Mezcla Infinita de Gaussianas

Rafael Eduardo Pacheco Góngora

Tesis presentada para obtener el grado de
maestro en ciencias computacionales y matemáticas industriales

Departamento de ciencias computacionales
Centro de investigación en matemáticas (CIMAT A.C.)
Guanajuato, Gto, México
Diciembre 2010

Dedicado a

Mis padres Mariana y Felipe quienes siempre estuvieron conmigo, que con su consejos, apoyo y comprensión me hicieron la persona que ahora soy. Siempre los llevo conmigo.

Mis tios Guadalupe y Luis quienes me abrieron las puertas de su hogar, me brindaron un gran amor , soporte y cariño siempre.

Mis hermanos Pilar, Felipe y Pablo por creer en mi y estar al pendiente en todo lo que realizaba en este viaje.

Mis primos Alex y Pau que siempre me motivaban para seguir adelante.

Mi novia Karina quien a pesar de su ausencia en este lugar, me hacía sentir como si estuviera a mi lado apoyandome en todo momento. Agradezco haber encontrado el amor y compartir mi existencia con ella.

Mis amigos quienes siempre nos dabámos palabras de aliento para no caer en los momentos difíciles. Sé que cuento con ellos siempre.

Mezcla Infinita de Gaussianas

Rafael Eduardo Pacheco Góngora

Presentada para obtener el grado de Maestro en Ciencias
Diciembre 2010

Resumen

Muchos problemas científicos modernos implican resultados complejos, objetos de dimensión infinita como las curvas y ciertas distribuciones. Con frecuencia la metodología estadística no es satisfactoria para la inferencia en este tipo de problemas. La investigación reciente en los métodos no paramétricos bayesianos se centra en la ampliación de los modelos existentes para dar cabida a las inferencias simultaneas para múltiples distribuciones dependientes.

Este trabajo de tesis se centra en el problema de la estimación de la densidad de datos tanto univariados como multivariados utilizando un modelo de mezcla infinita y contable de distribuciones gaussianas. Este modelo de mezcla infinita se desarrolla en el contexto de los metodos bayesianos no paramétricos de aprendizaje no supervisado. Dichos métodos no requieren en general llevar a cabo el proceso de selección de modelo (determinación del número de componentes de la mezcla) lo que los hace atractivos en algunas aplicaciones prácticas, ya que una parte fundamental es evitar el difícil problema de encontrar el número correcto de componentes. La inferencia de este número de componentes se hace de manera implícita usando cadenas de Markov y el muestreo Gibbs . En este trabajo se hace una revisión general de los métodos de estimación de mezcla infinita que utilizan una proceso de Dirichlet como modelo generativo a priori.

Reconocimientos

Al Dr. Salvador Ruiz Correa quien me brindó mucho de su tiempo a pesar de los momentos difíciles que pasó, y me transmitió todas sus enseñanzas, consejos e invaluable amistad e hicimos un gran equipo en el desarrollo de esta tesis.

A mis sinodales Johan Van Horebeek y Arturo Hernández Aguirre quienes no solo fueron mis sinodales sino fueron mis maestros en el cual aprendí mucho de ellos y me pulieron para ser un mejor profesionista.

Al Centro de Investigación en Matemáticas (CIMAT) por abrirme sus puertas y brindarme la oportunidad de continuar mi formación profesional de posgrado.

A toda la comunidad CIMAT, tanto maestros, alumnos y amigos que hicieron una estancia muy agradable en el estudio de la maestría.

Finalmente agradezco el apoyo económico recibido por parte del Consejo Nacional de Ciencia y Tecnología (CONACYT) para la realización de mis estudios de posgrado.

Índice general

Índice de figuras

Índice de cuadros

Capítulo 1

Introducción

Los métodos paramétricos son procedimientos de inferencia estadística, en los que se asume que los datos disponibles son generados a partir de una distribución de probabilidad que pertenecen a una familia de distribuciones específica. La utilidad de los métodos paramétricos se basa en su sencillez, y una serie de supuestos razonables sobre las familias paramétricas implicadas, definiendo distribuciones a priori y/o obteniendo distribuciones a posteriori que son relativamente sencillas, incluso para modelos complicados y muy estructurados. Por otro lado, los métodos no paramétricos evitan suposiciones acerca de la distribución de probabilidad que genera los datos, para establecer métodos que pueden utilizarse en entornos donde los supuestos regulares paramétricos no funcionan. Aunque se aplique de manera más general, los modelos no paramétricos pueden requerir técnicas aún en el caso de modelos simples.

En la estadística bayesiana, los modelos no paramétricos se construyen estableciendo distribuciones a priori a partir de familias de distribuciones de probabilidad. Por lo tanto el término modelo bayesiano no paramétrico realmente es un nombre inapropiado. Los modelos bayesianos no paramétricos contienen un número infinito de parámetros. Raiffa and Schlaifer [1961] y Ferguson [1973] en su trabajo sobre modelos bayesianos no paramétricos, menciona algunas características que debe tenerse en cuenta al construir los priors en espacios de distribuciones que se utilizan en el contexto no paramétrico:

- La clase debe ser analíticamente tratable, por lo tanto la distribución a posteriori debe ser fácil de calcular, ya se analíticamente o por medio de la simulación.

- La clase debe tener un soporte suficientemente grande, es decir el conjunto de datos no debe ser pequeño.
- La definición de los hiperparámetros a priori deben ser fácil de interpretar.

Aunque no siempre es posible satisfacer por completo todos los requisitos, se mencionó anteriormente que en esta tesis se hará incapie en la importancia de estas características cuando se presenten los modelos de mezclas finitas e infinitas, que son el punto central de este trabajo. Estos modelos han sido desarrollados en los últimos años en una serie de trabajos seminales. Una de los mayores ventajas de la metodología bayesiana es que no está sujeta al problema del sobreajuste, por lo tanto la tarea difícil, es hacer que la complejidad del modelo se desvanezca. Los métodos bayesianos no paramétricos ofrecieron sus primeros frutos y se popularizaron en el área de las ciencias de la computación cuando los trabajos de Neal [1996] (enfocados a redes neuronales) condujeron al desarrollo de los procesos gaussianos Williams and Rasmussen [1996]. En esta tesis se presenta la implementación de un modelo de mezclas tanto finita como infinita utilizando hiperparámetros que está basado en una cadena de Markov Monte Carlo. Modelos similares son conocidos en estadística como el proceso de Dirichlet (Ferguson [1973], Ferguson [1974]; Blackwell and MacQueen [1973]; Sethuraman [1994]; Antoniak [1974]).

El objetivo es encontrar los parámetros de la mezcla de gaussianas que para el caso parámetrico, se reduce en optimizar una función logaritmo de verosimilitud como puede verse en el **apéndice B**.

Esta tesis se centra en los modelos de mezclas pertenecientes a las técnicas no paramétricas que modelan la densidad de probabilidad asociada a un conjunto de datos como una superposición lineal de funciones denominados núcleos. A los modelos de mezclas que emplean como funciones núcleo funciones normales o gaussianas, se les conoce habitualmente como *modelos de mezclas de gaussianas.* Los modelos de mezclas de gaussianas que estudiaremos son el modelo de mezcla finita e infinita de gaussianas, donde estudiaremos que a partir del modelo de mezclas finita de gaussianas deduciremos el modelo de mezcla infinita de gaussianas, es por eso que es muy importante desarrollar el caso finito ya que de allí se deriva el modelo como el caso límite es decir al hacer que el número de componentes tienda al infinito, y convierte el modelo de mezcla finita de gaussianas en el modelo de

mezcla infinita de gaussianas, todo esto con el fin de que en el modelo infinito no se conozca el número de componentes y sea encontrado de manera implícita, esa es la gran diferencia que inclusive es la mayor aportación de este modelo. Los métodos bayesianos para las mezclas con un número desconocido (pero finito) de componentes ya han sido explorados por Richardson and Green [1997a] cuyos métodos no se extienden fácilmente para datos multivariados. También en el estado del arte se encuentra el proceso de Dirichlet que es la base para la comprensión de hacer la tendencia al infinito y encontrar el número indicado de componentes, para probar la existencia de este proceso existen varios métodos el cuál detallaremos más adelante, pero en especial estudiaremos el proceso del restaurante chino.

La importancia de estudiar este tipo de modelos finitos es precisamente las ventajas que nos ofrecen, estos son

- No importa la inicialización ya que con la cadena de Markov se va encontrando la distribución estacionaria para muestrear de la verdadera distribución.
- Puede caer en máximos locales pero también puede salir de ellos ya que la cadena se va incrementando hacia un cierto número de iteraciones y hace hace que pueda salir de ellos y dirigirse al máximo global.

Las desventajas para el modelo finito son

- Igual que el EM no estima el orden del modelo
- El número adecuado de iteraciones para la convergencia no es fácil de encontrar ya que puede ser que aún la cadena no haya convergido, se tendría que realizar un estudio para la convergencia de la cadena y eso es costoso computacionalmente.
- Algo menos grave pero importante es que en los modelos no paramétricos se tienen que encontrar ciertos valores a priori adecuados, y esto es mediante ensayo y error.

Las ventajas para el caso infinito son las mismas que el caso finito pero se agrega la estimación del orden del modelo por lo que no es necesario probar con distintos número de componentes ya que el proceso lo calcula de manera implícita. Las principales desventajas son las mismas que el caso finito e inclusive determinar la convergencia para la cadena de Markov se complica un poco más por el proceso del restaurante chino ya que el número de

componentes es muy variable y se toman las frecuencias de el.

El principal objetivo de esta tesis es estudiar bien estos métodos no paramétricos y realizar una aplicación a la segmentación de imágenes cerebrales por lo que esta tesis esta organizada de la siguiente manera: se divide en 8 capítulos donde en la primera parte comprende los capítulos 2, 3 y 4, que contiene el marco teórico importante como es el modelado gráfico, las técnicas de muestreo en especial la cadena de Markov Monte Carlo necesario para la aplicación de encontrar las densidades de un conjunto de datos, así como el muestreo Gibbs, y el paradigma bayesiano que es la esencia en este tipo de problemas de estadística no-paramétrica. La segunda parte comprende los capítulos 5 y 6, que son la parte fundamental ya que comprende las mezclas de gaussianas tanto finitas como infinitas, allí se encuentra toda la parte matemática estadística que demuestra como se hayan las distribuciones y la inferencia de los parámetros a estimar para encontrar las densidades apropiadas. Y la última parte que comprende los capítulos de 7 y 8, en el cuál se encuentran todas las aplicaciones realizadas con el método con experimentos sintéticos y la aplicación a la pre-segmentación de imágenes cerebrales, así como las conclusiones y el trabajo a futuro. También este trabajo cuenta con un manual de usuario del software implementado para esta tesis.

Capítulo 2

Modelos Gráficos

2.1. Introducción

La probabilidad juega un papel importante en el reconocimiento de patrones. La teoría de la probabilidad puede ser expresada en términos de 2 simples ecuaciones que corresponden a la regla de la suma y del producto. Nosotros procederemos a formular y resolver modelos probabilísticos complicados de una manera puramente algebraica. Encontraremos una gran ventaja al realizar análisis usando diagramas que representan distribuciones de probabilidad, estos son llamados modelos gráficos probabilísticos.

Estos modelos nos ofrecen útiles propiedades como:

1. Una manera simple para visualizar la estructura del modelo probabilístico así como para diseñar y motivar nuevos modelos.

2. Analizar las propiedades del modelo, incluyendo las propiedades de independencia condicional que es obtenida con solo inspeccionar el grafo.

3. Permite realizar cálculos complejos, necesarios para realizar la inferencia y el aprendizaje en sofisticados modelos, se puede expresar en términos de manipulaciones gráficas, en el cual las expresiones matemáticas son tomados de manera implícita.

Un grafo contiene nodos (también llamados vértices) conectados por medio de aristas (también llamados ejes). En un modelo gráfico probabilístico cada nodo representa un variable aleatoria (o grupo de variables aleatorias) y las aristas expresan relaciones entre dichas

variables. El grafo por lo tanto captura la manera en la cuál la distribución conjunta sobre todas las variables puede ser descompuesta en un producto de factores en donde cada una depende solamente de un subconjunto de variables.

Empezaremos discutiendo los modelos gráficos dirigidos también conocidos como redes bayesianas, en el cual las aristas de los grafos tiene una direccionalidad particular indicada por flechas. La otra clase de modelos gráficos son conocidos como modelos gráficos dirigidos o campos aleatorios de Markov, en el cual las aristas no contienen flechas y por lo tanto no tienen una dirección significante. Los grafos dirigidos son útiles para expresar relación causal entre variables aleatorias, mientras que los grafos no dirigidos se adaptan mejor a expresiones de restricciones simples entre las variables aleatorias. Para resolver problemas de inferencia es conveniente convertir de grafos dirigidos a grafos no dirigidos a diferentes representaciones llamados gráficas de factores.

Los modelos gráficos resultan una parte fundamental en el desarrollo de esta tesis porque trabaajaremos con probabilidades en donde es necesario conocer con solo inspeccionar el grafo, qué variables son independientes para poder eliminarlo de la probabilidad a calcular, tambien es importante estudiar la cobija de Markov porque cuando se requiera calcular la probabilidad condicional a posteriori de una variable, esta debe depender de las otras variables pertenecientes a la cobija de Markov, esta es la razón por lo cuál es importate estudiar los modelos probabilisticos.

2.2. Redes Bayesianas

Para motivar el uso de grafos dirigidos describiremos distribuciones de probabilidad, considerando primero una distribución conjunta arbitraria $p(a, b, c)$ sobre tres variables a, $b\,y\,c$. Note que no es necesario especificar nada más acerca de estas variables, como por ejemplo si son discretas o continuas. De hecho, uno de los poderosos aspectos de los modelos gráficos es que un gráfico específico puede hacer enunciados probabilísticos para una amplia clase de distribuciones. Aplicando la regla del producto, podemos escribir la distribución conjunta de la siguiente forma

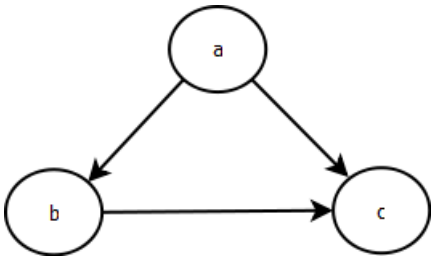


Figura 2.1: Modelo gráfico que representa la descomposición de la probabilidad conjunta sobre tres variables a, b y c

$$p(a, b, c) = p(c|a, b)p(a, b) \tag{2.1}$$

Una segunda aplicación de la regla del producto sería

$$p(a, b, c) = p(c|a, b)p(b|a)p(a) \tag{2.2}$$

Note que esta descomposición es válida para cualquier elección de distribuciones conjuntas. Ahora representemos la ecuación (2.2) en términos de un modelo gráfico de la siguiente manera

Primero introduciremos un nodo para cada variable aleatoria a, b, y c y asociemos cada nodo con la correspondiente distribución condicional, entonces para cada distribución condicional añadimos flechas dirigidas a la gráfica de los nodos correspondientes a las variables de la distribución condicionada, como muestra la figura (2.1).

Ahora extenderemos la representación de la ecuación (2.2) cuando tenemos K variables dadas por $p(x_1, ...x_K)$, repitiendo la aplicación de la regla del producto para la probabilidad de una distribución conjunta obtenemos

$$p(x_1, ..., x_K) = p(x_k|x_1, ..., x_{k-1})...p(x_2|x_1)p(x_1) \tag{2.3}$$

Para una determinada elección de K, podemos volver a representar como un grafo dirigido con K nodos, uno para cada distribución condicional en la parte derecha de (2.3), con cada nodo que tenga como entrada una arista de todos los nodos numerados descendentemente. Decimos que un grafo está totalmente conectado porque hay una relación entre cada par de nodos.

Ahora podemos decir que la distribución conjunta definida por un grafo está dado por el producto sobre todos los nodos del grafo, de una distribución condicional por cada nodo condicionado a las variables correspondientes de los padres de los nodos del grafo. Por lo tanto para un grafo con K nodos la distribución conjunta está dado por

$$p(x) = \prod_{k=1}^{K} p(x_k|pa_k) \tag{2.4}$$

donde pa_k denota el conjunto de padres de x_k y $x = \{x_1, ...x_K\}$.

La ecuación (2.4) expresa la propiedad de factorización de una distribución conjunta para los modelos gráficos dirigidos.

Una consideración importante es que los grafos dirigidos que estamos trabajando no deben tener ciclos, en otras palabras no hay caminos cerrados dentro de la gráfica de manera que podamos pasar de un nodo a otro a lo largo de los enlaces con la dirección de las flechas y terminar en el nodo de inicio. Estos grafos son llamados grafos acíclicos dirigidos o DAGS.

2.3. Independencia Condicional

Un importante concepto para las distribuciones de probabilidad sobre múltiples variables es la independencia condicional(Dawid [1980]). Consideremos tres variables a, b, y c y supongamos que la distribución condicional de a dado b y c es tal que no depende de la variable b esto es

$$p(a|b, c) = p(a|c) \tag{2.5}$$

decimos que a es condicionalmente independiente de b y c. Esto puede ser expresado de una manera diferente si consideramos la distribución conjunta de a y b condicionado a c, el cuál se puede escribir como

$$\begin{aligned} p(a, b|c) &= p(a|b, c)p(b|c) \\ &= p(a|c)p(b|c) \end{aligned} \tag{2.6}$$

Hemos usado la regla del producto de la probabilidad junto con la ecuación (2.5) para

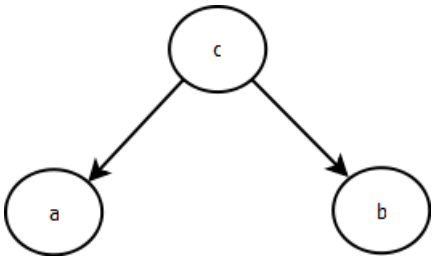


Figura 2.2: Modelo gráfico llamado cola-a-cola sin nodos observados

obtener ecuación (2.6). Entonces vemos que condicionado sobre c, la distribución conjunta de a y b factoriza en el producto de las distribuciones marginales de a y de la distribución marginal de b (ambos sobre c). Esto dice que las variables a y b son estadísticamente independientes dado c. Note que nuestra definición de independencia condicional requiere que (2.5) o el equivalente (2.6) debe ser válido para todos los valores posibles de c y no solo para algunos valores. La notación abreviada para denotar independencia condicional(Dawid [1979]) es como sigue

$$a \perp\!\!\!\perp b|c \tag{2.7}$$

denota que a es condicionalmente independiente de b dado c y es equivalente a (2.5).

Las propiedades de la independencia condicional es desempeñar un papel importante en el uso de modelos probabilístico para el reconocimiento de patrones mediante la simplificación tanto de la estructura de un modelo, así como de los cálculos necesarios para realizar la inferencia y aprendizaje en este modelo.

2.3.1. Distintos Casos de Independencia Condicional

Ejemplificaremos las propiedades de independencia condicional considerando tres simples casos que involucran grafos con tres nodos.

El primer grafo se muestra en la figura (2.2) y la distribución conjunta correspondiente al grafo es fácil de escribir usando la ecuación (2.4) como sigue

$$p(a, b, c) = p(a|c)p(b|c)p(c) \tag{2.8}$$

si ninguna de las variables son observadas, entonces podemos investigar si a y b son independientes marginalizando con respecto a c de la ecuación (2.8) , esto es

$$p(a,b) = \sum_c p(a|c)p(b|c)p(c) \tag{2.9}$$

En general, esto no factoriza en el producto $p(a)p(b)$ y por lo tanto

$$a \not\perp\!\!\!\perp b|\emptyset \tag{2.10}$$

Donde $\emptyset$ denota el conjunto vacio y el símbolo $\not\perp\!\!\!\perp$ significa que no cumple la propiedad de independencia condicional.

Ahora condicionemos la variable c como se muestra en la figura (2.3) y podemos escribirlo facílmente como la distribución condicional de a y b dado c como sigue

$$\begin{aligned} p(a,b|c) &= \frac{p(a,b,c)}{p(c)} \\ &= \frac{p(a|c)p(b|c)p(c)}{p(c)} \\ &= p(a|c)p(b|c) \end{aligned} \tag{2.11}$$

y por lo tanto obtenemos la propiedad de independencia condicional.

$$a \perp\!\!\!\perp b|c \tag{2.12}$$

Podemos realizar una simple interpretación gráfica del resultado, considerando el camino del nodo a al nodo b vía c. Se dice que el nodo c está de la forma cola-a-cola con respecto a ese camino porque el nodo está conectado a las colas de las dos flechas y la presencia de un camino que conecta a los nodos a y b hace que estos nodos sean dependientes. Sin embargo cuando condicionamos sobre el nodo c como vemos en la figura (2.3), el nodo condicionado bloquea el camino de a a b y esto causa que a y b sean condicionalmente independientes.

Podemos similarmente considerar el grafo de la figura (2.4), la distribución conjunta que corresponde a este grafo es nuevamente obtenida de la formula general (2.4) para obtener

$$p(a,b,c) = p(a)p(c|a)p(b|c) \tag{2.13}$$

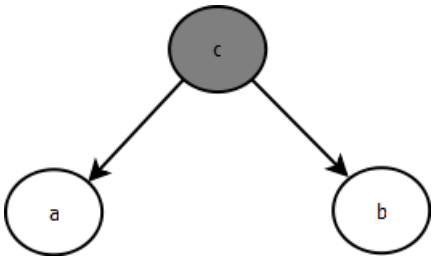


Figura 2.3: Modelo gráfico llamado cola-a-cola que está condicionado sobre la variable c

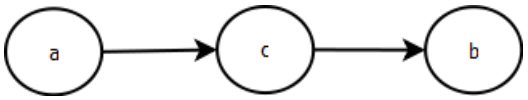


Figura 2.4: Modelo gráfico llamado cabeza-a-cola sin nodos observados

Primero supongamos que ninguna variable es observada, de nuevo podemos probar que si a y b son independientes y marginalizando sobre c obtenemos

$$p(a,b) = p(a)\sum_{c} p(c|a)p(b|c) = p(a)p(b|a) \tag{2.14}$$

el cuál no factoriza en $p(a)p(b)$ y por lo tanto

$$a \not\perp\!\!\!\perp b|\emptyset \tag{2.15}$$

Ahora supongamos que condicionamos u observamos el nodo c como muestra la figura (2.5), usando el teorema de bayes junto con la ecuación (2.13) obtenemos

$$\begin{aligned} p(a,b|c) &= \frac{p(a,b,c)}{p(c)} \\ &= \frac{p(a)p(c|a)p(b|c)}{p(c)} \\ &= \left(\frac{p(a)p(b|c)}{p(c)}\right)\left(\frac{p(a|c)p(c)}{p(a)}\right) \\ &= p(a|c)p(b|c) \end{aligned} \tag{2.16}$$

por lo que obtenemos la propiedad de independencia condicional

$$a \perp\!\!\!\perp b|c \tag{2.17}$$

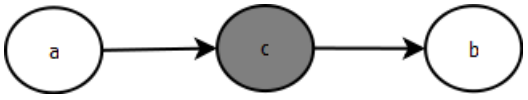


Figura 2.5: Modelo gráfico llamado cabeza-a-cola con el nodo c observado

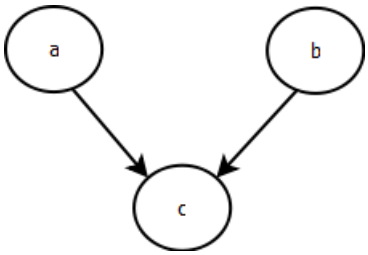


Figura 2.6: Modelo gráfico llamado cabeza-a-cabeza sin nodos observados

por lo que podemos interpretar gráficamente los resultados como sigue: se dice que el nodo c es cabeza-a-cola con respecto al camino que hay entre el nodo a y el nodo b. Este camino conecta los nodos a y b y los hace dependientes, si observamos ahora c como en la figura (2.5) esta observación bloquea el camino de a a b obteniendo independencia condicional sobre c.

Por último veamos el grafo de la figura (2.6), como vemos este tiene un comportamiento mas sutil que los anteriores, la distribución conjunta para este grafo es la siguiente

$$p(a, b, c) = p(a)p(b)p(c|a, b) \tag{2.18}$$

Igualmente que los anteriores consideremos el caso en donde ninguna de las variables son observadas, por lo que marginalizando nuevamente sobre c obtenemos

$$p(a, b) = p(a)p(b) \tag{2.19}$$

y por lo tanto a y b son independientes con ninguna variable observada, en contraste con los dos ejemplos anteriores. Por lo que podemos escribir el resultado como

$$a \perp\!\!\!\perp b|\emptyset \tag{2.20}$$

Ahora condicionemos sobre el nodo c como indica la figura (2.7), la distribución condicional de a y b está dada por

$$\begin{aligned} p(a, b|c) &= \frac{p(a,b,c)}{p(c)} \\ &= \frac{p(a)p(b)p(c|a,b)}{p(c)} \end{aligned} \tag{2.21}$$

el cual en general no factoriza como el producto $p(a)p(b)$ y $a \not\perp\!\!\!\perp b|c$.

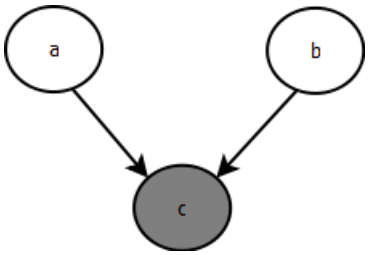


Figura 2.7: Modelo gráfico llamado cabeza-a-cabeza que observa el nodo c

Por lo tanto este caso tiene el comportamiento opuesto a los dos primeros. Gráficamente decimos que c es cabeza-a-cabeza con respecto al camino de a a b porque conecta a las cabezas de las flechas. Cuando el nodo c no es observado se dice que bloquea el camino y las variables a y b son independientes, sin embargo condicionado en el nodo c desbloquea el camino de a a b y los hace dependientes.

2.3.2. Separación D

Demos ahora un estado general a la propiedad d-separación (Pearl [1997]) para grafos dirigidos. Considere un grafo dirigido en general en la que A, B, y C son conjuntos de nodos arbitrarios. Queremos saber si $A \perp\!\!\!\perp B|C$ es condicionalmente independiente, teniendo en cuenta que es una grafo acíclico. Para saberlo consideremos todos los posibles caminos que hay en algún nodo de A a algún nodo de B, cualquiera de esas rutas de acceso se dice que está bloqueado si incluye un nodo de tal manera que sea

- Cuando las flechas del nodo tienen la forma de cabeza-a-cola o cola-a-cola y el nodo está en el conjunto C.
- Cuando las flechas del nodo tienen la forma de cabeza-a-cabeza y ni el nodo ni sus descendientes están en el conjunto C.

Si todos los caminos están bloqueados, entonces A se dice que está d-separado de B por C, y la distribución conjunta en todas las variables en el gráfico satisfacen que $A \perp\!\!\!\perp B|C$.

El concepto d-separación se ilustra en la Figura (2.8). En el primer grafo el camino de a a b no es bloqueado por el nodo d, ya que tiene un nodo cola-a-cola en este camino y no

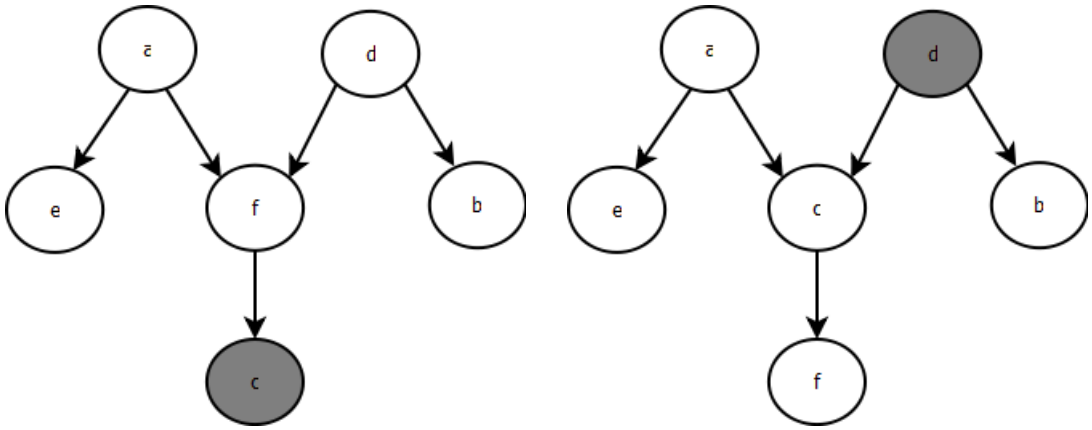


Figura 2.8: Modelos gráficos que ilustran la d-separación

se observa, ni es bloqueado por el nodo f, ya que, aunque este último es un nodo cabeza-a-cabeza, tiene un nodo c descendiente, debido a que está en el conjunto condicionado. Por lo que podemos decir que a es condicionalmente independiente de b dado c. En el segundo grafo el camino de a a b es bloqueado por el nodo d porque se trata de un nodo cola-a-cola que se observa, por lo que la propiedad de independencia condicional $(a \perp\!\!\!\perp b|d)$ serán satisfechas por cualquier distribución que se factoriza de acuerdo a este gráfico. Note que este camino también es bloqueado por el nodo c porque es un nodo cabeza-a-cabeza y ni él ni sus descendientes están en el conjunto condicionado. Una prueba formal puede encontrarlo en (Lauritzen [1996])

Hemos visto que un grafo dirigido representa una descomposición de una distribución de probabilidad conjunta en un producto de probabilidades condicionales. El grafo también expresa un conjunto de declaraciones de independencia condicional obtenidas a través del criterio d-separación y el teorema de la d-separación es en realidad una expresión de equivalencia de estas dos propiedades. Para aclarar este punto se puede pensar en un grafo como un filtro. Supongamos que consideramos una probabilidad conjunta, en particular una distribución $p(x)$ sobre las variables x que corresponde a los nodos no observados del grafo. El filtro permitirá a esta distribución de pasar si y solo si se puede expresar en términos de la autorización (2.4) que implica el grafo. Si se presenta el filtro como el conjunto de todas las posibles distribuciones $p(x)$ en el conjunto de variables x, entonces el subconjunto de las distribuciones que pasan por el filtro lo denominaremos DF, como se ilustra en la figura (2.9), también usamos el grafo como otro tipo de filtro es decir listando todas las propiedades de independencia condicional obteniendo el criterio de d-separación y permi-

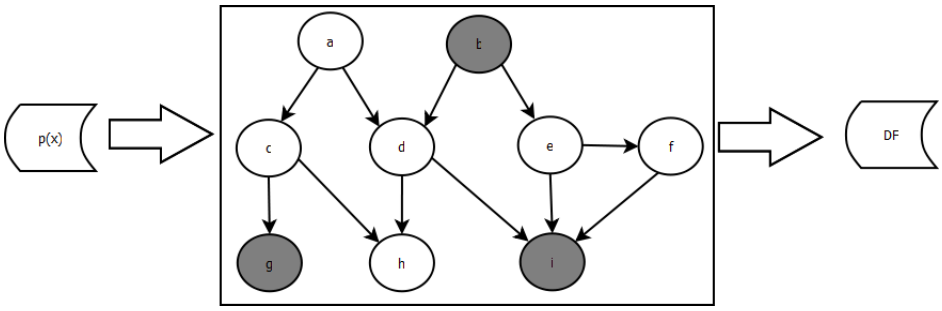


Figura 2.9: En esta figura podemos ver un modelo gráfico como un filtro, en el cual la distribución de probabilidad $p(x)$ pasa a través de un filtro si y solo si satisface la propiedad de factorización (2.4).

tiendo pasar solamente a los que cumplan estos criterios.

Cabe destacar que las propiedades de independencia condicional obtenida de la d-separación se aplican a cualquier modelo probabilístico que describe un grafo dirigido, es decir si las variables son discretas o continuas o una combinación de ellas, el grafo estará describiendo toda una familia de distribuciones de probabilidad.

Para finalizar este capítulo, terminaremos con las propiedades de independencia condicional mediante la exploración del concepto de la cobija de Markov. Consideremos una probabilidad conjunta $p(x_1, ...x_D)$ que representa un grafo dirigido con los nodos D y considere la distribución condicional de un nodo en particular con las variables x_i condicionado a todas las variables restantes $x_{j\neq i}$. Usando la propiedad de la factorización (2.4) se expresa de la siguiente forma

$$\begin{aligned} p(x_i|x_{\{j\neq i\}}) &= \frac{p(x_1,...x_D)}{\int \prod_k p(x_1...x_D)dx_i} \\ &= \frac{\prod_k p(x_k|pa_k)}{\int \prod_k p(x_k|pa_k)dx_i} \end{aligned} \tag{2.22}$$

el cuál la integral es reemplazada por una suma en caso de ser variables discretas.

Ahora considerando que cualquier factor de $p(x_k|pa_k)$ no tienen ningún tipo de dependencia funcional en x_i, se puede tomar fuera de la integral sobre x_i y por lo tanto se

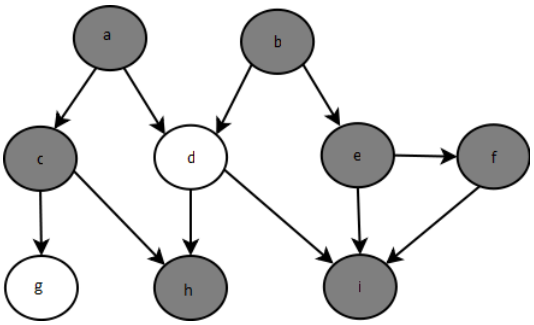


Figura 2.10: Cobija de Markov del nodo x_i que contiene a los padres, los hijos y los co-padres de los hijos

cancela entre el numerador y el denominador. Los únicos elementos que quedan serán las distribuciones condicionales $p(x_i|pa_i)$ para el nodo x_i, así mismo junto con las distribuciones condicionales para algún nodo en x_k tal que el nodo x_k es el conjunto condicionado de $p(x_k|pa_k)$, en otras palabras para el cuál x_i es padre de x_k. La condicional $p(x_k|pa_i)$ dependerá de los padres del nodo x_i mientras que las condicionales $p(x_k|pa_k)$ dependerá de los hijos de x_i y de los co-padres de los hijos. El conjunto de nodos que incluye a los padres, los hijos y los co-padres de los hijos se le llama cobija de Markov y se ilustra en la figura (2.10).

Resumen

En este capítulo estudiamos el modelado gráfico para tener un conocimiento más amplio de como identificar a las variables dependientes o independientes con solo inspeccionar el grafo, también ya sabemos que las variables involucradas en el cálculo de las probabilidades condicionales a posteriori de una determinada variable son los hijos, los padres y los co-padres, todo esto será de mucha ayuda al momento de deducir las ecuaciones de las probabilidades a posteriori involucradas en la mezcla finita e infinita de gaussianas, como veremos más adelante en el capítulo cinco.

Capítulo 3

Métodos de Muestreo

3.1. Introducción

Para la mayoría de modelos probabilísticos de interés práctico, la inferencia exacta es intratable y tenemos que recurrir a alguna forma de aproximación. Consideraremos métodos aproximados de inferencia basado en muestreo numérico, también conocido como técnicas Monte Carlo.

Formalmente una cadena de Markov es un proceso aleatorio discreto con la propiedad de Markov. Un proceso aleatorio discreto, es un sistema que puede tener varios estados y el cuál los cambios son aleatorios con pasos discretos, aunque estrictamente hablando el "paso" puede no tener nada que ver con el tiempo. Los estados de la propiedad de Markov de una distribución de probabilidad para el sistema en el siguiente paso solamente depende del estado actual del sistema y no de los estados restantes del sistema en pasos previos. Porque el sistema cambia aleatoriamente, es generalmente imposible predecir con exactitud el estado del sistema en el futuro. Sin embargo las propiedades estadísticas del sistema en un gran número de medidas en el futuro a menudo se puede describir. En muchas aplicaciones son estas propiedades estadísticas las mas importantes. Los cambios de estado del sistema se denominan transacciones y las probabilidades asociadas a diversos cambios de estado se llaman probabilidades de transición.

Los métodos Markov chain Monte Carlo (MCMC) son una clase de algoritmos para el muestreo de distribuciones de probabilidad basados en construcciones de cadenas de Markov

que tienen la deseada distribución como su distribución de equilibrio. El estado de la cadena después de un gran número de pasos es usado para muestrear la distribución deseada. La calidad de la muestras mejora en función del número de pasos. Por lo general, no es difícil construir una cadena de Markov con las propiedades deseadas. El problema mas dificil es determinar cuántos pasos son necesarios para converger a la distribución estacionaria con un error aceptable.

En esta tesis es importante estudiar algunos algoritmos de muestreo debido a estamos trabajando con métodos estadísticos no paramétricos por lo que la inferencia en los parámetros de la mezcla de gaussianas se realiza mediante cadenas de Markov Monte Carlo para obtener muestras de la verdadera distribución estacionaria, también estudiamos el muestreo Gibss para obtener muestras de las etiquetas para los datos en la mezcla, también en ocasiones no se puede determinar que distribución sigue una determinada variable pero sí se tiene caracteristicas como son la log-concavidad y derivables en todo su dominio, estas caracteristicas son las necesarias para tomar muestras por medio de algoritmo de rechazo adaptable, razón por la cuál estudiaremos a continuación.

3.2. Algoritmos de Muestreo

En esta sección, se consideran algunas estrategias simples para generar muestras aleatorias de una distribución dada.

3.2.1. Muestreo de Rechazo Adaptable (ARS)

El muestreo de rechazo adaptable reduce el número de evaluaciones de una función $g(x)$ de dos maneras:

Primero asumimos la log-concavidad de $f(x)$, evitamos la necesidad de localizar el óptimo de $g(x)$ en D. Segundo, después de cada rechazo, la probabilidad de necesitar evaluar $g(x)$ se reduce aun más actualizando las funciones de compresión y envolvente para incorporar la mas reciente información adquirida sobre $g(x)$. Ahora describamos nuestro método con mas detalle. Asumimos que D está conectado, que $g(x)$ es continua y diferenciable en todo D y que $h(x) = ln\, g(x)$ es cóncava en todo D (por ejemplo $h'(x) = \frac{dh(x)}{dx}$ decrementa

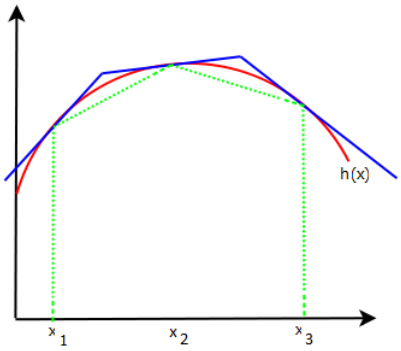


Figura 3.1: Función log-cóncava que muestra las líneas de intersección de las tangentes y de las líneas que se encuentran por debajo de la función, donde $x_{1,}$, x_2 y x_3 son las abscisas.

monótonamente cuando incrementa x en D) . La curva continua de la figura 3.1 ejemplifica una función $h(x)$ cóncava y diferenciable en todo su dominio D.

