conjunto de datos $x_1, ...x_n$ se encuentran dos grupos (J=2) y se define la regla de asignación como el mínimo de la distancia que hay del punto x_i al centro del grupo uno (C_1) y la distancia que hay del punto x_i al centro del grupo dos (C_2).

$$x_i \in C_1 \Leftrightarrow \{d(x_i, C_1) < d(x_i, C_2)\}$$

donde C_1 y C_2 son los centroides de cada grupo.

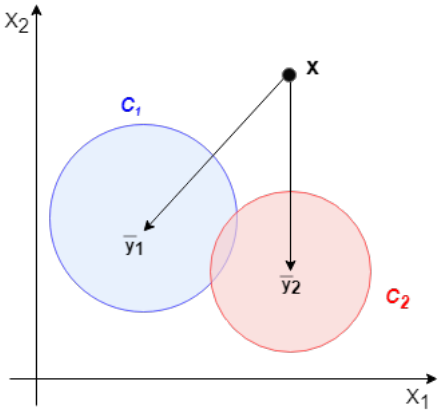


Figura 1.1: Segmentación basada en distancias

En la figura 1.1, se asigna la observación x_i al grupo C_2 ya que éste es el que posee la distancia mínima entre el centroide $\bar{y}_2$ y la observación x.

Este enfoque funciona siempre y cuando los datos sean escalares, pero en caso de que sean, por ejemplo, de tipo catégorico o de conteo, se vuelve un reto incorporar este tipo de variables, ya que no hay una forma para medir distancias a estas variables.

1.2. Segmentación basada en probabilidad

En este caso, la aplicación práctica se basa en el segundo método, el cual utiliza modelos de probabilidad para asignar las observaciones en los distintos grupos. Las reglas de asignación se hacen con base en la probabilidad de que x_i pertenezca a C_j; por lo que los atributos pueden ser escalares, categóricos o texto.

$$x_i \in C_{j^*} \Leftrightarrow j^* = \underset{j \in J}{\operatorname{argmax}}\, P(x_i \in C_j)$$

Para comenzar la segmentación, se suponen J grupos preexistentes en los cuales al menos una observación x_i se encuentra en cada uno de estos grupos $C_1, .., C_J$. Cabe destacar que cada una de las observaciones x_i's sólo pueden pertenecer a un sólo grupo a la vez.

$$C_j = \{x_i : x_i \sim F(x_i|\theta_j)\}$$

donde $\theta_j = T_{F_j}(x)$, y $T_{F_j}(x)$ se define como un atributo, por mencionar algunos ejemplos, la media, la varianza, la mediana, etc. Definimos e_j como el número de observaciones dentro del grupo C_j y se tiene que $P(x_i \in C_j) \approx \frac{e_j}{n}$.

Como se observa en la figura 1.2, un objeto se asigna al grupo con la mayor probabilidad de pertenencia $p(x_i \in C_j)$ con $j = 1, 2$.

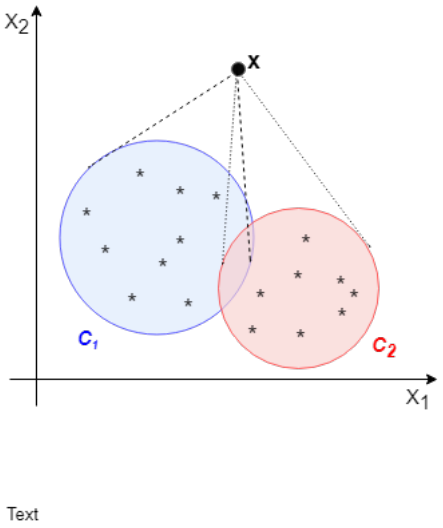


Figura 1.2: Segmentación basada en probabilidad

Bajo este enfoque, surgen varias problemáticas a resolver, que deben ser consideradas al momento de escoger el modelo de clasificación.

Al ser un problema de clasificación no supervisada, se desconoce el número J de grupos o clases en los que se segmentaran los datos. En consecuencia se desconocen cuáles son las x_j's que pertenecen a la clase C_j, así como los parámetros $\theta_1, ...\theta_j$ asociados con $F_1(x_1|\theta_1), ..., F_j(x_i|\theta_j)$, por lo que será necesario estimarlos.

En las próxima secciones se propone un modelo bajo el cual se abordan estas problemáticas.

Capítulo 1: Preliminares

.

Capítulo 2

Inferencia Bayesiana

2.1. Consideraciones de aleatoriedad

En inferencia estadística, hay dos enfoques ampliamente empleados en la práctica: la Inferencia Bayesiana y la Inferencia Frecuentista. La segunda, define la probabilidad como el límite de la frecuencia relativa de un evento en un gran número de intentos, bajo un contexto donde dichos experimentos son aleatorios y están perfectamente definidos. Por otro lado, la Inferencia Bayesiana, es capaz de asignar probabilidades a cualquier evento, aún cuando no hay un proceso aleatorio de por medio; se puede decir que, la probabilidad

se ve como una manera de representar el nivel de creencia sobre un evento, que en ocasiones, dicha creencia es dada una evidencia. Esto se debe a que muchas veces como es este el caso, los datos pueden no ser aleatorios en sí mismos; sin embargo, el desconocimiento de estos, antes de que sean observados nos puede conducir a consideraciones aleatorias.

La manera de cuantificar este desconocimiento es mediante una medida de probabilidad $P(x|\theta)$. Por lo que la forma en la cual se genera un aprendizaje sobre esta medida de probabilidad $P(x|\theta)$ para θ, es a partir de los datos, asociando variables aleatorias $X_1, ..., X_n \sim P(x|\theta)$ con los datos $x_1, ...x_n$; es decir, se supone que $X_1 = x_1, ..., X_n = x_n$.

La Inferencia bayesiana, como se menciona arriba, involucra un proceso de aprendizaje que consiste en modificar las creencias iniciales de los parámetros que fueron definidos antes de observados los datos, por un conocimiento posterior actualizado, que combine tanto el conocimiento previo como la información disponible (Bernardo and Smith, 2001). En otras palabras, un parámetro de valor desconocido θ se representa mediante la asignación de una medida de probabilidad $\pi(\theta)$ que se define con base en el nivel de información que se conoce de este parámetro.

En las siguientes secciones se hablará sobre los conceptos básicos para poder llevar a cabo el proceso de inferencia estadística bajo el enfoque bayesiano.

2.2. Intercambiabilidad

Al relajar el supuesto de independencia, se introduce el concepto de **intercambiabilidad**, el cual reconoce que el orden de las observaciones es invariante ante permutaciones de sus índices; es decir, toda la información relevante está contenida en los valores de las x_i's, de forma que sus índices no proporcionan información alguna. Obsérvese que el concepto de intercambiabilidad generaliza el de independencia condicional: un conjunto de observaciones independientes idénticamente distribuidas (iid) son siempre un conjunto de observaciones intercambiables (Bernardo, 1998).

Entonces, $\{X_j\}_{j=1}^{\infty}$ son intercambiables, si $\forall n < \infty$,

$$P(X_1 = x_1, ..., X_n = x_n) = P(X_{\sigma(1)} = x_1, ..., X_{\sigma(n)} = x_n) \qquad (2.1)$$

donde $\{\sigma(1), ..., \sigma(n)\}$ es cualquier permutación de {1,2,...,n}. La intercambiabilidad implica que la probabilidad es invariante al orden de las variables aleatorias y/o de los datos.

2.3. Teorema de Representación

Utilizando en concepto de intercambiabilidad, de Finetti demuestra su famoso teorema de representación para variables dicotómicas (De Finetti, 1930). En este caso en particular, la intercambiabilidad identifica las observaciones como una muestra aleatoria de un modelo probabilístico específico (Bernoulli) y garantiza la existencia, más no unicidad, de una distribución inicial sobre su parámetro. En el caso general, para variables aleatorias de cualquier rango y dimensión, la intercambiabilidad identifica las observaciones como una muestra aleatoria de algún modelo probabilístico y garantiza la existencia de una distribución inicial sobre el parámetro que lo describe (Bernardo, 1998).

Teorema de representación de Finetti. Si $\{X_j\}_{j=1}^{\infty}$ son variables aleatorias intercambiables, entonces existe un objeto estocástico θ, tal que:

$$P(X_1 = x_1, ..., X_n = x_n) = \int_{\Theta} \prod_{j=1}^{n} P(X_j = x_j|\theta)\pi(\theta)d\theta, \qquad (2.2)$$

donde $\theta \in \Theta$ se define como el límite (cuando $n \longrightarrow \infty$) de una función de las X_j's, y $\pi(\theta)$ es la función de distribución inicial sobre Θ.

En otras palabras, si la secuencia de observaciones es intercambiable, cual-

quier subconjunto de éstas, es una muestra aleatoria del modelo $P(X_j = x_j|\theta)$ y existe una distribución inicial $\pi(\theta)$ que describe la información inicial disponible del parámetro θ (Bernardo and Smith, 2001).

El teorema de representación permite ahondar en dos de los conceptos fundamentales en el aprendizaje estadístico via la actualización de la información; el primer concepto es la verosimilitud, que mediante la información dada por un conjunto de observaciones, permite que se pueda inferir el valor de los parámetros de una función bajo un modelo estadístico definido. El segundo concepto es la función de distribución previa, que antes de tomar en cuenta los datos, utiliza la información disponible para hacer inferencia sobre éstos.

Ambos conceptos son descritos a detalle en las siguientes seccioes.

2.4. Verosimilitud

En general, para realizar el aprendizaje acerca de un modelo $P(X = x|\theta)$ a partir de un conjunto de datos observados, $x_1, ..., x_n$, con $x_j \in \mathbb{R}^p$, se supone un modelo estocástico n-dimensional, tomando como supuesto adicional que son independientes e idénticamente distribuidos (iid), $(x_j, j = 1, ..., n)$.

Por lo que se define un modelo paramétrico $P(X = x) = F(x|\theta)$, en el cual $F(x|\theta)$ es una función de distribución dada. [1]. Utilizando el enfoque bayesiano, el modelo se puede ver como,

$$P(x_1, ..., x_n) = \int \prod_{i=1}^{n} P(x_i|\theta)\pi(d\theta)$$

donde, θ es el parámetro que indiza dicha distribución y toma valores en el espacio parametral $\Theta \in \mathbb{R}^p$ (con $p < \infty$); es decir, que hay un número finito de parámetros [2]. De esta manera tenemos bajo el supuesto de independencia, lo siguiente:

$$P(x_1, ..., x = n|\theta) = \prod_{j=1}^{n} P_j(x_j|\theta) \tag{2.3}$$

donde $X_j = (X_{j1}, ..., X_{jp})$, es un vector p-dimensional.

Al considerar $P(X_j = x_j)$ como una función de distribución o densidad, de las observaciones dado el parámetro θ, $P(X_j = x_j) = F(X_j = x_j|\theta)$, obtenemos la siguiente función de verosimilitud, que es aplicable a los casos

[1] ya sea por una función de masa de probabilidad en el caso discreto, o por una función de densidad en el caso absolutamente continuo

[2] al contrario del caso no paramétrico, donde el número de parámetros es infinito ($\Theta \in \mathbb{R}^{\infty}$)

anteriormente mencionados,

$$L(x_1, ..., x_n; \theta) = f(x_1, ..., x_n|\theta) = \prod_{j=1}^{n} f(x_j|\theta) \tag{2.4}$$

2.5. Distribución inicial o previa

La distribución inicial o previa, $\pi(\theta)$, se define como la medida de probabilidad sobre Θ que describe nuestra creencia acerca del parámetro θ, con base en el nivel de información disponible, que proviene no necesariamente de los datos.

Existen distintas maneras de obtener la distribución inicial para θ, en algunas situaciones es posible basarse en información que resulta a partir de evidencia acumulada de experimentos pasados. Asimismo, la previa puede determinarse de manera subjetiva con base en la experiencia de un experto. En el caso de no contar con información disponible, se puede recurrir a una distribución previa no informativa, la cual expresa sin distinción sobre diferentes valores específicos del parámetro (Congdon, 2007).

En la práctica, hay métodos que resultan tener una conveniencia analítica com es el caso cuando la especificación de $\pi(d\theta)$, es estructuralmente semejante a $p(x_1, ..., x_n|\theta)$, vista como una función de θ. Siendo así que $\pi(d\theta)$ y

$p(x_1, ..., x_n|\theta)$ definen la pareja conjugada que lleva a $\pi(\theta)$ y $\pi(\theta|x_1, ..., x_n)$ a ser estructuralmente semejantes.

Por ejemplo, si la función de verosimilitud es de la familia Gaussiana, entonces al elegir una distribución Gaussiana como distribución previa del parámetro, que en este caso es la media, nos asegurará que la distribución final sea una distribución Gaussiana. En general todas las distribuciones de probabilidad de la familia exponencial cuentan con distribuciones previas conjugadas (Congdon, 2007).

