

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

Tiago Lazcano L.

Profesor: Andrés Iturriaga C.

15 de julio de 2025

Tabla de contenidos I

1 Motivación

- Paradoja de la omnipotencia
- Marco teórico

2 Desarrollo teórico

- Distribución Dirichlet
- Proceso de Dirichlet (DP)
 - Definición
 - Representación Stick Breaking
 - Posteriori
 - Ejemplo
- Modelo de Mezcla del Proceso de Dirichlet (DPMM)
 - Representación como densidad
 - ¿Sobreajuste?
- Distribución Normal Multivariada
- Proceso Gaussiano (GP)
 - Definición de Proceso Gaussiano (GP)
 - Kernels
 - Proceso Gaussiano en regresión

Tabla de contenidos II

- Posteriori
- Ejemplo

3 Aplicación

- Regresión vía DPM
- Regresión vía GP
 - Aspectos a mejorar

4 Conclusión

5 Bibliografía

Paradoja de la omnipotencia

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

- Inferencia y aprendizaje Bayesiano
- Regresión
- Modelos no paramétricos
- Procesos estocásticos
- Teoría de la medida

¿Es posible obtener una estimación no paramétrica pero de manera Bayesiana?

Problema	Frecuentista	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

Distribución Dirichlet

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

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

donde:

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

Entonces $\mathbf{p} \sim \text{Dir}(\alpha)$ La distribución Dirichlet es una generalización multicategórica de la distribución Beta y una distribución continua asociada a la distribución multinomial.

Definición Proceso de Dirichlet (DP)

Sea $\alpha > 0$ un parámetro positivo y G_0 una medida de probabilidad definida en un espacio S . Un proceso de Dirichlet $DP(\alpha, G_0)$ es una medida de probabilidad aleatoria 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 $\{B_1, \dots, B_K\}$ del espacio S , la distribución conjunta de las probabilidades asignadas $\{G(B_1), \dots, G(B_K)\}$ sigue una distribución de Dirichlet con parámetros proporcionales a $\alpha G_0(B_i)$, es decir:

$$(G(B_1), G(B_2), \dots, G(B_K)) \sim \text{Dir}(\alpha G_0(B_1), \alpha G_0(B_2), \dots, \alpha G_0(B_K))$$

Decimos entonces:

$$\begin{cases} y \mid G \sim G \\ G \sim DP(\alpha, G_0) \end{cases} \quad ; \text{ se solía decir } \begin{cases} y \mid \theta \sim G \\ \theta \sim \pi(\theta) \end{cases}$$

Definición Proceso de Dirichlet (DP)

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.

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.

Definición Proceso de Dirichlet (DP)

Entonces podemos entender los parámetros correspondientes al proceso de Dirichlet como:

- 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.

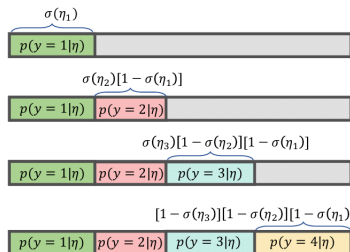
Representación Stick Breaking

Se extraen $S_1, S_2, \dots \sim F_0$, se extraen:

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

Luego

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



Representación Stick Breaking

Por lo tanto, podemos escribir la función de distribución $F(t)$ como una suma infinita de puntos o átomos:

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

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:

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

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

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

Con:

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

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

- Cuando $\alpha \rightarrow 0$, $\bar{F} \rightarrow \hat{F}_n$.
- Cuando $\alpha \rightarrow \infty$, $\bar{F} \rightarrow F_0$.
- Cuando $n \rightarrow \infty$, $\bar{F} \rightarrow \hat{F}_n$; sin embargo, como $\hat{F}_n \rightarrow F$ casi seguramente (Teorema de Glivenko-Cantelli), en realidad $\bar{F} \rightarrow F$.

