UNIVERSIDAD NACIONAL AUTÓNOMA
DE MÉXICO

FACULTAD DE CIENCIAS

Inferencia para series de tiempo estacionarias
desde una perspectiva bayesiana no paramétrica

T E S I S

QUE PARA OBTENER EL TÍTULO DE:

Actuario

PRESENTA:

Bernardo Flores López

TUTOR

Dr. Ramsés Humberto Mena Chávez

Ciudad Universitaria, Cd. Mx., 2019

Hoja de Datos del Jurado

1. Datos del alumno Flores López Bernardo (55) 1647 7607 Universidad Nacional Autónoma de México Facultad de Ciencias Actuaría 311152658
2. Datos del tutor Dr. Ramsés Humberto Mena Chávez
3. Datos del sinodal 1 Dra. Lizbeth Naranjo Albarrán
4. Datos del sinodal 2 Dr. Carlos Díaz Ávalos
5. Datos del sinodal 3 Dr. Carlos Erwin Rodríguez Hernández-Vela
6. Datos del sinodal 4 Dra. Michelle Anzarut Chacalo
7. Datos del trabajo escrito Inferencia para series de tiempo estacionarias desde una perspectiva bayesiana no paramétrica 99 p 2019

A mi familia y a mis amigos,
gracias por siempre apoyarme.

Índice general

Introducción

En los últimos años se ha realizado un esfuerzo por construir modelos flexibles para series de tiempo. Los modelos clásicos, como los ARMA o los GARCH, imponen supuestos distribucionales rígidos que resultan difíciles de cumplir en la mayoría de los casos, limitando su aplicabilidad.

El uso de procesos estacionarios es particularmente atractivo, debido a que las simetrías inducidas por la estacionariedad, como aquella descrita por el Teorema de Representación de Maitra[1], permiten obtener una regularidad en la serie que facilita la inferencia y da pie a la predicción.

El desarrollo de la estadística bayesiana no paramétrica (BNP) conllevó la creación de una serie de modelos inspirados por la construcción de Pitt *et al* (2002) de un proceso AR con distribución estacionaria Q dada, utilizando una medida aleatoria como distribución *a priori* para Q.

Las construcciones BNP aportan la flexibilidad necesaria para adaptarse a datos reales, manteniendo la estacionariedad que brinda la estabilidad en el tiempo requerida para el análisis de las series temporales.

Una manera natural de construir un modelo no paramérico para series temporales es construir un proceso autorregresivo tal que su distribución estacionaria sea una mezcla infinita de distribuciones, lo que le brinda flexibilidad. Sin embargo, un problema frecuente con este tipo de modelos es la dimensionalidad; la cual puede inducir expresiones complejas que involucren sumas infinitas.

El primer intento de contruir un modelo de esta clase fue en (Martínez-Ovando y Walker, 2011), en donde se utilizó un proceso beta-Stacy para generar los pesos que se utilizarían para el modelo de mezclas de la distribución estacionaria. Si bien esta construcción era flexible, el algoritmo de estimación resultaba tan complejo que la hacía muy ineficiente.

En (Mena y Walker, 2005) se utilizó una medida aleatoria para la transición de un proceso AR(1), restringiendo a la distribución invariante a una familia paramétrica de distribuciones dada, lo que limitaba la flexibilidad

[1](Maitra, 1977)

del proceso resultante.

Siguiendo la idea expuesta en (Martínez-Ovando y Walker, 2011), Antoniano-Villalobos y Walker (2016) utilizaron una mezcla de procesos Dirichlet para la distribución estacionaria, f, de un proceso de Markov:

$$f(y) = \int k(y \,|\, \theta)\, \mu(\mathrm{d}\theta)$$

$$\mu \sim \mathrm{PD}(G_0, \alpha_0)$$

en donde $k(\cdot \,|\, \theta)$ es densidad para todo θ; G_0 es la medida base y α_0 es el parámetro de concentración, ambos fijos. La diferencia clave con los enfoques previos es la utilización de variables latentes para eliminar los problemas de dimensionalidad en la verosimilitud, simplificando su estimación.

Para la comprensión de este último modelo, en el primer capítulo de esta tesis se expondrá la teoría básica necesaria de probabilidad y estadística; procesos estocásticos y series de tiempo.

Se presentarán las definiciones y propiedades básicas de los procesos estocásticos y posteriormente se tratarán dos de los modelos básicos de series temporales, los procesos ARMA y GARCH, así como sus propiedades, condiciones de estacionariedad y un esbozo de su estimación.

Posteriormente, en el segundo capítulo se dará una introducción a la estadística bayesiana no paramétrica, incluyendo métodos de estimación utilizando métodos Monte Carlo vía cadenas de Markov. Se comenzará con intercambiabilidad y la estadística bayesiana, para después pasar a medidas aleatorias y métodos numéricos.

En el tercer capítulo se expondrá el modelo desarrollado en (Antoniano-Villalobos y Walker, 2016), utilizando una reparametrización propia que evita problemas numéricos en la estimación.

Finalmente, con fin de mostrar su utilidad, en el cuarto capítulo se aplicará este modelo en los máximos diarios de 2016 a 2018 de partículas suspendidas menores a diez micrómetros en la Ciudad de México, las cuales representan una de las fuentes principales de contaminación del aire en la ciudad.

Capítulo 1

Preliminares

1.1. Procesos estocásticos

La teoría presentada en esta sección fue tomada de (Capasso y Bakstein, 2012), (Robert y Casella, 2004) y (Brockwell y Davis, 2016).

Un espacio medible es una pareja $(A, \mathscr{A})$ en donde A es un conjunto no vacío y $\mathscr{A}$ una σ-álgebra de subconjuntos de A. Un espacio de probabilidad es una tripleta $(A, \mathscr{A}, \mathbb{P})$, en donde $(A, \mathscr{A})$ es un espacio medible y $\mathbb{P}$ es una medida de probabilidad.

Definición 1 (Proceso estocástico)**.** Sea $(\Omega, \mathscr{F}, \mathbb{P})$ un espacio de probabilidad, τ un conjunto de índices y $(E, \mathscr{B})$ un espacio medible. Un proceso estocástico $(E, \mathscr{B})$-valuado sobre $(\Omega, \mathscr{F}, \mathbb{P})$ es una familia $\{X_t\}_{t\in\tau}$ de variables aleatorias $X_t : (\Omega, \mathscr{F}) \to (E, \mathscr{B})$ para todo $t \in \tau$.

El conjunto τ representa, generalmente, al tiempo. Cuando τ es un subconjunto numerable de $\mathbb{R}$, se dice que el proceso estocástico es de tiempo discreto; y cuando es algún intervalo de $\mathbb{R}$, a tiempo continuo.

Para cada valor de $t \in \tau$, la aplicación $\omega \to X_t(\omega)$ es una variable aleatoria, mientras que si se fija $\omega \in \Omega$, se obtiene una función $t \to X_\omega(t)$ llamada trayectoria del proceso.

Un modo de estudiar un proceso es mediante sus distribuciones finito dimensionales: en general, si se tiene un proceso estocástico $\{X_t\}_{t\in\tau}$, con τ discreto, a la ley de $(X_1, \ldots, X_n)$, para algún n, se le conoce como distribución finita dimensional, o, de manera abreviada, *fidi*.

Las distribuciones finito dimensionales poseen, en algunos casos, ciertos tipos de simetrías que simplifican su estudio. Una de ellas es la estacionariedad.

Definición 2 (Estacionariedad estricta)**.** Un proceso estocástico se dice

estrictamente estacionario si

$$(X_{t_1}, \ldots, X_{t_n}) \stackrel{d}{=} (X_{t_1+h}, \ldots, X_{t_n+h}), \quad \forall\, h > 0$$

en donde $\stackrel{d}{=}$ denota igualdad en distribución.

Esta propiedad hace que el proceso se mantenga estable a través del tiempo, facilitando la estimación al no importar en qué intervalo de tiempo se tomen las observaciones.

Este no es el único tipo de estacionariedad. Existe una versión menos restrictiva de dicha propiedad, la cual pide solamente igualdad en los primeros dos momentos.

Definición 3 (Estacionariedad débil)**.** Un proceso $\{X_t\}_{t\in\tau}$ se dice *estacionario de segundo orden* o *débilmente estacionario* si

$$\mathbb{E}\left[X_t\right] = \mu, \quad \text{para todo } t$$

y

$$\mathbb{E}\left[X_{t_1} X_{t_2}\right] = \mathbb{E}\left[X_{t_1+h}\, X_{t_2+h}\right], \quad \text{para todos } t_1,\, t_2,\, t_{1+h},\, t_{2+h} \in \tau$$

Cabe destacar que, si un proceso es estrictamente estacionario, entonces también lo es débilmente.

Cuando se modela usando un proceso estocástico, es importante definir la estructura que sigue la dependencia entre sus variables, de tal modo que se logre representar el fenómeno de interés.

Una forma de dependencia muy popular es la markoviana, la cual, de forma intuitiva, nos dice que lo que pasa hoy depende únicamente de lo que pasó ayer, es decir, la medición actual depende sólo de la medición anterior. Esto, de manera formal, se expresa de la siguiente forma:

Definición 4 (Proceso markoviano)**.** Un proceso estocástico $\{Z_t\}$ se dice markoviano si cumple la propiedad de Markov, es decir, si

$$\mathbb{P}(Z_t \in \mathrm{d}x \,|\, Z_{t-1} = z_{t-1}, \ldots, Z_1 = z_1) = \mathbb{P}(Z_t \in \mathrm{d}x \,|\, Z_{t-1} = z_{t-1})$$

En particular, cuando se tiene un proceso markoviano $\{X_t\}$ a tiempo discreto que toma valores en un espacio finito, se dice que $\{X_t\}$ es una cadena de Markov.
La obtención y el análisis de las distribuciones finito dimensionales de un proceso estocástico, y en general, de un proceso markoviano, no son tareas sencillas. La propiedad de Markov hace natural el uso de otro tipo de distribución asociada al proceso que permite caracterizarlo: la función de transición.

Definición 5 (Función de transición)**.** Sea $A \in \mathscr{F}$. A la función

$$P(x, s; t, A) = \mathbb{P}(X_t \in A \,|\, X_s = x), \quad t > s$$

se le llama función de probabilidad de transición o kérnel de transición y describe la probabilidad de que el proceso esté en el conjunto A en el tiempo t dado que en s estuvo en x. Cuando el proceso es a tiempo discreto se puede simplificar la notación eliminando los índices, haciendo

$$P(x, t-1; t, A) = P(x, A)$$

y

$$P(x, 0; n, A) = P^n(x, A) = \int K^{n-1}(y, A) K(x, \mathrm{d}y)$$

1.2. Series de tiempo

Cuando se toman observaciones a través del tiempo se obtiene una serie de tiempo. Estas pueden ser vistas como muestras de una trayectoria del proceso subyacente que genera los datos.

La modelación de una serie de tiempo consiste en describir la estructura probabilística subyacente, haciendo inferencia sobre las transiciones o, si es el caso, la distribución estacionaria del proceso estocástico. La teoría básica para ello presentada en esta sección se basará en (Shiryaev, 1999), (Brockwell y Davis, 2016) y (Bollerslev, 1986).

Si esta estructura se mantiene constante a través del tiempo, se habla

entonces de un proceso estacionario. Este tipo de procesos permite que el modelado sea relativamente simple debido a la estabilidad que presentan, haciéndolos deseables al construir modelos.

En general, es posible discernir tres componentes en cualquier serie de tiempo:

- Un componente de tendencia, que cambia de manera constante, (x)-
- Uno cíclico, ya sea periódico o aperiódico (y).
- Uno irregular, que fluctúa de manera caótica, llamado estocástico (z).

Estos componentes pueden ser combinados de distintas maneras, de manera que los datos observados $\{Xt\}$ pueden ser escritos como

$$X_t = x * y * z,$$

en donde $*$ denota suma, producto, o cualquier otra operación. Aquí se desarrollarán modelos lineales, es decir, modelos que corresponden a combinaciones lineales de estos componentes.

El componente estocástico en los modelos descritos en este capítulo está dado por un tipo especial de proceso estocástico conocido como ruido blanco. Un proceso $\{\varepsilon_t\}$ es un ruido blanco si $\mathbb{E}[\varepsilon_n] = 0$, $\mathbb{E}\left[\varepsilon_n^2\right] < \infty$ y $\mathbb{E}[\varepsilon_n \varepsilon_m] = 0$, para todo $n \neq m$.

Uno de los modelos básicos de series temporales son las medias móviles. En este, se asume que el estado X_t puede ser construido a partir del ruido blanco que genera la aleatoriedad del sistema del siguiente modo:

Definición 6 (Proceso MA)**.** $\{X_t\}$ es un proceso de medias móviles (MA) de orden q si

$$X_t = Z_t + \theta_1 Z_{t-1} + \cdots + \theta_q Z_{t-q} \tag{1.1}$$

en donde $\{Z_t\}$ es un ruido blanco y las $\{\theta_k\}_{k=1}^q$ son constantes.

La figura 1.1 muestra una realización de un proceso MA(1) con parámetro $\theta = 0.8$.

Como las Z_t son un ruido blanco, entonces $\mathbb{E}[X_t] =$, de donde se sigue que un proceso MA es siempre débilmente estacionario, y si las Z_t se toman independientes e idénticamente distribuidas, estrictamente estacionario.

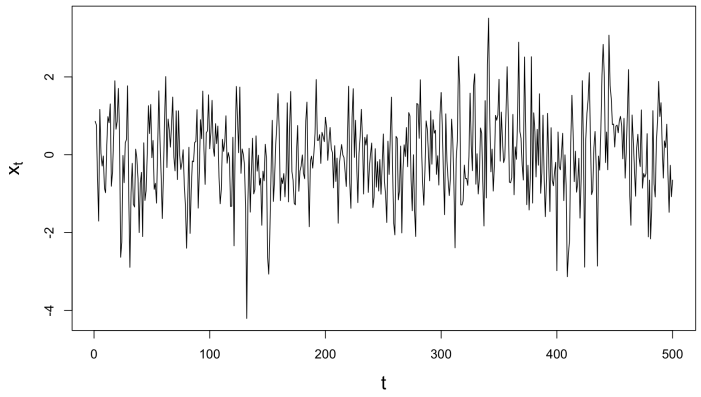

Figura 1.1: Realización de un proceso MA(1)

Poniéndolo en términos del operador de retardo, $B^j(X_t) = X_{t-j}$, la representación anterior puede ser escrita como

$$X_t = \theta(B)Z_t,$$

con $\theta(B) = \sum_0^q \theta_j B^j$.
Este modelo posee propiedades probabilísticas que lo hacen sencillo de manipular, para introducirlas primero enunciemos algunas definiciones.

Definición 7 (Función de autocovarianza)**.** La función de autocovarianza, acv, de un proceso estocástico $\{X_t\}_{t\in\tau}$ se define como

$$\gamma(t,s) = \mathbb{E}\left[(X_t - \mathbb{E}[X_t])(X_s - \mathbb{E}[X_s])\right] = \mathrm{Cov}\left(X_t\,,\,X_s\right)$$

para cualesquiera $t, s \in \tau$.

Esta función adquiere una forma más simple cuando se consideran procesos estacionarios. En particular, procesos débilmente estacionarios: si los

primeros dos momentos son constantes, entonces

$$\begin{aligned}\gamma(t,s) &= \mathrm{Cov}\,(X_t\,,\,X_s)\\ &= \mathrm{Cov}\,(X_{t+h}\,,\,X_{s+h})\end{aligned}$$

En particular, si tomamos $h=-s$,

$$\gamma(t,s) = \mathrm{Cov}\,(X_0\,,\,X_{s-t})$$

Es decir, la autocovarianza sólo depende de la separación que haya entre las variables, de la distancia entre s y t. Entonces, por simplicidad, cambiamos la notación a

$$\gamma(s) = \mathrm{Cov}\,(X_0\,,\,X_s) = \mathrm{Cov}\,(X_t\,,\,X_{t+s})$$

para todo $s,t\in\tau$, y con $t+s\in\tau$.

Nótese que $\gamma(0)=\mathrm{Cov}\,(X_0\,,\,X_0)=\mathrm{Var}\,(X_0)=\mathrm{Var}\,(X_t)=\sigma^2$, para todo $t\in\tau$.

Definición 8 (Función de autocorrelación)**.** Sea $\{X_t\}_{t\in\tau}$ un proceso estocástico. Su función de autocorrelación se define como

$$\rho(s)=\frac{\gamma(s)}{\gamma(0)}$$

Es inmediato que ρ es una función par si el proceso es estacionario:

$$\begin{aligned}\rho(-s) &= \frac{\gamma(-s)}{\gamma(0)}\\ &= \frac{\mathrm{Cov}\,(X_0\,,\,X_{-s})}{\gamma(0)}\\ &= \frac{\mathrm{Cov}\,(X_s\,,\,X_0)}{\gamma(0)}\\ &= \frac{\gamma(s)}{\gamma(0)}=\rho(s)\end{aligned}$$

De aquí sigue que no sea necesario calcular la autocorrelación para valores

negativos de s.
Así, para el caso MA(1), se tienen los siguientes resultados:

$$\gamma(h) = \begin{cases} \sigma^2(1+\theta^2), & \text{si } h = 0 \\ \sigma^2\theta, & \text{si } h = \pm 1 \\ 0, & \text{si } |h| > 1 \end{cases}$$

$$\rho(h) = \begin{cases} 1, & \text{si } h = 0 \\ \frac{\theta}{1+\theta^2}, & \text{si } h = \pm 1 \\ 0, & \text{si } |h| > 0 \end{cases}$$

con $h \geq 0$ y θ el parámetro del modelo.

Definición 9 (Proceso autorregresivo)**.** Un proceso $\{X_t\}$ se llama autorregresivo (AR) de orden p si cumple

$$X_t = \phi_0 + \phi_1 X_{t-1} + \phi_2 X_{t-2} + \cdots + \phi_p X_{t-p} + \varepsilon_t \tag{1.2}$$

en donde las $\{\phi_k\}_{k=1}^{p}$ son constantes dadas y $\{\varepsilon_t\}$ es un ruido blanco.

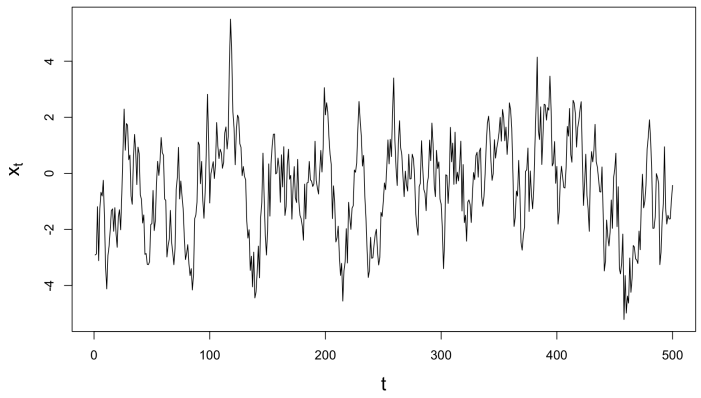

Figura 1.2: Realización de un proceso AR(1)

La figura 1.2 muestra una realización de un proceso AR(1) con parámetro $\phi = 0.8$.
Usando el operador de retardo, (1.2) puede ser reescrito como

$$(1 - \phi_1 B - \cdots - \phi_p B^p)X_n = \phi_0 + \varepsilon_n,$$

o, de manera compacta,

$$\phi(B)X_n = \omega_n,$$

con $\omega_n = \phi_0 + \varepsilon_n$. Para un AR(1), de manera recursiva se puede ver que

$$X_n = \phi_0(1 + X_1 + \cdots + X_1^{n-1}) + \phi_1^n X_0 + \varepsilon_n + \phi_1\varepsilon_{n-1} + \cdots + \phi_1^{n-1}\varepsilon_1,$$

en donde X_0 es una variable aleatoria independiente del ruido que modela el estado inicial del proceso. De esta manera se puede ver que las propiedades de X_n dependen del valor de ϕ_1, para el cual se distinguen tres casos: $|\phi_1| < 1$, $|\phi_1| = 1$ y $|\phi_1| > 1$.

De la representación anterior se obtiene[1] que

$$\mathbb{E}\left[X_n\right] = \phi_1^n \mathbb{E}\left[X_0\right] + \phi_0\left(1 + \phi_1 + \cdots + \phi_1^{n-1}\right)$$

$$\mathrm{Var}\left(X_n\right) = \phi_1^{2n}\mathrm{Var}\left(X_0\right) + \sigma^2\left(1 + \phi_1^2 + \cdots + \phi_1^{2(n-1)}\right)$$

$$\mathrm{Cov}\left(X_n\,,\,X_{n-k}\right) = \theta_1^{2n-k}\mathrm{Var}\left(X_0\right) + \sigma^2\phi_1^k\left(1 + \phi_1^2 + \cdots + \phi_1^{2(n-k-1)}\right),$$

para $n - k \geq 1$, y con σ^2 la varianza del ruido blanco.

De aquí se sigue que si $|\phi_1| < 1$ y $\mathbb{E}\left[|X_0|\right] < \infty$, entonces

$$\mathbb{E}\left[X_n\right] = \phi_1^n \mathbb{E}\left[X_0\right] + \frac{\phi_0(1 - \phi_1^n)}{1 - \phi_1} \to \frac{\phi_0}{1 - \phi_1}$$

cuando $n \to \infty$ y, si el segundo momento es finito,

$$\mathrm{Var}\left(X_n\right) = \phi_1^{2n}\mathrm{Var}\left(X_0\right) + \frac{\sigma^2(1 - \phi_1^{2n})}{1 - \phi_1^2} \to \frac{\sigma^2}{1 - \phi_1^2}$$

[1] (Shiryaev, 1999)

$$\mathrm{Cov}\,(X_n\,,\,X_{n-k}) \to \frac{\sigma^2\phi_1^k}{1-\phi_1^2}$$

Esto quiere decir que el proceso $\{X_n\}$ se aproxima a un estado estable cuando $n \to \infty$ y $|\phi_1| < 1$. Más aún, si la distribución inicial es gaussiana, es decir,

$$X_0 \sim \mathrm{N}\left(\frac{\phi_1}{1-\phi_1}, \frac{\sigma^2}{1-\phi_1^2}\right),$$

entonces $\{X_n\}$ es una secuencia gaussiana estacionaria, con

$$\mathbb{E}\,[X_n] = \frac{\phi_0}{1-\phi_1}, \quad \mathrm{Var}\,(X_n) = \frac{\sigma^2}{1-\phi_1^2}$$

$$\mathrm{Cov}\,(X_n\,,\,X_{n+k}) = \frac{\sigma^2\phi_1^k}{1-\phi_1^2}$$

De aquí se sigue que

$$\gamma(h) = \phi^h\gamma(0), \qquad \text{y}$$

$$\rho(h) = \phi^h$$

El caso $|\phi_1| = 0$ corresponde a una caminata aleatoria, y cuando $|\phi_1| > 1$ el proceso es explosivo, es decir, tanto la esperanza como la varianza incrementan exponencialmente a la par de n.

Para un modelo AR(p) general, consideremos la factorización

$$1 - \phi_1 B - \phi_2 B^2 - \cdots - \phi_p B^p = (1-\lambda_1 B)(1-\lambda_2 B)\cdots(1-\lambda_p B),$$

en donde las λ_i son todas distintas.

Si $|\lambda_i| < 1$ para toda i, la solución estacionaria de

$$(1 - \theta_1 B - \theta_2 B^2 - \cdots - \theta_p B^p)X_n = \omega_n$$

como sigue:

$$X_n = (1-\lambda_1 B)^{-1}\cdots(1-\lambda_p B)^{-1}\omega_n$$

Nótese que las $\lambda_1, \lambda_2, \ldots, \lambda_p$ son las raíces de la ecuación

$$\lambda^p - \phi_1\lambda^{p-1} - \cdots - \phi_{p-1}\lambda - \phi_p = 0$$

O, de otro modo, $\lambda_i = z_i^{-1}$, en donde las z_i son las raíces de la ecuación $1 - \phi_1 z - \phi_2 z - \cdots - \phi_p z_p = 0$. De aquí se sigue que $\{X_n\}$ es estacionaria si todas las raíces quedan fuera del círculo unitario.

Si hacemos

$$c_i = \frac{\lambda_i^{p-1}}{\prod_{\substack{1\leq k\leq p \\ k\neq i}}(\lambda_i - \lambda_k)},$$

entonces

$$X_n = \sum_{l=0}^{\infty}(c_1\lambda_1^l + \cdots + c_p\lambda_p^l)\omega_{n-l}, \tag{1.3}$$

de donde se sigue que

$$\gamma(0) = \phi_1\gamma(1) + \cdots + \phi_p\gamma(p) + \sigma^2$$

$$\gamma(k) = \phi_1\gamma(k-1) + \cdots + \phi_p\gamma(k-p)$$

Para $k = 1, 2, \ldots$. La función de autocorrelación, $\rho(k), k \geq 0$, satisface las mismas ecuaciones, conocidas como *ecuaciones de Yule-Walker*.

