Universidad de la República

Informe de trabajo final de grado de la Licenciatura en Estadística

---

Estimación de densidades mediante mezclas controladas por el Proceso de Dirichlet

---

Estudiantes:

Manuel Hernández Banadik
Mario Sierra Golombievski

Tribunal:

Ignacio Álvarez-Castro (*Tutor*)
Juan Kalemkerían
Marco Scavino

02 de julio de 2019.

Resumen

La estimación de densidades es un tema de mucha relevancia en el área de la Estadística. Existen al menos dos métodos para abordar este problema. Por un lado, el enfoque paramétrico asume un modelo de probabilidad para la muestra bajo estudio; mientras que el enfoque no–paramétrico busca relajar estos supuestos a costa de una modelización más compleja y flexible.

Independientemente de los métodos de estimación mencionados, se pueden considerar dos enfoques a la hora de abordar cualquier problema en Estadística, en particular el de la estimación de densidades. Estos son: el enfoque clásico y el bayesiano. En este trabajo haremos revisión de una técnica no paramétrica bayesiana.

Desde un punto de vista bayesiano, para estimar una distribución, se requiere establecer una distribución *a priori* en el espacio de las medidas de probabilidad. El Proceso de Dirichlet cumple con esta función.

Comenzamos presentando al Proceso de Dirichlet como una medida de probabilidad aleatoria, para luego estudiar en detalle los modelos de mezcla controlados por este proceso.

Analizamos en detalle la implementación computacional de esta técnica, comparando su desempeño con el de otras técnicas en datos simulados y reales. Como aplicación interesante realizamos un análisis de temperaturas máximas en el territorio Uruguayo.

Palabras claves: estimación de densidades; inferencia Bayesiana; proceso de Dirichlet; medida de probabilidad aleatoria; modelos de mezcla.

Índice general

Índice de figuras

1

Introducción

El modelado de la distribución de los datos es un asunto de mucha relevancia en la estadística. Si se tiene el supuesto de que lo observado proviene de algún modelo paramétrico conocido, el problema de estimar dicha distribución se reducirá a estimar los parámetros que la caracterizan. Sin embargo, por diversos motivos, esta hipótesis en muchos casos puede resultar muy restrictiva: a menudo los datos suelen tener complejos comportamientos, y no hay un modelo conocido que los describa adecuadamente. Son estos los casos en los que los modelos paramétricos resultan rígidos y limitantes, y debemos recurrir a métodos más flexibles. Wasserman (2006) caracteriza a la Estadística No Paramétrica como *un conjunto de técnicas modernas para hacer inferencia con la menor cantidad de hipótesis posibles.*

Si los datos, se consideran observaciones de una variable aleatoria absolutamente continua, entonces asumimos la existencia de una función de densidad de probabilidad. La estimación de dicha función es justamente una de las inquietudes que la estadística no paramétrica ha respondido en los últimos años, de forma cada vez más sofisticada, debido en gran parte, al acceso a una mayor capacidad de cálculo computacional.

Son muy conocidas y de uso extendido diversas técnicas para la estimación de densidades, desde las más sencillas como el histograma, hasta las más sofisticadas como los estimadores por núcleo, introducidos por Parzen (1962) y Rosenblatt (1956). Dada una muestra $x_1, \ldots, x_n$, una función núcleo $K(\cdot)$, y un parámetro de suavizado h (ancho de banda), se construyen estos estimadores de la siguiente manera: $\hat{f}(x) = \sum_{i=1}^{n} \frac{1}{nh} K(\frac{x-x_i}{h})$. Para una descripción detallada de esta técnica y la demostración de sus propiedades matemáticas

ver Wasserman (2006).

En el presente informe nos proponemos abordar una alternativa bayesiana no-paramétrica: la estimación a través de modelos de mezcla controlados por el Proceso de Dirichlet. Esta es una técnica concebida por Antoniak (1974), que modela la densidad como:

$$f(x) = \int f_\theta(x|\theta) d\mathrm{P}(\theta)$$

donde la mezcla estará controlada por una medida P que será una realización de un Proceso de Dirichlet. Escobar & West (1995) observan que este tratamiento, el de pensar en la función de densidad como una mixtura, subyace en todos los métodos más usados, incluidos los estimadores por núcleo.

Estructura del documento

En el capítulo 2 se hace una breve mención a conceptos básicos de la estadística Bayesiana, la descripción de una familia de modelos de mezcla y una mirada a la distribución de Dirichlet con su aplicación en el modelo Dirichlet-Multinomial.

En el capítulo 3 se introduce el Proceso de Dirichlet, su definición formal como una medida de probabilidad aleatoria y una definición constructiva (*Stick-Breaking*). Luego se definirán los modelos de mezcla donde la distribución que controla la mezcla es un Proceso de Dirichlet, y se verá la mezcla de normales como caso particular de dicha familia de modelos.

En el capítulo 4 se abordarán los aspectos computacionales de la implementación de esta técnica, la obtención del estimador así como un análisis de la convergencia del algoritmo de muestreo de Gibbs.

En el capítulo 5 se propone un estudio de simulación donde se compara el MISE estimado de esta técnica respecto del estimador por núcleos. En el capítulo 6 se aplica la técnica a un conjunto de datos de temperaturas máximas en Uruguay. Luego, se pueden encontrar comentarios finales y posibles líneas de trabajo a seguir.

2

Aspectos preliminares

"La irrisoria estadística .1 en un millón,
es un montón, si justo te toca a vos."

Riki Musso. *La antorcha humana*

2.1. Inferencia Bayesiana

Bajo la perspectiva de la estadística clásica, un parámetro es *algo* respecto de lo cual se tiene incertidumbre y que se busca estimar.

La inferencia bayesiana modela esta incertidumbre tratando al parámetro como una variable aleatoria. Antes de realizar un experimento, es decir, antes de contar con datos, a esta variable aleatoria se le asigna una distribución de probabilidad denominada distribución previa o *a priori*. La inferencia se hará estimando la distribución posterior o *a posteriori*, que será la distribución del parámetro, condicionada a la observación de un conjunto de datos.

Supongamos que se observan los datos $\mathbf{y} = (y_1, \ldots, y_n)$.

- $\mathbf{y} \sim f(\mathbf{y}|\theta)$ es lo que se denomina modelo para los datos
- $f(\theta)$ es la distribución previa para el parámetro
- $f(\theta|\mathbf{y})$ es la distribución posterior.

A través del Teorema de Bayes se obtiene la distribución posterior, $f(\theta|\mathbf{y}) = \frac{f(\mathbf{y}|\theta)f(\theta)}{\int f(\mathbf{y}|\theta)f(\theta)d\theta}$. En general, esto no se puede calcular analíticamente, siendo necesario recurrir a métodos computacionales para estimarla.

2.2. Inferencia de la posterior, MCMC y muestreo Gibbs

Los algoritmos *Markov Chain Monte Carlo*, generan secuencias de valores simulados que forman una cadena de Markov. Estas simulaciones resultarán ser muestras de la distribución posterior, que bajo un buen funcionamiento de estos métodos, permitirán hacer una buena estimación de la misma.

Cuando se tiene un parámetro multidimensional, uno de los algoritmos más utilizados es el *muestreo de Gibbs*.

Describiremos este método en forma sencilla:

1. Separar el vector de parámetros θ en sub-vectores $\theta_1, \ldots, \theta_d$
2. Hallar las distribuciones *full-conditional*. Esto es calcular:
   $f(\theta_i|\theta_{-i})$ donde $\theta_{-i} := (\theta_1, \ldots, \theta_{i-1}, \theta_{i+1}, \ldots, \theta_d)$
3. Dar valores iniciales $(\theta_1^{(0)}, \ldots, \theta_d^{(0)})$
4. Actualizar los valores, en la k-ésima iteración, según el siguiente criterio:
   $\theta_1^{(k)} \sim f(\theta_1|\theta_{-1}^{(k-1)})$
   $\theta_i^{(k)} \sim f(\theta_i|\theta_1^{(k)}, \ldots, \theta_{i-1}^{(k)}, \theta_{i+1}^{(k-1)}, \ldots, \theta_d^{(k-1)})$
   $\theta_d^{(k)} \sim f(\theta_d|\theta_{-d}^{(k)})$

Como resultado se obtiene una cadena de vectores $(\theta_1^{(k)}, \ldots, \theta_d^{(k)})_{k=1,\ldots,m}$. Hay ciertas precacuciones que hay que tener en cuenta de modo de evitar incurrir en posibles errores en la simulación. Siguiendo a Gelman et al. (2013), estos son:

- Implementar una cantidad suficiente de iteraciones, para permitir que la cadena explore la distribución deseada.

- Simular varias cadenas con distintos valores iniciales, de forma de evaluar si los estados que el proceso alcanza son sensibles respecto a su inicialización.
- Descartar las primeras iteraciones, ya que se considera que en su comienzo el proceso está en fase de *calentamiento* (*warm-up*): que aún no entró en régimen y guarda dependencia con sus valores iniciales.

Para ejemplificar un posible error imaginemos una Cadena de Markov con dos clases de comunicación y baja probabilidad de comunicación entre ellas. Si el proceso se observa durante un período de tiempo insuficiente, el mismo podría permanecer en una sola de las clases, sin llegar a alcanzar la otra. Como resultado se tendría una trayectoria que será dependiente del estado inicial del proceso y no representará adecuadamente al conjunto de estados.

2.3. Modelos de mezcla

" Albañil yo soy
y puedo con
el balde y la cuchara
taparte el sol. "

Jorge Lazaroff. *Albañil*

Los modelos de mezcla son modelos que se suelen utilizar para describir una población, cuando se considera que la misma es una agregación de subpoblaciones. Un esquema posible es considerar un modelo *base* $f(\cdot|\theta_h)$ para cada subpoblación (que depende de un parámetro θ_h) e indicar qué peso π_h tendrá la misma en el total.

Suponiendo que tenemos H componentes, y unos ciertos pesos $\pi_h \geq 0$ con $\sum_h \pi_h = 1$, el modelo poblacional estará dado por:

$$f(x) = \pi_1 f(x|\theta_1) + \cdots + \pi_H f(x|\theta_H). \tag{2.1}$$

Al observar un dato de una población con estas características, estaremos observando un dato de alguna de las H subpoblaciones que la componen, por tanto con probabilidad dada por el peso π_h, ese dato observado corresponderá al componente h, modelado según $f(\cdot|\theta_h)$.

La especificación de un modelo de este tipo se da mediante la especificación de un modelo base $f(\cdot|\theta)$ y la especificación de una distribución de mezcla, $\mathrm{P}(\theta)$. Desde un punto de vista bayesiano, donde θ se modela como variable aleatoria, $\mathrm{P}(\theta)$ controlará qué valores toma y con qué probabilidad. De esta forma el modelo explicitado en (2.1) se puede escribir de la siguiente forma:

$$f(x) = \int f(x|\theta) d\mathrm{P}(\theta)$$

con

$$\mathrm{P}(\theta) = \begin{cases} \pi_1 \ \ si\ \theta = \theta_1 \\ \vdots \\ \pi_H \ \ si\ \theta = \theta_H \end{cases}$$

donde $\theta_1, \ldots, \theta_H$ son los valores que toma θ en cada subpoblación.

Veamos que P es una distribución que está controlando la mezcla. Si P distribuye toda la probabilidad en un conjunto finito de átomos, entonces tendremos un modelo de mezcla finito. Si P es absolutamente continua, tendremos una mezcla continua.

Como ejemplo consideremos una de mezcla de dos componentes normales con medias μ_1, μ_2 y varianza conocida. El modelo resultante será $f(x) = \pi_1 f(x|\mu_1) + \pi_2 f(x|\mu_2)$, con $\pi_2 = 1 - \pi_1$. En la figura 2.1 se muestra la densidad resultante en el caso en que $\pi_1 > 1/2$ y $\mu_1 < \mu_2$.

En este caso $f(\cdot|\mu)$ es la distribución normal, y μ toma los valores

$$\begin{cases} \mu_1, & \text{con probabilidad } \pi_1 \\ \mu_2, & \text{con probabilidad } 1 - \pi_1 \end{cases}$$

2.3.1. Algoritmo de Gibbs en un ejemplo de mezcla finita

Los conceptos de esta sección están basados en Niemi (2017b). El código de **R** fue escrito por los autores de este trabajo.

Supongamos que los datos $x_1, \ldots, x_n$ son realizaciones de una variable aleatoria que tiene la siguiente función de densidad:

$$f(x|\pi, \mu_1, \mu_2) = \pi\phi(x; \mu_1, \sigma^2) + (1 - \pi)\phi(x; \mu_2, \sigma^2)$$

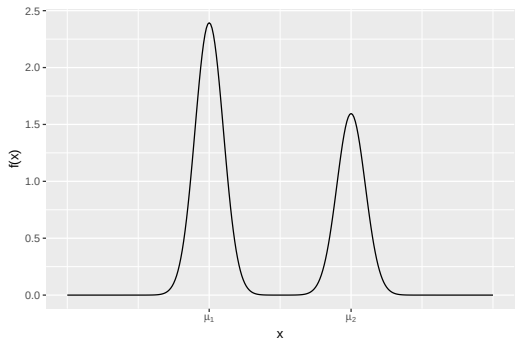


Figura 2.1: Ejemplo de mezcla de dos modelos normales

donde ϕ es la función de densidad normal con media μ y varianza σ^2, y supongamos también σ^2 es conocido.

La distribución de los datos, está dada por los parámetros (π, μ_1, μ_2). La mezcla es de dos componentes, sus pesos son π y $(1 - \pi)$.

Por tanto asumamos,

$$\begin{aligned} \pi &\sim Beta(a_1, a_2) \\ \mu_1, \mu_2 &\overset{ind}{\sim} N(m_h, v_h^2\sigma^2) \end{aligned}$$

π independiente de μ_1, μ_2.

Por otro lado definamos las variables ξ_i como variables latentes que indican la componente a la cual pertenece el i-ésimo dato. En este caso, $Rec(\xi_i) = 1, 2$ y $\mathbb{P}(\xi_i = 1) = \pi$, $\mathbb{P}(\xi_i = 2) = 1 - \pi$

Una previa sugerida (Niemi, 2017b) es considerar, para $h = 1, 2$:

$m_h = 0$
$v_h = 1$
$a_h = 1/2$

Una vez sorteados los valores iniciales, la iteración $k + 1$ del muestreo de Gibbs viene dada por el siguiente esquema:

1. Para $i = 1, \ldots, n$, sortear $\xi_i^{(k+1)}$ de su distribución *full conditional*:

$$\begin{aligned} \mathbb{P}(\xi_i^{(k+1)} = 1 | \pi^{(k)}, \mu^{(k)}) &\propto \pi^{(k)} N(x_i; \mu_h^{(k)}, \sigma^2) \\ \mathbb{P}(\xi_i^{(k+1)} = 2 | \pi^{(k)}, \mu^{(k)}) &\propto (1 - \pi)^{(k)} N(x_i; \mu_h^{(k)}, \sigma^2) \end{aligned}$$

2. Sortear π y μ:

   a) Sortear $\pi^{(k+1)} \sim Beta(a_1 + Z_1^{(k+1)}, a_2 + Z_2^{(k+1)})$ donde $Z_h^{(k+1)} = \sum_{i=1}^{n} \mathbb{1}(\xi_i^{(k+1)} = h)$.

   b) Para $h = 1, \ldots, 2$, sortear $\mu_h^{(k+1)}$ de su distribución *full conditional*:

$$\mu_h \overset{ind}{\sim} N(m_h^{(k+1)}, {v_h^{(k+1)}}^2 \sigma^2)$$

   donde

$$\begin{aligned} {v_h^{(k+1)}}^2 &= (1/{v_h^{(k)}}^2 + Z_h^{(k+1)})^{-1} \\ m_h^{(k+1)} &= {v_h^{(k+1)}}^2 (m_h^{(k)}/{v_h^{(k)}}^2 + Z_h^{(k+1)} \overline{x}_h) \\ \overline{x}_h &= \frac{1}{Z_h^{(k+1)}} \sum_{i:\xi_i^{(k+1)}=h} x_i \end{aligned}$$

En la figura 2.2 se presenta la evolución de las cadenas a lo largo de las iteraciones. En la figura 2.3 se presenta el gráfico de la densidad y la densidad estimada, así como también se señala la verdadera etiqueta de cada dato con diferente ordenada y la etiqueta estimada en distinto color.

