

INSTITUTO TECNOLÓGICO AUTÓNOMO DE MÉXICO

---



**UN MODELO BAYESIANO Y NO PARAMETRICO DE  
REGRESION SOBRE CUANTILES**

TESIS

QUE PARA OBTENER EL TITULO DE  
LICENCIADO EN MATEMATICAS APLICADAS

PRESENTA

**CARLOS OMAR PARDO GOMEZ**

CIUDAD DE MEXICO

2017

INSTITUTO TECNOLÓGICO AUTÓNOMO DE MÉXICO

---



**UN MODELO BAYESIANO Y NO PARAMETRICO DE  
REGRESION SOBRE CUANTILES**

TESIS

QUE PARA OBTENER EL TITULO DE  
LICENCIADO EN MATEMATICAS APLICADAS

PRESENTA

**CARLOS OMAR PARDO GOMEZ**

ASESOR: DR. JUAN CARLOS MARTINEZ OVANDO

CIUDAD DE MEXICO

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 tema de esta tesis es describir un modelo de regresión sobre cuantiles, debido a diversas bondades que presenta sobre el tradicional análisis de regresión a la media. Además, a través del paradigma Bayesiano permite incorporar conocimiento previo del fenómeno y presenta una gran flexibilidad, al contar con componentes no paramétricos. Asimismo, en este trabajo se abordan los modelos tradicionales de regresión, para entender las desventajas contrarrestadas por el nuevo modelo.

El capítulo 1 describe la importancia de las aproximaciones distintas a la regresión a la media, así como la evolución histórica de este tipo de modelos. El capítulo 2 introduce al paradigma Bayesiano y sus fundamentos 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 explica el algoritmo necesario para realizar inferencia y predicción. El capítulo 6 muestra simulaciones y aplicaciones del modelo, así como los resultados obtenidos de evaluarlo en diversos conjuntos de datos. Finalmente, el capítulo 7 hace referencia a las conclusiones de esta tesis, además de describir el trabajo futuro que se podría desarrollar.

# Índice general

<b>1. Introducción</b>	<b>6</b>
<b>2. Paradigma Bayesiano</b>	<b>9</b>
2.1. Inferencia de variables aleatorias . . . . .	9
2.2. Propiedad conjugada . . . . .	14
2.3. Inferencia con variables explicativas o predictivas . . . . .	15
<b>3. Modelos de regresión</b>	<b>17</b>
3.1. Concepto general . . . . .	17
3.2. Regresión a la media . . . . .	18
3.2.1. Modelo tradicional . . . . .	19
3.3. Regresión sobre cuantiles . . . . .	22
3.3.1. Modelo tradicional . . . . .	24
<b>4. Especificación no paramétrica</b>	<b>27</b>
4.1. Motivación . . . . .	27
4.2. En la distribución de $f_p$ , vía procesos Gaussianos . . . . .	30
4.2.1. Introducción a los procesos Gaussianos . . . . .	30
4.2.2. Definiciones y notación . . . . .	32
4.2.3. Funciones de covarianza . . . . .	34
4.2.4. Predicción . . . . .	36

4.3.	En la distribución de $\varepsilon_p$ , vía procesos de Dirichlet . . . . .	37
4.3.1.	Definición de los procesos de Dirichlet . . . . .	38
4.3.2.	Distribución posterior . . . . .	39
4.3.3.	Distribución predictiva . . . . .	40
4.3.4.	Proceso estocástico de rompimiento de un palo . . .	41
4.3.5.	Modelo general de mezclas infinitas de Dirichlet . . .	42
4.3.6.	Modelo de mezclas infinitas de Dirichlet para la distribución asimétrica de Laplace . . . . .	43
<b>5.</b>	<b>Modelo GPDP para regresión sobre cuantiles</b>	<b>46</b>
5.1.	Definición . . . . .	46
5.2.	Inferencia con el simulador de Gibbs . . . . .	48
5.2.1.	Actualización de la dispersión . . . . .	48
5.2.2.	Actualización de la tendencia . . . . .	50
5.3.	Predicción . . . . .	53
5.4.	Hiper-parámetros iniciales del modelo . . . . .	55
5.4.1.	Función de medias $m$ . . . . .	55
5.4.2.	<i>Gamma-Inversas</i> de $\lambda$ y el Proceso de Dirichlet . . .	56
5.4.3.	Parámetro de concentración $\alpha$ . . . . .	57
5.5.	Paquete <i>GPDPQuantReg</i> en R . . . . .	58
<b>6.</b>	<b>Aplicaciones</b>	<b>60</b>
6.1.	Simulación . . . . .	60
6.1.1.	Tendencia simple, dispersión simple . . . . .	61
6.1.2.	Tendencia compleja, dispersión simple . . . . .	64
6.1.3.	Tendencia simple, dispersión compleja . . . . .	65
6.1.4.	Tendencia compleja, dispersión compleja . . . . .	67
6.2.	Investigación de mercados . . . . .	69
6.2.1.	Conceptos iniciales . . . . .	69
6.2.2.	Caso real . . . . .	70
<b>7.</b>	<b>Conclusiones y trabajo futuro</b>	<b>74</b>



<b>Bibliografía</b>	<b>77</b>
<b>A. Distribuciones de referencia</b>	<b>80</b>
A.1. Distribución t-Student multivariada . . . . .	80
A.2. Distribución Normal condicional . . . . .	80
A.3. Distribución de Dirichlet . . . . .	81
<b>B. Algoritmos MCMC</b>	<b>84</b>
B.1. Introducción . . . . .	84
B.2. Simulador de Gibbs . . . . .	85
B.2.1. Simulador de Gibbs de dos pasos . . . . .	85
B.2.2. Simulador de Gibbs de múltiples pasos . . . . .	85
B.3. Monitoreo de convergencia y adaptación de los algoritmos MCMC . . . . .	87
B.3.1. Monitoreo de convergencia a la <i>estacionariedad</i> . . .	87
B.3.2. Monitoreo de convergencia a los promedios . . . . .	87
B.3.3. Monitoreo de convergencia a una muestra <i>iid</i> . . . .	88

# Capítulo 1

## Introducción

Detrás de cualquier modelo de regresión la intención es entender alguna característica asociada con una variable aleatoria en función de un conjunto de variables potencialmente explicativas o predictivas para tal característica. 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 y no lineal, simple y múltiple. Este tipo de modelos tiene un buen número de ventajas, entre las que destacan el bajo costo de estimación y la facilidad de interpretación. Sin embargo, como mencionan Hao & Naiman (2007), tienen tres grandes limitaciones, que a continuación se mencionan.

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 siempre es constante, sino cambia en sincronía con distintos valores de las covariables. También sucede que algunos fenómenos de estudio tienen distribuciones de colas pesadas, principalmente en las ciencias sociales. Esto da lugar a valores atípicos, mismos que pueden sesgar estimaciones de la media, mientras prácticamente no afectan a otros estadísticos, como la mediana.

La tercera es que al focalizar la intervención de las variables explicativas en la media, y fijar todo lo demás en la distribución del error, se suelen dejar de lado características del fenómeno que se está estudiando. Un ejemplo, dentro de otras características que podría estar dejando de lado, es la asimetría. Dicha propiedad es importante en estudios de ingreso, impuestos, esperanza de vida y, en general, 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 logaran 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 el  $p \times 100\%$  de los valores están por debajo de él, y el  $(1 - p) \times 100\%$ , por encima. 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 & Gelfland (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.

Con la finalidad de presentar un enfoque más flexible y totalmente probabilístico, esta tesis rescata las ideas de Kottas *et al.* (2007) y Kottas & Krnjajic (2005) para proponer un modelo Bayesiano y no paramétrico, útil en el contexto de regresión sobre cuantiles.

## Capítulo 2

# Paradigma bayesiano<sup>1,2</sup>

### 2.1. Inferencia de variables aleatorias

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 probabilidad condicional

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

---

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

<sup>2</sup>Esta tesis da como aceptados los axiomas de coherencia de la Teoría de la Decisión, mismos que pueden ser encontrados, por ejemplo, en Fishburn (1986). Por lo tanto, entiende al paradigma Bayesiano como el coherente para hacer estadística, cuando una toma de decisión con incertidumbre es el objetivo final del estudio.

pero esto requeriría conocer la función conjunta, misma que puede ser compleja por la estructura de dependencia de los datos.<sup>3</sup>

Este problema puede ser abordado mediante el Teorema de representación general de de Finetti. Para ello, antes se dará una definición.

**Definición.** Sea  $(y_1, y_2, \dots)$ , una sucesión de variables aleatorias, cuya distribución de probabilidad conjunta está dada por  $\mathbb{P}(y_1, y_2, \dots)$ . Sea  $\psi$  una función biyectiva que crea una permutación finita del conjunto  $\{1, 2, \dots\}$ , es decir, permuta un número finito de elementos y al resto los deja fijos. Se dice entonces que  $(y_1, y_2, \dots)$  es una **sucesión aleatoria infinitamente intercambiable** si se cumple que

$$\mathbb{P}(y_1, y_2, \dots) = \mathbb{P}(y_{\psi(1)}, y_{\psi(2)}, \dots),$$

para cualquier permutación  $\psi$ .

En pocas palabras, una sucesión  $(y_1, y_2, \dots)$  se considerará infinitamente intercambiable si el orden en que se etiquetan las variables no afecta su distribución conjunta. Es importante hacer notar que la comúnmente usada independencia implica intercambiabilidad, pero lo contrario no se cumple. Es decir, la intercambiabilidad es un supuesto menos rígido que la independencia.

Dicho esto, es momento de plantear el **Teorema de representación general de de Finetti**.<sup>4</sup>

---

<sup>3</sup>Cabe resaltar que en este trabajo se usará la notación  $\mathbb{P}$  como una forma general de definir una medida de probabilidad, independientemente de los asociados detalles teóricos sobre análisis y medibilidad.

