Universidad Nacional Autónoma de México

Facultad de Ciencias

Estimación basada en Cadenas de Markov Montecarlo para el análisis de Series de Tiempo Enteras

T E S I S

Que para obtener el título de
Actuario

Presenta
Eduardo Selim Martínez Mayorga.

Directora de tesis:
Dra. Ruth Selene Fuentes García.

2014

A mis papás,
a mis hermanos,
a mis amigos.

Agradecimientos

Con el mayor de los agradecimientos a mis papás, en especial a mi mamá, Alejandra Mayorga, quien es complice en cada una de mis locuras, y a José Eduardo Martínez. La lista no podría continuar sin mis cinco motivaciones más grandes, mis hermanos: Alejandro, Luis Adrián, Alonso, Nataly y Renata, por que los cinco me han mostrado lo más valioso de la vida.

A Silvia Maldonado y a los fuertes Mayorga.

Muchas sucesiones de eventos me cruzaron no sólo con dos excelentes profesores sino con dos grandes amigos (aunque en la Facultad nos gusta decirles papás académicos) y me refiero a la Dra. Gabriela Campero Arena y al Act. Jaime Vázquez Alamilla.

Mi suerte continuó y me siguió mandando excelentes profesores. Si alguien tuvo paciencia infinita, esa fue la Dra. Ruth Fuentes. Aguantó miles de dudas, preguntas tontas, procastinación por largos periodos, malos argumentos y a cambio me dió referencias, estrategias, observaciones, correcciones, prórrogas... todas con el mejor de los ánimos. Muchas gracias Ruth.

Repito... mi suerte fue grande. La vida me cruzó con la Dra. Ana Meda Guardiola. Quizá ella me enseñó las lecciones más duras pero también me dió las sonrisas más honestas y los consejos más francos. Gracias Ana.

Siento mucha admiración, respeto y cariño por Margarita Chávez Cano. Verla todos los días dando lo mejor que tiene a alumnos y profesores es motivante. Gracias por la revisión detallada de mi trabajo y las valiosas recomendaciones.

Este trabajo no sería el mismo sin las observaciones y recomendaciones del Dr. Fernando Baltazar Larios y la Dra. Silvia Ruiz-Velasco Acosta. Gracias por el tiempo y apoyo.

A las mujeres de mi vida: Paulina Barbosa, Casandra Ceballos y Berenice Domínguez.

A aquellos que me enseñaron el valor del trabajo: Jaime Vázquez, Erick Miér y Gabriel Holschneider.

A mi familia por elección: Chío Martínez, Benjamín López, Karina Morales, María Fernanda Agoitia y Emmanuel Juárez.

Se supone que yo tenía que enseñarles el camino, pero resulta que fue al revés. Muchas gracias Mónica González, Carlos Chida, Orlando García, Jesús Rodríguez, Karen Valencia, Oscar Pineda, Humberto Iturbe, Víctor Cruz y Ana Almanza.

Por su amistad sincera y esos grandes momentos, gracias Julio Martínez y Darío Díaz.

Sin todo el apoyo y trabajo compartido no hubiese sido fácil. Gracias Arrigo Coen, Jessica Jaurez, Carlos Lozano, Fernando Fernández, Marco Tolentino, Eduardo Ponce, Adriana Ramos, Fernando Daniel Pérez, Jonathán Jaimes, Fernando Díaz, Benito Díaz, Angélica Godínez, Israel Carbajal y Gabriela Suárez.

El agradecimiento más grande va para mis maestros: Carmen Rocío Vite González, Gabriela Posadas Durán, Jaqueline Cañetas Ortega, Ricardo Melgarejo, Jorge Humberto del Castillo Spíndola, Gonzálo Zubieta Russi, Juan Pablo Márquez Arias, Oscar Aranda Martínez, Ana Rosa Camacho Lombilla, Mariano Zerón Medina-Laris, Ricardo Méndez Fragoso, Rodrigo Martínez Viveros, Olga Tchebotareva Nikolaevna, Alberto Saldaña de Fuentes, María Emilia Caballero Acosta, Luis Antonio Rincón Solís, José Luis Pérez Garmendia, Adrián González Casanova, Gerónimo Uribe Bravo, Héctor García de la Cadena, Javier Páez Cárdenas, Judith Campos Cordero, Alejandro Mina Valdes, José de Jesús Echeverría Eguiluz, Diana Avella Alaminos, Ramón Plaza Villegas, Pablo Padilla Longoria, Melissa Gutiérrez Vivanco y Sergio Macías Álvarez.

Si bien no tomé clases con ellos, mi camino fue más ligero y divertido gracias a Elisa Viso Gurovich, Roberto Cánovas Theriot, Francisco Solsona y José Galaviz.

En mi paso como Consejero, descubrí a mucha gente preocupada por los estudiantes, pensando todos los días cómo hacer menos complicado el extraño paso por la Facultad: Emma Lam, Elena de Oteyza, Ramón Peralta, Paty Goldstein, Mauricio Aguilar, Fernando García, Dante Morán, Doña Josefina, Magnolia Ávalos, Adrián Girard, Alicia Cervantes, Patricia Magaña, Elena Abrín, Ivonne Gaspar y Rosa María Flores.

A mis amigos consejeros: Valeria Towns Alonso, Santiago Morato Sánchez de Tagle, Jesús Erasmo Batta, Emilio Roth Monzón Sergio Nicasio, Luis David Álvarez Corrales y Roberto Alatorre Bremont.

No pueden faltar mis yucas favoritos: David Contreras, Fernando Gual, Luis Vázquez y los hermanos Legorreta.

Finalmente, a mis todos alumitos (que algunos ya no son ni "itos" ni alumnos).

Eduardo Selim Martínez Mayorga

Índice general

Introducción

En este trabajo se analizarán algunas de las ideas más conocidas en series de tiempo que toman valores en un espacio a lo más numerable. Debido a esta característica discreta, algunas de las técnicas estándar en el análisis de series de tiempo (aquellas que pueden tomar cualquier valor real) no son adecuadas y las aproximaciones asintóticas se tienen que considerar de forma muy cuidadosa.

Dada la complejidad de la naturaleza de los datos en series de tiempo con valores discretos, este trabajo tiene como objetivo estudiar algunas de las principales técnicas para el análisis de series de tiempo de este tipo. Y más aún, se aprovecharán las ventajas de las técnicas de simulación actuales para resolver algunos problemas de inferencia y predicción asociados a este tipo de datos discretos.

En el capítulo 1 se hace un recorrido por las principales técnicas de simulación desde los principios básicos de transformada integral, hasta los algoritmos de cadenas de Markov Monte-Carlo con saltos reversibles, pasando por los algoritmos Metropolis-Hastings y Gibbs *sampler*. Se estudian las principales condiciones para la convergencia de estos algoritmos y algunas propuestas para utilizar algoritmos que tengan partes Gibbs-sampler pero también Metropolis-Hastings.

En el capítulo 2 simplemente se da un recordatorio de los conceptos básicos de series de tiempo y se analiza una importante clase de modelos: la estructura de correlación ARMA. Además de algunas de las propuestas más conocidas en la literatura para estimación, predicción y determinación del orden. Este capítulo no es un *cliché* de los trabajos en series de tiempo, en realidad este capítulo y el capítulo 4 se pueden ver como análogos continuo y discreto, respectivamente.

El primer acercamiento a modelos para series de tiempo con valores que toman valores en un conjunto a lo más numerable se hace en el capítulo 3. Se empieza con modelos muy sencillos de cadenas de Markov con espacio de estados numerables (principalmente los modelos de Raftery) y se finaliza con los modelos $DARMA$ (así como DAR y DMA) de Lewis & Jacobs.

El capítulo más amplio es el capítulo 4, en éste se discute extensamente acerca de las propuestas para modelar series de tiempo con valores discretos intentando extender (o redefinir) la clase de modelos $ARMA$. Sin embargo, la principal clase que se analiza es la de los modelos $INAR$. Se dan algunos resultados para la existencia de soluciones estacionarias para replicar las estructuras de correlación de los procesos AR. El factor clave para hacer estas extensiones es lo que se conoce como operador de adelgazamiento binomial. Se hacen dos propuestas, definiendo dos operaciones de adelgazamiento diferente; se generalizan a operadores de van Harn & Steutel para finalmente llevarlo a semigrupos de operadores de adelgazamiento (donde los conceptos de autodescomponibilidad y estabilidad toman un papel de mucha importancia).

Particularmente se profundiza los casos $INAR$ e $INMA$ cuando se tienen procesos de innovación con distribución Poisson, además se hacen algunos diagnósticos sencillos, junto con un ejercicio de simulación para calcular las potencias de algunas pruebas para determinar independencia serial.

Finalmente, en el capítulo 5 se fusionan las técnicas de simulación del capítulo 1 y el marco teórico de los capítulos 3 y 4 al analizar, definir principales propiedades, inferencia básica y predicción en los modelos $INARMA$. Además se analizará brevemente algunos básicos para determinar el orden de la estructura de correlación con un algoritmo basado en Cadenas de Markov con saltos reversibles.

Capítulo 1

Técnicas de simulación

1.1. Generación de números aleatorios

1.1.1. Métodos básicos

Generalmente se habla de manera indistinta de generación de números aleatorios ó de generación de variables aleatorias y son la base de los métodos de simulación estocástica. Dichos métodos están basados en la generación de variables aleatorias que se rigen mediante una ley de probabilidad F (una función de distribución), que no necesariamente se conoce explícitamente.

Hablando formalmente, en términos de teoría de análisis numérico, la generación de variables aleatorias se hace a partir de generadores de números pseudo-aleatorios; sin embargo, no habrá mayor discusión sobre este debate y se supondrá que se cuenta con un generador de números aleatorios aceptable.

1.1.1.1. Generación a partir de la transformada integral

Inicialmente, se discutirá acerca de algunos métodos de generación de variables aleatorias asociadas con la distribución $Unif[0,1]$ ya que éstas dan una "noción probabilística" básica de aleatoriedad.

Considérense un espacio de probabilidad $(\Omega, \mathscr{F}, \mathbb{P})$; una variable aleatoria U tal que $U \sim Unif[0,1]$ y el espacio de probabilidad $([0,1], \mathscr{B}([0,1]), F_U)$, donde F_U es la función de ditribución de U y $\mathscr{B}([0,1])$ son los Borelianos del intervalo [0,1].

Definición 1.1.1. *(Función inversa generalizada)*

Sea $g : A \subseteq \mathbb{R} \to \mathbb{R}$ una función creciente, se define la función inversa generalizada de g, denotada por g^-, como $g^- : Im(g) \to A$ tal que

$$g^-(y) = \text{ínf}\{z : g(z) \geq y\}$$

Lema 1.1.1. *(Transformada integral de probabilidad)*

Si $U \sim Unif[0,1]$, entonces la variable aleatoria $F^-(U)$ tiene función de distribución F

Demostración:

Sea $u \in [0,1]$ y $x \in F^-[0,1]$. Por definición de inversa generalizada

$$F(F^-(u)) = F(\text{ínf}\{z : F(z) \geq u\}) \geq u,$$

además como F es no decreciente (pues es función de distribución)

$$F^-(F(x)) = \text{ínf}\{z : F(z) \geq F(x)\} \leq x$$

Es decir, la inversa generalizada satisface

$$F(F^-(u)) \geq u \quad \text{y} \quad F^-(F(x)) \leq x$$

Por tanto,

$$\{(u,x) : F^-(u) \leq x\} = \{(u,x) : F(x) \geq u\}$$

Entonces, si $X = F^-(U)$,

$$F_X(x) = \mathbb{P}(X \leq x) = \mathbb{P}(F^-(U) \leq x) = \mathbb{P}(U \leq (F(x))) = F(x)$$

pues si $U \sim Unif[0,1]$, entonces

$$F_U(u) = \begin{cases} 0 & \text{si } u < 0; \\ u & \text{si } 0 \leq u \leq 1; \\ 1 & \text{si } u > 1; \end{cases}$$

$$\therefore F_X(x) = F(x)$$

$$\therefore F^-(U) \text{ tiene función de distribución } F$$

□

En virtud del lema anterior, si se quiere generar un número aleatorio x que provenga de una variable aleatoria X que tenga función de distribución F, es suficiente "seleccionar" (generar) $u \in sop(f_U)$ y hacer la transformación $x = F^-(u)$.

Ejemplo 1.1.1. *(Generación de variables exponenciales)*

Si $X \sim exp(\lambda)$, entonces $F_X(x) = 1 - e^{-\lambda x}$. Entonces, resolviendo para x la ecuación $u = 1 - e^{-\lambda x}$ se tiene que

$$u = 1 - e^{-\lambda x} \Leftrightarrow e^{-\lambda x} = 1 - u \Leftrightarrow -\lambda x = Ln(1-u) \Leftrightarrow x = \frac{Ln(1-u)}{-\lambda}$$

y se puede sustituir u por $1-u$ en la expresión anterior pues si $U \sim Unif(0,1)$, entonces $1 - U \sim Unif(0,1)$. Por tanto $x = \frac{Ln(u)}{-\lambda}$ es un número exponencial. Y de hecho, se puede probar que Si $U \sim Unif(0,1)$, entonces $X = \frac{Ln(U)}{-\lambda} \sim exp(\lambda)$, a saber

$$\begin{aligned} m_X(t) &= \mathbb{E}(e^{tX}) = \mathbb{E}\left[exp\left(t\frac{Ln(U)}{-\lambda}\right)\right] = \mathbb{E}\left[exp(Ln(U^{-t/\lambda}))\right] \\ &= \mathbb{E}\left[U^{-t/\lambda}\right] = \int_0^1 u^{-t/\lambda} du = \frac{u^{\frac{-t}{\lambda}+1}}{\frac{-t}{\lambda}+1} \,|_0^1 = \frac{1}{\frac{-t}{\lambda}+1} \\ &= \frac{1}{\frac{-t+\lambda}{\lambda}} = \frac{\lambda}{\lambda - t} \end{aligned}$$

$$\therefore m_X(t) = \frac{\lambda}{\lambda - t}$$

$\therefore X$ tiene la función generadora de momentos de una exponencial con parámetro λ

$$\therefore X \sim exp(\lambda)$$

▽

1.1.2. Generación de cantidades aleatorias discretas

La idea de esta sección es describir algunos métodos de generación de cantidades con distribución de probabilidad discreta. Dicha generación está basada en la generación de observaciones de una una variable aleatoria continua U tal que $U \sim Unif[0,1]$.

Supóngase que X es una variable aleatoria discreta que toma valores en $\{x_1, \ldots, x_k\}$. También supóngase que $\mathbb{P}(X = x_j) = p_j$ donde $\sum_{j=1}^{k} p_j = 1$. El intervalo [0,1] se divide en k intervalos $I_1, \ldots, I_k$, donde $I_j := (F_{j-1}, F_j]$ con $F_0 = 0$ y $F_j = p_1 + \ldots + p_j$, $j = 1, \ldots, k$. A cada valor x_j le corresponde uno y sólo uno de los I_j; entonces para generar observaciones de X, se genera un número uniforme u en el intervalo [0,1] y se observa en cuál de los intervalos I_j pertenece u. El valor generado de x será x_j. El método genera valores de la distribución de X pues

$$\mathbb{P}(X = x_j) = \mathbb{P}(U \in I_j) = \mathbb{P}(F_{j-1} < U \leq F_j) = F_j - F_{j-1} = p_j$$

1.2. Método de aceptación rechazo

Hay muchas distribuciones para las cuales es muy difícil, incluso imposible, simular directamente. Además, en algunos casos no es posible representar a la distribución en términos en que se pueda utilizar, por ejemplo en los casos de una transformación o de una mezcla. Ahora, se estudiará una clase de métodos en los que sólo se requiere conocer la forma funcional de la densidad de f (excepto quizá por una constante multiplicativa) sin un análisis analítico de dicha densidad tan profundo. La clave de este método es utilizar una densidad g de la cual sea sencillo simular, ésta se conoce como *densidad instrumental*.

Por tanto, hay una gran cantidad de funciones f (conocidas como *densidades objetivo*) de la cuales se podrá simular de esta forma.

El siguiente algoritmo se conoce como el *Algoritmo Aceptación-Rechazo* y lo primero que se debe cumplir es que si f es la densidad de interés, se deben encontrar g y M tales que para todo $x \in sop(f)$

$$f(x) \leq Mg(x)$$

El algoritmo se basa en el siguiente resultado:

Lema 1.2.1. *(Algoritmo Aceptación-Rechazo)*

El algoritmo

1. Generar $X \sim g$, $U \sim Unif(0,1)$
2. Aceptar $Y = X$ si $U \leq \frac{f(X)}{Mg(X)}$
3. Regresar a 1. en otro caso.

genera la variable Y que se distribuye de acuerdo a f.

Demostración:

La función de distribución de Y está dada por

$$\begin{aligned}
\mathbb{P}(Y \leq y) &= \mathbb{P}\left(X \leq y | U \leq \frac{f(X)}{Mg(X)}\right) = \frac{\mathbb{P}\left(X \leq y, U \leq \frac{f(X)}{Mg(X)}\right)}{\mathbb{P}\left(U \leq \frac{f(X)}{Mg(X)}\right)} \\
&= \frac{\int_{-\infty}^{y} \int_{0}^{f(x)/Mg(x)} f_U(u) g(x) du dx}{\int_{-\infty}^{\infty} \int_{0}^{f(x)/Mg(x)} f_U(u) g(x) du dx} \\
&= \frac{\frac{1}{M}\int_{-\infty}^{y} f(x) dx}{\frac{1}{M}\int_{-\infty}^{\infty} f(x) dx} = \int_{-\infty}^{y} f(x) dx
\end{aligned}$$

por tanto Y se distribuye de acuerdo a f. □

Este lema tiene dos consecuencias. Primero, proporciona un método genérico para simular de cualquier densidad f que sea conocida (excepto quizá por una constante multiplicativa), i.e. no necesariamente se conoce la constante de normalización de f ya que el método sólo necesita conocer el cociente $\frac{f}{M}$ (que no depende de la constante de normalización). Esta propiedad es particularmente importante en inferencia Bayesiana en la que generalmente se desea conocer la densidad posterior, que se obtiene a partir del Teorema de Bayes mediante

$$\pi(\theta|x) \propto \pi(\theta) f(x\theta)$$

por tanto, la densidad se puede especificar (excepto quizá por una constante de normalización), para usar el algoritmo de Aceptación-Rechazo, sin tener que encontrar la constante de normalización. Nótese que el acotamiento $f(x) \leq Mg(x)$ no es esctricto en el sentido de que el algoritmo sigue siendo válido si se reemplaza M por cualquier múltiplo de M.

Una segunda consecuencia del lema es que la probabilidad de aceptar en el algoritmo es exactamente $\frac{1}{M}$ cuando se evalúa en densidades propiamente normalizadas y el valor esperado de intentos hasta que se acepta la variable tiene media M, pues

$$\begin{aligned}\vartheta := \mathbb{P}(\text{Aceptar } Y = X) &= \mathbb{P}\left(U \leq \frac{f(X)}{Mg(X)}\right) \\ &= \int_{-\infty}^{\infty}\int_{0}^{f(x)/Mg(x)} f_U(u)g(x)dudx \\ &= \frac{1}{M}\int_{-\infty}^{\infty} f(x)dx = \frac{1}{M}\end{aligned}$$

y

$$\begin{aligned}\mathbb{E}(\text{Número de intentos hasta que se acepta } Y = X) &= \sum_{k=1}^{\infty} k(1-\vartheta)^{k-1}\vartheta \\ &= \vartheta\frac{1}{[1-(1-\vartheta)]^2} = \frac{\vartheta}{\vartheta^2} = \frac{1}{\vartheta} \\ &= \frac{1}{1/M} = M\end{aligned}$$

Por tanto, se pueden hacer comparaciones entre diferentes algoritmos Aceptación-Rechazo basados en diferentes densidades instrumentales $g_1, g_2, \ldots$ donde el criterio de comparación serán las respectivas $M_1, M_2 \ldots$. En particular, un primer método para optimizar la elección de g en $\{g_1, g_2, \ldots\}$ es considerar la que tenga la M_i más pequeña posible. Sin embargo esta forma de optimización es muy rudimentaria y tiene algunas limitaciones.

En el caso de que f y g sean funciones de densidad de probabilidad, la constante M necesariamente es mayor que 1; y por tanto el tamaño de M (y su eficiencia) se convierten en función del grado en que g puede imitar a f, especialmente en las colas de la distribución. Nótese que para que f/g esté acotado es necesario que g tenga colas más pesadas que f.

1.3. Integración Monte-Carlo clásica

Antes de empezar a analizar los métodos de simulación basados en cadenas de Makov que se utilizarán durante este trabajo, primero se desarrollarán algunas propiedades con más detalle. Esto se facilita si se estudia el problema genérico de calcular la integral

$$\mathbb{E}_f(h(X)) = \int_{\mathcal{X}} h(x)f(x)dx \tag{1.1}$$

Una de las primeras propuestas naturales es utilizar una muestra $(X_1, \ldots, X_m)$ generada de la densidad f y aproximar (1.1) mediante el promedio empírico, conocido como *estimador de Monte-Carlo*,

$$\bar{h}_m = \frac{1}{m} \sum_{j=1}^{m} h(x_j)$$

ya que $\bar{h}_m$ converge casi seguramente a $\mathbb{E}_f(h(X))$ (por la Ley Fuerte de Grandes Números).

Además, cuando h^2 tiene media finita (bajo f), la velocidad de convergencia de $\bar{h}_m$ se puede obtener ya que

$$Var(\bar{h}_m) = \frac{1}{m} \int_{\mathcal{X}} [h(x) - \mathbb{E}_f(h(X))]^2 f(x) dx$$

también se puede estimar a partir de la muestra $(X_1, \ldots, X_m)$ mediante

$$v_m = \frac{1}{m^2} \sum_{j=1}^{m} (h(x_j) - \bar{h}_m)^2$$

y para m suficientemente grande

$$\frac{\bar{h}_m - \mathbb{E}_f(h(X))}{\sqrt{v_m}}$$

tiene distribución asintótica $N(0,1)$ y permitiendo así la construcción de algunas pruebas de convergencia e intervalos de confianza de $\mathbb{E}_f(h(X))$.

1.4. Muestreo de importancia

Este método recibe su nombre porque se basa en lo que se conoce como "funciones de importancia", aunque algunos autores sugieren que un nombre más apropiado sería el de "muestreo ponderado".

El cálculo de (1.1) basado en simulaciones de f no necesariamente es óptimo (de hecho sería deseable un criterio de elección de distribuciones f). De hecho, la integral (1.1) se puede representar en un número infinito de formas mediante las ternas $(\mathcal{X}, h, f)$. Por tanto, la búsqueda de un estimador óptimo debería considerar "evitar" todas estas posibles representaciones, así como los problemas computacionales (principalmente en términos de tiempo).

Definición 1.4.1. *(Muestreo de importancia)*

El método de muestreo de importancia es un cálculo de

$$\mathbb{E}_f(h(X)) = \int_{\mathcal{X}} h(x)f(x)dx$$

basados en la generación de una muestra $(X_1, \ldots, X_m)$ de una distribución g dada y aproximando

$$\mathbb{E}_f(h(X)) \approx \frac{1}{m}\sum_{j=1}^{m} \frac{f(X_j)}{g(X_j)}h(X_j) \tag{1.2}$$

Observación 1.4.1.

El método de muestreo de importancia se basa en la representación alternativa de (1.1) como

$$\mathbb{E}_f(h(X)) = \int_{\mathcal{X}} h(x)\frac{f(x)}{g(x)}g(x)dx$$

y el hecho de que el estimador (1.2) converge a (1.1) por la misma razón que el estimador de Monte-Carlo, $\bar{h}_m$, converge para cualquier distribución tal que $sop(f) \subseteq sop(g)$. Este método es de considerable interés ya que es poco restrictivo en la elección de la distribución instrumental g (que se puede seleccionar de aquellas distribuciones que sean sencillas de simular o ya se cuente con algoritmo suficientemente aceptable para hacer simulaciones de ésta). Además, la misma muestra generada de g se puede utilizar para diferentes funciones f para h fija; lo cual es muy atractivo si se tienen fines de robustez o de análisis Bayesiano de sensitividad. ▽

Aunque se puede utilizar casi cualquier densidad g para que el estimador (1.2) converja, hay algunas elecciones que son mejores que otras e inmediatamente surge la necesidad de comparar diferentes distribuciones g para el cálculo de (1.1). Nótese que aunque (1.2) converge casi seguramente a (1.1) tendrá varianza finita sólo cuando

$$\begin{aligned}\mathbb{E}_g\left(h^2(X)\frac{f^2(X)}{g^2(X)}\right) &= \mathbb{E}_f\left(h^2(X)\frac{f(X)}{g(X)}\right)\\ &= \int_{\mathcal{X}} h^2(x)\frac{f^2(x)}{g(x)}dx < \infty\end{aligned}$$

Por tanto,

- Las distribuciones instrumentales con colas más ligeras que las de f no son apropiadas para el muestreo de importancia. De hecho, en estos casos las varianzas de los correspondientes estimadores (1.2) serán infintas para muchas funciones h.
- Si el cocientes f/g no está acotado, los pesos $f(x_j)/g(x_j)$ podrían variar abruptamente, dándole mucha importancia a pocos valores de x_j.

Las distribuciones con colas más pesadas que f garantizan que el cociente f/g no provoque la divergencia de $\mathbb{E}_f(h^2 f/g)$. Generalmente se divide en dos tipos de condiciones suficientes:

(i) Para todo $x \in \mathcal{X}$, $f(x)/g(x) < M$ y $Var_f(h) < \infty$.

(ii) $\mathcal{X}$ es compacto y para todo $x \in \mathcal{X}$, $f(x) < F$ y $g(x) > \varepsilon$.

Estas condiciones son muy restrictivas. En particular, $f/g < M$ implica que también se puede utilizar el método de Aceptación-Rechazo.

Teorema 1.4.1.

La elección de g que minimiza la varianza del estimador (1.2) es

$$g^*(x) = \frac{|h(x)|f(x)}{\int_{\mathcal{X}} |h(z)|f(z)dz}$$

Demostración:

Nótese que

$$Var_g\left(\frac{h(X)f(X)}{g(X)}\right) = \mathbb{E}_g\left(\frac{h^2(X)f^2(X)}{g^2(X)}\right) - \left[\mathbb{E}_g\left(\frac{h(X)f(X)}{g(X)}\right)\right]^2$$

como el segundo término del lado derecho no depende de g para minimzar la varianza sólo se necesita minimizar el primer término.

Por la desigualdad de Jensen se tiene que

$$\mathbb{E}_g\left(\frac{h^2(x)f^2(X)}{g^2(X)}\right) \geq \left[\mathbb{E}_g\left(\frac{|h(X)|f(X)}{g(X)}\right)\right]^2 = \left[\int_{\mathcal{X}} |h(x)|f(x)\right]^2$$

Obteniéndose una cota inferior que no depende de la elección de g. El denominador de g^* simplemente se obtiene al considerar la constante de normalización. □

Esta optimalidad es más una formalidad que un uso práctico ya que la elección óptima requiere que se conozca $\int_{\mathcal{X}} h(x)f(x)$, que es precisamente la integral de interés.

Una alternativa práctica es usar

$$\frac{\sum_{j=1}^{m} h(x_j)f(x_j)/g(x_j)}{\sum_{j=1}^{m} f(x_j)/g(x_j)} = \frac{\sum_{j=1}^{m} h(x_j)[h(x_j)]^{-1}}{\sum_{j=1}^{m} |h(x_j)|^{-1}} \tag{1.3}$$

donde $x_j \sim g \propto |h|f$; nótese que por la Ley Fuerte de Grandes Números (1.3) también converge a $\int h(x)f(x)dx$.

Desde un punto de vista práctico, el Teorema anterior sugiere que se deben buscar distribuciones g tales que $|h|f/g$ es casi constante con varianza finita. Es importante notar que aunque la restricción de varianza constante no es necesaria para la convergencia de (1.2) y (1.3), con el muestreo de importancia se obtienen malos resultados cuando

$$\int \frac{f^2(x)}{g(x)}dx = \infty \tag{1.4}$$

si en términos del comportamiento del estimador (saltos de alta amplitud, inestabilidad de la trayectoria del promedio, convergencia lenta, etc.) ó de la comparación con métodos de Monte-Carlo directos. Entonces, no se recomiendan las distribuciones g tales que (1.4) ocurra.

1.5. Métodos de Monte-Carlo basados en Cadenas de Markov

Ya se mencionó en las secciones anteriores que no es necesario usar una muestra de la distribución f para aproximar la integral

$$\int h(x)f(x)dx$$

ya que se pueden utilizar técnicas de muestreo de importancia. En esta sección se analizará una estrategia diferente y se mostrará que es posible obtener una muestra $X_1, \ldots, X_n$ aproximadamente distribuida como f sin simular directamente de f. El principio básico subyacente con estas estrategias es utilizar una cadena de Markov ergódica cuya distribución estacionaria sea f.