Supongamos que $h(x)$ y $h'(x)$ han sido evaluados en k abscisas en D: $x_1 \leq x_2 \leq ... \leq x_k$. Sea $T_k = \{x_i : i = 1, ..., k\}$. Definimos el rechazo sobre T_k como $u_k(x)$ donde $u_k(x)$ es una línea por partes formada de tangentes de $h(x)$ en las abscisas de T_k de manera en que la parte superior donde las tangentes se intersectan se tiene que

for $j = 1, ..., k-1$ las tangentes en x_j se intersectan en

$$z_j = \frac{h(x_{j+1}) - h(x_j) - x_{j+1}h'(x_{j+1}) + x_j h'(x_j)}{h'(x_j) - h'(x_{j+1})} \tag{3.1}$$

entonces para $x \in [z_{j-1}, z_j]$ y $j = 1, ..., k$ definimos

$$u_k(x) = h(x_j) + (x - x_j)h'(x_j) \tag{3.2}$$

donde z_0 es el límite inferior en D(o $-\infty$ si D no está acotada por abajo) y z_k si el límite superior en D (o $+\infty$ si D no está acotada por arriba). También definimos

$$s_k(x) = \frac{exp\, u_k(x)}{\int_D exp\, u_k(x')dx'} \tag{3.3}$$

finalmente, definimos la función de compresión sobre T_k como $exp\,(l(x))$, donde $l_k(x)$ es una línea por partes que está formado por debajo de la función en las abscisas de T_k de

manera que

$$l_k(x) = \frac{(x_{j+1} - x)h(x_j) + (x - x_j)h(x_{j+1})}{x_{j+1} - x_j} \tag{3.4}$$

donde $x \in [x_j, x_{j+1}]$, para $j = 1, ..., k-1$, $x < x_1$ o $x > x_k$ definimos $l_k(x) = -\infty$.

Por lo tanto las funciones de rechazo envolvente y de compresión son funciones exponenciales y la concavidad de $h(x)$ asegura que $l_k(x) \leq h(x) \leq u_k(x)$ para toda x en D. Para muestrear n puntos independientes de $f(x)$ por el método de muestreo de rechazo adaptable se tienen los siguientes pasos.

Paso de inicialización

Inicializa la abscisa en T_k. Si D está acotado por la izquierda entonces elegimos x_1 tal que $h'(x_1) > 0$. Si D está acotado por la derecha entonces elegimos x_k tal que $h'(x_k) < 0$. Teniendo definida k abscisas iniciales, calculando las funciones $u_k(x)$, $s_k(x)$ y $l_k(x)$ de las ecuaciones (3.1), (3.3) y (3.4) respectivamente.

Paso de muestreo

Se muestrea un valor de x^* de $s_k(x)$ y también de w independientemente de una distribución uniforme $(0, 1)$. Sea la siguiente prueba de compresión.

si

$$w \leq exp\{l_k(x^*) - u_k(x^*)\} \tag{3.5}$$

entonces aceptamos x^*, de lo contrario evaluamos $h(x^*)$ y $h'(x^*)$ para realizar la siguiente prueba de rechazo.

si

$$w \leq exp\{h(x^*) - u_k(x^*)\} \tag{3.6}$$

entonces acepta x^*, de otra manera rechaza x^*.

Paso de actualización

si $h(x)$ y $h'(x)$ fueron evaluados en el paso de muestreo, incluye x^* en T_k para formar T_{k+1} y se reetiqueta los elementos de T_{k+1} en orden ascendente, luego se construyen las funciones de $u_{k+1}(x)$, $s_{k+1}(x)$ y $l_{k+1}(x)$ de las ecuaciones (3.1), (3.3) y (3.4) respectivamente sobre las bases de T_{k+1}, posteriormente se incrementa k y se regresa al paso de muestreo, así hasta terminar con los n puntos deseados. La prueba de este método se encuentra en (Gilks and Wild [1992])

3.2.2. Cadena de Markov Monte Carlo

En la sección anterior, discutimos de la importancia de las estrategias del muestreo de rechazo para evaluar las expectativas de las funciones y notamos que sufren de severas limitaciones sobre todo en espacios de alta dimensionalidad.

En esta sección discutiremos sobre la cadena de Markov Monte Carlo que permite tomar muestras de una clase de distribuciones y realiza una buena escala con la dimensionalidad del espacio muestral. Las cadenas de Markov con los métodos Monte Carlo tiene sus orígenes en la física y solo hacia el final de la década de los ochenta comenzó a tener un gran impacto en el ámbito de la estadística. Al igual que el método de muestreo de rechazo, este método toma muestras de una distribución propuesta, solo que esta vez se mantiene un registro del estado actual z^τ y la distribución propuesta $q(z|z^\tau)$ depende del estado actual por lo que la secuencia de muestras $z^{(1)}, z^{(2)}, ...$ forma una cadena de Markov. Si se escribe $p(z) = \frac{\widetilde{p}(z)}{Z_p}$, asumiríamos que $p(z)$ puede ser fácilmente evaluado para cualquier valor dado de Z a pesar de que el valor de Z_p puede ser desconocido. La distribución propuesta en sí es elegido para ser suficientemente simple para tomar muestras de él directamente, en cada ciclo del algoritmo, se genera una muestra z^* como candidato que se toma de la distribución propuesta y entonces se acepta la muestra de acuerdo a un criterio apropiado. Ahora nos preguntamos bajo qué circunstancias la cadena de Markov converge a la distribución deseada. Como primera instancia una cadena de Markov se define como una serie de variables aleatorias $z^{(1)}, ..., z^{(M)}$ de tal manera que la siguiente propiedad de independencia condicional es válida para $m \in \{1, ..., M-1\}$

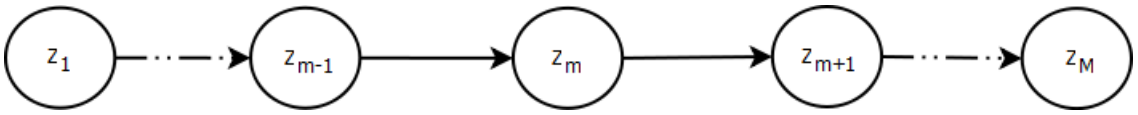


Figura 3.2: Modelo gráfico en forma de cadena

$$p(z^{(m+1)}|z^{(1)},...,z^{(m)}) = p(z^{(m+1)}|z^{m}) \tag{3.7}$$

Esto por supuesto puede ser representado como un grafo en forma de cadena, como en la figura 3.2. Ahora podemos especificar una cadena de Markov teniendo la distribución de probabilidad para la variable inicial $p(z^{(0)})$ junto con las probabilidades condicionales de las variables subsecuentes en forma de una probabilidad de transición esto es $T_m(z^{(m)}, z^{(m+1)}) \equiv p(z^{(m+1)}|z^{m})$. Una cadena de Markov es llamada homogénea si la transición de las probabilidades son la misma para todo m.

La probabilidad marginal para una variable en particular puede ser expresada en términos de la probabilidad marginal para los variables previas en la cadena de la forma

$$p(z^{(m+1)}) = \sum_{z^{(m)}} p(z^{(m+1)}|z^{(m)})p(z^{(m)}) \tag{3.8}$$

Una distribución se dice que es invariante o estacionaria con respecto a una cadena de Markov, si en cada paso de la cadena la distribución es invariante. Entonces para una cadena de Markov homogénea con probabilidades de transición $T(z', z)$, la distribución $p^*(z)$ es invariante si

$$p^*(z) = \sum_{z'} T(z', z)p^*(z') \tag{3.9}$$

Una cadena de Markov puede tener más de una distribución invariante, por ejemplo si las probabilidades de la transición están dadas por una transformación de la identidad, entonces alguna distribución será invariante.

Una sufiente pero no necesaria condición para asegurar que la distribución requerida $p(z)$ es invariante, es elegir una probabilidad de transición para satisfacer la propiedad de "balance detallado" definido por

$$p^*(z)T(z,z') = p^*(z')T(z',z) \tag{3.10}$$

para una particular distribución $p^*(z)$. Es fácil ver que una probabilidad de transición satisface el "balance detallado" con respecto a la distribución particular que saldrá de una distribución invariante, porque

$$\sum_{z'} p^*(z')T(z',z) = \sum_{z'} p^*(z)T(z',z) = p^*(z)\sum_{z'} p(z'|z) = p^*(z) \tag{3.11}$$

Una cadena de Markov que cumple con la propiedad de "balance detallado" se dice también que es reversible.

El objetivo de usar cadenas de Markov es muestrear de una distribución dada, ahora también se puede lograr si creamos una cadena de Markov tal que la distribución elegida sea invariante. Sin embargo también debemos exigir que para $m \to \infty$, la distribución $p(z^m)$ converge a la distribución invariante requerida $p*(z)$, independientemente de la elección de la distribución inicial $p(z^{(0)})$. Esta propiedad se llama "ergodicidad" y la distribución invariante es llamado distribución de "equilibrio". Claramente una cadena de Markov ergódica solo puede tener una distribución de equilibrio. Se puede demostrar que una cadena de Markov homogénea será ergódica, sujeto a las restricciones débiles de la distribución invariante y de las probabilidades de transición (Neal [1993]).

En la práctica con frecuencia construimos probabilidades de transición de un conjunto de transiciones $B_1, ..., B_k$. Esto puede ser obtenido a través de una mezcla de distribuciones de la forma

$$T(z',z) = \sum_{k=1}^{K} \alpha_k B_k(z',z) \tag{3.12}$$

para algún conjunto de mezcla de coeficientes $\alpha_1, ...\alpha_K$ satisface que $\alpha_k \geq 0$ y $\sum_k \alpha_k = 1$, ahora las transiciones base pueden ser combinados a través de aplicaciones sucesivas tal que

$$T(z',z) = \sum_{z_1} ... \sum_{z_{n-1}} B_1(z',z_1)...B_{K-1}(z_{K-2},z_{K-1})B_K(z_{K-1},z) \tag{3.13}$$

si una distribución es invariante con respecto a cada una de las transiciones base, en-

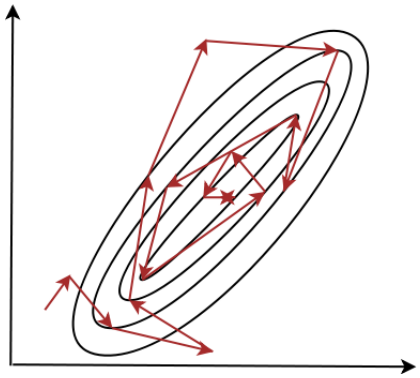


Figura 3.3: Muestra del algoritmo de Gibbs en dos dimensiones, comenzando en un punto inicial y completando en n iteraciones

tonces obviamente también será invariante con respecto a $T(z', z)$ dado por la ecuación (3.12) o (3.13). Para el caso de la mezcla (3.12) si cada una de las transiciones base satisfacen el "balance detallado", entonces la mezcla de transición T también satisface el "balance detallado". Uno de los algoritmos más conocidos para realizar el muestreo en una cadena de Markov usando técnicas de Monte Carlo es el conocido como gibbs sampling.

3.2.2.1. Muestreo Gibbs

El muestreo Gibbs (Geman and Geman [1984]) es un simple algoritmo altamente aplicable al MCMC y puede ser visto como un caso especial del algoritmo llamado "Metrópolis Hasting" (Hastings [1970]).

Considere la distribución $p(z) = p(z_1, ...z_M)$ del cuál queremos la muestra, y supongamos que hemos elegido un estado inicial de la cadena de Markov. Cada paso del procedimiento del muestreo Gibbs consiste en reemplazar el valor de una de las variables por un valor establecido de la distribución de esa variable condicionado a los valores restantes. Así sustituimos z_i por un valor establecido de la distribución $p(z_i|z_{-i})$, donde z_i denota el i -ésimo componente de z y z_{-i} denota $z_1, ..., z_M$ pero con z_i omitido. Este procedimiento se repite en ciclos a través de las variables en un orden en particular o eligiendo la variable que se actualiza en cada paso de una distribución aleatoria.

Por ejemplo, supongamos que tenemos una distribución $p(z_1, z_2, z_3)$ de tres variables y sea τ el paso del algoritmo, seleccionamos valores para $z_1^{(\tau)}$, $z_2^{(\tau)}$ y $z_3^{(\tau)}$. Primero reemplazamos $z_1^{(\tau)}$ por el nuevo valor $z_1^{(\tau+1)}$ obteniendo el muestreo de una distribución condicional como

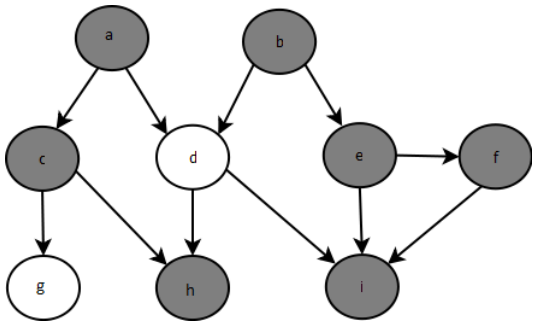


Figura 3.4: El método del muestreo Gibbs requiere de muestrear para las distribuciones condicionales de las variables condicionales restantes, en los modelos gráficos esta distribución es una función solamente de los estados de la cobija de Markov.

sigue

$$p(z_1^{(\tau+1)}|z_2^{(\tau)}, z_3^{(\tau)}) \tag{3.14}$$

luego reemplazamos $z_2^{(\tau)}$ por el valor $z_2^{(\tau+1)}$ obteniéndose de igual forma que el anterior como sigue

$$p(z_2^{(\tau+1)}|z_1^{(\tau+1)}, z_3^{(t)}) \tag{3.15}$$

y posteriormente actualizamos el nuevo valor para z_3 esto es

$$p(z_3^{(\tau+1)}|z_1^{(\tau+1)}, z_2^{(t+1)}) \tag{3.16}$$

y así sucesivamente se realizan los ciclos a través de las tres variables, como se puede ver en el algoritmo (3.1).

Para demostrar todo esto, primero notemos que la distribución $p(z)$ es invariante a cada uno de los pasos del muestreo Gibbs y por lo tanto en toda la cadena de Markov. Esto se deriva del hecho de que cuando se muestrea a partir de $p(z_i|z_{-i})$, la distribución marginal $p(z_{-i})$ es claramente invariante porque el valor de z_{-i} no se modifica. También en cada paso se define las muestras de la correcta distribución condicional $p(z_i|z_{-i})$. Porque estas distribuciones tanto condicionales como marginales especifican la distribución conjunta y por lo tanto vemos que esta distribución es invariante en sí misma.

El segundo requisito que debe cumplir para que el procedimiento del muestreo Gibbs

Algoritmo 3.1 Muestreo Gibbs

1.- Inicializa $\{z_i : i = 1, ..., M\}$
2.- For $\tau = 1, ..., T$
-Muestrea $z_1^{(\tau+1)} \sim p(z_1|z_2^{(\tau)}, z_3^{(\tau)}, ..., z_M^{(\tau)})$.
-Muestrea $z_2^{(\tau+1)} \sim p(z_2|z_1^{(\tau+1)}, z_3^{(\tau)}, ..., z_M^{(\tau)})$.
$\vdots$
-Muestrea $z_j^{(\tau+1)} \sim p(z_j|z_1^{(\tau+1)}, z_{j-1}^{(\tau+1)}, z_j^{(\tau)} ..., z_M^{(\tau)})$.
$\vdots$
-Muestrea $z_M^{(\tau+1)} \sim p(z_M|z_1^{(\tau+1)}, z_2^{(\tau+1)}, ..., z_{M-1}^{(\tau+1)})$.

sea correcto es que sea ergódica. Una condición suficiente de ergodicidad es que ninguna de las distribuciones condicionales sea cero. Si este es el caso entonces cualquier punto en el espacio z se puede llegar desde cualquier otro punto en un número finito que implican una actualización de cada una de las variables, si este no cumple este requisito, por lo que algunas de las distribuciones condicionales tienen ceros, si se aplica la ergodicidad debe ser probado en forma explicita. La distribución del estado inicial también se debe especificar para inicializar el algoritmo, a pesar de que las muestras tomadas después de muchas iteraciones serán independientes de esta distribución. Por supuesto, las muestras sucesivas de la cadena de Markov serán muy correlacionadas, por lo que para obtener muestras que sean casi independientes sera necesario submuestrear la secuencia.

La aplicación práctica del Muestreo Gibbs depende de la facilidad con la que las muestras pueden ser extraídas de la distribución condicional $p(z_k|z_{-k})$. En el caso de distribuciones de probabilidad especificadas usando modelos gráficos, las distribuciones condicionales para los nodos individuales dependen solamente de las variables que corresponden a la cobija de Markov, como se ilustra en la figura 3.4. Para grafos dirigidos existe una amplia selección de distribuciones condicionales de los nodos individuales condicionado a sus padres que se puede muestrear por Gibbs.

Resumen

Todo lo anterior visto en este capítulo es la base para encontrar los parámetros de la mezcla tanto finita como infinita de gaussianas ya que son estos algoritmos de muestreo los que nos proporcionan la inferencia en dichos parámetros, el muestreo de rechazo adaptable es un algoritmo de muestreo que encaja muy bien para tomar muestras que tienen funciones

son log-concavas y además son derivables en todo su dominio, por eso se optó por muestrear con ese algoritmo, como veremos más adelante en el capítulo cinco.

Capítulo 4

El Paradigma Bayesiano

4.1. Introducción

La estadística matemática utiliza dos grandes paradigmas, estas son la frecuentista y la bayesiana. La estadística que estamos acostumbrados a utilizar es la estadística frecuentista, que es la que se desarrolla a partir de los conceptos de probabilidad y que se centra en el cálculo de probabilidades y los contrastes de hipótesis. De alguna forma, la estadística frecuentista tiene como objetivo determinar una conclusión, sea en base al significado estadístico o aceptación y rechazo de hipótesis, siempre dentro del marco del estudio que se esté realizando. En el análisis estadístico que pretende comparar la eficacia de un nuevo tratamiento frente a otro conocido, se utiliza únicamente la información obtenida en el ensayo. No existen subjetividades referentes a parámetros, puesto que se han fijado los criterios de decisión a priori y estos permanecen estáticos durante todo el estudio.

Como enfoque alternativo a la estadística frecuentista, aparece cada vez más en escena la estadística bayesiana, basada como su nombre lo indica en el teorema de Bayes, y que se diferencia de la estadística frecuentista básicamente en la incorporación de información externa al estudio que se esté realizando, de manera que, tal como se ha explicado en la formulación del teorema de Bayes, si conocemos la probabilidad de que ocurra un suceso, su valor será modificado cuando dispongamos de esa información. Así pues, las fuentes de información "a priori" se ven trasformadas en probabilidad "a posteriori" y se utilizan posteriormente para realizar la inferencia.

Los métodos bayesianos ofrecen un paradigma completo, tanto para la inferencia estadística como para la toma de decisiones en condiciones de incertidumbre. Los métodos bayesianos pueden derivarse de un sistema axiomático, y por lo tanto proporcionar una metodología general y coherente. Los métodos bayesianos contienen como casos particulares muchos de los procedimientos frecuentistas que utilizan con mayor frecuencia y resuelve muchas de las dificultades que enfrentan los convencionales métodos estadísticos. En particular los métodos bayesianos hacen posible la incorporación del hipótesis científica en el análisis y se puede aplicar a los problemas cuya estructura es demasiado compleja para los métodos convencionales que no son capaz de manejar. El paradigma bayesiano se basa en una interpretación de la probabilidad como una medida condicional de incertidumbre que se acerca al sentido de la palabra probabilidad en el lenguaje ordinario. La inferencia estadística sobre una cantidad de interés se describe como la modificación de la incertidumbre acerca de su valor en función de las pruebas, y precisamente el teorema de bayes especifica como debe hacerse.

El enfoque bayesiano es la parte medular para la comprensión del calculo de probabilidades a posteriori de los distintos parámetros de la mezcla de gaussianas, es decir todas la probabilidades condicionales, la deducción de las ecuaciones o la manera de como se distribuyen ciertas variables siguen el teorema de bayes para calcular las probabilidades a posteriori y poder realizar la inferencia, también es muy importante conocer la manera en que se selecciona el modelo ya que en el caso de nuestras mezclas finita de gaussianas será necesario elegir de un conjunto de valores de componentes, cuál es el adecuado, por lo que es importante conocer varias técnicas de selección de modelo.

4.2. Modelo Bayesiano

El análisis estadístico de algunos datos observados D típicamente comienza con una evaluación informal descriptiva, el cual es usado como sugerencia tentativa, el modelo de probabilidad formal $\{p(D|w),\, w \in \Omega\}$ asume representar para algún (desconocido) valor de w, el mecanismo de probabilidad cual ha generado los datos observados D. Esto establece una necesidad lógica de distribuciones de probabilidad a priori sobre los parámetros del espacio Ω, describiendo el conocimiento disponible de K sobre el valor del prior de w de los

datos que están siendo observados. Esto sigue de la teoría de la probabilidad que menciona que sí el modelo de probabilidad es correcto toda la información disponible sobre el valor de w (después de que los datos D han sido observados) está contenido en la correspondiente distribución a posteriori cuya densidad de probabilidad, $p(w|D,K)$ es obtenido del teorema de bayes

$$p(w|D,A,K) = \frac{p(D|w,A,K)p(w|A,K)}{\int_{\Omega} p(D|w,A,K)p(w|A,K)dw} \tag{4.1}$$

donde A son las supuestos hechos en el modelo de probabilidad.

Otra forma de escribir esta probabilidad es como sigue:

Sea $p(D|\theta)$, donde $\theta \in \{\theta_1, ..., \theta_m\}$ que toman un número finito de valores, y D son los datos observados entonces usando el teorema de bayes obtenemos

$$p(\theta_i|D) = \frac{p(D|\theta_i)p(\theta_i)}{\sum_{j=1}^{m} p(D|\theta_j)p(\theta_j)} \tag{4.2}$$

para alguna distribución a priori de $p(\theta) = \{p(\theta_1), ..., p(\theta_m))\}$.

4.3. Proceso de Aprendizaje o Inferencia Bayesiana

En el paradigma bayesiano el proceso de aprendizaje de los datos es sistemáticamente implementado por el uso del teorema de bayes para combinar la información previa disponible con la información proporcionada por los datos para producir la distribución a posteriori. El cálculo de las densidades a posteriori suele verse facilitado por la siguiente expresión

$$p(w|D) \propto p(D|w)p(w) \tag{4.3}$$

La información proporcionada por los datos $p(D|w)$ es llamada función de verosimilitud (o likelihood), $p(w|D)$ es la distribución a posteriori y $p(w)$ es la información a priori.

4.3.1. Función de Verosimilitud

Las inferencias por verosimilitud se basan solamente en los datos s y el modelo $\{P_\theta : \theta \in \Omega\}$, es decir, un conjunto de posibles medidas de probabilidad para el sistema que es objeto de estudio. Con estos elementos obtenemos la entidad fundamental de la inferencia basada en la verosimilitud, denominada la función de verosimilitud.

Para fundamentar la definición de la función de verosimilitud suponga que tenemos un modelo estadístico en el que cada P_θ es discreta y viene dada por una función de probabilidad f_θ. Una vez se ha observado s, considere la función $L(\cdot|s)$ definida en el espacio de parámetros Ω y con valores de R^1, dada por $L(\theta|s) = f_\theta(s)$. Nos referiremos a la función $L(\cdot|s)$ determinada por el modelo y los datos como la función de verosimilitud, y al valor de $L(\theta|s)$ lo denominaremos verosimilitud de θ. Observe que para la función de verosimilitud fijamos los datos y variamos el valor del parámetro. Vemos que $f_\theta(s)$ no es más que la probabilidad de obtener los datos s cuando el valor verdadero del parámetro es θ. Esto implica establecer un orden de preferencia o confianza en Ω, es decir, creemos que θ_1 es el valor verdadero de θ, por encima de θ_2, siempre que $f_{\theta_1}(s) > f_{\theta_2}(s)$. Ello se debe a que la desigualdad indica que los datos son más probables si se considera θ_1 que si se considera θ_2. Por el contrario, si $f_{\theta_1} = f_{\theta_2}$ permanecemos indiferentes respecto a θ_1 y θ_2. La inferencia de θ por verosimilitud está basada en esta ordenación. Es importante recordar la interpretación correcta de $L(\theta|s)$. El valor de $L(\theta|s)$ es la probabilidad de s dado que θ es el valor verdadero, y no la probabilidad de θ dado que hemos observado s. También puede ocurrir que el valor de $L(\theta|s)$ sea muy pequeño para cada valor de θ. Sin embargo, no es el valor concreto de la verosimilitud el que nos informa de la confianza para cada valor de θ, sino el valor relativo de la verosimilitud de los diferentes valores posibles del parámetro.

EJEMPLO: Suponga que lanzamos una moneda n=10 veces y que obtenemos $s = 4$ caras. Sin ningún conocimiento sobre la probabilidad de obtener cara en un lanzamiento, el modelo estadístico adecuado para la situación es el modelo binomial(10, θ) con $\theta \in \Omega = [0, 1]$. La función de verosimilitud viene dada por

$$L(\theta|4) = \begin{pmatrix} 10 \\ 4 \end{pmatrix} \theta^4(1-\theta)^6 \tag{4.4}$$

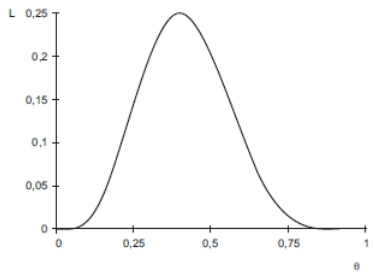


Figura 4.1: Función de verosimilitud del modelo binomial(10,θ) cuando $s = 4$.

representada en la figura (4.1).

4.3.1.1. Función de Verosimilitud para Distribuciones Continuas

En los casos anteriores, los eventos considerados tenían una probabilidad p estrictamente mayor que cero. Pero cuando la noción de verosimilitud se extiende a variables aleatorias con una función de densidad f sobre, por ejemplo, el eje real, la probabilidad de un evento cualquiera es nula. Por ejemplo, supóngase el caso de tener una variable aleatoria real de distribución desconocida X de la que se extrae una muestra $x_1, ..., x_n$ de observaciones independientes. Supóngase también que se dispone de una familia parametrizada de funciones de densidad $f_\theta(x)$ (es decir, que existe una función de densidad $f_\theta(x)$ para cada valor del parámetro $\theta(x)$). En este caso, $\theta(x)$ juega el papel de parámetro desconocido y es razonable definir la función de verosimilitud $L(\theta)$ de la siguiente manera

$$L(\theta) = L(\theta|x_1, ..., x_n) = \prod_i f_\theta(x_i) \tag{4.5}$$

Discusiones similares pueden aplicarse a los casos en que la variable aleatoria X tenga una distribución que sea un híbrido entre una variable continua y discreta.

4.3.2. Cómo Cuantificar la Información a Priori

En la cuantificación de la distribución de probabilidad a priori radica el principal punto de controversia de los métodos bayesianos ya que implica una, al menos aparente, pérdida de objetividad. Sin embargo está claro que, sobre todo en la toma de decisiones, los juicios sobre una técnica terapéutica, un nuevo fármaco, la posibilidad de aparición de efectos adversos etc.., nunca se fundamentan únicamente en los resultados de un solo estudio concreto. Hay que tener presente que el término a priori no implica necesariamente una relación

temporal en el sentido de que corresponda a una información obtenida con anterioridad a nuestro estudio, sino que se refiere, en un sentido más amplio, a la información externa a nuestro estudio.

Existen diferentes procedimientos para formalizar la distribución de probabilidad a priori y algunos autores recomiendan no limitarse a un sólo método para cuantificarla, sino utilizar varios de ellos con el fin de evaluar cómo se modifican la conclusiones en cada caso. Es lo que se conoce como análisis de sensibilidad.

Desde el punto de vista probabilístico o matemático en cuestión existen tres métodos fundamentales para establecer la distribución de probabilidad a priori:

- Distribución no informativa o de referencia, que corresponde a una ausencia de opinión o de conocimiento clínico a priori y por lo tanto no aporta información a lo que se observa en los datos.
- Distribución a priori escéptica, que considera que la probabilidad de que la hipótesis alternativa sea cierta (existe diferencia entre los grupos) es muy pequeña.
- Y distribución a priori entusiasta, que tiene razones fundadas para encontrar diferencias, por lo que determina que la probabilidad de que éstas sean 0 o peor en el grupo de interés tiene una probabilidad muy baja.

El problema radica en que la especificación y cuantificación de la distribución a priori no es una tarea sencilla, especialmente cuando se trata de modelos con más de un parámetro, como pueden ser los modelos de regresión. Por otro lado existe un cierta reticencia por parte de los investigadores a incorporar una distribución a priori con suficiente información, por temor a la posibilidad de que se les acuse de subjetividad.

4.4. Selección de Modelo

La selección de modelo(Burnham et al. [2002]) consiste en el problema de distinguir qué modelo es adecuado y se ajuste bien a los datos con diferentes parámetros. En la literatura existen dos grupos distintos, los basados en la teoría de la información como es el criterio de información Akaike (AIC)(Akaike [2003]) y los de la inferencia bayesiana como

son las pruebas bayesianas y el criterio de información bayesiano (BIC)(Liddle [2008]) , existe también un modelo que combina las dos anteriores este es llamado criterio de información de desviación (DIC)(Liddle [2008]) y es calculado del muestreo a posteriori de una MCMC(Gilks and Richardson [1995]). En general, un modelo es una elección de los parámetros (que varían) y una distribución de probabilidad a priori de los parámetros. El objetivo de la selección de modelo es equilibrar la calidad de ajuste de los datos observados contra la complejidad o el diagnostico de lograr que el modelo se ajuste. Esto se logra a través de estadísticas en general el mejor modelo que se ajuste se utiliza para la inferencia.

Las técnicas de selección de modelo puede considerarse como estimadores de alguna cantidad física. El sesgo y la varianza son medidas importantes de la calidad de un estimador, así como la eficiencia asintótica. La complejidad se mide generalmente contando el número de parámetros en el modelo. Un ejemplo común de selección de modelo es el ajuste de una curva, donde dado un conjunto de puntos y los conocimientos básicos de otro tipo (por ejemplo los puntos son el resultado de las muestras independientes, idénticamente distribuidas), debemos seleccionar un función que describe la mejor curva.

4.4.1. Criterios de Selección de Modelo

4.4.1.1. Criterio de Información Akaike (AIC)

El AIC (Akaike [2003]) está definido como

$$AIC = -2ln\left(\mathcal{L}_{max}\right) + 2k \tag{4.6}$$

donde $\mathcal{L}_{max}$ es la máxima verosimilitud obtenida del modelo y k el número de parámetros del modelo, el mejor modelo es aquel que tiene el valor mínimo de AIC.

El AIC es derivado de una aproximación a la minimización de la información de la entropia Kullback-Leibler que mide la diferencia entre la verdadera distribución de los datos y la distribución del modelo. El AIC no es una prueba del modelo en el sentido de las pruebas de hipótesis, sino que es una prueba entre los modelos (una herramienta para la selección del modelo). Dado un conjunto de datos, varios modelos de la competencia puede ser clasificados de acuerdo con su AIC, la que presentó el menor valor de AIC será el mejor.

A partir del valor de AIC se puede inferir que por ejemplo, los tres primeros modelos están en un empate y el resto son peores, pero sería arbitrario para asignar un valor por encima del cual un determinado modelo es "rechazado".

4.4.1.2. Modelo de Comparación Bayesiano

Factor Bayes

Es una alternativa para la clásica prueba de hipótesis, este método se basa principalmente en el teorema de bayes, es decir dado un problema de selección de modelo en el que tenemos que elegir entre dos modelos se prosigue a parametrizar el modelo y encontrar la verosimilitud marginal y con eso realizar un cociente. Si el cociente es mayor a uno entonces el modelo del numerador es el adecuado de lo contrario es el del denominador.

Sea M_1y M_2 dos diferentes modelos y sean θ_1 y θ_2 los vectores de los parámetros del modelo entonces el factor de bayes está dado por

$$K = \frac{p(D|M_1)}{p(D|M_2)} = \frac{\int p(\theta_1|M_1)p(D|\theta_1, M_1)d\theta_1}{\int p(\theta_2|M_2)p(D|\theta_2, M_2)d\theta_2} \tag{4.7}$$

donde D son los datos y $p(D|M_i)$ es llamado la verosimilitud marginal del modelo i.

Por lo tanto la comparación del modelo bayesiano no depende de los parámetros usados por cada modelo sino mas bien considera la probabilidad del modelo considerando todos los posibles valores de los parámetros.

Criterio de Información Bayesiano (BIC)

El BIC fue introducido por (Schwarz [1978]) y está definido como

$$BIC = -2ln\left(L_{max}\right) + kln(N) \tag{4.8}$$

donde N es el número de datos usados para el ajuste, $\mathcal{L}_{max}$ es la máxima verosimilitud obtenida del modelo y k el número de de parámetros del modelo.

Proviene de la aproximación de los coeficientes de pruebas de modelos, conocido como el factor bayes. El BIC asume que los datos son independientes e idénticamente distribuidos que puede o no, ser valido en función del conjunto de datos que se examinan. Cuando estimamos los parámetros del modelo usamos la estimación de la máxima verosimilitud, es posible incrementar la verosimilitud agregando parámetros al modelo, el cuál puede resultar un sobreajuste. El BIC resuelve este problema introduciendo un término de penalidad para el número de parámetros en el modelo. El modelo con menor valor de BIC es considerado el adecuado.

Criterio de Información de Desviación (DIC)

Es un modelo jerárquico que generaliza el BIC y el AIC y es útil en problemas donde distribuciones a posteriori del modelo han sido obtenidas por una cadena de Markov Monte Carlo(MCMC), esto es debido a que por cada cadena o iteración se toman las esperanzas de los parámetros para obtener la verosimilitud y luego se toman el conjunto de todos para obtener de igual manera la verosimilitud y con eso se realizan medidas. Es similar a realizar el diagnostico para la convergencia de la cadena de Markov(Gilks and Richardson [1995]). El valor con menor DIC es el adecuado.

4.4.1.3. Descripción de Longitud Mínima(MDL)

El principio de longitud de descripción mínima (minimum descripción length (MDL)) (Informatica) puede ser resumido como "elegir la explicación más corta a los datos observados". Incorporando conceptos básicos de teoría de la información tenemos

$$H = arg\,max\,\{p(D|h)p(h)\} \tag{4.9}$$

donde D son los datos y h son los parámetros, de forma equivalente, expresando esta ecuación en términos de la maximización de log

$$H = arg\,max\,\{log_2 p(D|h) + log_2 p(h)\} \tag{4.10}$$

o, alternativamente, minimizando el negativo de esta cantidad

$$H = arg\,max\,\{-log_2 p(D|h) - log_2 p(h)\} \tag{4.11}$$

Esta última ecuación puede ser interpretada como que se prefieren hipótesis cortas. Cada uno de estos términos se puede entender como la longitud de descripción de las distribuciones bajo una codificación óptima. No vamos a comentar los términos de teoría de información debido a que no es el objetivo de la tesis. El principio MDL recomienda la elección de las hipótesis que minimizan estas dos longitudes de descripción. Así, este principio se puede definir como elegir la hipótesis H_{MDL} dada

$$H_{MDL} = arg\,max\,\{L_{c_1} p(D|h) + L_{c_2} p(h)\} \tag{4.12}$$

siendo L_{Ci} la longitud de descripción del mensaje i con respecto a C, que es el número de bits requeridos para codificar el mensaje i utilizando el código C.

4.4.1.4. Validación Cruzada

Se refiere a una técnica para evaluar cómo los resultados de un análisis estadístico se generaliza a un conjunto de datos independientes. Lo que se hace es formar un subgrupo y llamarlo conjunto de entrenamiento y validarlo con otro subgrupo llamado conjunto de pruebas, y esto se realiza utilizando particiones diferentes y los resultados deben ser semejantes. El principal objetivo es la predicción.

4.4.1.5. Boostrap

Es un método de remuestreo y se utiliza principalmente para aproximar la distribución en el muestreo para un estadístico. Se usa frecuentemente para aproximar el sesgo o la varianza de un estadístico, así como construir intervalos de confianza o realizar pruebas de hipótesis. Un problema es que es costoso computacionalmente. Para mas información puede consultar (Efron [1979]).

Resumen

El cálculo de la probabilidades a posteriori se realiza obteniendo la verosimilitud y la probabilidad a priori, en este capítulo comprendimos mejor qué es la función verosimilitud y como cuantificar los datos a priori que son estos las partes importantes para realizar la probabilidad a posteriori utilizando el teorema de bayes, la aplicación de lo mencionado previamente se verá mas adelante cuando se realizen las deducciones de los parámetros a posteriori para realizar la inferencia de los parámetros mediante la cadena de Markov en el capitulo 5, también se revisaron y se comprendieron los distintos modelos de seleccion y se encontró que los más importantes para nuestro modelo son el BIC, AIC y el MDL, debido a que para ellos es muy importante la verosimilitud y para nosotros es fácil de obtener, ademas que evitan el sobreajuste, por lo que fueron los que se implementaron en el software realizado **(Ver apéndice A)**.

Capítulo 5

Modelo de Mezcla Finita de Gaussianas

5.1. Introducción

La mezcla finita de gaussianas puede ser escrita como una superposición lineal de gaussianas de la forma

$$p(x) = \sum_{k=1}^{K} \pi_k N(x|\mu_k, \Sigma_k) \tag{5.1}$$

donde π_k es llamado la proporción o peso de la mezcla y debe satisface que

$$0 \leq \pi_k \leq 1 \tag{5.2}$$

$$\sum_{k=1}^{K} \pi_k = 1 \tag{5.3}$$

Ahora para una representación gráfica de un modelo de mezclas de gaussianas, en el cual la distribución conjunta está expresada de la forma $p(x, z) = p(z)p(x|z)$ como se puede ver en la figura 1, la variable z se puede expresar de la siguiente manera

$$p(z) = \prod_{k=1}^{K} \pi_k^{z_k} \tag{5.4}$$

el cuál es una representación 1 de k , es decir solamente se enciende cuando $z_k = 1$. Por lo tanto para la distribución condicional de x dado el valor de z tenemos que

$$p(x|z_k = 1) = N(x|\mu_k, \Sigma_k) \tag{5.5}$$

Figura 5.1: Modelo gráfico donde la distribución conjunta es $p(x,z) = p(z)p(x|z)$

que puede ser escrito de la forma

$$p(x|z) = \prod_{k=1}^{K} N(x|\mu_k, \Sigma_k)^{z_k} \tag{5.6}$$

la distribución conjunta está dada por $p(z)p(x|z)$ y la distribución marginal de x, es obtenida sumando la distribución conjunta sobre todos los posibles estados de z es decir

$$p(x) = \sum_{z} p(z)p(x|z) = \sum_{k=1}^{K} \pi_k N(x|\mu_k, \Sigma_k) \tag{5.7}$$

El modelo de mezcla finita gaussiana con K componentes del cuál trabajaremos se escribe como

$$p(y|\mu_k, s_k, \pi_k) = \sum_{j=1}^{K} \pi_j N(\mu_j, s_j^{-1}) \tag{5.8}$$

donde:

$y = \{y_1, y_2, ...y_n\}$

$\mu_j = j - ésima\, media$

$s_j = precisión\, (inversa\, de\, las\, varianzas)$

$\pi_j = j - ésima\, proporción\, de\, la\, mezcla\, (Debe\, ser\, positiva\, y\, sumar\, a\, uno)$

La precisión de un estimador depende de su varianza; si la varianza es grande el valor del estimador puede tener grandes variaciones y por lo tanto en ocasiones su valor puede alejarse mucho del parámetro que se desea estimar; por el contrario, si la varianza de un estimador es pequeña la probabilidad de obtener valores del estimador alejados de valor del parámetro que se desea estimar es pequeña, es por eso que utilizamos la precisión.