2.6. Teorema de Bayes

El Paradigma Bayesiano se basa en un proceso de aprendizaje, en el cual los datos añaden nueva información al conocimiento previo y de esta forma, se actualizan las creencias sobre los parámetros de interés. Para realizar inferencias sobre cierta hipótesis, bajo un enfoque bayesiano, se deben especificar las creencias anteriores con base en información disponible antes de haber observado los datos y así describir el comportamiento del parámetro θ mediante una distribución inicial $\pi(\theta)$ definida como una medida de probabilidad sobre Θ. Esta información es entonces combinada con los datos para

producir la distribución a posteriori o final, $\pi(\theta|x_1, .., x_n)$, que expresa lo que se conoce de los parámetros, una vez que se introdujeron los datos. Se utiliza el teorema de Bayes como un mecanismo para combinar la información *a priori*, $\pi(\theta)$, con la información proporcionada por los datos, $P(x_1, ..., x_n|\theta)$; como se mencionó anteriormente, ésta última es la función de verosimilitud,

$$\pi(\theta|x_1, ..., x_n) = \frac{P(x_1, ..., x_n|\theta) \cdot \pi(\theta)}{P(x_1, ..., x_n)} \tag{2.5}$$

donde el denominador, $P(x_1, ..., x_n) = \int_\Theta P(x_1, ..., x_n|\theta) \cdot \pi(\theta) d\theta$, es una integral sobre todos los valores de θ del producto de la función de verosimilitud y la previa del parámetro θ y se toma como una constante de normalización para asegurar que $\pi(\theta|x_1, ..., x_n)$ sea una función densidad propia.

Simplificando, el teorema de Bayes puede ser expresado de la siguiente manera,

$$\pi(\theta|x_1, ..., x_n) \propto P(x_1, ..., x_n|\theta) \cdot \pi(\theta) \tag{2.6}$$

donde $\propto$ denota una relación de proporcionalidad, $P(x_1, ..., x_n)^{-1}$

2.7. Distribución final o posterior

La distribución posterior se obtiene aplicando el teorema de Bayes. Se combina la información inicial del parámetro $\theta = (\theta_1, ..., \theta_q)$, mediante la distribución $\pi(\theta) = \pi(\theta_1, ..., \theta_q)$ donde $(\theta \in \Theta \subseteq \mathbb{R}^q)$, y la distribución de las observaciones, mejor conocida como función de verosimilitud, todo esto, bajo el supuesto que las X_j's son intercambiables y condicionalmente independientes dado θ. Por lo que la distribución final de θ es:

$$\pi(\theta_1, ..., \theta_q | x_1, ..., x_n) = \frac{\prod_{j=1}^{n} P(X_j = x_j | \theta_1, ..., \theta_q) \cdot \pi(\theta_1, ..., \theta_q)}{\int_\Theta \prod_{j=1}^{n} P(X_j = x_j | \tilde{\theta_1}, ..., \tilde{\theta_q}) \cdot \pi(\tilde{\theta_1}, ..., \tilde{\theta_q}) d\tilde{\theta_1}, ..., \tilde{\theta_q}} \tag{2.7}$$

donde el denominador se define como la constante de normalización C, tal que $\pi(\theta_1, ..., \theta_q | x_1, ..., x_n)$ sea una densidad propia; es decir, $\int_\Theta \frac{1}{C} \cdot \prod_{j=1}^{n} P(X_j = x_j | \theta) \cdot \pi(\theta) d\theta = 1$. En la práctica, se conoce explícitamente sólo en términos del numerador (Congdon, 2007),

$$\pi(\theta_1, ..., \theta_q | x_1, ..., x_n) \propto \prod_{j=1}^{n} P(X_j = x_j | \theta) \cdot \pi(\theta) d\theta \tag{2.8}$$

El paso siguiente sería el cálculo de la distribución final, que en el caso de ser conjugados, resulta ser de una forma paramétrica conocida; y de no ser así, se recurre a métodos de aproximación, debido a la a imposibilidad de

obtener la constante de normalización.

2.8. Distribución predictiva

La distribución predictiva define como será el comportamiento de una nueva observación x_{n+1} con base en los datos observados $x_1, ..., x_n$.

$$\begin{aligned} P(X_{n+1} = x_{n+1} | X_1 = x_1, ..., X_n = x_n) &= \frac{P(x_1, ..., x_n, x_{n+1})}{P(x_1, ..., x_n)} \\ &= \frac{\prod_{j=1}^{n} P(X_j = x_j) \cdot P(X_{n+1} = x_{n+1})}{\prod_{j=1}^{n} P(X_j = x_j)} \\ &= P(X_{n+1} = x_{n+1}) \\ &= F(x_{n+1}|\theta) \quad ? \end{aligned} \tag{2.9}$$

Como se observa en (2.9), existe un problema de predictibilidad, ya que la distribución de la nueva observación no estaría tomando en cuenta la información proporcionada por las observaciones anteriores. Para obtener la distribución predictiva bajo el enfoque bayesiano es necesario hacer uso de los conceptos definidos en la sección anterior.

$$
\begin{aligned}
P(X_{n+1} = x_{n+1}|x_1, ..., x_n) &= \frac{P(X_{n+1} = x_{n+1}|x_1, ..., x_n, x_{n+1})}{P(x_1, ..., x_n)} \\
&= \frac{\int_\Theta \prod_{j=1}^n P(X_j = x_j|\theta) \cdot P(X_{n+1} = x_{n+1}|\theta) \cdot \pi(\theta) d\theta}{\int_\Theta \prod_{j=1}^n P(X_j = x_j|\tilde{\theta}) \cdot \pi(\tilde{\theta}) d\tilde{\theta}} \\
&= \int_\Theta P(X_{n+1} = x_{n+1}|\theta) \cdot \frac{\prod_{j=1}^n P(X_j = x_j|\theta) \cdot \pi(\theta) d\theta}{\int_\Theta \prod_{j=1}^n P(X_j = x_j|\tilde{\theta}) \cdot \pi(\tilde{\theta}) d\tilde{\theta}} \\
&= \int_\Theta P(X_{n+1} = x_{n+1}|\theta) \cdot \pi(\theta|x_1, ..., x_n) d\theta
\end{aligned}
\tag{2.10}
$$

donde $\pi(\theta|x_1, ..., x_n)$ es la distribución final o posterior de θ dado $x_1, ..., x_n$, que se define como la creencia sobre el comportamiento de θ después de observar los datos. La ecuación 2.10 se puede representar como:

$$
P(X_{n+1} = x_{n+1}|x_1, ..., x_n) = E_{\theta|x_1,..,x_n}[P(X_{n+1} = x_{n+1}|\theta)] \tag{2.11}
$$

De esa manera la predicción del comportamiento de los nuevos datos tambien considera la información proporcionada por los datos observados anteriormente.

2.8.1. Métodos de Aproximación

El problema de cálculo en la inferencia Bayesiana se reduce a la resolución de integrales que no tienen una solución analítica viable, por lo que se requieren algoritmos de aproximación numérica. Existen distintos tipos de métodos de aproximación como son la aproximación de Laplace o el método Monte Carlo vía Cadenas de Markov (MCMC) (Gelman and Stern, 2014), este último se describe a continuación.

I. Monte Carlo vía Cadenas de Markov

Para la aproximación de integrales, se utilizan técnicas de simulación estocástica. Uno de los métodos mas conocido se conoce como Monte Carlo, que consiste en generar simulaciones de los datos para aproximar la integral que en su caso puede ser una función de distribución.

Supongamos que se tiene una integral de la forma: $\int h(\theta)\pi(\theta)d\theta$, donde $\theta \in \Theta \subset \mathbb{R}^p$ es una variable aleatoria, $\pi(\theta)$ es la densidad de θ que está condicionada a la información relevante disponible al momento del análisis, y $h(\cdot)$ es una función real conocida e integrable con respecto a π.

Ahora, si se genera una muestra de simulaciones de tamaño T, $(\theta_1^{(t)}, ..., \theta_q^{(t)})$,

con $(t = 1...T)$ independientes e idénticamente distribuidas (iid) de la distribución $\pi(\theta)$, podemos aproximar el valor de la integral.

Lo puede ser interpretado como el valor esperado de h sobre π, $E_\pi[h(\theta)] = \int h(\theta)\pi(\theta)d\theta$, aproximado mediante un promedio de las simulaciones obtenidas, y que son evaluadas en h,

$$\hat{E}_\pi[h(\theta)] = \frac{1}{T}\sum_{i=1}^{T} h(\theta^{(i)})$$

Este estimador, que se conoce como el estimador de Monte Carlo, que es insesgado y converge casi seguramente con el valor de la integral de interés (Martínez Ovando, 2004).

Entonces, retomando la sección anterior, tenemos que la función de distribución predictiva (2.11) se puede ver de la siguiente manera:

$$E_{\theta|X_1,..,X_n}(P(X_{n+1}|\theta)) \approx \frac{1}{T}\sum_{t=1}^{T} P(X_{n+1}|\theta_1^{(t)}, ..., \theta_q^{(t)}) \tag{2.12}$$

donde la aproximación de (2.12), se puede interpretar como el promedio empírico de $P(X_{n+1}|\theta)$.

Bajo el contexto bayesiano, generalmente conocemos la densidad $\pi(\cdot)$, salvo por una constante de normalización, que usualmente resulta ser difícil de calcular, lo que lleva al problema inicial de resolver una integral; así pues, se

propone relajar el supuesto de *iid* de Monte Carlo.

En consecuencia a esta propuesta, $\theta_1^{(t)}, ..., \theta_q^{(t)}$ dependerá de $\theta_1^{(t-1)}, ..., \theta_q^{(t-1)}$; es decir, que el valor de $\theta_1^{(t)}, ..., \theta_q^{(t)}$ estará influenciado por la información que proporcione $\theta_1^{(t-1)}, ..., \theta_q^{(t-1)}$ y así sucesivamente para cada una de las T simulaciones. Esto se logra mediante una cadena de Markov, con una distribución de transición $g(\theta^{(t)}|\theta^{(t-1)}, X_1, ..., X_n)$. La unión del método de cadenas de Markov con el método de Monte Carlo, se denomina como el método Monte Carlo vía Cadenas de Markov (MCMC). Donde la cadena de Markov es un proceso estocástico $\{\theta^{(0)}, \theta^{(1)}, \theta^{(2)}, ...\}$, en el cual el próximo estado (t+1) depende solamente del estado actual (t) y no de los estados anteriores.

La idea central de este método es construir una cadena de transición definida por el *kernel* de transición que tenga a la distribución objetivo, $\pi(.)$,como invariante; es decir, que si $\theta^{(t)} \sim \pi$ implica que $\theta^{(t+1)} \sim \pi$.

El *kernel* $K : \Theta \times B(\Theta) \to [0, 1]$, es la función de transición de los estados, que denota la probabilidad de transición, $K(\theta, \theta') = P[\theta^{(t+1)} = \theta' | \theta^{(t)} = \theta]$, $\forall(\theta, \theta') \in \Theta$ entre las iteraciones t y t+1 (Martínez Ovando, 2004). Además, la cadena deberá cumplir con ciertas condiciones de regularidad que pueden ser revisadas a profundidad en (Congdon, 2007).

II. *Gibbs Sampler*

Un algoritmo en particular de cadenas de Markov que ha sido útil para problemas multidimensionales es *Gibbs Sampler*, que está definido en términos de subvectores del vector paramétrico θ.

Este método implica hacer una actualización parámetro a parámetro de cada uno de los componentes $\theta_1^{(t)}, ..., \theta_q^{(t)}$, mediante un muestreo sucesivo de las distribuciones condicionales, que al completarse nos da la transición de $\theta^{(t-1)}$ a $\theta^{(t)}$ (Gelman and Stern, 2014).

Este algoritmo genera una secuencia de números autocorrelacionados que cumplen con las condiciones de regularidad, que eventualmente olvida los valores iniciales, $\theta_1^{(0)}, ..., \theta_q^{(0)}$, usados para la cadena y que terminan por converger en una distribución estacionaria. De manera que, $\{(\theta_1^{(t)}, ..., \theta_q^{(t)})\}_{t=1}^{t}$ se define como una cadena de Markov (Congdon, 2007).

Capítulo 3

Mezclas de distribuciones

3.1. Modelos de mezclas

El método de clasificación bajo modelos de mezclas, surge en los años sesenta como consecuencia del desarrollo de procedimientos para estimar densidades flexibles, que permitan modelar datos de distintos tipos (continuos, categóricos, ordinales, etc.), y se definen de la forma,

$$f(x) = \sum_{j=1}^{J} w_j f(x|\theta_j)$$

De manera incidental, al estimar $(w_i, \theta_j)_{i=1}^{J}$ se calcula $p(x_i \in C_j) = w_j$, que permite utilizar este tipo de modelos en procedimientos de segmentación.

Por consiguiente, la estructura de los modelos de mezclas es utilizada en problemas de clasificación no supervisada, asumiendo que cada una de las x_i's puede tener una distribución f_j con probabilidad w_j.