Para un valor inicial arbitrario $x^{(0)}$, se generará una cadena, $\{X^{(t)}\}$ usando un kernel de transición con distribución estacionaria f, que asegure la convergencia en distribución de $\{X^{(t)}\}$ a una variable aleatoria cuya densidad sea f. Como la cadena será ergódica, en principio, el valor inicial $x^{(0)}$ no será importante. Por tanto, para un T_0 suficientemente grande, se puede considerar a $X^{(T_0)}$ como una variable aleatoria que tiene densidad f y no

sólo eso sino que se generó una a muestra dependiente $X^{(T_0)}, X^{(T_0+1)}$ que es generada de f.

Definición 1.5.1.

Un método de Cadena de Markov Monte-Carlo (MCMC) para la simulación de una distribución f es cualquier método que genere una cadena de Markov ergódica $\{X^{(t)}\}$ cuya distribución estacionaria es f.

En años recientes, se han incrementado las aplicaciones estadísticas que hacen uso de métodos de Cadenas de Markov Monte-Carlo (MCMC); dichos métodos se utilizan para simular distribuciones multivariadas no-estándar y complejas; de entre éstos, el algoritmo Metropolis-Hastings (hablando de él genéricamente) y el de muestreo de Gibbs es uno de los más conocidos, y su impacto en la estadística Bayesiana ha sido detallado en muchos artículos de investigación.

Con el objetivo de que este texto sea lo más auto-contenido posible, se dará una introducción acerca del algoritmo, desde sus principios teóricos de cadenas de Markov, algunos resultados relacionados con la implementación y el ajuste y finalmente algunos ejemplos.

Sólo con fines didácticos, la discusión se hará en términos de simulaciones para densidades absolutamente continuas; sin embargo, las ideas centrales también son válidas para los casos discretos y mixtos.

1.6. El algoritmo Metropolis-Hastings

La idea de simulación ahora se modificará, hasta ahora típicamente se generaron observaciones independientes de la misma distribución, dichas observaciones se simularon directamente de la densidad de interés f ó mediante alguna técnica de aceptación-rechazo. Sin embargo, ahora se discutirá acerca del algoritmo Metropolis-Hastings, que genera observaciones correlacionadas a partir de una cadena de Markov. Una de las razones por la cual se cambia de manera tan "radical" la perspectiva de simulación, y aún cuando esta dependencia existe, es que las cadenas de Markov tienen ciertas propiedades de convergencia que facilitan aquellos casos en los que, por ejemplo, el muestreo de importancia, no se puede aplicar fácilmente. Además de estas ventajas de convergencia, los hipótesis sobre la densidad objetivo f son pocas, por tanto permite hacer simulaciones aún con muy poca información acerca de f. Y finalmente, esta perspectiva de cadenas de Markov permite trabajar de forma fácil problemas de dimensiones altas al descomponer este tipo de problemas en pequeños problemas que son más fácil de resolver (en términos de simuación).

1.6.1. Aspectos técnicos de cadenas de Markov

En el Apéndice A, se hace un estudio extensivo de cadenas de Markov a tiempo discreto con espacio de estados continuo que justifica muchas de las afirmaciones que se hacen con respecto a los métodos de cadenas de Markov Monte-Carlo.

Ahora simplemente se dará la definición de cadena de Markov.

Definición 1.6.1. *(Cadena de Markov)*

Una cadena de Markov es una sucesión de variables aleatorias $\{X^{(t)}\}_{t=0}^{\infty}$ en un espacio de probabilidad $(\Omega, \mathscr{F}, \mathbb{P})$ tal que la distribución de $X^{(t)}$ dadas las variables pasadas depende sólo de $X^{(t-1)}$. Esta distribución condicional de probabilidad se conoce como kernel de trancisión o Kernel de Markov y se denota por K; es decir

$$X^{(t+1)}|X^{(0)} = x_0, X^{(1)} = x_1, \ldots, X^{(t)} = x_t \sim K(x_t, x_{t+1})$$

Ejemplo 1.6.1. *(Caminata aleatoria normal)*

El proceso conocido como caminata aleatoria normal (que es una cadena de Markov) se define como

$$X_{t+1} = X_t + \epsilon_t$$

donde $\epsilon_t \sim N(0,1)$ y además ϵ_t y X_t son independientes. El kernel de transición $K(x_t, x_{t+1})$ de esta cadena está dado por una densidad $N(x_t, 1)$. ▽

La mayoría de las cadenas de Markov que se utilizarán en metodologías MCMC satisfacen una propiedad muy fuerte de estabilidad: existe una distribución de probabilidad estacionaria para la cadena. Es decir, existe una distribución de probabilidad f tal que si $X^{(t)} \sim f$ entonces $X^{(t+1)} \sim f$ (Véase Apéndice A). De manera más formal, en las cadenas que se considerarán se debe satisfacer la ecuación

$$\int_{\mathcal{X}} K(x,y) f(x) dx = f(y)$$

La existencia de una distribución estacionaria presupone establecer una condición sobre K, conocida como irreductibilidad en la teoría de las cadenas de Markov (Véase Apéndice A). Dicha propiedad de irreductibilidad permite que el kernel "se mueva" libremente en todo el espacio de estados de la cadena, es decir, sin importar el valor inicial de $X^{(0)}$, la sucesión $\{X^{(t)}\}$ tiene probabilidad positiva de "llegar" a cualquier región del espacio de estados (una condicion suficiente es que $K(x, \cdot) > 0$. La existencia de una distribución estacionaria tiene

algunas otras consecuencias en el comportamiento de la cadena $\{X^{(t)}\}$ muy útiles en las metodologías MCMC; por ejemplo, una de estas consecuencias es la recurrencia, es decir, se puede regresar a casi cualquier subconjunto del espacio de estados un número infinito de veces.

En el caso de cadenas recurrentes, la distribución estacionaria también es la distribución límite en el sentido de que la distribución de $X^{(t)}$ es f para cualquier valor inicial $X^{(0)} = x_0$. Esta propiedad también se conoce como ergodicidad (Véase apéndice A) y por supuesto tiene consecuencias importantes desde el punto de vista de simulación como, por ejemplo, si un kernel K genera una cadena de Markov ergódica con distribución estacionaria f, generar una cadena de este kernel K eventualmente producirá simulaciones de f. En particular, para funciones integrables h

$$\frac{1}{T}\sum_{t=1}^{T} h(X^{(t)}) \to \mathbb{E}_f(h(X))$$

que significa que la Ley de los Grandes Números que se usa en los métodos de Monte-Carlo también se puede aplicar a las técnicas MCMC.

Sin embargo, hay un caso en el que la convergencia nunca ocurre: en un entorno Bayesiano, la distribución no es propia ya que la cadena puede no ser recurrente. Cuando se usan *aprioris* impropias f (muy comunes en modelos complejos) existe la posibilidad de que el producto de la verosimilitud y la apriori, $\ell(x) \cdot f(x)$, no sea integrable y, peor aún, esta situación sea indetectable (dada la complejidad de los modelos); en tales casos se pueden simular las cadenas de Markov, junto con $\ell(x) \cdot f(x)$, pero sin converger. En el mejor de los casos, las cadenas de Markov simuladas mostrarán rápidamente su comportamiento divergente, lo cual significará que hay algún problema y llevará a un análisis más detallado. Sin embargo, en los peores casos, estas cadenas de Markov presentan signos externos de estabilidad.

Ya se discutió acerca del principio detrás de las metodologías MCMC: dada una densidad objetivo f, se construirá un kernel de transición K con distribución estacionaria f y entonces se generará una cadena de Markov $\{X^{(t)}\}$ usando este kernel de tal forma que la distribución límite de $\{X^{(t)}\}$ sea f y que las integrales se puedan aproximar de acuerdo al teorema ergódico

$$\frac{1}{T}\sum_{t=1}^{T} h(X^(t)) \approx \mathbb{E}_f(h(X))$$

Por tanto, la dificultad estará en construir un kernel K que esté asociado con una densidad arbitraria f. Afortunadamente, existen métodos para obtener dichos kernel, que son universales y que son teóricamente válidos para cualquier f.

El algoritmo Metropolis-Hastings (y el Gibbs-sampler) es un ejemplo de estos métodos. Dada una densidad objetivo f, se le asocia una densidad condicional $q(y|x)$ (instrumental) que, en la práctica, sea fácil de simular. q puede ser casi arbitraria, de tal forma que $\frac{f(y)}{q(y|x)}$ sea conocida (excepto quizá por alguna constante que no depende de x) y que $q(\cdot|x)$ tenga "dispersión" suficiente para "explorar" todo el soporte de f. Es en este momento que parece pertinente mencionar que una de las ventajas del algoritmo Metropolis-Hastings es que, dado (casi) cualquier q, se puede construir un kernel Metropolis-Hastings tal que f sea su distribución estacionaria.

En esta sección se dará una breve descripción del algoritmo Metropolis-Hastings, el cual es una herramienta muy útil de Cadenas de Markov para hacer simulaciones de distribuciones multivariadas. Se revisarán algunos resultados teóricos que servirán de guía para la implementación de dicho algoritmo. A manera de ejemplo, se discutirá una aplicación del algoritmo para implementar un muestreo de aceptación-rechazo cuando no se cuenta con una función de cobertura (*blanketing function*).

1.6.1.1. Génesis desde el método de aceptación-rechazo

En contraste con los métodos MCMC, las técnicas de simulación clásicas generan muestras, generalmente independientes, no-Markovianas. Es decir, las observaciones sucesivas son independientes (estocásticamente), a menos que se introduzca correlación de manera artificial como con algún método de reducción de la varianza.

Un método importante es el método de aceptación-rechazo, cuyo objetivo es generar muestras de una densidad objetivo absolutamente continua $\pi(x) = f(x)/\kappa$, donde $f(x)$ es la densidad no-normalizada (se está usando la notación $\pi(\cdot)$ para esta distribución objetivo, sugierendo que ésta es la distribución invariante aunque aún no se prueba, sólo es una motivación), $x \in sop(f) \subseteq \mathbb{R}^d$ y $\kappa \in \mathbb{R}$ una constante de normalización (posiblemente desconocida). Sea $g(x)$ una densidad de la cual se pueden obtener simulaciones mediante algún método conocido (densidad instrumental), y supóngase que existe una constante conocida $c \in \mathbb{R}$ tal que para todo $x \in sop(f)$ $\ f(x) \leq c \cdot g(x)$. Entonces, para obtener un número aleatorio de $\pi(\cdot)$:

1. Generar un candidato z de $g(\cdot)$ y también un valor u de la distribución $Unif(0,1)$
2. Si $u \leq \frac{f(z)}{c \cdot g(z)}$, entonces $z = y$
3. Si $u > \frac{f(z)}{c \cdot g(z)}$, entonces regresar a (1).

Ya se demostró que el valor aceptado, y, efectivamente es un número proveniente de $\pi(\cdot)$ (en la sección donde se estudió el método de Aceptación-Rechazo). Para que este método

sea eficiente, se debe seleccionar c de manera cuidadosa. Ya que el número esperado de iteraciones de los pasos 1 y 2 para obtener un número aletorio (y) está dado por $\frac{1}{c}$, el método de rechazo se optimiza en

$$c = \sup_x \left\{ \frac{f(x)}{g(x)} \right\}$$

Sin embargo, aún con esta elección de c, el número de iteraciones para obtener números aleatorios de $\pi(\cdot)$ puede ser muy "grande".

Esta noción de una "densidad generadora" también aparece en el algoritmo Metropolis-Hastings, pero antes de revisar las diferencias y similitudes, se hará un análisis un poco más profundo acerca de los métodos de cadenas de Markov Monte-Carlo.

El método usual de la teoría de cadenas de Markov con espacio de estados continuo comienza con un kernel de transición $K(x, A)$ para $x \in \mathbb{R}^d$ y $A \in \mathcal{B}(\mathbb{R}^d)$ (los borelianos de $\mathbb{R}^d$). El kernel de transición es una función de distribución condicional que representa la probabilidad de moverse del punto x a algún punto del conjunto A. Ya que es una función de distribución, $K(x, \mathbb{R}^d) = 1$, donde está permitido que la cadena pueda transitar del punto x al punto x, es decir, $K(x, \{x\})$ no necesariamente es cero.

La distribución invariante satisface

$$\pi^*(dy) = \int_{\mathbb{R}^d} K(x, dy)\pi(x)dx$$

(este es un pequeño abuso de notación, pero piénsese en "dy" como un conjunto de tamaño infinitesimal) donde π es la densidad con respecto a la medida de Lebesgue de π^* (por tanto, $\pi^*(dy) = \pi(y)dy$). La n-ésima iteración está dada por $K^{(n)}(x, A) = \int_{R^d} K^{(n-1)}(x, dy)K(y, A)$, donde $K^{(1)}(x, dy) = K(x, dy)$. Bajo ciertas condiciones que se discutirán más adelante, se puede mostrar que la n-ésima iteración converge a la distribución invariante cuando $n \to \infty$.

La teoría de los métodos MCMC lo hace de manera inversa: involucra una densidad invariante $\pi(\cdot)$ conocida (excepto quizá por alguna constante), que "casualmente" es la densidad objetivo de la cual se desan obtener muestras; pero el kernel de transición es desconocido. Para generar muestras de $\pi(\cdot)$, los métodos encuentran y utilizan un kernel de transición cuya n-ésima iteración converge a $\pi(\cdot)$ para n muy "grande". El proceso empieza con un x arbitrario y se itera un gran número de veces. Después de un número suficientemente grande de iteraciones, la distribución de las observaciones generadas de la simulación es aproximadamente igual a la distribución objetivo.

Entonces, el problema ahora consiste en encontrar un $K(x, dy)$ apropiado. Esta búsqueda se puede simplificar si se hacen las siguientes consideraciones. Supóngase que el kernel

de transición se puede expresar de la forma

$$K(x, dy) = p(x, y)dy + r_1(x)\delta_x(dy)$$

para alguna función $p(x, y)$ tal que $p(x, x) = 0$, $\delta_x(dy) = 1$ si $x \in dy$ (de nuevo, este es un pequeño abuso de notación, pero piénsese en "dy" como un conjunto de tamaño infinitesimal) y $\delta_x = 0$ en otro caso; además $r_1(x) = 1 - \int_{\mathbb{R}^d} p(x, y)dy$ es la probabilidad de que la cadena permanezca en x. Dada la posibilidad de que $r_1(x) \neq 0$ es claro que la integral de $p(x, y)$ con respecto a y no necesariamente es 1.

Si la función $p(x, y)$ satisface la condición de reversibilidad (también conocida como "balance detallado", "reversibilidad microscópicà" ó "reversibilidad en el tiempo")

$$\pi(x)p(x, y) = \pi(y)p(y, x),$$

entonces $\pi(\cdot)$ es la densidad invariante de $K(x, \cdot)$.

Para verificar esto, considérese el siguiente cálculo

$$\begin{aligned}
\int K(x, A)\pi(x)dx &= \int \left(\int_A p(x, y)dy \right) \pi(x)dx + \int r_1(x)\delta_x(A)\pi(x)dx \\
&= \int_A \left(\int p(x, y)\pi(x)dx \right) dy + \int_A r_1(x)\pi(x)dx \\
&= \int_A \left(\int p(y, x)\pi(y)dx \right) dy + \int_A r_1(x)\pi(x)dx, \quad \text{pues se cumple la reversibilidad} \\
&= \int_A (1 - r_1(y))\pi(y)dy + \int_A r_1(x)\pi(x)dx \\
&= \int_A \pi(y)dy
\end{aligned}$$

Intuitivamente, el lado izquierdo de la condición de reversibilidad es la probabilidad (no condicional) de moverse de x hacia y, donde x es generado de $\pi(\cdot)$ y el lado derecho es la probabilidad (no condicional) de moverse de y hacia x. La condición de reversibilidad dice que ambos lados son iguales, y por tanto el resultado anterior muestra que $\pi^*(\cdot)$ es la distribución invariante de $K(\cdot, \cdot)$.

Este resultado da una condición suficiente (reversibilidad) que debe satisfacer $p(x, y)$. Ahora, se debe mostrar cómo el algoritmo Metropolis-Hastings encuentra $p(x, y)$ con esta propiedad.

Al igual que en el método de aceptación-rechazo, se supondrá que se tiene una densidad de la cual se pueden generar "candidatos". Sin embargo, ya que se está trabajando con cadenas de Markov, se permitirá que la densidad dependa del espacio de estados en el que se encuentra el proceso. La densidad "generadora de candidatos" se denota por $q(y|x)$ donde $\int q(y|x)dy = 1$. Se puede interpretar esta densidad diciendo que cuando un proceso está en el punto x, la densidad genera un valor y de $q(y|x)$.

La idea es encontrar $q(y|x)$ que satisfaga la condición de reversibilidad para cualesquiera x, y. Sin embargo, encontrar $q(y|x)$ que satisfaga dicha condición no es fácil.Por ejemplo, se puede encontrar q tal que para algunas x, y,

$$\pi(x)q(y|x) > \pi(y)q(x|y)$$

En este caso, de manera vaga, se puede decir que el proceso se mueve de x a y frecuentemente, pero de y a x raramente. Una forma conveniente de corregir esta situación es reduciendo el número de movimientos de x a y introduciendo una probabilidad $\rho(x, y) < 1$ de que se haga el movimiento. Se dice que $\rho(x, y)$ es la *probabilidad de movimiento*. Si el movimento no se hace, el proceso regresa de nuevo a x como un valor de la distribución objetivo. Nótese que en contraste con el método de aceptación-rechazo, en el cual cuando se rechaza un valor de y se genera un nuevo par (y, u) independientemente del valor previo de y. Por tanto, las transiciones de x a y (con $x \neq y$) se hacen de acuerdo a

$$p_{MH}(x, y) := q(y|x)\rho(x, y), \quad x \neq y,$$

donde $\rho(x, y)$ aún no se ha determinado.

Considérese de nuevo la desigualdad $\pi(x)q(y|x) > \pi(y)q(x|y)$. Ésta dice que el movimiento de x a y no se hace con suficiente frecuencia. Por tanto, se debe definir $\rho(x, y)$ tan "grande" como sea posible, y ya que es una probabilidad su límite superior es 1. La probabilidad de movimiento $\rho(x, y)$, se determinará haciendo que $p_{MH}(x, y)$ satisfaga la condición de reversibilidad, es decir,

$$\pi(x)q(y|x)\rho(x, y) = \pi(y)q(x|y)\rho(y, x) = \pi(y)q(x|y)$$

Por tanto,

$$\rho(x, y) = \frac{\pi(y)q(x|y)}{\pi(x)q(y|x)}$$

Sin embargo, si ahora ocurriera $\pi(x)q(y|x) < \pi(y)q(x|y)$, entonces $\rho(x, y) = 1$, y entonces la condición de reversibilidad establece

$$\pi(x)q(y|x) = \pi(x)q(y|x)\rho(x, y) = \pi(y)q(x|y)\rho(y, x)$$

y por tanto

$$\rho(y,x) = \frac{\pi(x)q(y|x)}{\pi(y)q(x|y)}$$

Por tanto, las probabilidades $\rho(x,y)$ y $rho(y,x)$ se introducen para asegurar que ambos lados de la desigualdad $\pi(x)q(y|x) > \pi(y)q(x|y)$ estén "balanceados", es decir $p_{MH}(x,y)$ satisfaga la reversibilidad. Por tanto, se ha mostrado que para que $p_{MH}(x,y)$ sea reversible, la probabilidad de movimiento debe hacerse

$$\rho(x,y) = \begin{cases} \text{mín}\left\{\frac{\pi(y)q(x|y)}{\pi(x)q(y|x)}, 1\right\} & \text{si } \pi(x)q(y|x) > 0; \\ 1 & \text{e.o.c} \end{cases}$$

Para terminar la definición del kernel de transición para la cadena de Metropolis-Hastings se debe considerar la probabilidad (posiblemente distinta de cero) de que el proceso permanezca en x. Como se definió anteriormente

$$r_1(x) = 1 - \int_{\mathbb{R}^d} p(x,y)dy = 1 - \int_{\mathbb{R}^d} q(y|x)\rho(x,y)dy$$

Por tanto el Kernel de transición de la cadena Metropolis-Hastings, que se denotará por K_{MH}, está dado por

$$K_{MH}(x,dy) = q(y|x)\rho(x,y)dy + \left(1 - \int_{\mathbb{R}^d} q(y|x)\rho(x,y)dy\right)\delta_x(dy)$$

Ya que por construcción p_{MH} es reversible, entonces el kernel Metropolis-Hastings tiene a $\pi(x)$ como densidad invariante.

Se deben hacer algunas observaciones acerca de este algoritmo:

1. El algoritmo Metropolis-Hastings está determinado por su densidad "generadora de candidatos" $q(y|x)$. La selección de $q(x,y)$ se discutirá más adelante.

2. Si un valor candidato se rechaza, el valor actual se toma como el siguiente en la sucesión.

3. El cálculo de $\rho(x,y)$ no requiere que se conozca la constante de normalización de π ya que aparece en el numerador y denominador de ρ.

4. Si la densidad "generadora de candidatos" es simétrica, i.e. $q(y|x) = q(x|y)$, entonces la probabilidad de movimiento se reduce a $\frac{\pi(y)}{\pi(x)}$; por tanto, si $\pi(y) \geq \pi(x)$ la cadena se mueve hacia y, y en otro caso se mueve con probabilidad $\frac{\pi(y)}{\pi(x)}$. En otras palabras, si el brinco va "cuesta arriba" siempre se acepta, si el brinco va "cuesta abajo" se acepta con una probabilidad distinta de cero.

Este algoritmo fue propuesto po Metropolis, N., Rosenbluth, A. W., Rosenbluth, M. N., Teller, A. H. & Teller, E. en 1954 en el artículo "Equations of State Calculations by Fast Computing Machines".

Para un valor arbitrario x_0 el Algoritmo Metropolis-Hastings se puede resumir como:

Algoritmo 1.6.1. *(Algoritmo Metropolis-Hastings I)*

1. Para $j \in \{1, 2, \ldots, N\}$
2. Generar y de $q(\cdot|x^{(j)})$ y u de $Unif(0,1)$
3. Si $u \leq \rho(x^{(j)}, y)$, hacer $x^{(j+1)} = y$
4. Si $u > \rho(x^{(j)}, y)$, hacer $x^{(j+1)} = x^{(j)}$

donde

$$\rho(x,y) = \begin{cases} \text{mín}\left\{\frac{\pi(y)q(x|y)}{\pi(x)q(y|x)}, 1\right\} & \text{si } \pi(x)q(y|x) > 0; \\ 1 & \text{e.o.c} \end{cases}$$

O bien,

Algoritmo 1.6.2. *(Algoritmo Metropolis-Hastings II)*

Dado $x^{(t)}$,

1. Generar y_t de $q(y|x^{(t)})$
2. Hacer

$$x^{(t+1)} = \begin{cases} y_t & \text{con probabilidad } \rho(x^{(t)}, y_t); \\ x^{(t)} & \text{con probabilidad } 1 - \rho(x^{(t)}, y_t) \end{cases}$$

donde

$$\alpha(x,y) = \begin{cases} \text{mín}\left\{\frac{\pi(y)q(x|y)}{\pi(x)q(y|x)}, 1\right\} & \text{si } \pi(x)q(y|x) > 0; \\ 1 & \text{e.o.c} \end{cases}$$

Para implementar el algoritmo Metropolis-Hastings se necesita especificar una densidad "generadora de candidatos" adecuada. Generalemente se selecciona esta densidad de una familia de distribuciones que frecuentemente se requiere conocer los parámetros asociados

a ellas (tales como parámetros de localización y escala).

Una familia de densidades "generadoras de candidatos" que aparece en el trabajo original de Metropolis et al. (1953) está dada por $q(y|x) = q_1(y - x)$, donde q_1 es una densidad multivariada. El candidato y se selecciona de acuerdo al proceso $y = x + Z$, donde Z se conoce incremento aleatorio con distribución q_1. Ya que el candidato es igual al valor actual más un "ruido" este caso se conoce como cadena de caminata aleatoria. Algunas elecciones comunes para q_1 son la normal multivariada y la t de Student multivariada. Nótese que cuando q_1 es simétrica, i.e. $q_1(z) = q_1(-z)$, la probabilidad de movimiento se reduce a

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

1.6.1.2. Una aplicación de los algoritmos de aceptación-rechazo y Metropolis-Hastings

Recuérdese que en el método de aceptación-rechazo se necesitaba una constante c y una densidad instrumestral g que domine ó cubra a la densidad objetivo f (posiblemente no normalizada). Encontrar dicha $c \in \mathbb{R}$ puede ser difícil en algunas aplicaciones. Además, si $f(x)$ depende de parámetros que se actualizan durante un ciclo iterativo, encontrar c para cada nuevo conjunto de parámetros hace que el algoritmo sea significativamente lento. Es po esto que vale la pena tener un método de aceptación-rechazo que no requiera de una función instrumental. Tierney (1994) propusó un algoritmo que usa un paso aceptación-rechazo para generar candidatos para un algoritmo Metropolis-Hastings.

Para fijar ideas recuérdese que se está interesado en simular muestras de la densidad objetivo $\pi(x) = f(x)/\kappa$, donde κ puede ser desconocida y se dispone de una densidad instrumental g para hacer muestreo. Supóngase que $c > 0$ es una constante conocida pero que $f(x)$ no necesariamente es menor que $c \cdot g(x)$ para todo $x \in sop(f)$, i.e. $c \cdot g(x)$ no necesariamente domina a $f(x)$. Sea

$$C := \{x \in \mathbb{R} : f(x) < c \cdot g(x)\}$$

En este algoritmo, dado $x^{(n)} = x$, el siguiente valor $x^{(n+1)}$ se obtiene como sigue: primero, se obtiene un valor candidato z independiente del valor actual x meiante el algoritmo de aceptación-rechazo usando $c \cdot g$ como la densidad "dominante".

Nótese que la densidad de una variable aleatoria que se genera mediante este procedimiento está dada por

$$\begin{aligned} q(y) &= \mathbb{P}(y|U \leq f(Z)/cg(Z)) \\ &= \frac{\mathbb{P}(U \leq f(Z)/cg(Z)|Z = y) \cdot g(y)}{\mathbb{P}(U \leq f(Z)/cg(Z))} \end{aligned}$$

Sin embargo $\mathbb{P}(U \leq f(Z)/cg(Z)|Z = y) = \text{mín}\left\{\frac{f(y)}{cg(y)}, 1\right\}$ y si se define $\varsigma := \mathbb{P}(U \leq f(Z)/cg(Z))$ entonces

$$q(y) = \frac{\text{mín}\left\{\frac{f(y)}{cg(y)}, 1\right\} \cdot g(y)}{\varsigma}$$

Simplificando un poco esta expresión se tiene que

$$q(y) = \begin{cases} \frac{f(y)}{c\cdot\varsigma} & \text{si } y \in C; \\ \frac{g(y)}{\varsigma} & \text{si } y \notin C \end{cases}$$

Nótese que en este caso no es necesario escribir $q(y|x)$ ya que el candidato y se simula independientemente de x.

Como $c \cdot g$ no domina a la densidad objetivo en C^c, la densidad objetivo no se simula adecuadamente en C^c. Esto se puede corregir con un paso Metropolis-Hastings aplicado a los valores y que vienen del paso aceptación-rechazo. Ya que x, y pueden estar en C ó en C^c se tienen cuatro casos:

(a) $x \in C$ y $y \in C$

(b) $x \notin C$ y $y \in C$

(c) $x \in C$ y $y \notin C$

(d) $x \notin C$ y $y \notin C$

El objetivo que se tiene ahora es encontrar una probabilidad de movimiento Metropolis-Hastings $\rho(x, y)$ tal que $q(y)\rho(x, y)$ satisface la condición de reversibilidad.

Se obtendrá $\rho(x, y)$ en los cuatro casos. Se considerará $\pi(x)q(y)$ y $\pi(y)q(x)$ (ó equivalentemente $f(x)q(y)$ y $f(y)q(x)$) para ver cómo se debe definir la probabilidad de moviemiento para asegurar la revesibilidad. Es decir, se necesita encontrar $\rho(x, y)$ y $\rho(y, x)$ tal que

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

en cada uno de los casos con

$$q(y) = \begin{cases} \frac{f(y)}{c\cdot\varsigma} & \text{si } y \in C; \\ \frac{g(y)}{\varsigma} & \text{si } y \notin C \end{cases}$$

(a) $x \in C$ y $y \in C$. En este caso se tiene que $f(x)q(y) = \frac{f(x)f(y)}{c\varsigma}$ es igual a $f(y)q(x)$, y por lo tanto $\rho(x, y) = \rho(y, x) = 1$ satisface la reversibilidad.

(b) $x \notin C$ y $y \in C$. En este caso $f(x) > cg(x)$ ó bien $g(x) < f(x)/c$ y multiplicando ambos lados por $f(y)/\varsigma$ se tiene que

$$\frac{f(y)g(x)}{\varsigma} < \frac{f(y)f(x)}{c\varsigma}$$

ó por la construcción de q, $f(y)q(x) < f(x)q(y)$. Nótese que hay relativamente pocas transiciones de y a x y demasiadas en la dirección opuesta. Haciendo $\rho(y,x) = 1$ se "arregla" este problema y entonces $\rho(x,y)$ se determina de tal fomra que

$$\frac{f(y)g(x)}{\varsigma} = \rho(x,y)\frac{f(y)f(x)}{c\varsigma}$$

obteniéndose $\rho(x,y) = cg(x)/f(x)$.