Implementación en R

```
# Definición de los parámetros del problema, y simulación de los datos
n <- 500 # cantidad de datos a simular
s <- .3 # desvío de las subpoblaciones
p <- .5 # peso del componente 1
mu1 <- 0 # media del componente 1
mu2 <- 1 # media del componente 2
set.seed(12348)
X <- c(rnorm(n*p,mu1,s),rnorm(n*(1-p),mu2,s)) # los datos
grupo <- c(rep(1,n*p),rep(2,n*(1-p))) # id. de componente

M <- 2000 # cantidad de iteraciones de Gibbs
# En los siguientes objetos se irá almacenando los valores de
# de los parámetros a lo largo de las iteraciones.
m <- matrix(0,M,2)
v2 <- matrix(0,M,2)
a <- matrix(0,M,2)
pi <- matrix(0,M,1)
```

```
muh <- matrix(0,M,2)
prob <- matrix(0,M,1)
z <- matrix(0,M,n)
# Especificación de la previa
m[1,] <- c(.5,.5) # media de la media de los componentes
v2[1,] <- c(1,1) # varianza de la media de los componentes
a[1,] <- c(.5,.5) # parámetro de la distribución Dirichlet/Beta
# Iteración
invisible(sapply(2:M,function(j){
  pi[j] <<- rbeta(1,a[j-1,1],a[j-1,2])
  muh[j,] <<- rnorm(2,m[j-1,],sqrt(v2[j-1,])*s)
  prob<<- pi[j]*dnorm(X,muh[j,1],s)/(pi[j]*(dnorm(X,muh[j,1],s))
                                     +(1-pi[j])*dnorm(X,muh[j,2],s))
  z[j,] <<- sapply(1:n, function(i)
    sample(c(1,2),size=1,prob=c(prob[i],1-prob[i])))
  a[j,1] <<- a[j-1,1]+sum(z[j,]==1)
  a[j,2] <<- a[j-1,2]+sum(z[j,]==2)
  v2[j,1] <<- 1/(1/v2[j-1,1]+sum(z[j,]==1))
  v2[j,2] <<- 1/(1/v2[j-1,2]+sum(z[j,]==2))
  m[j,1] <<- v2[j,1]*(m[j-1,1]/v2[j-1,1]+sum(X[z[j,]==1]))
  m[j,2] <<- v2[j,2]*(m[j-1,2]/v2[j-1,2]+sum(X[z[j,]==2]))}))
```

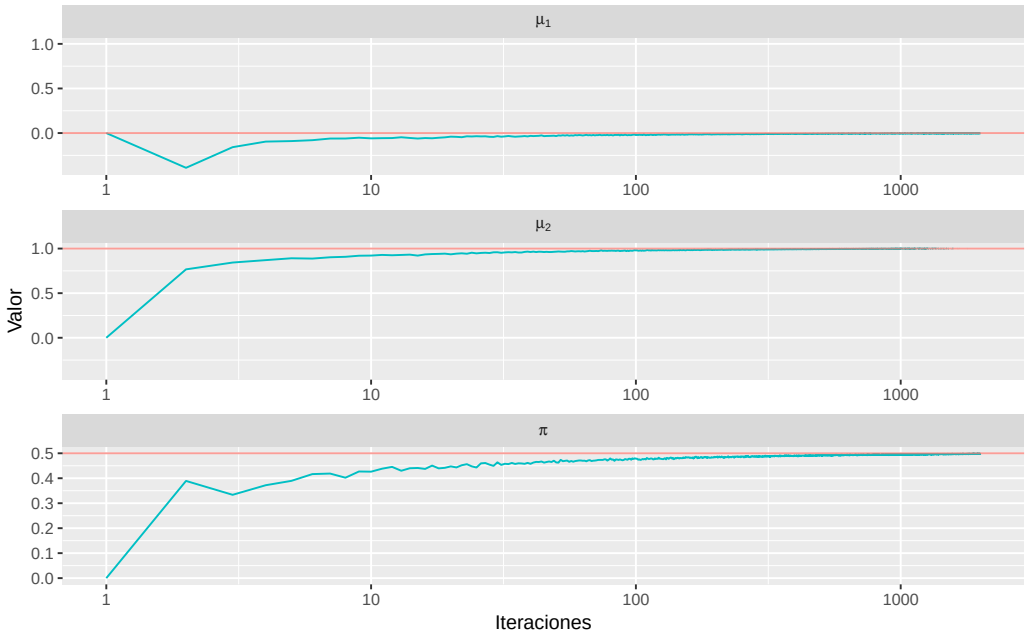


Figura 2.2: Evolución de las estimaciones de los parámetros en un modelo de mezcla finito

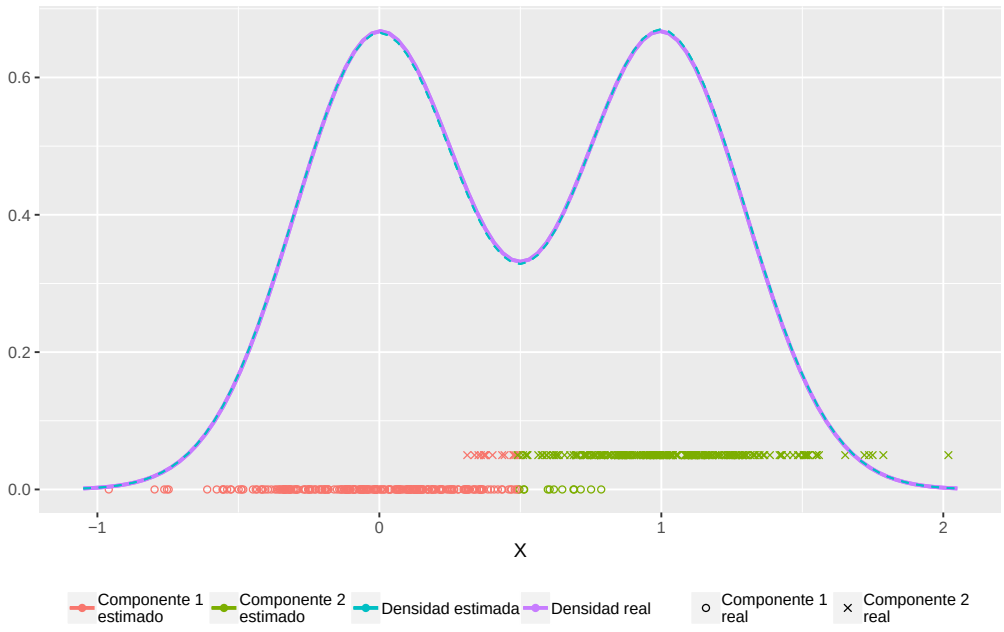


Figura 2.3: Densidad estimada, y representación de las componentes generadores de los datos en un modelo de mezcla finito

2.4. Modelo Dirichlet - Multinomial

Dado un experimento aleatorio cuyo espacio muestral es un conjunto finito de posibles resultados, el vector aleatorio que cuenta la cantidad de veces que se observó cada uno de los resultados al cabo de n intentos independientes, se puede modelar según la distribución Multinomial.

Sea $X_1, \ldots, X_n$ una sucesión de variables aleatorias i.i.d., categóricas que toman valores en $\mathcal{X} = \{1, 2, \ldots, k\}$ con probabilidades $p_1, p_2, \ldots, p_k$: $\sum_{j=1}^{k} p_j = 1$.

Si se define, $Y_j := \sum_{i=1}^{n} \mathbb{1}_{\{X_i=j\}}$ para $j = 1, \ldots, k$ se tiene que el vector aleatorio $(Y_1, \ldots, Y_k)$ se distribuye $Multinomial(p_1, \ldots, p_k)$.

Si se quiere hacer inferencia acerca de la distribución de probabilidad de la ocurrencia de dichos resultados, desde un enfoque bayesiano, se necesita asignar una distribución previa al vector de probabilidades. La distribución de Dirichlet, precisamente, cumple ese rol: es una distribución cuyo soporte son los vectores de probabilidad, es decir, todos los vectores cuyas entradas son valores reales en el intervalo [0,1] que suman 1.

2.4.1. Distribución de Dirichlet

Consideremos el conjunto de los vectores de probabilidad de dimensión k:

$$\mathcal{P}_k = \{(p_1, \ldots, p_k); p_i \geq 0, \forall i \in \{1, \ldots, k\} \wedge \sum_{i=1}^{k} p_i = 1\}$$

Diremos que un vector aleatorio $(P_1, \ldots, P_k)$ sigue una distribución Dirichlet de parámetros $(a_1, \ldots, a_k)$, con $a_i > 0$ para todo $i = 1, \ldots, k$, si:

$$f(p_1, \ldots, p_k | a_1, \ldots, a_k) = \frac{\Gamma\left(\sum_{j=1}^{k} a_j\right)}{\prod_{j=1}^{k} \Gamma(a_j)} \prod_{j=1}^{k} p_j^{a_j - 1} \mathbb{1}_{\{\mathcal{P}_k\}}(p_1, \ldots, p_k) \qquad (2.2)$$

y lo notaremos: $(P_1, \ldots, P_k) \sim Dirichlet(a_1, \ldots, a_k)$

Algunas observaciones acerca de esta distribución:

- Se trata de una distribución $(k-1)$-dimensional, pues $P_k = 1 - \sum_{j \neq k} P_j$

- Notar que si $k = 2$, por la observación anterior, $P_2 = 1 - P_1$, y la ecuación (2.2) se reduce a:

$$f(p_1|a_1, a_2) = \frac{\Gamma(a_1 + a_2)}{\Gamma(a_1)\Gamma(a_2)} p_1^{a_1-1}(1 - p_1)^{a_2-1} \mathbb{1}_{\{[0,1]\}}(p_1)$$

  Por tanto $P_1 \sim Beta(a_1, a_2)$

- Considerando $a_i = 1$ para todo i, se obtiene la distribución uniforme sobre los vectores de probabilidad.

Proposición 1. *Si* $P_i := \dfrac{Z_i}{\sum_{j=1}^{k} Z_j}$ *con* $Z_i \sim Gamma(a_i, 1)$ *independientes, entonces el vector* $(P_1, \ldots, P_k)$ *tendrá distribución Dirichlet* $(a_1, \ldots, a_k)$.

La demostración de esta proposición se encuentra en el apéndice A.

Si se observa $\mathbf{y}_1, \ldots, \mathbf{y}_N$, conteos independientes, donde $\mathbf{y}_i = (y_{1,i}, \ldots, y_{k,i})$, la posterior $(p_1, \ldots, p_k)|\mathbf{y}_1, \ldots, \mathbf{y}_N$ también tiene distribución de Dirichlet, por lo tanto, bajo el modelo Multinomial para los datos, ésta previa resulta ser conjugada natural. Esto se puede puede ver de la siguiente forma:

$$\begin{aligned}
f(p_1, \ldots, p_k|\mathbf{y}_1, \ldots, \mathbf{y}_N) \propto \; & \textstyle\prod_{i=1}^{N} f(\mathbf{y}_i|p_1, \ldots, p_k) f(p_1, \ldots, p_k) && = \\
& \textstyle\prod_{i=1}^{N} f(y_{1,i}, \ldots, y_{k,i}|p_1, \ldots, p_k) f(p_1, \ldots, p_k) && = \\
& \textstyle\prod_{i=1}^{N} n! \prod_{j=1}^{k} \left(\frac{p_j^{y_{j,i}}}{y_{j,i}!} \right) \frac{\Gamma\left(\sum_{j=1}^{k} a_j\right)}{\prod_{j=1}^{k} \Gamma(a_j)} \prod_{j=1}^{k} p_j^{a_j-1} && \propto \\
& \textstyle\prod_{j=1}^{k} \prod_{i=1}^{N} \{p_j^{y_{j,i}}\} \prod_{j=1}^{k} p_j^{a_j-1} && = \\
& \textstyle\prod_{j=1}^{k} \left(p_j^{a_j + \sum_{i=1}^{N} \{y_{j,i}\} - 1} \right)
\end{aligned}$$

Por tanto

$$(P_1, \ldots, P_k)|\mathbf{y}_1, \ldots, \mathbf{y}_N \sim Dirichlet(a_1 + \sum_{i=1}^{N} y_{1,i}, \ldots, a_k + \sum_{i=1}^{N} y_{k,i})$$

3

Proceso de Dirichlet

En este capítulo presentaremos el desarrollo teórico de este trabajo. Empezaremos describiendo los modelos y procedimientos más sencillos que permitirán, por ejemplo, estimar la distribución de una variable categórica. Luego definiremos el Proceso de Dirichlet y mostraremos cómo mediante su introducción, se puede extender el problema a dimensión infinita. Por último introduciremos una nueva clase de procesos: las Mezclas de Procesos de Dirichlet, que permitirán realizar la estimación de una función de densidad.

Lo desarrollado hasta este punto permite obtener la distribución posterior de un vector de probabilidades de dimensión finita, cuando tenemos datos con distribución multinomial. Esta inferencia, se puede considerar un primer paso en la estimación de la distribución de una variable aleatoria si la misma tiene recorrido finito; tal sería el caso de una variable categórica. En el caso de una variable aleatoria que tenga densidad, lo anterior no es posible.

La forma más simple de estimar una densidad es a través del histograma. Lo planteado hasta ahora permite la construcción de un histograma bayesiano, que no es otra cosa que un modelo de conteo, en el cual se le asigna a los intervalos, probabilidades *a priori* con distribución Dirichlet, la inferencia se hará con la distribución posterior de dicho vector de probabilidades.

Más allá de su simplicidad, son conocidas las limitaciones que tiene el histograma como estimador de una densidad: la función estimada resulta constante a trozos, presenta discontinuidades en los extremos de los intervalos y, a su vez, es muy sensible a la elección de la partición.

Este escenario es el que naturalmente conduce a generalizar la distribu-

ción de Dirichlet al caso infinito-dimensional, para estimar la distribución de probabilidad de una variable aleatoria, lo que puede ser considerado un parámetro dado por un vector de dimensión infinita.

Desde una perspectiva bayesiana, si se quiere estimar una medida de probabilidad, será necesario tener una distribución previa sobre el espacio de las medidas de probabilidad. El Proceso de Dirichlet, introducido por Ferguson (1973), es precisamente, una familia de procesos estocásticos cuyas trayectorias (realizaciones) son distribuciones de probabilidad.