Definición 10 (Proceso ARMA)**.** Al proceso $\{Y_t\}$ dado por

$$Y_t + \phi_i Y_{t-1} + \cdots + \phi_p Y_{t-p} = \varepsilon_t + \theta_1\varepsilon_{t-1} + \cdots + \theta_q\varepsilon_{t-q}$$

en donde $\{\varepsilon_t\}$ es un ruido blanco se le llama proceso ARMA(p,q).

Usando el operador de retardo, el proceso puede ser escrito como

$$\phi(B)Y_t = \theta(B)\varepsilon_t,$$

en donde $\phi(B)$ y $\theta(B)$ son polinomios de orden p y q, respectivamente.
En la figura 1.3 se muestra una trayectoria de un proceso ARMA(1,1) con parámetros $\phi = 0.5$ y $\theta = 0.5$.

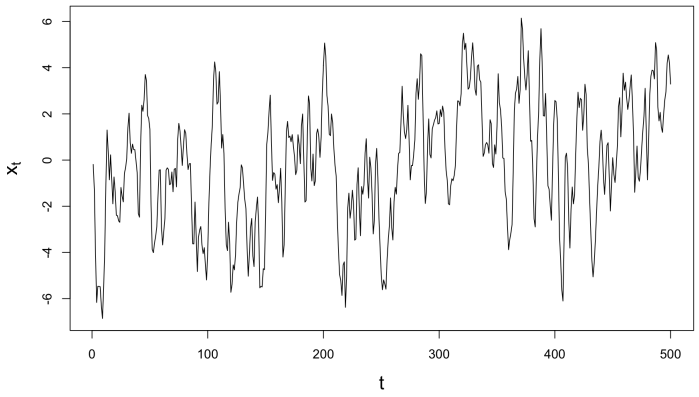

Figura 1.3: Realización de un proceso ARMA(1,1)

Para poder hacer predicciones con un modelo ARMA, es necesario que se cumpla una propiedad conocida como causalidad, la cual garantiza la predictibilidad del proceso.

Definición 11 (Causalidad)**.** Un proceso ARMA(p, q) $\{X_t\}$ es una función causal de $\{Z_t\}$, o simplemente, causal, si existen constantes $\{\psi_j\}$ tal que $\sum_{j=0}^{\infty} |\psi| < \infty$ y

$$X_t = \sum_{j=0}^{\infty} \psi_j Z_{t-j}, \quad \forall t$$

Lo que es equivalente a la condición

$$\phi(z) = 1 - \phi_1 z - \cdots - \phi_p z^p \neq 0, \text{para todo } |z| \leq 1$$

De manera intuitiva, que un proceso sea causal significa que sólo depende de la información pasada, o lo que es equivalente, que no depende del futuro, lo que lo hace previsible por $\{Z_t\}$.

La suposición de causalidad sirve para calcular la función de autocovarianza de un proceso ARMA, la cual puede ser derivada de la siguiente forma (Brockwell, 2003):

Primero tomamos el proceso ARMA(p, q) causal dado por

$$\phi(B)X_t = \theta(B)Z_t, \quad \{Z_t\} \sim \text{WN}(0, \sigma^2)$$

con $\phi(z) = 1 - \phi_1 z - \cdots - \phi_p z^p$ y $\theta(z) = 1 + \theta_1 z + \cdots + \theta_q z^q$. Como es causal, entonces

$$X_t = \sum_{j=0}^{\infty} \psi_j Z_{t-j},$$

en donde $\sum_{j=0}^{\infty} \psi_j z^j = \theta(z)/\phi(z),\ |z| \leq 1$. Los coeficientes ψ pueden calcularse de la manera siguiente:

$$\psi(z) = \sum_{j=0}^{\infty} \psi_j z^j = \theta(z)/\phi(z)$$

si y sólo si

$$(1 - \phi_1 z. \cdots - \phi_p z^p)(\psi_0 + \psi_1 z + \cdots) = 1 + \theta_1 z + \cdots + \theta_q z^q$$

Igualando a los coeficientes de las $z^j, j = 0, 1, \ldots,$ llegamos a que

$$\begin{aligned}
1 &= \psi_0 \\
\theta_1 &= \psi_1 - \psi_0\phi_1 \\
\theta_2 &= \psi_2 - \psi_1\phi_1 - \psi_0\phi_2 \\
&\vdots \\
\theta_j &= \psi_j - \sum_{k=1}^{p} \psi_k\psi_{j-k} \\
&\vdots
\end{aligned}$$

en donde $\theta_0 := 1, \theta_j := 0$ para $j > q$ y $\psi_j := 0$ para $j < 0$.

Una simetría importante dentro de los procesos estocásticos es la de

invertibilidad.

Definición 12 (Invertibilidad)**.** Un proceso ARMA $\{X_t\}$ es invertible si existe una sucesión absolutamente sumable $\{\pi_j\}$ tal que

$$\varepsilon_t = \sum_{i=1}^{\infty} \pi_i X_{t-i}$$

Esta condición puede verse como que el proceso pueda representarse como un AR(∞). De la siguiente proposición se desprende una condición más computable para saber si un proceso ARMA es invertible o no.

Proposición 1. *Un proceso ARMA(p, q) es invertible si y sólo si* $\theta(B) \neq 0$ *para* $|z| \leq 1$*, y los coeficientes* π_i *de la representación AR(*∞*) se determinan resolviendo*

$$\pi(z) = \sum_{j=1}^{\infty} \pi_j z^j = \frac{\phi(z)}{\theta(z)}, \quad |z| \leq 1$$

Es decir, un proceso ARMA es invertible cuando las raíces de $\theta(z)$ están fuera del círculo unitario.

De aquí se sigue que, claramente, los procesos autorregresivos siempre son invertibles, y los MA sólo bajo las condiciones de la proposición anterior.

Si nombramos a φ_i como la i-ésima raíz del polinomio $\phi(z)$, y a ϑ_i como la $i-$ésima raíz del polinomio $\theta(z)$,

Proceso	Invertible	Estacionario
AR(p)	Siempre	$\lvert\varphi_i\rvert < 1$, $\forall i$
MA(q)	$\lvert\vartheta_i\rvert < 1$, $\forall i$	Siempre

Se puede demostrar que la autocovarianza de un ARMA causal está dada por la siguiente fórmula.

$$\gamma(h) = \mathbb{E}\left[X_{t+h} X_t\right] = \sigma^2 \sum_{j=0}^{\infty} \psi_j \psi_{j+h}, \ h > 0$$

Definición 13 (Función de autocorrelación parcial)**.** La función de autoco-

rrelación parcial (PACF) de un proceso estocástico $\{X_t\}$ está definida como

$$\phi_{kk} = \text{Corr}\left(X_t, X_{t+k} \,|\, X_{t+1}, \ldots, X_{t+k-1}\right),$$

es decir, la PACF es el coeficiente de correlación entre X_t y X_{t+k} dadas todas las observaciones intermedias.

En el caso de un proceso ARMA, la PACF ϕ_{hh} es el último componente de $\phi_h = \Gamma_h^{-1}\gamma_h$, con $\Gamma_h = [\gamma(i-j)]_{i,j=1}^h$ y $\gamma_h = [\gamma(1), \gamma(2), \ldots, \gamma(h)]'$.

La utilidad de la función de autocorrelación parcial yace en la identificación de modelos. La siguiente tabla (Madsen, 2008) muestra el comportamiento de la ACF y de la PACF para procesos AR, MA y ARMA. Así,

	ACF $\rho(k)$	PACF ϕ_{kk}
AR(p)	Decrecimiento exponencial y/o sinusoidal	$\phi_{kk} = 0$ para $k > p$
MA(q)	$\rho(k) = 0$ para $k > q$	Dominada por funciones exponenciales y/o sinusoidales amortiguadas
ARMA(p, q)	Decaimiento exponencial y/o sinusoidal después del lag $q - p$	Dominada por funciones exponenciales y/o sinusoidales después del lag $p - q$

un método visual para identificar el modelo apropiado consiste en graficar la función de autocorrelación y la de autocorrelación parcial mediante un correlograma, y verificar los patrones que se forman. Un ejemplo de esto se ve en la figura 1.4, en donde se ve un comportamiento de ambas funciones acorde a un proceso ARMA(1,1).

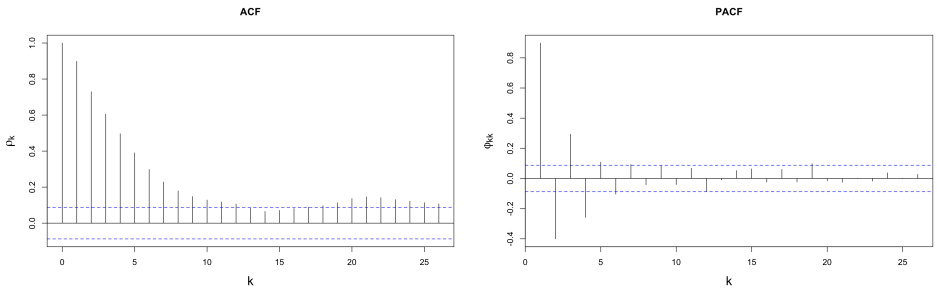

Figura 1.4: A la izquierda, el ACF de un proceso ARMA(1,1); a la derecha, su PACF.

Para realizar predicciones de un ARMA(p, q) $\{X_t\}$ causal e invertible, expresamos el proceso en forma MA:

$$X_t = \varepsilon_t + \psi_1 \varepsilon_{t-1} + \psi_2 \varepsilon_{t-2} + \cdots$$

Equivalentemente,

$$X_{t+k} = \varepsilon_{t+k} + \psi_1 \varepsilon_{t+k-1} + \psi_2 \varepsilon_{t+k-2} + \cdots \tag{1.4}$$

Y si lo escribimos en forma inversa,

$$\varepsilon_t = X_t + \pi_1 X_{t-1} + \pi_2 X_{t-2} + \cdots ,$$

lo que significa que los $\varepsilon_i, i = t, t-1, \ldots$ son conocidos cuando las $X_i, i = 1, 2, \ldots$ lo son. Entonces, si $\{\varepsilon_t\}$ es un ruido blanco se obtiene

$$\mathbb{E}\left[\varepsilon_{t+k} \,|\, X_t, X_{t-1}, \ldots\right] = \begin{cases} \varepsilon_{t+k} & \text{para } k \leq 0 \\ 0 & \text{para } k > 0 \end{cases} \tag{1.5}$$

Si usamos como predictor a la esperanza condicional, de (1.4) obtenemos

$$\hat{X}_{t+k\,|\,t} = \mathbb{E}\left[X_{t+k} \,|\, X_t, X_{t-1}, \ldots\right] = \psi_k \varepsilon_t + \psi_{k+1} \varepsilon_{t-1} \tag{1.6}$$

El error de predicción se obtiene restando (1.4) y (1.6),

$$e_{t+k\,|\,t} = X_{t+k} - \hat{X}_{t+k\,|\,t} = \varepsilon_{t+k} + \psi_1\varepsilon_{t+k-1} + \cdots + \psi_{k-1}\varepsilon_{t+1}$$

por lo que la varianza del error de predicción es

$$\sigma_k^2 = (1 + \psi_1^2 + \cdots + \psi_{k-1}^2)\sigma_\varepsilon^2$$

En particular, si asumimos que el ruido blanco es Gaussiano, se obtiene el siguiente intervalo de confianza para X_{t+k}

$$\hat{X}_{t+k\,|\,t} \pm z_{\alpha/2}\,\sigma_k = \hat{X}_{t+k\,|\,t} \pm z_{\alpha/2}\,\sigma_\varepsilon\sqrt{1 + \psi_1^2 + \cdots + \psi_{k-1}^2},$$

en donde $z_{\alpha/2}$ es el cuantil $\alpha/2$ de la normal estándar y α es el nivel de significancia. Las constantes ψ_i no son conocidas, sino estimadas con base en las observaciones.

GARCH

La dependencia introducida por los modelos ARMA indica que los datos x_t son generados de la densidad condicional $f(x_t\,|\,x_{t-1})$, lo que hace que la varianza dependa de los datos anteriores, $\mathrm{Var}\,(x_t\,|\,x_{t-1})$. Sin embargo, en estos modelos se asume una variabilidad constante a lo largo de la serie, lo que resulta restrictivo.

Definición 14 (ARCH (Engle, 1982))**.** Sea $\{Y_t\}$ una serie de tiempo y $\{\varepsilon_t\}$ un ruido blanco. Decimos que Y_t es un proceso autorregresivo condicionalmente heterocedástico de orden q, ARCH(q), si se cumplen las siguientes condiciones:

$$\begin{aligned} Y_t &= \sigma_t\varepsilon_t \\ \sigma_t^2 &= \alpha_0 + \sum_{i=1}^{q}\alpha_i Y_{t-i}^2 \end{aligned} \tag{1.7}$$

con $\alpha_0 > 0$ y $\alpha_i \geq 0$.

Es decir, las observaciones pueden ser desagregadas en un componente

estocástico y en uno no aleatorio dependiente del tiempo; este último dependiendo de manera autorregresiva del cuadrado de las observaciones.

Es posible extender las propiedades del AR a los ARCH(q).

Teorema 1. *Un proceso ARCH(q) es débilmente estacionario si y sólo si las raíces de su ecuación característica asociada caen todas dentro del círculo unitario.*

Sea ψ_{t-1} la información disponible a tiempo $t-1$. La esperanza y varianza condicional están dadas por las siguientes expresiones:

$$\mathbb{E}\left[Y_t \,|\, \psi_{t-1}\right] = \mathbb{E}\left[\sigma_t \varepsilon_t \,|\, \psi_{t-1}\right] = \sigma_t \,\mathbb{E}\left[\varepsilon_t \,|\, \psi_{t-1}\right] = 0$$

$$\mathrm{Var}\left(Y_t\right) = \frac{\alpha_0}{1 - \sum_{i=1}^{q} \alpha_i}$$

Ahora, para el caso de la esperanza, se tiene que

$$\mathbb{E}\left[Y_t\right] = \mathbb{E}\left[\mathbb{E}\left[Y_t \,|\, \psi_{t-1}\right]\right] = \mathbb{E}\left[0\right] = 0$$

La autocovarianza es cero siempre, pues

$$\begin{aligned} \mathrm{Cov}\left(Y_t\,,\, Y_s\right) &= \mathrm{Cov}\left(\sigma_t \varepsilon_t\,,\, \sigma_s \varepsilon_s\right) \\ &= \sigma_t\, \sigma_s\, \mathrm{Cov}\left(\varepsilon_t\,,\, \varepsilon_s\right) \\ &= 0 \end{aligned}$$

Con estos resultados e imponiendo el supuesto de distribucional $Y_t \,|\, \psi_{t-1} \sim \mathrm{N}\left(0, \sigma_t^2\right)$, se puede obtener la función de verosimilitud.

Como la autocovarianza del proceso es 0 siempre, la log-verosimilitud resulta simplemente la suma de las log-verosimilitudes gaussianas correspondientes a las fidis marginales. Entonces, si l_t es la log-verosimilitud de la t-ésima observación de una muestra de tamaño n de un proceso ARCH(q),

se tiene

$$\begin{aligned} l &= \frac{1}{n}\sum_{i=1}^{n} l_i \\ l_i &= -\frac{1}{2}\log\sigma_i^2 - \frac{1}{2}\frac{Y_i^2}{\sigma_i^2} + \text{constantes} \end{aligned} \tag{1.8}$$

Entonces para estimar los parámetros α, maximizamos l. El gradiente queda

$$\frac{\partial l_t}{\partial \alpha} = \frac{1}{2\sigma_t^2} z_t \left(\frac{Y_t^2}{\sigma_t^2} - 1 \right)$$

con $z_t = (1, Y_{t-1}^2, \ldots, Y_{t-q}^2)$.
Así, para encontrar los estimadores máximo verosímiles para λ, igualamos el gradiente a cero y se resuelve numéricamente el sistema de ecuaciones correspondiente.

La estructura relativamente simple del modelo ARCH lo hace sencillo de utilizar; sin embargo, en la práctica es común que se necesiten modelos más generales que permitan capturar la complejidad exhibida en la volatilidad de, por ejemplo, los activos financieros.

En 1986 Tim Bollerslev introdujo los *modelos autorregresivos condicionalmente heterocedásticos generalizados*, (GARCH), en los cuales se introduce un componente autorregresivo de las varianzas pasadas a la varianza actual.

Definición 15 (GARCH)**.** Se dice que una serie de tiempo $\{X_t\}$ es un proceso autorregresivo condicionalmente heterocedástico generalizado de órdenes p y q, GARCH(p, q), si

$$\begin{aligned} X_t &= \sigma_t \varepsilon_t \\ \sigma_t^2 &= \alpha_0 + \sum_{i=1}^{q} \alpha_i X_{t-i}^2 + \sum_{j=1}^{p} \beta_j \sigma_{t-j}^2 \end{aligned} \tag{1.9}$$

en donde $\{\varepsilon_t\}$ es un ruido blanco, $\alpha_0 > 0$ y $\alpha_i, \beta_j \geq 0$.

En la figura 1.5 se muestra la trayectoria de un proceso GARCH(1,1).

Nótese que al hacer las $\beta_j = 0$, el proceso se degenera a un ARCH(q).

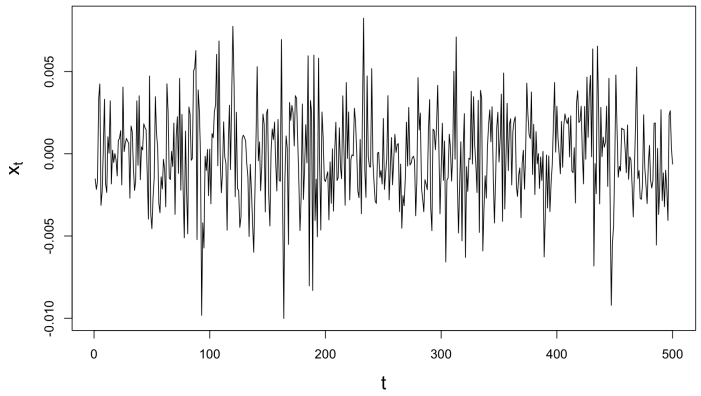

Figura 1.5: Trayectoria de un proceso GARCH(1,1)

Concentrémonos ahora en el modelo más simple, el GARCH(1,1). Este tiene la siguiente forma:

$$\begin{cases} X_t &= \sigma_t \, \varepsilon_t \\ \sigma_t^2 &= \alpha_0 + \alpha X_{t-1}^2 + \beta \sigma_{t-1}^2 \end{cases}$$

con $\{\varepsilon_t\}$ un ruido blanco de varianza unitaria, $\alpha_0 > 0$ y $\alpha, \beta > 0$. Las condiciones de estacionariedad del proceso está garantizada por el siguiente teorema:

Teorema 2. *Si* $0 \leq \gamma = \mathbb{E}\left[\log(\alpha \varepsilon_t^2 + \beta)\right] < \infty$, *entonces la serie*

$$h_t = \left\{ 1 + \sum_{i=1}^{\infty} a(\varepsilon_{t-1}) \cdots a(\varepsilon_{t-i}) \right\} \alpha_0,$$

con $a(x) = \alpha x^2 + \beta$, *converge casi seguramente y el proceso* $\{Z_t\}$ *definido por* $Z_t = \sqrt{h_t}\varepsilon_t$ *es la única solución estrictamente estacionaria del modelo GARCH(1,1).*

De aquí se sigue que, si $\gamma \geq 0$ entonces no existe ninguna solución estrictamente estacionaria.

Si bien los GARCH ofrecen flexibilidad de modelado, en la práctica resulta restrictivo asumir que una serie observada es producto de un ruido. Un forma popular de resolver este inconveniente es mediante el uso de un modelo ARMA-GARCH, en el cual el proceso GARCH no es observado directamente, sino que constituye la innovación de un proceso ARMA. Así, si se tiene una serie de observaciones $X_1, \ldots, X_n$, un modelo ARMA(P, Q)-GARCH(p, q) para ellas se plantea de la siguiente forma:

$$\begin{cases} X_t - c & = \sum_{i=1}^{P} a_{0i} \left(X_{t-i} - c\right) + e_t - \sum_{j=1}^{Q} b_{0j} e_{t-j} \\ e_t & = \sqrt{h_t}\eta_t \\ h_t & = \omega_0 + \sum_{i=1}^{q} \alpha_{0i} e_{t-i}^2 + \sum_{j=1}^{p} \beta_0 h_{t-j} \end{cases}$$

El vector de parámetros se denota como

$$\phi_0 = (\vartheta_0^{'}, \theta_0^{'}) = (c, a_{01}, \ldots, a_{0P}, b_{01}, \ldots, b_{0Q}, \theta_0^{'})^{'}$$

en donde $\theta_0^{'} = (\omega_0, \alpha_0)$.
La estimación de los parámetros se realiza mediante máxima cuasi-verosimilitud. Si $q \geq Q$, los valores iniciales son

$$X_0, \ldots, X_{1-(q-Q)-.P}, \tilde{\epsilon}_{-q+Q}, \ldots, \tilde{\epsilon}_{1-q}, \tilde{\sigma}_0^2, \ldots, \hat{\sigma}_{1-p}^2.$$

Para cualquier ϑ, los valores de $\tilde{\epsilon}_t(\vartheta)$, para $t = -q + Q + 1, \ldots, n$; y para toda θ, los de $\tilde{\sigma}_t^2(\phi)$, para $t = 1, \ldots, n$, pueden ser calculados de la siguiente manera:

$$\begin{cases} \tilde{\epsilon}_t & = \tilde{\epsilon}_t(\vartheta) = X_t - c - \sum_{i=1}^{P} a_i (X_{t-i} - c) + \sum_{j=1}^{Q} b_j \tilde{\epsilon}_{t-j} \\ \tilde{\sigma}_t^2 & = \tilde{\sigma}_t^2(\varphi) = \omega + \sum_{i=1}^{q} \alpha_i \tilde{\epsilon}_{t-i}^2 + \sum_{j=1}^{p} \beta_j \tilde{\sigma}_{t-j}^2 \end{cases}$$

Y en el otro caso, cuando $q < Q$, los valores iniciales son

$$X_0, \ldots, X_{1-(q-Q)-P}, \epsilon_0, \ldots, \epsilon_{1-Q}, \tilde{\sigma}_0^2, \ldots, \tilde{\sigma}_{1-p}^2$$

Entonces, condicionada en los valores iniciales, la log-verosimilitud Gaussiana está dada por

$$\tilde{I}_n(\varphi) = \frac{1}{n}\sum_{t=1}^{n}\tilde{l}_t, \quad \tilde{l}_t(\varphi) = \frac{\tilde{\epsilon}_t(\vartheta)}{\tilde{\sigma}_t^2(\varphi)} + \log\tilde{\sigma}_t^2(\varphi)$$

Por lo que el estimador por máxima cuasi-verosimilitud es una solución medible de la ecuación

$$\hat{\varphi}_n = \arg\min_{\varphi\in\Phi} \tilde{I}_n(\varphi)$$

Capítulo 2

Estadística bayesiana no paramétrica

2.1. Introducción a la estadística bayesiana

En (Goldstein, 2013) se expone que existen dos tipos de incertidumbre presentes en cualquier fenómeno: la incertidumbre aleatoria y la incertidumbre epistémica. La incertidumbre aleatoria es aquella inherente al sistema estudiado; mientras que la epistémica es aquella que se podría resolver si se contara con más información, es decir, es aquella que surge de la falta de conocimiento de ciertos valores fijos que, en teoría, se podrían conocer.

El estudio de esta falta de información es el objetivo de la estadística. En muchos casos, los modelos que usan muchas ciencias, como la física, asumen ambientes totalmente controlados en donde la evolución de los sistemas dinámicos es totalmente determinista, lo que dificulta su extensión a situaciones reales en donde los sistemas se ven perturbados por un sinnúmero de factores imposibles de cuantificar.

La estadística toma en cuenta estas perturbaciones insertando un ruido aleatorio a un modelo determinista, es decir, asume que el comportamiento del sistema puede ser descompuesto en una parte aleatoria y en otra no aleatoria. A la parte aleatoria se le asigna una distribución de probabilidad que permite estudiarla.

Cuando se toma un modelo paramétrico para la parte aleatoria, es decir, cuando se asume una distribución $\mathcal{P}_\theta$, la incertidumbre aleatoria queda capturada por la distribución, mientras que la epistémica yace sobre el parámetro, el cual se asume como una cantidad desconocida, pero estimable.

El enfoque frecuentista de la estadística asume que el parámetro es una cantidad fija que, si bien no se conoce de entrada, es posible estimar su

valor de los datos observados. El bayesiano, por el contrario, supone que el parámetro es aleatorio, por lo que se le da una distribución inicial de probabilidad, llamada distribución *a priori*.

La justificación para el uso de esta distribución inicial se da cuando se asume intercambiabilidad.

Definición 16 (Intercambiabilidad)**.** Una sucesión de variables aleatorias $\{X_i\}_{i=1}^{\infty}$ se dice intercambiable si

$$(X_1, \ldots, X_n) \stackrel{d}{=} (X_{\pi(1)}, \ldots, X_{\pi(n)})$$

para cualquier permutación π de los índices, es decir, si las distribuciones finito-dimensionales son invariantes ante permutaciones.

