Universidad Autónoma Metropolitana

División de Ciencias Básicas e Ingeniería

Un Modelo Bayesiano para Datos Circulares basado en Árboles de Pólya.

T E S I S

Que para Obtener el Grado de

Maestro en Ciencias Matemáticas Aplicadas e Industriales

Presenta

DANIEL EDUARDO ALLARD OROPEZA

Asesor: Dr. Gabriel Nuñez Antonio

Índice general

Capítulo 1

Introducción

En algunas áreas del conocimiento aparecen mediciones que representan direcciones, es decir, variables direccionales. Este tipo de datos aparecen de manera natural en diversas áreas como las ciencias biológicas, metereológicas y ecológicas. Dichos datos aparecen en estudios donde las observaciones corresponden a direcciones. Algunas aplicaciones de estos datos son, por ejemplo, el análisis de direcciones de viento, direcciones de migración de aves, propagación de fisuras en concreto y otros materiales, orientación de depósitos geológicos, análisis de datos axiales, etc. Para un análisis más detallado, se puede consultar Mardia y Jupp (2000). Los datos direccionales tienen que ver con el análisis de observaciones que son vectores unitarios en el espacio q-dimensional. Los datos direccionales en el plano 2-dimensional se denominan *datos circulares*, mientras que las direcciones en el plano 3-dimensional se denominan *datos esféricos*. Así, los espacios muestrales más comunes son el círculo unitario o la esfera unitaria.

Como lo señaló George Box (1979), *todos los modelos son erróneos, pero algunos son útiles*. Así, una manera de proponer modelos es restringirse a una familia de modelos que pueda ser indexada por un conjunto de parámetros de dimensión finita. Desde el enfoque Bayesiano de la estadística, el análisis de estos modelos incluye una distribución inicial o *a priori* de los parámetros. Sin embargo, como se señala en Müller y Mitra (2013), puede ser peligroso olvidar la simplificación derivada de este proceso. De hecho, en la práctica, pensar que un modelo paramétrico, por muy flexible que sea, pudiera modelar cualquier comportamiento presentado por los datos, puede ser equivocado. En los últimos años ha surgido una corriente Bayesiana que ha trabajado en

el marco teórico para ofrecer una clase más rica y más grande de modelos; lo anterior se logra considerando familias infinitas de modelos de probabilidad, por ejemplo, mediante procesos estocásticos cuyas trayectorias son distribuciones de probabilidad. Las distribuciones iniciales sobre estas familias son conocidas como *distribuciones iniciales Bayesianas no-paramétricas* (BNP, por sus siglas en inglés).

Por ejemplo, considérese el problema de estimar una función de densidad con datos observados $x_1, \ldots, x_n \,|\, F \sim F$. El proceso de inferencia bajo el enfoque Bayesiano requiere una especificación adicional del modelo, con una distribución inicial para la distribución desconocida, F. A menos que F esté restringida a una familia paramétrica de dimensión finita, esto conduce a un modelo BNP, con distribución inicial $p(F)$, que es un modelo de probabilidad para la distribución *infinito-dimensional* F. Para una discusión más exhaustiva de los modelos BNP, se puede consultar, por ejemplo, Hjort *et al.* (2010), Müller y Rodríguez (2013), Walker *et al.* (1999), Müller y Quintana (2004), y Walker (2013).

Aunque existen otras BNP, el Proceso Dirichlet (DP), Ferguson (1973), es, sin duda, la distribución inicial BNP más utilizada. Se denota como $F \sim \mathrm{DP}(\alpha F_0)$ para una distribución inicial DP sobre una medida aleatoria de probabilibad F. El modelo tiene dos parámetros: el parámetro α de masa total y la medida base F_0. La medida base determina la media del proceso, es decir, $\mathrm{E}(F) = F_0$. El parámetro de masa determina, entre otras implicaciones, la incertidumbre sobre F. Una característica del DP es que genera distribuciones de probabilidad discreta casi seguramente y puede escribirse como la suma de puntos de masa θ_j; por lo anterior, en muchas aplicaciones, la naturaleza discreta del DP es restrictiva e inadecuada.

Un proceso más general y menos restrictivo es el *Proceso de Árbol de Pólya* ($\mathcal{PT}$) que, a diferencia del DP, puede construirse de manera que genere, casi seguramente, tanto distribuciones de probabilidad discretas como continuas y absolutamente continuas, con la elección adecuada de parámetros; así que, a diferencia del DP, puede servir como un potencial candidato para especificar distribuciones iniciales sobre familias de funciones de densidad. El proceso genera sucesiones de variables aleatorias intercambiables definidas en un espacio muestral (separable) Ω. De acuerdo con el Teorema de Representación de Finetti, cada una de estas sucesiones es una mezcla de variables aleatorias indepentientes e idénticamente distribuidas respecto a una medida base que puede verse como una distribución inicial en el espacio de medidas de probabilidad sobre Ω. Esta medida fue caracterizada por

Mauldin *et al.* (1992) utilizando la noción de *estrategia*.
Así, el Árbol de Pólya puede identificarse mediante dos características:

- $\Pi = \{B_\epsilon,\, \epsilon = \varepsilon_1 \cdots \varepsilon_m : \varepsilon_j \in \{0,1\},\, m = 0,1,2,\ldots\}$, un conjunto de particiones binarias anidadas de Ω,
- $\mathcal{A} = \{\alpha_\epsilon,\, \epsilon = \varepsilon_1 \cdots \varepsilon_m : \varepsilon_j \in \{0,1\},\, m = 0,1,2,\ldots\}$, un conjunto de parámetros tal que cada α_ϵ está asociado al conjunto B_ϵ.

Para ciertos valores de α_ϵ, Ferguson (1974) demostró que el DP es un caso particular de $\mathcal{PT}$; mientras que Kraft (1964) y Metivier (1971) proporcionaron condiciones suficientes para garantizar que un Árbol de Pólya asigne probabilidad 1 a la clase de las distribuciones de probabilidad absolutamente continuas.

Tanto el DP como el $\mathcal{PT}$ pueden definirse mediante una sucesión de distribuciones condicionales o *modelo jerárquico*. Por ejemplo, para el modelo jerárquico de dos niveles

$$\begin{aligned} X_i &\sim f_i(x \,|\, \theta),\ \ i = 1,\ldots,n,\ \ \theta = (\theta_1,\ldots,\theta_p), \\ \theta_j &\sim \pi_j(\theta \,|\, \gamma),\ \ j = 1,\ldots,p,\ \ \gamma = (\gamma_1,\ldots,\gamma_s), \\ \gamma_k &\sim g(\gamma),\ \ k = 1,\ldots,s, \end{aligned}$$

la distribución conjunta está dada por

$$\prod_{i=1}^{n} f_i(x_i \,|\, \theta) \prod_{j=1}^{p} \pi_j(\theta_j \,|\, \gamma) \prod_{k=1}^{s} g(\gamma_k).$$

Para dicho modelo jerárquico, la distribución posterior conjunta sobre $(\theta,\, \gamma)$ está asociada con las distribuciones condicionales:

$$\begin{aligned} \theta_j &\propto \pi_j(\theta_j \,|\, \gamma) \prod_{i=1}^{n} f_i(x_i \,|\, \theta),\ \ j = 1,\ldots,p, \\ \gamma_k &\propto g(\gamma_k) \prod_{j=1}^{p} \pi_j(\theta_j \,|\, \gamma),\ \ k = 1,\ldots,s. \end{aligned}$$

La simulación de dichas distribuciones es posible, como se verá más adelante, mediante métodos Monte Carlo vía Cadenas de Markov (MCMC).

El objetivo de este trabajo de tesis es proponer un modelo Bayesiano

no-paramétrico que permita obtener la función de densidad de una muestra de ángulos $\psi_1, \ldots, \psi_n$, este modelo está basado en el Proceso de Árbol de Pólya y utiliza como medida base a la proyección radial de una distribución Normal bivariada.

La estructura de la tesis es la siguiente. En el Capítulo 2 se presenta una introducción a los Datos Círculares: su naturaleza, estadística descriptiva y modelos paramétricos. En este capítulo también se hace una breve introducción a la Estadística Bayesiana, se presenta una descripción de los modelos paramétricos y no-paramétricos. El capítulo concluye con la presentación del Proceso de Árbol de Pólya.

En el Capítulo 3 se exponen los resultados más importantes para el modelo Normal Proyectado y se aborda el caso particular en el que la matriz de covarianzas $\boldsymbol{\Lambda} = \boldsymbol{I}$. Para este último caso se detalla la inferencia paramétrica y se expone el modelo no-paramétrico; el capítulo concluye con la presentación del modelo propuesto.

En el Capítulo 4 se ejemplifican casos del modelo propuesto para datos reales y datos circulares, la implementación numérica se realizó en el ambiente `R` (R Core Team).

Finalmente, en el Capítulo de Conclusiones se resumen las ideas y contribuciones principales derivadas de este trabajo. También se mencionan futuros trabajos relacionados con esta tesis.

Capítulo 2

Preliminares

2.1. Datos Circulares

En la primera parte de este capítulo se presenta una introducción a los Datos Círculares: su naturaleza, estadística descriptiva y modelos paramétricos, se introduce el modelo Normal Proyectado. A lo largo de la sección, Ψ denota una variable aleatoria angular y ψ uno de los posibles valores que puede tomar Ψ.

2.1.1. Naturaleza

Los datos direccionales pueden representarse como puntos sobre la esfera unitaria $\mathbb{S}^{q-1} = \left\{\boldsymbol{u} \in \mathbb{R}^q : \boldsymbol{u}^T\boldsymbol{u} = 1\right\}$. En particular, cuando $q = 2$, a los datos direccionales se les denomina *datos circulares*, y se pueden definir empleando las coordenadas polares dadas por $\boldsymbol{u} = (\cos\psi, \sin\psi)^T$.

Desde el punto de vista teórico, existen tres enfoques para el análisis de datos direccionales:

1. Enfoque *embedding*: En este enfoque, $\mathbb{S}^{q-1}$ se reconoce como un subconjunto de $\mathbb{R}^q$.

2. Enfoque *wrapping*: En este enfoque, los vectores tangentes $\boldsymbol{u}$ a la esfera en $\boldsymbol{\mu}$ son "envueltos" sobre la esfera por medio de la transformación

$$\boldsymbol{u} \mapsto (\sin||\boldsymbol{u}||)\boldsymbol{\mu} + (\cos||\boldsymbol{u}||)\boldsymbol{\mu},$$

   donde $\boldsymbol{u}^T\boldsymbol{\mu} = 0$.

3. Enfoque *intrínseco*: En este enfoque, la esfera es reconocida propiamente como un subespacio en su propio contexto.

De manera particular, los datos circulares corresponden a direcciones en dos dimensiones. Dos de las principales formas en las que surgen los datos circulares están asociadas a dos de los instrumentos más importantes de medición circular:

1. La *brújula*. Observaciones típicas medidas por una brújula incluyen direcciones de viento y direcciones de migración de aves.

2. El *reloj*. Observaciones típicas medidas con un reloj corresponden, por ejemplo, a los tiempos de arribo de los pacientes a la unidad de terapia intensiva de un hospital, medidos sobre un reloj de 24 horas.

Los datos circulares pueden representarse mediante coordenadas cartesianas como puntos sobre el círculo unitario, o mediante coordenadas polares mediante vectores unitarios, es decir, en términos de un ángulo, ψ. Debido a lo anterior, la representación gráfica de este tipo de datos es a través de puntos sobre el círculo unitario; esta representación muestra los datos de manera desagrupada y se conoce como *diagrama circular*. Una manera de presentar los datos de manera agrupada es a través del *diagrama de rosa*, que se puede pensar como el análogo circular al histograma para datos en la recta real, y que consiste en reemplazar las barras del histograma por sectores circulares. A manera de ejemplo, la Figura 2.1 muestra las direcciones del viento, medidas cada 15 minutos entre las 3:00 a.m. y las 4:00 a.m., tomadas cada día entre el 29 de enero de 2001 y el 31 de marzo del 2001 en la región denominada "Col de la Roa" en los alpes italianos. Los datos están disponibles en el ambiente `R` (R Core Team) a través de la librería `circular`. En la parte superior izquierda aparece el histograma de los datos (representados de forma lineal), en la parte superior derecha aparece el diagrama circular, en la parte inferior izquierda aparece el diagrama de rosa y, por último, en la parte inferior derecha aparece el diagrama de rosa combinado con el diagrama circular. El ejemplo fue tomado de Pewsey *et al.* (2013).

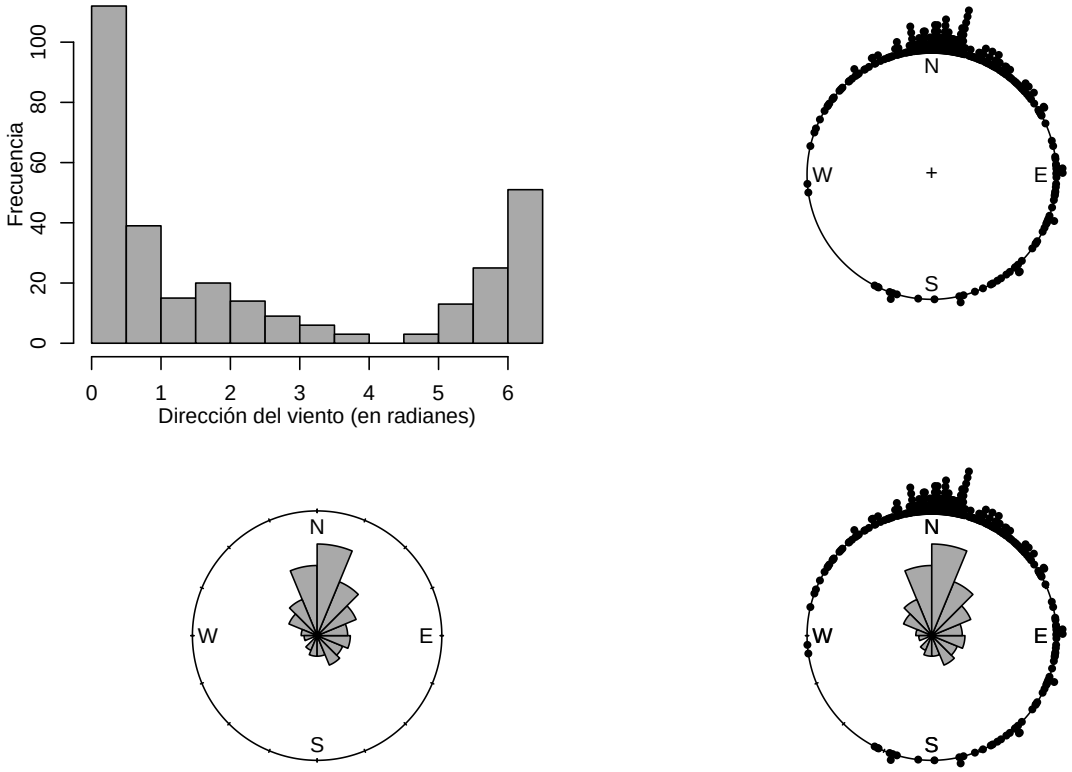

Figura 2.1: Direcciones del viento en la región de "Col de la Roa".

2.1.2. Estadística circular descriptiva

Debido a que el círculo y la recta tienen diferentes topologías, no es posible aplicar las técnicas lineales convencionales dada la periodicidad del círculo. Además, el círculo es una curva cerrada y acotada, mientras que la recta no lo es. Por lo tanto, se pueden anticipar diferencias entre la teoría estadística sobre la recta y sobre el círculo; es necesario definir, por ejemplo, funciones de distribución de probabilidad y medidas descriptivas que tomen en cuenta la topología del espacio muestral.

Un ejemplo de esta situación puede observarse en la Figura 2.2, en donde se tiene un conjunto de datos representados (en grados) por los ángulos 45 y 315. En este caso, la media aritmética resulta ser 180, sin embargo, no parece

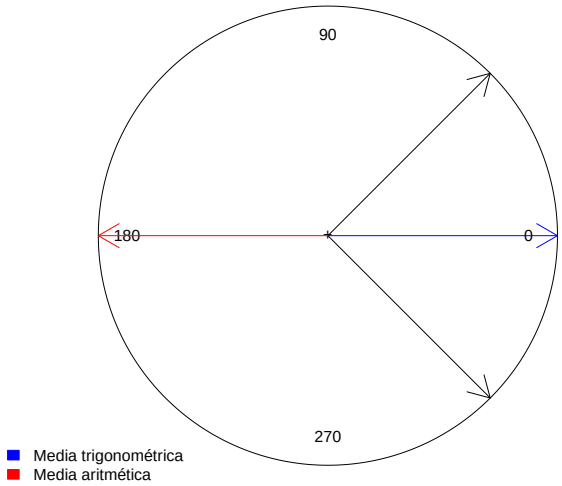

Figura 2.2: Distintas formas de calcular la media de un conjunto de ángulos.

ser una medida descriptiva adecuada; intuitivamente, la dirección 0 podría pensarse como una medida descriptiva más adecuada.

Como se mencionó anteriormente, al ser direcciones en el plano, los datos circulares pueden considerarse como vectores unitarios o como puntos sobre el círculo unitario. Sin embargo, una vez seleccionada una dirección y una orientación inicial, existen otras formas más útiles de considerar a estos datos:

- Como *ángulos*. Cada punto $\boldsymbol{u}$ sobre el círculo unitario se puede representar por un ángulo ψ, es decir,

  $$\boldsymbol{u} = (\cos\psi, \sin\psi).$$

  Esta representación puede verse como un enfoque *intrínseco* debido a que las direcciones son consideradas como puntos sobre el círculo mismo.

- Como *números complejos unitarios*. El punto $\boldsymbol{u}$ sobre el círculo unitario puede representarse mediante el número complejo

  $$z = e^{i\psi} = \cos\psi + i\sin\psi.$$

  Esta representación puede verse como un enfoque *embedding* debido a que las direcciones se consideran como puntos especiales sobre el plano.

Medidas de localización

Definición 2.1.1. *Sean $\boldsymbol{u}_1, \ldots, \boldsymbol{u}_n$ vectores unitarios, donde $\psi_1, \ldots, \psi_n$ son los ángulos asociados a dichos vectores. La* ***dirección media****, μ, de $\psi_1, \ldots, \psi_n$ se define como la dirección de la resultante $\boldsymbol{u}_1 + \cdots + \boldsymbol{u}_n$ de los vectores $\boldsymbol{u}_1, \ldots, \boldsymbol{u}_n$, la cual coincide con el centro de masa, $\bar{\boldsymbol{u}}$, de $\boldsymbol{u}_1, \ldots, \boldsymbol{u}_n$.*

Como las coordenadas cartesianas de $\boldsymbol{u}_i$ son $(\cos\psi_i, \sin\psi_i)$, $i = 1, \ldots, n$, entonces las coordenadas cartesianas del *centro de masa* están dadas por:

$$(\bar{C},\, \bar{S}) = \left(\frac{1}{n}\sum_{i=1}^{n} \cos\psi_i,\; \frac{1}{n}\sum_{i=1}^{n} \sin\psi_i \right). \tag{2.1}$$

Entonces, μ es solución de las ecuaciones

$$\bar{C} = \bar{R}\cos\bar{\psi}, \quad \bar{S} = \bar{R}\sin\bar{\psi}, \tag{2.2}$$

donde $\bar{R} = \left(\bar{C}^2 + \bar{S}^2\right)^{1/2}$.

Cuando $\bar{R} = 0$, μ no está definida. Si $\bar{R} > 0$, μ está dada por

$$\bar{\psi} = \begin{cases} \tan^{-1}(\bar{S}/\bar{C}) & \text{si } \bar{S} > 0,\, \bar{C} > 0, \\ \tan^{-1}(\bar{S}/\bar{C}) + \pi & \text{si } \bar{C} < 0, \end{cases}$$

donde $\tan^{-1}(\cdot) \in [-\pi/2,\, \pi/2]$. En el contexto de la estadística circular, μ *no* significa $(\psi_1 + \cdots + \psi_n)/n$, puesto dicha cantidad no está bien definida.

Definición 2.1.2. *Sean $\boldsymbol{u}_1, \ldots, \boldsymbol{u}_n$ vectores unitarios correspondientes a los ángulos $\psi_1, \ldots, \psi_n$. La* ***longitud de la resultante promedio****, $\bar{R}$, de los vectores $\boldsymbol{u}_1, \ldots, \boldsymbol{u}_n$, es decir, la longitud del centro de masa, $\bar{\boldsymbol{u}}$, está dada por*

$$\bar{R} = \left(\bar{C}^2 + \bar{S}^2\right)^{1/2},$$

donde $\bar{C}$ y $\bar{S}$ están dadas por (2.1).

Adicional a lo anterior, es conveniente definir una versión circular de la mediana muestral.

Definición 2.1.3. *La dirección* ***mediana muestral****,* $\tilde{\psi}$*, de los ángulos* $\psi_1, \ldots, \psi_n$ *es cualquier ángulo* ϕ *que cumple con:*

(i) la mitad de los datos están contenidos en el arco $[\phi, \phi + \pi)$*,*

(ii) la mayoría de los datos se encuentran más cerca de ϕ *que de* $\phi + \pi$*.*

Medidas de concentración y dispersión

La longitud de la resultante promedio $\bar{R}$ fue definida como la longitud del centro de masa $\bar{\boldsymbol{u}}$. Dado que los vectores $\boldsymbol{u}_1, \ldots, \boldsymbol{u}_n$ son unitarios, $0 \leq \bar{R} \leq 1$. Si las direcciones $\psi_1, \ldots, \psi_n$ están muy agrupadas, $\bar{R} \approx 1$; en cambio, si las direcciones están muy dispersas, $\bar{R} \approx 0$. Por lo tanto, $\bar{R}$ es una medida de la *concentración* del conjunto de datos. Nótese que para *cualquier* conjunto de datos de la forma $\psi_1, \ldots, \psi_n, \psi_1 + \pi, \ldots, \psi_n + \pi$, $\bar{R} = 0$, por lo que un valor de $\bar{R} \approx 0$ no implica que las direcciones estén dispersas alrededor del círculo unitario de manera uniforme.

De acuerdo con Mardia y Jupp (2000), $\bar{R}$ es una medida de dispersión más importante que cualquier otra medida. Sin embargo, cuando se desea comparar el análisis de datos sobre el círculo con el análisis sobre la recta real, es necesario considerar medidas de variabilidad similares para datos circulares.