Previo a definir el Proceso de Dirichlet es necesario introducir el concepto de Medida de Probabilidad Aleatoria.

3.1. Medida de probabilidad aleatoria

Una medida de probabilidad aleatoria, es un proceso estocástico, cuyas realizaciones son medidas de probabilidad. A diferencia de un proceso con trayectorias en $\mathbb{R}$, este tipo de procesos se define sobre los conjuntos medibles de un espacio $\mathcal{X}$. Se considera medible respecto de una σ-álgebra $\mathcal{B}(\mathcal{X})$.

Dado el espacio de probabilidad $(\Omega, \mathcal{B}(\Omega), \mathbb{P})$, y el espacio probabilizable $(\mathcal{X}, \mathcal{B}(\mathcal{X}))$, llamaremos medida de probabilidad aleatoria a una función $\mathcal{P} : \Omega \times \mathcal{B}(\mathcal{X}) \to [0, 1]$ tal que:

- Fijado $\omega \in \Omega$: para todo $B \in \mathcal{B}(\mathcal{X})$ el mapeo $B \mapsto \mathcal{P}(\omega, B)$ es una medida de probabilidad sobre el espacio probabilizable $(\mathcal{X}, \mathcal{B}(\mathcal{X}))$.
- Fijado $B \in \mathcal{B}(\mathcal{X})$, para todo $\omega \in \Omega$ el mapeo $\omega \mapsto \mathcal{P}(\omega, B)$ es una variable aleatoria que toma valores en $[0, 1]$. Intuitivamente podemos decir que $\mathbb{P}(B, \omega)$ será la probabilidad asignada al conjunto B, en el resultado ω.

Un ejemplo de medida de probabilidad aleatoria es el proceso empírico. Informalmente, se puede decir que es una función aleatoria pues depende de una muestra aleatoria (depende de un ω), y cualquier realización (fijado ω), induce una medida de probabilidad $\mathbb{P}_n : \mathbb{P}_n(B) = \frac{1}{n} \sum \mathbb{1}_{\{X_i \in B\}}$ para cualquier $B \in \mathcal{B}(\mathbb{R})$.

En la figura 3.1, se grafica la función de distribución empírica de cinco muestras aleatorias distintas. Fijando ω (lo cual es equivalente a determinar

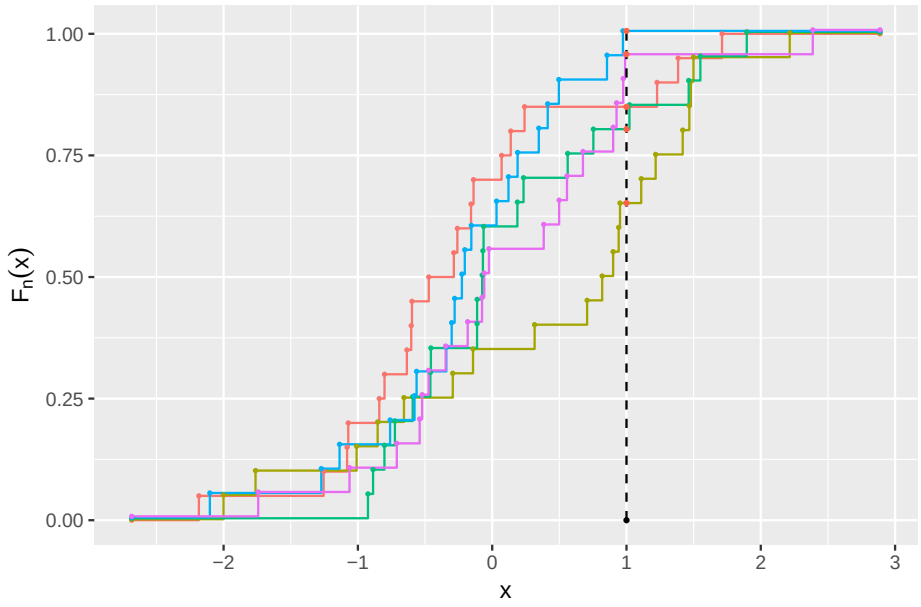


Figura 3.1: Cinco trayectorias de un proceso empírico.

un color), se tiene una función real, que representa una trayectoria del proceso indexado en los conjuntos de la forma $\{(-\infty, x] : x \in \mathbb{R}\}$. Fijando un conjunto, por ejemplo: (-∞, $\bullet$], se tiene la variable aleatoria $F_n(\omega, \bullet) \in [0, 1]$.

3.1.1. Definición

Sea $(\Omega, \mathcal{B}(\Omega), \mathbb{P})$ un espacio de probabilidad, $(\mathbb{R}, \mathcal{B}(\mathbb{R}))$ un espacio probabilizable y $\mathcal{P}$ una medida de probabilidad aleatoria sobre definida en $\Omega \times \mathcal{B}(\mathbb{R})$.

Diremos que $\mathcal{P}$ está definida por un proceso de Dirichlet de parámetros α y G_0 si para cualquier partición medible de $\mathbb{R}$ de la forma $\mathcal{A} = (A_1, \ldots, A_k)$, se cumple que:

$$\mathcal{P}(\mathcal{A}) = \Big(\mathcal{P}(A_1), \mathcal{P}(A_2), \ldots, \mathcal{P}(A_k)\Big) \sim Dirichlet(\alpha G_0(A_1), \ldots, \alpha G_0(A_k))$$

y lo notaremos: $\mathcal{P} \sim DP(\alpha, G_0)$, donde $\alpha > 0$ es un parámetro de precisión y G_0 es una medida de probabilidad base, definida en $\mathbb{R}$.

3.1.2. Estimación

Consideremos el siguiente modelo. Se tienen $X_1, \ldots, X_n \overset{ind}{\sim} \mathcal{P}$, y se desea estimar $\mathcal{P}$. Para ello, se le asigna un Proceso de Dirichlet (una medida de probabilidad aleatoria) como distribución previa: $\mathcal{P} \sim DP(\alpha, G_0)$.

En Ferguson (1973) se prueba que se trata de un modelo conjugado:

$$\mathcal{P}|\{X_1, \ldots, X_n\} \sim DP\left(\alpha + n, \frac{\alpha}{\alpha + n} G_0 + \frac{1}{\alpha + n} \sum_{i=1}^{n} \delta_{X_i}\right)$$

donde δ_x representa el *Delta de Dirac*, una distribución que concentra toda la masa de probabilidad en un átomo x.

Tomando $\mathcal{A} = (A_1, \ldots, A_k)$ una partición cualquiera, se tiene que

$$\mathcal{P}(\mathcal{A})|X_1, \ldots, X_n \sim Dirichlet\left(\alpha.G_0(A_1) + \sum_{j=1}^{n} \mathbb{1}_{\{X_j \in A_1\}}, \ldots, \alpha.G_0(A_k) + \sum_{j=1}^{n} \mathbb{1}_{\{X_j \in A_k\}}\right)$$

Consideremos particularmente la partición de la forma $\mathcal{A} = (A, A^c)$, para cualquier conjunto A medible, entonces:

$$\mathcal{P}(A) \sim Beta\left(\alpha G_0(A), \alpha G_0(A^c)\right)$$

por lo tanto

- $\mathbb{E}\left(\mathcal{P}(A)\right) = \dfrac{\alpha G_0(A)}{\alpha G_0(A) + \alpha G_0(A^c)} = \dfrac{\alpha G_0(A)}{\alpha} = G_0(A)$
- $\mathrm{Var}(\mathcal{P}(A)) = \dfrac{\alpha G_0(A) \alpha G_0(A^c)}{(\alpha G_0(A) + \alpha G_0(A^c))^2 (\alpha G_0(A) + \alpha G_0(A^c) + 1)} = \dfrac{G_0(A) G_0(A^c)}{\alpha + 1}$

Si $\alpha \to \infty$, $\mathrm{Var}(\mathcal{P}(A)) \to 0$, el proceso estará concentrado en $G_0(A)$, por eso el parámetro α es visto como un parámetro de confianza en la medida base G_0. Es común que en la práctica se interprete a α como un tamaño de muestra *a priori*.

La distribución posterior resulta:

$$\mathcal{P}(A)|X_1, \ldots, X_n \sim Beta\left(\alpha.G_0(A) + \sum_{j=1}^{n} \mathbb{1}_{\{X_j \in A\}}\ ,\ \alpha.G_0(A^c) + \sum_{j=1}^{n} \mathbb{1}_{\{X_j \in A^c\}}\right)$$

donde

$$\begin{aligned}\mathbb{E}\left(\mathcal{P}(A)|X_1, \ldots, X_n\right) &= \frac{\alpha.G_0(A) + \sum_{j=1}^{n} \mathbb{1}_{\{X_j \in A\}}}{\alpha + n} \\ &= \left(\frac{\alpha}{\alpha + n}\right) G_0(A) + \left(\frac{n}{\alpha + n}\right) \frac{\sum_{j=1}^{n} \mathbb{1}_{\{X_j \in A\}}}{n}\end{aligned}$$

Considerando particularmente $A_x = (-\infty, x]$, se tiene que $\mathcal{P}(A_x) = F(x)$ (bajo el esquema Bayesiano $F(x)$ es una variable aleatoria).

Si se quiere obtener una estimación puntual para $F(x)$ adoptando el criterio de minimizar el error cuadrático medio esperado del estimador: $\hat{F}(x) = \arg\min_G \mathbb{E}\left[F(x) - G(x)\right]^2$. Se prueba que $\hat{F}(x)$ resulta ser la esperanza de la distribución posterior:

$$\hat{F}(x) = E(F(x)|X_1, \ldots, X_n) = \left(\frac{\alpha}{\alpha + n}\right) G_0\big((-\infty, x]\big) + \left(\frac{n}{\alpha + n}\right) F_n(x) \tag{3.1}$$

que resulta ser una combinación lineal convexa entre la medida base y la función de distribución empírica.

Como se observa en (3.1), para valores grandes de n (o valores pequeños de α), la influencia de G_0 se ve disminuída, y $\hat{F}_n(x)$ se aproxima a $F_n(x)$.

Observemos esto último con un ejemplo didáctico, en el cual se simularán $x_1, \ldots, x_{100}$ con distribución exponencial de parámetro 1 y se elegirá como medida base (G_0) una distribución normal estándar (claramente, no es una buena elección). En la figura 3.2, se puede ver la influencia de la variación de α en la $\hat{F}$ resultante.

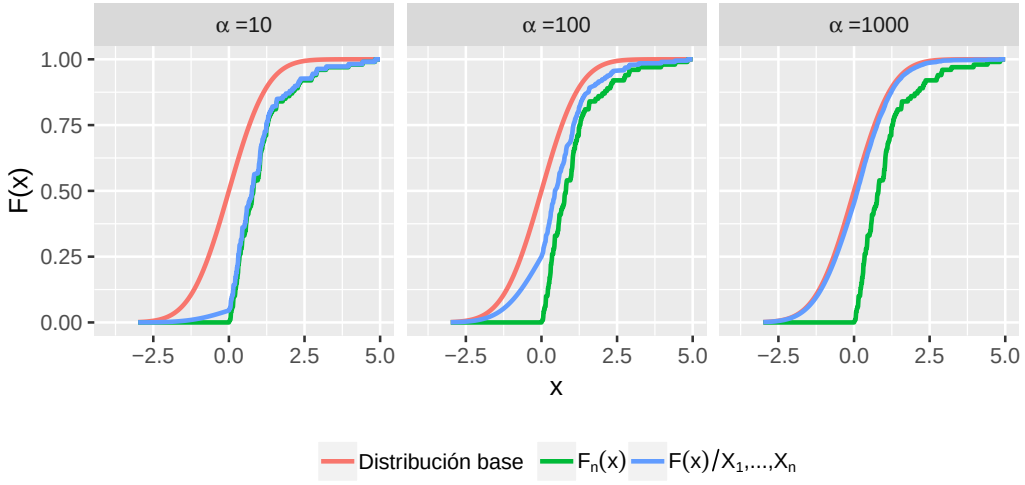


Figura 3.2: Influencia de α en la estimación puntual posterior.

3.1.3. Stick - Breaking

Las realizaciones de los procesos de Dirichlet son funciones discretas con probabilidad uno. Una prueba de esto, se debe a Sethuraman (1994), quien introdujo una forma constructiva de definir un proceso de Dirichlet, denominada *Stick-Breaking*.

Bajo esta construcción, se muestra que si $\mathcal{P}$ es un proceso de Dirichlet, entonces:

$$\mathcal{P} = \sum_{k=1}^{\infty} \pi_k \delta_{\theta_k} \tag{3.2}$$

donde $\theta_1, \theta_2, \ldots \overset{ind}{\sim} G_0$ y las probabilidades π_k se definen en base a variables auxiliares de la siguiente forma:

$$\begin{aligned} \pi_1 &:= V_1 \\ \pi_k &:= V_k \textstyle\prod_{j=1}^{k-1}(1 - V_j) \quad \text{si } k \geq 2 \end{aligned}$$

donde $V_1, V_2, \ldots$. iid, $V_i \sim Beta(1, \alpha)$. Se utilizará la notación $\boldsymbol{\pi} \sim Stick(\alpha)$ para describir la distribución conjunta del vector $\boldsymbol{\pi}$ cuando sus componentes π_h se definen a partir de la anterior construcción.

La denominación que recibe este proceso responde a una manera muy gráfica de imaginarlo. Se tiene un segmento de longitud 1, se sortea V_1, y al segmento unitario se le sustrae de uno de sus extremos, un segmento de longitud $\pi_1 = V_1$. Se repite el procedimiento, al segmento restante de longitud (1-V_1) se le sustrae, de un extremo, un segmento de longitud π_2 que representa una proporción V_2 de su longitud, y así sucesivamente. Cada segmento sustraído de longitud π_i representa la masa de probabilidad que se le asigna al punto θ_i.

Una representación gráfica de dos procesos *Stick-Breaking* distintos se puede ver en la figura 3.3, el código para generar dichos gráficos se encuentra en el anexo B.

Este proceso es una medida de probabilidad aleatoria casi seguramente discreta, ya que la probabilidad está (con probabilidad 1) concentrada en átomos.

Dado que este procedimiento es infinito, si se quieren obtener simulaciones del proceso, se asume una tolerancia ϵ y se trunca la simulación cuando $\sum_k \pi_k > 1 - \epsilon$.

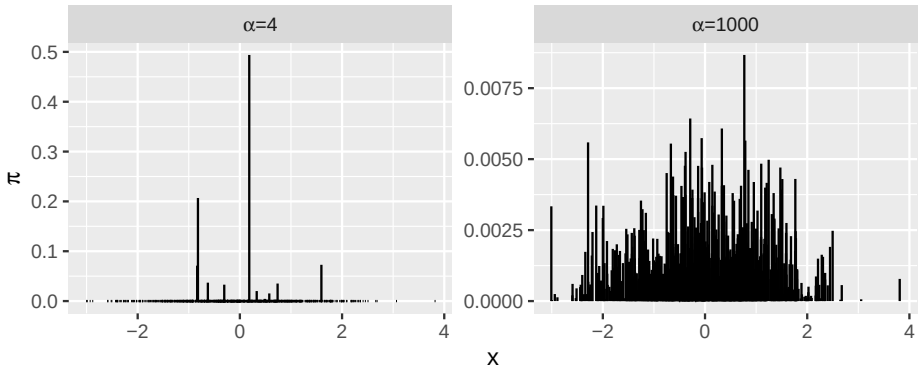


Figura 3.3: Simulaciones de dos procesos de Stick-Breaking, con $G_0 = N(0,1)$, a la izquierda: α=4; a la derecha: $\alpha = 1000$.