Tradicionalmente, las técnicas para la estimación de la densidad de probabilidad asociada a un conjunto de observaciones se han clasificado en técnicas paramétricas y no paramétricas. Cada una de ellas tiene sus propias ventajas e inconvenientes que describiremos a continuación.

- Las técnicas paramétricas asumen que la densidad de probabilidad asociada a los datos sigue una forma específica, que puede diferir bastante de la densidad real. Sin embargo este planteamiento permite que la función densidad sea evaluada rápidamente cada vez que surge una nueva observación.

- Los métodos no-paramétricos, por el contrario, permiten que la forma de la función densidad sea mucho más general, es decir no se puede asumir que los datos se ajusten a una distribución conocida, pero sufre el inconveniente de que el número de parámetros del modelo crece proporcionalmente al número de observaciones.

El modelo de mezcla de gaussianas son ampliamente utilizados en minería de datos, reconocimiento de patrones, máquinas de aprendizaje y análisis estadístico. En muchas aplicaciones, sus parámetros son determinados por la máxima verosimilitud usando el algoritmo de expectation-maximitation. Sin embargo existen ciertas limitaciones en este método por lo que utilizaremos un enfoque bayesiano en el que se emplea MCMC para estimar los parámetros de la mezcla.

El objetivo principal de la tesis es encontrar precisamente esos parámetros que hacen que la función se ajuste muy bien a los datos, es decir se pretende encontrar la densidad de los datos, pero la manera en que se realizará sera por medio de técnicas bayesianas con la cadena de Markov Monte Carlo útil para encontrar los parámetros deseados solo que en nuestro modelo agregaremos hiperparámetros para estimar los parámetros de la mezcla, también cabe mencionar que el modelo es finito dado que es necesario proporcionar el valor del número de componentes, a continuación detallaremos la representación de las mezclas finitas con hiperparámetros y detallaremos todas las ecuaciones implicadas para realizar la inferencia, incluyendo el muestreo Gibbs, necesario para obtener las etiquetas de los dato.

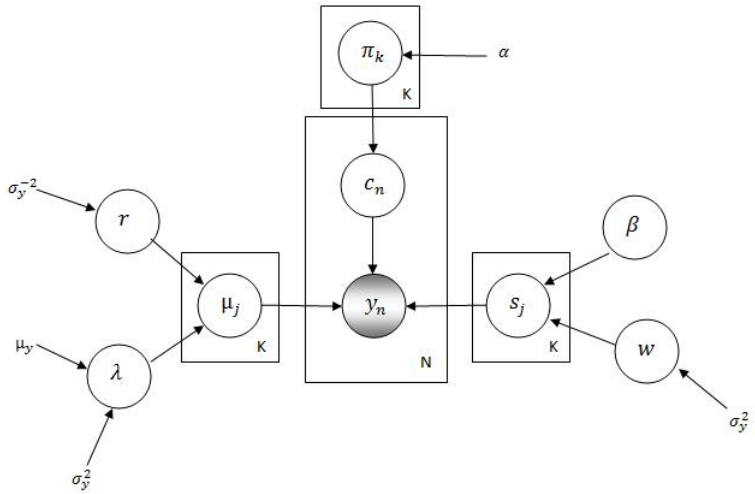


Figura 5.2: Modelo Gráfico propuesto para las mezclas finitas con hiperparámetros

5.2. Modelado de Mezcla Finita con Hiperparámetros

Las ecuaciones siguientes fueron propuestas por (Rasmussen [2000]) y el modelo gráfico propuesto es el que aparece en la figura (5.2). Donde $\mu_j, s_j \, y \, \pi_k$ son los parámetros de la mezcla de gaussianas, $r, \lambda, \beta, w \, y \, \alpha$ son los hiperparámetros, y son los datos de la población y la variable c son indicadores estocásticos uno por cada observación, su principal labor es de tener las clases que han generado la observación, estos indicadores toman valores de $1...K$ y $\mu_y, \sigma_y^{-2} \, y \, \sigma_y^2$ son los datos a priori de la población.

Los componentes de las medias μ_j están dadas por la gaussiana a priori:

$$p(\mu_j|\lambda, r) \sim N(\lambda, r^{-1}) \tag{5.9}$$

donde:

La media λ y la precisión r son hiperparámetros comunes a todos los componentes.

Los parámetros por sí mismos están distribuidos de forma normal y gamma a priori como:

$$p(\lambda) \sim N(\mu_y, \sigma_y^2) \tag{5.10}$$

$$p(r) \sim g(1, \sigma_y^{-2}) \propto r^{-1/2} exp\left(\frac{-r\sigma_y^2}{2}\right) \tag{5.11}$$

donde:

μ_y y σ_y^2 son la media y la varianza de las observaciones.

Para calcular la probabilidad condicional $p(\mu_j|c, y, s_j, \lambda, r)$ a posteriori utilizamos el teorema de bayes como sigue

$$p(\mu_j|c, y, s_j, \lambda, r) \propto p(y|\mu_j, c, s_j, \lambda, r)p(\mu_j|s_j, c, \lambda, r)$$

pero por la independencia en el modelo gráfico obtenemos

$$p(\mu_j|c, y, s_j, \lambda, r) \propto p(y|\mu_j, c, s_j)p(\mu_j|\lambda, r)$$

Ahora como podemos ver en la figura (5.1) haciendo $z = c$ obtenemos una ecuación similar a la ecuación (5.6) como sigue

$$p(y|\mu_j, c, s_j) = \prod_{j=1}^{K} N(y|\mu_j, s_j)^{c_{nj}} \tag{5.12}$$

pero como solo vamos a considerar cuando la variable c_{nj} está activa (su valor es uno y todos los demás son ceros) entonces el producto desaparece, es decir queda de la siguiente manera

$$p(y|\mu_j, c, s_j) = N(y|\mu_j s_j) \tag{5.13}$$

Ahora encontremos la probabilidad a posteriori, para ello multiplicaremos la verosimilitud de la ecuación (5.13) por la ecuación (5.9) como sigue

$$p(\mu_j|c, y, s_j, \lambda, r) \propto p(y|\mu_j, c, s_j)p(\mu_j|\lambda, r)$$

Aqui es muy importante mencionar un término que es de vital importancia para conocer las distribuciones a posteriori, cuando se tiene la distribución de la verosimilitud y la distribución del dato a priori, es posible obtener una distribución conjugada a priori que

detallaremos en la siguiente capítulo y se define de la siguiente manera

Definicion 1.- Dada una clase de funciones de verosimilitud $p(x|\theta)$, una clase de distribuciones de probabilidad a priori $p(\theta)$ se dice que será conjugado de la clase de funciones de verosimilitud $p(x|\theta)$ si el resultado de la distribución a posteriori $p(\theta|x)$ son de la misma familia que $p(\theta)$.

Por lo que el conjugado a priori de una gaussiana es otra gaussiana por lo que busquemos la media y varianza como sigue

$$p(\mu_j|c,y,s_j,\lambda,r) \propto \prod_{n=1}^{N}\left(\sqrt{\tfrac{s_j}{2\pi}}exp\left(-\tfrac{1}{2}\left[(y_n-\mu_j)^2 s_j\right]\right)\right)\left(\sqrt{\tfrac{r}{2\pi}}exp\left(-\tfrac{1}{2}\left[(\mu_j-\lambda)^2 r\right]\right)\right)$$

$$p(\mu_j|c,y,s_j,\lambda,r) \propto \left(\sqrt{\tfrac{s_j}{2\pi}}\right)^n exp\left(-\tfrac{1}{2}\left[s_j\sum_{n=1}^{N}(y_n-\mu_j)^2\right]\right)\left(\sqrt{\tfrac{r}{2\pi}}exp\left(-\tfrac{1}{2}\left[(\mu_j-\lambda)^2 r\right]\right)\right)$$

$$p(\mu_j|c,y,s_j,\lambda,r) \propto \left(\sqrt{\tfrac{s_j}{2\pi}}\right)^n \sqrt{\tfrac{r}{2\pi}}exp\left(-\tfrac{1}{2}\left[(\mu_j-\lambda)^2 r + s_j\sum_{n=1}^{N}(y_n-\mu_j)^2\right]\right)$$

Ahora apliquemos el logaritmo a esta expresión

$$log\, p(\mu_j|c,y,s_j,\lambda,r) \propto$$

$$\tfrac{n}{2}\left[log\, s_j - log\, 2\pi\right] + \tfrac{1}{2}\left[log\, r - log\, 2\pi\right] - \tfrac{1}{2}\left[(\mu_j-\lambda)^2 r + s_j\sum_{n=1}^{N}(y_n-\mu_j)^2\right]$$

Podemos desarrollar la expresión del exponente y acompletar los cuadrados para encontrar la media y la varianza a posteriori pero es complicado, existe una manera mas fácil de encontrarlo de acuerdo a (Thrun et al. [2005]) que dice que la media del producto de dos gaussianas es el primer momento y la inversa de la varianza es el segundo, por lo que derivemos para encontrar el primer momento con respecto a μ_j, y luego despejamos con respecto a la misma variable para encontrar la media, esto es

$$-\tfrac{1}{2}\left(2r(\mu_j-\lambda) + 2s_j\sum_{n=1}^{N}(y_n-\mu_j)(-1)\right) = 0$$

$$r\mu_j - r\lambda - s_j\sum_{n=1}^{N}y_n + s_j\sum_{n=1}^{N}\mu_j = 0$$

$$\mu_j \left(r + n_j s_j\right) = r\lambda + s_j \sum_{n=1}^{N} y_n$$

Ahora como solo estamos suponiendo para un cierto conjunto de valores de c entonces $\sum_{n=1}^{N} y_n$ se transforma en $\sum_{i:c_i=j} y_i$ por lo tanto

$$\mu_j = \frac{r\lambda + s_j \sum_{i:c_i=j} y_i}{r + n_j s_j}$$

Ahora para encontrar la varianza debemos encontrar el segundo momento o segunda derivada que de acuerdo con (Thrun et al. [2005]), que dice que al encontrar el segundo momento lo que resulte al derivar por segunda vez es la varianza inversa, esto es

$$r + s_j n_j = 0$$
$$varianza = \frac{1}{r + s_j n_j}$$

Por lo tanto

$$p(\mu_j | c, y, s_j, \lambda, r) \sim N \left(\frac{r\lambda + s_j \sum_{i:c_i=j} y_i}{r + n_j s_j}, \frac{1}{r + s_j n_j} \right) \tag{5.14}$$

donde:

$n_j = j$-ésimo número de observación que pertenece a la clase j.

Ahora hacemos lo mismo para encontrar como se distribuye $p(\lambda|\mu_k, r)$, de igual manera el conjugado a priori de una gaussiana es otra gaussiana y lo encontramos como sigue

$$p(\lambda|\mu_k, r) \propto p(\mu_k|\lambda, r)p(\lambda|r)$$
$$p(\lambda|\mu_k, r) \propto p(\mu_k|\lambda, r)p(\lambda)$$

sustituyendo la verosimilitud de (5.9) y la probabilidad de (5.10) tenemos que

$$p(\lambda|\mu_k, r) \propto \prod_{j=1}^{K} \left(\sqrt{\tfrac{r}{2\pi}} exp\left(-\tfrac{1}{2}\left[(\mu_j - \lambda)^2 r\right]\right) \right) \left(\sqrt{\tfrac{1}{2\pi\sigma_y}} exp\left(-\tfrac{1}{2}\left[\tfrac{(\lambda-\mu_y)^2}{\sigma_y^2}\right]\right) \right)$$

$$p(\lambda|\mu_k, r) \propto \left(\sqrt{\frac{r}{2\pi}}\right)^K exp\left(-\frac{1}{2}\left[r\sum_{j=1}^{K}(\mu_j - \lambda)^2\right]\right)\left(\sqrt{\frac{1}{2\pi\sigma_y}}exp\left(-\frac{1}{2}\left[\frac{(\lambda-\mu_y)^2}{\sigma_y^2}\right]\right)\right)$$

$$p(\lambda|\mu_k, r) \propto \left(\sqrt{\frac{r}{2\pi}}\right)^K \sqrt{\frac{1}{2\pi\sigma_y}}exp\left(-\frac{1}{2}\left[r\sum_{j=1}^{K}(\mu_j - \lambda)^2 + \frac{(\lambda-\mu_y)^2}{\sigma_y^2}\right]\right)$$

Aplicando el logaritmo obtenemos

$$log\, p(\lambda|\mu_k, r) \propto$$
$$\frac{K}{2}\left(log\, r - log\, 2\pi\right) + \frac{1}{2}(log\, 1 - log\, 2\pi\sigma_y) - \frac{1}{2}\left[r\sum_{j=1}^{K}(\mu_j - \lambda)^2 + \frac{(\lambda-\mu_y)^2}{\sigma_y^2}\right]$$

Aplicamos el primer momento con respecto a λ e igualamos a cero para encontrar la media

$$2r\sum_{j=1}^{K}(\mu_j - \lambda)(-1) + \frac{2(\lambda-\mu_y)}{\sigma_y^2} = 0$$
$$-\sigma_y^2 r\sum_{j=1}^{K}(\mu_j - \lambda) + \lambda - \mu_y = 0$$
$$-\sigma_y^2 r\sum_{j=1}^{K}\mu_j + \sigma_y^2 r\lambda K + \lambda - \mu_y = 0$$
$$\lambda\left(\sigma_y^2 rK + 1\right) = \mu_y + \sigma_y^2 r\sum_{j=1}^{K}\mu_j$$
$$\lambda = \frac{\mu_y + \sigma_y^2 r\sum_{j=1}^{K}\mu_j}{\sigma_y^2 rK + 1}$$

Encontramos el segundo momento (segunda derivada) para encontrar la varianza

$$\sigma_y^2 rK + 1 = var^{-1}$$
$$varianza = \frac{1}{\sigma_y^2 rK + 1}$$

Por lo tanto tenemos que

$$p(\lambda|\mu_k, r) \sim N\left(\frac{\mu_y + \sigma_y^2 r\sum_{j=1}^{K}\mu_j}{\sigma_y^2 rK + 1}, \frac{1}{\sigma_y^2 rK + 1}\right) \tag{5.15}$$

De manera similar encontremos como se distribuye $p(r|\mu_K, \lambda)$

$$p(r|\mu_k, \lambda) \propto p(\mu_k|\lambda, r)p(r|\lambda)$$
$$p(r|\mu_k, \lambda) \propto p(\mu_k|\lambda, r)p(r)$$

sustituyendo la verosimilitud de (5.9) y la probabilidad de (5.11) , aqui podemos ver que tenemos el producto de una normal por una gamma, pero al desarrollar el producto tenemos que

$$p(r|\mu_k, \lambda) \propto \prod_{j=1}^{K}\left(\sqrt{\tfrac{r}{2\pi}}exp\left(-\tfrac{1}{2}\left[(\mu_j-\lambda)^2 r\right]\right)\right)\left(r^{-1/2}exp\left(-\tfrac{1}{2}\left[r\sigma_y^2\right]\right)\right)$$
$$p(r|\mu_k, \lambda) \propto \left(\sqrt{\tfrac{r}{2\pi}}\right)^K exp\left(-\tfrac{1}{2}\left[r\sum_{j=1}^{K}(\mu_j-\lambda)^2\right]\right)\left(r^{-1/2}exp\left(-\tfrac{1}{2}\left[r\sigma_y^2\right]\right)\right)$$
$$p(r|\mu_k, \lambda) \propto (r)^{K/2}\, r^{-1/2}exp\left(-\tfrac{1}{2}\left[r\sum_{j=1}^{K}(\mu_j-\lambda)^2 + r\sigma_y^2\right]\right)$$
$$p(r|\mu_k, \lambda) \propto (r)^{\frac{K-1}{2}}\, exp\left(-\tfrac{1}{2}r\left[\sum_{j=1}^{K}(\mu_j-\lambda)^2 + \sigma_y^2\right]\right)$$

Ahora podemos ver que al parecer este producto es el conjugado a priori de una gamma, ajustemos las variables para obtener los parámetros de una distribución gamma.

$$p(r|\mu_k, \lambda) \propto (r)^{\frac{K+1}{2}-1}\, exp\left(-\tfrac{1}{2}r\left[\sum_{j=1}^{K}(\mu_j-\lambda)^2 + \sigma_y^2\right]\right)$$
$$p(r|\mu_k, \lambda) \propto (r)^{\frac{K+1}{2}-1}\, exp\left(-\tfrac{1}{2}\left[\frac{(K+1)r}{\frac{(K+1)}{\sum_{j=1}^{K}(\mu_j-\lambda)^2+\sigma_y^2}}\right]\right)$$

por lo tanto tenemos que

$$p(r|\mu_k, \lambda) \sim g\left(K+1, \frac{K+1}{\sum_{j=1}^{K}(\mu_j - \lambda)^2 + \sigma_y^2}\right) \tag{5.16}$$

Las componentes de las precisiones s_j estan distribuidas a priori como

$$p(s_j|\beta, w) \sim g(\beta, w^{-1}) \propto (s_j)^{\frac{\beta}{2}-1} exp\left(-\frac{1}{2}s_j w\beta\right) \tag{5.17}$$

donde:

β y w^{-1} son hiperparámetros que tienen a su vez distribuciones a priori como sigue

$$p(\beta^{-1}) \sim g(1,1) \Longrightarrow p(\beta) \propto \beta^{-3/2} exp\left(-\frac{1}{2\beta}\right) \tag{5.18}$$

$$p(w) \sim g(1, \sigma_y^2) \propto w^{-1/2} exp\left(\frac{-w}{2\sigma_y^2}\right) \tag{5.19}$$

Ahora encontremos como se distribuye $p(s_j|c, y, \mu_j, \beta, w)$, esto es

$p(s_j|c, y, \mu_j, \beta, w) \propto p(y|s_j c, \mu_j, \beta, w)p(s_j|c, \mu_j, \beta, w)$

$p(s_j|c, y, \mu_j, \beta, w) \propto p(y|\mu_j, s_j, c)p(s_j|\beta, w)$ ...por independencia condicional

sustituyendo la verosimilitud de (5.8) y la probabilidad de (5.17) tenemos

$$p(s_j|c, y, \mu_j, \beta, w) \propto$$
$$\left(\prod_{n=1}^{N}\sqrt{\frac{s_j}{2\pi}} exp\left(-\frac{1}{2}\left[(y_n - \mu_j)^2 s_j\right]\right)\right)\left((s_j)^{\frac{\beta}{2}-1} exp\left(-\tfrac{1}{2}s_j w\beta\right)\right)$$
$$p(s_j|c, y, \mu_j, \beta, w) \propto$$
$$\left(\sqrt{\tfrac{s_j}{2\pi}}\right)^{n_j} exp\left(-\tfrac{1}{2}\left[s_j\sum_{n=1}^{N}(y_n - \mu_j)^2\right]\right)\left((s_j)^{\frac{\beta}{2}-1} exp\left(-\tfrac{1}{2}s_j w\beta\right)\right)$$
$$p(s_j|c, y, \mu_j, \beta, w) \propto (s_j)^{n_j/2}(s_j)^{\frac{\beta}{2}-1} exp\left(-\tfrac{1}{2}\left[s_j\sum_{n=1}^{N}(y_n - \mu_j)^2 + s_j w\beta\right]\right)$$

Al igual que el anterior tenemos el producto de una distribución normal y una distribu-

ción gamma por lo que veamos si es posible encontrar la distribución a posteriori como sigue

$$p(s_j|c,y,\mu_j,\beta,w) \propto (s_j)^{\frac{n_j+\beta}{2}-1} exp\left(-\tfrac{1}{2}s_j\left[\sum_{n=1}^{N}(y_n-\mu_j)^2+w\beta\right]\right)$$

$$p(s_j|c,y,\mu_j,\beta,w) \propto (s_j)^{n_j/2}\,(s_j)^{\frac{\beta}{2}-1} exp\left(-\frac{1}{2}\left[\frac{s_j(n_j+\beta)}{\sum_{n=1}^{N}(y_n-\mu_j)^2+w\beta}\right]\right)$$

de igual manera que lo anterior como solo estamos tomando una clase de c entonces $\sum_{n=1}^{N}(y_n-\mu_j)^2 = \sum_{i:c_i=j}(y_i-\mu_j)^2$, por lo que vemos que sí encontramos la distribucion a posteriori el cual es una distribución gamma y queda de la siguiente manera

$$p(s_j|c,y,\mu_j,\beta,w) \sim g\left(n_j+\beta, \frac{n_j+\beta}{\sum_{i:c_i=j}(y_i-\mu_j)^2+w\beta}\right) \tag{5.20}$$

Ahora encontremos $p(w|s_k,\beta)$ como sigue

$$p(w|s_k,\beta) \propto p(s_k|\beta,w)p(\beta|w)$$
$$p(w|s_k,\beta) \propto p(s_k|\beta,w)p(w)$$

Sustituimos la verosimilitud de (5.17) y la probabilidad de (5.19), y ademas como nuestra variable es con respecto a w entonces

$$p(s_k|\beta,w) = \frac{(s_j)^{\frac{\beta}{2}-1}exp(-\frac{1}{2}\beta s_j w)}{\Gamma(\frac{\beta}{2})(\frac{2}{w\beta})^{\frac{\beta}{2}}} \propto (w\beta)^{\frac{\beta}{2}}(s_j)^{\frac{\beta}{2}-1}exp(-\tfrac{1}{2}\beta s_j w)$$

$$p(s_k|\beta,w) = \frac{(s_j)^{\frac{\beta}{2}-1}exp(-\frac{1}{2}\beta s_j w)}{\Gamma(\frac{\beta}{2})(\frac{2}{w\beta})^{\frac{\beta}{2}}} \propto (w\beta)^{\frac{\beta}{2}}(s_j)^{\frac{\beta}{2}-1}exp(-\frac{1}{2}\beta s_j w) \tag{5.21}$$

por lo tanto utilizando esta última ecuación (5.21) tenemos lo siguiente

$$p(w|s_k,\beta) \propto \left(\prod_{j=1}^{K}(w\beta)^{\frac{\beta}{2}}(s_j)^{\frac{\beta}{2}-1}exp\left(-\frac{1}{2}s_j w\beta\right)\right)\left(w^{-1/2}exp\left(-\tfrac{w}{2\sigma_y^2}\right)\right)$$

$$p(w|s_k,\beta) \propto \left(\prod_{j=1}^{K}(w\beta)^{\beta/2}\right)exp\left(-\tfrac{1}{2}w\beta\sum_{j=1}^{K}s_j\right)\left(w^{-1/2}exp\left(-\tfrac{w}{2\sigma_y^2}\right)\right)$$

$$p(w|s_k,\beta) \propto \left((s_j w)^{\frac{\beta K}{2}} w^{-\frac{1}{2}}\right) exp\left(-\tfrac{1}{2} w\left[\beta \sum_{j=1}^{K} s_j + \tfrac{1}{\sigma_y^2}\right]\right)$$
$$p(w|s_k,\beta) \propto \left((w)^{\frac{\beta K-1}{2}}\right) exp\left(-\tfrac{1}{2} w\left[\beta \sum_{j=1}^{K} s_j + \sigma_y^{-2}\right]\right)$$

de igual manera encontremos la distribucion a posteriori que al parecer es nuevamente una distribución gamma y queda de la siguiente manera

$$p(w|s_k,\beta) \propto \left((w)^{\frac{\beta K+1}{2}-1}\right) exp\left(-\tfrac{1}{2}\left[\frac{w}{\beta \sum_{j=1}^{K} s_j + \sigma_y^{-2}}\right]\right)$$
$$p(w|s_k,\beta) \propto \left((w)^{\frac{\beta K+1}{2}-1}\right) exp\left(-\tfrac{1}{2}\left[\frac{w(\beta K+1)}{\frac{(\beta K+1)}{\beta \sum_{j=1}^{K} s_j + \sigma_y^{-2}}}\right]\right)$$

por lo tanto

$$p(w|s_k,\beta) \sim g\left(\beta K+1, \frac{\beta K+1}{\beta \sum_{j=1}^{K} s_j + \sigma_y^{-2}}\right) \tag{5.22}$$

Encontremos ahora $p(\beta|s_k,w)$, de igual manera que el anterior tenemos que

$$p(\beta|s_k,w) \propto p(s_j|\beta,w)p(\beta|w)$$
$$p(\beta|s_k,w) \propto p(s_j|\beta,w)p(\beta)$$

como el parámetro que estamos buscando es β entonces será necesario tomar la verosimilitud de la igualdad de la ecuación (5.21) y el dato a priori de (5.18), esto es

$$p(\beta|s_k,w) \propto \left(\prod_{j=1}^{K} \frac{\left(\frac{w\beta}{2}\right)^{\frac{\beta}{2}} (s_j)^{\frac{\beta}{2}-1} exp\left(-\frac{1}{2} s_j w\beta\right)}{\Gamma\left(\frac{\beta}{2}\right)}\right)\left(\beta^{-\frac{3}{2}} exp\left(-\tfrac{1}{2\beta}\right)\right)$$

$$p(\beta|s_k,w) \propto \left(\Gamma\left(\frac{\beta}{2}\right)^{-K}\prod_{j=1}^{K}\left(\frac{w\beta}{2}\right)^{\frac{\beta}{2}}(s_j)^{\frac{\beta}{2}-1}exp\left(-\tfrac{1}{2}s_jw\beta\right)\right)\left(\beta^{-\frac{3}{2}}exp\left(-\tfrac{1}{2\beta}\right)\right)$$

$$p(\beta|s_k,w) \propto \left(\Gamma\left(\frac{\beta}{2}\right)^{-K}\left(\frac{w\beta}{2}\right)^{\frac{\beta K}{2}}(s_j)^{\frac{\beta K}{2}-1}\beta^{-\frac{3}{2}}\prod_{j=1}^{K}exp\left(-\tfrac{1}{2}s_jw\beta\right)\right)\left(exp\left(-\tfrac{1}{2\beta}\right)\right)$$

$$p(\beta|s_k,w) \propto \left(\Gamma\left(\frac{\beta}{2}\right)^{-K}exp\left(-\tfrac{1}{2\beta}\right)\left(\tfrac{w}{2}\right)^{\frac{\beta K}{2}}\left(\tfrac{\beta}{2}\right)^{\frac{\beta K}{2}}(s_j)^{\frac{\beta K}{2}}(s_j)^{-1}\beta^{-\frac{3}{2}}\prod_{j=1}^{K}exp\left(-\tfrac{1}{2}s_jw\beta\right)\right)$$

$$p(\beta|s_k,w) \propto \left(\Gamma\left(\frac{\beta}{2}\right)^{-K}exp\left(-\frac{1}{2\beta}\right)\left(\tfrac{ws_j}{2}\right)^{\frac{\beta K}{2}}\left(\tfrac{\beta}{2}\right)^{\frac{\beta K-3}{2}}\prod_{j=1}^{K}exp\left(-\tfrac{1}{2}s_jw\beta\right)\right)$$

$$p(\beta|s_k,w) \propto \left(\Gamma\left(\frac{\beta}{2}\right)^{-K}exp\left(-\frac{1}{2\beta}\right)\left(\frac{\beta}{2}\right)^{\frac{\beta K-3}{2}}\prod_{j=1}^{K}\left(\frac{ws_j}{2}\right)^{\frac{\beta}{2}}exp\left(-\frac{1}{2}s_jw\beta\right)\right) \tag{5.23}$$

Como podemos ver encontrar su respectivo conjugado a priori no es posible y podemos concluir que no son conjugados a priori, ahora debemos encontrar esta probabilidad a posteriori, como tenemos que $p(log(\beta)|s_k,w)$ es log-concavo, podemos generar muestras independientes usando el muestreo de rechazo adaptable(ARS) (Gilks and Wild [1992]) y obtener valores de β.

La mezcla de proporción π_j está dada por una distribución de Dirichlet (también conocida como multivariada beta) a priori con parámetro α/k, como sigue

$$p(\pi_k|\alpha) \sim Dirichlet(\alpha/K,...,\alpha/K) = \frac{\Gamma(\alpha)}{\Gamma(\alpha/K)^K}\prod_{j=1}^{K}\pi_j^{\alpha/K-1} \tag{5.24}$$

pero las π_j tienen un número de ocupación n_j que se distribuye multinomial y su distribución conjunta es

$$p(c_n|\pi_k) = \prod_{j=1}^{K}\pi_j^{n_j} \tag{5.25}$$

donde:

$$n_j = \sum_{i=1}^{n}\delta_{Kronecker}(c_i,j)$$

$$\delta_{ij} = \begin{cases} 1 & si\ i=j \\ 0 & si\ i\neq j \end{cases}$$

Ahora calculemos la $p(c_n|\alpha)$, de la siguiente manera

$$p(c_n|\pi_k,\alpha)p(\pi_k|\alpha) = p(c,\pi_k|\alpha)$$

Ahora marginalizando sobre π_k, sustituyendo (5.25) y (5.24), tenemos que

$$\begin{aligned}
p(c_n|\alpha) &= \int p(c_n|\pi_k,\alpha)p(\pi_k|\alpha)d\pi_k \\
p(c_n|\alpha) &= \int \left(\prod_{j=1}^{k}\pi_j^{n_j}\right)\left(\frac{\Gamma(\alpha)}{\Gamma(\alpha/K)^K}\prod_{j=1}^{K}\pi_j^{\alpha/K-1}\right)d\pi_j \\
p(c_n|\alpha) &= \frac{\Gamma(\alpha)}{\Gamma(\alpha/K)^K}\int\prod_{j=1}^{K}\pi_j^{n_j+\alpha/K-1}d\pi_j
\end{aligned}$$

Por las propiedades de la integral de Dirichlet (Ferguson [1973]), tenemos que debemos encontrar una constante de normalización para hacer que la integral sea igual a uno por lo que

$$\begin{aligned}
p(c_n|\alpha) &= \frac{\Gamma(\alpha)}{\Gamma(\alpha/K)^K}\int\frac{\Gamma(n+\alpha)}{\Gamma(n+a)}\frac{\prod_{j=1}^{K}(\Gamma(n_j+\alpha/K))}{\prod_{j=1}^{K}(\Gamma(n_j+\alpha/K))}\prod_{j=1}^{K}\pi_j^{n_j+\alpha/K-1}d\pi_j \\
p(c_n|\alpha) &= \frac{\Gamma(\alpha)\prod_{j=1}^{K}(\Gamma(n_j+\alpha/K))}{\Gamma(\alpha/K)^K\Gamma(n+a)}\int\frac{\Gamma(n+\alpha)}{\prod_{j=1}^{K}(\Gamma(n_j+\alpha/K))}\prod_{j=1}^{K}\pi_j^{n_j+\alpha/K-1}d\pi_j
\end{aligned}$$

como sabemos que la parte de la integral es igual a uno, entonces tenemos que

$$p(c_n|\alpha) = \frac{\Gamma(\alpha)}{\Gamma(n+\alpha)}\prod_{j=1}^{K}\frac{\Gamma(n_j+\alpha/K)}{\Gamma(\alpha/K)} \tag{5.26}$$

Ahora con esta última ecuación (5.26) tenemos que

$$p(c_i = j|c_{-i},\alpha) = \frac{p(c_i=j,c_{-i}|\alpha)}{p(c_{-i}|\alpha)} = \frac{p(c_n|\alpha)}{p(c_{-i}|\alpha)}$$

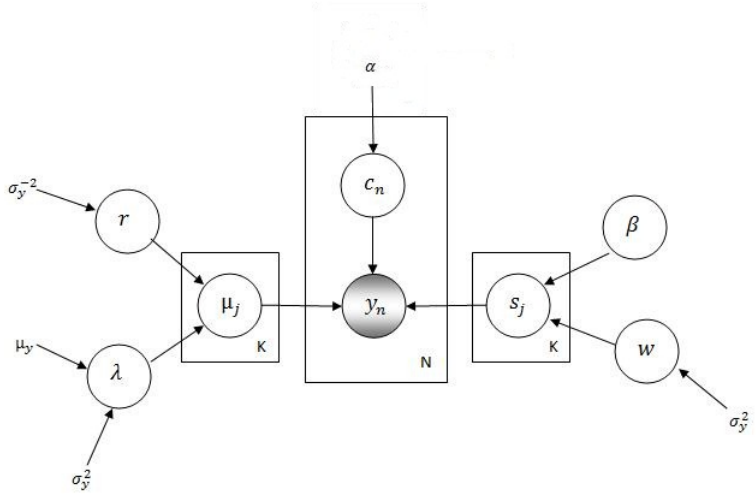


Figure 5.3: Modelo gráfico resultante cuando se integra o marginaliza sobre el parámetro π

$$p(c_i = j|c_{-i}\alpha) = \frac{\frac{\Gamma(\alpha)}{\Gamma(n+\alpha)}\prod_{j=1}^{K}\frac{\Gamma(n_j+\alpha/K)}{\Gamma(\alpha/K)}}{\sum_{c_{i=j}}\frac{\Gamma(\alpha)}{\Gamma(n+\alpha)}\prod_{j=1}^{K}\frac{\Gamma(n_j+\alpha/K)}{\Gamma(\alpha/K)}}$$

$p(c_i = j|c_{-i}\alpha) = \frac{\frac{\Gamma(\alpha)}{\Gamma(n+\alpha)}\frac{\Gamma(n_j+\alpha/K)}{\Gamma(\alpha/K)^K}}{\frac{\Gamma(\alpha)}{\Gamma(n-1+\alpha)}\frac{\Gamma(n_j-1+\alpha/K)}{\Gamma(\alpha/K)^K}}$, esto es debido a que se considera a una j en específico.

$$p(c_i = j|c_{-i}\alpha) = \frac{\Gamma(n-1+\alpha)\Gamma(n_j+\alpha/K)}{\Gamma(n+\alpha)\Gamma(n_j-1+\alpha/K)}$$

ahora para eliminar la función gamma utilizaremos una de sus propiedades

$$p(c_i = j|c_{-i}\alpha) = \frac{\Gamma(n-1+\alpha)\Gamma(n_j+\alpha/K+1-1)}{\Gamma(n+\alpha+1-1)\Gamma(n_j-1+\alpha/K)}$$
$$p(c_i = j|c_{-i}\alpha) = \frac{\Gamma(n-1+\alpha)(n_j+\alpha/K-1)\Gamma(n_j+\alpha/K-1)}{(n+\alpha-1)\Gamma(n+\alpha-1)\Gamma(n_j-1+\alpha/K)}$$
$$p(c_i = j|c_{-i}\alpha) = \frac{(n_j+\alpha/K-1)}{(n+\alpha-1)}$$

Ahora tenemos que $n_j - 1$ es equivalente a $n_{-i,j}$ y sustituyendo tenemos

$$p(c_i = j|c_{-i}\alpha) = \frac{n_{-i,j} + \alpha/K}{n - 1 + \alpha} \tag{5.27}$$

donde:

$-i$ = indica todos los índices excepto el i-ésimo

$n_{-i,j}$ = indica el número de observaciones, excluyendo a y_i asociado al componente j.

Las distribuciones a priori para α y $p(n_k|\alpha)$ estan dadas como

$$p(\alpha^{-1}) \sim g(1,1) \Longrightarrow p(\alpha) \propto \alpha^{-3/2} exp\left(-\frac{1}{2\alpha}\right) \tag{5.28}$$

$$p(n_k|\alpha) = \frac{\alpha^K \Gamma(\alpha)}{\Gamma(n+\alpha)} \tag{5.29}$$

Ahora para obtener $p(\alpha|K,n)$, tenemos que

$p(\alpha|K,n) \propto p(n|\alpha,K)P(\alpha|K)$
$p(\alpha|K,n) \propto p(n|\alpha,K)P(\alpha)$

Sustituyendo (5.29) y (5.28) tenemos que

$p(\alpha|K,n) \propto \left(\frac{\alpha^K \Gamma(\alpha)}{\Gamma(n+\alpha)}\right)\left(\alpha^{-3/2} exp\left(-\frac{1}{2\alpha}\right)\right)$

$$p(\alpha|K,n) \propto \frac{\alpha^{K-3/2} exp\left(-\frac{1}{2\alpha}\right)\Gamma(\alpha)}{\Gamma(n+\alpha)} \tag{5.30}$$

Note que las condicionales a posteriori para α, dependen solamente de las observaciones n y del número de componentes k, y no de como las observaciones estan distribuidas entre los componentes.

La distribución $p(log(a)|K,n)$ es log-concavo y podemos generar muestras independientes de dicha distribución mediante el muestreo de rechazo adaptable (ARS). Ahora para calcular los indicadores c obtenemos las variables de la cobija de Markov de acuerdo a la figura (5.3) , por lo tanto obtenemos

$$p(c=j|c_{-i},\mu_j,s_j,\alpha,y_i) \propto p(y_i|c,\mu_j,s_j,\alpha)p(c_i=j|c_{-i},\mu_j,s_j,\alpha)$$

por independecia condicional obtenemos

$$p(c=j|c_{-i},\mu_j,s_j,\alpha,y_i) \propto p(y_i|c,\mu_j,s_j)p(c_i=j|c_{-i},\alpha)$$

sustituyendo las ecuaciones (5.27) y (5.13) tenemos que

$$p(c = j|c_{-i}, \mu_j, s_j, \alpha, y_i) \propto \left(\frac{n_{-i,j} + \alpha/K}{n - 1 + \alpha}\right) \left((s_j)^{\frac{1}{2}} exp\left(-\frac{s_j(y_i - \mu_j)^2}{2}\right)\right) \tag{5.31}$$

y para este caso se ajusta bien el muestreo Gibbs para obtener las etiquetas (c) que son necesarias para saber del conjunto de datos cuales pertenecen a una determinada gaussiana. Todas las ecuaciones deducidas previamente útiliza una cadena de Markov Monte Carlo para obtener las propabilidades a posteriori y encontrar los valores adecuados cuando la cadena haya llegado a su distribución estacionaria. El algoritmo en pseudocódigo se encuentra en (5.1).