Comúnmente se realizan N muestras F tal que:

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

Y es posible obtener una curva promedio de forma que:

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

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:

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

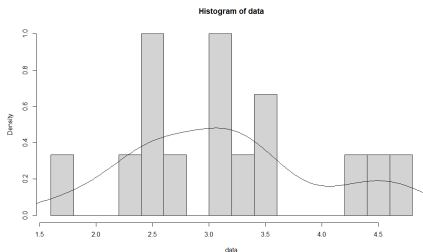


Figura 1: Histograma y densidad Kernel de los datos simulados.

Priors:

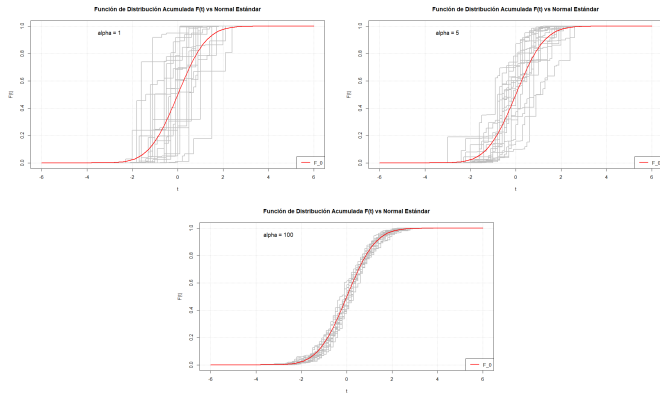


Figura 2: 20 simulaciones de $DP(\alpha, \mathcal{N}(0, 1))$

Ejemplo

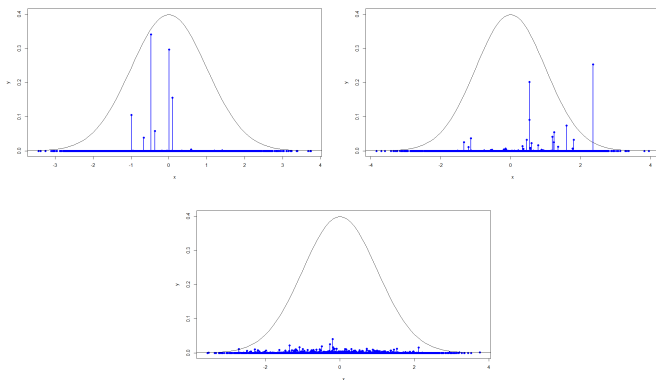


Figura 3: Muestras del proceso de Dirichlet para diferentes valores de α .

Ejemplo

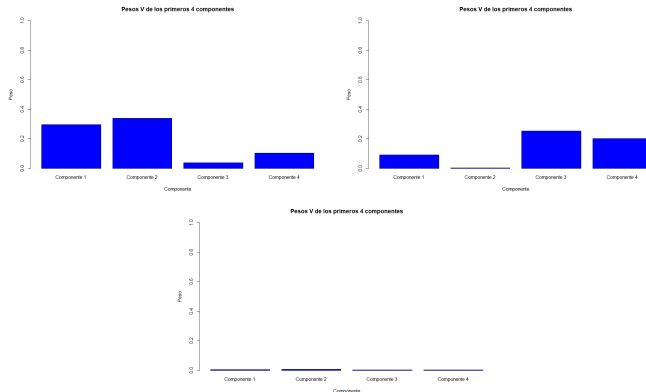


Figura 4: Valores de los primeros cuatro pesos para diferentes valores de α .

Posteriors:

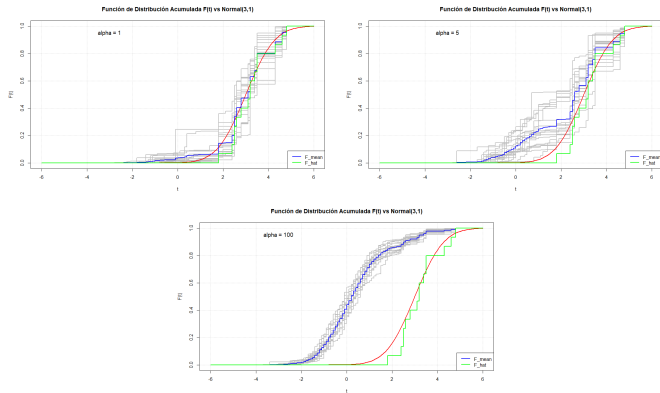


Figura 5: 20 simulaciones de $DP(\alpha + n, \bar{F})$

Representación como densidad

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:

$$\begin{cases} X_i|\theta_i \stackrel{\text{iid}}{\sim} p(x_i|\theta_i) \\ \theta_i | G \stackrel{\text{iid}}{\sim} G \\ G \sim DP(\alpha, G_0) \end{cases}$$