3.2. DP-Mixture Model

Como se mencionó en la sección anterior, los procesos de Dirichlet tienen la desventaja de tener realizaciones que son, casi seguramente, funciones discretas. Esto establece la imposibilidad de usar este método para estimar una función de densidad. Sin embargo, veremos cómo todo el herramental conceptual desarrollado hasta ahora, será de utilidad para tal propósito.

Podemos introducir una nueva clase de procesos: los *DP Mixture-Model* (DPMM), que son una familia de modelos de mezcla, donde los parámetros de la mezcla están controlados por un Proceso de Dirichlet. Si consideramos f como un DPMM entonces tenemos que:

$$f(y) = \int f_\theta(y|\theta) d\mathrm{P}(\theta) \tag{3.3}$$

donde P es un proceso de Dirichlet.

Veamos que esta formulación es equivalente a plantear el siguiente modelo jerárquico:

$$\begin{aligned} Y_i &\overset{ind}{\sim} F(\cdot|\theta_i) \\ \theta_i &\overset{ind}{\sim} G \\ G &\sim DP(\alpha, G_0) \end{aligned} \tag{3.4}$$

Este modelo indica que cada dato Y_i proviene de una subpoblación con distribución F, con parámetro θ_i, que son sorteados de una distribución G que será un proceso de Dirichlet con parámetros α y G_0, si caracterizamos el proceso de Dirichlet por el procedimiento de *Stick-Breaking*, podemos escribir

(3.4) como:

$$\begin{aligned} Y_i &\overset{ind}{\sim} F(\cdot|\theta_i) \\ \theta_i &\overset{ind}{\sim} \textstyle\sum_{h=1}^{\infty} \pi_h \delta_{\theta_h^*} \\ \theta_h^* &\overset{ind}{\sim} G_0 \\ \boldsymbol{\pi} &\sim Stick(\alpha) \end{aligned} \tag{3.5}$$

Por otro lado, podemos ver que como, $\sum_{h=1}^{\infty} \pi_h = 1$, dado $\epsilon > 0$ existe H tal que $\sum_{h=1}^{H} \pi_h > 1 - \epsilon$, por tanto podemos truncar el procedimiento de Stick-Breaking de forma tal de tener una cantidad finita de átomos θ_h y podemos aproximar la mezcla infinita por una mezcla finita.

3.2.1. DPMM-Normal

Si se toma $F(\cdot)$ la distribución normal, y se utiliza el esquema *normal-inversa gamma* como distribución base $G_0(\mu, \sigma)$ para el proceso de Dirichlet, el modelo queda definido de la siguiente forma:

$$\begin{aligned} Y_i &\overset{ind}{\sim} \mathcal{N}(\mu_i, \sigma_i^2) \\ (\mu_i, \sigma_i^2) &\sim G \\ G &\sim DP(\alpha, G_0) \\ G_0(\mu, \sigma^2) &= N(\mu|m, \tfrac{1}{k}\sigma^2) \times IG(\sigma^2|\nu, \psi) \\ m \in \mathbb{R},\ &\alpha, k, \nu, \psi \in \mathbb{R}^+ \end{aligned} \tag{3.6}$$

Usando la representación de un Proceso de Dirichlet a través del procedimiento de *Stick-Breaking* sobre la medida base G_0 se obtiene equivalentemente:

$$\begin{aligned} Y_i &\overset{ind}{\sim} \mathcal{N}(\mu_i, \sigma_i^2) \\ (\mu_i, \sigma_i^2) &\sim \textstyle\sum_{h=1}^{\infty} \pi_h \delta_{(\mu_h^*, \sigma_h^{2*})} \\ (\mu_h^*, \sigma_h^{2*}) &\sim N(\mu|m, \tfrac{1}{k}\sigma^2) \times IG(\sigma^2|\nu, \psi) \\ \boldsymbol{\pi} &\sim Stick(\alpha) \\ m \in \mathbb{R},\ &\alpha, k, \nu, \psi \in \mathbb{R}^+ \end{aligned} \tag{3.7}$$

Una realización de *Stick-Breaking* con la distribución de base G_0 que aparece en el modelo (3.6) se puede ver en la figura 3.4. En el anexo B se encuentra el código para generar dicho gráfico.

Tanto (3.6) como (3.7) resultan modelos jerárquicos. Tal como se expone en Hjort et al. (2010) y en Müller et al. (2015), se puede agregar un nivel

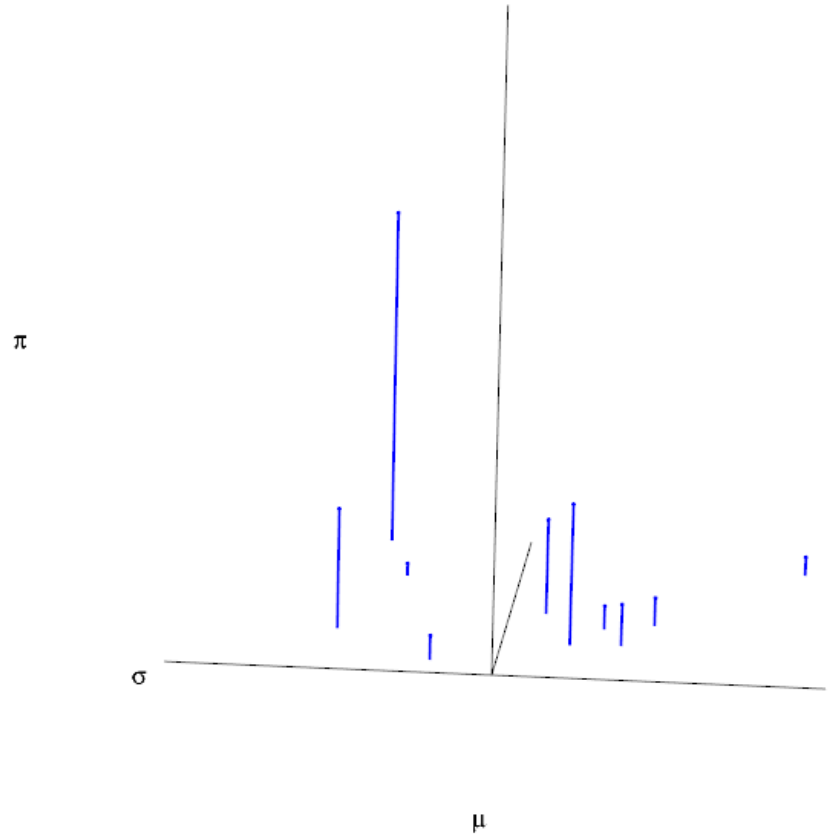


Figura 3.4: Una realización de *Stick-Breaking* bidimensional, para los parámetros μ y σ con la distribución de base indicada en el modelo (3.6)

más de aleatoriedad en los mismos, asignando distribuciones previas para los parámetros α, m, k, ν, ψ. De este modo lograremos mayor flexibilidad y se disminuirá el impacto que tenga la elección de la previa en la estimación, permitiendo que el modelo *aprenda* de los datos. El costo de este procedimiento se traduce en mayor complejidad y tiempo computacional.

4

Implementación computacional

"Horas, horas.
Colgados como dos computadoras."

Jorge Drexler. *Horas*

En este apartado mostraremos los aspectos computacionales de esta técnica. En la sección 2.3.1 se mostró la implementación del algoritmo de muestreo de Gibbs en un modelo de mezcla finito a través de un ejemplo sencillo; aquí veremos cómo se puede derivar una equivalencia entre el modelo (3.6) y un modelo de mezcla finito, que será una generalización de lo visto en 2.3.1. Luego ejemplificaremos la implementación a través de un conjunto con datos simulados,

El lenguaje utilizado fue **R** (R Core Team, 2018), y se utilizó el software **jags** (Plummer, 2003) y **rjags** (Plummer, 2018), para implementar el algoritmo de Gibbs.

4.1. Estimación de la posterior

Recordemos que en el modelo (3.7), se recurre a la caracterización de un Proceso de Dirichlet mediante *Stick-Breaking*. Por otro lado, dado $\epsilon > 0$,

existe H tal que $\sum_{h=1}^{H} \pi_h > 1 - \epsilon$ por tanto se puede hacer la aproximación

$$G = \sum_{h=1}^{\infty} \pi_h \delta_{\theta_h^*} \approx \sum_{h=1}^{H} \pi_h \delta_{\theta_h^*}$$

Este truncamiento implica que $(\mu_i, \sigma_i^2) \sim \sum_{h=1}^{H} \pi_h \delta_{(\mu_h^*, \sigma_h^{*2})}$ de donde tenemos que (μ_i, σ_i^2) provienen de una mezcla finita de distribuciones *Delta de Dirac*[1], por tanto provienen de una distribución discreta. Si miramos la i-ésima dupla (μ_i, σ_i^2) podemos decir que la misma proviene de algún componente de dicha mezcla. Para enfocar este concepto, introduciremos ciertas variables latentes ξ_i que indiquen para cada vector (μ_i, σ_i^2) cuál es el componente que lo generó. $Rec(\xi_i) = \{1, \ldots, H\}$. Condicionando en el valor de esta variable, se sabrá la distribución del i-ésimo átomo, que con probabilidad uno será un átomo (μ_h^*, σ_h^{2*}). Esto es, $(\mu_i, \sigma_i^2)|\{\xi_i = h\} \sim \delta_{(\mu_h^*, \sigma_h^{2*})}$, de donde se desprende que $(\mu_i, \sigma_i^2)|\{\xi_i = h\} \overset{(cs)}{=} (\mu_h^*, \sigma_h^{2*})$ y por tanto $Y_i|\{\xi_i = h\} \sim N(\mu_h^*, \sigma_h^{2*})$.

Para completar esta formulación resta decir que los ξ_i serán independientes y seleccionarán un componente de la mezcla con probabilidad equivalente al peso del componente, es decir, $\mathbb{P}(\xi_i = h) = \pi_h$ para $h = 1, \ldots, H$. Al cabo de todo este análisis, el modelo puede considerarse equivalente a un modelo de mezcla finito:

$$\begin{aligned} Y_i &\overset{ind}{\sim} \sum_{h=1}^{H} \pi_h N(\mu_h^*, \sigma_h^{2*}) \\ \mu_h^*|\sigma_h^* &\overset{ind}{\sim} N(m, \frac{1}{k}\sigma_h^{2*}) \\ \sigma_h^{2*} &\overset{ind}{\sim} IG(\nu, \psi) \\ \boldsymbol{\pi} &\sim Stick(\alpha) \end{aligned} \tag{4.1}$$

Es relevante, conceptualmente hablando, notar la importancia que tiene el truncamiento finito en el procedimiento de *Stick-Breaking* para la derivación de la equivalencia entre (3.7) y (4.1). También es pertinente observar que lo que era la distribución base (G_0) en los modelos (3.6) y(3.7), termina siendo una distribución previa en el modelo (4.1).

Dado que, tras cuidadosas derivaciones, finalmente se obtiene un modelo de mezcla finito de normales, la estimación de la distribución posterior de

[1]Una distribución donde toda la masa de probabilidad está concentrada en un átomo

los parámetros, se hace a través del algoritmo de Gibbs para tal caso. Un ejemplo sencillo (y simplificado) de implementación en **R** se detalla en la sección 2.3.1.

4.2. Obtención del estimador puntual e intervalos de credibilidad

Dado $y \in \mathbb{R}$, el estimador de $f(y)$, resultará:

$$\hat{f}(y) = \sum_{h=1}^{H} \pi_h \phi(y; \mu_h, \sigma_h)$$

Como μ_h, σ_h y π_h son de naturaleza aleatoria, también aleatorio será $\hat{f}(y)$.

A través del algoritmo de Gibbs se obtiene muestra de la distribución posterior de los parámetros μ_h, σ_h y π_h, (para $h = 1, \ldots, H$), es decir, tenemos: $\{\mu_h^{(k)}, \sigma_h^{(k)}, \pi_h^{(k)}\}_{k=1,2,\ldots,m}$, que da lugar a la cadena de estimaciones $\{\hat{f}(y)^{(k)}\}_{k=1,2,\ldots,m}$ donde $\hat{f}(y)^{(k)} = \sum_{h=1}^{H} \pi_h^{(k)} \phi(y; \mu_h^{(k)}, \sigma_h^{(k)})$.

Dado y, el estimador puntual $\hat{f}(y) = \hat{\mathbb{E}}\left[f(y)|y_1, \ldots, y_n\right]$ se puede obtener a través de *promediar en k*, es decir:

$$\hat{f}(y) = \frac{1}{m} \sum_{k=1}^{m} \hat{f}(y)^{(k)} = \frac{1}{m} \sum_{k=1}^{m} \left[\sum_{h=1}^{H} \pi_h^{(k)} \phi(y; \mu_h^{(k)}, \sigma_h^{(k)}) \right]$$

Si en lugar de tomar el promedio, se consideran los percentiles $\alpha/2$ y $1 - \alpha/2$ se obtienen intervalos de credibilidad $1 - \alpha$ para $f(y)$.

4.3. Un ejemplo con datos simulados

Simularemos datos bajo un modelo con la siguiente de densidad de probabilidad (su gráfico se muestra en la figura 4.1):

$$f(x) = \frac{1}{3}\phi(x; -4, 0.1) + \frac{1}{3}\phi(x; -2, 0.5) + \frac{1}{3}\phi(x; 2, 1)$$

y luego estimaremos la función de densidad a partir de ellos.

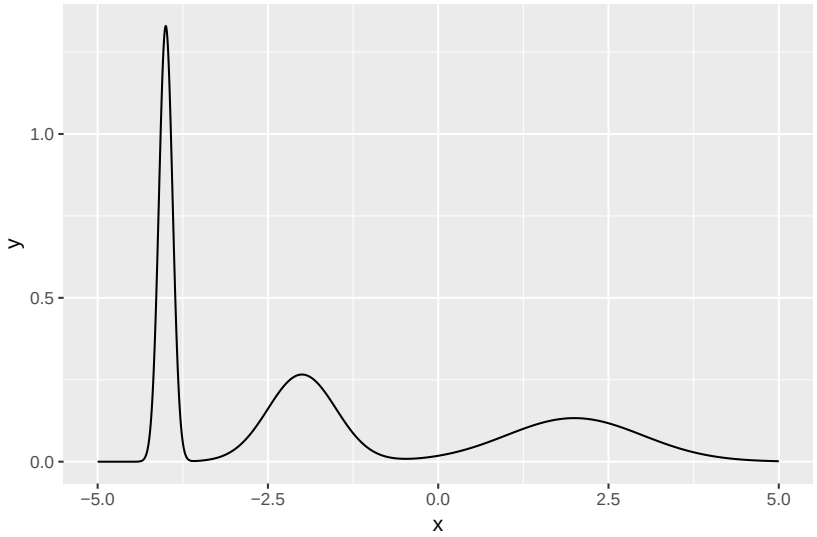


Figura 4.1: Gráfico de $f(x) = \frac{1}{3}\phi(x; -4, 0.1) + \frac{1}{3}\phi(x; -2, 0.5) + \frac{1}{3}\phi(x; 2, 1)$

Con este ejemplo se evidenciará, tal como se expone en Escobar & West (1995), una ventaja de DPM respecto del estimador por núcleos en su versión clásica: la posibilidad de tener un *ancho de banda* variable. Mientras que en la tradicional estimación por núcleos, la dispersión de las funciones núcleo es única, y su elección determina el suavizado de la estimación, esta técnica permite que la dispersión de los distintos componentes de la mezcla varíe. En la figura 4.2, se muestran dos estimadores por núcleo, cuyo ancho de banda fue elegido según dos criterios optimales distintos (en un caso validación cruzada, en otro la regla de referencia normal). Para computar dicha estimación se llamó a la función `density()` de **R**.

