UNIVERSIDAD
DE SANTIAGO
DE CHILE

Universidad de Santiago de Chile
Facultad de Ciencia
Departamento de Matemática y Ciencia de la Computación
Ingeniería Estadística

Procesos de Dirichlet y Procesos Gaussianos para la Inferencia en Modelos de Regresión Bayesianos

Primer Semestre 2025

Modelos Bayesianos en Machine Learning
Profesor - Andrés Iturriaga C.
Ayudante - Bastián Diaz .
Alumno - Tiago Lazcano L.

Índice general

Capítulo 1

Introducción

¿Puede un ser omnipotente crear una piedra tan pesada que ni siquiera él mismo pueda levantarla?

Esta paradoja nos habla de los límites que pueden aparecer incluso cuando parece que todo es posible. En estadística, algo parecido ocurre cuando construimos modelos extremadamente flexibles: si les damos demasiada libertad, corremos el riesgo de que no puedan decirnos nada útil.

En este trabajo exploraremos esa idea: cómo encontrar un equilibrio entre dejar que los datos hablen y evitar que el modelo se vuelva tan general que no aprenda nada concreto.

1. Motivación

En estadística, los datos se modelan como realizaciones de variables aleatorias que se asumen independientes y provenientes de una distribución desconocida F. Cuando esta distribución se describe mediante un conjunto finito de parámetros, hablamos de modelos paramétricos. Sin embargo, limitarse a este tipo de modelos puede restringir la capacidad de capturar la complejidad real de los datos y conducir a inferencias poco robustas. Por ello, surge la necesidad de modelos más flexibles, que no estén restringidos a un número finito de parámetros, sino que puedan considerarse en espacios de dimensión infinita. Estos modelos no paramétricos permiten una mayor adaptabilidad y robustez frente a posibles errores de especificación.

Abordar este problema desde una perspectiva bayesiana nos permite usar lo que ya sabemos y manejar la incertidumbre de forma clara y coherente. A medida que recibimos más datos, podemos actualizar nuestras ideas de manera flexible. Además, al permitir que los modelos sean tan complejos como los datos lo pidan, evitamos limitar las respuestas a formas rígidas, lo que mejora la forma en que entendemos y predecimos.

1.1. Fundamentos Teóricos

Inferencia y aprendizaje bayesiano

La inferencia bayesiana nos da un marco para manejar la incertidumbre y sumar conocimiento previo, algo vital cuando los datos son limitados o complejos.

Regresión

Los modelos clásicos de regresión suelen ser demasiado rígidos para captar patrones complejos o variables que cambian con el contexto.

Modelos no paramétricos

Los modelos no paramétricos permiten que la forma del modelo crezca y se adapte según lo que los datos realmente muestran, sin imponer reglas estrictas desde el principio.

Procesos estocásticos

Los procesos estocásticos son la base para definir modelos flexibles que pueden representar estructuras complejas y dependientes en los datos, como los que veremos más adelante.

Teoría de la medida

La teoría de la medida nos da las herramientas para manejar probabilidades en espacios complejos, como cuando trabajamos con funciones o conjuntos infinitos. Es la base matemática que hace posible definir modelos flexibles y rigurosos en estadística avanzada.

De este es importante estar levemente familiarizado con conceptos como Espacios Medibles, Medidas domintantes σ-aditivas, Medidas de Probabilidad y Medidas de Probabilidad Aleatorias.

La tabla a continuación compara los enfoques frecuentista y bayesiano para resolver tres problemas estadísticos comunes. En el enfoque frecuentista se utilizan técnicas como la estimación por Kernel y suavizado por núcleo. En contraste, el enfoque bayesiano utiliza modelos como el Proceso de Dirichlet y el Proceso Gaussiano para abordar estos mismos problemas.

Problema	Frecuestista	Bayesiano
Estimar F	$\hat{F}_n$	Proceso de Dirichlet
Estimar f	Kernel	Modelo de Mezcla del Proceso de Dirichlet
Regresión	Suavizado por núcleo	Proceso Gaussiano

En este trabajo, se abordarán principalmente los enfoques bayesianos, explorando el uso de procesos como el Proceso de Dirichlet y el Proceso Gaussiano para estimar funciones y realizar regresión. Se contrastarán con los métodos frecuentistas para entender las diferencias clave entre ambos enfoques.

Capítulo 2

Proceso de Dirichlet (DP)

1. Distribución Dirichlet

Sea $\boldsymbol{\alpha} = (\alpha_1, \ldots, \alpha_K)$ el vector de parámetros de forma, con $\alpha_i > 0$, la función de densidad de probabilidad de la distribución Dirichlet está dada por:

$$f(\boldsymbol{p}|\boldsymbol{\alpha}) = \frac{\Gamma\left(\sum_{i=1}^{K} \alpha_i\right)}{\prod_{i=1}^{K} \Gamma(\alpha_i)} \prod_{i=1}^{K} p_i^{\alpha_i - 1} = \frac{1}{\beta(\boldsymbol{\alpha})} \prod_{i=1}^{K} p_i^{\alpha_i - 1}$$

donde:

- $\boldsymbol{p} = (p_1, \ldots, p_K)$ son las probabilidades de las categorías, con $p_i \in (0, 1)$
- $\sum_{i=1}^{K} p_i = 1$

2. Definición formal

Definición (Proceso de Dirichlet): Sea $M > 0$ un parámetro positivo y G_0 una medida de probabilidad definida en un espacio S. Un proceso de Dirichlet $DP(M, G_0)$ es una medida de probabilidad aleatoria[1] G definida en S, que asigna una probabilidad $G(B)$ a cada conjunto medible $B \subseteq S$, cumpliendo que para cualquier partición finita y medible $\{B_1, \ldots, B_k\}$ del espacio S, la distribución conjunta de las probabilidades asignadas $\{G(B_1), \ldots, G(B_k)\}$ sigue una distribución de Dirichlet con parámetros proporcionales a $MG_0(B_i)$, es decir (Müller et al., 2015, Capítulo 2):

$$(G(B_1), G(B_2), \ldots, G(B_k)) \sim \text{Dirichlet}(MG_0(B_1), MG_0(B_2), \ldots, MG_0(B_k))$$

3. Intuición

Recordemos el caso paramétrico, donde tenemos una muestra de una densidad dada un parámetro θ, que es una variable aleatoria. Es decir, tenemos:

$$\begin{cases} \underset{\sim}{Y} \sim f(y|\theta) \\ \theta \sim \pi \end{cases}$$

[1] Es decir, una variable aleatoria con valores en el conjunto de medidas de probabilidad en S

La mecánica para obtener la posteriori es a través del uso del teorema de Bayes:

$$\pi(\theta|\underset{\sim}{Y}) \propto L(\theta)\pi(\theta)$$

Este procedimiento es bastante directo y nos permite obtener un estimador puntual y regiones de credibilidad para nuestro parámetro θ, lo cual se puede definir como un conjunto $\mathcal{A}$ tal que:

$$\int_{\mathcal{A}} \pi(\theta|\underset{\sim}{Y})d\theta = 1 - \alpha$$

Esto significa que, dado que θ es una cantidad aleatoria, podemos interpretarlo como un grado de creencia. Por lo tanto, la probabilidad de que θ esté en el conjunto $\mathcal{A}$ es:

$$\mathbb{P}(\theta \in \mathcal{A}|\underset{\sim}{Y}) = 1 - \alpha$$

Es importante no confundir esto con un intervalo de confianza. A diferencia de un intervalo de confianza, una región de credibilidad no dice que $\mathcal{A}$ "contiene" al valor verdadero de θ $(1 - \alpha)\,\%$ veces, sino que refleja nuestro grado de creencia.

El cálculo de la posteriori puede ser complicado en algunos casos, especialmente en modelos complejos, lo que da lugar a la necesidad de usar métodos MCMC (Markov Chain Monte Carlo) para simular de la distribución posterior y así obtener las cantidades que necesitamos.

El enfoque Bayesiano es intuitivo porque podemos expresar la posteriori en términos del producto de la verosimilitud y la priori. Este marco sencillo, aunque poderoso, nos permite construir modelos bien definidos y enfocar la estimación de parámetros de forma coherente.

Ahora bien, queremos extender este enfoque al caso no paramétrico, donde nuestro objetivo no es estimar un parámetro específico, sino una función de distribución acumulada (FDA), o bien, una función de densidad o incluso una función de regresión.

Supongamos que tenemos una muestra de valores reales provenientes de una función desconocida F, tal que:

$$(X_1, X_2, \ldots, X_n) \sim F$$

El objetivo aquí es estimar F, y lo único que sabemos es que es una FDA en la recta real.

Para resolver este problema desde un punto de vista frecuentista, la manera de estimar una FDA de forma no paramétrica es utilizando la función de distribución empírica (FDE) $\hat{F}_n(x)$, que se define como:

$$\hat{F}_n(x) = \frac{\#X_i \leq x}{n}$$

Este es un estimador consistente de F, lo que significa que (según el teorema de Glivenko-Cantelli):

$$\sup_F \sup_x |\hat{F}_n(x) - F(x)| \longrightarrow 0 \text{ casi seguramente}$$

Esto nos dice que, con probabilidad 1, la función de distribución empírica se aproxima a la verdadera distribución F conforme $n \to \infty$.

