Instituto Tecnológico Autónomo de México

UN MÉTODO BAYESIANO Y NO PARAMÉTRICO DE REGRESIÓN SOBRE CUANTILES, MEDIANTE EL USO DE PROCESOS GAUSSIANOS Y PROCESOS DE DIRICHLET

Tesis

que para obtener el título

Licenciado en Matemáticas Aplicadas

presenta

Carlos Omar Pardo Gómez

Asesor: Dr. Juan Carlos Martínez Ovando

México, D.F. 2017

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

Carlos Omar Pardo Gómez

Fecha

Firma

A Mago.

Agradecimientos

¡Muchas gracias a todos!

Prefacio

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

El capítulo 1 introduce la importancia de las aproximaciones disintas a la *regresión sobre la media*, así como la evolución histórica de este tipo de modelos. El capítulo 2 introduce al Paradigma bayesiano y sus métodos. El capítulo 3 se centra en los Procesos Gaussianos, como herramienta para describir la relación entre una variable dependiente y un conjunto de variables independientes. El capítulo 4 describe a los Procesos de Dirichlet, una distribución de distribuciones no paramétricas, que permiten modelar el error aleatorio. El capítulo 5 explora modelos bayesianos para la *regresión sobre cuantiles*, hasta llegar al modelo central de esta tesis, y su respectiva implementación computacional. El capítulo 6 describe aplicaciones del modelo, así como los resultados obtenidos de evaluarlo en diversos conjuntos de datos. Finalmente, el capítulo 7 hace referencia a las conclusiones finales de esta tesis, además de describir el trabajo futuro que se podría desarrollar, retomando las ideas de ésta.

Índice general

Capítulo 1

Introducción

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

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

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

La segunda es que los supuestos de este tipo de modelos no siempre se cumplen en el mundo real. Por ejemplo, el supuesto de homocedasti-

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

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

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

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

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

Los *modelos de regresión sobre cuantiles* fueron introducidos por Koenker & Bassett (1978), y han permitido concentrarse en valores de interés para los modeladores, sin importar que estén alejados de la me-

dia. 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 un perspectiva distinta y propusieron una aproximación de la verosimilitud para cuantiles.

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

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

Capítulo 2

Paradigma bayesiano[1]

2.1. Axiomas

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

2.2. Inferencia

Un problema clásico de la estadística es el de hacer predicción, utilizando la información de los datos que ya han sido observados. Por ejemplo, es posible pensar que ya se tiene el conjunto de n datos observados $\{y_1, ..., y_n\}$ 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

[1]Las ideas de este capítulo son retomadas de Denison (2002) y Bannerjee (2008).

probabilidad condicional

$$p(y_{n+1}|y_1, ..., y_n) = \frac{p(y_{n+1} \cap \{y_1, ..., y_n\})}{p(y_1, ..., y_n)} = \frac{p(y_1, ..., y_n, y_{n+1})}{p(y_1, ..., y_n)},$$

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

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

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

Siguiendo el mismo razonamiento, es posible obtener que

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

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

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

nados, es importante notar que es posible escribir

$$p(y_{n+1}|y_1, ..., y_n) = \int_\Theta p(y_{n+1}|\theta)p(\theta|y_1, ...y_n)d\theta,$$

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

$$\begin{aligned} p(\theta|y_1, ...y_n) &= \frac{p(y_1, ...y_n|\theta)p(\theta)}{p(y_1, ...y_n)} \\ &= \frac{\left[\prod_{i=1}^n p(y_i|\theta)\right] p(\theta)}{\int_\Theta \left[\prod_{i=1}^n p(y_i|\theta)\right] p(\theta)d\theta}, \end{aligned}$$

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

Cabe observar que el denominador no depende de θ, por lo que normalmente la probabilidad no se expresa como una igualdad, sino con la proporcionalidad

$$p(\theta|y_1, ..., y_n) \propto p(y_1, ..., y_n|\theta)p(\theta),$$

o en general,

$$Posterior \propto Verosimilitud \times Inicial.$$

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

2.3. Regresión lineal

Se piensa un modelo de *regresión a la media* tradicional, donde $y \in \mathbb{R}$ es la variable de respuesta y $x \in \mathbb{R}^n$ es el vector de covariables. La variable $\varepsilon \in \mathbb{R}$ representa el error aleatorio y se distribuye $\varepsilon \sim \mathcal{N}(0, \sigma^2)$, independiente respecto a x. De esta manera, entonces se tiene que:

$$y = \beta^T x + \varepsilon \sim \mathcal{N}(\beta^T x, \sigma^2),$$

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

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

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

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

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

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

Por el Teorema de Bayes,

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

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

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

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

donde M es la media inicial de los coeficientes, $\sigma^2 V$ su varianza, y a y

b son los parámetros iniciales de forma y medida de σ^2.

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

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

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 (β, σ^2) es *Normal - Gamma Inversa*, con parámetros $\mathcal{NGI}(\bar{M}, \bar{V}, \bar{a}, \bar{b})$.

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

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

Particularmente, si se continúa con el modelo conjugado *Normal -*

Gamma Inversa / Normal, es posible encontrar la solución analítica:

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

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

Definición 1. *Sea $X \in \mathbb{R}^p$ un vector aleatorio, con media, mediana y moda μ, matriz de covarianzas Σ, y ν grados de libertad, entonces $X \sim MVSt_\nu(\mu, \Sigma)$ si y sólo si su función de densidad es:*

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

Capítulo 3

Procesos Gaussianos

3.1. Motivación

3.1.1. Modelos de regresión no lineal

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

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

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

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

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

$$y = \beta_0 + \beta_1 \hat{x}_1 + \beta_2 \hat{x}_2 + \beta_3 \hat{x_1} \hat{x_2} + \varepsilon.$$

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

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

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

donde

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

y el procedimiento se convierte en el ya conocido.