Como se aprecia en dicho gráfico, ninguno de los dos estimadores logra el suavizado deseado. Esto es una limitante de estos métodos con ancho de banda constante. Tengamos en cuenta que esta población está formada por tres subpoblaciones con distinta dispersión.

4.3.1. Elección de la previa

Debemos especificar la distribución previa para los parámetros α, m, k, ν, ψ en el modelo (4.1). En el cuadro 4.1 se proponen 4 combinaciones distintas para los parámetros de dichas distribuciones previas.

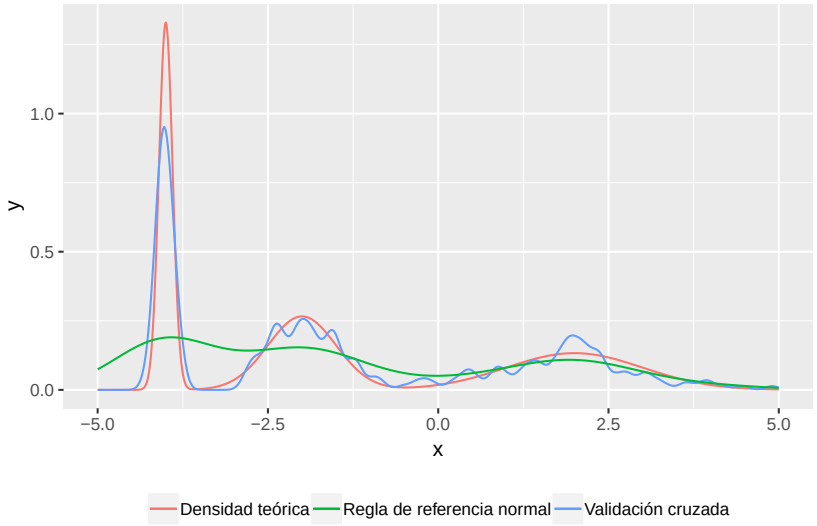


Figura 4.2: Estimadores kernel con ancho de banda óptimo según distintos criterios.

Previa 1	$\alpha = 1$	$m = -1$	$k = 1$	$\nu = 0.1$	$\psi = 0.1$
Previa 2	$\alpha = 1$	$m = -1$	$k = 0.01$	$\nu = 0.1$	$\psi = 0.1$
Previa 3	$\alpha = 1$	$m = -1$	$k = 20$	$\nu = 1000$	$\psi = 10$
Previa 4	$\alpha = 1$	$m = -1$	$k = 0.1$	$\nu = 1000$	$\psi = 10$

Cuadro 4.1: Previas usadas en el ejemplo

En la figura 4.3 se puede apreciar cómo varía el estimador según los distintos valores propuestos. Allí se ve que el estimador con la previa 1 resulta óptimo, con la previa 3 produce un estimador muy *rugoso*, mientras que las previas 2 y 4 resultan en un sobresuavizado.

Esto nos habla de la sensibilidad que tiene el estimador respecto de los parámetros de la previa elegida. No debemos alarmarnos por esto, en el caso de los estimadores por núcleo, se encuentra una fuerte dependencia del estimador respecto del parámetro h de suavizado, claro que hay métodos (validación cruzada, por ejemplo) para encontrar un valor óptimo. En nuestro contexto, tal como se expuso en la sección 3.2.1, se puede asignar a estos parámetros una distribución de probabilidad, agregando una escala más de jerarquía en el modelo.

Claramente, la previa 1 resulta adecuada para este caso. Con ella se tra-

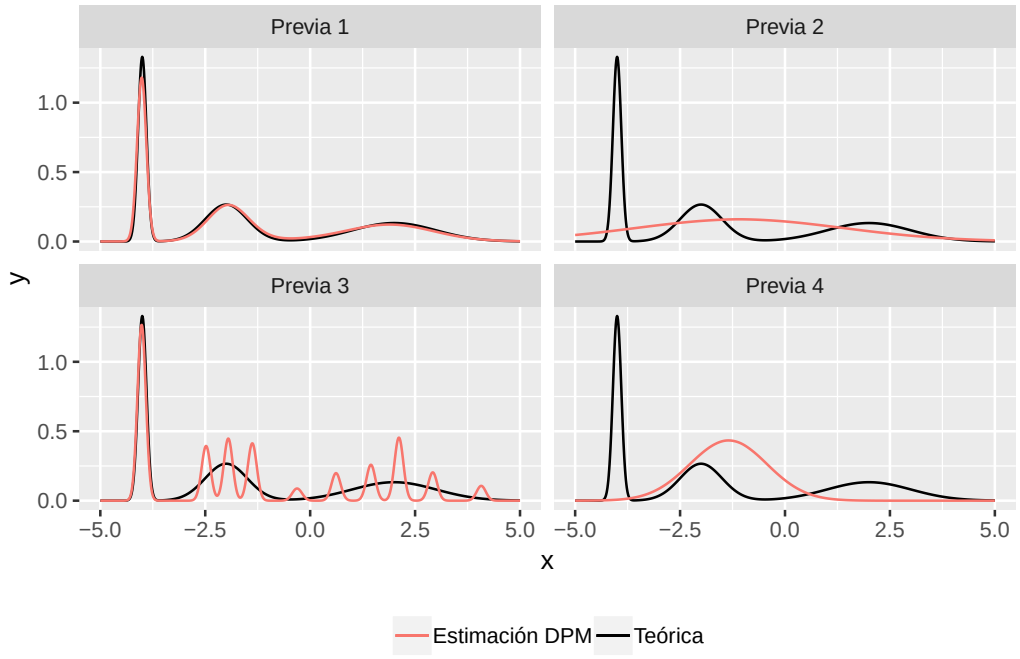


Figura 4.3: Estimaciones DPM con distintas previas

bajará en lo que resta del capítulo. En la figura 4.4 se muestran los intervalos de credibilidad al 95 % para la densidad.

4.4. Monitoreo de la convergencia

Como resultado del algoritmo de Gibbs se obtiene una muestra de tamaño m de cada parámetro. Para cada componente h de la mezcla se tienen 3 parámetros $\{\pi_h, \mu_h, \sigma_h\}$, así que en total tendremos $H \times 3$ parámetros. Adicionalmente se obtienen simulaciones de las variables latentes ξ_i que no tendrán relevancia a efectos de la estimación y más adelante explicaremos el interés por ellas.

Por cuestiones de identificabilidad el monitoreo de la convergencia no debe efectuarse para cada paramétro μ_h, σ_h, π_h. Pues a lo que se le llama la componente 1 en la iteración k, $\mu_1^{(k)}$, no tiene porqué ser la componente 1 en la iteración $k+1$, $\mu_1^{(k+1)}$. Dentro de cada iteración $\{\mu_h^{(k)}, \sigma_h^{(k)}, \pi_h^{(k)}\}$ funcionan como una terna pareada, cada una hace referencia a la posición, escala y

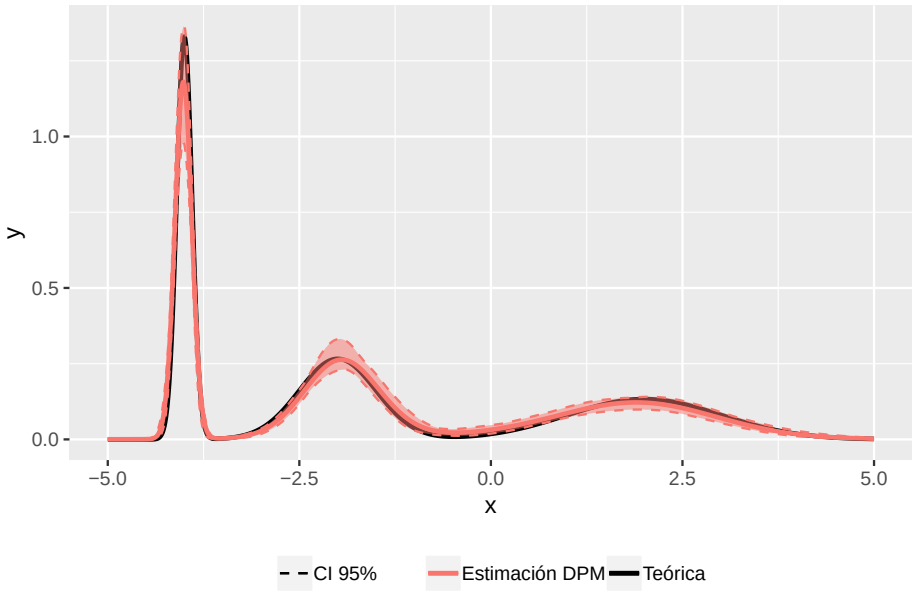


Figura 4.4: Bandas de confianza para la densidad

peso, respectivamente, de cada componente en la iteración. Consideremos un ordenamiento de los componentes según su peso, de modo que $\pi_1 \geq \pi_2 \geq \cdots \geq \pi_h$. En la figura 4.5 se muestran dos realizaciones de las cadenas $(\mu_h^{(k)}, \pi_h^{(k)})_{\substack{k=1,\ldots,1000 \\ h=1,2,3,4}}$ (la posición y el peso de los 4 átomos predominantes).

Es interesante ver que en la primer cadena la situación parece *muy ordenada*: hay 3 componentes predominantes (sus pesos sumados superan 0.9) y a lo largo de las iteraciones todos permanecen indexados en la misma forma. La cuarta componente ya presenta un peso significativamente menor, y se ve que su posición tiene una variablidad considerablemente mayor a las restantes, sin llegar a estabilizarse. Esto se puede interpretar como que las primeras tres componentes ya son suficientes para explicar distribución de los datos, y la cuarta permanece en una especie de búsqueda infructuosa. El caso de la segunda cadena es bien distinto, es curioso ver como algunos átomos *se cruzan* a lo largo de las iteraciones, se pueden reconocer las mismas tres componentes que en la otra cadena, pero no reciben el mismo índice a lo largo de las iteraciones. Ambas situaciones son normales. La situación aparentemente *caótica* del segundo caso indica en cierta forma un estado de *buena salud* del algoritmo, que explora la distribución posterior de los parámetros.

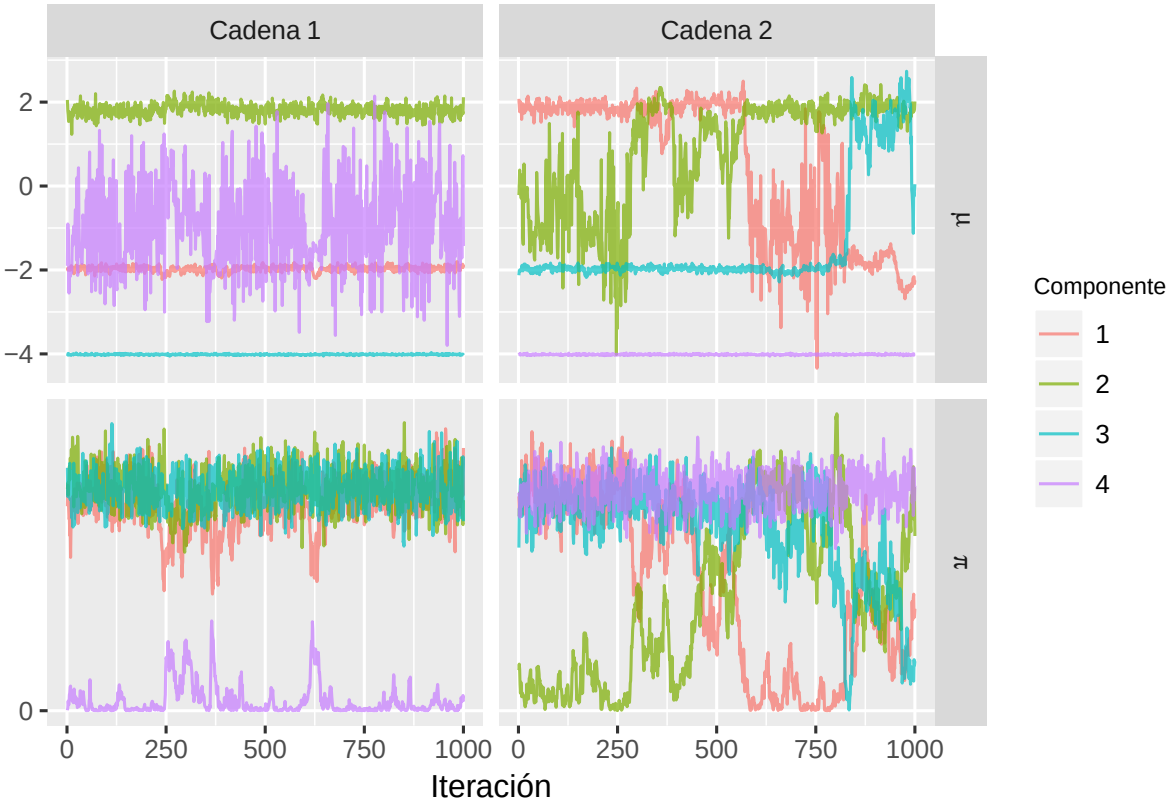


Figura 4.5: Gráfico de dos cadenas de muestras de la distribución posterior (μ_h, π_h) para los 4 componentes predominantes.

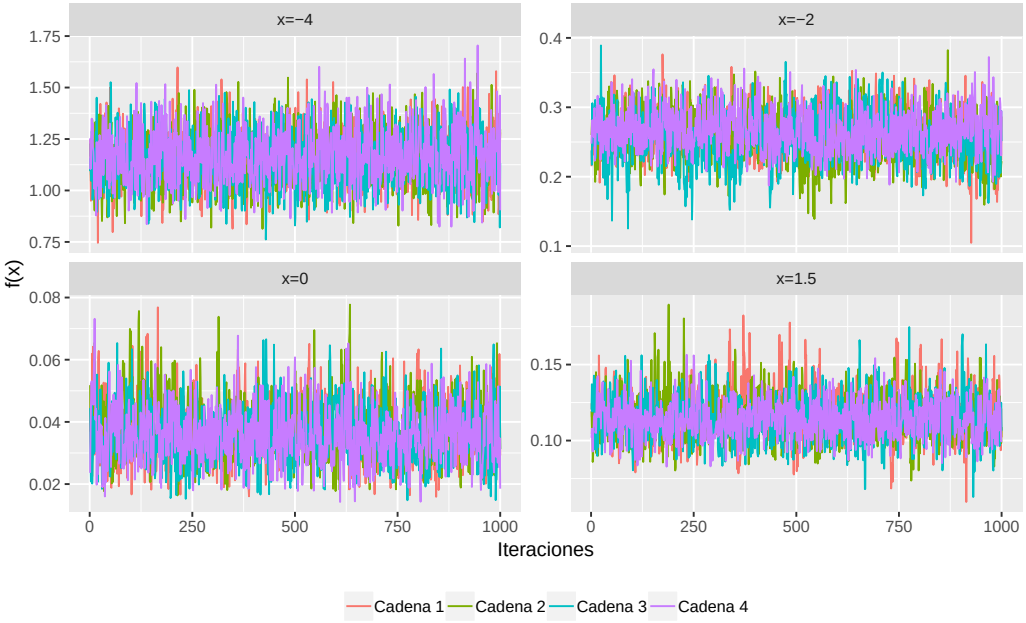


Figura 4.6: Monitoreo de convergencia en $x = -4, -2, 0, 1.5$

Para lidiar con esta situación y poder ver el comportamiento en conjunto de todos los parámetros es que no se sugiere observarlos por separado sino monitorear $f(y)$ que será una combinación lineal de ellos. Para ello basta con considerar algunos valores de y y observar el proceso $(f(y)^{(k)})_{k=1,\ldots,m}$. (Gelman et al., 2013) y (Niemi, 2017a). En la figura 4.6 se presenta la densidad estimada en los valores $x = -4, x = -2, x = 0, x = 1.5$ y no se observa indicio alguno de un problema en la convergencia.

