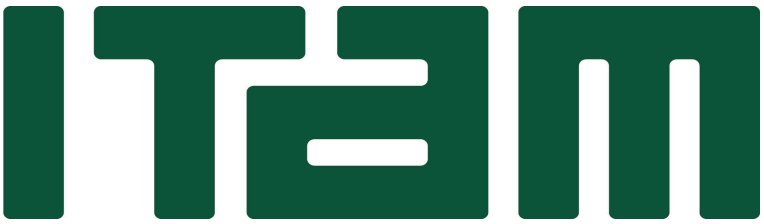
Instituto Tecnológico Autónomo de México



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. 2018

"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 tablas

Capítulo 1

Introducción

1.1. Estructura de la Tesis

La estructura de este trabajo de Tesis está compuesta de seis capítulos. El primer capítulo, que es en el que nos encontramos actualmente, da una breve introducción a los modelos de clasificación no supervisada y sus distintos enfoques. Más adelante en el capítulo dos, Inferencia Bayesiana, se exponen las bases y los principales conceptos necesarios para entender el modelo bayesiano propuesto para elaborar la clasificación, desde la definición de la función de verosimilitud, la distribución previa, la distribución final y la predictiva,

hasta los modelos que se utilizan para aproximar dichas funciones.

En el tercer capítulo, se describe a detalle la clasificación mediante modelos de probabilidad, se expone la metodología, las problemáticas que surgen al aplicar este modelo, como es el *label switching* y las alternativas que se proponen para resolverlas.

Una vez expuesta la metodología en el capítulo cuatro, se procede a describir como se aplica de manera específica a un conjunto de datos que posee variables tanto discretas como continuas. Asimismo, en este capítulo se detalla cómo se estiman las distribuciones y parámetros correspondientes de las variables antes mencionadas, mediante un algoritmo que aplica el método *Gibbs Sampler*.

El capítulo cinco contiene la aplicación práctica del modelo de clasificación, que se lleva a cabo con una base de datos de una empresa de empeño, cuyo objetivo es clasificar los clientes en distintos grupos con base en la información proporcionada por seis variables seleccionadas en el análisis exploratorio.

Por último en el capítulo seis, 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.

1.2. Introducción a los modelos de clasificación no supervisada

El problema en el que se enfoca la clasificación no supervisada es encontrar el número de grupos a los cuales se puede clasificar una observación.

Por lo que, para efectos del trabajo que se presenta a continuación, es necesario definir los dos enfoques bajo los cuales se aborda la clasificación no supervisada; la primera y más conocida es mediante argumentos geométricos o distancias, mientras que la segunda, en la que se basa este caso, basa en distribuciones o modelos de probabilidad.

En el enfoque de clasificación mediante distancias, es importante destacar, que la asignación a un grupo en específico se debe a que tan cercana, en términos de distancia, está una observación del centroide; es decir, la observación se clasifica en el grupo cuando la distancia entre el centroide y la observación es mínima. Este enfoque es el más utilizado, sin embargo, este trabajo se basa en el segundo método, el cual utiliza modelos de probabilidad que generan reglas para asignar observaciones a los distintos grupos.

De manera simultanea, en el modelo que definición de las reglas de asignación,

se aborda el problema de la partición en el número óptimo de grupos en los cuales se clasifican los datos.

En el algoritmo desarrollado en la aplicación práctica de la clafisicación no supervisada, se muestra como se estima el número de grupos en los que se clasifica la información de los clientes de una casa de empeño es clasificada en k grupos, utilizando un método Gibbs Sampler, que además, mediante iteraciones, estima las distribuciones predictivas que determinan las probabilidades de pertenencia de un individuo a un grupo, y así poder clasificarlo.

Capítulo 2

Inferencia Bayesiana

En inferencia estadística, hay dos enfoques que prevalecen en la práctica al momento de interpretar la probabilidad: la Inferencia Bayesiana y la Inferencia Frecuentista. Estos dos paradigmas inferenciales, suelen diferir con respecto a la naturaleza fundamental de la probabilidad. La alternativa frecuentista la define de una manera más restrictiva, como el límite de la frecuencia relativa de un evento en un gran número de intentos, bajo el 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, en ocaciones, dada una evidencia.

La Inferencia bayesiana 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 (Hall, 2012). 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.

2.1. Verosimilitud

En general, para realizar un análisis estadístico de un conjunto de datos observados, $x_1, ..., x_n$, con $x_j \in \mathbb{R}^p$, se supone un modelo estocástico de n-variables aleatorias independientes e idénticamente distribuidas, $(X_j, j = 1, ..., n)$.

Por lo que, se define un modelo paramétrico $P(X) = F(x|\theta)$, donde $F(\cdot|\theta)$

es una función de distribución dada [1]. En esta ocasión, 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 manera que, bajo el supuesto de independencia, tenemos lo siguiente:

$$P(X_1, ..., X_n) = \prod_{j=1}^{n} P_j(X_j) \tag{2.1}$$

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

Bajo el método de máxima verosimilitud, se utiliza la información disponible del conjunto de datos para encontrar los valores más probables del parámetro θ dentro del espacio parametral Θ, asociados con los datos $X_1, ..., X_n$ observados.

Al considerar $P(X_j)$ como una función de distribución o densidad, de las observaciones dado el parámetro θ, $P(X_j) = F(\cdot|\theta)$, obtenemos la siguiente función de verosimilitud:

$$L(x_1, ..., x_n; \theta) = f(x_1, ..., x_n|\theta) = \prod_{j=1}^{n} f(x_j|\theta) \tag{2.2}$$

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

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

de donde se obtiene el valor de θ que maximiza la función,

$$\hat{\theta} = argmax \prod_{j=1}^{n} f(x_j|\theta) \tag{2.3}$$

y así encontrar el Estimador de Máxima Verosimilitud (EVM)

2.2. Paradigma Bayesiano

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. Bajo este enfoque bayesiano,para realizar inferencias sobre cierta hipótesis, se deben especificar las creencias anteriores con base en información disponible antes de haber observado los datos y de esa forma 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,

$P(X_1, ..., X_n|\theta)$, que como se mencionó anteriormente, esta última es la función de verosimilitud.

$$\pi(\theta|X_1, ..., X_n) = \frac{P(X_1, ..., X_n|\theta) \cdot \pi(\theta)}{P(X_1, ..., X_n)} \tag{2.4}$$

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,

$$\pi(\theta|X_1, ..., X_n) \propto P(X_1, ..., X_n|\theta) \cdot \pi(\theta) \tag{2.5}$$

donde $\propto$ denota una relación de proporcionalidad, dada por $P(X_1, ..., X_n)^{-1}$

2.3. 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 son siempre un conjunto de observaciones intercambiables (Bernardo, 1998).

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

$$P(X_1 = x_1, ..., X_n = x_n) = P(X_{\sigma(1)} = x_1, ..., X_{\sigma(n)} = x_n) \qquad (2.6)$$

donde $\{\sigma(1), ..., \sigma(n)\}$ es cualquier permutación de {1,2,...,n}.

2.4. Teorema de Representación

Utilizando en concepto de intercambiabilidad, de Finetti demuestra su famoso teorema de representación para variables dicotómicas. En este caso en particular, la intercambiabilidad identifica las observaciones como una muestra aleatoria de un modelo probabilístico específico (Bernoulli) y garantiza la existencia 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:

$$P(X_1, ..., X_n) = \int_{\Theta} \prod_{j=1}^{n} P(X_j|\theta)\pi(\theta)d\theta, \tag{2.7}$$

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, cualquier subconjunto de éstas, es una muestra aleatoria de un modelo $P(X_j|\theta)$ y existe una distribución inicial $\pi(\theta)$ que describe la información inicial disponible del parámetro θ.

2.5. Distribución inicial o previa

La distribución inicial o previa, $\pi(\theta)$, se define como la medida de probabilidad sobre Θ que describe el comportamiento del parámetro θ, con base en el nivel de información disponible antes de observados los datos.

Existen distintas maneras de obtener la distribución inicial para θ, en algu-

nas situaciones es posible basarse en información que proviene de evidencia acumulada de experimentos pasados. De igual manera, 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 información de tipo objetivo sobre el parámetro, como por ejemplo, θ toma valores positivos (Congdon, 2007). En la práctica, hay métodos que resultan más convenientes matemáticamente hablando, como es el caso de escoger una distribución inicial conjugada, debido a que de esa forma la distribución final $\pi(\theta|\cdot)$ es de una forma paramétrica conocida.