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

Así surge la necesidad de buscar un método que permita encontrar

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

3.1.2. Introducción a los Procesos Gaussianos [1]

Para relajar el supuesto de linearidad, se puede pensar que la relación entre la variable de salida y y las covariables x se da mediante cierta función general $f : \mathbb{R}^n \to \mathbb{R}$. De esta forma, el modelo se plantea como:

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

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

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

donde

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

Cabe recordar que la función f es pensada constante, pero desconocida. De nueva cuenta, para reflejar la incertidumbre del modelador, es posible darle una distribución de probabilidad. Pero a diferencia del modelo anterior, ya no existe el parámetro β al cual canalizarle esta incertidumbre, por lo que ahora tendrá que ser sobre toda la función. Antes de continuar, es útil tener presente la siguiente definición.

[1] Las ideas de esta subsección y de lo que resta del capítulo son inspiradas por Rasmussen & Williams (2006).

Definición 2. *Un **proceso gaussiano** ($Y \in \mathbb{R}^m$), es una colección finita de m-variables aleatorias que tienen una distribución gaussiana (normal) conjunta.*

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

Visto de manera matricial y cometiendo un abuso de notación, dada una matriz de covariables $X \in \mathbb{R}^{m \times n}$, $f(X) \in \mathbb{R}^n$ es un vector aleatorio, que además depende de variables de entrada, por lo que $f(X)$ **es un proceso estocástico**. Además, dándole una estructura de covarianza entre los distintos valores de las covariables, $f(X)$ se distribuye Normal Multivariada, donde su vector de medias $M(X)$ y matriz de covarianzas $K(X, X)$ reflejan el conocimiento inicial del modelador.

Observación 1. *De acuerdo a como se acaba de describir el vector $f(X) \in \mathbb{R}^m$, y tomando en cuenta la Definición 2, además de ser un proceso estocástico, $f(X)$ **es un proceso gaussiano**.*

3.2. Inferencia sobre f

3.2.1. Definiciones iniciales

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

Definición 3. *Sean $x, x' \in \mathbb{R}^n$.*

*La **función de medias de f (m_f)** se define como*

$$m_f : \mathbb{R}^n \to \mathbb{R} \mid m_f(x) = \mathbb{E}[f(x)].$$

*La **función de covarianzas de f (k_f)** se define como*

$$k_f : \mathbb{R}^n \times \mathbb{R}^n \to \mathbb{R} \mid k_f(x, x') = \mathbb{E}[(f(x) - m(x))(f(x') - m(x'))].$$

Definición 4. *Sea $X \in \mathbb{R}^m \times \mathbb{R}^n$ y $X' \in \mathbb{R}^p \times \mathbb{R}^n$, es decir,*

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

$$X' = \begin{bmatrix} x_1 \\ ... \\ x_p \end{bmatrix}, x_i \in \mathbb{R}^n, \forall i \in \{1, ..., p\}.$$

*La **función matriz de medias de f (M_f)** se define como*

$$M_f : \mathbb{R}^m \times \mathbb{R}^n : \mathbb{R}^m \mid M_f(X) = \begin{bmatrix} m_f(x_1) \\ ... \\ m_f(x_m) \end{bmatrix}.$$

*La **función matriz de covarianzas de f (K_f)** se define como*

$$K_f : \mathbb{R}^m \times \mathbb{R}^n : \mathbb{R}^m \times \mathbb{R}^m \mid K_f(X, X') = \begin{bmatrix} k_f(x_1, x'_1) & ... & k_f(x_1, x'_p) \\ ... & ... & ... \\ k_f(x_m, x'_1) & ... & k_f(x_m, x'_p) \end{bmatrix}.$$

Dadas estas definiciones, se puede observar que el *proceso gaussiano* $f(X) \in \mathbb{R}^m$ está completamente caracterizado por su función matriz de medias $M_f(X)$ y su función matriz de covarianzas $K_f(X, X')$. Cabe re-

saltar que si se definen estas funciones de manera general para cualquier $X \in \mathbb{R}^{m \times n}$ que esté en el dominio del fenómeno a estudiar, en particular estarán definidas para cualquier matriz de datos observados o datos a predecir. 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 en las funciones recién definidas. Además, cuando se quiera referirse al proceso estocástico $f(X)$ que se distribuye como un *proceso gaussiano*, se hará con la siguiente notación:

$$f(X) \sim \mathcal{GP}(M(X), K(X, X)).$$

3.2.2. Predicción de observaciones sin ruido

Sea un conjunto de observaciones sin ruido, es decir, $\{(x_i, f_i)|i = 1, ..., m\}$, con $f_i = f(x_i)$. En otras palabras, para toda x_i, $y_i = f(x_i)$, sin estar sujeta a un error aleatorio. De forma matricial, se puede escribir como $\{(X, f(X))\}$, con $X \in \mathbb{R}^{m \times n}$ y $f(X) \in \mathbb{R}^m$.

Por otro lado, se tiene un conjunto de covariables $X_* \in \mathbb{R}^{p \times n}$, y se desea predecir $f(X_*)$, suponiendo que sigue la misma función f de los datos observados.