Por otro lado, en la sección 3.2.1 se vio que el modelo propuesto se podía considerar como la mezcla de una cantidad finita de componentes. Se debe evaluar si dicha aproximación induce a un error de estimación.

En ese entendido, se asume que de la mezcla de infinitas componentes, es relevante sólo una cantidad finita de ellas. En la formulación del modelo se introdujeron variables latentes ξ_i que indican de cuál de las H componentes proviene el dato y_i. Si denotamos $\xi_i^{(k)}$ al valor muestreado de dicha variable en la iteración k del algoritmo de Gibbs, la variable definida como: $\text{máx}_{1\leq i\leq n}\{\xi_i^{(k)}\}$ indicará el número de componentes *activas* en la k-ésima iteración, es decir, el número de componentes distintas a los cuales al menos un dato les es asignado.

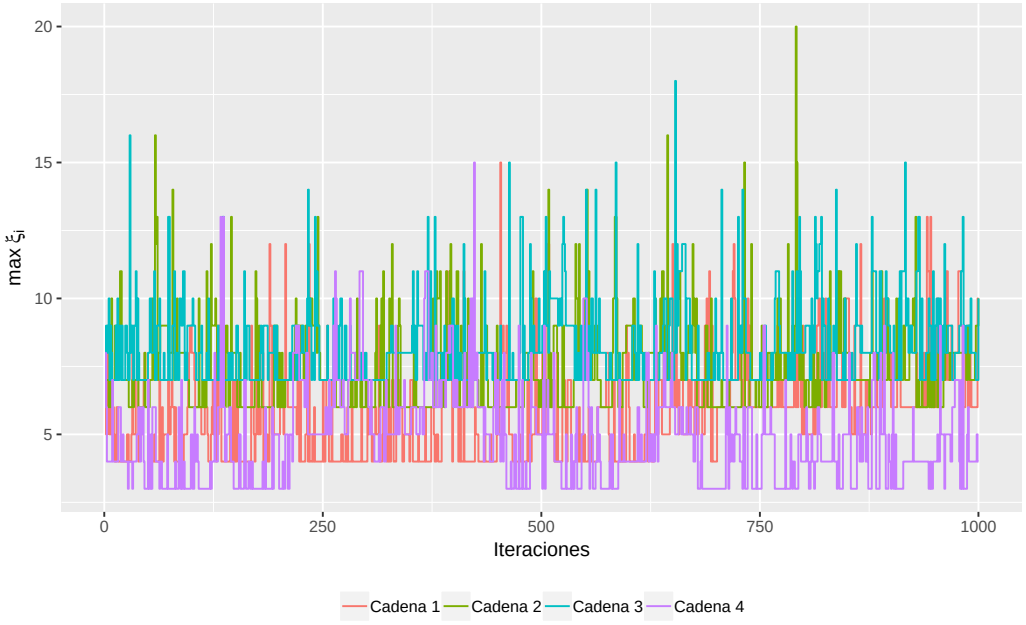
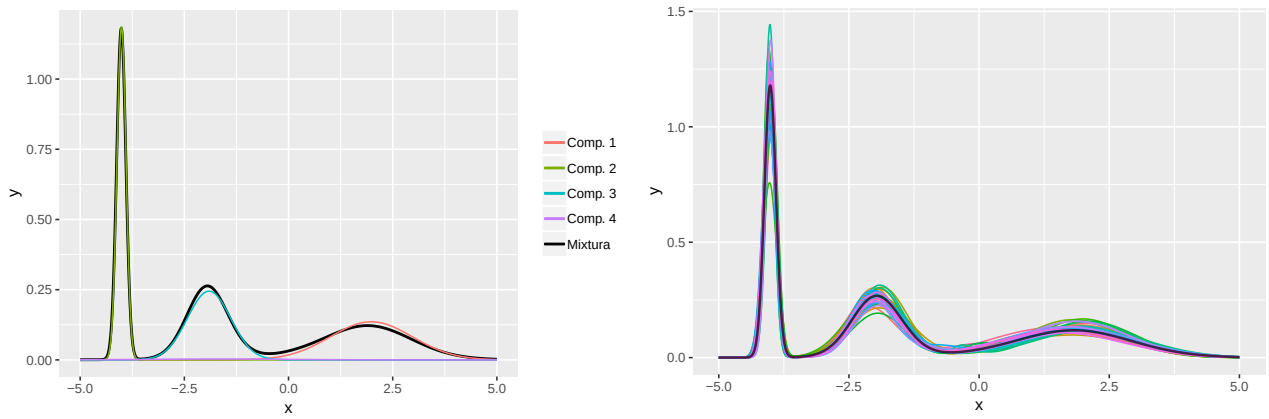


Figura 4.7: Número de átomos ocupados

De esta forma, si el número de componentes activas no es cercano a H, no hay indicios de que el truncamiento finito del proceso induzca a error, mientras que si el número de componentes activas es cercano a H, se sugiere incrementar el valor de H de forma de permitir la introducción de más componentes en la mezcla. En la figura 4.7 se presenta el gráfico de $\text{máx}_{1\leq i\leq n}\{\xi_i^{(k)}\}$ en cada iteración. Se ve que el número de componentes activas en ningún momento se concentra cerca del valor seleccionado para H $(H = 25)$.

En la figura 4.8a se puede ver en una iteración, cómo son las 4 componentes de mayor peso y cómo es la mezcla resultante de la combinación lineal de todas las componentes (no sólo esas 4). En la figura 4.8b se muestra cómo varía la función de densidad estimada a lo largo de 40 iteraciones. La convergencia anteriormente estudiada, es la convergencia puntual de esta sucesión de funciones. El promedio puntual de estas funciones es el estimador puntual de $f(y)$. Bien se podría tomar otro estimador, por ejemplo, definir una función de profundidad y elegir como estimador puntual a la curva de la sucesión que maximice dicha medida.

En el anexo B se encuentra el código en **R** para realizar todos estos procedimientos.



(a) Componentes de la mixtura en una iteración (b) Gráfico de 40 iteraciones del estimador $f(y)^{(k)}$

Figura 4.8

5

Estudio de simulación

Si supieras
un día serás de verdad
y habrá quien me quiera.

– Eduardo Mateo. *Quien te viera*

El objetivo de esta sección es evaluar la calidad de las estimaciones realizadas mediante la técnica descrita y su comparación con el método de los estimadores por núcleo. Para tal fin, siguiendo lo propuesto por Bourel & Cugliari (2018), simulamos computacionalmente distintos conjuntos de datos y estimamos su función de densidad con ambos métodos. Como en estos casos, el modelo de probabilidad que genera los datos es conocido, es posible medir el error que comete cada estimador. La medida de error que utiizaremos será el *error cuadrático medio integrado* (MISE). Dado que se está estimando un parámetro infinito-dimensional (una función), se integra en todo su dominio la diferencia cuadrática entre el verdadero valor f y el estimado $\hat{f}$. Dado que la función $\hat{f}(x)$ es aleatoria, dicha integral será aleatoria, es por eso se considera su valor esperado.

$$\text{MISE}(\hat{f}) = \mathbb{E}\left[\int_{\mathbb{R}} (\hat{f}(x) - f(x))^2 dx\right]$$

Considerando que se desconoce la distribución de esta variable aleatoria, su esperanza la estimaremos por simulación *Monte-Carlo*. A través de m (en este caso m=100) conjuntos de datos (cada uno de tamaño $n = 300$) simulados

bajo la misma distribución, obtendremos m estimadores de densidad, para cada uno de ellos computaremos el valor del error cuadrático integrado (ISE):

$$\text{ISE}(\hat{f}) = \int_{\mathbb{R}} (\hat{f}(x) - f(x))^2 dx.$$

La estimación del MISE entonces resulta de promediar los errores cuadráticos integrados.

Para el cálculo computacional en **R** de dichas integrales se creó una función auxiliar que llama a la función `integrate()`. Dicha función auxiliar se aplicó a cada estimador para cada conjunto de datos, y luego se promedió el valor de las integrales.

Para generar los datos se consideraron tres de las distribuciones presentadas en Bourel & Cugliari (2018) y la distribución presentada en la sección 4. Estas son:

- χ^2_{10}: Distribución $\chi^2_1 0$. $f(x) = \frac{x^4 e^{-\frac{x}{2}}}{2^5 \Gamma(5)} \mathbb{1}_{\{[0,+\infty)\}}(x)$
- Claw: $f(x) = \frac{1}{2}\phi(x;0,1) + \frac{1}{10}\sum_{j=0}^{4} \phi(x; \frac{j}{2} - 1, \frac{1}{10})$
- Mix_1: $f(x) = \frac{1}{3}\phi(x;-4,0.1) + \frac{1}{3}\phi(x;-2,0.5) + \frac{1}{3}\phi(x;2,1)$
- Mix_2: $f(x) = \frac{1}{2}\mathbb{1}_{\{[-2,-1]\}}(x) + \frac{1}{2}\mathbb{1}_{\{[1,2]\}}(x)$

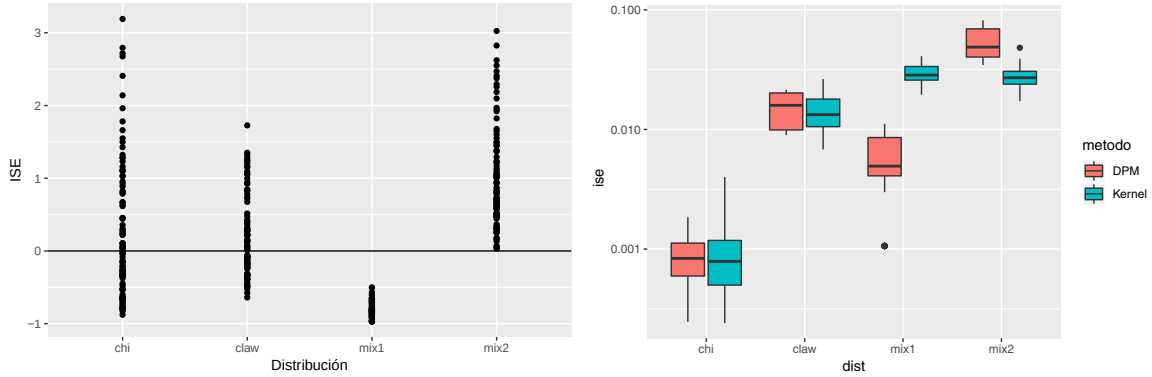
En la figura 5.2 se grafican estas funciones. $\phi(x;\mu,\sigma)$ hace referencia a la función de densidad normal, con media μ y desvío σ.

En el cuadro 5.1 se presenta, para cada caso, el valor de MISE estimado. También se presenta el valor estimado de MISE que se obtiene de los estimadores por núcleos. En la figura 5.1a se muestra la diferencia relativa en el error cuadrático integrado entre ambos métodos. Esta diferencia se ha computado de la siguiente forma: $\frac{ISE_{DPM} - ISE_{Ker}}{ISE_{Ker}}$. En la figura 5.1b se muestran los diagramas de caja con los valores de ISE de cada técnica (en escala logarítmica).

Es de esperar que el error del estimador, dependa de la distribución de los datos, para testear esta cuestión, se incluyen entre los casos de prueba algunos modelos bien distintos entre sí, algunos más *exigentes* que otros. Es bastante razonable pensar que se cometerá mayor error al estimar una densidad como *Claw*, que es multimodal y poco regular, que al estimar la densidad χ^2, que es unimodal y más *suave*.

	DPM	Kernel
χ^2_{10}	**8.53E-04**	9.42E-04
Claw	1.52E-02	**1.44E-02**
Mix_1	**5.71E-03**	2.94E-02
Mix_2	5.32E-02	**2.75E-02**

Cuadro 5.1: Estimación del MISE en cada uno de los casos



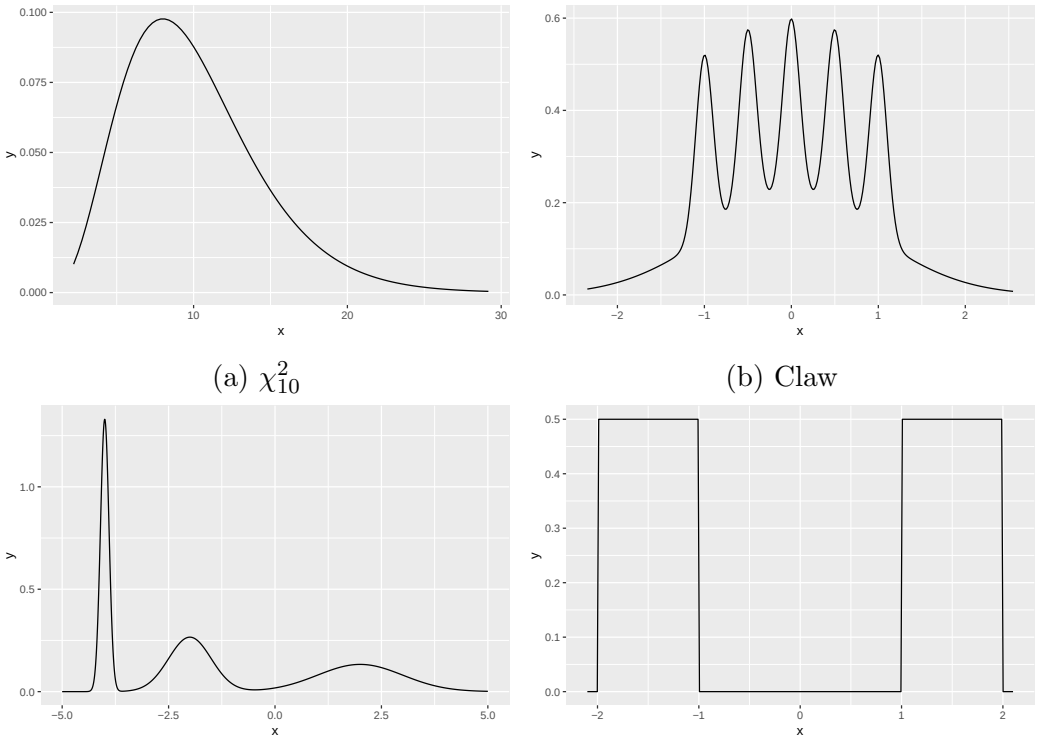
(a) Diferencia relativa entre el error de DPM y Kernel: $\frac{ISE_{DPM}-ISE_{Ker}}{ISE_{Ker}}$

(b) Gráficos de caja de los valores del error cuadrático integrado

Figura 5.1

Los casos de Mix_1 y Mix_2, son una mixtura de dos componentes, la densidad resultante en el primer caso es continua mientras que en el segundo no, y como las estimaciones resultan ser funciones continuas, es de esperar un mejor desempeño para la mezcla de normales que para la de uniformes. Por ejemplo, si en un punto x_0 la función teórica, presenta un salto finito de magnitud d, se puede asegurar que en un entorno de x_0 el estimador (continuo) cometerá un error mayor o igual a $d/2$, tal es el caso de Mix_2, para la cual esperamos un peor desempeño en relación a Mix_1.