Dado una distribución de probabilidad G definida en Θ , una mezcla de f_θ con respecto a G tiene como función de densidad:

$$f_G(x) = \int f(x|\theta) dG(\theta) = \sum_{j \geq 1} w_j p(x_j|\theta_j)$$

¿Sobreaajuste?

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**.

Distribución Normal Multivariada

Sea $\boldsymbol{\mu} = (\mu_1, \mu_2, \dots, \mu_K)'$ el vector de medias, con $\mu_i \in \mathbb{R}$ y $\boldsymbol{\Sigma}$ la matriz de covarianzas tal que,

$$\boldsymbol{\Sigma} = \begin{pmatrix} \text{Var}(X_1) & \text{Cov}(X_1, X_2) & \cdots & \text{Cov}(X_1, X_K) \\ \text{Cov}(X_2, X_1) & \text{Var}(X_2) & \cdots & \text{Cov}(X_2, X_K) \\ \vdots & \vdots & \ddots & \vdots \\ \text{Cov}(X_K, X_1) & \text{Cov}(X_K, X_2) & \cdots & \text{Var}(X_K) \end{pmatrix}$$

la cual es semidefinida positiva, entonces $\mathbf{X} = (X_1, X_2, \dots, X_K)$ tiene distribución Normal Multivariada si,

$$f_{\mathbf{X}}(\mathbf{x}) = \frac{1}{(2\pi)^{d/2} |\boldsymbol{\Sigma}|^{1/2}} \exp \left(-\frac{1}{2} (\mathbf{x} - \boldsymbol{\mu})' \boldsymbol{\Sigma}^{-1} (\mathbf{x} - \boldsymbol{\mu}) \right)$$

Entonces $\mathbf{X} \sim \mathcal{N}_K(\boldsymbol{\mu}, \boldsymbol{\Sigma})$

Definición de Definición Proceso Gaussiano (GP)

Un Proceso Gaussiano es una extensión de la distribución Gaussiana multivariante a dimensiones infinitas. Esto significa que es posible darle un vector $\mathbf{x} \in \mathbb{R}^n$ (para cualquier n) y el proceso devolverá un nuevo vector $\mathbf{y} \in \mathbb{R}^n$. Cada componente de \mathbf{y} representa la probabilidad de observar x_i según algún gaussiano en la dimensión i .

Sea $\mathcal{K}(\mathbf{x}_i, \mathbf{x}_j \mid \boldsymbol{\tau})$ un kernel bajo los parámetros $\boldsymbol{\tau}$, kernels muy usados son:

- Kernel lineal: $\mathbf{v}^T \mathbf{x} \mathbf{x}'$
- Kernel periódico: $\exp\left(\frac{2}{\ell^2} \sin^2\left(\frac{\pi}{\rho} \|\mathbf{x} - \mathbf{x}'\|\right)\right)$
- Kernel Gaussiano: $\sigma_g^2 \exp\left(-\frac{1}{2} \left(\frac{\|\mathbf{x} - \mathbf{x}'\|}{\ell}\right)^2\right)$

Es **computacionalmente** posible combinar kernels para obtener uno nuevo,

$$\mathcal{K}_c(\mathbf{x}, \mathbf{x}' \mid \boldsymbol{\tau}) = \mathcal{K}_a(\mathbf{x}, \mathbf{x}' \mid \boldsymbol{\tau}) + \mathcal{K}_b(\mathbf{x}, \mathbf{x}' \mid \boldsymbol{\tau})$$

$$\mathcal{K}_c(\mathbf{x}, \mathbf{x}' \mid \boldsymbol{\tau}) = \mathcal{K}_a(\mathbf{x}, \mathbf{x}' \mid \boldsymbol{\tau}) \mathcal{K}_b(\mathbf{x}, \mathbf{x}' \mid \boldsymbol{\tau})$$

No se puede construir una muestra de la función para \mathcal{K}_c multiplicando muestras de \mathcal{K}_a y \mathcal{K}_b , pero puede ser una aproximación útil.