<sup>4</sup>Una demostración de este teorema puede ser encontrada en Schervish (1996).

**Teorema.** *Sea  $(y_1, y_2, \dots)$  una sucesión aleatoria infinitamente intercambiable de valores reales. Entonces existe una distribución de probabilidad  $F$  sobre  $\mathcal{F}$ , el espacio de todas las distribuciones, de forma que la probabilidad conjunta de  $(y_1, y_2, \dots)$  se puede expresar como*

$$\mathbb{P}(y_1, y_2, \dots) = \int_{\mathcal{F}} \left[ \prod_{k=1}^{\infty} \mathbb{P}(y_k | G) \right] dF(G),$$

con

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

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

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

En otras palabras, el Teorema de de Finetti dice que  $\{y_1, y_2, \dots\}$  es un conjunto de variables aleatorias condicionalmente independientes, dada cierta distribución  $G$ . A su vez dicha  $G$  es desconocida y la incertidumbre respecto a ella se cuantifica mediante la distribución  $F(G)$ .

Cabe hacer notar que dicho teorema plantea la distribución conjunta de  $(y_1, y_2, \dots)$  como una mezcla de verosimilitudes condicionalmente independientes en  $G$ , donde el peso asociada a cada una depende de  $F(G)$ . Por lo tanto,  $F(G)$  expresa la creencia o conocimiento acerca de cuán probable es que  $G$  sea idónea para explicar el fenómeno, aún sin observar los datos.

Un subconjunto del espacio de todas las distribuciones  $\mathcal{F}$  es el espacio de las distribuciones paramétricas, es decir, aquellas que pueden ser descritas en su totalidad únicamente señalando el valor de un vector de parámetros

de tamaño finito  $\theta$ , mismo que puede tomar valores en todo un soporte  $\Theta$ <sup>5</sup>. Por lo tanto, si se hace el supuesto adicional que la distribución marginal de  $y_i$  es paramétrica, con vector de parámetros desconocido *theta*, se obtiene como corolario del Teorema de de Finetti que

$$\mathbb{P}(y_1, y_2, \dots) = \int_{\Theta} \left[ \prod_{k=1}^{\infty} \mathbb{P}(y_k | \theta) \right] \mathbb{P}(\theta) d\theta.$$

Siguiendo el razonamiento anterior,  $\mathbb{P}(\theta)$  indica la probabilidad de que  $\theta$  sea el vector de parámetros idóneo para explicar el fenómeno, antes de observar cualquier dato.

La intuición detrás de este resultado es que, 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 para  $\theta$ , sujeta a la información inicial o conocimiento previo (*CP*) que se tenga del fenómeno. Es decir,  $\mathbb{P}(\theta | CP)$ . Como una simplificación de la notación, en la literatura normalmente se escribe como  $\mathbb{P}(\theta) = \mathbb{P}(\theta | CP)$  y se conoce como la *probabilidad inicial* del parámetro.

Así, el Teorema de representación general garantiza que para variables aleatorias infinitamente intercambiables existe una distribución del vector de parámetros, tal que la probabilidad conjunta se puede expresar como la verosimilitud de variables independientes, condicionales a un vector de parámetros, multiplicada por la distribución inicial de dicho vector de pa-

---

<sup>5</sup>En lo que resta de esta sección las proposiciones y resultados harán referencia al conjunto restringido de distribuciones paramétricas, apelando a que así los nuevos conceptos tengan mayor claridad para el lector, considerando que este tipo de distribuciones son con las que se suele estar más familiarizado. Pero todos serán generalizables al conjunto de distribuciones  $\mathcal{F}$ .



rámetros, que es independiente de los valores que tomen las variables aleatorias. Con el teorema se prueba la existencia de dicha representación, mas no su unicidad. En los siguientes párrafos se hará uso de ella para realizar inferencia en variables aleatorias.

Regresando al problema inicial, y bajo los supuestos recién mencionados, es posible escribir

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

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  $\theta$ , y en realidad es una constante dada por la representación del Teorema de de Finetti de la probabilidad conjunta, por lo que normalmente la probabilidad condicional del vector de parámetros no se expresa como una igualdad, sino con la proporcionalidad

$$\mathbb{P}(\theta|y_1, \dots, y_n) \propto \mathbb{P}(y_{1,n}|\theta) \times \mathbb{P}(\theta),$$

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  $\mathbb{P}(y_{1,n}|\theta)$  es lo que se conoce también en otros paradigmas como *verosimilitud*, y que en caso de independencia condicio-

nal, dada por el Teorema de de Finetti, 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, el conocimiento final 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  $y_{n+1}$ . Ésta 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.

## 2.2. Propiedad conjugada

En los casos en los que la probabilidad posterior, que resulta del producto de la verosimilitud y la inicial, pertenece a la misma familia de la distribución inicial y únicamente difiere en el valor de los parámetros, se dice que la distribución de los parámetros con respecto a cierta verosimilitud pertenece a una **familia conjugada**.

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 son la *Normal-Normal*, *Normal-Gamma*, *Normal-Gamma Inversa*, *Bernoulli-Beta* o la *Poisson-Gamma*, donde la primera distribución representa a la verosimilitud y la segunda, la distribución de los parámetros.

Sin embargo, el rango de posibles modelos conjugados puede resultar limitado en algunos contextos prácticos debido a que el fenómeno en estudio puede ser mejor representado con ciertas distribuciones específicas, que usualmente no pertenecen a familias conjugadas.

## 2.3. Inferencia con variables explicativas o predictivas<sup>6</sup>

Como se verá en el siguiente capítulo de esta tesis, un problema común es estimar la distribución de cierta sucesión de variables aleatorias  $(y_1, y_2, \dots)$ , condicionadas a los valores  $(x_1, x_2, \dots)$  de otras variables comúnmente llamadas explicativas o predictivas. En este caso, la sucesión  $(y_1, y_2, \dots)$  ya no es intercambiable, porque cada valor  $y_i$  depende en alguna medida del valor de su respectiva  $x_i$ , por lo que no es posible aplicar de manera directa el Teorema de de Finetti. Para hacerlo de manera indirecta, se introducirá el término de intercambiabilidad parcial.

**Definición.** Sea  $(y_1, y_2, \dots)$  dada  $(x_1, x_2, \dots)$ , una sucesión numerable de variables aleatorias, asociadas con los correspondientes valores de ciertas variables predictivas, y cuya distribución de probabilidad conjunta está dada por  $\mathbb{P}(y_1, y_2, \dots | x_1, x_2, \dots)$ . Sea  $\psi$  una función biyectiva que crea una permutación finita de un conjunto, es decir, permuta un número finito de

---

<sup>6</sup>Esta sección carece de una formalidad completa, pero busca darle la intuición al lector para generalizar el resultado del Teorema de de Finetti en el contexto en el que se desarrollará este trabajo. Para una explicación más formal, consultar Dawid (2016).

elementos y al resto los deja fijos; y sean  $\tilde{x}_1, \tilde{x}_2, \dots$  los distintos valores únicos que toman las  $x$ 's. Entonces, se dice que  $(y_1, y_2, \dots)$  es una **sucesión aleatoria infinita y parcialmente intercambiable** si se cumple que

$$\mathbb{P}(y_{k_1}, y_{k_2}, \dots | \tilde{x}_k) = \mathbb{P}(y_{\psi(k_1)}, y_{\psi(k_2)}, \dots | \tilde{x}_k),$$

para cualquier permutación  $\psi$  y para todos los diferentes valores  $k$ .

Es decir, todas aquellas  $y$ 's cuyas  $x$ 's tienen el mismo valor, son infinitamente intercambiables entre sí.

Si además se cumpliera que el orden de los valores únicos  $\tilde{x}$ 's es intercambiable, entonces, intuitivamente se podría tomar la  $G$  del Teorema de de Finetti como dependiente de las  $x_i$ 's,  $G(x_i)$ , y se obtendría que

$$\mathbb{P}(y_1, y_2, \dots | x_1, x_2, \dots) = \int_{\mathcal{F}} \left[ \prod_{k=1}^{\infty} \mathbb{P}(y_k | G(x_i)) \right] dF(G(x_1, \dots, x_n)),$$

donde las  $y$ 's resultarían ser independientes entre sí, condicionadas a una distribución que depende la  $x_i$  asociada.

El tema del siguiente capítulo será la discusión de métodos de inferencia sobre las variables  $y$ , dentro de este contexto específico de dependencia de una variable explicativa o predictiva  $x$ , normalmente conocidos como **modelos de regresión**.

## Capítulo 3

# Modelos de regresión

### 3.1. Concepto general

Los modelos de regresión tienen como objetivo describir la distribución de una variable aleatoria  $y \in \mathbb{R}$ , generalmente llamada *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 utilizan un número finito de parámetros para describirla. Además, la interpretación de dichos parámetros suele tener relevancia para el modelador, como es el caso de la media o la mediana.

También es importante resaltar que este trabajo tiene interés en modelar a  $y$ , dado que ya se observaron los valores de  $x$ . Sin embargo, se podría pensar en modelos donde haga sentido la distribución conjunta de  $y$  y  $x$ , misma que se podría obtener como

$$\mathbb{P}(y, x) = \mathbb{P}(y|x) \times \mathbb{P}(x).$$

## 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 que requiere su estimación, particularmente porque familias conjugadas suelen hacer sentido en este contexto. De hecho, fuera del paradigma Bayesiano suelen ser comúnmente usados debido a que se asocian con la minimización de términos cuadrados, con las bondades de diferenciabilidad que eso implica. Además se valora la capacidad interpretativa que suelen tener los parámetros estimados.