Bruno de Finetti desarrolló a principios del siglo pasado un teorema de representación para sucesiones de variables aleatorias intercambiables que provee un argumento para justificar las distribuciones *a priori*.

Teorema 3 (Teorema de de Finetti)**.** *Sea $\mathbb{X}$ un espacio polaco dotado con su σ-álgebra de Borel $\mathcal{X}$. Denotemos por $\mathcal{P}_{\mathbb{X}}$ al espacio de todas las medidas de probabilidad sobre $(\mathbb{X}, \mathcal{X})$. Una sucesión infinita de variables aleatorias $\mathbb{X}$-valuadas $\{X_i\}_{i=1}^{\infty}$ es intercambiable si existe una medida de probabilidad Q sobre $\mathcal{P}_{\mathbb{X}}$ tal que estas, dada Q, son independientes, es decir,*

$$\mathbb{P}(X_1 \in A_1, \ldots, X_n \in A_n) = \int_{\mathcal{P}_{\mathbb{X}}} \prod_{i=1}^{n} P(A_i)\, Q(\mathrm{d}P),$$

para todo $n \geq 1$, y en donde Q es el límite de las distribuciones empíricas, es decir, $Q = \lim_{n\to\infty} \frac{1}{n} \sum_{i=1}^{n} \delta_{X_i}(A)$, y $P \sim Q$.

Así, bajo el supuesto de intercambiabilidad, este teorema garantiza la existencia de la distribución inicial.

El proceso de inferencia bayesiana, de acuerdo a lo expuesto en (Schervish, 1996), se resume en, dados los datos $x_1, \ldots, x_n$ condicionalmente independientes dada una medida *a priori*, se obtiene la distribución posterior $\theta \,|\, x_1, \ldots, x_n$ usando el teorema de Bayes. Si se asume que todas las medidas involucradas

tienen densidad,

$$f(\theta \,|\, x_1, \ldots, x_n) = \frac{f(x_1, \ldots, x_n \,|\, \theta)\; f(\theta)}{\int f(x_1, \ldots, x_n \,|\, \theta)\, f(\theta)\, \nu(\mathrm{d}\theta)},$$

en donde $\nu(\cdot)$ es la medida de referencia del espacio.
Si se omite la constante de normalización, se obtiene

$$f(\theta \,|\, x_1, \ldots, x_n) \propto f(x_1, \ldots, x_n \,|\, \theta)\; f(\theta) \tag{2.1}$$

Una vez obtenida la distribución posterior, se pueden obtener estimadores puntuales para el parámetro mediante la elección de una función de pérdida. Si se escoge la función de pérdida cuadrática, $L(\theta, \hat{\theta}) = (\theta - \hat{\theta})^2$, el estimador de bayes para el parámetro se vuelve la esperanza de la distribución posterior, $\mathbb{E}\left[\theta \,|\, x_1, \ldots, x_n\right]$.
La probabilidad de obtener nuevas observaciones está dada por la distribución predictiva.

$$f(x^*_{n+1} \,|\, x_n, \ldots, x_1) = \int f(x^*_{n+1} \,|\, \theta)\, f(\theta \,|\, x_1, \ldots, x_n)\, \nu(\mathrm{d}\theta) \tag{2.2}$$

En la práctica, la elección de la distribución *a priori* se conjuga con la de la verosimilitud, es decir, con la distribución de los datos dado el parámetro. El cálculo de la distribución posterior no siempre es numéricamente sencillo, por lo que se buscan modelos que simplifiquen su obtención. Una propiedad particularmente útil para ello es la de conjugación.

Definición 17 (Familia conjugada (Gelman y cols., 2013))**.** Una familia de distribuciones sobre el espacio parametral Θ, $\zeta = \{q_\phi(\theta); \phi \in \psi\}$ se dice conjugada para la familia paramétrica $\mathcal{P}_\theta$ si la distribución posterior $q_n \in \zeta$ para la verosimilitud. Es decir, si $q(\theta \,|\, x) \in \zeta$.

Ejemplo de un modelo con esta propiedad es el beta-Bernoulli.

Ejemplo 1. Se toma una muestra iid $\underline{x} = x_1, \ldots, x_n$ de una distribución Bernoulli(p). Como p es una probabilidad, se debe elegir una distribución inicial con soporte en el $[0, 1]$. Una distribución con esta propiedad es la distribución beta.

Como la muestra es independiente, la verosimilitud es el producto de las marginales, es decir,

$$\begin{aligned} f(\underline{x} \,|\, p) &= \prod_{i=1}^{n} f(x_i \,|\, p) \\ &= \prod_{i=1}^{n} p^{x_i}(1-p)^{1-x_i} \\ &= p^{\,\Sigma\, x_i}(1-p)^{n-\Sigma\, x_i} \end{aligned}$$

Entonces, tomando como distribución *a priori* una distribución beta con parámetros α y β, la distribución posterior queda

$$\begin{aligned} f(p \,|\, \underline{x}) &\propto f(p) f(\underline{x} \,|\, p) \\ &\propto p^{\alpha-1}(1-p)^{\beta-1} p^{\,\Sigma\, x_i}(1-p)^{n-\Sigma\, x_i} \\ &\propto p^{\alpha+\Sigma\, x_i-1}(1-p)^{\beta-1+n-\Sigma\, x_i} \end{aligned}$$

Identificamos el kérnel de una distribución Beta($\alpha + \sum x_i$, $\beta + n - \sum x_i$).

Entonces como tanto la distribución posterior como la inicial son betas, decimos que esta familia de distribuciones es conjugada para la Bernoulli.

La distribución posterior del modelo anterior es un ejemplo en donde se aprecia claramente cómo se actualiza las creencias subjetivas iniciales con los datos obtenidos, agregándose de manera lineal. Por ejemplo, al simular 100 realizaciones independientes de una v.a. Bernoulli(0.8), y fijar los hiperparámetros de la distribución *a priori* Beta(10,10), los parámetros de la posterior se vuelven 31 y 89, centrándola alrededor del 0.8. En la figura 2.1 se muestran las densidades de ambas distribuciones.

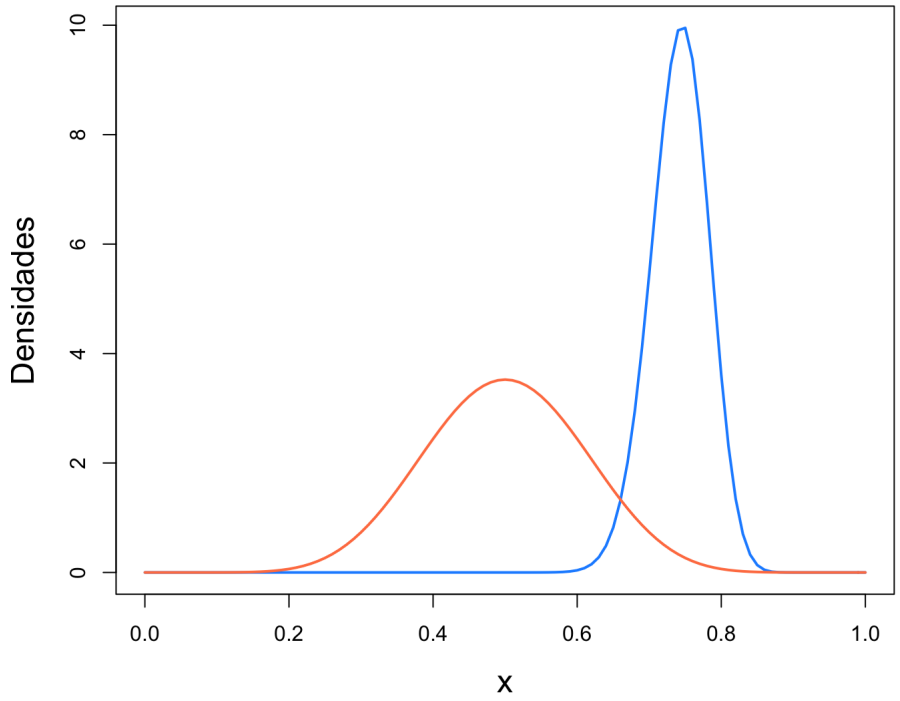

Figura 2.1: En rojo, distribución *a priori* Beta$(10, 10)$; en azul, la distribución posterior.

En este ejemplo la propiedad de conjugación hace que sea sencillo reconocer la familia distribucional a la que pertenece la distribución posterior, pero este no siempre es el caso. Por ejemplo, si intentamos estimar la media de una normal con varianza 1 mediante una distribución Cauchy de parámetros 0 y 1, es decir,

$$x \,|\, \mu \sim \mathrm{N}\,(\mu, 1)$$

$$\mu \sim \mathrm{Cauchy}\,(0, 1),$$

la posterior es proporcional a

$$\begin{aligned} f(x) &\propto f(x \,|\, \mu) f(\mu) \\ &\propto \frac{1}{\sqrt{2\pi}} \exp\left\{-\frac{1}{2}(x-\mu)^2\right\} \cdot \frac{1}{\pi(1+\mu^2)}, \end{aligned}$$

expresión que, de antemano, no corresponde a ninguna distribución conocida. Naturalmente, el siguiente paso sería intentar calcular la constante de normalización, que es igual a la integral

$$\int_{\mathbb{R}} \frac{1}{\sqrt{2\pi}} \exp\left\{-\frac{1}{2}(x-\mu)^2\right\} \cdot \frac{1}{\pi(1+\mu^2)} \, \mathrm{d}\mu$$

la cual, de nuevo, resulta incalculable en términos de una antiderivada. Esto hace complicada la inferencia al sólo conocer la densidad posterior en proporcionalidad.

2.2. Cadenas de Markov Monte Carlo

La complejidad que exhiben datos reales requiere de modelos que, en muchos casos, no pueden ser manipulados de manera analítica. Existen varios métodos de simulación que permiten obtener información sobre alguna distribución objetivo de manera numérica.

Una serie de técnicas particularmente útiles para la estadística bayesiana son los métodos de cadenas de Markov Monte Carlo[1]. Son una serie de algoritmos que permiten muestrear de manera aproximada una distribución cuando sólo se conoce su densidad a proporcionalidad, como es el caso del ejemplo expuesto Normal-Cauchy. Una vez realizada la simulación, se utilizan resultados asintóticos para obtener probabilidades y momentos de la distribución objetivo.

Los métodos MCMC tienen ese nombre porque se basan en la idea de construir una cadena de Markov de tal modo que converja a la distribución deseada, para después muestrear de esta simulando una trayectoria de la

[1] En inglés, *Markov chain Monte Carlo*, MCMC.

cadena.

El concepto de convergencia de una cadena de Markov, y en general de cualquier proceso markoviano, se traduce en la propiedad de ergodicidad. Para comprender esta propiedad, primero expondremos algunos conceptos básicos que permitan desarrollar la teoría atrás, de acuerdo a lo expuesto en (Robert y Casella, 2004).

El diseño de un algoritmo MCMC requiere que la cadena de Markov resultante cumpla con ciertas propiedades de regularidad que permitan que el algoritmo sea tan eficiente y preciso como se pueda. Una de ellas es la de irreducibilidad.

Definición 18 (Irreducibilidad)**.** Dada una medida ϕ, una cadena de Markov $\{X_t\}$ con kérnel de transición $K(x, y)$ se dice ϕ-irreducible si para todo $A \in \mathcal{X}^2$ con $\phi(A) > 0$, existe una n tal que $K^n(x, A) > 0$ para todo $x \in \mathbb{X}$, o equvalentemente, que $\mathbb{P}(\tau_A < \infty) > 0$, con $\tau_A = \text{ínf}\{n \geq 1; X_n \in A\}$. La cadena es fuertemente ϕ-irreducible si $n = 1$ para todo A medible.

Intuitivamente, esta propiedad nos dice que todos los estados de la cadena se comunican, es decir, que hay una probabilidad positiva de llegar a cualquier estado desde cualquier estado. Esto es importante pues si la cadena construida no fuera irreducible, esta podría ser inicializada en un estado del cual nunca podría llegar a la convergencia, volviendo inútil al algoritmo.

Ahora, la irreducibilidad sólo garantiza que todo A será visitado por la cadena, no dice nada sobre la frecuencia de las visitas; la convergencia no puede ser garantizada sin un barrido más o menos estable del espacio, situación que se formaliza con el concepto de recurrencia.

Definición 19 (Recurrencia)**.** En un espacio finito $\mathbb{X}$, un estado $\omega \in \mathbb{X}$ es transitorio si el número esperado de visitas a este,

$$\mathbb{E}_\omega[\,\eta_\omega\,] = \mathbb{E}_\omega\left[\sum_{n=1}^{\infty} \mathbb{1}_\omega(X_n)\right],$$

es finito, y es recurrente si $\mathbb{E}_\omega[\,\eta_\omega\,] = \infty$.

[2]De aquí en adelante se denotará al espacio muestral por $\mathbb{X}$, y a su σ-álgebra de Borel por $\mathcal{X}$.

Existe un concepto un poco más fuerte que permite construir cadenas más estables: la Harris-recurrencia.

Definición 20 (Harris-recurrencia)**.** Un conjunto A es Harris-recurrente si $\mathbb{P}(\eta_A = \infty) = 1$. Una cadena $\{X_T\}$ es Harris-recurrente si existe una medida ψ tal que es ψ-irreducible y todo boreliano $A \in \mathcal{X}$ es Harris-recurrente.

Esta propiedad sólo tiene sentido cuando $\mathbb{X}$ es no numerable, pues cuando es numerable o finito, $\mathbb{E}[\eta_x] = \infty$ si y sólo si $\mathbb{P}(\eta_A = \infty) = 1$.

Algunas veces, el comportamiento de una cadena de Markov se ve restringido por condiciones deterministas, una propiedad que ejemplifica estas condiciones es la periodicidad. Cuando el espacio de estados es contable, el periodo d de un estado $\omega \in \mathbb{X}$ es el máximo común denominador de $\{m \geq 1; K^m(\omega, \omega) > 0\}$. Esto fuerza a que, dado que se visitó el estado ω, el retorno a este pueda darse sólo en múltiplos de d. Para generalizar esta propiedad es necesario introducir el concepto de *conjunto pequeño*.

Definición 21 (Conjunto pequeño)**.** Un conjunto C es pequeño si existe $m \in \mathbb{N}$ y una medida positiva ν_m tal que

$$K^m(x, A) \geq \varepsilon\nu(A), \quad \forall x \in C, \forall A \in \mathcal{X}$$

Entonces se extiende el concepto de periodicidad a un espacio de estados general de la siguiente forma.

Definición 22 (Periodicidad)**.** Una cadena ψ-irreducible posee un ciclo de longitud d si existen un conjunto pequeño C, un entero M y una medida de probabilidad ν_M tal que d es el máximo común denominador de

$$\{m \geq 1; \exists\, \delta_m > 0 \text{ tal que } C \text{ es pequeño para } \nu_m > \delta_m \nu_M\}$$

Si $d = 1$ se dice que la cadena es aperiódica.

El concepto de estacionariedad para cadenas de Markov se da en forma de la distribución estacionaria, la cual no es más que la marginal del proceso evaluada en cada tiempo. Esta puede ser expresada en términos del kérnel de transición de la manera siguiente:

Definición 23. Una medida σ-finita π es invariante o estacionaria para el kérnel de transición $K(\cdot,\cdot)$ y para su cadena asociada si

$$\pi(B) = \int_{\mathbb{X}} K(x, B)\, \pi(\mathrm{d}x), \quad \forall B \in \mathcal{X}$$

En general se busca que las cadenas construidas mediante los algoritmos MCMC posean una distribución invariante.

Finalmente, la última condición de regularidad que se pide es el de invertibilidad. Una cadena de Markov estacionaria es invertible si sus distribuciones finito-dimensionales son invariantes al flujo del tiempo. Es decir, si la distribución de $X_n \,|\, X_{n+1}$ es igual a la de $X_{n+1} \,|\, X_n$.

Existen una serie de ecuaciones conocidas como de *balance detallado* que caracterizan a la invertibilidad y a la estacionariedad.

Teorema 4 (Balance detallado)**.** *Si una cadena de Markov con kérnel de transición K satisface la siguiente ecuación conocida como balance detallado*

$$K(y, x)\, \pi(y) = K(x, y)\, \pi(x) \tag{2.3}$$

para alguna medida de probabilidad π y para cada (x, y), entonces

1. *π es la medida estacionaria de la cadena.*
2. *La cadena es reversible.*

Demostración. Tenemos que

$$K(y, B)\, \pi(y) = \int_B K(y, x)\, \pi(y)\, \mathrm{d}x,$$

por lo que

$$\int_{\mathcal{Y}} K(y, B)\, \pi(y)\, \mathrm{d}y = \int_{\mathcal{Y}} \int_B K(y, x)\, \pi(y)\, \mathrm{d}x \mathrm{d}y$$

Si se cumple balance detallado,

$$= \int_{\mathcal{Y}} \int_{B} K(x,y)\, \pi(x)\, \mathrm{d}x \mathrm{d}y$$

$$= \left(\int_{B} K(x,y)\, \mathrm{d}y \right) \left(\int_{\mathcal{Y}} \pi(x)\, \mathrm{d}x \right)$$

Como, por ser densidad, $\int_{\mathcal{Y}} K(x,y)\, \mathrm{d}y = 1$,

$$= \int_{B} \pi(x)\, \mathrm{d}x$$

Lo cual corresponde a la medida de B bajo π, por lo que π es la medida invariante de la cadena. La reversibilidad se sigue de manera directa, pues $K(y,x)$ es la condicional de x dado y, y $\pi(x)$ es la marginal de x. □

Con estos conceptos es posible dar una noción de convergencia para cadenas de Markov. La existencia de la distribución invariante hace que sea natural el pensar que, si una cadena estacionaria converge, converja a su distribución estacionaria. Esta idea se ve cristalizada en el concepto de ergodicidad.

Definición 24 (Ergodicidad geométrica)**.** Si una cadena $\{X_n\}$ es Harris-recurrente, con distribución estacionaria π y cumple que $\mathbb{E}_{\pi}[h] < \infty$ y que existe $r_h > 1$ tal que

$$\sum_{n=1}^{\infty} r_h^n \parallel K^n(x,\cdot) - \pi \parallel_h \; < \infty$$

para cada $x \in \mathcal{A}$, con $h \geq 1$ y $\parallel \cdot \parallel_h = \sup_{|g| \leq h} | \int g(x)\, \mu(\mathrm{d}x)|$, entonces se dice que es geométricamente h-ergódica.

Propiedad que implica que $| K^n(x,\cdot) - \pi |_h$ decrece a velocidad geométrica.

Existe otra forma de ergodicidad más fuerte, la ergodicidad uniforme.

Definición 25. Una cadena $\{X_n\}$ es uniformemente ergódica si

$$\lim_{n\to\infty} \sup_{x\in\mathcal{A}} \| K^n(x,\cdot) - \pi \|_{VT} = 0,$$

con $\|\nu\|_{VT} = \sup_A |\, \nu(A) \,|$.

Finalmente, existe un teorema que justifica el uso de las trayectorias de una cadena de Markov para hacer inferencia como si fueran muestras iid.

Teorema 5 (Teorema ergódico)**.** *Si $\{X_n\}$ tiene una medida invariante σ-finita π, los siguientes enunciados son equivalentes:*

1. *Si $f, g \in L^1(\pi)$, con $\int g(x)\,\pi(\mathrm{d}x) \neq 0$, y $S_n(h) = \frac{1}{n}\sum_{i=1}^n h(X_i)$, entonces*
$$\lim_{n\to\infty} \frac{S_n(f)}{S_n(g)} = \frac{\int f(x)\,\pi(\mathrm{d}x)}{\int g(x)\,\pi(\mathrm{d}x)}$$
2. *La cadena $\{X_n\}$ es Harris-recurrente.*

Este teorema nos garantiza que los promedios empíricos convergen a la esperanza, dando una versión de la Ley de los Grandes Números para trayectorias de cadenas de Markov, justificando el uso de los algoritmos MCMC.

Ahora, dados estos resultados básicos, es posible diseñar un algoritmo que construya una cadena de Markov ergódica con distribución estacionaria f^*, la densidad que se desea aproximar. El primero de ellos es conocido como Metropolis-Hastings.

Algoritmo Metropolis-Hastings

Supóngase que se tiene una densidad objetivo f. Se escoge una densidad condicional $q(y\,|\,x)$ de tal modo que sea sencilla de simular y que o se conozca a proporcionalidad (independiente de x), o que sea simétrica, es decir, que $q(x\,|\,y) = q(y\,|\,x)$. La única condición sobre f es que el cociente $f(x)/q(y\,|\,x)$ sea conocido a proporcionalidad, independiente de x, también. Así, se produce una cadena de Markov $\{X_t\}$ con la transición dada por el siguiente algoritmo:

Algoritmo 1 Metropolis-Hasting

Dado x_t,

1: Generar $Y_t \sim q(y \mid x_t)$

2: Tomar

$$x_{t+1} = \begin{cases} Y_t & \text{con probabilidad } \rho(x_t, Y_t), \\ x_t & \text{con probabilidad } 1 - \rho(x_t, Y_t) \end{cases}$$

con

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

La estacionariedad de la cadena generada está dada por las ecuaciones de balance detallado.

Demostración. El kérnel de transición dado por el algoritmo es el siguiente:

$$K(x, y) = \rho(x, y) q(y \mid x) + (1 - r(x)) \delta_x(y),$$

en donde $r(x) = \int \rho(x, y) q(y \mid x) \, \mathrm{d}x$, y $\delta_x(y)$ es la delta de Dirac en x. Entonces, si partimos en sumandos, como las densidades son positivas siempre, para el primero se tiene que

$$\begin{aligned} \rho(x, y) q(y \mid x) f(x) &= \text{mín} \left\{ \frac{f(y)}{f(x)} \frac{q(x \mid y)}{q(y \mid x)}, 1 \right\} q(y \mid x) f(x) \\ &= \text{mín} \left\{ \frac{f(y)}{f(x)} q(x \mid y), q(y \mid x) \right\} f(x) \\ &= \text{mín} \left\{ f(y) q(x \mid y), q(y \mid x) f(x) \right\} \end{aligned}$$

Nótese que la expresión anterior es simétrica, por tanto es igual a

$$= \rho(y, x) q(x \mid y) f(y)$$

Ahora, para el segundo sumando, la delta de Dirac da masa 1 cuando $x = y$, por lo que es inmediato que $(1 - r(x))\delta_x(y) = (1 - r(y))\delta_y(x)$. Con esto

concluimos que

$$K(x, y)f(y) = K(y, x)f(x),$$

por lo que Metropolis-Hasting cumple las ecuaciones de balance detallado y f es la distribución estacionaria. □

Dado este resultado, la ergodicidad puede ser garantizada si la cadena es Harris-recurrente. La demostración de esta puede ser encontrada en (Robert y Casella, 2004), y está ligada al hecho de que la cadena generada por Metropolis-Hasting es f-irreducible.

Como una de las propiedades claves de este algoritmo es que permite aproximar mediante una densidad propuesta arbitraria cualquier densidad que se conozca a proporcionalidad, el ejemplo de la sección anterior, el Normal-Cauchy, puede ser solucionado con Metropolis-Hasting.

Recordemos que el modelo consistía en asignar una distribución inicial Cauchy(μ; 0, 1) a la media de una Normal(x; μ,1). La posterior resultante no pareciera ser alguna distribución conocida, y la integral de la constante de normalización no posee antiderivada, por lo que un enfoque MCMC para simular la posterior resulta apropiado.

De manera arbitraria asignemos como distribución propuesta una normal, es decir,

$$q(x^*_{t+1} \,|\, x_t) = \mathrm{N}\left(x^*_{t+1}; x_t, 1\right)$$

Entonces, denotando como $\tilde{x}$ al valor muestral, la densidad objetivo es

$$f(x) \propto \mathrm{Cauchy}(x; 0, 1)\, \mathrm{N}\left(\tilde{x}; x, 1\right)$$

Y como la densidad propuesta es simétrica, el algoritmo queda:

Dado x_t,

1: Generar $Y_t \sim \mathrm{N}\left(x^*_{t+1}; x_t, 1\right)$

2: Tomar

$$x_{t+1} = \begin{cases} Y_t & \text{con probabilidad } \rho(x_t, Y_t), \\ x_t & \text{con probabilidad } 1 - \rho(x_t, Y_t) \end{cases}$$

con

$$\rho(x, y) = \text{mín}\left\{ \frac{\text{Cauchy}(y; 0, 1)\,\mathrm{N}\left(\tilde{x}; y, 1\right)}{\text{Cauchy}(x; 0, 1)\,\mathrm{N}\left(\tilde{x}; x, 1\right)}, 1 \right\}$$

Al implementarse el algoritmo es importante tomar en cuenta que la cadena de Markov puede tardarse un número grande de iteraciones en converger a su distribución invariante, por lo que se deben tomar las simulaciones realizadas después del periodo de *burn in* que se proponga.