Dependiendo del escenario que se plantee, la meta puede ser reconstruir los grupos a los que pertenecen las observaciones para proveer de estimadores para los parámetros de los diferentes grupos, o incluso, estimar el número de grupos.

Las distribuciones mixtas pueden contener un modelo finito o infinito de componentes, que posiblemente pueden ser de distintos tipos de distribuciones, y describen distintas características de los datos.

En este caso, nos enfocaremos al modelo de mezclas con un número finito de componentes, que se define así,

$$P(X = x) = \sum_{j=1}^{J} w_j f_j(x|\theta_j) \tag{3.1}$$

donde w_j es la probabilidad de pertenecer al componente o clase C_j, y

$$\sum_{j=1}^{J} w_j = 1$$

Sin embargo, la manera en como esta representado el modelo en 3.1 vuelve complicado derivar el estimador de máxima verosimilitud (cuando existe) y los estimadores bayesianos.

Por ejemplo, si se considera el caso de n observaciones *iid*, $x = (x_1, ...x_n)$ y definimos los parámetros $w = (w_1, ...w_J)$ y $\theta = (\theta_1, ..., \theta_J)$, se puede observar que a pesar de utilizar previas conjugadas para cada parámetro, si se desea obtener la distribución previa de manera explícita, es necesario que se realice la expansión de la verosimilitud en k^n términos, que para la práctica resulta ser muy costoso computacionalmente hablando (Marin Jean-Michel, 2005).

3.2. Variables Latentes

Como se menciona la sección anterior, estimar la verosimilitud bajo un modelo de mezclas puede resultar complicado, así que en los años setenta se plantea la introducción una variable de asignación, mejor conocida como variable latente, $\underline{z} = (z_1, ..., z_n)$, que identifican a que componente j, $(j = 1, ...J)$ pertenece cada una de las observaciones $x = (x_1, ..., x_n)$.

$$z_i = \begin{cases} j, & \text{si } x_i \in C_j \\ 0, & \text{si } e.oc. \end{cases}$$

Donde $Z_i \sim M_j(1; w_1, ...w_J)$.

Ahora, $P(x_i)$ puede definirse en términos de (x_i, z_i) dados los parámetros

$(\underline{\theta}, \underline{w})$,

$$\begin{aligned} P(x_i) &= \sum_{z_i \in \mathbb{Z}} P(x_i, z_i | \underline{\theta}, \underline{w}) \\ &= \sum_{j=1}^{J} P(x_i, z_i | \underline{\theta}, \underline{w}) \\ &= \sum_{j=1}^{J} P(x_i | z_i = j, \underline{\theta}, \underline{w}) \cdot P(z_i = j | \underline{\theta}, \underline{w}) \\ &= \sum_{j=1}^{J} f(x_i | \theta_j) \cdot w_j \cdot \mathbb{1}(z_i = j) \\ &= \sum_{j=1}^{J} w_j f(x_i | \theta_j) \end{aligned} \tag{3.2}$$

Esto permite redefinir la función de verosimilitud en términos de la variable latente, para la clase j.

$$L(\underline{\theta}, \underline{w}, \{z_i\}_{i=1}^n | \{x_i\}_{i=1}^n) = \prod_{i=1}^{n} w_j f(x_i | \theta_j) \mathbb{1}_{(z_i = j)}$$

donde se define $p(z_i = j = w_j, \forall i = 1, ..., n$ como la función previa y así tener una distibución posterior para cada observación x_i,

$$\begin{aligned} P(x_i, z_i | \underline{\theta}, \underline{w}) &= P(x_i | z_i = j, \underline{\theta}, \underline{w}) P(z_i = j | \underline{w}) \\ &= f(x_i | \theta_j) \cdot w_j \cdot \mathbb{1}(z_i = j) \end{aligned} \tag{3.3}$$

Lo que muestra que la variable latente se encuentra integrada dentro del modelo y no es necesario realizar ninguna otra transformación o procedimineto para calcular los demás parámetros.

Ahora, la distribución posterior de forma general es la siguiente,

$$\pi(\theta, w|X = x) \propto (\prod_{i=1}^{n} \sum_{j=1}^{J} w_i f_j(X_i = x_i|\theta_j))\pi(\theta, w) \tag{3.4}$$

Por lo que se define Z como el conjunto de todos los J_n vectores de asignación z, que puede ser descomponpuesta en una partición de J conjuntos. Para un vector de asignación dado $(n_1, ..., n_J)$ donde $n_1 + ... + n_J = n$ definimos el conjunto,

$$Z_i = \{\underline{z} : \sum_{i=1}^{n} \mathbb{1}_{z_i=1} = n_1, ..., \sum_{i=1}^{n} \mathbb{1}_{z_i=J} = n_J\}$$

que consiste en todas las asignaciones dadas por una partición, que en este caso es determinada por el vector de asignación $(n_1, ..., n_J)$. Existe un número r de soluciones enteras no negativas de las n observaciones en las J clases, en las que se cumple que $\sum_{j=1}^{J} n_j = n$.

$$r = \binom{n + k - 1}{n} \tag{3.5}$$

De esa manera tenemos la partición $Z = \cup_{i=1}^{r} Z_i$. Aunque el número total de elementos de Z no sea manejable en términos computacionales J^n, el número de conjuntos de particiones es mucho más manejable al ser del orden $\frac{n^{k-1}}{(k-1)!}$

Ahora, la distribución posterior se puede descomponer de esta manera,

$$\pi(\theta, w|x) = \sum_{i=1}^{r} \sum_{\underline{z} \in Z_i} p(\underline{z}) \pi(\theta, w|x, \underline{z}) \tag{3.6}$$

donde $p(\underline{z})$ se define como la probabilidad posterior, dada la asignación $\underline{z}$. Esta descomposición hace que la distribución posterior le asigne una probabilidad posterior $p(\underline{z})$ a cada posible asignación $\underline{z}$ de los datos, para luego construir la distribución posterior de los parámetros, condicional a esa asignación (Marin Jean-Michel, 2005).

Entonces retomando lo anterior $p(z_j = k) = p(x_j \in C_k) = p_k$ para $k = 1, ..., J$, se define la verosimilitud extendida como,

$$L(\theta_j, w_j)_{j=1}^{J}, (z_i)_{i=1}^{n} | x_1, ...x_n) = \prod_{i=1}^{n} w_j f(x_i|\theta_k) \mathbb{1}_{(x_i \in C_k)} \tag{3.7}$$

en k^n términos, que para la práctica resulta ser muy costoso computacionalmente hablando (George, 1989).

3.3. Mezclas gausianas

Para ejemplificar la complejidad de la estimación del modelo basado en mezclas de distribuciones, consideraremos el caso de una mezcla de normales con dos componentes. El modelo se define de la siguiente manera,

$$\pi(j, \theta_j|x_1, ..., x_n) = \frac{P(x_1, ..., x_n|\theta_j, j) \cdot \pi(\theta_j|j) \cdot \pi(j)}{\sum_{j=1}^{2}\{\int_{\Theta_j} P(x_1, ..., x_n|\theta_j, j) \cdot \pi(\theta_j|j)d\theta_j\}\}\pi(j)} \tag{3.8}$$

Bajo el enfoque bayesiano, la ecuación (3.8) se puede representar como,

$$\pi(j, \theta_j|x_1, ..., x_n) \propto P(x_1, ..., x_n|\theta_j, j) \cdot \pi(\theta_j|j) \cdot \pi(j)$$

donde el denominador de (3.8) es la constante de normalización, y como se ha mencionado, su obtención es muy compleja, por lo que en esta tesis se propone una alternativa a este método.

Primero, definimos a la distribución posterior del parámetro θ y a la función de distribución final del modelo respectivamente,

$$\pi(\theta_j|x_1, ..., x_n, j) = \frac{P(x_1, ..., x_n|\theta_j, j) \cdot \pi(\theta_j, j)}{\int_{\Theta_j} P(x_1, ..., x_n|\theta_j, j) \cdot \pi(\theta_j|j)d\theta_j} \tag{3.9}$$

$$\pi(j|x_1, ..., x_n) = \frac{P(x_1, ..., x_n|j) \cdot \pi(j)}{\sum_{j=1}^{J} P(x_1, ..., x_n|j) \cdot \pi(j)} \tag{3.10}$$

donde $P(x_1, ..., x_n|j) = \int_{\Theta_j} P(x_1, ..., x_n|\theta_j, j) \cdot \pi(\theta_j|j)d\theta_j$.

Ahora, tomando en cuenta que es un modelo de mezclas gausianas, se definen las siguentes funciones de verosimilitud y de distribución inicial de los parámetros,

$$P(x_1, ..., x_n|\theta_j, j) = \prod_{i=1}^{n} N(x_i|\theta_j, 1) \tag{3.11}$$

$$\pi(\theta_j|j) = N(\theta_j|\mu_0, \sigma_0^2 = 1) \tag{3.12}$$

Para obtener la distribución posterior de los parámetros, sustituimos en (3.9) tanto la función de verosimilitud como la distribución inicial de los parámetros.

$$\begin{aligned}\pi(\theta_j|x_1, ..., x_n, j) =& \frac{\prod_{i=1}^{n} N(x_i|\theta_j, 1) \cdot N(\theta_j|\mu_0, \sigma_0^2 = 1)}{\int_{\Theta_j} \prod_{i=1}^{n} N(x_i|\theta_j, 1) \cdot N(\theta_j|\mu_0, \sigma_0^2 = 1) d\theta_j} \\ &\propto \prod_{i=1}^{n} N(x_i|\theta_j, 1) \cdot N(\theta_j|\mu_0, \sigma_0^2 = 1)\end{aligned}$$

y se proceda a expandir los cuadrados con el fin de factorizar con respecto a θ_j,

$$\begin{aligned}=& \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi}} \exp\{-\frac{1}{2}(x_i - \theta_j)^2\} \cdot \frac{1}{\sqrt{2\pi}} \exp\{-\frac{1}{2}(\theta_i - \mu_0)^2\} \\ =& \frac{1}{2\pi} \exp\left\{-\frac{1}{2}(\sum_{i=1}^{n}(x_i - \theta_j)^2 + (\theta_j - \mu_0)^2)\right\} \\ =& \frac{1}{2\pi} \exp\left\{-\frac{1}{2}(\sum_{i=1}^{n} x_i^2 - 2\theta_j n\bar{x} + (n+1)\theta_j^2 - 2\theta_j\mu_0 + \mu_0^2)\right\}\end{aligned}$$

El siguiente paso es sumar y restar el término $\left(\frac{n\bar{x}+\mu_0}{n+1}\right)$ para completar el cuadrado,

$$=\frac{1}{2\pi}\exp\left\{-\frac{1}{2}\left(\theta_j-\frac{n\bar{x}+\mu_0}{n+1}\right)^2\right\}\cdot\exp\left\{-\frac{1}{2}\left(-\left(\frac{n\bar{x}+\mu_0}{n+1}\right)^2+\frac{\mu_0^2+\sum_{i=1}^n x_i^2}{n+1}\right)\right\} \tag{3.13}$$

De esta manera se puede ver que el primer término de (3.13) es el kernel de una distribución normal con media $\frac{n\bar{x}+\mu_0}{n+1}$ y varianza 1,

$$\pi(\theta_j|x_1,...,x_n,j)\sim N\left(\frac{n\bar{x}+\mu_0}{n+1},1\right) \tag{3.14}$$

además, con el propósito de garantizar que $\pi(\theta_j|x_1,...,x_n,j)$ sea una función propia, se calcula su constante de normalización,

$$\left[\frac{1}{\sqrt{2\pi}}\exp\left\{-\frac{1}{2}\left(-\left(\frac{n\bar{x}+\mu_0}{n+1}\right)^2+\frac{\mu_0^2+\sum_{i=1}^n x_i^2}{n+1}\right)\right\}\right]^{-1}$$

Los resultados obtenidos en la parte de arriba, se utlizarán para estimar la función de distribución final del modelo (3.10).

$$\pi(j|x_1,...,x_n)=\frac{\int_{\Theta_j}\prod_{i=1}^n N(x_i|\theta_j,1)\cdot N(\theta_j|\mu_0,\sigma_0^2=1)d\theta_j\cdot\pi(j)}{\sum_{j=1}^2\int_{\Theta_j}\prod_{i=1}^n N(x_i|\theta_j,1)\cdot N(\theta_j|\mu_0,\sigma_0^2=1)d\theta_j\cdot\pi(j)}$$

donde a $\int_{\Theta_j}\prod_{i=1}^n N(x_i|\theta_j,1)\cdot N(\theta_j|\mu_0,\sigma_0^2=1)d\theta_j$ se le llama verosimilitud integrada y se puede sustituir de la siguiente manera,

$$\pi(j|x_1,...,x_n)=\frac{\frac{1}{\sqrt{2\pi}}\exp\left\{-\frac{1}{2}\left(-\left(\frac{n\bar{x}+\mu_0}{n+1}\right)^2+\frac{\mu_0^2+\sum_{i=1}^n x_i^2}{n+1}\right)\right\}\cdot\pi(j)}{\sum_{j=1}^2\int_{\Theta_j}\prod_{i=1}^n N(x_i|\theta_j,1)\cdot N(\theta_j|\mu_0,\sigma_0^2=1)d\theta_j\cdot\pi(j)}$$

En la figura 3.1 se ejemplifica como se verían las distribuciones del modelo de mezclas de normales cuando el número de componentes es $J = 2$.