En notación probabilística, retomando el hecho de que  $y|x \sim \mathbb{P}(y|x)$ , el modelo de regresión a la media 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 todo  $\tilde{\varepsilon} \neq \dot{\varepsilon}$ ,  $\tilde{\varepsilon}$  y  $\dot{\varepsilon}$  son independientes. Por lo tanto, sean  $\tilde{x}$  las covariables asociadas a la variable de respuesta  $\tilde{y}$ , y  $\dot{x}$  las asociadas a  $\dot{y}$ , se tiene que  $\tilde{y}|\tilde{x}, f, \sigma^2$  es condicionalmente independiente a  $\dot{y}|\dot{x}, f, \sigma^2$ .

### 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  $CP$  la hipótesis o el conocimiento previo al que tiene acceso el modelador, se tiene que

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

Como ya se mencionó en el capítulo anterior, se omitirá escribir la distri-

---

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

bución condicional respecto a  $CP$  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  $\{X, Y | X \in \mathbb{R}^{m \times n}, Y \in \mathbb{R}^m\}$ .<sup>2</sup> 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:

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

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, debido a la independencia condicional, 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)*

---

<sup>2</sup>Por simplificación y limpieza de notación en este trabajo se escribirán de igual manera variables aleatorias y los datos en efecto observados, considerando que en cada caso el contexto será suficiente para saber de cuál se está hablando, siendo asociadas las letras minúsculas a una única observación y las mayúsculas a una matriz de observaciones.



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  $a, b$  son los parámetros iniciales de forma y escala, respectivamente, 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 como se detalla a continuación.

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

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 en el Apéndice A.

### 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.** Sea  $F_y$  la función de distribución de la variable aleatoria  $y$ , entonces el cuantil  $p$ -ésimo de dicha variable aleatoria 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, se buscará 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  $q_p(\varepsilon) = 0$ .

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  $\varepsilon$ s, y, por lo tanto, hay independencia condicional entre las observaciones.

Otro aspecto importante a resaltar es que en este contexto la interpretación de  $\varepsilon_p$  como el *error aleatorio* ya no hace tanto sentido, y tendría que ser entendido más como la dispersión que siguen los datos alrededor de  $f_p$ .

### 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.*

Data esta definición, 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 variable de dispersión, entonces es posible definir

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

El modelo, entonces, 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, debido a la independencia condicional, 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 flexibles, que capturen con mayor precisión las particularidades de cada fenómeno y la interacción entre las variables de salida y las covariables. En el siguiente capítulo se discutirá la importancia de capturar mayor complejidad en la distribuciones, tanto de  $f_p$ , como de  $\varepsilon_p$ , mediante el uso de métodos no paramétricos.

## Capítulo 4

# Especificación no paramétrica

### 4.1. Motivación

En el capítulo anterior se analizaron métodos 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 en los parámetros. Pero las funciones lineales sólo son un 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  $\dot{x}$  es un

buen predictor de la media de  $y$ , pero como polinomio de orden 3, es decir,

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

Entonces, se puede definir el vector  $x$  de covariables como  $x = (1, \dot{x}, \dot{x}^2, \dot{x}^3)$  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

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

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

$$\begin{aligned}\dot{y} &= \dot{\beta}_0\dot{x}_1^{\beta_1}\dot{x}_2^{\beta_2}e^\varepsilon \\ \implies \ln(\dot{y}) &= \ln(\dot{\beta}_0) + \beta_1\ln(\dot{x}_1) + \beta_2\ln(\dot{x}_2) + \varepsilon \\ \implies y &= \beta_0 + \beta_1x_1 + \beta_2x_2 + \varepsilon,\end{aligned}$$

con

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

y el procedimiento se convierte en el ya conocido.



Si bien estos ejemplos amplían el conjunto de funciones que es posible cubrir usando el modelo tradicional de regresión lineal, también 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 de relación entre  $y$  y  $x$ , sin restringirla a un 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 a la dispersión  $\varepsilon_p$ , 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 arriba, el  $(1 - p) \times 100\%$ .

Si bien esta es una característica necesaria, puede no ser suficiente debido a que la forma de la distribución de la dispersión podría ser distinta a la asimétrica de Laplace, por ejemplo, en el peso que le asigna a las colas. Dicha problemática podría ser mitigada mediante el uso de una mezcla de distribuciones como aproximación de la distribución de la dispersión. Particularmente es posible usar asimétricas de Laplace con diferentes valores para  $\sigma$  y probabilidad asociada a cada uno de esos valores 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 óp-

tima, ya que cualquier mezcla con número fijo de parámetros sería un caso particular.

En resumen, tanto la estimación de la distribución 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 los métodos estadísticos tradicionales es imposible hacerlo, y más si el tiempo es finito. Pero esto abre la puerta a una visión menos explorada para hacer estadística: los **métodos no paramétricos**.

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 la distribución de $f_p$ , vía procesos Gaussianos <sup>1</sup>

### 4.2.1. Introducción a los procesos Gaussianos

Retomando las ideas del capítulo anterior, los modelos de regresión tienen como objetivo describir la distribución de una variable aleatoria  $y$ , condi-

---

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

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

Es importante recordar que la función  $f_p$  es pensada fija, 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.

Para ello, se puede pensar a  $f_p$  como aleatoria. Pero como  $f_p$  usualmente está definida para múltiples valores  $x$  ya no se podrá pensar como una variable aleatoria, sino como un conjunto de variables aleatorias que depende de variables de entrada, es decir, un proceso estocástico. Para el caso particular de esta tesis, se pensará que sigue un proceso *proceso Gaussiano*, concepto que se introduce a continuación.

**Definición 3.** *Un **proceso Gaussiano** es un conjunto de variables aleatorias, tal que todo subconjunto finito de ellas tendrá una distribución Gaussiana (Normal) conjunta.*

Visto de esa manera, cada  $f_p(x)$  tiene una distribución Normal univariada, con media  $m(x)$  y covarianza  $k(x, x')$ , las cuales reflejan el conocimiento previo que se tiene del fenómeno de estudio. Cabe resaltar que dicha media  $m(x)$  y covarianza  $k(x, x')$  están en función de  $x$ , es decir, varían 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

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

$$x_i \in \mathbb{R}^n, \forall i \in \{1, \dots, m\}.$$

Y debido a la definición de proceso Gaussiano,  $f_p(X) \in \mathbb{R}^n$  es un vector aleatorio que se distribuye de forma Normal multivariada, con vector de medias  $M_{f_p}(X)$  y matriz de covarianzas  $K_{f_p}(X, X)$ .

#### 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*

$$M_{f_p} : \mathbb{R}^m \times \mathbb{R}^n \rightarrow \mathbb{R}^m \mid$$

$$M_{f_p}(X) = \begin{bmatrix} m_{f_p}(x_1) \\ \dots \\ m_{f_p}(x_m) \end{bmatrix}.$$

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

$$K_{f_p} : \mathbb{R}^m \times \mathbb{R}^n \times \mathbb{R}^m \times \mathbb{R}^r \rightarrow \mathbb{R}^m \times \mathbb{R}^m \mid$$

$$K_{f_p}(X, X') = \begin{bmatrix} k_{f_p}(x_1, x'_1) & \dots & k_{f_p}(x_1, x'_r) \\ \dots & \dots & \dots \\ k_{f_p}(x_m, x'_1) & \dots & k_{f_p}(x_m, x'_r) \end{bmatrix}.$$

Dadas estas definiciones, se puede observar que el proceso Gaussiano  $f_p$  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  $f_p$  que se distribuye como un proceso Gaussiano, se hará con la notación

$$f_p \sim \mathcal{GP}(m, k).$$

### 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 variables aleatorias que son elementos del proceso estocástico  $f_p$ . Por ejemplo,  $f_p(x)$  y  $f_p(x')$ . Para explicar de mejor manera este punto, se

puede pensar en el siguiente ejemplo.

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

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

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

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  $x'$ , escrita usualmente como  $\|x, x'\|$ . Es decir,  $k(x, x') = k(\|x, x'\|)$ . 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  $f_p(x) \approx f_p(x')$ , por lo que se 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:

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

donde  $\lambda$  es un parámetro de escala,  $\tau$  de rango y  $\gamma$  especifica el tipo de norma euclidiana a usar.

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.

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

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

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

#### 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  $\{(X, f_p(X)) | X \in \mathbb{R}^{m \times n}, f_p(X) \in \mathbb{R}^m\}$ . Por otro lado, se tiene un nuevo conjunto de covariables  $X_* \in \mathbb{R}^{r \times n}$ , y se desea predecir  $f_p(X_*) \in \mathbb{R}^r$ , subconjunto del proceso Gaussiano  $f_p$ .

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

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



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<sup>2</sup>, se obtiene que:

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

con

$$\begin{aligned}\bar{M}(X, X_*) &= M(X_*) + K(X_*, X)K(X, X)^{-1}(f_p(X) - M(X)), \\ \bar{K}(X, X_*) &= K(X_*, X_*) - K(X_*, X)K(X, X)^{-1}K(X, X_*).\end{aligned}$$

De esta manera quedan sentadas las bases de la distribución no paramétrica de  $f_p$ . A continuación se analizarán las de la distribución de  $\varepsilon_p$ , y en el próximo capítulo se estudiará cómo hacer inferencia conjuntando ambas, mediante el uso del modelo central de esta tesis.

### 4.3. En la distribución de $\varepsilon_p$ , vía Procesos de Dirichlet <sup>3</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 este trabajo y de su misión de encontrar un modelo Bayesiano y no paramétrico para la regresión sobre cuantiles, los procesos

---

<sup>2</sup>La especificación de la distribución Normal condicional se puede encontrar en el Apéndice A.