Si se realizan 10,000 simulaciones con un punto inicial de 5, y se toma un periodo de *burn in* de 5,000 iteraciones, la esperanza posterior, es decir, el estimador bayesiano con pérdida cuadrática, resulta de 1.3115. En la figura 2.2 se muestran las trayectorias de cuatro simulaciones bajo distintos puntos iniciales para el modelo Cauchy-normal.

Muestreo de Gibbs

El muestreo de Gibbs es otro algoritmo MCMC en el cual se utilizan variables aleatorias auxiliares para aproximar a la densidad objetivo.

Supóngase que se quiere simular de la densidad de la v.a. X, $f(x)$; y supóngase que X y la variable aleatoria Y poseen densidad conjunta $f(x, y)$. Entonces para el caso bidimensional, el muestreo de Gibbs genera la siguiente cadena de Markov (X_t, Y_t):

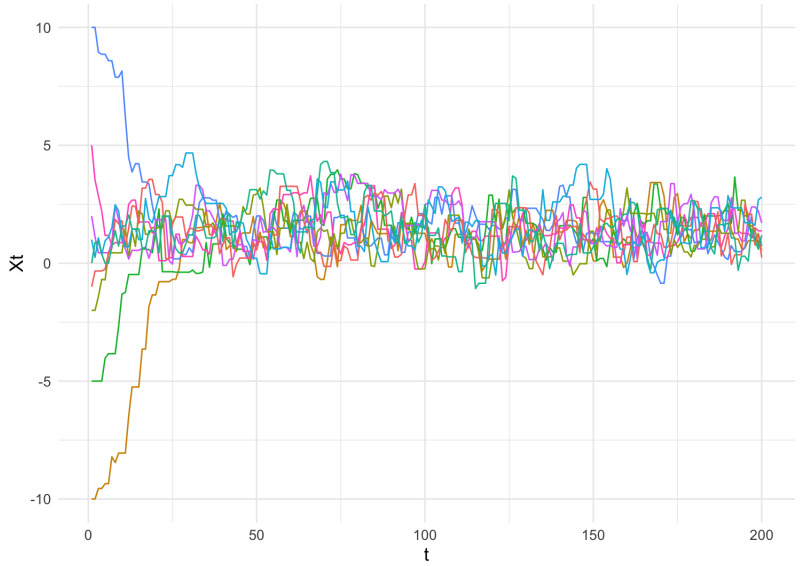

Figura 2.2: Cuatro trayectorias de 200 simulaciones cada una. En azul, punto inicial 10; en verde, 0; en morado, -5; y en rojo, -10.

Algoritmo 2 Muestreo de Gibbs bietápico

1: Tomar $X_0 = x_0$
2: **for** $t = 1, 2, \ldots$ **do**
3: Generar $Y_t \sim f_{Y \mid X}(\cdot \mid x_{t-1})$
4: Generar $X_t \sim f_{X \mid Y}(\cdot \mid y_t)$
5: **end for**

Los componentes marginales de la cadena bidimensional generada son a su vez cadenas de Markov, pues sus transiciones son

$$K(x, x^*) = \int f_{Y \mid X}(y \mid x)\, f_{X \mid Y}(x^* \mid y)\, \mathrm{d}y$$

$$K(y, y^*) = \int f_{Y \mid X}(y^* \mid x)\, f_{X \mid Y}(x \mid y)\, \mathrm{d}x,$$

las cuales dependen solamente de los valores pasados inmediatos de $\{X_t\}$ y de $\{Y_t\}$.
El muestreo de Gibbs se basa en que las distribuciones estacionarias de las

subcadenas son las marginales de X y de Y, pues

$$\begin{aligned}
f_X(x^*) &= \int f_{X,Y}(x^*, y)\,\mathrm{d}y \\
&= \int f_{X\,|\,Y}(x^*\,|\,y)\, f_Y(y)\,\mathrm{d}y \\
&= \int f_{X\,|\,Y}(x^*\,|\,y) \left(\int f_{Y\,|\,X}(y\,|\,x)\, f_X(x)\,\mathrm{d}x \right) \mathrm{d}y \\
&= \int \int f_{X\,|\,Y}(x^*\,|\,y)\, f_{Y\,|\,X}(y\,|\,x)\, f_X(x)\,\mathrm{d}x\,\mathrm{d}y \\
&= \int \int f_{X\,|\,Y}(x^*\,|\,y)\, f_{Y\,|\,X}(y\,|\,x)\, f_X(x)\,\mathrm{d}y\,\mathrm{d}x \\
&= \int \left(\int f_{X\,|\,Y}(x^*\,|\,y)\, f_{Y\,|\,X}(y\,|\,x)\,\mathrm{d}y \right) f_X(x)\,\mathrm{d}x \\
&= \int K(x, x^*)\, f_X(x)\,\mathrm{d}x
\end{aligned}$$

De manera análoga para la subcadena $\{Y_t\}$.

Para el caso general, supóngase que, dada una distribución multidimensional $f(x_1, \ldots, x_p)$, se puede simular de las condicionales

$$X_i\,|\,X_1, \ldots, X_{i-1}, X_{i+1}, \ldots, X_p, \quad \text{para } i = 1, 2, \ldots, p$$

Entonces el muestreo de Gibbs es el siguiente:

Algoritmo 3 Muestreo de Gibbs

1: Tomar $x^{(t)} = (x_1^{(t)}, \ldots, x_p^{(t)})$
2: **for** $k = 1, 2, \ldots, p$ **do**
3: Generar $X_1^{(t+1)} \sim f_1(x_1\,|\,x_2^{(t)}, \ldots, x_p^{(t)})$
4: Generar $X_2^{(t+1)} \sim f_2(x_2\,|\,x_1^{(t+1)}, x_3^{(t)} \ldots, x_p^{(t)})$
$\vdots$
5: Generar $X_p^{(t+1)} \sim f_p(x_p\,|\,(x_1^{(t+1)}, \ldots, x_{p-1}^{(t+1)})$
6: **end for**

A las densidades $f_1, \ldots, f_p$ se les llama condicionales completas[3], y al ser unidimensionales, le dan la ventaja a este método de simplificar la simulación aun en situaciones de dimensiones altas.

Puede demostrarse (Robert y Casella, 2004) mediante una extensión de los argumentos expuestos arriba que, la cadena, y cada subcadena, generada por este algoritmo es estacionaria con distribución invariante igual a la distribución de cada marginal X_k del vector aleatorio original. Además, son Harris-recurrentes y f_k-irreducibles si se cumplen ciertas condiciones de regularidad sobre el kérnel de transición y las marginales f_k, y más aún, son ergódicas, lo que demuestra la correctez del algoritmo.

Slice sampler

Supóngase que se quiere simular de una variable aleatoria continua, X. Su densidad puede ser vista como

$$f(x) = \int_0^{f(x)} \mathrm{d}u,$$

lo que nos lleva al siguiente resultado:

Teorema 6 (Teorema fundamental de la simulación)**.** *Simular* $X \sim f(x)$ *es equivalente a simular*

$$(X, U) \sim U(\{(x, u) : 0 < u < f(x)\})$$

La interpretación geométrica de este resultado es simple: basta con simular una uniforme en el área abajo del gráfico de la densidad objetivo, y posteriormente quedarnos sólo con la primera coordenada de los vectores simulados, como se muestra en la figura 2.3.

La ventaja de este enfoque es que, en $\mathbb{R}^2$, la segunda coordenada, el eje Y, se deshecha y por ende sólo es necesario conocer de manera proporcional la densidad objetivo, es decir, no es necesaria la constante de normalización.

El problema de usar este teorema es que simular abajo de una curva, si bien es conceptualmente simple, computacionalmente no lo es.

[3]En inglés, *full conditionals*.

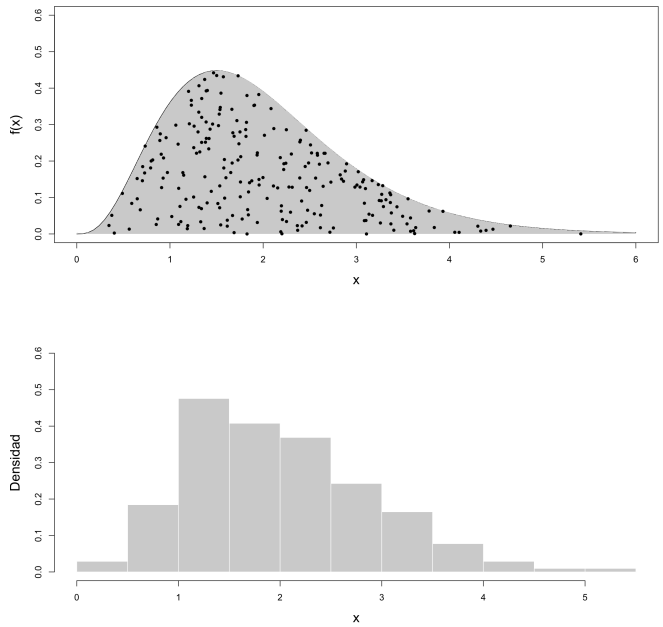

Figura 2.3: Arriba, simulación de una v.a. uniforma debajo del gráfico de la densidad de una distribución gamma; abajo, el histograma de la primera coordenada de los vectores muestreados.

Una forma es usar alguna variante de un algoritmo de aceptación-rechazo, en el cual se busca un conjunto $\mathscr{A}$ tal que $\mathscr{S} = \{(x, u) : 0 < u < f(x)\} \subset \mathscr{A}$ y en el cual sea sencillo simular una uniforme, se simulan los puntos y eliminas aquellos que no estén en $\mathscr{S}$. Esto, para densidades con formas más complejas que una campana, es muchas veces sumamente ineficiente, pues puede que se tenga que deshechar muchas simulaciones antes de llegar a alguna que cumpla la condición.

Radford Neal propuso en 1997 una solución a este problema. En vez de descartar valores usa una cadena de Markov que explore el conjunto $\mathscr{S}$.

Algoritmo 4 *Slice sampler* bidimensional

1: Tomar $X_0 = x_0$, $U_0 = u_0$
2: **for** $t = 1, 2, \ldots$ **do**
3: Generar $U_t \sim U\left([0, f(x_{t-1})]\right)$
4: Generar $X_t \sim U\left(\{x : f(x) \geq u_t\}\right)$
5: **end for**

Cabe destacar que estas distribuciones corresponden a las condicionales $U \,|\, X$ y $X \,|\, U$, por lo que el *slice sampler* es un caso particular del muestreo de Gibbs. Por supuesto que la validez de este algoritmo depende de que la distribución estacionaria de la cadena sea una uniforme sobre $\mathscr{S}$. En efecto,

$$\begin{aligned}
&\int\int \mathbb{P}\left((x,v) \,|\, (x,u)\right) \mathbb{P}\left((y,v) \,|\, (x,v)\right) \mathbb{1}_{[0,f(x)]}(u)\,\mathrm{d}u\,\mathrm{d}x = \\
&= \int\int \frac{\mathbb{1}_{[0,f(x)]}(v)}{f(x)} \frac{\mathbb{1}_{f(y)\geq v}}{\mu(\{a|f(a)\geq v\})} \mathbb{1}_{[0,f(x)]}(u)\,\mathrm{d}u\,\mathrm{d}x \\
&= \int f(x) \frac{\mathbb{1}_{[0,f(x)]}(v)}{f(x)} \frac{\mathbb{1}_{f(y)\geq v}}{\mu(\{a|f(a)\geq v\})}\,\mathrm{d}x \\
&= \mathbb{1}_{[0,f(y)]}(v) \int \frac{\mathbb{1}_{f(x)\geq v}}{\mu(\{a|f(a)\geq v\})}\,\mathrm{d}x \\
&= \mathbb{1}_{[0,f(y)]}(v),
\end{aligned}$$

en donde $\mu(.)$ es la medida de referencia, en este caso, la de Lebesgue. Esto muestra que la estacionaria es uniforme en $\mathscr{S}$, validando el algoritmo. Es importante destacar que esta prueba funciona aún cuando sólo se conoce f a proporcionalidad, es decir, cuando en vez de f, se tiene f^* tal que $f = kf^*$ para alguna constante k, por lo que, de nuevo, no es necesaria la constante de normalización.

La figura 2.3 muestra un ejemplo de un *slice sampler* para una distribución normal truncada. Una de las desventajas de este algoritmo es que no siempre resulta sencillo ni computacionalmente factible simular una unifor-

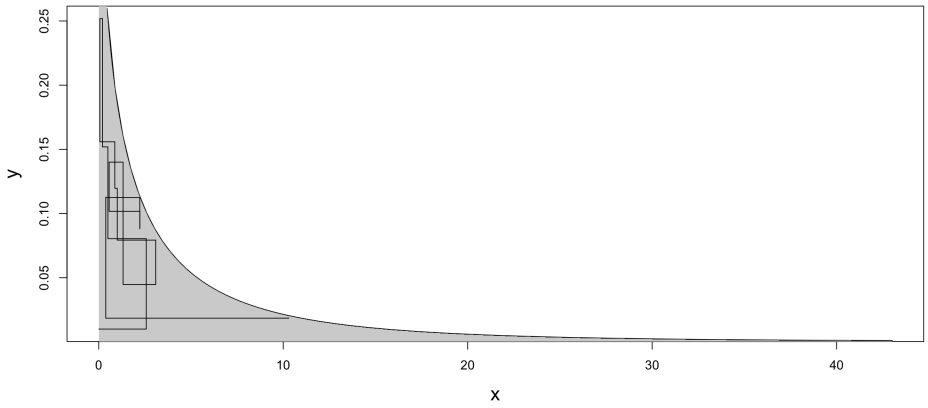

Figura 2.4: Quince primeras iteraciones de un *slice sampler* para una normal $\mathrm{N}(3,1)$ truncada al $[0,1]$.

me en

$$\mathscr{A}_t = \{x : f(x) \geq u_t\}.$$

Radford Neal propuso en 2003 una modificación que recreaba el *slice sampler* original eliminando el problema de la simulación: en vez de generar la uniforme en $\mathscr{A}_t$, ésta se genera en un intervalo que contenga a $\mathscr{A}_t$. Este intervalo se construye inciando con uno de longitud ω que contenga a x_{t-1}, después se va agrandando el intervalo en pasos de tamaño ω, es decir,

$$[x_t - \omega, x_t + \omega], [x_t - 2\omega, x_t + 2\omega], [x_t - 3\omega, x_t + 3\omega], \ldots,$$

hasta que ambos extremos del intervalo estén fuera de $\mathscr{A}_t$. Una vez que esto ocurra se simula una uniforme en ese intervalo, aceptándose sólo si está dentro de $\mathscr{A}_t$, y rechazándose si pasa lo contrario.

La dificultad de este enfoque es que encontrar el parámetro ω que funcione bien para todas las iteraciones puede ser complicado, ya que si el intervalo resultante es muy grande, se pueden llegar a rechazar muchos valores antes de que alguno caiga en el conjunto de interés.

2.2.1. Diagnóstico de convergencia

La ergodicidad de los procesos generados por los algoritmos anteriormente descritos garantiza la convergencia a la distribución objetivo cuando el número de iteraciones tiende a infinito. Es claro que, en la práctica, al sólo poder hacer simulaciones finitas, no sólo es necesario saber que la cadena va a converger, también se deben tener criterios para determinar cuándo esta ha convergido.

Grosso modo existen tres sentidos en los que hay que verificar la convergencia de la cadena:

- Convergencia a la distribución estacionaria.
- Convergencia en media.
- Convergencia a una muestra *iid*.

La primera, debido a que la ergodicidad sólo garantiza la estacionariedad de manera asintótica, se refiere a la velocidad con la que la cadena explora el soporte de la distribución estacionaria y al grado de correlación que la muestra exhibe a lo largo del tiempo.

La segunda se refiere a la convergencia de los promedios empíricos

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

hacia $\mathbb{E}_f[h(\theta)]$; en donde $\theta^{(t)}$ es el t-ésimo valor simulado de la cadena θ y f es la distribución invariante.

El propósito de la verificación de esta convergencia es el saber si el proceso generado por el algoritmo exploró todo el soporte de f. Esto es particularmente útil cuando la distribución objetivo es multimodal, pues puede que la cadena se estanque en una de las modas solamente, sin visitar el resto del espacio.

La convergencia a una muestra *iid* mide qué tanto se aproximan los valores generados a la independencia. Un método para alcanzar esto es el *subsampling*, el cual toma una muestra de tamaño k de la cadena $\{\theta^{(t)}\}$,

de modo que se seleccionen sólo los valores $\eta^k = \theta^{(tk)}$. Este método está justificado si la autocovarianza de la cadena decrece con el tiempo, pero, en general, esto no siempre pasa, además de que la elección de k no siempre es sencilla. La reducción del tamaño muestral perjudica a la convergencia en media, por lo que, en muchas ocasiones, es mejor tener una muestra muy correlacionada que una pequeña.

El hecho de que el uso del *subsampling* empeora la convergencia en media se puede ver con la siguiente proposición.

Proposición 2. *Supóngase que $h \in \mathcal{L}^2(f)$[4] y $\{\theta^{(t)}\}$ es una cadena de Markov con distribución estacionaria f. Si*

$$\delta_1 = \frac{1}{Tk}\sum_{t=1}^{Tk} h(\theta^{(t)}) \quad y \quad \delta_k = \frac{1}{T}\sum_{l=1}^{T} h(\theta^{(kl)}),$$

entonces la varianza de δ_1 satisface que

$$Var\,(\delta_1) \leq Var\,(\delta_k)$$

para cada $k > 1$.

Demostración. Sean $\delta_k^1, \ldots, \delta_k^{k-1}$ las versiones desplazadas de $\delta_k = \delta_k^0$. Es decir,

$$\delta_k^i = \frac{1}{T}\sum_{t=1}^{T} h(\theta^{(tk-i)})$$

Entonces, δ_1 puede ser escrito como $\delta_1 = \frac{1}{k}\sum_{i=0}^{k-1}\delta_k^i$, y en consecuencia,

$$\begin{aligned}\mathrm{Var}\,(\delta_1) &= \mathrm{Var}\left(\frac{1}{k}\sum_{i=0}^{k-1}\delta_k^i\right) \\ &= \mathrm{Var}\left(\delta_k^0\right)/k + \sum_{i\neq j}\mathrm{Cov}\left(\delta_k^i\,,\,\delta_k^j\right)/k^2\end{aligned}$$

La estacionariedad del proceso implica que, $\mathrm{Var}\left(\delta_k^0\right) = \mathrm{Var}\left(\delta_1^0\right)$, y si se utiliza la desigualdad de Cauchy-Schwarz, $\mathrm{Cov}\,(X\,,\,Y)^2 \leq \mathrm{Var}\,(X)\,\mathrm{Var}\,(Y)$,

[4] Es decir, cuyo cuadrado es f-integrable.

se tiene que

$$\begin{aligned}\mathrm{Cov}\left(\delta_k^i\,,\,\delta_k^j\right)^2 &\leq \mathrm{Var}\left(\delta_k^i\right)\mathrm{Var}\left(\delta_k^j\right)\\ &= \mathrm{Var}\left(\delta_k^0\right)\mathrm{Var}\left(\delta_k^0\right)\end{aligned}$$

Por lo que

$$\left|\mathrm{Cov}\left(\delta_k^i\,,\,\delta_k^j\right)\right| \leq \mathrm{Var}\left(\delta_k\right)$$

Entonces,

$$\begin{aligned}\mathrm{Var}\left(\delta_1\right) &= \mathrm{Var}\left(\delta_k^0\right)/k + \sum_{i\neq j}\mathrm{Cov}\left(\delta_k^i\,,\,\delta_k^j\right)/k^2\\ &\leq \mathrm{Var}\left(\delta_k^0\right)/k + \sum_{i\neq j}\mathrm{Var}\left(\delta_k^0\right)/k^2\\ &= \mathrm{Var}\left(\delta_k^0\right)/k + k(k-1)\mathrm{Var}\left(\delta_k^0\right)/k^2\\ &= \mathrm{Var}\left(\delta_k\right)\end{aligned}$$

□

De aquí se concluye que el hecho de usar una muestra más pequeña hace que se pierda información, por lo que las estimaciones resultarían de menor calidad que aquellas realizadas con la muestra completa. Sin embargo, cuando el cómputo de $h(\theta)$ es costoso, esta técnica se vuelve sumamente útil.

Hay varios enfoques para monitorear la convergencia a la estacionariedad. El empírico consiste en graficar las trayectorias simuladas de modo que puedan detectarse comportamientos erráticos que indiquen que no se ha convergido a la distribución invariante. Claro está que este método sólo funciona para detectar grandes desviaciones de la estacionariedad.

Otro método muy utilizado es el uso de pruebas no paramétricas como la Kolmogorov-Smirnov. Si se tiene una muestra $\theta^{(t)}$, de una cadena estacionaria, entonces $\theta^{(t_1)}$ tiene la misma distribución que $\theta^{(t_2)}$, por lo que basta con

dividir la muestra en dos submuestras, $\theta_1^{(t)}$ y $\theta_2^{(t)}$, y verificar que distribuyen igual.

Las pruebas no paramétricas, y en particular la Kolmogorov-Smirnov, están diseñadas para muestras independientes, por lo que se debe hacer una corrección que tome en cuenta la correlación en las observaciones. Un modo de hacer esto es el *subsampling*. Se toman G observaciones de cada submuestra de modo que G sea lo suficientemente grande como para que se desvanescan las autocorrelaciones y se alcance la cuasi independencia, y se evalúa la prueba con estas nuevas submuestras.

También es posible encontrar un buen punto de *burn-in*, es decir, encontrar el tiempo a partir del cual se alcanza la estacionariedad. Un modo de hacer es realizar la prueba iterativamente para diferentes subconjuntos de la muestra, por ejemplo, para las últimas 1000 observaciones, luego para las últimas 1,100, y así sucesivamente; hasta que el p-value alcance algún nivel arbitrario.

En contraste con la estacionariedad, la convergencia en media no es sencilla de detectar. En la práctica, es necesario primero que el proceso haya explorado varias regiones del espacio de estados para poder concluir que se ha llegado a la convergencia.

Una manera de monitorear que la media ha convergido es calcular distintos estimadores de esta de manera simultanea hasta que todos coincidan con una precisión dada. Se puede, por ejemplo, usar la media aritmética y el estimador Rao-Blackwellizado, que no es más que utilizar la esperanza de un estimador base condicional a algún estadístico. Esto, de acuerdo al teorema de Rao-Blackwell, reduce la varianza del estimador original.

Existen muchas maneras de Rao-Blackwellizar un algoritmo MCMC. Para el caso de un muestreo de Gibbs bietápico, por ejemplo, una manera clásica es reemplazar a

$$\delta_0 = \frac{1}{T}\sum_{t=1}^{T} h(y_1^{(t)})$$

Por

$$\delta_{rb} = \frac{1}{T}\sum_{t=1}^{T} \mathbb{E}\left[h(Y_1) \,|\, y_2^{(t)}\right],$$

lo cual es factible debido a que se conocen las formas distribucionales de las condicionales completas.

2.3. Medidas aleatorias

En la práctica, uno de los problemas más comunes en el quehacer estadístico es la elección de una distribución para los datos. Antes del desarrollo del cómputo, el uso de técnicas bayesianas estaba restringido a las familias conjugadas que simplificaban los cálculos.

El aumento de poder de cómputo y la creación de métodos como las cadenas de Markov Monte Carlo permitieron la utilización de modelos más complejos que fueran más adaptables a datos reales. Aún así, la elección de una familia de distribuciones continúa siendo un problema central cuya solución dista de alcanzarse, y por ello, se utilizan criterios de conveniencia y simplicidad.

El teorema de de Finetti caracteriza a la distribución conjunta de los datos mediante el uso de una medida sobre un espacio de medidas, el cual, en el caso paramétrico, se contrae a una familia de distribuciones, por ejemplo, las gaussianas con media θ y varianza 1. El caso no paramétrico se da cuando se eliminan las restricciones sobre el espacio de medidas, haciendo que la distribución *a priori* tenga como soporte a todo el espacio de distribuciones de probabilidad.

Comencemos primero con la teoría básica para el estudio de estos modelos. De aquí en adelante, sea $\mathbb{X}$ un espacio polaco con $\mathcal{X}$ su correspondiente σ-álgebra de Borel. Sea $M(\mathbb{X})$ el espacio de todas las medidas de probabilidad sobre el espacio medible $(\mathbb{X}, \mathcal{X})$. A una variable aleatoria que toma valores sobre $M(\mathbb{X})$ se le llama medida aleatoria. De manera más precisa,

Definición 26 ((Kallenberg, 2017))**.** Una medida aleatoria ν es una función $\nu : M(\mathbb{X}) \times \mathbb{X} \to \mathbb{X}$ tal que

1. $\nu(m, \cdot)$ es una variable aleatoria para cada $m \in M(\mathbb{X})$.

2. $\nu(\cdot, \omega)$ es una medida de probabilidad para cada $\omega \in \mathbb{X}$

2.3.1. Proceso Dirichlet

Existen muchas maneras de construir medidas aleatorias. Dos de las más comunes son la normalización de procesos estocásticos y el uso de construcciones *stick breaking*; del cual, una versión particular, el proceso Dirichlet, se utilizará en este texto. Comencemos definiendo la versión finita dimensional de esta medida, la distribución Dirichlet.

