

INSTITUTO TECNOLÓGICO AUTÓNOMO DE MÉXICO



UN MODELO BAYESIANO Y NO PARAMÉTRICO  
DE REGRESIÓN SOBRE CUANTILES

TESIS

QUE PARA OBTENER EL TÍTULO

LICENCIADO EN MATEMÁTICAS APLICADAS

PRESENTA

CARLOS OMAR PARDO GÓMEZ

ASESOR: DR. JUAN CARLOS MARTÍNEZ OVANDO

México, D.F.

2017

Con fundamento en los artículos 21 y 27 de la Ley Federal del Derecho de Autor y como titular de los derechos moral y patrimonial de la obra titulada **"UN MODELO BAYESIANO Y NO PARAMÉTRICO DE REGRESIÓN SOBRE CUANTILES"**, otorgo de manera gratuita y permanente al Instituto Tecnológico Autónomo de México y a la Biblioteca Raúl Baillères Jr., la autorización para que fijen la obra en cualquier medio, incluido el electrónico, y la divulguen entre sus usuarios, profesores, estudiantes o terceras personas, sin que pueda percibir por tal divulgación una contraprestación.

CARLOS OMAR PARDO GÓMEZ

---

FECHA

---

FIRMA

*A Mago.*

# Agradecimientos

¡Muchas gracias a todos!

# Prefacio

El centro de esta tesis es describir un modelo de *regresión sobre cuantiles*, debido a las bondades que tiene sobre el comúnmente usado análisis de *regresión sobre la media*. Además, aceptando los axiomas de la Estadística Bayesiana, permite incorporar conocimiento previo del modelador. Por otra parte, el modelo es no paramétrico, aumentando la flexibilidad en su forma.

El capítulo 1 describe la importancia de las aproximaciones distintas a la *regresión sobre la media*, así como la evolución histórica de este tipo de modelos. El capítulo 2 introduce al paradigma bayesiano y sus metodologías generales. El capítulo 3 se centra en los modelos bayesianos tradicionales de regresión, tanto a la media, como sobre cuantiles. El capítulo 4 plantea la especificación no paramétrica particular del modelo de esta tesis, separándolo de los tradicionales. El capítulo 5 introduce las variables latentes y algoritmos necesarios para realizar inferencia y predicción con lo expuesto en los capítulos anteriores. El capítulo 6 muestra algunas aplicaciones del modelo, así como los resultados obtenidos de evaluarlo en diversos conjuntos de datos. Finalmente, el capítulo 7 hace referencia a las conclusiones finales de esta tesis, además de describir el trabajo futuro que se podría desarrollar, retomando las ideas de ésta.

# Índice general

<b>1. Introducción</b>	<b>7</b>
<b>2. Paradigma bayesiano</b>	<b>10</b>
2.1. Axiomas . . . . .	10
2.2. Inferencia . . . . .	10
2.3. Propiedad conjugada . . . . .	13
<b>3. Modelos de regresión</b>	<b>14</b>
3.1. Concepto general . . . . .	14
3.2. Regresión a la media . . . . .	15
3.2.1. Modelo tradicional . . . . .	16
3.3. Regresión sobre cuantiles . . . . .	19
3.3.1. Modelo tradicional . . . . .	20
<b>4. Especificación no paramétrica</b>	<b>23</b>
4.1. Motivación . . . . .	23
4.2. En $f_p$ , vía Procesos Gaussianos . . . . .	26
4.2.1. Procesos Gaussianos . . . . .	26
4.2.2. Definiciones y notación . . . . .	28
4.2.3. Funciones de covarianza . . . . .	29

4.2.4.	Predicción . . . . .	32
4.3.	En $\varepsilon_p$ , vía Procesos de Dirichlet . . . . .	34
4.3.1.	Procesos de Dirichlet . . . . .	34
4.3.2.	Distribución posterior . . . . .	37
4.3.3.	Distribución predictiva . . . . .	38
4.3.4.	Esquema de urna de Blackwell-MacQueen . . . .	39
4.3.5.	Proceso estocástico de rompimiento de un palo .	41
4.3.6.	Modelo general de mezclas infinitas de Dirichlet .	42
4.3.7.	Modelo de mezclas infinitas de Dirichlet para la regresión sobre cuantiles . . . . .	43
<b>5.</b>	<b>Modelo GPD<sub>P</sub> para regresión sobre cuantiles</b>	<b>45</b>
5.1.	Definición . . . . .	45
5.2.	Inferencia con <i>Gibbs Sampler</i> . . . . .	47
5.2.1.	Actualización del error . . . . .	47
5.2.2.	Actualización de la tendencia . . . . .	49
5.3.	Predicción . . . . .	50
5.4.	Parámetros iniciales . . . . .	51
5.4.1.	Función de medias $m_{f_p}$ . . . . .	52
5.4.2.	<i>Gamma-Inversas</i> de $\lambda$ y el Proceso de Dirichlet .	52
5.4.3.	Parámetro de concentración $\alpha$ . . . . .	53
5.5.	Paquete <i>GPDPQuantReg</i> en R . . . . .	54
	<b>Bibliografía</b>	<b>56</b>

# Capítulo 1

## Introducción

Detrás de cualquier modelo de regresión, la intención es explicar una variable dependiente en función de un conjunto de variables independientes, suponiendo cierto error aleatorio. Ha sido común resumir esta dependencia mediante alguna medida de tendencia central, condicionada a los valores de las covariables.

La medida de tendencia central tradicionalmente usada ha sido la media, dando lugar a los modelos de *regresión sobre la media*, con sus variantes lineal o no lineal, simple o múltiple. Este tipo de modelos tiene un buen número de ventajas, entre las que destacan el bajo costo de calcularlos y la facilidad de interpretación. Sin embargo, como mencionan Hao & Naiman (2007), tienen tres grandes limitaciones.

La primera es que al resumir la relación entre la variable dependiente y las independientes con el valor esperado, no necesariamente se puede extender la inferencia a valores lejanos a la media, que suelen ser de interés en ciertos contextos, como los seguros o las finanzas.

La segunda es que los supuestos de este tipo de modelos no siempre se cumplen en el mundo real. Por ejemplo, el supuesto de homocedasticidad; es decir, la varianza no es constante, sino cambia en sincronía con



distintos valores de las covariables. También es posible que fenómenos de estudio tengan distribuciones de colas pesadas, principalmente en las ciencias sociales. Esto da lugar a valores atípicos, mismos que no suelen ser manejados como se desearía por los modelos de *regresión sobre la media*.

La tercera es que no permiten conocer las propiedades y forma de la distribución completa. Por ejemplo, la asimetría es una característica importante en estudios de ingreso, impuestos, esperanza de vida y, en general, en estudios de desigualdad.

Debido a esto, desde mitades del siglo XVIII han surgido alternativas a este tipo de modelos, siendo la primera los modelos de *regresión sobre la mediana*. De nueva cuenta se buscó una medida de tendencia central, pero con otras bondades. Por ejemplo, ser una mejor medida informativa para distribuciones asimétricas y menos susceptible a valores atípicos.

Así como los *modelos de regresión sobre la media* son comúnmente relacionados con la minimización de los errores cuadráticos, los *modelos de regresión sobre la mediana* lo son con la minimización de los errores absolutos. Debido a la no diferenciabilidad, tuvieron que pasar muchos años para que lograran ser viables, hasta que el poder computacional y los algoritmos de Programación Lineal lo permitieron.

Cabe recordar que el cuantil  $p$ -ésimo es aquel valor tal que una proporción  $p$  de los valores están por debajo de él, y una proporción  $1 - p$ , por arriba. Así, la mediana es un caso particular de un cuantil, específicamente el 0.5-ésimo. Esto abre la idea de que otros cuantiles también podrían ser modelados en función de las covariables, y no necesariamente tienen que ser una medida de tendencia central.

Los *modelos de regresión sobre cuantiles* fueron introducidos por Koenker & Bassett (1978), y han permitido concentrarse en valores de interés para los modeladores, sin importar que estén alejados de la media. Además, el cálculo de diversos cuantiles para un mismo fenómeno ha

permitido entender mejor la forma y propiedades de las distribuciones condicionales de la variable de respuesta.

En el paradigma bayesiano, el desarrollo de este tipo de modelos ha sido lento. Walker & Mallick (1999), Kottas & Gelfand (2001) y Hanson & Johnson (2002) desarrollaron modelos para la mediana, suponiendo una distribución no paramétrica del error. Yu & Moyeed (2001) y Tsionas (2003) desarrollaron inferencia paramétrica, basados en la distribución asimétrica de Laplace para los errores. Por otro lado, Lavine (1995) y Dunson & Taylor (2005) usaron una perspectiva distinta y propusieron una aproximación de la verosimilitud para cuantiles.

Las limitantes de estos trabajos han sido que, aunque han dado formas flexibles a la distribución del error, han estado basados en funciones lineales para describir la relación entre la variable de respuesta y las covariables, o han tenido que recurrir a estimaciones no probabilísticas o no bayesianas, para resolver alguna parte del problema.

Entendiendo como *modelo no paramétrico* a aquel en el que el número de parámetros no está previamente definido, sino que depende de los datos, esta tesis rescata las ideas de Kottas *et al.* (2007) para proponer un modelo bayesiano y no paramétrico, útil en el contexto de *regresión sobre cuantiles*.

## Capítulo 2

# Paradigma bayesiano<sup>1</sup>

### 2.1. Axiomas

Esta tesis da como aceptados los axiomas de la Estadística Bayesiana, mismos que pueden ser encontrados, por ejemplo, en Fishburn (1986). Por lo tanto, entiende a dicho paradigma como el coherente para hacer estadística, cuando una toma de decisión con incertidumbre es el objetivo final del estudio.

### 2.2. Inferencia

Un problema clásico de la estadística es el de hacer predicción, utilizando la información de los datos que ya han sido observados. Por ejemplo, es posible pensar que ya se tiene el conjunto de  $n$  datos observados  $\{y_1, \dots, y_n\}$  y se desea hacer predicción acerca del valor del dato  $y_{n+1}$ , que aún no ha sido observado. Para esto, se podría usar la

---

<sup>1</sup>Las ideas de este capítulo son retomadas de Denison (2002).

probabilidad condicional

$$\mathbb{P}(y_{n+1}|y_1, \dots, y_n) = \frac{\mathbb{P}(y_{n+1} \cap \{y_1, \dots, y_n\})}{\mathbb{P}(y_1, \dots, y_n)} = \frac{\mathbb{P}(y_1, \dots, y_n, y_{n+1})}{\mathbb{P}(y_1, \dots, y_n)},$$

pero esto requeriría conocer la función conjunta, misma que puede ser compleja por la estructura de dependencia de los datos.

No tiene mucho sentido suponer una estructura de independencia entre ellos, porque entonces el conjunto de observaciones  $\{y_1, \dots, y_n\}$  no daría información alguna para  $y_{n+1}$ . Pero se puede suponer una distribución condicionalmente independiente. Es decir, se supone que cada una de las  $y_i$ 's tiene una misma distribución paramétrica, con vector de parámetros  $\theta$ , y se cumple que

$$\mathbb{P}(y_{k+1}, y_k|\theta) = \mathbb{P}(y_k|\theta) \times \mathbb{P}(y_{k+1}|\theta).$$

Siguiendo el mismo razonamiento, es posible obtener que

$$\mathbb{P}(y_1, \dots, y_n|\theta) = \prod_{i=1}^n \mathbb{P}(y_i|\theta).$$

Al igual que en otros paradigmas, se supone a  $\theta$  como constante, pero desconocido, y la tarea es estimarlo. Una particularidad del paradigma bayesiano es expresar la incertidumbre que tiene el modelador acerca del valor verdadero mediante la asignación de una distribución a  $\theta$ , sujeta la información inicial o conocimiento previo que se tenga del fenómeno ( $H$ ). Es decir,  $\mathbb{P}(\theta|H)$ . Como una simplificación de la notación, en la literatura normalmente se escribe como  $\mathbb{P}(\theta) = \mathbb{P}(\theta|H)$  y se conoce como la *probabilidad inicial* del parámetro.