Definición 2.1.4. *Sea* $\psi_1, \ldots, \psi_n$ *un conjunto de ángulos. Se define la* ***varianza circular****,* V*, como*

$$V = 1 - \bar{R}.$$

Debido a que $0 \leq \bar{R} \leq 1$*,* $0 \leq V \leq 1$*.*

Una medida útil de la *distancia* entre dos ángulos ψ y ξ está dada por

$$1 - \cos(\psi - \xi).$$

Así, una manera de medir la dispersión de un conjunto de ángulos $\psi_1, \ldots, \psi_n$ alrededor de cierto ángulo α está dada por

$$D(\alpha) = \frac{1}{n}\sum_{i=1}^{n}\{1 - \cos(\psi_i - \alpha)\}. \tag{2.3}$$

Una propiedad importante es que $D(\alpha)$ alcanza su valor mínimo en $\alpha = \bar{\psi}$ y $D(\bar{\psi}) = V$, lo anterior se puede consultar en Mardia y Jupp (2000) y en Nuñez-Antonio (2010). La propiedad anterior es análoga al caso de la recta real, donde dadas las observaciones $x_1, \ldots, x_n$, la cantidad

$$\frac{1}{n}\sum_{i=1}^{n}(x_i - u)^2$$

alcanza su mínimo en $u = \bar{x}$, y este valor mínimo corresponde a la varianza muestral (con divisor n).

Momentos muestrales

Como se puede observar en la sección anterior, los momentos

$$\bar{C} = \frac{1}{n}\sum_{i=1}^{n}\cos\psi_i, \quad \bar{S} = \frac{1}{n}\sum_{i=1}^{n}\sin\psi_i$$

tienen un papel importante en la definición de la dirección media y la varianza muestral. A partir de estos momentos, se puede definir el primer momento trigonométrico alrededor de la dirección cero.

Definición 2.1.5. *Dado un conjunto de ángulos, $\psi_1, \ldots, \psi_n$, el **p-ésimo momento trigonométrico alrededor de la dirección cero**, m'_p, se define como*

$$m'_p = a_p + ib_p,\ p = 1,\ 2, \ldots,$$

donde

$$a_p = \frac{1}{n}\sum_{i=1}^{n}\cos p\psi_i,\ b_p = \frac{1}{n}\sum_{i=1}^{n}\sin p\psi_i.$$

Es decir,

$$m'_p = \bar{R}_p e^{i\bar{\psi}_p},$$

donde $\bar{\psi}_p$ y $\bar{R}_p$ denotan la dirección media muestral y la longitud de la resultante promedio de los ángulos $p\psi_1, \ldots, p\psi_n$.

De la definición anterior, se puede observar que

$$m_1' = \bar{C} + i\bar{S} = \bar{R}e^{i\bar{\psi}}.$$

De manera análoga se pueden definir los momentos trigonométricos alrededor de la dirección media.

Definición 2.1.6. *Dado un conjunto de ángulos,* $\psi_1, \ldots, \psi_n$, *el* ***p-ésimo momento trigonométrico alrededor de la dirección media*** μ, m_p, *se define como*

$$m_p = \bar{a}_p + i\bar{b}_p,\ p = 1,\ 2, \ldots,$$

donde

$$\bar{a}_p = \frac{1}{n}\sum_{i=1}^{n} \cos p(\psi_i - \bar{\psi}),\ \bar{b}_p = \frac{1}{n}\sum_{i=1}^{n} \sin p(\psi_i - \bar{\psi}).$$

En particular,

$$m_1 = \bar{R}.$$

Las versiones poblacionales de los momentos trigonométricos juegan un papel importante en la teoría de distribuciones sobre el círculo. A continuación se presenta una breve exposición de estos conceptos.

La función característica y los momentos poblacionales

Una manera de especificar una distribución sobre el círculo unitario es a través de la *función de distribución*. La función de distribución, F, es la función de un ángulo aleatorio Ψ definida sobre la recta real, que cumple:

1. $\mathrm{F}(x) = \mathrm{P}(0 < \Psi \leq x),\ 0 \leq x \leq 2\pi$,
2. $\mathrm{F}(x + 2\pi) - \mathrm{F}(x) = 1,\ -\infty \leq x \leq \infty$.

La última condición quiere decir que, sobre el círculo unitario, cualquier arco de longitud 2π tiene probabilidad 1. Más aún, a diferencia de las funciones de distribución sobre la recta real,

$$\lim_{x\to\infty} \mathrm{F}(x) = \infty,\ \lim_{x\to-\infty} \mathrm{F}(x) = -\infty.$$

Dadas las observaciones anteriores, puede definirse la función característica.

Definición 2.1.7. *La* ***función característica****, ψ_p, de un ángulo aleatorio Ψ con respecto a la distribución* F*, se define como la doble sucesión infinita de números complejos $\{\phi_p : p = 0, \pm 1, \ldots\}$ dada por*

$$\psi_p = \mathrm{E}[e^{ip\psi}] = \int_0^{2\pi} e^{ip\psi}\, d\,\mathrm{F}(\psi),\ p = 0, \pm 1, \ldots,$$

donde $\mathrm{F}(\psi)$ es la función de distribución, definida sobre el círculo unitario, $\mathbb{S}$, del ángulo aleatorio Ψ.

Nótese que

$$\psi_p = \alpha_p + i\beta_p,$$

donde

$$\alpha_p = \mathrm{E}[\cos p\psi] = \int_0^{2\pi} \cos p\psi\, d\,\mathrm{F}(\psi)$$

y

$$\beta_p = \mathrm{E}[\sin p\psi] = \int_0^{2\pi} \sin p\psi\, d\,\mathrm{F}(\psi).$$

De esta manera, ϕ_p, α_p y β_p son las versiones poblacionales de los momentos trigonométricos muestrales, m'_p, a_p y b_p, respectivamente.

Definición 2.1.8. *El* ***p-ésimo momento poblacional trigonométrico alrededor de la dirección cero****, ϕ_p, de un ángulo aleatorio Ψ se define como*

$$\phi_p = \alpha_p + i\beta_p,\ \ p = 0, \pm 1, \ldots,$$

donde

$$\alpha_p = \mathrm{E}[\cos p\psi]\ \ y\ \ \beta_p = \mathrm{E}[\sin p\psi].$$

La sucesión $\{(\alpha_p, \beta_p) : p = 0, \pm 1, \ldots\}$ de momentos trigonométricos es equivalente a la función característica de Ψ, ψ_p.

Para $p \geq 0$, escribimos

$$\phi_p = \rho_p e^{i\mu_p}, \;\; \rho_p \geq 0$$

como la versión poblacional de m'_p. Debido a que el caso $p = 1$ es el más utilizado, se escribe

$$\alpha_1 = \alpha, \, \beta_1 = \beta, \, \rho_1 = \rho, \, \mu_1 = \mu.$$

Definición 2.1.9. *El* ***p-ésimo momento poblacional trigonométrico alrededor de la dirección media*** μ, $\bar{\phi}_p$, *se define como*

$$\bar{\phi}_p = \bar{\alpha}_p + i\bar{\beta}_p,$$

donde

$$\bar{\alpha}_p = \mathrm{E}[\cos p(\psi - \mu)], \, \bar{\beta}_p = \mathrm{E}[\sin p(\psi - \mu)].$$

Para el caso $p = 1$, se tiene que $\phi_1 = \rho e^{i\mu}$, donde μ es llamada la *dirección media* y ρ es la *longitud de la resultante media*. Nótese que μ y ρ son las versiones poblacionales de μ y $\bar{R}$, respectivamente.

2.1.3. Modelos paramétricos

Como se señaló anteriormente, los modelos paramétricos de probabilidad para datos direccionales se pueden clasificar en tres grandes categorías.

1. Modelos wrapped

Dada una distribución de probabilidad definida sobre la recta real, esta puede "envolverse" alrededor del círculo unitario. Esto es, si X es una variable aleatoria sobre la recta real, la variable aleatoria X_ω de la correspondiente distribución *wrapped* está dada por

$$X_\omega = X(\text{mód } 2\pi).$$

Si se identifica el círculo unitario con los números complejos de norma 1, la transformación *wrapped*, $X \mapsto X_\omega$, puede escribirse como

$$X \mapsto e^{iX}.$$

Si F es la función de distribución de X, la función de distribución, F_ω, de X_ω está dada por

$$F_\omega(\psi) = \sum_{k=-\infty}^{\infty} \{F(\psi + 2k\pi) - F(2k\pi)\},\ 0 \leq \psi \leq 2\pi.$$

En particular, si f es la función de densidad de X, la función de densidad de X_ω, f_ω, está dada por

$$f_\omega(\psi) = \sum_{k=-\infty}^{\infty} f(\psi + 2k\pi).$$

Algunos de los modelos más importantes dentro de esta familia son la distribución Normal *wrapped*, la distribución Cauchy *wrapped* y la distribución Poisson *wrapped*.

2. Modelos tipo von Mises-Fisher

La construcción de estos modelos considera de manera natural la topología del correspondiente espacio muestral, es decir, estos modelos corresponden al enfoque *embedding* del análisis de datos direccionales. Algunos modelos importantes dentro de esta familia son el modelo *lattice*, el modelo uniforme y el modelo von Mises-Fisher. En el caso de los datos circulares, a la distribución von Mises-Fisher se le conoce como distribución *von Mises*.

Definición 2.1.10. *Se dice que un ángulo aleatorio Ψ tiene una distribución* ***von Mises con dirección media μ y parámetro de concentración*** *κ, si su función de densidad está dada por*

$$\mathrm{M}(\psi \,|\, \mu,\, \kappa) = \frac{1}{2\pi\, \mathrm{I}_0(\kappa)} e^{\kappa \cos(\psi - \mu)},$$

donde I_0 denota la función de Bessel modificada de primer tipo y orden cero, dada por

$$\mathrm{I}_0 = \frac{1}{2\pi}\int_0^{2\pi} e^{\kappa\cos\psi}\, d\psi.$$

Desde el punto de vista de la inferencia estadística es, quizá, la distribución más útil en el análisis de datos circulares. Más aún, es el análogo natural sobre el círculo de la distribución normal sobre la línea real. La distribución von Mises es unimodal y simétrica alrededor de $\psi = \mu$, con moda en $\psi = \mu$.

Mardia y Jupp (2000) muestran que la distribución von Mises puede aproximarse por una distribución normal *wrapped* cuando $\kappa \to \infty$. Por otro lado, Kent (1978) muestra que

$$f_{vM}(\psi \,|\, \mu,\, \kappa) - f_{Nw}(\psi \,|\, \mu,\, A(\kappa)) = \mathrm{O}(\kappa^{-1/2}),\ \ \kappa \to \infty,$$

donde $f_{vM}(\psi \,|\, \mu,\, \kappa)$ y $f_{WN}(\psi \,|\, \mu,\, A(\kappa))$ denotan las densidades de la distribución von Mises, $\mathrm{M}(\psi \,|\, \mu,\, \kappa)$, y la distribución normal *wrapped* que la aproxima, $\mathrm{WN}(\psi \,|\, \mu,\, A(\kappa))$.

3. Modelos Proyectados

Las distribuciones sobre el círculo pueden obtenerse mediante la proyección radial de las distribuciones sobre la recta real. Es decir, dado un vector aleatorio $\boldsymbol{X}$ en $\mathbb{R}$ tal que $\mathrm{P}(\boldsymbol{X} = \boldsymbol{0}) = 0$, entonces $||\boldsymbol{X}||^{-1}\boldsymbol{X}$ corresponde a una variable aleatoria sobre el círculo unitario. El Capítulo 3 está dedicado al Modelo *Normal Proyectado*, que se obtiene cuando el vector aleatorio $\boldsymbol{X}$ sigue una distribución normal bivariada.

2.2. Estadística Bayesiana

En la segunda parte de este capítulo se hace una breve introducción a la Estadística Bayesiana, pues en esta tesis se empleará el enfoque Bayesiano de la Estadística, se presenta una descripción de los modelos paramétricos y no-paramétricos. El capítulo concluye con la presentación del Proceso de Árbol de Pólya. En esta sección, $\boldsymbol{\Theta} \subset \mathbb{R}^p$ denota al espacio de parámetros.

2.2.1. Introducción al Enfoque Bayesiano

A lo largo de la historia se han propuesto modelos para tratar de explicar los fenómenos reales en los que la humanidad se ha visto involucrada, por ejemplo, modelos o leyes físicas del movimiento de los cuerpos y modelos de ecuaciones diferenciales para el crecimiento de poblaciones. Sin embargo, también existen fenómenos que presentan variabilidad, es decir, fenómenos cuyos resultados no siempre son los mismos y las condiciones iniciales sólo determinan de manera probabilística el resultado final. La metodología Bayesiana está basada en la interpretación subjetiva de la probabilidad y tiene como punto central el Teorema de Bayes.

En algunos textos como Bernardo y Smith (1994), Hoff (2009) y Koch (2007), se presenta la filosofía bayesiana tomando como punto de partida la *Teoría de la Decisión*, mostrando la dualidad entre los conceptos de probabilidad y utilidad, demostrando que la maximización de la utilidad esperada es el único criterio de decisión compatible bajo un sistema axiomático.

De acuerdo con Gutiérrez-Peña y Erderly (2006), en el *enfoque subjetivo* de la probabilidad, la probabilidad de un evento A es una medida del grado de creencia que tiene un individuo en la ocurrencia de A con base en la información K que dicho individuo posee. Bajo este enfoque, toda probabilidad es condicional en la información de la cual se dispone. Este enfoque de la probabilidad es ampliamente aprovechado por la metodología Bayesiana y es por ello que se puede decir que la estadística Bayesiana va más allá que la estadística frecuentista, pues busca aprovechar toda la información disponible.

En el enfoque Bayesiano de la estadística, la incertidumbre presente en un modelo de probabilidad, $p(x \mid \boldsymbol{\theta})$, es representada a través de lo que se conoce como una *distribución inicial*, $p(\boldsymbol{\theta})$, sobre los posibles valores del parámetro desconocido $\boldsymbol{\theta}$ (típicamente multidimiensional), que define al modelo. Así, desde la perspectiva Bayesiana de la estadística, un modelo queda especificado por una verosimilitud y una distribución inicial, es decir, por el conjunto $\{p(x \mid \boldsymbol{\theta}),\, p(\boldsymbol{\theta})\}$. El Teorema de Bayes,

$$\begin{aligned} p(\boldsymbol{\theta} \mid \boldsymbol{x}) &= \frac{p(\boldsymbol{x} \mid \boldsymbol{\theta}) p(\boldsymbol{\theta})}{p(\boldsymbol{x})} \\ &= \frac{p(\boldsymbol{x} \mid \boldsymbol{\theta}) p(\boldsymbol{\theta})}{\displaystyle\int_{\Theta} p(\boldsymbol{x} \mid \tilde{\boldsymbol{\theta}}) p(\tilde{\boldsymbol{\theta}})\, d\tilde{\boldsymbol{\theta}}} \\ &\propto p(\boldsymbol{x} \mid \boldsymbol{\theta}) p(\boldsymbol{\theta}) \end{aligned}$$

permite entonces incorporar la información contenida en un conjunto de datos $\boldsymbol{x} = (x_1, \ldots, x_n)$, produciendo una descripción actualizada de la incertidumbre sobre los valores de los parámetros del modelo, a través de la *distribución final*, $p(\boldsymbol{\theta} \mid \boldsymbol{x})$.

De acuerdo con Wasserman (2010), la inferencia Bayesiana, en general, se lleva a cabo de la siguiente manera:

1. Se elige un modelo estadístico $p(x \mid \boldsymbol{\theta})$ que refleje nuestra creencia respecto a x, dado $\boldsymbol{\theta}$.

2. Se determina una distribución inicial $p(\boldsymbol{\theta})$ que exprese nuestra creencia respecto a $\boldsymbol{\theta}$, antes de observar los datos.

3. Una vez observados los datos $\boldsymbol{x} = (x_1, \ldots, x_n)$, se actualiza nuestra creencia y se calcula la distribución final $p(\boldsymbol{\theta} \mid \boldsymbol{x})$.

La distribución final, $p(\boldsymbol{\theta} \mid \boldsymbol{x})$, nos permite construir medidas descriptivas tales como la esperanza *a posteriori* o la varianza *a posteriori*, dadas por:

$$\begin{aligned} \mathrm{E}(\boldsymbol{\theta} \mid \boldsymbol{x}) &= \int_{\Theta} \tilde{\boldsymbol{\theta}}\, p(\tilde{\boldsymbol{\theta}} \mid \boldsymbol{x})\, d\tilde{\boldsymbol{\theta}}, \\ \mathrm{Var}(\boldsymbol{\theta} \mid \boldsymbol{x}) &= \int_{\Theta} (\tilde{\boldsymbol{\theta}} - \mathrm{E}(\tilde{\boldsymbol{\theta}} \mid \boldsymbol{x}))^2\, p(\tilde{\boldsymbol{\theta}} \mid \boldsymbol{x})\, d\tilde{\boldsymbol{\theta}}. \end{aligned}$$

Una manera formal de obtener estimadores puntuales de $\boldsymbol{\theta}$ es mediante la *Teoría de la Decisión* (ver Wasserman (2010)), la cual es una teoría formal que permite comparar procedimientos estadísticos. Sea $\boldsymbol{\theta} \in \boldsymbol{\Theta}$ y $\hat{\boldsymbol{\theta}}$ un estimador de $\boldsymbol{\theta}$, en el lenguaje de la Teoría de la Decisión, a dicho estimador se le conoce como **regla de decisión** y al conjunto de valores posibles de la regla de decisión se le conoce como **acciones**.

La magnitud de la diferencia entre $\boldsymbol{\theta}$ y $\hat{\boldsymbol{\theta}}$ puede medirse mediante una **función de pérdida** $L(\boldsymbol{\theta}, \hat{\boldsymbol{\theta}})$, donde $L : \boldsymbol{\Theta} \times \boldsymbol{\Theta} \to \mathbb{R}$. Algunos ejemplos de funciones de pérdida,para $\boldsymbol{\Theta} \subset \mathbb{R}$, son:

$$\begin{aligned} L(\boldsymbol{\theta}, \hat{\boldsymbol{\theta}}) &= (\boldsymbol{\theta} - \hat{\boldsymbol{\theta}})^2, \text{ pérdida de error cuadrático,} \\ L(\boldsymbol{\theta}, \hat{\boldsymbol{\theta}}) &= |\boldsymbol{\theta} - \hat{\boldsymbol{\theta}}|, \text{ pérdida de error absoluto,} \\ L(\boldsymbol{\theta}, \hat{\boldsymbol{\theta}}) &= |\boldsymbol{\theta} - \hat{\boldsymbol{\theta}}|^p, \text{ pérdida } L_p \end{aligned}$$

Debido a que el estimador es una función de los datos, también suele denotarse por $\hat{\boldsymbol{\theta}}(\boldsymbol{X})$. Para evaluar un estimador, se analiza la *pérdida promedio* o *riesgo*.

Definición 2.2.1. *El **riesgo** de un estimador $\hat{\boldsymbol{\theta}}$ se define como:*

$$R(\boldsymbol{\theta}, \hat{\boldsymbol{\theta}}) = \mathrm{E}_{\boldsymbol{\theta}}(L(\boldsymbol{\theta}, \hat{\boldsymbol{\theta}})) = \int L(\boldsymbol{\theta}, \hat{\boldsymbol{\theta}}(x))\, p(x \mid \boldsymbol{\theta})\, dx.$$

Por lo tanto, para comparar dos estimadores, basta con comparar sus funciones de riesgo. Dicha comparación se puede hacer mediante alguna medida descriptiva de la función de riesgo, dos ejemplos de estas medidas son el *riesgo máximo* y el riesgo de Bayes.

Definición 2.2.2. *El **riesgo máximo** se define como:*

$$\bar{R}(\hat{\boldsymbol{\theta}}) = \sup_{\theta} R(\boldsymbol{\theta}, \hat{\boldsymbol{\theta}}).$$

*El **riesgo de Bayes** se define como:*

$$r(p, \hat{\boldsymbol{\theta}}) = \int R(\boldsymbol{\theta}, \hat{\boldsymbol{\theta}})\, p(\boldsymbol{\theta})\, d\boldsymbol{\theta},$$

en donde $p(\boldsymbol{\theta})$ es una distribución inicial para $\boldsymbol{\theta}$.

De la definición anterior, se puede concluir que una manera de elegir un estimador puntual es tomando aquel estimador que minimice el riesgo de Bayes, lo cual tiene como consecuencia el siguiente teorema.

Teorema 2.2.3. *Dada la función de pérdida* $L(\boldsymbol{\theta}, \hat{\boldsymbol{\theta}}) = (\boldsymbol{\theta} - \hat{\boldsymbol{\theta}})^2$*, entonces el estimador Bayesiano es*

$$\hat{\boldsymbol{\theta}}(x) = \int \boldsymbol{\theta}\, p(\boldsymbol{\theta} \,|\, x)\, d\boldsymbol{\theta} = \mathrm{E}(\boldsymbol{\theta} \,|\, X = x).$$

Si $L(\boldsymbol{\theta}, \hat{\boldsymbol{\theta}}) = |\boldsymbol{\theta} - \hat{\boldsymbol{\theta}}|$*, entonces el estimador Bayesiano es la mediana de la distribución posterior* $p(\boldsymbol{\theta} \,|\, x)$*.*

En algunos casos, además de un interés respecto a las características de $\boldsymbol{\theta}$, también se desea describir el comportamiento de observaciones futuras del fenómeno aleatorio que se está estudiando, es decir, hacer predicción. Dicho problema puede ser atacado desde el enfoque Bayesiano: el modelo $p(x \,|\, \boldsymbol{\theta})$ y la distribución inicial $p(\boldsymbol{\theta})$ inducen una distribución conjunta para el vector aleatorio (X, Θ) mediante la probabilidad condicional, pues

$$p(x, \boldsymbol{\theta}) = p(x \,|\, \boldsymbol{\theta}) p(\boldsymbol{\theta}).$$

Así, al obtener la densidad marginal, se tiene que

$$\begin{aligned} p(x) &= \int_{\Theta} p(x, \tilde{\boldsymbol{\theta}})\, d\tilde{\boldsymbol{\theta}} \\ &= \int_{\Theta} p(x \,|\, \tilde{\boldsymbol{\theta}}) p(\tilde{\boldsymbol{\theta}})\, d\tilde{\boldsymbol{\theta}}. \end{aligned} \tag{2.4}$$

A la ecuación (2.4) se le conoce como **distribución predictiva inicial** y describe el conocimiento acerca de una observación futura X, basado únicamente en la información contenida en $p(\boldsymbol{\theta})$; dicha expresión no depende de $\boldsymbol{\theta}$.