Definición 27 (Distribución Dirichlet)**.** Sea $Q = (Q_1, \ldots, Q_k)$ un vector aleatorio y sea $\alpha = (\alpha_1, \ldots, \alpha_k) \in \mathbb{R}^k$ tal que $\alpha_i > 0$ para toda i. Denotemos $\alpha_0 = \sum_{i=1}^{k} \alpha_i$. Decimos que Q distribuye Dirichlet con parámetro α si su función de densidad está dada por:

$$f(q; \alpha) = \frac{\Gamma(\alpha_0)}{\prod_{i=1}^{k} \Gamma(\alpha_i)} \prod_{i=1}^{k} q_i^{\alpha_i - 1} \mathbb{1}_{\Delta_k}(q),$$

en donde Δ_k es el simplejo k-dimensional.

El soporte de esta distribución, el simplejo k-dimensional, se define como el conjunto de puntos en $\mathbb{R}^k$ tal que la suma de sus entradas es igual a uno, *i.e.*,

$$\Delta_k := \left\{ (x_1, \ldots, x_k) \in \mathbb{R}^k : \sum_{i=1}^{k} x_i = 1 \right\}$$

En la figura 2.4 se muestra la función de densidad de una distribución Dirichlet con parámetro $\alpha = (2, 3, 4)$. Esta distribución posee las siguientes propiedades:

1. $\mathbb{E}[X_i] = \frac{\alpha_i}{\alpha_0}$.

2. $\mathrm{Var}(X_i) = \frac{\alpha_i(\alpha_0 - \alpha_i)}{\alpha_0^2(\alpha_0 + 1)}$.

3. $X_i \sim \mathrm{Beta}(\alpha_i, \alpha_0 - \alpha_i)$.

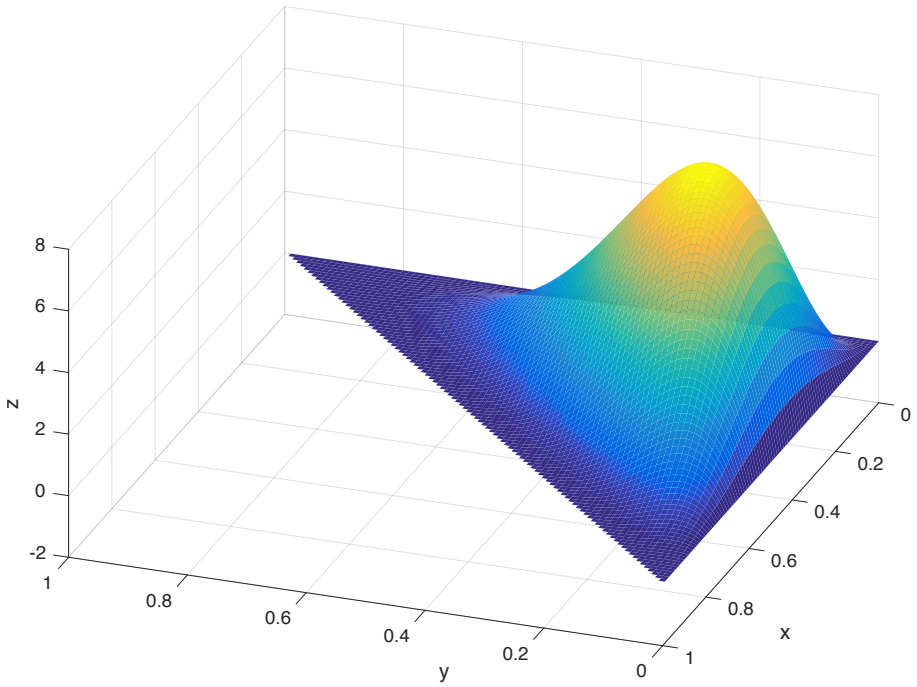

Figura 2.5: Densidad de la distribución Dirichlet con $\alpha = (2, 3, 4)$.

4. Si se toma una partición $\{A_1, \ldots, A_n\}$ de $\{1, \ldots, k\}$, entonces

$$\left(\sum_{i \in A_1} X_i, \sum_{i \in A_2} X_i, \ldots, \sum_{i \in A_n} X_i \right) \sim \operatorname{Dir}\left(\sum_{i \in A_1} \alpha_i, \sum_{i \in A_2} \alpha_i, \ldots, \sum_{i \in A_n} \alpha_i \right).$$

Además, la distribución Dirichlet es conjugada para la multinomial, pues si se tiene una muestra independiente de una multinomial de parámetros n_0 y $q = (q_1, \ldots, q_k)$, con n fijo, y se toma como distribución *a priori* para q una Dirichlet con híperparámetros $\alpha = (\alpha_1, \ldots, \alpha_k)$, la posterior queda

$$f(q \,|\, x) \propto f(q)\, f(x \,|\, q)$$

$$\propto \frac{n!}{x_1!x_2!\cdots x_k!}\prod_{i=1}^{k} q_i^{x_i}\left(\frac{\Gamma(\alpha_1+\cdots+\alpha_k)}{\prod_{i=1}^{k}\Gamma(\alpha_i)}\prod_{i=1}^{k} q_i^{\alpha_i-1}\right)$$

$$\propto \prod_{i=1}^{k} q_i^{\alpha+x-1}$$

El cual corresponde al kérnel de una Dirichlet con parámetros $\alpha_1+x_1,\ldots,\alpha_k+x_k$, por lo que la distribución Dirichlet es conjugada para la multinomial.

Es importante recalcar que, como el soporte de la Dirichlet es el simplejo k-dimensional, cada muestra de esta distribución puede considerarse como una función de masa de probabilidad, es decir, como cada muestra debe ser positiva y sumar 1, estas pueden verse como distribuciones discretas.

Resulta natural entonces el pensar en esta distribución al intentar construir medidas de probabilidad sobre $M(\mathbb{X})$. Thomas Ferguson introdujo en 1973 el proceso Dirichlet, el cual cumple la propiedad de que sus distribuciones finito dimensionales poseen esa misma distribución:

Definición 28 (Proceso Dirichlet (Ferguson, T. (1973)))**.** Sea α una medida finita sobre un espacio polaco $\mathbb{X}$ y $M>0$. Una medida aleatoria P sobre $\mathbb{X}$ es un proceso Dirichlet de parámetros α y M si, para cada partición medible $\{B_1,\ldots,B_n\}$ de $\mathbb{X}$, la distribución conjunta de $(P(B_1),\ldots,P(B_k))$ es una distribución Dirichlet k-dimensional con parámetros $M\alpha(B_1),\ldots,M\alpha(B_k)$.

El proceso Dirichlet puede ser visto entonces como una extensión de la distribución con el mismo nombre, lo que hace que ciertas propiedades se hereden. Si se considera la partición $\{A,A^c\}$, entonces $P(A)\sim \text{Beta}(\alpha(A),\alpha(A^c))$. La esperanza

$$\mathbb{E}\left[P(A)\right]=\frac{\alpha(A)}{(\alpha(A)+\alpha(A^c))}=G(A),$$

en donde $G(A)=\frac{\alpha(A)}{\alpha(\mathbb{X})}$. La varianza es

$$\text{Var}\left(P(A)\right)=G(A)G(A^c)/(1+M)$$

De esta expresión se sigue que, al hacer M más grande, la variabilidad disminuye, concentrándose sobre la medida $G(A)$. Por ello a M se le conoce

como parámetro de precisión.

Una construcción del proceso Dirichlet particularmente útil para la simulación es la llamada construcción *stick breaking*. Esta fue derivada por Jayaram Sethuraman en 1994 y se basa en la discretez del proceso Dirichlet.

Si se toma una muestra iid $m_1, m_2, \ldots$ de la medida centradora α, y si se genera una sucesión de pesos $\{w_i\}_{i=1}^{\infty}$ de la siguiente forma

$$w_1 = v_1 \quad \text{y} \quad w_i = v_i \prod_{j=1}^{i-1}(1 - v_j), \quad i \geq 2, \tag{2.4}$$

con $v_i \overset{iid}{\sim} \text{Beta}(1, M)$, entonces el proceso

$$P(B) = \sum_{i=1}^{\infty} w_i \delta_{m_i}(B), \quad B \in \mathcal{X}$$

es un proceso Dirichlet con medida centradora α y parámetro de concentración M.

En la figura 2.5 se muestran dos muestreos de cien simulaciones cada uno del proceso Dirichlet, cada uno con distinto parámetro de precisión.

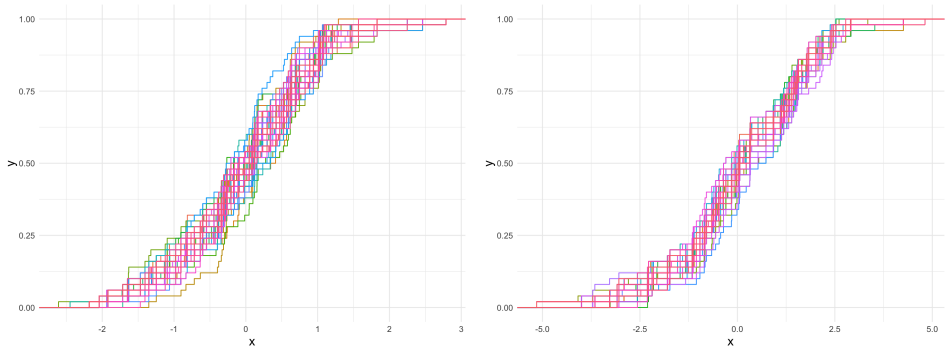

Figura 2.6: Cien muestras de un proceso Dirichlet con medida centradora $\text{N}(0, 1)$. A la izquierda, parámetro de precisión 1, a la derecha, 0.5.

La discretez del proceso Dirichlet hace que su soporte sólo contenga medidas discretas. Además, el soporte depende exclusivamente del soporte

de la medida centradora de la manera siguiente:

$$\mathrm{Sop}(PD(\alpha, M)) = \{\mu : \mathrm{Sop}(\mu) \subseteq \mathrm{Sop}(G)\},$$

es decir, el soporte del proceso Dirichlet son todas las medidas de probabilidad cuyos soportes estén contenidos en el soporte de la medida centradora, haciendo que, por ejemplo, si ésta es una distribución gaussiana, el soporte sea todas las distribuciones discretas real-valuadas.

Una propiedad clave de esta medida aleatoria es que ésta es conjugada para cualquier distribución P, es decir, si se tiene el modelo

$$x_1, \ldots, x_n \sim P$$
$$P \sim PD(\alpha, M)$$

entonces la distribución posterior es también un proceso Dirichlet con los siguientes parámetros:

$$P \,|\, x_1, \ldots, x_n \sim PD\left(\frac{M}{M+n}G + \frac{n}{M+n}\frac{\sum_{i=1}^{n}\delta_{x_i}}{n}, M+n\right) \tag{2.5}$$

Esto quiere decir que la posterior es una media ponderada entre la medida centradora *a priori* y la función de distribución empírica. Esto indica que, conforme se hace tender a M a 0, la distribución se vuelve cada vez más y más no informativa, pues en el límite, la medida centradora consiste solamente de la distribución empírica.

Otra caracterización del proceso Dirichlet que facilita muestrear de él es el esquema de urnas de Pólya desarrollado por Blackwell y MacQueen en 1973. Éste expresa las distribuciones finito-dimensionales del proceso de la siguiente forma:

$$\mathbb{P}\left(x_i \,|\, x_1, \ldots, x_{n-1}\right) = \frac{\sum_{j=1}^{i-1}\delta_{x_j}(x_i) + MG(x_i)}{M+i-1},$$

y si $x_1 \sim \frac{\alpha}{\alpha(\mathbb{X})} = G$, entonces

$$\mathbb{P}(x_n, \ldots, x_1) = \prod_{i=1}^{n} \frac{\sum_{j=1}^{i-1} \delta_{x_j}(x_i) + MG(x_i)}{M + i - 1}, \quad n \geq 1$$

Este esquema genera el siguiente algoritmo para muestrear del proceso Dirichlet.

Algoritmo 5 Muestreo del proceso Dirichlet

1: Simula $x_1 \sim \frac{\alpha}{\alpha(\mathbb{X})}$.
2: Con probabilidad $\frac{n}{n\alpha(\mathbb{X})}$, muestrea un valor de las ya obtenidos, y con probabilidad $\frac{\alpha(\mathbb{X})}{n+\alpha(\mathbb{X})}$ regresar al paso 1.

Como ya se había mencionado, una de las desventajas del proceso Dirichlet es que sus muestras son discretas casi seguramente, impidiendo su uso para estimar densidades continuas.

Una solución para esta problemática es mezclar sus trayectorias mediante alguna distribución continua, usando un proceso Dirichlet como medida mezcladora para algún kérnel paramétrico. Es decir, expresando a la densidad objetivo, f_θ, del siguiente modo:

$$\begin{aligned} f_\theta(x) &= \int k(\cdot \,|\, \theta)\, G(\mathrm{d}\theta) \\ G &\sim PD(\alpha, M) \end{aligned} \tag{2.6}$$

en donde $k(\cdot \,|\, \theta)$ es una densidad para todo θ.

Esto puede expresarse como un modelo jerárquico del siguiente modo:

$$x_i \,|\, \theta_i \sim f_{\theta_i}$$

$$\theta_i \,|\, G \overset{\text{iid}}{\sim} G$$

$$G \,|\, \eta, M \sim PD(M, G_\eta)$$

$$(\eta, M) \sim \pi$$

Por ejemplo, podemos tomar $G_\eta = \mathrm{N}\left(\mu, \sigma^2\right)$, es decir, $\eta = (\mu, \sigma^2)$; como distribución inicial se escoge a la distribución normal-gamma inversa.[5]

Una de las aplicaciones de este tipo de modelos es el análisis de conglomerados. Como el proceso Dirichlet es discreto, cada θ_i latente posee una probablidad positiva de empates, así, sea θ_j^*, con $j = 1, \ldots, k$ y $k \leq n$, el número de valores únicos. Sea $S_j = \{i : \theta_i = \theta_j^*\}$ y $n_j = |S_j|$ el número de θ_is empatadas con las θ_j^*s. De esta manera, el conjunto $\rho_n = \{S_1, \ldots, S_k\}$ forma una partición sobre los índices bajo los cuales los datos están etiquetados. Cabe destacar que, como las θ_is son aleatorias, entonces los S_j también lo son.

Por conveniencia, denotemos la pertenencia a algún *cluster* como $s_i = j$ si $i \in S_j$. Por definición, $s_1 = 1$. Sea k_i el número de θ_ls únicos de entre $\{\theta_1, \ldots, \theta_i\}$, y sea $n_{i,j}$ la multiplicidad del j-ésimo de estos valores únicos. Así,

$$\mathbb{P}\left(s_i = j \,|\, s_1, \ldots, s_{i-1}\right) = \begin{cases} \frac{n_{i-1,j}}{M+i-1} & \text{para } j = 1, \ldots, k_{i-1} \\ \frac{M}{M+i-1} & j = k_{i-1} + 1 \end{cases} \tag{2.7}$$

Denotemos $\mathbf{s_{-i}} = (s_i, \ldots, s_{i-1}, s_{i+1}, \ldots, s_n)$. Entonces, la probabilidad inicial $p(\rho_n)$ puede ser escrita como

$$p(s) = \prod_{i=2}^{n} p(s_i \,|\, s_1, \ldots, s_{i-1}) = \frac{M^{k-1} \prod_{j=1}^{k} (n_j - 1)!}{(M+1) \cdots (M+n-1)} \tag{2.8}$$

Sea $\theta_{i,j}^*$ el j-ésimo valor único de entre las $\{\theta_1, \ldots, \theta_i\}$. Nótese que $s_i = j$ implica que $\theta_i = \theta_{i-1,j}^*$ y $s_i = k_{i-1} + 1$ implica que $\theta_i \sim G_0$. De esto se tiene que

$$p(\theta_i \,|\, \theta_1, \ldots, \theta_{i-1}) \propto \sum_{j=1}^{k_{i-1}} n_{i-1,j}\, \delta_{\theta_{i-1,j}^*}(\theta_i) + M G_0(\theta_i)$$

Esta expresión puede simplificarse usando que el modelo es intercambiable sobre las θ_is.

$$p(\theta_i \,|\, \theta_{-\mathbf{i}}) \propto \sum_{k=1}^{k^-} n_j^-\, \delta_{\theta_j^{*-}}(\theta_i) + M G_0(\theta_i), \tag{2.9}$$

[5]Distribución de cuatro parámetros que es conjugada para la normal cuando ninguno de sus dos parámetros son conocidos.

en donde $\theta_{-\mathbf{i}}$ denota al vector θ sin el i-ésimo elemento θ_i; k^- es el número de valores únicos en $\theta_{-\mathbf{i}}$ y θ_j^{*-} es el j-ésimo elemento único.

La inferencia para este modelo puede realizarse mediante métodos MCMC.

Actualizamos las θ_j^* condicionalmente a la partición $\mathbf{s}$ mediante

$$p(\theta_j^* \,|\, \mathbf{s}, \mathbf{y}) \propto G_0(\theta_j^*) \prod_{i \in S_j} f_{\theta_j^*}(y_i),$$

en donde $f_{\theta_j^*}$ es la distribución de la cual se muestrean los datos.

Para muestrear de las s_i, construimos la distribución conjunta de s_i y θ_i condicional a $\theta_{-\mathbf{i}}$ y a $\mathbf{y}$. Primero, obtenemos la distribución posterior de θ_i, la cual es proporcional a

$$p(\theta_i \,|\, \theta_{-\mathbf{i}}, \mathbf{y}) \propto \sum_{j=1}^{k^-} n_j^- \, f_{\theta_j^{*-}}(y_i) \delta_{\theta_j^{*-}}(\theta_i) + M f_{\theta_i}(y_i) G_0(\theta_i)$$

Nótese que $f_{\theta_i}(y_i)G_0(\theta_i)$ no está normalizada, por lo que hacemos $H_0(\theta_i) \propto f_{\theta_i}(y_i)G_0(\theta_i)$ con constante de normalización $h_0(y_i) = \int f_\theta(y_i)\, G_0(\mathrm{d}\theta)$.

Entonces, como $\theta_i = \theta_i^{*-}$ implica que $s_i = j$, entonces la condicional pasada puede ser escrita como la conjunta entre θ_i y s_i:

$$p(\theta_i, s_i \,|\, \theta_{-\mathbf{i}}, \mathbf{y}) \propto \sum_{j=1} k^- n_j^- \, f_{\theta_j^{*-}}(y_i) \delta_j(s_i) \delta_{\theta_j^{*-}}(\theta_i) + M h_0(y_i) \delta_{k^-+1}(s_i) H_0(\theta_i)$$

Finalmente, marginalizamos con respecto a θ_i y con respecto a $\theta_{-\mathbf{i}}$, y obtenemos

$$p(s_i = j \,|\, \mathbf{s}_{-\mathbf{i}}.\mathbf{y}) \propto \begin{cases} n_j^- \, p(y_i \,|\, s_i = j, \mathbf{y}^{*-}{}_j) & \text{para } j = 1, \ldots, k^- \\ M \, h_0(y_i) & j = k^- + 1 \end{cases} \tag{2.10}$$

Con $p(y_i \,|\, s_i = j, \mathbf{y}^{*-}{}_j) = \int f_{\theta_j^*}(y_i)\, \mathrm{d}p(\theta_j^{*-} \,|\, \mathbf{y}_\mathbf{j}^{*-})$.

Ahora, para actualizar los hiperparámetros, nótese que la representación *stick-breaking* del proceso Dirichlet implica que, *a priori*, $\theta_j^* \,|\, \eta, k \overset{\text{iid}}{\sim} G_\eta$, para

$j = 1, \ldots, k$, por lo que

$$p(\eta \mid \theta^*) \propto p(\eta) \prod_{j=1}^{k} G_\eta(\theta_j^*),$$

en donde p es la distribución inicial para η, la cual es condicionalmente independiente de $\mathbf{s}$ y $\mathbf{y}$ dado θ^*.

Para M, fijamos una *a priori* gamma, es decir, $M \sim \text{Ga}(a, b)$. Usando la *a priori* para ρ_n, la verosimilitud para M es

$$\begin{aligned} p(k \mid M) &\propto M^k \Gamma(M)\Gamma(M+n) \\ &= M^k \frac{\Gamma(M+n)}{M\,\Gamma(n)} \int_0^n \eta^M (1-\eta)^{n-1}\,\mathrm{d}\eta \end{aligned}$$

Luego, introducimos una variable latente ϕ tal que

$$p(\phi \mid M, k, \ldots) = \text{Be}(M+1, n),$$

lo que nos lleva a que

$$p(M \mid \phi, k) = \pi \text{Ga}(a+k, b-\log(\phi)) + (1-\pi)\text{Ga}(a+k-1, b-\log(\phi)),$$

con

$$\frac{\pi}{1-\pi} = \frac{a+k-1}{n(b-\log(\phi))}$$

Finalmente, el algoritmo queda:

Algoritmo 6 Algoritmo MCMC para el modelo de mezclas del proceso Dirichlet

1: **for** i = 1,. . . ,n **do**
2: Muestrear $s_i \sim p(s_i \mid s_{-i}, \mathbf{y})$.
3: **end for**
4: **for** j = 1,. . . ,k **do**
5: Generar $\theta_j^* \sim p(\theta_j^* \mid \mathbf{s}, \mathbf{y})$
6: **end for**
7: Actualizar los hiperparámetros mediante alguna distribución de transición para η basada en la posterior condicional $p(\eta \mid \theta^*) \propto p_\eta(\eta) \prod_{j=1}^{k} G_\eta(\theta_j^*)$.
8: Para la el parámetro de precisión, generar $\phi \sim \text{Be}(M+1, n)$; después, evaluar $\frac{\pi}{1-\pi} = \frac{a+k-1}{n(b-\log(\phi))}$, y finalmente, generar

$$M \mid \phi, k \sim \begin{cases} \text{Ga}(a+k, b-\log(\phi)) & \text{con probabilidad } \pi \\ \text{Ga}(a+k-1, b-\log(\phi)) & \text{con probabilidad } 1-\pi \end{cases}$$

Otro modo de estimar un modelo del tipo (2.6) es aprovechando la representación *stick-breaking*, mediante la cual es posible reescribir a

$$f_\theta(x) = \int k(\cdot \mid \theta)\, G(\mathrm{d}\theta)$$

como

$$f_\theta(x) = \sum_{i=1}^{\infty} w_i k(x \mid \theta).$$

Más aún, como los pesos son productos de variables aleatorias independientes con distribución beta,

$$w_l = v_l \prod_{j<l} (1 - v_j),$$

con

$$v_i \overset{\text{iid}}{\sim} \text{Beta}(1, c),$$

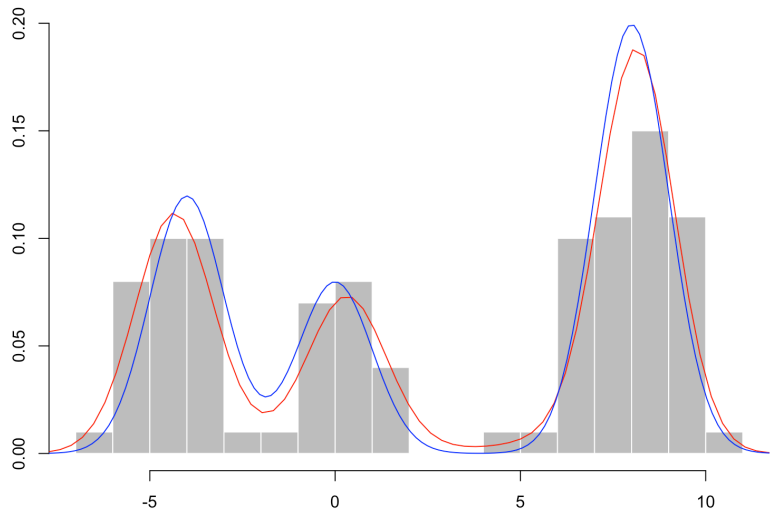

Figura 2.7: Estimación de la densidad mediante una mezcla de procesos Dirichlet de una mezcla de normales de varianza unitaria y con medias -8, 0 y 8. En azul, la densidad verdadera; en rojo, la estimada.

entonces su valor esperado es

$$\mathbb{E}\left[w_l\right] = \frac{1}{c+1}\frac{c}{c+1}^{l-1},$$

el cual decrece de manera geométrica al aumentar el índice. Esto implica que, aunque el número de componentes sea infinito, los pesos eventualmente se hacen tan pequeños que la contribución de sus componentes a la mezcla sea insignificante. En este espíritu, en (Ishwaran y Zarepour, 2000) se expuso un algoritmo de estimación que trunca la mezcla en N componentes, de modo que se aproxima

$$\sum_{i=1}^{\infty} w_i k(x \mid \theta_i)$$

mediante

$$\sum_{i=1}^{N} w_i k(x \,|\, \theta_i),$$

para una N fija.