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

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

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

Propiedad 1. *Sea $X \in \mathbb{R}^p$ un vector aleatorio que tiene distribución*

Normal conjunta y está particionado de la siguiente manera:

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

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

$$\mu = \begin{bmatrix} \mu_1 \\ \mu_2 \end{bmatrix}, \text{ dimensiones } \begin{bmatrix} (p-q) \\ q \end{bmatrix},$$
$$\Sigma = \begin{bmatrix} \Sigma_{11} & \Sigma_{12} \\ \Sigma_{21} & \Sigma_{22} \end{bmatrix}, \text{ dimensiones } \begin{bmatrix} (p-q)\times(p-q) & (p-q)\times q \\ q \times (p-q) & q \times q \end{bmatrix}.$$

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

$$\bar{\mu} = \mu_2 + \Sigma_{2,1}\Sigma_{11}^{-1}(a - \mu_1)$$
$$\bar{\Sigma} = \Sigma_{22} - \Sigma_{21}\Sigma_{11}^{-1}\Sigma_{12}$$

De regreso al modelo, tomando en cuenta que existen datos conocidos, es posible condicionar la distribución conjunta, dadas esas observaciones. Utilizando las propiedades de la distribución Normal condicional, se obtiene que:

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

con

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

Observación 2. *$f(X_*)|f(X)$ es una colección finita de p-variables aleatorias que tienen una distribución Normal conjunta, por lo tanto,* $\mathbf{f}(\mathbf{X}_*)|\mathbf{f}(\mathbf{X})$

es un proceso gaussiano.

Para confirmar que no existe ruido en las observaciones, es posible sustituir $X_* = X$ y ver los posible valores para $f(X_*)$.

$$\begin{aligned}\mathbb{E}[f(X_*)|f(X)] &= \bar{M}(X, X_*) \\ &= \bar{M}(X, X) \\ &= M(X) + K(X, X)K(X, X)^{-1}(f(X) - M(X)) \\ &= M(X) + f(X) - M(X) \\ &= f(X),\end{aligned}$$

$$\begin{aligned}Var(f(X_*)|f(X)) &= \bar{K}(X, X_*) \\ &= \bar{K}(X, X) \\ &= K(X, X) - K(X, X)K(X, X)^{-1}K(X, X) \\ &= K(X, X) - K(X, X) \\ &= 0.\end{aligned}$$

Es decir, si $X_* = X$, $f(X_*) \sim \mathcal{N}(f(X), 0)$. En otras palabras, la media es el vector de valores ya obtenidos $f(X)$ y varianza 0, por lo que se cumple que para cualquier X, $f(X)$ tendría siempre un único valor.

3.2.3. Predicción de observaciones con ruido

Ahora se supone un conjunto de observaciones con ruido, es decir, $\{(x_i, y_i)|i = 1, ..., m\}$, con $y_i = f(x_i) + \epsilon$, donde $\epsilon \sim \mathcal{N}(0, \sigma^2)$ y $\sigma^2 > 0$. Con notación matricial se puede describir a este conjunto como $\{(X, Y)\}$, con $X \in \mathbb{R}^{m \times n}$ y $Y \in \mathbb{R}^m$, y el vector de errores aleatorios es $\mathcal{E} \sim \mathcal{N}(0, \sigma^2 I)$. Así, se tiene que

$$Y = f(X) + \mathcal{E}.$$

Es posible observar que al ser suma de dos Normales, la distribución inicial de Y será Normal. Por lo tanto,

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

Observación 3. *Y es una colección finita de m-variables aleatorias que tienen una distribución Normal conjunta, por lo tanto, Y* ***es un proceso gaussiano****.*

A partir de este punto, en esta sección se supondrá a σ^2 **como constante y conocida**, y la atención principal estará sobre la función f.

Ahora se piensa en un conjunto de covariables $X_* \in \mathbb{R}^{p \times n}$, y se busca predecir $f(X_*)$, suponiendo que sigue la misma función f de los datos observados. La distribución inicial conjunta de los datos observados Y y los datos a predecir $f(X_*)$ es:

$$\begin{bmatrix} Y \\ f(X_*) \end{bmatrix} \sim \mathcal{N}\left(\begin{bmatrix} M(X) \\ M(X_*) \end{bmatrix}, \begin{bmatrix} K(X,X) + \sigma^2 I_m & K(X,X_*) \\ K(X_*,X) & K(X_*,X_*) \end{bmatrix} \right),$$

donde I_m es la matrz identidad de dimensión m.

Considerando que ya se cuenta con datos conocidos, se toma la distribución condicional de la Normal y se obtiene que:

$$f(X_*)|Y \sim \mathcal{N}(\bar{M}(X, X_*, \sigma^2), \bar{K}(X, X_*, \sigma^2)),$$

con

$$\bar{M}(X, X_*, \sigma^2) = M(X_*) + K(X_*, X)(K(X,X) + \sigma^2 I_m)^{-1}(Y - M(X))$$
$$\bar{K}(X, X_*, \sigma^2) = K(X_*, X_*) - K(X_*, X)(K(X,X) + \sigma^2 I_m)^{-1} K(X, X_*).$$

Observación 4. *$f(X_*)|Y$ es una colección finita de p-variables aleato-*

rias que tienen una distribución Normal conjunta, por lo tanto, $\mathbf{f}(\mathbf{X}_*)|\mathbf{Y}$ ***es un proceso gaussiano.***

Es posible observar que aunque $X_* = X$, no necesariamente se cumple que $f(X_*)|Y = Y$. En primer lugar, porque $\bar{K}(X, X, \sigma^2) \neq 0$, debido al efecto de σ^2. En segundo lugar, y de nueva cuenta por causa de σ^2, $\bar{M}(X, X, \sigma^2) \neq Y$.