Se dice que, $\pi(\theta)$ y $\pi(\theta|\cdot)$ son distribuciones conjugadas, cuando la distribución final pertenece a la misma familia que la distribución inicial. Por ejemplo, si la función de verosimilitud es de la familia Gaussiana, entonces al elegir una distribución Gaussiana como distribución previa del parámetro, la media en este caso, nos asegurará que la distribución final se una distribución Gaussiana. En general todas las distribuciones de probabilidad de la familia exponencial cuentan con distribuciones previas conjugadas (Hall, 2012).

2.6. 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,

$$P(X_1, ..., X_n|\theta) = \prod_{j=1}^{n} P(X_j|\theta)$$

en la que se dice que las X_j's son intercambiables y condicionalmente independientes dado θ. Por lo que la distribución final de θ es:

$$\pi(\theta_1, ..., \theta_q|X_1, ..., X_n) = \frac{\prod_{j=1}^{n} P(X_j|\theta_1, ..., \theta_q) \cdot \pi(\theta_1, ..., \theta_q)}{\int_\Theta \prod_{j=1}^{n} P(X_j|\tilde{\theta_1}, ..., \tilde{\theta_q}) \cdot \pi(\tilde{\theta_1}, ..., \tilde{\theta_q}) d\tilde{\theta_1}, ..., \tilde{\theta_q}} \tag{2.8}$$

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,

$\int_\Theta \frac{1}{C} \cdot \prod_{j=1}^{n} P(X_j|\theta) \cdot \pi(\theta) d\theta = 1$. En la práctica, se conoce explícitamente sólo en términos del numerador (Congdon, 2007),

$$\pi(\theta_1, ..., \theta_q|X_1, ..., X_n) \propto \prod_{j=1}^{n} P(X_j|\theta) \cdot \pi(\theta) d\theta \tag{2.9}$$

El paso siguiente sería el cálculo de la distribución final, que puede derivar en dos escenarios:

a) Distribuciones conjugadas; es decir, la distribución previa y la distribución posterior son la misma forma funcional, se facilita el cálculo, ya que la distribución final resulta ser de una forma paramétrica conocida.

b) No conjugadas, donde la distribución final puede no conocerse explícitamente debido a la imposibilidad de obtener la constante de normalización correspondiente.

En este último, es necesario recurrir a distintos métodos de aproximación para calcular la distribución final de θ.

2.7. Distribución predictiva

Ahora, al utilizar argumentos probabilísticos, cómo es el supuesto de independencia, se busca obtener la distribución predictiva. Esta distribución es la que define como será el comportamiento de una nueva observación X_{n+1} con base en los datos observados $X_1, ..., X_n$.

$$
\begin{aligned}
P(X_{n+1}|X_1 = x_1, ..., X_n = x_n) &= \frac{P(X_1, ..., X_n, X_{n+1})}{P(X_1, ..., X_n)} \\
&= \frac{\prod_{j=1}^{n} P(X_j) \cdot P(X_{n+1})}{\prod_{j=1}^{n} P(X_j)} \\
&= P(X_{n+1}) \\
&= F(X_{n+1}|\theta) \quad ?
\end{aligned}
\tag{2.10}
$$

Como se observa en (2.10), este supuesto tiene como consecuencia un problema de predictibilidad, ya que la distribución de la nueva observación no estaría tomando en cuenta la información proporcionada por las observaciones anteriores. Por lo que, tomando el estimador de máxima verosimilitud (EMV) se recurre a lo siguiente:

$$
P(X_{n+1}|X_1, ..., X_n) \approx P[X_{n+1}|\hat{\theta}(X_1, ..., X_n)] \tag{2.11}
$$

Sin embargo, la ecuación (2.11) no puede ser considerada como una distribución predictiva, debido a que estimador del parámetro θ es un valor que se obtuvo con respecto a la información de $(X_1, ..., X_n)$ y X_{n+1} es independiente de estas observaciones anteriores.

Por lo que es momento de retomar los conceptos de la sección anterior, y proceder a obtener la distribución predictiva bajo el enfoque bayesiano.

$$\begin{aligned} P(X_{n+1}|X_1,...,X_n) &= \frac{P(X_{n+1}|X_1,...,X_n,X_{n+1})}{P(X_1,...,X_n)} \\ &= \frac{\int_\Theta \prod_{j=1}^n P(X_j|\theta)\cdot P(X_{n+1}|\theta)\cdot\pi(\theta)d\theta}{\int_\Theta \prod_{j=1}^n P(X_j|\tilde{\theta})\cdot\pi(\tilde{\theta})d\tilde{\theta}} \\ &= \int_\Theta P(X_{n+1}|\theta)\cdot\frac{\prod_{j=1}^n P(X_j|\theta)\cdot\pi(\theta)d\theta}{\int_\Theta \prod_{j=1}^n P(X_j|\tilde{\theta})\cdot\pi(\tilde{\theta})d\tilde{\theta}} \\ &= \int_\Theta P(X_{n+1}|\theta)\cdot\pi(\theta|X_1,...,X_n)d\theta \end{aligned} \tag{2.12}$$

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. A su vez, la ecuación (2.12), también se puede ver de la siguiente manera,

$$P(X_{n+1}|X_1,...,X_n) = E_{\theta|X_1,..,X_n}[P(X_{n+1}|\theta)] \tag{2.13}$$

Como la esperanza de la distribución $X_{n+1}|\theta$, condicional a los datos observados, $(X_1 = x_1,...,X_n = x_n)$. De esa manera la predicción del comportamiento

de los nuevos datos considera la información proporcionada por los datos observados anteriormente.

Para llegar a este punto de definir una distribución predictiva, es necesario como primer paso, encontrar la distribución inicial, y después la distribución posterior de los parámetros, como se explica en las secciones anteriores.

2.7.1. 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) (Hall, 2012).

I. Monte Carlo vía Cadenas de Markov

Para la aproximación de integrales, se utilizan en técnicas de simulación estocástica, empleando simulaciones de los datos para aproximar dicha integral que en su caso puede ser una función de distribución.

Supongamos que se tiene una integral de la forma:

$$\int h(\theta)\pi(\theta)d\theta \tag{2.14}$$

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.

Si se es capaz de generar 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 (2.14), que tambíen puede ser interpretada como el valor esperado de h sobre π,

$$E_\pi[h(\theta)] = \int h(\theta)\pi(\theta)d\theta$$

mediante un promedio de dichas simulaciones obtenidas, evaluadas en h

$$\hat{E}_\pi[h(\theta)] = \frac{1}{T}\sum_{i=1}^{T} h(\theta^{(i)}) \tag{2.15}$$

el estimador (2.15) que se conoce como el estimador de Montecarlo, es insesgado y converge casi seguramente con el valor de la integral de interés. Los resultados de esta sección son aplicables en los casos en los que θ represente algunos parámetros asociados a un modelo, o cuando represente variables aleatorias observables (Martínez Ovando, 2004).

Entonces, retomando la sección anterior, tenemos que la función de distribución predictiva (2.13) se puede ver de la siguiente manera:

$$E_{\theta|X_1,..,X_n}(P(X_{n+1}|\theta)) \approx \frac{1}{T}\sum_{t=1}^{T} P(X_{n+1}|\theta_1^{(t)},...,\theta_q^{(t)}) \tag{2.16}$$

donde la aproximación de (2.16), se puede interpretar como el promedio empírico de $P(X_{n+1}|\theta)$.

Este método es conocido como Monte Carlo, sin embargo, bajo el contexto bayesiano, generalmente conocemos la densidad $\pi(\cdot)$ salvo por una constante de normalización, que usualmente es difícil de calcular, y de hecho nos remonta al problema inicial de resolver una integral, que en este caso nos es difícil generar datos de la distribución $\pi(\cdot)$ directamente. Lo que lleva a que la estimación de funciones de distribución, como es la final o posterior, $\pi(\theta_1,...,\theta_q|X_1,...,X_n)$, llegue a ser muy compleja (Martínez Ovando, 2004). Se propone relajar el supuesto de *iid* de Monte Carlo. De esta forma, $\theta_1^{(t)},...,\theta_q^{(t)}$ 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:

$$g(\theta^{(t)}|\theta^{(t-1)},X_1,...,X_n) \tag{2.17}$$

La unión del método de cadenas de Markov con el método de Monte Carlo, es a lo que se le denomina método Monte Carlo vía Cadenas de Markov (MCMC).

La cadena de Markov es un proceso estocástico $\{\theta^{(0)}, \theta^{(1)}, \theta^{(2)}, ...\}$, donde $\theta^{(i)}$ es el estado del proceso en el tiempo i, y se define como una variable aleatoria cuyos valores se encuentran en un espacio de estados $\Theta \subset I\!R^p$.