Después de observar los datos, el modelo $p(x \,|\, \boldsymbol{\theta})$ y la distribución final $p(\boldsymbol{\theta} \,|\, \boldsymbol{x})$ inducen una distribución conjunta para el vector aleatorio (X, Θ), condicional dados los datos $\boldsymbol{x} = \{x_1, \ldots, x_n\}$, pues

$$\begin{aligned} p(x, \boldsymbol{\theta} \,|\, \boldsymbol{x}) &= \frac{p(x, \boldsymbol{\theta}, \boldsymbol{x})}{p(\boldsymbol{x})} \\ &= \frac{p(x \,|\, \boldsymbol{\theta}, \boldsymbol{x})}{p(\boldsymbol{x})} \\ &= p(x \,|\, \boldsymbol{\theta}, \boldsymbol{x}) p(\boldsymbol{\theta} \,|\, \boldsymbol{x}) \\ &= p(x \,|\, \boldsymbol{\theta}) p(\boldsymbol{\theta} \,|\, \boldsymbol{x}), \end{aligned}$$

en donde $p(x \,|\, \boldsymbol{\theta}, \boldsymbol{x}) = p(x \,|\, \boldsymbol{\theta})$ se sigue de la independencia condicional de X y $\boldsymbol{X} = (X_1, \ldots, X_n)$ dado Θ. Al marginalizar, se obtiene

$$\begin{aligned} p(x \,|\, \boldsymbol{x}) &= \int_{\Theta} p(x, \tilde{\boldsymbol{\theta}} \,|\, \boldsymbol{x}) \, d\tilde{\boldsymbol{\theta}} \\ &= \int_{\Theta} p(x \,|\, \tilde{\boldsymbol{\theta}}) p(\tilde{\boldsymbol{\theta}} \,|\, \boldsymbol{x}) \, d\tilde{\boldsymbol{\theta}}. \end{aligned} \tag{2.5}$$

A la ecuación (2.5) se le conoce como **distribución predictiva final** y describe el conocimiento acerca de una observación futura X basado tanto en la información contenida en $p(\boldsymbol{\theta})$ como en la información muestral $\boldsymbol{x} = \{x_1, \ldots, x_n\}$. Dicha distribución no depende de $\boldsymbol{\theta}$.

Definición 2.2.4. *Sea $p(x_1, \ldots, x_n)$ la densidad conjunta de las variables aleatorias $X_1, \ldots, X_n$. Decimos que $X_1, \ldots, X_n$ son* ***intercambiables*** *si $p(x_1, \ldots, x_n) = p(x_{\pi(1)}, \ldots, x_{\pi(n)})$, para todas las permutaciones π de $\{1, \ldots, n\}$.*

Se puede ver que si $\boldsymbol{\theta} \sim p(\boldsymbol{\theta})$ y $X_1, \ldots, X_n$ son i.i.d condicionalmente, dado $\boldsymbol{\theta}$, entonces, marginalmente, $X_1, \ldots, X_n$ son intercambiables, pues

$$\begin{aligned} p(x_1, \ldots, x_n) &= \int_\Theta p(x_1, \ldots, x_n \,|\, \tilde{\boldsymbol{\theta}}) p(\tilde{\boldsymbol{\theta}})\, d\tilde{\boldsymbol{\theta}} \\ &= \int_\Theta \left\{ \prod_{i=1}^n p(x_i \,|\, \tilde{\boldsymbol{\theta}}) \right\} d\tilde{\boldsymbol{\theta}} \\ &= \int_\Theta \left\{ \prod_{i=1}^n p(x_{\pi(i)} \,|\, \tilde{\boldsymbol{\theta}}) \right\} d\tilde{\boldsymbol{\theta}} \\ &= p(x_{\pi(1)}, \ldots, x_{\pi(n)}). \end{aligned}$$

A continuación, el *Teorema de Representación de de Finetti* nos dice que la afirmación anterior también es cierta en el sentido inverso (véase Hoff (2009) y Bernardo y Smith (1994)).

Teorema 2.2.5. *Sea $X_1, X_2, \ldots$ una sucesión infinita de variables aleatorias. Supóngase que, para cualquier $n \in \mathbb{N}$, la sucesión $X_1, \ldots, X_n$ es intercambiable. Entonces, se tiene que*

$$p(x_1, \ldots, x_n) = \int_\Theta \left\{ \prod_{i=1}^n p(x_i \,|\, \boldsymbol{\theta}) \right\} d\boldsymbol{\theta},$$

para algún parámetro $\boldsymbol{\theta}$, alguna distribución inicial para $\boldsymbol{\theta}$ y algún modelo $p(x) \,|\, \boldsymbol{\theta})$.

Se verá ahora que existen algunos casos donde la distribución final se puede encontrar de manera relativamente sencilla y sin necesidad de calcular integrales.

Definición 2.2.6. *Una clase $\mathcal{P}$ de distribuciones iniciales para $\boldsymbol{\theta}$ se llama* ***conjugada*** *respecto al modelo $p(\boldsymbol{x} \,|\, \boldsymbol{\theta})$ si*

$$p(\boldsymbol{\theta}) \in \mathcal{P} \Rightarrow p(\boldsymbol{\theta} \,|\, \boldsymbol{x}) \in \mathcal{P}.$$

Ejemplo. Sea $X \sim \text{Bin}(n, \theta)$, donde

$$\mathrm{P}(X = x \,|\, \theta) = \binom{n}{x}\theta^x(1-\theta)^{n-x},\ x \in \{1, \ldots, n\}.$$

Una vez que se han observado los datos $\boldsymbol{x} = x_1, \ldots, x_n$, se quiere obtener una distribución inicial para θ. Una opción es $\theta \sim \text{Beta}(a, b)$, pues esta variable aleatoria toma sus valores en el intervalo $[0, 1]$. Recordemos que una variable aleatoria Beta tiene función de densidad dada por

$$f(\theta \,|\, a, b) = \frac{\Gamma(a+b)}{\Gamma(a)\Gamma(b)}\theta^{a-1}(1-\theta)^{b-1}.$$

Por lo tanto, se tiene que

$$\begin{aligned} p(\theta \,|\, \boldsymbol{x}) &\propto \left[\prod_{i=1}^{n}\binom{n}{x_i}\theta^{x_i}(1-\theta)^{n-x_i}\right]\frac{\Gamma(a+b)}{\Gamma(a)\Gamma(b)}\theta^{a-1}(1-\theta)^{b-1} \\ &\propto \theta^{\left\{a+\sum_{i=1}^{n}x_i-1\right\}}(1-\theta)^{\left\{b+n-\sum_{i=1}^{n}x_i-1\right\}}. \end{aligned}$$

Por lo tanto, como $p(\theta \,|\, \boldsymbol{x})$ es proporcional al kernel de una distribución Beta con parámetros $(a + \sum_{i=1}^{n} x_i,\ b + n - \sum_{i=1}^{n} x_i)$, se tiene que $p(\theta \,|\, \boldsymbol{x}) \sim \text{Beta}\,(a + \sum_{i=1}^{n} x_i,\ b + n - \sum_{i=1}^{n} x_i)$, es decir, la clase de distribuciones iniciales Beta es conjugada para el modelo Binomial.
□

Hoff (2009) proporciona expresiones para encontrar distribuciones iniciales conjugadas para modelos de la familia exponencial.

2.2.2. Métodos MCMC

La implementación de las técnicas Bayesianas usualmente requiere de un esfuerzo computacional muy alto. La mayor parte de este esfuerzo se concentra en el cálculo de ciertas características de la distribución final del parámetro de interés (denominados comúnmente resúmenes inferenciales). Así, por ejemplo, para pasar de una distribución conjunta a una colección de distribuciones y momentos marginales que sean útiles para hacer inferencias sobre subconjuntos de parámetros, se requiere integrar. En la mayoría de los casos los resúmenes inferenciales básicos se reducen a integrales de la forma

$$S_I\{g(\boldsymbol{\theta})\} = \int g(\boldsymbol{\theta})p(\boldsymbol{x} \,|\, \boldsymbol{\theta})p(\boldsymbol{\theta})\, d\boldsymbol{\theta}_{I^c},$$

donde $g : \mathbb{R}^d \to \mathbb{R}$, $\boldsymbol{\theta} = (\theta_1, \ldots, \theta_d)$, $I \subseteq \{1, \ldots, d\}$, $I^c = \{1, \ldots, d\} \setminus I$ y $\boldsymbol{\theta}_{I^c} = \{\theta_i : i \in I^c\}$.

Se vio que en el caso en el que se tiene una clase de distribuciones iniciales conjugadas para un modelo, el cálculo de la distribución posterior no requiere de integrales complicadas; en otros casos es posible realizar una evaluación explícita de dichas integrales. Sin embargo, en muchos casos la evaluación de las integrales es muy complicada o no existe una forma cerrada para resolverlas, por lo que es necesario el uso de aproximaciones numéricas o de técnicas de simulación como, por ejemplo, los métodods *Monte Carlo vía Cadenas de Markov* (MCMC, por sus siglas en inglés). Además de ayudarnos a implementar una solución numérica, dichas técnicas nos permiten obtener muestras de alguna distribución final.

El material de esta sección está basado en los textos de Gamerman y Lopes (2006), Gelman *et al.* (1995), Gilks *et al.* (1996), Hoff (2009), Rincón (2012), Robert y Casella (2004 y 2010) y Wasserman (2010).

Integración Monte Carlo

Supóngase que se desea evaluar una integral del estilo

$$I = \int_a^b h(x)\, dx$$

para alguna función h. En el caso en el que h sea, por ejemplo, una función trigonométrica o un polinomio, la integral puede resolverse de forma cerrada. La idea básica del método Monte Carlo consiste en escribir la integral I como el valor esperado de la función h con respecto a alguna distribución de probabilidad, es decir, dada una variable aleatoria X cuyo soporte contiene al intervalo (a, b), se escribe

$$I = \mathrm{E}_f[w(X)] = \int_a^b w(x) f(x)\, dx,$$

con $w(x) = h(x)/f(x)$, con $f(x) > 0$ para $x \in [a, b]$, y se genera una muestra $X_1, \ldots, X_N \sim f$. Por la Ley de los Grandes Números, se tiene

$$\hat{I} \equiv \frac{1}{N} \sum_{i=1}^{N} w(X_i) \xrightarrow{P} \mathrm{E}_f[w(X)] = I.$$

Cuando $\mathrm{E}_f[w^2(X)] < \infty$, la varianza asintótica de la aproximación está dada por

$$\mathrm{Var}(\hat{I}) = \frac{1}{N}\int_a^b (w(x) - \mathrm{E}_f[w(X)])^2\, dx,$$

la cual también puede aproximarse a partir de la muestra $X_1, \ldots, X_N \sim f$ como

$$v_N = \frac{1}{N^2}\sum_{i=1}^{N}\left[w(x_i) - \hat{I}\right]^2.$$

Muestreo por Importancia

Este método recibe su nombre porque se apoya en las llamadas *funciones de importancia*, las cuales se utilizan en lugar de las funciones originales. Considérese de nuevo la integral $I = \int w(x)f(x)\,dx$, con w y f como en la sección anterior. El método Monte Carlo básico consiste en generar una muestra $X_1, \ldots, X_N \sim f$ para evaluar la integral. Sin embargo, existen casos en los que no sabemos cómo generar una muestra de f. Por ejemplo, en el cálculo de la distribución posterior $p(\boldsymbol{\theta} \mid \boldsymbol{x})$, no hay garantía de que se trate de una distribución conocida.
El Muestreo por Importancia es una generalización del método Monte Carlo básico que nos permite resolver este problema. Sea g una densidad de probabilidad de la cual sabemos cómo generar fácilmente muestras. Entonces,

$$I = \int w(x)f(x)\,dx = \int \frac{w(x)f(x)}{g(x)}g(x)\,dx = \mathrm{E}_g\left[\frac{w(X)f(X)}{g(X)}\right].$$

Se puede generar una muestra $X_1, \ldots, X_N \sim g$ y estimar I mediante la expresión

$$\hat{I} = \frac{1}{N}\sum_{i=1}^{N}\frac{w(X_i)f(X_i)}{g(X_i)}.$$

Nuevamente, por la Ley de los Grandes Números, $\hat{I} \xrightarrow{P} I$. La varianza del estimador $\hat{I}$ depende de la selección de g, lo cual puede ocasionar que $\hat{I}$ tenga un error estándar infinito, pues al calcular el segundo momento de $w(X_i)f(X_i)/g(X_i)$, se tiene

$$\mathrm{E}_g\left[\left(\frac{w(X)f(X)}{g(X)}\right)^2\right] = \int\left(\frac{w(x)f(x)}{g(x)}\right)^2 g(x)\,dx = \int\frac{w^2(x)f^2(x)}{g(x)}\,dx,$$

cuyo valor puede ser infinito si g tiene colas más ligeras que f. Para evitar esto, se recomienda que g

- Sea fácil de simular;
- Tenga una forma similar a h, la función que se desea integrar;
- Tenga colas más pesadas que f.

De hecho, puede decirse cuál es la elección óptima para g (véase Wasserman (2010)).

Teorema 2.2.7. *La elección de g que minimiza la varianza de $\hat{I}$ es*

$$g^*(x) = \frac{|w(x)|f(x)}{\int |w(s)|f(s)\,ds}.$$

El teorema 2.2.7 únicamente tiene interés teórico pues, si no sabemos cómo simular de f, puede ser más complicado simular de g^*.

Algunos Conceptos de Procesos Estocásticos

En general, no es posible generar directamente muestras de una función arbitraria, sobre todo en el caso en el que se trabaja en dimensiones altas. La idea de los métodos MCMC consiste en una aproximación indirecta en la que se requiere simular Cadenas de Markov, por lo anterior se necesita definir algunos conceptos de la teoría de los Procesos Estocásticos y las Cadenas de Markov.

Definición 2.2.8. *Un **proceso estocástico** es una colección de variables aleatorias $\{X_t : t\}$ parametrizada por un conjunto T, llamado espacio parametral, en donde las variables toman valores en un conjunto $\mathcal{X}$, llamado espacio de estados.*

El espacio parametral, T, puede ser continuo, es decir, $T \subset \mathbb{R}$, o discreto, es decir $T = \{0, 1, 2, \ldots\}$. En este último caso, se dice que el proceso es *a tiempo discreto*, y este tipo de procesos puede denotarse explícitamente por

X_0, X_1, X_2, ..., o por $\{X_n : n \in \mathbb{N}\}$.

A continuación se definirán las cadenas de Markov (con espacio parametral discreto), que serán de gran utilidad para el resto de los métodos utilizados en esta sección.

Definición 2.2.9. *El proceso* $\{X_n : n \in \mathbb{N}\}$ *es una* ***cadena de Markov*** *si se cumple que*
$P(X_{n+1} = x_{n+1} \,|\, X_0 = x_0, \ldots, X_n = x_n) = P(X_{n+1} = x_{n+1} \,|\, X_n = x_n)$, *para cualquier* $n \in \mathbb{N}$ *y cualesquiera estados* $x_0, x_1, \ldots, x_{n+1} \in \mathcal{X}$.

Esta propiedad es equivalente a

$$p(x_0, \ldots, x_{n+1}) = p(x_0)p(x_1 \,|\, x_0) \cdots p(x_{n+1} \,|\, x_n).$$

Lo anterior quiere decir que la probabilidad del evento futuro $(X_{n+1} = x_{n+1})$ depende únicamente del evento $(X_n = x_n)$. A la probabilidad condicional $P(X_{n+1} = x_{n+1} \,|\, X_n = x_n)$ se le denomina **probabilidad de transición** del estado x_n en el tiempo n al estado x_{n+1} en el tiempo $n + 1$.

Definición 2.2.10. *Una cadena de Markov se llama* ***homogénea*** *si*

$$P(X_{n+1} = j \,|\, X_n = i) = P(X_1 = j \,|\, X_0 = i),$$

es decir, la probabilidad de transición del estado i *al estado* j *no cambia con el tiempo. Por simplicidad,* $P(X_{n+1} = j \,|\, X_n = i)$ *suele denotarse como* $p_{ij}(n, n + 1)$.

Definición 2.2.11. *Se dice que el estado* j *es* ***accesible*** *desde el estado* i *si existe un entero* $n \geq 0$ *tal que* $P(X_n = j \,|\, X_0 = i) = p_{ij}(n) > 0$, *esto se escribe simplemente como* $i \to j$. *Se dice además que los estados* i *y* j *son* ***comunicantes****, y se escribe* $i \leftrightarrow j$, *si se cumple* $i \to j$ *y* $j \to i$.

La comunicación resulta ser una relación de equivalencia en el espacio de estados de una cadena de Markov y, por lo tanto, induce una partición dada por los subconjuntos de estados comunicantes. Se dice, además, que una cadena de Markov es **irreducible** si todos los estados se comunican entre sí.

Definición 2.2.12. *El* ***periodo*** *es un estado i es un número entero no negativo, denotado por $d(i)$, y definido como sigue:*

$$d(i) = m.c.d.\{n \geq 1 : p_{ii}(n) > 0\}.$$

Cuando $p_{ii}(n) = 0$ para todo $n \geq 1$, se define $d(i) = 0$. En particular, se dice que un estado i es ***aperiódico*** *si $d(i) = 1$. Cuando $d(i) = k \geq 2$, se dice que i es periódico de periodo k.*

El periodo es una propiedad de clase, es decir, todos los estados de una misma clase de comunicación tienen el mismo periodo.

Proposición 2.2.13. *Si los estados i y j pertenecen a la misma clase de comunicación, entonces tienen el mismo periodo.*

También resultará de interés estudiar el primer momento en el que una cadena de Markov visita un estado particular o un conjunto de estados.

Definición 2.2.14. *Sea A un subconjunto del espacio de estados de una cadena de Markov $\{X_n : n \geq 0\}$. El* ***tiempo de primera visita*** *al conjunto A es la variable aleatoria*

$$\tau_A = \begin{cases} \text{mín}\{n \geq 1 : X_n \in A\} & \text{si } X_n \in A \text{para algún } n \geq 1, \\ \infty & \text{en otro caso.} \end{cases}$$

En otras palabras, τ_A es el primer momento positivo en el cual la cadena toma un valor dentro de la colección de estados A, en caso de que ello suceda. Cuando $A = \{j\}$, y si suponemos que la cadena inicia en el estado i, se escribirá τ_{ij}.

Definición 2.2.15. *Para cada $n \geq 1$, el número $f_{ij}(n)$ denota la probabilidad de que una cadena que inicia en el estado i llegue al estado j por primera vez en exactamente n pasos, es decir,*

$$f_{ij}(n) = P(Xn = j,\, X_{n-1} \neq j,\, \ldots,\, X_1 \neq j \,|\, X_0 = 1).$$

Adicionalmente se define $f_{ij}(0) = 0$, incluyendo el caso $i = j$.

Nótese que $f_{ij}(n) = P(\tau_{ij} = n)$ y, en particular, $f_{ii}(n)$ es la probabilidad de regresar por primera vez al mismo estado i en el n-ésimo paso.

Los estados de una cadena de Markov pueden ser clasificados en dos tipos, dependiendo si la cadena es capaz de regresar, con certeza, al estado de partida.

Definición 2.2.16. *La cadena irreducible X_t se dice* ***recurrente****, si para todo estado i,*

$$P\left(\text{mín}\{t > 0 \,:\, X_t = i \,|\, X_0 = i\} < \infty\right) = 1.$$

Un estado que no es recurrente será ***transitorio****, y en tal caso la probabilidad anterior es estrictamente menor que 1.*

Entonces, un estado es recurrente si con probabilidad 1 la cadena es capaz de regresar eventualmente a ese estado, y cuando ello ocurre en algún momento finito, por la propiedad de Markov, se puede regresar a él una y otra vez con probabilidad 1.

Teorema 2.2.17. *El estado i es*

1. *Recurrente, si y sólo si* $\displaystyle\sum_{n=1}^{\infty} P(X_n = i \,|\, X_0 = i) = \infty.$
2. *Transitorio, si y sólo si* $\displaystyle\sum_{n=1}^{\infty} P(X_n = i \,|\, X_0 = i) < \infty.$

En Rincón (2012) se demuestra que toda cadena de Markov finita tiene, por lo menos, un estado recurrente.

Definición 2.2.18. *El **tiempo medio de recurrencia** de un estado recurrente, j, a partir del estado i, se define como el valor esperado de τ_{ij}, y se denota por μ_{ij}, es decir,*

$$\mu_{ij} = \mathrm{E}(\tau_{ij}) = \sum_{n=1}^{\infty} n f_{ij}(n).$$

Definición 2.2.19. *Se dice que un estado recurrente, i, es:*

1. ***Recurrente positivo**, si $\mu_{ii} < \infty$.*
2. ***Recurrente nulo**, si $\mu_{ii} = \infty$.*

La recurrencia positiva es una condición necesaria para garantizar que una cadena irreducible tenga una distribución estacionaria.

Definición 2.2.20. *Una distribución de probabilidad $\pi = (\pi_0, \pi_1, \ldots)$ es **estacionaria o invariante** para una cadena de Markov con probabilidad de transición p_{ij}, si*

$$\pi_j = \sum_i \pi_i p_{ij}.$$

Resulta que en el caso de una cadena de Markov recurrente positiva (todos sus estados son recurrentes positivos), la distribución estacionaria (que es nuestra distribución objetivo) también es una *distribución límite* de iteraciones sucesivas de la cadena. Esto se cumple para cualquier valor inicial de la cadena. El siguiente teorema es de gran utilidad para las implementaciones numéricas sucesivas (véase Gilks *et al.* (1996)).

Teorema 2.2.21. *Si X_t es una cadena de Markov irreducible, aperiódica y recurrente positiva, entonces tiene una única distribución estacionaria, $\pi(\cdot)$. En este caso, llamamos a X_t* ***ergódica****, y se cumplen las siguientes condiciones:*

1. *$p_{ij}(n) \to \pi_j$, cuando $n \to \infty$, para cualesquiera estados i y j.*
2. *Si $\mathrm{E}_\pi[|f(X_t)|] < \infty$, donde $f(\cdot)$ es una función que toma valores en los reales, entonces*

$$P\left(\frac{\sum_{i=1}^n f(X_i)}{n} \to \mathrm{E}_\pi[f(X_t)]\right) = 1,$$

donde $\mathrm{E}_\pi[f(X_t)] = \sum_i f(X_i)\pi(X_i)$ es la esperanza de $f(X_t)$ con respecto a $\pi(\cdot)$.

Diremos, además, que se cumple la condición de **balance detallado**, si para una cadena de Markov irreducible y con distribución estacionaria π se tiene $\pi_i p_{ij} = \pi_j p_{ji}$.