3.3. Varianza

3.3.1. Funciones de covarianza

Hasta el momento, no se han descrito las características de la función de covarianzas de f (k_f). Se empezará por recordar que la **función de covarianzas de f ($\mathbf{k_f}$)** se define como

$$k_f : \mathbb{R}^n \times \mathbb{R}^n \to \mathbb{R} \mid k_f(x, x') = \mathbb{E}[(f(x) - m_f(x))(f(x') - m_f(x'))],$$

donde $f(x)$ es una variable aleatoria que se distribuye Normal con media $m_f(x)$. Por lo tanto, se deduce que

$$Cov(f(x), f(x')) = k_f(x, x').$$

Cabe resaltar que k_f no es una *covarianza* en general, ni cumple con todas las propiedades, sino únicamente describe la covarianza entre dos vectores aleatorios $f(x)$ y $f(x')$, con la misma f, sin la intervención, por ejemplo, de constantes. Para explicar de mejor manera este punto, se da el siguiente ejemplo:

$$\begin{aligned} Cov(af(x) + f(x'), f(x')) &= Cov(af(x), f(x')) + Cov(f(x), f(x')) \\ &= a \times Cov(f(x), f(x')) + Cov(f(x'), f(x')) \\ &= a \times k_f(x, x') + k_f(x', x') \end{aligned}$$

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

$$k_f(x, x') = k_f(x', x) \text{ (simetría)},$$
$$k_f(x, x) = Var(f(x)) \geq 0.$$

Si bien es cierto que dadas esas restricciones hay una variedad muy grande de funciones con las que se puede describir $k_f(x, x')$, por practicidad, y tomando en cuenta que es un supuesto sensato para la mayoría de los casos, es común describir a la función k_f en relación a la distancia entre x y x', $\|x, x'\|_p$. Es decir, $k_f(x, x') = k_f(\|x, x'\|_p)$. 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 $f(x) \approx f(x')$ en la mayoría de los casos, lo que tiene cierto supuesto implícito de que f es una función continua.

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

$$k(x, x') = k(\|x, x'\|_\gamma; \gamma, \lambda) = exp\left(-\frac{1}{\gamma}\left(\frac{\|x, x'\|_\gamma}{\lambda}\right)^\gamma\right),$$

donde λ es un parámetro de rango.

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

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

$$k(x, x') = k(\|x, x'\|_1; \nu, \lambda) = \frac{2^{1-\nu}}{\Gamma(\nu)}\left(\frac{\sqrt{2\nu}\|x, x'\|_1}{\lambda}\right)^\nu K_\nu\left(\frac{\sqrt{2\nu}\|x, x'\|_1}{\lambda}\right)^\nu,$$

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

$$k\left(\|x,x'\|_1;\nu=\frac{3}{2},\lambda\right)=\left(1+\frac{\sqrt{3}\|x,x'\|_1}{\lambda}\right)exp\left(-\frac{\sqrt{3}\|x,x'\|_1}{\lambda}\right),$$

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

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

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

con $\alpha,\lambda>0$.

Existen otro tipo de funciones estacionarias que no guardan una relación inversa entre distancia y covarianza, sino que a cierta distancia aumenta la covarianza, y esto sucede de forma cíclica. En otras palabras, este tipo de funciones capturan un componente **estacional**, normalmente usado en series de tiempo. De esta manera, y siendo t la covariable del tiempo, es posible pensar en una función de la forma

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

donde $\bar{k}$ es alguna de las funciones estacionarias antes mencionadas, δ es la *delta de Kroenecker* y E es el periodo de estacionalidad. Por ejemplo, $E=12$ para una serie mensual.

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

$$k(x,x',t,t';E,\lambda)=\bar{k}(x,x')+exp\left(-\frac{1}{\lambda^2}\frac{E}{\pi}sin^2\left(\frac{\pi}{E}|t'-t|\right)\right).$$

3.3.2. Varianza del error aleatorio

Una vez dicho todo lo anterior, el único pendiente restante es dejar de suponer a σ^2 (la varianza del error aleatorio ε) como una constante conocida. Como ya se mencionó anteriormente, si bien se piensa en ella como constante, es posible reflejar la incertidumbre del modelador respecto al valor verdadero con una distribución de probabilidad. Es decir, si H son las creencias o la información previa con la que se cuenta, entonces

$$\sigma^2 \sim P(\sigma^2|H).$$

Es claro que esta distribución tiene que tener soporte en algún subconjunto de $\mathbb{R}^+$, por lo que la distribución Gamma o Gamma Inversa son las comúnmente utilizadas. Suponiendo que es la primera, el conocimiento de H se tendrá que traducir en los parámetros α y β.

Así, el modelo de Procesos Gaussianos queda especificado como

$$\begin{aligned} y - f(x)|f(x), \sigma^2 &\sim \mathcal{N}(0, \sigma^2) \\ f(x) &\sim \mathcal{GP}(m(x), k(x,x)) \\ \sigma^2 &\sim Gamma(\alpha, \beta). \end{aligned}$$

Capítulo 4

Procesos de Dirichlet[1]

4.1. Motivación

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.