Estas cadenas de Markov cumplen con una propiedad en la 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 la distribución invariante; es decir, que si $\theta^{(t)} \sim \pi$ implica que $\theta^{(t+1)} \sim \pi$. Donde 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.

Se dice que $\pi(.)$ es la distribución invariante de la cadena de Markov, si se define un *kernel* de transición que satisfaga la condición de balance $K(\theta, \theta')\pi(\theta) =$

$K(\theta', \theta)\pi(\theta')$ (Martínez Ovando, 2004).

De igual manera, esta cadena deberá cumplir con ciertas condiciones de regularidad (Congdon, 2007):

a) Irreductible, si para cualquier pareja de estados $(\theta^{(t)}, \theta^{(s)}) \in \theta$ existe una probabilidad distinta de que la cadena se puede mover de $\theta^{(t)}$ a$\theta^{(s)}$ en un número finito de pasos.

b) Aperiódica. Un estado tiene una periodicidad k, $(k > 1)$, si puede ser revisitado solamente después de un número de pasos que sea múltiplo de k, de otra manera el estado es aperiódico. por lo que si todos los estados son aperiódicos, la cadena es aperiódica.

c) Recurrencia positiva. Si el número de pasos para visitar cualquier estado de una cadena tiene media positiva.

d) Ergódica. Se asegura que el promedio aritmético (2.16) converja, casi seguramente, con la esperanza calculada bajo la distribución invariante conforme T se va haciendo más grande $(T \longrightarrow \infty)$. Si la cadena cumple con las condiciones anteriores, se dice que tiene ergodicidad.

En la siguiente sección se describe un método para construir cadenas de Markov que cumple con estas características.

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 de θ.

Supongamos que el parámetro vector θ está dividido en subvectores, $\theta = (\theta_1, ..., \theta_q)$. Cada una de las T iteraciones del Gibbs sampler, compuestas por q pasos, recorren los subvectores de θ haciendo que cada subconjunto sea condicional al valor de todos los demás. En cada una de estas iteraciones t, se elige un ordenamiento de los q subvectores de θ y a su vez, se genera una muestra de cada θ_j^t a partir de la distribución condicional dado los demás componentes de θ (Gelman et al., 2014).

Por lo que, método *Gibbs Sampler*, 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 $\theta^{(t)}$.

1). $\theta_1^{(t)}|\theta_2^{(t-1)},...,\theta_q^{(t-1)},X_1,...,X_n \sim \prod_{j=1}^{n} P(X_j|\theta_1,\theta_2^{(t-1)},...,\theta_q^{(t-1)}) \cdot \pi(\theta_1|\theta_2^{(t-1)},...,\theta_q^{(t-1)})$;

2). $\theta_2^{(t)}|\theta_1^{t},\theta_3^{(t-1)},...,\theta_q^{(t-1)},X_1,...,X_n \sim \prod_{j=1}^{n} P(X_j|\theta_1^{k},\theta_2,\theta_3^{(t-1)},...,\theta_q^{(t-1)}) \cdot \pi(\theta_2|\theta_1^{t},\theta_3^{(t-1)},...,\theta_q^{(t-1)})$;

.

.

q). $\theta_q^{(t)}|\theta_1^{t},...,\theta_{q-1}^{(t)},X_1,...,X_n \sim \prod_{j=1}^{n} P(X_j|\theta_1^{(t)},...,\theta_{q-1}^{(t)},\theta_q) \cdot \pi(\theta_q|\theta_1^{(t)},...,\theta_{q-1}^{(t)})$

(2.18)

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

Clasificación basada en Modelos

3.1. Modelos de Clasificación

Los tipos de clasificación se dividen en dos grandes grupos, la clasificación supervisada y la no supervisada. Para efectos de este análisis se utilizará la clasificación no supervisada, debido a que la información no tiene clases previas en las cuales se puedan ubicar las observaciones.

El problema de la clasificación no supervisada consiste en 'adivinar' o encon-

trar el número de grupos J con $J = 1, ..., n$ en los cuales se pueden clasificar las observaciones. Con base en los grupos definidos, se crean reglas de asociación con las cuales X_i es asignado a una clase $C_j \in \{1, 2, ..., J\}$

Dichas reglas de asignación pueden definirse bajo dos corrientes:

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

$$X_i \in C_{j^*} \text{ si y sólo si } j^* = \text{argmin}_j \, d(X_i, C_j).$$

Por ejemplo, para clasificar 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).

$$X_i \in C_1 \text{ si y sólo si } \{d(X_i, C_1) < d(X_i, C_2)\}$$

donde C_1 y C_2 son los centroides de cada grupo.

Este enfoque funciona siempre y cuando los datos sean escalares, pero en caso de que sean de tipo catégorico surge un problema, ya que no hay manera de medir las distancias.

Por otro lado, ha surgido una nueva escuela en la que 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. $X_i \in C_{j^*}$ si y sólo si $j^* = \text{argmax}_{j \in J} P(X_i \in C_j)$

Para comenzar la clasificación, 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).

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

donde $\theta_j = T_{F_j}(X)$, definiendo a $T_{F_j}(X)$ 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 $P(X_i \in C_j) \approx \frac{e_j}{n}$

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

Primero, al ser un problema de clasificación no supervisada, se desconoce el número K de grupos o clases en los que se clasificaran los datos, asi que también se desconoce cuáles son X_j's pertencen a la clase C_j. Los parámetros

$\theta_1, ...\theta_j$ asociados con $F_1(\cdot|\theta_1), ..., F_j(\cdot|\theta_j)$, son desconocidos, por lo que será necesario estimarlos.

En la próxima sección se propone un modelo, bajo el cual se abordan estas problemáticas.

3.2. Modelo basado en Mezclas de distribuciones

La estructura del modelo tipo mezcla aparece debido a que, como es común en los problemas de clasificación no supervisada, no cuenta con información sobre la pertenencia de cada observación a una subpoblación o clase específicas. Por lo tanto, se asume que cada una de las X_i's puede tener una distribución f_j con probabilidad p_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 com-

ponentes, 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,

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

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

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

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 $p = (p_1, ...p_J)$ y $\theta = (\theta_1, ..., \theta_J)$, vemos que aunque se hayan utilizado previas conjugadas para cada parámetro, para obtener la dostribución previa de manera explícita, requiere que se realice la expanción de la verosimilitud

$$L(\theta, p|x) = \prod_{i=1}^{n} \sum_{j=1}^{J} p_i f_j(x_i|\theta_j) \tag{3.2}$$

en k^n términos, que para la práctica resulta ser muy costoso computacionalmente hablando (Marin et al., 2005).

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

$$\pi(j, \theta_j|x_1, ..., x_n) = \frac{P(x_1, ..., x_n|\theta_j, j) \cdot \pi(\theta_j|j) \cdot \pi(j)}{\sum_{j=1}^{2}\{\int_{\Theta_j} P(x_1, ..., x_n|\theta_j, j) \cdot \pi(\theta_j|j)d\theta_j\}\}\pi(j)} \tag{3.3}$$

Bajo el enfoque bayesiano la ecuación (3.3) se puede ver representar así,

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

donde el denominador de (3.3) es la constante de normalización y que su obtención es muy compleja, por lo que se propone una alternativa a este método.

Definimos a la distribución posterior del parámetro θ y a la función de distribución final del modelo respectivamente como,

$$\pi(\theta_j|x_1, ..., x_n, j) = \frac{P(x_1, ..., x_n|\theta_j, j) \cdot \pi(\theta_j, j)}{\int_{\Theta_j} P(x_1, ..., x_n|\theta_j, j) \cdot \pi(\theta_j|j)d\theta_j} \tag{3.4}$$

$$\pi(j|x_1, ..., x_n) = \frac{P(x_1, ..., x_n|j) \cdot \pi(j)}{\sum_{j=1}^{J} P(x_1, ..., x_n|j) \cdot \pi(j)} \tag{3.5}$$

donde $P(x_1, ..., x_n|j) = \int_{\Theta_j} P(x_1, ..., x_n|\theta_j, j) \cdot \pi(\theta_j|j)d\theta_j$, y se tienen las siguentes funciones de verosimilitud y de distribución inicial de los parámetros,

$$P(x_1, ..., x_n|\theta_j, j) = \prod_{i=1}^{n} N(x_i|\theta_j, 1) \tag{3.6}$$

$$\pi(\theta_j|j) = N(\theta_j|\mu_0, \sigma_0^2 = 1) \tag{3.7}$$

Para obtener la distribución posterior de los parámetros, sustituimos en (3.4) tanto la función de verosimilitud como la distribución inicial de los parámetros.