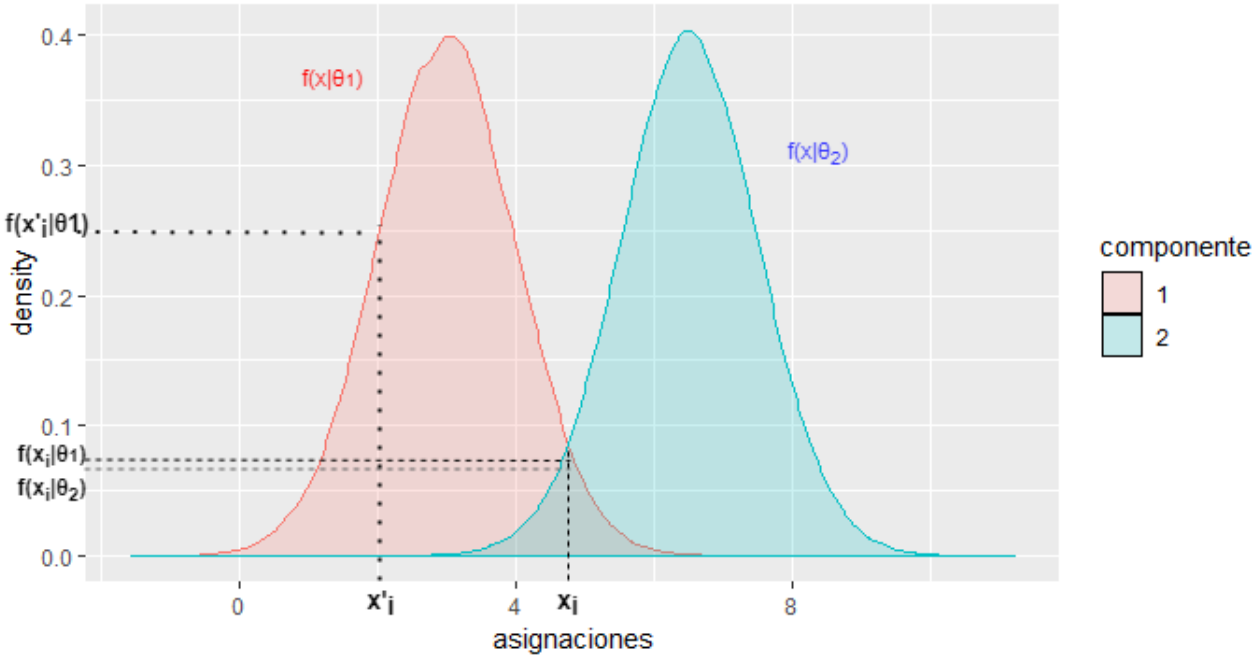


Figura 3.1: Modelo de mezcla de normales con dos componentes

El modelo asigna una regla de pertenencia a los grupos mediante una probabilidad; sin embargo, en los puntos donde las distribuciones se intersectan como se en la figura 3.1, no está claro a que componente pertenece una observación x_i, ya que las probabilidades $f(x_i|\theta_1)$ y $f(x_i|\theta_2)$ son similares, por lo que, tanto la acción de encontrar la distribución del modelo como la

asignación de las medidas de probabilidad, pueden resultar problemático.

Como solución a lo antes mencionado, a demás de la introducción de la variable latente, en los setenta se creó el algoritmo $E-M$, (Dempster et al., 1977), que estima los parámetros $(\underline{\theta}, \underline{w})$ de forma iterativa sobre $\{z_i\}_{i=1}^{n}$. El algoritmo consiste en calcular multiples veces la probabilidad de que la variable latente z_i tome el valor de k $(k = 1, ..., J)$, dada la observación x_i.

$$P(z_i = k|x_i) = \frac{w_k(x_i|\theta_k)}{\sum_{k'=1}^{J} w_{k'} \cdot f(x_i|\theta_{k'})} \tag{3.15}$$

De esta forma se recopila mayor información y se genera una mejor estimación de las probabilidades de asignación.

Gracias a este algoritmo $(E-M)$, se observó que al definir $P(z_i = k|x_i) = P(x_i \in C_k)$ con $k = 1, ..., K$, el modelo de mezclas puede ser utilizado como herramienta de segmentación para problemas de clasificación no supervisada, ya que esta probabilidad se puede ver como una probabilidad de pertenecer a un grupo k, dados los valores de la observación i.

3.4. *Label Switching Problem*

Como se menciona en la sección anterior, el modelo de mezclas no fue desarrollado como una herramienta de segmentación probabilística, de modo que al ser aplicado para este fin, se encontraron ciertas complicaciones como lo es el problema de *label switching*, (Stephens, Stephens).

El término *label switching* se utiliza para describir la invarianza de la función de verosimilitud extendida al momento de reetiquetar a los componentes del modelo de mezclas. En otras palabras, para cualquier permutación σ de 1,...,k, se define la permutación correspondiente al parámetro θ como,

$$\sigma(\theta) = ((\pi_{\sigma(1)}, ..., \pi_{\sigma(k)}), (\theta_{\sigma(1)}, ..., \theta_{\sigma(k)}))$$

La raíz del problema de *label switching*, que surge durante el proceso de estimación del modelo, ya que la función de verosimilitud es la misma para todas las permutaciones de θ. Asimismo, bajo el enfoque bayesiano, la distribución previa $\pi(\theta)$ será la misma para todas las permutaciones si no se cuenta con la información previa que distinga entre los componentes del modelo de mezclas, y trae como consecuencia que la distribución posterior sera simétrica.

Dicha simetría puede causar problemas cuando se busca estimar algún atributo relacionado a los componentes del modelo de manera individual. Visto otra forma, las funciones de densidad predictivas son las mismas para cada componente, de forma que las probabilidades de asignación no son de utilidad para la segmentación de las observaciones en grupos, ya que son las mismas para cada observación $(1/k)$.

De manera similar, al tener la misma distribución posterior, la media de un parámetro dentro de un componente específico será la misma media que se utilice para ese parámetro en los demás componentes del modelo, por lo que en general resulta ser una estimación muy pobre para esos parámetros (Stephens, Stephens).

Se han propuesto diversas alternativas para resolver el problema, como es la de imponer una restricción (*identifiability constraint*) sobre los parámetros; por ejemplo, ordenar las medias (o las varianzas o pesos), que desde el punto de vista Bayesiano, equivale a truncar la distribución previa original. Sin embargo, esto puede llevar a modificar de manera radical el modelo de la distribución previa. Una alternativa es seleccionar una de las $k!$ regiones modales de la distribución posterior y realizar el reetiquetadoçon base en la proximidad a esta región. (Marin Jean-Michel, 2005)

Capítulo 4

Clasificación con datos discretos y continuos

4.1. Inferencia bayesiana en modelos de mezclas

Originalmente, los modelos de mezclas fueron desarrollados para datos categóricos; sin embargo, actualmente se ha propuesto extenderlos a datos mixtos.

Más adelante en la sección 4.4, se demuesta como la función de verosimulitud

marginal puede ser factorizada en el producto de los componentes discretos y continuos para poder ser tratados por separado. En el caso de la distribución de probabilidad de los datos continuos puede deducirse si la funcion de verosimilitud cuenta con una función de distribución previa conjugada, de manera que la distribución conjunta pueda ser determinada analíticamente. (Blomstedt et al., 2015)

Tomando en cuenta todos los conceptos definidos anteriormente, suponemos un conjunto N de n observaciones, $(i \in N)$, caracterizados como vectores d-dimensionales $x^{(i)} = (x_1^{(i)}, ..., x_d^{(i)}) \in \mathbb{R}^d$ y el conjunto de datos se define como $x = (x^1, ..., x^n)^T$. Posteriormente, se lleva a cabo la partición N en k conjuntos no vacíos y que a su vez no se traslapen, representados por $S = \{s_1, ..., s_k\}$ tal que, $s_c \cup_{c=1}^{k} = N$, $s_c \cap s_{c'} = \emptyset \forall c, c' = 1, ..., k$ con $c \neq c'$.

Una vez se define lo anterior, la distribución de mezclas se ve de la siguiente forma:

$$F(x) = P(x \in D)F(x|X \in D) + P(x \in D^C)F(x|X \in D_C), \tag{4.1}$$

donde D es un conjunto de puntos discontinuos con respecto a F, finito y con cardinalidad r.

Aún sin conocer la forma de la distribución del componente discreto, supo-

nemos un vector aleatorio $(Y_1, ..., Y_r)$ de manera que:

$$Y_l = \begin{cases} 1, & \text{si } X = d_l \\ 0, & \text{si } x \neq d_l \end{cases}$$

donde $l = 1, ..., r$. La distribución de probabilidad del componete discreto se define como sigue:

$$f_D(y_1, ..., y_r|\psi_1, ..., \psi_r) = \prod_{l=1}^{r} \psi_l^{y^l} \tag{4.2}$$

donde $\psi_l = P(Y_l = 1$ y $\sum_{i=1}^{r} \psi_l = 1$.

Ahora, para el componente continuo del modelo se define una función de densidad $f_C(x|\lambda)$.

Los datos del grupo C con características j, que suponemos observados, se definen: $x_{cj} = (x^{(1)}, ..., x^{(n_c)})^T$, así pues tenemos la función de verosimilitud,

$$L_{x_{cj}}(w, \psi_1, ...\psi_r, \lambda) = \prod_{i=1}^{n_c} \left[(1-w) \prod_{l=1}^{r} \psi_l^{y_l^{(i)}} + w f_C(x^{(i)}|\lambda) \right] \tag{4.3}$$

donde $w = P(X^{(i)} \in D^c)$, $\forall i = 1, ..., n_c$, asumiendo que las observaciones son condicionalmente independientes.

Al expandir la ecuación 4.3 se obtienen 2^{n_c} términos, que vuelven complicados los cálculos analíticos, debido a esto, se propone introducir la siguiente variable aleatoria:

$$Y_{r+1} = \begin{cases} 1, & \text{si } X \in D^c \\ 0, & \text{si } X \in D \end{cases}$$

y definiendo,

$$w_l = \begin{cases} (1-w)\psi_l, & \text{si } l = 1, ..., r \\ w, & \text{si } l = r+1 \end{cases}$$

donde $\sum_{l=1}^{r+1} w_l = 1$, lo que permite que la función de verosimilitud se reescriba así,

$$L_{x_{cj}}(w_1, ..., w_{r+1}, \lambda) = \prod_{i=1}^{n_c} \left[w_1^{y_1^{(i)}}, ..., w_r^{y_r^{(i)}} + w_{r+1} f_c(x^i|\lambda)^{y_{r+1}^{(i)}} \right] \tag{4.4}$$

Finalmente, al definir $Z = X|X \in D^c$ reescribimos la función de verosimilitud separando la parte continua z_{cj} y la parte discreta y_{cj}.

$$L_{x_{cj}}(w_1, ..., w_{r+1}, \lambda) = \prod_{l=1}^{r+1} w_l^{n_{cl}} \prod_{m=1}^{n_{c,r+1}} f_c(z_{(m)}|\lambda) = L_{y_{cj}}(w_1, ..., w_{r+1}) L_{z_{cj}}(\lambda) \tag{4.5}$$

donde $n_{cl} = \sum_{i=1}^{n_c} y_l^{(i)}$, de forma que $\sum_{l=1}^{r+1} n_{cl} = n_c$ y $z^{(m)}$ es el m-ésimo valor observado en D^c.