(c) $x \in C$ y $y \notin C$. Es análogo al caso anterior intercambiando x por y.

(d) $x \notin C$ y $y \notin C$. En este caso se tiene que $f(x)q(y) = \frac{f(x)g(y)}{\varsigma}$ y $f(y)q(x) = \frac{f(y)g(x)}{d}$ y hay dos posibilidades. Habrá muy pocas transiciones de y a x para satisfacer la reversibilidad si

$$\frac{f(x)g(y)}{\varsigma} > f(y)q(x).$$

En este caso si se hace $\alpha(y,x) = 1$ y $\alpha(x,y)$ tal que

$$\alpha(x,y)\frac{f(x)g(y)}{\varsigma} = \frac{f(y)g(x)}{\varsigma}$$

que implica que

$$\alpha(x,y) = \text{mín}\{\frac{f(y)g(x)}{f(x)g(y)}, 1\}$$

Si hay muy pocas transiciones de x a y, sólo se intercambian x y y en el argumento anterior.

Nótese que en los casos en los que $x \in C$, la probabilidad de moverse hacia y es 1, sin importar si $y \in C$ o $y \notin C$.

Se puede resumir el algoritmo, de la siguiente forma:

Algoritmo 1.6.3. *(Algoritmo Aceptación-Rechazo/Metropolis-Hastings)*

1. Sean $C_1 = \{f(x) < cg(X)\}$ y $C_2 = \{f(x) < cg(y)\}$
2. Genérese u de la distribución $Unif(0,1)$ y
   - Si $C_1 = 1$, hágase $\alpha = 1$
   - Si $C_1 = 0$ y $C_2 = 1$, hágase $\alpha = \frac{cg(x)}{f(x)}$
   - Si $C_1 = 0$ y $C_2 = 0$, hágase $\alpha = \text{mín}\{\frac{f(y)g(x)}{f(x)g(y)}, 1\}$
3. Si $u \leq \alpha$, regrésese y
4. Si $u > \alpha$, regrésese x

Ejemplo 1.6.2. *(Simulaciones de la normal bivariada)*

Considérese el ejemplo tradicional de hacer simulaciones de una distribución normal bivariada $N_2(\mu, \Sigma)$ con $\mu = (1,2)'$ y

$$\Sigma = \begin{pmatrix} 1 & 0.9 \\ 0.9 & 1 \end{pmatrix}$$

Si bien esto se puede hacer mediante una descomposición de Choleski haciendo $y = \mu + P'u$ donde u es una simulación de la distribución $N_2(0, I_2)$ y P satisface $P'P = \Sigma$ este es un ejemplo tradicional de la implementación del algorimo Metropolis-Hastings.

A partir de la expresión de la densidad normal multivariada, la probabilidad de movimiento (para una densidad generadora simétrica) es

$$\rho(x,y) = \text{mín}\left\{\frac{exp\{-\frac{1}{2}(y-\mu)'\Sigma(y-\mu)\}}{exp\{-\frac{1}{2}(x-\mu)'\Sigma(x-\mu)\}}, 1\right\}$$

∇

Ejemplo 1.6.3. *(Simulaciones de una serie de tiempo $AR(2)$)*

Ahora se analizará el uso del algoritmo Metropolis-Hastings para hacer simulaciones de ua distribición analíticamente intratable que surge en un modelo de series de tiempo $AR(2)$. Se desea hacer 100 simulaciones del modelo

$$Y_t = \phi_1 Y_{t-1} + \phi_2 Y_{t-2} + \epsilon_t, \quad t = 1, \ldots, 100$$

donde $\phi_1 = 1$ y $\phi_2 = -1/2$ y $\epsilon_t \sim N(0,1)$. Los valores de $\phi = (\phi_1, \phi_2)$ pertenece a la región $S \subseteq \mathbb{R}^2$ que satisface las restricciones de estacionariedad

$$\phi_1 + \phi_2 < 1, \quad \phi_2 - \phi_1 < 1, \quad \phi_2 > -1$$

Según Box & Jenkins (1976) se tiene que la función de verosimilitud para este modelo está dada por

$$l(\phi, \sigma^2) = \Psi(\phi, \sigma^2)(\sigma^2)^{-(n-2)/2} \exp\left\{-\frac{1}{2\sigma^2}\sum_{t=3}^{n}(y_t - w_t'\phi)^2\right\}$$

donde $n = 100$, los datos muestrales son $Y_n = (y_1, \ldots, y_n)'$, $w_t := (y_{t-1}, y_{t-2})'$,

$$\Psi(\phi, \sigma^2) = \frac{\sigma^2}{|V^{-1}|^{1/2}} \exp\left\{-\frac{1}{2\sigma^2 Y_2' V^{-1} Y_2}\right\}$$

es la densidad de $Y_2 = (y_1, y_2)'$ y

$$V^{-1} = \begin{pmatrix} 1-\phi_2^2 & -\phi_1(1+\phi_2) \\ -\phi_1(1+\phi_2) & 1-\phi_2^2 \end{pmatrix}$$

El término $\exp\left\{-\frac{1}{2\sigma^2}\sum_{t=3}^{n}(y_t - w_t'\phi)^2\right\}$ es proporcional a la densidad de $(y_3, \ldots, y_n)$ dado Y_2.

Si la única información *a priori* disponible es que el proceso es estacionario, entonces la distribución posterior de los parámetros está dada por

$$\pi(\phi, \sigma^2 | Y_n) \propto l(\phi, \sigma^2) I_S(\phi)$$

Para simular de esta densidad posterior se debe notar que

$$\exp\left\{-\frac{1}{2\sigma^2}\sum_{t=3}^{n}(y_t - w_t'\phi)^2\right\} \propto \exp\{-\frac{1}{2\sigma^2}(\phi - \hat{\phi})' G (\phi - \hat{\phi})\}$$

donde $\hat{\phi} = G^{-1}\sum_{t=3}^{n} w_t y_t$ y $G = \sum_{t=3}^{n} w_t w_t'$. Este es el kernel de una densidad normal con media $\hat{\phi}$ y matriz de varianzas-covarianzas $\sigma^2 G^{-1}$. Con todas estas observaciones se puede concluir que la densidad de σ^2 dado ϕ y Y_n es una gamma inversa con parámetros $n/2$ y $Y_2 V^{-1} Y_2 + \sum_{t=3}^{n}(y_t - w_t'\phi)^2$ y que la densidad de ϕ dado σ^2 y Y_n es

$$\pi(\phi | Y_n, \sigma^2) \propto \Psi(\phi, \sigma^2) f_{nor}(\phi | \phi, \sigma^2 G^{-1}) I_S(\phi)$$

donde f_{nor} es la función de densidad de la distribución normal.

Se pueden hacer simulaciones de $\pi(\sigma^2, \phi|Y_n)$ obteniendo una simulación de $\pi(\phi|Y_n, \sigma^2)$ y dados este valor, simular σ^2 de $\pi(\sigma^2|Y_n, \phi)$.

ϕ se simula de la siguiente forma: en la j-ésima iteración (dado el valor de $\sigma^{2(j)}$, se simula un candidato ϕ^{j+1} de la densidad normal con media $\hat{\phi}$ y covarianza $\sigma^{2(j)}G^{-1}$; si se satisface la estacionariedad se mueve a este punto con probabilidad mín$\{\frac{\Psi(\phi^{(j+1)},\sigma^{2(j)})}{\Psi(\phi^{(j)},\sigma^{2(j)})}, 1\}$ y en otro caso $\phi^{(j+1)} = \phi^{(j)}$.

Se puede aplicar el método de aceptación-rechazo tradional simulando candidatos $\phi^{(j+1)}$ de la densidad normal $f_{nor}(\phi|\phi, \sigma^2 G^{-1})I_S(\phi)$ hasta que $U \leq \Psi(\phi^{(j+1)}, \sigma^{2(j)})$. Se necesitarán muchos intentos de ϕ antes de que se acepte uno ya que $\Psi(\phi, \sigma^2)$ se puede convertir en una cantidad extremadamente pequeña. Por tanto, aunque se puede usar el algoritmo de aceptación-rechazo no es tan poderoso como la implementación con el algoritmo Metropolis-Hastings

∇

1.6.2. Descripción formal del algoritmo

Antes de exhibir formalmente la universalidad de los algoritmos Metropolis-Hastings, primero se analizará su validez teórica.

A manera de resumen de lo que se hizo en la sección previa, el algoritmo Metropolis-Hastings comienza con la densidad objetivo f. Entonces se selecciona una densidad condicional $q(y|x)$ que se define con respecto a la medida dominante del modelo. El algoritmo Metropolis-Hastings se puede implementar en la práctica cuando se puede simular fácilmente de $q(\cdot|x)$ y se conoce explícitamente (excepto quizá por una constante multiplicativa que no depende de x) ó bien se sabe que es simétrica, i.e. $q(x|y) = q(y|x)$. Ya se dijo que un requisito más general es que el cociente $f(y)/q(y|x)$ sea conocido (excepto quizá por un factor multiplicativo) y no dependa de x.

El algoritmo Metropolis-Hastings asociado con la densidad objetivo f y la densidad condicional q genera una cadena de Markov $\{X^{(t)}\}$ mediante la siguiente transición:

Algoritmo 1.6.4. *(Algoritmo Metropolis-Hastings)*

Dado $x^{(t)}$,

1. Generar $Y_t \sim q(y|x^{(t)})$
2. Hacer
$$X^{(t+1)} = \begin{cases} Y_t & \text{con probabilidad } \rho(x^{(t)}, Y_t); \\ x^{(t)} & \text{con probabilidad } 1 - \rho(x^{(t)}, Y_t) \end{cases}$$

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

La distribución q se conoce como *distribución instrumental.*

Este algoritmo siempre acepta valores y_t tales que el cociente $f(y_t)/q(y_t|x^{(t)})$ sea mayor en comparación con el cociente previo $f(x^{(t)})/q(x^{(t)}|y_t)$. Sólo en el caso simétrico la aceptación se debe al cociente de verosimilitudes $f(y_t)/f(x^{(t)})$. Una característica importante del algoritmo (1.6.4) es que puede aceptar valores y_t tales que el cociente haya disminuido. Como en el método Aceptación-Rechazo, el algoritmo Metropolis-Hastings sólo depende de los cocientes

$$f(y_t)/f(x^{(t)}) \quad \text{y} \quad q(x^{(t)}|y_t)/q(y_t|x^{(t)})$$

Y por tanto, no depende de constantes de normalización suponiendo que $q(\cdot|y)$ se conoce (excepto quizá por una constante de normalización) que es independiente de y (que no siempre es el caso).

La probabilidad $\rho(x^{(t)}, y_t)$ se define sólo cuando $f(x^{(t)}) > 0$. Sin embargo, si la cadena empieza con un valor $x^{(0)}$ tal que $f(x^{(0)}) > 0$ entonces para todo $t \in \mathbb{N}^+$ $f(x^{(t)}) > 0$ ya que los valores de y_t tales que $f(y_t) = 0$ inducen $\rho(x^{(t)}, y_t) = 0$ y por tanto el algoritmo los rechaza.

Notación 1.6.1.

Para evitar dificultades teóricas, se hará la convención de que el cociente $\rho(x, y)$ es igual a cero si $f(x)$ y $f(y)$ son cero. ▽

Hay similitudes entre (1.6.4) y el método de Aceptación-Rechazo y es posible utilizar (1.6.4) como una alternativa del algoritmo Aceptación-Rechazo para una pareja (f, g) dada.

Una muestra generada por (1.6.4) no es una muestra independiente e idénticamente distribuida. Por un lado, tal muestra podría involucrar repeticiones del mismo valor ya que el rechazo de Y_t lleva a la repetición de $X^{(t)}$ al tiempo $t+1$. Por tanto, para calcular una media tal como

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

los Y_t's generados por el algoritmo (1.6.4) se pueden asociar con pesos de la forma m_t/T ($m_t = 0, 1, \ldots$) donde m_t cuenta el número de veces que se rechazaron valores subsecuentes.

Se necesitan condiciones de regularidad mínimas en f y q para que f sea la distribución límite de la cadena, $\{X^{(t)}\}$, generada por (1.6.4); por ejemplo, que $sop(f) = \mathcal{E}$ sea conexo. En la literatura generalmente se omite el hecho de que si $\mathcal{E}$ no es conexo, el algoritmo Metropolis-Hastings puede ya no ser válido. Para tales casos se tiene que trabajar en una componente conexa a la vez y probar que las diferentes componentes conexas se "unen" mediante el kernel de (1.6.4). Si q trunca al espacio de estados, i.e. si existe $A \subseteq \mathcal{E}$ tal que

$$\int_A f(x)dx > 0 \quad \text{y} \quad \int_A q(y|x)dy = 0$$

el algoritmo (1.6.4) no tendrá a f como distribución límite ya que para $x^{(0)} \notin A$ la cadena $\{X^{(t)}\}$ nunca visitará a A. Por tanto, una condición necesaria minimal es que

$$sop(f) \subseteq \bigcup_{x \in sop(f)} sop(q(\cdot|x))$$

Para ver de manera más artificial al de la motivación que f es la distribución estacionaria de la cadena Metropolis se analizará el kernel y se probará que satisface una pequeña condición equivalente a la condición de balance detallado.

Definición 1.6.2. *(Condición de balance detallado)*

Una cadena de Markov con función de transición K satisface la condición de balance detallado si existe una función f tal que para cualesquira x, y

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

Teorema 1.6.1.

Supóngase que una cadena de Markov con función de transición K satisface la condición de balance detallado con f una función de densidad de probabilidad. Entonces,

(i) La densidad f es la densidad invariante de la cadena.

(ii) La cadena es reversible.

Demostración:

Según la condición de balance detallado, para cualquier B medible se tiene que

$$\begin{aligned}\int_{\mathcal{Y}} K(y,B)f(y)dy &= \int_{\mathcal{Y}}\int_{B} K(y,x)f(y)dxdy \\ &= \int_{\mathcal{Y}}\int_{B} K(x,y)f(x)dxdy = \int_{B} f(x)dx\end{aligned}$$

pues $\int K(x,y)dy = 1$. Dada la existencia del kernel K y la densidad invariante f, la condición de balance detallado y la reversibilidad son la misma propiedad. □

Teorema 1.6.2.

Para toda distribución condicional q tal que $\mathcal{E} \subseteq sop(q)$, f es una distribución estacionaria de la cadena $\{X^{(t)}\}$ generada por (1.6.4)

Demostración:

El kernel de transición asociada con (1.6.4) es

$$K(x,y) = \rho(x,y)q(y|x) + [1 - r(y)]\delta_x(y) \tag{1.5}$$

donde $r(x) = \int \rho(x,y)q(y|x)dy$ y δ_x es la masa de Dirac en x. Entonces,

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

y

$$[1 - r(x)]\delta_x(y)f(x) = [1 - r(y)]\delta_y(x)f(y)$$

por tanto, se cumple la condición de balance detallado para la cadena Metropolis-Hastings. □

En virtud del resultado anterior, la estacionariedad de f se cumple para casi cualquier distribución condicional q, hecho que exhibe la universalidad de los Algoritmos Metropolis-Hastings.

1.6.2.1. Propiedades de convergencia

Para probar que la cadena de Markov (1.6.4) en verdad converge, se utilizarán los resultados del Apéndice A.

Por construcción, la cadena de Markov Metropolis-Hastings tiene una distribución de probabilidad invariante, si esta distribución invariante es Harris-recurrente y aperiódica entonces (por el Teorema ergódico) se tiene que

$$\frac{1}{T}\sum_{t=1}^{T} h(X^{(t)}) \to \mathbb{E}_f(h(X))$$

Una condición suficiente para que la cadena de Markov Metropolis-Hastings sea aperiódica es que el algoritmo (1.6.4) permita eventos tales que tales como $(X^{(t+1)} = X^{(t)})$, i.e. que la probabilidad de dichos eventos no sea cero y por tanto

$$\mathbb{P}\left(f(X^{(t)})q(Y_t|X^{(t)}) \leq f(Y_t)q(X^{(t)}|Y_t)\right) < 1$$

Esta condición implica que q no es el kernel de transición de una cadena de Markov reversible con distribución estacionaria f. Nótese que q no es el kernel de transición de la cadena Metropolis-Hastings dada por (1.6.4) que SÍ es reversible.

La propiedad de irreductibilidad de la cadena Metropolis-Hastings $\{X^{(t)}\}$ es consecuencia de la condición de positividad de la densidad condicional q, i.e. para cualesquiera $x, y \in \mathcal{E}$

$$q(y|x) > 0$$

Entonces, todo subconjunto de $\mathcal{E}$ con medida de Lebesgue positiva se puede alcanzar en un sólo paso. Como la densidad f es la medida invariante para la cadena, entonces la cadena es positiva y recurrente. Sin embargo, se tiene en resultado más fuerte para la cadena Metropolis-Hastings.

Lema 1.6.1.

Si la cadena Metropolis-Hastings es f-irreducible, entonces es Harris-recurrente.

Demostración:

Se probará este lema usando el hecho de que lo que caracteriza a la Harris-recurrencia es que sólo las únicas funciones harmónicas acotadas son las constantes. (Véase Apéndice A.)

Si h es una función harmónica, entonces

$$h(x_0) = \mathbb{E}(h(X^{(1)})|x_0) = \mathbb{E}(h(X^{(t)})|x_0)$$

ya que la cadena Metropolis-Hastings es positiva recurrente y aperiódica, entonces se puede concluir que h es constante casi donde sea (con respecto a f) y es igual a $\mathbb{E}_f(h(X))$. Como

$$\mathbb{E}(h(X^{(1)})|x_0) = \int \rho(x_0, x_1)q(x_1|x_0)h(x_1)dx_1 + [1 - r(x_0)]h(x_0)$$

entonces

$$\mathbb{E}_f(h(X))r(x_0) + [1 - r(x_0)]h(x_0) = h(x_0)$$

i.e. para cualquier $x_0 \in \mathcal{E}$, $[h(x_0) - \mathbb{E}_f(h(X))]r(x_0) = 0$.

Finalmente, como para todo $x_0 \in \mathcal{E}$, $r(x_0) > 0$ y la f-irreductibilidad, h necesariamente es constante. □

Entonces, ya se puede establecer un primer resultado de convergencia para las cadenas de Markov Metropolis-Hastings.

Teorema 1.6.3.

Supóngase que una cadena de Markov Metropolis-Hastings, $\{X^{(t)}\}$, es f-irreducible.

(i) Si $h \in L^1(f)$, entonces

$$\lim_{T\to\infty} \frac{1}{T}\sum_{t=1}^{T} h(X^{(t)}) = \int h(x)f(x)dx \quad \text{para casi todo } f$$

(ii) Si además $\{X^{(t)}\}$ es aperiódica, entonces

$$\lim_{n\to\infty} \left\| \int K^n(x,\cdot)\nu(dx) - f \right\|_{TV} = 0$$

para cualquier distribución inicial ν, donde $K^n(x,\cdot)$ es el kernel para n transiciones.

Demostración:

Si $\{X^{(t)}\}$ es f-irreducible, por el lema anterior $\{X^{(t)}\}$ es Harris-recurrente y por el Teorema ergódico se cumple (i). El inciso (ii) es consecuencia del Teorema (5.2) del apéndice A.

□

Como la f-irreductibilidad de la cadena Metropolis-Hastings se debe a la propiedad de positividad de la densidad condicional q, se tiene el siguiente corolario.

Corolario 1.6.1.

Las conclusiones del Teorema anterior se cumplen si la cadena de Markov Metropolis-Hastings, $\{X^{(t)}\}$, tiene densidad condicional $q(x|y)$ que satisface $\mathbb{P}\left(f(X^{(t)})q(Y_t|X^{(t)}) \leq f(Y_t)q(X^{(t)}|Y_t)\right) < 1$ y $q(y|x) > 0$.

Para terminar de estudiar estos aspectos técnicos del Algoritmo (1.6.4) se dará un resultado de Roberts & Tweedie (1996) que no requiere que $q(y|x)$ se estrictamente positivo para todo $x, y \in \mathcal{E}$.

Lema 1.6.2.

Supóngase que f está acotada y es positiva en todo compacto de $\mathcal{E}$. Si existen $\varepsilon, \delta > 0$ tales que

$$|x - y| < \delta \implies q(y|x) > \varepsilon$$

entonces la cadena de Markov Metropolis-Hastings, $\{X^{(t)}\}$, es f-irreducible y aperiódica. Además, todo compacto no vacío es un conjunto pequeño (véase Apéndice A).

Demostración:

Sea $x^{(0)} \in A \subseteq \mathcal{E}$ un valor inicial arbitrario. La conexidad de $\mathcal{E}$ implica que existen $m \in \mathbb{N}^+$ y una sucesión $\{x^{(i)}\}_{i=1}^m$ tales que $x^{(m)} \in A$ y $|x^{(i+1)} - x^{(i)}| < \delta$. Por tanto, se puede unir a $x^{(0)}$ con A mediante una sucesión de bolas de radio δ. Las suposiciones sobre f implican que la probabilidad de aceptación de un punto $x^{(i)}$ de la i-ésima bola, comenzando de la bola $i-1$, es positiva y por tanto $\mathbb{P}^m_{x^{(0)}}(A) = \mathbb{P}(X^{(m)} \in A | X^{(0)} = x^{(0)}) > 0$

$$\therefore \{X^{(t)}\} \text{ es } f - \text{irreducible}$$

Para $x^{(0)} \in \mathcal{E}$ arbitrario, y para $y \in B(x^{(0)}, \delta/2)$ (la bola con centro en $x^{(0)}$ y radio $\delta/2$) se tiene que

$$\begin{aligned} \mathbb{P}_y(A) &\geq \int_A \rho(y,z) q(z|y) dz \\ &= \int_{A \cap D_y} \frac{f(z)}{f(y)} q(z|y) dz + \int_{A \cap D_y^c} q(z|y) dz, \end{aligned}$$

donde $D_y := \{z : f(z)q(y|z) \leq f(y)q(z|y)\}$. Por tanto,

$$\begin{aligned} \mathbb{P}_y(A) &\geq \int_{A \cap D_y \cap B} \frac{f(z)}{f(y)} q(z|y) dz + \int_{A \cap D_y^c \cap B} q(z|y) dz \\ &\geq \frac{\text{ínf}_B f(x)}{\sup_B f(x)} \int_{A \cap D_y \cap B} q(z|y) dz + \int_{A \cap D_y^c \cap B} q(z|y) dz \\ &\geq \frac{\text{ínf}_B f(x)}{\sup_B f(x)} \lambda(A \cap B) \end{aligned}$$

donde $\lambda(\cdot)$ es la medida de Lebesgue en $\mathcal{E}$. Las bolas $B(x^{(0)}, \delta/2)$ son conjuntos pequeños asociados con distribuciones uniformes sobre $B(x^{(0)}, \delta/2)$. Esto simultáneamente la aperiodicidad de $\{X^{(t)}\}$ y el hecho de que cada compacto es un conjunto pequeño (véase Apéndice

A). □

Observación 1.6.1.

El razonamiento detrás de este resultado es el siguiente: si la distribución condicional $q(y|x)$ permite movimientos en una vecindad (con diámetro acotado por abajo) de $x^{(t)}$ y si f es tal que $\rho(x^{(t)}, y)$ es positiva en esta vecindad, entonces cualquier subconjunto de $\mathcal{E}$ se puede visitar en k pasos, para k suficientemente grande. Se debe suponer que $\mathcal{E}$ es conexo. ▽

Corolario 1.6.2.

Supóngase que una cadena de Markov Metropolis-Hastings, $\{X^{(t)}\}$, es f-irreducible con f densidad de probabilidad invariante y $q(x|y)$ que satisface la condiciones del Lema anterior. Entonces,

(i) Si $h \in L^1(f)$, entonces

$$\lim_{T\to\infty} \frac{1}{T}\sum_{t=1}^{T} h(X^{(t)}) = \int h(x)f(x)dx \quad \text{para casi todo } f$$

(ii) Si además $\{X^{(t)}\}$ es aperiódica, entonces

$$\lim_{n\to\infty} \left\| \int K^n(x,\cdot)\nu(dx) - f \right\|_{TV} = 0$$

para cualquier distribución inicial ν, donde $K^n(x,\cdot)$ es el kernel para n transiciones.

1.6.3. Algunos algoritmos Metropolis-Hastings particulares

Uno de los aspectos más interesantes del algoritmo (1.6.4) es su universalidad, i.e. el hecho de que con una distribución condicional arbitraria q con soporte en $\mathcal{E}$ se pueda simular de una distribución arbitraria f sobre $\mathcal{E}$.

Por otro lado, esta universalidad puede ser sólo una formalidad si la distribución instrumental q simula valores raramente en la parte "principal" de $\mathcal{E}$, i.e. en la región donde se localiza la mayoría de la masa de la densidad f.

Ahora se verán algunas metodologías específicas estudiadas en la literatura con algunas propiedades probabilísticas. Una clasificación completa de algoritmos Metropolis-Hastings

es imposible dada la versatilidad del método y la posibilidad de crear métodos híbridos con varios de estos algoritmos.

1.6.3.1. El caso independiente

Este método se puede considerar como una generalización del algoritmo de Aceptación-Rechazo (ya en la motivación se intuía esto, pero ahora se hará con un poco más de formalidad) en el sentido de que la distribución instrumental q no depende de $x^{(t)}$ y por analogía se denota por g. Entonces, el Algoritmo (1.6.4) genera la siguiente transición de $x^{(t)}$ a $X^{(t+1)}$.

Algoritmo 1.6.5. *(Algoritmo Metropolis-Hastings Independiente)*

Dado $x^{(t)}$,

1. Generar $Y_t \sim g(y)$
2. Hacer
$$X^{(t+1)} = \begin{cases} Y_t & \text{con probabilidad mín}\{\frac{f(Y_t)g(x^{(t)})}{f(x^{(t)})g(Y_t)}, 1\}; \\ x^{(t)} & \text{en otro caso} \end{cases}$$

Aunque en este caso las Y_t's se generan independientemente, la muestra resultante no es independiente ya que la probabilidad de aceptar Y_t depende de $X^{(t)}$.

Las propiedades de convergencia de la cadena $X^{(t)}$ son consecuencia de las propiedades de la densidad g en el sentido de que $X^{(t)}$ es irreducible y aperiódica si y sólo si g es positiva casi en todo el soporte de f. Algunas propiedades más fuertes de convergencia como la ergodicidad uniforme o ergodicidad geométrica se enuncian en el siguiente resultado de Mengersen & Tweedie (1996).

Teorema 1.6.4.

El Algoritmo (1.6.5) genera una cadena uniformente ergódica si existe una constante M tal que para todo $x \in sop(f)$,

$$f(x) \leq Mg(x)$$

y en este caso

$$\|K^n(x,\cdot) - f\|_{TV} \leq 2\left(1 - \frac{1}{M}\right)^n$$

donde $\|\cdot\|_{TV}$ es la norma de variación total (véase Apéndice A). Por otro lado, si para cada M existe un conjunto de medida positiva tal que $f(x) > Mg(x)$, $\{X^{(t)}\}$ no es geométricamente ergódica.

Demostración:

Si $f(x) \leq Mg(x)$ se cumple, entonces el kernel de transición satisface

$$\begin{aligned} K(x,x') &\geq g(x')\,\text{mín}\left\{\frac{f(x')g(x)}{f(x)g(x')}, 1\right\} \\ &= \text{mín}\left\{f(x')\frac{g(x)}{f(x)}, g(x')\right\} \geq \frac{f(x')}{M} \end{aligned}$$

Por tanto, el conjunto $\mathcal{E}$ es pequeño y la cadena es uniformemente ergódica.