Para la simulación de los datos usamos las funciones del paquete **stats** de **R**. Para obtener la estimación por núcleos se utilizó la librería **ks** (Duong, 2019) de **R**. En la figura 5.3 se presentan los resultados de la estimación por ambos métodos, así como los intervalos de credibilidad par el estimador *DPM* y las verdaderas densidades.



(a) χ^2_{10} (b) Claw

(c) Mix_1 (d) Mix_2

Figura 5.2: Funciones de densidad a ser estimadas

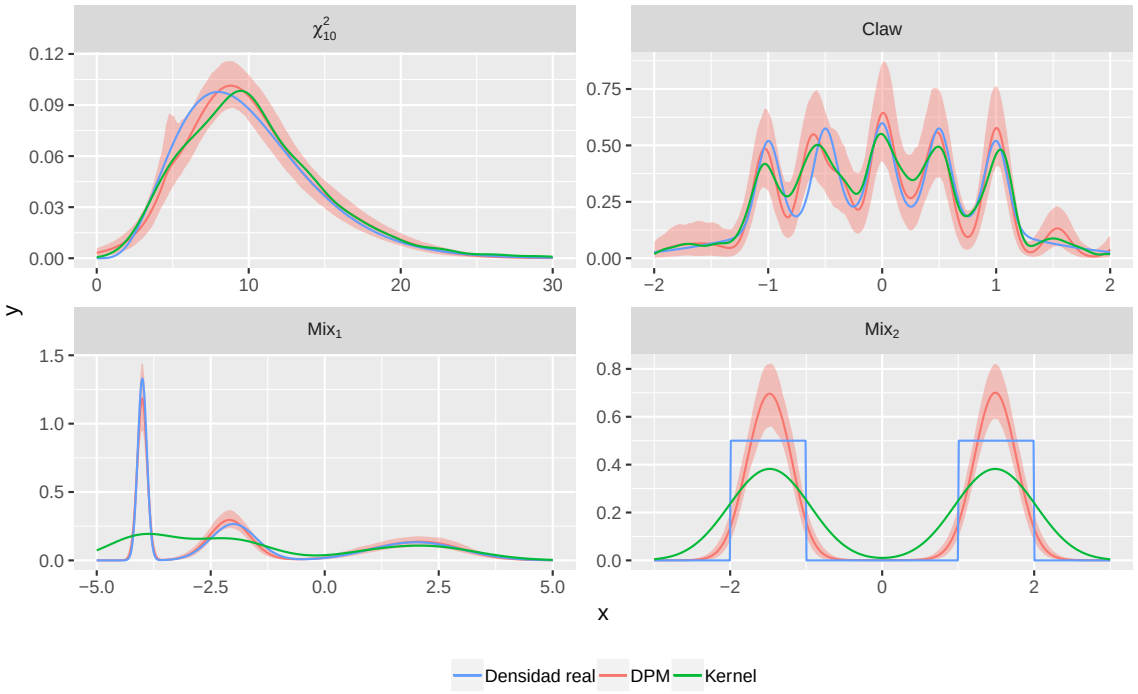


Figura 5.3: Comparación de estimaciones producidas por el método de Kernel y por DPM

6

Distribución de temperaturas máximas en Uruguay

"En el invierno
Cuando los campos tristecen
Tramo esta canción
Pensando en vos
(...)
En el verano
Cuando los días son largos
Siento que tal vez la
Terminaré"

Eduardo Darnauchans. *1959*

En el presente capítulo aplicamos la técnica descrita a lo largo del trabajo, a un conjunto de datos reales. Contamos con datos de temperatura máxima en Uruguay, relevados en la estación meteorológica de Estanzuela, desde 1950 a 2014, a resolución diaria.

Se consideran dos períodos de tiempo, el comprendido entre los años 1950 y 1980, y el comprendido entre 1990 y 2014. Se buscará estimar la densidad de la temperatura máxima en dichos períodos para efectuar una comparación.

Los datos temporales, tienen una componente de dependencia muy importante. Hay múltiples metodologías para lidiar con este asunto, para un desarrollo teórico del tratamiento de valores extremos ver Coles et al. (2001)

y Cardarello & Luraghi (2019).

En este caso, se utilizó el método del umbral y posteriormente se aplicó el procedimiento de *declustering*. Se consideraron como excedencias a las observaciones que superaran el percentil 90 de la distribución empírica de los datos, y se consideraron los clusters como las rachas de excedencias ocurridas en días consecutivos. Dicho de otra forma, para que dos excedencias no pertenezcan al mismo cluster, debe haber al menos un día en los cuales la temperatura observada fue inferior al umbral fijado. Finalmente, dentro de cada cluster se tomó la excedencia máxima (señaladas en color rojo en la figura 6.1), y el conjunto de dichos máximos fue el conjunto de datos con los cuales se estimó la densidad.

Para implementar dicho procedimiento en **R** se utilizó la función `decluster` del paquete **extRemes** (Gilleland & Katz, 2016).

En la figura 6.2 se presenta el gráfico de las densidades estimadas (con sus intervalos de credibilidad al 95 %) para los períodos mencionados.

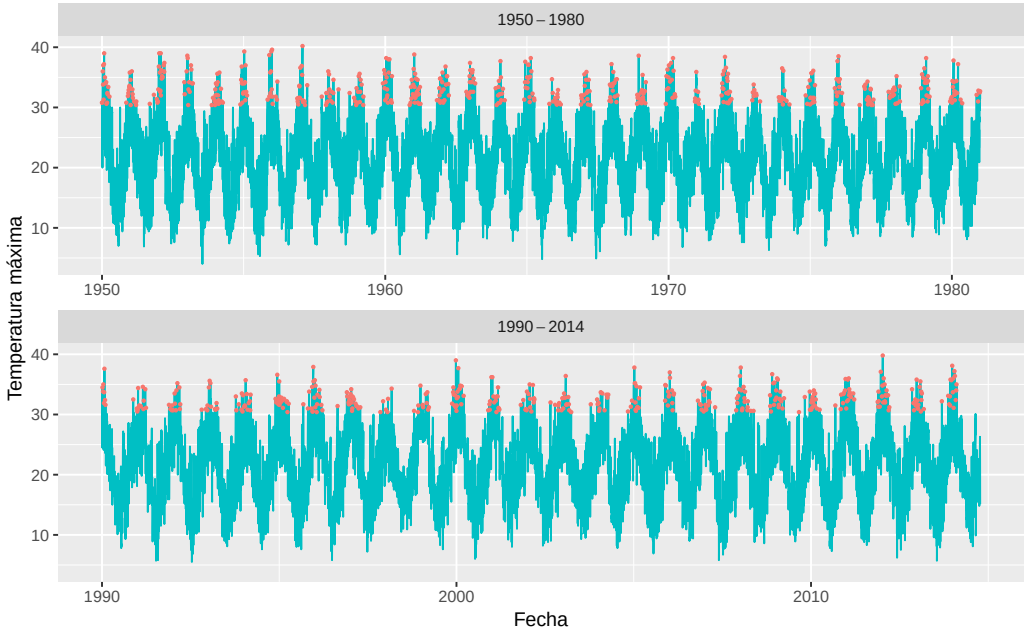


Figura 6.1: Gráfico de la serie de temperaturas por período. En • se señalan los máximos de cada cluster de excedencias.

A través de la estimación de estas densidades se pretende analizar posibles

diferencias entre la distribución de las temperaturas máximas en los distintos períodos.

La cola izquierda de la distribución para ambos períodos es similar. Podemos encontrar un valor, cercano a los 31º, tal que la probabilidad de observar un valor menor al mismo es prácticamente igual en ambas distribuciones. Por otro lado, entre los 34º y los 35º hay un punto de corte, que determina que la probabilidad de observar temperaturas máximas entre 31º y 34º es mayor en la actualidad que en el período anterior. Finalmente, la diferencia más notoria se encuentra en la cola derecha, la densidad en el período 1990-2014 decae más rápidamente a 0 que en el período 1950-1980. Se estima una acumulación de datos en la cola derecha para ese primer período.

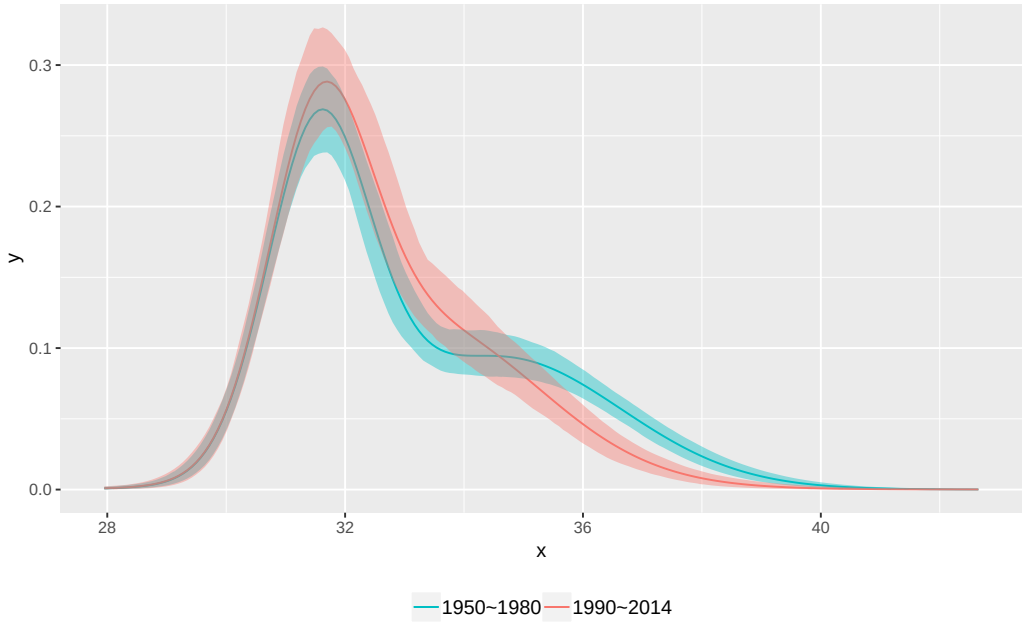


Figura 6.2: Comparación de densidad de temperaturas máximas entre los períodos de tiempo.

7

Conclusiones y consideraciones finales

En el presente trabajo se hizo una revisión, desde el punto de vista teórico y computacional, de una técnica de estimación de una función de densidad de probabilidad.

En el estudio de simulación (ver sección 5), la técnica se comparó con la técnica clásica de estimación por núcleos. Como se vio en el cuadro 5.1 y en las figuras 5.1a y 5.1b, los resultados fueron auspiciosos. El error cuadrático medio integrado (MISE) de ambos métodos es comparable, en algunos casos uno resulta mejor que otro, dependiendo de la distribución de los datos. No se estudió cómo incide el tamaño de muestra en el error. Para el caso del estimador por núcleos, existen resultados sobre la consistencia asintótica del estimador, que involucran el cálculo del AMISE (*Asymptotic* MISE). Una evaluación del desempeño empírico teniendo en cuenta el tamaño de muestra como un factor adicional es algo que se podría efectuar en caso de ahondar en este análisis.

Mientras que el estimador por núcleos es muy rápido de computar, la implementación del estimador DPMM es más costosa en tiempo computacional, pero brinda más flexibilidad dado que permite especificar un modelo más complejo. Esta característica se hará más notable en la medida que en el modelo jerárquico (ecuación 4.1) se agreguen más niveles de aleatoriedad. Este es otro punto a profundizar.

Por otro lado, existen ligeras extensiones de esta técnica con aplicación en

clustering. En la sección 3.2.1 se describió la introducción de ciertas variables latentes ξ_i, para $i = 1, \ldots, n$. Estas variables indican con qué componente de la mezcla se corresponde cada uno de los datos de la muestra, se sortean como variable categórica donde las probabilidades son proporcionales a la verosimilitud de cada dato bajo las distintas componentes. Como resultado se obtiene una partición de la muestra en grupos disjuntos, que estarán conformados por aquellos datos que comparten la etiqueta asignada (ver 2.3 del apéndice 2.3.1).

Bibliografía

Antoniak, C. E. (1974). Mixtures of dirichlet processes with applications to bayesian nonparametric problems. The annals of statistics, pages 1152–1174.

Bourel, M. and Cugliari, J. (2018). Bagging of density estimators. arXiv preprint arXiv:1808.03447.

Cardarello, M. and Luraghi, L. (2019). Análisis de valores extremos: una aplicación a temperaturas mínimas en uruguay.

Coles, S., Bawa, J., Trenner, L., and Dorazio, P. (2001). An introduction to statistical modeling of extreme values, volume 208. Springer.

Duong, T. (2019). ks: Kernel Smoothing. R package version 1.11.4.

Escobar, M. D. and West, M. (1995). Bayesian density estimation and inference using mixtures. Journal of the american statistical association, 90(430):577–588.

Ferguson, T. S. (1973). A bayesian analysis of some nonparametric problems. The annals of statistics, pages 209–230.

Gelman, A., Stern, H. S., Carlin, J. B., Dunson, D. B., Vehtari, A., and Rubin, D. B. (2013). Bayesian data analysis. Chapman and Hall/CRC.

Gilleland, E. and Katz, R. W. (2016). extRemes 2.0: An extreme value analysis package in R. Journal of Statistical Software, 72(8):1–39.

Hjort, N. L., Holmes, C., Müller, P., and Walker, S. G. (2010). Bayesian nonparametrics, volume 28. Cambridge University Press.

Müller, P., Quintana, F. A., Jara, A., and Hanson, T. (2015). Bayesian nonparametric data analysis, volume 18. Springer.

Niemi, J. (2017a). Bayesian nonparametrics. http://www.jarad.me/courses/stat615/slides/Nonparametrics/nonparametrics.pdf.

Niemi, J. (2017b). Finite mixture models. http://www.jarad.me/courses/stat615/slides/Nonparametrics/finiteMixtures.pdf.

Parzen, E. (1962). On estimation of a probability density function and mode. The annals of mathematical statistics, 33(3):1065–1076.

Plummer, M. (2003). Jags: A program for analysis of bayesian graphical models using gibbs sampling.

Plummer, M. (2018). rjags: Bayesian Graphical Models using MCMC. R package version 4-8.

R Core Team (2018). R: A Language and Environment for Statistical Computing. R Foundation for Statistical Computing, Vienna, Austria.

Rosenblatt, M. (1956). Remarks on some nonparametric estimates of a density function. The Annals of Mathematical Statistics, pages 832–837.

Sethuraman, J. (1994). A constructive definition of dirichlet priors. Statistica sinica, pages 639–650.

Wasserman, L. (2006). All of nonparametric statistics. Springer Science & Business Media.

Anexos

" No encontré ninguna frase célebre que hablara de [anexos]."