<sup>3</sup>Las ideas de esta sección son retomadas de Teh (2010).

de Dirichlet serán utilizados para ajustar la distribución de la dispersión  $\varepsilon_p$  alrededor de  $f_p$ .

#### 4.3.1. Definición de los procesos de Dirichlet

En términos generales, para que una distribución de probabilidad  $G$  se distribuya de acuerdo a un proceso de Dirichlet, toda partición finita de su soporte tiene que tener una distribución Dirichlet<sup>4</sup>. A continuación se enuncia una definición más detallada.

**Definición 4.** Sean  $G$  y  $H$  dos distribuciones cuyo soporte es el conjunto  $\Theta$  y sea  $\alpha \in \mathbb{R}_+$ . Si se toma una partición finita cualquiera  $(A_1, \dots, A_r)$  del conjunto  $\Theta$ , se entenderá que  $G(A_i)$  es la probabilidad de que una realización de  $G$  pertenezca al conjunto  $A_i$ . A su vez,  $G$  será realización de otra variable aleatoria, por lo que el vector  $(G(A_1), \dots, G(A_r))$  también será aleatorio.

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 concentración  $\alpha$ , si

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

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

---

<sup>4</sup>Antes de revisar la definición formal de los Procesos de Dirichlet, es conveniente recordar la definición de la distribución de Dirichlet, misma que se ubica en el Apéndice A.

de la distribución de Dirichlet, entonces

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

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

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, es decir, es un parámetro de precisión. 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  $H$ .

### 4.3.2. Distribución posterior

Sea  $(\phi_1, \dots, \phi_n)$  una sucesión de realizaciones independientes provenientes de la función de distribución  $G$ , y toman valores dentro de  $\Theta$ .  $G$  es desconocida, y para reflejar dicha incertidumbre se le asigna a su vez una distribución, particularmente la de un proceso de Dirichlet.

Sea de nuevo  $(A_1, \dots, A_r)$  una partición finita cualquiera del conjunto  $\Theta$ , y  $n_k = |\{i : \phi_i \in A_k\}|$  el número de valores  $\phi$  observados dentro del conjunto  $A_k$ . Por la propiedad conjugada entre la distribución de Dirichlet

y la distribución Multinomial, se obtiene que

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

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  $\phi_i, \dots, \phi_n$  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, casi seguramente. Por lo tanto,  $G$  es una distribución discreta también casi seguramente.

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

Dado que  $G \sim DP(\alpha, H)$  es una distribución discreta casi seguramente, se puede expresar como una suma de centros de masa de la siguiente manera:

$$\begin{aligned}
 G(\phi) &= \sum_{k=1}^{\infty} \pi_k \delta_{\phi_k^*}(\phi), \\
 \phi_k^* &\sim H,
 \end{aligned}$$

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

Dicha probabilidad de ocurrencia será generada con la siguiente metáfora.<sup>5</sup> 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  $\sum_{k=1}^{\infty} \pi_k = 1$ . En algunas ocasiones se nombra a esta distribución  $\pi \sim \text{GEM}(\alpha)$ , en honor a Griffiths, Engen y McCloskey.

#### 4.3.5. 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*:

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

En este *modelo de mezclas* es posible que existan  $y$ 's que comparten un mismo valor para  $\phi_i$  (por la propiedad discreta de  $G$ ), por lo que pueden ser consideradas pertenecientes a una misma subpoblación.

---

<sup>5</sup>Una demostración de la equivalencia puede ser encontrada en Paisley (2010).

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  $z_i$  la subpoblación a la que pertenece  $y_i$  entre las  $\Phi_1^*, \Phi_2^*, \dots$  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

$$\begin{aligned} y_i | z_i, \phi_k^* &\sim F(y_i | \phi_{z_i}), \\ z_i | \pi &\sim Mult_\infty(\pi)^6, \\ \pi | \alpha &\sim GEM(\alpha), \\ \phi_k^* | H &\sim H. \end{aligned}$$

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 subpoblaciones, pero donde intuitivamente la importancia realmente recae sólo en aquellas que tienen un peso  $\pi$  posterior mayor a cierto umbral; pero dichos pesos son detectados hasta después de observar los datos, a diferencia de los modelos de mezclas finitas, que ya tienen un número de subpoblaciones definidas previamente.

#### 4.3.6. Modelo de mezclas infinitas de Dirichlet para la distribución asimétrica de Laplace

Aterrizando las ideas anteriores al caso particular de los modelos de regresión sobre cuantiles, se busca describir la distribución de  $\varepsilon_p$  como producto de una mezcla infinita de distribuciones asimétricas de Laplace, de la mane-

---

<sup>6</sup>Se usará la notación  $Mult_\infty$  para denotar al límite de la distribución Multinomial, cuando el número de posibles categorías tiende a infinito.

ra 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

$$\begin{aligned}\varepsilon_{p_i}|z_i, \sigma_k^* &\sim AL_p(\varepsilon_{p_i}|\sigma_{z_i}), \\ z_i|\pi &\sim Mult_\infty(\pi), \\ \pi|\alpha &\sim GEM(\alpha), \\ \sigma_k^*|H &\sim H.\end{aligned}$$

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 hiper-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.

En el siguiente capítulo se retomará este modelo para especificar la dispersión de la regresión sobre cuantiles, y conjuntándolo con los procesos Gaussianos (vistos antes en este capítulo), se obtendrá el modelo GPDP,



centro de esta tesis.

## Capítulo 5

# Modelo GPDP para regresión sobre cuantiles

### 5.1. Definición

Después de analizar la introducción de componentes no paramétricos en las distribuciones, tanto de  $f_p$ , como de  $\varepsilon_p$ , 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 tendencia y  $\varepsilon_p \in \mathbb{R}$  es la dispersión, ambos desconocidos.

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

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

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

La razón de fijar  $\tau = 1$  es para simplificar el proceso de inferencia que se verá en la siguiente sección, pero bien podría también tener una distribución inicial que refleje la incertidumbre acerca de su valor.

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:

$$\begin{aligned} y_i | f_p(x_i), z_i, \sigma_k^* &\sim AL_p(\varepsilon_{p_i} = y_i - f_p(x_i) | \sigma_{z_i}), \\ f_p | m, k, \lambda &\sim \mathcal{GP}(m, k(\lambda) | \lambda), \\ \lambda &\sim GI(c_\lambda, d_\lambda), \\ z_i | \pi &\sim Mult_\infty(\pi), \\ \pi | \alpha &\sim GEM(\alpha), \\ \sigma_k^* | c_{DP}, d_{DP} &\sim GI(\sigma_k | c_{DP}, d_{DP}), \\ k(x_i, x_j | \lambda) &= \lambda \exp\{-\|x_i - x_j\|_2\}. \end{aligned}$$

## 5.2. Inferencia con el simulador de Gibbs

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 simulador de Gibbs.<sup>1</sup>

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. Es momento de recordar que dichas distribuciones posteriores resultan de multiplicar la verosimilitud y la distribución inicial, como se revisó en el capítulo 2 de este trabajo.

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 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 de la dispersión

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 del *slice sampling*, propuesto

---

<sup>1</sup>En caso de que el lector no esté familiarizado con este tipo de algoritmos, puede consultar una breve descripción de ellos en el Apéndice B.

por Kalli *et al.* (2009), y adaptado para el modelo propuesto en esta tesis. A grandes rasgos consiste en truncar las posibles subpoblaciones a las que puede pertenecer cada observación a un número finito, el cual variará de forma aleatoria, de acuerdo a lo que vaya aprendiendo de los datos.

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 números naturales, una variable auxiliar incorporada al modelo.

### Actualización de los centros de masa

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

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

### Actualización de los pesos

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

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

Dado que existe un número finito de posibles subpoblaciones, ya no se sigue

propiamente la distribución  $GEM$ , sino un truncamiento de ella hasta la  $N$ -ésima subpoblación. Posteriormente se realiza un cambio de escala a las probabilidades para que sumen 1. Es decir, se calcula

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

### Actualización de las clases y variables de truncamiento

Siguiendo el algoritmo de Kalli *et al.* (2009), 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  $k \in \{1, 2, \dots, N\}$ ,

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

Posteriormente se actualiza

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

#### 5.2.2. Actualización de la tendencia

Se define la siguiente variable aleatoria auxiliar, con la finalidad de anticipar si  $\varepsilon_p = y - f_p(x)$  será positiva o negativa, y así simplificar el cálculo de la actualización de  $f_p$ .