Regresando al problema inicial, y bajo los supuestos recién mencio-

nados, es importante notar que es posible escribir

$$\mathbb{P}(y_{n+1}|y_1, \dots, y_n) = \int_{\Theta} \mathbb{P}(y_{n+1}|\theta) \mathbb{P}(\theta|y_1, \dots, y_n) d\theta,$$

donde a su vez, usando el **Teorema de Bayes**, se obtiene que

$$\mathbb{P}(\theta|y_1, \dots, y_n) = \frac{\mathbb{P}(y_1, \dots, y_n|\theta) \times \mathbb{P}(\theta)}{\mathbb{P}(y_1, \dots, y_n)},$$

que en el paradigma bayesiano se conoce como la *probabilidad posterior* del parámetro.

Se puede observar que el denominador no depende de  $\theta$ , por lo que normalmente la probabilidad no se expresa como una igualdad, sino con la proporcionalidad

$$\mathbb{P}(\theta|y_1, \dots, y_n) \propto \mathbb{P}(y_1, \dots, y_n|\theta) \times \mathbb{P}(\theta),$$

y sólo difiere de la igualdad por una constante que permita que, al integrar sobre todo el soporte de  $\theta$ , el resultado sea igual a 1.

Cabe resaltar que el factor  $\mathbb{P}(y_1, \dots, y_n|\theta)$  es lo que se conoce también en otros paradigmas como *verosimilitud*, y que en caso de independencia condicional puede ser reescrito como

$$\mathbb{P}(y_1, \dots, y_n|\theta) = \prod_{i=1}^n \mathbb{P}(y_i|\theta).$$

Por lo tanto, es posible afirmar que el aprendizaje en el paradigma bayesiano se obtiene como

$$Posterior \propto Verosimilitud \times Inicial,$$

es decir, surge de conjuntar el conocimiento inicial con la información contenida en los datos.

Es importante notar que bajo este enfoque se obtiene una distribución de probabilidad completa para el pronóstico de  $y_{n+1}$ . Esta se puede utilizar para el cálculo de estimaciones puntuales o intervalos (que en el caso del paradigma bayesiano son llamados de *probabilidad*) mediante funciones de utilidad o pérdida, y haciendo uso de la Teoría de la Decisión.

## 2.3. Propiedad conjugada

En los casos en los que la probabilidad posterior tiene la misma familia de distribución que la inicial, sólo siendo distintas en el valor de los parámetros, se dice que la distribución inicial y la posterior son **conjugadas**.

Esta propiedad es conveniente, porque permite a la distribución posterior tener forma analítica cerrada, evitando tener que usar métodos numéricos para aproximarla. Además permite ver de forma más clara cómo afectan los datos a la actualización, respecto a la distribución inicial.

Algunas de las familias conjugadas más conocidas en el caso continuo son la *Normal-Normal*, *Normal-Gamma* o la *Normal-Gamma Inversa*, donde la primer distribución es la de los datos y la segunda la de los parámetros. También en el caso discreto es popular el uso de la *Bernoulli-Beta* o la *Poisson-Gamma*.

## Capítulo 3

# Modelos de regresión

### 3.1. Concepto general

Los modelos de regresión tienen como objetivo describir la distribución de una variable aleatoria  $y \in \mathbb{R}$ , normalmente conocida como la *variable de respuesta*, condicional a los valores de las variables  $x \in \mathbb{R}^n$ , conocidas como *covariables* o *variables de entrada*. Visto en términos matemáticos, se puede expresar como

$$y|x \sim \mathbb{P}(y|x).$$

Si bien esta relación se da por hecha y es fija, normalmente es desconocida. Por lo tanto, la intención de estos modelos es realizar alguna aproximación de ella. Dado que es complicado aproximar con exactitud toda la distribución, comúnmente se enfocan en una medición particular, como la media o la mediana.

## 3.2. Regresión a la media

La *regresión a la media* es el caso particular más usado de los modelos de regresión, tanto en el paradigma bayesiano, como en otros. Esto sucede debido al bajo uso de recursos, además de su capacidad interpretativa.

En notación probabilística, retomando el hecho de que  $y|x \sim \mathbb{P}(y|x)$ , busca aproximar a la función  $f$ , tal que

$$\mathbb{E}(y|x) = f(x).$$

Para hacer esto, normalmente se vale del supuesto que

$$y = f(x) + \varepsilon,$$

con  $\varepsilon \in \mathbb{R} \sim \mathcal{N}(0, \sigma^2)$  (denominado comúnmente como el *error aleatorio*), y siendo  $f : \mathbb{R}^n \rightarrow \mathbb{R}$  y  $\sigma^2 \in \mathbb{R}^+$  desconocidas, de forma que

$$y|x, f, \sigma^2 \sim \mathcal{N}(f(x), \sigma^2).$$

Además, se supone independencia entre  $\varepsilon_s$ , es decir, para toda  $\bar{\varepsilon} \neq \hat{\varepsilon}$ ,  $\bar{\varepsilon}$  y  $\hat{\varepsilon}$  son independientes. Por lo tanto, sean  $\bar{x}$  las covariables asociadas a la variable de respuesta  $\bar{y}$ , y  $\hat{x}$  las asociadas a  $\hat{y}$ , se tiene que  $\bar{y}|\bar{x}, f, \sigma^2$  es condicionalmente independiente a  $\hat{y}|\hat{x}, f, \sigma^2$ .



### 3.2.1. Modelo tradicional <sup>1</sup>

La *regresión lineal a la media* es el caso particular más usado en el contexto de *regresión a la media*. Consiste en definir

$$f(x) = x^T \beta,$$

donde  $\beta \in \mathbb{R}^n$  se piensa con valores constantes, pero desconocidos, y la tarea es estimarlos, al igual que  $\sigma^2$ .

Para hacer esto, el enfoque bayesiano le asigna una distribución inicial de probabilidad a ambos parámetros, reflejando la incertidumbre que tiene el modelador acerca de su valor real. Es decir, sea  $H$  la hipótesis o el conocimiento previo al que tiene acceso el modelador, se tiene que

$$\beta, \sigma^2 \sim \mathbb{P}(\beta, \sigma^2 | H).$$

A partir de este momento se omitirá escribir la distribución condicional respecto a  $H$  por simplificación de la notación, pero es importante no olvidar su existencia.

Sea  $\{(x_i, y_i) | x_i \in \mathbb{R}^n, y_i \in \mathbb{R}, i \in \{1, \dots, m\}\}$  el conjunto de datos observados de las variables de respuesta y de las covariables. Es posible representar este mismo conjunto con la notación matricial  $\{X, Y | X \in \mathbb{R}^{m \times n}, Y \in \mathbb{R}^m\}$ . Sea  $\mathcal{E} \in \mathbb{R}^m$  el vector de errores aleatorios, tal que  $\mathcal{E} \sim \mathcal{N}(0, \sigma^2 I)$ . El modelo se puede reescribir como:

$$Y = X\beta + \mathcal{E} \sim \mathcal{N}(X\beta, \sigma^2 I).$$

---

<sup>1</sup>Algunas ideas de esta subsección son retomadas de Denison (2002) y Bannerjee (2008).

Por el Teorema de Bayes,

$$\begin{aligned}\mathbb{P}(\beta, \sigma^2 | Y, X) &= \frac{\mathbb{P}(Y|X, \beta, \sigma^2) \times \mathbb{P}(\beta, \sigma^2 | X)}{P(Y|X)} \\ &= \frac{\mathbb{P}(Y|X, \beta, \sigma^2) \times \mathbb{P}(\beta, \sigma^2)}{\mathbb{P}(Y|X)} \\ &\propto \mathbb{P}(Y|X, \beta, \sigma^2) \times \mathbb{P}(\beta, \sigma^2),\end{aligned}$$

donde  $\mathbb{P}(Y|X, \beta, \sigma^2)$  es la verosimilitud de los datos observados y se puede calcular como  $\mathbb{P}(Y|X, \beta, \sigma^2) = \mathcal{N}(X\beta, \sigma^2 I) = \prod_{i=1}^m \mathcal{N}(x_i^T \beta, \sigma^2)$ . Por otro lado,  $\mathbb{P}(\beta, \sigma^2)$  es la distribución inicial de los parámetros.

Por conveniencia analítica, hay una distribución inicial comúnmente usada para los parámetros  $\beta$  y  $\sigma$  debido a que es conjugada respecto a la distribución Normal de los datos. Su nombre es *Normal-Gamma Inversa (NGI)* y se dice que  $\beta, \sigma^2 \sim \mathcal{NGI}(M, V, a, b)$ , si

$$\begin{aligned}\mathbb{P}(\beta, \sigma^2) &= \mathbb{P}(\beta | \sigma^2) \times \mathbb{P}(\sigma^2) \\ &= \mathcal{N}(\beta | M, \sigma^2 V) \times \mathcal{GI}(\sigma^2 | a, b) \\ &\propto (\sigma^2)^{-(a+(n/2)+1)} \exp \left( -\frac{(\beta - M)^T V^{-1} (\beta - M) + 2b}{2\sigma^2} \right),\end{aligned}$$

donde  $M$  es la media inicial de los coeficientes,  $\sigma^2 V$  su varianza, y  $a$  y  $b$  son los parámetros iniciales de forma y escala de  $\sigma^2$ .

Aprovechando la propiedad conjugada, es posible escribir la probabilidad posterior de los parámetros como:

$$\begin{aligned}\mathbb{P}(\beta, \sigma^2 | Y, X) &\propto \mathbb{P}(Y|X, \beta, \sigma^2) \times \mathbb{P}(\beta, \sigma^2), \\ &\propto (\sigma^2)^{-(\bar{a}+(n/2)+1)} \exp \left( -\frac{(\beta - \bar{M})^T \bar{V}^{-1} (\beta - \bar{M}) + 2\bar{b}}{2\sigma^2} \right),\end{aligned}$$

donde

$$\begin{aligned}\bar{M} &= (V^{-1} + X^T X)^{-1}(V^{-1}M + X^T Y), \\ \bar{V} &= (V^{-1} + X^T X)^{-1}, \\ \bar{a} &= a + n/2, \\ \bar{b} &= b + \frac{\bar{M}^T V^{-1}M + Y^T Y - \bar{M}^T \bar{V}^{-1} \bar{M}}{2}.\end{aligned}$$

Es decir, la distribución posterior de  $(\beta, \sigma^2)$  es *Normal - Gamma Inversa*, con parámetros  $\mathcal{NGI}(\bar{M}, \bar{V}, \bar{a}, \bar{b})$ .

Si se tiene una nueva matriz de covariables  $X_*$  y se desea hacer predicción de las respectivas variables de salida  $Y_*$ , es posible hacer inferencia con los datos observados de la siguiente manera:

$$\begin{aligned}\mathbb{P}(Y_*|X_*, Y, X) &= \int \int \mathbb{P}(Y_*|X_*, \beta, \sigma^2) \times \mathbb{P}(\beta, \sigma^2|Y, X) d\sigma^2 d\beta \\ &= \int \int \mathcal{N}(Y_*|X_*\beta, \sigma^2 I) \times \mathbb{P}(\beta, \sigma^2|Y, X) d\sigma^2 d\beta.\end{aligned}$$

Particularmente, si se continúa con el modelo conjugado *Normal - Gamma Inversa / Normal*, es posible encontrar la solución analítica:

$$\begin{aligned}\mathbb{P}(Y_*|X_*, Y, X) &= \int \int \mathcal{N}(Y_*|X_*\beta, \sigma^2 I) \times \mathcal{NGI}(\beta, \sigma^2|\bar{M}, \bar{V}, \bar{a}, \bar{b}) d\sigma^2 d\beta \\ &= MVSt_{2\bar{a}}\left(X_*\bar{M}, \frac{\bar{b}}{\bar{a}}\left(I + X_*\bar{V}X_*^T\right)\right),\end{aligned}$$

donde *MVSt* es la distribución *t-Student* multivariada, y cuya definición se describe a continuación.