También es posible realizar ciertas transformaciones para ajustar esta.

Proceso Gaussiano en regresión

Sean los datos de entrenamiento $\{(\mathbf{x}_i, y_i)\}_{i=1}^N$, asumiendo que $\mathbb{E}(y_i) = 0$, $\forall i$, además:

$$y_i = f(\mathbf{x}_i) + \varepsilon_i \quad ; \quad \varepsilon_i \sim \mathcal{N}(0, \sigma_\varepsilon^2)$$

Queremos obtener distribuciones predictivas en los datos de prueba $\{\mathbf{x}_i^*\}_{i=1}^M$, por lo que vamos a recolectar estos en matrices/vectores $\mathbf{X}, \mathbf{X}^*, \mathbf{y}, \mathbf{f}, \mathbf{f}^*$, donde:

- \mathbf{X} es una matriz donde cada fila corresponde a un vector de datos de entrenamiento.
- \mathbf{X}^* es una matriz donde cada fila corresponde a un vector de datos de prueba.
- \mathbf{y} es un vector de todas las salidas observadas.
- \mathbf{f} es la salida de la función verdadera no observada para nuestras entradas de entrenamiento.
- \mathbf{f}^* es la salida de la función verdadera no observada para nuestras entradas de prueba.

Sea $\mathbf{K}_{\mathbf{X},\mathbf{X}}$ una matriz de $N \times N$ de todas las similaridades $\mathcal{K}(\mathbf{x}_i, \mathbf{x}_j \mid \boldsymbol{\tau})$, donde $\mathbf{K}_{\mathbf{X},\mathbf{X}^*}$, $\mathbf{K}_{\mathbf{X}^*,\mathbf{X}}$ y $\mathbf{K}_{\mathbf{X}^*,\mathbf{X}^*}$ están definidos de manera similar. Entonces \mathbf{y} y \mathbf{f}^* están distribuidas como una Normal Mutlivariada $N + M$ dimensional,

$$\begin{pmatrix} \mathbf{y} \\ \mathbf{f}^* \end{pmatrix} \sim \mathcal{N} \left(\begin{pmatrix} \mathbf{0} \\ \mathbf{0} \end{pmatrix}, \begin{pmatrix} \hat{\mathbf{K}}_{\mathbf{X},\mathbf{X}} & \mathbf{K}_{\mathbf{X},\mathbf{X}^*} \\ \mathbf{K}_{\mathbf{X}^*,\mathbf{X}} & \mathbf{K}_{\mathbf{X}^*,\mathbf{X}^*} \end{pmatrix} \right)$$

donde $\hat{\mathbf{K}}_{\mathbf{X},\mathbf{X}} = \mathbf{K}_{\mathbf{X},\mathbf{X}} + \sigma_{\varepsilon}^2 \mathbf{I}$

La distribución posteriori sobre \mathbf{f}^* proviene de condicionar:

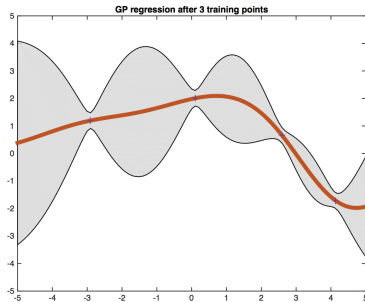
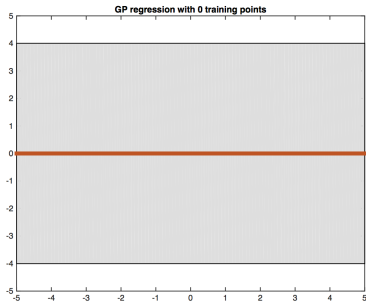
$$\mathbf{f}^* | \mathbf{X}^*, \mathcal{D} \sim \mathcal{N}(\boldsymbol{\mu}_{\mathbf{f}^*}, \boldsymbol{\Sigma}_{\mathbf{f}^*})$$