Para evaluar cuán lejos estamos de la FDA real, usamos la desigualdad de Dvoretzky–Kiefer–Wolfowitz (DKW) en el peor caso:

$$\mathbb{P}(\sup_x |\hat{F}_n(x) - F(x)| > \varepsilon) \leq 2e^{-2n\varepsilon^2} = \alpha \qquad \forall \varepsilon > 0$$

Y la distancia de confianza ε se puede expresar como:

$$\varepsilon = \sqrt{\frac{1}{2n} \log \frac{1}{\alpha}}$$

Esto nos da una banda confidencial para la FDA, es decir:

$$\hat{F}_n(x) \pm \varepsilon_n = [L, U]$$

Finalmente, con probabilidad al menos $1 - \alpha$, tenemos que:

$$\mathbb{P}(L(x) \leq F(x) \leq U(x) \quad \forall x) \geq 1 - \alpha$$

La pregunta ahora es: ¿cuál es el espacio de parámetros en el caso no paramétrico?

Supongamos que $\mathcal{F}$ es el conjunto de todas las funciones de distribución acumulada F en la recta real. Este es un conjunto infinito-dimensional, lo que significa que no podemos parametrizarlo utilizando un número finito de parámetros.

Cada punto en $\mathcal{F}$ corresponde a una posible FDA.

Para trabajar con este conjunto, necesitamos especificar una priori. Después de resolver este problema y tener los datos, nos enfrentamos a la pregunta: ¿Cómo obtenemos la posteriori?

Aquí surge una limitación importante: no podemos obtener la posteriori directamente usando el teorema de Bayes, porque este requiere la existencia de una medida dominante σ-finita, y en este caso el conjunto $\mathcal{F}$ no tiene una medida dominante σ-finita. Sin embargo, esto no significa que no exista una posteriori.

La manera más común de definir una priori sobre el espacio de las FDA's es mediante el Proceso de Dirichlet (DP), que se denota como $DP(\alpha, F_0)$, donde α y F_0 son los hiperparámetros:

- F_0 es la creencia inicial sobre la distribución F.
- α Una medida que cuantifica cuánta confianza tenemos en F_0, es decir, un parámetro que controla la concentración de la distribución.

Aunque no podemos escribir una fórmula explícita para esta priori, es posible simular muestras de ella. De forma análoga a cómo simulamos de una Normal o Gamma, podemos simular de un Proceso de Dirichlet. Esto nos lleva a la siguiente pregunta: ¿Cómo simulamos una muestra del proceso de Dirichlet?

Supongamos que tenemos una priori $\theta \sim \mathcal{N}(0, 1)$. Podríamos escribir su función de densidad (pdf) y usar un algoritmo para simular de ella. De manera similar, podemos realizar el mismo proceso para el Proceso de Dirichlet. Por lo tanto, no nos preocupa tanto encontrar la fórmula explícita, sino entender cómo simular muestras de esta distribución.

4. Stick Breaking

Comenzamos extrayendo una secuencia infinita de variables aleatorias independientes e idénticamente distribuidas $S_1, S_2, \ldots$ de nuestra creencia inicial F_0.

A continuación, extraemos otra secuencia infinita de variables $V_1, V_2, \ldots$, provenientes de una distribución Beta con parámetros 1 y α, es decir:

$$V_i \overset{\text{iid}}{\sim} \text{Beta}(1, \alpha) \quad \text{tal que} \quad p(v) \propto (1-v)^{\alpha-1}, \quad v \in (0,1), \quad \alpha > 0$$

Luego, generamos una nueva secuencia de variables aleatorias $W_1, W_2, \ldots$ tal que:

$$W_1 = V_1 \quad \text{y} \quad W_j = V_j \prod_{k<j} (1 - V_k)$$

Esto nos da una secuencia de números no negativos que suman 1, es decir:

$$\sum_{j \geq 1} w_j = 1$$

De esta forma, asignamos una masa w_j a cada S_j que extrajimos, creando así una distribución discreta.

Por lo tanto, podemos escribir la función de distribución $F(t)$ como:

$$F(t) = \sum_{j \geq 1} w_j \mathbb{1}(S_j \leq t)$$

O equivalentemente:

$$F(t) = \sum_{j \geq 1} w_j \delta_{S_j}$$

Donde δ_{S_j} es la medida de Dirac que se define como:

$$\delta_x(A) = \begin{cases} 0 & \text{si } x \notin A \\ 1 & \text{si } x \in A \end{cases}$$

Algoritmo

1. Extraer muestras $S_1, S_2, \ldots$ de F_0

2. Extraer muestras $W_1, W_2, \ldots$ tal que $V_i \overset{\text{iid}}{\sim} \text{Beta}(1, \alpha)$, $W_1 = V_1$ y $W_j = V_j \prod_{k<j}(1 - V_k)$

3. Representar $F(t)$ como $\sum_{j \geq 1} w_j \mathbb{1}(s_j \leq t)$

4.1. Ejemplo

Para el siguiente ejemplo se simuló 15 datos provenientes de una distribución normal de media 3 y varianza 1, además se decide usar la distribución o medida base de una normal estándar $\mathcal{N}(0,1)$ entonces:

$$\begin{cases} X|F \sim F \\ F \sim DP(\alpha, \mathcal{N}(0,1)) \end{cases}$$

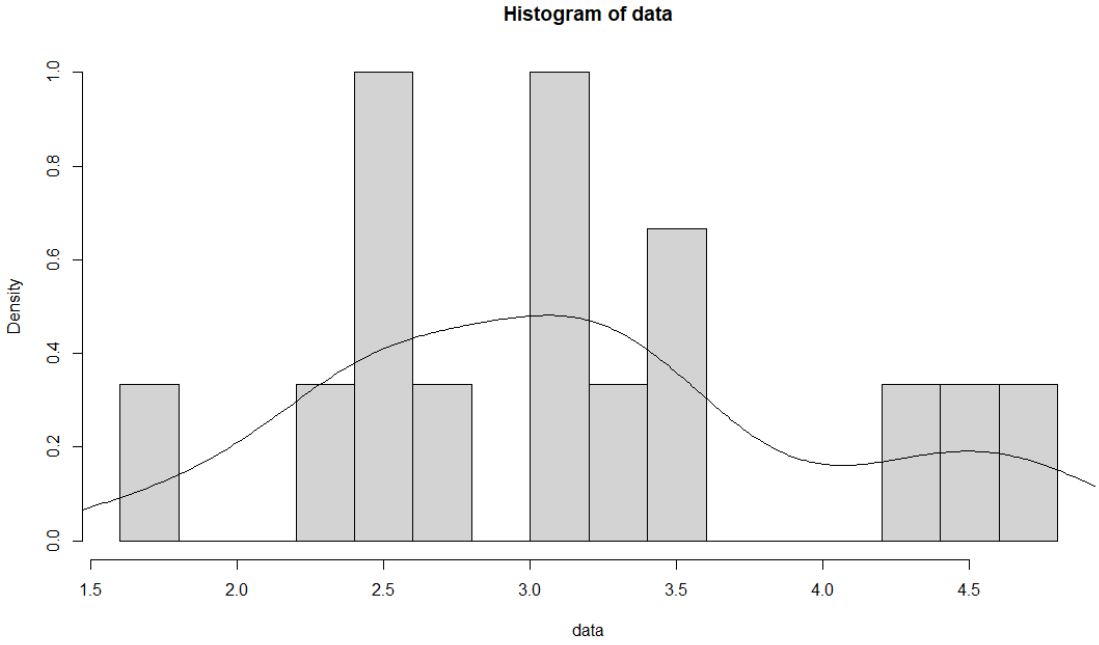


Figura 2.1: Histograma y densidad Kernel de los datos simulados

Notar que los datos nos instan a usar una medida base distinta, dado que sospechamos de forma clara que estos pueden estar centrados en 3 (suponiendo que no sabemos de donde provienen estos mismos), pero se realizará el análisis de todas maneras para mostrar como se comporta una medida base parcialmente equivocada.

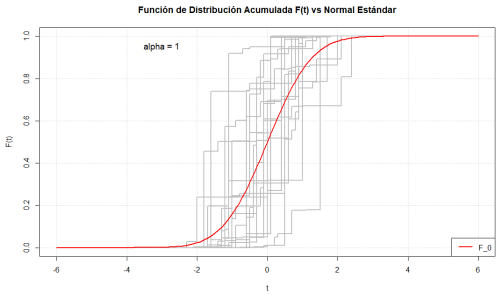


Figura 2.2: $\alpha = 1$

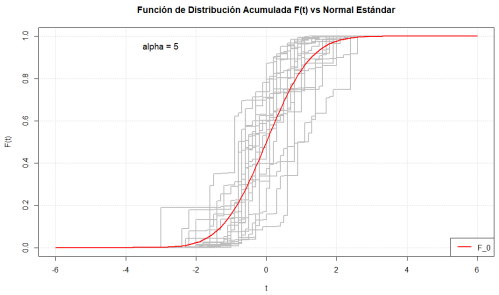


Figura 2.3: $\alpha = 5$

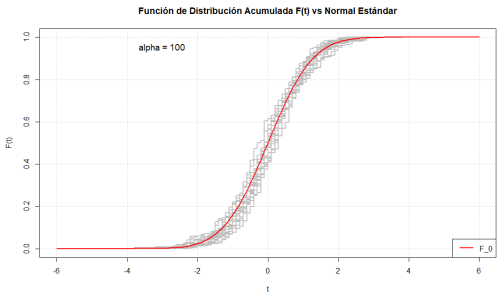


Figura 2.4: $\alpha = 100$

Figura 2.5: 20 simulaciones de $DP(\alpha, \mathcal{N}(0,1))$

Se obtienen entonces 20 funciones de distribución acumuladas, donde se aprecia de manera evidente como se comporta F a medida que se aumenta el valor de α o equivalentemente, a medida que aumentamos nuestra creencia respecto a esa medida base.