Con base en la función de verosimilitud obtenida para un grupo se puede obtener la función marginal de verosimilitud para una partición S dada. Se define θ_{cj} de manera conjunta como $(w_{cj,1}, ..., w_{cj,r+1}, \lambda_{cj})$ y análogamente, θ_S se define como el conjunto de parámetros de la partición S. Suponiendo independencia de los grupos dados los parámetros, la función de verosimilitud conjunta para θ_S,

$$L_x(\theta_S) = \prod_{c=1}^{k} \prod_{j=1}^{d} L_{y_{cj}}(w_{cj,1}, ..., w_{cj,r+1}) L_{z_{cj}}(\lambda_{cj}) \tag{4.6}$$

Si se supone independencia entre λ_{cj} y $(w_{cj,1}, ..., w_{cj,r+1})$, se puede factorizar la previa $\pi(\theta_{cj})$,

$$\pi(\theta_{cj}) = \pi(w_{cj,1}, ..., w_{cj,r}) \pi(\lambda_{cj}) \tag{4.7}$$

lo que lleva a una distribución de θ sobre la partición S,

$$\pi(\theta_S) = \prod_{c=1}^{k} \prod_{j=1}^{d} \pi(w_{cj,1}, ..., w_{cj,r+1}) \pi(\lambda_{cj}) \tag{4.8}$$

Una vez definido esto, la función de verosimilitud marginal se puede ver así,

$$
\begin{aligned}
p(x|S) &= \int_{\Theta_S} L_X(\theta_S)\pi(\theta_S)d\theta_S \qquad (4.9)\\
&= \int_{P_S} \left[\prod_{c=1}^{k} \prod_{j=1}^{d} L_{y_{cj}}(w_{cj,1}, ..., w_{cj,r+1})\pi(w_{cj,1}, ..., w_{cj,r+1}) \right] d_{P_S} \\
&\quad \cdot \int_{H_S} \prod_{c=1}^{k} \prod_{j=1}^{d} L_{z_{cj}}(\lambda_{cj})\pi(\lambda_{cj})d_{\lambda_S} \\
&= p(y|S)p(z|S)
\end{aligned}
$$

Esta definición de la verosimilitud permite separar los componetes discretos de los continuos, y así obtener las funciones de verosimilitud para cada componente (Blomstedt et al., 2015), como se muestra en la próxima sección.

4.2. Revisión de la propuesta

En esta sección se describe el modelo propuesto para clasificar datos mixtos aplicado en una base de datos que posee, tanto información categórica como continua. Este conjunto de datos se compone de p variables de conteo y d variables continuas, con n observaciones.

El modelo $p(x) = p(y, z) = \sum_{j=1}^{J} p_j P(y|\theta_j) P(z|\gamma_j)$, donde y representa las variables discretas y z las variables continuas. Las p variables continuas se distribuyen de la forma Normal-Multivariada, $y = X^c \sim N(\mu_j, \Sigma_j)$, mientras que cada una de las d variables de conteo tiene una distribución Poisson, $z = X_{l=1}^{d} \sim Po(\lambda_{lj})$.

Por lo que nuestro modelo de mezclas propuesto con K componentes se define de la siguiente manera:

$$p(z, y) = \sum_{j=1}^{K} \pi_j \cdot N(\mu_j, \Sigma_j) \cdot \prod_{l=1}^{p} Po(\lambda_{lj}) \tag{4.10}$$

donde $\pi_j = p((y, z) \in C_j)$.

Ahora, como se menciona el capítulo anterior, para facilitar el cálculo de las distribuciones, se incluye la variable latente z_j al modelo, la cual actúa como variable de asignación. Debido a que x_i sólo puede pertenecer a un componente, se puede interpretar a la probabilidad de asignación como la esperanza de z_{ij} condicional a los datos observados,

$$E[z_{ij}|x] = \frac{\pi_j^{(t)}|x \cdot N(\mu_j, \Sigma_j) \cdot \prod_{l=1}^{p} Po(\lambda_{lj})}{\sum_{j=1}^{K} \pi_j^{(t)}|x \cdot N(\mu_j, \Sigma_j) \cdot \prod_{l=1}^{p} Po(\lambda_{lj})} \tag{4.11}$$

Para obtener la función de distribución conjunta, tenemos que, $P(x, z; \theta) = P(z, \theta) \cdot P(x|z, \theta)$ con el parámetro $\theta = (\pi, \mu, \Sigma, \lambda)$, entonces definimos lo siguiente,

$$P(z,\theta) = \prod_{j=1}^{K} \pi_j^{z_j} \tag{4.12}$$

$$P(x|z_j,\theta) = (N(\mu_j,\Sigma_j) \cdot \prod_{l=1}^{p} Po(\lambda_{lj}))^{z_j} \tag{4.13}$$

Por definición obtenemos la distribución conjunta del modelo incluyendo a la variable latente,

$$P(x,z;\theta) = \prod_{j=1}^{K}\prod_{i=1}^{n} \left(\pi_j \cdot N(\mu_j,\Sigma_j) \cdot \prod_{l=1}^{p} Po(\lambda_{lj}) \right)^{z_{ij}} \tag{4.14}$$

4.3. Parámetros y distribución inicial

Para las distribuciones iniciales, se elegirán distribuciones conjugadas que aseguren una distribución final con una forma paramétrica conocida.

Empezando por las variables continuas, las previas que se eligieron tomaran una distribución Normal Multivariada para el parámetro de medias y una Wishart Inversa para el parámetro de la varianza. Esta decisión se tomo con base en la distribución de la función de verosimilitud, que al tomar la forma de una Normal Multivariada, se produce una distribución posterior Normal Multivariada para el parámetro de medias y una distribución posterior Wishart Inversa para el parámetro de varianza y facilita los cálculos posteriores.

En el caso de las variables discretas, se eligió la distribución previa como una Gamma ya que se facilitan los cálculos, debido a que la función de verosimilitud al estar compuesta de distribuciones Poisson, produce una distribución posterior Gamma. Para el parámetro de proporción de la variable latente que tiene como previa una distribución Dirichlet, le corresponde una Dirichlet como distribución posterior. De manera que, las distribuciones previas de los parámetros y sus hiperparámetros correspondientes, se definen así,

- $\Sigma_j \sim W^{-1}(\Lambda_j, v_j)$ con una función de densidad,

$$P(\Sigma_j|\Lambda_j, v_j) \propto |\Sigma_j^{-1}|^{-(v_j+d+1)/2} \exp\left\{-\frac{1}{2} trace\left[\Sigma_j^{-1}\Lambda_j\right]\right\}$$

  donde Λ_j es la matriz de covarianzas de los datos observados y v_j es el número de variables continuas mas uno, $(d+1)$

- $\mu_j|\Sigma_j \sim N(\varepsilon_j, \frac{\Sigma_j}{n_j})$ con función de densidad,

$$P(\mu_j|\varepsilon_j, \frac{\Sigma_j}{n_j}) \propto |\Sigma_j^{-1}|^{-\frac{1}{2}} \exp\left\{-\frac{n_j}{2} trace\left[(\mu_j - \varepsilon_j)'\Sigma_j^{-1}(\mu_j - \varepsilon_j)\right]\right\}$$

  donde ε_j es un vector de las medias observas de las variables continuas, y n_j para fines prácticos se tomo como el número total de observaciones n entre el número de componentes K

- $\lambda_{lj} \sim G(a_{lj}, S_{lj})$ con función de densidad,

$$P(\lambda_{lj}|a_{lj}, S_{lj}) \propto \lambda_{lj}^{a_{lj}} \exp\{-S_{lj}\lambda_{lj}\}$$

donde S_{lj} es la varianza observada de la l-ésima variable discreta y a_{lj} el parámetro de forma.

- $\pi \sim Dir(\alpha_1 = 1/k, ..., \alpha_k = 1/k)$ con función de densidad,

$$P(\pi|\alpha_1, ..., \alpha_k) \propto \prod_{j=1}^{K} \pi_j^{\alpha_j}$$

4.4. Especificación del kernel

Tomando en cuenta lo anterior, se tiene la siguiente función de distribución posterior, que está compuesta por la función de verosimilitud y las distribuciones previas de los parámetros.

$$P(\theta|x, z) = P(x|z, \theta) \cdot P(\pi|\alpha_1, ..., \alpha_k) \cdot \prod_{j=1}^{K} \left[P(\Sigma_j|\Lambda_j, v_j) \cdot P(\mu_j|\varepsilon_j, \frac{\Sigma_j}{n_j}) \cdot \prod_{l=1}^{p} P(\lambda_{lj}|a_{lj}, S_{lj}) \right] \tag{4.15}$$

Con el fin de ser más claros en el desarrollo de la distribución posterior, se separa la parte de las variables continuas y la parte discreta.

4.4.1. Parte discreta

Se tienen p variables de conteo son iid con una distribución Poisson, de las cuales se obtiene la parte discreta de la función de verosimilitud:

$$\begin{aligned} P(x|z,\theta) &= \prod_{i=1}^{n}\left(\prod_{l=1}^{p} Po(\lambda_{lj})\right)^{z_{ij}} \\ &\propto \prod_{i=1}^{n}\left(\prod_{l=1}^{p} \lambda_{lj}^{x_{li}} \exp\{-\lambda_{lj}\}\right)^{z_{ij}} \\ &= \prod_{l=1}^{p}\prod_{i=1}^{n} \lambda_{lj}^{x_{li}\cdot z_{ij}} \exp\{-\lambda_{lj}\cdot z_{ij}\} \\ &= \prod_{l=1}^{p} \lambda_{lj}^{\sum_{i=1}^{n} x_{li}\cdot z_{ij}} \exp\left\{-\lambda_{lj}\sum_{i=1}^{n} z_{ij}\right\} \\ &= \prod_{l=1}^{p} \lambda_{lj}^{\bar{z}_j \bar{x}_{lj}} exp\{-\bar{z}_j \lambda_{lj}\} \end{aligned} \tag{4.16}$$

donde $\bar{z}_j = \sum_{i=1}^{n} z_{ij}$ y $\bar{x}_{lj} = \sum_{i=1}^{n} \frac{z_{ij}x_{li}}{\bar{z}_j}$

Ahora sustituimos dentro de la función posterior la función de verosimilitud (4.16) y las distribuciones previas conjugadas de los parámetros $P(\lambda_j|a_j, S_j)$,

que serán definidas en la sección 4.3,

$$
\begin{aligned}
P(\theta|x,z) &= P(x|z,\theta)\cdot\prod_{l=1}^{p}P(\lambda_{lj}|a_{lj},S_{lj}) \qquad (4.17)\\
&\propto \prod_{l=1}^{p}\lambda_{lj}^{\bar{z}_j\bar{x}_{lj}}exp\left\{-\bar{z}_j\lambda_{lj}\right\}\cdot\lambda_{lj}^{a_{lj}}\exp\left\{-S_{lj}\lambda_{lj}\right\}\\
&= \prod_{l=1}^{p}\lambda_{lj}^{\bar{z}_j\bar{x}_{lj}+a_{lj}}\cdot exp\left\{-\lambda_{lj}(\bar{z}_j+S_{lj})\right\}\\
&= \prod_{l=1}^{p}Ga(\tilde{a}_{lj},\tilde{S}_{lj})
\end{aligned}
$$

4.4.2. Parte continua

Para la parte continua se tienen d variables que iid con una distribución normal-mutivariada, para desarollar la parte continua de la función de vero-

similitud,

$$
\begin{aligned}
P(x|z,\theta) &= \prod_{i=1}^{n} \left(N(x_i|\mu_j,\Sigma_j)\right)^{z_{ij}} \qquad (4.18)\\
&\propto \left(|\Sigma_j^{-1}|^{\frac{1}{2}} \exp\left\{-\frac{1}{2}(x_i-\mu_j)'\Sigma_j^{-1}(x_i-\mu_j)\right\}\right)^{z_{ij}}\\
&= |\Sigma_j^{-1}|^{\frac{\sum_{i=1}^{n} z_{ij}}{2}} \exp\left\{-\frac{1}{2}\sum_{i=1}^{n} z_{ij} trace\left[\Sigma_j^{-1}(x_i-\mu_j)(x_i-\mu_j)'\right]\right\}\\
&= |\Sigma_j^{-1}|^{\frac{\sum_{i=1}^{n} z_{ij}}{2}} \exp\left\{-\frac{1}{2} trace\left[\Sigma_j^{-1}\sum_{i=1}^{n} z_{ij}(x_i x_i' - \mu_j x_i' - x_i\mu_j' + \mu_j\mu_j')\right]\right\}\\
&= |\Sigma_j^{-1}|^{\frac{\sum_{i=1}^{n} z_{ij}}{2}} \exp\left\{-\frac{1}{2} trace\left[\Sigma_j^{-1}\sum_{i=1}^{n} z_{ij}(x_i x_i' - \bar{x}_j\bar{x}_j' + \bar{x}_j\bar{x}_j' - \mu_j x_i' - x_i\mu_j' + \mu_j\mu_j')\right]\right\}\\
&\propto \exp\left\{-\frac{1}{2} trace\left[\Sigma_j^{-1}\sum_{i=1}^{n} z_{ij}(x_i x_i' - \bar{x}_j\bar{x}_j' + \bar{x}_j x_i' - x_i\bar{x}_j') + \bar{z}_j(\bar{x}_j-\mu_j)(\bar{x}_j-\mu_j)'\right]\right\}\\
&\propto |\Sigma_j^{-1}|^{-\frac{\bar{z}_j}{2}} \exp\left\{-\frac{1}{2} trace\left[\Sigma_j^{-1}\left(\bar{z}_j(\bar{x}_j-\mu_j)(\bar{x}_j-\mu_j)' + \sum_{i=1}^{n} z_{ij}(x_i-\bar{x}_j)(x_i-\bar{x}_j)'\right)\right]\right\}
\end{aligned}
$$