donde

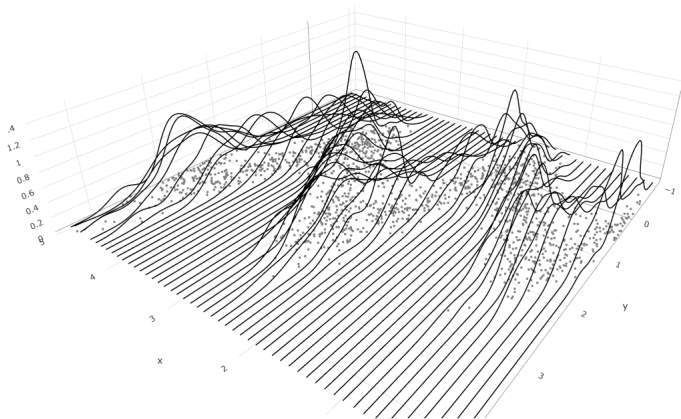
$$\boldsymbol{\mu}_{\mathbf{f}^*} = \mathbf{K}_{\mathbf{X}^*, \mathbf{X}} \hat{\mathbf{K}}_{\mathbf{X}, \mathbf{X}}^{-1} \mathbf{y}$$

$$\boldsymbol{\Sigma}_{\mathbf{f}^*} = \mathbf{K}_{\mathbf{X}^*, \mathbf{X}^*} - \mathbf{K}_{\mathbf{X}^*, \mathbf{X}} \hat{\mathbf{K}}_{\mathbf{X}, \mathbf{X}}^{-1} \mathbf{K}_{\mathbf{X}, \mathbf{X}^*}$$

Ejemplos visuales



Ejemplos visuales



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 $[0, 5]$.

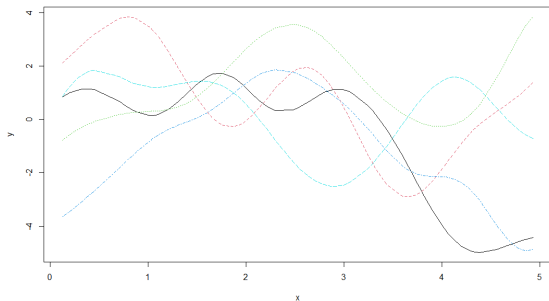


Figura 6: 5 realizaciones a priori de un proceso gaussiano con kernel RBF.

Ejemplo

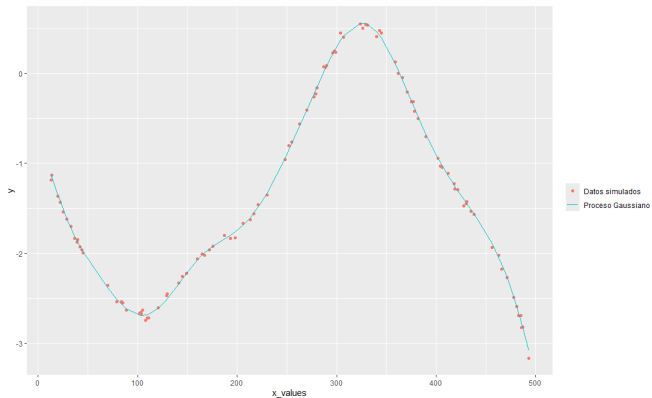


Figura 7: Datos

Modelo

```
model_string <- '
model{
  gp ~ dmnorm(mu,Sigma.inv)
  Sigma.inv <- inverse(Sigma)

  for(i in 1:n_obs)
  {
    mu[i] <- 0
    Sigma[i,i] <- sigma_g^2 + 0.00001
    for(j in (i+1):n_obs) {
      Sigma[i,j] <- sigma_g^2*exp(-(phi^2)*(d[i,j]^2))
      Sigma[j,i] <- Sigma[i,j]
    }

    y[i]~dnorm(gp[i],sigma_y^-2)
  }

  sigma_g ~ dt(0,10^-2,1)T(0,)
  phi ~ dt(0,4^-2,1)T(0,)
  sigma_y ~ dt(0,10^-2,1)T(0,)
}',
```