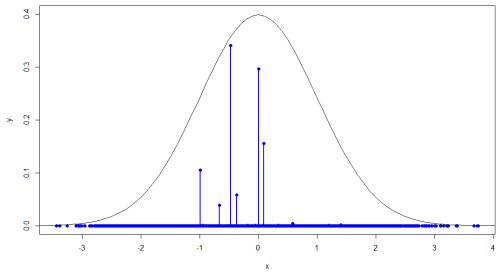


Figura 2.6: $\alpha = 1$

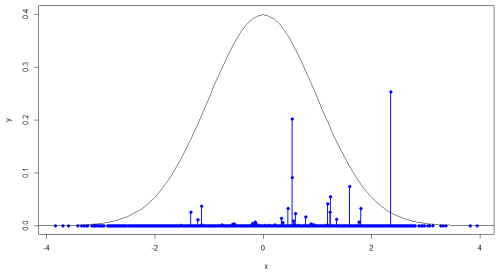


Figura 2.7: $\alpha = 5$

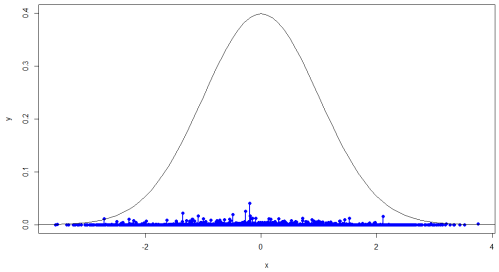


Figura 2.8: $\alpha = 100$

Figura 2.9: Muestras del proceso de Dirichlet para diferentes valores de α.

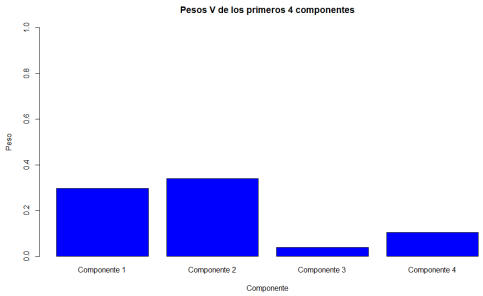


Figura 2.10: $\alpha = 1$

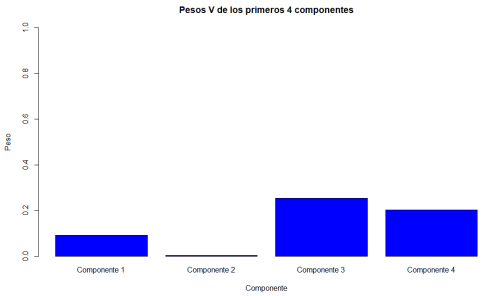


Figura 2.11: $\alpha = 5$

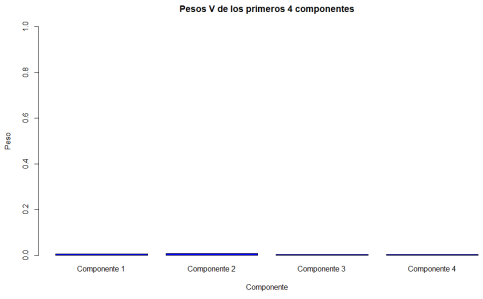


Figura 2.12: $\alpha = 100$

Figura 2.13: Valores de los primeros cuatro pesos para diferentes valores de α.

Como se puede ver en los gráficos, las muestras de un proceso de Dirichlet son distribuciones discretas y se vuelven menos concentradas (más dispersas) a medida que aumenta α. Los gráficos fueron generados utilizando la vista del proceso de ruptura de palos del proceso de Dirichlet.

En cuanto a los pesos, se puede observar que tienden a equilibrarse progresivamente entre sí a medida que el valor de α aumenta, reflejando una mayor dispersión y uniformidad en la distribución. Este comportamiento indica que, con valores más altos de α, el proceso de Dirichlet se aproxima a una distribución más equitativa, donde los pesos se distribuyen de manera más homogénea entre las distintas categorías.

4.2. Posteriori

Resulta que la distribución posterior también es un proceso de Dirichlet, dado que este es conjugado. Por lo tanto, la distribución $F|\underset{\sim}{X}$ es tal que:

$$F|\underset{\sim}{X} \sim DP(\alpha + n, \bar{F})$$

Con:

$$\bar{F} = \frac{n}{n+\alpha}\hat{F}_n + \frac{\alpha}{n+\alpha}F_0$$

Donde los parámetros del proceso se actualizan en función del número de observaciones y de la concentración previa. La función $\bar{F}$ es una combinación convexa entre la distribución empírica acumulada $\hat{F}_n$ y la distribución base F_0.

Notablemente, obtenemos esta forma de la posterior sin aplicar explícitamente el teorema de Bayes.

- Cuando $\alpha \to 0$ (es decir, no confiamos en la distribución base), se tiene que $\bar{F} \to \hat{F}_n$: la posteriori se guía completamente por los datos observados.

- Cuando $\alpha \to \infty$ (confiamos plenamente en la distribución base), entonces $\bar{F} \to F_0$: los datos tienen poco o ningún efecto sobre la posterior.

- Cuando $n \to \infty$ (es decir, se tiene mucha información), también $\bar{F} \to \hat{F}_n$; sin embargo, como $\hat{F}_n \to F$ casi seguramente (Teorema de Glivenko-Cantelli), en realidad $\bar{F} \to F$. En este caso, el análisis posterior pierde relevancia práctica, ya que la distribución empírica por sí sola es una excelente aproximación a la verdadera distribución subyacente.

Lo que comúnmente se realiza es tomar una muestra de tamaño N de este proceso tal que:

$$F_1, \ldots, F_N \sim DP(\alpha + n, \bar{F})$$

De esa forma podemos obtener algún estimador para F como la media entre estas, osea:

$$\frac{1}{N}\sum_{i=1}^{N} F_i$$

4.3. Ejemplo

Continuando con el análisis previo podemos visualizar entonces $F_j|\underset{\sim}{X}$, donde j va desde 1 hasta 20.

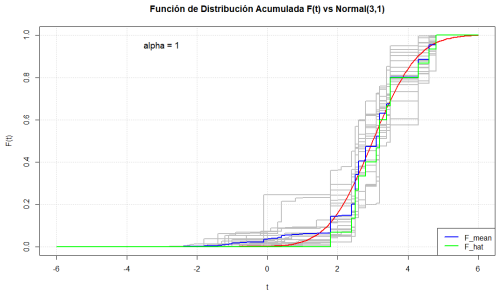


Figura 2.14: $\alpha = 1$

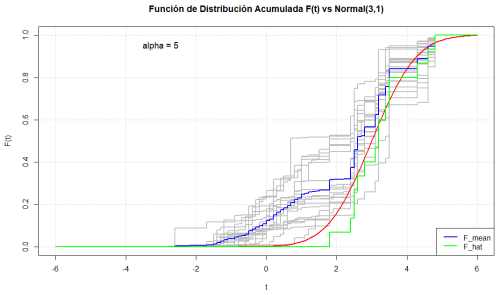


Figura 2.15: $\alpha = 5$

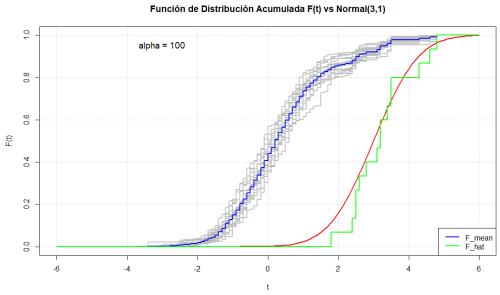


Figura 2.16: $\alpha = 100$

Figura 2.17: 20 simulaciones de $DP(\alpha + n, \bar{F})$

Finalmente se aprecia como se ajusta las F a medida que se aumenta el valor de α, denotando claramente como estas se equivoca cada vez más con respecto a la función de distribución acumulada empírica y la teórica.

Capítulo 3

Modelo de Mezcla del Proceso de Dirichlet (DPMM)

Sabemos como tomar una muestra de una función de distribución acumnulada a través del proceso de Dirichlet a través del método mencionado anteriormente.

Dada la naturaleza discreta de F podemos llegar a pensar que no nos es útil para poder estimar densidades, pero en realidad esto nos presenta una ventaja importante dado que en este caso es posible encontrar una representación continua para F

Lo que hacemos ahora es tomar una muestra n de parámetros θ de F, por lo que no estamos tomando una muestra de nuestros datos de F, sino que de nuestros parámetros, uno por cada observación que tengamos, por lo que cada x_i pertenecerá a la distribución $p(x_i|\theta_i)$, por lo que obtenemos:

$$\begin{cases} F \sim DP(\alpha, F_0) \\ \theta_1, \ldots, \theta_n | F \overset{\text{iid}}{\sim} F \\ X_i | \theta_i \overset{\perp\!\!\!\perp}{\sim} p(x_i|\theta_i) \end{cases}$$

Escribir la densidad como una mezcla infinita de densidades:

$$p(x) = \sum_{j=1}^{+\infty} w_j p(x_j|\theta_j)$$

Primero obtenemos una función de distribución acumulada aleatoria, después obtenemos n muestras de un vector de parámetros $\boldsymbol{\theta}$ de aquella fda y después tomamos la densidad $p(x_j|\theta_j)$ con ese θ_j en particular. Se está creando entonces una mezcla de densidades, pero permitiendo un componente por cada observación y eso define una densidad, por lo que de esta manera estamos definiendo una priori en el espacio de densidades.

La pregunta es, ¿No resulta esto en un sobreajuste dado que estamos asociando una densidad distinta por cada observación que tenemos? Pero eso es lo lindo de esto, recordemos que F es discreto, entonces, ¿Qué sucede cuando extraemos valores de esta? Obtendremos empates de forma regular, por lo que estamos obteniendo un número menor a n de parámetros, estamos generando clusters.