El algoritmo consiste en, después de truncar la suma, introducir variables de localización d_i, de modo que, para cada observación x_i, d_i tome un valor entero que indique a qué componente de la mezcla ésta pertenece. Así, el modelo de mezclas truncado puede verse como

$$\begin{aligned} x_i \,|\, \theta_1, \ldots, \theta_N, d_i, \ldots, d_n &\sim \pi_{d_i}(x_i \,|\, \theta_{d_i}) \\ d_i \,|\, w_i &\sim \sum_{k=1}^{N} w_k \delta_k(\cdot) \\ (w_1, \ldots, w_N, \theta_1, \ldots, \theta_N) &\sim f(w) f(\theta,) \end{aligned} \tag{2.11}$$

en donde $f(\theta)$ denota la *prior* para θ y $f(w)$ es el truncamiento del proceso Dirichlet, de modo que los pesos sean calculados como en (2.4).

De este modo se implementa un muestreo de Gibs en el que en cada iteración se muestrea de las distribuciones

$$f(w, \theta \,|\, d_i, x_1, \ldots, x_n) \quad \text{y} \quad f(d_i \,|\, w, \theta, x_1, \ldots, x_n)$$

La elección del límite de truncamiento, N, resulta del análisis de la suma

$$\sum_{i=N}^{\infty} w_i \tag{2.12}$$

Para esto se puede utilizar la siguiente función

$$U_N(r) = \left(\sum_{i=N}^{\infty} w_i \right)^r$$

pues mediante esta se obtienen los momentos de (2.12):

$$\mathbb{E}\left[\sum_{i=N}^{\infty} w_i\right] = \mathbb{E}\left[U_N(1)\right] = \mathbb{E}\left[W_N(1)\right]$$

y

$$\text{Var}\left(UN(1)\right) = \mathbb{E}\left[U_N(2)\right] - \mathbb{E}\left[U_N(1)\right]$$

En (Ishwaran y Zarepour, 2000) se obtiene que, si las v_is de la represntación *stick breaking* siguen una distribución Beta$(1, M)$, entonces

$$\mathbb{E}\left[U_N(r)\right] = \left(\frac{M}{M+r}\right)^{N-1},$$

la cual es claramente una función creciente de M, de donde se sigue que, para valores de M pequeños, la esperanza también disminuye y se puede tomar un valor de truncamiento menor al que se tomaría si la M fuera más grande.

Aún así, en el artículo se obtiene empíricamente que el valor de esta suma es notablemente más sensible al de N que al de M, siendo pertinente la elección de un truncamiento relativamente grande para asegurar una buena aproximación a la suma infinita, sin importar el valor de M.

Label Switching

Un problema inherente a este tipo de modelos que puede resultar en malas estimaciones si no se le toma en cuenta es el *label switching*.

En 1982 Richard Redner y Homer Walker introdujeron el término como el fenómeno en el cual la verosimilitud de un modelo de mezclas permanece invariante cuando se cambian las etiquetas de los componentes.

Asúmase un modelo de mezclas

$$p(x \mid \mathbf{w}, \theta) = w_1 f(x, \theta_1) + w_2 f(x, \theta_2) + \cdots + w_k f(x, \theta_k)$$

Si se toma una permutación π de los parámetros, es decir,

$$\pi(\phi) = \pi(\theta, \mathbf{w}) = ((w_{\pi(1)}, \ldots, w_{\pi(k)}), (\theta_{\pi(1)}, \ldots, \theta_{\pi(k)})),$$

entonces la verosimilitud

$$L(\theta \,|\, \mathbf{x}) = \prod_i \left(w_1 f(x, \theta_1) + \cdots + w_k f(x, \theta_k)\right)$$

es idéntica para cualquier permutación de los parámetros. Esto crea el problema que, cuando la distribución *a priori* no lleva algún modo para distinguir entre los distintos componentes de la mezcla, entonces la posterior se vuelve simétrica entre componentes, lo que dificulta estimar sus parámetros individuales.

Uno de los intentos de solucionar el *label switching* es el imponer restricciones de identificabilidad sobre los parámetros, por ejemplo, hacer $\mu_1 < \mu_2 < \cdots < \mu_k$ para las medias. Esto puede ser hecho directamente sobre el espacio parametral, modificando la distribución *a priori*, o post procesando la muestra obtenida mediante el algoritmo MCMC, de modo que, si se tienen $\theta_{\mathbf{1}}, \theta_{\mathbf{2}}, \ldots$, se aplican permutaciones $\{\nu_i\}$ de modo que $\nu_1(\theta_{\mathbf{1}}), \nu_2(\theta_{\mathbf{2}}), \ldots$ satisfagan la condición.

En su tesis doctoral publicada en 1997, Mathew Stephens demuestra la justificación de esta equivalencia usando la siguiente proposición:

Proposición 3. *Considérese la restricción $\theta \in A$, en donde A es un conjunto tal que para toda θ, $\nu(\theta) \in A$ para una única permutación $\nu = \nu_\theta$. Sea g la función dada por $g(\theta) = \nu_\theta(\theta)$. Entonces*

$$\mathbb{E}\left[F(g(\Theta)) \,|\, \mathbf{x}\right] = \mathbb{E}\left[F(\Theta) \,|\, \mathbf{x}, \Theta \in A\right].$$

Esto garantiza que las medias muestrales permutadas converjan a la esperanza del modelo bajo la restricción, pues si se tienen N iteraciones del MCMC,

$$\frac{1}{N}\sum_{t=1}^{N} F(g(\mathbf{\Theta}_t)) \to \mathbb{E}\left[F(g(\mathbf{\Theta})) \,|\, \mathbf{x}\right] = \mathbb{E}\left[F(\mathbf{\Theta}) \,|\, \mathbf{x}, \mathbf{\Theta} \in A\right]$$

En la figura 2.7 se muestra el resultado de aplicar este método para un muestreo de Gibbs para la base de datos *fish* del paquete `bayesmix` de R.

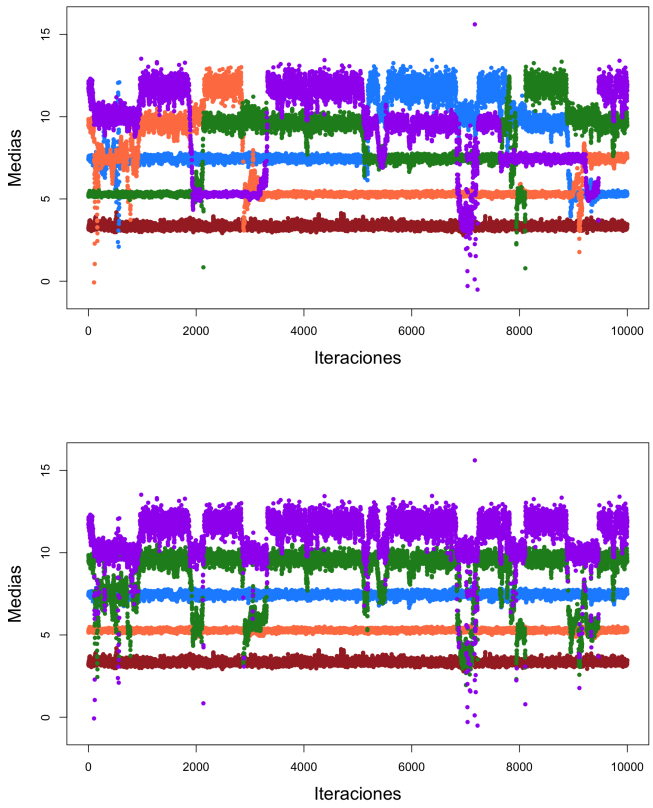

Figura 2.8: Diez mil iteraciones de un muestreo de Gibbs para las medias de la base de datos *fish* del paquete `bayesmix` de R. Arriba, etiquetado original; abajo, etiquetado tras ordenar las medias.

Muestreo de especies

En (Pitman, 1996) se utilizó el esquema de urnas propuesto por Blackwell y MacQueen para solucionar un problema común en biología: el muestreo de especies, en el cual, dada una muestra de n individuos de j diferentes especies, se intenta averiguar cuántas especies desconocidas existen y estimar

la probabildad de, si se observan m nuevos individuos, que estos provengan de k nuevas especies.

La formalización de este problema permite caracterizar al proceso Dirichlet mediante su distribución predictiva, lo que a su vez facilita la construcción de medidas aleatorias más generales que puedan funcionar como alternativa para desarrollar modelos.

Siguiendo la notación de (Hjort, Holmes, Müller, y Walker, 2010), un modelo de muestreo de especies es de la siguiente forma:

Definición 29 (Modelo de muestreo de especies)**.** Sea $(\tilde{p}_j)_{j\geq 1}$ una secuencia de pesos aleatorios no negativos tal que $\sum_{j\geq 1}\tilde{p}_j \leq 1$ y supóngase que $(\xi_n)_{n\geq 1}$ es una sucesión de variables aleatorias iid con función de distribución no atómica P_0. Si se toma a ξ_i como independiente de $\tilde{p}_j$, entonces la medida aleatoria

$$\tilde{p} = \sum_{j\geq 1}\tilde{p}_j\delta_{\xi_j} + \left(1 - \sum_{j\geq 1}\tilde{p}_j\right)P_0$$

es un modelo de muestreo de especies.

Cuando se da la igualdad $\sum_{j\geq 1}\tilde{p}_j = 1$, se dice que $\tilde{p}$ es propio, y es el caso que se considerará de aquí en adelante.

Cuando el número total de especies, j, es finita, el soporte de esta medida se colapsa al simplejo $j-1$-ésimo dimensional. Para el caso infinito dimensional, puede pensarse a $\tilde{p}_i$ como la proporción de la i-ésima especie, y a ξ_i como su etiqueta asignada. Como las ξ_i provienen de una distribución no atómica, entonces son distintas entre sí casi seguramente, lo que implica que especies distintas tengan etiquetas distintas con probabilidad 1. Además, un modelo de muestreo de especies genera una serie de distribuciones predictivas, las cuales se pueden caracterizar del siguiente modo:

Teorema 7 (Pitman, 1996)**.** *Sea $(\xi_n)_{n\geq 1}$ una sucesión iid de variables aleatorias condicional a una medida no atómica P_0. Entonces, condicionalmente a $\tilde{p}$, $(X_n)_{n\geq 1}$ es una sucesión de v.a. iid si y sólo si existe una colección de pesos $\{p_{j,n}(n_1,\ldots,n_k : 1\leq j\leq k, 1\leq k\leq n, n\geq 1\}$ tal que $X_1 = \xi_1$ y que,*

para todo $n \geq 1$,

$$X_{n+1} \,|\, X_1, \dots X_n = \begin{cases} \xi_{n+1} & \textit{con probabilidad } p_{k_n+1,n}(n_1, \dots, n_{k_n}, 1) \\ X^*_{n,j} & \textit{con probabilidad } p_{k_n,n}(n_1, \dots, n_j + 1, \dots, n_{k_n}) \end{cases},$$

en donde k_n *es el número de valores distintos* $X^*_{n,1}, \dots, X^*_{n,k_n}$ *en las observaciones,* X_i.

Para poder hacer uso de esta caracterización es necesario tener un modo de hacer inferencia sobre los pesos, para lo cual es posible usar un proceso *stick breaking*,

$$\tilde{p}_1 = V_1, \quad \tilde{p}_i = V_i \prod_{j=1}^{i-1}(1 - V_j), \quad \text{para } i = 2, \dots$$

en donde (V_i) es una sucesión de variables aleatorias iid en el $[0, 1]$. En particular, si $V_i \overset{\text{iid}}{\sim} \text{Beta}(1, \alpha(\mathbb{X}))$, estos son generados por un proceso Dirichlet con medida de concentración α.

Esto permite hacer inferencia y utilizar este enfoque en diversas aplicaciones en biología y lingüística. En (Lijoi, Mena, y Prünster, 2007) se generaliza este modelo a una familia de medidas aleatorias más generales, las *priors* tipo Gibbs, las cuales constituyen a las particiones aleatorias inducidas por la siguiente distribución conjunta:

$$\mathbb{P}\left(K_n = k, N_{1,n} = n_1, \dots, N_{k,n} = n_k\right) = V_{n,k} \prod_{j=1}^{k}(1 - \sigma)_{n_j - 1},$$

en donde K_n es el número de distintas especies en una muestra de n individuos; las $N_{i,n}$ el número de individuos de la muestra que pertenecen a la i-ésima especie; $\{V_{n,k} : n \geq 1, 1 \leq k \leq n\}$ es la sucesión de pesos no negativos; $\sigma \in (0, 1)$ y $(a)_n$ es el factorial ascendente.[6]

Las distribuciones tipo Gibbs inducen una forma simple de la probabilidad de muestrear una nueva especie en el muestreo $n + m + 1$, condicional a las observaciones $X_j^{1:n}$, y a que en éstas se observaron j especies distintas

[6] $(a)_n = a(a + 1) \cdots (a + n - 1)$.

es

$$D_m^{n:j} = \sum_{k=0}^{m} \frac{V_{n+n+1,j+k+1}}{V_{n,j}} \frac{1}{\sigma^k} \mathscr{C}(m,k;\sigma,-n+j\sigma),$$

en donde $\mathscr{C}(n,k;\sigma,\gamma)$ es un coeficiente factorial generalizado no centrado,

$$\mathscr{C}(n,k;\sigma,\gamma) = \frac{1}{k!} \sum_{j=0}^{k} (-1)^j \binom{k}{j} (-j\sigma-\gamma)_n$$

Además, si $\sigma \to 0$, esta familia de medidas colapsa en un proceso Dirichlet con medida de concentración α, de modo que si $\alpha(\mathbb{X}) = \theta \in (0,\infty)$, entonces $V_{n,k} = \theta^k/(\theta)_n$, y $D_m^{n:j}$ toma la forma

$$D_m^{n:j} = \frac{\theta}{(\theta+n)_{m+1}} \sum_{k=0}^{m} \theta^k \sum_{l=k}^{m} \binom{m}{l} \left|s(l,k)\right| (n)_{m-l},$$

en donde $\left|s(l,k)\right|$ es un número de Stirling sin signo de primer orden.

Proceso geométrico

Otra de las formas de resolver el *label switching*, además de reestricciones de identificabilidad, es usar medidas que lleven la información del orden integrada, de modo que este fenómeno no se presente.

En (Fuentes-García, Mena, y Walker, 2010) se expuso una medida aleatoria con una construcción simple y con esta propiedad que sirve como alternativa al proceso Dirichlet: el proceso geométrico.

Se toma el siguiente modelo

$$f(y \mid N) = \frac{1}{N} \sum_{l=1}^{N} K(y;\theta_l)$$

$$N \sim q_N,$$

en donde q_N es una distribución sobre los enteros positivos, y se tiene una N para cada observación.

Si se marginaliza N, se tiene

$$f(y) = \sum_{N=1}^{\infty} \frac{1}{N} \sum_{l=1}^{N} K(y;\theta_l) q_N$$

$$= \sum_{l=1}^{\infty} \sum_{N=l}^{\infty} \frac{q_N}{N} K(y;\theta_l)$$

Haciendo $w_i = \sum_{N=l}^{\infty} \frac{q_N}{N}$,

$$f(y) = \sum_{l=1}^{\infty} w_l K(y;\theta_l) = \int K(y,\theta)\, P(\mathrm{d}\theta),$$

en donde

$$P(\mathrm{d}\theta) = \sum_{l=1}^{\infty} w_l\, \delta_{\theta_l}(\mathrm{d}\theta)$$

Estos pesos tienen dos características importantes:

1. Suman 1.
2. Son decrecientes, lo que hace que este modelo no presente los problemas identificabilidad que ocurren con otras mediads aleatorias, como el proceso Dirichlet.

La elección de la distribución para q_N repercute de manera directa en la complejidad de la estimación, pues no todas las elecciones llevan a formas simples de los pesos. En (Fuentes-García y cols., 2010) se toma

$$q_N \sim \mathrm{BinNeg}(2, \lambda),$$

con una distribución $\mathrm{Beta}(a, b)$ como híper-*a priori* para λ, lo que lleva a que los pesos sigan una distribución geométrica, es decir son de la forma

$$w_i = \lambda(1-\lambda)^{i-1}$$

Además, se introduce una variable latente de localización, d_i.

Si a las θ_i se les asigna una distribución *a priori* g, las condicionales completas para el modelo son de la siguiente forma:

- $f(\theta_j \mid \cdots) \propto g(\theta_j) \prod_{d_i=j} K(y_i; \theta_j)$.
- $\mathbb{P}(d_i = l \mid \cdots) \propto K(y_i; \theta_l) \mathbb{1}(l \in \{1, \ldots, N_i\})$
- $\mathbb{P}(N_i = N \mid \cdots) = (1-\lambda)^{N-1}(N \geq d_i)$
- $\mathbb{P}(\lambda \mid \cdots) = \text{Beta}(a + 2n, b - n + \sum_{i=1}^{n} N_i)$

Este modelo está fuertemente conectado con el proceso Dirichlet, pues manipulando el valor esperado de los pesos obtenidos mediante su representación *stick breaking*,

$$\begin{aligned} \mathbb{E}[w_l] &= \frac{1}{c+1}\frac{c}{c+1}^{l-1} \\ &= \lambda(1-\lambda)^{l-1}, \end{aligned}$$

haciendo $\lambda = \frac{1}{c+1}$. Esto significa que el proceso geométrico puede ser interpretado como una variación de la construcción del proceso Dirichlet, en la que se reemplaza la aleatoriedad de las v_i por sus valores esperados.

Proceso Poisson-Dirichlet de dos parámetros

Una generalización del proceso Dirichlet es el proceso Poisson-Dirichlet de dos parámetros, PPD, introducida en (Pitman y Yor, 1997).

Definición 30 (Distribución Poisson-Dirichlet de dos parámetros (Pitman y Yor, 1997))**.** Sea $\{V_i\}$ una sucesión de variables aleatorias independientes tal que, para $0 \leq \alpha < 1$ y $\theta > -\alpha$,

$$V_k \sim \text{Beta}(1-\alpha, \theta + k\alpha).$$

Sea

$$\begin{aligned} w_1 &= V_1 \\ w_k &= (1-V_1)\cdots(1-V_{k-1})V_k \qquad (k \geq 2), \end{aligned}$$

entonces, a la distribución del vector $(\tilde{w}_1, \tilde{w}_2, \dots)$, en donde $\tilde{w}_i$ denota a las w_i ordenadas de manera decreciente, se le conoce como distribución Poisson-Dirichlet de dos parámetros, $PD(\alpha, \theta)$.

De aquí se deriva la definición del PPD, la cual es muy similar a la del proceso Dirichlet.

Definición 31 (Proceso Poisson-Dirichlet de dos parámetros)**.** Sea $\{X_i\}_{i=1}^{\infty}$ una sucesión de variables aleatorias iid con distribución H. Si se toma $w_i \sim \text{PD}(\alpha, \theta)$, para $i = 1, \dots,$ entonces

$$\sum_{i=1}^{\infty} w_i \, \delta_{X_i}$$

es un proceso Poisson-Dirichlet con parámetros α y θ, y con medida base H, PPD(α,θ,H).

Nótese que cuando $\alpha = 0$, se recupera al proceso Dirichlet. Además, $w_1 > w_2 > \cdots > 0$ y $\sum w_i = 1$, casi seguramente.

La distribución predictiva de esta medida aleatoria también admite una representación como un esquema de urnas de Pólya, de modo que, si se tiene una muestra, $(X_1, \dots, X_n)$, con k valores distintos, $X_1^*, X_2^*, \dots,$ y tal que n_j de ellos son iguales a X_j^*, entonces

$$\mathbb{P}\left(X_{n+1} \in \mathrm{d}x \,|\, X_1, \dots, X_N\right) = \frac{\theta + k\alpha}{\theta + n} H(\mathrm{d}x) + \frac{1}{\theta + n} \sum_{j=1}^{k} (n_j - \alpha) \delta_{X_j^*}(\mathrm{d}x)$$

Nótese que, en este caso, a diferencia del Proceso Dirichlet, hay una dependencia del número de grupos, lo que hace que sea más adaptable a escenarios en los que la distribución posterior difiera mucho de la distribución inicial. Además, su simplicidad computacional la hace una buena alternativa al proceso Dirichlet al construir modelos.

Capítulo 3

Modelo no paramétrico para series de tiempo estacionarias

Una de las limitaciones de los modelos clásicos de series temporales son los supuestos distribucionales que, en muchos casos, no se cumplen al trabajar con datos reales. De este modo, las bondades de los modelos que utilizan procesos estacionarios pueden resultar opacadas por una mala elección de la distribución invariante de los mismos, por lo que sería deseable contar con un modelo cuya distribución estacionaria fuera lo suficientemente flexible como para adaptarse a cualquier conjunto de datos a los que se ajuste.
Una solución natural a este problema está dada dentro de la estadística bayesiana no paramétrica. Isadora Antoniano - Villalobos y Stephen Walker propusieron en 2015 el usar una mezcla de procesos Dirichlet *a priori* para la distribución estacionaria de un proceso de Markov, ganando así adaptabilidad sin perder las facilidades de estimación que la estacionariedad provee.

En (Antoniano-Villalobos y Walker, 2016) se propone un modelo con transiciones y densidad estacionaria no paramétricas, de manera que se obtenga un modelo sencillo de estimar pero lo suficientemente flexible como para poder utilizarse en muchos contextos.

Si bien es posible extender la siguiente construcción a procesos con una estructura de dependencia de orden más alto, por simplicidad se comienza con un modelo normal autorregresivo de primer orden, AR(1).

Denotemos por

$$K_\theta(y, x) = \mathrm{N}_2\left((y, x) \,|\, (\mu, \mu), \Sigma\right)$$

a la densidad de una normal bivariada con media $\mu \in \mathbb{R}$ y matriz de covarianzas

$$\Sigma = \sigma^2 \begin{pmatrix} 1 & \rho \\ \rho & 1 \end{pmatrix}$$

para alguna $-1 < \rho < 1$ y $\sigma^2 > 0$, por lo que $\theta = (\mu, \rho, \sigma)$. Claramente,

$$K_\theta(y) = \int K_\theta(y \,|\, x)\, K_\theta(x)\, \mathrm{d}x = \mathrm{N}\left(y \,|\, \mu, \sigma^2\right)$$

por lo que es posible definir un proceso de Markov homogéneo en el tiempo con densidad estacionaria $K_\theta(y)$ a través de la probabilidad de transición

$$\mathbb{P}(x_{n+1} \in A \,|\, x_n) = \int_A K_\theta(y \,|\, x_n)\, \mathrm{d}y; \quad A \in \mathcal{A}$$

con su correspondiente densidad de transición dada por

$$K_\theta(y \,|\, x) = \mathrm{N}\left(y \,|\, \mu + \rho(x - \mu), (1 - \rho^2)\sigma^2\right)$$

Nótese que esto corresponde al modelo AR(1) usual, con una parametrización escogida para garantizar estacionariedad sin condiciones adicionales sobre los parámetros.

Se define entonces una mezcla no paramétrica sobre la densidad bivariada $K_\theta(y, x)$, lo que preserva la estacionariedad. En el caso general tomamos

$$f_P(y, x) = \int K_\theta(y, x)\, \mathrm{d}P(\theta)$$

En particular, f puede ser representada como

$$f_P(y, x) = \sum_{j=1}^{\infty} w_j K_{\theta_j}(y, x),$$

en donde P es una medida de probabilidad discreta dada por

$$P = \sum_{j=1}^{\infty} w_j \delta_{\theta_j}$$

En (Antoniano-Villalobos y Walker, 2016) se toma $P \sim PD$, pero esta pudiera ser remplazada por cualquier medida aleatoria con soporte sobre el espacio de medidas de probabilidad discretas, como el proceso geométrico o el Poisson-Dirichlet.

Análogo al caso paramétrico, se define a la densidad de transición como la densidad condicional

$$f_P(y \,|\, x) = \frac{\sum_{j=1}^{\infty} w_j K_{\theta_j}(y, x)}{\sum_{j=1}^{\infty} w_j K_{\theta_j}(x)}$$

Entonces la probabilidad de transición está dada por

$$\mathbb{P}(x_{n+1} \in A \,|\, x_n) = \int_A f_P(y \,|\, x_n)\, \mathrm{d}y; \quad A \in \mathcal{A}$$

la cual define un proceso temporalmente homogéneo estacionario de primer orden con densidad invariante

$$f_P(y) = \sum_{j=1}^{\infty} w_j K_{\theta_j}(y)$$

La transición puede ser expresada como una mezcla no paramétrica de densidades de transición con pesos dependientes

$$f_P(y \,|\, x) = \sum_{j=1}^{\infty} w_j(x) K_{\theta_j}(y \,|\, x) \tag{3.1}$$

donde

$$w_j(x) = \frac{w_j \, K_{\theta_j}(y \,|\, x)}{\sum_{i=1}^{\infty} w_i K_{\theta_i}(x)} \tag{3.2}$$

Una vez construido el modelo, pasemos a su estimación.