Convergencia de cadenas

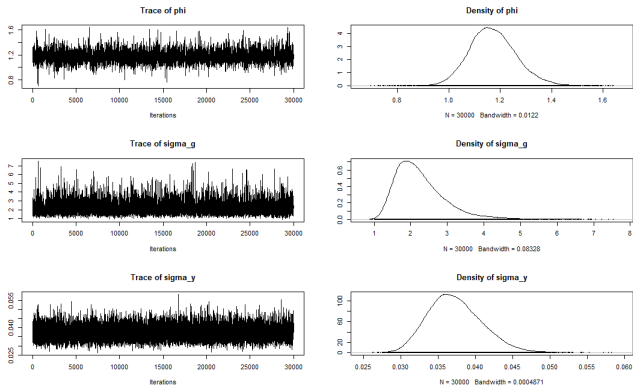


Figura 8: Convergencia de los parámetros

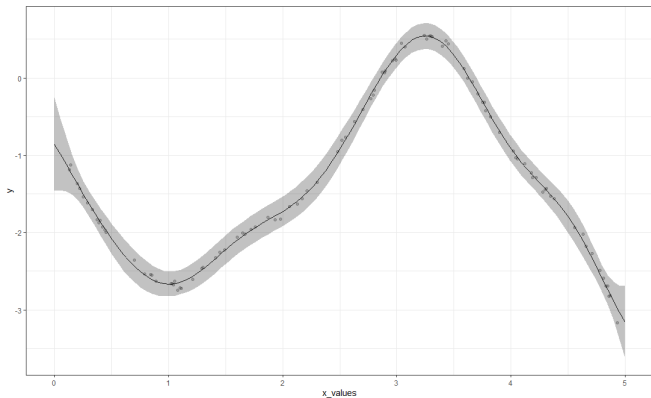


Figura 9: Curva predictiva

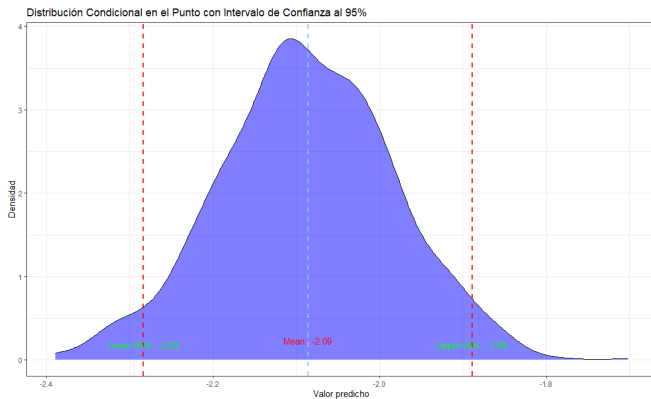


Figura 10: Distribución predictiva del punto $x = 0.5$

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

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

Donde \mathcal{SN} corresponde a la distribución normal asimétrica centrada en -2 , con desviación estándar 10 y asimetría 5.