**Definición.** Sea  $X \in \mathbb{R}^p$  un vector aleatorio, con media, mediana y moda  $\mu$ , matriz de covarianzas  $\Sigma$ , y  $\nu$  grados de libertad, entonces  $X \sim MVSt_\nu(\mu, \Sigma)$  si y sólo si su función de densidad es:

$$w(x|\mu, \sigma, \nu) = \frac{\Gamma((\nu + p)/2)}{\Gamma(\nu/2)\nu^{p/2}\pi^{p/2}|\Sigma|^{1/2}} \left[1 + \frac{1}{\nu}(x - \mu)^T \Sigma^{-1}(x - \mu)\right]^{-\frac{\nu+p}{2}}.$$

### 3.3. Regresión sobre cuantiles

La *regresión sobre cuantiles* es una alternativa que se ha desarrollado recientemente y que permite enfocarse en aspectos alternativos de la distribución, como lo que pasa en las colas. Además relaja supuestos de la *regresión a la media*, como la simetría inducida por el error normal.

**Definición 1.** *El cuantil  $p$ -ésimo de la variable aleatoria  $Y$  es aquel valor  $q_p$  tal que*

$$F_Y(q_p) = p.$$

*Equivalentemente, la función que regresa el cuantil  $p$ -ésimo de la variable aleatoria  $Y$  se escribe*

$$q_p(Y) = F_Y^{-1}(p),$$

*cuando  $F_Y^{-1}$  está bien definida.*

Dicho en otras palabras, si se tiene un conjunto grande de realizaciones de una variable aleatoria  $Y$ , se esperará que el  $p \times 100\%$  esté por debajo de  $q_p(Y)$  y el  $(1 - p) \times 100\%$  esté por arriba. Por ejemplo, la mediana es un caso particular de un cuantil, específicamente el 0.5-ésimo.

En notación probabilística, busca aproximar a la función  $f$ , tal que

$$q_p(y|x) = f(x),$$

para  $p \in (0, 1)$  fijo arbitrario.

Para hacer esto, normalmente se vale del supuesto que

$$y = f_p(x) + \varepsilon_p,$$

con  $\varepsilon_p \in \mathbb{R} \sim E_p(\theta)$ , de manera que  $E_p$  es una variable aleatoria con

vector de parámetros  $\theta$ , tal que  $q_p(\varepsilon_p) = 0$ .

Es importante aclarar que  $f_p(x) \in \mathbb{R}$  y  $\theta$  son desconocidos. Asimismo, al igual que con la *regresión a la media*, se supone independencia entre los errores aleatorios, y por lo tanto, hay independencia condicional entre las observaciones.

### 3.3.1. Modelo tradicional

Cuando surgió entre la comunidad estadística el problema de *regresión sobre cuantiles*, inicialmente fue modelado bajo un enfoque no bayesiano, como se describe en Yu & Moyeed (2001). Posteriormente, Koenker & Bassett (1978) retomaron esas ideas, y las aplicaron en el paradigma bayesiano.

Al igual que en la *regresión a la media*, el primer y más popular modelo ha sido el lineal. Es decir, para  $p \in (0, 1)$  fijo arbitrario, se define

$$f_p(x) = x^T \beta_p,$$

donde  $\beta_p$  es el vector de coeficientes, dependiente de  $p$ .

**Definición 2.** *Se define a la función*

$$\rho_p(u) = u \times [pI_{(u>0)} - (1-p)I_{(u<0)}].$$

*Se dice que una variable aleatoria  $U$  sigue una distribución asimétrica de Laplace ( $U \sim AL_p(\sigma)$ ) si su función de densidad se escribe como*

$$w_p^{AL}(u|\sigma) = \frac{p(1-p)}{\sigma} \exp \left[ -\rho_p \left( \frac{u}{\sigma} \right) \right],$$

*con  $\sigma$  parámetro de escala.*

Es posible darse cuenta que si  $\varepsilon_p \sim AL_p(\sigma)$ , entonces  $q_p(\varepsilon_p) = 0$ . Recordando que esta es la única característica necesaria para la distri-

bución del error aleatoria, entonces se definirá

$$\varepsilon \sim AL_p(\sigma).$$

El modelo se puede reescribir como:

$$y|x, \beta_p, \sigma \sim AL_p(y - x^T \beta_p | \sigma).$$

Sea  $\{(X, Y) | X \in \mathbb{R}^{m \times n}, Y \in \mathbb{R}^m\}$  el conjunto de datos observados. Por el Teorema de Bayes,

$$\mathbb{P}(\beta_p, \sigma | Y, X) \propto \mathbb{P}(Y | X, \beta_p, \sigma) \times \mathbb{P}(\beta_p, \sigma),$$

donde  $\mathbb{P}(Y | X, \beta_p, \sigma)$  es la verosimilitud de los datos observados y se puede calcular como

$$\mathbb{P}(Y | X, \beta_p, \sigma) = \prod_{i=1}^m AL_p(y_i - x_i^T \beta_p | \sigma).$$

Por otro lado,  $\mathbb{P}(\beta_p, \sigma^2)$  es la distribución inicial de los parámetros, para los que normalmente se usa

$$\beta_p, \sigma \sim \mathcal{NGI}(M, V, a, b).$$

A diferencia del modelo tradicional de la *regresión a la media*, este modelo no es conjugado. Por lo tanto se requieren métodos computacionales (como los que serán descritos en el capítulo 5) para aproximar la distribución posterior.

En el caso de la predicción, si se tiene una nueva matriz de covariables  $X_* \in \mathbb{R}^{r \times n}$ , la inferencia con los datos observados se realiza de la

siguiente manera:

$$\begin{aligned}\mathbb{P}(Y_*|X_*, Y, X) &= \int \int \mathbb{P}(Y_*|X_*, \beta_p, \sigma) \times \mathbb{P}(\beta_p, \sigma|Y, X) d\sigma d\beta_p \\ &= \int \int \prod_{i=1}^r AL_p(y_i - x_i^T \beta_p | \sigma) \times \mathbb{P}(\beta_p, \sigma|Y, X) d\sigma d\beta_p,\end{aligned}$$

que tampoco tiene solución analítica.

Si bien este modelo representa un gran avance, aún queda la posibilidad de retomar estas ideas y crear modelos más precisos. La intención de esta tesis es encontrar un modelo para la *regresión sobre cuantiles* que sea completamente bayesiano y no paramétrico, con la intención de poder representar distribuciones más complejas.

## Capítulo 4

# Especificación no paramétrica

### 4.1. Motivación

En el capítulo anterior se analizaron para realizar regresión hacia una variable de respuesta  $y$ , dado un cierto conjunto de covariables  $x$ . Si bien son modelos con muchas ventajas, es relevante no olvidar que cuentan con un supuesto fuerte: la relación entre la variable dependiente  $y$  y las variables independientes  $x$  únicamente se da de forma lineal. Pero las funciones lineales sólo son un pequeño subconjunto del conjunto infinito no-numerable de funciones existentes. Por ello, valdría la pena analizar si es posible relajar este supuesto y tener un modelo más general.

Una idea inicial para darle la vuelta es redefinir variables, de tal manera que se pueda obtener un polinomio. Por ejemplo, se supone que  $\hat{x}$  es un buen predictor de  $y$ , pero como polinomio de orden 3, es decir:

$$y = \beta_0 + \beta_1 \hat{x} + \beta_2 \hat{x}^2 + \beta_3 \hat{x}^3 + \varepsilon.$$



Entonces, se puede definir el vector  $x$  de covariables como  $x = (1, \hat{x}, \hat{x}^2, \hat{x}^3)$  y aplicar las técnicas de regresión lineal ya mencionadas.

Otra crítica que se le podría hacer a este modelo es la rigidez en la interacción entre variables. Para ejemplificar esto, se podría pensar en un modelo de la forma:

$$y = \beta_0 + \beta_1 \hat{x}_1 + \beta_2 \hat{x}_2 + \beta_3 \hat{x}_1 \hat{x}_2 + \varepsilon.$$

Es posible entonces declarar el vector  $x$  de variables de entrada de la forma  $x = (1, \hat{x}_1, \hat{x}_2, \hat{x}_1 \hat{x}_2)$ , y el procedimiento sería análogo.

Y aún es posible dar un siguiente paso, saliendo del terreno de los polinomios y entrando en el de las funciones biyectivas. Se podría pensar en un caso como el siguiente (donde siempre se cumpla que  $\hat{y} > 1$ ):

$$\begin{aligned} \ln(\hat{y}) &= \hat{\beta}_0 \hat{x}_1^{\beta_1} \hat{x}_2^{\beta_2} e^{\varepsilon} \\ \implies \ln(\ln(\hat{y})) &= \ln(\hat{\beta}_0) + \beta_1 \ln(\hat{x}_1) + \beta_2 \ln(\hat{x}_2) + \varepsilon \\ \implies y &= \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \varepsilon, \end{aligned}$$

con

$$\begin{aligned} y &= \ln(\ln(\hat{y})), \\ \beta_0 &= \ln(\hat{\beta}_0), \\ x_1 &= \ln(\hat{x}_1), \\ x_2 &= \ln(\hat{x}_2), \end{aligned}$$

y el procedimiento se convierte en el ya conocido.

Si bien estos ejemplos permiten ampliar el conjunto de funciones que es posible cubrir usando el modelo tradicional de regresión lineal, permiten darse cuenta de cómo se puede complicar la relación de dependencia entre  $y$  y las covariables  $x$ , de tal manera que muchas funciones pueden no ser descritas con el método antes planteado.

Así surge la necesidad de buscar un método que permita encontrar

cualquier tipo relación entre  $y$  y  $x$ , sin restringirla a un pequeño subconjunto de funciones. El reto es que únicamente se tiene tiempo finito para encontrar la mejor estimación, entre una infinidad no-numerable de opciones.

Por otro lado, en cuanto al error aleatorio  $\varepsilon_p$ , la distribución asimétrica de Laplace cumple el cometido de que el cuantil  $p$ -ésimo sea igual a 0, es decir, implícitamente provoca la asimetría necesaria para que el valor esperado de los valores por debajo de  $f_p(x)$  sean el  $p \times 100\%$ , y por encima, el  $(1 - p) \times 100\%$ .

Si bien esta es una característica necesaria, puede no ser suficiente debido a que la forma de la distribución queda totalmente determinada por un único parámetro  $\sigma$ . Por lo que si bien puede coincidir con la distribución *real* en que el cuantil  $p$ -ésimo  $q_p$  sea el mismo para ambas e igual a 0, la forma de la distribución podría ser totalmente distinta. Por ejemplo, en el peso que le asigna a las colas.

Dicha problemática podría ser mitigada mediante el uso de una mezcla de distribuciones, particularmente asimétricas de Laplace, en la que se usen varios valores para  $\sigma$  y la probabilidad asociada a cada valor vaya de acuerdo a su factibilidad. Entonces surgen algunas preguntas como ¿cuántos valores de  $\sigma$  debería de contener el modelo y cuáles deberían ser esos valores? Normalmente no existe una respuesta definitiva a ambas preguntas y se deja la decisión arbitraria al modelador. Pero, ¿qué pasaría si se planteara un modelo de mezclas infinitas de distribuciones? Así, se podría encontrar la mezcla óptima, ya que cualquier mezcla con número fijo de parámetros sería un caso particular.

En resumen, tanto la estimación de  $f_p$ , como la de  $\varepsilon_p$  podrían mejorarse usando modelos de infinitos parámetros, que generalizan a los modelos con un número de parámetros predefinido. Con el paradigma estadístico tradicional es imposible hacerlo, y más si el tiempo es finito. Pero esto abre la puerta a una visión menos explorada para hacer

estadística: el **paradigma no paramétrico**.