Resumen

Se deducieron todas las probabilidades para realizar la inferencia en los parámetros mediante el uso de la mezcla finita de gaussianas agregando hiperpámetros, todo esto desarrollado se obtiene a partir del valor del número de componentes proporcionado a priori, sin embargo el objetivo final de esta tesis es no proporcionar el número de componentes sino que el propio método lo calcule y para esto lo que se realizará en el siguiente capítulo es hacer que el valor del número de componentes(K) sea reemplazarla por infinito es decir hacer que tienda al infinito y con esto poder decir que el modelo es un modelo de mezcla de gausssinas infinitas, no es trivial comprender que significa tender al infinito pero esto se verá con más detalle en el siguiente capitulo.

Resumen de las ecuaciones a priori

1. $p(y|\mu_k, s_k, \pi_k) = \sum_{j=1}^{K} \pi_j N(\mu_j, s_j^{-1})$
2. $p(\mu_j|\lambda, r) \sim N(\lambda, r^{-1})$
3. $p(\lambda) \sim N(\mu_y, \sigma_y^2)$
4. $p(r) \sim g(1, \sigma_y^{-2}) \propto r^{-1/2} exp\left(\frac{-r\sigma_y^2}{2}\right)$

Algoritmo 5.1 Pseudocódigo principal para la mezcla finita de gaussianas

Entradas: Y, k, e $itera$, donde Y son los datos de la población, k es el número de componentes e $itera$ es el número de iteraciones.
Salidas: μ^t, s^t y π^t son los parámetros de una distribución normal con vectores para la media, inversa de la varianza y pesos respectivamente.

function [μ^t, s^t π^t]=kfinito(Y, k, $itera$)
{
//Datos a priori de la población
$\mu_y = media(Y)$
$\sigma_y^2 = varianza(Y)$
$\sigma_y^{-2} = inversa(varianza(Y))$

//Inicializa las variables a priori
$p(\lambda) \sim N(\mu_y, \sigma^2)$
$p(r) \sim G(1, \sigma_y^{-2})$
$p(\beta) \sim \frac{1}{G(1,1)}$
$p(w) \sim G(1, \sigma_y^2)$
$p(\alpha) \sim \frac{1}{G(1,1)}$
for $j = 1$ to k
$p(s_j|\beta, w) \sim G(\beta, w^{-1})$
$p(\mu_j|\lambda, r) \sim N(\lambda, r^{-1})$
end

//Inicializa los valores de c todos iguales a 1
$c_{1\ldots n} = 1$

Encuentra las frecuencias de las clases n_{ij}

//Realiza la cadena de Markov
for $t = 2$ to $t < itera$
//Encuentra las probabilidades a posteriori
for $j = 1$ to k
$p(\mu_j^t|c, Y, s_j^{t-1}, \lambda^{t-1}, r^{t-1})$ de acuerdo a la ecuación (5.14)
$p(s_j^t|c, Y, \mu_j^{t-1}, \beta^{t-1}, w^{t-1})$ de acuerdo a la ecuación (5.20)
end
$p(\lambda^t|\mu_k^{t-1}, r^{t-1})$ de acuerdo a la ecuación (5.15)
$p(r^t|\mu_k^{t-1}, \lambda^{t-1})$ de acuerdo a la ecuación (5.16)
$p(w^t|s_k^{t-1}, \beta^{t-1})$ de acuerdo a la ecuación (5.22)

//Muestrea por medio del muestreo de rechazo adaptable las variables α y β
$p(\beta^t|s_k^{t-1}, w^{t-1})$ de acuerdo a la ecuación (5.23)
$p(\alpha^t|k, n)$ de acuerdo a la ecuación (5.30)

//Utiliza el muestreo gibbs
for $j = 1$ to k
Obtiene $p(c_i^t = j|c_{-i}^{t-1}, \mu_j^{t-1}, s_j^{t-1}, \alpha^{t-1})$ de acuerdo a la ecuación (5.31)
end

//Encuentra los valores de c de acuerdo al muestreo de una distribución multinomial
$j = arg\,min_{j'=1}^{k} \left(\sum_{h=1}^{j'} p_h \geq X \right)$ donde $X \sim Uniforme(0, 1)$
Encuentra las nuevas frecuencias de las clases n_{ij}
Encuentran los pesos $\pi^t = \frac{n_{ij}}{\sum_{k=1}^{k} n_{ij}}$
Actualizar todas las variables a la siguiente iteración $t = t + 1$
end
}

5. $p(\beta^{-1}) \sim g(1,1) \Longrightarrow p(\beta) \propto \beta^{-3/2} exp\left(-\frac{1}{2\beta}\right)$
6. $p(w) \sim g(1,\sigma_y^2) \propto w^{-1/2} exp\left(\frac{-w}{2\sigma_y^2}\right)$
7. $p(c_i = j|c_{-i}\alpha) = \frac{n_{-i,j}+\alpha/K}{n-1+\alpha}$
8. $p(\alpha^{-1}) \sim g(1,1) \Longrightarrow p(\alpha) \propto \alpha^{-3/2} exp\left(-\frac{1}{2\alpha}\right)$

Ecuaciones a posteriori

1. $p(\lambda|\mu_k, r) \sim N\left(\frac{\mu_y + \sigma_y^2 r \sum_{j=1}^{K} \mu_j}{\sigma_y^2 rK+1}, \frac{1}{\sigma_y^2 rK+1}\right)$
2. $p(\mu_j|c, y, s_j, \lambda, r) \sim N\left(\frac{r\lambda + s_j \sum_{i:c_i=j} y_i}{r+n_j s_j}, \frac{1}{r+s_j n_j}\right)$
3. $p(r|\mu_k, \lambda) \sim g\left(K+1, \frac{K+1}{\sum_{j=1}^{K}(\mu_j-\lambda)^2+\sigma_y^2}\right)$
4. $p(s_j|c, y, \mu_j, \beta, w) \sim g\left(n_j+\beta, \frac{n_j+\beta}{\sum_{i:c_i=j}(y_i-\mu_j)^2+w\beta}\right)$
5. $p(w|s_k, \beta) \sim g\left(\beta K+1, \frac{\beta K+1}{\beta\sum_{j=1}^{K} s_j+\sigma_y^{-2}}\right)$
6. $p(\beta|s_k, w) \propto \left(\Gamma\left(\frac{\beta}{2}\right)^{-K} exp\left(-\frac{1}{2\beta}\right)\left(\frac{\beta}{2}\right)^{\frac{\beta K-3}{2}} \prod_{j=1}^{K}\left(\frac{ws_j}{2}\right)^{\frac{\beta}{2}} exp\left(-\frac{1}{2}s_j w\beta\right)\right)$
7. $p(\alpha|K, n) \propto \frac{\alpha^{K-3/2} exp\left(-\frac{1}{2\alpha}\right)\Gamma(\alpha)}{\Gamma(n+\alpha)}$
8. $p(c = j|c_{-i}, \mu_j, s_j, \alpha, y_i) \propto \left(\frac{n_{-i,j}+\alpha/K}{n-1+\alpha}\right)\left((s_j)^{\frac{1}{2}} exp\left(-\frac{s_j(y_i-\mu_j)^2}{2}\right)\right)$

Capítulo 6

Modelo de Mezcla Infinita de Gaussianas

6.1. Introducción

Hasta ahora hemos considerado el número de componentes como una cantidad finita, en esta sección detallaremos todo lo que respecta a la parte infinita es decir cuando $k \to \infty$ y haremos las derivaciones finales con respecto a las condicionales a posteriori de los indicadores c. Para todas las variables del modelo finito excepto los indicadores, las condicionales a posteriori de las ecuaciones del capítulo anterior, se obtienen sustituyendo para el valor de k, el valor de k_{rep} que se vaya encontrando de acuerdo al número de clases obtenidas por la cadena de Markov.

Una forma de explicar cuando k es infinito es mediante el uso del proceso de Dirichlet que es semejante a tomar el limite $k \to \infty$. Modelar una distribución como una mezcla de distribuciones es útil para estimar la densidad no paramétrica o para una manera de identificar clases latentes que pueden explicar las dependencias observadas entre las variables. La mezcla con un contable número de componentes infinitos es manejado como un marco bayesiano, empleando una distribución a priori para las proporciones de la mezcla tal como un proceso de Dirichlet. Usar la mezcla infinita tiene como finalidad determinar el número correcto de componentes a diferencia de la mezcla finita que esos componentes son a priori y siempre se tienen dificultades para encontrar el adecuado.

Utilizar el proceso de Dirichlet es computacionalmente factible con el desarrollo de métodos de las cadenas de Markov para muestrear la distribución a posteriori de los parámetros de los componentes y la asociación de la mezcla de los componentes con las observaciones. El método muestreo Gibbs puede fácilmente ser implementado para modelos basados en distribuciones conjugadas a priori, pero cuando no son conjugados a priori son apropiados en muchos contextos, frecuentemente el muestreo Gibbs requiere de una integral numérica difícil de tratar. West et al. [1994] usa aproximaciones Monte Carlo para esta integral. Maceachern and MÃŒller [1998] han ideado una manera para el manejo de prioris no conjugados que utiliza un conjunto de parámetros auxiliares.

6.2. Familia Exponencial

La distribución de probabilidad que hemos estudiado en los capítulos anteriores son ejemplos específicos de una amplia clase de distribuciones llamado la "familia exponencial" (Duda and Hart [1973]; Bernardo and Smith [1994]). Los miembros de la familia exponencial tienen muchas importantes propiedades en común y es apropiado discutir estas propiedades con cierta generalidad. En este capítulo es importante estudiar la familia exponencial ya que juega un papel importante debido a que es la base para comprender qué es un conjugado a priori y de donde proviene.

La familia exponencial de distribuciones en x dado los parámetros η, se define como el conjunto de las distribuciones de la forma

$$p(x|\eta) = h(x)g(\eta)exp\{\eta^T u(x)\} \tag{6.1}$$

donde x puede ser un escalar o un vector, y puede ser discreto o continuo. La variable η es llamado el parámetro natural de la distribución y $u(x)$ es alguna función de x. La función $g(\eta)$ puede ser interpretada como el coeficiente que asegura que la distribución es normalizada y por lo tanto satisface

$$g(\eta)\int h(x)exp\{\eta^T u(x)\}dx = 1 \tag{6.2}$$

donde la integral es reemplazada por sumatoria si x es una variable discreta.

Para aclarar más las ideas realizemos el siguiente ejemplo.

Ejemplo.- Consideremos primero la distribución de Bernoulli

$$p(x|\mu) = Bern(x|\mu) = \mu^x(1-\mu)^{1-x} \tag{6.3}$$

Expresando el lado derecho como el logaritmo del exponencial tenemos

$$\begin{aligned} p(x|\mu) &= exp(x\,ln\,\mu + (1-x)\,ln(1-\mu)) \\ &= (1-\mu)exp\left\{ln\left(\frac{\mu}{1-\mu}\right)x\right\} \end{aligned} \tag{6.4}$$

ahora comparando las ecuaciones (6.1) y (6.4) identificamos lo siguiente

$$\eta = ln\left(\frac{\mu}{1-\mu}\right) \tag{6.5}$$

el cuál podemos resolver la ecuación $\mu = \sigma(\eta)$ donde

$$\sigma(\eta) = \frac{1}{1+exp(-\eta)} \tag{6.6}$$

es llamado la función "logística sigmoide". Entonces podemos escribir la distribución de Bernoulli usando la representación estándar (6.1) por lo tanto tenemos que

$$u(x) = x \tag{6.7}$$

$$h(x) = 1 \tag{6.8}$$

$$g(\eta) = \sigma(-\eta) \tag{6.9}$$

6.2.1. Conjugados a Priori

Creamos una distribución a priori $p(\theta)$ y estamos interesados en la distribución a posteriori $p(\theta|x)$. Será deseable si $p(\theta|x)$ es de la misma familia exponencial de $p(\theta)$, porque nos permitirá encadenar las observaciones $X = \{x_i\}_{i=1}^n$ y actualizar la posteriori $p(\theta|x)$ sucesivamente. Esta relación entre $p(x|\theta)$ y $p(\theta)$ es llamado el conjugado a priori el cuál

Verosimilitud	Priori	Posteriori
Multinomial$(\theta_1, ..., \theta_k)$	Dirichlet$(\alpha_1, ..., \alpha_k)$	Dirichlet$(\alpha + x_1, ..., \alpha_k + x_k)$
Binomial(p)	Beta(α, β)	Beta$\left(\alpha + \sum_{i=1}^{n} x_i, \beta + \sum_{i=1}^{n} N_i - \sum_{i=1}^{n} x_i\right)$
Normal$(\mu, \sigma^2 conocida)$	Normal(μ_0, σ_0^2)	Normal$\left(\frac{\left(\frac{\mu_0}{\sigma_0^2} + \frac{\sum_{i=1}^{n} x_i}{\sigma^2}\right)}{\left(\frac{1}{\sigma_0^2} + \frac{n}{\sigma^2}\right)}, \left(\frac{1}{\sigma_0^2} + \frac{n}{\sigma^2}\right)^{-1}\right)$
Normal$(\mu, \tau\, conocida)$	Normal(μ_0, τ_0)	Normal$\left(\frac{\left(\tau\mu_0 + \tau \sum_{i=1}^{n} x_i\right)}{\tau_0 + n\tau}, \tau_0 + n\tau\right)$
Normal$(\mu, \tau\, conocida)$	gamma(α, β)	gamma$\left(\alpha + \frac{n}{2},\ \beta + \frac{\sum_{i=1}^{n}(x_i - \mu)^2}{2}\right)$
gamma$(\alpha\, conocida, \beta)$	gamma(α_0, β_0)	gamma$\left(\alpha_0 + n\alpha, \beta_0 + \sum_{i=1}^{n} x_i\right)$

Cuadro 6.1: Distribuciones conjugadas más comunes, donde τ es la precisión y β es la inversa de la escala

está definido de la siguiente manera:

Definición 1 (Conjugado a priori).- Dada una clase de funciones de verosimilitud $p(x|\theta)$, una clase de distribuciones de probabilidad a priori $p(\theta)$ se dice que es conjugado a la clase de funciones de verosimilitud $p(x|\theta)$ si el resultado de las distribuciones a posteriori $p(\theta|x)$ pertenecen a la misma familia que $p(\theta)$.

6.3. Distribución de Dirichlet

La distribución de Dirichlet de orden n está definido sobre el espacio simplex n-dimensional

$$\Delta_n = \left\{ x \in \mathbb{R}^n : \sum_i x_i = 1, x_i \geq 0 \right\} \tag{6.10}$$

La distribución está parametrizada para n valores positivos $\{\alpha\}_{i=1}^{n}\,(a_i > 0)$

Definicion 2 (Distribución de Dirichlet).- Una variable aleatoria $x \in \Delta_n$ se dice que tiene una distribución de Dirichlet si su función de densidad de probabilidad con respecto a la medida de Lebesgue está dado por

$$p(x_1, ..., x_n) = \frac{\Gamma(\sum_{i=1}^{n} \alpha_i)}{\prod_{i=1}^{n} \Gamma(\alpha_i)} \prod_{i=1}^{n} x_i^{\alpha_i - 1} \tag{6.11}$$

y es denotado como $x \sim Dir(\alpha_1, ..., \alpha_n)$ o simplemente $x \sim Dir(\alpha)$.

6.4. Proceso de Dirichlet

El proceso de Dirichlet(DP) (Ferguson [1973],Ferguson [1974]; Blackwell and MacQueen [1973]; Sethuraman [1994]) es la base de los modelos no paramétricos ampliamente usados para distribuciones aleatorias en estadística bayesiana, debido principalmente a la disponibilidad de eficientes técnicas computacionales. Algunas aplicaciones recientes del proceso de Dirichlet incluye las finanzas (Kacperczyk et al. [2005]), epidemiología (Dunson [2005]), genética (Medvedovic and Sivaganesan [2002]) y medicina (Bigelow and Dunson [2007]).

Definición 3 (Proceso de Dirichlet).- Sea H una distribución sobre Θ y α un número real positivo. Entonces para alguna medida de partición finita $A_1, ..., A_r$ de Θ, el vector $(G(A_1), ...G(A_r))$ es aleatorio debido a que G es aleatorio. Decimos que G es un proceso de Dirichlet distribuido con la distribución base H y parámetro de concentración α, y se escribe como $G \sim DP(\alpha, H)$ si

$$(G(A_1), ..., G(A_r)) \sim Dir(\alpha H(A_1), ..., \alpha H(A_r)) \tag{6.12}$$

para cada medida de partición finita $A_1, ..., A_r$ de Θ.

Para alguna medida del conjunto A tenemos que

$$E(G(A)) = H(A) \tag{6.13}$$

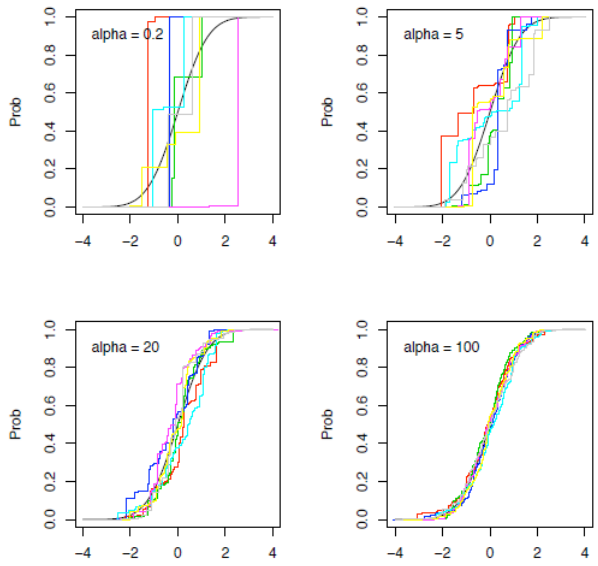


Figura 6.1: Muestras del proceso de Dirichlet centrada en una distribución gaussiana estándar con diferentes parámetros de precisión.

$$Var(G(A)) = \frac{H(A)(1 - H(A))}{\alpha + 1} \tag{6.14}$$

La prueba se puede encontrar en (Xinhua [2008]).

6.4.1. Distribución a Posteriori

Debido a que G es una distribución aleatoria procedente de un proceso de Dirichlet y no es observable, podemos hacer mejores estimaciones de G tomando muestras de G. Esta es la idea de las distribuciones a posteriori.

Supongamos que hemos observado valores $\theta_1, ..., \theta_n$. Sea $A_1, ..., A_r$ una medida de partición finita de Θ y sea $n_k = \#\{i : \theta_i \in A_k\}$ el número de valores observados en A_k. Utilizando la ecuación (6.12) como dato a priori entonces el modelo de la verosimilitud es multinomial porque ha particionado el espacio completo Θ en un conjunto fijo de subconjuntos y $r < \infty$. Por conjugación de las distribuciones Dirichlet y multinomial, tenemos

$$(G(A_1), ..., G(A_r)|\theta_1, ..., \theta_n \sim Dir(\alpha H(A_1) + n_1, ..., \alpha H(A_r) + n_r)) \tag{6.15}$$

Debido a que ecuación (6.15) es verdadera para todo medida de partición finita, podemos predecir que la distribución a posteriori sobre G es también un proceso de Dirichlet. Note que los parámetros tienen una suma constante $\sum_{i=1}^{n} \alpha H(A_r) + n_r = \alpha + n$ donde este último

termino es constante y no depende de la partición. Si podemos encontrar una distribución H' y un número real positivo α' tal que para todas las particiones, $\alpha H(A_i)+n_i = \alpha' H'(A_i)$ para toda $i = 1, ...r$ entonces $G|\theta_1, ..., \theta_n$ deberá ser un proceso de Dirichlet. Afortunadamente es fácil ver que el posteriori de un proceso de Dirichlet tiene un parámetro de concentración actualizado $\alpha' = \alpha + n$ y la distribución base $H = \frac{\alpha H + \sum_{i=1}^{n} \delta_{\theta_i}}{\alpha+n}$, donde δ_{θ_i}es un punto masa localizado en el átomo θ_i es decir es la probabilidad de Dirac concentrado en θ_i. En otras palabras

$$G|\theta_1, ..., \theta_{n-1} \sim DP\left(\alpha + n - 1, \frac{\alpha}{\alpha + n - 1}H + \frac{n-1}{\alpha + n - 1}\frac{\sum_{i=1}^{n}\delta_{\theta_i}}{n-1}\right) \tag{6.16}$$

Ahora estudiamos la distribución predictiva θ_n después de observar $\theta_1, ..., \theta_{n-1}$ y marginalizando sobre G

$$p(\theta_n) = \int_G p(\theta_n, G|\theta_1, ..., \theta_{n-1}) \tag{6.17}$$

Podemos ver que $\theta_n \perp\!\!\!\perp \theta_1, ..., \theta_{n-1}|G$ es decir θ_n es condicionalmente independiente de $\theta_1, ..., \theta_{n-1}$ dado G.

Para todo el conjunto de medidas A tenemos

$$\begin{aligned} p(\theta_n \in A|\theta_1, ..., \theta_{n-1}) &= E[G(A)|\theta_1, ..., \theta_{n-1}] \\ &= \frac{\alpha H(A) + \sum_{i=1}^{n-1}\delta_{\theta_i}}{\alpha+n-1} \end{aligned} \tag{6.18}$$

donde la primera igualdad es por definición y la segunda igualdad viene de la distribución base a posteriori de G.

Debido a que en la ecuación (6.18) tiene una A arbitraria, tenemos que

$$\theta_n|\theta_1, ..., \theta_{n-1} \sim \frac{\alpha H + \sum_{i=1}^{n-1}\delta_{\theta_i}}{\alpha + n - 1} \tag{6.19}$$

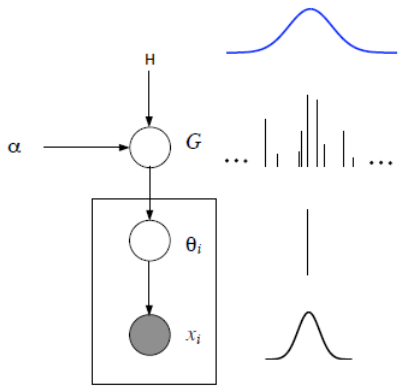


Figura 6.2: Modelo gráfico del proceso de Dirichlet

Ahora vemos que la distribución base a posteriori dado $\theta_1, ..., \theta_{n-1}$ es también la distribución predictiva de θ_n.

Ahora surgen dos preguntas fundamentales:

1. ¿Es la definición 3 razonable?

2. ¿Existe tal manera que se ajuste a la definición?

Hay dos maneras de probarlo, una es mostrando la intercambiabilidad de los procedimientos y luego recurrir al teorema de Finneti, y el otro es mas directo que consiste en construir un proceso de manera explicita y demostrar que cumple las condiciones de la definición 3. El modelo gráfico lo podemos ver en la figura (6.2). Estudiaremos el primer enfoque en las siguientes secciones.

6.5. Esquema de la Urna de Pólya

La ecuación (6.19) provee una conveniente manera de tomar muestras de G, tomando en cuenta que G no es observable por si mismo.

Suponga que cada valor en Θ es un único color, y tomar $\theta \sim G$, que son pelotas con el valor (siendo el color de la pelota) tomado. Ademas tenemos una urna conteniendo las pelotas previamente vistas. En el inicio no hay pelotas en la urna, y tomamos un color de H por ejemplo tomar $\theta_1 \sim H$ una pelota pintada con ese color, y colocarlo en la urna. En subsecuentementes n pasos, con probabilidad $\frac{\alpha}{\alpha+n-1}$ toma un nuevo color (tomar $\theta_n \sim H$),

pintar la pelota con ese color y colocarlo en la urna, o con probabilidad $\frac{n}{\alpha+n-1}$ tomar una pelota aleatoriamente de la urna, registrar su color θ_n y pintar una nueva pelota con ese color y depositar las dos pelotas en la urna.

Es importante interpretar el mecanismo de probabilidad de las pelotas tomadas. Porque las valores de los colores de las pelotas $\{\theta_k\}$ son repetidas, sea $\theta_1^*, ..., \theta_{m-1}^*$ los únicos valores entre $\theta_1, ...\theta_{n-1}$ y n_k son los números de repeticiones de θ_k^* . Entonces la distribución predictiva de la ecuación (6.19) puede ser equivalente escrita como sigue

$$\theta_n|\theta_1, ..., \theta_{n-1} \sim \frac{1}{\alpha + n - 1}\left(\alpha H + \sum_{i=1}^{m-1} n_k \delta_{\theta_k^*}\right) \tag{6.20}$$

Note que el valor de θ_k^* será repetido θ_n con probabilidad proporcional a n_k, que es el número de veces que ya ha sido observado. Mientras más grande es n_k mayor es la probabilidad de que crecerá. Esto es un fenómeno donde grandes grupos(un conjunto de $\theta_i's$ con idénticos valores θ_k^*) se hacen cada vez mas grandes. Esto es bueno y malo, por el lado bueno esto nos lleva al hecho de tomar G de $DP(\alpha, H)$ con probabilidad uno. El número de nuevos colores puede mostrarse que crece logarítmica mente en el número de muestras. Sin embargo por el otro lado puede limitarse a la capacidad del modelo. Finalmente note que este fenómeno no tiene nada que ver con la medida base H el cuál se asume que es suave (continuo) en la mayoría de los casos.

6.5.1. Aplicación del Teorema de Finetti

El esquema de la urna de Pólya ha sido usado para mostrar la existencia del proceso de Dirichlet (Blackwell and MacQueen [1973]). La herramienta básica para eso es teorema de Finneti, quien garantiza un modelo de mezcla que dice que las observaciones son intercambiables, cuál es claramente satisfecha por el esquema de la urna de Pólya, el modelo de mezcla es exactamente el proceso de Dirichlet, y se justifica en la siguiente definición

Definición 4 (Intercambiabilidad de secuencias aleatorias).- Un proceso aleatorio $(\theta_1, \theta_2, ...)$ es llamado intercambiabilidad infinita si para algún $n \in \mathbb{N}$ y alguna permutación σ de $1, ...n$, la probabilidad de generar $(\theta_1, ..., \theta_n)$ es igual a la probabilidad de tomarlos en diferente orden $(\theta_{\sigma(1)}, ..., \theta_{\sigma(n)})$

$$P(\theta_1, ..., \theta_n = p(\theta_{\sigma(1)}, ..., \theta_{\sigma(n)})) \tag{6.21}$$

Teorema 1(Teorema de Finneti).- Suponga un proceso aleatorio $(\theta_1, \theta_2, ...)$ es infinítamente intercambiable, entonces la probabilidad conjunta $p(\theta_1, \theta_2, ..., \theta_N)$ tiene una representación como una mezcla

$$p(\theta_1, \theta_2, ..., \theta_N) = \int \left(\prod_{i=1}^{N} G(\theta_i) \right) dP(G) \tag{6.22}$$

para alguna variable aleatoria G.

Según la definición del esquema de la urna de Pólya especialmente la ecuación (6.19) , construimos una distribución sobre secuencias $\theta_1, \theta_2, ...$ iterativamente tomamos cada θ_i dado $\theta_1, ..., \theta_{n-1}$. Note que la probabilidad condicional de la ecuación (6.19) esta siempre bien definida independientemente si existe un DP.

Para $n \geq 1$ sea

$$p(\theta_1, ..., \theta_n) = \prod_{i=1}^{n} p(\theta_i | \theta_1, ..., \theta_{i-1}) \tag{6.23}$$

la distribución conjunta sobre las primeros n observaciones donde $p(\theta_i | \theta_1, ..., \theta_{i-1})$ está dado por la ecuación (6.19). Es sencillo verificar que esta secuencia aleatoria es infinítamente intercambiable, en efecto si son C colores (por ejemplo la acción de tomar un nuevo color $\theta_n \sim H$ ocurre C veces) y n_c pelotas son tomadas por cada color θ_k^*, entonces el estadístico $\{n_c\}_c$ no cambia con la permutación $(\theta_1, ..., \theta_{n-1})$. Por otra parte $p(\theta_1, ..., \theta_n)$ depende solamente de $\{n_c\}_c$ en

$$p(\theta_1, ..., \theta_n) = \frac{\alpha^C \prod_{c=1}^{C} H(\theta_c^*)(n_c - 1)!}{(\alpha + n - 1)(\alpha + n - 2)...\alpha} \tag{6.24}$$

donde $n = \sum_{i=1}^{C} n_c$.

Ahora los estados del teorema de Finetti que existen a priori sobre las distribuciones aleatorias $p(G)$ es exactamente el proceso de Dirichlet $DP(\alpha, H)$ por lo tanto se establece la existencia.

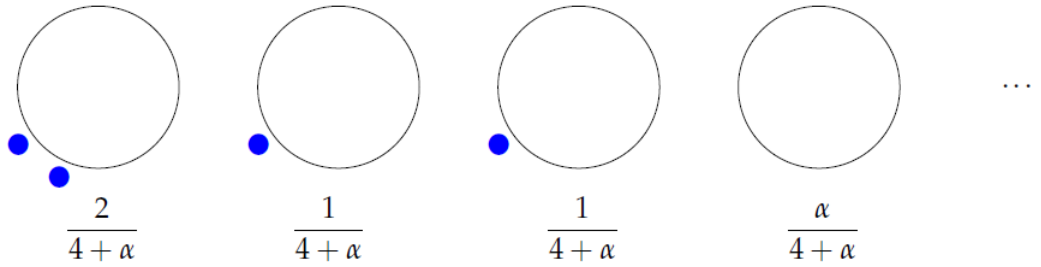


Figura 6.3: Proceso del restaurante chino

6.6. Proceso del Restaurante Chino

También es bien conocido que el proceso de Dirichlet puede ser representado como el proceso del restaurante chino (CRP).

Suponga que se tiene un restaurante chino con un número infinito de mesas, donde en cada mesa se pueden sentar un infinito número de clientes. El primer cliente entra al restaurante y se sienta en la primera mesa, el segundo cliente entra y decide sentarse ya sea con el primer cliente o el solo en una nueva mesa. En general el $(n+1)$-ésimo cliente si se sienta en una mesa k ya ocupada previamente, la probabilidad al número de clientes sentados allí es $\frac{n_k}{\alpha+n-1}$ y si decide sentarse solo en una nueva mesa la probabilidad es $\frac{\alpha}{\alpha+n-1}$. Identificando los clientes con enteros $1, 2, ...$, las mesas como clusters, y los n clientes que se sentaron en las mesas definen una partición de n con la distribución sobre las particiones siendo la misma que la de la Pólya urn scheme como lo muestra la figura (6.3). El hecho de que la mayoría de los restaurantes chinos tienen mesas redondas es un aspecto importante en el CRP, esto se debe a que no se limita a definir una lista de las particiones de n, sino también define una distribución sobre las particiones de n, donde cada mesa corresponde a un ciclo de la permutación.

Es fácil establecer una correspondencia entre el esquema de la urna de Pólya y el proceso del restaurante chino. Por ejemplo, abrir una nueva mesa corresponde a tomar un nuevo color de la pelota y las probabilidades relevantes son claramente las mismas. La única diferencia superficial es que los colores de la Pólya urn scheme son tomados aleatoriamente de H y abrir una nueva mesa parece ser más una acción determinista. En ocasiones se usa el proceso del restaurante chino como metáfora cuando la gente asocia un platillo θ_k^* en cada

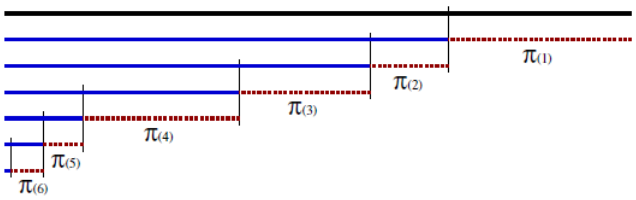


Figura 6.4: Construcción de la varilla quebrada

mesa k y que el platillo es tomado independiente e idénticamente distribuido de H.

Es interesante considerar el número esperado de mesas(cluster) entre los n clientes (observaciones). Entonces el número promedio de m mesas es

$$E[m|n] = \sum_{i=1}^{n} \frac{\alpha}{\alpha + i - 1} \in O(\alpha \log n) \tag{6.25}$$

donde i es la cantidad de datos u observaciones.

También es importante considerar que el número de clusters solo crece logarítmicamente con el número de observaciones, este lento crecimiento tiene sentido por el fenómeno que ocurre en el esquema de la urna de Pólya, que implica que mientras más grande sea α más grande será el número de cluster a priori.

6.7. Construcción de la Varilla Quebrada

Es intuitivo pensar que tomar muestras en este proceso de Dirichlet está compuesto de la suma de pesos de los puntos de masa. Sethuraman [1994] hizo esta precisión al proporcionar una definición constructiva del proceso de Dirichlet llamado la construcción de la varilla quebrada. Esta construcción es significativamente más fácil y general que previas pruebas de la existencia del proceso de Dirichet. Está formada de la siguiente manera

$$\begin{aligned} \beta_k &\sim Beta(1,\alpha) & \theta_k^* &\sim H \\ \pi_k &= \beta_k \prod_{l=1}^{k-1}(1-\beta_k) & G &= \sum_{k=1}^{\infty} \pi_k \delta_{\theta_k^*} \end{aligned} \tag{6.26}$$

Entonces $G \sim Dir(\alpha, H)$. La construcción de π puede ser comprendida metafóricamente como sigue: Comenzando con una varilla de longitud uno, rompemos la varilla en β_1

asignando π_1 la longitud de la varilla que acabamos de romper, ahora rompemos recursivamente la varilla en otras proporciones para obtener π_2, π_3 y así sucesivamente, como lo muestra la figura(6.4). La varilla quebrada se distribuye sobre π y es escrito de esta manera $\pi \sim GEM(\alpha)$ donde las letras son las iniciales de Giffiths, Engen y McCloskey. Debido a su simplicidad, la construcción de la varilla quebrada ha llevado a una variedad de extensiones, así como de nuevas técnicas de inferencia para el proceso de Dirichlet.

Esto definitivamente no es trivial para superar las representaciones precedentes a las otras definiciones del proceso de Dirichlet presentadas en las secciones anteriores. Tratamos de hacerlo en una dirección, derivando la construcción de la varilla quebrada directamente de la definición 2, esta prueba o derivación puede verse en (Xinhua [2008]).

6.8. Modelo de Mezcla Infinita

La aplicación más común del proceso de Dirichlet es el modelo de mezcla infinita, aquí la naturaleza no-paramétrica se traduce en modelo de mezcla con un número infinito de componentes. Modelamos el conjunto de observaciones $\{x_1, ..., x_n\}$ usando un conjunto de parámetros latentes $\{\theta_1, ..., \theta_n\}$, cada θ_i es tomada independientemente e idénticamente distribuido de G, mientras cada x_i tiene una distribución $F(\theta_i)$ parametrizado por θ_i

$$\begin{aligned} x_i|\theta_i &\sim F(\theta_i) \\ \theta_i &\sim G \\ G|\alpha, H &\sim DP(\alpha, H) \end{aligned} \tag{6.27}$$

porque G es discreto, múltiples θ_i pueden tomar el mismo valor simultaneamente y el modelo de arriba puede ser visto como un modelo de mezcla, donde las x_i con el mismo valor de θ_i pertenece a el mismo cluster. La perspectiva de esta mezcla puede estar más de acuerdo con la representación usual representado en los modelos que utilizan la varilla quebrada como muestra la ecuación (6.26). Sea z_i un cluster asignado a una variable, el cuál toma el valor de k con probabilidad π_k. Entonces la ecuación (6.27) puede ser expresado equivalentemente como sigue

$$\begin{aligned} \pi|\alpha &\sim GEM(\alpha) & \theta_k^*|H &\sim H \\ z_i|\pi &\sim Multi(\alpha) & x_i|z_i, \{\theta_k^*\} &\sim F(\theta_{z_i}) \end{aligned} \tag{6.28}$$

La terminología del modelo de mezcla es el siguiente

$G = \sum_{k=1}^{\infty} \pi_k \delta_{\theta_k}$ y $\theta_i = \theta^*_{z_i}$.
π es la proporción de la mezcla.
θ^*_k son los parámetros del cluster k.
$F(\theta^*_k)$ es la distribución sobre los datos del cluster k.
y H son los datos a priori sobre los parámetros de los cluster.

Aunque el modelo del proceso de Dirichlet puede ser visto como un modelo de mezcla infinito, sin embargo porque las π_k decresen exponencialmente, solo un número pequeño de clusters serán usados para el modelo de los datos a priori (en efecto como vimos previamente ecuación (6.25), la esperanza del número de componentes usados a priori es logarítmico en el número de observaciones). Esto es diferente cuando el modelo de mezcla es finito, ya que usa un número fijo de clusters en el modelo de los datos, en el modelo de mezcla del proceso de Dirichlet, el número actual de clusters usados para el modelo no es fijo y puede inferirse automáticamente de los datos usando inferencia bayesiana a posteriori.

6.8.1. Mezcla Infinita de Gaussianas con Hiperparámetros

Todo lo visto anteriormente con el proceso de Dirichlet fueron la base para saber que significa que hagamos que el valor de $k \to \infty$ ya que ahora dadas las ecuaciones analizadas en la sección anterior donde se encontraba las posteriores de los valores de c haremos que el valor de k tienda al infinito para encontrar equivalencias con el proceso de Dirichlet. Ahora siguiendo con las ecuaciones deducidas en el capítulo anterior, tenemos que

Sea la ecuación

$$p(c_i = j | c_{-i}, \alpha) = \frac{n_{-i,j} + \alpha/k}{n - 1 + \alpha} \tag{6.29}$$

Ahora haciendo $k \to \infty$ tenemos que

$$p(c_i = j | c_{-i}, \alpha) = \frac{n_{-i,j}}{n - 1 + \alpha} \tag{6.30}$$

con $n_{-i,j} > 0$.

Ahora cuando $n_{-i,j} = 0$ que es el caso cuando se va a agregar una nueva clase se tiene, se busca el complemento de la siguiente manera

$$P(c_i \neq c_{i'} \forall\, i' \neq i | c_{-i}, \alpha) = 1 - \frac{\sum_{j=1}^{k} n_j - 1}{n - 1 + \alpha} \tag{6.31}$$

$$P(c_i \neq c_{i'} \forall\, i' \neq i | c_{-i}, \alpha) = \frac{n - 1 + \alpha - n + 1}{n - 1 + \alpha} \tag{6.32}$$

$$P(c_i \neq c_{i'} \forall\, i' \neq i | c_{-i}, \alpha) = \frac{\alpha}{n - 1 + \alpha} \tag{6.33}$$

esto es debido a que cuando se sumen todos los componentes la nueva clase solamente tendrá el valor de α y esto hará que la suma sea igual a uno lo cuál se considera la proporción de la nueva clase.

La ecuaciones (6.30) son las clases condicionales a priori para componentes que están asociadas con observaciones que son proporcionales al número de tales observaciones, para la ecuación (6.33) de todas las demás clases dependen únicamente de α y de n.

Ahora para encontrar las condicionales a posteriori hagamos lo siguiente

$p(c_i = j | c_{-i}, \alpha, y_n, \mu_j, s_j) \propto p(y_n | c_n, \alpha, \mu_j, s_j) p(c_i = j | c_{-i}, \alpha, \mu_j, s_j)$

$p(c_i = j | c_{-i}, \alpha, y_n, \mu_j, s_j) \propto p(y_n | c_n, \mu_j, s_j) p(c_i = j | c_{-i}, \alpha)$ Por independencia condicional.