Además,

$$\begin{aligned} \|K(x,\cdot) - f\|_{TV} &= 2\sup_A \left|\int_A [K(x,y) - f(y)]dy\right| \\ &= 2\int_{\{y:f(y)\geq K(x,y)\}} [f(y) - K(x,y)]dy \\ &\leq 2\left(1 - \frac{1}{M}\right)\int_{\{y:f(y)\geq K(x,y)\}} f(y)dy \\ &\leq 2\left(1 - \frac{1}{M}\right) \end{aligned}$$

Pero también,

$$\int_A [K^2(x,y) - f(y)]dy = \int_{\mathcal{E}} \left(\int_A [K(u,y) - f(y)]dy \right) [K(x,u) - f(u)]du$$

Con un argumento como el anterior, se tiene que

$$\|K^2(x,\cdot) - f\|_{TV} \leq 2\left(1 - \frac{1}{M}\right)^2$$

En general, se tiene la relación de recursión

$$\int_A [K^{n+1}(x,y) - f(y)]dy = \int_{\mathcal{E}} \left(\int_A [K^n(x,y) - f(y)]dy \right) [K(x,u) - f(u)]du$$

Si $f(x) > Mg(x)$, entonces los conjuntos

$$D_n := \left\{ x : \frac{f(x)}{g(x)} \geq n \right\}$$

satisfacen que para cualquier $n \in \mathbb{N}^+$, $\mathbb{P}_f(D_n) > 0$. Entonces,

$$\begin{aligned} P(x,\{x\}) &= 1 - \mathbb{E}_g \left[\text{mín} \left\{ \frac{f(Y)g(x)}{g(Y)f(x)}, 1 \right\} \right] \\ &= 1 - \mathbb{P}_g \left(\frac{f(Y)}{g(Y)} \geq \frac{f(x)}{g(x)} \right) - \mathbb{E}_g \left[\frac{f(Y)g(x)}{g(Y)f(x)} I_{\{\frac{f(Y)}{g(Y)} < \frac{f(x)}{g(x)}\}} \right] \\ &\geq 1 - \mathbb{P}_g \left(\frac{f(Y)}{g(Y)} \geq n \right) - \frac{g(x)}{f(x)} \geq 1 - \frac{2}{n} \end{aligned}$$

Pero por la desigualdad de Markov,

$$\mathbb{P}_g \left(\frac{f(Y)}{g(Y)} \geq n \right) \leq \mathbb{E}_g \left[\frac{f(Y)}{g(Y)} \right] = \frac{1}{n}$$

Finalmente, considérese un conjunto C tal que $D_n \cap C^c \neq \varnothing$, para n suficientemente grande. Sea $x_0 \in D_n \cap C^c$. El tiempo de regreso a C, τ_C, satisface

$$\mathbb{P}_{x_0}(\tau_C > k) \geq \left(1 - \frac{2}{n}\right)^k$$

por tanto, el radio de convergencia en la serie (en κ) $\mathbb{E}_{x_0}(\kappa^{\tau_C})$ es más pequeño que $\frac{n}{n-2}$ para cualquier $n \in \mathbb{N}^+$. Es decir, $\{X^{(t)}\}$ no puede ser geométricamente ergódica. □

Esta clase particular de algoritmos Metropolis-Hastings sugiere una comparación natural con los métodos Aceptación- Rechazo ya que cada pareja (f, g) satisface $f(x) \leq Mg(x)$ (el requisito fundamental para usar el algoritmo Aceptación-Rechazo).

Teorema 1.6.5.

Si para todo $x \in sop(f)$, $f(x) \leq Mg(x)$, entonces la probabilidad de aceptación esperada en el Algoritmo (1.6.5) es al menos $1/M$ cuando la cadena es estacionaria.

Demostración:

El valor esperado de la probabilidad de aceptación es

$$\begin{aligned}
\mathbb{E}\left[\text{mín}\left\{\frac{f(Y_t)g(X^{(t)})}{f(X^{(t)})g(Y_t)}, 1\right\}\right] &= \int I_{\{\frac{f(y)g(x)}{g(y)f(x)}\geq 1\}} f(x)g(y)dxdy \\
&+ \int \frac{f(y)g(x)}{g(y)f(x)} I_{\{\frac{f(y)g(x)}{g(y)f(x)}\geq 1\}} f(x)g(y)dxdy \\
&= 2\int I_{\{\frac{f(y)g(x)}{g(y)f(x)}\geq 1\}} f(x)g(y)dxdy \\
&\geq 2\int I_{\{\frac{f(y)}{g(y)}\geq \frac{f(x)}{g(x)}\}} f(x)\frac{f(y)}{M}dxdy \\
&= \frac{2}{M}\mathbb{P}\left(\frac{f(X_1)}{g(X_1)} \geq \frac{f(X_2)}{g(X_2)}\right) = \frac{1}{M}
\end{aligned}$$

ya que en la última probabilidad X_1 y X_2 se generan de f. □

Por tanto, el algoritmo Metropolis-Hastings independiente (1.6.5) es más eficiente que el algoritmo de Aceptación-Rechazo ya que en promedio acepta más valores propuestos. Sin embargo, medir esta eficiencia en términos de varianzas de estimadores empíricos puede ser complicado. Por esto, para fines de este trabajo bastará con la comparación vía este número promedio de aceptaciones.

Como una pequeña disgresión, la etiqueta de "Hastings" para el algoritmo (1.6.5) es un poco inapropiada ya que Hastings (1970) consideró el algoritmo (1.6.4) en general, utilizando caminatas aleatorias (como se verá justo a continuación) en vez de distribuciones independientes. También, Hastings (1970) propuso una justificación teórica de estos métodos para cadenas Markov con espacios de estados finitos basado en la representación finita de los reales con fines computacionales. Sin embargo, una justificación completa de esta

discretización física necesita tomar en cuenta el efecto de la aproximación en el análisis entero. En particular, se necesite revisar que la elección computacional de la aproximación discreta a la distribución continua no tenga efecto en la distribución estacionaria resultante o en la irreductibilidad de la cadena. Como Hastings (1970) no estudió a detalle (se concentró más bien en detalles computacionales de simulación) se centrará la atención en las propiedades teóricas de estos algoritmos sin considerar la representación finita de números con computadoras y suponiendo que se tienen generadores de números pseudo-aleatorios suficientemente aceptables.

1.6.3.2. Caminatas aleatorias

Un segundo método natural para la construcción práctica de un algoritmo Metropolis-Hastings es tomar en cuenta el valor previamente simulado para generar el siguiente valor.

Ya que en el algoritmo (1.6.4) se permite que g dependa del estado actual, $X^{(t)}$, una primera elección a considerar es simular Y_t de acuerdo a

$$Y_t = X^{(t)} + \varepsilon_t$$

donde ε_t es una perturbación aleatoria con distribución g, independiente de $X^{(t)}$. En términos del algoritmo (1.6.4), $q(y|x)$ ahora es de la forma $g(y-x)$. La cadena de Markov asociada con q es una caminata aleatoria sobre $\mathcal{E}$.

Las hipótesis de los resultados de convergencia se cumplen en este caso particular. Por el Lema (1.6.2), si g es positiva en una vecindad de 0 entonces la cadena $\{X^{(t)}\}$ es f-irreducible y aperiódica, por tanto es ergódica. Las distribuciones más comunes desde esta perspectiva son distribuciones uniformes sobre esferas centradas en el origen o distribuciones estándar como la normal y la t Student (algunas de estas distribuciones necesitan ser escaladas para que efectivamente sean distribuciones de probabilidad). Nótese que la elección de una función simétrica g hace que el Algoritmo (1.6.4) (en la forma originalmente propuesta por Metropolis *et al.* (1953)) se reduzca a:

Algoritmo 1.6.6. *(Algoritmo Metropolis-Hastings de caminata aleatoria)*

Dado $x^{(t)}$,

1. Generar $Y_t \sim g(y - x^{(t)})$.
2. Hacer
$$X^{(t+1)} = \begin{cases} Y_t & \text{con probabilidad mín}\{\frac{f(Y_t)}{f(x^{(t)})}, 1\}; \\ x^{(t)} & \text{en otro caso} \end{cases}$$

Aún cuando el Metropolis-Hastings de caminata aleatoria es muy simple y tiene características bastante naturales, éste no tiene propiedades de ergodicidad uniforme. Sin embargo, se obtendrán condiciones suficientes y necesarias para garantizar la ergodicidad geométrica. Mengersen & Tweedie (1996) propusieron una condición basada en la log-concavidad de f en las colas, i.e. si existe $\alpha > 0$ y x_1 tal que

$$Log(f(x)) - Log(f(y)) \geq \alpha|y - x| \tag{1.6}$$

para $y < x < -x_1$ ó $x_1 < x < y$.

Teorema 1.6.6.

Considerése una densidad simétrica f log-cóncava con una constante asociada α en (1.6) para x suficientemente grande. Si la densidad g es positiva y simétrica, la cadena $X^{(t)}$ de (1.6.6) es geométricamente ergódica. Si f no es simétrica, una condición suficiente para la ergodicidad geométrica es que $g(t)$ esté acotada por $b \cdot exp(-\alpha|t|)$ para una constante b suficientemente grande.

La demostración de este resultado utiliza a la que se conoce como función de deriva $V(x)=e^{\alpha|x|/2}$ y verificar la condición de deriva geométrica de la forma

$$\Delta V(x) \leq -\lambda V(x) + bI_{[-x^*,x^*]}(x)$$

para alguna cota adecuada x^*. Mengersen & Tweedie (1996) probaron que esta condición sobre g es también necesaria en el sentido de que si $\{X^{(t)}\}$ es geométricamente ergódica, entonces existe $s > 0$ tal que

$$\int e^{s|x|} f(x)dx < \infty$$

Tierney (1994) propuso una modificación de los algoritmos anteriores con una densidad propuesta de la forma $g(y - a - b(x - a))$, i.e.

$$y_t = a + b(x^{(t)} - a) + z_t, \quad z_t \sim g$$

esta representación auto-regresiva se puede ver como un intermedio entre la versión independiente ($b = 0$) y la versión de caminata aleatoria ($b = 1$) del algoritmo Metropolis-Hastings. Si $b < 0$, $X^{(t)}$ y $X^{(t+1)}$ están correlacionadas negativamente y esto facilita visitar la superficie de f si se selecciona adecuadamente el punto (de simetría) a. Hastings (1970) tomó como alternativa de la distribución uniforme sobre $[x^{(t)} - \delta, x^{(t)} + \delta]$ a la distribución uniforme sobre $[-x^{(t)} - \delta, -x^{(t)} + \delta]$. La convergencia del promedio empírico a 0 es muy

rápida en este caso, sin embargo, la elección de 0 como punto de simetría es crucial y requiere información a priori de la distribución f. En general, a y b se pueden calibrar durante las primeras iteraciones.

1.6.3.3. ARMS: Un algoritmo Metropolis-Hastings general

El algoritmo ARS es un método de Aceptación-Rechazo para densidades log-cóncavas y se puede generalizar al metodo ARMS (por sus siglas en inglés *Adaptive Rejection Metropolis Sampling*) siguiendo la metodología que aplicaron Gilks *et al.* (1995). Esta generalización se aplica para simular de densidades arbitrarias)no necesariamente log-concavas como lo hace el algoritmo ARS) simplemente adaptando el ARS para densidades f que no son log-cóncavas. El algoritmo se ajusta progesivamente a una función g, que juega el papel de pseudo-cubierta de la densidad f. En general, está función g no acota superiormente a f, pero la incorporación de un paso Metropolis-Hastings justifica el procedimiento.

Sea $h(x) = Log(f_1(x))$ con f_1 proporcional a la densidad f. Para una muestra $S_n := \{x_i : i \in \{0, 1, \ldots, n+1\}\}$ y $y = L_{i,i+1}(x)$ las ecuaciones de las rectas que unen a $(x_i, h(x_i))$ y $(x_{i+1}, h(x_{i+1}))$. Para $x_i \leq x < x_{i+1}$, considérese

$$\tilde{h}_n(x) = \text{máx}\{L_{i,i+1}(x), \text{mín}\{L_{i-1,i}(x), L_{i+1,i+2}(x)\}\}$$

con

$$\tilde{h}_n(x) = \begin{cases} L_{0,1}(x) & \text{si } x < x_0 \\ \text{máx}\{L_{0,1}(x), L_{1,2}(x)\} & \text{si } x_0 \leq x < x_1 \\ \text{máx}\{L_{n,n+1}(x), L_{n-1,n}(x)\} & \text{si } x_n \leq x < x_{n+1} \\ L_{n,n+1}(x) & \text{si } x \geq x_{n+1} \end{cases}$$

La distribución propuesta resultante es $g_n(x) \propto \exp\{\tilde{h}_n(x)\}$. El algoritmo ARMS está basado en g_n y se puede descomponer en dos partes, un primer paso de un Aceptación-Rechazo estándar para la simulación de la distribución instrumental

$$\psi_n(x) \propto \text{mín}\{f_1(x), \exp\{\tilde{h}_n(x)\}\}$$

basada en g_n y una segunda parte que consiste en simular de un procedimiento de Metropolis-Hastings:

Algoritmo 1.6.7. *(Metropolis-Hastings ARMS)*

1. Simular Y de $g_n(y)$ y $U \sim Unif(0,1)$ hasta que

$$U \leq \frac{f_1(Y)}{\exp\{\tilde{h}_n(Y)\}}$$

2. Generar $V \sim Unif(0,1)$ y hacer

$$X^{(t+1)} = \begin{cases} Y & \text{si } V < \text{mín}\{\frac{f_1(Y)\psi_n(x^{(t)})}{f_1(x^{(t)})\psi_n(Y)}, 1\}; \\ x^{(t)} & \text{en otro caso} \end{cases}$$

De hecho, con el paso de Aceptación-Rechazo se obtiene una variable aleatoria con distribución $\psi_n(x)$ y esto justifica la expresión de la probabilidad de aceptación en el paso Metropolis-Hastings. Nótese que el algoritmo (1.6.7) es un caso particular de los algoritmos de Aceptación-Rechazo considerados por Tierney (1994). La probabilidad en el paso 2 del ARMS se puede reescribir como:

$$\begin{cases} \text{mín}\left\{1, \frac{f_1(Y)\exp\{\tilde{h}_n(x^{(t)})\}}{f_1(x^{(t)})\exp\{\tilde{h}_n(Y)\}}\right\} & \text{si } f_1(Y) > \exp\{\tilde{h}_n(Y)\}; \\ \text{mín}\left\{1, \frac{\exp\{\tilde{h}_n(x^{(t)})\}}{f_1(x^{(t)})}\right\} & \text{en otro caso} \end{cases}$$

lo cual implica una aceptación segura de Y si $f_1(x^{(t)}) > \exp\{\tilde{h}_n(x^{(t)})\}$, i.e. cuando la cota es correcta.

Cada simulación de $Y \sim g_n$ en el paso 1 del Algortimo (1.6.7) proporciona una actualización de S_n en $S_{n+1} = S_n \cup \{y\}$ y por tanto de g_n cuando Y se rechaza. Como en el caso del algoritmo ARS, el conjunto inicial S_n se debe seleccionar de tal forma que g_n sea efectivamente una densidad de probabilidad. Si el soporte de f no está acotado, $L_{0,1}$ debe ser creciente y $L_{n,n+1}$ decreciente.

Como el Algoritmo (1.6.7) parece un caso particular del Metropolis-Hastings independiente, podría parecer que los resultados de convergencia y ergodicidad se cumplen; sin embargo, no hay homogeneidad en el tiempo de la cadena. El kernel de transición, basado en g_n, con cambio en cada paso con probabilidad positiva. Como el estudio de las cadenas es delicado, el Algoritmo (1.6.7) sólo se puede justificar revirtiendo al caso homogéneo, i.e. fijando la función g_n y el conjunto S_n después de un periodo de calentamiento de longitud n_0. La constante n_0 no puede ser fija para que con este periodo de calentamiento (variable) se pueda concluir cuando la aproximación de f_1 y g_n es satisfactoria, por ejemplo cuando

la tasa de rechazo en el paso 1 del Algoritmo (1.6.7) es suficientemente pequeña. Entonces, este algoritmo debe empezar con un paso de inicialización (ó calibración) que adapte los parámetros de g_n para f_1.

El algoritmo ARMS es muy útil cuando es imposible analizar f analíticamente, como por ejemplo en modelos lineales generalizados. De hecho, sólo se necesita calcular f (ó f_1) en unos poco puntos para iniciar el algoritmo, entonces no se necesitará buscar una "buena" densidad g que aproxime a f. Esta característica se debe contrastar con los casos del algoritmo Metropolis-Hastings y de velocidad suficiente como en los casos de las caminatas aleatorias.

1.6.4. Algoritmo Metropolis-Hastings con saltos reversibles

El pequeño catálogo de algoritmos que se han estudiado proporciona métodos de simulación en espacios de dimensión fija, pero no se pueden utilizar cuando la dimensión es otro parámetro a estimar (y con los cuales se está haciendo simulación). Tales problemas se presentan naturalmente en situaciones de elección de modelo en los cuales para cada modelo hay una dimensión diferente y éstos se estan comparando. Otros ejemplos pueden ocurrie en regresión, en la estimación del número de componentes de una mezcla, el orden de un modelo de series de tiempo $ARMA(p,q)$ e $INARMA(p,q)$, ó el número de puntos de cambio en una sucesión estacionaria por pedazos.

Considérese una distribución con densidad

$$f(k,\theta^{(k)}) = f(\theta^{(k)}|k)p(k)$$

donde $k \in \mathbb{N}^+$ y $\theta^{(k)} \in \Xi_k$, i.e. la dimensión de $\theta^{(k)}$ cambia cuando k cambia. La función $p(k)$ es una densidad con respecto a la medida de conteo sobre $\mathbb{N}$ y $f(\theta^{(k)}|k)$ es una densidad con respecto a la medida de Lebesgue sobre Ξ_k. La densidad $f(k,\theta^{(k)}) = f(\theta^{(k)}|k)p(k)$ es una densidad con respecto a la medida de la suma de espacios

$$\Xi = \bigoplus \Xi_k$$

El problema en la implementación de un algoritmo de Cadena de Markov Monte-Carlo es que se necesita poder moverse de un submodelo $\mathcal{H}_k = \{k\} \times \Xi_k$ otro submodelo $\mathcal{H}_{k'}$, donde $k \neq k'$ (es particularmente difícil cuando $k < k'$). Si se hace $x = (k, \theta^{(k)})$ un solución que se propuso en Green (1995) basada en un kernel de transición reversible, K, i.e. un kernel que satisface

$$\int_A \int_B K(x,dy)\pi(x)dx = \int_B \int_A K(y,dx)\pi(y)dy$$

para alguna densidad invariante π. Si q_m es una medida transición al submodelo m y ρ_m la probabilidad de aceptar el salto, el kernel se puede descomponer como

$$K(x,B) = \sum_{m=1}^{\infty} \int_B \rho_m(x,y') q_m(x,dy') + \omega(x) I_B(x)$$

donde

$$\omega(x) := 1 - \sum_m q_m(x,\mathcal{H}), \quad \mathcal{H} = \bigcup_k \mathcal{H}_k$$

es la probabilidad de que no haya salto entre dimensiones. Generalmente, los saltos están limitados a movimientos entre modelos con dimensiones similares a la dimensión de $\mathcal{H}_k$, posiblemente incluyendo $\mathcal{H}_k$. La definición de ρ_m (y la verificación de la suposición de reversibilidad) depende de la siguiente restricción: La medida conjunta $\pi(dx)q_m(x,dy)$ debe ser absolutamente continua con respecto a la medida simétrica $\xi_m(x,dy)\pi(dx)$ sobre $\mathcal{H}^2$. Si $g_m(x,y)$ es la densidad de $q_m(x,dy)\pi(dx)$ y ρ_m se escribe en la forma usual

$$\rho_m = \text{mín}\left\{1, \frac{g_m(y,x)}{g_m(x,y)}\right\}$$

entonces por la simetría de la medida ξ_m se garantiza la reversibilidad

$$\begin{aligned}
\int_A \int_B \rho_m(x,y) q_m(x,dy)\pi(dx) &= \int_A \int_B \rho_m(x,y) g_m(x,y) \xi_m(dx,dy) \\
&= \int_A \int_B \rho_m(y,x) g_m(y,x) \xi_m(dy,dx) \\
&= \int_A \int_B \rho_m(y,x) q_m(y,dx)\pi(dy)
\end{aligned}$$

pues por construcción $\rho_m(x,y)g_m(x,y) = \rho_m(y,x)g_m(y,x)$.

La principal dificultad de esta metodología consiste en determinar la medida ξ_m, dada la restricción de simetría. Si los saltos se pueden descomponer en movimientos entre sólo dos modelos $\mathcal{H}_{k_1}$ y $\mathcal{H}_{k_2}$, la idea de Green (1995) consistió en complementar $\mathcal{H}_{k_1}$ y $\mathcal{H}_{k_2}$ y hacer una biyección entre ambos complementos. Por ejemplo, si $dim(\mathcal{H}_{k_1}) > dim(\mathcal{H}_{k_2})$ y si el movimiento de $\mathcal{H}_{k_1}$ a $\mathcal{H}_{k_2}$ se puede representar por una transformación determinista de $\theta^{(k_1)}$, i.e. $\theta^{(k_2)} = T(\theta^{(k_1)})$; Green (1995) impuso una condición de correspondencia entre dimensiones que consiste en que el movimiento opuesto de $\mathcal{H}_{k_2}$ a $\mathcal{H}_{k_1}$ se concentra en la curva

$$\{\theta^{(k_1)} : \theta^{(k_2)} = T(\theta^{(k_1)})\}$$

En el caso general, si $\theta^{(k_1)}$ se completa a $(\theta^{(k_1)}, u_1)$ y $\theta^{(k_2)}$ se completa a $(\theta^{(k_2)}, u_2)$ de tal forma que el mapeo entre $(\theta^{(k_1)}, u_1)$ y $(\theta^{(k_2)}, u_2)$ sea inyectivo, por ejemplo

$$(\theta^{(k_2)}, u_2) = T(\theta^{(k_1)}, u_1),$$

la probabilidad de aceptar un movimiento de $\mathcal{H}_{k_1}$ a $\mathcal{H}_{k_2}$ es

$$\text{mín}\left\{1, \frac{af(k_2, \theta^{(k_2)})\pi_{12}g_2(u_2)}{f(k_1, \theta^{(k_1)})\pi_{21}g_1(u_1)} \left|\frac{\partial T(\theta^{(k_1)}, u_1)}{\partial(\theta^{(k_1)}, u_1)}\right|\right\}$$

que involucra a las probabilidad π_{ij} de seleccionar $\mathcal{H}_{k_j}$ en vez de $\mathcal{H}_{k_i}$, y g_i, la densidad de u_i.

La construcción de la transformación para la correspondencia entre las dimensiones puede ser difícil; incluso se puede decir que es una desventaja del método de saltos reversibles. En particular, las trasnformación T debe ser derivable. Además, la libertad total del principio de saltos reversibles acerca de la elección de los saltos (que a veces se conocen como movimientos separación-unión) genera una ineficiencia potencial y requiere "afinar" aquellos pasos que sean muy demandantes.

Green (1995) menciona que la densidad f no se necesita normalizar, pero las diferentes densidades componentes $f(\theta^{(l)}|k)$ se deben conocer (excepto quizá por una constante multiplicativa).

En conclusión, nótese que la complejidad del método se incrementa por las restricciones de reversibilidad que se imponen sobre los saltos, aunque éstas no son necesarias para la covergencia de los algoritmos Metropolis-Hastings. En particular, se puede dejar de considerar la simetría entre los movimientos.

1.7. El Gibbs-sampler

En las secciones anteriores se estudiaron métodos de simulación que se pueden llamar "genéricos" ya que sólo requieren poca información acerca de la distribución de la que se desea simular. Sin embargo, los algoritmos Metropolis-Hastings pueden ser más eficientes cuando se consideran algunas particularidades de f. En esta dirección, ahora se estudiará el Gibbs-sampling, cuya eficiencia está muy relacionada con las propiedades de f. Esto se debe a que la elección de la distribución instrumental esencialmente se reduce a un número finito de posibilidades.

El Gibbs-sampler es muy popular desde su descripción en el artículo de Geman & Geman (1984) en el estudio de modelos de procesamiento de imágenes. Sin embargo, la génesis del método se debe a los trabajos de Metropolis, Rosenbluth, Rosenbluth, Teller & Teller (1953) y Hastings (1970); notando el interés estadístico desde los trabajos de Gelfand &

Smith (1990).

El Gibbs-sampler es una técnica para generar variables aleatorias indirectamente de una distribución (marginal) sin tener que calcular la densidad.

Mediante el Gibbs-sampler se pueden evitar algunas dificultades de cálculo haciendo algunas operaciones más sencillos.

1.7.1. Motivación

Supóngase que dada una densidad conjunta $f(x, y_1, \ldots, y_p)$ y se está interesado en obtener características de la densidad marginal

$$f(x) = \int \ldots \int f(x, y_1, \ldots, y_p) dy_1, \ldots, dy_p$$

tales como media y varianza. Sin embargo, calcular alguna ó algunas integrales puede ser complicado de forma analítica.

En vez de calcular ó aproximar $f(x)$ directamente, el Gibbs-sampler genera una muestra $x_1, \ldots, x_m$ de $f(x)$ sin que se requiera a f. Y como ya se mencionó, al simular una muestra grande se puede aproximar con cierto grado de exactitud a la media, la varianza ó alguna otra característica numérica de la variable de interés.

Es importante notar que efectivamente el resultado final de cualquier cálculo basado en simulaciones es precisamente una característica numérica de la población. Por ejemplo, para calcular la media de $X \sim f$, se puede usar el hecho de que

$$\lim_{m\to\infty} \frac{1}{m} \sum_{i=1}^{m} X_i = \int_{-\infty}^{\infty} x f(x) dx = \mathbb{E}(X)$$

Por tanto, tomando m suficientemente grande, se puede obtener cualquier característica poblacional (incluso la densidad misma) con cierto grado de exactitud.

Para entender el algoritmo Gibbs-sampler primero se estudiará el caso de dos variables. Se empieza con una pareja de variables aleatorias (X, Y), el algoritmo de Gibbs genera una muestra de $f(x)$ a través de $f(x|y)$ y $f(y|x)$. Esto se hace generando lo que se conoce como una "sucesión de Gibbs" de variables aleatorias

$$Y_0', X_0', Y_1', X_1', \ldots, Y_k', X_k'$$

Se especifica el valor $Y_0' = y_0'$ y el resto se obtiene iterativamente generado valores

$$X_j' \sim f(x|Y_j' = y_j')$$

$$Y'_{j+1} \sim f(y|X'_j = x'_j)$$

Bajo condiciones suficientemente razonables, la distribución de X'_k converge a $f(x)$ (la marginal de X) cuando $k \to \infty$. Por tanto, para k suficientemente grande $X'_k = x'_k$ efectivamente es un punto muestral de $f(x)$.

La convergencia (en distribución) de la sucesión de Gibbs se puede aprovechar de varias formas para obtener una muestra aproximada de f. Por ejemplo, Gelfand & Smith (1990) sugieren generar m sucesiones de Gibbs independientes de longitud k y usar el valor final de X'_k de cada sucesión. Si se selecciona k suficientemente grande, se obtiene aproximadamente una muestra independiente e identicamente distribuida de f.

Ejemplo 1.7.1.

Considérese la siguiente distribución para (X, Y)

$$f(x,y) \propto \binom{n}{x} y^{x+\alpha-1}(1-y)^{n-x+\beta-1} I_{\{0,1,\ldots,n\}}(x) I_{(0,1)}(y)$$

Supóngase que se tiene interés en calcular alguna característica de la distribución marginal $f(x)$ de X. El Gibbs-sampler genera una muestra de esta marginal de la siguiente manera: Primero nótese que $f(x|y)$ es una densidad binomial con parámetros (n, y) y que $f(y|x)$ es una densidad $Beta(x+\alpha, n-x+\beta)$ (este ejemplo es un tanto artificial debido a que ya se tiene en mente la famosa conjugación beta-binomial).

Si se construye la sucesión de Gibbs a la densidad conjunta se puede generar una sucesión $X_1, X_2, \ldots, X_m$ de f y se usa ésta para estimar la característica numérica deseada. Nótese que en este ejemplo no se necesita forzosamente un esquema de Gibbs ya que se puede obtener analíticamente que

$$f(x) = \binom{n}{x} \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)} \frac{\Gamma(x+\alpha)\Gamma(n-x+\beta)}{\Gamma(\alpha+\beta+n)} I_{\{0,1,\ldots,n\}}(x)$$

la distribución beta-binomial.

Aquí, las características numéricas de $f(x)$ se pueden obtener directamente de esta ecuación ya sea analíticamente ó generando una muestra de la marginal sin necesidad de las condicionales.

En este ejemplo, se quiere hacer notar que el Gibbs-sampler no se necesita forzosamente en una situación bivariada en la que la distribución conjunta $f(x, y)$ se puede calcular (pues $f(x) = \frac{f(x,y)}{f(y|x)}$). Sin embargo, en el siguiente ejemplo se analizará una situación en la que no se pueden calcular $f(x, y)$, $f(x)$ y $f(y)$. ▽

Ejemplo 1.7.2.

Supóngase que X y Y tienen distribuciones condicionales que son exponenciales restringidas al intervalo $(0, B)$, i.e.

$$f(x|y) \propto ye^{-yx} I_{(0,B)}(x)$$
$$f(y|x) \propto xe^{-xy} I_{(0,B)}(y)$$

donde $B \in (0, \infty)$ es una constante conocida. La restricción al intervalo $(0, B)$ asegura que la marginal $f(x)$ exista. La forma de esta marginal no es fácil de calcular, sin embargo mediante el Gibbs-sampler y con las condicionales se pueden obtener algunas características numéricas de $f(x)$.