Emplear distribuciones de este tipo permite combatir, por un lado, el *subajuste*, debido a que cualquier distribución se puede representar de manera no paramétrica. Por otro lado, combate al *sobreajuste* utilizando un enfoque bayesiano para calcular la probabilidad posterior, dando como distribución inicial a aquella que se percibe como la más factible.

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

[1] Las ideas de este capítulo son retomadas de Teh (2010).

4.2. Inferencia

4.2.1. Definición formal

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

Definición 5. *Se dice que un vector aleatorio $x \in \mathbb{R}^p$ se distribuye de acuerdo a la* ***distribución de Dirichlet*** *($\mathbf{x} \sim \mathbf{Dir}(\alpha)$) con vector de parámetros α, específicamente,*

$$x = \begin{pmatrix} x_1 \\ \cdots \\ x_p \end{pmatrix}, \qquad \alpha = \begin{pmatrix} \alpha_1 \\ \cdots \\ \alpha_p \end{pmatrix},$$

para los cuales se cumplen las restricciones

$$\begin{aligned} x_i > 0, \forall i \in \{1, ..., p\} \\ \sum_{i=1}^{p} x_i = 1 \\ \alpha_i > 0, \forall i \in \{1, ..., p\}, \end{aligned}$$

si su función de densidad es

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

donde B *es la función Beta multivariada, y puede ser expresada en términos de la función* Γ *como*

$$\mathrm{B}(\alpha) = \frac{\prod_{i=1}^{p} \Gamma(\alpha_i)}{\Gamma\left(\sum_{i=1}^{p} \alpha_i\right)}, \qquad \alpha = (\alpha_1, \cdots, \alpha_p).$$

La esperanza y varianza de cada x_i son los siguientes:

$$\mathbb{E}[x_i] = \frac{\alpha_i}{\sum_{k=1}^{p} \alpha_k}$$
$$Var(x_i) = \frac{\alpha_i \left(\sum_{k=1}^{p} \alpha_k - \alpha_i\right)}{\left(\sum_{k=1}^{p} \alpha_k\right)^2 \left(\left(\sum_{k=1}^{p} \alpha_k + 1\right)\right)}$$

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

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

Definición 6. *Sean G y H dos distribuciones cuyo soporte es el conjunto Θ y sea $\alpha \in \mathbb{R}^+$. Entonces, si se toma una partición finita cualquiera $A_1, ..., A_r$ del conjunto Θ, el vector $(G(A_1), ..., G(A_r))$ es aleatorio, porque G también lo es.*

Se dice que G se distribuye de acuerdo a un ***Proceso de Dirichlet*** *($\mathbf{G} \sim \mathbf{DP}(\alpha, \mathbf{H})$), con distribución media H y parámetro de concentración α, si*

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

para cualquier partición finita $A_1, ..., A_r$ del conjunto Θ.

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

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

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

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 α tiene una relación inversa con la varianza. Así, a una mayor α, corresponde una menor varianza del Proceso de Dirichlet, y, por lo tanto, una mayor concentración respecto a la distribución media H.

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

4.2.2. Distribución posterior

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

distribución de Dirichlet y la distribución Multinomial, se obtiene que

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

Es momento de analizar el parámetro $\alpha H(A_k)+n_k$ de la distribución de Dirichlet, correspondiente a $G(A_k|\phi_1, ..., \phi_n)$, donde $k \in \{1, ..., r\}$. Pero antes, es importante observar que 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.

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

donde

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

Por lo tanto, $G|\phi_1, ..., \phi_n \sim DP(\bar{\alpha}, \bar{H})$. 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 α, y la distribución empírica de los datos, con peso proporcional al número de observaciones n. Por otro lado, el parámetro de concentación posterior es estrictamente más grande que el inicial, por lo que a medida que crecen las n observaciones, la varianza del Proceso de Dirichlet se reduce y el Proceso se concentra alrededor de la distribución empírica.

4.2.3. Distribución predictiva

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

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

es decir,

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

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

4.3. Problemas equivalentes y aplicaciones

En esta sección se revisará la equivalencia entre los Procesos de Dirichlet y otros problemas famosos en la literatura, lo que facilitará el entendimiento de la intiuición que hay detrás, así como la resolución y demostración de algunas propiedades pendientes.

4.3.1. Esquema de urna de Blackwell-MacQueen

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

Se oprime *Play* a la máquina, y se obtiene una pelota, la cual se arroja a la urna *resultados*. A ϕ_1 se le aginará el color de dicha pelota. Posteriormente se añade una pelota de color blanca a la urna *probabilidades* y se pasa a la segunda ronda.

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

Así, después de n rondas, se obtiene la secuencia $\phi_1, ..., \phi_n$. Es importante notar que cada ϕ_{k+1} es una variable aleatoria que depende de las k anteriores, y cuya distribución es

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

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

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

Antes de continuar, es importante repasar una definición.

Definición 7. *Sea $\phi_1, ..., \phi_n$, una secuencia de n variables aleatorias, cuya distribución de probabilidad conjunta está dada por $P(\phi_1, ..., \phi_n)$. Sea ψ una función biyectiva, que va de $\{1, ..., n\} \to \{1, ..., n\}$, es decir, una función que crea una permutación del conjunto $\{1, ..., n\}$. Entonces, se dice que $\phi_1, ..., \phi_n$ es una* ***secuencia aleatoria infinitamente intercambiable*** *si se cumple que*

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

para cualquier permutación ψ.