Es realmente el mismo procedimiento que hicimos anteriormente pero añadiendo una capa adicional, obteniendo entoncesun modelo jerárquico.

Algoritmo

1. Extraer muestras $W_1, W_2, \ldots$ tal que $V_i \overset{\text{iid}}{\sim} \text{Beta}(1, \alpha)$, $W_1 = V_1$ y $W_j = V_j \prod_{k<j}(1 - V_k)$

2. Escribir $F = \sum_{j \geq 1} w_j \mathbb{1}(\theta_j \leq t) \overset{d}{=} DP(\alpha, F_0)$

3. $\theta_n^* \overset{\text{iid}}{\sim} F$

4. $x_n \overset{\perp\!\!\!\perp}{\sim} F(\theta_n^*)$

Repetimos este procedimiento para la posteriori.

1. Ejemplo

Para el siguiente ejemplo se simuló 50 datos provenientes de una mezcla de distribuciónes normales tal que:

$$f(x_i) = \frac{1}{2} \cdot \mathcal{N}(x_i| - 2, 1) + \frac{1}{2} \cdot \mathcal{N}(x_i|2, 1)$$

Además se decide usar la distribución o medida base de una normal estándar $\mathcal{N}(0, 1)$.

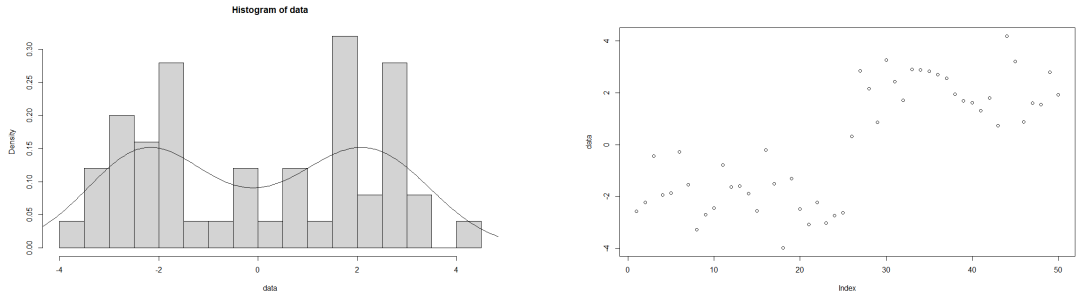


Figura 3.1: Histograma y densidad Kernel, Scatterplot

De igual manera que el ejemplo 2.1 previamente analizado, los datos nos instan a usar una medida base distinta, dado que sospechamos de forma clara que estos pueden estos se **pueden** comportar como una mezcla de dos distribuciones normales, pero de todas maneras se realizará el ejercicio.

En este caso el análisis se realizará para la media, donde la varianza es constante y solo con $\alpha = 1$ dado que en este caso ya sabemos como se comportarán estas densidades a medida que aumentamos el valor de α.

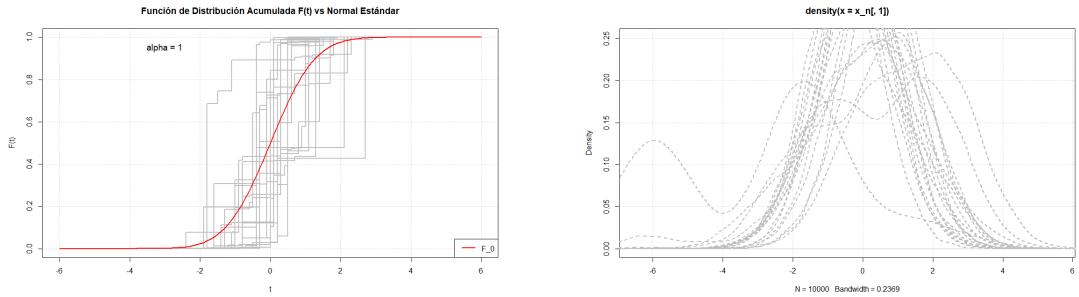


Figura 3.2: FDA y Densidad $\theta \mid F$

Aquí podemos visualizar las prioris F y sus densidades. Estas funciones muestran cómo la distribución de los clústeres se define inicialmente antes de la observación de los datos. Cada curva refleja una posible realización del proceso de Dirichlet, y la forma de la priori influye en la formación de los clústeres a medida que se incorporan los datos observados.

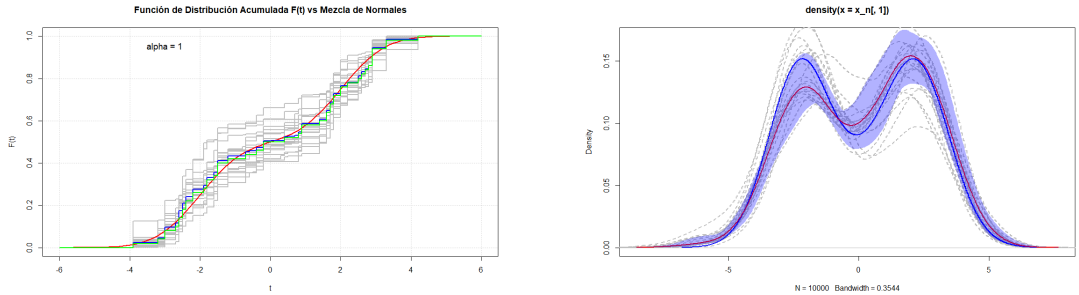


Figura 3.3: FDA y Densidad $F \mid \theta$ con banda confidencial a nivel de confianza 95 %

Aquí podemos visualizar las posterioris F y sus densidades. Estas curvas reflejan cómo las distribuciones de los clústeres evolucionan después de observar los datos. Las bandas de confianza al 95 % muestran la incertidumbre asociada con las estimaciones de los clústeres, brindando un rango en el cual se espera que se ubiquen las distribuciones verdaderas, dadas las observaciones.

Este mismo análisis puede extenderse a una dimensión multivariada, y en el caso bidimensional, se puede observar cómo las nubes de puntos se colorean según el clúster al que pertenecen, permitiendo una visualización más clara de cómo los datos se agrupan y se asignan a diferentes categorías dentro del modelo.

Es importante señalar que este proceso se vuelve más complejo a medida que se añaden nuevos hiperparámetros, los cuales controlan los parámetros previamente especificados de manera arbitraria. La introducción de estos hiperparámetros adicionales complica aún más el modelo. Como resultado, se hace necesario utilizar el algoritmo de Gibbs para abordar la inferencia en estos modelos. Sin embargo, el objetivo de este trabajo no es abordar esta complejidad, por lo que no se profundizará en estos aspectos.

Capítulo 4

Proceso Gaussiano (GP)

Otro priori bayesiano no paramétrico (BNP) comúnmente utilizado para una función media aleatoria $f(\cdot)$ es el proceso gaussiano (GP). Sea $\mu(x)$, con $x \in \mathbb{R}^d$, una función dada, y sea $r(x_1, x_2)$, para $x_j \in \mathbb{R}^d$, una función de covarianza, es decir, la matriz $R \in \mathbb{R}^{n\times n}$ con entradas $R_{ij} = r(x_i, x_j)$ es definida positiva para cualquier conjunto de puntos distintos $x_i \in \mathbb{R}^d$.

Definición (Proceso Gaussiano): Una función aleatoria $f(x)$, con $x \in \mathbb{R}^d$, tiene un priori GP si, para cualquier conjunto finito de puntos $x_i \in \mathbb{R}^d$, $i = 1, \ldots, n$, la evaluación de la función en esos puntos forma un vector aleatorio normal multivariado:

$$(f(x_1), \ldots, f(x_n))' \sim \mathcal{N}\left((\mu(x_1), \ldots, \mu(x_n))', R\right)$$

donde $R = [r(x_i, x_j)]$ es la matriz de covarianza de tamaño $n \times n$. Denotamos esto como:

$$f \sim \mathcal{GP}(\mu(x), r(x, y))$$

El proceso gaussiano puede utilizarse como priori para la función media desconocida en la ecuación (4.1). Asumiendo residuos normales, la distribución posterior para

$$\mathbf{f} = (f(x_1), \ldots, f(x_n))$$

Entonces:

$$\begin{cases} y \sim \mathcal{N}(f(x), \sigma_y^2) \\ f(x) \sim \mathcal{GP}(\mu, r) \end{cases}$$

es nuevamente normal multivariada. De forma similar, la distribución de $f(x)$ en nuevas ubicaciones x_{n+i}, que no se encuentran en los datos observados, también está caracterizada por una distribución normal multivariada.

La regresión bayesiana completamente basada en GPs puede ser computacionalmente exigente, ya que es necesario invertir una matriz de covarianza de alta dimensión en cada iteración de un muestreador de Gibbs o rutina de maximización.

1. Ejemplo

Se simularán datos provenientes de un proceso gaussiano (GP) utilizando una matriz de covarianza definida por un kernel específico. El conjunto de valores de entrada x se obtiene a partir de una muestra aleatoria ordenada de 100 puntos en el intervalo $[0, 5]$.

El kernel utilizado para modelar la correlación entre los puntos corresponde a un kernel Gaussiano, el cual corresponde a:

$$\mathcal{K}(x, x') = \sigma_g^2 e^{-\frac{(x-x')^2}{2\ell^2}} = \sigma_g^2 e^{-\phi^2(x-x')^2}$$

donde σ_g^2 es la varianza del proceso y ϕ es el parámetro de longitud de escala, que controla la suavidad del proceso. La matriz de covarianza K se calcula utilizando el kernel $\mathcal{K}(x, x')$ evaluado en todas las combinaciones de puntos x.