El algoritmo de Gibbs se puede usar para estimar la densidad misma promediando las densidades condicionales de cada sucesión de Gibbs. Los valores de $X'_k = x'_k$ son una realización de $X_1, \ldots, X_m \sim f(x)$ y los valores de $X'_k = x'_k$ son una realización de $Y_1, \ldots, Y_m \sim f(y)$. El promedio de las densidades condicionales $f(x|Y'_k = y'_k)$ será aproximación de $f(x)$, y se puede estimar $f(x)$ mediante

$$\hat{f}(x) = \frac{1}{m} \sum_{i=1}^{m} f(x|y_i)$$

donde $y_1, \ldots, y_m$ son simulaciones de Y que se obtienen a partir de la sucesión de Gibbs. Una pequeña justificación de lo anterior es que

$$\mathbb{E}(f(x|Y)) = \int f(x|y) f(y) dy = f(x)$$

∇

1.7.2. Descripción del algoritmo

Supóngase que para algún $p > 1$, el vector aleatorio $\mathbf{X}$ se puede escribir como $\mathbf{X} = (X_1, \ldots, X_p)$ donde $X_1, \ldots, X_p$ pueden ser unidimensionales ó multidimensionales. Además, supóngase que se puede simular de las correspondientes densidades condicionales $f_1, \ldots, f_p$, i.e. se puede simular

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

para $i = 1, 2, \ldots, p$. El algoritmo de Gibbs asociado (ó *Gibbs-sampler*) está dado por la siguiente transición de $X^{(t)}$ a $X^{(t+1)}$.

Algoritmo 1.7.1. *(Gibbs-sampler)*

Dado $\mathbf{x}^{(t)} = (x_1^{(t)}, \ldots, x_p^{(t)})$, generar

1. $X_1^{(t+1)} \sim f_1(x_1 | x_2^{(t)}, \ldots, x_p^{(t)})$
2. $X_2^{(t+1)} \sim f_2(x_2 | x_1^{(t+1)}, x_3^{(t)}, \ldots, x_p^{(t)})$

   $\vdots$

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

Las densidades $f_1, \ldots, f_p$ se conocen como *condicionales completas* y una característica particular del Gibbs-sampler es que son las únicas densidades que el algoritmo utiliza para la simulación. Por tanto, aún en problemas de dimensión alta, todas las simulaciones pueden ser univariadas (que generalmente es una ventaja).

Ejemplo 1.7.3. *(Gibbs-sampler bivariado)*

El Gibbs sampler bivariado genera una cadena de Markov de una distribución conjunta de la siguiente forma: si dos variables aleatorias X y Y tienen densidad conjunta $f(x, y)$, cuyas correspondientes densidades condicionales son $f_{Y|X}$ y $f_{X|Y}$, el Gibbs sampler bivariado genera una cadena de Markov (X_t, Y_t) de acuerdo al siguiente algoritmo:

Algoritmo 1.7.2. *(Algoritmo Gibbs-sampler bivariado)*

Tomar $X_0 = x_0$
Para $t = 1, 2, \ldots$ generar

1. $Y_t \sim f_{Y|X}(\cdot | x_{t-1})$
2. $X_t \sim f_{X|Y}(\cdot | y_t)$

El algoritmo Gibbs sampler bivariado es sencillo de implementar siempre y cuando simular de ambas condicionales sea factible, ya que si se tiene a $f(x, y)$ en forma explícita (excepto quizá por alguna constante de normalización) entonces se puede obtener explícitamente $f_{Y|X}$ y $f_{X|Y}$. Si no es posible simular de estas condicionales se utilizarán algunas aproximaciones que se verán más adelante. Es fácil ver por qué si (X_t, Y_t) tiene densidad conjunta f, entonces (X_{t+1}, Y_{t+1}) también tienen densidad f; esto se debe a que en ambos

pasos de la t-esima iteración se simula de las condicionales. La convergencia de la cadena de Markov (y por tanto del algoritmo) se asegura, a menos que los soportes de las condicionales no sean conexos. Las sucesiones $\{(X_t, Y_t)\}$, $\{X_t\}$ y $\{Y_t\}$ son cadenas de Markov. Por ejemplo, el kernel de transición de $\{X_t\}$ es

$$K(x, x^*) = \int f_{Y|X}(y|x) f_{X|Y}(x|y) dy$$

con densidad invariante $f_X(\cdot)$. ▽

Aunque formalmente el Gibbs sampler es una caso particular del algoritmo Metropolis-Hastings, este algoritmo tiene algunas características diferentes

(i) La tasa de aceptación del Gibbs-sampler es uniformemente igual a 1. Por tanto, todo valor simulado se acepta. Esto también significa que la convergencia de este algoritmo se debe tratar con técnicas diferentes a las que se utilizaron para el algoritmo Metropolis-Hastings.

(ii) El uso del Gibbs-sampler implica limitaciones en la elección de las distribuciones instrumentales y requiere conocimiento previo de algunas propiedades analíticas o probabilísticas de f

(iii) Por construcción, el Gibbs-sampler es multidimensional. Aunque algunas componentes del vector simulado puedan ser artificiales para el problema de interés ó innecesarios para el problema de inferencia, la construcción es de al menos dos dimensiones.

(iv) En principio, el Gibbs-sampler no aplica en problemas donde el número de parámetros no sea fijo y esto se debe a la obvia falta de irreductibilidad de la cadena.

1.7.3. Completación

El algoritmo Gibbs-sampler se puede generalizar mediante "demarginalización" ó construcción de la completación, y de hecho ya se mencionó durante la motivación.

Definición 1.7.1. *(Completación)*

Dada una densidad de probabilidad f, a una densidad g que satisface

$$\int g(x, z) dz = f(x)$$

se le conoce como completación de f

La densidad g se selecciona de tal forma que las condicionales completas de g sean fáciles de simular y el algoritmo de Gibbs sea implementado sobre g en vez de sobre f. Para $p > 1$ sea $y = (x, z)$ y denótese a las densidades completas de $g(y) = g(y_1, \ldots, y_p)$ por

$$Y_1|y_2, \ldots, y_p \sim g_1(y_1|y_2, \ldots, y_p)$$

$$Y_2|y_1, y_3, \ldots, y_p \sim g_2(y_2|y_1, y_3, \ldots, y_p)$$

$$\vdots$$

$$Y_p|y_1, \ldots, y_{p-1} \sim g_p(y_p|y_1, \ldots, y_{p-1})$$

Entoces, el movimiento de $Y^{(t)}$ a $Y^{(t+1)}$ se define como sigue

Algoritmo 1.7.3. *(Gibbs-sampler de completación)*

Dado $\mathbf{x}^{(t)} = (x_1^{(t)}, \ldots, x_p^{(t)})$, generar

1. $Y_1^{(t+1)} \sim g_1(y_1|y_2^{(t)}, \ldots, y_p^{(t)})$
2. $Y_2^{(t+1)} \sim g_2(y_2|y_1^{(t+1)}, y_3^{(t)}, \ldots, y_p^{(t)})$

   $\vdots$

p. $Y_p^{(t+1)} \sim g_p(y_p|y_1^{(t+1)}, \ldots, y_{p-1}^{(t+1)})$

La completación de una densidad f en una densidad g tal que f sea la densidad marginal de g es una de las primeras apariciones históricas del muestreo de Gibbs en estadística y se considera como introducción del *data-augmentation* de Tanner & Wong (1987).

En aquellos casos en los que la completación parece necesaria (por ejemplo cuando toda distribución condicional asociada con f no es explícita ó cuando f es unidimensional) existe un número infinito de densidades de las cuales f es marginal. No se discutirá acerca de esta elección en términos de optimalidad pues hay pocos resultados prácticos con respecto a esto. Además, en general, existe una completación natural de f en g conocido como modelo de datos faltantes con similaridades de esta metodología con el algoritmo EM para maximizar una verosimilitud con datos faltantes.

En principio, el Gibbs-sampler no requiere que la completación de f en g y de x en $y = (x, z)$ esté relacionada con el problema de interés. De hecho, hay algunas aplicaciones

en las que el vector z no tiene interpretación estadística.

Ahora, se verá una versión más general del Gibbs-sampler, que se conoce como *slice sampler* (Wakefield *et al.* (1991), Besag & Green (1993), Damien & Walker (1996), Higdon (1996), Damien *et al.* (1999) y Tierney & Mira (1999)). Éste se aplica a la mayoría de las distribuciones y se basa en la simulación de variables aleatorias uniformes con parámetros particulares. Si $f(\theta)$ se puede escribir como el producto

$$\prod_{i=1}^{k} f_i(\theta)$$

donde $f_1, \ldots, f_k$ son funciones positivas, no necesariamente densidades. f se puede completar (ó demarginalizar) como

$$\prod_{i=1}^{k} I_{\{0 \leq \omega_i \leq f_i(\theta)\}}$$

obteniendo el siguiente algoritmo de Gibbs:

Algoritmo 1.7.4. *(Slice-sampler)*

Simular

1. $\omega_1^{(t+1)} \sim Unif[0, f_1(\theta^{(t)})]$

   $\vdots$

k. $\omega_k^{(t+1)} \sim Unif[0, f_k(\theta^{(t)})]$

k+1. $\theta^{(t+1)}$ con distribución uniforme en el conjunto $A^{(t+1)}$ donde

$$A^{(t+1)} = \{y : f_i(y) \geq \omega_i^{(t+1)}, i = 1, \ldots, k\}$$

Roberts & Rosenthal (1998) estudiaron el *slice-sampler* (1.7.4) y probaron que generalmente tiene buenas propiedades teóricas. En particular, ellos probaron que el *slice-sampler* permite conjuntos pequeños (ver Apéndice A) y que la cadena asociada sea geométricamente ergódica cuando la densidad f está acotada. Tierney & Mira (1999) establecieron la ergodicidad uniforme en el caso $k = 1$.

En la práctica puede haber varios problemas. Si k crece, se puede complicar significativamente encontrar (determinar) al conjunto $A^{(t+1)}$. También, cuando k es grande, se está forzando al Gibbs-sampler a que trabaje en un espacio altamente restrictivo (y no lineal) y puede alentar la convergencia.

1.7.4. Propiedades de convergencia

Ahora se estudiarán las propiedades de convergencia del Gibbs-sampler. Primero, se mostrará que la mayoría de los Gibbs-samplers satisfacen condiciones mínimas para garantizar la ergodicidad. Como ya se mencionó, el Gibbs-sampler es un caso particular del algoritmo Metropolis-Hastings, sin embargo este hecho no será de mucha ayuda para cuestiones de convergencia (por el momento).

Se trabajará con el Gibbs-sampler general (1.7.3) y se mostrará que la cadena de Markov $\{Y^{(t)}\}$ converge a la distribución g y que la subcadena $\{X^{(t)}\}$ converge a la distribución f. Es muy importante notar que por construcción $\{Y^{(t)}\}$ es una cadena de Markov, sin embargo la subcadena $\{X^{(t)}\}$ generalmente no lo es (un caso muy usado en el que sí lo es, es el algoritmo *Data Augmentation* que se estudiará más adelante).

Teorema 1.7.1.

En el Gibbs-sampler (1.7.3), si $\{Y^{(t)}\}$ es ergódica, entonces la distribución g es una distribución estacionaria para dicha cadena y f es la distribución límite de la subcadena $\{X^{(t)}\}$.

Demostración:

El kernel de la cadena $\{Y^{(t)}\}$ es el producto

$$K(\mathbf{y}, \mathbf{y}') = g_1(y_1'|y_2, \ldots, y_p) g_2(y_2'|y_1', y_3, \ldots, y_p) \cdots g_p(y_p'|y_1', \ldots, y_{p-1}') \tag{1.7}$$

para el vector $\mathbf{y} = (y_1, y_2, \ldots, y_p)$, sea $g^i(y_1, \ldots, y_{i-1}, y_{i+1}, \ldots, y_p)$ la densidad marginal del vector $\mathbf{y}$ sin y_i (i.e se obtuvo integrando con respecto a y_i). Si $\mathbf{Y} \sim g$ y A es medible bajo la medida dominante, entonces

$$\begin{aligned}
\mathbb{P}(\mathbf{Y}' \in A) &= \int_A I_A(\mathbf{y}')K(\mathbf{y},\mathbf{y}')g(\mathbf{y})d\mathbf{y}'d\mathbf{y} \\
&= \int_A I_A(\mathbf{y}')[g_1(y_1'|y_2,\ldots,y_p)\cdot\ldots\cdot g_p(y_p'|y_1',\ldots,y_{p-1}')] \\
&\times \; [g_1(y_1|y_2,\ldots,y_p)g^1(y_2,\ldots,y_p)]dy_1\ldots dy_p dy_1'\ldots dy_p' \\
&= \int_A I_A(\mathbf{y}')g_2(y_2'|y_1',\ldots,y_p)\cdot\ldots\cdot g_p(y_p'|y_1',\ldots,y_{p-1}') \\
&\times \; g(y_1',y_2,\ldots,y_p)dy_2\ldots dy_p dy_1'\ldots dy_p'
\end{aligned}$$

donde se combinó $g_1(y_1'|y_2,\ldots,y_{p-1})g^1(y_2,\ldots,y_p) = g(y_1',y_2,\ldots,y_p)$ para y_1. Ahora, se escribirá $g(y_2'|y_1',y_3,\ldots,y_p)g^2(y_1',y_3,\ldots,y_p) = g(y_1',y_2,\ldots,y_p)$ para y_2, obteniendo

$$\begin{aligned}
\mathbb{P}(\mathbf{Y}' \in A) &= \int_A I_A(\mathbf{y}')g_3(y_3'|y_1',y_2',\ldots,y_p)\cdot\ldots\cdot g_p(y_p'|y_1',\ldots,y_{p-1}') \\
&\times \; g(y_1',y_2',y_3,\ldots,y_p)dy_3\ldots dy_p dy_1'\ldots dy_p'
\end{aligned}$$

continuando de esta manera se obtiene

$$\mathbb{P}(\mathbf{Y}' \in A) = \int_A g(y_1',y_p')d\mathbf{y}$$

probando que g es la distribución estacionaria. Por tanto, el Teorema (5.1) del Apéndice A implica que $Y^{(t)}$ se distribuye asintóticamente como g y $X^{(t)}$ se distribuye asintóticamente como f (integrando sucesivamente). □

Una prueba más corta se puede hacer utilizando la estacionariedad de g después de cada paso en los p pasos de (1.7.3).

Ahora, se analizará la irreductibilidad de la cadena de Markov generada por el Gibbs-sampler.

Una condición suficiente para la irreductibilidad de la cadena de Markov generada por el Gibbs-sampler es la que propuso Besag (1974) y se establece a continuación.

Definición 1.7.2. *(Condición de positividad)*

Sea $(Y_1,Y_2,\ldots,Y_p) \sim g(y_1,y_2,\ldots,y_p)$, donde $g^{(i)}$ denota la distribución marginal de Y_i. Se dice que g satisface la condición de positividad cuando $g^{(1)}(y_1),g^{(2)}(y_2),\ldots,g^{(p)}(y_p) > 0$ implica que $g(y_1,y_2,\ldots,y_p) > 0$

Observación 1.7.1.

Cuando se cumple la condición de positividad el soporte de g es el producto cartesiano de los soportes de los $g^{(i)}$'s. Además, las distribuciones condicionales no reducirán el rango de los posibles valores de Y_i cuando se comparan con g. En este caso, dos Borelianos arbitrarios del soporte se pueden juntar en una sola iteración de (1.7.3).

Teorema 1.7.2.

Para el Gibbs-sampler (1.7.3), si la densidad g satisface la condición de positividad, entonces la cadena generada es irreducible.

Sin embargo, la condición del Teorema anterior es difícil de verificar. Tierney (1994) estableció una condición un poco más fácil de verificar.

Lema 1.7.1.

Si el kernel de transición asociado con (1.7.3) es absolutamente continuo con respecto a la medida dominante, entonces la cadena resultante es Harris-recurrente.

En este caso, la condición sobre el kernel de transición proporciona una cadena irreducible y por tanto se cumple la Harris-recurrencia. Una vez que se determina la irreductibilidad de la cadena asociada con (1.7.3), se pueden analizar propiedades de convergencia más complejas.

Esta condición de continuidad absoluta sobre el kernel (1.7) se satisface en la mayoría de las descomposiciones. Sin embargo, si uno de los i pasos ($i \in \{1, \ldots, p\}$) se reemplaza por una simulación del algoritmo Metropolis-Hastings (como en el caso de los algoritmos híbridos que se comentarán más adelante) se pierde la continuidad absoluta y es necesario estudiar las propiedades de recursión de la cadena ó bien instroducir un paso de simulación adicional que garantice la Harris-recurrencia.

De la propiedad de Harris-recurrencia se puede establecer el siguiente resultado.

Teorema 1.7.3.

Supóngase que el Kernel de transición de la cadena de Markov $\{Y^{(t)}\}$ del Gibbs-sampler es absolutamente continuo con respecto a la medida dominante.

(i) Si $h_1, h_2 \in L^1(g)$ tales que $\int h_2(y)dg(y) \neq 0$, entonces

$$\lim_{n\to\infty} \frac{\sum_{t=1}^{T} h_1(Y^{(t)})}{\sum_{t=1}^{T} h_2(Y^{(t)})} = \frac{\int h_1(y)dg(y)}{\int h_2(y)dg(y)}, \quad \text{para casi toda } g$$

(ii) Si además, $\{Y^{(t)}\}$ es aperódica, entonces

$$\lim_{n\to\infty} \int \|K^n(y,\cdot)\nu(dx) - g\|_{TV} = 0$$

para cualquier distribución inicial ν

También hay una condición más general que permite movimientos en una vecindad mínima.

Lema 1.7.2.

Sean $\mathbf{y} = (y_1, \ldots, y_p)$ y $\mathbf{y'} = (y'_1, \ldots, y'_p)$ y supóngase que existe $\delta > 0$ tal que para $\mathbf{y}, \mathbf{y'} \in sop(g)$, $|\mathbf{y} - \mathbf{y'}| < \delta$ y

$$g_i(y_i|y_1, \ldots, y_{i-1}, y'_{i+1}, \ldots, y'_p) > 0, \quad i = 1, 2, \ldots, p$$

Si existe $\delta' < \delta$ para el cual casi cualquier pareja $(\mathbf{y}, \mathbf{y'})$ en el soporte de g se pueda conectar mediante una sucesión finita de bolas de radios δ' con medida (estacionaria) positiva de la intersección de dos bolas consecutivas, entonces la cadena generada por (1.7.3) es irreducible y aperiódica.

Observación 1.7.2.

La formulación tan laboriosa de esta condición se necesita adaptar a aquellas situaciones en las que el soporte de g no es conexo. ▽

Corolario 1.7.1.

Las conlusiones del Teorema (1.7.3) se satisfacen si las cadena de Markov $\{Y^{(t)}\}$ del Gibbs-sampler tiene densidades condicionales $g_i(y_i|y_1,\ldots,y_{i-1},y'_{i+1},\ldots,y'_p)$ que satisfagan las suposiciones del Lema (1.7.2)

1.7.5. Gibbs-sampler y Metropolis-Hastings

Ahora se estudiará la relación exacta en los algoritmos Gibbs-sampler y Metropolis-Hastings, considerando (1.7.3) como la composición de p kernel Markovianos. Como se mencionó antes, esta representación no es suficiente para establecer la irreductibilidad ya que cada uno de estos algoritmos Metropolis-Hastings por separado no genera una cadena de Markov (dado que éste sólo modifica una componente de y). De hecho, estos kernel nunca son irreducibles ya que están restringidas a subespacios de dimensiones menores.

Teorema 1.7.4.

El método de Gibbs-sampler (1.7.3) es equivalente a la composición de p algoritmos Metropolis-Hastings con probabilidades de aceptación uniformemente igual a 1.

Demostración:

Si se escribe (1.7.3) como la composición de p algoritmos "elementales" que correspondan a los p pasos de simulación de la distribución condicional, es suficiente demostrar que cada uno de estos algoritmos tiene una probabilidad de aceptación igual a 1. Para $i \in \{1,\ldots,p\}$ la distribución instrumental en el i-ésimo paso está dada por

$$q_i(y'|y) = g_i(y'_i|y_i,\ldots,y_{i-1},y_{i+1},y_p)I_{(y_i,\ldots,y_{i-1},y_{i+1},y_p)}(y'_i,\ldots,y'_{i-1},y'_{i+1},y'_p)$$

Y el cociente que define la probabilidad $\rho(y,y')$ es

$$\begin{aligned}\frac{g(y')q_i(y|y')}{g(y)q_i(y'|y)} &= \frac{g(y')g_i(y_i|y_1,\ldots,y_{i-1},y_{y+1},y_p)}{g(y)g_i(y'_i|y_1,\ldots,y_{i-1},y_{y+1},y_p)} \\ &= \frac{g'_i(y_i|y_1,\ldots,y_{i-1},y_{y+1},y_p)g_i(y_i|y_1,\ldots,y_{i-1},y_{y+1},y_p)}{g_i(y_i|y_1,\ldots,y_{i-1},y_{y+1},y_p)g_i(y'_i|y_1,\ldots,y_{i-1},y_{y+1},y_p)} \\ &= 1\end{aligned}$$

□

Nótese que el Gibbs-sampler no es el único algoritmo MCMC que tiene esta propiedad. En la sección (1.5) se mencionó que todo kernel q asociado con una cadena de Markov reversible con distribución invariante g tiene una probabilidad de aceptación uniformemente igual a 1. Sin embargo, la probabilidad de aceptación global para el vector $(y_1 \ldots, y_p)$ generalmente es diferente de 1 y un procesamiento de (1.7.3) como un algoritmo Metropolis-Hastings con la que se obtiene una probabilidad de rechazo positiva.

En este análisis global de (1.7.3) es posible rechazar algunos valores generados por la sucesión generada por los pasos $1, \ldots, p$ de este algoritmo. Esta versión del algoritmo Metropolis-Hastings se puede comparar con el Gibbs-sampler original.

Una de las características más sorprendentes del Gibbs-sampler es que las distribuciones condicionales $g_i(y_i|y_{j\neq q})$ contienen suficiente información para generar una muestra de g. Visto como problema de optimización esta metodología se puede ver como la maximización de una función objetivo sucesivamente en cada dirección de una base dada. Es bien sabido que este método no necesariamente lleva a un máximo global pero puede terminar en un punto de equilibrio.

Entonces es notable que las distribuciones condicionales completas g_i resuman perfectamente la densidad conjunta $g(y)$, sin embargo, el conjunto de distribuciones marginales $g^{(i)}$ no hacen este resumen adecuadamente. Antes de analizar el fenómeno subyacente primero se considerará el caso particular $p = 2$, donde (1.7.3) se implementa a partir de dos distribuciones condicionales $g_1(y_1|y_2)$ y $g_2(y_2|y_1)$. El siguiente resultado muestra cómo se puede obtener directamente la densidad conjunta a partir de las densidades condicionales.

Teorema 1.7.5.

La distribución conjunta asociada con las densidades condicionales g_1 y g_2 tiene la densidad

$$g(y_1, y_2) = \frac{g_2(y_2|y_1)}{\int g_2(v|y_1)/g_1(y_1|v)dv}$$

Demostración:

Recuérdese que $g^{(1)}$ y $g^{(2)}$ denotan las densidades marginales de Y_1 y Y_2, respectivamente. Entonces, se puede escribir

$$g(y_1, y_2) = g_1(y_1|y_2)g^{(2)}(y_2) = g_2(y_2|y_1)g^{(1)}(y_1)$$

y por tanto, para cualquier $y_1 \in \mathcal{Y}$

$$g^{(2)}(y_2) = \frac{g_2(y_2|y_1)}{g_1(y_1|y_2)} g^{(1)}(y_1)$$

Como $g^{(2)}(\cdot)$ es densidad, $\int g^{(2)}(y_2)dy_2 = 1$ y por tanto

$$\int \frac{g_2(y_2|y_1)}{g_1(y_1|y_2)} dy_2 = \frac{1}{g^{(1)}(y_1)}$$

□

Por supuesto que la obtención de $g(y_1, y_2)$ requiere la existencia y cálculo de la integral

$$\int \frac{g_2(v|y_1)}{g_1(y_1|v)} dv$$

Entonces, este resultado exhibe claramente la característica fundamental de que $g_1(y_1|y_2)$ y $g_2(y_2|y_1)$ son suficientemente informativas para recuperar la densidad conjunta. Nótese que en este teorema (y en lo que queda de esta sección) se está haciendo la suposición implícita de que la densidad conjunta $g(y_1, y_2)$ existe.

El caso general para enteros $p > 2$ es análogo al que se hizo en el desarrollo anterior y lleva a lo que se conoce como el Teorema de Hammersley-Clifford.

Teorema 1.7.6. *(Teorema de Hammersley-Clifford)*

Bajo la condición de positividad, la distribución conjunta g satisface

$$g(y_1, \ldots, y_p) \propto \prod_{j=1}^{p} \frac{g(y_{\ell_j}|y_{\ell_1}, \ldots, y_{\ell_{j-1}}, y'_{\ell_{j+1}}, \ldots, y'_{\ell_p})}{g(y'_{\ell_j}|y_{\ell_1}, \ldots, y_{\ell_{j-1}}, y'_{\ell_{j+1}}, \ldots, y'_{\ell_p})}$$

para toda permutación ℓ sobre $\{1, 2, \ldots, p\}$ y todo $y' \in \mathcal{Y}$

Este resultado es interesante desde el punto de vista de métodos de cadenas de Markov Monte-Carlo ya que muestra que la densidad g se conoce (excepto quizá por una constante multiplicativa) cuando las densidades $g_i(y_i|y_{j\neq i})$ se conocen. Por tanto, es posible comparar al Gibbs-sampler con alternativas como el método de Aceptación-Rechazo ó algoritmos Metropolis-Hastings. Esto también implica que el Gibbs-sampler nunca es el único método disponible para simular de g y que siempre es posible incorporar pasos Metropolis-Hastings en el Gibbs.

1.7.5.1. Estructuras jerárquicas

Ahora se analizará brevemente un estructura para la cual el Gibbs-sampler se puede adaptar bien: los modelos jerárquicos. Éstas son estructuras en las cuales (por razones computacionales o estructurales) f se puede descomponer (para $I \geq 0$) como

$$f(x) = \int f_1(x|z_1) f_2(z_1|z_2) \cdots f_I(z_I|z_{I+1}) f_{I+1}(z_{I+1}) dz_1 \ldots dz_{I+1}$$

Dichos modelos aparecen naturalemente en análisis Bayesiano de modelos complejos en los que se tiene información previa o la variabilidad de las observaciones requiere considerar varios niveles de distribuciones *a priori*.

Para algunos casos de modelos jerárquicos se puede probar que las cadenas de Gibbs asociadas son uniformemente ergódicas. Generalmente se tiene que hacer un análisis caso por caso para probar esto y aparecen cadenas $\{X^{(t)}\}$ y $\{Z^{(t)}\}$ tales que $X^{(t+1)}$ se genera condicionalmente sobre $z^{(t)}$ y $Z^{(t+1)}$ se genera condicionalmente sobre $x^{(t+1)}$. Esto corresponde a la técnica que se conoce como *data augmentation* que se describirá más adelante. Este caso particular del Gibbs-sampler simplifica el estudio de las propiedades probabilísticas del algoritmo ya que cada cadena se analiza por separado sin perder relación con las demás. Sin embargo, esta característica ya no se cumple cuando $p \geq 3$ ya que las subcadenas $\{Y_1^{(t)}\}, \ldots, \{Y_p^{(t)}\}$ no satisfacen la propiedad de Markov aunque $\{\mathbf{Y}^{(t)}\}$ sí. Por tanto, no hay kernel de transición asociado con $\{Y_i^{(t)}\}$ y el estudio de la ergodicidad uniforme sólo puede considerar a un vector de $p-1$ de las p componentes de $\{\mathbf{Y}^{(t)}\}$ ya que en general, el kernel original puede no estar uniformemente acotado por abajo. Si $\mathbf{z}_1 := (y_2, \ldots, y_p)$, la transición de $\mathbf{z}_1$ a $\mathbf{z}_1'$ tiene el kernel

$$K_1(\mathbf{z}_1, \mathbf{z}_1') = \int_{\mathcal{Y}_1} g_1(y_1'|\mathbf{z}_1) g_2(y_2'|y_1', y_3, \ldots, y_p) \cdots g_p(y_p' 1|y_1', y_2', \ldots, y_{p-1}') dy_1'$$

aunque se pueda acotar uniformemente a K_1, generalmente es imposible lograr una minorización uniforme de K.

1.7.6. El Gibbs-sampler de dos etapas

Un problema común cuando se estudian propiedades probabilisticas del Gibbs-sampler es que complicado hacer un análisis por componentes, por lo que se prefiere un estudio de manera global. Como se mencionó en la sección anterior, un análisis por componentes del Algoritmo (1.7.3) muy pocas veces es interesante (factible) ya que los correspondientes algoritmos Metropolis-Hastings no son irreducibles y las correspondientes sucesiones $\{Y_i^{(t)}\}$ no son cadenas de Markov. Sin embargo, en el caso particular $p = 2$ sí se puede hacer una descomposición en dos pasos y por tanto se puede estudiar de forma más detallada el Gibbs-sampler.

1.7.6.1. Estructuras duales de probabilidad

El método de *data augmentation* es un caso particular de algoritmo (1.7.3), donde f es completada en g y x en $y = (y_1, y_2)$ de tal forma que ambas densidades condicionales, $g_1(y_1|y_2)$ y $g_2(y_2|y_1)$ sean conocidas (recuérdese que y_1 y y_2 pueden ser vectores o escalares).