Ahora como sabemos por la deducción en el capítulo anterior que

$$p(y_n | c_n, \mu_j, s_j) \propto s_j^{1/2} exp\left(\frac{-s_j (y_i - \mu_j)^2}{2}\right) \tag{6.34}$$

y de la ecuación (6.30) sustituyendo tenemos

$$p(c_i = j | c_{-i}, \alpha, y_n, \mu_j, s_j) \propto \frac{n_{-i,j}}{n - 1 + \alpha} s_j^{1/2} exp\left(\frac{-s_j (y_i - \mu_j)^2}{2}\right) \tag{6.35}$$

de la misma manera encontramos para

$$p(c_i \neq c_{i'} \forall\, i' \neq i | c_{-i}, \alpha, \lambda, r, \beta, w) \propto$$
$$p(y_n | c_n, \lambda, r, \beta, w, \alpha) p(c_i \neq c_{i'} \forall\, i' \neq i | c_{-i}, \alpha, \lambda, r, \beta, w)$$

$$p(c_i \neq c_{i'} \forall\, i' \neq i | c_{-i}, \alpha, \lambda, r, \beta, w) \propto p(y_n | c_n, \lambda, r, \beta, w, \alpha) p(c_i \neq c_{i'} \forall\, i' \neq i | c_{-i}, \alpha) \quad (6.36)$$

Ahora analizando únicamente $p(y_n | c_n, \lambda, r, \beta, w, \alpha)$ tenemos que

$$p(y_n | c_n, \lambda, r, \beta, w, \alpha) = \int p(y_n, \mu_j, s_j | c_n, \lambda, r, \beta, w, \alpha) d\mu_j ds_j$$
$$p(y_n | c_n, \lambda, r, \beta, w, \alpha) = \int p(y_n | \mu, s_j) p(\mu_j, s_j | c_n, \lambda, r, \beta, w) d\mu_j ds_j$$
$$p(y_n | c_n, \lambda, r, \beta, w, \alpha) = \int p(y_n | \mu, s_j) p(\mu_j, s_j | \lambda, r, \beta, w) d\mu_j ds_j \text{ por independencia condicional}$$
$$p(y_n | c_n, \lambda, r, \beta, w, \alpha) = \int p(y_n | \mu, s_j) p(\mu_j | \lambda, r) p(s_j | \beta, w) d\mu_j ds_j$$

por lo tanto tenemos que

$$p(y_n | c_n, \lambda, r, \beta, w, \alpha) = \int p(y_n | \mu, s_j) p(\mu_j | \lambda, r) p(s_j | \beta, w) d\mu_j ds_j \quad (6.37)$$

Ahora sustituyendo de las ecuaciones (6.33 y 6.37) en la ecuación (6.36) tenemos

$$p(c_i \neq c_{i'} \forall\, i' \neq i | c_{-i}, \alpha, \lambda, r, \beta, w) \propto \frac{\alpha}{n - 1 + \alpha} \int p(y_n | \mu, s_j) p(\mu_j | \lambda, r) p(s_j | \beta, w) d\mu_j ds_j \quad (6.38)$$

Esta última ecuación es llamada clase "no representada" y los valores para los parámetros son tomados aleatoriamente de los parámetros a priori de las cuales se distribuyen gaussiano para μ_j y gamma para s_j.

Ahora vemos que la integral que tenemos no es una integral tratable, una manera de resolverlo es de acuerdo a (Neal [2000]) quien sugiere muestrear de los prioris (en este caso

es una gaussiana y una gamma) para generar una estimación Monte Carlo de la probabilidad de generar una nueva clase. Note que este enfoque genera parámetros (muestreando de los prioris) para las clases no representadas porque la estimación Monte Carlo es imparcial, el resultado de la cadena va a muestrear exactamente la distribución deseada, sin importar cuantas muestras son usadas para aproximar la integral, se ha encontrado que usando una sola muestra trabajan muy bien en muchas aplicaciones.

Existen tres posibilidades cuando el cálculo de las probabilidades condicionales a posteriori dependen del número de observaciones asociadas a la clase.

- Si $n_{-i,j} > 0$ hay observaciones asociadas con la clase j y la clase a posteriori está dada por la ecuación (6.35).
- Si $n_{-i,j} = 0$ no hay observaciones asociadas a la clase j y la clase a posteriori está dada por la ecuación (6.38)
- Si $c_i = j$ la observación y_i es actualmente la única observación asociada con la clase j, esta es una peculiar situación porque no hay otras observaciones asociadas con esa clase, pero como la clase tiene parámetros, resulta que esta situación puede ser manejada como una clase no representada.

Cuando una clase no representada es elegida, una nueva clase es introducida en el modelo y las clases son borradas si llegan a ser vacíos.

Ahora dadas las ecuaciones (6.35) y (6.38) vemos que obtenemos el proceso del restaurante chino esto es

$$p(c_i = j|c_{-i}, \mu_j, s_j, \alpha) = \begin{cases} \frac{n_{-i,j}}{n-1+\alpha} s_j^{1/2} exp\left(\frac{-s_j(y_i-\mu_j)^2}{2}\right) & si\ j\ es\ rep \\ \frac{\alpha}{n-1+\alpha} \int p(y_n|\mu, s_j)p(\mu_j|\lambda, r)p(s_j|\beta, w)d\mu_j ds_j & si\ j\ no\ es\ rep \end{cases} \tag{6.39}$$

por lo que podemos deducir que (6.39) es un proceso de Dirichlet. El algoritmo en pseudocódigo se encuentra en (6.1)

Algoritmo 6.1 Pseudocódigo principal para la mezcla infinita de gaussianas

Entradas: Y e $itera$, donde Y son los datos de la población e $itera$ es el número de iteraciones.
Salidas: μ^t, s^t y π^t son los parámetros de una distribución normal con vectores para la media, inversa de la varianza y pesos respectivamente.

function [μ^t, s^t π^t]=kfinito(Y, k, $itera$)
{
//Datos a priori de la población
$\mu_y = media(Y)$
$\sigma_y^2 = varianza(Y)$
$\sigma_y^{-2} = inversa(varianza(Y))$

//Inicializa las variables a priori
$p(\lambda) \sim N(\mu_y, \sigma^2)$
$p(r) \sim G(1, \sigma_y^{-2})$
$p(\beta) \sim \frac{1}{G(1,1)}$
$p(w) \sim G(1, \sigma_y^2)$
$p(\alpha) \sim \frac{1}{G(1,1)}$
$k^t = 1$
for $j = 1$ to k^t
$p(s_j|\beta, w) \sim G(\beta, w^{-1})$
$p(\mu_j|\lambda, r) \sim N(\lambda, r^{-1})$
end

//Inicializa los valores a priori de c igual a 1
$c_{1...n} = 1$

Encuentra las frecuencias de las clases n_{ij}

//Realiza la cadena de Markov
for $t = 2$ to $t < itera$
//Encuentra las probabilidades a posteriori
for $j = 1$ to k^t
$p(\mu_j^t|c, Y, s_j^{t-1}, \lambda^{t-1}, r^{t-1})$ de acuerdo a la ecuación (5.14)
$p(s_j^t|c, Y, \mu_j^{t-1}, \beta^{t-1}, w^{t-1})$ de acuerdo a la ecuación (5.20)
end
$p(\lambda^t|\mu_k^{t-1}, r^{t-1})$ de acuerdo a la ecuación (5.15)
$p(r^t|\mu_k^{t-1}, \lambda^{t-1})$ de acuerdo a la ecuación (5.16)
$p(w^t|s_k^{t-1}, \beta^{t-1})$ de acuerdo a la ecuación (5.22)

//Muestrea por medio del muestreo de rechazo adaptable las variables α y β
$p(\beta^t|s_k^{t-1}, w^{t-1})$ de acuerdo a la ecuación (5.23)
$p(\alpha^t|k, n)$ de acuerdo a la ecuación (5.30)

//Utiliza el muestreo gibbs y de acuerdo al restaurante chino obtiene
$p(c_i = j|c_{-i}, \alpha) = \frac{n_{-ij}}{n-1-\alpha}$ si $n_{-ij} > 0$
$p(c_i \neq c_{i'} \, para \, todo \, i' \neq i|c_{-i}, \alpha) = \frac{\alpha^{t-1}}{n-1-\alpha^{t-1}}$ en otro caso

//Encuentra los valores a posteriori de acuerdo a lo siguiente
for $j = 1$ to k^t
Obtiene $p(c_i^t = j|c_{-i}^{t-1}, \mu_j^{t-1}, s_j^{t-1}, \alpha^{t-1})$ de acuerdo a la ecuación (6.39) y utiliza el algoritmo de (Neal [2000]) para resolver la integral
end

//Encuentra los valores de c de acuerdo al muestreo de una distribución multinomial
$j = arg\,min_{j'=1}^{k} \left(\sum_{h=1}^{j'} p_h \geq X \right)$ donde $X \sim Uniforme(0, 1)$
Encuentra las nuevas frecuencias de las clases n_{ij}
Encuentra los pesos $\pi^t = \frac{n_{ij}}{\sum_{k=1}^{k} n_{ij}}$
Actualizar todas las variables a la siguiente iteración $t = t + 1$
end
}

Resumen

En este capítulo deducimos la manera en transladar la parte finita a la parte infinita comprendiendo bien la tendencia al infinito que nos proporciona el proceso de Dirichlet en especial el proceso del restaurante chino, con este esquema ya no es necesario tener a priori el valor del número de componentes ya que de eso se encarga este modelo de mezcla infinita de gaussianas.

Capítulo 7

Experimentos y Resultados

7.1. Introducción

El objetivo del modelo de mezcla infinita de gaussianas es calcular la densidad de un conjunto de datos, una parte importante es que el modelo nos proporciona de manera implícita el número adecuado de componentes, en este capítulo mostraremos las distintas pruebas que se realizaron, las comparaciones con los metodos de mezcla finita e infinita de gaussianas y la maximización de la esperanza (EM), la aplicación en el cuál se centra este trabajo es encontrar una pre-segmentación de imágenes cerebrales y encontrar su densidad por medio de la mezcla de gaussianas.

La parte fundamental de este capítulo de la tesis, se encuentra en analizar datos univariados, sin embargo también se realizaron experimentos con datos multivariados, en la primera parte mostraremos el análisis de datos univariados estos son: análisis de datos sintéticos, análisis de imagenes en blanco y negro de tamaño 256x256 y el análisis de imágenes cerebrales. En la segunda parte mostraremos el análisis de datos bivariados y el análisis de datos en tres dimensiones y en la última parte se harán las discusiones sobre los resultados obtenidos.

Las pruebas para el caso finito se realizaron 5,000 iteraciones de las cuales se tomo 2000 para desecharlas y 10 de desfase y para la maximización de la esperanza (EM) el número máximo de iteraciones es de 5,000 con $1e^{-6}$ de tolerancia.

Después de desechar las 5,000 iteraciones se toma la frecuencia del valor de k de todas las muestras obtenidas, y esa nos indica el valor de k adecuado, sin embargo este valor de k, se compone de dos clases que son llamadas las clases representadas y las clases no representadas, las clases representadas son aquellas que su distribución a posteriori fueron tomados con los parámetros estimados por la cadena de Markov porque el valor de j(número de componentes) es representado, en términos del restaurante chino se diría que son las mesas que ya tienen a los clientes y hay un número considerado de clientes por mesa que pueden representar una gaussiana, por lo contrario los componentes no representados son aquellos en donde el valor de j (número de componentes) no es conocido y el valor a posteriori se calcula a partir de los parámetros a priori, en términos del restaurante chino seria las nuevas mesas que se abrieron y que no hay un número considerable de clientes sentados en ellos en ocasiones en esas mesas solo hay 5, 7, 10 o 15, y no se consideran que puedan formar una gaussiana.

El porcentaje de masa se calcula por la formula

$$pm = \frac{n_{total\,de\,datos}}{n_{total\,de\,datos} + \alpha} \tag{7.1}$$

y la autocovarianza se calcula de la siguiente manera

$$AC(desfase) = E\left[(X_i - \mu)(X_{i+desfase} - \mu)\right] \tag{7.2}$$

donde

E: es el valor esperado.

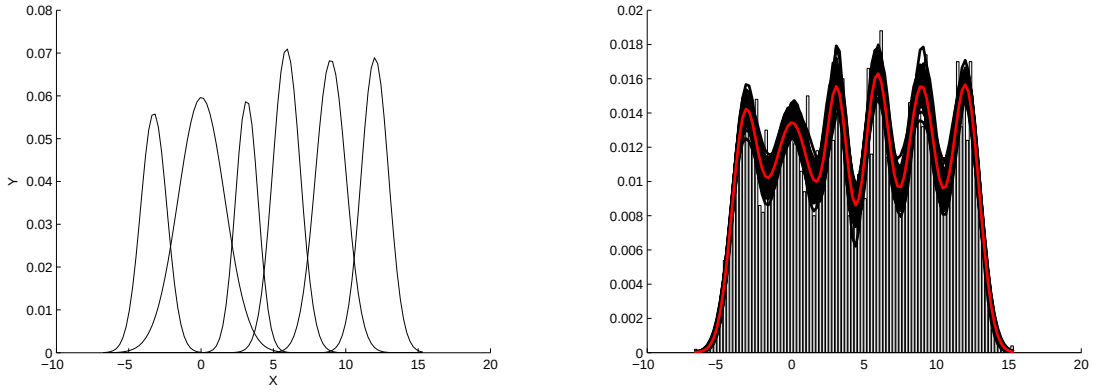
X_i : representa un proceso estacionario.

μ: es la media.

$desfase$: es el desplazamiento.

7.2. Experimentos con Datos Sintéticos

Para estos experimentos lo que se pretendió es saber si nuestro algoritmo realiza una buena aproximación a los parámetros y a la densidad por lo que se generaron datos sintéticos y se compararon con algoritmos: Infinito, finito y EM



(a) (b)

Figura 7.1: (a) Gaussianas encontradas con el modelo de mezclas infinitas (b) Ajuste de la densidad encontrada por el modelo de mezclas infinitas con el histrograma de los datos, los coloreados en negro son todas las muestras de las densidades obtenidas y el coloreado en rojo es la media de ellas, ambos para el primer experimento sintético.

7.2.1. Primer Experimento

Se generó el primer experimento con 5,000 datos con 6 gaussianas con medias $[-3, 0, 3, 6, 9, 12]$, varianzas $[1, 1, 1, 1, 1, 1]$ y componentes $\left[\frac{1}{6}, \frac{1}{6}, \frac{1}{6}, \frac{1}{6}, \frac{1}{6}, \frac{1}{6}\right]$, y se dejo correr el algoritmo con 10,000 iteraciones, los cuales las primeras 5,000 iteraciones se desechan para luego tomar las muestras en intervalos de 10 iteraciones de desfase.

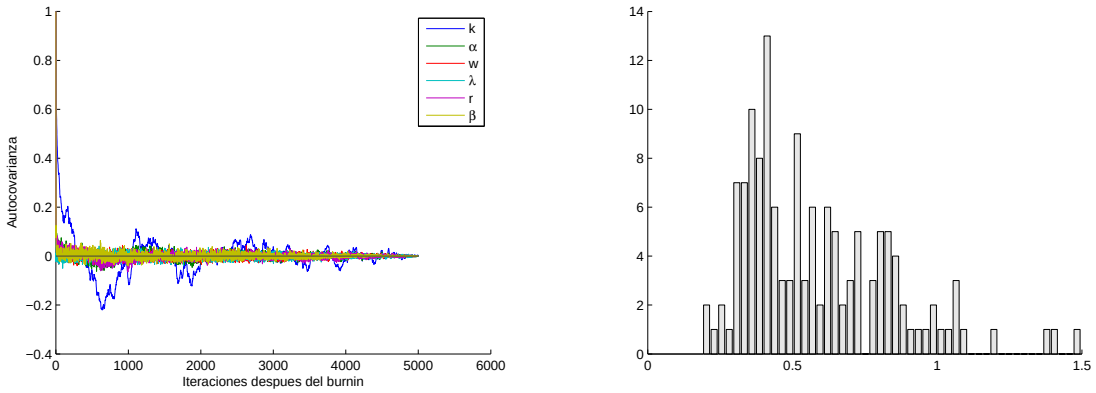
Se tomaron 500 muestras despues de desechar las primeras 5,000 iteraciones en intervalos de 10 iteraciones de desfase y se tomó el número de componentes más frecuente que fue el 6, la gráfica de componentes después de desechar las primeras 5,000 iteraciones puede verse en la figura (7.3)(a), donde puede verse que hay una fuerte tendencia al valor de 6.

Los 6 componentes representan el 99.9 % de la masa de los datos con $\alpha = 0.3860$ como puede versen en la figura (7.2)(b), por lo que las mesas esperadas para el restaurante chino es de 5.1282.

Podemos ver en la tabla (7.1) y en la figura (7.3)(b) que los procesos que involucran la cadena de Markov ajustaron mejor la densidad que el EM quien no hizo un buen ajuste.

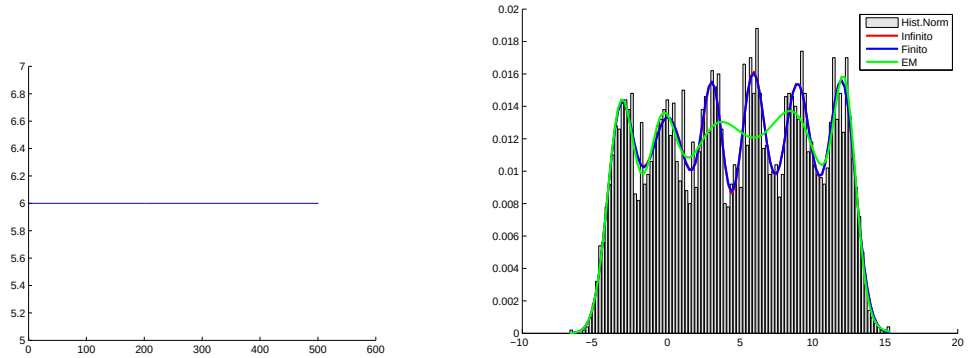
7.2.1.1. Segundo Experimento

Se generaron 10,000 datos con 5 gaussianas con medias $[-3, 0, 3, 5, 7]$, varianzas $[1, 2, 1, 2, 1]$ y componentes $\left[\frac{1}{5}, \frac{1}{5}, \frac{1}{5}, \frac{1}{5}, \frac{1}{5}\right]$, y se dejo correr el algoritmo con 10,000 iteraciones, los cuales se desecharon las primeras 5,000 iteraciones para luego tomar muestras en intervalos



(a) (b)

Figura 7.2: (a) Gráfica de la autocovarianza de los hiperparámetros (b) Histograma del hiperparámetro α, ambos para el primer experimento sintético.



(a) (b)

Figura 7.3: (a) Gráfica del número de componentes despues de desechar las primeras 5,000 iteraciones, (b) Comparación de la densidad encontrada con los tres métodos, ambos para el primer experimento sintético.

Parámetros	Infinito	Finito	EM
Medias	$\begin{bmatrix} -3.2699 \\ 0.0284 \\ 3.1590 \\ 5.9010 \\ 8.9657 \\ 12.0179 \end{bmatrix}$	$\begin{bmatrix} -3.1987 \\ -0.197 \\ 3.1102 \\ 5.9172 \\ 8.9620 \\ 12.0385 \end{bmatrix}$	$\begin{bmatrix} -3.3598 \\ -0.5269 \\ 3.4387 \\ 6.3770 \\ 9.7503 \\ 12.2854 \end{bmatrix}$
Varianzas	$\begin{bmatrix} 0.8234 \\ 2.5985 \\ 0.6098 \\ 0.9063 \\ 1.1330 \\ 0.8964 \end{bmatrix}$	$\begin{bmatrix} 0.8980 \\ 1.8928 \\ 0.8229 \\ 0.8149 \\ 1.2258 \\ 0.8701 \end{bmatrix}$	$\begin{bmatrix} 0.7206 \\ 3.0211 \\ 6.3804 \\ 6.8318 \\ 2.8456 \\ 0.5671 \end{bmatrix}$
Pesos	$\begin{bmatrix} 0.1275 \\ 0.2411 \\ 0.1155 \\ 0.1696 \\ 0.1827 \\ 0.1636 \end{bmatrix}$	$\begin{bmatrix} 0.1416 \\ 0.2072 \\ 0.1410 \\ 0.1601 \\ 0.1894 \\ 0.1606 \end{bmatrix}$	$\begin{bmatrix} 0.1065 \\ 0.1911 \\ 0.2133 \\ 0.2196 \\ 0.1679 \\ 0.1016 \end{bmatrix}$

Cuadro 7.1: Comparación de los parámetros obtenidos con los métodos infinito, finito y EM, todos para el primer experimento sintético.

de 10 iteraciones de desfase, dando un total de 500 muestras tomadas, los resultados fueron

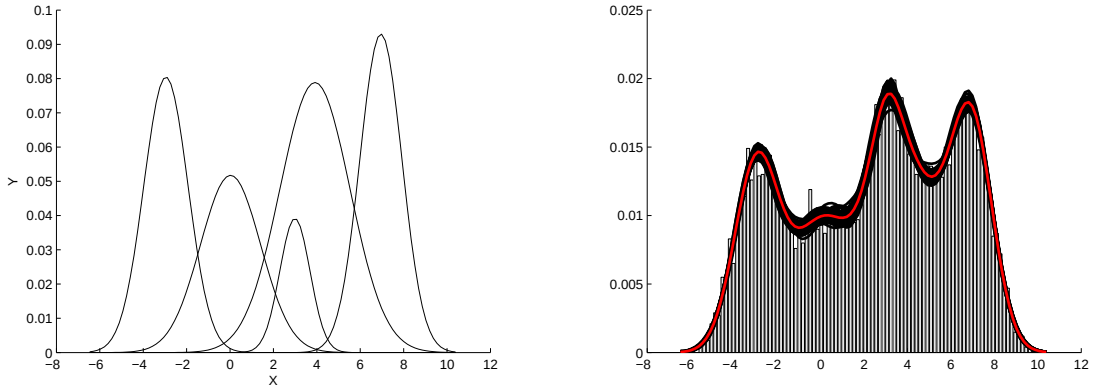
El número de componente mas frecuente fue el 5 como se puede apreciar en la gráfica (7.6)(a), después de desechar las primeras 5,000 iteraciones se puede ver una gran tendencia al valor de 5.

Los 5 componentes representan el 99.9 % de la masa de los datos con $\alpha = 0.3325$ como puede versen en la figura (7.5)(b), por lo que las mesas esperadas para el restaurante chino es de 4.89.

Podemos ver en la tabla (7.2), y la figura (7.6)(b) que los tres procesos nos dan una buena aproximación a la densidad.

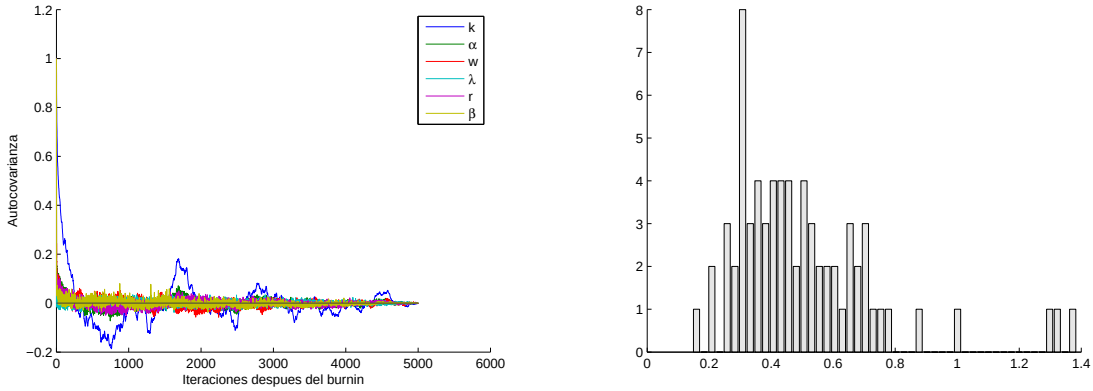
7.2.2. Imágenes Sintéticas con Ruido

En estos experimentos se utilizaron imágenes sintéticas de 128x128 con 5 componentes con medias $[1, 2, 3, 4, 5]$ y varianzas iguales a 0.5, pero proporcionándole ruido gaussiano, estas pruebas se realizaron con la finalidad de probar que tan bien funciona nuestro método ya que nuestra aplicación final tiene que ver con imágenes.



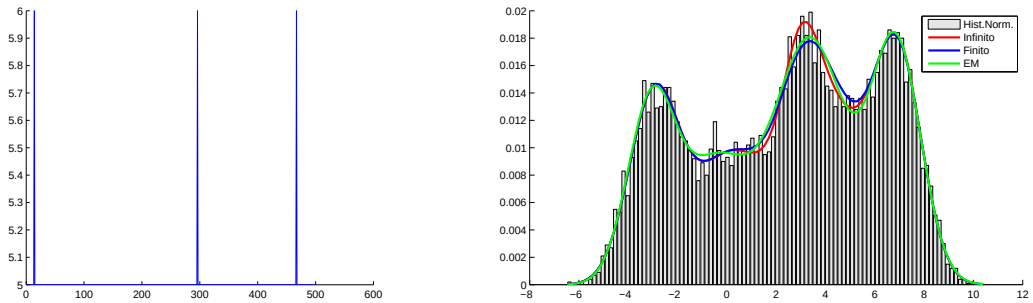
(a) (b)

Figura 7.4: (a) Gaussianas encontradas con el modelo de mezclas infinitas (b) Ajuste de la densidad encontrada por el modelo de mezclas infinitas con el histrograma de los datos, los coloreados en negro son todas las muestras de las densidades obtenidas y el coloreado en rojo es la media de ellas, ambos para el segundo experimento sintético.



(a) (b)

Figura 7.5: (a) Gráfica de la autocovarianza de los hiperparámetros, (b) Histograma del hiperparámetro α, ambos para el segundo experimento sintético.



(a) (b)

Figura 7.6: (a) Gráfica del número de componentes despues de desechar las primeras 5,000 iteraciones (b) Comparación de la densidad encontrada con los tres métodos, ambos para el segundo experimento sintético

Parámetros	Infinito	Finito	EM
Medias	$\begin{bmatrix} -2.9525 \\ 0.0253 \\ 2.9907 \\ 3.9157 \\ 6.9527 \end{bmatrix}$	$\begin{bmatrix} -2.9876 \\ 0.3284 \\ 2.9585 \\ 4.3011 \\ 6.9278 \end{bmatrix}$	$\begin{bmatrix} -3.0558 \\ -0.4451 \\ 2.4575 \\ 3.5090 \\ 6.8170 \end{bmatrix}$
Varianzas	$\begin{bmatrix} 1.0389 \\ 1.9339 \\ 0.4870 \\ 2.6078 \\ 0.9421 \end{bmatrix}$	$\begin{bmatrix} 1.0019 \\ 2.9161 \\ 0.5519 \\ 1.7084 \\ 0.9709 \end{bmatrix}$	$\begin{bmatrix} 0.487 \\ 2.2287 \\ 4.8092 \\ 1.4280 \\ 1.0506 \end{bmatrix}$
Pesos	$\begin{bmatrix} 0.2055 \\ 0.1805 \\ 0.0685 \\ 0.3192 \\ 0.2263 \end{bmatrix}$	$\begin{bmatrix} 0.1930 \\ 0.2460 \\ 0.0974 \\ 0.2238 \\ 0.2398 \end{bmatrix}$	$\begin{bmatrix} 0.1818 \\ 0.1511 \\ 0.1685 \\ 0.2343 \\ 0.2643 \end{bmatrix}$

Cuadro 7.2: Comparación de los parámetros obtenidos por los metodos infinito, finito y EM, todos para el segundo experimento sintetico.

7.2.2.1. Primer Experimento Agregandole Ruido Gaussiano

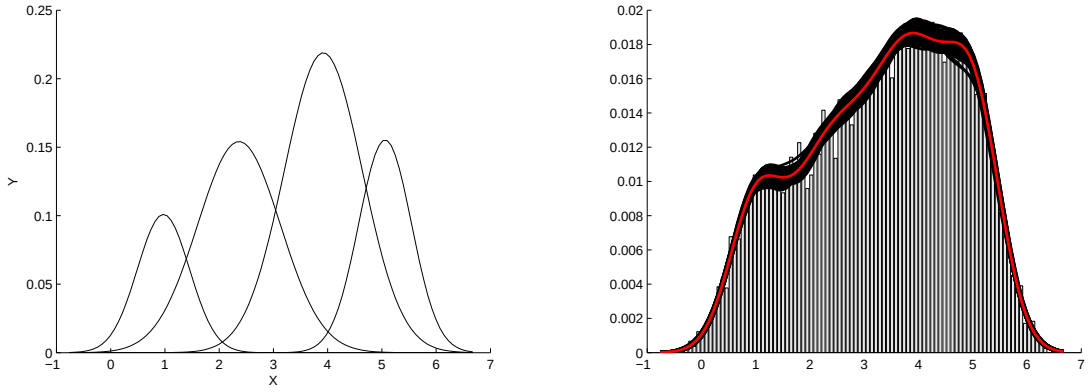
Se corrieron 10,000 iteraciones, desechando las primeras 5,000 iteraciones y obteniendo muestras en intervalos de 10 iteraciones de desfase, los resultados son los siguientes

En la figura (7.7)(a) se puede ver que encontro 4 gaussianas, como puede verse en la figura (7.9)(a) que el número de componentes para este experimento es de 4, el número correcto de componentes es 5, pero el quinto componente no lo encuentra, esto es debido a que el color gris del quinto componente es parecido a los tonos de grises restantes por lo que nuestro algoritmo lo modela como si perteneciera a ellos, esto puede verse en la figura (7.10) donde está la pre-segmentación y la imagen original, alli puede verse el parecido de los tonos de grises.

También puede verse en la figura (7.8)(b) que el alfa mas frecuente se encuentra en 0.3353 por lo que el número de mesas esperadas es de 4.0164 y el porcentaje de masa es del 99.9 %.

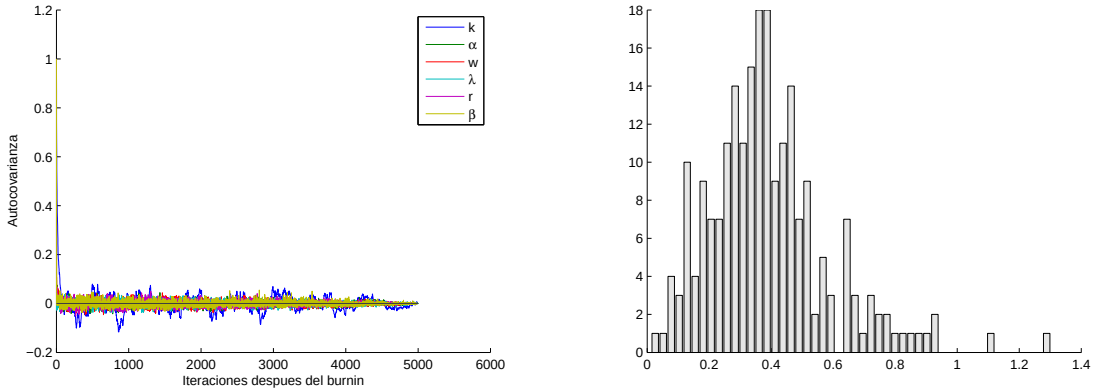
En la figura (7.9)(b) se pueden ver las comparaciones realizadas y se puede ver que el algoritmo del EM no se ajusta bien a los datos pero los demás tanto finito e infinito no tiene problemas para realizar un buen ajuste.

También se puede ver en la figura (7.10) que nuestro algoritmo modela el ruido y por eso la imagen resultante queda tambien ruidosa, lo que se puede ver es que si encuentra



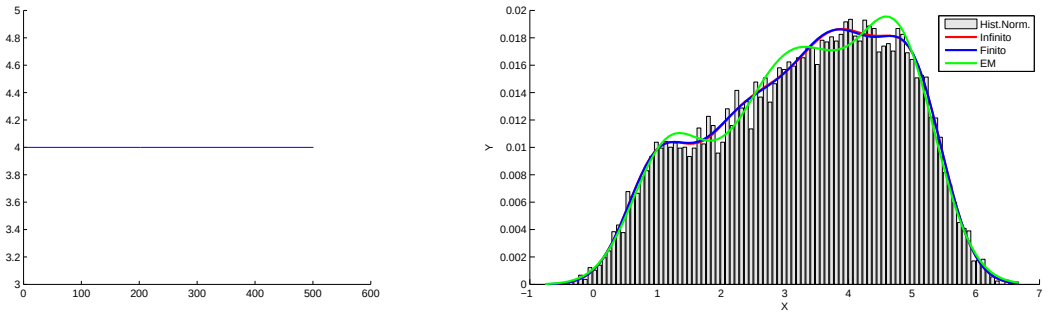
(a) (b)

Figura 7.7: (a) Gaussianas encontradas con el modelo de mezclas infinitas (b) Ajuste de la densidad encontrada por el modelo de mezclas infinitas con el histrograma de los datos, los coloreados en negro son todas las muestras de las densidades obtenidas y el coloreado en rojo es la media de ellas, ambos para la primera imagen sintética.



(a) (b)

Figura 7.8: (a) Gráfica de la autocovarianza de los hiperparámetros (b) Histograma del hiperparámetro α, ambos para la primera imagen sintética.

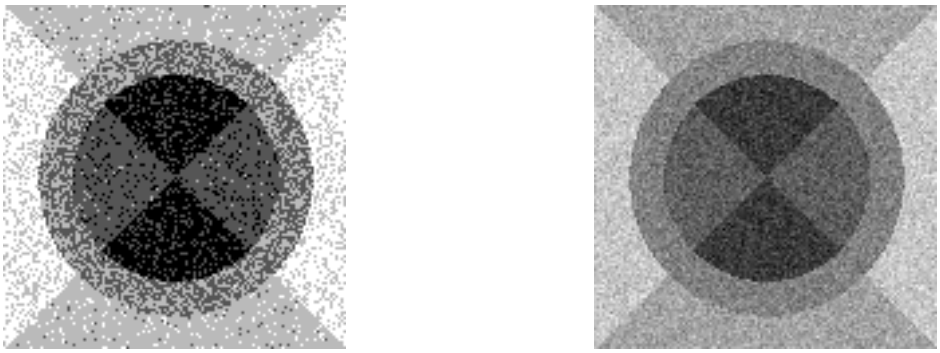


(a) (b)

Figura 7.9: (a) Gráfica de los componentes despues de desechar las primeras 5,000 iteraciones (b) Comparación de los tres métodos, ambos para la primera imagen sintética.

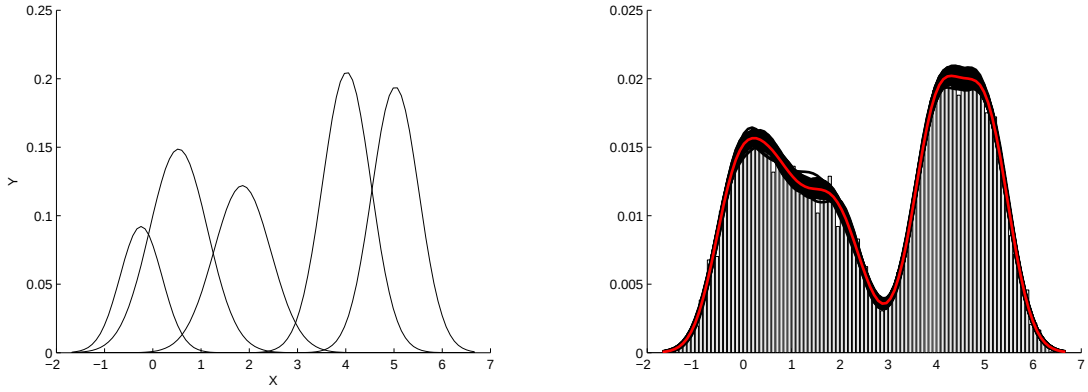
Parámetros	Infinito	Finito	EM
Medias	$\begin{bmatrix} 0.9730 \\ 2.3659 \\ 3.9197 \\ 5.0585 \end{bmatrix}$	$\begin{bmatrix} 0.9730 \\ 2.2770 \\ 3.8913 \\ 5.0701 \end{bmatrix}$	$\begin{bmatrix} 1.1634 \\ 2.9299 \\ 3.4727 \\ 4.8516 \end{bmatrix}$
Varianzas	$\begin{bmatrix} 0.2295 \\ 0.5810 \\ 0.5179 \\ 0.2368 \end{bmatrix}$	$\begin{bmatrix} 0.2316 \\ 0.5274 \\ 0.6158 \\ 0.2252 \end{bmatrix}$	$\begin{bmatrix} 0.3215 \\ 0.9454 \\ 0.8008 \\ 0.3283 \end{bmatrix}$
Pesos	$\begin{bmatrix} 0.1212 \\ 0.2945 \\ 0.3949 \\ 0.1894 \end{bmatrix}$	$\begin{bmatrix} 0.1225 \\ 0.2572 \\ 0.4505 \\ 0.1698 \end{bmatrix}$	$\begin{bmatrix} 0.1668 \\ 0.2695 \\ 0.2881 \\ 0.2755 \end{bmatrix}$

Cuadro 7.3: Comparación de los parámetros obtenidos por los metodos infinito, finito y EM, todos para la primera imagen sintética.



(a) (b)

Figura 7.10: (a) Imagen Segmentada (b) Imagen Original, ambos para la primera imagen sintética.



(a) (b)

Figura 7.11: (a) Gaussianas encontradas con el modelo de mezclas infinitas (b) Ajuste de la densidad encontrada por el modelo de mezclas infinitas con el histrograma de los datos, los coloreados en negro son todas las muestras de las densidades obtenidas y el coloreado en rojo es la media de ellas, ambos para la segunda imagen sintética.

bien los tonos de gris que mas sobresalen e incluso unió un tono de gris que se parecen a los restantes y eso es bueno debido a que con menos componentes está realizando una buena estimación de la densidad. Ahora la segmentacion puede mejorar quitandole el ruido encontrado, es decir nuestro algoritmo puede servir como entrada para otros algoritmos que quitan ruido y la imagen quedaria mejor segmentada. En pocas palabras nuestro algoritmo no toma en cuenta la correlación espacial.