Como menciona Wasserman (2006): *La idea básica de la inferencia no paramétrica es usar los datos para inferir una medida desconocida, haciendo los menos supuestos posibles. Normalmente esto significa usar modelos estadísticos de dimensión infinita. De hecho, un mejor nombre para la inferencia no paramétrica podría ser inferencia de dimensión infinita.*

Y si bien esto puede sonar irreal, la idea intuitiva que está detrás de este tipo de modelos es que el modelador no debería fijar el número de parámetros antes de analizar la información, sino que los datos deben ser los que indiquen cuántos y cuáles son los parámetros significativos.

## 4.2. En $f_p$ , vía Procesos Gaussianos <sup>1</sup>

### 4.2.1. Procesos Gaussianos

Retomando las ideas del capítulo anterior, los modelos de regresión tienen como objetivo describir la distribución de una variable aleatoria  $y$ , condicional a los valores de las covariables  $x$ , es decir  $y|x \sim \mathbb{P}(y|x)$ . Dado que es complicado aproximar con exactitud toda la distribución, comúnmente se enfocan en una medición particular representada por la función  $f_p$ , que en el caso de la *regresión sobre cuantiles* se define como  $q_p(y|x) = f_p(x)$ .

Con el objetivo de ajustar un modelo, se utiliza el supuesto que

$$y = f_p(x) + \varepsilon_p,$$

tal que  $q_p(\varepsilon_p) = 0$ .

En el modelo tradicional se utiliza el supuesto de relación lineal

---

<sup>1</sup>Las ideas de esta sección son inspiradas por Rasmussen & Williams (2006).

$f_p(x) = x^T \beta_p$ , mismo que se buscará relajar en esta sección, para obtener un modelo más general.

Es importante recordar que la función  $f_p$  es pensada constante, pero desconocida. De nueva cuenta, para reflejar la incertidumbre del modelador, es posible darle una distribución de probabilidad. Pero a diferencia del modelo lineal, ya no existirá el parámetro  $\beta_p$  al cual canalizarle esta incertidumbre, por lo que ahora tendrá que ser sobre toda la función.

Es de utilidad, entonces, pensar a  $f_p(x)$  como una variable aleatoria. Particularmente se le puede asignar una distribución *Normal*, donde la media  $m(x)$  y la covarianza  $k(x, x')$  reflejen el conocimiento previo que se tenga del fenómeno de estudio. Cabe resaltar que dicha media  $m(x)$  y covarianza  $k(x, x')$  están en función de  $x$ , es decir, podrían variar de acuerdo al valor de las covariables.

Para continuar con la notación matricial del capítulo anterior, sean  $Y \in \mathbb{R}^m$  y  $X \in \mathbb{R}^{m \times n}$ , y  $\mathcal{E}_p \in \mathbb{R}^m$  el vector de errores aleatorios, es posible describir al modelo como

$$Y = f_p(X) + \mathcal{E}_p$$

donde

$$f_p(X) = \begin{bmatrix} f_p(x_1) \\ \dots \\ f_p(x_m) \end{bmatrix}, x_i \in \mathbb{R}^n, \forall i \in \{1, \dots, m\}.$$

Por lo tanto, bajo el supuesto de que cada  $f_p(x_i)$  es una variable aleatoria,  $f_p(X) \in \mathbb{R}^n$  es un vector aleatorio. Además, depende de variables de entrada, por lo que  **$f_p(X)$  es un proceso estocástico**. Asimismo, debido a que cada  $f_p(x_i)$  tiene una distribución *Normal univariada*, dándole una estructura de covarianza,  $f_p(X)$  se distribuirá *Normal Multivariada*, donde el vector de medias  $M_{f_p}(X)$  y la matriz de covarianzas  $K_{f_p}(X, X)$  reflejarán el conocimiento inicial del modelador.

**Definición 3.** Un *proceso gaussiano* ( $Y \in \mathbb{R}^m$ ), es una colección finita de  $m$ -variables aleatorias que tienen una distribución gaussiana (normal) conjunta.

**Observación.** De acuerdo a la construcción del vector  $f_p(X) \in \mathbb{R}^m$ , y tomando en cuenta la Definición 3, además de ser un proceso estocástico,  $f_p(X)$  es un *proceso gaussiano*.

#### 4.2.2. Definiciones y notación

Para las siguientes definiciones se supondrá que  $f_p(x)$  es una variable aleatoria y  $f_p(X)$  un vector aleatorio, con medias y covarianzas conocidas y finitas.

**Definición.** Sean  $x, x' \in \mathbb{R}^n$ .

La *función de medias de  $f_p$*  ( $m_{f_p}$ ) se define como

$$m_{f_p} : \mathbb{R}^n \rightarrow \mathbb{R} \mid m_{f_p}(x) = \mathbb{E}[f_p(x)].$$

La *función de covarianzas de  $f_p$*  ( $k_{f_p}$ ) se define como

$$k_{f_p} : \mathbb{R}^n \times \mathbb{R}^n \rightarrow \mathbb{R} \mid k_{f_p}(x, x') = \text{Cov}(f_p(x), f_p(x')).$$

**Definición.** Sea  $X \in \mathbb{R}^m \times \mathbb{R}^n$  y  $X' \in \mathbb{R}^r \times \mathbb{R}^n$ , es decir,

$$X = \begin{bmatrix} x_1 \\ \dots \\ x_m \end{bmatrix},$$

$$X' = \begin{bmatrix} x_1 \\ \dots \\ x_r \end{bmatrix}.$$

La **función vector de medias de  $f_p$**  ( $M_{f_p}$ ) se define como

$$M_{f_p} : \mathbb{R}^m \times \mathbb{R}^n : \mathbb{R}^m \mid M_{f_p}(X) = \begin{bmatrix} m_{f_p}(x_1) \\ \dots \\ m_{f_p}(x_m) \end{bmatrix}.$$

La **función matriz de covarianzas de  $f_p$**  ( $K_{f_p}$ ) se define como

$$K_{f_p} : \mathbb{R}^m \times \mathbb{R}^n : \mathbb{R}^m \times \mathbb{R}^m \mid K_{f_p}(X, X') = \begin{bmatrix} k_{f_p}(x_1, x'_1) & \dots & k_{f_p}(x_1, x'_r) \\ \dots & \dots & \dots \\ k_{f_p}(x_m, x'_1) & \dots & k_{f_p}(x_m, x'_r) \end{bmatrix}.$$

Dadas estas definiciones, se puede observar que el *proceso gaussiano*  $f_p(X) \in \mathbb{R}^m$  está completamente caracterizado por su función de medias  $m_{f_p}$  y su función de covarianzas  $k_{f_p}$ . Por lo tanto, la manera en que se definan estas dos funciones representará el conocimiento inicial que se tiene del objeto de estudio.

A partir de este punto, y cuando el contexto lo permita, por simplicidad de notación se omitirá el uso del subíndice  $f_p$  en las funciones recién definidas. Además, cuando se desee referirse al proceso estocástico  $f_p(X)$  que se distribuye como un *proceso gaussiano*, se hará con la siguiente notación:

$$f_p(X) \sim \mathcal{GP}(m_{f_p}, k_{f_p}).$$

### 4.2.3. Funciones de covarianza

Hasta el momento, no se han descrito las características de la función de covarianzas  $k$ . Cabe resaltar que  $k$  no es una *covarianza* en general, ni cumple con todas las propiedades, sino únicamente describe la covarianza entre dos vectores aleatorios  $f_p(x)$  y  $f_p(x')$ , con la misma  $f_p$ , sin la intervención, por ejemplo, de constantes. Para explicar de mejor

manera este punto, se da el siguiente ejemplo:

$$\begin{aligned} Cov(af_p(x) + f_p(x'), f_p(x')) &= Cov(af_p(x), f_p(x')) + Cov(f_p(x), f_p(x')) \\ &= a \times Cov(f_p(x), f_p(x')) + Cov(f_p(x'), f_p(x')) \\ &= a \times k(x, x') + k(x', x') \end{aligned}$$

En este orden de ideas, las propiedades que  $k(x, x')$  tiene que cumplir son

$$\begin{aligned} k(x, x') &= k(x', x) \text{ (simetría),} \\ k(x, x) &= Var(f_p(x)) \geq 0. \end{aligned}$$

Si bien es cierto que dadas esas restricciones hay una variedad muy grande de funciones con las que se puede describir  $k(x, x')$ , por practicidad, y tomando en cuenta que es un supuesto sensato para la mayoría de los casos, es común describir a la función  $k$  en relación a la distancia entre  $x$  y  $x'$ ,  $\|x, x'\|_\gamma$ . Es decir,  $k(x, x') = k(\|x, x'\|_\gamma)$ . A este tipo de funciones de covarianza se les denomina **estacionarias**.

Además, esta relación entre covarianza y distancia suele ser inversa, es decir, entre menor sea la distancia, mayor será la covarianza, y viceversa. De esta manera, para valores  $x \approx x'$ , se obtendrá que  $f_p(x) \approx f_p(x')$  en la mayoría de los casos, lo que tiene el supuesto implícito de que  $f_p$  es una función continua.

Un ejemplo de este tipo de funciones son las  **$\gamma$ -exponencial**, mismas que se definen de la siguiente manera:

$$k(x, x') = k(\|x, x'\|_\gamma; \gamma, \lambda, \tau) = \lambda \times \exp\left(-\tau\|x, x'\|_\gamma\right),$$

donde  $\lambda$  es un parámetro de escala y  $\tau$  de rango.

Las de uso más común suelen ser la 1 y 2-*exponencial*. Ambas tienen la ventaja de ser continuas, pero la 2-*exponencial* tiene además la peculiaridad de ser infinitamente diferenciable y, por lo tanto, es suave.

El siguiente ejemplo de funciones estacionarias es la **clase de Matérn**, descrita como

$$\begin{aligned} k(x, x') &= k(\|x, x'\|_1; \nu, \lambda, \tau) \\ &= \lambda \times \frac{2^{1-\nu}}{\Gamma(\nu)} \left( \tau \sqrt{2\nu} \|x, x'\|_1 \right)^\nu \times \left[ K_\nu \left( \tau \sqrt{2\nu} \|x, x'\|_1 \right) \right]^\nu, \end{aligned}$$

donde  $K_\nu$  es la función modificada de Bessel y  $\Gamma(\cdot)$  es la función *gamma*. Los casos más utilizados son

$$\begin{aligned} k \left( \|x, x'\|_1; \nu = \frac{3}{2}, \lambda, \tau \right) &= \lambda \times \left( 1 + \tau \sqrt{3} \|x, x'\|_1 \right) \times \exp \left( -\tau \sqrt{3} \|x, x'\|_1 \right), \\ k \left( \|x, x'\|_1; \nu = \frac{5}{2}, \lambda, \tau \right) &= \lambda \times \left( 1 + \tau \sqrt{3} \|x, x'\|_1 + \tau^2 \frac{5}{3} \|x, x'\|_1^2 \right) \\ &\quad \times \exp \left( -\tau \sqrt{5} \|x, x'\|_1 \right). \end{aligned}$$

Otra posible función de covarianza es la **racional cudrática**, caracterizada como

$$k(x, x') = k(\|x, x'\|_2; \alpha, \lambda, \tau) = \lambda \times \left( 1 + \tau \frac{\|x, x'\|_2^2}{2\alpha} \right)^{-\alpha},$$

con  $\alpha, \lambda, \tau > 0$ .

Existen otro tipo de funciones estacionarias que no guardan una relación inversa entre distancia y covarianza, sino que capturan un componente **estacional**, normalmente usado en series de tiempo. De esta manera, y siendo  $t$  la covariable del tiempo, es posible pensar en una función de la forma

$$k(x, x', t, t'; E, \lambda) = \bar{k}(x, x') + \lambda \times \delta_{\{(|t' - t| \bmod E) = 0\}},$$

donde  $\bar{k}$  es alguna de las funciones estacionarias antes mencionadas,  $\delta$  es la *delta de Kroenecker* y  $E$  es el periodo de estacionalidad. Por ejemplo,



$E = 12$  para una serie mensual.