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

se expanden los cuadrados con el fin de factorizar con respecto a θ_j,

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

se suma y resta el término $\left(\frac{n\bar{x} + \mu_0}{n+1}\right)$ para completar el cuadrado,

$$=\frac{1}{2\pi}\exp\left\{-\frac{1}{2}\left(\theta_j-\frac{n\bar{x}+\mu_0}{n+1}\right)^2\right\}\cdot\exp\left\{-\frac{1}{2}\left(-\left(\frac{n\bar{x}+\mu_0}{n+1}\right)^2+\frac{\mu_0^2+\sum_{i=1}^n x_i^2}{n+1}\right)\right\} \tag{3.8}$$

De esta manera se puede ver que el primer término de (3.8) es el kernel de una distribución normal con media $\frac{n\bar{x}+\mu_0}{n+1}$ y varianza 1,

$$\pi(\theta_j|x_1,...,x_n,j)\sim N\left(\frac{n\bar{x}+\mu_0}{n+1},1\right) \tag{3.9}$$

además, la constante de normalización de $\pi(\theta_j|x_1,...,x_n,j)$ que garantiza que sea una función propia es,

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

Los resultados obtenidos en la parte de arriba, se utlizarán para estimar la función de distribución final del modelo (3.5).

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

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,

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

Como se puede ver en este ejemplo encontrar la distribución del modelo puede resultar engorroso, por lo que se proponen diferentes enfoques abordar el problema como la introducción de variables latentes que se explica en la siguiente sección.

I. Variables Latentes

Una manera de facilitar la estimación es introducir, dentro del modelo, variables aleatorias no observadas (variables latentes) $\underline{z} = (z_1, ..., z_n)$, que identifican a que componente j, (j=1,...J) pertenece cada una de las observaciones $x = (x_1, ..., x_n)$.

$$X_i|Z_i = z \sim f(x|\theta_z) \tag{3.10}$$

Donde $Z_i \sim M_j(1; p_1, ...p_J)$. Si tomamos la ecuación 3.1 y definimos a $\pi(\theta, p)$ como la distribución inicial de (θ, p). La distribución posterior es la siguiente,

$$\pi(\theta, p|X) \propto (\prod_{i=1}^{n} \sum_{j=1}^{J} p_i f_j(X_i|\theta_j))\pi(\theta, p) \tag{3.11}$$

Definimos Z como el conjunto de todos los J_n vectores de asignación z, que podemos descomponer 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 $\sum_{j=1}^{J} n_j = n$.

$$r = \binom{n+k-1}{n} \tag{3.12}$$

De esa manera tenemos la partición $Z = \cup_{i=1}^{r} Z_i$. Y 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,

$$\pi(\theta, p|x) = \sum_{i=1}^{r} \sum_{\underline{z} \in Z_i} w(\underline{z}) \pi(\theta, p|x, \underline{z}) \tag{3.13}$$

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

3.3. *Label Switching Problem*

El término *label switching* se utiliza para describir la invarianza de la función de verosimilitud 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,

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

Lo que a continuación nos lleva a la raíz del problema de *label switching*. La función de verosimilitud es la misma para todas las permutaciones de θ. De igual manera, bajo el enfoque bayesiano, si no se cuenta con la información previa que distinga entre los componentes del modelo de mezclas, la distribución previa $\pi(\theta)$ será la misma para todas las permutaciones y por consiguiente 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. Por ejemplo, gracias a esta simetría, las funciones de densidad predictivas son las mismas para cada componente, de forma que las probabilidades de clasificación no son se utilidad para la clasificación 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 en específico será la misma que la de media de ese parámetro en los demás componentes del modelo, por lo que en general suele ser una estimación muy pobre para esos parámetros (Stephens, 2000).

Se han propuesto diversas alternativas para resolver el problema, como es la de imponer una restricción (*identifiability constraint*) sobre los parámetros, como 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 et al., 2005)

3.4. Inferencia bayesiana en modelos de mezclas

La clasificación no supervisado se basa en modelar los datos bajo un enfoque probabilístico de particiones aleatorias, generando distintos grupos, que están definidos por los individuos / observaciones contenidas dentro de éstos. Por lo tanto el número de grupos o clases, depende también de las particiones. Originalmente, este enfoque fue desarrollado para datos categóricos, sin embargo ahora se propone extenderlo a datos mixtos. Más adelante 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. 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 analitícamente. (Blomstedt et al., 2015)

Tomando en cuenta todos los conceptos definidos anteriormente, suponemos un conjunto N de n observaciones $i \in N$ que se caracterizan como vectores d-dimensionales $x^{(i)} = (x_1^{(i)}, ..., x_d^{(i)}) \in \mathbb{R}^d$. El conjunto de datos se define como $x = (x^1, ..., x^n)^T$. Ahora, se lleva a cabo la partición N en k conjuntos

no vacios 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 distribución de mezclas, puede ser definida como:

$$F(x) = P(x \in D)F(x|X \in D) + P(x \in D^C)F(x|X \in D_C), \tag{3.14}$$

donde D es un conjunto de puntos discontinuos con respecto a F y finito con cardinalidad r. Aún sin conocer la forma de la distribución del componente discreto, suponemos un vector aleatorio $(Y_1, ..., Y_r)$ de manera que:

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

con $l = 1, ..., r$, por lo que la distribución de probabilidad del componete discreto se define como sigue:

$$f_D(y_1, ..., y_r|\psi_1, ..., \psi_r) = \prod_{l=1}^{r} \psi_l^{y^l} \tag{3.15}$$

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 $f_C(x|\lambda)$.

Con los datos del grupo c y características j, que suponemos observados, se

definen $x_{cj} = (x^{(1)}, ..., x^{(n_c)})^T$, tenemos la función de verosimilitud

$$L_{x_{cj}}(w, \psi_1, ...\psi_r, \lambda) = \prod_{i=1}^{n_c} \left[(1-w) \prod_{l=1}^{r} \psi_l^{y_l^{(i)}} + w f_C(x^{(i)}|\lambda) \right] \tag{3.16}$$

donde $w = P(X^{(i)} \in D^c) \forall i = 1, ..., n_c$, y asumiendo que las observaciones son condicionalmente independientes. Al expandir la ecuación 3.16 se obtienen 2^{n_c} términos, que complican 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,

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

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

$$L_{x_{cj}}(p_1, ..., p_{r+1}, \lambda) = \prod_{i=1}^{n_c} \left[p_1^{y_1^{(i)}}, ..., p_r^{y_r^{(i)}} + p_{r+1} f_c(x^i|\lambda)^{y_{r+1}^{(i)}} \right] \tag{3.17}$$

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}.

$$L_{x_{cj}}(p_1, ..., p_{r+1}, \lambda) = \prod_{l=1}^{r+1} p_l^{n_{cl}} \prod_{m=1}^{n_{c,r+1}} f_c(z_{(m)}|\lambda) = L_{y_{cj}}(p_1, ..., p_{r+1}) L_{z_{cj}}(\lambda) \tag{3.18}$$

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 o *cluster* se puede obtener la función marginal de verosimilitud para una partición S dada. Se define θ_{cj} de manera conjunta como $(p_{cj,1}, ..., p_{cj,r+1}, \lambda_{cj})$ y análogamente, θ_S se define como el conjunto de parámetros de la partición S. Suponiendo independencia de los clusters dados los parámetros, la función de verosimilitud conjunta para θ_S

$$L_x(\theta_S) = \prod_{c=1}^{k} \prod_{j=1}^{d} L_{y_{cj}}(p_{cj,1}, ..., p_{cj,r+1}) L_{z_{cj}}(\lambda_{cj}) \tag{3.19}$$

y si suponemos independencia entre λ_{cj} y $(p_{cj,1}, ..., p_{cj,r+1})$, se puede factorizar la previa $\pi(\theta_{cj})$,

$$\pi(\theta_{cj}) = \pi(p_{cj,1}, ..., p_{cj,r}) \pi(\lambda_{cj}) \tag{3.20}$$

lo que lleva a que,

$$\pi(\theta_S) = \prod_{c=1}^{k} \prod_{j=1}^{d} \pi(p_{cj,1}, ..., p_{cj,r+1}) \pi(\lambda_{cj}) \tag{3.21}$$

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

$$
\begin{aligned}
p(x|S) &= \int_{\Theta_S} L_X(\theta_S)\pi(\theta_S)d\theta_S \qquad (3.22)\\
&= \int_{P_S} \left[\prod_{c=1}^{k} \prod_{j=1}^{d} L_{y_{cj}}(p_{cj,1}, ..., p_{cj,r+1})\pi(p_{cj,1}, ..., p_{cj,r+1}) \right] d_{P_S} \\
&\quad \cdot \int_{H_S} \prod_{c=1}^{k} \prod_{j=1}^{d} L_{z_{cj}}(\lambda_{cj})\pi(\lambda_{cj})d_{\lambda_S} \\
&= p(y|S)p(z|S)
\end{aligned}
$$

lo que 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 el próximo capítulo.