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

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

### Actualización de $f_p(X)$

Es pertinente recordar que la función de densidad de una observación  $y_i$ , debido a que sigue la distribución asimétrica de Laplace, se escribe

$$P(y_i|f_p(x_i), \sigma_i) = \frac{p(1-p)}{\sigma_i} \exp \left\{ -\rho_p \left( \frac{y_i - f_p(x_i)}{\sigma_i} \right) \right\} \mathbb{1}_{(-\infty, \infty)}.$$

Una vez calculada la variable auxiliar  $b$  que recién se acaba de definir, dicha densidad se puede expresar de forma condicional como

$$P(y_i|f_p(x_i), \sigma_i, b_i) \propto \begin{cases} \exp \left\{ \frac{-p(y_i - f_p(x_i))}{\sigma_i} \right\} \mathbb{1}_{\{y_i - f(x_i) \geq 0\}} & \text{si } b_i > 0 \\ \exp \left\{ \frac{(1-p)(y_i - f_p(x_i))}{\sigma_i} \right\} \mathbb{1}_{\{y_i - f(x_i) < 0\}} & \text{si } b_i < 0 \end{cases}.$$

Por lo tanto, la verosimilitud de las observaciones se puede calcular como

$$\begin{aligned}
 & P(y|f_p(X), \sigma, b) \\
 & \propto \exp \left\{ - \sum_{\{i|b_i>0\}} \frac{p}{\sigma_i} (y_i - f_p(x_i)) - \sum_{\{i|b_i<0\}} -\frac{1-p}{\sigma_i} (y_i - f_p(x_i)) \right\} \\
 & \quad \prod_{\{i|b_i>0\}} \mathbb{1}_{\{y_i \geq f(x_i)\}} \prod_{\{i|b_i<0\}} \mathbb{1}_{\{y_i < f(x_i)\}} \\
 & = \exp \left\{ - \sum_{\{i|b_i>0\}} b_i (y_i - f_p(x_i)) - \sum_{\{i|b_i<0\}} b_i (y_i - f_p(x_i)) \right\} \\
 & \quad \prod_{\{i|b_i>0\}} \mathbb{1}_{\{y_i \geq f(x_i)\}} \prod_{\{i|b_i<0\}} \mathbb{1}_{\{y_i < f(x_i)\}} \\
 & = \exp \left\{ -b^T (y - f_p(X)) \right\} \prod_{\{i|b_i>0\}} \mathbb{1}_{\{y_i \geq f(x_i)\}} \prod_{\{i|b_i<0\}} \mathbb{1}_{\{y_i < f(x_i)\}}.
 \end{aligned}$$

Es importante recalcar que en esta peculiar verosimilitud, cada  $f_p(x_i)$  estará condicionada de manera excluyente a estar por arriba o por abajo de  $y_i$ . Por ese motivo, al multiplicar la verosimilitud por la distribución inicial Gaussiana de  $f_p(X)$ , se obtendrá como distribución posterior una Normal Truncada, misma que se detalla a continuación.

$$f_p(X)|Y, X, M, b, \lambda \sim \text{TruncNormal}(\bar{M}(X, b), K(X, X|\lambda), \gamma, \eta),$$

$$\bar{M}(X, b) = M(X) + K(X, X|\lambda)b,$$

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

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

donde  $\gamma$  es el vector de límites inferiores y  $\eta$  es el vector de límites superiores



de la distribución Normal truncada.

Cabe notar que debido a que  $f_p(X)$  se encuentra en la verosimilitud únicamente como un elemento de primer orden, al hacer la multiplicación con la distribución inicial Normal, la varianza queda exactamente igual. La actualización se da únicamente en la media, como una perturbación de la media inicial, dada por las covarianzas que tiene cada observación respecto a las demás, así como el signo de  $b_i$  para cada una.

### Actualización del parámetro de escala

Condicional a los demás valores obtenidos, se obtiene la distribución posterior de  $\lambda$  como

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

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

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

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

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

## 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 simulador de Gibbs, 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

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

donde  $K(X_1, X_2) = K(X_1, X_2|\lambda = 1)$ , y  $X_1$  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. Hiper-parámetros iniciales del modelo

Una complicación que puede tener un modelo con la complejidad del GPDP es que los hiper-parámetros que tiene que definir 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 $m$

Para asignar la función de medias del proceso Gaussiano, se puede partir de la hipótesis que  $m$  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  $f_p$  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. Si el modelador tiene una idea del nivel donde espera los datos, puede asignar la constante  $c$ . En caso de no tenerla, un valor que se suele usar en diversos contextos es el de  $c = 0$ , de forma que

$$\begin{aligned} m : \mathbb{R}^n &\rightarrow \mathbb{R} \mid \\ m(x) &= c. \end{aligned}$$

### 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

$$\mathbb{E}[U] = \frac{d}{c-1}, \quad c > 1,$$

$$Var(U) = \frac{d^2}{(c-1)^2(c-2)}, \quad c > 2.$$

Por lo tanto, eligiendo  $c = 2$ ,  $Var(U)$  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 tal 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  $Var(f_p(x)) = \frac{1}{2}$  y  $Var(\varepsilon_p) = \frac{1}{2}$ , 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, des-

pejando 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 de subpoblaciones.

De hecho, sea  $\bar{m}$  el número de subpoblaciones 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 esperado de subpoblaciones 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/GPDP-QuantReg**.