Si se desea suavizar esta componente de estacionalidad para que no sea únicamente puntual, es posible describir la covarianza con una función como la siguiente:

$$k(x, x', t, t'; E, \lambda, \tau) = \bar{k}(x, x') + \lambda \times \exp\left(-\tau \frac{E}{\pi} \sin^2\left(\frac{\pi}{E}|t' - t|\right)\right).$$

#### 4.2.4. Predicción

Para esta subsección se supondrá que se cuenta con datos de  $f_p(X)$ , mismos que en la práctica son imposibles de observar directamente y únicamente se pueden aproximar con el modelo descrito anteriormente. La intención de este supuesto es sentar las bases teóricas para realizar predicción con el modelo central de esta tesis (GPDP), tema que será explorado con más detalle en el siguiente capítulo.

Sea un conjunto de observaciones  $\{(x_i, f_p(x_i)) | i = 1, \dots, m\}$ . De forma matricial, se puede escribir como  $\{(X, f_p(X))\}$ , con  $X \in \mathbb{R}^{m \times n}$  y  $f_p(X) \in \mathbb{R}^m$ . Por otro lado, se tiene un conjunto de covariables  $X_* \in \mathbb{R}^{r \times n}$ , y se desea predecir  $f_p(X_*) \in \mathbb{R}^r$ , suponiendo que sigue la misma función  $f_p$  de los datos observados.

La distribución inicial conjunta de los datos de entrenamiento  $f_p(X)$  y los datos a predecir  $f_p(X_*)$  es:

$$\begin{bmatrix} f_p(X) \\ f_p(X_*) \end{bmatrix} \sim \mathcal{N}\left(\begin{bmatrix} M(X) \\ M(X_*) \end{bmatrix}, \begin{bmatrix} K(X, X) & K(X, X_*) \\ K(X_*, X) & K(X_*, X_*) \end{bmatrix}\right)$$

Es momento oportuno para recordar algunas propiedades de la distribución *Normal condicional*.

**Propiedad 1.** *Sea  $X \in \mathbb{R}^m$  un vector aleatorio que tiene distribución Normal conjunta y está particionado de la siguiente manera:*

$$X = \begin{bmatrix} X_1 \\ X_2 \end{bmatrix}, \text{ con dimensiones } \begin{bmatrix} (m-q) \\ q \end{bmatrix},$$

Entonces, la media  $\mu \in \mathbb{R}^m$  y varianza  $\Sigma \in \mathbb{R}^{m \times m}$  de  $X$  se pueden escribir

$$\mu = \begin{bmatrix} \mu_1 \\ \mu_2 \end{bmatrix}, \text{ dimensiones } \begin{bmatrix} (m-q) \\ q \end{bmatrix},$$

$$\Sigma = \begin{bmatrix} \Sigma_{11} & \Sigma_{12} \\ \Sigma_{21} & \Sigma_{22} \end{bmatrix}, \text{ dimensiones } \begin{bmatrix} (m-q) \times (m-q) & (m-q) \times q \\ q \times (m-q) & q \times q \end{bmatrix}.$$

La distribución condicional de  $X_2$ , sujeta a que  $X_1 = a$  es Normal con  $X_2|X_1 = a \sim \mathcal{N}(X_2|\bar{\mu}, \bar{\Sigma})$ , donde

$$\bar{\mu} = \mu_2 + \Sigma_{2,1}\Sigma_{11}^{-1}(a - \mu_1)$$

$$\bar{\Sigma} = \Sigma_{22} - \Sigma_{21}\Sigma_{11}^{-1}\Sigma_{12}$$

De regreso al modelo, bajo el supuesto que ya se conocen los valores de  $f_p(X)$ , es posible condicionar la distribución conjunta, dadas esas observaciones. Utilizando las propiedades de la distribución Normal condicional, se obtiene que:

$$f_p(X_*)|f_p(X) \sim \mathcal{N}(\bar{M}(X, X_*), \bar{K}(X, X_*)),$$

con

$$\bar{M}(X, X_*) = M(X_*) + K(X_*, X)K(X, X)^{-1}(f_p(X) - M(X)),$$

$$\bar{K}(X, X_*) = K(X_*, X_*) - K(X_*, X)K(X, X)^{-1}K(X, X_*).$$

**Observación.**  $f(X_*)|f(X)$  es una colección finita de  $r$ -variables aleatorias que tienen una distribución Normal multivariada conjunta, por lo tanto,  $\mathbf{f}(X_*)|\mathbf{f}(X)$  es un proceso gaussiano.

### 4.3. En $\varepsilon_p$ , vía Procesos de Dirichlet <sup>2</sup>

Un Proceso de Dirichlet, visto de manera general, es una distribución sobre distribuciones. Es decir, cada realización de él es en sí misma una distribución de probabilidad. Además, cada una de esas distribuciones será no paramétrica, debido a que no será posible describirla con un número finito de parámetros.

En el caso particular de esta tesis y de su misión de encontrar un modelo bayesiano y no paramétrico para la *regresión sobre cuantiles*, los Procesos de Dirichlet serán utilizados para ajustar la distribución del error aleatorio  $\varepsilon_p$ .

#### 4.3.1. Procesos de Dirichlet

Antes de revisar la definición formal de los Procesos de Dirichlet, es conveniente recordar la definición de la distribución de Dirichlet.

**Definición 4.** *Se dice que un vector aleatorio  $x \in \mathbb{R}^n$  se distribuye de acuerdo a la **distribución de Dirichlet** ( $\mathbf{x} \sim \mathbf{Dir}(\alpha)$ ) con vector de parámetros  $\alpha$ , específicamente,*

$$x = \begin{pmatrix} x_1 \\ \cdots \\ x_n \end{pmatrix}, \quad \alpha = \begin{pmatrix} \alpha_1 \\ \cdots \\ \alpha_n \end{pmatrix},$$

*para los cuales se cumplen las restricciones*

$$x_i > 0, \forall i \in \{1, \dots, n\}$$

$$\sum_{i=1}^n x_i = 1$$

$$\alpha_i > 0, \forall i \in \{1, \dots, n\},$$

---

<sup>2</sup>Las ideas de esta sección son retomadas de Teh (2010).

si su función de densidad es

$$f(x|\alpha) = \frac{1}{B(\alpha)} \prod_{i=1}^n x_i^{\alpha_i-1},$$

donde  $B$  es la función Beta multivariada, y puede ser expresada en términos de la función  $\Gamma$  como

$$B(\alpha) = \frac{\prod_{i=1}^n \Gamma(\alpha_i)}{\Gamma(\sum_{i=1}^n \alpha_i)}, \quad \alpha = (\alpha_1, \dots, \alpha_n).$$

La esperanza y varianza de cada  $x_i$  son los siguientes:

$$\begin{aligned} \mathbb{E}[x_i] &= \frac{\alpha_i}{\sum_{k=1}^n \alpha_k} \\ \text{Var}(x_i) &= \frac{\alpha_i (\sum_{k=1}^n \alpha_k - \alpha_i)}{(\sum_{k=1}^n \alpha_k)^2 ((\sum_{k=1}^n \alpha_k + 1))} \end{aligned}$$

Es común que esta distribución sea usada como la inicial conjugada de la distribución multinomial, debido a que el vector  $x$  tiene las mismas propiedades de una distribución de probabilidad discreta (elementos positivos y que en conjunto suman 1).

Retomando el tema central, en términos generales, para que una distribución de probabilidad  $G$  se distribuya de acuerdo a un Proceso de Dirichlet, sus distribuciones marginales tienen que tener una distribución Dirichlet. A continuación se enuncia una definición más detallada.

**Definición 5.** Sean  $G$  y  $H$  dos distribuciones cuyo soporte es el conjunto  $\Theta$  y sea  $\alpha \in \mathbb{R}^+$ . Entonces, si se toma una partición finita cualquiera  $A_1, \dots, A_r$  del conjunto  $\Theta$ , el vector  $(G(A_1), \dots, G(A_r))$  es aleatorio, porque  $G$  también lo es.

Se dice que  $G$  se distribuye de acuerdo a un **Proceso de Dirichlet** ( $G \sim DP(\alpha, H)$ ), con distribución media  $H$  y parámetro de concen-

tración  $\alpha$ , si

$$(G(A_1), \dots, G(A_r)) \sim \text{Dir}(\alpha H(A_1), \dots, \alpha H(A_r)),$$

para cualquier partición finita  $A_1, \dots, A_r$  del conjunto  $\Theta$ .

Es momento de analizar el papel que juegan los parámetros. Sea  $A_i \subset \Theta$ , uno de los elementos de la partición anterior, y recordando las propiedades de la distribución de Dirichlet, entonces

$$\begin{aligned} E[G(A_i)] &= \frac{\alpha H(A_i)}{\sum_{k=1}^p \alpha H(A_k)} \\ &= H(A_i) \\ \text{Var}(G(A_i)) &= \frac{\alpha H(A_i) (\sum_{k=1}^p (\alpha H(A_k)) - \alpha H(A_i))}{(\sum_{k=1}^p \alpha H(A_k))^2 (\sum_{k=1}^p (\alpha H(A_k)) + 1)} \\ &= \frac{\alpha^2 [H(A_i)(1 - H(A_i))]}{\alpha^2 (1)^2 (\alpha + 1)} \\ &= \frac{H(A_i)(1 - H(A_i))}{\alpha + 1}. \end{aligned}$$

En este orden de ideas, es posible darse cuenta que la distribución  $H$  representa la *distribución media* del Proceso de Dirichlet. Por otro lado, el parámetro  $\alpha$  tiene una relación inversa con la varianza. Así, a una mayor  $\alpha$ , corresponde una menor varianza del Proceso de Dirichlet, y, por lo tanto, una mayor concentración respecto a la distribución media  $H$ .

Siguiendo la secuencia lógica, si  $\alpha \rightarrow \infty$ , entonces  $G(A_i) \rightarrow H(A_i)$  para cualquier elemento  $A_i$  de la partición. Es decir,  $G \rightarrow H$  en distribución. Sin embargo, cabe aclarar que esto no es lo mismo que  $G \rightarrow H$ . Por un lado,  $H$  puede ser una distribución de probabilidad continua, mientras que, como se verá más adelante,  $G$  puede arrojar dos muestras iguales con probabilidad mayor a 0, por lo que es una distribución

discreta.

### 4.3.2. Distribución posterior

Sea  $G \sim DP(\alpha, H)$ . Dado que  $G$  es (aunque aleatoria) una distribución, es posible obtener realizaciones de ella. Sean  $\phi_1, \dots, \phi_n$  una secuencia de realizaciones independientes de  $G$ , que toman valores dentro de su soporte  $\Theta$ . Sea de nuevo  $A_1, \dots, A_r$  una partición finita cualquiera del conjunto  $\Theta$ , y sea  $n_k = |\{i : \phi_i \in A_k\}|$  el número de valores observados dentro del conjunto  $A_k$ . Por la propiedad conjugada entre la distribución de *Dirichlet* y la distribución *Multinomial*, se obtiene que

$$(G(A_1), \dots, G(A_r)) | \phi_1, \dots, \phi_n \sim Dir(\alpha H(A_1) + n_1, \dots, \alpha H(A_r) + n_r).$$

Es posible reescribir  $n_k = \sum_{i=1}^n \delta_i(A_k)$ , donde  $\delta_i(A_k) = 1$  si  $\phi_i \in A_k$ , y 0 en cualquier otro caso. Así,

$$\begin{aligned} \alpha H(A_k) + n_k &= \alpha H(A_k) + \sum_{i=1}^n \delta_i(A_k) \\ &= (\alpha + n) \left[ \frac{\alpha \times H(A_k) + n \times \frac{\sum_{i=1}^n \delta_i(A_k)}{n}}{\alpha + n} \right] \\ &= \bar{\alpha} \bar{H}(A_k), \end{aligned}$$

con

$$\begin{aligned} \bar{\alpha} &= \alpha + n \\ \bar{H}(A_k) &= \left( \frac{\alpha}{\alpha + n} \right) H(A_k) + \left( \frac{n}{\alpha + n} \right) \frac{\sum_{i=1}^n \delta_i(A_k)}{n}. \end{aligned}$$

Por lo tanto,  $G | \phi_1, \dots, \phi_n \sim DP(\bar{\alpha}, \bar{H})$ . Es decir, la probabilidad posterior de  $G$  sigue distribuyéndose mediante un Proceso de Dirichlet, con parámetros actualizados. Asimismo, se puede interpretar a la distribución media posterior  $\bar{H}$  como una mezcla entre la distribución media

inicial, con peso proporcional al parámetro de concentración inicial  $\alpha$ , y la distribución empírica de los datos, con peso proporcional al número de observaciones  $n$ .