Proposición 2.2.22. *Dada una cadena de Markov irreducible para la cual existe una distribución π que satisface la condición de balance detallado, entonces π es una distribución estacionaria.*

Considérese nuevamente el problema de estimar la integral $I = \int h(x)f(x)\,dx$. La idea principal de los métodos MCMC es construir una cadena de Markov $X_1, X_2, \ldots$, cuya distribución estacionaria sea f. De acuerdo con el Teorema 2.2.21,

$$P\left(\frac{\sum_{i=1}^n h(X_i)}{n} \to \mathrm{E}_f[h(X_t)]\right) = 1.$$

En este trabajo se considerarán los algoritmos de Metropolis-Hastings y Muestreo de Gibbs, pertenecientes a la familia de métodos MCMC.

El algoritmo de Metropolis-Hastings

Sea $X \sim f(x)$, se desea generar una muestra de $f(x)$.La idea detrás de este método es la siguiente. Dada una distribución de transición $q(y \mid x)$, de la que se sepa cómo generar muestras, se crea una sucesión de observaciones $X_0, X_1, \ldots,$ como sigue.

Algoritmo 1 Algoritmo de Metropolis-Hastings

Elegir X_0 de manera arbitraria. Supóngase que se han generado $X_0, X_1, \ldots, X_i$. Para generar X_{i+1} se hace lo siguiente:

1: Generar $Y \sim q(y \mid X_i)$.
2: Evaluar $r \equiv r(X_i, Y)$, donde

$$r(x, y) = \text{mín}\left\{\frac{f(y)}{f(x)}\frac{q(x \mid y)}{q(y \mid x)}, 1\right\}.$$

3: Generar $U \sim \text{Unif}(0, 1)$.
4: Elegir

$$X_{i+1} = \begin{cases} Y, & U < r, \\ X_i, & U \geq 1 - r. \end{cases}$$

Este proceso genera una Cadena de Markov con una distribución de transición $P(X_{i+1} \mid X_i) = r(X_{i+1}, X_i)q(X_{i+1} \mid X_i)$.

El paso (2) del algoritmo 1 suele simplificarse al tomar la distribución de transición *simétrica*; esto quiere decir $q(x \mid y) = q(y \mid x)$. Con lo anterior, el cálculo de r queda como

$$r(x, y) = \text{mín}\left\{\frac{f(y)}{f(x)}, 1\right\}.$$

Una elección común de una distribución de transición simétrica suele ser $q(y \mid x) \sim \text{N}(x, b^2)$, con $b > 0$.

Para ver por qué el algoritmo funciona, recordemos que una distribución π satisface la condición de balance detallado si $\pi_i p_{ij} = \pi_j p_{ji}$. De acuerdo con la Proposición 2.2.22, si π satisface la condición de balance detallado, entonces es una distribución estacionaria.
Se cambiará la notación puesto que se está trabajando con cadenas de Markov con espacio de estados continuo; se denotará por $p(x, y)$ a la probabilidad de pasar del estado x al estado y, y por $f(x)$ en lugar de π a la distribución. Con esta nueva notación, f es una distribución estacionaria si

$f(x) = \int f(y)p(y,\, x)\, dy$, y la condición de balance detallado se cumple para f si $f(x)p(x,\, y) = f(y)p(y,\, x)$.
Si se cumple la condición de balance detallado, entonces

$$\begin{aligned}\int f(y)p(y,\, x)\, dy &= \int f(x)p(x,\, y)\, dy \\ &= f(x)\int p(x,\, y)\, dy \\ &= f(x),\end{aligned}$$

es decir, $f(x) = \int f(y)p(y,\, x)\, dy$. Considérense los puntos x y y, entonces se cumple que

$$f(x)q(y \,|\, x) < f(y)q(x \,|\, y) \quad \text{o} \quad f(x)q(y \,|\, x) > f(y)q(x \,|\, y),$$

el caso $f(x)q(y \,|\, x) = f(y)q(x \,|\, y)$ se excluye porque dos distribuciones continuas son iguales con probabilidad cero. Sin pérdida de generalidad, supóngase que $f(x)q(y \,|\, x) > f(y)q(x \,|\, y)$; esto implica que

$$r(x,\, y) = \frac{f(y)}{f(x)}\frac{q(x \,|\, y)}{q(y \,|\, x)}$$

y que $r(y,\, x) = 1$. Como $p(x,\, y)$ es la probabilidad de pasar del estado x al estado y, se requieren dos cosas: (i) la distribución de transición debe generar a y, y (ii) se debe aceptar el valor de y. Entonces,

$$\begin{aligned}p(x,\, y) &= q(y \,|\, x)r(x,\, y) \\ &= q(y \,|\, x)\frac{f(y)}{f(x)}\frac{q(x \,|\, y)}{q(y \,|\, x)} \\ &= \frac{f(y)}{f(x)}q(x \,|\, y).\end{aligned}$$

Por lo tanto,

$$f(x)p(x,\, y) = f(y)q(x \,|\, y). \tag{2.6}$$

Por otro lado, $p(y,\, x)$ es la probabilidad de pasar de y a x; se requieren dos cosas: (i) la distribución de transición debe generar a x, y (ii) se debe aceptar el valor de x. Lo anterior ocurre con probabilidad

$$p(y,\, x) = q(x \,|\, y)r(y,\, x) = q(x \,|\, y).$$

Por lo tanto,

$$f(y)p(y,\, x) = f(y)q(x \,|\, y).$$

Al comparar las ecuaciones (2.6) y (2.2.2), se verifica que f satisface la condición de balance detallado y, por lo tanto, es una distribución estacionaria.

El muestreo de Gibbs

La idea de este método es convertir un problema multidimensional en varios problemas unidimensionales. Dada $p > 1$ y la variable aleatoria $\boldsymbol{X} = (X_1, \ldots, X_p)$, donde las X_i's pueden ser unidimensionales y multidimensionales, supóngase que se pueden simular muestras de las densidades condicionales completas $f_1, \ldots, f_p$, es decir, se puede simular,

$$X_i \,|\, X_1, X_2, \ldots, X_{i-1}, X_{i+1}, \ldots, X_p \sim f_i(x_i \,|\, x_1, x_2, \ldots, x_{i-1}, x_{i+1}),$$

para $i = 1, 2, \ldots, p$. El algoritmo de muestreo de Gibbs está dado por la siguiente transición de X_t a X_{t+1}.

Algoritmo 2 Muestreo de Gibbs

En la iteración $t = 1, 2, \ldots$, dado $x^{(t)} = (x_1^{(t)}, \ldots, x_p^{(t)})$:

1: Generar $X_1^{(t+1)} \sim f_1(x_1 \,|\, x_2^{(t)}, \ldots, x_p^{(t)})$.

2: Generar $X_2^{(t+1)} \sim f_2(x_2 \,|\, x_1^{(t+1)}, x_3^{(t)}, \ldots, x_p^{(t)})$.

$\vdots$

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

Este proceso genera una Cadena de Markov con una distribución de transición

$$P(X_{t+1} \,|\, X_t) = \prod_{i=1}^{p} P(X_i^{(t+1)} \,|\, X_1^{(t+1)}, \ldots, X_{i-1}^{(t+1)}, X_{i+1}^{(t+1)}, \ldots, X_{p-1}^{(t+1)}).$$

Análisis de Convergencia

Una vez que se ha implementado alguno de los algoritmos anteriores, pueden surgir dificultades numéricas que provoquen que la convergencia del algoritmo sea muy lenta.

1. Una de estas dificultades puede darse cuando la cadena no ha tenido un número suficiente de iteraciones, pues, de manera teórica, la distribución estacionaria se obtiene cuando el número de iteraciones tiende a infinito; en este caso, las simulaciones parecerían no ser una muestra representativa de la distribución objetivo. Inclusive si las simulaciones parecen haber alcanzado la convergencia, las iteraciones iniciales reflejan de mayor manera a la aproximación inicial que a la distribución objetivo.

2. Otra dificultad se da al observar la correlación entre iteraciones sucesivas o *autocorrelación* de la cadena; además de los problemas de convergencia, la inferencia que se obtiene sobre observaciones correlacionadas suele ser menos precisa que la inferencia que se obtendría sobre el mismo número de observaciones no-correlacionadas.

En el primer caso, para disminuir la influencia de las primeras iteraciones, se desecha un número suficientemente grande de iteraciones y se observa el comportamiento de la cadena después de este número de iteraciones. Al número de iteraciones desechado se le conoce como *periodo de calentamiento*. En Gamerman y Lopes (2006), Gelman *et al.* (1995) y Gilks *et al.* (1996), se pueden consultar algunos métodos para estimar el periodo de calentamiento.

En el segundo caso, de acuerdo con Hoff (2009), mientras mayor sea autocorrelación de la cadena, mayor será la varianza del algoritmo MCMC y la aproximación a la distribución estacionaria será menos eficiente. Para observar qué tan autocorrelacionada está la cadena, suele analizarse la función de autocorrelación muestral con *retraso* o (*lag*, en inglés) t, que para una sucesión $\{\phi_1, \ldots, \phi_S\}$, estima la correlación entre elementos de la sucesión que se encuentran a t pasos de distancia. Dicha función se calcula de la siguiente manera:

$$\text{acf}_t(\phi) = \frac{\frac{1}{S-t}\sum_{s=1}^{S-t}(\phi_s - \bar{\phi})(\phi_{s+t} - \bar{\phi})}{\frac{1}{S-1}\sum_{s=1}^{S}(\phi_s - \bar{\phi})^2},$$

esta función se puede calcular en `R` mediante el comando `acf()`, que permite observar gráficamente el valor de $\text{acf}_t(\phi)$ para varios valores de t. Se necesita buscar un valor de t tal que la autocorrelación de la cadena sea razonablemente baja, y quedarse con las t-ésimas iteraciones de la cadena.

2.2.3. Enfoque No-Paramétrico de la Estadística Bayesiana

Los conceptos de estadística paramétrica y no-paramétrica se refieren, esencialmente, la las hipótesis que se plantean respecto a la distribución de las observaciones disponibles. Por ejemplo, dada una muestra $y_1, \ldots, y_n$, una suposición común es que las y_i's son obtenidas de manera independiente de una distribución de probabilidad F. El problema estadístico comienza cuando existe incertidumbre respecto a F. Sea f la función de densidad de probabilidad de F; un modelo estadístico surge cuando se sabe que f es un elemento f_θ de una familia $\mathcal{F} = \{f_\theta : \theta \in \Theta\}$, indexada por un conjunto de parámetros θ de un conjunto de índices Θ.

Los modelos que son descritos por un vector θ conformado por un número finito de valores, por lo general, reales, son denominados como modelos finito-dimensionales o *paramétricos*. Los modelos paramétricos pueden describirse como $\mathcal{F} = \{f_\theta : \theta \in \mathbf{\Theta} \subset \mathbb{R}^p\}$. El objetivo del análisis es, entonces, utilizar la muestra observada para reportar un valor aceptable para θ, o por lo menos determinar un conjunto de $\mathbf{\Theta}$ que, de manera ideal, contenga a θ.

En muchas situaciones, sin embargo, restringir la inferencia a una forma paramétrica específica puede limitar el alcance y tipo de inferencias que se pueden obtener a partir de dichos modelos. Por lo tanto, sería necesario debilitar las hipótesis paramétricas para tener mayor flexibilidad, en este caso, y procediendo de manera Bayesiana, se necesita especificar una distribución inicial para un espacio de parámetros de dimensión infinita. Una forma de especificar dichas distribuciones iniciales es mediante procesos estocásticos cuyas trayectorias son distribuciones de probabilidad.

Los parámetros de interés infinito-dimensionales suelen ser funciones, por ejemplo, funciones de densidad de probabilidad, funciones de distribución, o funciones de regresión. El considerar distribuciones de probabilidad requiere la definición de medidas de probabilidad sobre una colección de funciones de distribución. Dichas medidas de probabilidad se denominan *medidas de probabilidad aleatorias*. De manera muy breve, supóngase que se tiene un espacio de probabilidad $(\Omega, \mathcal{A}, \mu)$ y S, un espacio métrico completo y separable, con la σ-álgebra de Borel, $\mathcal{B}$. Sea $M(S)$ el espacio de medidas de probabilidad sobre S, dotado con la topología de la convergencia débil, lo cual lo convierte en un espacio métrico completo y separable. Las distribuciones iniciales Bayesianas no-paramétricas que se presentan a continuación son distribuciones sobre $M(S)$. Es decir, si d es una métrica que induce una topología en $\mathcal{F}$, donde $\mathcal{F}$ es la familia de funciones de distribución acumulada, para una medi-

da de probabilidad G, se busca segurarse que $\mathrm{P}\{F \in \mathcal{F} : d(F, G) < \epsilon\} > 0$, para toda $\epsilon > 0$.

El Proceso Dirichlet

La primera construcción importante para asignar distribuciones iniciales a espacios de parámetros de dimensión infinita fue propuesta por Ferguson (1973), y lleva el nombre de *Proceso Dirichlet*; como su nombre lo indica, tiene como punto de partida a la Distribución Dirichlet. Esta distribución es una generalización multivariada de la Distribución Beta, y sirve como un modelo conjugado para la Distribución Multinomial.

Sea $\boldsymbol{\alpha} = (\alpha_1, \ldots, \alpha_k)$, con $\alpha_j > 0$ y $\alpha_0 \equiv \sum_{j=1}^{k} \alpha_j$, entonces

$$\begin{aligned} \theta &\sim \mathrm{Dir}(\alpha_1, \ldots, \alpha_k) \\ p(\theta) &= \mathrm{Dir}(\theta \,|\, \alpha_1, \ldots, \alpha_k) \\ &= \frac{\Gamma(\alpha_1 + \cdots + \alpha_k)}{\Gamma(\alpha_1) \cdots \Gamma(\alpha_k)} \theta_1^{\alpha_1 - 1} \cdots \theta_k^{\alpha_k - 1}, \end{aligned}$$

con $\theta_j \geq 0$ y $\sum_{j=1}^{k} \theta_j = 1$.

De manera formal, el Proceso Dirichlet es un proceso estocástico cuyas realizaciones son, con probabilidad uno, medidas de probabilidad. Dado un espacio de probabilidad $(\Omega, \mathcal{B}, P)$ y $B_1, \ldots, B_k$ una partición de Ω, decimos que P se distribuye de acuerdo con un Proceso Dirichlet si

$$(P(B_1), \ldots, P(B_k)) \sim \mathrm{Dir}(\alpha P_0(B_1), \ldots, \alpha P_0(B_k)),$$

donde P_0 es una medida de probabilidad base y $\alpha > 0$ es un parámetro de concentración que controla la aproximación de P hacia P_0. Lo anterior se denota como $P \sim \mathrm{DP}(\alpha P_0)$.
La definición del Proceso Dirichlet y las propiedades de la distribución Dirichlet implican que para cualquier $B \in \mathcal{B}$:

$$\begin{aligned} P(B) &\sim \mathrm{Beta}(\alpha P_0(B), \alpha(1 - P_0(B))), \\ \mathrm{E}(P(B)) &= P_0(B), \\ \mathrm{Var}(P(B)) &= \frac{P_0(B)(1 - P_0(B))}{1 + \alpha}. \end{aligned}$$

Además, dada una muestra aleatoria $Y_1, \ldots, Y_n$, con $Y_i \sim P$, entonces para cualquier partición $B_1, \ldots, B_k$ de Ω,

$$P(B_1), \ldots, P(B_k) \,|\, \boldsymbol{y} \sim \mathrm{Dir}\left(\alpha P_0(B_1) + \sum_{i=1}^{n} \mathbb{1}_{y_i \in B_1}, \ldots, \alpha P_0(B_k) + \sum_{i=1}^{n} \mathbb{1}_{y_i \in B_k}\right),$$

donde $\boldsymbol{y} = (y_1, \ldots, y_n)$.
De lo anterior se obtiene que

$$P \,|\, \boldsymbol{y} \sim \mathrm{DP}\left(\alpha P_0 + \sum_i \delta_{y_i}\right).$$

El nuevo parámetro de concentración es $\alpha + n$ y la esperanza posterior de P tiene la expresión

$$\mathrm{E}(P(B) \,|\, \boldsymbol{y}) = \left(\frac{\alpha}{\alpha + n}\right) P_0(B) + \left(\frac{n}{\alpha + n}\right) \frac{1}{n} \sum_{i=1}^{n} \delta_{y_i}.$$

Este proceso es restrictivo, pues genera distribuciones discretas con probabilidad uno.

Procesos *Libres de Cola*

La noción de Proceso Libre de Cola (*Tail-Free*, TF) fue introducida por Freedman (1963) Y Fabius (1964) y antecede al Proceso Dirichlet. Supóngase que Ω es un espacio muestral completo y separable, $E = \{0, 1\}$, denotemos al producto cartesiano de E consigo mismo como $E^m = E \times \ldots \times E$, $E^0 = \emptyset$ y $E^* = \bigcup_{m=0}^{\infty} E^m$. Se escribirá a $\epsilon \in E^m$ como un enterio binario, es decir, $\epsilon = \varepsilon_1 \cdots \varepsilon_m$, con $\varepsilon_i \in E$. Un Proceso Libre de Cola se define mediante la asignación de probabilidades aleatorias a conjuntos que conforman una sucesión de particiones anidadas del espacio muestral, utilizando a $\epsilon \in E^*$ para indexar a los miembros de la partición, B_ϵ, de la siguiente manera. Sea $\pi_0 = \{\Omega\}$, $\pi_1 = \{B_0, B_1\}$, $\pi_2 = \{B_{00}, B_{01}, B_{10}, B11\}$, ..., una sucesión de particiones anidadas de Ω, tales que $B_\epsilon = B_{\epsilon 0} \cup B_{\epsilon 1}$ y $B_{\epsilon 0} \cap B_{\epsilon 1} = \emptyset$ para cada $\epsilon = \varepsilon_1 \cdots \varepsilon_m \in E^*$, es decir, π_n es una partición de Ω y π_{n+1} un refinamiento de π_n que se obtiene de la descomposición de cada conjunto $B_\epsilon \in \pi_n$ dada por $B_\epsilon = B_{\epsilon 0} \cup B_{\epsilon 1}$.

Supóngase que B_ϵ es un intervalo abierto por la izquierda y cerrado por la derecha, salvo en el caso $\epsilon = 1 \cdots 1$ y que $\bigcup_{m=0}^{\infty} \pi_m$ es un generador de la σ-álgebra de Borel sobre Ω; la última condición se asegura si la colección de extremos derechos de B_ϵ es densa en Ω. Se puede describir una medida de probabilidad F al especificar todas las probabilidades condicionales $\{Y_\epsilon = F(B_{\epsilon 0} \,|\, B_\epsilon) \,:\, \epsilon \in E^*\}$. Una distribución inicial para F puede definirse, por lo tanto, al especificar las distribuciones conjuntas de las probabilidades condicionales Y_ϵ, esta especificación puede entenderse como un árbol en donde las distribuciones de probabilidad en los distintos niveles de jerarquía pueden interpretarse como las especificaciones iniciales en

los distintos niveles del árbol. Una distribución inicial para F se dice *libre de cola* con respecto a la sucesión de particiones $\{\pi_m\}_{m=0}^{\infty}$ si las colecciones $\{Y_\emptyset\}$, $\{Y_0, Y_1\}$, $\{Y_{00}, Y_{01}, Y_{10}, Y_{11}\}$, ..., son mutuamente independientes. Nótese que las variables dentro del mismo nivel de jerarquía no necesariamente deben ser independientes, únicamente las variables que se encuentran en niveles de jerarquía diferentes necesitan serlo.

La familia de Procesos Libres de Cola incluye al Proceso Dirichlet como un caso particular muy importante, pues es el único proceso que es Libre de Cola con respecto a cualquier sucesión de particiones (véase Ferguson, 1974).

2.3. Árboles de Pólya

Este proceso se basa en particiones binarias del espacio muestral Ω (generalmente $\mathbb{R}$); es un proceso menos restrictivo, pues permite generar tanto distribuciones discretas como continuas, con probabilidad uno, y tiene como caso particular al Proceso Dirichlet. Fue propuesto por Ferguson (1974) y estudiado posteriormente por Mauldin *et al.* (1992) y Lavine (1992, 1994). Son dos las características que definen a un Árbol de Pólya:

- $\Pi = \{B_\epsilon,\, \epsilon = \varepsilon_1 \cdots \varepsilon_m : \varepsilon_j \in \{0,1\},\, m = 0,1,2,\ldots\}$, un conjunto de particiones binarias anidadas de Ω,
- $\mathcal{A} = \{\alpha_\epsilon,\, \epsilon = \varepsilon_1 \cdots \varepsilon_m : \varepsilon_j \in \{0,1\},\, m = 0,1,2,\ldots\}$, un conjunto de parámetros tal que cada α_ϵ está asociado al conjunto B_ϵ.

Sea $\Pi = \{B_{\varepsilon_1 \cdots \varepsilon_m}\}$ el conjunto de particiones binarias anidadas de $\Omega = \mathbb{R}$, tal que en el nivel $m = 1, 2, \ldots,$ se tiene una partición de $\mathbb{R}$ con 2^m elementos, y donde el índice $j = 1, 2, \ldots, 2^m$ identifica al j-ésimo elemento de la partición en el nivel m. Es decir, en el nivel m, $B_{\varepsilon_1 \cdots \varepsilon_m}$ se particiona en $(B_{\varepsilon_1 \cdots \varepsilon_m 0},\, B_{\varepsilon_1 \cdots \varepsilon_m 1})$ para el nivel $m+1$, con $B_{\varepsilon_1 \cdots \varepsilon_m 0} \cap B_{\varepsilon_1 \cdots \varepsilon_m 1} = \emptyset$.

Sea $\mathcal{A} = \{\alpha_{\varepsilon_1 \cdots \varepsilon_m}\}$ el conjunto de parámetros tal que cada $\alpha_{\varepsilon_1 \cdots \varepsilon_m}$ está asociado al conjunto $B_{\varepsilon_1 \cdots \varepsilon_m}$. El parámetro $\alpha_{\varepsilon_1 \cdots \varepsilon_m}$ es tal que define las probabilidades condicionales

$$Y_{\epsilon 0} = \mathrm{P}(Y \in B_{\epsilon 0} \,|\, Y \in B_\epsilon) \;\text{ y }\; Y_{\epsilon 1} = \mathrm{P}(Y \in B_{\epsilon 1} \,|\, Y \in B_\epsilon) = 1 - Y_{\epsilon 0}.$$

Por lo tanto, para los conjuntos en el nivel m, la probabilidad de que Y pertenezca al conjunto $B_{\varepsilon_1 \cdots \varepsilon_m}$ es el producto de todas las probabilidades

condicionales $\alpha_{\varepsilon_1 \cdots \varepsilon_m}$, una para cada nivel, a donde pertenece el conjunto $B_{\varepsilon_1 \cdots \varepsilon_m}$. Es decir,

$$\mathrm{P}(Y \in B_{\varepsilon_1 \cdots \varepsilon_m}) = \prod_{j=1}^{m} Y_{\varepsilon_1 \cdots \varepsilon j},$$

donde $Y_{\varepsilon_1 \cdots \varepsilon j-1,0} \sim \text{Beta}(\alpha_{\varepsilon_1 \cdots \varepsilon j-10}, \alpha_{\varepsilon_1 \cdots \varepsilon j-1,1})$ y $Y_{\varepsilon_1 \cdots \varepsilon j-11} = 1 - Y_{\varepsilon_1 \cdots \varepsilon j-10}$; debido a que las variables aleatorias $Y_{\epsilon 0}$ y $Y_{\epsilon 0}$, la distribución inicial Beta resulta conveniente para dichas variables. Lo anterior se expresa de manera formal en la siguiente definición.