Guillermo Lamolle.
Cual Retazo de los Suelos

A. Demostración de la proposición 1

Proposición. *Si* $P_i := \dfrac{Z_i}{\sum_{j=1}^{k} Z_j}$ *con* $Z_i \sim Gamma(a_i, 1)$ *independientes, entonces el vector* $(P_1, \ldots, P_k)$ *tendrá distribución Dirichlet*$(a_1, \ldots, a_k)$.

Demostración. Sea $Z_i \sim Gamma(a_i, 1)$, por tanto, $f_{Z_i}(z_i) = \frac{e^{-z_i} z_i^{a_i - 1}}{\Gamma(a_i)} \mathbb{1}_{(0,+\infty)}(z_i)$

Por la independencia dos a dos de $Z_1, \ldots, Z_n$:

$$f_{Z_1,\cdots,Z_k}(z_1, \cdots, z_k) = \frac{e^{-\sum_{i=1}^{k} z_i} \prod_{i=1}^{k} z_i^{a_i - 1}}{\prod_{i=1}^{k} \Gamma(a_i)} \prod_{i=1}^{k} \mathbb{1}_{(0,+\infty)}(z_i)$$

Se tiene la transformación: $p_i := \frac{z_i}{\sum_{j=1}^{k} z_j}, i = 1, \ldots, k$. El rango de esta transformación es $k - 1$, por tanto no es biyectiva. Entonces definamos:

- $\mathbf{z} = (z_1, \ldots, z_k)$
- $\mathbf{p} = (p_1, p_2, \ldots, p_{k-1}, \tilde{z})^T$ donde $\tilde{z} = \sum_{j=1}^{k} z_j$
- $T(\mathbf{z}) = \mathbf{p}$

Definida así, T resulta invertible y se tiene que:

$$\mathbf{z} = T^{-1}(\mathbf{p}) = \big(\tilde{z}p_1, \tilde{z}p_2, \ldots, \tilde{z}(1 - \sum_{j=1}^{k} p_j)\big)$$

Por el teorema de transformación de un vector aleatorio:

$$f(\mathbf{p}) = f_{\mathbf{z}}\big(T^{-1}(\mathbf{p})\big)|J_T^{-1}| \tag{1}$$

$$J_T^{-1} = \left(\begin{array}{c|c} \mathbf{p}_{-k} & \tilde{z}\mathbf{I}_{k-1} \\ \hline 1 - \sum p_j & -\tilde{z}\mathbf{1}_{k-1}^T \end{array}\right)$$

Donde

- $\mathbf{p}_{-k} = (p_1, p_2, \ldots, p_{k-1})^T$
- $\mathbf{I}_{k-1}$ es la matriz identidad de dimensión $k-1$
- $\mathbf{1}_{k-1}$ es un vector de dimensión $k-1$ cuyas entradas son 1.

Para calcular $|J_T^{-1}|$, definamos previamente la secuencia de matrices cuadradas de dimensión $k-1$, $\{A_j\}_{j=1}^{k-1}$ de la siguiente forma:

$$((A_j))_{il} = \begin{cases} -\tilde{z} & \text{si } i = k-1 \\ \tilde{z} & \text{si } i = l \quad , \quad l < j \\ & \text{si } i = l-1, \quad j < l \leq k-1 \\ 0 & \text{en otro caso} \end{cases}$$

$A_1 = \left(\begin{array}{c|c} 0\mathbf{1}_{k-1} & \tilde{z}\mathbf{I}_{k-1} \\ \hline -\tilde{z}\mathbf{1}_k^T & \end{array}\right)$, se construye como una matriz diagonal con entradas $\tilde{z}$, a la que se le agrega una columna de 0 a su izquierda, y luego, se le agrega una última fila con entradas $-\tilde{z}$.

De esta forma, haciendo un desarrollo por la primer columna:

$$J_T^{-1} = \sum_{j=1}^{k-1}(-1)^{j+1}(p_j)|A_j| + (-1)^k(1 - \sum_{j=1}^{k} p_j)\tilde{z}^{k-1} \tag{2}$$

Otra observación importante es que A_{j+1} resultará de hacer un intercambio entre dos columnas de A_j. Por tanto $|A_{j+1}| = -|A_j|$, en particular $|A_j| = \begin{cases} |A_1| & j \text{ impar} \\ -|A_1| & j \text{ par} \end{cases}$, y por otro lado $|A_1| = (-1)^k(-\tilde{z})\tilde{z}^{k-2} = (-1)^{k-1}\tilde{z}^{k-1}$

Sustituyendo en (2):

$$\begin{aligned} J_T^{-1} &= \sum_{j=1}^{k-1}(-1)^{j+1}p_j(-1)^{j+1}|A_1| + (-1)^k(1-\sum_{j=1}^{k-1}p_j)\tilde{z}^{k-1} \\ &= \sum_{j=1}^{k-1}p_j|A_1| + (-1)^k(1-\sum_{j=1}^{k-1}p_j)\tilde{z}^{k-1} \\ &= (\sum_{j=1}^{k-1}p_j)(-1)^{k-1}\tilde{z}^{k-1} + (-1)^k(1-\sum_{j=1}^{k-1}p_j)\tilde{z}^{k-1} = \\ &= (-1)^{k-1}\tilde{z}^{k-1} \end{aligned}$$

Por tanto

$$|J_T^{-1}| = \tilde{z}^{k-1}$$

. Sustituyendo en (1) y usando que $1 - \sum_{i=1}^{k-1} p_i = p_k$:

$$\begin{aligned} f(\mathbf{p}) &= f_{\mathbf{z}}\big(T^{-1}(\mathbf{p})\big)|J_T^{-1}| \\ &= \frac{e^{-\sum_{i=1}^{k-1}\tilde{z}p_i}\prod_{i=1}^{k-1}(\tilde{z}p_i)^{a_i-1}}{\prod_{i=1}^{k-1}\Gamma(a_i)}\frac{e^{-\tilde{z}p_k}(\tilde{z}p_k)^{a_k-1}}{\Gamma(a_k)}\tilde{z}^{k-1}\prod_{i=1}^{k}\mathbb{1}_{\{\tilde{z}p_i>0\}} \\ &= \frac{e^{-\sum_{i=1}^{k}\tilde{z}p_i}\prod_{i=1}^{k}(\tilde{z}p_i)^{a_i-1}}{\prod_{i=1}^{k}\Gamma(a_i)}\tilde{z}^{k-1}\prod_{i=1}^{k}\mathbb{1}_{\{\tilde{z}p_i>0\}} \\ &= \frac{e^{-\tilde{z}}\tilde{z}^{(\sum_{i=1}^{k}a_i-1)}\tilde{z}^{k-1}\prod_{i=1}^{k}p_i^{a_i-1}}{\prod_{i=1}^{k}\Gamma(a_i)}\prod_{i=1}^{k}\mathbb{1}_{\{\tilde{z}p_i>0\}} \\ &= \frac{e^{-\tilde{z}}\tilde{z}^{(\sum_{i=1}^{k}a_i)-1}\prod_{i=1}^{k}p_i^{a_i-1}}{\prod_{i=1}^{k}\Gamma(a_i)}\prod_{i=1}^{k}\mathbb{1}_{\{\tilde{z}p_i>0\}} \end{aligned}$$

$$\begin{aligned} f(p_1,\ldots,p_{k-1}) &= \int_{\mathbb{R}} f(\mathbf{p})d\tilde{z} \\ &= \frac{\prod_{i=1}^{k}p_i^{a_i-1}}{\prod_{i=1}^{k}\Gamma(a_i)}\int_{\mathbb{R}} e^{-\tilde{z}}\tilde{z}^{(\sum_{i=1}^{k}a_i)-1}\prod\mathbb{1}_{\{\tilde{z}p_i>0\}}d\tilde{z} \end{aligned}$$

Notando que $\int_{\mathbb{R}} e^{-\tilde{z}} \tilde{z}^{(\sum_{i=1}^{k} a_i)-1} \mathbb{1}_{\{\tilde{z}>0\}} d\tilde{z} = \Gamma(\sum_{i=1}^{k} a_i)$ se obtiene lo que queríamos probar:

$$f(p_1, \ldots, p_{k-1}) = \frac{\prod_{i=1}^{k} p_i^{a_i-1}}{\prod_{i=1}^{k} \Gamma(a_i)} \Gamma(\sum_{i=1}^{k} a_i) \mathbb{1}_{\{(p_1,\ldots,p_k \in \mathcal{P}_k\}}$$

□

B. Código de R

Código para simular los procesos de *Stick-Breaking*

```
set.seed(1)
N <- 1000
x <- rnorm(N,0,1)
eps <- 1e-4
alfa <- 4
b <- rbeta(N,1,alfa)
p <- b[1]
p[2:N] <- sapply(2:N, function(i) b[i]*prod(1 - b[1:(i-1)]))
id1 <- cumsum(p)<1-eps

alfa <- 1000
b <- rbeta(N,1,alfa)
p2 <- b[1]
p2[2:N] <- sapply(2:N, function(i) b[i]*prod(1 - b[1:(i-1)]))
p2 <- p2/sum(p2)
id2 <- cumsum(p2)<1-eps
# graficos
plot(x[id1],p[id1],"h")
plot(x[id2],p2[id2],"h")
```

```
set.seed(12)
N <- 1000
alfa <- 4
b <- rbeta(N,1,alfa)
p3 <- b[1]
```

```
p3[2:N] <- sapply(2:N, function(i) b[i]*prod(1 - b[1:(i-1)]))
eps <- 1e-2
id3 <- cumsum(p3)<1-eps
NN <- length(which(id3!=0))
Y <-1/rgamma(NN,10,10)
X <- rnorm(NN,0,2)
X <- rnorm(NN,0,Y)
cbind(X,Y,p3[id3])
# grafico 3d
library(rgl)
plot3d(0,0,0,type="n",axes=F,
       xlim=c(-3,3),ylim=c(0,2),zlim=c(0,1),
       theta=10,
       zlab=expression(pi),
       ylab=expression(sigma),
       xlab=expression(mu))
axis3d('x', pos = c(NA, 0, 0),labels = F,tick=F,lwd=2.5)
axis3d('y', pos = c(0, NA, 0),labels=F,tick=F,lwd=2.5)
axis3d('z', pos = c(0, 0, NA),labels=F,tick=F,lwd=2.5)
points3d(X,Y,3*p3,col=4,pch=20)
for (i in 1:NN)
  segments3d(c(X[i],X[i]),c(Y[i],Y[i]),c(3*p3[i],0),col=4,lwd=2)
```

Código para la estimación de densidad con rjags

La siguiente función `r.jags()` permite obtener las simulaciones de la distribución posterior. Como argumentos requiere los datos cuya densidad se desea estimar, los parámetros especificados de la distribucion previa (ver ecuación 4.1), `H`: la cantidad de componentes de la mezcla, `n.iter` la cantidad de iteraciones a realizar, `n.chains` la cantidad de cadenas a simular.

La función retornará una lista de 4 elementos, con las simulaciones obtenidas para μ, σ, π, ξ. En el caso de los tres primeros, se tienen arreglos de dimensiones $H \times n.iteraciones \times n.chains$, que contienen para cada cadena, todas las iteraciones de μ_h, σ_h, π_h respectivamente, para $h = 1, \ldots, H$. Para el caso de ξ, lo que se tiene es un arreglo de dimensiones $n \times n.iter \times n.chains$ donde n es el tamaño de la muestra. Este arreglo, indicará para cada cadena, en cada iteración, a qué componente es asignada cada observación de la

muestra.

El modelo escrito en **jags** fue tomado de Niemi (2017a).

```
r.jags <- function(x,H=25,a=1,m,k,nu,psi,n.chains=1,n.iter=1e2){
  modelo="model {
for (i in 1:n) {
y[i] ~ dnorm(mu[zeta[i]], tau[zeta[i]])
zeta[i] ~ dcat(pi[])
}
for (h in 1:H) {
mu[h] ~ dnorm(m,1/k)
tau[h] ~ dgamma(nu,psi)
sigma[h] <- 1/sqrt(tau[h])
}
# Stick breaking
for (h in 1:(H-1)) { V[h] ~ dbeta(1,a)T(0.001,0.999) }
V[H] <- 1
pi[1] <- V[1]
for (h in 2:H) {
pi[h] <- V[h] * (1-V[h-1]) * pi[h-1] / V[h-1]
}
}"
dat_temp = list(n=length(x),H=H,y=x,a=a,m=m,k=k,nu=nu,psi=psi)
jm_temp = jags.model(textConnection(modelo), data = dat_temp,
                     n.chains = n.chains)
r_temp = jags.samples(jm_temp, c('mu','sigma','pi','zeta'), n.iter)
return(r_temp)
}
```

Las funciones `estimar_densidad()` y `estimar_densidadCI()` toman como argumento la lista devuelta por `r.jags()` y retornan la estimación puntual y la estimación del cuantil q, respectivamente.

```
estimar_densidad <- Vectorize(function(x,mod,cad=1)
  mean(apply(
    mod$pi[,,cad]*dnorm(x,mean=mod$mu[,,cad],sd=mod$sigma[,,cad]),
    2,sum)),
  vectorize.args = "x")
```

```
estimar_densidad_q <- Vectorize(function(x,mod,q,cad=1)
  as.numeric(quantile(apply(
    mod$pi[,,1]*dnorm(x,mean=mod$mu[,,cad],sd=mod$sigma[,,cad]),
    2,sum),q)),vectorize.args = "x")
```

Código para monitoreo de la convergencia

El objeto `mod_11` debe contener el modelo devuelto por `r.jags`. El siguiente código da el valor de cada iteración de las 4 cadenas generadas para la densidad estimada en el punto $x = 4$.

```
iter=1:1000
cad=1:4
x=-4
apply(mod_11$pi[,iter,cad]*
        dnorm(x,
              mean=mod_11$mu[,iter,cad],
              sd=mod_11$sigma[,iter,cad]),
      2:3,sum) %>%
  as_data_frame %>%
  set_names("Cadena 1","Cadena 2","Cadena 3","Cadena 4") %>%
  mutate(iter=iter) %>% gather(cadena,valor,-iter) %>%
  ggplot(aes(x=iter,y=valor,col=cadena))+geom_line()+
  theme(legend.title = element_blank())+ggtitle("x=-4")
```

El siguiente código monitorea las iteraciones de los parámetros μ_h y π_h para $h = 1, \ldots, 4$ en la cadena 1.

```
cadena <- 1
data.frame(mu=t(mod_11$mu[1:4,,cadena]),
                    pi=t(mod_11$pi[1:4,,1])) %>%
  mutate(iter=1:1000) %>% gather(par,valor,-iter) %>%
  separate(par,into=c("par","comp")) %>%
  ggplot(aes(x=iter,y=valor,col=comp))+geom_line(alpha=.7)+
  facet_grid(par~.,scales = "free_y")
```

El siguiente código permite monitorear en cada iteración el número de componentes activas.

```
apply(mod_11$zeta[,,],2:3,max) %>% as_data_frame %>%
  set_names("Cadena 1","Cadena 2","Cadena 3","Cadena 4") %>%
  mutate(iter=iter)  %>% gather(cadena,max,-iter) %>%
  ggplot(aes(x=iter,y=max,col=cadena))+geom_step()+
  theme(legend.title=element_blank(),legend.position = "bottom")+
  xlab("Iteraciones")+ylab(expression(paste("max ",xi[i])))
```