A continuación, se genera una realización del proceso gaussiano utilizando esta matriz de covarianza y se añade ruido aleatorio $\varepsilon \sim \mathcal{N}(0, \sigma_y^2)$, donde $\sigma_y^2 = 0{,}03$, para obtener los datos simulados $y = g + \varepsilon$,

Esta simulación produce una serie de curvas correspondientes a las diferentes realizaciones del proceso gaussiano, sobre las cuales se introduce ruido para reflejar las observaciones en un contexto real.

Para este procedimiento vamos a particionar la base de datos en datos de entrenamiento y en datos de testeo, donde van a corresponder al 80 % y 20 % del tamaño de esta respectivamente.

Podemos entonces explicitar la jerarquía:

$$\begin{cases} y \sim \mathcal{N}(f(x), \sigma_y^2) \\ f(x) \sim \mathcal{GP}(0, \mathcal{K}(x, x')) \end{cases}$$

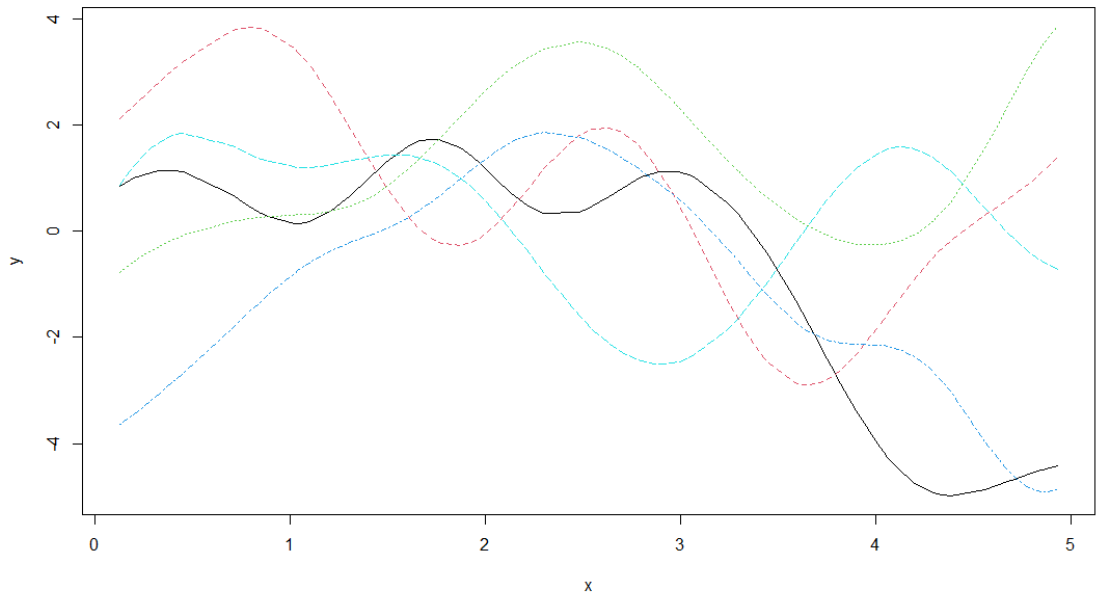
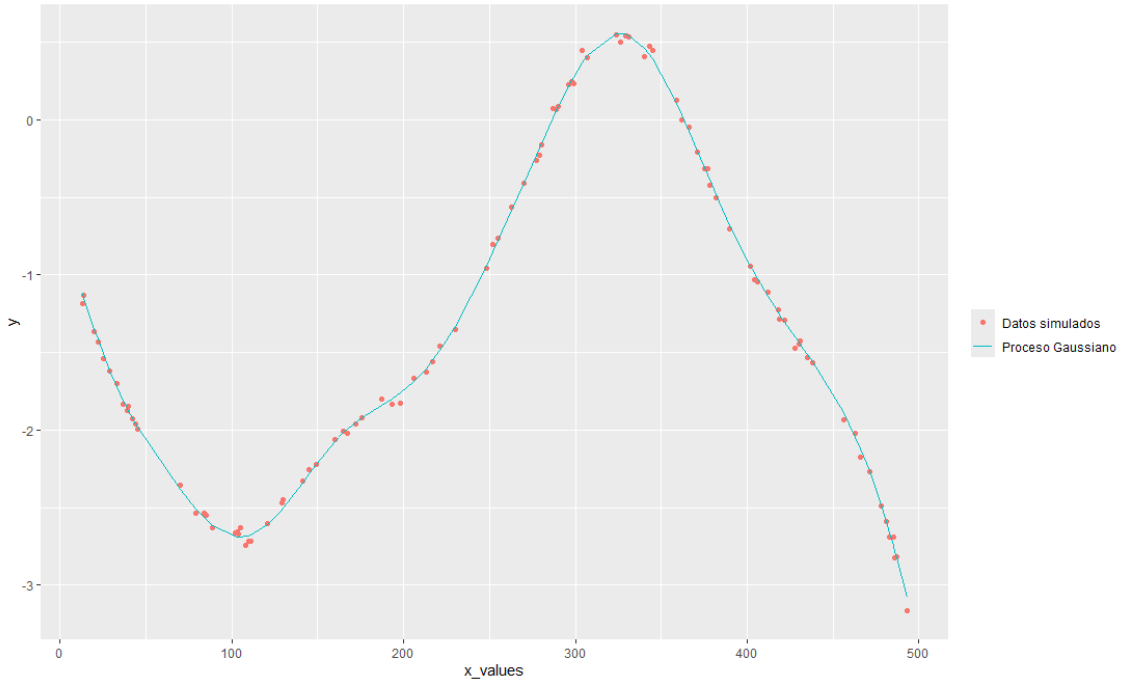
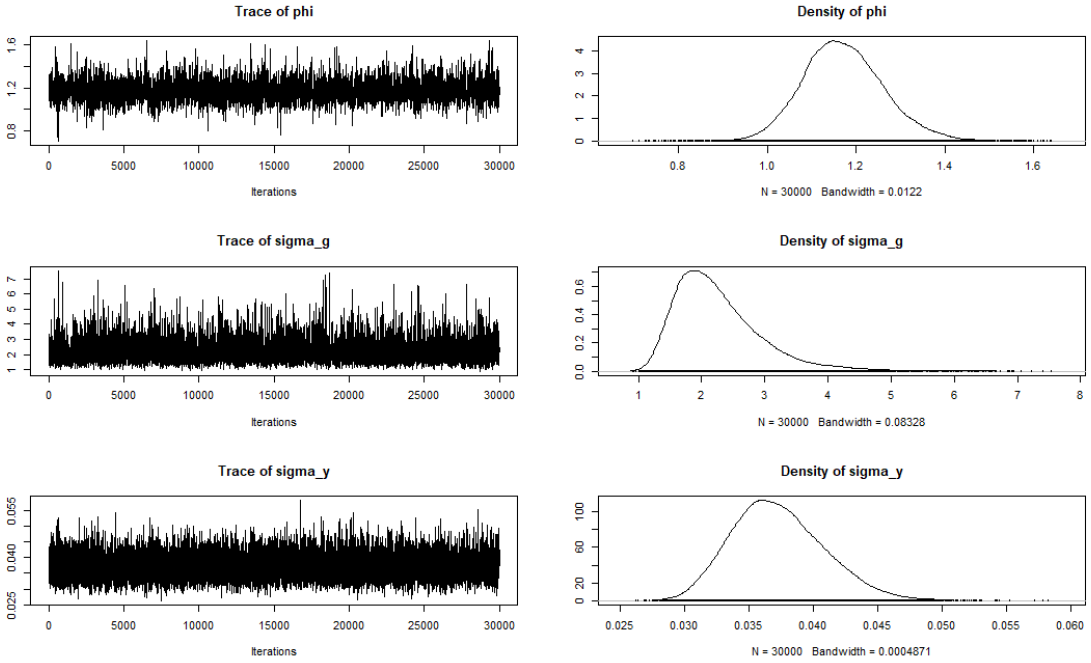


Figura 4.1: 5 realizaciones a priori de un proceso gaussiano con kernel RBF.

Visualizando los datos simulados de aquellos parámetros y el proceso gaussiano que se ajusta a ella de manera perfecta:



Podemos hacer uso del algoritmo de Gibbs para obtener las estimaciones a posteriori de los parámetros σ_g^2, ϕ y σ_y^2, obteniendo entonces las cadenas: Además es posible ver el comportamiento de la convergencia de estas:



Tenemos que el estimador de Bayes para estos 3 parámetros son:

$$\hat{\phi} = 1{,}17211 \quad ; \quad \hat{\sigma}_g^2 = 2{,}26469 \quad ; \quad \hat{\sigma}_y^2 = 0{,}03731$$

Ademas bajo ciertos criterios de convergencia es posible concluir que las cadenas han convergido correctamente:

Convergencia de Gelman y Rubin		
	Point.Esd	Upper C.I
ϕ	1	1.01
σ_g^2	1	1.01
σ_y^2	1	1.00

Los resultados de la convergencia de Gelman y Rubin muestran que el modelo ha convergedido adecuadamente, con valores cercanos a 1 para todos los parámetros.

Tamaño de muestra efectivo		
ϕ	σ_g^2	σ_y^2
1921.422	3334.172	8239.724

Los tamaños de muestra efectivos para los parámetros ϕ, σ_g^2 y σ_y^2 son elevados, lo que indica que la estimación de estos parámetros es precisa y estable.