Una vez obtenida la función de verosimilitud (4.16), se procede a desarrollar la función posterior,

$$
\begin{aligned}
P(\theta|x,z) &= P(x|z,\theta)\cdot P(\Sigma_j|\Lambda_j,v_j)\cdot P(\mu_j|\varepsilon_j,\frac{\Sigma_j}{n_j}) \qquad (4.19)\\
&\propto|\Sigma_j^{-1}|^{-\frac{\bar{z}_j}{2}}\exp\left\{-\frac{1}{2}trace\left[\Sigma_j^{-1}\left(\bar{z}_j(\bar{x}_j-\mu_j)(\bar{x}_j-\mu_j)'+\sum_{i=1}^{n}z_{ij}(x_i-\bar{x}_j)(x_i-\bar{x}_j)'\right)\right]\right\}\\
&\quad\cdot|\Sigma_j^{-1}|^{-\frac{1}{2}}\exp\left\{-\frac{n_j}{2}trace\left[(\mu_j-\varepsilon_j)'\Sigma_j^{-1}(\mu_j-\varepsilon_j)\right]\right\}\\
&\quad\cdot|\Sigma_j^{-1}|^{-(v_j+d+1)/2}\exp\left\{-\frac{1}{2}trace\left[\Sigma_j^{-1}\Lambda_j\right]\right\}\\
&=|\Sigma_j^{-1}|^{-(\tilde{v}_j+d+1)/2}\exp\left\{-\frac{1}{2}trace\left[\Sigma_j^{-1}\tilde{\Lambda}_j\right]\right\}\\
&\quad\cdot|\Sigma_j^{-1}|^{-\frac{1}{2}}\exp\left\{-\frac{\tilde{n}_j}{2}trace\left[(\mu_j-\tilde{\varepsilon}_j)'\Sigma_j^{-1}(\mu_j-\tilde{\varepsilon}_j)\right]\right\}\\
&=N(\mu_j|\tilde{\varepsilon}_j,\frac{\Sigma_j}{\tilde{n}_j})\cdot W^{-1}(\Sigma_j|\tilde{v}_j,\tilde{\Lambda}_j)
\end{aligned}
$$

El desarrollo completo se encuentra en el anexo.

4.5. Distribución final completa

Retomando de la sección 4.2, al juntar la parte discreta con la parte continua, incluyendo a la variable latente, se obtiene la siguiente función de distribución

final, compuesta por :

$$P(\theta|x,z) = P(x|z,\theta)\cdot P(\pi|\alpha_1,...,\alpha_k) \cdot \prod_{j=1}^{K}\left[P(\Sigma_j|\Lambda_j,v_j)\cdot P(\mu_j|\varepsilon_j,\frac{\Sigma_j}{n_j})\cdot\prod_{l=1}^{p}P(\lambda_{lj}|a_{lj},S_{lj})\right] = \prod_{j=1}^{k} Dir(\pi_1,\pi_k|\tilde{\alpha}_1,...,\tilde{\alpha}_k)\cdot G(\tilde{a}_j,\tilde{S}_j)\cdot N(\tilde{\varepsilon}_j,\frac{\Sigma_j}{\tilde{n}_j})\cdot W^{-1}(\tilde{\Lambda}_j,\tilde{v}_j) \tag{4.20}$$

4.6. Gibbs Sampler

Se aplica el método de *Gibbs Sampler* al Modelo basado en Mezclas, mediante el siguiente algoritmo, que toma como referencia el Algortimo 6 propuesto por (Liang, 2009) alicando los siguientes pasos.

1. Obtener los valores iniciales $\{\theta_j^{(0)} = (\pi_j^{(0)},\mu_j^{(0)},\Sigma_j^{(0)},\lambda_j^{(0)})\}_{j=1}^{K}$ de los parámetros con base en las distribuciones previas definidas en la sección anterior.

   - $\Sigma_j^{(0)} \sim W^{-1}(\Lambda_j, v_j)$.
   - $\mu_j^{(0)}|\Sigma_j^{(0)} \sim N(\varepsilon_j, \frac{\Sigma_j}{n_j})$.
   - $\lambda_{lj}^{(0)} \sim G(a_{lj}, S_{lj})$
   - $\pi^{(0)} \sim Dir(\alpha_1 = 1/k, ..., \alpha_k = 1/k)$

2. Repetir para $t = 1, 2, ..., T$, siendo T el número de iteraciones.

a) Generar las simulaciones de la variable latente $z_{ij}^{(t)} \in \{0, 1\}$ para las n observaciones, con $i = 1, ..., n$ y

$$z_{ij}^{(t)} \sim M_k(1; p_1, ..., p_k)$$

donde $p_j = \left(\frac{\pi_j^{(t-1)} \cdot N(\mu_j, \Sigma_j) \cdot \prod_{l=1}^{p} Po(\lambda_{lj})}{\sum_{j=1}^{K} \pi_j^{(t-1)} \cdot N(\mu_j, \Sigma_j) \cdot \prod_{l=1}^{p} Po(\lambda_{lj})} \right)$

b) Generar las simulacines de las distribuciones posteriores de los parámetros para cada componente j, con $j = 1, ..., K$.

- $\Sigma_j^{(t+1)} \sim W^{-1}(\tilde{\Lambda}_j, \tilde{v}_j)$, se definen $\tilde{\Lambda}_j$ y $\tilde{v}_j$ como,

  $$\tilde{\Lambda}_j = \Lambda_j + \sum_{i=1}^{n} z_{ij}(x_i - \bar{x}_j)(x_i - \bar{x}_j)' + \frac{n_j \bar{z}_j}{n_j + \bar{z}_j}(\bar{x}_j - \varepsilon_j)(\bar{x}_j - \varepsilon_j)'$$

  $$\tilde{v}_j = v_j + \bar{z}_j$$

  donde $\bar{z}_j = \sum_{i=1}^{n} z_{ij}$ y $\bar{x}_j = \sum_{i=1}^{n} \frac{z_{ij} x_i}{\bar{z}_j}$

- $\mu_j^{(t+1)} \sim N(\tilde{\varepsilon}_j, \frac{\Sigma_j}{\tilde{n}_j})$ donde,

  $$\tilde{\varepsilon}_j = \frac{\bar{z}_j \bar{x}_j + n_j \varepsilon_j}{\bar{z}_j + n_j}$$

  $$\tilde{n}_j = \bar{z}_j + n_j$$

- $\lambda_{lj}^{(t+1)} \sim G(\tilde{a}_j, \tilde{S}_j)$, se definen $\tilde{a}_{lj}$ y $\tilde{S}_{lj}$ como,

  $$\tilde{a}_{lj} = \bar{z}_j \bar{x}_{lj} + a_{lj}$$

  $$\tilde{S}_{lj} = \bar{z}_j + S_{lj}$$

donde $\bar{x}_{lj} = \sum_{i=1}^{n} \frac{z_{ij} x_{li}}{\bar{z}_j}$

- $\pi_j^{(t+1)} \sim Dir(\tilde{\alpha}_1, ..., \tilde{\alpha}_k)$ donde $\tilde{\alpha}_j = \bar{z}_j + \alpha_j$

3. Repetir el paso 2 hasta que la distribución conjunta de $(z^{(t)}, \theta^{(t)})$ no cambie.

Descripción del código

Para la aplicación práctica del modelo propuesto, se deserrollo un código en R, que está construido de forma funcional, esto con el propósito de facilitar la modificación o la revisión de cada una de las funciones que integran al algoritmo. El código consta de una función principal, en la cual se introduce la información, el número T de iteraciones que hará el algortmo de Gibbs Sampler y el número J de grupos bajo el cual se va a realizar la segmentación. Dicha función llama a las demás funciones que estiman las distribuciones previas de los parámetros, para variables continuas y de conteo por separado; de manera que con base en estas estimaciones se comiencen las iteraciones, para estimar las distribuciones posteriores y generar las probabilidades de asignación que segmentan las observaciones en los J grupos.

Este código se encuentra disponible para su consulta en el sitio web de *Git Hub* en la dirección https://github.com/montserratvizcayno/Tesis1

Capítulo 5

Aplicación práctica

5.1. Aspectos generales de la información

Los datos que se utilizaron como fuente de información, provienen de una empresa de empeño y microcréditos que comenzó en febrero del 2006, con diez sucursales en el Estado de México y Querétaro.

Primero comenzaremos explicando la forma en la se realizan los pagos del préstamo prendario. A diferencia de otras empresas del ramo, ésta ofrece tres esquemas de pago entre los cuales el cliente puede elegir, según le resulte más conveniente, para recuperar los objetos que ha empeñado:

1. **Tradicional** es la clásica forma de pago en la cual se pagan interés y se cuenta con 5 refrendos, al llegar al último refrendo se tiene que pagar el monto total del préstamo.

2. **Pagos Fijos** consiste en dividir el monto de la deuda más los intereses entre el número de semanas o meses que tiene el cliente como plazo para liquidar la deuda.

3. **Flexible** es una mezcla de los dos esquemas anteriores, consiste en ir pagando intereses o capital según le convenga hasta cubrir el monto total de la deuda en un plazo acordado.

El inventario de objetos que son admitidos para realizar un empeño, es limitado y es necesario que se encuentren dentro de la siguente clasificación, de otra forma, el objeto no será admitido:

Metales

- oro
- plata

Electrónicos

- Televisores
- Minicomponentes

- Celulares
- Dvd´s
- Consolas de video juegos
- Computadoras
- Camaras digitales
- Reproductores de mp3

Otros

- Relojes

5.1.1. Descripción de los datos

La base datos que nos compartió la Compañía se conforma de la información histórica total de 29,822 clientes. Cada una de estas observaciones (clientes) cuenta con información de 11 variables categóricas y 8 de escala de razón, que se describen a continuación:

1. **Cliente.desde** ($\mathtt{v}_1$) es la fecha de registro en la que el cliente se dio de alta en el sistema.

2. **Edad** ($\mathtt{v}_2$) es el número de años cumplidos a la fecha en la que se realizó la extracción de la informacióón.

3. **Sexo** (v_3) es una variable categórica binaria que se codificó con uno en caso de ser mujer y cero en caso de ser hombre.

4. **Ciudad** (v_4) es la población donde el cliente reside.

5. **Código postal** (v_5) es el código postal que pertenece al domicilio registrado como la residencia del cliente.

6. **Colonia** (v_6) es la dirección que el cliente ha indicado como su lugar de residencia; normalmente se toma de la credencial de elector.

7. **Suc** (v_7) es el número de la sucursal en la cual el cliente fue dado de alta en el sistema.

8. **Créditos** (v_8) es el número total de créditos que se le han otorgado al cliente ha la fecha corte o de extracción.

9. **Vigente** (v_9) número de créditos que se encuentran activos, es decir, que el monto no ha sido saldado y el cliente se encuentra al corriente con los pagos.

10. **Monto.prom** (v_{10}) es la cantidad promedio de dinero otorgada al cliente de los créditos que se le ha otorgado.

11. **Cred.perd** (v_{11}) es el número de créditos que el cliente ha perdido, es decir, que no ha podido pagar para recuperar su prenda, lo que trae como consecuencia que su prenda sea adjudicada, para que después, sea vendida.

12. **Int.pag** (v_{12}) es el número de créditos en los que se realizaron pagos fuera del esquema de pagos pactado. Esto puede ocurrir cuando el cliente se atrasa y realiza un pago a destiempo con tal de no perder su prenda.

13. **Metal** (v_{13}) es el número de prendas metálicas que ha empreñado en el total de las transacciónes realizadas.

14. **Electrónico** (v_{14}) es el número de prendas de tipo electrónico que ha empreñado el cliente en las transacciones realizadas a la fecha de corte.

15. **Cred.trad** (v_{17}) es el número de operaciones en las que el cliente ha escogido esquema tradicional de pagos para liquidar los prestamos / empeños.

16. **Cred.PF** (v_{18}) es el número operaciones en las que el cliente ha escogido el esquema de pagos fijos.

17. **Cred.Flex** (v_{19}) es el número de operaciones en las que el cliente ha seleccionado el esquema de pagos flexible.

18. **Ing.tot** (v_{20}) es la suma de los pagos por concepto de ingresos que ha recibido la empresa por cliente.

19. **Ing.prom** (v_{21}) es el ingreso promedio que la empresa ha recibido por el total los pagos que el cliente ha realizado desde su alta hasta la fecha de corte.

5.1.2. Análisis Exploratorio

Antes de comenzar a analizar los datos contenidos en la base, se realizó una depuración de la base de datos. Los principales puntos encontrados son los siguientes:

- Errores en captura de fechas de nacimiento. En la base original, se encontraron 119 clientes con errores en las fechas de nacimiento, debido a errores en la captura de éstas desde que fueron dados de alta en el sistema.
- Información faltante. Se encontraron 761 individuos que presentan datos faltantes en variables como código postal, muncipio, sucursal, monto

promedio e ingreso promedio.

Al total de 880 registros mencionados anteriormente, se decidió eliminarlos de la base para no generar un sesgo al momento del análisis de los datos.