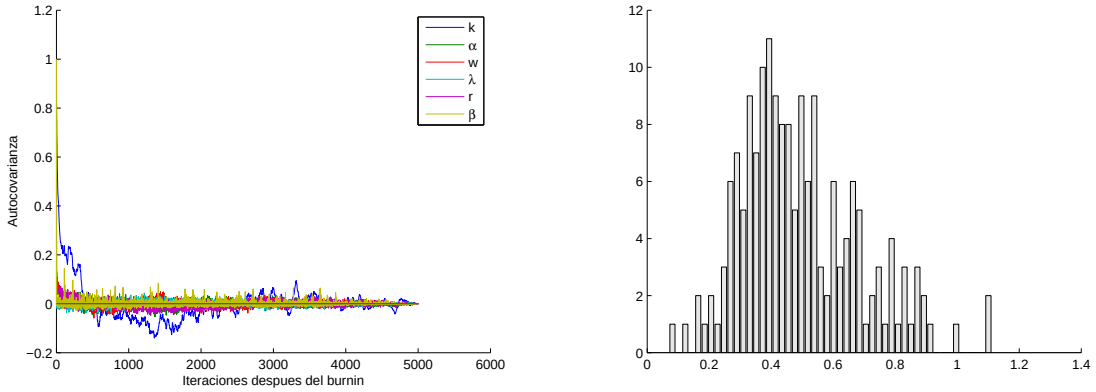
En la tabla (7.3) se puede ver que los valores del infinito y del finito realizaron una buena aproximación, pero en el caso del EM no encontró una buena aproximación de la varianza y de los pesos por lo que el ajuste es malo.

7.2.2.2. Segundo Experimento Agregando Ruido Gaussiano pero con un Arillo Diferenciable.

Se corrieron 10,000 iteraciones, desechando las primeras 5,000 iteraciones y obteniendo muestras en intervalos de 10 iteraciones de desfase, los resultados son los siguientes

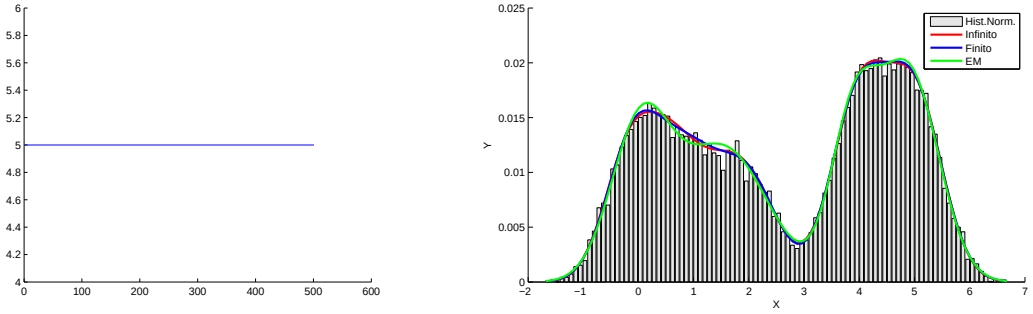
En este experimento lo que se intentó probar es que si el arillo de la imagen no se pareciera a los tonos de grises restantes deberia encontrar sin problemas el número adecuado de componentes, se realizó la prueba y se vió que efectivamente el algoritmo encontró sin ningun problema el quinto componente y los resultados estuvieron muy bien, como puede verse en la aproximación de la densidad en la figura (7.11)(b).

Para el valor de α se puede ver en la figura (7.12)(b) que el valor más frecuente de



(a) (b)

Figura 7.12: (a) Gráfica de la autocovarianza de los hiperparámetros (b) Histograma del hiperparámetro α, ambos para la segunda imagen sintética.



(a) (b)

Figura 7.13: (a) Gráfica de los componentes despues de desechar las primeras 5,000 iteraciones (b) Comparación de los tres métodos, ambos para la segunda imagen sintética.

$\alpha = 0.4351$ por lo que el porcentaje de masa es del 99.9 % y las mesas esperadas del restaurante chino es de 4.38.

La comparación de los tres métodos se puede ver en la figura (7.13)(b) donde se puede apreciar que los tres métodos realizan una buena aproximación a la densidad, en este caso el método EM realizó una mejor estimación que el anterior como puede verse en la tabla(7.4)

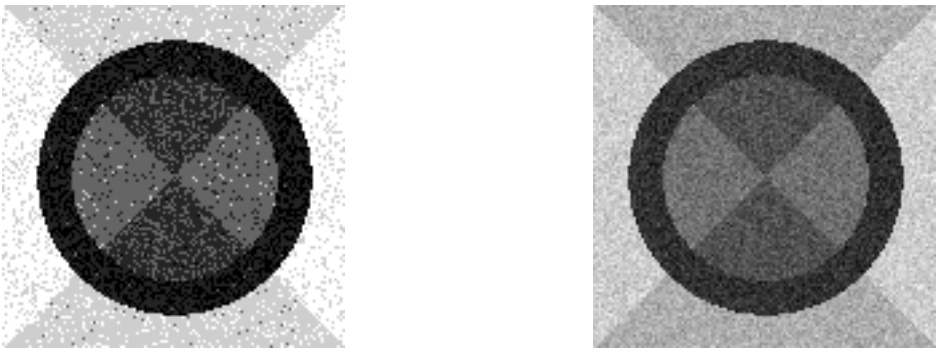
Para el caso de la pre-segmentación al igual que el anterior siempre detecta el ruido, incluyendo el ruido que contiene el arillo negro, porque nuestro algoritmo no modela la correlación espacial.

7.3. Experimentos con Imágenes en Blanco y Negro

En estos experimentos lo que se pretende es saber si nuestro algoritmo puede encontrar la densidad de una imagen sin ruido, pero sobre todo conocer cual es la pre-segmentación

Parámetros	Infinito	Finito	EM
Medias	$\begin{bmatrix} -0.2456 \\ 0.5336 \\ 1.8560 \\ 4.0211 \\ 5.0279 \end{bmatrix}$	$\begin{bmatrix} -0.0808 \\ 1.0313 \\ 2.0911 \\ 3.9068 \\ 4.9424 \end{bmatrix}$	$\begin{bmatrix} -0.1798 \\ 0.6086 \\ 1.4983 \\ 1.8239 \\ 4.4866 \end{bmatrix}$
Varianzas	$\begin{bmatrix} 0.1886 \\ 0.3587 \\ 0.3513 \\ 0.2517 \\ 0.2388 \end{bmatrix}$	$\begin{bmatrix} 0.2398 \\ 0.5098 \\ 0.2133 \\ 0.2569 \\ 0.2830 \end{bmatrix}$	$\begin{bmatrix} 0.2070 \\ 0.2462 \\ 0.2861 \\ 0.6667 \\ 0.5280 \end{bmatrix}$
Pesos	$\begin{bmatrix} 0.1003 \\ 0.2232 \\ 0.1813 \\ 0.2575 \\ 0.2377 \end{bmatrix}$	$\begin{bmatrix} 0.1690 \\ 0.2473 \\ 0.0856 \\ 0.2211 \\ 0.2771 \end{bmatrix}$	$\begin{bmatrix} 0.1451 \\ 0.1471 \\ 0.0765 \\ 0.1313 \\ 0.4999 \end{bmatrix}$

Cuadro 7.4: Comparación de los parámetros obtenidos por los metodos infinito, finito y EM, todos para el segundo experimento sintético.

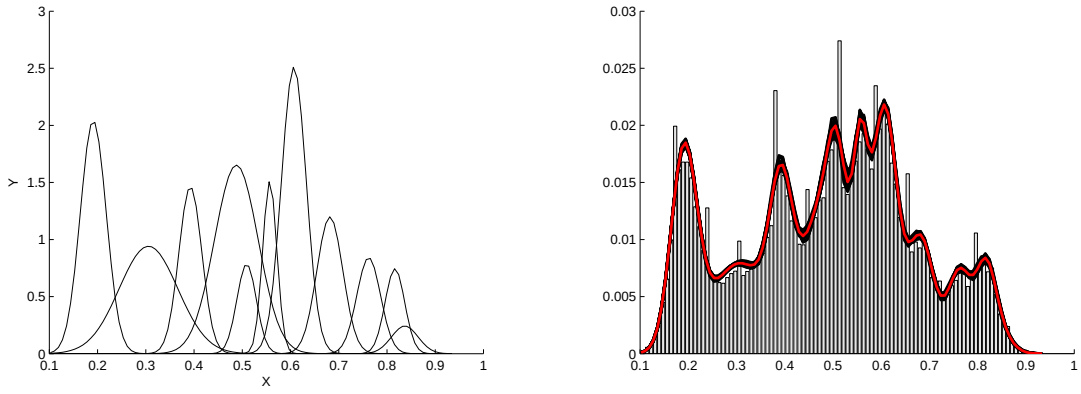


(a) (b)

Figura 7.14: a) Imagen Segmentada b) Imagen Original, ambos para la segunda imagen sintética.

Algoritmo 7.1 Algoritmo de pre-segmentación a imágenes en blanco y negro

1: Procesar la imagen con el algoritmo de mezcla infinita de gaussianas.
2: Encontrar los parámetros(medias, varianzas y pesos) y las etiquetas de los pixeles de cada muestra, proporcionadas por el modelo de mezcla infinita de gaussianas.
3: Ordenar las medias de menor a mayor y hacer coincidir las varianzas y pesos de acuerdo al ordenamieto de ellas.
4: Hallar las medias de los parámetros encontrados.
5: Hallar las etiquetas de los pixeles, obteniendo el adecuado al tomar el más frecuente de cada pixel de todas las muestras.
6: Visualizar la imágen de cada componente con su respectiva etiqueta.
7: Visualizar la imagen segmentada tomando el tono de gris de acuerdo a sus respectivas medias.



(a) (b)

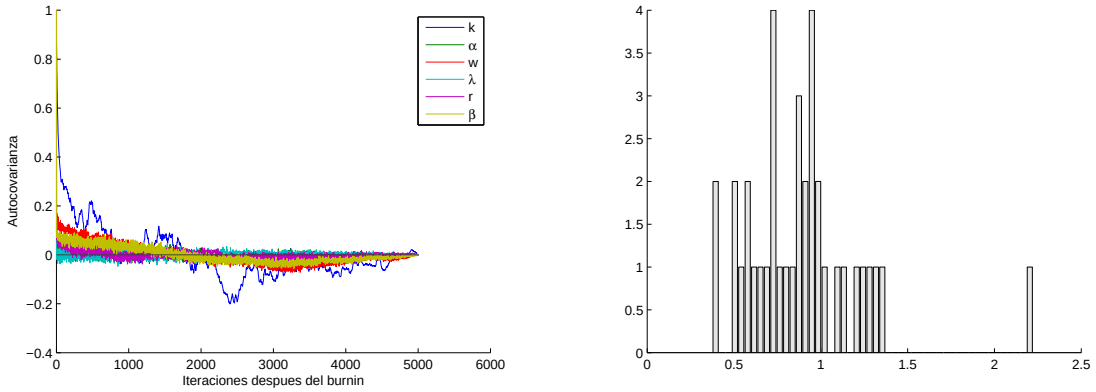
Figura 7.15: (a) Gaussianas encontradas con el modelo de mezclas infinitas (b) Ajuste de la densidad encontrada por el modelo de mezclas infinitas con el histrograma de los datos, los coloreados en negro son todas las muestras de las densidades obtenidas y el coloreado en rojo es la media de ellas, ambos para la imagen de Lena.

que nos proporciona el algoritmo y cuales serian los detalles cuando esto sucede, el algoritmo de segmentacion se puede ver en (7.1).

7.3.1. Imagen de la Modelo Lena

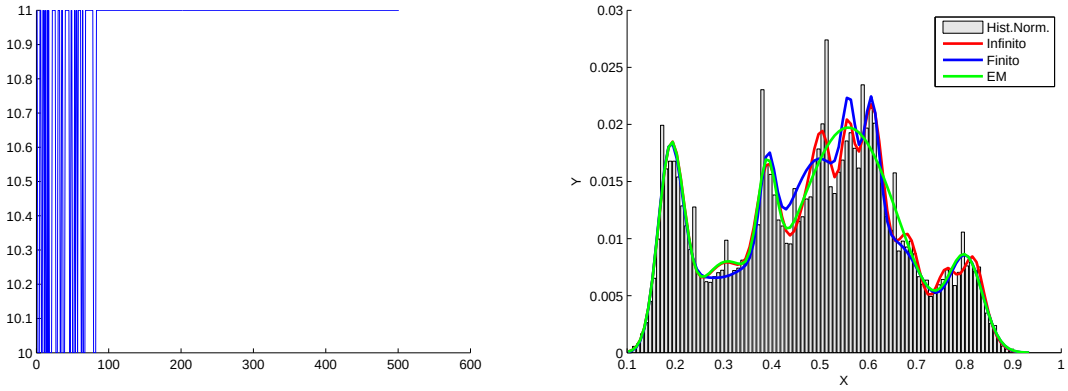
Se corrieron 10,000 iteraciones, desechando las primeras 5,000 iteraciones y obteniendo muestras en intervalos de 10 iteraciones de desfase, obteniendo 500 muestras, los resultados son los siguientes

La figura (7.15)(b) tiene ciertos picos que produce que al momento en que nuestro algoritmo realize la busqueda del número de gaussianas tome en cuenta estos picos y los modele de una buena manera, estos tipos de histogramas con tanto detalle hace que nuestro algoritmo aproxime muy bien la densidad con el número adecuado de componentes tomando



(a) (b)

Figura 7.16: (a) Gráfica de la autocovarianza de los hiperparámetros, b) Histograma del hiperparámetro α, ambos para la imagen de Lena.



(a) (b)

Figura 7.17: (a) Gráfica de los componentes encontrados despues de desechar las primeras 5,000 iteraciones, (b) Comparación de las densidades con los tres métodos, ambos para la imagen de Lena.

en cuenta los detalles como picos o ciertas monticulos dado que el histograma no es suave, y por eso podemos deducir que nuestro algoritmo estima bien la densidad.

Para el valor de α se puede ver en la figura (7.16)(b) que existen dos valores más frecuentes, se toma el mas grande ya que este es el que puede afectar los resultados del porcentaje de masa y del número de mesas, ya que mientras más grande es el valor de α el valor del porcentaje de masa es menor, por lo que $a = 0.90$, por lo que el porcentaje de masa es del 99.8 % y el número de mesas esperadas del restaurante chino es de 10.66.

El número de componentes encontrados es de 11 como puede verse en la figura (7.17)(a), como puede verse en las primeras 100 muestras despues de desechar las primeras 5,000 iteraciones el valor oscila entre 10 y 11 sin embargo despues de las 100 muestras el valor de

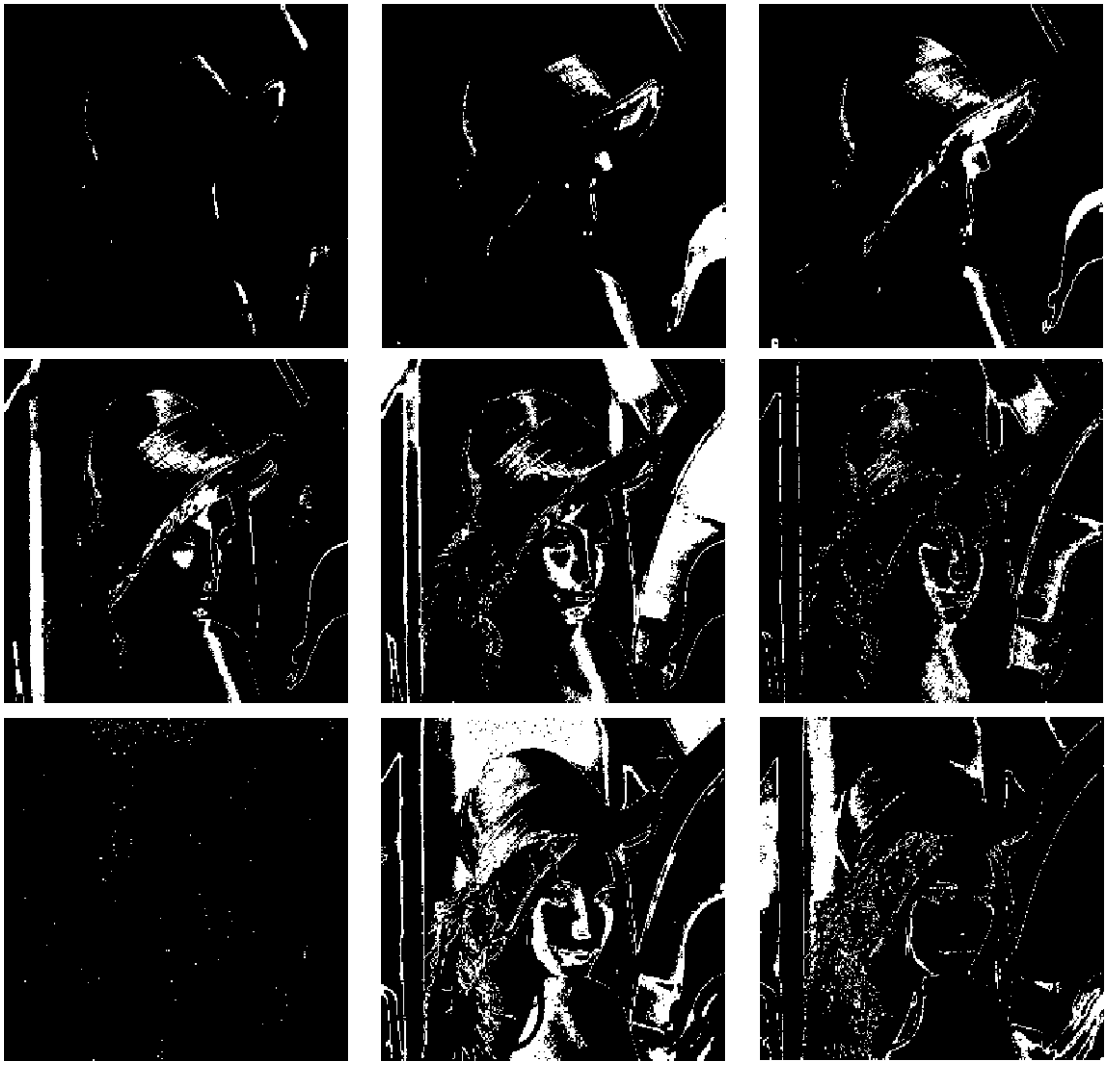


Figura 7.18: Visualización de los primeros 9 componentes de la imagen Lena.

k se estabiliza en 11 y de alli ya no se mueve por lo que nos indica una convergencia.

Ahora comparando las densidades encontradas con otros métodos se puede ver que el infinito ajusta muy bien la densidad, el caso finito nos proporciona una buena aproximación pero el caso del EM la estimación a la densidad es mala.

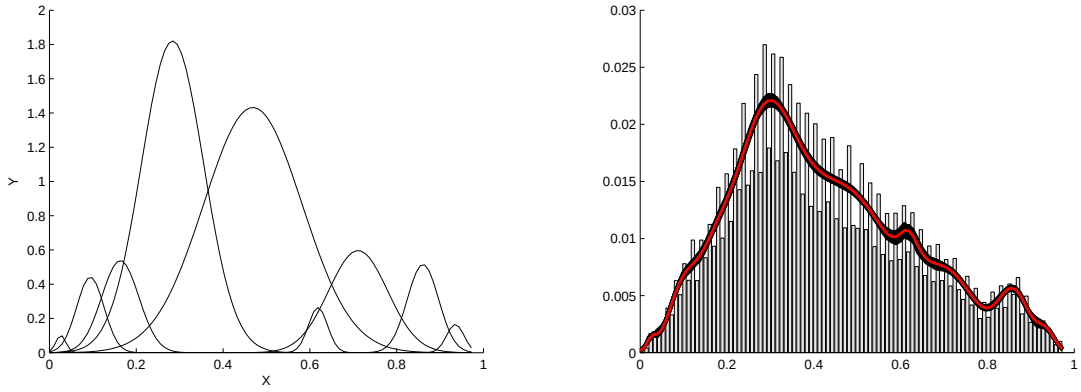
Ahora veamos como realiza la pre-segmentación de la imagen, podemos ver en las figuras (7.18) y (7.19) los 11 componentes encontrados y como habiamos mencionado este método calcula muy bien los detalles por lo que podemos ver en las diferentes imágenes de los componentes algunas tienen cierto parecido que en la imágen original se puede apreciar que tiene otro tono de gris y por lo tanto es otro componente.

La pre-segmentación final para la imagen de Lena puede verse en la figura (7.20) y la original en la misma, allí se puede ver que la pre-segmentación aparece bastante bien y es muy parecida a la original, y como habiamos dicho antes contiene todos los detalles o los

Figura 7.19: Visualización de los últimos 2 componentes de la imagen Lena.

(a) (b)

Figura 7.20: (a) Imagen Segmentada (b) Imagen Original, ambos de la imagen de Lena.



(a) (b)

Figura 7.21: (a) Gaussianas encontradas con el modelo de mezclas infinitas (b) Ajuste de la densidad encontrada por el modelo de mezclas infinitas con el histrograma de los datos, los coloreados en negro son todas las muestras de las densidades obtenidas y el coloreado en rojo es la media de ellas, ambos para la imagen del puente.

distintos tonos de grises que aparecen en la original.

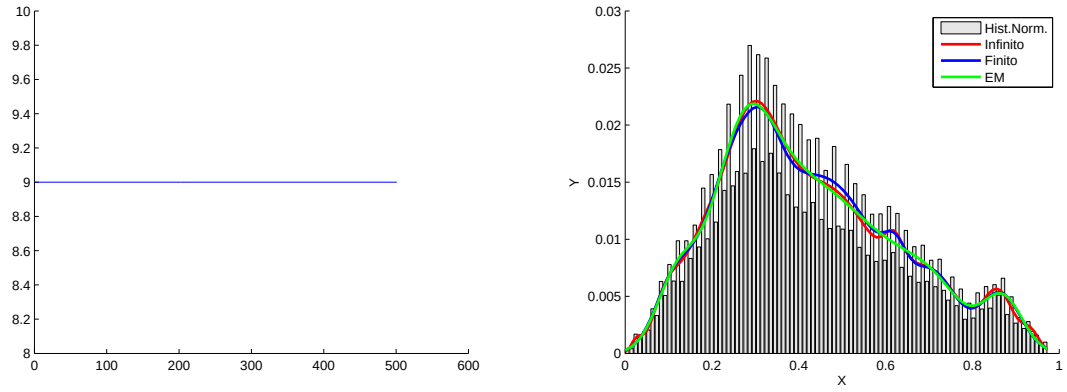
Para concluir podemos decir que nuestro algoritmo es muy bueno para ver los distintos detalles contenidos en el histograma y lo modela con la cantidad necesaria de componentes con el fin de acaparar todos los tonos de grises posibles, por lo contrario si en lo que se requiere hacer no es necesario tanto detalle porque quieres una densidad mas suave por ejemplo, nuestro algoritmo será inapropiado, porque puede encontrar muchos componentes que a lo mejor no sirva.

7.3.2. Imagen de un Puente

Se corrieron 20,000 iteraciones, desechando las primeras 10,000 iteraciones y obteniendo muestras en intervalos de 20 iteraciones de desfase, obteniendo 500 muestras, los resultados son los siguientes

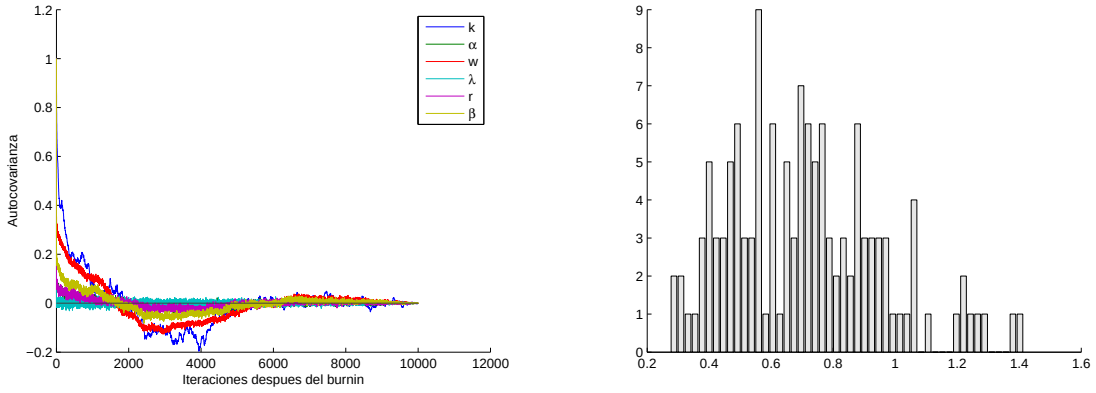
En esta prueba se encontraron 9 componentes para la imagen y se puede notar en la figura(7.22)(a) que en todas las muestras este fue el número indicado para crear 9 gaussianas que se muestran en la figura (7.21)(a) y todas las muestras fueron muy parecidas ya que muchas inclusive estuvieron encimadas.

En el cálculo de la densidad los tres métodos nos proporcionan una buena aproximación en el caso del infinito y del finito se ve que ajusta bien los detalles como son pequeñas montañas o picos esto puede verse en la figura (7.22)(a) y el método EM su ajuste es mas suave.



(a) (b)

Figura 7.22: (a) Gráfica de los componentes después de desechar las primeras 10,000 iteraciones (b) Comparación de las densidades con los tres métodos, ambos para la imagen del puente.



(a) (b)

Figura 7.23: (a) Gráfica de la autocovarianza de los hiperparámetros, (b) Histograma del hiperparámetro α, ambos para la imagen del puente.

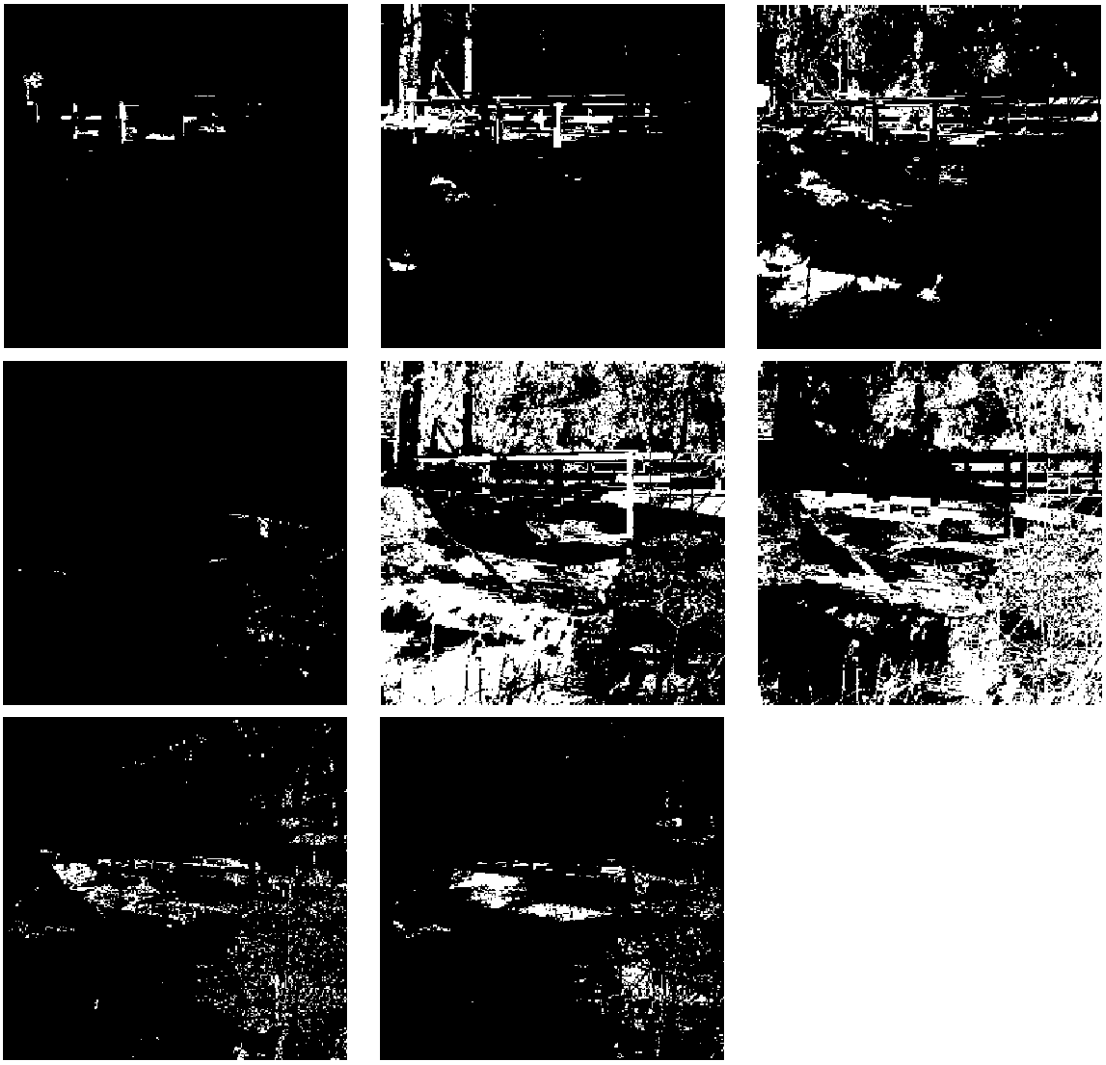


Figura 7.24: Visualización de los componentes de la imagen del Puente

El valor de $\alpha = 0.575$, por lo que el porcentaje de masa es del 99.8 % y el número promedio de mesas del restaurante chino es de 7.3169.

Las imágenes mostradas en la figura (7.24) son las distintos matizes de la imagen original, cada componente simboliza un determinado tono de gris o blanco o negro y como habíamos mencionado encuentra muy bien los tonos suaves o bajos que es lo que ocasiona las montañas en el histograma.

Al final se presenta la pre-segmentación y la imagen original (figura (7.25)) el cuál comparamos y se ve bastante bien la agrupación de los distintos matizes de la imagen original a escala de gris.

Con respecto a la pre-segmentación puede existir un componentes que este vacío esto se debe a que al momento de encontrar todas las muestras, cada una tiene las etiquetas respectivas de la imagen sin embargo al realizar en este caso las 500 muestras entonces se

(a) (b)

Figura 7.25: (a) Imagen Segmentada (b) Imagen Original, ambos de la imagen del puente

tienen 500 etiquetas por cada pixel por lo que se opta es tomar de todas ellas la más frecuente, sin embargo suele ocurrir que se pierda componentes ya que existen puntos o pixeles que no tienen bien definido que etiqueta le pertenece dicho en otras palabras son los pixeles que se comparten gaussianas y que en ciertas muestran pertenecen a una determinada pero en otra muestra pertenece a otra gaussiana y al final como se toma la frecuencia, resulta que se inclina mas por una y la otra etiqueta se pierde por lo que el algoritmo nos está indicando que del número encontrado se puede segmentar con uno menos en este caso, esto puede verse gráficamente ya que el componente que hace falta es la gaussiana más pequeña en la figura (7.21) y podemos considerar a esta como un dato espurio.

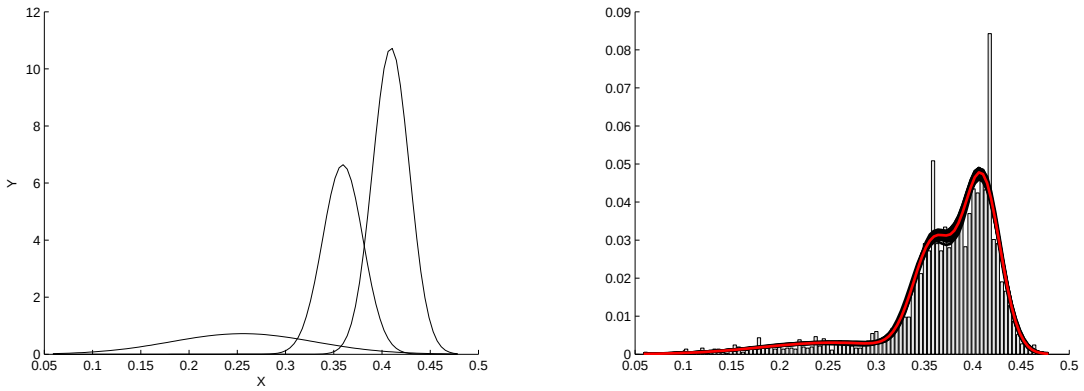
7.4. Experimentos con Imágenes Cerebrales

Esta es la aplicación más importante que queríamos realizar desde el inicio de esta tesis, ya que la segmentación en imágenes del tipo MRI (Magnetic Resonance Imaging) no son fáciles de realizar, y queríamos realizar un algoritmo para la pre-segmentación de este tipo de imágenes, para eso utilizamos tres cortes de cerebro que son imágenes del tipo T1, que la imagen original de todos ellos son de 881x881 o sea muy grandes y como ya habíamos visto sus histogramas tienen mucho detalle por lo que nuestro algoritmo encontrará mas componentes de los 3 componentes que deben ser ya que en el cerebro solo se encuentran 3 componentes que son: materia blanca, materia gris y líquido encefaloraquidio.

Ahora como habíamos mencionado previamente lo que queremos es encontrar estos elementos contenidos en el cerebro por lo que no necesitamos detalle en el histograma, entonces para quitar este detalle vamos a reducir la imagen en una imagen pequeña y que

Algoritmo 7.2 Algoritmo de pre-segmentación de imágenes cerebrales MRI del tipo T1.

1: Reducir el tamaño de la imagen original en una más pequeña de tal manera que el histograma pierda detalle, pero sin perder los limites de los valores de la imagen original.
2: Procesar la imagen con el algoritmo de mezcla infinita de gaussianas.
3: Realizar la pre-segmentación de acuerdo al algoritmo(7.1).
3: Encontrar los umbrales de las gaussianas encontradas que serán las intersecciones de las gaussianas que se encuentran juntas.
4: Con estos umbrales segmentar imágenes más grandes, tomando cada umbral como referencia para etiquetar los pixeles.



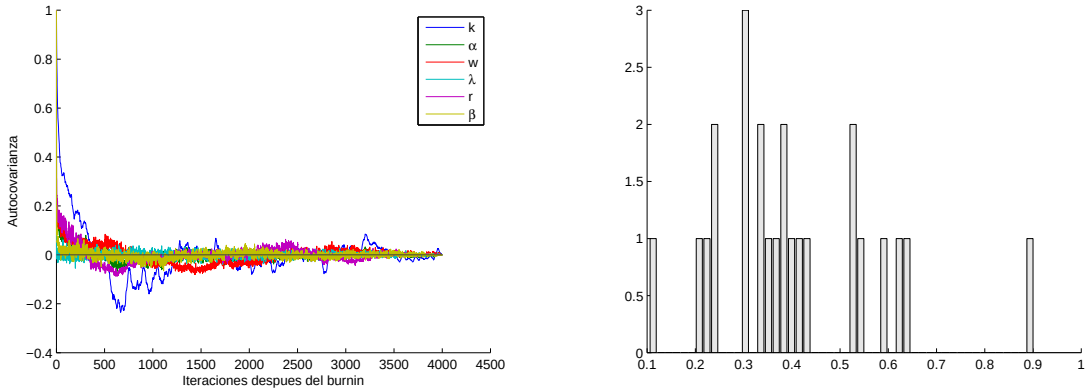
(a) (b)

Figura 7.26: (a) Gaussianas encontradas con el modelo de mezclas infinitas (b) Ajuste de la densidad encontrada por el modelo de mezclas infinitas con el histograma de los datos, los coloreados en negro son todas las muestras de las densidades obtenidas y el coloreado en rojo es la media de ellas, ambos para la segunda imagen sintética.

el histograma quede más suave, con esa imagen la procesaremos y encontraremos ciertos umbrales que nos indicarán cuales serán los intervalos que pertenecerán a los distintos componentes, el algoritmo para la pre-segmentación puede verse en (7.2).

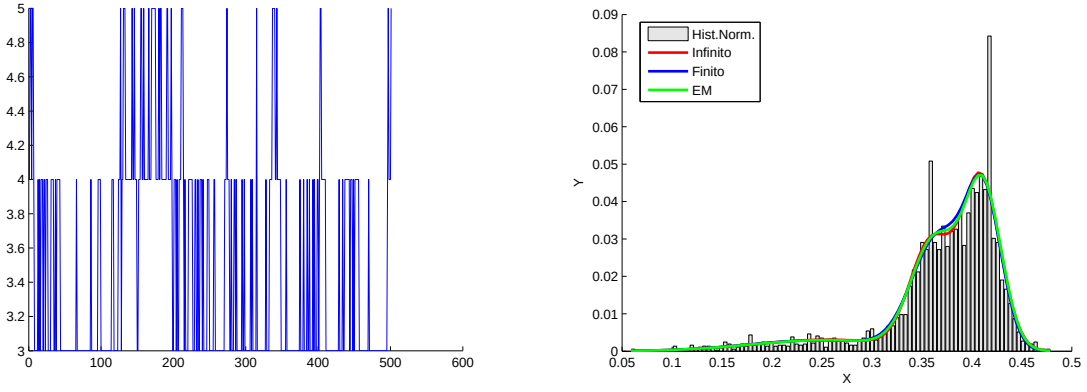
De acuerdo al algoritmo se redujo la imagen original de 881x881 a una imagen de 100x100, se proceso con el algoritmo, se obtuvieron los umbrales (que son la intersección de las gaussianas) y posteriormente segmentamos de acuerdo a dichos umbrales imágenes del doble y del original.

Todas las pruebas ser realizaron con 5000 iteraciones, desechando las primeras 1,000 iteraciones y tomando muestras en intervalos de 8 iteraciones de desfase para obtener 500 muestras.



(a) (b)

Figura 7.27: (a) Gráfica de la autocovarianza de los hiperparámetros (b) Histograma del hiperparámetro α, ambos para la segunda imagen sintética.



(a) (b)

Figura 7.28: (a) Gráfica de los componentes después de desechar las primeras 1,000 iteraciones (b) Comparación de las densidades con los tres métodos, ambos para el primer corte cerebral.

7.4.0.1. Primer Corte del Cerebro.

Para este primer corte podemos ver que en la figura (7.26)(a) que encuentra muy bien tres gaussianas esto es que encontró 3 componentes como puede verse en la figura(7.28)(a), donde el valor de k oscila entre 3, 4 y 5 pero el más frecuente resultó ser el valor de 3, las muestras que se tomaron ajustaron muy bien la densidad y la media de ellas se puede ver en la figura (7.26)(b) que es la que tomamos como resultado final.

Para el valor más frecuente de $\alpha = 0.3$ por lo que el porcentaje de masa es del 99.9 % y el número de mesas esperadas es de 4.66.