Al momento de escribir este trabajo, cuenta con tres funciones públicas: *GPDPQuantReg*, para ajustar el modelo con el simulador de Gibbs; *predict*, para realizar predicción en un nuevo 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)
```

## Capítulo 6

# Aplicaciones

A continuación se exponen los resultados de utilizar el paquete *GPDP-QuantReg* en R, mismo que, como se detalló en el capítulo anterior, implementa el modelo *GPDP* para la regresión sobre cuantiles.

En primer lugar se presenta el ajuste del modelo en datos simulados, con el fin de comparar los resultados con los valores conocidos de antemano. Posteriormente se presenta para un conjunto de datos reales, con la intención de obtener conclusiones en aplicaciones prácticas, mediante el uso del modelo central de esta tesis.

### 6.1. Simulación

Los datos de esta sección se obtuvieron de la siguiente manera. Sea  $y \in \mathbb{R}$  el valor de la variable de respuesta,  $x \in \mathbb{R}$  su respectiva covariable,  $g : \mathbb{R} \rightarrow \mathbb{R}$  la función denominada *tendencia* y  $E \in \mathbb{R}$  una dispersión aleatoria, se



simuló:

$$y = g(x) + E.$$

Las diferencias entre las subsecciones siguientes radican en variaciones del valor de  $g$  y la distribución de  $E$ .

Dada esta construcción, la función real del cuantil  $p$ -ésimo de  $y|x$  se puede obtener como

$$q_p(y|x) = g(x) + q_p(E),$$

la cual se estimará para diversos valores de  $p$ .

En todos los casos se ajustó el modelo para los cuantiles 0.5-ésimo, por ser la mediana y una medida de tendencia central; el 0.95-ésimo, dado que es un valor extremo, y el 0.25-ésimo, debido a que es el primer cuartil, y no es ni medida de tendencia central, pero tampoco un valor extremo.

Por otro lado, para todos los casos se simularon 40 datos sin reemplazo, dentro del intervalo  $(-15, 15)$ , con un refinamiento de 3 decimales.

### 6.1.1. Tendencia simple, dispersión simple

Para obtener este conjunto de datos se utilizó una tendencia  $g$  considerada simple: la cuadrática

$$g(x) = \frac{1}{40}x^2 - \frac{1}{20}x - 2.$$

Por otro lado, la distribución de la dispersión  $E \sim \mathcal{N}(0, 1)$  también fue sencilla, debido a que fue simétrico y no acotado. Los datos simulados se pueden observar en la figura 6.1.

Posteriormente se ajustó el modelo y se realizó predicción sobre un refina-

miento del intervalo simulado de  $x$ , obteniendo buenos resultados (figura 6.2), ya que las funciones reales de los diversos cuantiles cayeron en su totalidad dentro del intervalo de probabilidad al 90 %, estimado por el modelo. Además, las medianas de la distribuciones posteriores siguieron un comportamiento similar a las originales.

Figura 6.1: Datos simulados y cuantiles de referencia, para tendencia simple y dispersión simple.

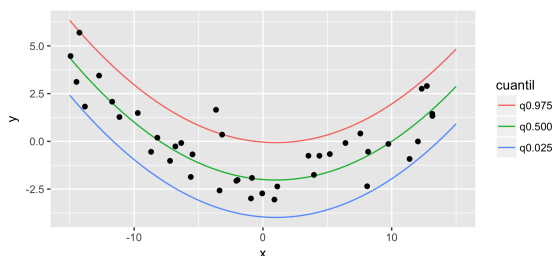
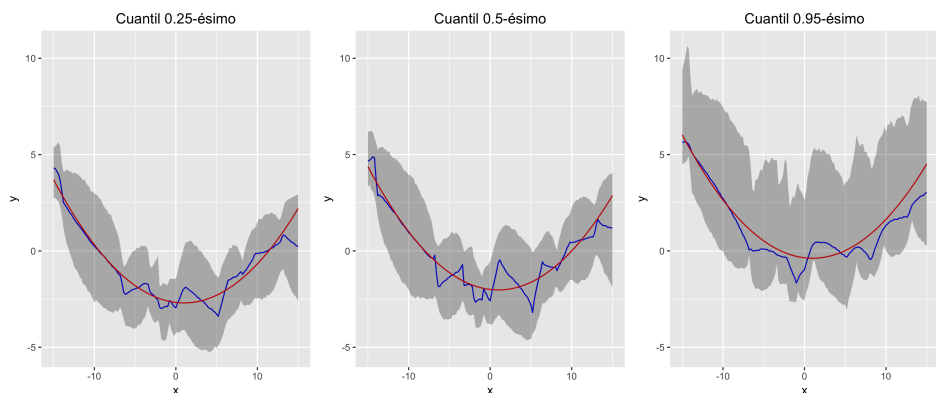


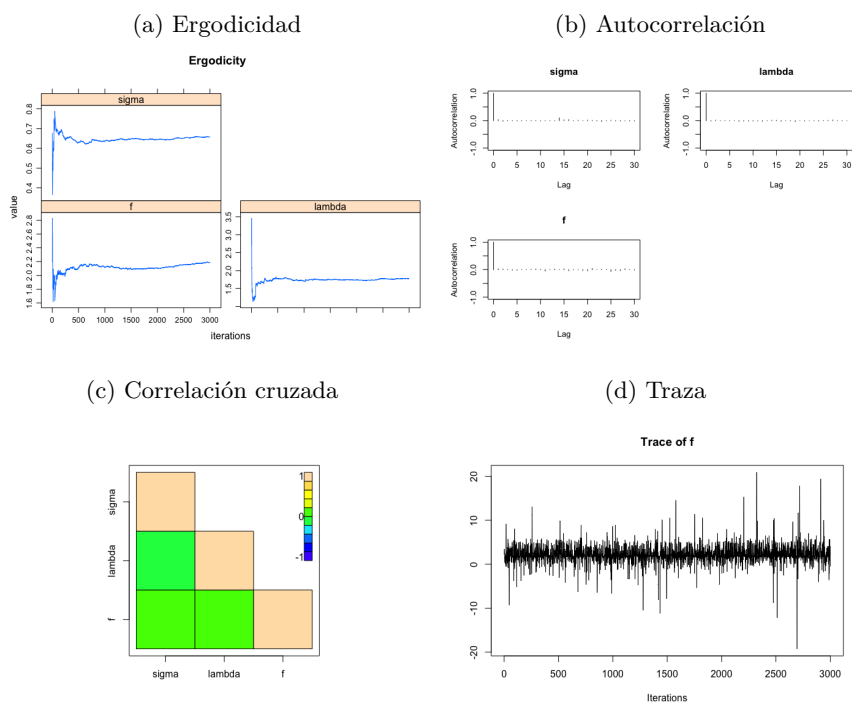
Figura 6.2: Ajuste del modelo *GPDP*, para tendencia simple y dispersión simple.



Nota: La línea roja representa el valor real de cada cuantil, la línea azul representa la mediana de la distribución posterior predictiva y el área gris su intervalo de probabilidad al 90 %.

Por otro lado, como se detalló en el capítulo anterior, con el uso del paquete *GPDPQuantReg* también es posible algunos de los diagnósticos de las cadenas de Markov, los cuales se detallan en el Apéndice B. Por ejemplo, se presentan los del cuántil 0.5-ésimo en la figura 6.3, mismos que reflejan un buen desempeño del algoritmo.

Figura 6.3: Diagnósticos de las cadenas de Markov del cuántil 0.5-ésimo, para tendencia simple y dispersión simple.



### 6.1.2. Tendencia compleja, dispersión simple

En este caso, se mantuvo que  $E \sim \mathcal{N}(1, 0)$ , pero la tendencia  $g$  usada fue más compleja:

$$g(x) = \frac{1}{2}x \cos(x) - \exp\left(\frac{1}{10}x\right).$$

Los datos simulados se pueden observar en la figura 6.4, los cuales al ajustar el modelo mostraron de nuevo buenos resultados (figura 6.5), apareciendo la tendencia original adentro del intervalo de probabilidad al 90 % en prácticamente todos los valores de  $x$  en los que se realizó predicción, a excepción de la última zona, en la que no hubo datos de entrenamiento.

Figura 6.4: Datos simulados y cuantiles de referencia, para tendencia compleja y dispersión simple.

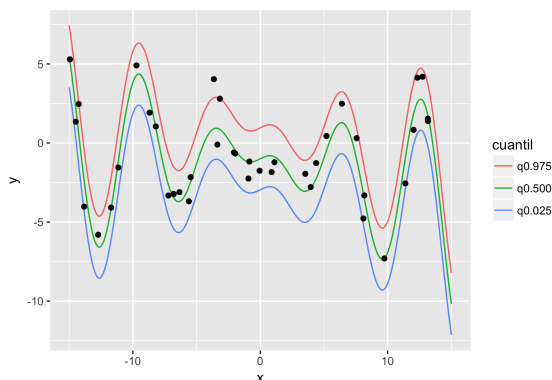
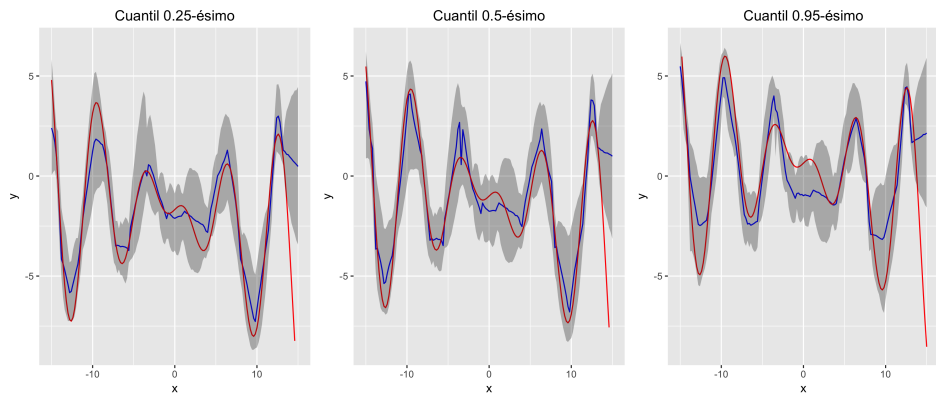


Figura 6.5: Ajuste del modelo *GPDP*, para tendencia compleja y dispersión simple.



Nota: La línea roja representa el valor real de cada cuantil, la línea azul representa la mediana de la distribución posterior predictiva y el área gris su intervalo de probabilidad al 90 %.

### 6.1.3. Tendencia simple, dispersión compleja

En este caso, se usó la tendencia lineal:

$$g(x) = \frac{1}{2}x.$$

La complejidad se introdujo en  $E \sim \text{Gamma}(\alpha = 2, \beta = 1)$ , debido a que la dispersión no fue simétrica y fue acotada por la izquierda.

El conjunto de datos usado para este modelo aparece en la figura 6.6, y a pesar de la complejidad de la dispersión, se obtuvieron buenos resultados (figura 6.7), debido a que nuevamente las funciones reales de los diversos cuantiles cayeron en su totalidad dentro del intervalo de probabilidad al 90 %, estimado por el modelo.

Un detalle notable es que, al igual que el caso de tendencia simple y dispersión simple, la estimación de la función del cuantil 0.95-ésimo muestra una varianza más grande que los otros dos cuantiles. Esto debido a que en valores extremos el modelo refleja mayor incertidumbre de lo que en realidad podría estar ocurriendo.

Figura 6.6: Datos simulados y cuantiles de referencia, para tendencia simple y dispersión compleja.

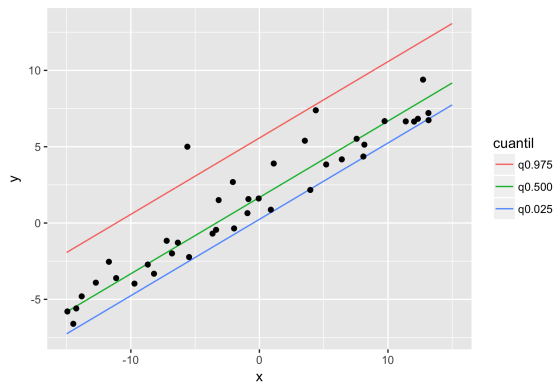
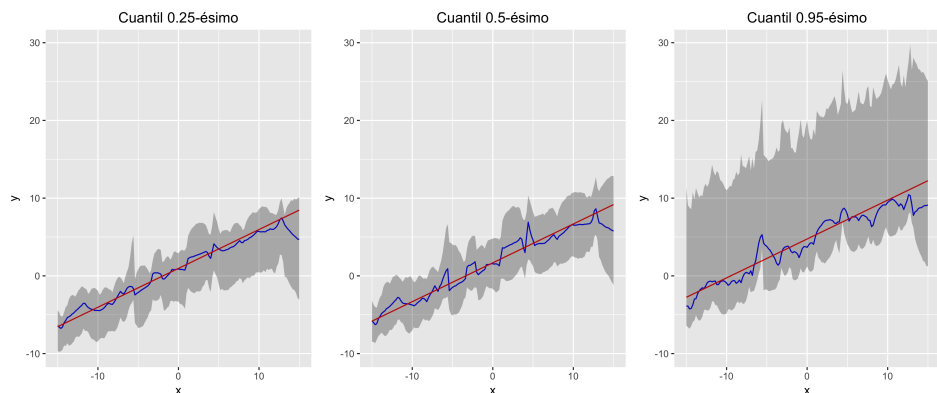


Figura 6.7: Ajuste del modelo *GPDP*, para tendencia simple y dispersión compleja.



Nota: La línea roja representa el valor real de cada cuantil, la línea azul representa la mediana de la distribución posterior predictiva y el área gris su intervalo de probabilidad al 90 %.

#### 6.1.4. Tendencia compleja, dispersión compleja

En este modelo se usaron datos (figura 6.8) provenientes tanto de una dispersión compleja,  $E \sim \text{Gamma}(\alpha = 2, \beta = 1)$ , como de una tendencia  $g$  compleja:

$$g(x) = \frac{1}{2}x \cos(x) - \exp\left(\frac{1}{10}x\right).$$

Después del ajuste (figura 6.9), el balance fue positivo, debido a que las funciones reales de los diversos cuantiles cayeron en su totalidad dentro del intervalo de probabilidad al 90 %, del estimado por el modelo, a excepción de la zona de la que no se tuvieron datos, como en el caso de tendencia compleja y dispersión simple. Pero a diferencia de ese modelo, la dispersión compleja produce una mayor incertidumbre, particularmente notoria en el caso del cuantil 0.95-ésimo.

Figura 6.8: Datos simulados y cuantiles de referencia, para tendencia compleja y dispersión compleja.

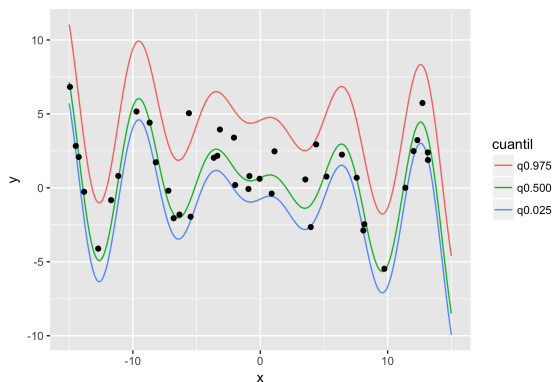
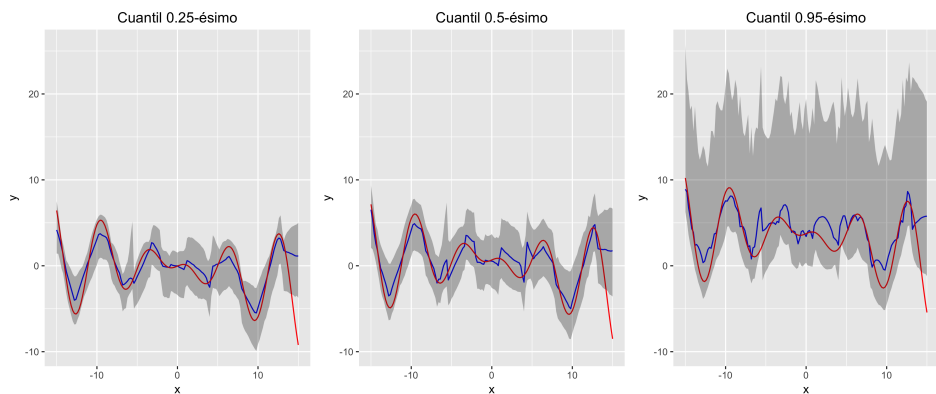


Figura 6.9: Ajuste del modelo *GPDP*, para tendencia compleja y dispersión compleja.



Nota: La línea roja representa el valor real de cada cuantil, la línea azul representa la mediana de la distribución posterior predictiva y el área gris su intervalo de probabilidad al 90 %.

Una vez corroborado que el modelo funciona bien con datos simulados, a continuación se presentan los resultados de su aplicación a un problema real, dentro de las tareas laborales del autor.



## 6.2. Investigación de mercados

### 6.2.1. Conceptos iniciales

En el contexto de la investigación de mercados una de las métricas que se consideran más importantes es la del *conocimiento de marca*, misma que se define como el porcentaje de una cierta población que declara conocer el nombre de la marca en cuestión.

Dentro de dicho ambiente, la teoría dice que esa métrica normalmente depende de la publicidad pautaada semana a semana. Cuando una marca únicamente se publicita en televisión, el valor comúnmente usado para medir esa inversión se denomina *adstocked GRPs*, y, en esencia, se compone de cuántas veces se transmitió el comercial y cuánta gente estaba viendo la televisión cuando se transmitió, ponderado por qué tan lejana en el tiempo sucedió dicha transmisión, respecto al día de hoy. Asimismo, también se suelen usar las inversiones de los competidores para explicar el *conocimiento de marca*, debido a que es común que las personas se confundan y asocien a la marca un comercial del competidor.

En el pasado, inversiones iguales han representado resultados ligeramente distintos en los niveles de *conocimiento de marca*, volatilidad que normalmente es asociada a la calidad de los comerciales, tanto los propios, como los del competidor. En otras palabras, comerciales más memorables han generado un mayor *conocimiento de marca* a la que los pauta, y una aportación muy pequeña al competidor, cuando se han transmitido en aproximadamente la misma cantidad ocasiones y a una similar audiencia.

Además, también es importante revisar el concepto de *marca madre* y *subvariantes*, para nuestra siguiente aplicación del modelo GPDP. Una *marca*

*madre* es aquella que tiene fama por su propio nombre, pero que se ofrece al consumidor mediante productos (también llamados *subvariantes*) que son publicitados por sí solos. Por ejemplo, se puede pensar en la empresa de tecnología *Pera* que tiene comerciales que posicionan su nombre, pero tiene también comerciales donde promociona únicamente el celular que producen, y en otros, únicamente la tableta.

Dicho esto, normalmente se piensa que los comerciales de la *marca madre* contribuyen más al *conocimiento de marca* que aquellos de las *subvariantes*, que tienen como propósito vender los productos específicos, más que posicionar la marca.