Como resultado, la base con la cual se llevó a cabo el análisis exploratorio, que se presenta a continuación, está compuesta de 32551 observaciones y 18 variables.

Primero se realizó un análisis de correlaciones en el que se observó que las variables más significativas son: Saldo contra Ingreso Total con 0.9, Créditos contra Creditos a Pagos Fijos con 0.8 y Creditos contra Electrónicos con .7, como se pueden observar en la gráfica 5.1.

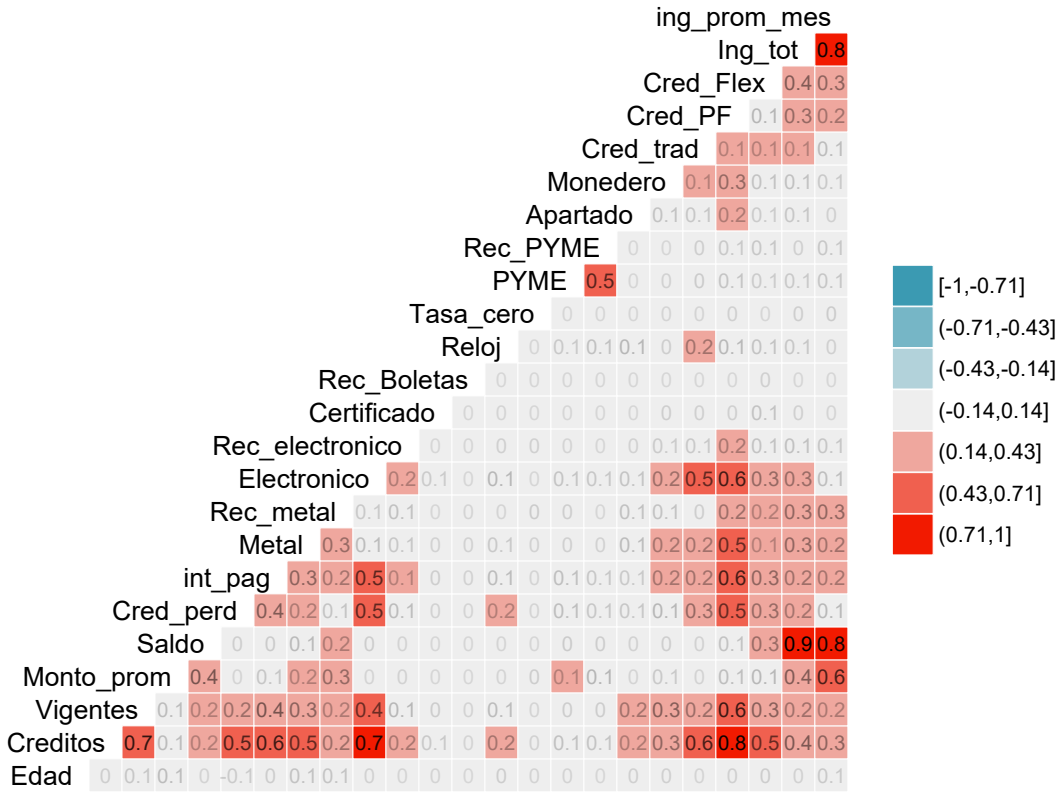


Figura 5.1: Gráfica de correlaciones

Con base en la gráfica anterior, se ha decidido enforcarse en sólo seis variables, 3 variables continuas que son **Saldo**, **Monto.prom** y **Ing.tot**, y en tres variables de conteo **Electrónico**, **Cred.perd** y **Cred.PF**. De esta manera podremos utilizar estas variables para los componentes continuos y discretos del modelo que se definió en el capítulo anterior.

A continuación, con el fin de realizar un análisis de las variables seleccionadas, se muestran las tablas con los estadísticos descriptivos.

Tabla 5.1: Tabla de Media y Varianza

variable	media	varianza
Cred_perd	0.67	1.36
Cred_PF	1.56	9.04
Electronico	16.23	1565.86
Ing_tot	5175.00	4990698327.63
Monto_prom	1296.89	7208400.90
Saldo	640.16	236844874.62

Tabla 5.2: Estadísticos descriptivos

	Cred_perd	Cred_PF	Electronico	Ing_tot	Monto_prom	Saldo
Min. :	0	0	0	-251	0	0
1st Qu.:	0	0	1	100	500	0
Median :	0	1	7	407	800	0
Mean :	0.6707	1.56	16.23	5175	1297	640.2
3rd Qu.:	1	2	18	1631	1350	0
Max. :	59	168	3439	11024580	143881	2674250

En la figura 5.2 de abajo, se puede apreciar como están altamente correlacionadas las tres variables: Saldo, Monto.prom e Ing.tot. Esto implica que al momento de modelar las variables bajo el modelo seleccionado se podrán distribuir como:

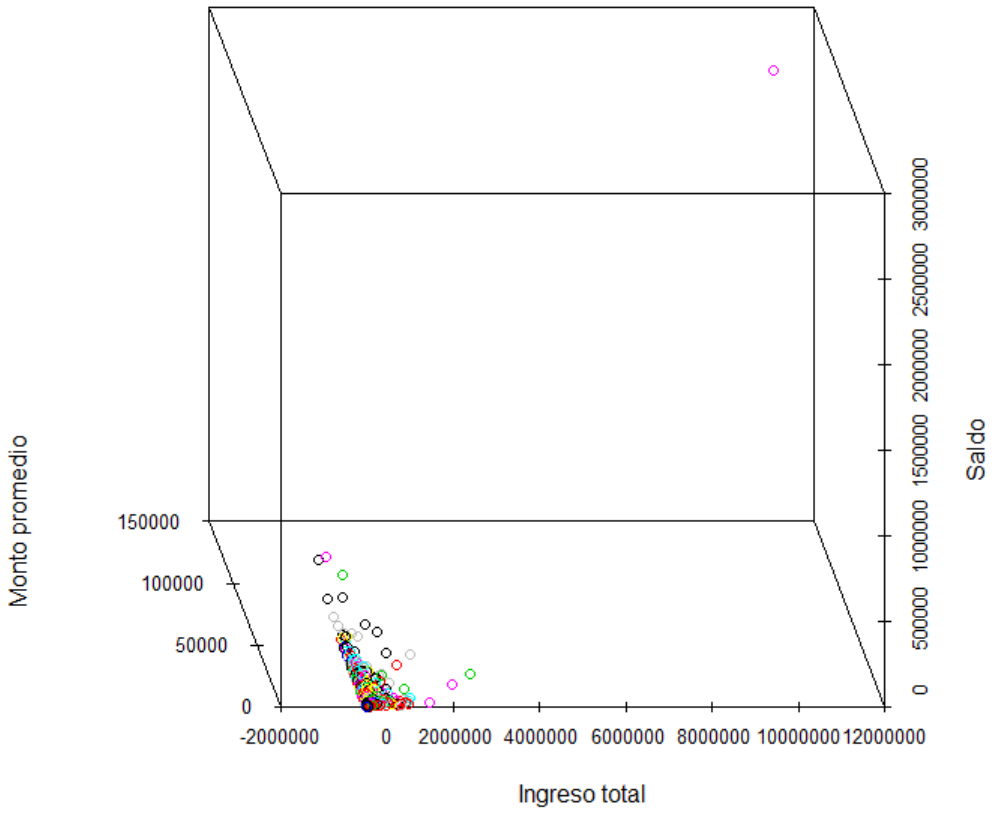


Figura 5.2: Dispersión de las variables Ing tot, Monto prom y Saldo

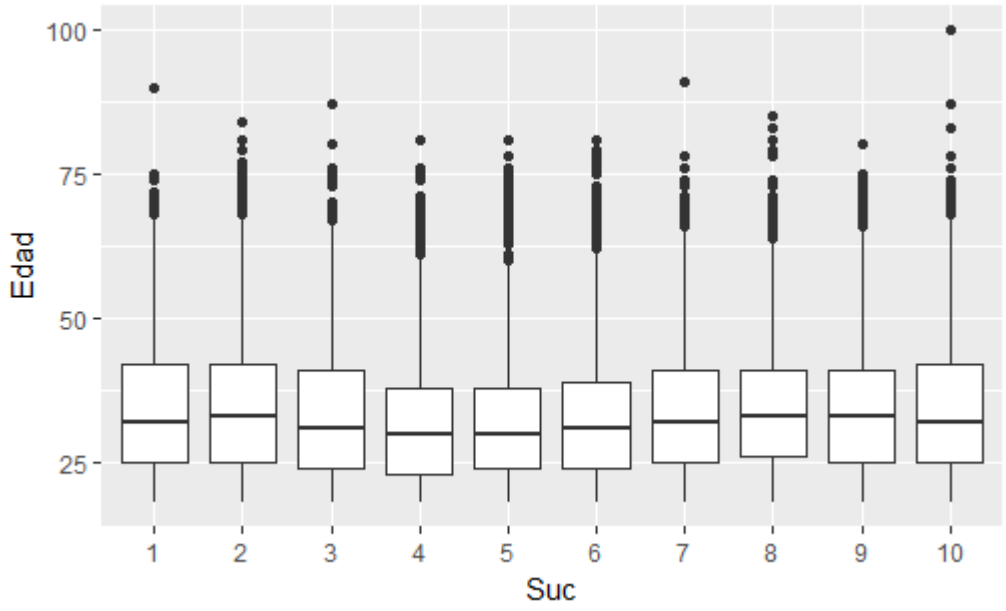


Figura 5.3: Distribución de las edades de los clientes por sucursal

Ahora, en la gráfica de caja y brazos (5.3) muestra como se distribuyen a las edades según la sucursal.

En cuanto a las variables elegidas para el análisis y aplicación del modelo, a continuación se presentan los

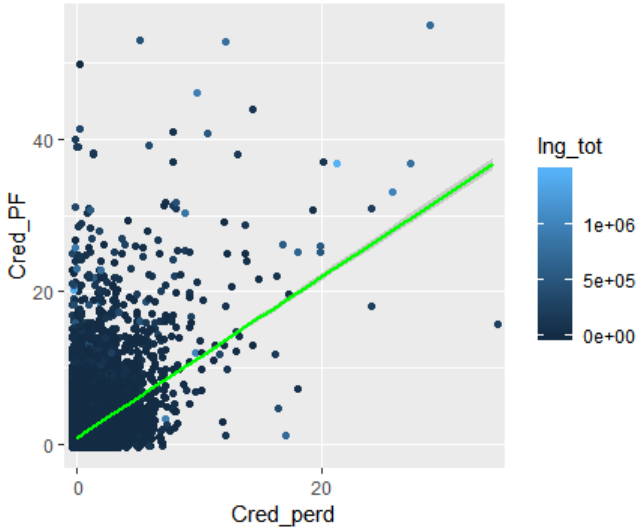


Figura 5.4: Gráfica de dispersión crédito en pagos fijos vs créditos perdidos

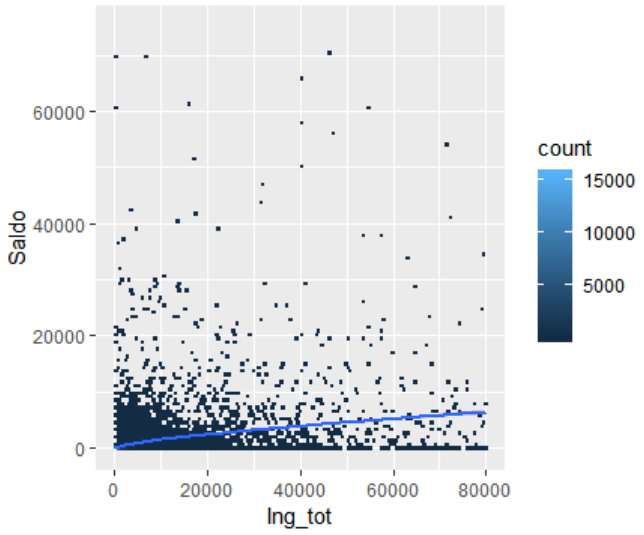


Figura 5.5: Gráfica de dispersión de Ingreso Total vs Saldo

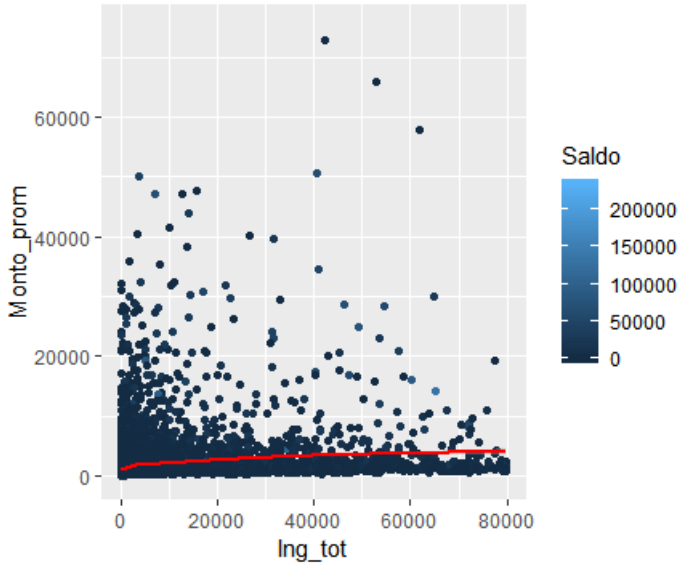


Figura 5.6: Gráfica de dispersión de Monto Promedio vs Ingreso Total

5.2. Resultados de la aplicación

Una vez realizado el análisis exploratorio de los datos, y contar con una base de datos consistente compuesta de cinco variables, Saldo, Monto.prom, Ing.tot, Cred.PF y Cred.perd, donde las primeras tres son continuas, mientras que las últimas dos son discretas.