Función de autocorrelación para cadenas de Markov			
	ϕ	σ_g^2	σ_y^2
Lag 0	1.00000000	1.00000000	1.00000000
Lag 1	0.87915605	0.75143105	0.49186045
Lag 5	0.52974647	0.30445098	0.07057849
Lag 10	0.28924867	0.12735634	0.02775408
Lag 50	0.02178455	0.01105234	0.01128360

La autocorrelación para las cadenas de Markov disminuye rápidamente con el aumento del lag, indicando que las cadenas han convergido y que los valores de los parámetros son independientes después de ciertos lags.

1.1. Predicciones

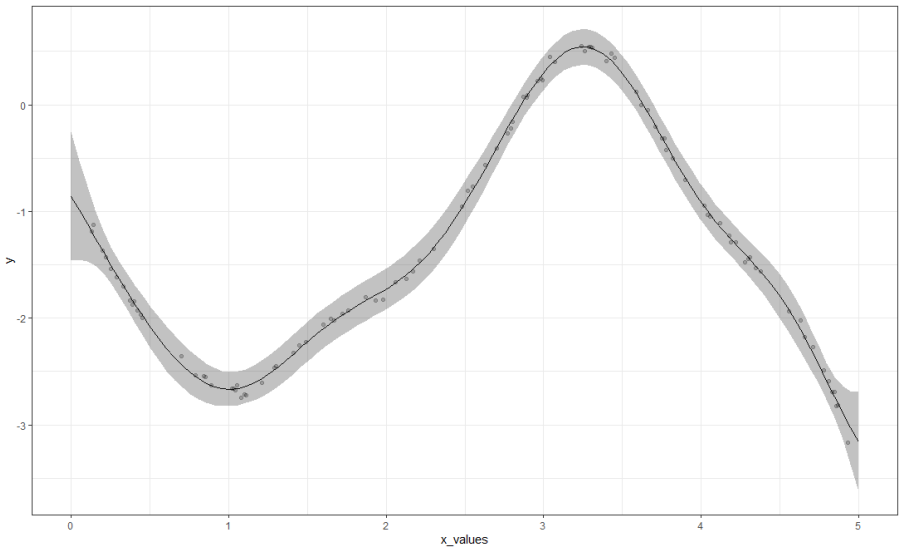


Figura 4.2: Curva a posteriori con ajustada vía kernel Gaussiano

La curva predictiva que hemos obtenido mediante el Proceso Gaussiano no solo se ajusta muy bien a los datos observados, sino que también logra capturar de manera intuitiva el patrón subyacente. La suavidad de la línea y la forma en que sigue la tendencia general, sin sobreajustar ni ignorar las variaciones importantes, nos da a entender que el modelo ha aprendido bien la estructura de los datos de entrenamiento.

Además, las bandas de incertidumbre alrededor de la predicción reflejan de manera transparente dónde el modelo tiene más seguridad y dónde las cosas son menos claras, lo cual es clave para tomar decisiones informadas. Esto no solo es útil desde el punto de vista técnico, sino que también hace que el modelo sea más interpretable y cercano a nuestra intuición.

Ahora, para entender aún mejor cómo se comporta el modelo en puntos específicos, vamos a explorar la distribución predictiva a posteriori de un punto en particular. Esto nos permitirá ver no solo su valor esperado, sino también toda la gama de posibilidades que el modelo considera plausibles, junto con sus probabilidades.

1.2. Un solo punto y su distribución

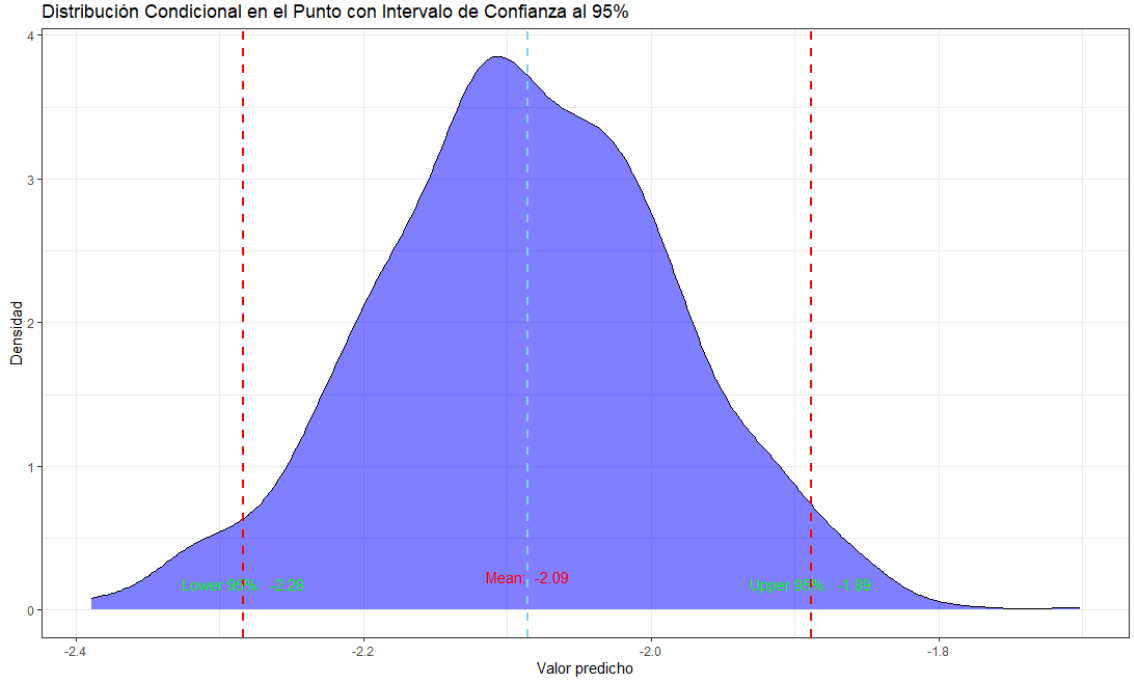


Figura 4.3: Densidad predictiva a posteriori del punto x = 0.5

En el punto x = 0.5, la distribución predictiva a posteriori presenta una media de -2.09, con un intervalo de confianza del 95 % que abarca desde -2.28 hasta -1.89. El intervalo representa el margen de variación que el modelo considera normal para esta predicción.

Esta capacidad de proporcionar estimaciones probabilísticas completas, en lugar de valores puntuales, convierte al modelo en una herramienta particularmente valiosa para aplicaciones donde la evaluación rigurosa de riesgos y la toma de decisiones bajo incertidumbre son aspectos críticos. La transparencia en la comunicación de la incertidumbre asociada a cada predicción constituye un estándar deseable en modelado predictivo, particularmente en dominios científicos y técnicos donde la sobreinterpretación de resultados puntuales puede llevar a conclusiones erróneas.

Capítulo 5

Regresión

Consideremos el problema genérico de regresión, en el cual se busca explicar un resultado y_i como función de una covariable $x_i \in \mathcal{X}$. Por el momento, asumimos que tanto el resultado como la covariable son univariados, y planteamos el problema de regresión como:

$$y_i = f(x_i) + \varepsilon_i, \quad i = 1, \ldots, n$$

Aquí, f es una función de centrado desconocida y ε_i representa los residuos, que usualmente se asumen independientes. En muchos casos, f representa la respuesta media, aunque alternativamente puede representar un cuantil específico, como la mediana. Si tanto la función f como la distribución de los residuos están indexadas por un vector de parámetros de dimensión finita θ, el problema se reduce a una regresión paramétrica tradicional, por ejemplo, una regresión lineal normal.

Esta definición de regresión no paramétrica como una relajación del modelo paramétrico hace natural distinguir entre tres tipos de modelos de regresión no paramétrica:

- Cuando se usa un priori no paramétrico para modelar la distribución de los residuos, es decir, $\varepsilon_i \mid G \overset{\text{iid}}{\sim} G$, con un priori bayesiano no paramétrico (BNP) sobre G;

- Cuando se asume un priori no paramétrico sobre la función de media f;

Nos referimos a estas tres aproximaciones como modelos con distribuciones residuales no paramétricas, funciones de media no paramétricas y regresión completamente no paramétrica, respectivamente. Esta última también se conoce como o regresión de densidades.

Capítulo 6

Aplicaciones

1. Regresión vía DPM

Para realizar regresión vía Modelos de Mezcla de Procesos de Dirichlet vamos a asumir una cierta forma para $f(x_i)$ como por ejemplo:

$$f(x_i) = \sum_{k=0}^{3} \beta_k x_i^k$$

Además vamos a considerar una priori no paramétrica para modelar la distribución de los residuos, por lo que finalmente se obtiene:

$$y_i = \sum_{k=0}^{3} \beta_k x_i^k + \varepsilon_i \quad ; \quad \varepsilon_i \mid G \overset{\text{iid}}{\sim} G$$

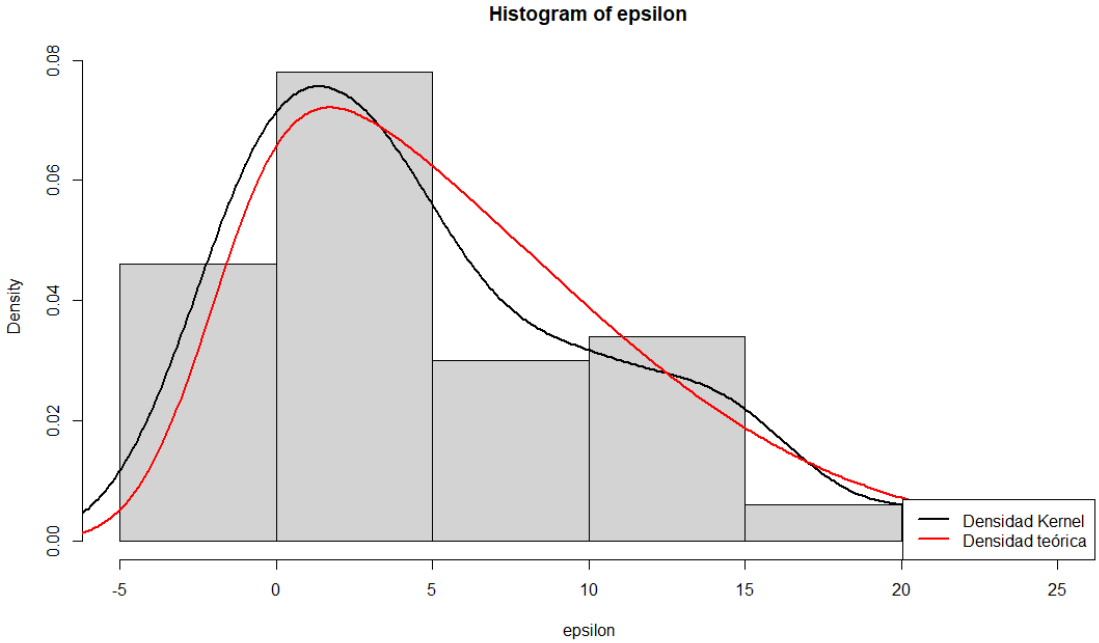
Donde no se tiene por qué asumir una normalidad de estos, o en otras palabras, uno quiere que los errores tengan forma flexible, no necesariamente normal.