Las comparaciones con los otros métodos se puede ver en la figura(7.28)(b) donde se puede ver que los tres métodos ajustan bien y están casi encimados por lo que la densidad

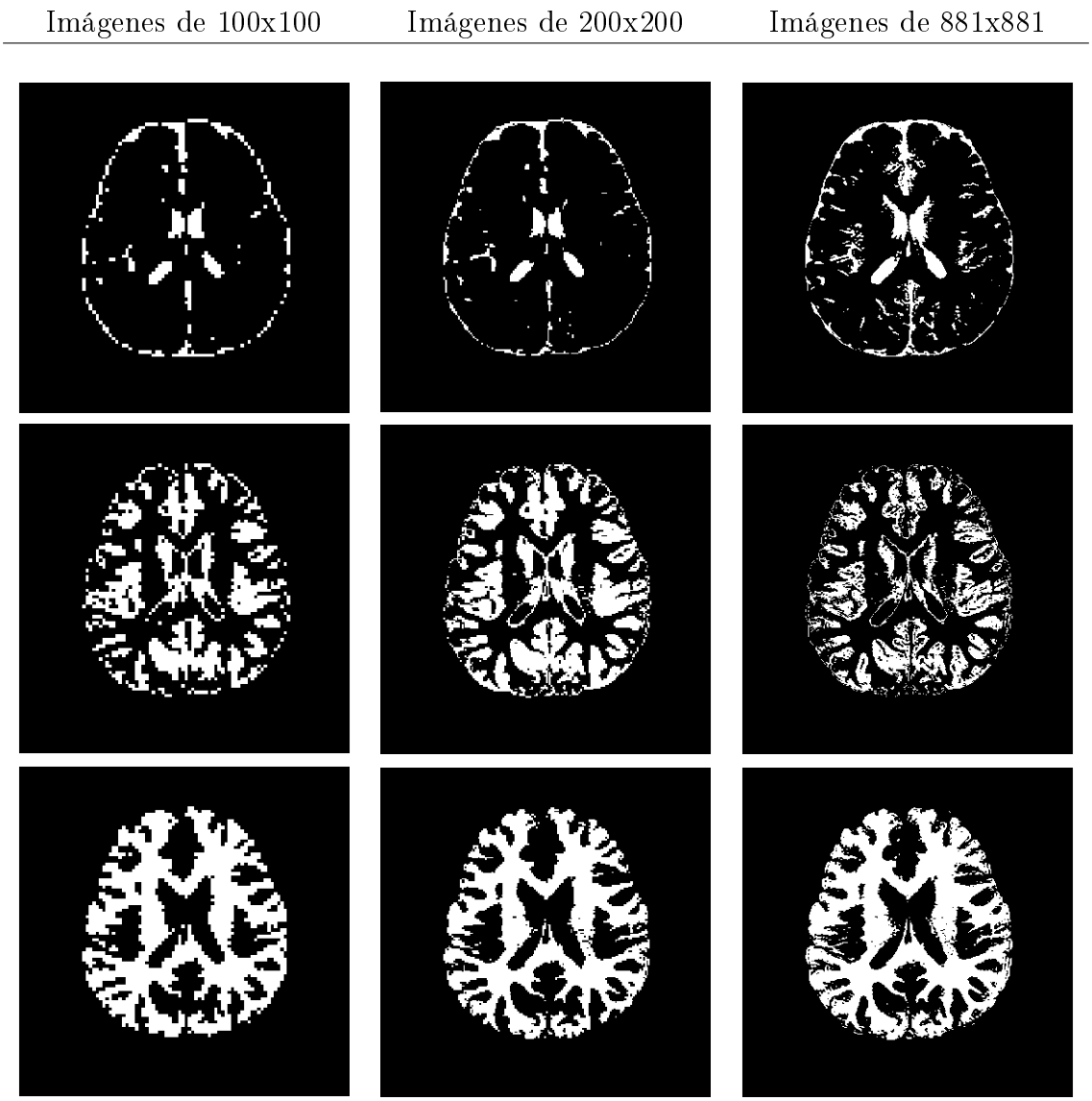


Figura 7.29: Visualización de los componentes para el primer corte cerebral con los distintos tamaños, en la parte de arriba se encuentra el líquido encefaloraquidio, en la parte de enmedio se encuentra la materia gris y en la parte de abajo está la materia blanca.

encontrada no tiene ningún problema.

Después de procesar la imagen de 100x100 con la misma escala que la original se encontraron los siguientes umbrales:

- Para el componente 1: [0, 0.34)
- Para el componente 2: [0.34, 0.3809)
- Para el componente 3: [0.3809, 0.5]

Los componentes encontrados pueden verse en la figura (7.28). Posteriormente se procedió a realizar la pre-segmentación con el doble de tamaño y con la original, esto se realizó

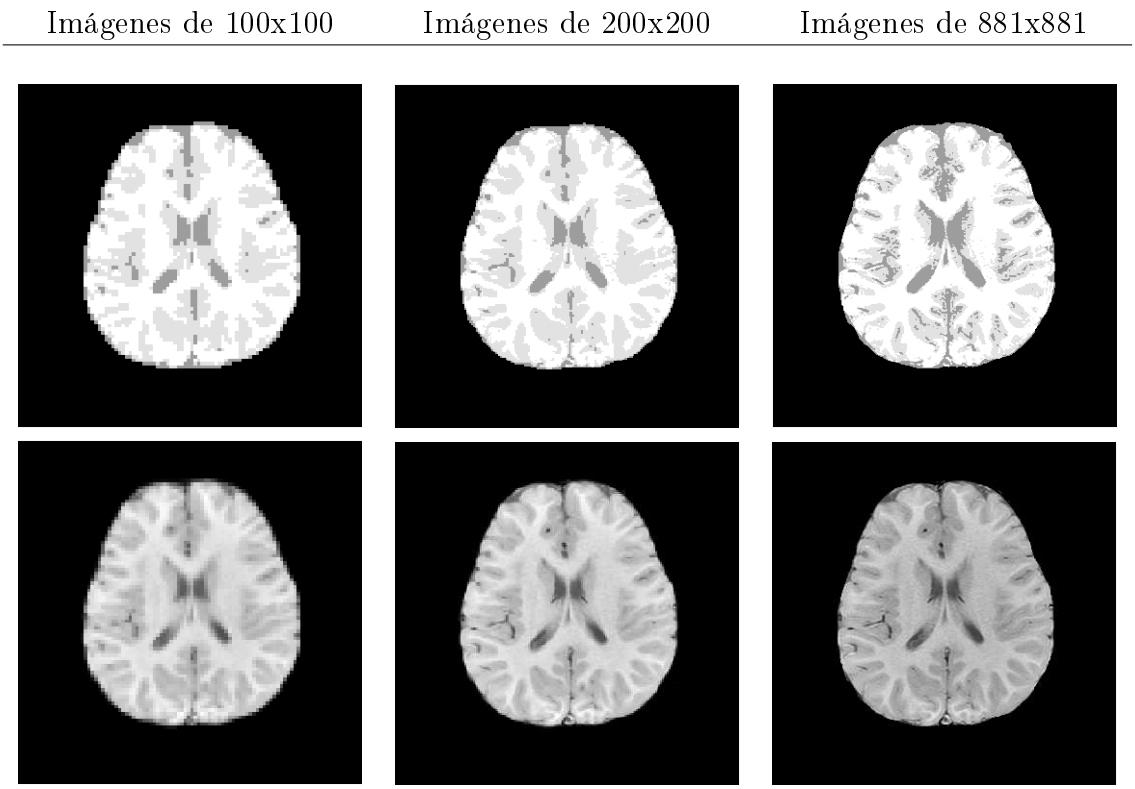


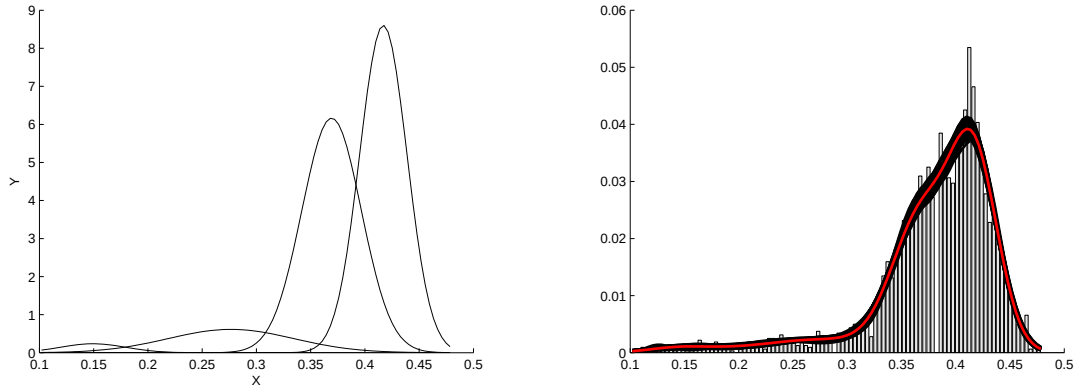
Figura 7.30: En la parte de arriba se encuentran las imágenes segmentadas del corte 1 y en la parte de abajo se encuentran las imágenes originales del mismo corte cerebral.

etiquetando dichas imágenes en los pixeles que pertenezcan a los umbrales mencionados previamente, es decir en las imágenes más grandes todos lo valores de los pixeles que están entre [0,0.34) se le dio la etiqueta 1, a los pixeles que tienen valores entre [0.34 y 0.3809) se le dio la etiqueta 2 y así sucesivamente, y por lo tanto los resultados pueden verse en las figuras (7.29) y la pre-segmentación resultante puede verse en la figura (7.30).

Ahora como puede notar la pre-segmentación realizada a la imagen pequeña se hizo muy bien, que al pasar a la imagen que la duplica el resultado es muy similar pero al tratar de segmentar la imagen original de 881x881 se pierde cierto detalle, esto es debido a que al minimizar la imagen los pequeños detalles de los puntos cercanos a los umbrales no serian los mismos para la imagen agrandado es allí donde no se puede tener un control cuando intentamos tener los mismos umbrales para imágenes más grandes.

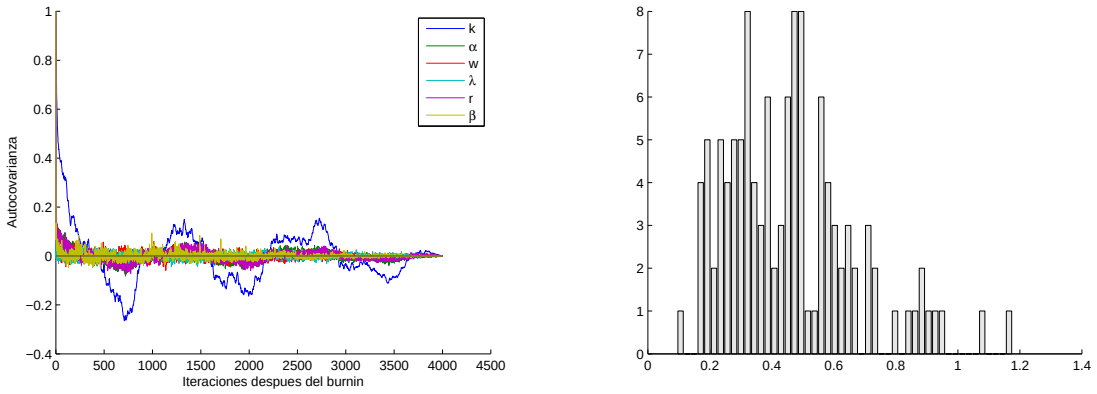
7.4.0.2. Segundo Corte del Cerebro.

Para el segundo corte de cerebro el valor del número de componentes oscila entre 4 y 5 pero el más frecuente resultó ser el 4 como puede verse en la figura (7.33)(a) por lo que se encontraron 4 gaussianas que aparecen en la figura (7.31)(a) y los muestras tomadas y la



(a) (b)

Figura 7.31: (a) Gaussianas encontradas con el modelo de mezclas infinitas (b) Ajuste de la densidad encontrada por el modelo de mezclas infinitas con el histograma de los datos, los coloreados en negro son todas las muestras de las densidades obtenidas y el coloreado en rojo es la media de ellas, ambos para el segundo corte cerebral.



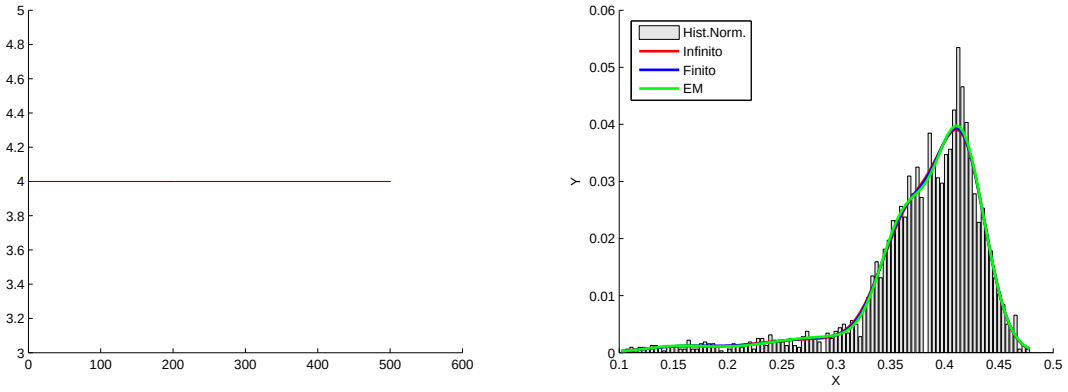
(a) (b)

Figura 7.32: (a) Gráfica de la autocovarianza de los hiperparámetros (b) Histograma del hiperparámetro α, ambos para la segunda imagen cerebral.

media de ellas aparece en la misma figura, como se puede observar la densidad se encuentra muy bien estimada con estos 4 componentes y la comparación de ellos se puede ver en la figura (7.33)(b) donde se aprecia que los tres métodos ajustan bien la densidad sin embargo el nuestro es el que se encarga de saber cuantos son necesarios.

Para el valor de alfa mas frecuente resulta ser en este caso el mayor de los frecuentes este es $\alpha = 0.49$ como puede verse en la figura (7.32) y el porcentaje de masa es del 99.9 % y el número de mesas esperadas es 5.33 por lo que al encontrar 4 no está tan mal de la media calculada.

Después de procesar la imagen de 100x100 con la misma escala que la original se encontraron los siguientes umbrales:



(a) (b)

Figura 7.33: (a) Gráfica de los componentes después de desechar las primeras 1,000 iteraciones (b) Comparación de las densidades con los tres métodos, ambos para el segundo corte cerebral.

- Para el componente 1: [0, 0.178)
- Para el componente 2: [0.178 0.3073)
- Para el componente 3: [0.3073, 0.391)
- Para el componente 4: [0.391,0.5]

Los componentes de las imágenes cerebrales se pueden ver en la figura (7.34) allí se pueden apreciar los 3 componentes del cerebro que son la materia blanca, la materia gris, el líquido encefaloraquidio y un componente más que es considerado como la mezcla que existe entre ellos y que lo consideraremos como espurio. Al igual que el corte anterior se puede ver que la pre-segmentación para la imagen de 200x200 es muy buena y para la imagen original no lo está tanto ya que se pierde en los bordes de los umbrales encontrados como habíamos mencionado antes, lo que podemos observar es que este corte es similar al corte anterior con la diferencia que contiene más materia gris en la parte del centro que era lo que pretendíamos saber si lo encuentra y lo realizó muy bien.

La pre-segmentación final puede verse en la figura (7.35).

Esta aplicación se realizó con la finalidad de encontrar una pre-segmentación de imágenes cerebrales (MRI) del tipo T1 que pueda servir a otros algoritmos de segmentación como punto inicial o usarlo directamente, la desventaja que tiene este método es que se tiene que reducir la imagen original y al segmentar dicha imagen puede perder información por la reducción.

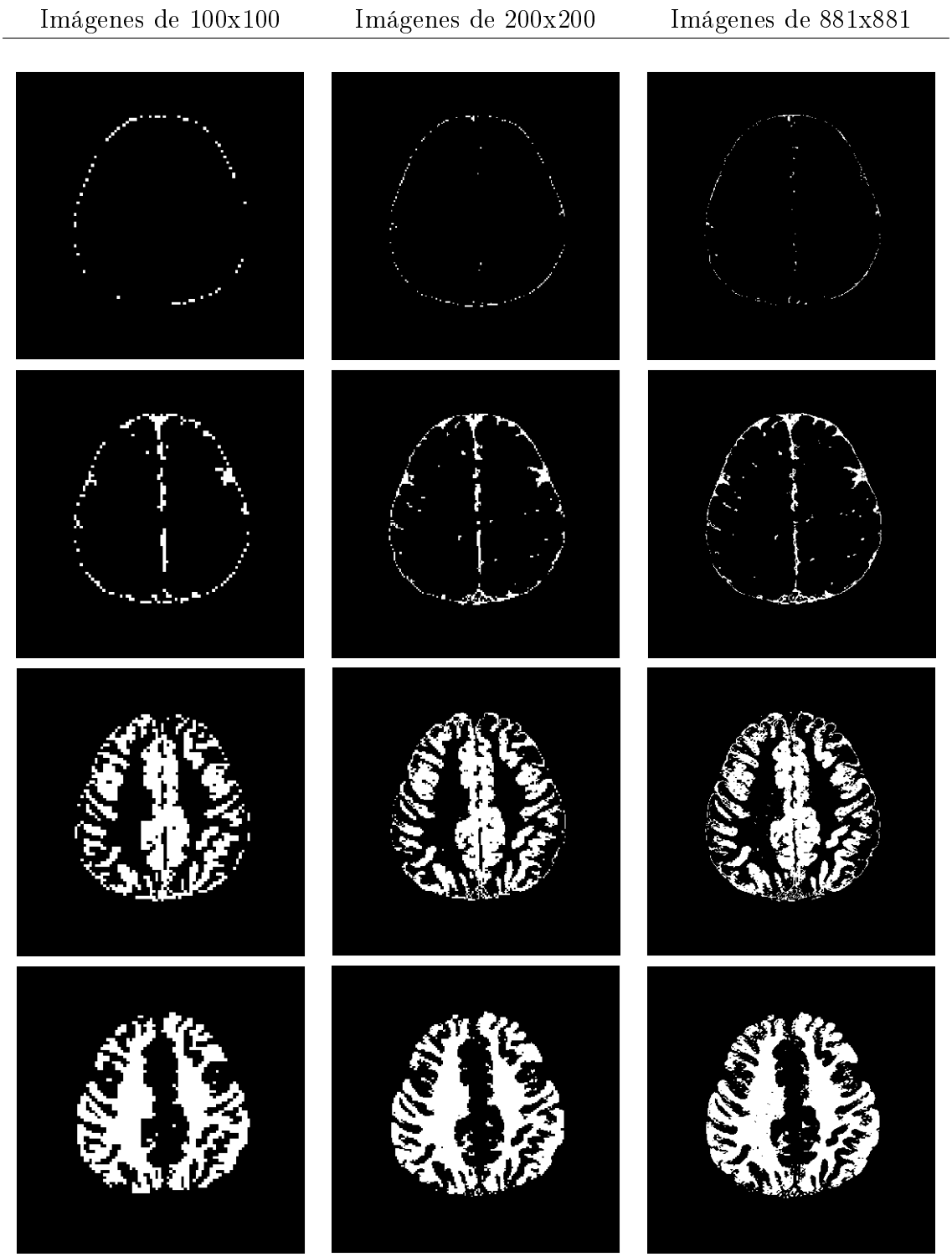


Figura 7.34: Visualización de los componentes para el segundo corte cerebral con los distintos tamaños, en la parte de arriba se encuentra el tejido mezclado, posteriormente el líquido encefaloraquidio, luego la materia gris y por último en la parte de abajo la materia blanca.

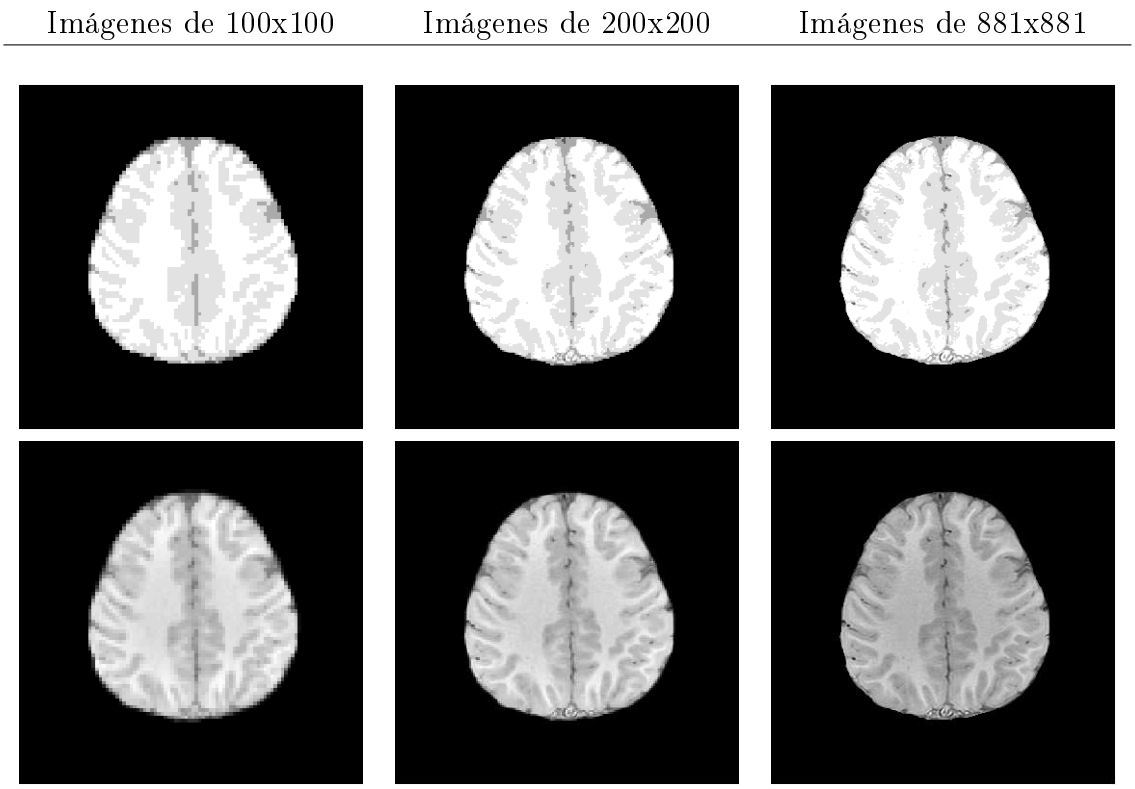


Figura 7.35: En la parte de arriba se encuentran las imágenes segmentadas del segundo corte cerebral y en la parte de abajo se encuentran las imágenes originales del mismo corte cerebral.

7.4.1. Experimento con Datos en dos Dimensiones

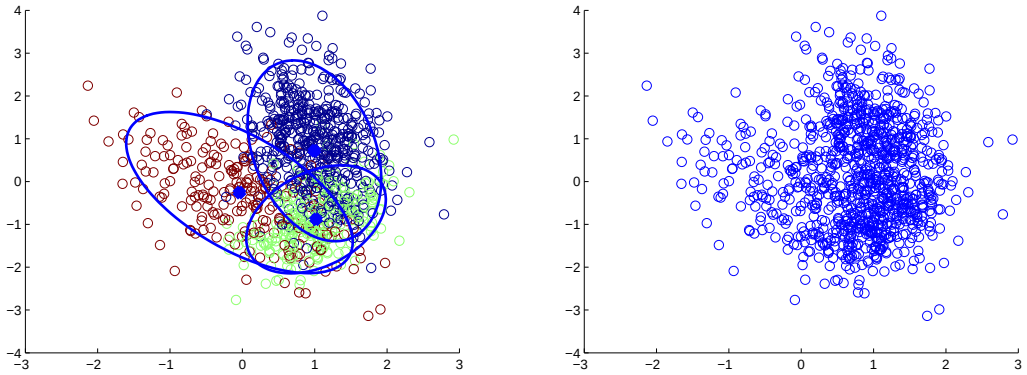


Figura 7.36: En la gráfica de la izquierda se muestra la figura de la clasificación indicando en colores y encerrando con elipses los datos que pertenecen a una determinada gaussiana y en la gráfica de la derecha se muestra la figura original de los datos totalmente mezclados.

Se puede ver de la figura(7.36) de la parte derecha, que este conjunto de datos es un conjunto bastante complicado para los algoritmos, ya que los datos de las gaussianas se encuentran bastante mezclados y básicamente a mano sería muy difícil clasificarlos. El pun-

Infinito Multivariado	Finito Multivariado	EM
$\hat{\mu} = \begin{pmatrix} 0.2729 & -0.1964 \\ 0.9957 & 1.1189 \\ 1.0642 & -0.8968 \end{pmatrix}$	$\hat{\mu} = \begin{pmatrix} -0.0034 & -0.1937 \\ 1.0196 & 1.0504 \\ 1.1393 & -0.8164 \end{pmatrix}$	$\hat{\mu} = \begin{pmatrix} 0.1681 & -0.0717 \\ 0.9822 & 1.1188 \\ 1.0288 & -0.9384 \end{pmatrix}$
$\hat{\Sigma}_1 = \begin{pmatrix} 0.8551 & -0.2520 \\ -0.2520 & 0.8052 \end{pmatrix}$	$\hat{\Sigma}_1 = \begin{pmatrix} 0.7146 & -0.2636 \\ -0.2636 & 0.8224 \end{pmatrix}$	$\hat{\Sigma}_1 = \begin{pmatrix} 0.8037 & -0.1810 \\ -0.1810 & 0.7712 \end{pmatrix}$
$\hat{\Sigma}_2 = \begin{pmatrix} 0.2026 & -0.1286 \\ -0.1286 & 0.9371 \end{pmatrix}$	$\hat{\Sigma}_2 = \begin{pmatrix} 0.2234 & -0.1032 \\ -0.1032 & 0.8529 \end{pmatrix}$	$\hat{\Sigma}_2 = \begin{pmatrix} 0.1975 & -0.1215 \\ -0.1215 & 0.9118 \end{pmatrix}$
$\hat{\Sigma}_3 = \begin{pmatrix} 0.1871 & 0.1508 \\ 0.1508 & 0.5024 \end{pmatrix}$	$\hat{\Sigma}_3 = \begin{pmatrix} 0.2387 & 0.1306 \\ 0.1306 & 0.3864 \end{pmatrix}$	$\hat{\Sigma}_3 = \begin{pmatrix} 0.2074 & 0.1521 \\ 0.1521 & 0.4997 \end{pmatrix}$
$\hat{\pi} = \begin{pmatrix} 0.3225 \\ 0.4500 \\ 0.2275 \end{pmatrix}$	$\hat{\pi} = \begin{pmatrix} 0.4760 \\ 0.2500 \\ 0.2740 \end{pmatrix}$	$\hat{\pi} = \begin{pmatrix} 0.2996 \\ 0.4348 \\ 0.2656 \end{pmatrix}$

Cuadro 7.5: Comparación del algoritmo infinito y finito con el algoritmo EM

to azul indica donde se encuentra la media y la elipse indica donde se encuentra la mayor parte de los datos de acuerdo a su matriz de covarianza.

Los datos fueron creados con los siguientes parámetro

$$\mu = \begin{pmatrix} 0 & -0.1 \\ 1 & 1 \\ 1 & -1 \end{pmatrix}$$

$$\Sigma = \left[\begin{pmatrix} 0.625 & -0.2165 \\ -0.2165 & 0.875 \end{pmatrix}, \begin{pmatrix} 0.2241 & -0.1368 \\ -0.1368 & 0.9759 \end{pmatrix}, \begin{pmatrix} 0.2375 & 0.1516 \\ 0.1516 & 0.4125 \end{pmatrix} \right]$$

$$\pi = \begin{bmatrix} 0.5 & 0.25 & 0.25 \end{bmatrix}$$

Ejecutamos los algoritmos que para el caso del nuestro infinito lo corrimos con 10,000 iteraciones, obtenemos los siguientes resultados

Lo que podemos observar de este conjunto de datos es que el parámetro mas difícil de estimar es el $[0, -0.1]$ ya que sus varianzas son muy grandes (0.625 y 0.875) en comparación de las medias, esto genera un error muy grande en el cálculo de las medias y posteriormente en el de las matrices de covarianzas. Para los otros parámetros los algoritmos fueron más certeros y no tuvieron problemas al encontrar una buena estimación.

Para el cálculo de las matrices de covarianzas vemos que hay ciertos errores esto es debido a que los datos están bastante mezclados y ocasiona que se generen errores numéricos al intentar encontrar las matrices de covarianza y como podemos observar en los datos, nuestra

aproximación es muy buena e inclusive en varias ocasiones mejora la precisión del EM y nuestro algoritmo es mejor porque no tiene el problema del punto inicial que tiene el EM.

7.4.2. Experimento con Datos en Forma de Espiral

Se generaron datos aleatorios para formar una espiral en 3 dimensiones con 6 componentes y el resultado fue el siguiente

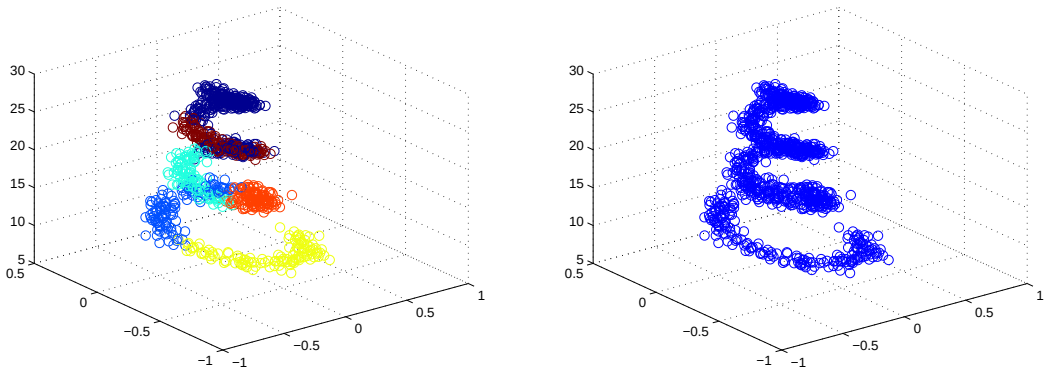


Figura 7.37: En la gráfica de la izquierda se muestra la figura de la espiral indicando en colores los datos que pertenecen a una determinada gaussiana y en la gráfica de la derecha se muestra la figura original de los datos en 3 dimensiones.

En este tipo de datos no es fácil obtener el número indicado de componentes por lo que nuestro algoritmo nos puede dar una idea de cuantos componentes seria adecuado para realizar una clasificación o por lo menos nos otorga una cota superior del número de componentes.

En la gráfica (7.37) el proceso encuentra 6 componentes con 10,000 iteraciones para la espiral y se puede deducir que es apropiado tomando en cuenta como clasificó los puntos al ver que forman una determinada gaussiana los puntos cercanos y no puntos muy distantes como puede ser los puntos de la punta de la espiral con los puntos de abajo, y esto nos da un buen indicador de que el ajuste es correcto. Encontrar los componentes en 3 dimensiones no es nada trivial.

Discusión de los Resultados

En la gráficas de la autocovarianza lo que se pretende visualizar es la tendencia a cero que nos indican que las muestras de los hiperparámetros sean muestras independientes e identicamente distribuidas(i.i.d.), la literatura menciona que la tendencia a cero indica una buena independencia y si se encuentra entre el intervalo -0.2 y 0.2 se puede considerar buena en otros caso no. En muchas gráficas se pudo notar que el hiperparámetro k varia u oscila al inicio en dicho intervalo esto es porque este valor es muy cambiante en el modelo, pero logra encontrar la estabilidad posteriormente.

Como se pudo observar la pre-segmentación realizada para las distintas imágenes sean cerebrales o no, se comporta muy bien y nos proporciona una buena aproximación, decimos que es buena porque este algoritmo no es un algoritmo de segmentación sin embargo nos proporciona una buena aproximación en donde talvez otros algoritmos les pueda servir como dato inicial, sin embargo el método encuentra muy bien la densidad de los datos y realiza un buen ajuste por encima del EM que es un algoritmo bastante clásico.

Como se vio, para realizar una pre-segmentación cerebral en donde la imagen es grande y contiene mucho detalle es necesario suavizarla para realizar una buena aproximación a la segmentación, esto se puede aplicar a cualquier imagen, sin embargo no es recomendable ya que si la imagen es grande y se disminuye mucho, la pre-segmentación tiende a perderse en la imagen original ya que esos detalles que se perdieron al minimizar la imagen no se verán reflejado y la pre-segmentación tiende a fallar.

Las imágenes cerebrales (MRI) del tipo T2 se sometieron a pruebas con este algoritmo pero el resultado no fue bueno ya que este tipo de imágenes tienes otros tonos de grises y por lo tanto no resultaron útiles para la pre-segmentación.

Podemos concluir que nuestro algoritmo funciona bien para encontrar las densidades de un conjunto de datos sea imagen o no, y es posible hacerlo para datos univariados como para datos multivariados, no era la intensión abarcar demasiado la parte multivariada sino simplemente se realizó la implementación y se hicieron ciertas pruebas sencillas para saber si se comporta de buena manera.

Uno de los inconvenientes para la parte multivariada es que en lugar de trabajar con la distribución gamma, se trabaja con la distribución wishart debido a que es la generalización de la gamma cuando los grados de libertad no necesariamente son enteros, y como nosotros trabajamos con datos reales entonces esta distribución resulta bastante bien pero la matriz que recibe debe ser positiva definida por lo que en ocasiones esto no sucede y se tiene que realizar cálculos numéricos para hacer que una matriz sea positiva definida y eso hace más tardado el proceso y en ocasiones la verdadera matriz se va perdiendo y no nos proporciona una buena estimación.

La mayoría de los experimentos se realizaron con 10,000 iteraciones para encontrar los parámetros dentro de la cadena de Markov, esto fue porque se encontró que con ese número de iteraciones resultó bastante adecuado, ya que se probó con mas iteraciones pero el resultado fue muy similar encontrando siempre el mismo número de componentes por lo que deducimos que la cadena ya había convergido, las pruebas se hicieron con 20,000 y 30 000 del cuál observamos que en muchos casos con 10,000 era suficiente para asegurar una posible convergencia.

También se realizaron pruebas con imágenes homogéneas pero los resultados no fueron satisfactorios porque la misma homogeneidad en los datos hacía que la pre-segmentación produzca mucho ruido debido a que encuentra muchos componentes con varianza iguales a cero y por lo tanto la densidad sufría altibajos, por lo que deducimos que nuestro proceso no es adecuado para este tipo de imágenes.

Para las imágenes con ruido el algoritmo encuentra el ruido contenido en la imagen esto es que dependiendo lo que se requiera hacer al final se puede decir si es adecuado o no,

ya que si la intensión es realizar una pre-segmentación entonces no eliminara los puntos ruidosos pero si lo encontrará y lo clasificará de acuerdo a una determinada gaussiana, esto puede ser benéfico para algunos pero puede ser perjudicial para otros, es por eso que se lo dejamos al lector que tan conveniente es procesar imágenes con el proceso de mezcla finita con imágenes con ruido.

También no es muy adecuado a imágenes que solo tienen número enteros o son constantes como etiquetas, ya que nuestro algoritmo es continuo y le cuesta trabajo aproximar valores enteros.

También se notó que en la pre-segmentación de imágenes para los tonos muy parecidos al blanco o al negro no tiene problemas en agruparlos como blanco o negro pero si le cuesta más trabajo clasificar a los distintos tonos de gris, es por eso que tiende a confundir ciertos datos en los distintos tonos de gris ya que cada tono de gris la diferencia con otro es muy pequeña y los datos en ocasiones se encuentra enmedio de esos valores por lo que confunde en realidad a cuál tono pertenece y es cuando se puede confundir del verdadero tono que le pertenece a cierto dato.

La principal desventaja de este algoritmo es el número adecuado de iteraciones que se tienen que hacer para saber que la cadena de Markov haya llegado a su distribución estacionaria, lo que se hizo en este caso fue realizar varias corridas y ver que tan parecido son los resultados se realizaron pruebas de 5,000, 10,000 y 20,000 iteraciones y se notó en ocasiones que cuando no había una cierta convergencia la densidad hallada no ajustaba bien los datos por lo que se proseguía a incrementar el número de iteraciones.

Nuestro algoritmo es en general un algoritmo de estimación de densidad y los resultados obtenidos fueron bastante buenos, encontrando en el proceso el número de componentes adecuados que es lo en otros algoritmos se tiene que hacer un estudio aparte o una selección de modelo y eso lleva tiempo computacionalmente, el principal objetivo de este proceso y de la tesis fue hallar ese número que pueda servir como base en estudios posteriores. En general el proceso de mezcla infinita de gaussianas nos proporciona una cota superior del número de componentes adecuados para los datos, por lo que puede ser que con otros procesos o

técnicas se mejore este número de componentes, como es el caso de la segmentación que existen diversas técnicas para poder quitar el ruido u otras técnicas de post-proceso para mejorar el resultado.

Capítulo 8

Conclusiones y Trabajo a Futuro

El principal objetivo de esta tesis fue aprender todo lo relacionado con el modelo de mezcla finita e infinita con hiperparámetros, comprender bien como se derivan todas las ecuaciones y como se produce el proceso de Dirichlet entre ellas el proceso del restaurante chino. Posteriormente se realizó la aplicación con imágenes cerebrales y con datos multidimensionales.

El modelo de mezcla bayesiano infinito ha sido mostrado con un buen rendimiento (evitando el sobreajuste) y puede ser logrado también con datos multidimensionales. Un algoritmo eficiente y practico como es el MCMC ha sido mostrado en este método, el método es automático sin necesidad de especificar los parámetros sino únicamente proporcionar el número de iteraciones, que tal vez pueda decirse que es su punto débil pero en otros algoritmos además de el número de iteraciones tienes que ajustarle otros parámetros.

Muchas pruebas en la variedad de problemas revelan que la mezcla infinita produce densidades cuya generalización es muy competitiva con otros métodos comunes.

El trabajo actual se llevo a cabo para explorar el rendimiento en problemas de búsqueda de las densidades en imágenes que por el mismo hecho de ser imágenes se cuenta con información pesada por lo que probamos la eficiencia computacional y la generalización.

El modelo de mezcla infinita tiene varias ventajas con su contraparte el modelo de mezcla finita:

- En muchas aplicaciones, puede ser más apropiado no limitar el número de clases.
- El número de clases se determina automáticamente
- El uso del MCMC evita eficazmente mínimos locales que muchos otros algoritmos se quedan atorados en ello o dependen de su inicialización como por ejemplo el EM(Nakano et al. 1999).
- Es mucho más sencillo de manejar el límite infinito que trabajar con modelos finitos con tamaños desconocidos como en (Richardson and Green 1997b) o los enfoques tradicionales sobre la base de la validación cruzada.

El modelo de mezcla infinito bayesiano resuelve de manera simultanea varios problemas de grandes cantidades de datos y de la estimación o inferencia de la densidad.

Se realizaron varios experimentos para validar la implentación del modelo, entre ellos se implementó el algoritmo de pre-segmentación como base para futuros algoritmos de segmentación de imágenes que tomen en cuenta la correlación espacial y le sirva como punto de partida.

Mi contribución es haber aplicado el modelo de mezcla infinita de gaussianas para segmentar imágenes medicas y naturales. Los resultados obtenidos son razonables pero las limitaciones de tiempo no nos permitieron realizar un estudio mas exhaustivo del algoritmo.

Finalmente se desarolló una herramienta de procesamiento y visualización de los resultados obtenidos por el proceso de mezcla finita e infinita de gaussianas el cuál el usuario puede configurar sobre como quiere hacer el procesamiento y cuál sea su visualización, el cuál su manual de usuario se encuentra en el **apéndice A**.