### 4.3.3. Distribución predictiva

Continuando con la idea de la sección anterior de que ya se conoce el valor de  $\phi_1, \dots, \phi_n$  realizaciones provenientes de la distribución aleatoria  $G$ , se desea hacer predicción de la observación  $\phi_{n+1}$ , condicionada a los valores observados. Así,

$$\begin{aligned} P(\phi_{n+1} \in A_k | \phi_1, \dots, \phi_n) &= \int P(\phi_{n+1} \in A_k | G) P(G | \phi_1, \dots, \phi_n) dG \\ &= \int G(A_k) P(G | \phi_1, \dots, \phi_n) dG \\ &= \mathbb{E}[G(A_k) | \phi_1, \dots, \phi_n] \\ &= \bar{H}(A_k), \end{aligned}$$

es decir,

$$\phi_{n+1} | \phi_1, \dots, \phi_n \sim \left( \frac{\alpha}{\alpha + n} \right) H(\phi_{n+1}) + \left( \frac{n}{\alpha + n} \right) \frac{\sum_{i=1}^n \delta_i(\phi_{n+1})}{n}.$$

Cabe resaltar que dicha distribución predictiva tiene puntos de masa localizados en  $\phi_1, \dots, \phi_n$ . Esto significa que la probabilidad de que  $\phi_{n+1}$  tome un valor que ya ha sido observado es mayor a 0, independientemente de la forma de  $H$ . Yendo aún más allá, es posible darse cuenta que si se obtienen realizaciones infinitas de  $G$ , cualquier valor obtenido será repetido eventualmente, con probabilidad igual a 1. Por lo tanto,  $G$  es una distribución discreta.

#### 4.3.4. Esquema de urna de Blackwell-MacQueen

En esta sección se revisará la equivalencia entre los Procesos de Dirichlet y otro problema famoso en la literatura, lo que ayudará con la intuición que hay detrás de los Procesos de Dirichlet, así como la resolución y demostración de algunas propiedades pendientes.

Sea  $\Theta$  un conjunto (finito o infinito) cuyos elementos son colores distintos al negro y al blanco, y donde cada color es distinto entre sí. Existe una máquina llamada  $H$  que cada que se le oprime *Play* arroja de manera aleatoria una pelota con algún color perteneciente al conjunto  $\Theta$ , siguiendo una regla de probabilidad dada previamente. Se tienen 2 urnas: una llamada *probabilidades*, que contiene  $\alpha$  bolas negras. Otra llamada *resultados*, que en un principio se encuentra vacía.

Se oprime *Play* a la máquina, y se obtiene una pelota, la cual se arroja a la urna *resultados*. A  $\phi_1$  se le aginará el color de dicha pelota. Posteriormente se añade una pelota de color blanca a la urna *probabilidades* y se pasa a la segunda ronda.

Las siguientes rondas, por ejemplo la ronda  $n+1$ , comienza tomando al azar una pelota de la urna *probabilidades*. Si el color de la pelota es negra (probabilidad proporcional a  $\alpha$ ), se obtiene una nueva pelota de la máquina  $H$  y se repite lo sucedido en la primera ronda, incluyendo el asignar el color de la pelota a  $\phi_{n+1}$ . Si es blanca, se toma al azar una pelota de la urna *resultados*, se asigna el color de esa pelota a  $\phi_{n+1}$  y se regresa a la urna de *resultados* esa misma pelota, así como una nueva pintada del mismo color. En ambos casos, después de hacer lo antes mencionado, se introduce una nueva pelota blanca a la urna *probabilidades* y se pasa a la siguiente ronda.

Así, después de  $n$  rondas, se obtiene la secuencia  $\phi_1, \dots, \phi_n$ . Es importante notar que cada  $\phi_{k+1}$  es una variable aleatoria que depende de



las  $k$  anteriores, y cuya distribución es

$$\phi_{k+1}|\phi_1, \dots, \phi_k \sim \left( \frac{\alpha}{\alpha + k} \right) H(\phi_{k+1}) + \left( \frac{k}{\alpha + k} \right) \frac{\sum_{i=1}^k \delta_i(\phi_{k+1})}{k}.$$

La distribución conjunta de  $\phi_1, \dots, \phi_n$  se puede obtener como

$$P(\phi_1, \dots, \phi_n) = P(\phi_1) \prod_{i=2}^n P(\phi_i|\phi_1, \dots, \phi_{i-1}).$$

**Definición.** Sea  $\phi_1, \dots, \phi_n$ , una secuencia de  $n$  variables aleatorias, cuya distribución de probabilidad conjunta está dada por  $P(\phi_1, \dots, \phi_n)$ . Sea  $\psi$  una función biyectiva, que va de  $\{1, \dots, n\} \rightarrow \{1, \dots, n\}$ , es decir, una función que crea una permutación del conjunto  $\{1, \dots, n\}$ . Entonces, se dice que  $\phi_1, \dots, \phi_n$  es una **secuencia aleatoria infinitamente intercambiable** si se cumple que

$$P(\phi_1, \dots, \phi_n) = P(\phi_{\psi(1)}, \dots, \phi_{\psi(n)}),$$

para cualquier permutación  $\psi$ .

Regresando al juego de urnas, es importante observar que si bien  $\phi_{k+1}$  es dependiente de las  $k$  observaciones anteriores, esta dependencia sólo se da en términos de los valores observados previamente y la frecuencia de dichas observaciones, pero el orden en que hayan sido obtenidos no es relevante. Por lo tanto, es posible afirmar que  $\phi_1, \dots, \phi_n$  es una secuencia aleatoria infinitamente intercambiable. Dicho esto, es conveniente recordar el **Teorema de representación general de de Finetti**.<sup>3</sup>

**Teorema.** Sea  $\phi_1, \dots, \phi_n$  una secuencia aleatoria infinitamente intercambiable de valores reales. Entonces existe una distribución de proba-

---

<sup>3</sup>Una demostración de este teorema puede ser encontrada en Schervish (1996).

bilidad  $G$  sobre  $\mathcal{F}$ , el espacio de todas las distribuciones, de forma que la probabilidad conjunta de  $\phi_1, \dots, \phi_n$  se puede expresar como

$$P(\phi_1, \dots, \phi_n) = \int_{\mathcal{F}} \left[ \prod_{k=1}^n G(\phi_k) \right] dP(G),$$

con

$$P(G) = \lim_{n \rightarrow \infty} P(G_n),$$

donde  $P(G_n)$  es una función de distribución evaluada en la función de distribución empírica definida por

$$G_n = \frac{1}{n} \sum_{i=1}^n I(y_i \leq y).$$

En otras palabras, el Teorema de de Finetti dice que existe una distribución  $G$  tal que  $\phi_1, \dots, \phi_n$  son condicionalmente independientes, dada dicha  $G$ . A su vez dicha  $G$  es aleatoria y sigue una distribución  $P(G)$ .

Una vez dicho esto, y sean  $\phi_1, \dots, \phi_n$  una secuencia de colores obtenida con la rutina de esta sección, es posible darse cuenta que cada  $\phi_k \sim G$ . Además  $P(G) = DP(\alpha, H)$ , según lo visto en la sección anterior. Con esto, queda demostrada la existencia de los Procesos de Dirichlet.

#### 4.3.5. Proceso estocástico de rompimiento de un palo

Es importante recordar que una realización  $G$  de un Proceso de Dirichlet es discreta con probabilidad 1, debido a que toda muestra tiene probabilidad mayor a 0 de ser repetida. Por lo tanto, se puede expresar a  $G$  como una suma de centros de masa, de la siguiente manera:

$$G(\phi) = \sum_{k=1}^{\infty} \pi_k \delta_{\phi_k^*}(\phi),$$

$$\phi_k^* \sim H,$$

siendo  $\pi_k$  la probabilidad de ocurrencia de  $\phi_k$ .

Dicha probabilidad de ocurrencia será generada con la siguiente metáfora. Se piensa un palo de longitud 1. Se genera un número aleatorio  $\beta_1 \sim \text{Beta}(1, \alpha)$ , mismo que estará en el intervalo  $(0, 1)$ . Esa será la magnitud del pedazo que será separado del palo de longitud 1, y le será asignado a  $\pi_1 = \beta_1$ . Así, quedará un palo de magnitud  $(1 - \beta_1)$  a repartir. Posteriormente se vuelve a generar un número aleatorio  $\beta_2 \sim \text{Beta}(1, \alpha)$ , que representará la proporción del palo restante que le será asignada a  $\pi_2$ . Es decir,  $\pi_2 = \beta_2(1 - \beta_1)$ . En general, para  $k \geq 2$ ,

$$\begin{aligned}\beta_k &\sim \text{Beta}(1, \alpha), \\ \pi_k &= \beta_k \prod_{i=1}^{k-1} (1 - \beta_i).\end{aligned}$$

Dada su construcción, es inmediato darse cuenta que  $\sum_{k=1}^{\infty} \pi_k = 1$ . Algunas ocasiones se nombra a esta distribución  $\pi \sim \text{GEM}(\alpha)$ , en honor a Griffiths, Engen y McCloskey.

#### 4.3.6. Modelo general de mezclas infinitas de Dirichlet

Sean  $\{y_1, \dots, y_n\}$  un conjunto de observaciones con distribución  $F$ , condicionalmente independientes, y que se suponen vienen del *Modelo de mezclas de Dirichlet*:

$$\begin{aligned}y_i | \phi_i &\sim F(y_i | \phi_i), \\ \phi_i | G &\sim G(\phi_i), \\ G | \alpha, H &\sim DP(\alpha, H).\end{aligned}$$

Se dice que este es un *modelo de mezclas* debido a que existen  $y'_i$ 's que comparten un mismo valor para  $\phi_i$  (por la propiedad discreta de  $G$ ), y entonces estas  $y'_i$ 's pueden ser consideradas pertenecientes a una misma

subpoblación.

Es posible reescribir este modelo usando la equivalencia entre los Procesos de Dirichlet y el Proceso estocástico de rompimiento de un palo, visto anteriormente. Sea  $z_i$  el *cluster* al que pertenece  $y_i$  entre los  $\Phi_1^*, \Phi_2^*, \dots$  posibles, se tiene entonces que  $P(z_i = \Phi_k^*) = \pi_k$ . Y si  $\phi_k^*$  es el valor que comparten los miembros de  $\Phi_k^*$ , se usará la notación  $\phi_{z_i} = \phi_k^*$ , cuando  $z_i = \Phi_k^*$ . Por lo tanto, el modelo se puede ahora escribir como

$$\begin{aligned} y_i | z_i, \phi_k^* &\sim F(y_i | \phi_{z_i}), \\ z_i | \pi &\sim Mult(\pi), \\ \pi | \alpha &\sim GEM(\alpha), \\ \phi_k^* | H &\sim H. \end{aligned}$$

De esta manera, el Modelo de mezclas de Dirichlet es un modelo de mezclas infinitas, debido a que tiene un número infinito numerable de posibles *clusters*, pero donde intuitivamente la importancia realmente recae sólo en aquellos que tienen un peso  $\pi$  posterior mayor a cierto umbral, pero qson detectados hasta después de observar los datos; a diferencia de los modelos de mezclas finitas, que ya tienen un número de *clusters* definidos previamente.