3.1. Función de verosimilitud

Consideremos una muestra $\mathbf{x_n} = (x_0, \ldots, x_n)$. La función de verosimilitud para el modelo está dada por

$$\begin{aligned} f_P(\mathbf{x_n}) &= f_P(x_0) \prod_{i=1}^{n} f_P(x_i \,|\, x_{i-1}) \\ &= f_P(x_0) \prod_{i=1}^{n} \left(\sum_{j=1}^{\infty} w_j(x_{i-1}) K_{\theta_j}(x_i \,|\, x_{i-1}) \right), \end{aligned} \tag{3.3}$$

en donde los pesos dependientes están dados por (3.2), y se asume que la primera observación proviene de la densidad estacionaria $f_P(x_0) = \sum_{j=1}^{\infty} w_j K_{\theta_j}(x_0)$.

Para simplificar la notación, consideremos, sin pérdida de generalidad, la verosimilitud condicional

$$f_P(\mathbf{x_n} \,|\, x_0) = \prod_{i=1}^{n} f_P(x_i \,|\, x_{i-1}) = \prod_{i=1}^{n} \left(\sum_{j=1}^{\infty} w_j(x_{i-1}) K_{\theta_j}(x_i \,|\, x_{i-1}) \right), \tag{3.4}$$

asumiendo un punto inicial fijo $X_0 = x_0$.

Se han propuesto diferentes métodos de inferencia para el tipo de modelo de verosimilitud utilizado en (3.4). En (Antoniano-Villalobos y Walker, 2016) se utilizó la idea propuesta por Muliere y Tardella, 1998; Ishwaran y Zarepour, 2000; Papaspiliopoulos y Roberts, 2008; Kalli *et al.*, 2011; en la cual se muestrea una cantidad finita, pero suficiente, de variables en cada iteración de la simulación de una cadena de Markov con la distribución estacionaria deseada.

Se introduce entonces, para cada i, una variable de asignación $d_i \in \{1, 2, \ldots\}$, y se utiliza el siguiente modelo latente:

$$\begin{aligned} f_P(\mathbf{x_n}, \mathbf{d_n}) &= \prod_{i=1}^{n} w_{d_i}(x_{i-1}) K_{\theta_{d_i}}(x_i \,|\, x_{i-1}) \\ &= \frac{\prod_{i=1}^{n} w_{d_i} K_{\theta_{d_i}}(x_{i-1}) K_{\theta_{d_i}}(x_i \,|\, x_{i-1})}{\prod_{i=1}^{n} \sum_{j=1}^{\infty} w_j K_{\theta_j}(x_{i-1})} \end{aligned} \tag{3.5}$$

Ahora, de aquí en adelante considérese un kérnel Gaussiano bivariado y mézclese sobre la media y el coeficiente de correlación, manteniendo fija la varianza a través de lo componentes de la mezcla. Tómese entonces

$$K_{\theta_j}(y \,|\, x) = \mathrm{N}\left(y \,|\, \mu_j + \rho_j(x - \mu_j), (1 - \rho_j^2)\sigma^2\right)$$

$$K_{\theta_j}(x) = \mathrm{N}\left(x \,|\, \mu_j, \sigma^2\right)$$

Para este caso el denominador en (3.5) puede ser reescrito como

$$\sigma^{-n} \prod_{i=1}^{n} \left(\sum_{j=1}^{\infty} w_j \exp\left\{ -\frac{1}{2}(x_{i-1} - \mu_j)^2/\sigma^2 \right\} \right)$$

Como los términos del producto están acotados por 1, se puede utilizar que

$$\sum_{k=1}^{\infty} (1 - c)^k = c^{-1},$$

igualdad que se satisface para cualquier $0 < c < 1$. Por ende podemos escribir el denominador como

$$\frac{1}{\sigma^{-n} \left(\sum_{j=1}^{\infty} w_j \exp\left\{ -\frac{1}{2}(x_{i-1} - \mu_j)^2/\sigma^2 \right\} \right)} =$$

$$\sigma^n \sum_{k=0}^{\infty} \left(1 - \sum_{j=1}^{\infty} w_j \exp\left\{ -\frac{1}{2}(x_{i-1} - \mu_j)^2/\sigma^2 \right\} \right)^k$$

Para poder trabajar con esta suma, de manera análoga a lo expuesto en (Ishwaran y Zarepour, 2000), truncamos a m pesos; alternativamente se podrían utilizar variables latentes para formar un *slice sampler*, pero esto resulta computacionalmente poco eficiente. Usando la primera alternativa, el denominador queda

$$\sigma^n \sum_{k=0}^{\infty} \left(1 - \sum_{j=1}^{m} w_j \exp\left\{ -\frac{1}{2}(x_{i-1} - \mu_j)^2/\sigma^2 \right\} \right)^k$$

Pero como la suma converge, la sucesión debe converger a cero, por lo que podemos truncar para un número lo suficientemente grande de términos, ya que todos los subsecuentes debieran ser tan pequeños que sean irrelevantes, en particular, al ser una suma geométrica, entonces si truncamos la suma en l sumandos, el error que se tendría sería

$$\sum_{k=1}^{\infty}(1-c)^k - \sum_{k=1}^{l}(1-c)^k = \frac{1}{c} - \frac{1-(1-c)^l}{c} = \frac{(1+c)^l}{c},$$

y si se agregaran otros t términos,

$$\text{error}(t+l) = \frac{(1+c)^(l+t)}{c} = (1+c)^t\frac{(1+c)^l}{c},$$

es decir, decrecería el error por un factor de c^t, pues $|c|<1$, lo que puede interpretarse como que el número de decimales correctos en la suma es proporcional al número de términos que se suman[1], lo que da un criterio para escoger el valor de truncamiento, l.
Truncando, queda

$$\sigma^n \sum_{k=0}^{l}\left(1-\sum_{j=1}^{m} w_j \exp\left\{-\frac{1}{2}(x_{i-1}-\mu_j)^2/\sigma^2\right\}\right)^k \tag{3.6}$$

Por lo que la verosimilitud queda

$$\sigma^n \prod_{i=1}^{n} w_{d_i}\mathrm{N}\left((x_i,x_{i-1})|(\mu_{d_i},\mu_{d_i}),\Sigma_{d_i}\right)$$

$$\sum_{k=0}^{l}\left(1-\sum_{j=1}^{m} w_j \exp\left\{-\frac{1}{2}(x_{i-1}-\mu_j)^2/\sigma^2\right\}\right)^k$$

$$\propto \sigma^n \prod_{i=1}^{n}\frac{w_{d_i}}{\sigma\sqrt{1-\rho_{d_i}^2}}\exp\left\{-\frac{1}{2}(x_i-\mu_{d_i}-\rho_{d_i}(x_{i-1}-\mu_{d_i}))^2/(1-\rho_{d_i}^2)\sigma^2)\right\}$$

[1] (Bradie, 2006)

$$\exp\left\{-\frac{1}{2}(x_{i-1}-\mu_{d_i})^2/\sigma^2\right\}\sum_{k=0}^{l}\left(1-\sum_{j=1}^{m} w_j \exp\left\{-\frac{1}{2}(x_{i-1}-\mu_j)^2/\sigma^2\right\}\right)^k$$

$$=\prod_{i=1}^{n} w_{d_i}\exp\left\{-\frac{1}{2\sigma^2}\left(\frac{(x_i-\mu_{d_i}-\rho_{d_i}(x_{i-1}-\mu_{d_i}))^2}{1-\rho_{d_i}^2}+(x_{i-1}-\mu_{d_i})^2\right)\right\}$$

$$(1-\rho_{d_i}^2)^{-\frac{1}{2}}\sum_{k=0}^{l}\left(1-\sum_{j=1}^{m} w_j \exp\left\{-\frac{1}{2}(x_{i-1}-\mu_j)^2/\sigma^2\right\}\right)^k$$

Con $d_i \in \{1,2,\ldots,k\}$.

3.2. Inferencia posterior vía cadenas de Markov Monte Carlo

Para hacer inferencia sobre el modelo latente anterior, es necesario asignar las distribuciones *a priori* a P, a σ y a $(w_j,\mu_j,\rho_j)_{j=1}^m$. Para P se escoge un proceso Dirichlet.

Para $\tau=\sigma^{-2}$, se usa una a priori gamma, y para cada ρ_j, una uniforme discreta en $R\subset(-1,1)$, todas independientes entre las j. Las (μ_j) son tomadas iid de una medida base, en este caso de una distribución normal.

Las distribuciones a priori determinan una densidad conjunta para todas las variables, la cual necesita ser muestrada mediante un algoritmo MCMC.

En cada iteración del MCMC, los $(w_j)_{j=1}^m$ pueden ser calculados mediante $w_1=v_1$ y $w_j=v_j\prod_1^{j-1}(1-v_i)$. Las (v_i) deben ser muestreadas de manera independiente tomando una distribución Beta$(1,b)$ como prior, y la distribución condicional completa queda

$$\mathbb{P}\left(v_j \mid \cdots\right) \propto v_j^{n_j}(1-v_j)^{n_j^+}(1-v_j)^{b-1}$$

$$\prod_{i=1}^{n}\sum_{k=0}^{l}\left(1-\sum_{j=1}^{m} w_j \exp\left\{-\frac{\tau}{2}(x_{i-1}-\mu_j)^2\right\}\right)^k$$

$$\propto \text{Beta}(n_j + 1, n_j^+ + b) \prod_{i=1}^{n} \sum_{k=0}^{l} \left(1 - \sum_{j=1}^{m} w_j \exp\left\{ -\frac{\tau}{2} (x_{i-1} - \mu_j)^2 \right\} \right)^k , \tag{3.7}$$

con $n_j^+ = \sum \mathbb{1}(d_i > j)$. Como no es práctico simular directamente de esta distribución, usamos un paso Metropolis-Hastings con *proposal distribution* igual al factor beta y con probabilidad de aceptación dada por el mínimo entre 1 y

$$\frac{Q}{\prod_{i=1}^{n} \sum_{k=0}^{l} \left(1 - \sum_{j=1}^{m} w_j \exp\left\{ -\frac{\tau}{2} (x_{i-1} - \mu_j)^2 \right\} \right)^k},$$

en donde

$$Q = \prod_{i=1}^{n} \sum_{k=0}^{l} \left(1 - \sum_{l<j} w_l \exp\left\{ -\frac{\tau}{2} (x_{i-1} - \mu_l)^2 \right\} - \sum_{l \geq j} w_l^* \exp\left\{ -\frac{\tau}{2} (x_{i-1} - \mu_l)^2 \right\} \right)^k ,$$

y con w_l^* los pesos calculados con el v_j propuesto.

Para las variables de localización, (d_i), se tiene la siguiente condicional completa:

$$\mathbb{P}\left(d_i = j \mid \cdots\right) \propto w_j K_{\theta_j}(x_i, x_{i-1}); \quad j = \{1, 2, \dots\} \tag{3.8}$$

Una *a priori* discreta, π, para el coeficiente de correlación entre componentes, ρ_j, resulta en una distribución condicional discreta, con

$$\mathbb{P}\left(\rho_j = r \mid \cdots\right) \propto (1 - r^2)^{-n_j/2} \pi(r) \exp\left\{ \sum_{d_i = j} -\frac{1}{2\sigma^2} \left(\frac{(x_i - \mu_{d_i} - r(x_{i-1} - \mu_{d_i}))^2}{1 - r^2} \right) \right\}, \tag{3.9}$$

con $n_j = \sum_i \mathbb{1}(d_i = j)$.

para cada $r \in R$. Para las medias, μ_l, tomando como prior una normal con media m y precisión t, la condicional completa es

$$\mathbb{P}\left(\mu_l \mid \cdots\right) \propto \mathrm{N}\left(\mu_l, \mid m, t\right)$$

$$\prod_{d_i=l} \mathrm{N}\left(x_i \mid \mu_{d_i} + \rho_{d_i}(x_{i-1} - \mu_{d_i}), (1-\rho_{d_i}^2)\sigma^2\right) \mathrm{N}\left(x_{i-1} \mid \mu_{d_i}, \sigma^2\right)$$

$$\prod_{i=1}^{n} \sum_{k=0}^{l} \left(1 - \sum_{j=1}^{m} w_j \exp\left\{-\frac{1}{2}(x_{i-1} - \mu_j)^2/\sigma^2\right\}\right)^k$$

Usando que la normal es conjugada para la media de otra gaussiana,

$$\mathbb{P}\left(\mu_l \mid \cdots\right) \propto \mathrm{N}\left(\mu_l \mid \frac{mt + \frac{\tau}{1-\rho_{d_i}^2} \sum_{d_i=l} \frac{x_i - \rho_{d_i} x_{i-1}}{1+\rho_{d_i}}}{t + n_l \frac{\tau}{1-\rho_{d_i}^2}}, t + n_l \frac{\tau}{1-\rho_{d_i}^2}\right)$$

$$\prod_{d_i=l} \mathrm{N}\left(x_{i-1} \mid \mu_{d_i}, \tau\right)$$

$$\prod_{i=1}^{n} \sum_{k=0}^{l} \left(1 - \sum_{j=1}^{m} w_j \exp\left\{-\frac{1}{2}(x_{i-1} - \mu_j)^2/\sigma^2\right\}\right)^k$$

Usando de nuevo la misma propiedad,

$$\mathbb{P}\left(\mu_l \mid \cdots\right) \propto \mathrm{N}\left(\frac{mt + \frac{\tau}{1-\rho_{d_i}^2} \sum_{d_i=l} \frac{x_i - \rho_{d_i} x_{i-1}}{1+\rho_{d_i}} + \sum_{d_i=l} x_{i-1}}{S}, S\right)$$

$$\prod_{i=1}^{n} \sum_{k=0}^{l} \left(1 - \sum_{j=1}^{m} w_j \exp\left\{-\frac{1}{2}(x_{i-1} - \mu_j)^2/\sigma^2\right\}\right)^k$$

De donde se sigue que,

$$\mathbb{P}\left(\mu_l \mid \cdots\right) \propto \mathrm{N}\left(\frac{mt + \frac{\tau}{1-\rho_{d_i}^2}\sum_{d_i=l}\frac{x_i+x_{i-1}}{1+\rho_{d_i}}}{S}, S\right)$$

$$\prod_{i=1}^{n}\sum_{k=0}^{l}\left(1-\sum_{j=1}^{m} w_j \exp\left\{-\frac{1}{2}(x_{i-1}-\mu_j)^2/\sigma^2\right\}\right)^k, \qquad (3.10)$$

en donde

$$S = t + n_l\frac{\tau}{1-\rho_{d_i}^2} + n_l\tau$$

Entonces usamos un paso slice sampler para muestrear de esta distribución en cada iteración del Gibbs Sampler general.

Para $\tau = \sigma^{-2}$, como se le fue asignada una a priori $\mathrm{Gamma}(\tau \mid a, c)$, entonces la condicional es:

$$\mathbb{P}\left(\tau \mid \cdots\right) \propto \sigma^n \mathrm{Gamma}(\tau \mid a, b)$$

$$\prod_{i=1}^{n}\mathrm{N}\left(x_i \mid \mu_{d_i} + \rho_{d_i}(x_{i-1}-\mu_{d_i}, (1-\rho_{d_i}^2)\sigma^2\right)\mathrm{N}\left(x_{i-1} \mid \mu_{d_i}, \sigma^2\right)$$

$$\prod_{i=1}^{n}\sum_{k=0}^{l}\left(1-\sum_{j=1}^{m} w_j \exp\left\{-\frac{1}{2}(x_{i-1}-\mu_j)^2/\sigma^2\right\}\right)^k$$

Usando que la gamma es conjugada para la precisión de la gaussiana,

$$\mathbb{P}\left(\tau \mid \cdots\right) \propto \sigma^n \mathrm{Gamma}\left(\tau \mid a + \frac{n}{2},\, b + \frac{\sum_{i=1}^{n}(x_{i-1}-\mu_{d_i})^2}{2}\right)$$

$$\prod_{i=1}^{n}\mathrm{N}\left(x_i \mid \mu_{d_i} + \rho_{d_i}(x_{i-1}-\mu_{d_i}, (1-\rho_{d_i}^2)\sigma^2\right)$$

$$\prod_{i=1}^{n}\sum_{k=0}^{l}\left(1-\sum_{j=1}^{m} w_j \exp\left\{-\frac{1}{2}(x_{i-1}-\mu_j)^2/\sigma^2\right\}\right)^k$$

$$\propto \sigma^n \tau^{a+n/2-1} \exp\left\{-\tau\left(b + \frac{\sum_{i=1}^n (x_{i-1} - \mu_{d_i})^2}{2}\right)\right\}$$

$$\sigma^{-n} \exp\left\{\sum_{i=1}^n -\frac{\tau}{2}\left(\frac{(x_i - \mu_{d_i} - \rho_{d_i}(x_{i-1} - \mu_{d_i}))^2}{1-\rho_{d_i}^2}\right)\right\}$$

$$\prod_{i=1}^n \sum_{k=0}^l \left(1 - \sum_{j=1}^m w_j \exp\left\{-\frac{1}{2}(x_{i-1} - \mu_j)^2/\sigma^2\right\}\right)^k$$

$$\propto \exp\left\{-\tau\left(b + \frac{\sum_{i=1}^n (x_{i-1} - \mu_{d_i})^2}{2} + \right.\right.$$

$$\left.\left.\tau^{a+n/2-1}\frac{1}{2}\sum_{i=1}^n \frac{(x_i - \mu_{d_i} - \rho_{d_i}(x_{i-1} - \mu_{d_i}))^2}{1-\rho_{d_i}^2}\right)\right\}$$

$$\prod_{i=1}^n \sum_{k=0}^l \left(1 - \sum_{j=1}^m w_j \exp\left\{-\frac{1}{2}(x_{i-1} - \mu_j)^2/\sigma^2\right\}\right)^k$$

Así, la condicional completa es

$$\Gamma\left(\tau \,|\, a + \frac{n}{2},\ b + \frac{\sum_{i=1}^n (x_{i-1} - \mu_{d_i})^2}{2} + \frac{1}{2}\sum_{i=1}^n \frac{(x_i - \mu_{d_i} - \rho_{d_i}(x_{i-1} - \mu_{d_i}))^2}{1-\rho_{d_i}^2}\right)$$

$$\prod_{i=1}^n \sum_{k=0}^l \left(1 - \sum_{j=1}^m w_j \exp\left\{-\frac{\tau}{2}(x_{i-1} - \mu_j)^2\right\}\right)^k, \tag{3.11}$$

en donde $\Gamma(a, b)$ denota una distribución gamma con parámetro de forma a y parámetro de medida[2] b.

En este caso, de nuevo volvemos a usar un paso *slice sampler* para muestrear de esta condicional completa.

De este modo, el muestreo de Gibbs para el modelo queda

[2]En inglés, *rate parameter*, el inverso del parámetro de escala.

1: Inicializa los valores iniciales $d_i^{(1)}$, $v_i^{(1)}$, $\mu_i^{(1)}$, $\tau^{(1)}$ y $\rho_i^{(1)}$.
2: **for** k = 2,. . . ,n **do**
3: Simula $d_i^{(k)}$ de la distribución discreta (3.8).
4: Simula $v_i^{(k)} \sim$ (3.7) usando un paso Metropolis-Hastings.
5: Calcula los pesos haciendo $w_1^{(k)} = v_1^{(k)}$, $w_i^{(k)} = v_i^{(k)} \prod_{l<j}(1 - v_i^{(k)})$.
6: Simula $\mu_i^{(k)} \sim$ (3.10) usando un paso *slice sampler*.
7: Simula $\tau^{(k)} \sim$ (3.11) usando un paso *slice sampler*.
8: Simula $\rho_i^{(k)}$ de la distribución discreta (3.9).
9: **end for**

Es importante recalcar que, aunque en el artículo original se expuso un kérnel Gaussiano para el modelo latente y el algoritmo MCMC, es posible usar otras distribuciones. Si se considera un espacio medible $(\mathbb{X}, \mathcal{A})$ y denotamos por $K_\theta(y, x)$ a cualquier densidad bivariada sobre $\mathbb{X} \times \mathbb{X}$ con respecto a alguna medida de referencia, para la cual las marginales son idénticas, es decir,

$$K_\theta(y) = \int K_\theta(y, x)\,\mathrm{d}x \quad \text{y} \quad K_\theta(x) = \int K_\theta(y, x)\,\mathrm{d}y$$

De donde se sigue que

$$K_\theta(y) = \int K_\theta(y \,|\, x) K_\theta(x)\,\mathrm{d}x,$$

y por lo tanto, la construcción del modelo seguiría manteniendo la estacionariedad y la flexibilidad.

En (Antoniano-Villalobos y Walker, 2016) se menciona que la condición que las dos marginales sean iguales sólo es necesitada para garantizar la estacionariedad del proceso de mezclas de Markov, por lo que si se prescindiera de ella, aún sería posible construir un modelo autorregresivo más general en el cual no importe la estacionariedad.

Capítulo 4

Análisis de concentraciones de contaminantes

En los capítulos anteriores se expuso tanto la viabilidad teórica del modelo, como un algoritmo que puede ser usado para estimarlo. Ahora se hará la comprobación de su utilidad con datos reales, así como su comportamiento numérico.

En 2013, el Centro Internacional de Investigación del Cáncer, IARC, por sus siglas en inglés, lanzó un comunicado de prensa en el cual se anunciaba que la Organización Mundial de la Salud comenzaría a clasificar a la contaminación del aire como cancerígena para el ser humano.

Tan sólo en 2010 se estima que, globalmente, 223,000 muertes por cáncer de pulmón se debieron a la polución, y el número de personas expuestas a niveles riesgosos de contaminación tendrá a incrementarse conforme la densidad de población en megaciudades aumente debido a la migración de zonas rurales a urbanas.

La naturaleza dinámica de la contaminación hace que sea difícil diseñar políticas públicas que mitiguen el deterioro de salud y que, al mismo tiempo, tengan un impacto económico mínimo. La utilización de modelos no paramétricos permite acomodar al dinamismo de los datos históricos de contaminación sin tener que preocuparse con comportamientos multimodales o colas pesadas, abriendo la puerta a un sinnúmero de aplicaciones.

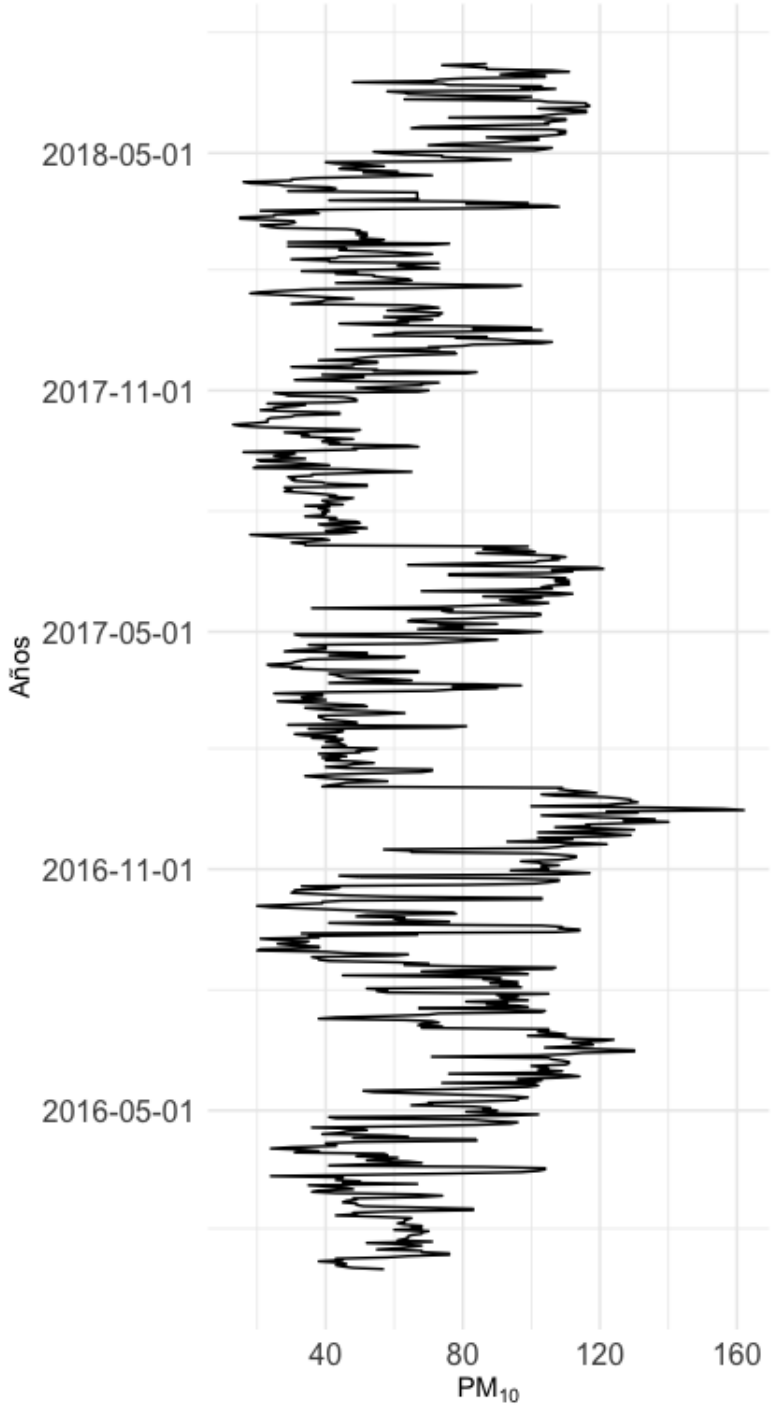

Figura 4.1: Históricos de PM_{10}

Se descargaron los máximos diarios medido en IMECAs de partículas suspendidas menores a 10 micrómetros en la Ciudad de México del 2016 a junio de 2018 (figura 4.1), esto para tener una cantidad suficiente de datos como para alcanzar a capturar la estructura de dependencia entre estos.