Definición 2.3.1. *Sea* $\mathcal{A} = \{\alpha_\varepsilon, \varepsilon \in \bigcup_{m=0}^{\infty} \{0,1\}^m\}$ *un conjunto de números reales no negativos,* $m = 1, 2, \ldots$, *y*
$\Pi = \{B_\epsilon, \epsilon = \varepsilon_1 \cdots \varepsilon_m : \varepsilon_j \in \{0,1\}, m = 0, 1, 2, \ldots\}$ *un conjunto de particiones binarias anidadas de* Ω. *Se dice que una medida aleatoria de probabilidad* F *sobre* $(\mathbb{R}, \mathcal{B})$ *tiene* ***una distribución inicial de Árbol de Pólya*** *con parámetros* $(\Pi, \mathcal{A})$, *si para* $m = 1, 2, \ldots$ *y para* $\varepsilon = \varepsilon_1 \cdots \varepsilon_m \in \{0,1\}^m$:

1. $\mathrm{F}(B_{\varepsilon_1 \cdots \varepsilon_m}) = \prod_{j=1}^{m} Y_{\varepsilon_1 \cdots \varepsilon_j}$.
2. *Las probabilidades condicionales* $Y_{\varepsilon_1 \cdots \varepsilon_{j-1} 0}$ *son variables aleatorias independientes tales que*

$$\begin{aligned} Y_{\varepsilon_1 \cdots \varepsilon_{j-1},0} &\sim \text{Beta}(\alpha_{\varepsilon_1 \cdots \varepsilon_{j-1},0}, \alpha_{\varepsilon_1 \cdots \varepsilon_{j-1},1}), \\ Y_{\varepsilon_1 \cdots \varepsilon_{j-1},1} &= 1 - Y_{\varepsilon_1 \cdots \varepsilon_{j-1},0}. \end{aligned}$$

Se escribe $\mathrm{F} \sim \mathcal{PT}(\Pi, \mathcal{A})$.

El Proceso Dirichlet es un caso particular del Proceso de Árbol de Pólya, que se caracteriza por la condición $\alpha_{\varepsilon_0} + \alpha_{\varepsilon_1} = \alpha_\varepsilon$, para todo $\varepsilon \in \bigcup_{m=0}^{\infty} \{0,1\}^m$, sin embargo, a diferencia del Proceso Dirichlet, el Árbol de Pólya puede generar distribuciones de probabilidad absolutamente continuas, con probabilidad 1. Kraft (1964) y Metivier (1971) demostraron que $\alpha_{\varepsilon_1 \cdots \varepsilon_m} = m^2$ es una condición suficiente para garantizar que un Árbol de Pólya asigne probabilidad 1 a la clase de las distribuciones de probabilidad absolutamente continuas. El

Proceso de Árbol de Pólya pertenece a la clase de Procesos Libres de Cola, sin embargo, a diferencia del Proceso Dirichlet, **el Árbol de Póya depende de la partición del espacio muestral que se especifique**.

Debido a que las variables aleatorias Y_ε siguen una distribución de probabilidad Beta, es posible encontrar expresiones para los momentos de $F(B_\varepsilon)$ Para cada $\varepsilon \in \bigcup_{m=0}^{\infty} \{0,1\}^m$, sean $\mu_{\varepsilon 0} = \mathrm{E}(Y_{\varepsilon 0})$, $\mu_{\varepsilon 1} = 1 - \mathrm{E}(Y_{\varepsilon 0})$, $s_{\varepsilon 0} = \mathrm{E}(Y_{\varepsilon 0}^2)$ y $s_{\varepsilon 1} = \mathrm{E}([1 - Y_{\varepsilon 0}]^2)$ el primero y segundo momentos de una distribución Beta($\alpha_{\varepsilon 0}$, $\alpha_{\varepsilon 1}$), entonces

$$\begin{aligned} \mathrm{E}\left\{F(B_{\varepsilon_1 \cdots \varepsilon_m})\right\} &= \prod_{j=1}^{m} \mu_{\varepsilon_1 \cdots \varepsilon_j} \\ \mathrm{Var}\left\{F(B_{\varepsilon_1 \cdots \varepsilon_m})\right\} &= \prod_{j=1}^{m} s_{\varepsilon_1 \cdots \varepsilon_j} - \prod_{j=1}^{m} \mu_{\varepsilon_1 \cdots \varepsilon_j}^2. \end{aligned}$$

Las condiciones anteriores permiten centrar a la medida $F \,|\, \Pi, \mathcal{A} \sim \mathcal{PT}(\Pi, \mathcal{A})$ en alguna media inicial deseada, $\mathrm{E}(F) = F_0$. La distribución inicial de Árbol de Pólya está definida en términos de la partición Π y el conjunto de parámetros no-negativos $\mathcal{A}$. Estos dos conjuntos deben reflejar nuestro conocimiento inicial respecto a la medida de probabilidad $F(\cdot)$. Si supiéramos que el valor verdadero de $F(\cdot)$ es cercano a una distribución $F_0(\cdot)$, por ejemplo N(0, 1), podemos hacer que la distribución inicial satisfaga $\mathrm{E}(F) = F_0$ de la siguiente manera (ver, por ejemplo, Hanson y Johnson (2002)). Sea π_m la partición de Ω en el nivel m del árbol y sea $B_{\varepsilon_1 \cdots \varepsilon_m} \in \pi_m$ el cuantil $F_0^{-1}(k/2^m)$, $k = 0, 1, \ldots, 2^m$; sea $N(\varepsilon)$ el entero con representación en base 2 tal que $\varepsilon = \varepsilon_1 \cdots \varepsilon_m \in \bigcup_{m=0}^{\infty} \{0,1\}^m$, se define

$$B_{\varepsilon_1 \cdots \varepsilon_m} = \left[F_0^{-1}\left(\frac{N(\varepsilon)}{2^m}\right),\, F_0^{-1}\left(\frac{N(\varepsilon)+1}{2^m}\right)\right),$$

con $F_0^{-1}(0) = -\infty$, $F_0^{-1}(1) = \infty$.

Si se toma $\alpha_{\varepsilon 0} = \alpha_{\varepsilon 1}$ para todo $\varepsilon \in \bigcup_{m=0}^{\infty} \{0,1\}^m$, entonces

$$\mathrm{E}\left[F\left(B_{\varepsilon_1 \cdots \varepsilon_m}\right)\right] = \prod_{j=1}^{m} \mathrm{E}\left(Y_{\varepsilon_1 \cdots \varepsilon_m}\right) = \frac{1}{2^m} = F_0\left(B_{\varepsilon_1 \cdots \varepsilon_m}\right).$$

La distribución F_0 juega un papel similar a la medida base en el Proceso Dirichlet y se escribe $F \sim \mathcal{PT}(F_0, \mathcal{A})$.

Considérese la construcción $F \sim \mathcal{PT}(F_0, \mathcal{A})$, una vez que el Árbol de Pólya ha sido centrado alrededor de F_0, la familia $\mathcal{A} = \{\alpha_\varepsilon,\, \varepsilon \in \bigcup_{m=0}^{\infty} \{0,1\}^m\}$ determina qué tanto F puede distar de F_0, tal como el parámetro α del Proceso Dirichlet. El Árbol de Pólya tiene un número infinito de parámetros de la familia $\mathcal{A}$, para evitar una búsqueda extensiva de una distribución inicial, α_ε suele elegirse dependiendo solamente de la longitud de la cadena binaria ε. Walker y Mallick proponen $_{\varepsilon_1 \cdots \varepsilon_m} = \alpha m^2$, en donde $\alpha > 0$ es un parámetro de precisión: mayores valores de α hacen que la distribución inicial se concentre más cerca de F_0 y valores menores de α ocasionarán una mayor variabilidad alrededor de F_0. Al considerar valores distintos de α, Hanson y Johnson (2002) encontraron que la familia $\alpha_{\varepsilon_1 \cdots \varepsilon_m} = \alpha m^2$ es lo suficientemente rica para capturar características interesantes de las distribuciones bajo consideración, se denotará a esta familia como

$$\mathcal{A}_\alpha = \left\{ \alpha_\varepsilon = \alpha m^2,\, \varepsilon = \varepsilon_1 \cdots \varepsilon_m \in \bigcup_{m=0}^{\infty} \{0,1\}^m \right\}.$$

Mauldin *et al.* (1992) y Lavine (1992, 1994) demuestran varias propiedades del Proceso de Árbol de Pólya, incluyendo un resultado de conjugación, es decir, el Árbol de Pólya es la distribución inicial conjugada de una medida de probabilidad F bajo una muestra aleatoria $x_1, \ldots, x_n \,|\, F \sim F$. La distribución posterior es, de nuevo, un Árbol de Pólya con parámetros Beta actualizados. El parámetro α_ε se incrementa en 1 por cada $x_i \in B_\varepsilon$. Sea $n_\varepsilon = \sum_{i=1}^{n} \mathbb{1}(x_i \in B_\varepsilon)$ el número de observaciones dentro del conjunto B_ε.

Proposición 2.3.2. *Dada una muestra aleatoria* $x_1, \ldots, x_n \,|\, F \sim F$ *y* $F \sim \mathcal{PT}(\Pi, \mathcal{A})$*, entonces*

$$F \,|\, x_1, \ldots, x_n \sim \mathcal{PT}(\Pi, \mathcal{A}^*),$$

donde los parámetros Beta actualizados $\alpha_\varepsilon^* \in \mathcal{A}^*$ *están dados por*

$$\alpha_\varepsilon^* = \alpha_\varepsilon + n_\varepsilon.$$

Lavine (1992) encontró una expresión para el modelo marginal $p(x_1, \ldots, x_n)$ cuando $x_1, \ldots, x_n \,|\, F \sim F$ y $F \sim \mathcal{PT}(\Pi, \mathcal{A})$. Supóngase que se tiene un Árbol de Pólya centrado, con $F \,|\, F_0, \mathcal{A} \sim \mathcal{PT}(F_0, \mathcal{A})$ o con la especificación $F \,|\, \Pi, F_0 \sim \mathcal{PT}(\Pi, F_0)$. Sea f_0 la función de densidad de probabilidad de F_0; a continuación se introduce notación para localizar una observación, x, perteneciente a los conjuntos que forman parte de la partición del espacio muestral

para los diferentes niveles de la partición anidada. Para cada $m = 1, 2, \ldots$, sea $\epsilon_m(x_i) = \varepsilon_1 \cdots \varepsilon_m \in \bigcup_{m=0}^{\infty} \{0,1\}^m$ el índice del subconjunto del nivel m del árbol que contiene a x_i, es decir, $x_i \in B_{\varepsilon_1 \cdots \varepsilon_m}$. También, sea $m^*(x_i)$ el menor nivel m tal que x_i es el único punto en $B_{\epsilon_m(x_i)}$, entre $x_1, \ldots, x_i$. Finalmente, sean

$$\begin{aligned} n_{\epsilon}^{(j)} &= \sum_{i=1}^{j-1} \mathbb{1}(x_i \in B_{\varepsilon}), \\ \alpha_{\varepsilon}^{*(j)} &= \alpha_{\varepsilon} + n_{\epsilon}^{(j)} \end{aligned}$$

el número de observaciones $x_1, \ldots, x_{j-1}$ en B_ε y los parámetros Beta que se actualizan con dichos conteos, respectivamente.

Proposición 2.3.3. *La distribución marginal de una muestra aleatoria $x_1, \ldots, x_n$ está dada por*

$$p(x_1, \ldots, x_n) = \left\{ \prod_{i=1}^{n} f_0(x_i) \right\} \prod_{j=2}^{n} \prod_{m=1}^{m^*(x_j)} \frac{\alpha_{\epsilon_m(x_j)}^{*(j)}}{\alpha_{\epsilon_m(x_j)}} \cdot \frac{\alpha_{\epsilon_{m-1}(x_j)0} + \alpha_{\epsilon_{m-1}(x_j)1}}{\alpha_{\epsilon_{m-1}(x_j)0}^{*(j)} + \alpha_{\epsilon_{m-1}(x_j)1}^{*(j)}}. \tag{2.7}$$

Nótese que el producto $\prod_m \alpha_{\epsilon_m(x_j)}^{*(j)} / \left\{ \alpha_{\epsilon_{m-1}(x_j)0}^{*(j)} + \alpha_{\epsilon_{m-1}(x_j)1}^{*(j)} \right\}$ es igual a la probabilidad posterior predictiva $p(x_i \in B_\varepsilon \,|\, x_1, \ldots, x_{j-1})$ para $\varepsilon = \epsilon_m(x_j)$. Se sigue de la Proposición 2.3.2, aplicada a $p(F \,|\, x_1, \ldots, x_{j-1})$, y de las expresiones para los momentos de $F(B_{\varepsilon_1 \cdots \varepsilon_m})$, sustituyendo $\mathrm{E}(Y_\varepsilon) = \alpha_{\varepsilon 0}/(\alpha_{\varepsilon 0} + \alpha_{\varepsilon 1})$. Condicionado sobre $x_j \in B_\varepsilon$, la distribución condicional inicial predictiva es F_0, es decir, tiene densidad $f_0(x_j)\,\{1/F_0(B_\varepsilon)\}$. Por construcción, se tiene que $F_0(B_\varepsilon) = \mathrm{E}[F(B_\varepsilon)]$, la media inicial, que resulta poder escribirse similarmente como un producto de factores, $\prod_{m=1}^{m^*(x_j)} \alpha_{\epsilon_m(x_j)} / \left\{ \alpha_{\epsilon_{m-1}(x_j)0} + \alpha_{\epsilon_{m-1}(x_j)1} \right\}$, que ahora involucra a los parámetros iniciales α_ε.

El argumento anterior también es válido para la construcción $\mathcal{PT}(\Pi,\, F_0)$, en este caso, de acuerdo con Müller *et al.* (2015), la expresión (2.7) puede simplificarse a

$$p(x_1, \ldots, x_n) = \left\{ \prod_{i=1}^{n} f_0(x_i) \right\} 2^m \prod_{j=2}^{n} \prod_{m=1}^{m^*(x_j)} \frac{\alpha m^2 + n_{\epsilon_m(x_j)}^{(j)}}{2\alpha m^2 + n_{\epsilon_{m-1}(x_j)}^{(j)}}. \tag{2.8}$$

Las ecuaciones (2.7) y (2.8) son válidas cuando la medida F_0 depende de hiperparámetros desconocidos η, es decir, cuando la medida base es $F_{0,\eta}$.

Aunque la función de distribución inicial de la media tenga una densidad

de Lebesgue diferenciable, las muestras aleatorias de las densidades de un Árbol de Pólya pueden no ser diferenciables en ninguna parte. Barron *et al.* (1999) notaron que las las densidades posteriores predictivas de observaciones futuras calculadas medianet una distribución inicial de Árbol de Pólya tienen saltos notables en las fronteras de los subconjuntos que forman la partición del árbol y que una elección de la medida base F_0 que no sea parecida a la distribución de los datos hará que la convergencia de la distribución final sea muy lenta. Para superar estas dificultades, es natural considerar una medida base con parámetros no especificados, $F_{0,\eta}$, y una distribución inicial π de dichos hiperparámetros; el modelo jerárquico resultante es una *Mezcla de Árboles de Pólya.*

Por ejemplo, en los modelos de regresión, en el caso de las distribuciones residuales, es de gran interés el caso en el que η es un parámetro de escala y F está forzada a tener mediana cero. Hanson y Johnson (2002) demostraron que un modelo de Mezcla de Árboles de Pólya con una distribución inicial sobre un parámetro de escala η para la medida base $F_{0,\eta}$ implica una distribución predictiva continua, excepto en el cero. El resultado anterior es válido para una muestra aleatoria $x_1, \ldots, x_n \,|\, F \sim F$ con una distribución inicial de Árbol de Pólya $F \,|\, \eta \sim \mathcal{PT}\left(\Pi_{F_{0,\eta}}, \mathcal{A}\right)$, con $\alpha_{\varepsilon_1 \cdots \varepsilon_m} = \alpha m^2$ y una distribución inicial $\pi(\eta)$ para los hiperparámetros.

Otra manera de crear un modelo de Mezcla de Árboles de Pólya es mantener fija la partición del espacio muestral y variar los parámetros α_ε. Como la partición no varía, la densidad resultante es discontinua en todos lados, como en el Árbol de Pólya usual.

Una de las limitantes computacionales de los Árboles de Pólya es la necesidad de actualizar una cantidad infinita de parámetros, una alternativa a esta limitante son los *Árboles de Pólya finitos* (ver Lavine (1994)). La construcción del Árbol de Pólya finito es idéntica a la del Árbol de Pólya hasta un nivel J previamente especificado, también los parámetros en el conjunto $\{\alpha_\varepsilon, \varepsilon \in \bigcup_{m=0}^{\infty} \{0,1\}^m\}$ se actualizan hasta el nivel J, este modelo se denota por $\mathcal{PT}\left(F_{0,\eta}, \mathcal{A}, J\right)$.

Así, el algoritmo para la inferencia posterior para modelos con distribución inicial de Árbol de Pólya es:

Algoritmo 3 Simulación de la distribución posterior dado $\mathcal{PT}(\Pi, \mathcal{A})$.

1: Construir Π: Para cada $m = 1, \ldots, M$ y para $\epsilon \in \bigcup_{m=0}^{\infty} \{0,1\}^m$, se define $R_\epsilon = F_0^{-1}\left(\frac{N(\epsilon)+1}{2^m}\right)$, el extremo derecho del intervalo de la partición.

2: Evaluar $\mathcal{A}^*$. Para cada $m = 1, \ldots, M$:
Para $\epsilon \in \bigcup_{m=0}^{\infty} \{0,1\}^m$, sea $\alpha_\epsilon = \alpha_\epsilon^* = \alpha m^2$ (valor inicial de α_ϵ^*). Para $i = 1, \ldots, n$:

- $\epsilon_m(x_i) = \text{mín}\{\epsilon \in \bigcup_{m=0}^{\infty} \{0,1\}^m : R_\epsilon \geq x_i\}$ (índice del nivel m que contiene a x_i).
- $\alpha^*_{\epsilon_m(x_i)} = \alpha^*_{\epsilon_m(x_i)} + 1$ (construyendo $\alpha_\epsilon + n_\epsilon$ para todos los niveles de la partición).

3: Simulación posterior.
Para cada $m = 0, \ldots, M-1$ y para $\epsilon \in \bigcup_{m=0}^{\infty} \{0,1\}^m$, sea

$$Y_{\epsilon 0} \sim \text{Beta}(\alpha^*_{\epsilon 0}, \alpha^*_{\epsilon 1}), \text{ y } Y_{\epsilon 1} = 1 - Y_{\epsilon 0}.$$

Para cada $\epsilon \in \bigcup_{m=0}^{\infty} \{0,1\}^m$,

$$\text{F}(B_\epsilon) = \prod_{m=1}^{M} Y_{\varepsilon_1 \cdots \varepsilon_m}.$$

De igual manera, el algoritmo para encontrar la esperanza posterior para un modelo con distribución inicial de Árbol de Pólya es el siguiente:

Algoritmo 4 Simulación de la esperanza posterior dado $\mathcal{PT}(\Pi, \mathcal{A})$.

1: Ejecutar los pasos 1 y 2 del Algoritmo 3 para evaluar α_ϵ^*. Posteriormente evaluar:

$$\text{E}\left[F\left(B_{\varepsilon_1 \cdots \varepsilon_J}\right) \mid \boldsymbol{x}\right] = \prod_{m=1}^{J} \text{E}(Y_{\varepsilon_1 \cdots \varepsilon_m} \mid \boldsymbol{x}) = \prod_{m=1}^{J} \frac{\alpha^*_{\varepsilon_1 \cdots \varepsilon_m}}{\alpha^*_{\varepsilon_1 \cdots \varepsilon_{m-1} 0} + \alpha^*_{\varepsilon_1 \cdots \varepsilon_{m-1} 1}}.$$

El modelo jerárquico obtenido de la partición dada por los cuantiles diádicos de $F_{0,\eta}$ es:

$$\begin{aligned} x_1, \ldots, x_n \,|\, F &\sim F, \\ F \,|\, \alpha, \eta &\sim \mathcal{PT}(\Pi^\eta, \mathcal{A}) \\ (\alpha, \eta) &\sim \pi. \end{aligned} \tag{2.9}$$

Es posible marginalizar la ecuación (2.9) para hacer inferencia sobre la distribución predictiva $p(x_1, \ldots, x_n \,|\, \alpha, \eta)$, de acuerdo con la ecuación (2.7). Nótese que $\alpha^{*(j)}_{\epsilon_m}$ en la ecuación (2.7) es ahora una función de α y que la partición Π^η es una función de los hiperparámetros η. Así, la inferencia vía métodos MCMC para la distribución marginal posterior $p(\eta, \alpha \,|\, \boldsymbol{x})$ está dada por:

$$\begin{aligned} p(\eta, \alpha \,|\, \boldsymbol{x}) &\propto p(x_1, \ldots, x_n \,|\, \alpha, \eta) \\ &= \left\{ \prod_{i=1}^{n} f_{0,\eta}(x_i) \right\} \prod_{j=2}^{n} \prod_{m=1}^{m^*(x_j)} \frac{\alpha^{*(j)}_{\epsilon_m(x_j)}}{\alpha_{\epsilon_m(x_j)}} \cdot \frac{\alpha_{\epsilon_{m-1}(x_j)0} + \alpha_{\epsilon_{m-1}(x_j)1}}{\alpha^{*(j)}_{\epsilon_{m-1}(x_j)0} + \alpha^{*(j)}_{\epsilon_{m-1}(x_j)1}} \pi(\alpha, \eta), \end{aligned} \tag{2.10}$$

y puede simplificarse, utilizando la ecuación (2.8), como:

$$p(\eta, \alpha \,|\, \boldsymbol{x}) \propto \left\{ \prod_{i=1}^{n} f_{0,\eta}(x_i) \right\} \prod_{j=2}^{n} \prod_{m=1}^{m^*(x_j)} \frac{2\alpha m^2 + 2n^{(j)}_{\epsilon_m(x_j)}}{2\alpha m^2 + n^{(j)}_{\epsilon_{m-1}(x_j)}} \pi(\alpha, \eta). \tag{2.11}$$

A continuación se formula un algoritmo para evaluar la ecuación (2.10) para un Árbol de Pólya Finito. Sea $p(x_1, \ldots, x_n \,|\, \alpha, \eta) = \left\{ \prod_{i=1}^{n} f_{0,\eta}(x_i) \right\} \psi(\boldsymbol{x})$, con

$$\psi(\boldsymbol{x}) = \left\{ \prod_{i=1}^{n} f_{0,\eta}(x_i) \right\} \prod_{j=2}^{n} \prod_{m=1}^{m^*(x_j)} \frac{\alpha^{*(j)}_{\epsilon_m(x_j)}}{\alpha^{*(j)}_{\epsilon_{m-1}(x_j)0} + \alpha^{*(j)}_{\epsilon_{m-1}(x_j)1}} \cdot \frac{\alpha_{\epsilon_{m-1}(x_j)0} + \alpha_{\epsilon_{m-1}(x_j)1}}{\alpha_{\epsilon_m(x_j)}}.$$

Dado que los conjuntos $B_{\varepsilon_1 \cdots \varepsilon_m}$ están formados por los cuantiles de $F_{0,\eta}$, por lo que los valores $\alpha^{*(j)}_{\epsilon_m}$ son funciones de η y de α. Tomando en cuenta lo anterior, se describe el algoritmo para calcular la distribución marginal $p(\boldsymbol{x})$ con una distribución inicial de Mezcla de Árboles de Pólya Finitos.