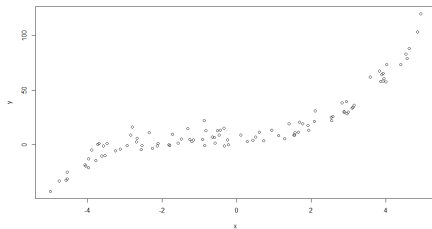
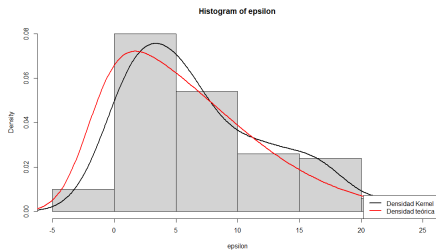


Figura 11: $\alpha = 5$

Regresión vía DPM

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)
```

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

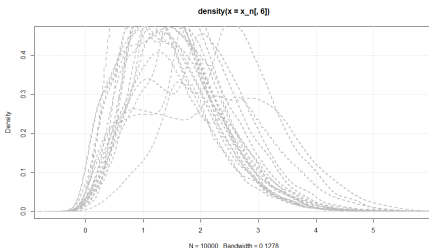
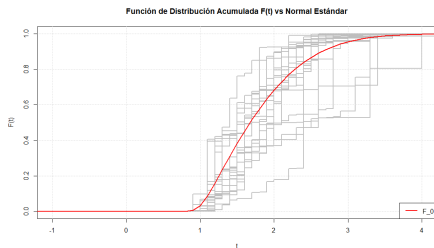


Figura 12: Histograma errores y scatterplot datos

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

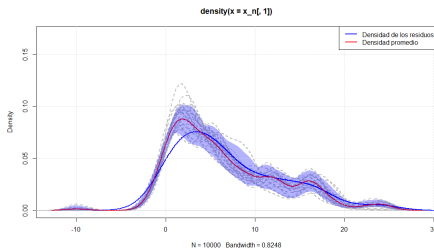
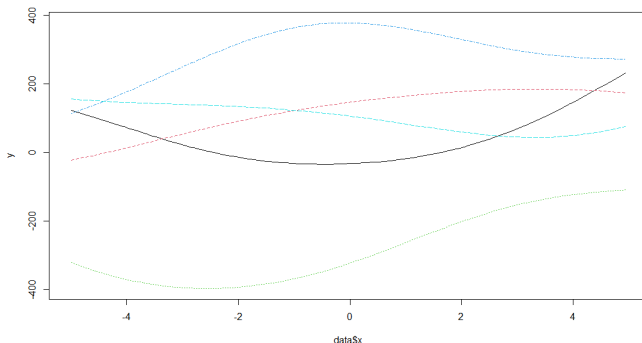


Figura 13: Densidad a posteriori de los residuos

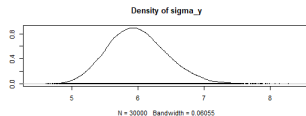
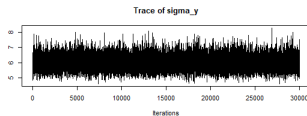
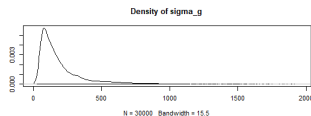
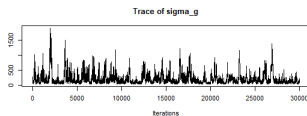
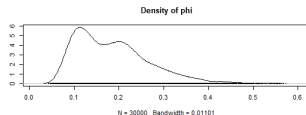
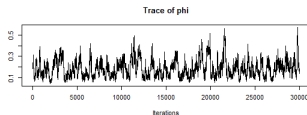
Regresión vía GP

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:

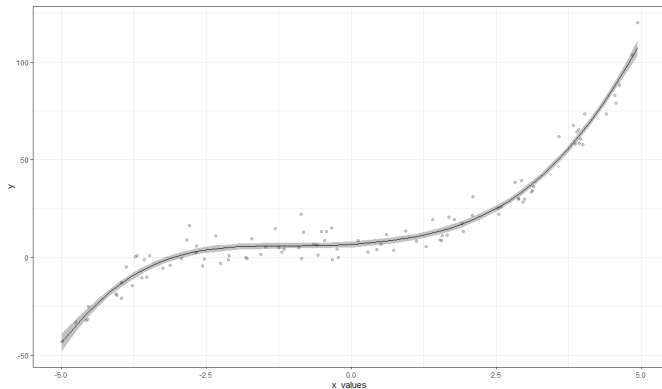
$$\sigma_g^2 \exp\left(-\frac{1}{2} (\phi\|\mathbf{x} - \mathbf{x}'\|)^2\right)$$



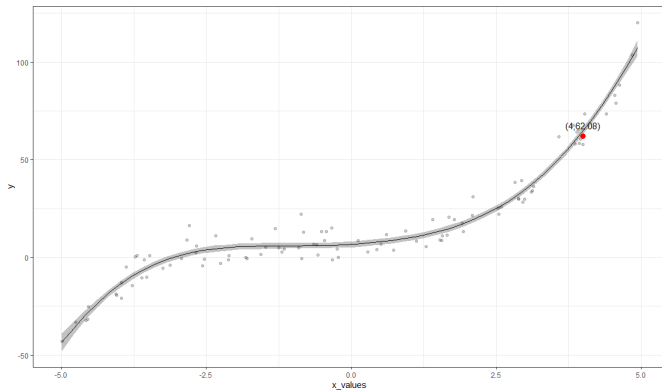
Regresión vía GP



Regresión vía GP



Regresión vía GP



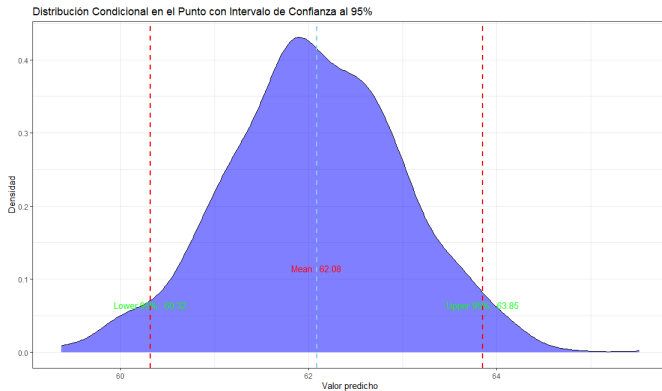