Regresando al juego de urnas, es importante observar que si bien ϕ_{k+1} es dependiente de las k observaciones anteriores, esta dependencia sólo se da en términos de los valores observados previamente y la frecuencia de dichas observaciones, pero el orden en que hayan sido obtenidos no es relevante. Por lo tanto, es posible afirmar que $\phi_1, ..., \phi_n$ es una secuencia aleatoria infinitamente intercambiable. Dicho esto, es conveniente recordar el ***Teorema de represesentación general de de Finetti***.[2]

Teorema 1. *Sea $\phi_1, ..., \phi_n$ una secuencia aleatoria infinitamente intercambiable de valores reales. Entonces existe una distribución de probabilidad G sobre $\mathcal{F}$, el espacio de todas las distribuciones, de forma que*

[2]Una demostración de este teorema puede ser encontrada en Schervish (1996).

la probabilidad conjunta de $\phi_1, ..., \phi_n$ *se puede expresar como*

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

con

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

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

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

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

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

4.3.2. Proceso estocástico del restaurante chino

Sean $\phi_1, ..., \phi_n$ una secuencia de realizaciones de G, con $G \sim DP(\alpha, H)$. Recordando lo mencionado anteriormente, cada valor obtenido tiene una probabilidad mayor a 0 de ser repetido en una nueva observación.

Sean $\phi_1^*, ..., \phi_m^*$ los m valores únicos observados, y sea n_k^* sea el número de veces que se repite cada valor ϕ_k^*. Entonces, la distribución predictiva se puede reescribir como

$$\phi_{n+1} | \phi_1, ..., \phi_n \sim \left(\frac{\alpha}{\alpha + n} \right) H(\phi_{n+1}) + \left(\frac{n}{\alpha + n} \right) \frac{\sum_{k=1}^{m} n_k^* \delta_{\phi_k^*}(\phi_{n+1})}{n}.$$

A partir de este momento se definirá como el *cluster* Φ_k^* al conjunto cuyos elementos son todos los $\phi_i's$ idénticos que tomen el valor ϕ_k^*. Es inmediato observar que la probabilidad de que una nueva observación ϕ_{n+1} se ubique dentro del *cluster* Φ_k^* es proporcional a n_k^*. Es decir, se da el fenómeno de *los ricos se vuelven más ricos*, ya que entre mayor sea el número de elementos de un *cluster*, mayor será la probabilidad de que una nueva observación sea parte de él.

Así, queda inducida una partición sobre el conjunto $N = \{1, ..., n\}$, debido a que cada uno de dichos números naturales pertenece a un, y sólo un, Φ_k^*. Además, el *cluster* al que pertenecerá cada uno es aleatorio, por lo que la partición inducida también es aleatoria, y encapsula todas las propiedades de los Procesos de Dirichlet.

Para ver cómo sucede esto, únicamente hay invertir el proceso generador. En este caso, primero se obtiene de manera aleatoria una partición del conjunto N en *clusters* $\Phi_1^*, ..., \Phi_m^*$. Después, para cada Φ_k^* se encuentra su respectivo valor mediante una realización de $\phi_k^* \sim H$, y se asigna $\phi_i = \phi_k^*$, para toda $i \in \Phi_k^*$.

La distribución sobre las particiones ha sido nombrada el *Proceso estocástico del restaurante chino*, y es un problema que ha sido estudiado de manera independiente a los Procesos de Dirichlet, siendo descubierta posteriormente su equivalencia. Su nombre lo toma de la siguiente metáfora.

Se supone un restaurante con infinito número de mesas e infinitas sillas en cada una de ellas. El primero consumidor entra y se sienta en la mesa Φ_1^*. El segundo entra y con probabilidad $\frac{1}{\alpha+1}$ se sienta en la misma mesa Φ_1^* del consumidor anterior, mientras que con probabilidad $\frac{\alpha}{\alpha+1}$ se sienta en una nueva mesa Φ_2^*.

En general, después de que han entrado n personas, han sido ocupadas m mesas. Sea n_k^* el número de personas que están sentadas en la mesa Φ_k^*, una nueva persona $n+1$ se sienta en una mesa con las

siguientes probabilidades:

$$P(n+1 \in \Phi^*_{m+1}) = \frac{\alpha}{\alpha+n},$$
$$P(n+1 \in \Phi^*_k) = \frac{n^*_k}{\alpha+n},$$

siendo Φ^*_{m+1} una mesa que aún no ha sido ocupada.

Para conectar esta metáfora con los Procesos de Dirichlet, se podría pensar que todos los integrantes de cada mesa comerán el mismo platillo, mismo que sería elegido aleatoriamente mediante la distribución H, entre un infinito número de platillos.

4.3.3. Proceso estocástico de rompimiento de un palo

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

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

siendo π_k la probabilidad de ocurrencia de ϕ_k.

Dicha probabilidad de ocurrencia será generada con la siguiente metáfora. Se piensa un palo de longitud 1. Se genera una número aleatorio $\beta_1 \sim Beta(1, \alpha)$, 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 $\beta_2 \sim Beta(1, \alpha)$, que representará la proporción del palo restante que le será asignada a

π_2. Es decir, $\pi_2 = \beta_2(1-\beta_1)$. En general, para $k \geq 2$,

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

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

4.3.4. Modelo de mezclas infinitas de Dirichlet

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

$$y_i|\phi_i \sim F(y_i|\phi_i),$$
$$\phi_i|G \sim G(\phi_i),$$
$$G|\alpha, H \sim DP(\alpha, H).$$

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

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

cuando $z_i = \Phi_k^*$. Por lo tanto, el modelo se puede ahora escribir como

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

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

Capítulo 5

Regresión sobre cuantiles

5.1. Motivación

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).