Algoritmo 1.7.5. *(Algoritmo Data Augmentation)*

Dado $y^{(t)}$,

1. Simular $Y_1^{(t+1)} \sim g(y_1|y_2^{(t)})$
2. Simular $Y_2^{(t+1)} \sim g(y_2|y_1^{(t+1)})$

Este método lo propusieron Tanner & Wong (1987) al margen del Gibbs-sampler, aunque quizá si inspirado en el algoritmo EM de Dempster *et al.* (1977) y métodos de restauración estocástica. Desde esta perspectiva de estructuras de datos faltantes, cuando la distribución se puede simular se puede escribir como la distribución marginal

$$\pi(\theta|x) \propto \int f(x, z|\theta)\pi(\theta)dz$$

Las aplicaciones de este método no están restringidas a problemas de datos faltantes.

Lema 1.7.3.

Las sucesiones $\{Y_1^{(t)}\}$ y $\{Y_2^{(t)}\}$ generadas por el Algoritmo (1.7.5) es una cadena de Markov con distribuciones estacionarias

$$g^{(2)}(y_1) = \int g(y_1, y_2)dy_2, \quad g^{(1)}(y_2) = \int g(y_1, y_2)dy_1,$$

respectivamente. Si la restricción de positividad sobre g se cumple, ambas cadenas son fuertemente irreducibles.

Demostración:

Nótese que $\{Y_1^{(t)}\}$ está asociada con el kernel de transición

$$K_1(y_1, y_1') = \int g_1(y_1'|y_2)g_2(y_2|y_1)dy_2$$

por tanto, $Y_1^{(t)}$ se simula condicionalmente sólo sobre $y_1^{(t-1)}$ y es independiente de $Y_1^{(t-2)}, Y_1^{(t-3)}, \ldots$ (dado $y_1^{(t-1)}$). Por tanto, $\{Y_1^{(t)}\}$ es una cadena de Markov.

Bajo la restricción de positividad, si g es positiva, entonces $g_1(y_1'|y_2)$ es positiva sobre el soporte proyectado de g y todo Boreliano de $\mathcal{Y}_1$ se puede visitar en una sola iteración de (1.7.5) y por tanto se cumple la irreductibilidad fuerte.

Análogamente se prueba lo correspondiente para $\{Y_2^{(t)}\}$ □

Este razonamiento tan sencillo prueba que si sólo la cadena $\{Y_1^{(t)}\}$ es de interés y se cumple la condición $g_1(y_1'|y_2) > 0$ para cada pareja (Y_1', Y_2), entonces la irreductibilidad se cumple. La cadena dual $\{Y_2^{(t)}\}$ se puede utilizar para establecer algunas propiedades probabilísticas de $\{Y_1^{(t)}\}$.

1.7.6.2. Cadenas reversibles y entrelazadas

Liu et al. (1994) mostraron que el Gibbs-sampler de dos etapas, ó equivalentemente el *data augmentation*, tiene una propiedad estructural muy fuerte. Las dos cadenas, la cadena de interés $\{X^{(t)}\}$ y la cadena instrumental (ó dual), $\{Y^{(t)}\}$, satisface una propiedad de dualidad conocida como *entrelazamiento*. Esta propiedad es característica del *data augmentation* y las únicas cadenas de Gibbs-sampling que satisfacen esta propiedad son las asociadas con el kernel de tipo *data augmentation*.

Definición 1.7.3. *(Propiedad de entrelazamiento)*

Se dice que dos cadenas de Markov $\{X^{(t)}\}$ y $\{Y^{(t)}\}$ son conjugadas con la propiedad de entrelazamiento si

(i) Dado $Y^{(t)}$, $X^{(t)}$ y $X^{(t+1)}$ son independientes.

(ii) Dado $X^{(t)}$, $Y^{(t-1)}$ y $Y^{(t)}$ son independientes.

(iii) $(X^{(t)}, Y^{(t-1)})$ y $(X^{(t)}, Y^{(t)})$ son idénticamente distribuidos bajo estacionariedad.

En la mayoría los casos cuando se tienen cadenas entrelazadas hay más interés en una de las cadenas. La propiedad de entrelazamiento siempre se cumple en el método de *data augmentation*. Nótese que la cadena $(X^{(t)}, Y^{(t)})$ no necesariamente es reversible en el tiempo.

Lema 1.7.4.

Cada una de las cadenas $\{X^{(t)}\}$ y $\{Y^{(t)}\}$ generadas por el algoritmo de data augmentation es reversible y la cadena $(X^{(t)}, Y^{(t)})$ satisface la propiedad de entrelazamiento.

Demostración:

La reversibilidad de cada cadena (propiedad que es independiente del entrelazamiento) es inmediata de la condición de balance detallado. Considérese (X_0, Y_0) que se distribuye como g (la distribución estacionaria), con distribuciones marginales $g^{(2)}(x)$ y $g^{(1)}(y)$. Entonces, si

$$K_1(x_0, x_1) = \int g_2(y_0|x_0) g_1(x_1|y_0) dy_0$$

es el kernel de transición de $\{X^{(t)}\}$,

$$\begin{aligned} g^{(2)}(x_0) K_1(x_0, x_1) &= g^{(2)}(x_0) \int g_2(y_0|x_0) g_1(x_1|y_0) dy_0 \\ &= \int g(x_0, y_0) g_1(x_1|y_0) dy_0, \quad \text{pues } g^{(2)}(x_0) = \int g(x_0, y_0) dy_0 \\ &= \int g(x_0, y_0) \frac{g_2(y_0|x_1) g^{(2)}(x_1)}{g^{(1)}(y_0)}, \quad \text{pues } g_1(x_1|y_0) = \frac{g_2(y_0|x_1) g^{(2)}(x_1)}{g^{(1)}(y_0)} \\ &= g^{(2)}(x_1) K_1(x_1, x_0) \end{aligned}$$

Por tanto $\{X^{(t)}\}$ es reversible. Análogamente se prueba que $\{Y^{(t)}\}$, i.e. si K_2 es el kernel asociado, se satisface la condición de balance detallado

$$g^{(1)}(y_0) K_2(y_0, y_1) = g^{(1)}(y_1) K_2(y_1, y_0)$$

La construcción de cada cadena satisface (i) y (ii) de la definición de cadenas entrelazadas. Para probar la propiedad (iii) nótese que la densidad conjunta e X_0 y Y_0 es

$$\mathbb{P}(X_0 \leq x, Y_0 \leq y) = \int_{-\infty}^{x} \int_{-\infty}^{y} g(v, u) du dv$$

y la densidad conjunta de X_1 y Y_0 es

$$\mathbb{P}(X_1 \leq x, Y_1 \leq y) = \int_{-\infty}^{x} \int_{-\infty}^{y} g_1(v|u) g^{(1)}(u) du dv$$

ya que

$$\begin{aligned} g_1(v|u)g^{(1)}(u) &= \int g_1(v|u)g(v',u)dv' \\ &= \int g_1(v|u)g_1(v'|u)g^{(1)}(u)dv' \\ &= g(v,u) \end{aligned}$$

$$\therefore \{X^{(t)}\},\ \{Y^{(t)}\} \text{ están entrelazadas}$$

□

Para garantizar que $(X^{(t)}, Y^{(t)})$ es reversible se debe agregar un paso adicional en (1.7.5).

Algoritmo 1.7.6. *(Data augmentation reversible)*

Dado $y_2^{(t)}$,

1. Simular $W \sim g_1(w|y_2^{(t)})$
2. Simular $Y_2^{(t+1)} \sim g_2(y_2|w)$
3. Simular $Y_1^{(t+1)} \sim g_1(y_1|y_2^{(t+1)})$

En este caso,

$$\begin{aligned} (Y_1, Y_2, Y_1', Y_2') &\sim g(y_1,y_2)\left(\int g_1(w|y_2)g_2(y_2'|w)dw\right)g_1(y_1'|y_2') \\ &= g(y_1',y_2')\left(\int \frac{g_2(y_2'|w)}{g_1(y_2')}g(w,y_2)dw\right)g_1(y_1|y_2) \\ &= g(y_1',y_2')\left(\int g_1(w|y_2')g_2(y_2|w)dw\right)g_1(y_1|y_2) \end{aligned}$$

se distribuye como (Y_1', Y_2', Y_1, Y_2).

El Gibbs-sampler de (1.7.3) también se conoce como Gibbs-sampler de *escaneo sistemático* ó de *barrido sistemático* ya que la trayectoria de la iteración se sigue sistemáticamente

en una dirección. Tal metodología lleva a una cadena de Markov no reversible, pero se puede construir un Gibbs-sampler reversible con escaneo simétrico. El siguiente algoritmo garantiza la reversibilidad de la cadena $\{Y^{(t)}\}$.

Algoritmo 1.7.7. *(Gibbs-sampler reversible)*

Dado $(y_2^{(t)}, \ldots, y_p^{(t)})$, generar

1. $Y_1^* \sim g_1(y_1|y_2^{(t)}, \ldots, y_p^{(t)})$
2. $Y_2^* \sim g_2(y_2|y_1^*, y_3^{(t)}, \ldots, y_p^{(t)})$

$\vdots$

p-1. $Y_{p-1}^* \sim g_{p-1}(y_{p-1}|y_1^*, \ldots, y_{p-2}^*, y_p^{(t)})$

p. $Y_p^{(t+1)} \sim g_p(y_p|y_1^*, \ldots, y_{p-1}^*)$

p+1. $Y_{p-1}^{(t+1)} \sim g_{p-1}(y_{p-1}|y_1^*, \ldots, y_{p-2}^*, y_p^{(t+1)})$

$\vdots$

2p+1. $Y_1^{(t+1)} \sim g_1(y_1|y_2^{(t+1)}, \ldots, y_p^{(t+1)})$

Liu *et al.* (1995) propusieron una alternativa a (1.7.7) que se conoce como Gibbs-sampler con *escaneo aleatorio* en donde la simulación de cada componente de y se hace en orden aleatorio en cada transición. Esta modificación de (1.7.3) genera una cadena reversible con distribución estacionaria g y se usa cada valor que se simula.

Algoritmo 1.7.8. *(Gibbs-sampler de barrido aleatorio)*

1. Generar una permutación $\sigma \in \mathcal{G}_p$
2. Simular $Y_{\sigma_1}^{(t+1)} \sim g_{\sigma_1}(y_{\sigma_1}|y_j^{(t)}, j \neq \sigma_1)$

$\vdots$

p+1. Simular $Y_{\sigma_p}^{(t+1)} \sim g_{\sigma_p}(y_{\sigma_p}|y_j^{(t+1)}, j \neq \sigma_p)$

Este algoritmo mejora al algoritmo (1.7.7) en el sentido de que usa sólo una simulación de las dos.

1.7.6.3. Covarianza monótona y Rao-Blackwellización

Algunas veces, condicionando a un subconjunto de las variables simuladas se puede mejorar significativamente al estimador empírico en términos de varianza mediante un "reciclado" de las variables rechazadas. Desde su génesis, el Gibbs-sampler no permite este tipo de reciclaje ya que se acepta todo valor simulado. Sin embargo, Gelfand & Smith (1990) propusieron un tipo de condicionamiento, que sí "reocupa" variables, conocido como *Rao-Blackwellización*. La *Rao-Blackwellización* se basa en la identidad

$$g_1(y_1) = \int g_1(y_1|Y_2, \ldots, y_p) g^1(y_2, \ldots, y_p) dy_2 \ldots dy_p$$

y reemplaza

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

con

$$\delta_{rb} = \frac{1}{T} \sum_{t=1}^{T} \mathbb{E}(h(Y_1)|y_2^{(t)}, \ldots, y_p^{(t)})$$

ambos estimadores convergen a $\mathbb{E}(h(Y_1))$ y bajo la distribución estacionaria también son insesgados. Con la identidad de varianza iterada, $Var(U) = Var(\mathbb{E}(U|V)) + \mathbb{E}(Var(U|V))$ se tiene que

$$Var[\mathbb{E}(h(Y_1)|Y_2^{(t)}, \ldots, Y_p^{(t)})] \leq Var(h(Y_1))$$

Esto permitió que Gelfand & Smith (1990) sugirieran usar δ_{rb} en vez de δ_0. No obstante, la desigualdad anterior no es suficiente para concluir acerca de la dominiación de δ_{rb} cuando se compara con δ_0 pues falla al comparar la correlación entre las $Y^{(t)}$'s. La dominación de δ_0 por δ_{rb} se puede establecer en muy pocos casos; Liu *et al.* (1994) mostraron que esta dominación sí ocurre en el caso de *data augmentation*.

La idea es dar un criterio de dominación. Para esto, se necesitarán algunos resultados previos.

Lema 1.7.5.

Si $h \in \mathcal{L}^2(g(2))$ y si $\{X^{(t)}\}$ se entrelaza con $\{Y^{(t)}\}$, entonces

$$Cov(h(Y^{(1)}), h(Y^{(2)})) = Var(\mathbb{E}(h(Y)|X))$$

Demostración:

Suponiendo que $\mathbb{E}_{g_2}(h(\theta)) = 0$,

$$\begin{aligned} Cov(h(Y^{(1)}), h(Y^{(2)})) &= \mathbb{E}\left[h(Y^{(1)})h(Y^{(2)})\right] \\ &= \mathbb{E}\left[\mathbb{E}(h(Y^{(1)})|X^{(2)})\mathbb{E}(h(Y^{(2)})|X^{(2)})\right] \\ &= \mathbb{E}\left[\mathbb{E}^2(h(Y^{(1)})|X^{(2)})\right] = Var(\mathbb{E}(h(Y)|X)) \end{aligned}$$

donde la segunda igualdad se sigue del Teorema de esperanza iterada y la independencia del la propiedad de entrelazamiento. La última igualdad simplemente es una aplicación del la propiedad (iii) de la definición de cadenas entrelazadas. □

Proposición 1.7.1.

Si $\{Y^{(t)}\}$ es una cadena de Markov con la propiedad de entrelazamiento, entonces las covarianzas

$$Cov(h(Y^{(1)}), h(Y^{(t)}))$$

son positivas y decrecientes como función de t para todo $h \in \mathcal{L}^2(g(2))$.

Demostración:

Por el lema anterior, inductivamente se tiene que

$$\begin{aligned} Cov(h(Y^{(1)}), h(Y^{(t)})) &= \mathbb{E}\left[\mathbb{E}(h(Y)|X^{(2)})\mathbb{E}(h(Y)|X^{(t)})\right] \\ &= Var\left(\mathbb{E}\left[\cdots\mathbb{E}(\mathbb{E}(h(Y)|X)|Y)\cdots\right]\right) \end{aligned}$$

donde el último término involucra $t-1$ esperanzas condicionales alternadamente en X y en Y. La desigualdad en t se sigue directamente de la desigualdad sobre las esperanzas

condicionales $Var[\mathbb{E}(h(Y_1)|Y_2^{(t)},\ldots,Y_p^{(t)})] \leq Var(h(Y_1))$ □

Teorema 1.7.7.

Si $\{X^{(t)}\}$ y $\{Y^{(t)}\}$ son dos cadenas de Markov entrelazadas con distribuciones g_1 y g_2, respectivamente. Entonces el estimador δ_{rb} domina al estimador δ_0 para todo función $h \in \mathcal{L}(g_i)$ $(i = 1, 2)$.

Demostración:

Suponiendo que $\mathbb{E}_{g_2}(h(X)) = 0$ y considerando los estimadores

$$\delta_0 = \frac{1}{T}\sum_{t=1}^{T} h(X^{(t)}), \quad \delta_{rb} = \frac{1}{T}\sum_{t=1}^{T} \mathbb{E}[h(X)|Y^{(t)}]$$

se tiene que

$$\begin{aligned} Var(\delta_0) &= \frac{1}{T^2}\sum_{t,s} Cov(h(X^{(t)}), h(X^{(s)})) \\ &= \frac{1}{T^2}\sum_{t,s} Var\left(\mathbb{E}\left[\cdots \mathbb{E}(h(X)|Y)\cdots\right]\right) \end{aligned}$$

y

$$\begin{aligned} Var(\delta_{rb}) &= \frac{1}{T^2}\sum_{t,s} Cov\left(\mathbb{E}[h(X)|Y^{(t)}], \mathbb{E}[h(X)|Y^{(s)}]\right) \\ &= \frac{1}{T^2}\sum_{t,s} Var\left(\mathbb{E}\left[\cdots \mathbb{E}(\mathbb{E}(h(X)|Y)|X)\cdots\right]\right) \end{aligned}$$

con $|t - t'|$ esperanzas condicionales en la expresión de $Var(\delta_0)$ y $|t - t'| + 1$ esperanzas condicionales en la expresión de $Var(\delta_{rb})$, por tanto es suficiente comparar $Var(\delta_0)$ y $Var(\delta_{rb})$ término a término y finalmente $Var(\delta_0) \geq Var(\delta_{rb})$. □

Una primera pregunta natural en *Rao-Blackwellización* es cuándo habrá una reducción de varianza considerable si el tamaño muestra se incrementa (ó bien el número de iteraciones de Monte-Carlo). Esta idea fue estudiada por Levine (1996) y formuló este problema en términos de la eficiencia relativa asintótica de δ_0 con respecto a su versión *Rao-Blackwellizada*,

$\delta_{rb} = \frac{1}{T}\sum_{t=1}^{T}\mathbb{E}[h(X)|Y^{(t)}]$ en el que las parejas $(X^{(t)}, Y^{(t)})$ se genera a partir de un Gibbs-sampler bivariado. La eficiencia relativa asintótica es el cociente de las varianzas de la distribución límite para los dos estimadores, que están dadas por

$$\sigma^2_{\delta_0} = Var(h(X^{(0)})) + 2\sum_{k=1}^{\infty} Cov\left(h(X^{(0)}), h(X^{(k)})\right)$$

y

$$\sigma^2_{\delta_{rb}} = Var\left(\mathbb{E}[h(X)|Y]\right) + \sum_{k=1}^{\infty} Cov\left(\mathbb{E}[h(X^{(0)})|Y^{(0)}], \mathbb{E}[h(X^{(k)})|Y^{(k)}]\right)$$

Levine (1996) concluyó que $\frac{\sigma^2_{\delta_0}}{\sigma^2_{\delta_{rb}}} \geq 1$ y la igualdad se cumple si y sólo si $Var(h(X)) = 0$.

Liu *et al.* (1995) extendieron la proposición anterior al caso del Gibbs-sampler aleatorio, donde todo paso sólo actualiza una sola componente de y. En (1.7.3) se define una distribución multinomial $\sigma = (\sigma_1, \ldots, \sigma_p)$.

Algoritmo 1.7.9. *(Gibbs-sampler aleatorio)*

Dado $y^{(t)}$

1. Seleccionar una componente $\nu \sim \sigma$
2. Generar $Y_\nu^{(t+1)} \sim g_\nu(y_\nu|y_j^{(t)}, j \neq \nu)$ y tomar $y_j^{(t+1)} y_j^{(t)}$ para $j \neq \nu$

Nótese que aunque (1.7.9) sólo genera una componente de y en cada iteración, la cadena resultante es fuertemente irreducible debido a la elección aleatoria de ν. Está también satisface la siguiente propiedad.

Proposición 1.7.2.

La cadena $\{Y^{(t)}\}$ que genera el algoritmo (1.7.9) tiene la propiedad de que para cualquier función $h \in \mathcal{L}_2(g)$, la covarianza $Cov(h(Y^{(0)}), h(Y^{(t)}))$ es positiva y decreciente en t.

Demostración:

Si se supone que $\mathbb{E}_g(h(Y)) = 0$, entonces

$$\begin{aligned}\mathbb{E}\left[h(Y^{(0)})h(Y^{(t)})\right] &= \sum_{v=1}^{p}\mathbb{E}[\mathbb{E}[h(Y^{(0)})h(Y^{(1)})|\nu=v,(Y_j^{(0)},j\neq v)]]\\ &= \sum_{v=1}^{p}\mathbb{E}[\mathbb{E}^2[h(Y^{(0)})|(Y_j^{(0)},j\neq v)]]\\ &= \mathbb{E}[\mathbb{E}^2[h(Y)|\nu,(Y_j,j\neq\nu)]]\end{aligned}$$

gracias a la reversibilidad de la cadena y la independencia de $Y^{(0)}$ y $Y^{(1)}$ condicionalmente con ν y $(y_j^{(0)}, j\neq\nu)$. Recursivamente se tiene que

$$\mathbb{E}\left[h(Y^{(0)})h(Y^{(t)})\right] = Var\left(E[\cdots\mathbb{E}[\mathbb{E}[h(Y)|\nu,(Y_j,j\neq\nu)]|Y]\cdots]\right)$$

donde el segundo término involucra t esperanzas condicionales sucesivamente en $(\nu,(y_j,j\neq\nu)$ y en Y. □

Otro beneficio considerable de la *Rao-Blackwellización* es un método elegante para la aproximación de densidades de diferentes componentes de y. Como

$$\frac{1}{T}\sum_{t=1}^{T} g_i(y_i|y_j^{(t)}, j\neq i)$$

es insesgado y converge a la densidad marginal $g_i(y_i)$ si éstas densidades condicionales son conocidas (en una forma cerrada) y por tanto es innecesario (ineficiente) utilizar métodos no parámetricos para la estimación de densidades (tal como los métodos de kernel).

1.7.6.4. El principio de dualidad

La *Rao-Blackwellización* muestra una diferencia interesante entre la pespectiva estadística y el uso de la simulación en el sentido de que las aproximaciones que se usan en el estimador no involucran directamente a la cadena de interés. Ahora se analizará este fenómeno desde un punto de vista más fuerte que simplemente mejorar la varianza de algunos estimadores; ahora se tendrá como objetivo estudiar las propiedades de convergencia de la cadena de interés, $\{X^{(t)}\}$, basados en la cadena instrumental $\{Y^{(t)}\}$ aún cuando está no está relacionada con el problema de inferencia. Diebolt & Robert (1993, 1994) se refieren al uso de esta cadena dual como *Principio de dualidad*. Aunque este principio se motivó desde una perspectiva del algoritmo *data augmentation*, éste se extiende a otros métodos de Gibbs ya que es suficiente que la cadena de interés $\{X^{(t)}\}$ se genere condicionalmente a otra cadena $\{Y^{(t)}\}$ (que tiene algunas propiedades probabilísticas interesantes).

Teorema 1.7.8.

Considérese una cadena de Markov $\{X^{(t)}\}$ y una sucesión de variables aleatorias $\{Y^{(t)}\}$ generada a partir de las distribuciones condicionales

$$X^{(t)}|y^{(t)} \sim \pi(x|y^{(t)}), \quad Y^{(t+1)}|x^{(t)}, y^{(t)} \sim f(y|x^{(t)}, y^{(t)}).$$

Si la cadena $\{Y^{(t)}\}$ es ergódica (ó geométricamente o uniformemente ergódica) y $\pi^t_{y^{(0)}}$ denota la distribución de $\{X^{(t)}\}$ asociada con el valor inicial $y^{(0)}$, entonces la norma $\|\pi^t_{y^{(0)}} - \pi\|_{TV}$ se va a cero conforme t se va a infinito (ó bien a cero geométrica o uniformente con tasa acotada).

Demostración:

El kernel de transición es

$$K(y, y') = \int_{\mathcal{X}} \pi(x|y) f(x'|x, y) dx$$

Si $f^t_{y^{(0)}}$ es la densidad marginal de $Y^{(t)}$, entonces la marginal de $X^{(t)}$ es

$$\pi^t_{y^{(0)}}(x) = \int_{\mathcal{Y}} \pi(x|y) f^t_{y^{(0)}}(y) dy$$

y la convergencia de la cadena $\{X^{(t)}\}$ se deba a la convergencia de $\{Y^{(t)}\}$. La variación total de la cadena $\{X^{(t)}\}$ está dada por

$$\begin{aligned}
\|\pi^t_{y^{(0)}} - \pi\|_{TV} &= \frac{1}{2} \int_{\mathcal{Y}} \left| \pi^t_{y^{(0)}}(x) - \pi(x) \right| dx \\
&= \frac{1}{2} \int_{\mathcal{X}} \left| \int_{\mathcal{Y}} \pi(x|y)(f^t_{y^{(0)}}(x) - f(y)) dy \right| dx \\
&\leq \frac{1}{2} \int_{\mathcal{X}\times\mathcal{Y}} \left| f^t_{y^{(0)}}(y) - f(y) \right| \pi(x|y) dx dy \\
&= \|f^t_{y^{(0)}} - f\|_{TV}
\end{aligned}$$

Por tanto, las propiedes de convergencia de $\|f^{t+1}_{y^{(0)}} - f\|_{TV}$ se "transfieren" a $\|\pi^t_{y^{(0)}} - \pi\|_{TV}$. Ambas sucesiones tiene la misma velocidad de convergencia ya que

$$\|f^{t+1}_{y^{(0)}} - f\|_{TV} \leq \|\pi^t_{y^{(0)}} - \pi\|_{TV} \leq \|f^t_{y^{(0)}} - f\|_{TV}$$

□

Cuando la espacio de estados de la cadena $\{Y^{(t)}\}$ es finito, el principio de dualidad es un poco más sencillo.

Teorema 1.7.9.

Si $\{Y^{(t)}\}$ es una cadena de Markov con espacio de estados finito $\mathcal{Z}$ tal que para cualesquiera $(k, x) \in \mathcal{X} \times \mathcal{Y}$

$$\mathbb{P}(Y^{(t+1)} = j | x, y^{(t)}) > 0$$

Entonces la sucesión $\{X^{(t)}\}$ que se obtiene de $\{Y^{(t)}\}$ mediante el kernel de transición $\pi(x|y^{(t)})$ es uniformemente ergódica.

Demostración:

Sea

$$p_{ij} = \int \mathbb{P}(Y^{(t+1)} = j | x, y^{(t)} = i)\pi(x|y^{(t)} = i)dx, \quad i, j \in \mathcal{Y}$$

Si se define la cota inferior $\rho := \text{mín}_{i,j\in\mathcal{Y}}\{p_{ij}\} > 0$, entonces

$$\mathbb{P}(Y^{(t)} = i | y^{(0)} = i) \geq \rho \sum_{k\in\mathcal{Y}} \mathbb{P}(Y^{(t-1)} = k | y^{(0)} = i) = \rho$$

y por tanto, para cualquier $i \in \mathcal{Y}$

$$\sum_{t=1}^{\infty} \mathbb{P}(Y^{(t)} = i | y^{(0)} = i) = \infty$$

Es decir, la cadena es positiva recurrente con distribución límite

$$q_i = \lim_{t\to\infty} \mathbb{P}(Y^{(t)} = i | y^{(0)} = i)$$

Entonces, por el teorema anterior $\{X^{(t)}\}$ es ergódica. Y por la finitud de $\mathcal{Y}$, $\{X^{(t)}\}$ es uniformemente ergódica. □

Observación 1.7.3.

Nótese que este resultado de convergencia no impone restricciones en la transición $\pi(x|y)$, por ejemplo que no necesariemente sea positiva. ▽

Las afirmaciones de los dos teoremas anteriores se complican por el hecho de que $\{X^{(t)}\}$ no necesariamente es una cadena de Markov y por tanto no se pueden aplicar los resultados de ergodicidad y ergodicidad geométrica. Sin embargo, sigue habiendo una distribución límite π para $\{X^{(t)}\}$ que se obtiene a partir de una transformación de la distribución límite de $\{Y^{(t)}\}$ mediante la transición $\pi(x|y)$. En el algoritmo *data augmentation* esta formlación es menos complicada.

Corolario 1.7.2.

Sean $\{X^{(t)}\}$ y $\{Y^{(t)}\}$ cadenas de Markov entrelazadas. Si $\{X^{(t)}\}$ es ergódica (ó bien geométricamente ergódica), entonces $\{Y^{(t)}\}$ es ergódica (ó bien geométricamente ergódica).

Finalmente, se analizará brevemente la tasa de convergencia y se verá que el Principio de dualidad se sigue cumpliendo y la tasa es la misma entre la cadena original y su dual.

Proposición 1.7.3.

Si $\{Y^{(t)}\}$ converge geométricamente en un espacio de estados compacto y con tasa de convergencia ρ, entonces existe $C_h \in \mathbb{R}$ tal que

$$\|\mathbb{E}(h(X^{(t)})|y^{(0)}) - \mathbb{E}^{\pi}(h(X))\| \leq C_h \rho^t$$

para cualquier funcion $h \in L^1(\pi(\cdot|x))$ uniformemente en $y^{(0)} \in \mathcal{Y}$.