3.5. Predicción(numérica)

Para obtener la distribución predictiva, primero es necesario obtener una posterior mediante un ratio de probabilidades de la previa $\frac{\pi(p')}{\pi(p)}$ donde p' se obtiene de p al reasignar un objeto de dicha partición.

3.5.1. Métodos para elegir el número de clases

Existen varios métodos para definir el número de clases latentes de modelos basados en mezclas; sin embargo aún no existe un consenso general para elegir el mejor enfoque.

A continuación se presentan las distintas propuestas para abordar el tema:

1. *Deviance information criterion.* Este enfoque se basa en una compensación entre la bondad de ajuste (devianza) y la complejidad del modelo (el número de párametros P_D) y se calcula de la siguiente manera:

$$D(\theta) = -2logf(y|\theta) + 2logh(y) \tag{3.23}$$

donde $h(y)$ es un término estandarizado en función de los datos y el número estimado de parámetros se define como $P_D = \bar{D}(\theta) - D(\hat{\theta})$ donde $\bar{D}(\theta)$ es la devianza media posterior y $\hat{\theta} = \mathbf{E}[\theta|y]$ es la media posterior de los parámetros del modelo.

$$DIC = -4E_\theta[logf(y|\theta)|\theta] + 2logf(y|\hat{\theta}) \tag{3.24}$$

2. *Reversable jump MCMC algorithm.* Este método permite hacer *sampling* de la distribución posterior en distintos espacios con diferentes dimensiones.

3. *Rosseau and Mengersen's criterion.* Consiste en definir un modelo sobre ajustado con $K_{max}(K_{max} > K)$ clases latentes para los datos.

4. *Bayesian information criterion.* Se define como:

$$BIC = -2\left[log f(y|\hat{\theta})\right] + g\dot{l}og(n), \tag{3.25}$$

donde $\hat{\theta}$ es el estimado de máxima verosimilitud del parámetro θ, g es el número de parámetros libres en el modelo, y n es el número de observaciones en los datos.

Capítulo 4

Clasificación con datos discretos y continuos

4.1. Revisión de la propuesta

En esta sección se describe a detalle cual es la propuesta de modelo que utiliza para 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.

Las d variables continuas se distribuyen de la forma Normal-Multivariada,

$X^c \sim N(\mu_j, \Sigma_j)$, mientras que cada una de las p variables de conteo tiene una distribución Poisson, $X_l^d \sim Po(\lambda_{lj})$.

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

$$\sum_{j=1}^{K}(\pi_j \cdot N(\mu_j, \Sigma_j) \cdot \prod_{l=1}^{p} Po(\lambda_{lj})) \tag{4.1}$$

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

$$E[z_{ij}|x] = \frac{\pi_j^{(t)} \cdot N(\mu_j, \Sigma_j) \cdot \prod_{l=1}^{p} Po(\lambda_{lj})}{\sum_{j=1}^{K} \pi_j^{(t)} \cdot N(\mu_j, \Sigma_j) \cdot \prod_{l=1}^{p} Po(\lambda_{lj})} \tag{4.2}$$

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

$$P(z, \theta) = \prod_{j=1}^{K} \pi_j^{z_j} \tag{4.3}$$

$$P(x|z_j, \theta) = (N(\mu_j, \Sigma_j) \cdot \prod_{l=1}^{p} Po(\lambda_{lj}))^{z_j} \tag{4.4}$$

Por definición obtenemos la distribución conjunta del Modelo incluyendo la

variable latente,

$$P(x, z; \theta) = \prod_{j=1}^{K} \prod_{i=1}^{n} \left(\pi_j \cdot N(\mu_j, \Sigma_j) \cdot \prod_{l=1}^{p} Po(\lambda_{lj}) \right)^{z_{ij}} \tag{4.5}$$

4.2. 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.

$$\begin{aligned} P(\theta|x, z) = P(x, z|\theta) \cdot P(\pi|\alpha_1, ..., \alpha_k) \\ \cdot \prod_{j=1}^{K} \left[P(\Sigma_j|\Lambda_j, v_j) \cdot P(\mu_j|\varepsilon_j, \frac{\Sigma_j}{n_j}) \cdot \prod_{l=1}^{p} P(\lambda_{lj}|a_{lj}, S_{lj}) \right] \end{aligned} \tag{4.6}$$

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.

4.2.1. 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:

$$
\begin{aligned}
P(x,z|\theta) &= \prod_{i=1}^{n}\left(\prod_{l=1}^{p} Po(\lambda_{lj})\right)^{z_{ij}} \\
&\propto \prod_{i=1}^{n}\left(\prod_{l=1}^{p} \lambda_{lj}^{x_{li}} \exp\{-\lambda_{lj}\}\right)^{z_{ij}} \\
&= \prod_{l=1}^{p}\prod_{i=1}^{n} \lambda_{lj}^{x_{li}\cdot z_{ij}} \exp\{-\lambda_{lj}\cdot z_{ij}\} \\
&= \prod_{l=1}^{p} \lambda_{lj}^{\sum_{i=1}^{n} x_{li}\cdot z_{ij}} \exp\left\{-\lambda_{lj}\sum_{i=1}^{n} z_{ij}\right\} \\
&= \prod_{l=1}^{p} \lambda_{lj}^{\bar{z}_j \bar{x}_{lj}} exp\{-\bar{z}_j \lambda_{lj}\}
\end{aligned}
\tag{4.7}
$$

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 (4.7) y las distribuciones previas conjugadas de los parámetros $P(\lambda_j|a_j, S_j)$, que serán definidas en la sección 4.3,

$$
\begin{aligned}
P(\theta|x,z) &= P(x,z|\theta)\cdot\prod_{l=1}^{p} P(\lambda_{lj}|a_{lj},S_{lj}) \\
&\propto \prod_{l=1}^{p} \lambda_{lj}^{\bar{z}_j\bar{x}_{lj}} exp\{-\bar{z}_j\lambda_{lj}\}\cdot\lambda_{lj}^{a_{lj}}\exp\{-S_{lj}\lambda_{lj}\} \\
&= \prod_{l=1}^{p} \lambda_{lj}^{\bar{z}_j\bar{x}_{lj}+a_{lj}}\cdot exp\{-\lambda_{lj}(\bar{z}_j+S_{lj})\} \\
&= \prod_{l=1}^{p} Ga(\tilde{a}_{lj},\tilde{S}_{lj})
\end{aligned}
\tag{4.8}
$$

4.2.2. 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 verosimilitud,

$$
\begin{aligned}
P(x,z|\theta) &= \prod_{i=1}^{n} \left(N(x_i|\mu_j,\Sigma_j)\right)^{z_{ij}} \qquad (4.9)\\
&\propto \left(|\Sigma_j^{-1}|^{\frac{1}{2}} \exp\left\{-\frac{1}{2}(x_i-\mu_j)'\Sigma_j^{-1}(x_i-\mu_j)\right\}\right)^{z_{ij}}\\
&= |\Sigma_j^{-1}|^{\frac{\sum_{i=1}^{n} z_{ij}}{2}} \exp\left\{-\frac{1}{2}\sum_{i=1}^{n} z_{ij} trace\left[\Sigma_j^{-1}(x_i-\mu_j)(x_i-\mu_j)'\right]\right\}\\
&= |\Sigma_j^{-1}|^{\frac{\sum_{i=1}^{n} z_{ij}}{2}} \exp\left\{-\frac{1}{2} trace\left[\Sigma_j^{-1}\sum_{i=1}^{n} z_{ij}(x_i x_i' - \mu_j x_i' - x_i\mu_j' + \mu_j\mu_j')\right]\right\}\\
&= |\Sigma_j^{-1}|^{\frac{\sum_{i=1}^{n} z_{ij}}{2}} \exp\left\{-\frac{1}{2} trace\left[\Sigma_j^{-1}\sum_{i=1}^{n} z_{ij}(x_i x_i' - \bar{x}_j\bar{x}_j' + \bar{x}_j\bar{x}_j' - \mu_j x_i' - x_i\mu_j' + \mu_j\mu_j')\right]\right\}\\
&\propto \exp\left\{-\frac{1}{2} trace\left[\Sigma_j^{-1}\sum_{i=1}^{n} z_{ij}(x_i x_i' - \bar{x}_j\bar{x}_j' + \bar{x}_j x_i' - x_i\bar{x}_j') + \bar{z}_j(\bar{x}_j-\mu_j)(\bar{x}_j-\mu_j)'\right]\right\}\\
&\propto |\Sigma_j^{-1}|^{-\frac{\bar{z}_j}{2}} \exp\left\{-\frac{1}{2} trace\left[\Sigma_j^{-1}\left(\bar{z}_j(\bar{x}_j-\mu_j)(\bar{x}_j-\mu_j)' + \sum_{i=1}^{n} z_{ij}(x_i-\bar{x}_j)(x_i-\bar{x}_j)'\right)\right]\right\}
\end{aligned}
$$