Sea q_p el cuantil p-ésimo de X, es decir, $P(X \leq q_p) = p$. Entonces el cuantil p-ésimo condicional de y, dado x, se supondrá es posible escribirlo como

$$q_p(y|x) = x^T \beta(p),$$

donde $\beta(p)$ es el vector de coeficientes, dependientes de p.

Se define una función de pérdida como

$$\rho_p(u) = u \times (p - I_{(u<0)}),$$

misma que se puede reescribir como

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

o

$$\rho_p(u) = \frac{|u| + (2p-1)u}{2}.$$

Siguiendo este orden de ideas, se puede demostrar que para el problema de minimización

$$\min_{q_p} \mathbb{E}[\rho_p(y_i - q_p)],$$

la solución q_p^* cumple que $P(X \leq q_p^*) = p$.

Así, la primera idea para resolver el problema de *regresión sobre cuantiles* fue resolver el problema de minimización

$$\min_{\beta(p)} \sum_i \rho_p(y_i - x_i^T \beta(p)).$$

Posteriormente, Koenker & Bassett (1978) retomaron esta idea, aplicándola en el paradigma bayesiano.

Definición 8. *Se dice que una variable aleatoria U sigue una distribución asimétrica de Laplace si su función de densidad se escribe como*

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

con μ parámetro de localización y σ parámetro de escala.

Es fácilmente demostrable que el anterior problema de minimización de los errores es equivalente a maximizar la función de verosimilitud formada como producto de densidades independientes asimétricas de Laplace.

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

representar distribuciones más complejas.

5.2. Modelos [1]

Esta sección buscará desarrollar modelos que tomen en cuenta los aprendizajes previos, para realizar análisis de *regresión sobre cuantiles*. A lo largo de ella se supondrá lo que se anuncia a continuación.

Definición 9. *El* ***modelo de regresión del cuantil p-ésimo****, para una variable de respuesta* $y \in \mathbb{R}$*, y un conjunto de covariables* $x \in \mathbb{R}^n$*, será aquel que se pueda escribir como*

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

donde $h : \mathbb{R}^n \to \mathbb{R}$*, y* ε *es el error aleatorio, independiente de* x*, y cuyo cuantil p-ésimo es igual a* 0*. Es decir, si* h_p *es la función de densidad de* ε*, se tiene que*

$$\int_{-\infty}^{0} h_p(\varepsilon) d\epsilon = p.$$

De la definición anterior es posible darse cuenta que el problema de regresión por cuantiles se puede reinterpretar como el problema de encontrar la forma de $f(x)$ y la de $h_p(\varepsilon)$.

Es inmediato darse cuenta que $h_p(\varepsilon)$ tendrá una forma simétrica si, y sólo si, $p = 0.5$. Es decir, sera simétrica únicamente para el modelo de *regresión sobre la mediana*, y asimétrica en cualquier otro caso. Intuitivamente, esto provocará que el modelo espere una proporción de errores negativos similar a p, y de errores positivos similar a $1 - p$, que coincide con lo buscado al realizar un modelo de este tipo.

[1] Los primeros dos son retomados de Kottas & Krnjajic (2005), y el tercero de Kottas *et al.* (2007).

A continuación se plantean diversos modelos que realizan supuestos adicionales acerca de la forma de $f(x)$ y de $h_p(\varepsilon)$, para resolver el problema.

5.2.1. Mezcla asimétrica de densidades de Laplace

En estos primeros modelos se supondrá una forma lineal para f, es decir, f se podrá escribir como $f(x) = x^T\beta$. Por otro lado, se supondrá como forma paramétrica de h_p a la familia de distribuciones asimétricas de Laplace, sin parámetro de localización, y cuya densidad se escribe como

$$w_p^{AL}(\varepsilon, \sigma) = \frac{p(1-p)}{\sigma} exp\left(-\frac{|\varepsilon| + (2p-1)\varepsilon}{2\sigma}\right),$$

con parámetro de escala $\sigma \in \mathbb{R}^+$ y $0 < p < 1$. Dado que es poco común que se conozca el valor de σ, se recurrirá a un modelo de mezclas infinitas de Dirichlet (visto en el capítulo anterior), con parámetro de concentración α, y distribución media H, con soporte en $\mathbb{R}^+$. Se define entonces la función de densidad del error aleatorio como

$$h_p^{AL}(\varepsilon|G) = \int w_p^{AL}(\varepsilon|\sigma) dG(\sigma),$$
$$G \sim DP(\alpha, H).$$

Cabe resaltar que a pesar de la mezcla, se sigue cumpliendo la condición de que $\int_{-\infty}^{0} h_p^{AL}(\varepsilon|G)d\varepsilon = p$, para cualquier G. En cuanto a los parámetros del Proceso de Dirichlet, se tomará una $\alpha > 0$, y H será una Gamma-Inversa, con parámetro de forma $c > 0$, y de razón $d > 0$.