### 6.2.2. Caso real

Cierta *marca madre* es cliente de la empresa de investigación de mercados en la que solía trabajar. Dicha marca registró semana a semana los valores de inversión en los comerciales que buscaban posicionar su nombre, los realizados para los de sus *subvariantes*, los de su competencia y el *conocimiento de marca* reportado durante 2014 y 2015. Estos sirvieron para entrenar un modelo GPDP, bajo el supuesto que las mediciones semanales fueron independientes condicionalmente a la inversión. Los valores predichos por el GPDP comparados contra los que en efecto se observaron se presentan a continuación.

Figura 6.10: Modelo de *conocimiento de marca* para datos de entrenamiento (2014-2015).



Nota: El intervalo de credibilidad se construyó usando las estimaciones de la mediana de los cuantiles 0.05-ésimo y 0.95-ésimo.

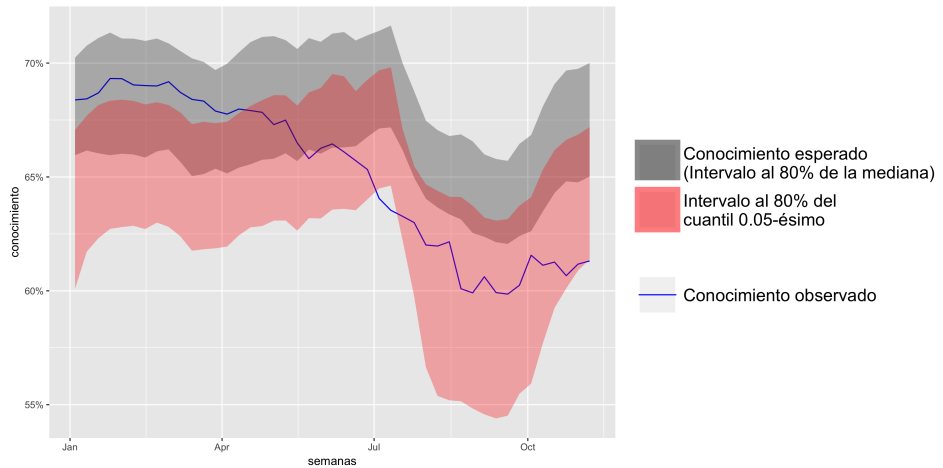
Es verificable que el *conocimiento de marca* sigue un movimiento muy similar a la mediana que predijo el modelo durante el primer año y medio, y, de hecho, en los últimos se ha despegado positivamente.

Todo lo anterior se hizo ignorando el hecho de que también se tenían los datos de 2016, con la intención de ver cómo funcionaría el modelo. Al cliente particularmente le interesaba ver esto porque la métrica tuvo una estrepitosa caída durante el 2016 y tenía la duda si era por una estrategia desafortunada de su inversión o por el hecho de que su competidor había cambiado completamente el concepto de sus comerciales, situación que podría estar provocando que la gente ya no se confundiera y los relacionara erróneamente a los de la marca de nuestro cliente.

Traslado al lenguaje del modelo, se deseaba ver si el valor realmente ob-

servado pudo haber sido predicho por el modelo o si lo consideraba poco probable, situación en la que efectivamente se podría hablar de un cambio estructural ocurrido dentro de este contexto. Los resultados obtenidos fueron los siguientes.

Figura 6.11: *Conocimiento de marca* en 2016, comparado con el modelo GPDP.



Como se puede observar, hasta el mes de abril el *conocimiento de marca* se comportó de acuerdo a lo esperado, pero después tuvo una caída estrepitosa que, si bien el modelo había anticipado para después de julio, coincidió en mayor medida con lo que se hubiera esperado para el cuantil 0.05-ésimo. Es decir, suponiendo que no hubiera cambio estructural, se habría presenciado el peor de cada 20 casos.

En otras palabras, confiando en la construcción del modelo, el supuesto de independencia entre las observaciones y un error modesto en la medición del *conocimiento de marca*, hay información suficiente para pensar que, en efecto, el cambio de concepto en los comerciales del competidor impactó la

métrica del cliente.

## Capítulo 7

# Conclusiones y trabajo futuro

Si bien los modelos de regresión a la media han sido de mucha utilidad en las últimas décadas, principalmente cuando el poder computacional era menor, es importante recalcar que actualmente existen contextos en los que resultan insuficientes, tanto porque se quiere estudiar qué tan factible es un valor atípico, o porque se necesita modelar algún fenómeno asimétrico, por mencionar algunos ejemplos.

De manera similar, la aproximación lineal y la distribución Normal del error han sido fundamentales para que los modelos de regresión hayan proliferado en una gran cantidad de industrias, tanto por su interpretabilidad, como por su bajo costo. Pero es imposible ignorar que únicamente representan un subconjunto del universo infinito de funciones y dispersiones posibles. Crear modelos que permitan una mayor flexibilidad, como los surgidos de los métodos no paramétricos, acercarán más a la Estadística a una

representación certera de la realidad.

Asimismo, utilizar el paradigma Bayesiano para realizar este tipo de modelado tiene la ventaja de poder introducir información de las y los expertos en el fenómeno a estudiar, así como ponderar cuándo fiarse más de los datos y cuándo de dicho conocimiento. Además, de fondo tiene una construcción axiomática, que todo aquel que disfrute de la formalidad en las Matemáticas, valorará.

Un reto importante que presentó este trabajo fue la realización del paquete para implementarlo en R. Esto debido a que se tuvieron que realizar funciones lo suficientemente generales para funcionar con los insumos básicos definidos. Si bien el tiempo que tarda en correr es elevado, es entendible porque la simulación de las cadenas de Markov no se puede hacer en paralelo, y para lograr una mejor estimación, requiere incrementar el número de iteraciones. Por ello el paquete deja ese parámetro a decisión del modelador, de tal manera que él o ella evalúe si prefiere precisión o velocidad.

Si bien estos avances son significativos, aún existe mucho que explorar respecto a lo expuesto en este trabajo. Por ejemplo, el modelo planteado en este trabajo no es capaz de darle un peso distinto a cada variable de entrada, sino que las toma por igual al momento de calcular la distancia entre observaciones. Para mejorar esta situación se podría plantear una descomposición de la función estimada del cuantil en muchos procesos Gaussianos, uno por covariable, lo que brindaría un mayor peso a aquellas covariables que en efecto sean más significativas para explicar el fenómeno en cuestión.

Además, sería conveniente la inclusión de un parámetro de rango que regule dinámicamente la relación entre la distancia y la covarianza entre observaciones. Por ejemplo, aún cuando estén estandarizados los datos, una distancia de 1 podría significar una covarianza grande entre observaciones

para alguna covariable o fenómeno, pero covarianza casi nula para otro. Lograr implementar este parámetro dinámico brindará mayor flexibilidad, y por ende, un mejor ajuste al modelo.



# Bibliografía

- Bannerjee, S. 2008. *Bayesian Linear Models: The Gory Details*. Descargado de <http://www.biostat.umn.edu/ph7440/>.
- Dawid, A. Philip. 2016. Exchangeability and Its Ramifications. *Chap. 2, pages 19–29 of: Damien, Paul, Dellaportas, Petros, Polson, Nicholas, & Stephens, David A. (eds), Bayesian Theory and Applications*. Oxford University Press.
- Denison, David G. T., Holmes, Christopher C., Mallick, Bani K., & Smith, Adrian F. M. 2002. *Bayesian Methods for Nonlinear Classification and Regression*. Wiley Series in Probability. 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 series, 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.
- Paisley, J. 2010. *A Simple Proof of the Stick-Breaking Construction of the Dirichlet Process*. Tech. rept. MIT.
- Rasmussen, C.E., & Williams, C.K.I. 2006. *Gaussian Processes for Machine Learning*. Adaptive Computation and Machine Learning series. University Press Group Limited.
- Robert, Christian P., & Casella, George. 2009. *Introducing Monte Carlo Methods with R*. Berlin, Heidelberg: Springer-Verlag.
- 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. Springer-Verlag.
- Yu, K., & Moyeed, Rana A. 2001. Bayesian Quantile Regression. *Statistics & Probability Letters*, **54**(4), 437–447.

# Apéndice A

## Distribuciones de referencia

### A.1. Distribución t-Student multivariada

**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:

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

### A.2. Distribución Normal condicional

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

$$X = \begin{bmatrix} X_1 \\ X_2 \end{bmatrix},$$

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

$$\mu = \begin{bmatrix} \mu_1 \\ \mu_2 \end{bmatrix},$$

con dimensiones  $\begin{bmatrix} (m-q) \\ q \end{bmatrix}, y$

$$\Sigma = \begin{bmatrix} \Sigma_{11} & \Sigma_{12} \\ \Sigma_{21} & \Sigma_{22} \end{bmatrix},$$

con 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

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

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

### A.3. Distribución de Dirichlet

**Definición 5.** Se dice que un vector aleatorio  $x \in \mathbb{R}^n$  se distribuye de acuerdo a la **distribución de Dirichlet** ( $\mathbf{x} \sim \text{Dir}(\alpha)$ ) con vector de pa-

rámetros  $\alpha$ , específicamente,

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

para los cuales se cumplen las restricciones

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

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

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

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:

$$\mathbb{E}[x_i] = \frac{\alpha_i}{\sum_{k=1}^n \alpha_k}$$

$$Var(x_i) = \frac{\alpha_i (\sum_{k=1}^n \alpha_k - \alpha_i)}{(\sum_{k=1}^n \alpha_k)^2 (\sum_{k=1}^n \alpha_k + 1)}$$

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).

# Apéndice B

## Algoritmos MCMC<sup>1</sup>

### B.1. Introducción

Los algoritmos MCMC son utilizados para aproximar distribuciones de probabilidad, normalmente complejas. La idea es lograr simular una muestra de la distribución, para poder aproximar sus características. Entre más grande sea la muestra, mejor será la estimación.

Para hacer esto simula cadenas de markov de los distintos elementos de la distribución compleja, y, bajo el supuesto de que se alcanza la distribución estacionaria, toma al conjunto de dichas esas simulaciones como una muestra de la distribución original. De hecho, el nombre MCMC viene del inglés *Monte Carlo Markov Chains*, haciendo también referencia a la simulación de Monte Carlo para cada iteración.

---

<sup>1</sup>Las ideas de este apéndice son retomadas de Robert & Casella (2009)



## B.2. Simulador de Gibbs

Se trata de un caso particular de los algoritmos *MCMC*, y a continuación se analizan dos tipos, siendo el segundo una generalización del primero.

### B.2.1. Simulador de Gibbs de dos pasos

Funciona de la siguiente manera: si dos variables aleatorias  $X$  y  $Y$  tienen una densidad conjunta  $f(x, y)$ , con sus correspondientes densidades condicionales  $f_{Y|X}$  y  $f_{X|Y}$ , se genera una cadena de markov  $(X_t, Y_t)$  de acuerdo al siguiente algoritmo:

---

**Algoritmo 1:** Simulador de Gibbs de dos pasos

---