Algoritmo 5 Distribución marginal $p(\boldsymbol{x})$ con una distribución inicial de Mezcla de Árboles de Pólya Finitos.

Comenzar con los pasos 1 y 2 del Algoritmo 3 para evaluar α_ϵ^* y α_ϵ. Posteriormente, para la distribución marginal, se evalúa únicamente $\psi(\boldsymbol{x})$.

1: Inicializar $L = 0$; para cada $m = 1, \ldots, J$ y cada $j = 2, \ldots, n$ y para cada $\epsilon \in \{0, 1\}^m$ inicializar $n_\epsilon^{(j)} = \mathbb{1}\{\epsilon = \epsilon_m(x_1)\}$.
2: Para $j = 2, \ldots, n$
   Para $m = 1, \ldots, J$
   - Si $n_{\epsilon_{m-1}(x_j)}^{(j)} > 0$, entonces:
     - Sea $\epsilon_0 = \epsilon_{m-1}(x_j)0$ y $\epsilon_1 = \epsilon_{m-1}(x_j)1$
     - Sea $p0 = \alpha_{\epsilon_m(x_j)}^{*(j)} / (\alpha_{\epsilon_{m-1}(x_j)0}^{*(j)} + \alpha_{\epsilon_{m-1}(x_j)1}^{*(j)})$
     - Sea $pr = \alpha_{\epsilon_m(x_j)} / (\alpha_{\epsilon_{m-1}(x_j)0} + \alpha_{\epsilon_{m-1}(x_j)1})$
     - Se actualiza $L = \log(p0/pr)$
   - Para $\ell = j + 1, \ldots, n$, actualizar $n_{\epsilon_m(x_j)}^{(\ell)} = n_{\epsilon_m(x_j)}^{(\ell)} + 1$
3: $\log(\psi(\boldsymbol{x})) = L$.

De acuerdo con Hanson (2006), y utilizando la notación de la descripción anterior, la densidad predictiva para una nueva observación, para un árbol con M niveles, está dada por

$$mj(x) = \left[\prod_{m=1}^{M} \frac{2\alpha_{mj} + 2n_m(x)}{2\alpha_{mj} + n_{m-1}(x)} \right] f_0(x), \tag{2.12}$$

donde $n_m(x)$ es el número de elementos en el conjunto $B_{m,k(x)}$, con $k(x) = \lfloor 2^m F_0(x) + 1 \rfloor$ y $n_0(x) = n$.

Ejemplo. Siguiendo los pasos de los algoritmos 3, 4 y 5, se obtiene la densidad predictiva de una mezcla de las siguientes distribuciones normales:

$$\begin{aligned} \mathrm{N}_1 &\sim \mathrm{N}(-1,\, 0.0625) \\ \mathrm{N}_2 &\sim \mathrm{N}(0,\, 0.0625) \\ \mathrm{N}_3 &\sim \mathrm{N}(1,\, 0.0625), \end{aligned}$$

con probabilidades $(0.2, 0.5, 0.3)$, respectivamente.
Los datos fueron simulados y graficados en el ambiente `R` (R Core Team), generando 1000 observaciones.

Para la distribución inicial, se utilizó un Árbol de Pólya con $m = 8$ niveles y las particiones de los niveles se formaron mediante los cuantiles diádicos de una distribución $\mathrm{N}(0, 1)$ y 20 simulaciones de la distribución predictiva bajo un Árbol de Pólya con dichas características, se tomó el parámetro $\alpha = 0.5$. Debido a que el Árbol de Pólya depende de la partición del espacio muestral que se elija, los parámetros de esta medida fueron establecidos como la media y varianza de las observaciones generadas.

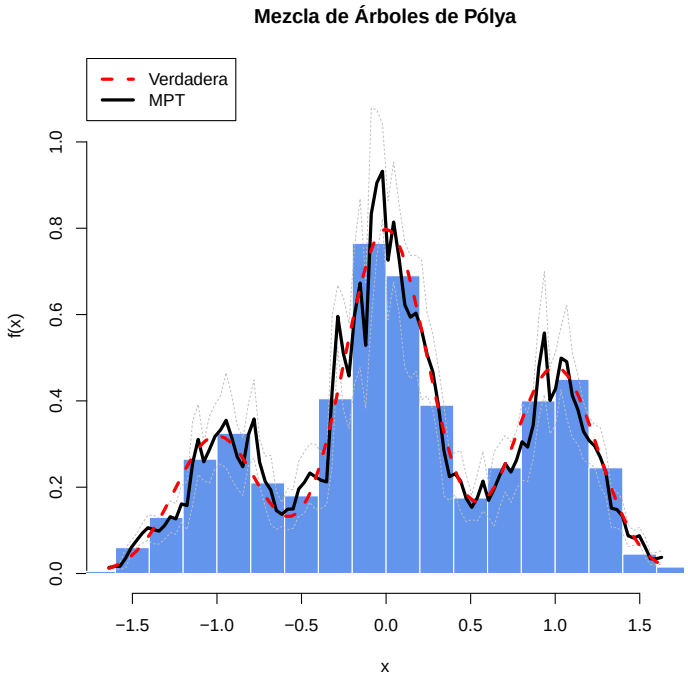

Figura 2.3: Densidad predictiva de una mezcla de distribuciones normales, ajustada con una distribución inicial de Árbol de Pólya.

□

Capítulo 3

El Modelo Normal Proyectado

Un caso importante de los modelos proyectados es cuando $\boldsymbol{X}$ sigue una distribución normal bivariada, con vector de medias $\boldsymbol{\mu}$ y matriz de precisión $\boldsymbol{\Lambda}$ (la inversa de la matriz de covarianzas), $\mathrm{N}_2(\boldsymbol{\mu},\, \boldsymbol{\Lambda})$. En este caso, se dice que que el vector aleatorio $||\boldsymbol{X}||^{-1}\boldsymbol{X}$ tiene una distribución **Normal proyectada**, y se denota por $\mathrm{PN}(\boldsymbol{\mu},\, \boldsymbol{\Lambda})$. En este capítulo se exponen los resultados más importantes para el modelo Normal Proyectado y se aborda el caso particular en el que la matriz de covarianzas $\boldsymbol{\Lambda} = \boldsymbol{I}$. Para este último caso se detalla la inferencia paramétrica y se expone el modelo no-paramétrico; el capítulo concluye con la presentación del modelo propuesto.

3.1. Un caso especial: el modelo $\mathrm{PN}(\boldsymbol{\mu},\, \boldsymbol{I})$

A continuación se presentan algunos resultados asociados al modelo Normal Proyectado y, posteriormente, se presentarán resultados importantes para el caso $\boldsymbol{\Lambda} = \boldsymbol{I}$. Mardia y Jupp (2000) presentan la siguiente definición.

Definición 3.1.1. *La* ***función de densidad de probabilidad de una distribución Normal proyectada****, para un ángulo aleatorio* Ψ*, está dada por*

$$\mathrm{PN}(\boldsymbol{\mu},\, \boldsymbol{\Lambda}) = \frac{\varphi(\psi \,|\, \mathbf{0},\, \boldsymbol{\Lambda}) + |\boldsymbol{\Lambda}|^{-1/2} d(\psi)\Phi(d(\psi))\phi(|\boldsymbol{\Lambda}|^{-1/2}(\boldsymbol{u}^T\boldsymbol{\Lambda}\boldsymbol{u})^{-1/2}\boldsymbol{\mu} * \boldsymbol{u})}{\boldsymbol{u}^T\boldsymbol{\Lambda}\boldsymbol{u}} \mathbb{1}_{(0,\, 2\pi]}(\psi),$$

donde $\varphi(\cdot \,|\, \mathbf{0},\, \boldsymbol{\Lambda})$ *denota la función de densidad de una* $\mathrm{N}_2(\cdot \,|\, \mathbf{0},\, \boldsymbol{\Lambda})$*,* $\Phi(\cdot)$ *y* $\phi(\cdot)$ *denotan las funciones de distribución y de densidad de una Normal estándar, respectivamente,* $\boldsymbol{u} = (\cos\psi,\, \sin\psi)^T$*,*

$$d(\psi) = \frac{\boldsymbol{\mu}^T \boldsymbol{\Lambda}^{-1} \boldsymbol{u}}{(\boldsymbol{u}^T \boldsymbol{\Lambda}^{-1} \boldsymbol{u})^{1/2}}$$

y $\boldsymbol{\mu} * \boldsymbol{u} = \mu_1 \sin\psi + \mu_2 \cos\psi$, *con* $\boldsymbol{\mu} = (\mu_1, \mu_2)$.

La distribución Normal proyectada puede modelar comportamientos unimodales, multimodales, simétricos y/o asimétricos (véase Nuñez-Antonio, 2010).

Definición 3.1.2. *Sea* $\boldsymbol{Y}$ *un vector aleatorio bivariado con distribución* $\mathrm{N}_2(\boldsymbol{\mu}, \boldsymbol{\Lambda})$. *Entonces, la* ***función de densidad de probabilidad*** *de* $\boldsymbol{Y}$ *está dada por*

$$f(\boldsymbol{y} \,|\, \boldsymbol{\mu}, \boldsymbol{\Lambda}) = \frac{|\boldsymbol{\Lambda}|^{1/2}}{2\pi} \exp\left\{ -\frac{1}{2} (\boldsymbol{y} - \boldsymbol{\mu})^T \boldsymbol{\Lambda} (\boldsymbol{y} - \boldsymbol{\mu}) \right\}.$$

Proposición 3.1.3. *Sea* $\boldsymbol{Y}$ *un vector aleatorio bivariado con distribución* $\mathrm{N}_2(\boldsymbol{\mu}, \boldsymbol{\Lambda})$. *Si se define la transformación*

$$\boldsymbol{y} = r(\cos\psi, \sin\psi)^T = r\boldsymbol{v}^T,$$

donde $\psi \in (0, 2\pi]$ *y* $r \in \mathbb{R}^+$. *Entonces, la* ***función de densidad conjunta*** *de la transformación* (r, ψ) *está dada por*

$$f(r, \psi \,|\, \boldsymbol{\mu}, \boldsymbol{\Lambda}) = r\, C_6\, d^2 \exp\left\{ -\frac{1}{2} d^2 r^2 + d^2 b r \right\}, \tag{3.1}$$

donde $d^2 = \boldsymbol{v}^T \boldsymbol{\Lambda} \boldsymbol{v}$, $b = \frac{\boldsymbol{v}^T \boldsymbol{\Lambda} \boldsymbol{\mu}}{\boldsymbol{v}^T \boldsymbol{\Lambda} \boldsymbol{v}}$ *y*

$$C_6 = \frac{|\boldsymbol{\Lambda}|^{1/2}}{2\pi} \frac{\exp\left\{ -\frac{1}{2} \boldsymbol{\mu}^T \boldsymbol{\Lambda} \boldsymbol{\mu} \right\}}{\boldsymbol{v}^T \boldsymbol{\Lambda} \boldsymbol{v}}.$$

Demostración. El determinante de la matriz jacobiana de la transformación $\boldsymbol{y}$ está dado por

$$|J_{\boldsymbol{y}}(r, \psi)| = \begin{vmatrix} \cos\psi & -r\sin\psi \\ \sin\psi & r\cos\psi \end{vmatrix} = r.$$

Así, utilizando el Teorema de cambio de variable:

$$\begin{aligned}
f(r, \psi \,|\, \boldsymbol{\mu}, \boldsymbol{\Lambda}) &= f(r\boldsymbol{v}^T \,|\, \boldsymbol{\mu}, \boldsymbol{\Lambda}) \times |J_{\boldsymbol{y}}(r, \psi)| \\
&= r(2\pi)^{-1}|\boldsymbol{\Lambda}|^{1/2} \exp\left\{-\frac{1}{2}(r\boldsymbol{v} - \boldsymbol{\mu})^T\boldsymbol{\Lambda}(r\boldsymbol{v} - \boldsymbol{\mu})\right\} \\
&= r(2\pi)^{-1}|\boldsymbol{\Lambda}|^{1/2} \exp\left\{-\frac{1}{2}\left[(\boldsymbol{v}^T\boldsymbol{\Lambda}\boldsymbol{v})r^2 - 2(\boldsymbol{v}^T\boldsymbol{\Lambda}\boldsymbol{\mu})r + (\boldsymbol{\mu}^T\boldsymbol{\Lambda}\boldsymbol{\mu})\right]\right\} \\
&= r(2\pi)^{-1}|\boldsymbol{\Lambda}|^{1/2} \exp\left\{-\frac{1}{2}\boldsymbol{\mu}^T\boldsymbol{\Lambda}\boldsymbol{\mu}\right\} \exp\left\{-\frac{1}{2}\boldsymbol{v}^T\boldsymbol{\Lambda}\boldsymbol{v}\left[r^2 - 2\frac{\boldsymbol{v}^T\boldsymbol{\Lambda}\boldsymbol{\mu}}{\boldsymbol{v}^T\boldsymbol{\Lambda}\boldsymbol{v}}r\right]\right\} \\
&= r(2\pi)^{-1}|\boldsymbol{\Lambda}|^{1/2} \exp\left\{-\frac{1}{2}\boldsymbol{\mu}^T\boldsymbol{\Lambda}\boldsymbol{\mu}\right\} \frac{1}{\boldsymbol{v}^T\boldsymbol{\Lambda}\boldsymbol{v}} \times \\
&\quad (\boldsymbol{v}^T\boldsymbol{\Lambda}\boldsymbol{v}) \exp\left\{-\frac{1}{2}(\boldsymbol{v}^T\boldsymbol{\Lambda}\boldsymbol{v})r^2 + (\boldsymbol{v}^T\boldsymbol{\Lambda}\boldsymbol{v})\left(\frac{\boldsymbol{v}^T\boldsymbol{\Lambda}\boldsymbol{\mu}}{\boldsymbol{v}^T\boldsymbol{\Lambda}\boldsymbol{v}}\right)r\right\} \\
&= r\, C_6\, d^2 \exp\left\{-\frac{1}{2}d^2r^2 + d^2br\right\},
\end{aligned}$$

obteniéndose así la ecuación (3.1).

□

Proposición 3.1.4. *Bajo las mismas condiciones de la Proposición 3.1.3, la función de densidad del ángulo aleatorio Ψ, es decir, la densidad de probabilidad de la correspondiente Normal proyectada, está dada por*

$$\begin{aligned}
\mathrm{PN}(\psi \,|\, \boldsymbol{\mu}, \boldsymbol{\Lambda}) &= \int_0^\infty f(r, \psi \,|\, \boldsymbol{\mu}, \boldsymbol{\Lambda})\, dr \\
&= C_6\left[1 + \frac{db}{\phi(db)}\Phi(db)\right]\mathbb{1}_{(0,\,2\pi]}(\psi), \qquad (3.2)
\end{aligned}$$

donde d, b y C_6 son como en la Proposición 3.1.3, $\Phi(\cdot)$ y $\phi(\cdot)$ denotan las funciones de distribución y densidad de una Normal estándar, respectivamente.

Demostración. Se tiene que

$$\begin{aligned} \mathrm{PN}(\psi \,|\, \boldsymbol{\mu},\, \boldsymbol{\Lambda}) &= \int_0^\infty f(r,\, \psi \,|\, \boldsymbol{\mu},\, \boldsymbol{\Lambda})\, dr \\ &= C_6 \int_0^\infty r d^2 \exp\left\{-\frac{1}{2} d^2 r^2\right\} \exp\left\{d^2 b r\right\}\, dr. \end{aligned} \tag{3.3}$$

La ecuación (3.3) puede integrarse por partes, haciendo el cambio de variable

$$u = \exp\left\{d^2 b r\right\} \;\text{ y }\; dv = r d^2 \exp\left\{-\frac{1}{2} d^2 r^2\right\},$$

lo cual implica

$$v = \int r d^2 \exp\left\{-\frac{1}{2} d^2 r^2\right\}\, dr = -\exp\left\{-\frac{1}{2} d^2 r^2\right\}.$$

Por lo tanto,

$$\begin{aligned} \mathrm{PN}(\psi \,|\, \boldsymbol{\mu},\, \boldsymbol{\Lambda}) &= C_6 \int_0^\infty r d^2 \exp\left\{-\frac{1}{2} d^2 r^2\right\} \exp\left\{d^2 b r\right\}\, dr \\ &= \left[-\exp\left\{d^2 b r\right\} \exp\left\{-\frac{1}{2} d^2 r^2\right\}\right]_0^\infty \\ &\quad + \int_0^\infty \exp\left\{-\frac{1}{2} d^2 r^2\right\} \exp\left\{d^2 b r\right\} d^2 b\, dr \\ &= 1 + \int_0^\infty d^2 b \exp\left\{-\frac{(dr - db)^2}{2} + \frac{d^2 b^2}{2}\right\}\, dr. \end{aligned} \tag{3.4}$$

La integral de (3.4) puede simplificarse de la siguiente manera:

$$\begin{aligned} \int_0^\infty d^2 b \exp\left\{-\frac{(dr - db)^2}{2} + \frac{d^2 b^2}{2}\right\}\, dr &= d^2 b \exp\left\{\frac{d^2 b^2}{2}\right\} \int_0^\infty \exp\left\{-\frac{(dr - db)^2}{2}\right\}\, dr \\ &= d^2 b \exp\left\{\frac{d^2 b^2}{2}\right\} \sqrt{2\pi} \int_{db}^\infty \frac{1}{\sqrt{2\pi}} \exp\left\{-\frac{s^2}{2}\right\}\, ds \\ &= db \Phi(db) \left[\phi(db)\right]^{-1}. \end{aligned}$$

Por lo tanto,

$$\mathrm{PN}(\psi \,|\, \boldsymbol{\mu},\, \boldsymbol{\Lambda}) = C_6 \left[1 + \frac{db}{\phi(db)}\Phi(db)\right] \mathbb{1}_{(0,\,2\pi]}(\psi),$$

que coincide con la ecuación (3.2).

□

De las ecuaciones (3.1) y (3.2) puede obtenerse la densidad condicional de R dado Ψ, la cual es necesaria para los procedimientos de inferencia que se necesitarán más adelante.

Corolario 3.1.5. *Bajo las mismas condiciones de la Proposición 3.1.3, la función de densidad condicional de R dado Ψ está dada por*

$$f(r \,|\, \psi,\, \boldsymbol{\mu},\, \boldsymbol{\Lambda}) = \frac{d^2\, r \exp\left\{-\frac{1}{2}d^2[r^2 - 2br]\right\}}{1 + \frac{db}{\phi(db)}\Phi(db)} \mathbb{1}_{(0,\,\infty)}(r). \qquad (3.5)$$

Demostración. Utilizando la definición de probabilidad condicional y las ecuaciones (3.1) y (3.2), se tiene:

$$\begin{aligned} f(r \,|\, \psi,\, \boldsymbol{\mu},\, \boldsymbol{\Lambda}) &= \frac{f(r,\, \psi \,|\, \boldsymbol{\mu},\, \boldsymbol{\Lambda})}{f(\psi \,|\, \boldsymbol{\mu},\, \boldsymbol{\Lambda})} \\ &= \frac{r\, C_6\, d^2 \exp\left\{-\frac{1}{2}d^2 r^2 + d^2 br\right\}}{C_6\left[1 + \frac{db}{\phi(db)}\Phi(db)\right]} \\ &= \frac{d^2\, r \exp\left\{-\frac{1}{2}d^2[r^2 - 2br]\right\}}{1 + \frac{db}{\phi(db)}\Phi(db)}, \end{aligned}$$

con $\psi \in (0,\, 2\pi]$ y $r \in (0,\, \infty)$.

□

Un modelo particularmente importante es el modelo Normap Proyectado cuando $\boldsymbol{\Lambda} = \boldsymbol{I}$. Por lo anterior, a continuación se presentan resultados asociados. Los parámetros del modelo Normal Proyectado no son identificables, pues si $a > 0$ y se toma $\boldsymbol{\mu}^* = a\boldsymbol{\mu}$ y $\boldsymbol{\Lambda}^* = \boldsymbol{\Lambda}/a^2$, las direcciones observadas siguen una distribución $\mathrm{PN}(\boldsymbol{\mu},\, \boldsymbol{\Lambda})$. Para evitar el problema de falta de identificabilidad se pueden tomar matrices de precisión tales que $|\boldsymbol{\Lambda}| = 1$, por ejemplo, $\boldsymbol{\Lambda} = \boldsymbol{I}$ (ver Nuñez-Antonio, 2010).

La distribución Normal proyectada PN($\boldsymbol{\mu}, \boldsymbol{I}$) es unimodal y rotacionalmente simétrica alrededor de su dirección media, $\boldsymbol{\eta}$, la cual se puede mostrar que es igual a $\boldsymbol{\mu}/||\boldsymbol{\mu}||$. Esta afirmación y la siguiente proposición pueden consultarse en Nuñez-Antonio (2010).