#### 4.3.7. Modelo de mezclas infinitas de Dirichlet para la *regresión sobre cuantiles*

Aterrizando las ideas anteriores al caso particular de los modelos de *regresión sobre cuantiles*, se busca describir a  $\varepsilon_p$  como producto de una mezcla infinita de distribuciones *Asimétricas de Laplace*, de la manera siguiente. Sea  $w_p^{AL} | \sigma$  la función de densidad de la distribución *Asimétrica de Laplace*, condicional en el valor del parámetro  $\sigma$ . Sea  $h_p | G$  la función de densidad de  $\varepsilon_p$  condicional en una distribución  $G(\sigma)$ , realización de

un proceso de Dirichlet con parámetro de concentración  $\alpha$  y distribución media  $H$ . Se tiene entonces que

$$h_p(\varepsilon|G) = \int w_p^{AL}(\varepsilon|\sigma)dG(\sigma),$$

$$G \sim DP(\alpha, H).$$

Cabe resaltar que a pesar de la mezcla, se sigue cumpliendo la condición de que  $q_p(\varepsilon_p|G) = 0$ , para toda  $G$ .

Además, por construcción, esta formulación es equivalente al modelo de mezclas infinitas de Dirichlet (visto en la subsección anterior), por lo que se puede reescribir como

$$\varepsilon_{p_i}|z_i, \sigma_k^* \sim AL_p(\varepsilon_{p_i}|\sigma_{z_i}),$$

$$z_i|\pi \sim Mult(\pi),$$

$$\pi|\alpha \sim GEM(\alpha),$$

$$\sigma_k^*|H \sim H.$$

En este orden de ideas, la tarea del modelador únicamente consistirá en definir el valor del parámetro de concentración  $\alpha$ , así como a la distribución de  $H$  y sus respectivos parámetros, con la restricción de que su soporte deberá ser un subconjunto de  $\mathbb{R}^+$ . Por lo tanto, la distribución *Gamma* o la *Gamma-Inversa* se postulan como opciones convenientes.

## Capítulo 5

# Modelo GPDP para regresión sobre cuantiles

### 5.1. Definición

Después de analizar la introducción de componentes no paramétricos, tanto para la función  $f_p$ , como para el error  $\varepsilon_p$ , a continuación se enunciará el modelo central de esta tesis, con sus especificaciones correspondientes.

A partir de este punto, a dicho modelo se le denominará **Modelo GPDP** (por las siglas en inglés de Procesos Gaussianos y Procesos de Dirichlet).

Sea  $\{(y_i, x_i) | i = 1, \dots, m\}$  el conjunto de observaciones de la variable de respuesta y sus respectivas covariables, cuya relación se supone como

$$y = f_p(x) + \varepsilon_p,$$

donde  $f_p : \mathbb{R}^n \times \mathbb{R}$  es la función base y  $\varepsilon_p \in \mathbb{R}$  es el error aleatorio, ambos desconocidos.

Para reflejar la incertidumbre y el conocimiento previo del modelador, se supone a  $f_p(X) \sim \mathcal{GP}(m_{f_p}, k_{f_p})$ , con función de medias  $m_{f_p}$  dada por el modelador y función de covarianza  $k_{f_p}$  2-exponencial, con parámetro de rango fijo  $\tau = 1$ . Es decir,

$$k_{f_p}(x_i, x_j | \lambda, \tau = 1) = \lambda \times \exp\{-\|x_i - x_j\|_2\},$$

con  $\lambda \sim GI(c_\lambda, d_\lambda)$ ; siendo  $c_\lambda$  y  $d_\lambda$  los parámetros de forma y escala de una *Gamma-Inversa*.

La razón de fijar  $\tau = 1$  es para simplificar el proceso de inferencia que se verá en el siguiente capítulo, pero bien podría también ser una variable aleatoria.

En cuanto a la distribución inicial de  $\varepsilon_p$ , se supondrá un modelo de mezclas infinitas de Dirichlet, cuya distribución media  $H$  del proceso de Dirichlet será una *Gamma-Inversa*, con parámetros de forma  $c_{DP}$  y escala  $d_{DP}$ .

En resumen, el **Modelo GPDP** queda descrito de la siguiente forma:

$$\begin{aligned} y_i | f_p(x_i), z_i, \sigma_k^* &\sim w_p^{AL}(\varepsilon_{p_i} = y_i - f_p(x_i) | \sigma_{z_i}), \\ f_p(X) | m_{f_p}, \lambda &\sim \mathcal{GP}(m_{f_p}, k_{f_p} | \lambda), \\ \lambda &\sim GI(c_\lambda, d_\lambda), \\ z_i | \pi &\sim Mult(\pi), \\ \pi | \alpha &\sim GEM(\alpha), \\ \sigma_k^* | c_{DP}, d_{DP} &\sim GI(\sigma_k | c_{DP}, d_{DP}), \\ k_{f_p}(x_i, x_j | \lambda) &= \lambda \times \exp\{-\|x_i - x_j\|_2\}. \end{aligned}$$

## 5.2. Inferencia con *Gibbs Sampler*

Dado que el modelo descrito no es conjugado, las distribuciones posteriores tienen que ser aproximadas mediante métodos computacionales. Para hacer esto, se puede hacer uso de algoritmos MCMC (Monte Carlo Markov Chains), y particularmente del *Gibbs Sampler*. En caso de que el lector no esté familiarizado con este tipo de algoritmos, puede consultar el Anexo 1 de este trabajo.

En este orden de ideas, a continuación se detallan las distribuciones condicionales posteriores de los parámetros del modelo, así como la inclusión de algunas variables latentes para permitir el funcionamiento del algoritmo.

Cabe aclarar que antes de correr los algoritmos, resulta conveniente primero estandarizar los datos. En primer lugar, para que la estructura de covarianza tenga más sentido, ya que la escala de las covariables afectaría la correlación que existe entre los datos, al depender esta de la distancia absoluta entre ellas. Además, estandarizar los datos suele mejorar el rendimiento computacional de este tipo de algoritmos. Asimismo, vuelve más sencillo definir el valor inicial de los parámetros, como se detallará más adelante.

### 5.2.1. Actualización del error

Recordando que los centros de masa y los pesos del Proceso de Dirichlet son independientes, pueden ser actualizados por separado, con el inconveniente de que hay un número infinito de parámetros que actualizar. Para resolverlo, se utilizará el algoritmo de truncamiento mediante el *slice sampling*, propuesto por Kalli *et al.* (2009), y adaptado para el modelo propuesto en esta tesis.

Sea  $\xi_1, \xi_2, \xi_3, \dots$  una secuencia positiva, generalmente elegida determinista y decreciente. Sea  $N$  una variable aleatoria con soporte en los



enteros positivos, una variable auxiliar incorporada al modelo.

### Actualización de los centros de masa

Para cada  $k \in \{1, 2, \dots, N\}$ ,

$$\begin{aligned}\sigma_k | \{\varepsilon_{p_i}, z_i | z_i = k\}, c, d &\sim GI(\bar{c}_{DP}, \bar{d}_{DP}), \\ \bar{c}_{DP} &= c_{DP} + |\{i | z_i = k\}|, \\ \bar{d}_{DP} &= d_{DP} + p \left[ \sum_{\{i | z_i = k, \varepsilon_{p_i} \geq 0\}} \varepsilon_{p_i} \right] + (1 - p) \left[ \sum_{\{i | z_i = k, \varepsilon_{p_i} < 0\}} -\varepsilon_{p_i} \right].\end{aligned}$$

### Actualización de los pesos

Sea  $\hat{\pi}_k = \beta_k \prod_{j=1}^{k-1} (1 - \beta_j)$ , de modo que para cada  $k \in \{1, 2, \dots, N\}$ ,

$$\begin{aligned}\beta_k | \{z_i\}, a, b &\sim \text{Beta}(\bar{a}, \bar{b}), \\ \bar{a} &= 1 + |\{i | z_i = k\}|, \\ \bar{b} &= \alpha + |\{i | z_i > k\}|.\end{aligned}$$

Entonces, se calcula

$$\pi_k = \frac{\bar{\pi}_k}{\sum_{j=1}^N \bar{\pi}_j}$$

### Actualización de las clases y variables de truncamiento

Para cada observación  $i \in \{1, \dots, m\}$ , se obtiene

$$u_i \sim U(0, \xi_{z_i}),$$

valor que se utiliza para actualizar la probabilidad de pertenencia a cada clase de la siguiente forma. Para cada  $k \in \{1, 2, \dots, N\}$ ,

$$P(z_i = k | \varepsilon_{p_i}, \pi_k, \sigma_k) \propto \mathbb{1}(\xi_k > u_i) \cdot \frac{\pi_k}{\xi_k} \cdot AL_p(\varepsilon_{p_i} | \sigma_k).$$

Posteriormente se actualiza

$$N = \text{máx}\{N_i | N_i = \text{máx}\{j | \xi_j > u_i\}, i \in \{1, \dots, m\}\}.$$

### 5.2.2. Actualización de la tendencia

Se define la variable aleatoria auxiliar:

$$b_i \sim \begin{cases} \frac{p}{\sigma_i} & \text{prob} = P(\varepsilon_{p_i} \geq 0) = 1 - p \\ -\frac{1-p}{\sigma_i} & \text{prob} = P(\varepsilon_{p_i} < 0) = p \end{cases},$$

de forma que  $b = [b_1, \dots, b_m]^T$ .

#### Actualización de $f_p(x)$

Es posible calcular que

$$f_p(x) | Y, X, M_{f_p}(X), b, \lambda \sim \text{TruncNormal}(\bar{M}_{f_p}(X, b), K_{f_p}(X, X | \lambda), \rho, \eta),$$

$$\bar{M}_{f_p}(X, b) = M_{f_p}(X) + K_{f_p}(X, X | \lambda)b,$$

$$\rho_i = \begin{cases} -\infty & \text{si } b_i > 0 \\ y_i & \text{si } b_i < 0 \end{cases},$$

$$\eta_i = \begin{cases} y_i & \text{si } b_i > 0 \\ \infty & \text{si } b_i < 0 \end{cases},$$

donde  $\rho$  es el vector de límites inferiores y  $\eta$  es el vector de límites superiores de la distribución *Normal truncada*.

## Actualización del parámetro de escala

Por otro lado, se puede obtener que

$$P(\lambda|X, M_{f_p}(X), f_p(X), b, c_\lambda, d_\lambda) \propto \lambda^{-\bar{c}_\lambda - 1} \cdot \exp\left\{-\frac{\bar{d}_\lambda}{\lambda}\right\} \cdot \exp\{-\bar{B}\lambda\},$$

$$\bar{c}_\lambda = c_\lambda + \frac{p}{2},$$

$$\bar{d}_\lambda = d_\lambda + \bar{F},$$

$$\bar{F} = \frac{1}{2}(f_p(X) - M_{f_p}(X))^T [K_{f_p}(X, X|\lambda = 1)^{-1}](f_p(X) - M_{f_p}(X)),$$

$$\bar{B} = \frac{1}{2}b^T [K_{f_p}(X, X|\lambda = 1)]b.$$

## 5.3. Predicción

Una de las desventajas de los modelos no paramétricos es que, a diferencia de los modelos paramétricos, es complicado interpretar los resultados del ajuste del modelo.

Por ello, resulta particularmente importante la faceta de la predicción, que es la que más explota sus ventajas, y en la que los modelos paramétricos normalmente se quedan cortos. Específicamente esta sección se enfocará en la predicción de  $f_p$ , que es el parámetro de mayor interés del modelo.