En este ejemplo se considerarán datos simulados de una distribución de mezcla, en este caso de:

$$y = 1 + 2x + x^2 + 0{,}5x^3 + \varepsilon_i \quad ; \quad \varepsilon_i \sim \mathcal{SN}(\xi = 0, \omega = 10, \alpha = 5)$$

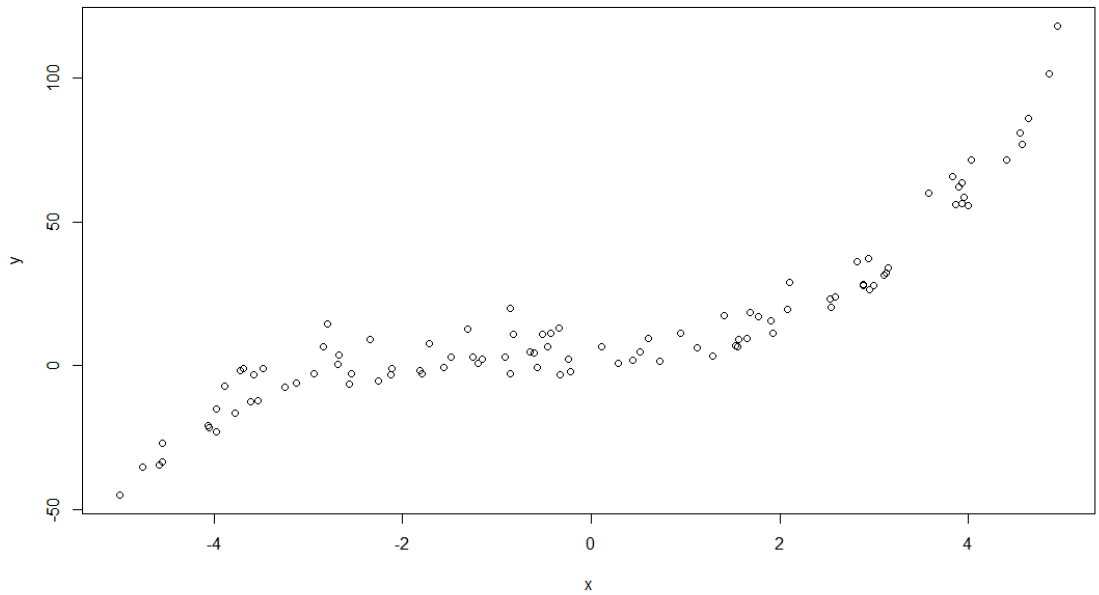
Donde $\mathcal{SN}$ corresponde a la distribución normal asimétrica centrada en 0, con desviación estándar 10 y asimetría 5.

Visualizando el histograma de los errores con su densidad kernel y su densidad teórica:



En la realidad no es posible visualizar estos datos dado que los errores son considerados desconocidos y estimables vía residuos.

Con nuestra estructura ya definida para $f(x_i)$ y para los errores podemos ver entonces la gráfica de puntos de nuestros pares de datos (x, y):



Se evidencia la forma cúbica de los datos dado el patrón que siguen estos mismos. Se estimaron los errores a través de los residuos del modelo ajustado por la instrucción:

```
ml=lm(formula = y ~ I(x) + I(x^2) + I(x^3), data = data)
```

Donde nos entrega un ajuste bastante deplorable pero de todas formas esta nos ayuda a darnos una idea de como se pueden comportar los errores a través de los resiudos de esta.

Se especifica la jerarquía tal que:

$$\begin{cases} \varepsilon_i \mid F \sim F \\ F \sim DP(5, \mathcal{SN}(0, 1, 10)) \end{cases}$$

Entonces se obtinen las densidades a priori:

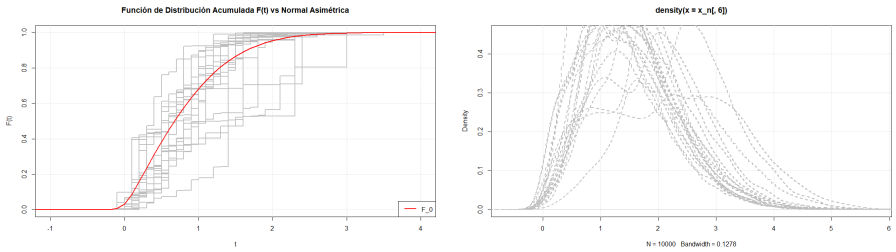


Figura 6.1: Histograma errores y scatterplot datos

Y por último la densidad a posteriori de los residuos realizando un promedio de las 20 densidades a posteriori calculadas. Se introduce además un intervalo de confianza para estas a un nivel de significancia del 5 %

$$\begin{cases} \varepsilon_i \mid F \sim F \\ F \sim DP(5, \mathcal{SN}(0, 1, 10)) \end{cases}$$

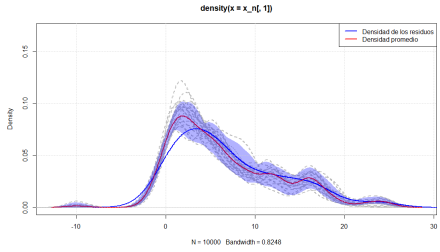


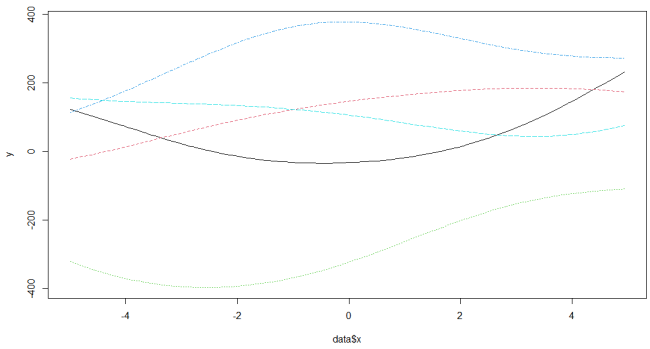
Figura 6.2: Densidad a posteriori de los residuos

2. Regresión vía GP

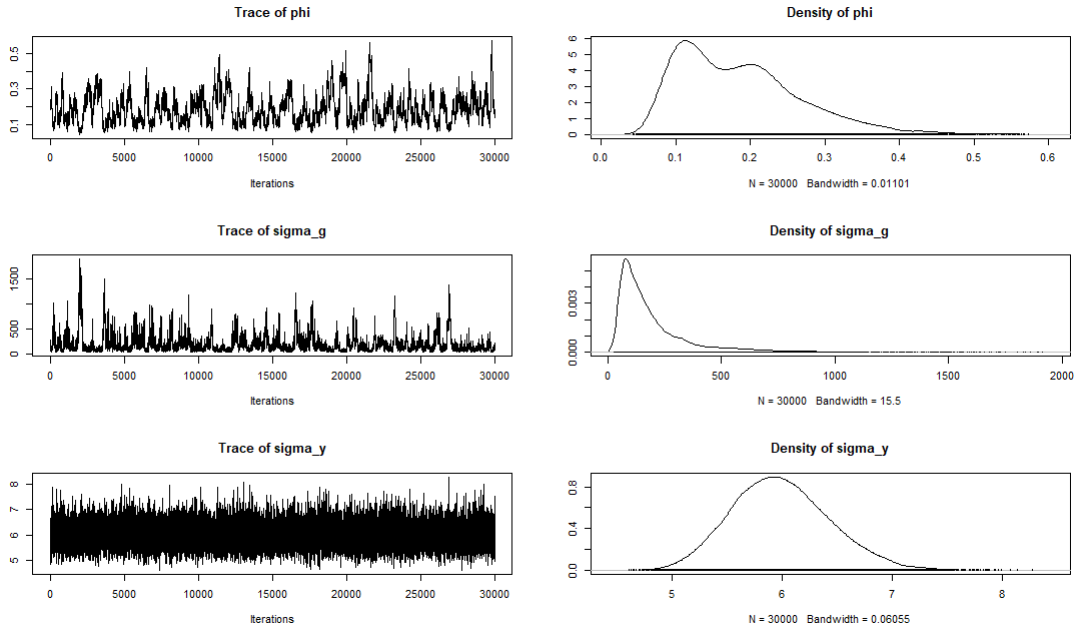
Ahora se quiere estimar $f(x_i)$ en $y_i = f(x_i) + \varepsilon_i$ si es que suponemos que $\varepsilon_i \sim \mathcal{N}(0, \sigma_y^2)$, es este caso se hace uso de un kernel Gaussiano tal que:

$$\sigma_g^2 \exp\left(-\frac{1}{2}\left(\phi||\boldsymbol{x} - \boldsymbol{x}'||\right)^2\right)$$

Entonces es posible visualizar algunas curvas a priori con este tipo de kernel,



Con el modelo ya especificado realizamos métodos tipo MCMC como Gibbs Sampling para poder estimar la densidad de los parámetros ϕ, σ_g^2 y σ_y^2 En cuanto a la convergencia de σ_y^2



se puede ver a simple vista que esta pudo hacerlo sin nignún problema, no así ϕ y σ_g^2, si realizamos los diagnósticos correspondientes,

Convergencia de Gelman y Rubin		
	Point.Esd	Upper C.I
ϕ	1.04	1.10
σ_g^2	1.01	1.02
σ_y^2	1	1.00

Los resultados de la convergencia de Gelman y Rubin muestran que solamente σ_y^2 pudo converger sin problema alguno, no así ϕ y σ_g^2, los cuales tuvieron ciertas dificultades.