Una vez obtenida la función de verosimilitud (4.7), se procede a desarrollar la función posterior,

$$
\begin{aligned}
P(\theta|x,z) &= P(x,z|\theta)\cdot P(\Sigma_j|\Lambda_j,v_j)\cdot P(\mu_j|\varepsilon_j,\frac{\Sigma_j}{n_j}) \qquad (4.10)\\
&\propto|\Sigma_j^{-1}|^{-\frac{\bar{z}_j}{2}}\exp\left\{-\frac{1}{2}trace\left[\Sigma_j^{-1}\left(\bar{z}_j(\bar{x}_j-\mu_j)(\bar{x}_j-\mu_j)'+\sum_{i=1}^{n}z_{ij}(x_i-\bar{x}_j)(x_i-\bar{x}_j)'\right)\right]\right\}\\
&\quad\cdot|\Sigma_j^{-1}|^{-\frac{1}{2}}\exp\left\{-\frac{n_j}{2}trace\left[(\mu_j-\varepsilon_j)'\Sigma_j^{-1}(\mu_j-\varepsilon_j)\right]\right\}\\
&\quad\cdot|\Sigma_j^{-1}|^{-(v_j+d+1)/2}\exp\left\{-\frac{1}{2}trace\left[\Sigma_j^{-1}\Lambda_j\right]\right\}\\
&=|\Sigma_j^{-1}|^{-(\tilde{v}_j+d+1)/2}\exp\left\{-\frac{1}{2}trace\left[\Sigma_j^{-1}\tilde{\Lambda}_j\right]\right\}\\
&\quad\cdot|\Sigma_j^{-1}|^{-\frac{1}{2}}\exp\left\{-\frac{\tilde{n}_j}{2}trace\left[(\mu_j-\tilde{\varepsilon}_j)'\Sigma_j^{-1}(\mu_j-\tilde{\varepsilon}_j)\right]\right\}\\
&=N(\mu_j|\tilde{\varepsilon}_j,\frac{\Sigma_j}{\tilde{n}_j})\cdot W^{-1}(\Sigma_j|\tilde{v}_j,\tilde{\Lambda}_j)
\end{aligned}
$$

El desarrollo completo se encuentra en el anexo.

4.3. 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, se elegiran las previas de una Normal Multivariada para el parámetro de medias y una Wishart Inversa para el parámetro de la varianza, ya que al tener una función de verosimilitud

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. En el caso de las variables discretas, se elegirán la distribución previa como una Gamma ya que la función de verosimilitud esta compuesta de distribuciones Poisson y se 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. 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.

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 :

$$\begin{aligned} P(\theta|x,z) &= P(x,z|\theta) \cdot P(\pi|\alpha_1, ..., \alpha_k) \\ &\quad \cdot \prod_{j=1}^{K} \left[P(\Sigma_j|\Lambda_j, v_j) \cdot P(\mu_j|\varepsilon_j, \frac{\Sigma_j}{n_j}) \cdot \prod_{l=1}^{p} P(\lambda_{lj}|a_{lj}, S_{lj}) \right] \\ &= \prod_{j=1}^{k} Dir(\pi_1, \pi_k|\tilde{\alpha}_1, ..., \tilde{\alpha}_k) \cdot G(\tilde{a}_j, \tilde{S}_j) \cdot N(\tilde{\varepsilon}_j, \frac{\Sigma_j}{\tilde{n}_j}) \cdot W^{-1}(\tilde{\Lambda}_j, \tilde{v}_j) \end{aligned} \tag{4.11}$$

Capítulo 5

Aplicación práctica

5.1. Objetivo

Obtener una clasificación de los clientes de la empresa de empeño en cuestión con base en los modelos descritos en los capítulos anteriores, utilizado la información contenida en las variables de la base de datos, de modo que a estos clientes se les asigne un grupo y así, con base en esta clasificación, ofrecerles distintos productos.

5.2. Descripción 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

5.2.1. Descripción de la base

La base 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** (v_1) es la fecha de registro en la que el cliente se dio de alta en el sistema.

2. **Edad** (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.

5.2.2. 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.

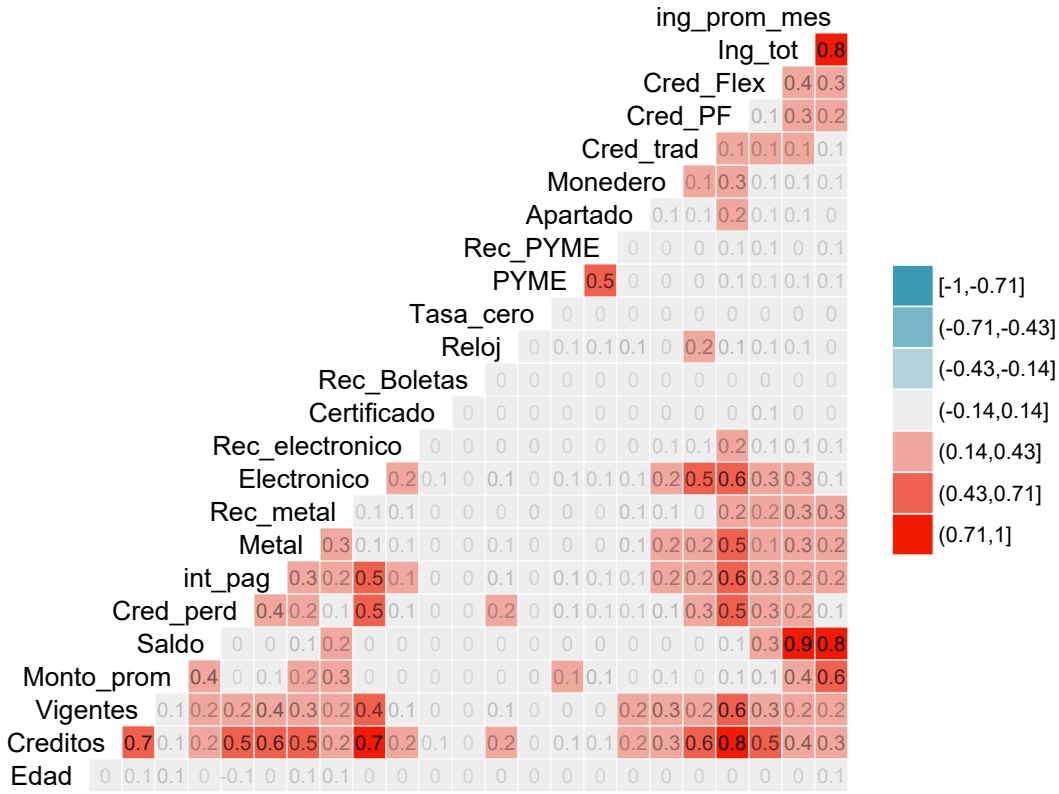


Figura 5.1: Gráfica de correlaciones

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:

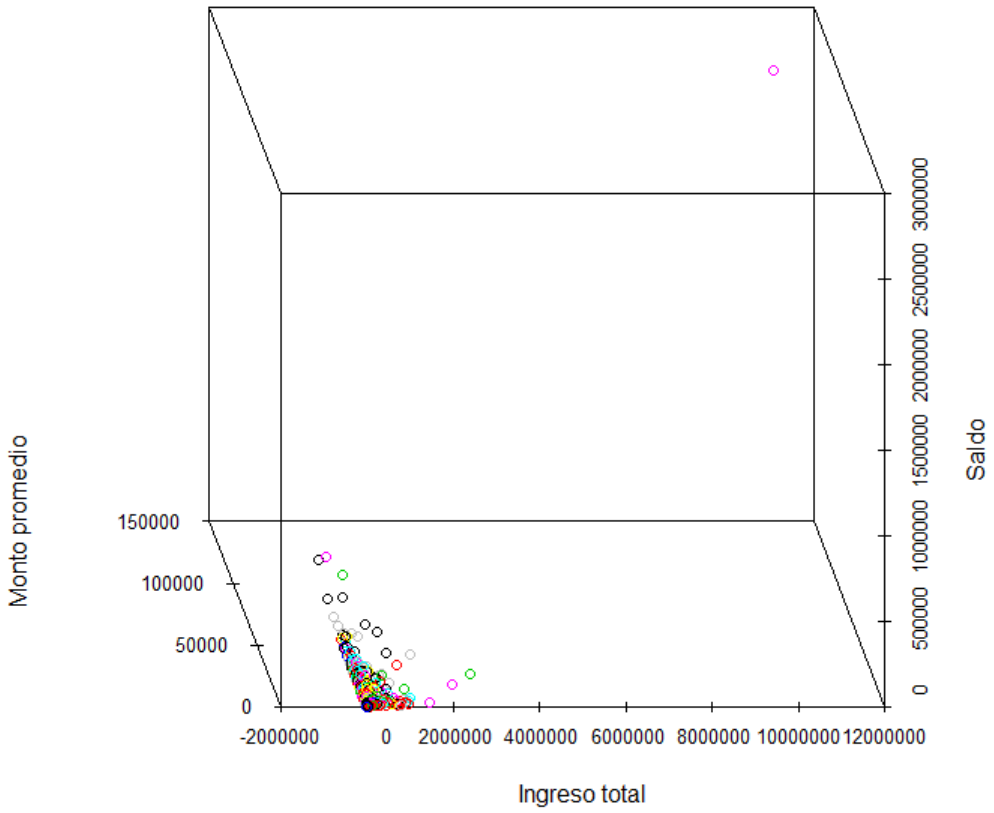


Figura 5.2: Dispersión de las variables Ing tot, Monto prom y Saldo

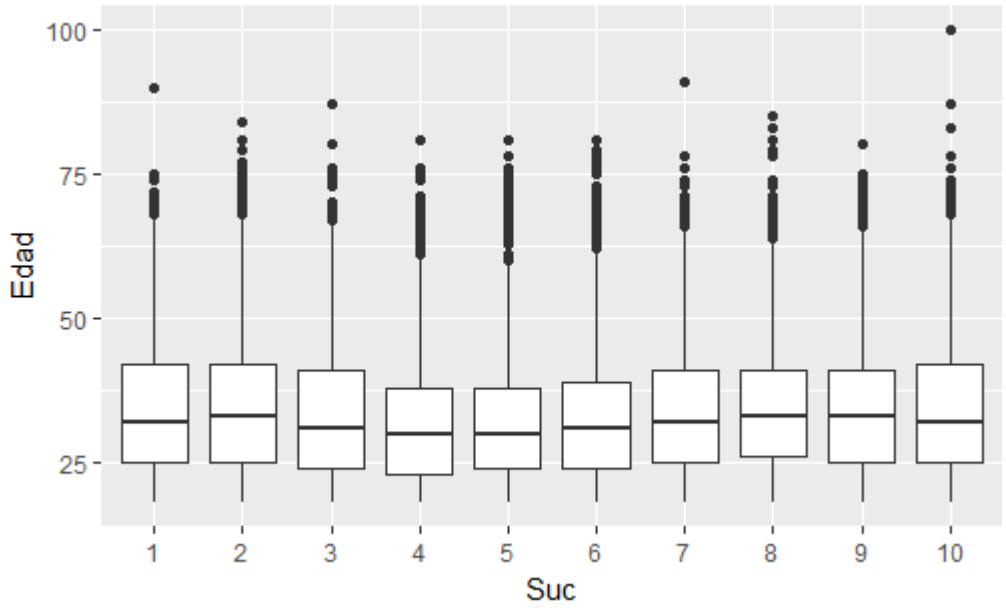


Figura 5.3: Distribución de las edades de los clientes por sucursal

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

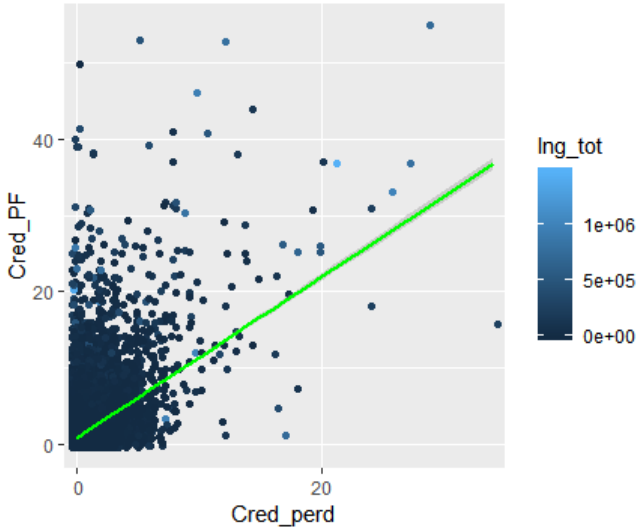


Figura 5.4: Gráfica de dispersión crédito en pagos fijos vs créditos perdidos

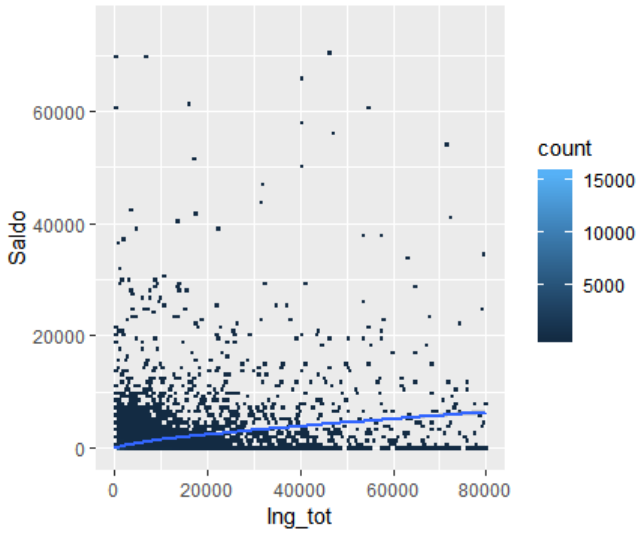


Figura 5.5: Gráfica de dispersión de Ingreso Total vs Saldo

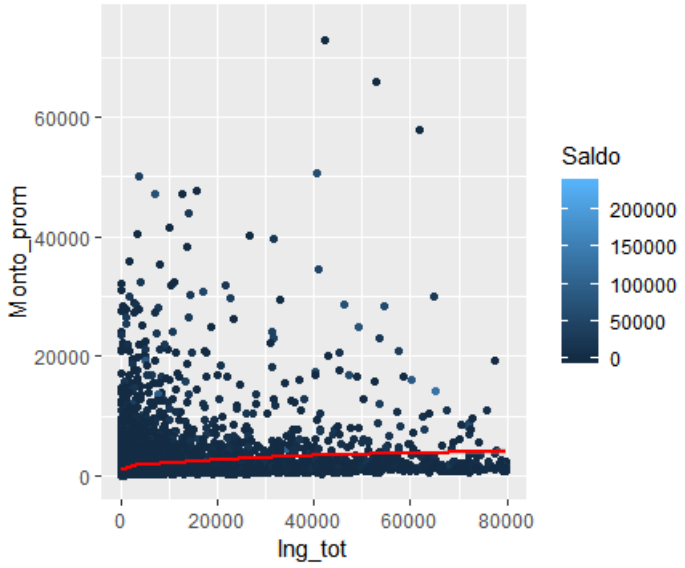


Figura 5.6: Gráfica de dispersión de Monto Promedio vs Ingreso Total

5.3. 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.