Debido al uso del *Gibbs sampler*, después de realizar el ajuste se cuenta con un conjunto grande de realizaciones aproximadas de  $f_p(X)$ , provenientes de las cadenas de markov.

Recordando lo visto en la sección 4.2.4, cuando se tienen valores de  $f_p(X)$ , es posible usar la propiedad de la *Normal condicional* para realizar predicción. Sea  $X \in \mathbb{R}^m \times \mathbb{R}^n$  la matriz de datos originales,  $X_* \in \mathbb{R}^r \times \mathbb{R}^n$  la matriz de datos a predecir,  $f_p(X)$  una realización de la distribución posterior correspondiente a  $X$ , y  $f_p(X_*)$  el vector aleatorio

de los datos a predecir. Se tiene entonces que

$$f_p(X_*)|f_p(X) \sim \mathcal{N}(\bar{M}(X, X_*), \bar{K}(X, X_*|\lambda)),$$

con

$$\begin{aligned}\bar{M}(X, X_*) &= M(X_*) + K(X_*, X)K(X, X)^{-1}(f_p(X) - M(X)), \\ \bar{K}(X, X_*|\lambda) &= \lambda \times \left[ K(X_*, X_*) - K(X_*, X)K(X, X)^{-1}K(X, X_*) \right].\end{aligned}$$

con  $K(X_1, X_2) = K(X_1, X_2|\lambda = 1)$ , donde  $X_1$  y  $X_2$  pueden ser  $X$  o  $X_*$ .

Por lo antes descrito, es posible obtener una realización de  $f_p(X_*)$  simulando de dicha distribución *Normal*. De esta manera, por cada valor de  $f_p(X)$  y  $\lambda$  en la cadena de markov, se simula una realización de  $f_p(X_*)$ , y entonces es posible aproximar la distribución posterior de  $q_p(y|x)$ , para los datos  $X_*$ .

## 5.4. Parámetros iniciales

Una complicación que puede tener un modelo con la complejidad del GPDP, es que los parámetros que tiene que asignar el modelador no son inmediatos, sino están en la profundidad de un conjunto jerárquico de distribuciones. Por ello, no resulta sencillo asignarles valores iniciales.

Para mitigar este problema, a continuación se proponen una serie de heurísticas para su cálculo, mismas que se derivan de algunas ideas que me parecen sensatas, pero no se originan de ningún cuerpo axiomático y bien podrían ser mejoradas. También es importante aclarar que por lo comentado al inicio de la sección 5.2, para todas ellas se pensará que los datos están estandarizados.

### 5.4.1. Función de medias $m_{f_p}$

Para asignar la función de medias del proceso gaussiano, se puede partir de la hipótesis que  $m_{f_p}$  es constante, y por lo tanto, las variaciones son únicamente producto de la varianza de  $f_p$  y  $\varepsilon_p$ . Dada la estructura de probabilidad posterior, la media de  $f_p(x)$  podrá actualizarse si los datos cuentan con información suficiente para suponer lo contrario.

Una vez aceptada esta estructura para definir a la función de medias, resta asignar el valor constante que tomará, siendo una idea el asignar el cuantil muestral  $Q_p(y)$  de los datos de la variable de respuesta  $y$ , es decir,

$$m_{f_p} : \mathbb{R}^n \rightarrow \mathbb{R} | m_{f_p}(x) = Q_p(y).$$

### 5.4.2. *Gamma-Inversas* de $\lambda$ y el Proceso de Dirichlet

Tanto  $c_\lambda$  y  $d_\lambda$ , como  $c_{DP}$  y  $d_{DP}$  son parámetros de distribuciones *Gamma-Inversa*. Es oportuno recordar que si  $U \sim \mathcal{GI}(c, d)$ , entonces

$$\begin{aligned} \mathbb{E}[U] &= \frac{d}{c-1}, \quad c > 1 \\ \text{Var}(U) &= \frac{d^2}{(d-1)^2(d-2)}, \quad c > 2. \end{aligned}$$

Por lo tanto, eligiendo  $c = 2$ ,  $\text{Var}(U)$  será infinita y  $\mathbb{E}[U] = d$ . Asignar a  $c_\lambda$  y  $c_{DP}$  de esta manera permitirá darle a  $d_\lambda$  y  $d_{DP}$  el valor que se piense como el mejor estimador puntual *a priori* de  $\lambda$  y  $\sigma$ , pero con una varianza grande y cola pesada, que permitirá a los datos tener el peso principal en el ajuste del modelo.

Debido a la estandarización de los datos, la varianza muestral de  $y$  es igual a 1. Es posible pensarla como el resultado de sumar la varianza de  $f_p(x)$  y la de  $\varepsilon_p$ , que además se suponen independientes. Entonces, se puede definir una heurística tal que  $\text{Var}(f_p(x)) = \frac{1}{2}$  y  $\text{Var}(\varepsilon_p) = \frac{1}{2}$ ,

a falta de mayor información.

La varianza de  $f_p(x)$  es igual a  $\lambda$ , por lo que lo coherente con lo dicho en los párrafos anteriores será asignar  $d_\lambda = \frac{1}{2}$ .

Por el otro lado, si únicamente para este ejercicio, y con el afán de volver analítico el cálculo, se piensa a  $\varepsilon_p \sim AL_p(\sigma = d_{DP})$ . Entonces, su varianza estaría dada por

$$Var(\varepsilon_p) = \left[ \frac{d_{DP}}{p(1-p)} \right]^2 (1 - 2p(1-p)).$$

Dado que se fijará  $Var(\varepsilon_p) = \frac{1}{2}$ , por la heurística antes mencionada, despejando es posible obtener que

$$d_{DP} = \frac{p(1-p)}{\sqrt{2(1-2p(1-p))}}.$$

### 5.4.3. Parámetro de concentración $\alpha$

Este es el parámetro más difícil de definir, por su complejidad de interpretación. Pero cabe recordar que el valor de  $\alpha$  tiene una relación positiva con el número *clusters*.

De hecho, sea  $\bar{m}$  el número de clusters y  $m$  el número de datos de entrenamiento, Teh (2010) expone que

$$\mathbb{E}[\bar{m}|\alpha, m] \simeq \alpha \log \left( 1 + \frac{m}{\alpha} \right), \text{ para } m, \alpha \gg 0.$$


Si se define  $\alpha = \frac{\sqrt{m}}{2}$ , se tiene que

$$\begin{aligned} \mathbb{E}[\bar{m}|m] &\simeq \frac{\sqrt{m}}{2} \times \log(1 + 2\sqrt{m}) \\ &\simeq \frac{m}{7}, \text{ para } m \approx 100. \end{aligned}$$

Es decir, si se tienen alrededor de 100 observaciones, el número espe-

rado de *clusters* será alrededor de la séptima parte de las observaciones. Valor que a falta de mayor exploración en este tema, parece sensato.

## 5.5. Paquete *GPDPQuantReg* en R

Todas las ideas expuestas en este capítulo han sido implementadas en el paquete *GPDPQuantReg* del lenguaje de programación R, mismo que puede ser encontrado en el repositorio de Github  titulado: **opardo/GPDPQuantReg**.

Hasta el momento de escribir este trabajo, cuenta con tres funciones públicas: *GPDPQuantReg*, para ajustar el modelo con el *Gibbs sampler*; *predict*, para realizar predicción en un conjunto de datos del modelo ajustado; y *diagnose*, para realizar el diagnóstico de la ergodicidad, la autocorrelación, la correlación cruzada y la traza de las cadenas de markov, para los distintos parámetros.

A continuación se expone un ejemplo de uso, el cual es similar a lo que se realizó para obtener los resultados del capítulo siguiente.

```
1 # Instalación del paquete
2 install.packages("devtools")
3 library(devtools)
4 install_github("opardo/GPDPQuantReg")
5 library(GPDPQuantReg)
6
7 # Simulación de datos
8 set.seed(201707)
9 f_x <- function(x) return(0.5 * x * cos(x) - exp(0.1 * x))
10 error <- function(m) rgamma(m, 2, 1)
11 m <- 20
12 x <- sort(sample(seq(-15, 15, 0.005), m))
13 sample_data <- data.frame(x = x, y = f_x(x) + error(m))
14
15 # Ajuste del modelo
16 GPDP_MCMC <- GPDPQuantReg(y ~ x, sample_data, p = 0.250)
```

```
17 |
18 | # Predicción, usando el modelo ajustado
19 | pred_data <- data.frame(x = seq(-15, 15, 0.25))
20 | credibility <- 0.90
21 | prediction <- predict(GPDP_MCMC, pred_data, credibility)
22 |
23 | # Diagnóstico de las cadenas de markov
24 | diagnose(GPDP_MCMC)
```



# Bibliografía

- Bannerjee, S. 2008. *Bayesian Linear Models: The Gory Details*. Descargado de <http://www.biostat.umn.edu/ph7440/>.
- Denison, D.G.T. 2002. *Bayesian Methods for Nonlinear Classification and Regression*. Wiley Series in Probability and Statistics. Wiley.
- Dunson, D.B., & Taylor, J.A. 2005. Approximate Bayesian inference for quantiles. *Journal of Nonparametric Statistics*, **17**, 385–400.
- Fishburn, Peter C. 1986. The Axioms of Subjective Probability. *Statistical Science*, **1**(3), 335–345.
- Hanson, T., & Johnson, W.O. 2002. Modeling regression error with a mixture of Polya trees. *Journal of the American Statistical Association*, **97**, 1020–1033.
- Hao, L., & Naiman, D.Q. 2007. *Quantile Regression*. Quantile Regression, no. 149. SAGE Publications.
- Kalli, Maria, Griffin, Jim E., & Walker, Stephen G. 2009. Slice sampling mixture models. *Statistics and Computing*, **21**(1), 93–105.
- Koenker, Roger, & Bassett, Gilbert. 1978. Regression Quantiles. *Econometrica*, **46**(1), 33–50.

- Kottas, A., & Gelfland, A.E. 2001. Bayesian semiparametric median regression modeling. *Journal of the American Statistical Association*, **96**, 1458–1468.
- Kottas, A., & Krnjajic, M. 2005. *Bayesian Nonparametric Modeling in Quantile Regression*. Technical Report AMS 2005-06. University of California, Santa Cruz.
- Kottas, A., Krnjajic, M., & Taddy, M. 2007. Model-Based Approaches to Nonparametric Bayesian Quantile Regression. *Pages 1137–1148 of: Proceedings of the 2007 Joint Statistical Meetings*.
- Lavine, M. 1995. On an approximate likelihood for quantiles. *Biometrika*, **82**, 220–222.
- Pavlidis, Marios G., & Wellner, Jon A. 2012. Nonparametric estimation of multivariate scale mixtures of uniform densities. *Journal of Multivariate Analysis*, **107**, 71–89.
- Rasmussen, C.E., & Williams, C.K.I. 2006. *Gaussian Processes for Machine Learning*. Adaptive computation and machine learning series. University Press Group Limited.
- Schervish, M.J. 1996. *Theory of Statistics*. Springer Series in Statistics. Springer New York.
- Teh, Yee Whye. 2010. Dirichlet Process. *Pages 280–287 of: Sammut, C, & Webb, GI (eds), Encyclopedia of Machine Learning*. Springer.
- Tsionas, E.G. 2003. Bayesian quantile inference. *Journal of Statistical Computation and Simulation*, **73**, 659–674.
- Walker, S.G., & Mallick, B.K. 1999. A bayesian semiparametric accelerated failure time model. *Biometrics*, **55**(2), 477–483.

- Wasserman, Larry. 2006. *All of Nonparametric Statistics (Springer Texts in Statistics)*. Secaucus, NJ, USA: Springer-Verlag New York, Inc.
- Yu, K., & Moyeed, Rana A. 2001. Bayesian quantile regression. *Statistics & Probability Letters*, **54**(4), 437–447.