Tamaño de muestra efectivo		
ϕ	σ_g^2	σ_y^2
97.60229	170.06629	15897.65000

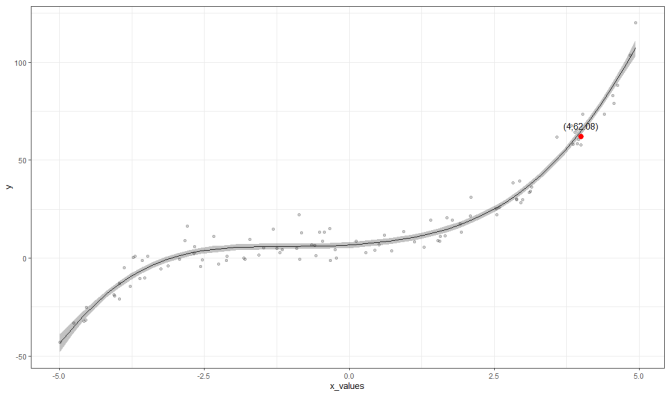
Los tamaños de muestra efectivos para σ_y^2 es elevado, lo que indica que la estimación de estos parámetros es precisa y estable.

En cuanto a ϕ y σ_g^2 se indica una precisión moderada-alta.

Función de autocorrelación para cadenas de Markov			
	ϕ	σ_g^2	σ_y^2
Lag 0	1.00000000	1.00000000	1.00000000
Lag 1	0.9935140	0.9800286	0.310051302
Lag 5	0.9678672	0.9189487	0.004451091
Lag 10	0.9360487	0.8625831	0.011951510
Lag 50	0.7050428	0.5760672	-0.005549717

La alta autocorrelación en ϕ y σ_g^2 (especialmente en Lag 1 y Lag 50) indica mezcla lenta en las cadenas MCMC, lo que sugiere ineficiencia en la sampling y posible necesidad de ajustar el modelo (reparametrizar, aumentar iteraciones o afinar el sampler). En contraste, σ_y^2 muestra una autocorrelación baja y rápida decadencia, señalando un comportamiento aceptable.

Es posible evaluar gráficamente el desempeño del modelo mediante la curva predictiva a posteriori, superpuesta a los datos observados.



Como se muestra en la figura, la curva captura la tendencia subyacente de los datos, mostrando un ajuste efectivo.

Además se introduce un punto $x = 4$ junto con su predicción, es posible obtener su distribucón predictiva a posteriori,

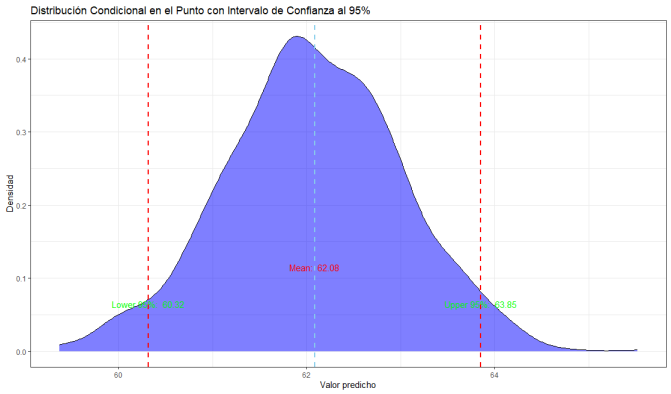
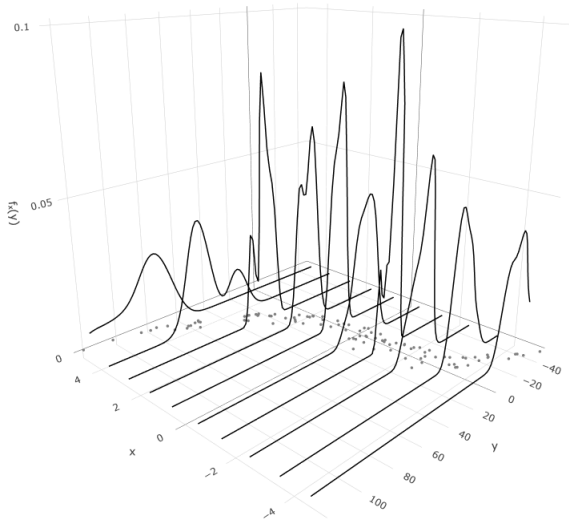


Figura 6.3: Distribución predictiva del punto x = 4

La densidad predictiva a posteriori en $x = 4$ inserta incertidumbre del modelo, con una media estimada en 62.08 y un intervalo del 95 % entre 60.32 y 83.85. Al observar el abanico



de curvas predictivas, podemos apreciar cómo el modelo captura tanto las tendencias dominantes como las discrepancias en los datos. El modelo muestra mayor confianza en las zonas donde las curvas predictivas convergen, indicando patrones consistentes en los datos. En cambio, donde las curvas divergen, revela áreas con mayor incertidumbre.

El método empleado demostró ser altamente efectivo para ajustar los datos, capturando tanto la tendencia central como la variabilidad inherente al conjunto de observaciones. A través de su enfoque probabilístico, no solo proporcionó estimaciones puntuales precisas, sino que también cuantificó de manera rigurosa la incertidumbre asociada a las predicciones.

Capítulo 7

Conclusión

Este trabajo ha explorado los fundamentos y aplicaciones de los modelos bayesianos no paramétricos, centrándose en el Proceso de Dirichlet (DP) y el Proceso Gaussiano (GP). A través de ejemplos y simulaciones, hemos demostrado cómo estos métodos permiten modelar estructuras complejas en los datos sin tener que depender de supuestos paramétricos potentes. El DP, con su capacidad para generar distribuciones discretas y adaptarse a los datos resulta especialmente útil en problemas de estimación de densidades. Por otro lado, el GP ofrece un marco flexible para modelar funciones suaves y capturar patrones en regresión, donde se nos proporcionan intervalos para las distribuciones predictivas de los datos.

La elección entre estos enfoques depende del problema específico: mientras el DP es ideal para datos con agrupamientos naturales, el GP brinda ventajas en contextos donde la relación entre variables es continua y suave. Ambos métodos comparten la capacidad de incorporar incertidumbre de manera explícita, lo que las hace herramiebntas robustas y transparentes para la toma de decisiones. Sin embargo, su implementación requiere buena especificación

Por otro lado las implementaciones de modelos tipo DP o DDP son escazas en la comunidad estadística, dificultando los procedimientos para los menos entendidos del tema. Estas malas prácticas o "egoísmo científico" lo único que se consiguen es un retroceso en conocimiento respecto a estos temas, eventualmente estos métodos dejarán de ser usados de forma convencional dado que existirán métodos más sofisticados y por sobretodo, más documentados.

En conclusión, los modelos presentados aquí representan herramientas poderosas para el análisis estadístico. Futuras líneas de trabajo podrían explorar extensiones como los procesos dependientes de Dirichlet (DDP) o mejoras computacionales para escalar estos métodos a conjuntos de datos más grandes.

Hagamos de la comunidad estadística un espacio de aprendizaje y no de egoísmo, un espacio de prueba y error y no de superioridad y elitismo. La preservación de estos métodos los condena a ser olvidados.

Capítulo 8

Bibliografía

Dahl, David (dic. de 2005). "Sequentially-allocated merge-split sampler for conjugate and nonconjugate Dirichlet process mixture models".

Hanada, M. y S. Matsuura (2022). MCMC from Scratch: A Practical Introduction to Markov Chain Monte Carlo. Springer Nature Singapore. ISBN: 9789811927157. URL: https://books.google.cl/books?id=1nmWEAAAQBAJ.

Jain, Sonia y Radford M. Neal (2004). "A Split-Merge Markov Chain Monte Carlo Procedure for the Dirichlet Process Mixture Model". En: Journal of Computational and Graphical Statistics 13.1, págs. 158-182. ISSN: 10618600. URL: http://www.jstor.org/stable/1391150.

Müller, Peter et al. (2015). Bayesian Nonparametric Data Analysis. 1st. Springer Series in Statistics. Springer. ISBN: 978-3-319-18967-1. DOI: https://doi.org/10.1007/978-3-319-18968-8.

Rizzo, M.L. (2007). Statistical Computing with R. Chapman & Hall/CRC The R Series. Taylor & Francis. ISBN: 9781584885450. URL: https://books.google.cl/books?id=BaHhdqOugjsC.

Capítulo 9

Anexo

https://tiagolazcano.github.io/Codes.R/codes.html