8.1. Trabajo a Futuro

El trabajo a futuro inmediato sería encontrar una manera de como encontrar convergencia en la cadena de Markov para que no se pierda tiempo y cálculo computacional y se tengan los resultados más precisos, no es una tarea sencilla debido a la forma del proceso

de restaurante chino, pero sería de mucha utilidad contar con eso.

Una de las cuestiones importates es realizar pruebas exhaustivas del modelo que combinen la pre-segmentacion para realizar una mejor validación al modelo, encontrar un algoritmo que tome en cuenta la correlacion espacial en las imágenes y crear un modelo en el que se combine el uso de la correlación espacial y la estimacion no parámetrica del número de componentes.

También el trabajo a futuro de esta tesis sería proponer otros métodos como puede ser un modelo bayesiano jerárquico que puede ser el proceso de Dirichlet jerárquico(Teh et al. [2006]) que pueda servir para otras aplicaciones y también proponer en lugar del restaurante chino otros métodos como el proceso de la franquisia del restaurante chino (CRF) o el proceso del buffet indio (IBP)(Teh et al. [2007]) entre otros, es decir hacer una extensión de este tipo de modelos, a modelos más complicados donde involucran más variables latentes o más variables observadas y explotar toda la gama de procesos jerárquicos.

Apéndice A

Manual de Usuario para el Software Desarrollado sobre el Proceso de Mezcla Finita e Infinita de Gaussianas

En este apéndice se indica la manera de de utilizar las interfaces del software desarrollado en este trabajo de tesis, de tal manera que se lleve a cabo el proceso deseado.

A.1. Plataforma del Software

Se desarrolló la herramienta para el proceso y visualización de las mezclas finitas e infinitas con la opción de que el usuario pueda configurar tanto la manera del proceso como el de la visualización.

El software fue realizado bajo la plataforma de Matlab versión R2009a de 32 bits, conteniendo el toolbox de estadística, fue desarrollado bajo el ambiente windows sin embargo puede correr en cualquier maquina que tenga linux o mac siempre y cuando este instalado el Matlab con el respectivo toolbox.

Para hacer correr con mas rapidez este software se recomienda tener como mínimo

- Memoria ram de 2 GB
- Disco duro de 1 GB
- Procesador de 2 núcleos.

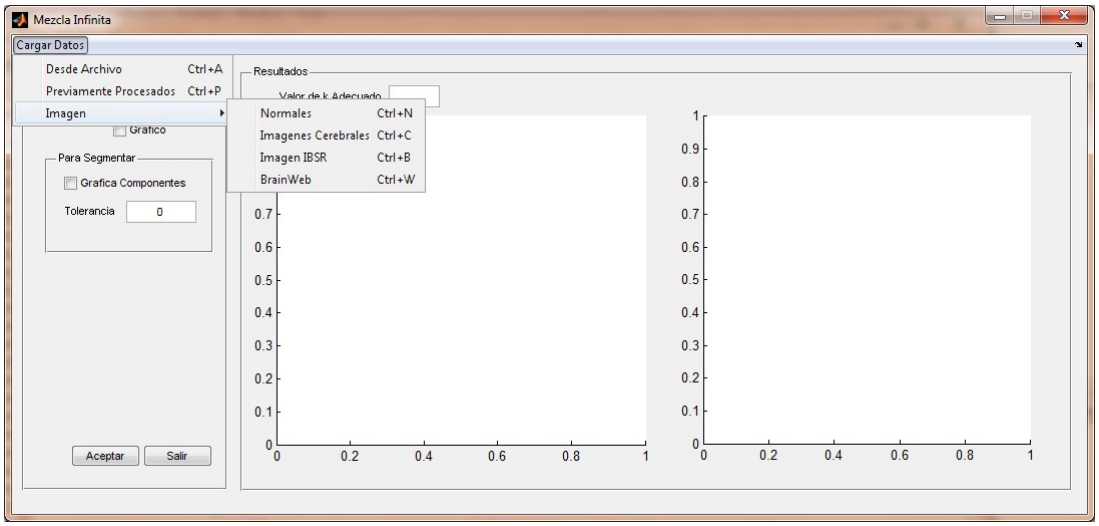


Figura A.1: Menú Principal

A.2. Interfaz del Software de la Mezcla Infinita de Gaussianas

A.2.1. Modo de Cargar Datos al Programa

El software de mezclas infinitas tiene las siguientes características principales:

Contiene un menú que se utiliza para cargar los datos que se van a procesar, como puede verse en la figura (A.1), tal menú contiene las distintas formas que se pueden cargar los datos como son:

- ***Desde archivo:*** Significa que puedes cargar datos en formato de texto(.txt) o en el formato de Matlab(.mat).
- ***Previamente procesados:*** Significa que puedes cargar los datos que antes ya hayas procesado y esto hace que ya no proceses los datos sino que solo te muestre los resultados y puedes configurar con las opciones de la izquierda para obtener los resultados que desees. Es importante mencionar que la manera de hacer esto es mediante la elección del archivo que desees cargar como se puede ver en la figura (A.2) y ese archivo es obtenido cuando procesas los datos y el programa automáticamente lo guarda en un archivo .mat con el mismo nombre que tiene con la diferencia de anteponerle una letra "Y", ejemplo si el archivo se llamaba "cimat.txt" al guardarlo se llamará "Ycimat.mat"
- El submenu imagen tiene las siguientes formas de cargar una imagen:

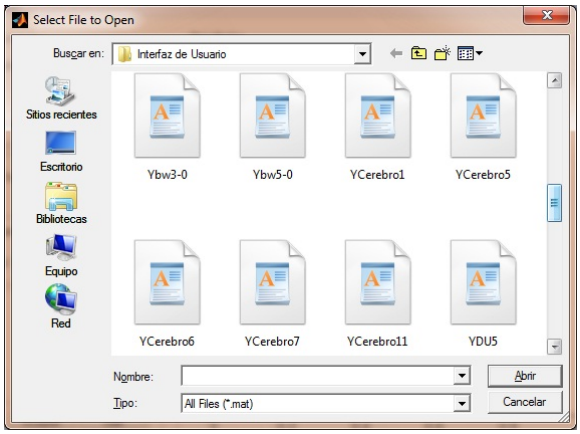


Figura A.2: Ventana para seleccionar un archivo

- ***Normales:*** Significa que puedes cargar cualquier tipo de imagen en los formatos .jpg, .tiff, .bmp, .gif entre otros.

- ***Cerebrales:*** Significa que puedes cargar cualquier imagen cerebral que tenga los mismo formatos que el de normales.

- ***IBSR:*** Significa que puedes cargar imágenes obtenidas de la página web IBSR ya que tienen un formato especial de 512x512 y que además no se considera la parte del fondo por lo que el software lo hace de manera automática.

- ***Brainweb:*** Significa que puedes cargar imágenes obtenidas de la pagina web "brain-web" ya que también tiene un formato especial de imágenes de 256x256 y de igual manera no se considera el fondo además que se realiza una normalización de los datos.

A.2.2. Modo de Configuración del Proceso

Una vez que ya se hayan cargado los datos se procede a configurar el proceso, esto consiste en llenar los campos de la parte izquierda como aparece en la figura (A.3), estos son:

- ***Iteraciones:*** Indica el número de iteraciones para la cadena de Markov el cuál nos indicará los resultados y es por default el valor de 50,000.

- ***Burnin:*** Indica el número desde el cuál se van a considerar las muestras, es decir a partir del burnin se estaria considerando una convergencia y se tomarian muestras a partir de ese número y todo lo anteriro se considera desechado. Por default es 30,000.

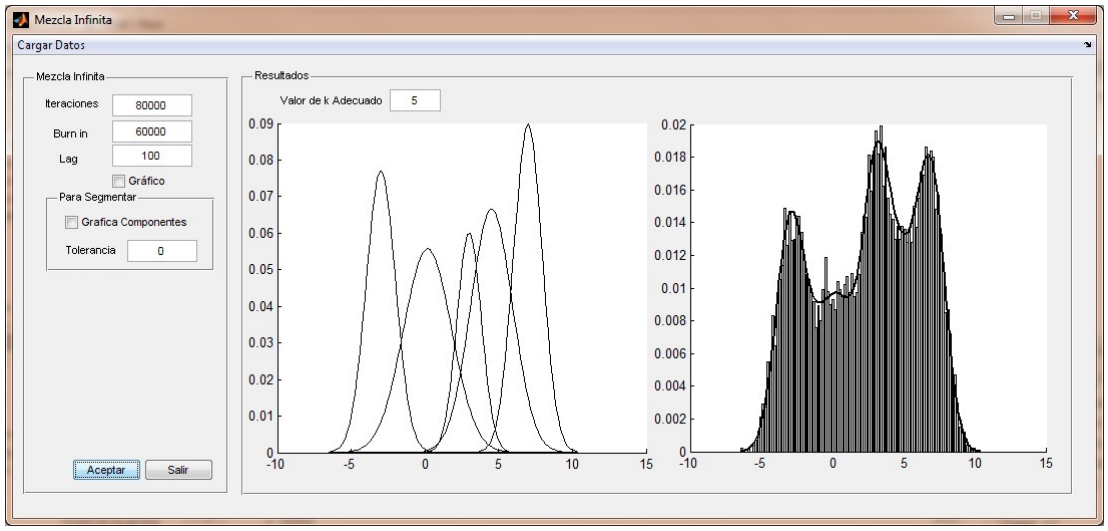


Figura A.3: Visualización de datos univariados

- ***Lag:*** Indica el intervalo por el cuál se toman las muestras, esto quiere decir que a partir del burnin se toman muestras en intervalos (lag) hasta llegar al número de iteraciones. Por default este valor es 100.

- ***Gráfico:*** Significa que para los datos bivariados es posible poder visualizarlo de manera que mientras se está corriendo el programa tu puedas ver como va avanzando y como se va realizando la clasificación que es indicada por colores, cabe mencionar que esto solo se puede realizar para los datos bivariados.

- ***Gráfica componentes:*** Pertenece a la parte de pre-segmentación que consiste en mostrar en pantalla con ventanas emergentes las gráficas de los distintos componentes encontrados, en el caso de las imágenes cerebrales muestra las distintas zonas donde se encontró una determinada gaussiana.

- ***Tolerancia:*** Pertenece a la parte de pre-segmentación y se utiliza para indicar que la varianza es considerada como cero, es decir si la varianza de un componente es 10^{-5} entonces podemos decir que este valor es cero si le ponemos en este caso el valor de 0.05, o sea a partir de el valor de la tolerancia o menos se considerara como varianza cero.

Posteriormente se procede hacer click al botón de aceptar y en ese momento comienza el proceso de la mezcla infinita de gaussianas.

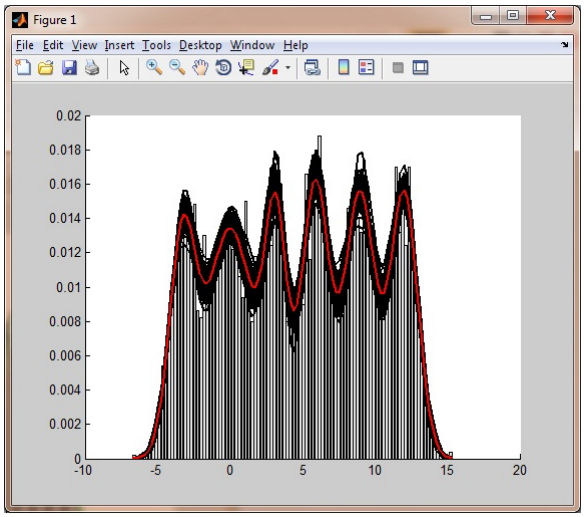


Figura A.4: Ventana Emergente que visualiza las muestras tomadas y la media de ellas.

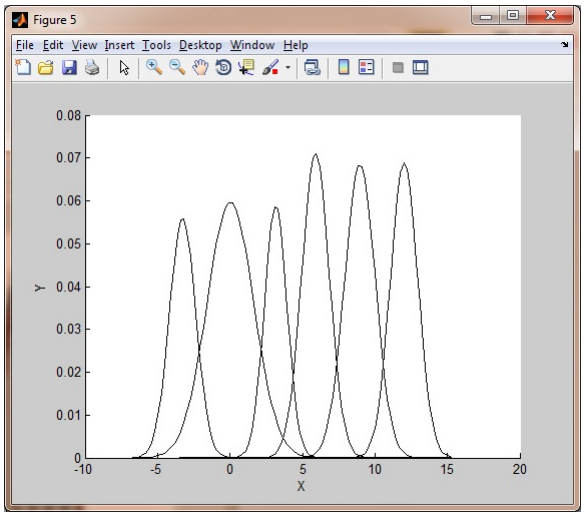


Figura A.5: Ventana Emergente que muestra las gaussianas encontradas.

A.2.2.1. Guardado de Datos

Ahora después de que realize el proceso, automáticamente guarda el archivo en un formato .mat con el nombre anteponiendole una "Y" como había mencionado y también guarda los parámetros obtenidos como son la media, la varianza y los pesos en un formato .txt llamado "Output.txt" pero la forma de guardarlo es mediante un historial es decir si corres dos veces el proceso te guardará los parámetros de esas dos corridas en ese mismo archivo. Para el caso de las imágenes también guarda un archivo binario llamado "prueba" de 256x256.

Nota: Todo lo guardado en el archivo .txt es posible verlo desde la consola de Matlab.

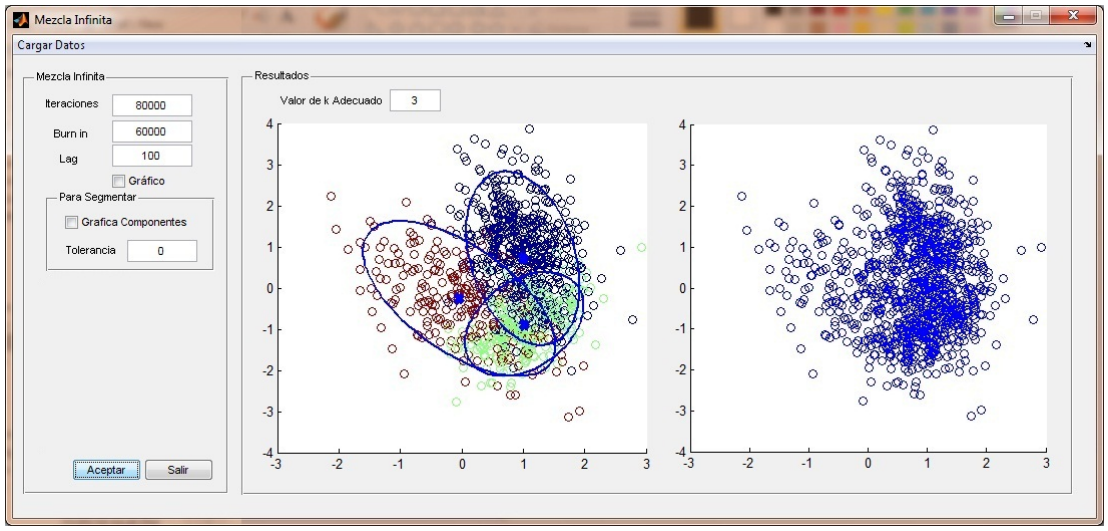


Figura A.6: Visualización de datos Bivariados

A.2.3. Visualización de los Resultados

La visualización de los resultados se caracteriza por mostrar en el apartado "resultados" (que se encuentra en la parte derecha de la ventana principal), lo obtenido por el proceso (como puede verse en la figura (A.3)) .

Primeramente se muestra el número de componentes denominado "k" que encontró el proceso, posteriormente se muestra lo siguiente:

- ***Para los datos univariados*** se visualiza en la parte izquierda las gráficas de las gaussianas encontradas por el método estas son graficadas una por una y en la parte derecha se muestra el histograma normalizado con la estimación de la densidad que genera la mezcla de las gaussianas, como se muestra en la figura (A.3).

- ***Para los datos bivariados*** se visualizan en la parte izquierda los puntos con la clasificación hecha de acuerdo a las gaussianas encontradas y lo pinta de un cierto color para indicar que puntos pertenecen a una determinada gaussiana, también muestra con un punto azul la media y con una elipse los puntos encerrados por su determinada covarianza y en la parte derecha se muestran los puntos originales, como se muestra en la figura (A.6).

- ***Para los datos en 3 dimensiones*** se visualiza en la parte izquierda los datos en 2 dimensiones y en la parte derecha los datos en 3 dimensiones, y de igual manera indicando con colores los que pertenecen a una determinada gaussiana, como se muestra en la figura (A.7).

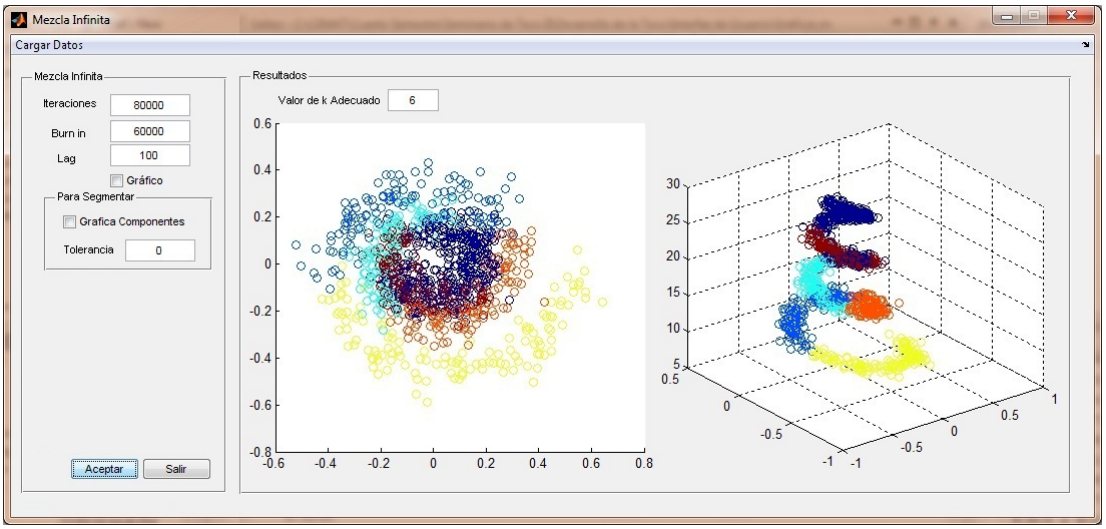


Figura A.7: Visualización de datos en 3 dimensiones, en donde la parte izquierda se muestran los datos en dos dimensiones y en la parte derecha se muestran los datos en 3 dimensiones.

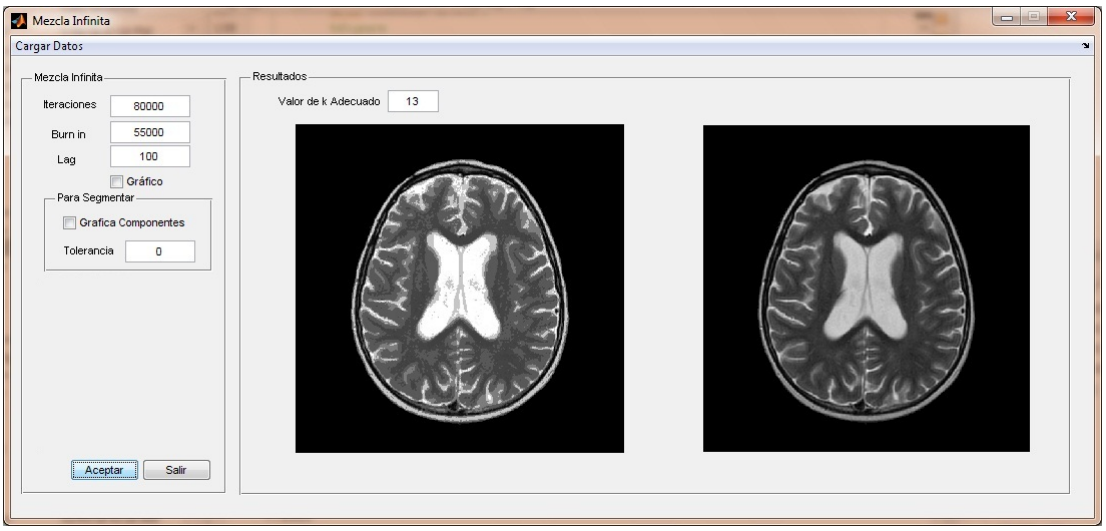


Figura A.8: Visualización de los cerebros, en la parte izquierda se muestra el cerebro segmentado y en la parte derecha el original

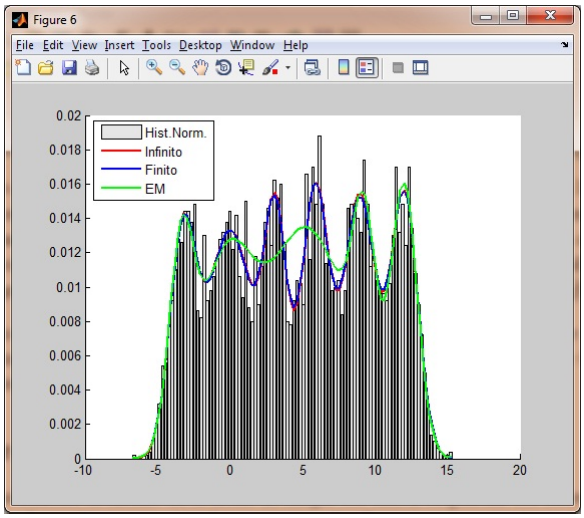


Figura A.9: Ventana Emergente que muestra la comparación de las densidades de los métodos finito, infinito y el EM.

- ***Para las imágenes tanto cerebrales como normales*** la visualización se realiza de la siguiente manera: En la ventana principal en la parte izquierda se muestra la imagen segmentada y en la parte derecha se muestra la imagen original, como puede verse en la figura (A.8), pero además de esto se muestran en ventanas emergentes (como en la figuras(A.4) y (A.5)) el histograma original de los datos así como el histograma normalizado, y la estimación de la densidad con la mezcla de gaussianas, estas ventanas emergentes contiene las mismas gráficas de la estimación de las densidades que las anteriores por lo que omití agregar más imágenes sobre lo mismo, como lo muestra la figura (A.8).

Nota importante: Para el caso univariado cuando se termina el proceso aparecen ventanas de visualización de los histogramas de los hiperparámetros $\alpha\, y\, k$, asi como también la gráfica de autocovarianza, de las gaussianas encontradas, la densidad encontrada de todas las muestras y la media de ellas y por último la comparación de las densidades con los métodos finito, infinito y el EM como aparece en la figura (A.9).

A.3. Interfaz del Software de la Mezcla Finita de Gaussianas

El software de mezcla finita contiene básicamente lo mismo en el modo de cargar los datos, en guardar los datos y en la visualización de los resultados, en donde hay una gran

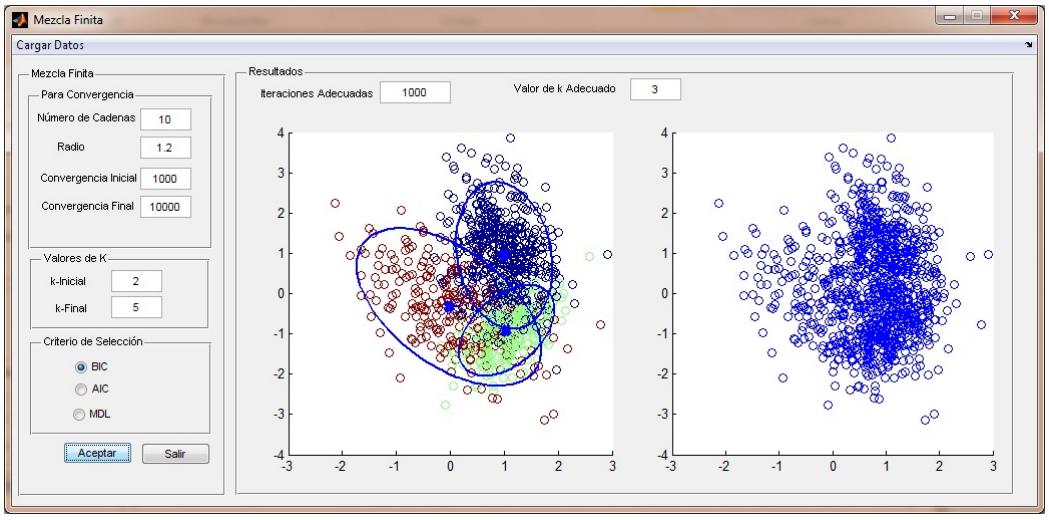


Figura A.10: Visualización y configuración del proceso de los datos para el caso finito

diferencia es en la configuración para correr el proceso, que detallaré a continuación.

El apartado para convergencia se utiliza para realizar la convergencia en la cadena de Markov ya que es posible determinar cuando la cadena está convergiendo y eso nos ahorrará tener que adivinar cuál es el número de iteraciones adecuado, esto es diferente en el caso infinito ya que por la manera de la implementación donde se considera el proceso del restaurante chino no es posible hacer esto, por lo tanto tenemos que:

- ***Número de cadenas:*** Indica cuantas cadenas de Markov se útilizaran para hacer converger, en la literatura se recomienda 10 por lo que en el programa este número está por default.

- ***Radio:*** Para saber que una cadena converge es necesario que al realizar este proceso los parámetros que se requieren que converjan tienden a ser 1 pero con no siempre es exacto por lo que hay un margen de error que en la literatura consideran entre 1.2 y 1.3 y es por ello que lo dejo abierto a su consideración.

- ***Convergencia Inicial:*** Indica el inicio del número de iteraciones para probar la convergencia en caso que no se logre se va duplicando hasta alcanzar la convergencia final y si aun así no converge se dirá que no existe convergencia en los parámetros.

- ***Convergencia final:*** Como ya indique previamente es el número de iteraciones límite para saber si hubo convergencia en los parámetros.

- ***k-inicial:*** Indica el número de k inicial, esto es posible si al ver los datos tienes alguna intuición de más o menos entre que intervalo se encontrará la k adecuada.

- ***k-final:*** Indica el número de k límite y de igual manera que el anterior se utiliza para encerrar en un cierto intervalo la k que se considera adecuada, aquí es importante mencionar que si se considera una k muy grande, el programa al detectar que no está habiendo convergencia después de analizar 3 número de k, inmediatamente genera un resultado, esto es porque el proceso de mezcla finita restringe a encontrar componentes nulos y esto haré restricción al número de k adecuado.

- ***Criterio de selección:*** Posteriormente después de tener los resultados de que la cadena de Markov ya ha convergido en los distintos número de k adecuados, se procede a realizar el criterio de selección que nos indicará de todas la k encontradas cuál es el valor correcto, esto lo realiza de manera automática de acuerdo a 3 posibles criterios que se le indicara por el usuario, los posibles criterios son el BIC, AIC y MDL que ya hemos hablado de ellos.

Como podrán darse cuenta aquí no existe la opción para segmentar porque no se desarrollo especialmente para ello, otra diferencia es que además de mostrar el valor de k adecuado, también muestra las iteraciones adecuadas que se obtuvieron mediante el proceso. Todo lo mencionado puede verse en la figura (A.10).

Para el caso de tomar muestras, lo hace de manera automática, a partir del total de iteraciones va tomando en intervalos (lag) de 10 hasta obtener 300 muestras y de alli saca el promedio de ellas.

Apéndice B

Función Verosimilitud de la Mezcla de Gaussianas

La función verosimilitud de la mezcla de gaussianas está dado por

$$\ell(Y|\mu,\Sigma) = log\, p(Y|\mu,\Sigma) = \sum_{n=1}^{N} log \sum_{k=1}^{K} \pi_k p(y_k|\mu_k,\Sigma_k) \tag{B.0.1}$$

donde

$Y = \{y_1, ..., y_n\}$ es el conjunto de observaciones.

por lo que la función a optimizar es

$$\Theta^*_{ML} = arg \max_{\Theta} \{\ell(\mu,\Sigma)\} \tag{B.0.2}$$

donde

Θ^*_{ML} son los parámetros optimizados por máxima verosimilitud.

Es importante destacar que la maximización de la función B.0.2 no es trivial por varias razones

- Existen valores del conjunto de parámetros para los que la verosimilitud es infinita. Esto ocurre cuando la media coincide con el dato y la matriz de covarianza o la varianza es nula o muy cercana a cero.
- La presencia de grupos de observaciones muy próximas unas de otras puede provocar un máximo local de la función. Como consecuencia, se obtendría una pobre representación de la densidad de probabilidad real.

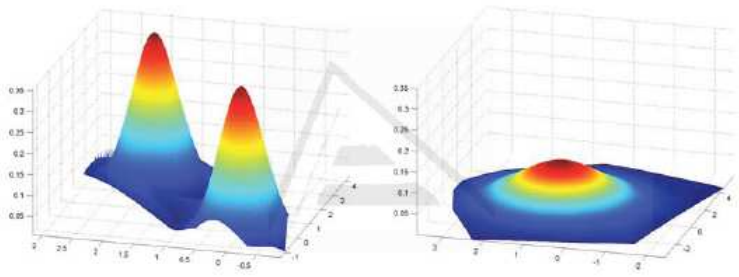


Figura B.1: Si el número de componentes del modelo no es establecido correctamente, los datos pueden describirse erróneamente. Dos distribuciones gaussianas con media $\mu_1 = [0, 0]$ y $\mu_2 = [3, 2]$ (izquierda) son modeladas con un único componente con media $\mu = [1.5, 1]$(derecha).

Por lo tanto, es necesario la utilización de algún método que aproxime muy bien el conjunto óptimo de parámetros, un algoritmo iterativo muy conocido es la maximización de la esperanza (EM)(Bishop [2007]), el cuál permite encontrar soluciones de máxima verosimilitud a problemas en los que existen variables ocultas, el algoritmo consta de dos pasos principales que son el paso E que es donde realiza una estimación del valor esperado de las variables ocultas del problema a partir de los datos observados y la estimación actual de los parámetros del modelo. Ahora a partir del paso E realiza el paso M que una vez que se tenga el valor esperado, el nuevo conjunto de parámetros se maximiza de cierta manera ya definida, para una descripción más detallada del algoritmo puede consultar (Bishop [2007] y Redner and Walker [1984]).

La mayor parte de los algoritmos existentes para ajustar los parámetros de una mezcla con un número desconocido de componentes a priori utilizan el algoritmo del EM. Aunque el funcionamiento del algoritmo es satistactorio, presenta algunos inconvenientes que se detallan a continuación

- ***Inicialización:*** El éxito del algoritmo EM depende en gran medida de los valores iniciales del conjunto de parámetros. En general se utiliza una clusterización previa por medio de un algoritmo llamado k-medias.

- ***Convergencia hacia un máximo local:*** Cuando se ajustan los parámetros de un modelo de mezcla de gaussiana sin ningún tipo de restricciones sobre las matrices de covarianza de los componentes, alguno de los pesos (π_i) podría aproximarse a cero y por consiguiente, el componente podría estar próximo a la singularidad. Cuando

el número de componentes es superior al óptimo esto puede ocurrir con relativa frecuencia, convirtiéndose por lo tanto en un serio problema para métodos que requieren estimaciones de los parámetros de la muestra para varios valores de los componentes (k).

- ***No se estima el orden del modelo:*** Como se ha descrito anteriormente, el algoritmo por si mismo no permite la estimación del número de componentes del modelo, por lo que es necesario realizar una selección de modelo que detallaremos en capítulos posteriores. Esto es necesario ya que pueden ocurrir errores como aparece en la figura (B.1).

Bibliografía

H. Raiffa and R. Schlaifer. *Applied Statistical Decision Theory.* MIT Press, Cambridge, 1961. 1

T. S. Ferguson. A bayesian analysis of some nonparametric problems. *The Annals of Statistics*, 1(2):209–230, 1973. ISSN 00905364. doi: 10.2307/2958008. 1, 2, 52, 62

Radford M. Neal. *Bayesian Learning for Neural Networks (Lecture Notes in Statistics).* Springer, 1 edition, August 1996. ISBN 0387947248. 2

Christopher K. I. Williams and Carl Edward Rasmussen. Gaussian processes for regression. In *Advances in Neural Information Processing Systems 8.* MIT press, 1996. 2

T. S. Ferguson. Prior distributions on spaces of probability measures. *The Annals of Statistics*, 1(2):615–629, 1974. 2, 62

David Blackwell and James B. MacQueen. Ferguson distributions via polya urn schemes. *The Annals of Statistics*, 1(2):353–355, 1973. ISSN 00905364. doi: 10.2307/2958020. 2, 62, 66

J. Sethuraman. A constructive definition of dirichlet priors. *Statistica Sinica*, 4:639–650, 1994. 2, 62, 69

Charles E. Antoniak. Mixtures of dirichlet processes with applications to bayesian nonparametric problems. *The Annals of Statistics*, 2(6):1152–1174, 1974. ISSN 00905364. doi: 10.2307/2958336. 2

Sylvia Richardson and Peter J. Green. On bayesian analysis of mixtures with an unknown number of components (with discussion). *Journal of the Royal Statistical Society: Series B (Methodological)*, 59(4):731–792, 1997a. ISSN 1467-9868. doi: 10.1111/1467-9868.00095. 3

A. P. Dawid. Condicional independence for statistical operations. *The Annals of Statistics*, 1(8):589–617, 1980. 8

A. P. Dawid. Conditional independence in statistical theory. *Journal of the Royal Statistical Society. Series B (Methodological)*, 41(1):1–31, 1979. ISSN 00359246. doi: 10.2307/2984718. 9

Judea Pearl. *Probabilistic reasoning in intelligent systems : networks of plausible inference*. Morgan Kaufmann, September 1997. ISBN 1558604790. 13

Steffen L. Lauritzen. *Graphical Models (Oxford Statistical Science Series)*. Oxford University Press, USA, July 1996. ISBN 0198522193. 14

W. R. Gilks and P. Wild. Adaptive rejection sampling for Gibbs sampling. *Journal of the Royal Statistical Society. Series C (Applied Statistics)*, 41(2):337–348, 1992. ISSN 00359254. doi: 10.2307/2347565. 21, 51

Radford M. Neal. Probabilistic inference using markov chain monte carlo methods, 1993. 23

Stuart Geman and Donald Geman. Stochastic relaxation, gibbs distributions and the bayesian restoration of images. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 6(6):721–741, November 1984. doi: 10.1080/02664769300000058. 24

W. K. Hastings. Monte carlo sampling methods using markov chains and their applications. *Biometrika*, 57(1):97–109, April 1970. doi: 10.1093/biomet/57.1.97. 24

Kenneth P. Burnham, David R. Anderson, and Kenneth P. Burnham. *Model selection and multimodel inference: a practical information-theoretic approach*. Springer, 2nd edition, December 2002. ISBN 978. 33

H. Akaike. A new look at the statistical model identification. *Automatic Control, IEEE Transactions on*, 19(6):716–723, January 2003. 33, 34

A. Liddle. Information criteria for astrophysical model selection. *astro-ph/0701113*, (6), June 2008. 34

W. Gilks and S. Richardson. *Markov Chain Monte Carlo in Practice: Interdisciplinary Statistics (Chapman Hall Interdisciplinary Statistics)*. Chapman Hall and CRC, 1 edition, December 1995. ISBN 0412055511. 34, 36

Gideon Schwarz. Estimating the dimension of a model. *The Annals of Statistics*, 6(2): 461–464, 1978. 35

Centrum Wiskunde Informatica. A tutorial introduction to the minimum description length principle. 36

B. Efron. Bootstrap methods: Another look at the jackknife. *The Annals of Statistics*, 7 (1):1–26, 1979. ISSN 00905364. doi: 10.2307/2958830. 37

Carl E. Rasmussen. The infinite gaussian mixture model. In *In Advances in Neural Information Processing Systems 12*, volume 12, pages 554–560, 2000. 42

Sebastian Thrun, Wolfram Burgard, and Dieter Fox. *Probabilistic Robotics (Intelligent Robotics and Autonomous Agents)*. The MIT Press, September 2005. ISBN 0262201623. 44, 45

Mike West, Peter MÃŒller, Peter M, and Michael D. Escobar. Hierarchical priors and mixture models, with application in regression and density estimation, 1994. 59

Steven N. Maceachern and Peter MÃŒller. Estimating mixture of dirichlet process models. *Journal of Computational and Graphical Statistics*, 7(2):223–238, 1998. ISSN 10618600. doi: 10.2307/1390815. 59

Richard O. Duda and Peter E. Hart. *Pattern Classification and Scene Analysis*. John Wiley & Sons Inc, June 1973. ISBN 0471223611. 59

José M. Bernardo and Adrian F. M. Smith. *Bayesian Theory*. Wiley, 1 edition, 1994. ISBN 047149464X. 59

Marcin T. Kacperczyk, Paul Damien, and Stephen G. Walker. A New Class of Bayesian Semiparametric Models with Applications to Option Pricing. *SSRN eLibrary*, 2005. doi: 10.2139/ssrn.416583. 62

David B. Dunson. Bayesian semiparametric isotonic regression for count data. *Journal of the American Statistical Association*, 100:618–627, June 2005. 62

M. Medvedovic and S. Sivaganesan. Bayesian infinite mixture model-based clustering of gene expression profiles. *Bioinformatics*, pages 1194–1206, 2002. 62

J.L Bigelow and D.B Dunson. Posterior simulation across nonparametric models for functional clustering. *Journal of the Royal Statistical society*, 2007. 62

Zhang Xinhua. A very note on the construction of dirichlet process. Technical report, September 2008. 63, 70

Radford M. Neal. Markov chain sampling methods for dirichlet process mixture models. *Journal of Computational and Graphical Statistics*, 9(2):249–265, 2000. ISSN 10618600. doi: 10.2307/1390653. 73, 75

Ryohei Nakano, Zoubin Ghahramani, Georey E. Hinton, Georey E. Hinton, Naonori Ueda, and Naonori Ueda. Smem algorithm for mixture models. *Neural Computation*, 12(12): 200–0, 1999. 112

Sylvia Richardson and Peter J. Green. On bayesian analysis of mixtures with an unknown number of components. In *Institute of International Economics Project on International Competition Policy; COM/DAFFE/CLP/TD(94)42*, 1997b. 112

Y. W. Teh, M. I. Jordan, M. J. Beal, and D. M. Blei. Hierarchical Dirichlet processes. *Journal of the American Statistical Association*, 101(476):1566–1581, 2006. 113

Y. W. Teh, D. Görür, and Z. Ghahramani. Stick-breaking construction for the Indian buffet process. In *Proceedings of the International Conference on Artificial Intelligence and Statistics*, volume 11, 2007. 113

Christopher M. Bishop. *Pattern Recognition and Machine Learning (Information Science and Statistics)*. Springer, 1 edition, October 2007. ISBN 0387310738. 125

R. Redner and H. Walker. Mixture densities, maximum likelihood and the em algorithm. *SIAM Review*, 26(2):195–239, 1984. 125