Proposición 3.1.6. *Sea $\boldsymbol{Y}$ un vector aleatorio bivariado con distribución de probabilidad* $\mathrm{N}_2(\boldsymbol{\mu}, \boldsymbol{I})$. *Entonces,*

1. $\mathrm{PN}(\psi \mid \boldsymbol{\mu}, \boldsymbol{I}) = \frac{1}{2\pi} \exp\left\{-\frac{1}{2}||\boldsymbol{\mu}||^2\right\} \left[1 + \frac{\boldsymbol{v}^T\boldsymbol{\mu}}{\phi(\boldsymbol{v}^T\boldsymbol{\mu})}\Phi(\boldsymbol{v}^T\boldsymbol{\mu})\right] \mathbb{1}_{(0,\,2\pi]}(\psi)\mathbb{1}_{\mathbb{R}^2}(\boldsymbol{\mu}),$

2. $f(r \mid \psi, \boldsymbol{\mu}, \boldsymbol{I}) = \frac{r \exp\left\{-\frac{1}{2}\left[r^2 - 2(\boldsymbol{v}^T\boldsymbol{\mu})r\right]\right\}}{1 + \frac{\boldsymbol{v}^T\boldsymbol{\mu}}{\phi(\boldsymbol{v}^T\boldsymbol{\mu})}\Phi(\boldsymbol{v}^T\boldsymbol{\mu})} \mathbb{1}_{(0,\,\infty)}(r)\mathbb{1}_{\mathbb{R}^2}(\boldsymbol{\mu}),$

3. $f(r \mid \psi, \boldsymbol{\mu}, \boldsymbol{I})$ *pertenece a una familia exponencial con parámetro canónico* $b = \boldsymbol{v}^T\boldsymbol{\mu}$ *y*

$$\mathrm{E}(R \mid \Psi) = b + \frac{\Phi(b)}{\phi(b) + b\Phi(b)}.$$

3.2. Inferencia vía métodos MCMC para la distribución PN($\psi \mid \boldsymbol{\mu}, \boldsymbol{I}$)

Dada $\boldsymbol{X} \sim \mathrm{N}_2(\boldsymbol{\mu}, \boldsymbol{I})$, puede obtenerse, vía una transformación a coordenadas polares, la densidad conjunta de (Ψ, R), donde $R = ||\boldsymbol{X}||$, de donde se obtiene que $\Psi \sim \mathrm{PN}(\psi \mid \boldsymbol{\mu}, \boldsymbol{I})$. Dada una muestra de ángulos $\{\psi_1, \ldots, \psi_n\}$ que sigue la distribución PN($\psi \mid \boldsymbol{\mu}, \boldsymbol{I}$), el objetivo de la inferencia bayesiana es hacer inferencias sobre $\boldsymbol{\mu}$, basadas en la distribución final $f(\boldsymbol{\mu} \mid \psi_1, \ldots, \psi_n)$. De acuerdo con Nuñez-Antonio y Gutiérrez-Peña (2005), lo anterior podría lograrse si pudiera observarse una muestra $(\Psi_1, R_1), \ldots, (\Psi_n, R_n)$. En la realidad, únicamente se cuenta con la muestra $\{\psi_1, \ldots, \psi_n\}$; La estructura propuesta por Nuñez-Antonio y Gutiérrez-Peña (2005) sugiere tratar a los radios no obsevados $R_1, \ldots, R_n$ como variables latentes.

Utilizando los resultados conocidos para distribuciones normales bivariadas con varianza conocida (ver, por ejemplo, Gelman *et al.*, 1995), se puede proponer para $\boldsymbol{\mu}$ una distribución inicial $\mathrm{N}_2(\boldsymbol{\mu} \mid \boldsymbol{\mu}_0, \lambda_0\boldsymbol{I})$, con $\boldsymbol{\mu}_0 \in \mathbb{R}^2$ y $\lambda_0 > 0$. De esta manera, la densidad condicional completa para $\boldsymbol{\mu}$ está dada por

$$f(\boldsymbol{\mu} \,|\, \psi_1, \ldots, \psi_n, \boldsymbol{r}, \boldsymbol{I}) = \mathrm{N}_2(\cdot \,|\, \boldsymbol{\mu_F}, \boldsymbol{\Lambda_F}), \tag{3.6}$$

donde:

$$\begin{aligned} \boldsymbol{\Lambda_F} &= \lambda_0 \boldsymbol{I} + n\boldsymbol{I}, & (3.7)\\ \boldsymbol{\mu_F} &= \boldsymbol{\Lambda_F}^{-1}(\lambda_0 \boldsymbol{I}\boldsymbol{\mu}_0 + n\boldsymbol{I}\overline{X_{r,\psi}}), & (3.8) \end{aligned}$$

y $\boldsymbol{r} \in \mathbb{R}^{n+}$; $\overline{X_\psi}$ corresponde al vector los promedios aritméticos de las columnas de la matriz $X_{r,\psi}$, con las columnas dadas por:

$$\begin{aligned} [X_{r,\psi}]_{i,1} &= r_i \cos\psi_i \\ [X_{r,\psi}]_{i,2} &= r_i \sin\psi_i, \end{aligned}$$

con $i = 1, \ldots, n$.

De la Proposición 3.1.6, se sabe que

$$f(r \,|\, \psi, \boldsymbol{\mu}, \boldsymbol{I}) = \frac{r \exp\left\{-\frac{1}{2}\left[r^2 - 2(\boldsymbol{v}^T\boldsymbol{\mu})r\right]\right\}}{1 + \frac{\boldsymbol{v}^T\boldsymbol{\mu}}{\phi(\boldsymbol{v}^T\boldsymbol{\mu})}\Phi(\boldsymbol{v}^T\boldsymbol{\mu})} \mathbb{1}_{(0,\infty)}(r)\mathbb{1}_{\mathbb{R}^2}(\boldsymbol{\mu}), \tag{3.9}$$

y de esa manera se puede simular un vector aleatorio $\boldsymbol{R}$ de $f(\boldsymbol{r} \,|\, \psi_1, \ldots, \psi_n, \boldsymbol{\mu})$.

Con las ecuaciones (3.6) y (3.9) se puede implementar un muestreo de Gibbs para obtener una muestra de la distribución posterior conjunta $f(\boldsymbol{\mu}, \boldsymbol{r} \,|\, \psi_1, \ldots, \psi_n)$ y, a su vez, se puede obtener una muestra de la distribución marginal $f(\boldsymbol{\mu} \,|\, \psi_1, \ldots, \psi_n)$.

Para inicializar el Algoritmo de Metropolis-Hastings, se utilizará la *Aproximación Normal Asintótica* (ver, por ejemplo, Christensen *et al.*, 2011), y como densidad de transición a una Normal. Debido a que la distribución Normal tiene como soporte a $\mathbb{R}$ y $r_i \geq 0$, $i = 1, \ldots, n$, se utilizará la transformación $Y = \log R$. Con esto, la función de densidad de $f_Y((y \,|\, \psi, \boldsymbol{\mu}, \boldsymbol{I}))$ está dada por:

$$
\begin{aligned}
f_Y((y \,|\, \psi,\, \boldsymbol{\mu},\, \boldsymbol{I})) &= f_R(e^y \,|\, \psi,\, \boldsymbol{\mu},\, \boldsymbol{I})) \left|\frac{\partial r}{\partial y}\right| \\
&= \frac{e^y \exp\left\{-\frac{1}{2}\left[e^{2y} - 2(\boldsymbol{v}^T\boldsymbol{\mu})e^y\right]\right\}}{1 + \frac{\boldsymbol{v}^T\boldsymbol{\mu}}{\phi(\boldsymbol{v}^T\boldsymbol{\mu})}\Phi(\boldsymbol{v}^T\boldsymbol{\mu})} e^y \\
&= \frac{e^{2y} \exp\left\{-\frac{1}{2}\left[e^{2y} - 2(\boldsymbol{v}^T\boldsymbol{\mu})e^y\right]\right\}}{1 + \frac{\boldsymbol{v}^T\boldsymbol{\mu}}{\phi(\boldsymbol{v}^T\boldsymbol{\mu})}\Phi(\boldsymbol{v}^T\boldsymbol{\mu})}.
\end{aligned}
$$

Por lo tanto,

$$
\log f_Y((y \,|\, \psi,\, \boldsymbol{\mu},\, \boldsymbol{I})) = 2y - \frac{1}{2}\left[e^{2y} - 2(\boldsymbol{v}^T\boldsymbol{\mu})e^y\right] - \log\left\{1 + \frac{\boldsymbol{v}^T\boldsymbol{\mu}}{\phi(\boldsymbol{v}^T\boldsymbol{\mu})}\Phi(\boldsymbol{v}^T\boldsymbol{\mu})\right\}.
$$

Lo anterior implica que

$$
\begin{aligned}
\frac{\partial}{\partial y}\log f_Y((y \,|\, \psi,\, \boldsymbol{\mu},\, \boldsymbol{I})) &= \frac{\partial}{\partial y}\left[2y - \frac{1}{2}\left[e^{2y} - 2(\boldsymbol{v}^T\boldsymbol{\mu})e^y\right] - \log\left\{1 + \frac{\boldsymbol{v}^T\boldsymbol{\mu}}{\phi(\boldsymbol{v}^T\boldsymbol{\mu})}\Phi(\boldsymbol{v}^T\boldsymbol{\mu})\right\}\right] \\
&= 2 - \frac{1}{2}\left[2e^{2y} - 2(\boldsymbol{v}^T\boldsymbol{\mu})e^y\right] \\
&= 2 - e^{2y} + (\boldsymbol{v}^T\boldsymbol{\mu})e^y.
\end{aligned}
$$

A continuación, se procede a encontrar la moda, m, de $\log f_Y((y \,|\, \psi,\, \boldsymbol{\mu},\, \boldsymbol{I}))$, que coincide con el punto en donde $\frac{\partial}{\partial y}\log f_Y((y \,|\, \psi,\, \boldsymbol{\mu},\, \boldsymbol{I})) = 0$. Este punto se utilizará como media de la distribución Normal que se emplee en la Aproximación Normal Asintótica, y está dado por

$$
\begin{aligned}
e^m &= \frac{(\boldsymbol{v}^T\boldsymbol{\mu}) + \sqrt{(\boldsymbol{v}^T\boldsymbol{\mu})^2 + 8}}{2} \\
&\Rightarrow \\
\hat{m} &= \log\left[\frac{(\boldsymbol{v}^T\boldsymbol{\mu}) + \sqrt{(\boldsymbol{v}^T\boldsymbol{\mu})^2 + 8}}{2}\right].
\end{aligned}
$$

Por otro lado, se puede mostrar que

$$
\mathrm{Var}(\hat{m}) = \frac{1}{2 + e^{2\hat{m}}}.
$$

3.3. Inferencia no-paramétrica: el modelo propuesto

Sea $\Omega = \mathbb{S}^1$, el círculo unitario. Debido a que $\mathbb{S}^1$ es separable y completo, es posible especificar un Proceso Libre de Cola, en particular, un Árbol de Pólya, sobre $\mathbb{S}^1$. En este trabajo se propone definir dicho árbol en términos de una medida base $F_0 = \mathrm{PN}(\psi \,|\, \boldsymbol{\mu}_{\boldsymbol{F_0}}, \boldsymbol{I})$, determinando al conjunto de particiones binarias de $\mathbb{S}^1$ en el nivel m, $\Pi = \{B_{\varepsilon_1 \cdots \varepsilon_m}\}$, a partir de los cuantiles de la medida base $F_0^{-1}(k/2^m)$, $k = 0, 1, \ldots, 2^m$. Si $N(\varepsilon)$ denota el entero con representación en base 2 tal que $\varepsilon = \varepsilon_1 \cdots \varepsilon_m \in \bigcup_{m=0}^{\infty} \{0,1\}^m$, se tiene que:

$$B_{\varepsilon_1 \cdots \varepsilon_m} = \left[F_0^{-1}\left(\frac{N(\varepsilon)}{2^m}\right),\, F_0^{-1}\left(\frac{N(\varepsilon)+1}{2^m}\right)\right),$$

al igual que en el caso $\Omega = \mathbb{R}$.

Se define al cuantil de orden $p \in [0,\, 1]$ de $\mathrm{PN}(\psi \,|\, \boldsymbol{\mu}_{\boldsymbol{F_0}}, \boldsymbol{I})$ como el ángulo $\tilde{\psi}_p$ tal que

$$\int_0^{\tilde{\psi}_p} \mathrm{PN}(\psi \,|\, \boldsymbol{\mu}_{\boldsymbol{F_0}}, \boldsymbol{I})\, d\psi = p,$$

pues, de acuerdo con la proposición (3.1.6), $\psi \in (0,\, 2\pi]$. Dicha integral puede aproximarse numéricamente mediante métodos Monte Carlo, por ejemplo, al generar una muestra de ángulos $\psi_1, \ldots, \psi_n \sim \mathrm{PN}(\psi \,|\, \boldsymbol{\mu}_{\boldsymbol{F_0}}, \boldsymbol{I})$ y obtener el cuantil muestral de orden p correspondiente.

Como se mencionó anteriormente, la estructura de un Árbol de Pólya depende de la partición del espacio muestral, por lo que se propone estimar $\boldsymbol{\mu}_{\boldsymbol{F_0}}$ mediante un muestreo de Gibbs basado en las ecuaciones (3.6) y (3.9). Lo anterior no sugiere que los datos observados sigan una distribución $\mathrm{PN}(\psi \,|\, \boldsymbol{\mu}_{\boldsymbol{F_0}}, \boldsymbol{I})$, pero ofrece una estructura conveniente para definir las particiones del espacio muestral que dan lugar al proceso de Árbol de Pólya.

Así, utilizando las ideas del capítulo anterior, se puede definir una medida aleatoria de probabilidad, F_Ψ sobre $(\mathbb{S}^1, \mathcal{B}_{\mathbb{S}^1})$ con una distribución inicial de Árbol de Pólya, denotado por $\mathcal{PT}_\Psi(\Pi_{\mathrm{PN}}, \mathcal{A})$, para una muestra de ángulos $\psi_1, \ldots, \psi_n$, de la siguiente manera:

1. Dada una muestra de ángulos $\psi_1, \ldots, \psi_n$, obtener las condicionales comple-

tas:

$$f(\boldsymbol{\mu} \,|\, \psi_1, \ldots, \psi_n, \boldsymbol{r}, \boldsymbol{I}) = \mathrm{N}_2(\cdot \,|\, \boldsymbol{\mu_F}, \boldsymbol{\Lambda_F})$$

$$f(r \,|\, \psi, \boldsymbol{\mu}, \boldsymbol{I}) = \frac{r \exp\left\{-\frac{1}{2}\left[r^2 - 2(\boldsymbol{v}^T\boldsymbol{\mu})r\right]\right\}}{1 + \frac{\boldsymbol{v}^T\boldsymbol{\mu}}{\phi(\boldsymbol{v}^T\boldsymbol{\mu})}\Phi(\boldsymbol{v}^T\boldsymbol{\mu})} \mathbb{1}_{(0,\infty)}(r)\mathbb{1}_{\mathbb{R}^2}(\boldsymbol{\mu}),$$

con $\boldsymbol{\Lambda_F}$ y $\boldsymbol{\mu_F}$ dadas por las ecuaciones (3.7) y (3.8), respectivamente.

2. Con las condicionales completas obtenidas en el paso anterior, estimar $\boldsymbol{\mu_{F_0}}$ mediante un muestreo de Gibbs.

3. Dada la distribución $F_0 = \mathrm{PN}(\psi \,|\, \boldsymbol{\mu_{F_0}}, \boldsymbol{I})$, para cada $m = 1, \ldots, M$ y para $\epsilon \in \bigcup_{m=0}^{\infty} \{0,1\}^m$, se define $R_\epsilon = F_0^{-1}\left(\frac{N(\epsilon)+1}{2^m}\right)$, el extremo derecho del intervalo de la partición.

4. Evaluar $\mathcal{A}^*$. Para cada $m = 1, \ldots, M$:
   Para $\epsilon \in \bigcup_{m=0}^{\infty} \{0,1\}^m$, sea $\alpha_\epsilon = \alpha_\epsilon^* = \alpha m^2$ (valor inicial de α_ϵ^*). Para $i = 1, \ldots, n$:

   - $\epsilon_m(\psi_i) = \text{mín}\{\epsilon \in \bigcup_{m=0}^{\infty} \{0,1\}^m : R_\epsilon \geq \psi_i\}$ (índice del nivel m que contiene a ψ_i).
   - $\alpha^*_{\epsilon_m(\psi_i)} = \alpha^*_{\epsilon_m(\psi_i)} + 1$ (construyendo $\alpha_\epsilon + n_\epsilon$ para todos los niveles de la partición).
   - Simulación posterior.
     Para cada $m = 0, \ldots, M-1$ y para $\epsilon \in \bigcup_{m=0}^{\infty} \{0,1\}^m$, sea

     $$Y_{\epsilon 0} \sim \mathrm{Beta}(\alpha^*_{\epsilon 0}, \alpha^*_{\epsilon 1}), \text{ y } Y_{\epsilon 1} = 1 - Y_{\epsilon 0}.$$

     Para cada $\epsilon \in \bigcup_{m=0}^{\infty} \{0,1\}^m$,

     $$\mathrm{F}(B_\epsilon) = \prod_{m=1}^{M} Y_{\varepsilon_1 \cdots \varepsilon_m}.$$

Posteriormente, el cálculo de la distribución marginal, $p(\boldsymbol{\psi})$, es análogo al del Algoritmo 5.

Capítulo 4

Aplicaciones

En este capítulo se muestra la implementación numérica del modelo propuesto mediante ejemplos con datos simulados y datos reales.

4.1. Datos Simulados

Como un primer ejemplo se generan 150 observaciones de una mezcla de dsitribuciones normales proyectadas, con vectores de medias $\boldsymbol{\mu_1} = (-1, -1)$, $\boldsymbol{\mu_2} = (0, 0)$, $\boldsymbol{\mu_3} = (1, 1)$, y probabilidades $(0.3, 0.4, 0.3)$, respectivamente. Los datos fueron simulados en el ambiente `R` (R Core Team).

Para la distribución inicial se utilizó un Árbol de Pólya con $m = 8$ niveles y una medida base $F_0 = \text{PN}((-0.0947366, -0.0141922), \boldsymbol{I})$, con lo cual se obtuvieron 10 simulaciones de la densidad predictiva $p(\boldsymbol{\psi})$; se utilizó el parámetro de precisión $\alpha = 1.5$.

4.2. Datos Reales

Se llevó a cabo un estudio de la interacción de especies en la reserva de biósfera "El Triunfo" en México en 2015. El uso de cámaras permitió a ecologistas generar información de actividad temporal (hora del día) para tres especies de animales: pecaríes, tapires y ciervos. Los datos fueron analizados por Nuñez *et al.* utilizando una mezcla de Procesos Dirichlet de distribuciones normales proyectadas para calcular el índice de traslape de las tres especies. De igual manera, Nuñez y Nieto-Barajas (2019) realizaron un análisis de estos datos mediante un Proceso de Árbol de Pólya que consiste en estudiar las componentes direccionales de un Árbol de Pólya Bivariado, en este último artículo

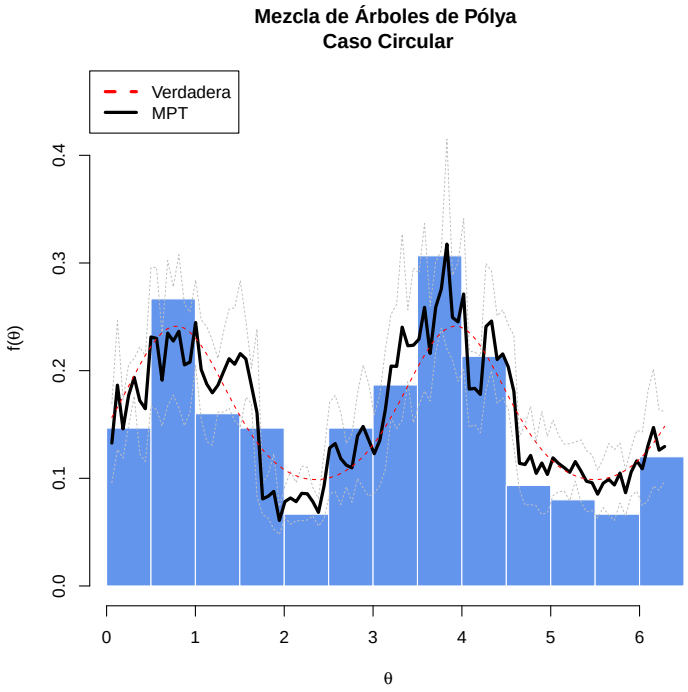

Figura 4.1: Densidad predictiva de una mezcla de distribuciones normales proyectadas, ajustada con una distribución inicial de Árbol de Pólya Proyectado.

se pueden encontrar los datos.

El primer conjunto de datos analizado corresponde a las obervaciones de venados. Se utilizó una distribución inicial de Árbol de Pólya con $m = 8$ niveles y una medida base $F_0 = \text{PN}((0.0746105, -0.2486351), \boldsymbol{I})$; el parámetro de precisión utilizado fue $\alpha = 0.5$.

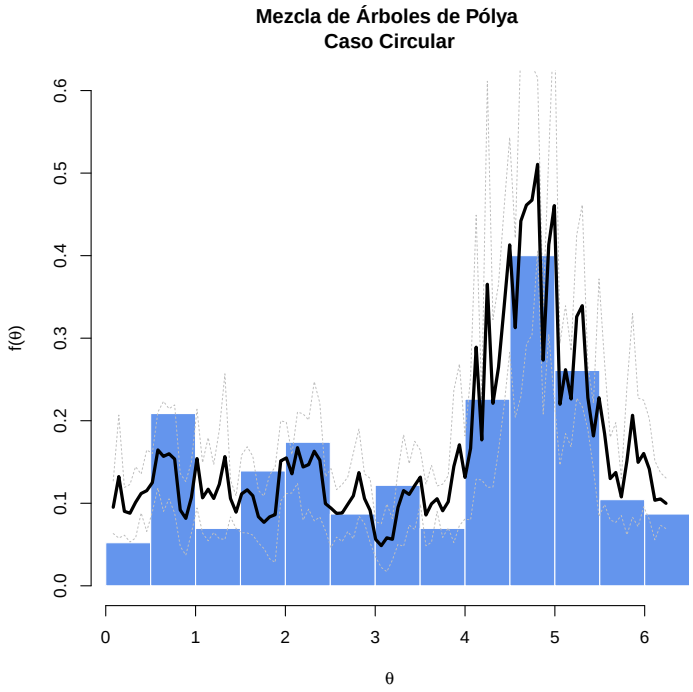

Figura 4.2: Densidad predictiva de los datos correspondientes a observaciones de venados, ajustada con una distribución inicial de Árbol de Pólya Proyectado.