Un análisis exploratorio inicial de los datos reveló que había cinco observaciones faltantes, por lo que se decidió imputarlas usando el promedio de $\pm$ 4 días. Ninguna otra modificación se le hizo a los datos.

Además, un análisis visual indica la presencia de un componente periódico, aproximadamente cada tres meses, y haciendo un acercamiento al periodo de abril a julio de 2018 (figura 4.2), la media móvil de orden cuatro muestra otro comportamiento periódico a menor escala, aproximadamente cada semana y media, por lo que parecieran no ser estacionarios en el sentido de una distribución simétrica, como la normal o la t de Student.

Para ajustar los hiperparámetros del modelo se tomaron en cuenta tanto los estadísticos muestrales como los posibles problemas numéricos que pudieran aparecer: tener un rango tan amplio, tanto las probabilidades de los componentes individuales como la condicional completa de la precisión pueden llegar a hacer *underflow*, lo que entorpecería la estimación. Por esto se decidieron las siguientes distribuciones iniciales:

1. Para las medias, una distribución normal $\mathrm{N}(60, 900)$. Esta se centra en la media muestral y se le impone una desviación estándar de 30 para hacerla no informativa.

2. Para la precisión, es decir, la inversa de la varianza, una distribución $\mathrm{Gamma}(0.2, 1)$. Se centra la esperanza en la precisión muestral y se concentra la masa en el intervalo $[0, 0.5]$ para evitar problemas de *overfitting*.

3. Para el proceso Dirichlet que genera los pesos, una distribución $\mathrm{Beta}(1, 0.7)$.

4. A los coeficientes de correlación se les da una distribución *prior* uniforme sobre el conjunto $\{0.001, 0.002, 0.003, \ldots, 0.999\}$.

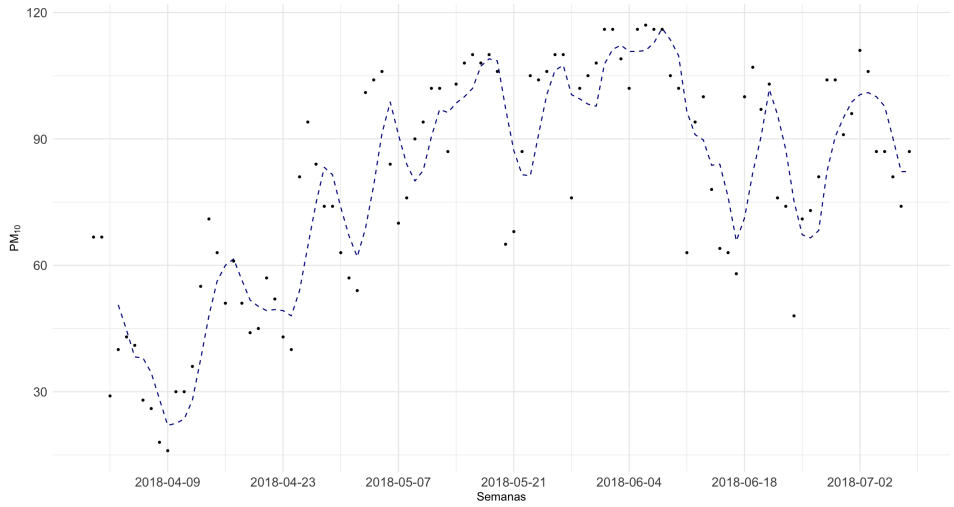

Figura 4.2: Históricos de PM_{10} del 31 de marzo al 8 de julio de 2018. Los puntos son los datos observados; la línea punteada azul, la media móvil de orden cuatro.

Además, se trunca el número de componentes a quince y la suma infinita a diez, la cual al ser una serie geométrica nos garantiza que el valor tenga aproximadamente diez dígitos correctos.

Se corren 100,000 iteraciones del muestreo de Gibbs con un *burn in* de 90,000. Posteriormente se hace un adelgazamiento de la cadena, tomando un valor cada diez iteraciones, tras lo que se procede a hacer un reordenamiento de las subcadenas ordenando las medias para solucionar el *label switching*, lo cual pudiera dificultar el cálculo de la distribución predictiva, (3.1).

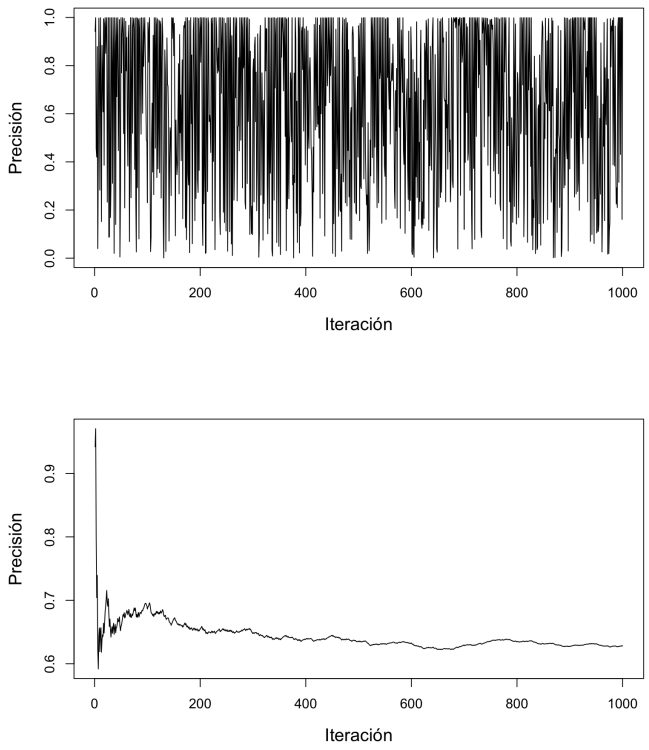

Figura 4.3: Arriba, la cadena generada para la precisión después del adelgazamiento; abajo, el promedio acumulado de dicha cadena.

En la figura 4.3 se muestran los *trace plots* de las cadenas generadas para el promedio de los coeficientes de correlación y de las medias después del *burn in* y del adelgazamiento, en donde una inspección visual pareciera indicar la presencia de estacionariedad. Debido al *label swtiching* es difícil monitorear la convergencia para los demás parámetros, por lo que se utiliza la media entre cada iteración, es decir, por ejemplo, para las medias, $\sum \mu_i/15$.

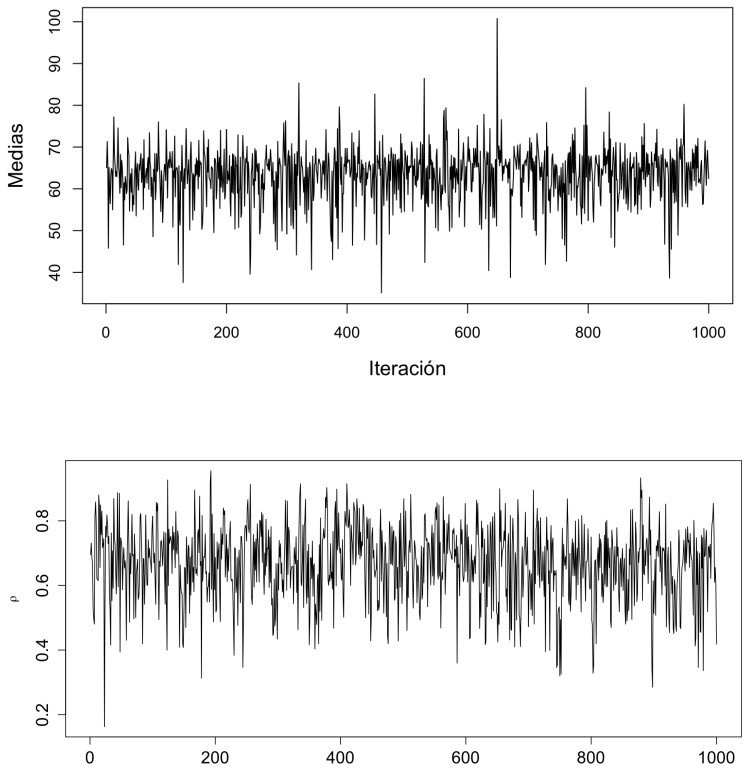

Figura 4.4: Arriba, los promedios acumulados para las medias; abajo, para los coeficientes de correlación.

Para los pesos resulta poco informativa la gráfica (figura 4.4), por lo que, para monitorear la estacionariedad de las cadena, después del *burn in* y el *thinning*, se calculó el promedio de los pesos en cada iteración y posteriormente se muestrearon aleatoriamente dos subconjuntos de quinientos puntos muestrales cada uno, de modo que la sucesión resultante se partiera en dos, y posteriormente se realizó la prueba de Kolmogorov-Smirnov al 95 % de confianza. El muestreo aleatorio y la prueba se repitieron mil veces, tras lo cual se llegó a que el 95.8 % de las veces, no se rechazaba la hipótesis de que las dos submuestras tuvieran la misma distribución, por lo que se puede

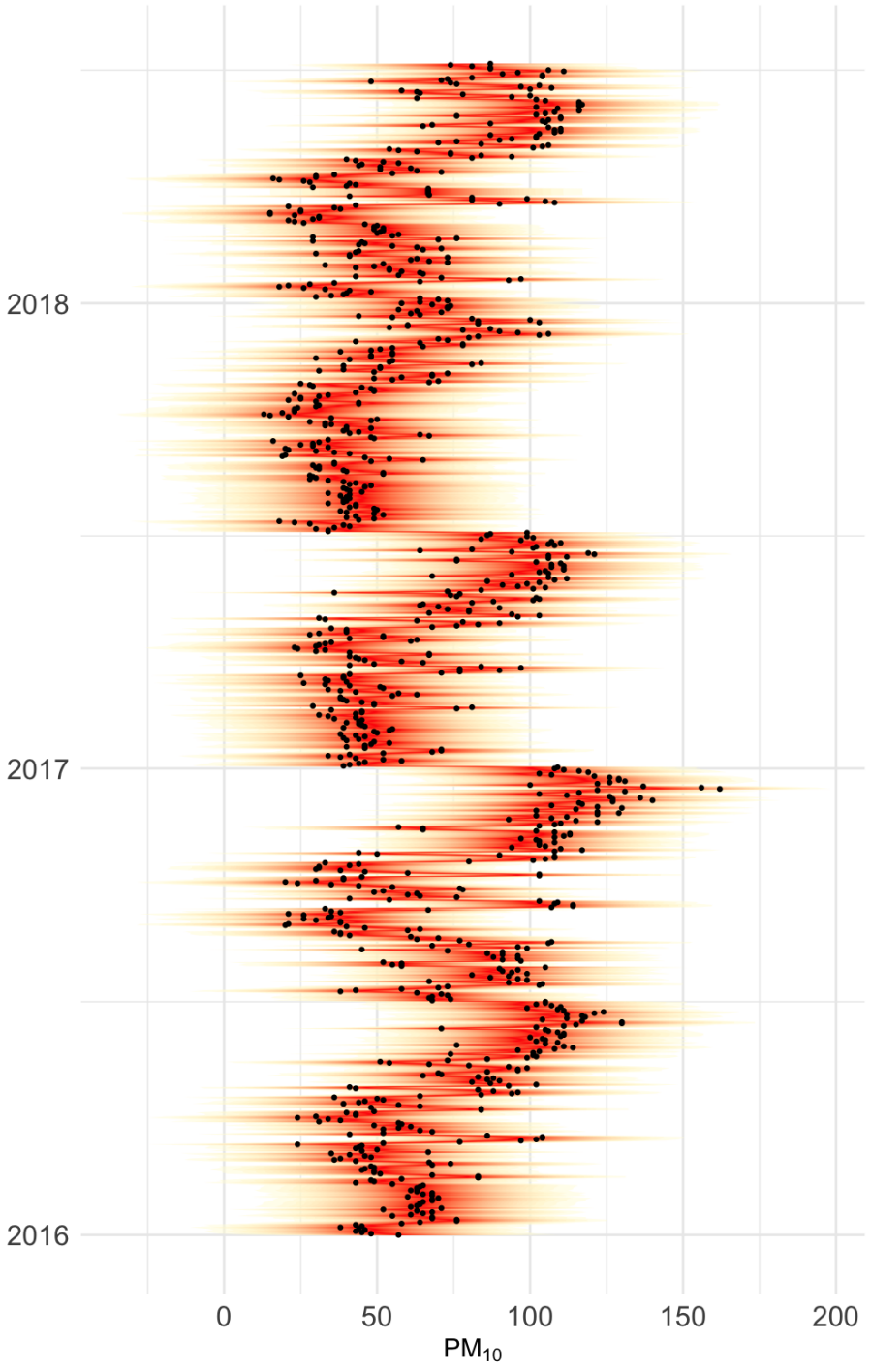

Figura 4.5: Densidad predictiva a un paso estimada para los máximos de PM_{10}. Colores más oscuros indican mayor probabilidad, $\mathbb{P}\left(x_t \mid x_{t-1}\right)$; los puntos, los datos observados.

concluir que parece haberse alcanzado la estacionariedad.

Una vez realizadas las pruebas de estacionariedad de las cadenas generadas por el algoritmo MCMC, tanto las visuales como usando el procedimiento descrito en el párrafo anterior, se procedió a estimar la transición a un paso para ilustrar al modelo resultante (figura 4.5). También se predijeron los niveles de contaminación a cincuenta días, (figura 4.6).

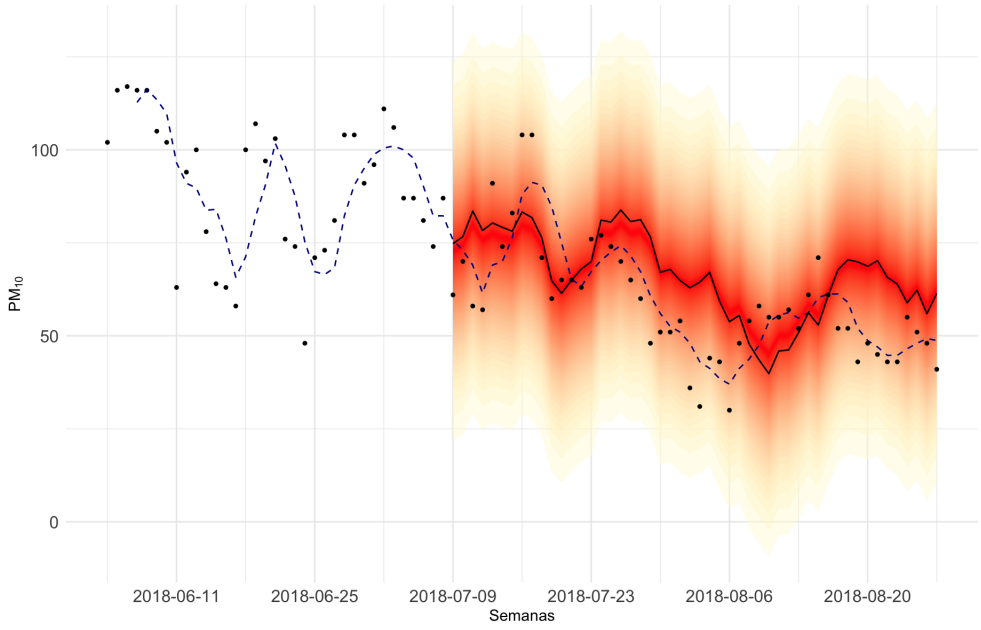

Figura 4.6: Predicción de 50 días de contaminación. Colores más oscuros indican mayor probabilidad; la línea negra, la predicción puntual; los puntos, los valores reales registrados; la línea azul punteada, la media móvil de orden 4.

La gráfica en la figura 4.5 indica que el modelo fue lo suficientemente flexible como para estimar de manera satisfactoria la probabilidad de transición de los datos. La predicción, comparada con la media móvil de orden 4, muestra que el modelo parece recuperar la estacionalidad semanal de manera adecuada, mostrando así su utilidad para este tipo de análisis. Se identifican

aún así dos debilidades: la estacionariedad y la estimación.

Si bien la estacionariedad es en muchos casos ventajosa por la simplicidad que aporta a la manipulación del proceso estocástico estimado, también conlleva el problema de no poder conocer probabilidades sin condicionar a datos pasados, pues para cualesquiera tiempos s, t, las marginales en ambos son idénticas, lo que, en este caso, hace que se ignore la estacionalidad natural que las concentraciones de contaminantes poseen.

También resulta problemática la sensibilidad del algoritmo de estimación, pues cualquier cambio, por ligero que sea, en alguno de los hiperparámetros causa problemas o de sobreestimación o de subestimación, lo que indica que sería mejor utilizar métodos con mecanismos que les permitan evitar que la cadena quede atorada en zonas de probabilidad alta, como el Monte Carlo hamiltoniano u otros algoritmos adaptativos. Sin embargo esto causaría a su vez un declive notorio en la eficiencia computacional, por lo que resultaría ventajoso un estudio más a profundidad del espacio parametral y la implementación de los métodos elegidos en un lenguaje de programación compilado, en vez de uno interpretado, como `R`. Además, la dificultad del monitoreo de la estacionariedad, así como la complejidad del algoritmo, hace que, si se quisiera aplicar a datos más irregulares, probablemente se requiera la utilización de muchas más iteraciones que las que se utilizaron para el presente trabajo, lo cual se traduciría en un largo tiempo de cómputo, lo cual limita el uso del modelo en situaciones donde se tengan límites de tiempo, como se da en las entidades financieras.

Capítulo 5

Discusión

La estadística bayesiana no paramétrica ofrece modelos altamente flexibles con construcciones potencialmente más simples que aquellas que se tendrían que hacer en casos paramétricos para alcanzar el mismo nivel de adaptabilidad. Sin embrago, su uso se ve limitado por la eficiencia de su estimación.

El aumento exponencial de poder de cómputo ha coadyuvado al desarrollo de la estadística computacional, en particular, de los métodos Monte Carlo, los cuales se aprovechan en el enfoque bayesiano no paramétrico para hacer inferencia.

El modelo aquí expuesto posee la virtud de ser una extensión del modelo clásico autorregresivo, lo que lo hace altamente interpretable, y el uso de una mezcla infinita para la distribución estacionaria del proceso le brinda flexibilidad, manteniendo las propiedades que la estabilidad le da otorga a los procesos estacionarios.

Mediante pruebas computacionales se comprobó que el algoritmo MCMC tansdimensional expuesto en (Antoniano-Villalobos y Walker, 2016) era numéricamente muy inestable debido a la presencia de probabilidades muy pequeñas, lo que finalmente ocasionaba una pérdida sustancial de eficiencia debido a la propagación de los errores numéricos.

Para solucionar este problema se aprovechó que la parametrización original transformaba mediante variables latentes las sumas infinitas presentes en series geométricas, cuya velocidad de convergencia hace factible el truncarlas a una precisión dada, estabilizando la estimación.

El algoritmo resultante resultó notoriamente más eficiente, aunque con la desventaja de ser altamente sensible a cambios en los hiperparámetros, lo que entorpecía la convergencia de las cadenas, por lo que se optó por el uso de métodos no aproximantes como el *slice sampler*.

Los resultados empíricos descritos en el capítulo anterior demuestran la validez y utilidad práctica del método resultante, dejándolo abierto a optimización, sujeto a la implementación de métodos computacionales más eficientes.

Uno de los posibles cambios que se le pudieran hacer al modelo para hacer más eficiente su estimación es el uso de medidas aleatorias distintas al proceso Dirichlet. En particular, el uso de pesos geométricos reduciría el fenómeno de *label switching* que en este trabajo se minimizó ordenando las medias.

El uso de kérnels distintos al gaussiano también pudiera ayudar en datos con colas muy pesadas, en donde se necesitan muchos componentes para capturar el comportamiento en los extremos de la distribución; en cambio, si se utilizaran distribuciones como la t, se necesitaría estimar menos componentes, lo que haría su estimación más eficiente computacionalmente.

Otro tipo de comportamientos más complejos pudieran ser mejor capturados utilizando estructuras de correlación más intrincadas, como sería el uso de órdenes superiores de dependencia markoviana, cuya construcción no cambiaría drásticamente a la expuesta en este trabajo.

Referencias

Antoniano-Villalobos, I., y Walker, S. (2016). A nonparametric model for stationary time series. *Journal of Time Series Analysis*, *37*(1), 126-142.

Bollerslev, T. (1986). *Generalized autoregressive conditional heteroskedasticity* (3a ed.). Springer.

Bradie, B. (2006). *A friendly introduction to numerical analysis*. Pearson Education.

Brockwell, P., y Davis, R. (2016). *Introduction to time series and forecasting*. Springer International Publishing.

Buntine, W. L., y Hutter, M. (2010). *A bayesian view of the poisson-dirichlet process* (Inf. Téc.).

Capasso, V., y Bakstein, D. (2012). *An introduction to continuous time stochastic processes* (2a ed.). Brikhäuser Basel.

Damien, P., Dellaportas, P., Polson, N., y Stephens, D. (Eds.). (2013). *Bayesian theory and applications* (1a ed.). Oxford University Press.

Engels, R. (1982). Autoregressive conditional heteroskedasticity with estimates of variance of united kingdom inflation. *Econometrica*, *50*(4), 987-1008.

Francq, C. (2010). *Garch models: structure, statistical inference and financial applications* (1a ed.). Wiley.

Fuentes-García, R., Mena, R. H., y Walker, S. G. (2010). A New Bayesian Nonparametric Mixture Model. *Communications in Statistics - Simulation and Computation*, *39*(04), 669-682.

Gelman, A., Carlin, J., Stern, H., Dunson, D., Vehtari, A., y Rubin, D. (2013). *Bayesian data analysis* (3a ed.). Taylor & Francis.

Goldstein, M. (2013). Observables and models: exchangeability and the inductive argument. En P. Damien, P. Dellaportas, N. Polson, y D. Stephens (Eds.), *Bayesian theory and applications* (p. 3-18). Oxford University Press.

Hjort, N., Holmes, C., Müller, P., y Walker, S. (Eds.). (2010). *Bayesian nonparametrics* (1a ed.). Cambridge University Press.

International Agency for Research in Cancer. (s.f.). *Iarc: Outdoor air pollution a leading environmental cause of cancer deaths.* Descargado de `https://www.iarc.fr/en/media-centre/iarcnews/pdf/pr221_E.pdf`

Ishwaran, H., y Zarepour, M. (2000, 06). Markov chain monte carlo in approximate dirichlet and beta two-parameter process hierarchical models. *Biometrika*, *87*. doi: 10.1093/biomet/87.2.371

Kallenberg, O. (2017). *Random measures, theory and applications* (1a ed.). Springer.

Lijoi, A., Mena, R. H., y Prünster, I. (2007). Bayesian nonparametric estimation of the probability of discovering new species. *Biometrika*, *94*(4), 769–786.

MacEachern, S. N., y Berliner, L. M. (1994). Subsampling the gibbs sampler. *The American Statistician*, *48*(3), 188-190.

Madsen, H. (2007). *Time series analysis* (1a ed.). CRC Press.

Maitra, A. (1977). Integral representations of invariant measures. *Transactions of the American Mathematical Society*, *229*, 209-225.

Martínez-Ovando, J. C., y Walker, S. G. (2011, septiembre). *Time-series modelling, stationarity and bayesian nonparametric methods* (Working Papers n.º 2011-08). Banco de México.

Mena, R., y Walker, S. (2005, 1 de 11). Stationary autoregressive models via

a bayesian nonparametric approach. *Journal of Time Series Analysis*, *26*(6), 789–805.

Müller, P., Quintana, F., Jara, A., y Hanson, T. (2015). *Bayesian nonparametric data analysis* (1a ed.). Springer.

Pitman, J. (1996). Some developments of the blackwell-macqueen urn scheme. En T. S. Ferguson, L. S. Shapley, y J. B. MacQueen (Eds.), *Statistics, probability and game theory* (Vol. Volume 30, pp. 245–267). Hayward, CA: Institute of Mathematical Statistics. doi: 10.1214/lnms/1215453576

Pitman, J., y Yor, M. (1997, 04). The two-parameter poisson-dirichlet distribution derived from a stable subordinator. *Ann. Probab.*, *25*(2), 855–900.

Robert, C., y Casella, G. (2004). *Monte carlo statistical methods* (2a ed.). Springer.

Schervish, M. (1996). *Theory of statistics*. Springer New York.

Shiryaev, A. (1999). *Essentials of stochastic finance* (1a ed.). World Scientific.

Stephens, M. (1997). *Bayesian methods for mixtures of normal distributions* (Tesis Doctoral no publicada). Magdalen College, Oxford.

Stephens, M. (2000). Dealing with label switching in mixture models. *Journal of the Royal Statistical Society. Series B (Statistical Methodology)*, *62*(4), 795-809.