Siguiendo este orden de ideas, se puede reescribir el modelo como

$$\begin{aligned} y - x^T\beta|\beta, \sigma &\sim w_p^{AL}(\varepsilon|\sigma), \\ \beta &\sim \mathcal{N}(\beta|m, v), \\ \sigma|G &\sim G(\sigma), \\ G|\alpha, c, d &\sim DP(\alpha, H(c, d)) \end{aligned}$$

5.2.2. Mezcla de densidades uniformes, con $f(x)$ lineal

Como se puede corroborar en Pavlides & Wellner (2012), para cualquier función de densidad $h(\cdot)$ no creciente y con soporte en R^+, existe una distribución G, tal que

$$h(x|G) = \int \theta^{-1} I_{[0,\theta)}(x) dG(\theta).$$

En otras palabras, $h(\cdot)$ puede ser escrita como una mezcla de densidades uniformes.

Siguiendo este orden de ideas, y suponiendo que $G \sim (\alpha, H)$, dicho resultado puede ser usado para definir la parte positiva de la función de densidad $h_p(\varepsilon)$. Por otro lado, la parte negativa se puede definir de la misma manera usando un signo negativo. Entonces, se tiene que

$$h_p(\varepsilon) = \int \int w_p^{DU}(\varepsilon|\sigma_1, \sigma_2) dG_1(\sigma_1) dG_2(\sigma_2),$$

donde G_1 y G_2 representan a las partes negativas y positivas, respectivamente.

Además, vale la pena recordar que el cuantil p-ésimo tiene que ser igual a 0, por lo que el peso de cada uno de los dos lados estará dado por

$$w_p^{DU}(\varepsilon|\sigma_1, \sigma_2) = \frac{p}{\sigma_1} I_{(-\sigma_1,0)}(\varepsilon) + \frac{(1-p)}{\sigma_2} I_{(0,\sigma_2)}(\varepsilon).$$

De esta manera se tiene un modelo más general que el de Mezcla asimétrica de densidades de Laplace, ya que el único supuesto es que la moda está en 0 y la densidad es no creciente hacia las colas. En resumen, el modelo obtenido es

$$\begin{aligned} y - x^T\beta|\beta,\sigma &\sim w_p^{DU}(\varepsilon|\sigma_1,\sigma_2),\\ \beta &\sim \mathcal{N}(\beta|m,v),\\ \sigma_r|G_r &\sim G_r(\sigma_r),\\ G_r|\alpha_r,c_r,d_r &\sim DP(\alpha_r,H(c_r,d_r)),\\ r &\in \{1,2\}. \end{aligned}$$

5.2.3. Mezcla de densidades uniformes, con Procesos Gaussianos

Hasta ahora, se ha supuesto que $f(x)$ es lineal. Para darle mayor flexibilidad al modelo, ahora se supondrá que $f(x) \sim \mathcal{GP}(m(x), k(x,x))$, con m función de medias dada por el modelador, y k función de covarianzas *exponencial cuadrada*, es decir,

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

donde, a su vez, $\lambda \sim GI(\lambda_\alpha, \lambda_\beta)$.

Añadiendo esta idea al modelo anterior, se obtiene

$$\begin{aligned} y - f(x)|f(x), \sigma &\sim w_p^{DU}(\varepsilon|\sigma_1, \sigma_2), \\ f(x) &\sim \mathcal{GP}(m(x), k(x, x|\lambda)), \\ \lambda &\sim GI(\lambda_\alpha, \lambda_\beta), \\ \sigma_r|G_r &\sim G_r(\sigma_r), \\ G_r|\alpha_r, c_r, d_r &\sim DP(\alpha_r, H(c_r, d_r)), \\ r &\in \{1, 2\}. \end{aligned}$$

Bibliografía

Bannerjee, S. 2008. *Bayesian Linear Models: The Gory Details*. Descargado de http://www.biostat.umn.edu/ ph7440/.

Denison, D.G.T. 2002. *Bayesian Methods for Nonlinear Classification and Regression*. Wiley Series in Probability an. Wiley.

Dunson, D.B., & Taylor, J.A. 2005. Approximate Bayesian inference for quantiles. *Journal of Nonparametric Statistics*, **17**, 385–400.

Fishburn, Peter C. 1986. The Axioms of Subjective Probability. *Statistical Science*, **1**(3), 335–345.

Hanson, T., & Johnson, W.O. 2002. Modeling regression error with a mixture of Polya trees. *Journal of the American Statistical Association*, **97**, 1020–1033.

Hao, L., & Naiman, D.Q. 2007. *Quantile Regression*. Quantile Regression, no. no. 149. SAGE Publications.

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.

Pavlides, Marios G., & Wellner, Jon A. 2012. Nonparametric estimation of multivariate scale mixtures of uniform densities. *Journal of Multivariate Analysis*, **107**, 71–89.

Rasmussen, C.E., & Williams, C.K.I. 2006. *Gaussian Processes for Machine Learning*. Adaptative computation and machine learning series. University Press Group Limited.

Schervish, M.J. 1996. *Theory of Statistics*. Springer Series in Statistics. Springer New York.

Teh, Yee Whye. 2010. Dirichlet Process. *Pages 280–287 of:* Sammut, C, & Webb, GI (eds), *Encyclopedia of Machine Learning*. Springer.

Tsionas, E.G. 2003. Bayesian quantile inference. *Journal of Statistical Computation and Simulation*, **73**, 659–674.

Walker, S.G., & Mallick, B.K. 1999. A bayesian semiparametric accelerated failure time model. *Biometrics*, **55**(2), 477–483.

Yu, K., & Moyeed, Rana A. 2001. Bayesian quantile regression. *Statistics & Probability Letters*, **54**(4), 437–447.