```

Tomar  $X_0 = x_0$  arbitraria ;
para  $t = 1, 2, \dots, n$  hacer
    | 1.  $Y_t \sim f_{Y|X}(y|x_{t-1})$ 
    | 2.  $X_t \sim f_{X|Y}(x|y_t)$ 
fin

```

---

La convergencia de la cadena de markov está asegurada, a menos que los soportes de las condicionales no estén conectados.

### B.2.2. Simulador de Gibbs de múltiples pasos

Sea  $\mathbb{X} \in \mathcal{X}$  una variable aleatoria que puede ser escrita como  $\mathbb{X} = (X_1, \dots, X_p)$ , con  $p \in \mathbb{Z}^+$ , y donde las  $X_i$ 's bien pueden ser unidimensionales o multidimensionales. Además, es posible encontrar las distribuciones condicionales,

de forma que

$$X_i | x_1, \dots, x_{i-1}, x_{i+1}, \dots, x_p \sim f_i(x_i | x_1, \dots, x_{i-1}, x_{i+1}, \dots, x_p),$$

$$i \in \{1, \dots, p\}.$$

El correspondiente algoritmo de Gibbs está dado por:

---

**Algoritmo 2:** Simulador de Gibbs de múltiples pasos

---

Tomar  $\mathbf{x}^{(0)} = (x_1^{(0)}, \dots, x_p^{(0)})$  arbitraria;

**para**  $t = 1, 2, \dots, n$  **hacer**

1.  $X_1^{(t)} \sim f_1(x_1 | x_2^{(t-1)}, \dots, x_p^{(t-1)})$

2.  $X_2^{(t)} \sim f_2(x_2 | x_1^{(t)}, x_3^{(t-1)}, \dots, x_p^{(t-1)})$

...

$k$ .  $X_k^{(t)} \sim f_k(x_k | x_1^{(t)}, \dots, x_{k-1}^{(t)}, x_{k+1}^{(t-1)}, \dots, x_p^{(t-1)})$

...

$p$ .  $X_p^{(t)} \sim f_p(x_p | x_1^{(t)}, \dots, x_{p-1}^{(t)})$

**fin**

---

Cabe resaltar que el desempeño puede estar fuertemente afectado por la parametrización del modelo. Por ello puede resultar una buena idea reparametrizar el modelo, buscando que las componentes sean lo más independientes posible.

## B.3. Monitoreo de convergencia y adaptación de los algoritmos MCMC

### B.3.1. Monitoreo de convergencia a la *estacionariedad*

El primer requisito de convergencia de un algoritmo MCMC es que la distribución de la cadena  $(x^{(t)})$  sea la distribución estacionaria  $f$ . Una meta menos ambiciosa sería que sea independiente del punto inicial  $x^{(0)}$ , después de muchas realizaciones de la cadena. La principal herramienta para verificar *estacionariedad* es correr varias cadenas en paralelo, para poder comparar sus rendimientos.

Un primer acercamiento empírico al control de convergencia es el dibujar gráficas de las cadenas simuladas (componente a componente o juntas), para detectar valores muy desviados y comportamientos no estacionarios.

Otro diagnóstico gráfico que se puede utilizar es la *traza*, es decir, la gráfica de cada uno de los valores de la cadena en el eje  $y$ , contra su respectivo número de iteración en el eje  $x$ . Así será posible observar cuando la cadena tiene un comportamiento repetitivo en ciertos valores y a partir de qué momento se distribuye sobre todo el soporte, es decir, a partir de qué iteración alcanza la distribución estacionaria.

### B.3.2. Monitoreo de convergencia a los promedios

Una vez cubierta la distribución estacionaria, se verifica la convergencia del promedio aritmético

$$\frac{1}{T} \sum_{t=1}^T h(x^{(t)})$$

a la esperanza  $\mathbb{E}_f[h(x)]$ , para una función  $h$  arbitraria. Esta propiedad se denomina comúnmente *ergodicidad*.

La herramienta inicial y más natural suele ser el graficar la evolución del estimador del promedio, conforme crece  $T$ . Si dicha curva no se ha estabilizado después de  $T$  iteraciones, habría que incrementar la longitud de la cadena de markov.

### B.3.3. Monitoreo de convergencia a una muestra *iid*

Para finalizar, idealmente, la aproximación de  $f$  obtenida de los algoritmos MCMC se debería extender a la producción (aproximada) de muestras *iid* de  $f$ . La técnica más usada para lograr esto es el *submuestreo o refinamiento*, donde se consideran sólo los valores  $y^{(t)} = x^{(kt)}$ , para cierta  $k$ .

Como medidas diagnósticas normalmente se usan las siguientes: la autocorrelación dentro de cada variable aleatoria que es parte del simulador de Gibbs; y la correlación cruzada entre las distintas variables aleatorias, dado que se busca independencia entre ellas.