Demostración:

Sea $h(x) = (h_1(x), \ldots, h_d(x))$,

$$\begin{aligned}
\|\mathbb{E}(h(X^{(t)})|y^{(0)}) - \mathbb{E}^{\pi}(h(X))\| &= \sum_{i=1}^{d} \left(\mathbb{E}(h_i(X^{(t)})|y^{(0)}) - \mathbb{E}^{\pi}(h_i(X))\right)^2 \\
&= \sum_{i=1}^{d} \left(\int h_i(x)[\pi^t(x|y^{(0)}) - \pi(x)]dx\right)^2 \\
&= \sum_{i=1}^{d} \left(\int h_i(x)\pi(x|y) \int [f^t(y|y^{(0)}) - f(y)]dydx\right)^2 \\
&\leq \sum_{i=1}^{d} \sup_{y\in\mathcal{Y}} \mathbb{E}^2[|h_i(X)||y]4\|f^t_{y^{(0)}} - f\|^2_{TV} \\
&\leq 4d \max_i \sup_{y\in\mathcal{Y}} \mathbb{E}^2[|h_i(X)||y]\|f^t_{y^{(0)}} - f\|^2_{TV}
\end{aligned}$$

□

1.7.7. Gibbs samplers híbridos

1.7.7.1. Comparación con algoritmos Metropolis Hastings

A priori, cuando se compara un método de Gibbs con un algoritmo de Metropolis-Hastings se favorecerá al primero, ya que éste se obtiene de las distribuciones condicionales de la verdadera distribución f, mientras que un kernel Metropolis-Hastings se basa en una aproximación de su distribución f. En particular, el Gibbs-sampler es más claro (desde su construcción) ya que no tiene una "mala" elección de la distribución instrumental y por tanto no permite rechazos.

En esta sección se intentará observar que la disponibilidad y aparente objetividad del Gibbs-sampler no necesariamente justifica su uso. Visto como una combinación de algoritmos Metropolis-Hastings, el Gibbs-sampler está compuesto de subalgoritmos que no son válidos ya que no necesariamente generan cadenas de Markov irreducibles, $\{Y^{(t)}\}$. Sólo una combinación de un número suficiente de algoritmos Metropolis-Hastings permite asegurar la validez del Gibbs-sampler. Esta estructura de composición también es una debilidad del método ya que una descomposición de la distribución conjunta f dado un sistema de coordenadas particular no necesariamente coincide con la forma de f, i.e. una elección incorrecta del sistema de coordenadas puede "atrapar" al correspondiente Gibbs-sampler en una de las componentes conexas del soporte de f. Hills & Smith (1992, 1993) analizaron ejemplos en los que una parametrización incorrecta del modelo incrementa significativamente el tiempo

de convergencia del Gibbs-sampler.

Para mostrar una analogía, recuérdese que cuando una función $v(y_1, \ldots, y_p)$ se maximiza vía una componente a la vez, la solución que se obtiene no siempre es satisactoria ya que ésta puede corresponder a un punto de equilibrio ó a un máximo local de v. Análogamente, la simulación de una sola componente en cada iteración de (1.7.3) restringe las posibles excursiones de la cadena $\{Y^{(t)}\}$ y esto implica que el Gibbs-sampler generalmente convergerán lentamente.

Este defecto intrínseco del Gibbs-sampler lleva a un fenómeno parecido al de convergencia a máximos locales en algoritmos de optimización, que se expresan mediante fuertes atracciones a las modas locales y por tanto es difícil explorar todo el soporte de f.

1.7.7.2. Mezclas y ciclos

Las desventajas entre los algoritmos Metropolis-Hastings son diferentes de las del Gibbs-sampler y generalmente están relacionados con la relación entre f y su distribución instrumental. Además, la libertad que tienen los métodos Metropolis-Hastings a veces permite "resolver" esos inconvenientes mediante la modificación de alguna escala (parámetros o hiper-parámetros).

Comparado con el Gibbs-sampler, un falla de los algoritmos Metropolis-Hastings es que se pasan por alto algunos detalles de la distribución f si la simulación es muy burda. Sin embargo, según Tierney (1994), se pueden aprovechar las ventajas de ambos algoritmos para implementar un metodología híbrida, que utilice algoritmos Metropolis-Hastings y Gibbs-sampler.

Definición 1.7.4.

Un algoritmo MCMC híbrido es un método de Cadena de Markov Monte-Carlo que utiliza simultáneamente pasos de Gibbs-sampler y pasos de Metropolis-Hastings. Si $K_1, \ldots, K_n$ son los kernel correspondientes a estos diferentes pasos y si $(\alpha_1, \ldots, \alpha_n)$ es una distribución de probabilidad, una mezcla de $K_1, \ldots, K_n$ es un algoritmo asociado con el kernel

$$\tilde{K} = \alpha_1 K_1 + \ldots + \alpha_n K_n$$

y un ciclo de $K_1, \ldots, K_n$ es un algoritmo asociado con el kernel

$$K^* = \alpha_1 K_1 \circ \ldots \circ \alpha_n K_n$$

donde "$\circ$" es la composición usual de funciones.

Esta definición es un poco ambigua pues ya se mencionó que el Gibbs-sampler es una composición de varios kernel Metropolis-Hastings, i.e. un ciclo según la definición anterior. Sin embargo, la esta definición se debe entender como un procesamiento de algoritmos MCMC donde la cadena bajo estudio es una subcadena, $\{Y^{(kt)}\}$, de la cadena generada por el algoritmo.

Desde una perspectiva de velocidad de convergencia del Gibbs-sampler, la definición anterior permite considerar modificaciones del algoritmo original en el cual cada m iteraciones, (1.7.3) se reemplaza por un paso Metropolis-Hastings con dispersión más grande; ó alternativamente, en cada iteración, el paso Metropolis-Hastings se selecciona con probabilidad $1/m$. Estas modificaciones son particularmente útiles para evitar "efectos de captura" relacionados con las modas locales de f.

Las metodologías híbridas también son válidas desde un punto de vista teórico, cuando la heterogeneidad de las cadenas generadas por ciclos se quita al considerar sólo las subcadenas $\{Y^{(mt)}\}_t$. Una composición de kernels asociados con distribuciones estacionarias idénticas a f, lleva a un kernel con distribución estacionaria f. La irreductibilidad y la aperiodicidad de $\tilde{K}$ se sigue directamente de la irreductibilidad y aperiodicidad de los kernels K_i. En el caso de un ciclo, ya se dijo que la irreductibilidad de K^* para el Gibbs-sampler no requiere de la irreductibilidad de sus kernels componentes y generalmente se necesita un estudio más profundo del algoritmo. Tierney (1994) propuso condiciones suficientes para la ergodicidad uniforme.

Proposición 1.7.4.

Si K_1 y K_2 son dos kernel con la misma distribución estacionaria f y K_1 genera una cadena de Markov uniformemente ergódica, entonces $\tilde{K} = \alpha K_1 + (1-\alpha)K_2$, $(\alpha \in (0,1))$ es uniformemente ergódico. Además, si $\mathcal{X}$ es un conjunto pequeño para K_1 con $m = 1$, entonces los ciclos $K_1 \circ K_2$ y $K_2 \circ K_1$ son uniformemente ergódicos.

Demostración:

Si K_1 genera una cadena de Markov uniformente ergódica, entonces existe $m \in \mathbb{N}^+$, $\epsilon_m > 0$ y una medida de probabilidad ν_m tal que para cualesquiera $x \in \mathcal{X}$, $A \in \mathscr{B}(\mathcal{X})$

$$K_1^m(x, A) \geq \epsilon_m \nu_m(A)$$

por tanto se tiene una condición de minorización

$$(\alpha K_1 + (1-\alpha)K_2)^m(x, A) \geq \alpha^m K_1^m(x, A) \geq \alpha^m \epsilon_m \nu_m(A)$$

Si $\mathcal{X}$ es un conjunto pequeño para K_1 con $m = 1$ de nuevo se tiene condiciones de minorización

$$\begin{aligned}(K_1 \circ K_2)(x, A) &= \int_A \int_{\mathcal{X}} K_2(x, dy) K_1(y, dz) \\ &\geq \epsilon_1 \nu_1(A) \int_{\mathcal{X}} K_2(x, dy) = \epsilon_1 \nu_1(A)\end{aligned}$$

y

$$\begin{aligned}(K_2 \circ K_1)(x, A) &= \int_A \int_{\mathcal{X}} K_1(x, dy) K_2(y, dz) \\ &\geq \epsilon_1 \int_A \int_{\mathcal{X}} \nu_1(dy) K_2(x, dy) = \epsilon_1 (K_2 \circ \nu_1)(A)\end{aligned}$$

Como en ambos casos se tienen condiciones de minorización, se cumple la ergodicidad uniforme. □

Con estos resultados se observa que es posible generar un kernel uniformemente ergódico de un algoritmo Metropolis-Hastings independiente con distribución instrumental g tal que f/g esté acotado. Por tanto, los algoritmos MCMC híbridos se pueden utilizar para imponer las ergodicidad uniforme casi de forma automática.

1.7.7.3. "Metropolización" del Gibbs-sampler

Los algoritmos MCMC híbridos frecuentemente son útiles en un nivel elemental del proceso de simulación, i.e. cuando algunas componentes de (1.7.3) no se puede simular fácilmente. En vez de buscar un algoritmo más particular como el de Aceptación-Rechazo en cada uno de estos casos, se puede utilizar un Gibbs-sampler alternativo propuesto por Müller (1991, 1993). En el i-ésimo paso de (1.7.3) donde es difícil simular de $g_i(y_i|y_j, j \neq i)$ se sustituye por una simulación de la distribución instrumental q_i. En (1.7.3), la modificación de Müller es

Algoritmo 1.7.10. *(MCMC híbrido)*

Para $i = 1, \ldots, p$, dado $(y_1^{(t+1)}, \ldots, y_{i-1}^{(t+1)}, y_i^{(t)}, \ldots, y_p^{(t)})$

1. Simular $\tilde{y}_i \sim q_i(y_i | y_1^{(t+1)}, \ldots, y_i^{(t)}, y_{i+1}^{(t)}, \ldots, y_p^{(t)})$
2. Hacer

$$y_i^{(t+1)} = \begin{cases} y_i^{(t)} & \text{con probabilidad } 1-\rho \\ \tilde{y}_i & \text{con probabilidad } \rho \end{cases}$$

donde

$$\rho := \text{mín}\left\{1, \frac{\left(\frac{g_i(\tilde{y}_i | y_1^{(t+1)}, \ldots, y_{i-1}^{(t+1)}, y_{i+1}^{(t)}, \ldots, y_p^{(t)})}{q_i(\tilde{y}_i | y_1^{(t+1)}, \ldots, y_{i-1}^{(t+1)}, y_i^{(t)}, y_{i+1}^{(t)}, \ldots, y_p^{(t)})}\right)}{\left(\frac{g_i(y_i^{(t)} | y_1^{(t+1)}, \ldots, y_{i-1}^{(t+1)}, y_{i+1}^{(t)}, \ldots, y_p^{(t)})}{q_i(y_i^{(t)} | y_1^{(t+1)}, \ldots, y_{i-1}^{(t+1)}, \tilde{y}_i, y_{i+1}^{(t)}, \ldots, y_p^{(t)})}\right)}\right\}$$

Se debe notar que en esta sustitución, el paso Metropolis-Hastings sólo se usa vez en una iteración de (1.7.3). Por tanto, el paso modificado genera una sola simulación de $\tilde{y}_i$ en vez de interntar aproximar $g_i(y_i | y_j, j \neq i)$ con más precisión que generando T simulaciones de q_i. Hay al menos dos razones para esta elección:

- El algoritmo híbrido resultante es válido ya que g es su distribución estacionaria.
- El Gibbs-sampler también intenta una aproximación de g. Para dar una aproximación más precisa de $g_i(y_i | y_j, j \neq i)$ en (1.7.10) no necesariamente lleva a una mejor aproximación de g y el reemplazo de g_i por q_i puede mejorar la velocidad de la cadena.

Cuando varios pasos Metropolis-Hastings aparecen en un algoritmo de Gibbs ya sea por distribuciones condicionales completas complicadas ó por problemas de convergencia, Müller (1993) propuso implementar un sólo paso de aceptación después de p simulaciones condicionales. Este método es más tardado (en términos de tiempo) pero el algoritmo resultante se puede escribir como un Metropolis-Hastings en vez de una combinación de algoritmos Metropolis-Hastings. Además, se tendrá una aproximación, $q(y)$, de la distribución $g(y)$ en vez de aproximaciones locales de las distribuciones condicionales de los subvectores y_i.

Se terminará esta sección con un resultado de Liu (1995) que establece que en espacio de estados discreto, se puede mejorar al Gibbs-sampler con pasos Metropolis-Hastings. Esta mejora se expresa en términos de una reducción de la varianza de la media empírica de los $h(y^{(t)})$'s de método de Metropolis-Hastings. Esta metodología se conoce como

Metropolización y se basa en el siguiente resultado de Peskum (1973).

Lema 1.7.6.

Considérese dos cadenas de Markov reversibles con espacios de estados numerables, con matrices de probabilidades de transición T_1 y T_2 tales que $T_1 \ll T_2$. La cadena asociada con T_2 domina a la cadena asociada con T_1 en términos de varianzas.

Notación 1.7.1.

$T_1 \ll T_2$ significa que las entradas de T_2 fuera de la diagonal son más grandes que las entradas de T_1 fuera de la diagonal. ▽

Dada una distribución condicional $g_i(y_i|y_j, j \neq i)$ sobre un espacio discreto, la modificación propuesta por Liu (1995) se usa para incorporar un paso Metropolis-Hastings adicional.

Algoritmo 1.7.11. *("Metropolización" del Gibbs-sampler)*

Dado $y^{(t)}$,

1. Simular $z_i \neq y_i^{(t)}$ con probabilidad

$$\frac{g_i(z_i|y_j^{(t)}, j \neq i)}{1 - g_i(y_i^{(t)}|y_j^{(t)}, j \neq i)}$$

2. Aceptar $y_i^{(t+1)}$ con probabilidad

$$\text{mín}\left\{1, \frac{1 - g_i(y_i^{(t)}|y_j^{(t)}, j \neq i)}{1 - g_i(z_i|y_j^{(t)}, j \neq i)}\right\}$$

La probabilidad de movimiento de $y_i^{(t)}$ a un valor diferente es más alta en (1.7.11) que en el Gibbs-sampler original. El lema anterior implica el siguiente resultado de dominación.

Teorema 1.7.10.

La modificación (1.7.11) del Gibbs-sampler es más eficiente en términos de varianzas.

1.7.8. *A prioris* impropias

Ahora se discutirá acerca de un problema cuando se utilizan mal los algoritmos Metropolis-Hastings, en particular el Gibbs-sampler.

Se sabe que el Gibbs-sampler está basado en las distribuciones condicionales que se obtienen de $f(x_1, \ldots, x_q)$ ó $g(y_1, \ldots, y_p)$. Esto se complica particularmente cuando se pueden hacer simulaciones pero estas no corresponden a una distribución conjunta g, i.e. la función g en el Teorema de Hammersley-Clifford puede que no sea integrable. Puede ocurrir el mismo problema cuando se utiliza una relación de proporcionalidad como $\pi(\theta|x) \propto \pi(\theta)f(x|\theta)$ para obtener un algoritmo del tipo Metropolis-Hastings para $\pi(\theta)f(x|\theta)$.

Este no es un "defecto" del Gibbs-sampler, incluso tampoco del filosófico concepto de simulación; pues la "distribución" g no existe en estos casos de no integrabilidad. El problema en realidad es que algunas veces se utiliza sin cuidado el Gibbs-sampler en aquellas situaciones en las cuales las suposiciones subyacentes no se cumplen. Entonces, se debe tener cuidado cuando se utilizan algoritmos MCMC en situaciones en las cuales, en una perspectiva Bayesiana, se están utilizando *a proris* no informativas.

Cuando se utilizan distribuciones posteriores a prori, se puede analizar desde un punto de vista teórico el comportamiento del Gibbs-sampler como se hace en Hobert & Casella (1996, 1998). Por ejemplo, las cadenas de Markov asociadas no pueden ser positivo-recurrentes ya que la medida σ-finita asociada con las distribuciones condicionales en el Teorema de Hammersley-Clifford es invariante. Sin embargo, el reto más grande en dichos problemas se presenta cuando se tiene sospecha que algo anda mal con la medida estacionaria y generalmente se puede "resolver" revisando de nuevo la distribución *a priori*.

Una primera propuesta para verificar que el Gibbs-sampler se comporte adecuadamente es un simple monitoreo gráfico que muestre un comportamiento de desviación del Gibbs-sampler. Sin embargo, hay muchos ejemplos que se presentaron en Casella (1996) en los que los resultados del Gibbs-sampler no difieren del comportamiento de una cadena de Markov positivo-recurrente; esto puede ocurrir cuando la divergencia de la densidad posterior ocurre "en 0", i.e. en una vecindad de un punto específico que raramente se visita.

Bajo algunas condiciones de regularidad del kernel de transición, Hobert & Casella (1996) probaron que si existe una función positiva b, $\epsilon > 0$ y un conjunto compacto tal que para todo $x \in C^c$ $b(x) < \epsilon$, entonces la cadena $\{y^{(t)}\}$ satisface

$$\liminf_{t\to\infty} \frac{1}{t} \sum_{s=1}^{t} b(y^{(s)}) = 0$$

Un inconveniente en la condición anterior es que obtener una función b mientras se está monitoreando el límite inferior es delicado en un experimento de simulación. Otras referencias en el análisis del Gibbs-samplers impropios son Besag et al. (1995), Roberts & Sahu (1997) y Natarajan & McCulloch (1998).

Capítulo 2

Modelos Tradicionales de Series de Tiempo

2.1. Notación, definiciones e inferencia básica

Las expresiones "datos en series de tiempo" o "simplemente series de tiempo" se refieren a conjuntos de observaciones tomadas de forma secuencial a través del tiempo. Generalmente dichas observaciones se podrán observar en puntos del tiempo igualmente espaciados. En este caso se utilizará la notación $\{x_t\}_{t\in\mathbb{Z}}$, i.e. se indexa mediante t al conjunto de observaciones. Si las observaciones no se toman en puntos del tiempo igualmente espaciados se utilizará la notación $\{x_{t_j}\}_{j\in\mathbb{N}^+}$.

Se dará una definición formal del concepto de serie de tiempo.

Definición 2.1.1. *(Proceso de serie de tiempo)*

Un proceso de serie de tiempo es un proceso estocástico, es decir, es una colección de variables aleatorias $\{X_t\}_{t\in\mathcal{T}}$ en un mismo espacio de probabilidad $(\Omega, \mathscr{F}, \mathbb{P})$.

Durante el desarrollo de este trabajo se intentará ser ortodoxo en el uso de la notación, se utilizará $\{X_t\}_{t\in\mathcal{T}}$ para el proceso de serie de tiempo y $\{x_t\}_{t\in\mathcal{T}}$ para una realización de dicha serie de tiempo, es decir para observaciones recolectadas como serie de tiempo; sin embargo, en algunas ocasiones si la notación es excesiva (en particular en un entorno bayesiano) se utilizarán ambas notaciones indistintamente.

Notación 2.1.1.

Si $\mathcal{T}$ es de la forma $\{t_j : j \in \mathbb{N}\}$, entonces se dirá que la serie de tiempo es un proceso a tiempo discreto. Si $\mathcal{T}$ es un intervalo de $\mathbb{R}$ ó $\mathbb{R}$ mismo, entonces se dirá que la serie de tiempo es un proceso a tiempo continuo. Desde esta perspectiva, un conjunto de datos en series de

tiempo $\{x_t\}_{t=1}^T$ también se denotará mediante $x_{1:T}$ que es precisamente una colección de T realizaciones de algún proceso de serie de tiempo.

En algunos modelos estadísticos tradicionales es fundamental la suposición de que las observaciones son realizaciones de variables aleatorias independientes, sin embargo en el análisis de series de tiempo no se hará esta suposición y de hecho, se tiene interés por la estructura de dependencia de los elementos de una sucesión de variables aleatorias.

Notación 2.1.2.

Se supondrá que para cada $t \in \mathcal{T}$:

(i) $F_t(a) := \mathbb{P}(X_t \leq a)$

(ii) $\mathbb{E}(X_t) = \int_{-\infty}^{\infty} x dF_t(x) = \mu_t$

(iii) $Var(X_t) = \int_{-\infty}^{\infty} (x - \mu_t)^2 dF_t(x) = \sigma_t^2$

Y que para cada $t \in \mathcal{T}$ las integrales en (ii) y (iii) existen y son finitas. Y como consecuencia de esto se tendrá que

(iv) $\forall t \in \mathcal{T}\ \mathbb{E}(X_t^2) < \infty$

(v) Para cada $n \in \mathbb{N}^+, t_1, t_2, \ldots, t_n \in \mathcal{T}$ la distribución conjunta del vector $(X_{t_1}, X_{t_2}, \ldots, X_{t_n})$ es $F_{t_1,t_2,\ldots,t_n}(a_1, a_2, \ldots, a_n) := \mathbb{P}(X_{t_1} \leq a_1, X_{t_2} \leq a_2, \ldots, X_{t_n} \leq a_n)$

2.2. Procesos estocásticos y estacionariedad

Muchos modelos de series de tiempo están basados en la suposición de estacionariedad. Intuitivamente, un proceso de serie de tiempo es estacionario si el desarrollo de la serie no depende del momento en el cual se empezó a observar esta. En otras palabras, diferentes secciones de la serie se verán aproximadamente igual en intervalos de la misma longitud.

Definición 2.2.1. *(Estacionariedad fuerte)*

Un proceso estocástico $\{X_t\}_{t\in\mathcal{T}}$ se llama completamente estacionario ó fuertemente estacionario si $\forall n \in \mathbb{N}^+\ \forall t_1, t_2, \ldots, t_n \in \mathcal{T}\ \forall t \in \mathbb{R}$ tal que $t_1 + t, t_2 + t, \ldots, t_n + t \in \mathcal{T}$

$$F_{t_1,t_2,\ldots,t_n}(a_1, a_2, \ldots, a_n) = F_{t_1+t,t_2+t,\ldots,t_n+t}(a_1, a_2, \ldots, a_n)$$

En general, pedir que un proceso sea fuertemente estacionario es bastante ambicioso y en la práctica resulta bastante complicado verificar que un proceso es fuertemente estacionario. Dada esta condición excesiva surge el concepto de estacionariedad débil.

Definición 2.2.2. *(Estacionariedad débil)*

Un proceso estocástico $\{X_t\}_{t\in\mathcal{T}}$ tal que $\forall t \in \mathcal{T}$se llama débilmente estacionario ó estacionario de segundo orden si $\forall n \in \mathbb{N}^+ \; \forall t_1, t_2, \ldots, t_n \in \mathcal{T} \; \forall t \in \mathbb{R}$ tal que $t_1+t, t_2+t, \ldots, t_n+t \in \mathcal{T}$ todos lo momentos de orden 1 y 2 del vector $(X_{t_1}, X_{t_2}, \ldots, X_{t_n})$ son iguales a los correspondientes de orden 1 y 2 del vector $(X_{t_1+t}, X_{t_2+t}, \ldots, X_{t_n+t})$. Es decir,

$$\mathbb{E}(X_{t_1}^{r_1}, X_{t_2}^{r_2}, \ldots, X_{t_n}^{r_n}) = \mathbb{E}(X_{t_1+t}^{r_1}, X_{t_2+t}^{r_2}, \ldots, X_{t_n+t}^{r_n})$$

para $r_1, r_2, \ldots, r_n \in \mathbb{N}$ tales que $\sum_{k=1}^{n} r_k \leq 2$.

Ejemplo 2.2.1. *(Variables aleatorias independientes e idénticamente distribuidas)*

Si $X_1, X_2, \ldots$ son variables aleatorias independientes e idénticamente distribuidas, entonces $\{X_t\}_{t\in\mathbb{N}^+}$ es un proceso fuertemente estacionario. Si además, para todo $t \in \mathbb{N}^+ \; \mathbb{E}(X_t^2) < \infty$, $\{X_t\}_{t\in\mathbb{N}^+}$ también es débilmente estacionario. ▽

Proposición 2.2.1. *(Estacionariedad fuerte implica estacionariedad débil)*

Sea $\{X_t\}_{t\in\mathcal{T}}$ un proceso estocástico tal que $\forall t \in \mathcal{T} \; \mathbb{E}(X_t^2) < \infty$. Entonces, si $\{X_t\}_{t\in\mathcal{T}}$ es fuertemente estacionario, entonces $\{X_t\}_{t\in\mathcal{T}}$ es débilmente estacionario. □

No necesariamente ocurre que si $\{X_t\}_{t\in\mathcal{T}}$ es débilmente estacionario entonces $\{X_t\}_{t\in\mathcal{T}}$ es fuertemente estacionario. Considérese el siguiente contraejemplo.

Ejemplo 2.2.2. *(Estacionariedad débil no implica estacionaridad fuerte)*

Sea $\{X_t\}_{t\in\mathbb{N}^+}$ una sucesión de variables aleatorias independientes tales que para cualquier $k \in \mathbb{N}^+$ $X_{2k+1} \sim N(1,1)$ y $X_{2k} \sim exp(1)$; nótese que $\forall t \in \mathbb{N}^+ \; \mathbb{E}(X_t) = 1 = Var(X_t)$ y además si $i \neq j$, $Cov(X_{t_i}, X_{t_j}) = 0$ (por la independencia). Es decir, $\{X_t\}_{t\in\mathbb{N}^+}$ es débilmente estacionario pero como no hay idéntica distibución no se da la estacionariedad fuerte, i.e. para t impar, X_n y X_{n+t} no tienen la misma distribución. ▽

Observación 2.2.1.

Sea $\{X_t\}_{t\in\mathcal{T}}$ un proceso estocástico de segundo orden débilmente estacionario. Entonces

$$Cov(X_t, X_s) = Cov(X_{t+\tau}, X_{s+\tau})$$

para culquier τ tal que $t+\tau, s+\tau \in \mathcal{T}$. Si se toma $\tau = -t$, entonces

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

ó bien si $\tau = -s$, entonces

$$Cov(X_t, X_s) = Cov(X_{t-s}, X_0).$$

En otras palabras, para un proceso estacionario de segundo orden la covarianza entre X_s y X_t es una función de $|t-s|$ cuando este es débilmente estacionario. Adicionalmente, para un proceso estacionario de segundo orden débilmente estacionario la esperanza y la varianza es constante (como función de t), i.e.

$$\forall t \in \mathcal{T}\ \mathbb{E}(X_t) = \mu_t = \mu,\ Var(X_t) = \sigma_t^2 = \sigma^2$$

2.3. Autocorrelación, autocorrelación parcial y correlación cruzada

De manera general, muchas veces el primer paso cuando se realiza un análisis estadístico es prácticamente descriptivo, es decir, hacer algunas gráficas, obtener algunas estadísticas importantes, hacer algunas primeras afirmaciones sencillas, etc. Una de las técnicas descriptivas más utilizadas en el análisis de series de tiempo consiste en explorar los patrones de correlación en la serie o par de series en diferentes puntos del tiempo. Esto se hace mediante gráficos de la autocorrelación muestral y correlación cruzada muestral que son estimados de las funciones de autocorrelación y correlación cruzada, respectivamente. Por supuesto se tiene que empezar definiendo dichos conceptos.

Definición 2.3.1. *(Función de autocovarianza)*

Sea $\{X_t\}_{t\in\mathcal{T}}$ un proceso de serie de tiempo. Se define la función de autocovarianza, γ_X, de $\{X_t\}_{t\in\mathcal{T}}$ como

$$\gamma_X(s,t) := Cov(X_s, X_t) = \mathbb{E}[(X_s - \mathbb{E}(X_s))(X_t - \mathbb{E}(X_t))]$$

Ya se mencionó que para procesos estacionarios $\mu_t = \mu$ para cualquier $t \in \mathcal{T}$ y la función de autocovarianza depende sólo de $|s-t|$. En este caso se puede escribir a la función de autocovarianza como función de una variable de retraso (lag), h, i.e. $\gamma_X(h) := Cov(X_t, X_{t+h})$ y como depende de $|s-t| := h$ esta función es par, i.e. $\gamma_X(-h) = \gamma_X(h)$.

Definición 2.3.2. *(Función de autocorrelación)*

Sea $\{X_t\}_{t\in\mathcal{T}}$ un proceso de serie de tiempo débilmente estacionario de segundo orden. Se define la función de autocorrelación, ρ_X, de $\{X_t\}_{t\in\mathcal{T}}$ como

$$\rho_X(h) := \frac{\gamma_X(h)}{\gamma_X(0)} = \frac{Cov(X_s, X_{s+h})}{\sqrt{Var(X_s)Var(X_s)}} = \frac{Cov(X_s, X_{s+h})}{\sqrt{Var(X_s)Var(X_{s+h})}}$$

La función de autocorrelación mide la dependencia lineal entre un valor de la serie de tiempo al tiempo t y los valores pasados o futuros de tal proceso. Esta función hereda algunas propiedades de las función de autocovarianza como:

- $\rho_X(h) \in [-1, 1]$, por definirse como una correlación, i.e. $|\gamma_X(h)| \leq \gamma_X(0)$
- $\rho_X(h) = \rho_X(-h)$, i.e. ρ_X es una función par.
- Si X_t y X_{t+h} son independientes, entonces $\rho_X(h) = 0$.