Capítulo 5

Conclusiones

Apéndice. Códigos en R

Representaciones gráficas de Datos Circulares

```
##########################################################
# Ejemplo de diagrama de rosa y diagrama circular.       #
# Direcciones del viento en la región "Col de la "Roa". #
# Tomado del libro de Pewsey, Neuhäuser y Ruxton (2013) #
##########################################################

library(circular)

windc<-circular(wind, type="angles",units="radians",
                template="geographics")

par(mai = c(0.6, 0.8, 0.03, 0.03),
    mgp = c(1.8, 0.9, 0), xpd = T,
    mfrow=c(2,2))

#Histograma sobre la recta real.
hist(wind, col = "darkgray",
     xlab = "Dirección del viento (en radianes)",
     ylab = "Frecuencia", main = "")

#Diagrama circular.
plot(windc, cex=1, bin=720, stack=TRUE,
     sep=0.05, shrink=1.3, pch=20)

#Diagrama de rosa.
rose.diag(windc, bins=16, col="darkgrey",
          cex=1, prop=1.3, add=FALSE, shrink = 1.3)

#Diagrama de rosa y diagrama circular.
```

```
plot(windc, cex=1, bin=720, stack=TRUE,
     sep=0.05, shrink=1.3, pch=20)
rose.diag(windc, bins=16, col="darkgrey", cex=1,
          prop=1.3, add=TRUE, shrink = 1.3)
```

Diferencia entre la Media Aritmética y la Media Trigonométrica

```
#Ángulos de 45 y 315.
ang<-c(pi/4, 7*pi/4)

#Objeto de clase circular.
cir<-circular(ang, units = "degrees", zero = circular(0),
              rotation = "counter")

par(mai = c(0.6, 1, 0.03, 0.03),
    mgp = c(1.8, 1, 0),
    xpd = T, mfrow=c(1,1))

#Se grafica una circunferencia.
plot(cir,stack = TRUE, bins = 1, cex = 0)

#Se agregan flechas con los ángulos deseados.
arrows(x0=0,y0=0,x1=cos(ang[1]),y =sin(ang[1]))
arrows(x0=0,y0=0,x1=cos(ang[2]),y1= sin(ang[2]))

colores<-c("blue", "red")

#Media Trigonométrica.
arrows(x0=0,y0=0,x1=cos(0),y1=sin(0),col=colores[1])

#Media Aritmética.
arrows(x0=0,y0=0,x1=cos(pi),y1=sin(pi),col=colores[2])

legend("bottomleft",
       c("Media trigonométrica", "Media aritmética"),
       fill = colores, border = colores, bty = "n")
```

Algoritmo para generar variables aleatorias de una distribución que sigue un Árbol de Pólya

```
#######################################################
#   Generación de observaciones de v.a.'s de una      #
#   distribución que sigue un Árbol de Pólya.         #
#   La partición del dominio se hace a partir         #
#   de los cuantiles de una N(m,s).                   #
#######################################################

      #Función para generar las realizaciones del árbol
      #Únicamente se graficarán dos cuantiles de las
      #realizaciones
      graf_polya<-function(P,B){#1
        valx<-NULL #Valores eje X
        valy<-NULL #Valores eje Y
        #Se discretiza el eje x
        malla_x<-seq(from=min(datos),to=max(datos),
                     length=tam_malla)
        #Se crea una malla para discretizar los
        #valores de la función de densidad
        malla_f<-rep(NA,tam_malla)
        for(j in 1:2^Nniv){#2
          #Intervalo B[e]=[x0,x1]
          if (j==1){#3
            x0<-min(datos)
                   }#3
          else{#4
            x0<-B[[Nniv]][j-1]
              }#4

          x1<-B[[Nniv]][j]

          #Dens: Probab/Long intervalo
          y01<-P[[Nniv]][j]/(x1-x0)
          #Longitudes de intervalo
          valx<-c(valx,x0,x1)
          #Probabilidades
          valy<-c(valy,y01,y01)
          #Se actualizan las probs condicionales
          idx<-(malla_x>=x0)&(malla_x<=x1)
          malla_f[idx]<-y01
                        }#2
```

```
        return(malla_f)
                                }#1

PT<-function(Nniv,m,s,C,datos,mu,sig,K,niter){#1

  #-----------------------------------------------------#
  #      Mezcla de Árboles de Pólya (MPT)               #
  # Nniv: Número de niveles del árbol                   #
  # m: Media de la medida base                          #
  # s: SD de la medida base                             #
  # C: Parámetro de precisión                           #
  # datos: Muestra para la que se hace la estimación    #
  # mu: Hiperparámetro para la media de m               #
  # sig: Hiperparámetro para la SD de m                 #
  # K: Número de iteraciones del algoritmo MH           #
  # niter: Núm. de iteraciones de la distr. predictiva  #
  #-----------------------------------------------------#

    #----------Histograma----------#
    malla_x<-seq(from=min(datos),to=max(datos),
                 length=tam_malla)
    hist_aux<-hist(datos,nclass=20,plot=F)
    yl<-max(hist_aux$density)*1.5

    hist(datos,probability=TRUE,ylim=c(0,yl),
         xlim=c(min(datos),max(datos)),nclass=20,
         main="Mezcla de Árboles de Pólya",
         col="cornflowerblue",border="white",
         xlab="x",ylab="f(x)")

    malla_F<-matrix(NA,nrow=niter,ncol=tam_malla)

    for (i in 1:niter){#2
      #-------------------------------------------------#
      #Intervalos de la partición
      B<-list()
      for(b in 1:Nniv){#3
        q<-1/(2^b)*(1:(2^b-1)) #Extremos derechos
        B[[b]]<-c(qnorm(q,m,s),max(datos)+0.0001)
        #Se reemplazan "-Inf" e "Inf", para evitar
```

```
  #errores
  B[[b]][B[[b]]==-Inf]<-min(datos)-0.0001
  B[[b]][B[[b]]==Inf]<-max(datos)+0.0001
                }#3
#-----------------------------------------------------#
#Parámetros alpha
n_int<-list() #Número de obs. en el intervalo
alp<-list() #Parámetros
for(k in 1:Nniv){#4
  n_int[[k]]<-rep(0,2^k) #Vector de conteos
  alp[[k]]<-rep(C*k^2,2^k) #Parámetros iniciales
  for(t in 1:length(datos)){#5
    #Conjunto donde se encuentra la
    #t-ésima observación
    j<-min(which(B[[k]]>datos[t]))
    #Actualizar vector de conteos
    n_int[[k]][j]<-n_int[[k]][j]+1
                            }#5
  #Actualizar parámetros
  alp[[k]]<-alp[[k]]+n_int[[k]]
                }#4
#-----------------------------------------------------#
#Simulación posterior y
#actualización de probabilidades
Y<-as.list(1:Nniv) #Probabilidades condicionales
P<-as.list(1:Nniv) #Probabilidad de cada intervalo
#Generar probabilidades condicionales Y y medida F
for(k in 0:(Nniv-1)){#6
  for(j in 1:2^k){#7
    #Cuando m=0, j=1 es el espacio muestral
    j0<-(j-1)*2+1  #B[e_0]
    j1<-(j-1)*2+2  #B[e_1]
    #Actualiza parámetros alfa
    a0<-alp[[k+1]][j0]
    a1<-alp[[k+1]][j1]

    Y0<-rbeta(1,a0,a1)

    #Probabilidades condicionales
    Y[[k+1]][j0]<-Y0
```

```
      Y[[k+1]][j1]<-1-Y0

      #Probabilidades de cada intervalo
      if (k>0){#8
        P[[k+1]][j0]<-Y[[k+1]][j0]*P[[k]][j]
        P[[k+1]][j1]<-Y[[k+1]][j1]*P[[k]][j]
              }#8
      else{#9
        P[[k+1]][j0]<-Y[[k+1]][j0]
        P[[k+1]][j1]<-Y[[k+1]][j1]
          }#9
                }#7
                   }#6
#-------------------------------------------------#
#Probabilidad marginal, psi
#Igual a probabilidad marginal/medida base
#p(x1...xn|eta)/F_0(x1...xn)
neps<-list() #Núm obs en cada nivel
a<-list()
a_star<-list()
for(k in 1:Nniv){#10
  neps[[k]]<-rep(0,2^k) #Inicia conteo
  #Parámetros beta iniciales
  a[[k]]<-rep(C*k^2,2^k)
  a_star[[k]]<-a[[k]]
  j<-min(which(B[[k]]>datos[1]))
  neps[[k]][j]<-1
  a_star[[k]][j]<-a[[k]][j]+1
                }#10
lmarg<-0
for(t in 2:length(datos)){#11
  #Núm de x[j], j<i en el subconjunto
  lag_i<-(i-1)
  for (k in 1:Nniv){#12
    j<-min(which(B[[k]]>datos[t]))
    if (lag_i>0){#13
      lag_i<-neps[[k]][j]
      #Índices de los conjuntos que forman B[m,j]
      j1<-((j+1)%/%2)*2
      j0<-j1-1
```

```
        #Parámetros iniciales
        a0<-a[[k]][j0]
        a1<-a[[k]][j1]
        #Parámetros finales
        as0<-a_star[[k]][j0]
        as1<-a_star[[k]][j1]
        po<-a_star[[k]][j]/(as0+as1)
        pr<-a[[k]][j]/(a0+a1)
        lmarg<-lmarg+log(po/pr)
                  }#13
      #Actualiza conteos y parámetros finales
      neps[[k]][j]<-neps[[k]][j]+1
      a_star[[k]][j]<-a_star[[k]][j]+1
                    }#12
                          }#11
#-----------------------------------------------------#
#Algoritmo MH para los hiperparámetros
lf<-sum(dnorm(datos,m,s,log=TRUE))
for (k in 1:K){#14
  m<-rnorm(1,mu,sig)
  B_mh<-list()
  for(b in 1:Nniv){#15
    q<-1/(2^b)*(1:(2^b-1))
    B_mh[[b]]<-c(qnorm(q,m,s),
                 max(datos)+0.00001)
    #Se reemplazan "-Inf" e "Inf", para evitar
    #errores
    B_mh[[b]][B_mh[[b]]==-Inf]<-min(datos)-0.00001
    B_mh[[b]][B_mh[[b]]==Inf]<-max(datos)+0.00001
                  }#15
  #Probabilidad marginal, psi
  #Igual a probabilidad marginal/medida base
  #p(x1...xn|eta)/F_0(x1...xn)
  neps_mh<-list() #Núm obs en cada nivel
  a_mh<-list()
  a_star_mh<-list()
  for(mm in 1:Nniv){#16
    neps_mh[[mm]]<-rep(0,2^mm) #Inicia conteo
    #Parámetros beta iniciales
    a_mh[[mm]]<-rep(C*mm^2,2^mm)
```

```
      a_star_mh[[mm]]<-a_mh[[mm]]
      j<-min(which(B[[mm]]>datos[1]))
      neps_mh[[mm]][j]<-1
      a_star_mh[[mm]][j]<-a_mh[[mm]][j]+1
                    }#16
    lmarg_mh<-0
    for(ii in 2:length(datos)){#17
      #Núm de x[j], j<i en el subconjunto
      lag_ii<-(ii-1)
      for (mmm in 1:Nniv){#18
        j_mh<-min(which(B[[mmm]]>datos[ii]))
        if (lag_ii>0){#19
          lag_ii<-neps_mh[[mmm]][j_mh]
          #Índices de los conjuntos que forman B[m,j]
          j1_mh<-((j_mh+1)%/%2)*2
          j0_mh<-j1_mh-1
          #Parámetros iniciales
          a0_mh<-a_mh[[mmm]][j0_mh]
          a1_mh<-a_mh[[mmm]][j1_mh]
          as0_mh<-a_star_mh[[mmm]][j0_mh]
          as1_mh<-a_star_mh[[mmm]][j1_mh]
          po_mh<-a_star_mh[[mmm]][j_mh]/(as0_mh+as1_mh)
          pr_mh<-a_mh[[mmm]][j_mh]/(a0_mh+a1_mh)
          lmarg_mh<-lmarg_mh+log(po_mh/pr_mh)
                    }#19
        #Actualiza conteos y parámetros finales
        neps_mh[[mmm]][j_mh]<-neps_mh[[mmm]][j_mh]+1
        a_star_mh[[mmm]][j_mh]<-a_star[[mmm]][j_mh]+1
                        }#18
                              }#17
    lf_mh<-sum(dnorm(datos,m,s,log=T))
    r<-lf_mh+lmarg_mh-lf-lmarg
    #Aceptación MH
    if (log(runif(1))<r){#20
      B<-B_mh
      lmarg<-lmarg_mh
      lf<-lf_mh
                        }#20
                  }#14
  f<-graf_polya(P,B)
```

```
      malla_F[i,]<-f
      #-------------------------------------------------#
                    }#2

      #Se graficarán únicamente los cuantiles de orden
      #0.2 y 0.8 de las realizaciones
      F_aux<-apply(malla_F,2,quantile,
                   probs=c(0.2,0.8),
                   type=8)
      lines(malla_x,F_aux[1,],lwd=0.5,col="gray",lty=2)
      lines(malla_x,F_aux[2,],lwd=0.5,col="gray",lty=2)


      Fbar<-apply(F_aux,2,mean,na.rm=TRUE)
      lines(malla_x,Fbar,col="black",lwd=3)
                                          }#1

##########################################################

##########################################################
### Cómo correr el ejemplo de la mezcla de normales ###
##########################################################

### Parámetros ###
m0<-c(-1,0,1)
s0<-rep(0.25,4)
p0<-c(0.2,0.5,0.3)

N1<-1e3

z<-hist(runif(N1),breaks=c(0,cumsum(p0)),plot=F)$counts

a<-rnorm(z[1],m0[1],s0[1])
b<-rnorm(z[2],m0[2],s0[2])
d<-rnorm(z[3],m0[3],s0[3])

datos<-c(a,b,d)

Nniv<-8
m<-round(mean(datos),0)
```

```r
s<-round(sd(datos),0)
C<-0.5
mu<-m
sig<-1e-2
K<-20
niter<-20
tam_malla<-100

### Ejemplo ###
PT(Nniv=Nniv,m=m,s=s,C=C,datos=datos,
   mu=mu,sig=sig,K=K,niter=niter)

#Graficar verdadera densidad
mezcla_norm<-function(m,s,p,x){
  salida<-as.numeric(crossprod(p,dnorm(x,m,s)))
  return(salida)
                               }

curve(sapply(x,FUN=mezcla_norm,m=m0,s=s0,p=p0),
      add=TRUE,col="red",lwd=3,lty=2)

legend("topleft",c("Verdadera","MPT"),
       col=c("red","black"),lwd=3,
       lty=2:1)

##########################################################
```

Bibliografía

[1] Barron, A., Schervish, M. J., Wasserman, L., *Posterior distributions in nonparametric problems*, Ann Stat, 27, 536–561.

[2] Bernardo, J. M., Smith, A. F. M., *Bayesian Theory*, Wiley: Chichester, 1994.

[3] Box, G. E. P., *Some problems of statistics and everyday life*, Journal of the American Statistical Association, 74, 365, 1–4, 1979.

[4] Christensen, R., Johnson, W., Branscum, A., Hanson, T,, *Bayesian Ideas and Data Analysis: An Introduction for Scientists and Statisticians*, CRC Press, 2011.

[5] Fabius, J., *Asymptotic behavior of Bayes' estimates*, Ann Math Stat, 35:846–856, 1964.

[6] Ferguson, T. S., *A Bayesian analysis of some nonparametric problems*, The annals of statistics, 209–230, 1973.

[7] Ferguson, T. S., *Prior distributions on spaces of probability measures*, The annals of statistics, 615–629, 1974.

[8] Fisher, N. I., *Statistical Analysis of Circular Data*, Cambridge, University Press, 1993.

[9] Freedman, D, *On the asymptotic distribution of Bayes' estimates in the discrete case*, Ann Math Stat, 34:1386–1403, 1963.

[10] Gamerman, D., Lopes, H. F., *Markov Chain Monte Carlo: Stochastic Simulation for Bayesian Inference*, 2a ed. Chapman–Hall, CRC: NY, 2006.

[11] Gelman, A., Carlin, J.B., Stern, H.S., Rubin, D.B., *Bayesian Data Analysis*, Chapman and Hall, 1995.

[12] George, B., Ghosh, K., *A Semiparametric Bayesian Model for Circular-Linear Regression*, Communications in Statistics - Simulation and Computation, 35, 4, 911–923, 2006.

[13] Ghosh, J., Delampady, M., Samanta, T., *An introduction to Bayesian Analysis, Theory and Methods*, Springer Texts in Statistics, 2007.

[14] Gilks, W.R., Richardson, S., Spiegelhalter, D., *Markov Chain Monte Carlo in Practice*, Chapman and Hall, 1996.

[15] Gutiérez-Peña, E. Erderly, A., *Monografía de Estadística Bayesiana*, Universidad Nacional Autónoma de México, 2006.

[16] Hanson, T., Johnson, W. O., *Modeling regression error with a mixture of Pólya trees*, Journal of the American Statistical Association, 97, 460, 1020–1033, 2002.

[17] Hanson, T, *Inference for Mixtures of Finite Pólya Tree Models*, Journal of the American Statistical Association, 101, 476, 1548–1565, 2006.

[18] Hjort, N. L., Holmes, C., Müller, P., Walker, S. G, *Bayesian Nonparametrics*, Cambridge University Press, 2010.

[19] Hoff, Peter D., *A First Course in Bayesian Statistical Methods*, Springer Publishing Company, Incorporated, 2009.

[20] Kraft, C.M., *A class of distribution function processes which have derivatives*, Journal of Applied Probability, 1, 385–388, 1964.

[21] Kent, J. T., *Limiting behaviour of the von Mises-Fisher distribution*, Math. Proc. Cambridge Phil. Soc, 84, 531–536.

[22] Koch, K. R., *Introduction to Bayesian Statistics*, Springer Publishing Company, Incorporated, 2007.

[23] Lavine, M., *Some Aspects of Pólya Tree Distributions for Statistical Modelling*, Annals of Statistics, 20, 3, 1222–1235, 1992.

[24] Lavine, M., *More Aspects of Pólya Tree Distributions for Statistical Modelling*, Annals of Statistics, 22, 3, 1161–1176, 1994.

[25] Mauldin, R. D., Sudderth, W. D., Williams, S. C., *Pólya Trees and Random Distributions*, Annals of Statistics, 20, 3, 1203–1221, 1992.

[26] Mardia, K. V., Jupp, P. E., *Directional Statistics,* Wiley: Chichester, 2000.

[27] Metivier, M., *Sur la construction des mesures aleatoirespresque surement absolument continues par rapport à une mesure donée*, Zeitschrift fur Wahrscheinlichkeitstheorie und Verwandte Gebiete, 20, 332–334, 1971.

[28] Müller, P., Quintana, F., Jara, A., Hanson, T., *Bayesian Nonparametric Data Analysis,* Springer Series in Statistics, 2015.

[29] Müller, P., Mitra, R., *Bayesian Nonparametric Inference - Why and How*, Bayesian Analysis, 8, 2, 269–302, 2013.

[30] Müller, P., Quintana, F., *Nonparametric Bayesian Data Analysis*, Statistical Science, 19,1, 95–110, 2004.

[31] Müller, P., Rodríguez, A., *Nonparametric Bayesian Inference*, NSF-CBMS Regional Conference Series in Probability and Statistics, 9, 1, i–110, 2013.

[32] Nieto-Barajas, L. E., De Alba, E., *Predictive Modeling Applications in Actuarial Science (International Series on Actuarial Science)*, Cambridge: Cambridge University Press, 2014.

[33] Nieto-Barajas, L. E., Müller, P., *Rubbery Pólya Tree*, Annals of Statistics, 20, 1222–1235.

[34] Nuñez-Antonio, G., Gutiérrez-Peña, E., *A Bayesian analysis of directional data using the projected normal distribution*, Journal of Applied Statistics, 32, 10, 995–1001, 2005.

[35] Nuñez-Antonio, G., *Análisis Bayesiano de Modelos Lineales para Datos Direccionales considerando la Distribución Normal bajo Proyección*, Tesis Doctoral, Universidad Autónoma Metropolitana, 2010.

[36] Nuñez-Antonio, G., Gutiérrez-Peña, E., Escarela, G., *A Bayesian regression model for circular data based on the projected normal distribution*, Statistical Modelling, 11, 3, 185–201, 2011.

[37] Nuñez-Antonio, G., Mendoza, M., Contreras-Cristán, A., Gutiérrez-Peña, E., Mendoza, E., *Bayesian nonparametric inference for the overlap of daily animal activity patterns*, Environmental and Ecological Statistics, 25, 471–494.

[38] Nuñez-Antonio, G., Nieto-Barajas, L. E., *Projected Pólya Tree*, arXiv:1902.06020 [stat.ME].

[39] Pewsey, A., Neuhäuser, M., Ruxton, G. D., *Circular Statistics in* R, OUP Oxford, 2013.

[40] R Core Team, *R: A Language and Environment for Statistical Computing*, R Foundation for Statistical Computing, Vienna, Austria, 2017. URL `http://www.R-project.org`.

[41] Rincón, L., *Introducción a los procesos estocásticos*, UNAM, Facultad de Ciencias, 2012.

[42] Robert, C., Casella, G., *Monte Carlo Statistical Methods*, Springer-Verlag New York, 2004.

[43] Robert, C., Casella, G., *Introducing Monte Carlo Methods with R*, Springer-Verlag New York, 2010.

[44] Tanner, M.A., Wong, W.H., *The calculation of posterior distributions by data augmentation*, Journal of the American Statistical Association, 82, 398, 528–540.

[45] Walker, S., Mallick, B.K., *Hierarchical generalized linear models and frailty models with Bayesian nonparametric mixing.*

[46] Walker, S., *Bayesian Nonparametrics (In Damien, P., Dellaportas, P., Polson, N. G., Stephens, D. A.)*, Bayesian Theory and Applications, Oxford University Press, 249–270, 2013.

[47] Walker, S., Damien, P., Laud, P., Smith, A., *Bayesian nonparametric inference for distributions and related functions (with discussion)*, Journal of the Royal Statistical Society, Series B, 61, 3, 485–527, 1999.

[48] Wasserman, L., *All of Statistics: A Concise Course in Statistical Inference*, Springer Publishing Company, Incorporated, 2010.