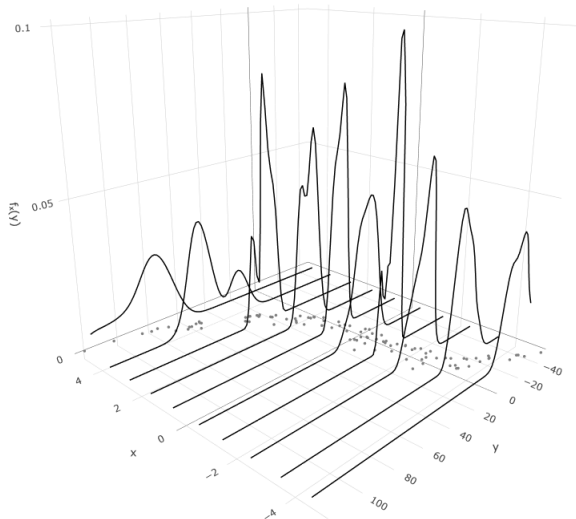






Figura 14: Distribución predictiva del punto $x = 4$

Regresión vía GP



- Implementación de la distribución Normal Asimétrica en JAGS.
- Uso de un Kernel más adecuado para el problema, como por ejemplo uno polinomial.
- Cómo evaluar el ajuste de los modelos, usando métricas como el RMSE o la validación cruzada.
- Ser más flexible a la hora de modelar hiperparámetros en DP y GP.

A veces, los modelos excesivamente complejos dificultan la interpretación y pueden generar sobreajuste, mientras que los modelos demasiado rígidos no logran capturar toda la variabilidad de los datos. Este trabajo demuestra que los Procesos de Dirichlet y los Procesos Gaussianos logran un buen equilibrio, siendo lo suficientemente flexibles para ajustarse a los datos sin complicar su comprensión. Lo importante es encontrar un modelo que sea flexible pero lo suficientemente claro para obtener resultados fáciles de interpretar.

-  Dahl, David (dic. de 2005). “Sequentially-allocated merge-split sampler for conjugate and nonconjugate Dirichlet process mixture models”. En.
-  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=BaHhdq0ugjsC>.