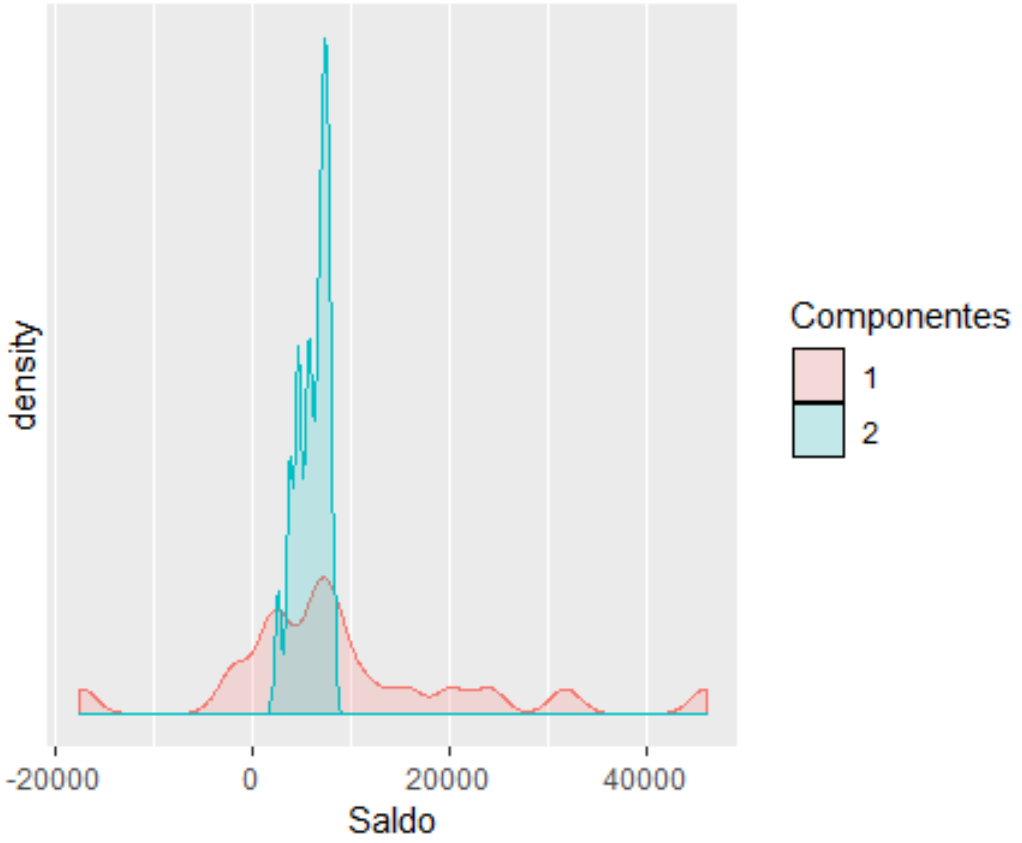


Figura 5.7: Primer resultado bajo 100 observaciones con dos componentes de Saldo

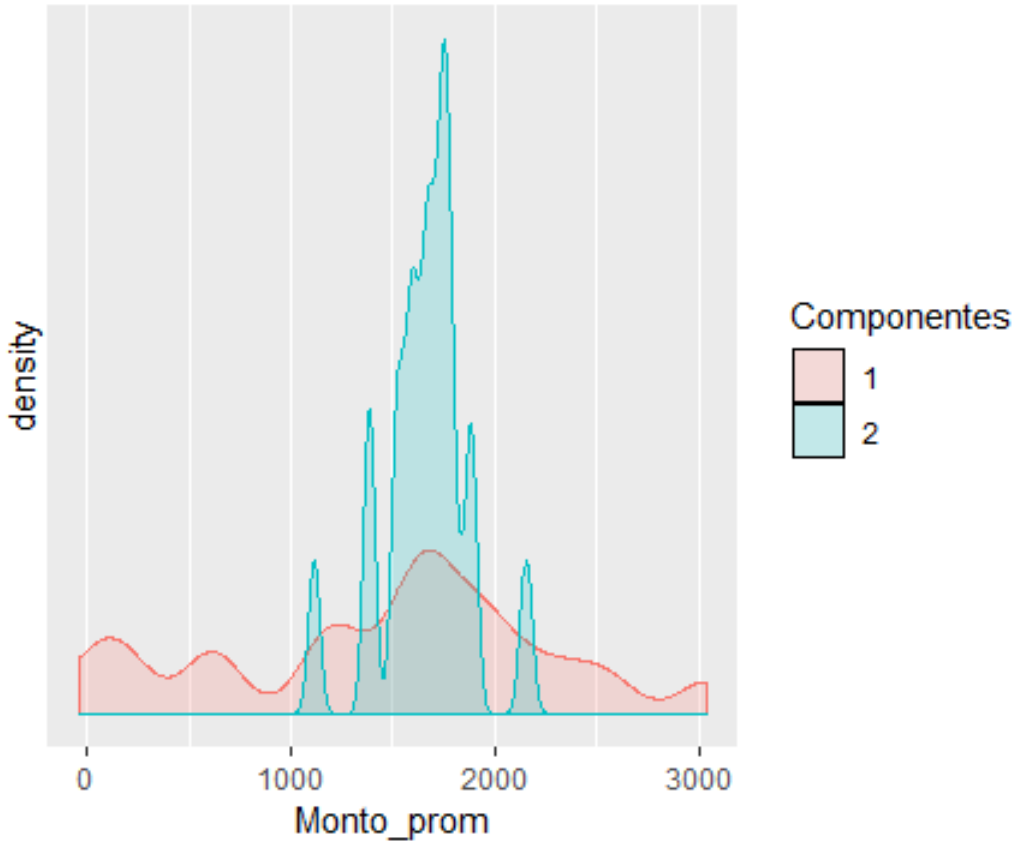


Figura 5.8: Primer resultado bajo 100 observaciones con dos componentes de Monto promedio

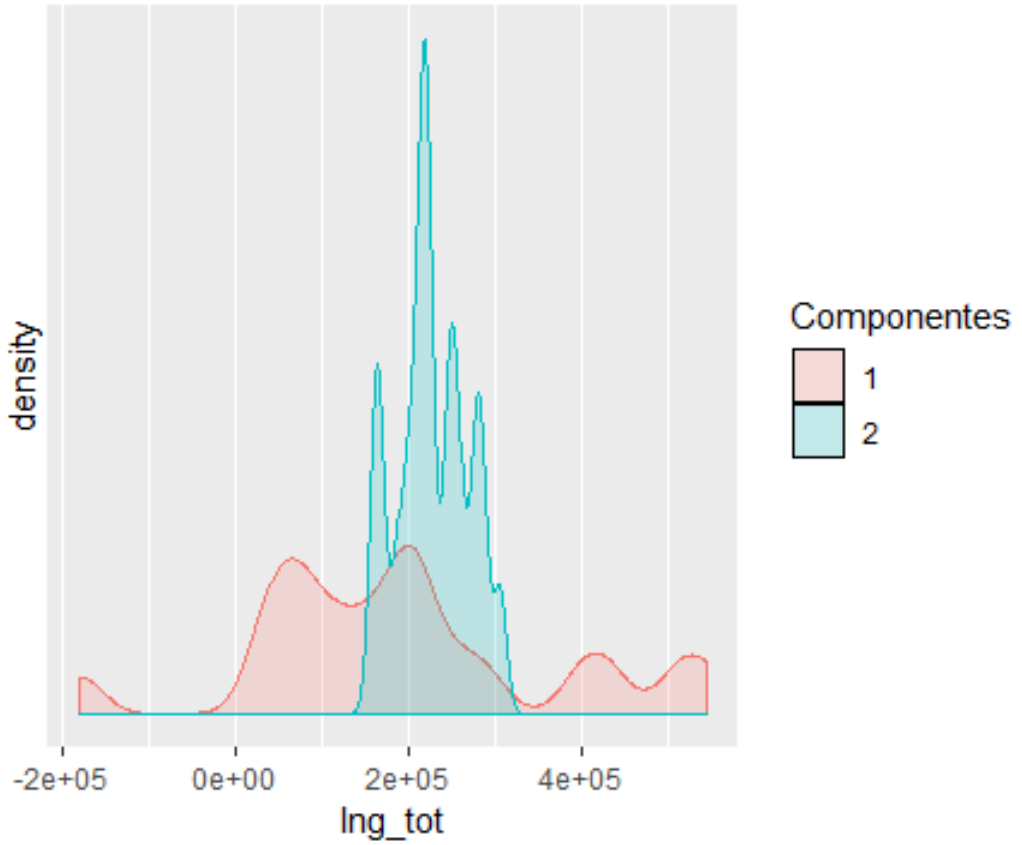


Figura 5.9: Primer resultado bajo 100 observaciones con dos componentes de Ingreso total

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 contemporanea. *Historia de la Matématica en el siglo XX, S. Rios (ed.), Real Academia de Ciencias, Madrid*, 63–80.

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 pattern analysis and machine intelligence*@(3), 489–498.

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

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

Hall, B. (2012). Bayesian inference. *Statisticat, LLC*.

Bibliografía

Liang, L. (2009). On simulation methods for two component normal mixture models under bayesian approach. *Uppsala Universitet, Project ReportUppsala Universitet, Project Report*.

Marin, J.-M., K. Mengersen, and C. P. Robert (2005). Bayesian modelling and inference on mixtures of distributions. *Handbook of statistics 25, 459–507.*

Martínez Ovando, J. C. (2004). Un criterio predictivo de selección de modelos para series de tiempo.

Stephens, M. (2000). Dealing with label switching in mixture models. Journal of the Royal Statistical Society: Series B (Statistical Methodology).