Para un primer acercamiento y con el fin de realizar unas pruebas sobre el desempeño del modelo, se corrió el algoritmo dearrollado en R (incluido

en el Anexo 1) sobre una muestra de 100 observaciones y 30 iteraciones. A continuación se muestran las distribuciones por componente de la variables variables utilizadas.

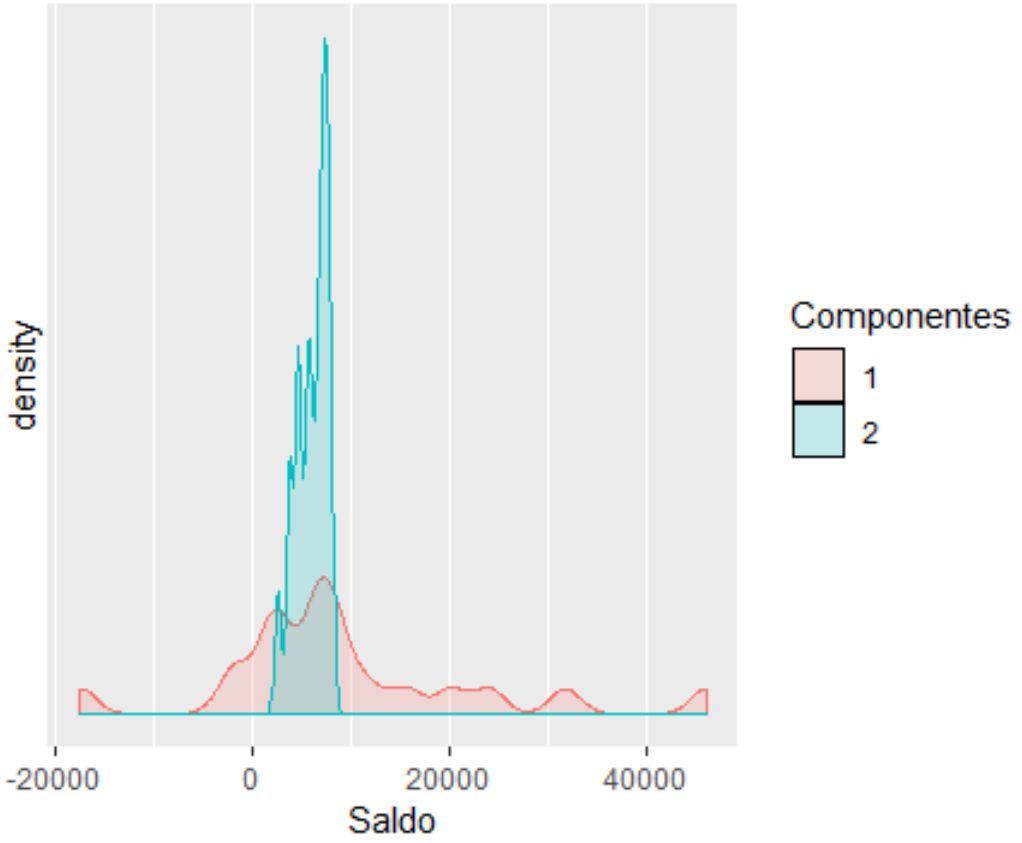


Figura 5.7: Primer resultado bajo 100 observaciones con dos componentes de Saldo

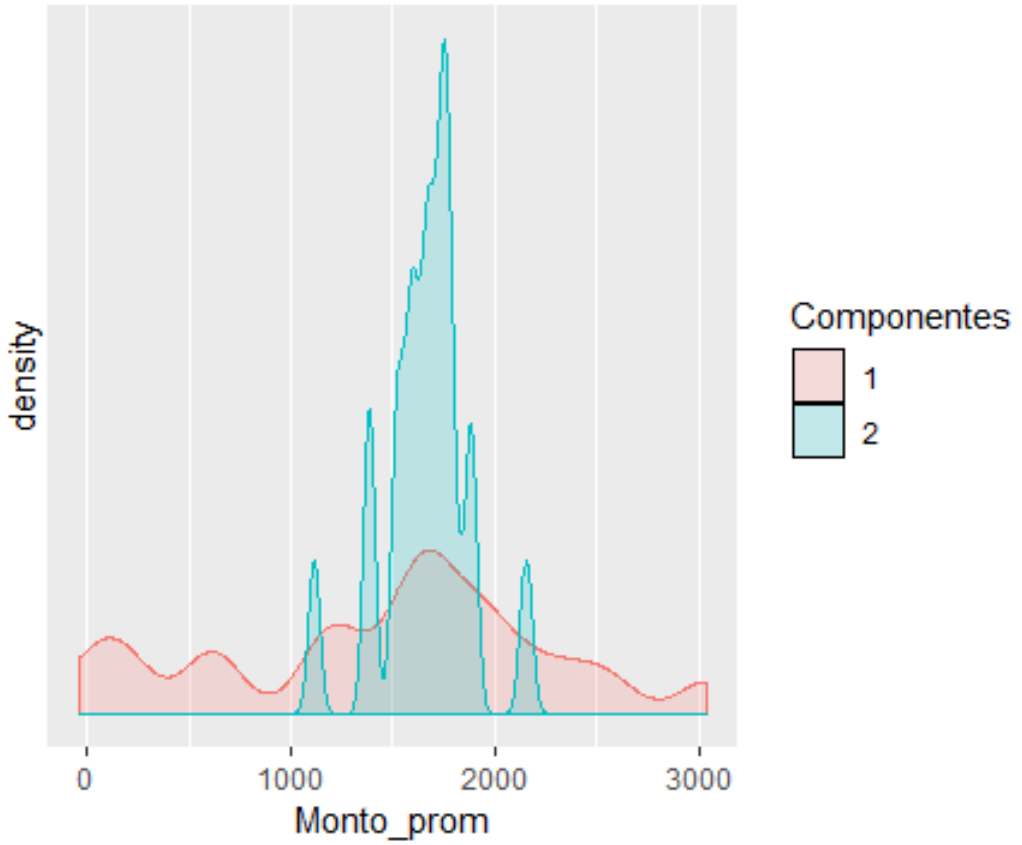


Figura 5.8: Primer resultado bajo 100 observaciones con dos componentes de Monto promedio

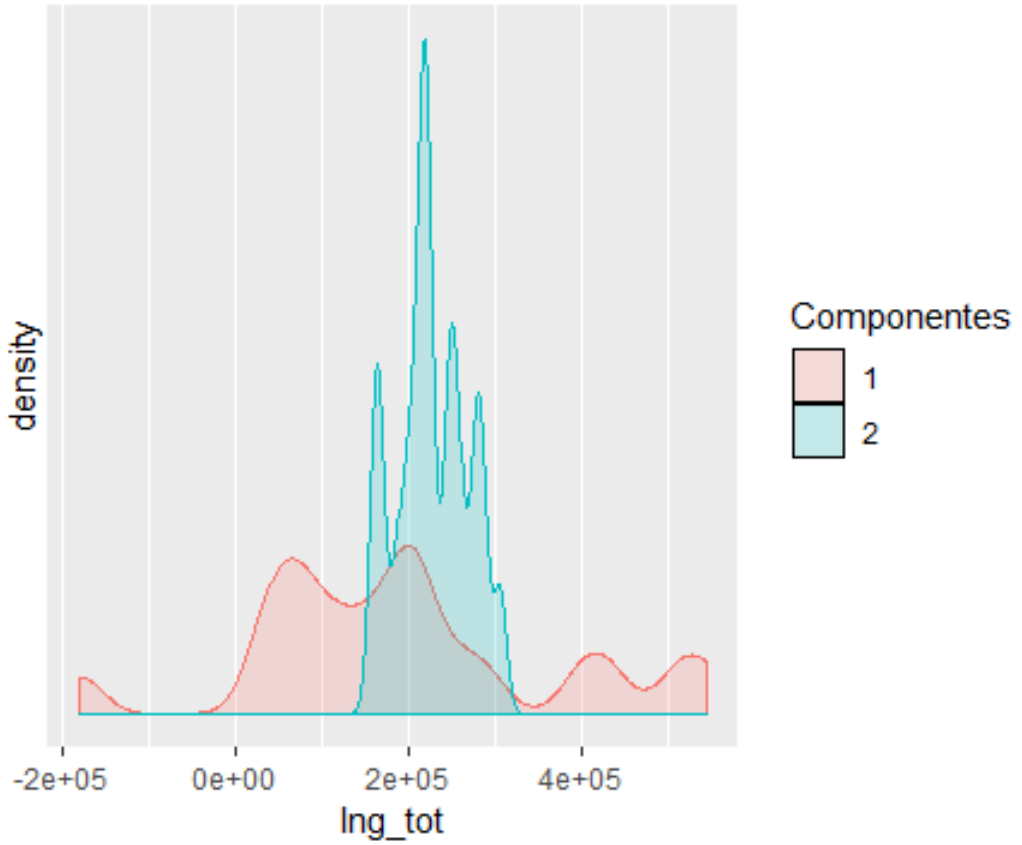


Figura 5.9: Primer resultado bajo 100 observaciones con dos componentes de Ingreso total

En este primer acercamiento se puede ver como no las distribuciones de ambos componentes se traslapan, porlo que no es posible hacer una distinción clara sobre la asignación probabilistica de las observaciones, por lo que se decidió realizar más pruebas con muestras más grandes y mayor número de

iteraciones, para obsevar si con estos ajustes, el algoritmo presenta un mejor resultado para asignar a los grupos de clasificación.

Capítulo 6

Conclusiones

6.1. Observaciones del modelo

1. Parametrización de la distribución previa $Ga(\alpha, \beta)$ para las variables de conteo. se designó para el parámetro de forma α un valor arbitrareo $a_{lj} \in N$, mientras que para el parametro de escala β se tomó la varianza observada S_{lj} de la variable en cuestión, y una distribución posterior $Ga(\tilde{\alpha}, \tilde{\beta})$ tiene como parámetros $\tilde{\alpha} = \bar{z}_j \bar{x}_j + a_{lj}$ y $\tilde{\beta} = S_{lj} + \bar{z}_j$. Por lo que en cada iteración los parámetros van creciendo muy rápidamente y pronto, los valores observados se encuentran fuera de la densidad y

al valuarlas nos da cero. Aún cambiando el parámetro de escala a un valor arbitrareo pequeño para la previa, sigue creciendo muy rápido.

Bibliografía

Bernardo, J.-M. (1998). Bruno de Finetti en la estadística contemporánea. *Historia de la Matématica en el Siglo XX, S. Rios (ed.), Real Academia de Ciencias, Madrid*, 63 – 80.

Bernardo, J. M. and A. F. Smith (2001). Bayesian theory.

Blomstedt, P., J. Tang, J. Xiong, C. Granlund, and J. Corander (2015). A Bayesian Predictive Model for Clustering Data of Mixed Discrete and Continuous Type. *IEEE Transactions on Patteren Analysis and Machine Intelligence*@(3), 489–498. VK: Kaski, S.; COIN; HIIT.

Congdon, P. (2007). *Bayesian Statistical Modelling*, Volume 704. John Wiley & Sons.

De Finetti, . (1930). *De Finetti*.

Bibliografía

Dempster, A. P., N. M. Laird, and D. B. Rubin (1977). Maximum likelihood from incomplete data via the em algorithm. *Journal of the royal statistical society. Series B (methodological)*, 1–38.

Gelman, Andrew y Carlin, J. B. and D. B. Stern, Hal S y Rubin (2014). *Bayesian Data Analysis*, Volume 2. Chapman y Hall/CRC Boca Raton, FL, USA.

George, B. (1989). *Verosimilitud Extendida*.

Liang, L. (2009). On simulation methods for two component normal mixture models under Bayesian approach.

Marin Jean-Michel, Mengersen Kerrie, R. C. P. (2005). Bayesian modelling and inference on mixtures of distributions. In D. Dey and C. Rao (Eds.), *Bayesian Thinking*, Volume 25 of *Handbook of Statistics*, pp. 459 – 507. Elsevier.

Martínez Ovando, J. C. (2004). Un criterio predictivo de selección de modelos para series de tiempo. Master's thesis, IIMAS–UNAM.

Stephens, M. Dealing with label switching in mixture models. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*@(4), 795–809.