Ejemplo 2.3.1. *(Ruido iid)*

Considérese una sucesión de variables aleatorias independientes e idénticamente distribuidas $\{X_t\}_{t\in\mathbb{Z}}$ tal que para cualquier $t \in \mathbb{Z}$, $\mathbb{E}(X_t) = 0$ y $\mathbb{E}(X_t^2) =: \sigma^2 < \infty$. Nótese que

$$Cov(X_{t+h}, X_t) = \gamma_X(t+h, t) = \begin{cases} \sigma^2 & \text{si } h = 0 \\ 0 & \text{si } h \neq 0 \end{cases}$$

que no depende de t. A esta sucesión se le conoce como *ruido iid* y se denota $\{X_t\}_{t\in\mathbb{Z}} \sim IID(0, \sigma^2)$. Por tanto, un ruido iid con segundo momento finito es débilmente estacionario. $\triangledown$

Ejemplo 2.3.2. *(Ruido blanco)*

Considérese una sucesión de variables aleatorias no correlacionadas $\{X_t\}_{t\in\mathbb{Z}}$ tal que para cualquier $t \in \mathbb{Z}$, $\mathbb{E}(X_t) = 0$ y $\mathbb{E}(X_t^2) =: \sigma^2 < \infty$. Nótese que dicha sucesión tiene la misma función de autocorrelación que el ejemplo anterior. Esta sucesión se conoce como *ruido blanco* (con media 0 y varianza σ^2) y se denota $\{X_t\}_{t\in\mathbb{Z}} \sim WN(0, \sigma^2)$ (por las siglas en inglés *white noise*). Claramente toda sucesión $IID(0, \sigma^2)$ es $WN(0, \sigma^2)$ pero no al revés. $\triangledown$

Ejemplo 2.3.3. *(Proceso de medias móviles de orden 1)*

Considérese una sucesión de variables aleatorias $\{X_t\}_{t\in\mathbb{Z}}$ definido por la expresión

$$X_t = \epsilon_t + \theta\epsilon_{t-1}$$

donde $\{\epsilon_t\}_{t\in\mathbb{Z}} \sim WN(0,\sigma^2)$ y $\theta \in \mathbb{R}$ constante. Nótese que $\mathbb{E}(X_t) = \mathbb{E}(\epsilon_t) + \theta\mathbb{E}(\epsilon_{t-1}) = 0$ y $Var(X_t) = Var(\epsilon_t) + Var(\theta\epsilon_{t-1}) + 2Cov(\epsilon_t, \epsilon_{t-1}) = \sigma^2 + \theta^2\sigma^2$. Es decir, $\mathbb{E}(X_t^2) = \sigma^2(1+\theta^2) < \infty$, y

$$Cov(X_{t+h}, X_t) = \gamma_X(t+h,t) = \begin{cases} \sigma^2(1+\theta^2) & \text{si } h = 0 \\ \sigma^2\theta & \text{si } h = -1, 1 \\ 0 & \text{si } |h| > 1. \end{cases}$$

Por tanto, $\{X_t\}_{t\in\mathbb{Z}}$ es estacionario de segundo orden. La función de autocorrelación de $\{X_t\}_{t\in\mathbb{Z}}$ está dada por

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

∇

Ejemplo 2.3.4. *(Proceso autoregresivo de orden 1)*

Supóngase que $\{X_t\}_{t\in\mathbb{Z}}$ es una serie estacionaria que satisface las ecuaciones

$$X_t = \phi X_{t-1} + \epsilon_t$$

donde $\{\epsilon_t\}_{t\in\mathbb{Z}} \sim WN(0,\sigma^2)$, $|\phi| < 1$ y ϵ_t está no correlacionado con X_s para todo $s < t$ (más adelante se probará que efectivamente hay solución de la ecuación anterior). Tomando esperanza de ambos lados se tiene que $\mathbb{E}(X_t) = 0$. Para encontrar la función de autocorrelación se multiplicará ambos lados de la ecuación por X_{t-h} se tomarán esperanzas, es decir

$$\mathbb{E}(X_t X_{t-h}) = \mathbb{E}(\phi X_{t-1} X_{t-h}) + \mathbb{E}(\epsilon_t X_{t-h})$$

Obteniéndose

$$\begin{aligned} \gamma_X(h) &= Cov(X_t, X_{t-h}) = Cov(\phi X_{t-1}, X_{t-h}) + Cov(\epsilon_t, X_{t-h}) \\ &= \phi\gamma_X(h-1) + 0 = \ldots = \theta^h\gamma_X(0) \end{aligned}$$

Y entonces,

$$\rho_X(h) = \frac{\gamma_X(h)}{\gamma_X(0)} = \phi^{|h|}, \quad h \in \mathbb{Z}$$

También nótese que

$$\gamma_X(0) = Cov(X_t, X_t) = Cov(\phi X_{t-1} + \epsilon_t, \phi X_{t-1} + \epsilon_t) = \phi^2 \gamma_X(0) + \sigma^2$$

y por tanto, $\gamma_X(0) = \frac{\sigma^2}{1-\phi^2}$. ▽

Definición 2.3.3.

Se dice que una función $\kappa : \mathbb{Z} \to \mathbb{R}$ es definida no-negativa si para cualquier $n \in \mathbb{N}^+$ y $\mathbf{a} = (a_1, \ldots, a_n)' \in \mathbb{R}^n$

$$\sum_{i,j=1}^{n} a_i \kappa(i-j) a_j \geq 0$$

Teorema 2.3.1.

Una función $\kappa : \mathbb{Z} \to \mathbb{R}$ es la función de autocovarianza de una serie de tiempo estacionaria si y sólo si es una función par ó no-negativa definida.

Además de la autocorrelación entre X_t y X_{t+k} se desea estudiar la correlación entre X_t y X_{t+k} después de su dependencia lineal mutua si se hubiese removido $X_{t+1}, X_{t+2}, ..., X_{t+k-1}$. La correlación condicional

$$Corr(X_t, X_{t+k} | X_{t+1}, \ldots, X_{t+k-1})$$

se conoce como *autocorrelación parcial*.

Considérese un proceso estocástico $\{X_t\}$ y sin pérdida de generalidad supóngase que $\mathbb{E}(X_t) = 0$.

Se define la dependencia lineal de X_{t+k} sobre $X_{t+1}, \ldots, X_{t+k-1}$ como el mejor estimador lineal (en el sentido de media cuadrática) como una función de $X_{t+1}, \ldots, X_{t+k-1}$. Es decir, si $\tilde{X}_{t+k}$ es el mejor estimador lineal de X_{t+k}, entonces

$$\tilde{X}_{t+k} = \alpha_1 X_{t+k-1} + \alpha_2 X_{t+k-2} + \ldots + \alpha_{k-1} X_{t+1}$$

donde $\alpha_1, \alpha_2, \ldots, \alpha_{k-1}$ son los coeficientes por mínimos cuadrados, i.e. los que se obtienen minimizando

$$\mathbb{E}[(X_{t+k} - \tilde{X}_{t+k})^2] = \mathbb{E}[(X_{t+k} - \alpha_1 X_{t+k-1} - \ldots - \alpha_{k-1} X_{t+1})^2]$$

La rutina de minimización (mediante diferenciación) lleva al siguiente sistema de ecuaciones lineales.

$$\gamma_i = \alpha_1\gamma_{i-1} + \alpha_2\gamma_{i-2} + \ldots + \alpha_{k-1}\gamma_{i-k+1} \quad \text{y}$$

$$\rho_i = \alpha_1\rho_{i-1} + \alpha_2\rho_{i-2} + \ldots + \alpha_{k-1}\rho_{i-k+1} \quad i = 1, 2, \ldots, k-1$$

En notación matricial,

$$\begin{pmatrix} \rho_1 \\ \rho_2 \\ \vdots \\ \rho_{k-1} \end{pmatrix} = \begin{pmatrix} 1 & \rho_1 & \rho_2 & \cdots & \rho_{k-2} \\ \rho_1 & 1 & \rho_1 & \cdots & \rho_{k-3} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ \rho_{k-2} & \rho_{k-3} & \rho_{k-4} & \cdots & 1 \end{pmatrix} \begin{pmatrix} \alpha_1 \\ \alpha_2 \\ \vdots \\ \alpha_{k-1} \end{pmatrix} \tag{2.1}$$

Análogamente,

$$\tilde{X}_t = \beta_1 X_{t+1} + \beta_2 X_{t+2} + \ldots + \beta_{k-1} X_{t+k+1}$$

donde $\beta_1, \ldots, \beta_{k-1}$ se obtienen minimizando

$$\mathbb{E}[(X_t - \tilde{X}_t)^2] = \mathbb{E}[(X_t - \beta_1 X_{t+1} - \beta_2 X_{t+2} - \ldots - \beta_{k-1} X_{t+k-1})^2]$$

Y por tanto,

$$\begin{pmatrix} \rho_1 \\ \rho_2 \\ \vdots \\ \rho_{k-1} \end{pmatrix} = \begin{pmatrix} 1 & \rho_1 & \rho_2 & \cdots & \rho_{k-2} \\ \rho_1 & 1 & \rho_1 & \cdots & \rho_{k-3} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ \rho_{k-2} & \rho_{k-3} & \rho_{k-4} & \cdots & 1 \end{pmatrix} \begin{pmatrix} \beta_1 \\ \beta_2 \\ \vdots \\ \beta_{k-1} \end{pmatrix} \tag{2.2}$$

lo cual implica que $\alpha_1 = \beta_1, \ldots, \alpha_{k-1} = \beta_{k-1}$.

Por tanto, la autocorrelación parcial entre X_t y X_{t+k} será igual a la autocorrelación ordinaria entre $(X_t - \tilde{X}_t)$ y $(X_{t+k} - \tilde{X}_{t+k})$. Sea Λ_k la autocorrelación parcial estre X_t y X_{t+k}, entonces

$$\Lambda_k = \frac{Cov[(X_t - \tilde{X}_t), (X_{t+k} - \tilde{X}_{t+k})]}{\sqrt{Var(X_t - \tilde{X}_k) Var(X_{t+k} - \tilde{X}_{t+k})}}$$

Sin embargo, nótese que

$$\begin{aligned} Var(X_{t+k} - \tilde{X}_{t+k}) &= \mathbb{E}[(X_{t+k} - \alpha_1 X_{t+k-1} - \ldots - \alpha_{k-1} X_{t+1})^2] \\ &= \mathbb{E}[X_{t+k}(X_{t+k} - \alpha_1 X_{t+k-1} - \ldots - \alpha_{k-1} X_{t+1})] \\ &- \alpha_1 \mathbb{E}[X_{t+k-1}(X_{t+k} - \alpha_1 X_{t+k-1} - \ldots - \alpha_{k-1} X_{t+1})] \\ &- \ldots - \alpha_{k-1} \mathbb{E}[X_{t+1}(X_{t+k} - \alpha_1 X_{t+k-1} - \ldots - \alpha_{k-1} X_{t+1})] \\ &= \mathbb{E}[X_{t+k}(X_{t+k} - \alpha_1 X_{t+k-1} - \ldots - \alpha_{k-1} X_{t+1})] \end{aligned}$$

Por tanto,

$$Var(X_{t+k} - \tilde{X}_{t+k}) = Var(X_t - \tilde{X}_t) = \gamma_0 - \alpha_1\gamma_1 - \ldots - \alpha_{k-1}\gamma_{k-1}$$

Y ahora usando el hecho de que $\alpha_1 = \beta_1, \ldots, \alpha_{k-1} = \beta_{k-1}$

$$\begin{aligned} Cov((X_t - \tilde{X}_t), (X_{t+k} - \tilde{X}_{t+k})) &= \mathbb{E}[(X_t - \alpha_1 X_{t+1} - \ldots - \alpha_{k-1} X_{t+k-1})(X_{t+k} - \alpha_1 X_{t+k-1} - \ldots - \alpha_{k-1} X_{t+} \\ &= \mathbb{E}[(X_t - \alpha_1 X_{t+1} - \ldots - \alpha_{k-1} X_{t+k-1}) X_{t+k}] \\ &= \gamma_k - \alpha_1\gamma_{k-1} - \ldots - \alpha_{k-1}\gamma_1 \end{aligned}$$

$$\therefore Cov((X_t - \tilde{X}_t), (X_{t+k} - \tilde{X}_{t+k})) = \gamma_k - \alpha_1\gamma_{k-1} - \ldots - \alpha_{k-1}\gamma_1$$

Por tanto,

$$\Lambda_k = \frac{\gamma_k - \alpha_1\gamma_{k-1} - \ldots - \alpha_{k-1}\gamma_1}{\gamma_0 - \alpha_1\gamma_1 - \ldots - \alpha_{k-1}\gamma_{k-1}} = \frac{\rho_k - \alpha_1\rho_{k-1} - \ldots - \alpha_{k-1}\rho_1}{1 - \alpha_1\rho_1 - \ldots - \alpha_{k-1}\rho_{k-1}}$$

Para encontrar $\alpha_1, \ldots, \alpha_{k-1}$ se hará mediante el método de Cramer:

$$\alpha_i = \frac{\begin{vmatrix} 1 & \rho_1 & \cdots & \rho_{i-2} & \rho_1 & \rho_i & \cdots & \rho_{k-2} \\ \rho_1 & 1 & \cdots & \rho_{i-3} & \rho_2 & \rho_{i-1} & \cdots & \rho_{k-3} \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \ddots & \vdots \\ \rho_{k-2} & \rho_{k-3} & \cdots & \rho_{k-i} & \rho_k & \rho_{k-i-2} & \cdots & 1 \end{vmatrix}}{\begin{vmatrix} 1 & \rho_1 & \cdots & \rho_{i-2} & \rho_{i-1} & \rho_i & \cdots & \rho_{k-2} \\ \rho_1 & 1 & \cdots & \rho_{i-3} & \rho_{i-2} & \rho_{i-1} & \cdots & \rho_{k-3} \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \ddots & \vdots \\ \rho_{k-2} & \rho_{k-3} & \cdots & \rho_{k-i} & \rho_{k-i-1} & \rho_{k-i-2} & \cdots & 1 \end{vmatrix}} \tag{2.3}$$

Después de una larga aritmética se puede demostrar que

$$\Lambda_k = \frac{\begin{vmatrix} 1 & \rho_1 & \rho_2 & \cdots & \rho_{k-2} & \rho_1 \\ \rho_1 & 1 & \rho_1 & \cdots & \rho_{k-3} & \rho_2 \\ \vdots & \vdots & \vdots & \ddots & \vdots & \vdots \\ \rho_{k-1} & \rho_{k-2} & \rho_{k-3} & \cdots & \rho_1 & \rho_k \end{vmatrix}}{\begin{vmatrix} 1 & \rho_1 & \rho_2 & \cdots & \rho_{k-2} & \rho_{k-1} \\ \rho_1 & 1 & \rho_1 & \cdots & \rho_{k-3} & \rho_{k-2} \\ \vdots & \vdots & \vdots & \ddots & \vdots & \vdots \\ \rho_{k-1} & \rho_{k-2} & \rho_{k-3} & \cdots & \rho_1 & 1 \end{vmatrix}} \tag{2.4}$$

La autocorrelación parcial también se puede obtener de la siguiente forma: Considérese el modelo de regresión donde la variable dependiente X_{t+k} (de un proceso estacionario con media cero) que se hará regresión en k variables "atrasadas" $X_{t+k-1}, X_{t+k-2}, \ldots, X_t$, i.e.

$$X_{t+k} = \phi_{k1} X_{t+k-1} + \phi_{k2} X_{t+k-2} + \ldots + \phi_{kk} X_t + e_{t+k}$$

donde ϕ_{ki} denota al i-ésimo parámetro de regresión y e_{t+k} es un término de error con media cero y no-correlacionado con X_{t+k-j} para $j = 1, 2, \ldots, k$. Multiplicando la ecuación anterior por X_{t+k-j}

$$X_{t+k} X_{t+k-j} = \phi_{k1} X_{t+k-1} X_{t+k-j} + \phi_{k2} X_{t+k-2} X_{t+k-j} + \ldots + \phi_{kk} X_t X_{t+k-j} + e_{t+k} X_{t+k-j}$$

tomando esperanzas, se tiene que

$$\gamma_j = \phi_{k1} \gamma_{j-1} + \phi_{k2} \gamma_{j-2} + \ldots + \phi_{kk} \gamma_{j-k}$$

Y por tanto

$$\rho_j = \phi_{k1} \rho_{j-1} + \phi_{k2} \rho_{j-2} + \ldots + \phi_{kk} \rho_{j-k}$$

De aquí que para $j = 1, 2, \ldots, k$ se tiene el siguiente sistema de ecuaciones

$$\begin{aligned} \rho_1 &= \phi_{k1}\rho_0 + \phi_{k2}\rho_1 + \ldots + \phi_{kk}\rho_{k-1} \\ \rho_2 &= \phi_{k1}\rho_1 + \phi_{k2}\rho_0 + \ldots + \phi_{kk}\rho_{k-2} \\ &\cdots \\ \rho_k &= \phi_{k1}\rho_{k-1} + \phi_{k2}\rho_{k-2} + \ldots + \phi_{kk}\rho_0 \end{aligned}$$

Utilizando de forma sucesiva la regla de Cramer para $k = 1, 2, \ldots$ se tiene que $\phi_{11} = \rho_1$,

$$\phi_{22} = \frac{\begin{vmatrix} 1 & \rho_1 \\ \rho_1 & \rho_2 \end{vmatrix}}{\begin{vmatrix} 1 & \rho_1 \\ \rho_1 & 1 \end{vmatrix}} \tag{2.5}$$

$$\phi_{33} = \frac{\begin{vmatrix} 1 & \rho_1 & \rho_1 \\ \rho_1 & 1 & \rho_2 \\ \rho_2 & \rho_1 & \rho_3 \end{vmatrix}}{\begin{vmatrix} 1 & \rho_1 & \rho_2 \\ \rho_1 & 1 & \rho_1 \\ \rho_2 & \rho_1 & 1 \end{vmatrix}} \tag{2.6}$$

$$\phi_{kk} = \frac{\begin{vmatrix} 1 & \rho_1 & \rho_2 & \cdots & \rho_{k-2} & \rho_1 \\ \rho_1 & 1 & \rho_1 & \cdots & \rho_{k-3} & \rho_2 \\ \vdots & \vdots & \vdots & \ddots & \vdots & \vdots \\ \rho_{k-1} & \rho_{k-2} & \rho_{k-3} & \cdots & \rho_1 & \rho_k \end{vmatrix}}{\begin{vmatrix} 1 & \rho_1 & \rho_2 & \cdots & \rho_{k-2} & \rho_{k-1} \\ \rho_1 & 1 & \rho_1 & \cdots & \rho_{k-3} & \rho_{k-2} \\ \vdots & \vdots & \vdots & \ddots & \vdots & \vdots \\ \rho_{k-1} & \rho_{k-2} & \rho_{k-3} & \cdots & \rho_1 & 1 \end{vmatrix}} \tag{2.7}$$

Inmediatamente se puede notar que $\phi_{kk} = \Lambda_k$. Por tanto, la autocorrelación parcial entre X_t y X_{t+k} también se puede obtener como el coeficiente de regresión asociado con X_t cuando se está "regresando" X_{t+k} sobre sus k variables "atrasadas" $X_{t+k-1}, \ldots, X_t$. ϕ_{kk} es la notación común para la autocorrelación parcial entre X_t y X_{t+k} (en la literatura de series de tiempo). Como función de k. ϕ_{kk} se conoce como función de autocorrelación parcial.

Es posible definir las funciones de covarianza cruzada y correlación cruzada para dos series de tiempo univariadas.

Definición 2.3.4. *(Funciones de covarianza cruzada y de correlación cruzada)*

Sean $\{X_t\}_{t\in\mathcal{T}}$ y $\{Y_t\}_{t\in\mathcal{T}}$ dos procesos de serie de tiempo. Se define la función de covarianza cruzada entre $\{X_t\}_{t\in\mathcal{T}}$ y $\{Y_t\}_{t\in\mathcal{T}}$, $\gamma_{X,Y}$, como

$$\gamma_{X,Y}(s,t) := E[(X_t - \mu_{X_t})(Y_s - \mu_{Y_s})]$$

para cualesquiera $s,t \in \mathcal{T}$. Y la función de correlación cruzada entre $\{X_t\}_{t\in\mathcal{T}}$ y $\{Y_t\}_{t\in\mathcal{T}}$, $\rho_{X,Y}$, se define como

$$\rho_{X,Y}(s,t) = \frac{\gamma_{X,Y}(s,t)}{\sqrt{\gamma_{X,X}(s,s)\gamma_{Y,Y}(t,t)}}$$

2.4. Procesos lineales

La clase de modelos lineales de series de tiempo, que incluye a la clase de modelos autoregresivos y de medias moviles ($ARMA$), proporciona un marco general para el estudio de procesos estacionarios. De hecho, todo proceso estacionario de segundo orden es un proceso lineal ó se puede transformar en un proceso lineal al revomer una componente determinista. Este resultado se conoce como descomposición de Wold.

Definición 2.4.1. *(Proceso lineal)*

La serie de tiempo $\{X_t\}$ es un proceso lineal si tiene la representación

$$X_t = \sum_{j=-\infty}^{\infty} \psi_j \epsilon_{t-j} \tag{2.8}$$

donde $\{\epsilon_t\}$ es un ruido blanco con varianza σ^2 y $\{\psi_j\}$ es una sucesión de números reales tal que $\sum_{j=-\infty}^{\infty} |\psi_j| < \infty$

Notación 2.4.1.

En términos del operador de retraso B, se puede escribir (2.8) como $X_t = \psi(B)(\epsilon_t)$, donde $\psi(B) = \sum_{j=-\infty}^{\infty} \psi_j B^j$. ▽

Se dice que $\{X_t\}$ proceso lineal es un *proceso de medias móviles de órden infinito* ó $MA(\infty)$ si $\psi_j = 0$ para todo $j \in \mathbb{Z}^-$, i.e. si

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

Observación 2.4.1.

La condición $\sum_{j=-\infty}^{\infty} |\psi_j| < \infty$ garantiza que la suma infinita (2.8) converja c.s. ya que $\mathbb{E}(|\epsilon_t|) \leq \sigma^2$ y

$$\mathbb{E}(|X_t|) \leq \sum_{j=-\infty}^{\infty} |\psi_j| \mathbb{E}(|\epsilon_{t-j}|) \leq \left(\sum_{j=-\infty}^{\infty} |\psi_j| \right) \sigma < \infty$$

Esta condición también garantiza que $\sum_{j=-\infty}^{\infty} \psi_j^2 < \infty$ y por tanto que la serie en (2.8) converja en media cuadrática, i.e. que X_t es el límite en media cuadrático de las sumas parciales $\sum_{j=-n}^{n} \psi_j \epsilon_{t-j}$. La condición $\sum_{j=-n}^{n} |\psi_j| < \infty$ también asegura que la serie en la siguiente proposición converja casi seguramente y en error cuadrático medio. ▽

El operador $\psi(B)$ se puede pensar como un filtro lineal, que cuando se le introduce un ruido blanco $\{\epsilon_t\}$ como "input" se obtiene un proceso $\{X_t\}$ como "output". Y de hecho, esto se generaliza en la siguiente proposición.

Proposición 2.4.1.

Sea $\{Y_t\}$ una serie de tiempo estacionaria con media 0 y función de covarianza γ_Y. Si $\sum_{j=-\infty}^{\infty} |\psi_j| < \infty$, entonces la serie de tiempo

$$X_t = \sum_{j=-\infty}^{\infty} \psi_j Y_{t-j} = \psi(B)(Y_t) \tag{2.9}$$

es estacionaria con media cero y función de autocovarianza

$$\gamma_X(h) = \sum_{j=-\infty}^{\infty} \sum_{k=-\infty}^{\infty} \psi_j \psi_k \gamma_Y(h+k-j) \tag{2.10}$$

En el caso especial de que $\{X_t\}$ sea un proceso lineal

$$\gamma_X(h) = \sum_{j=-\infty}^{\infty} \psi_j \psi_{j+h} \sigma^2 \tag{2.11}$$

Demostración:

$$\mathbb{E}(|X_t|) \leq \sum_{j=-\infty}^{\infty} |\psi_j| \mathbb{E}(|Y_{t-j}|) \leq \left(\sum_{j=-\infty}^{\infty} |\psi_j| \right) \sqrt{\gamma_Y(0)} < \infty$$

Como $\mathbb{E}(Y_t) = 0$, se tiene que

$$\mathbb{E}(X_t) = \mathbb{E}\left[\sum_{j=-\infty}^{\infty} \psi_j Y_{t-j} \right] = \sum_{j=-\infty}^{\infty} \psi_j \mathbb{E}(Y_{t-j}) = 0$$

y además

$$\begin{aligned} \mathbb{E}(X_{t+h} X_t) &= \mathbb{E}\left[\left(\sum_{j=-\infty}^{\infty} \psi_j Y_{t+h-j} \right) \left(\sum_{k=-\infty}^{\infty} \psi_k Y_{t-k} \right) \right] \\ &= \sum_{j=-\infty}^{\infty} \sum_{k=-\infty}^{\infty} \psi_j \psi_k \mathbb{E}(Y_{t+h-j} Y_{t-k}) \\ &= \sum_{j=-\infty}^{\infty} \sum_{k=-\infty}^{\infty} \psi_j \psi_k \gamma_Y(h-j+k) \end{aligned}$$

que no depende de t, lo cual prueba que $\{X_t\}$ es estacionario y con la función de autocovarianza correspondiente. El intercambio entre la suma infinita y el operador de esperanza se justifica dada la convergencia absoluta de $\{\psi_j\}$. □

Observación 2.4.2.

La convergencia absoluta de (2.9) implica que los filtros de la forma $\alpha(B) = \sum_{j=-\infty}^{\infty} \alpha_j B^j$ y $\beta(B) = \sum_{j=-\infty}^{\infty} \beta_j B^j$ con coeficientes absolutamente sumables se puedan aplicar sucesivamente a series estacionarias $\{Y_t\}$ para generar una nueva serie estacionaria

$$W_t = \sum_{j=-\infty}^{\infty} \psi_j Y_{t-j}$$

donde

$$\psi_j = \sum_{k=-\infty}^{\infty} \alpha_k \beta_{j-k} = \sum_{k=-\infty}^{\infty} \beta_k \alpha_{j-k} \tag{2.12}$$

Ó equivalentemente, $W_t = \psi(B)Y_t$, donde $\psi(B) = \alpha(B)\beta(B) = \beta(B)\alpha(B)$ y los productos se definen mediante (2.12) ó equivalentemente, multiplicando las series $\sum_{j=-\infty}^{\infty} \alpha_j B^j$ y $\sum_{j=-\infty}^{\infty} \beta_j B^j$ término a término y agrupando las potencias de B. Es claro que con ψ_j así obtenido es indistinto el orden con el cual se aplican los filtros $\alpha(B)$ y $\beta(B)$. $\triangledown$

Ejemplo 2.4.1. *(Proceso AR(1))*

Considérese el proceso de serie de tiempo $\{X_t\}$ definido por

$$X_t = \phi X_{t-1} + \epsilon_t, \quad t \in \mathbb{Z} \tag{2.13}$$

donde $\{\epsilon_t\}$ es un ruido blanco con media cero y varianza σ^2, $|\phi| < 1$ y para todo $s < t$ X_s y ϵ_t están no correlacionados. Para mostrar que existe solución estacionaria de (2.13) y es única considérese el proceso lineal definido por

$$X_t = \sum_{j=0}^{\infty} \phi^j \epsilon_{t-j} \tag{2.14}$$

Nótese que los coeficientes ϕ^j para $j \geq 0$ son absolutamente sumables ya que $|\phi| < 1$. Obsérvese que

$$\begin{aligned} X_t - \phi X_{t-1} &= \sum_{j=0}^{\infty} \phi^j \epsilon_{t-j} - \phi \sum_{j=0}^{\infty} \phi^j \epsilon_{t-1-j} \\ &= \epsilon_t + \sum_{j=1}^{\infty} \phi^j \epsilon_{t-j} - \sum_{j=0}^{\infty} \phi^{j+1} \epsilon_{t-1-j} \\ &= \epsilon_t + \sum_{j=1}^{\infty} \phi^j \epsilon_{t-j} - \sum_{j=1}^{\infty} \phi^j \epsilon_{t-j} = \epsilon_t \end{aligned}$$

Es decir, $X_t = \sum_{j=0}^{\infty} \phi^j \epsilon_{t-j}$ es solución de (2.13). Y por la proposición 2.4.1 es estacionario y con función de autocovarianza

$$\gamma_X(h) = \sum_{j=0}^{\infty} \phi^j \phi^{j+h} \sigma^2 = \phi^h \sigma^2 \sum_{j=0}^{\infty} (\phi^2)^j = \frac{\phi^h \sigma^2}{1 - \phi^2}, \quad \text{para } h \in \mathbb{N}$$

Para mostrar que (2.14) es la única solución estacionaria de (2.13) considérese $\{Y_t\}$ una solución estacionaria de (2.13). De forma iterativa se puede observar que

$$\begin{aligned} Y_t &= \phi Y_{t-1} + \epsilon_t \\ &= \epsilon_t + \phi \epsilon_{t-1} + \phi^2 Y_{t-2} \\ &= \ldots = \epsilon_t + \phi \epsilon_{t-1} + \ldots + \phi^k \epsilon_{t-k} + \phi^{k+1} Y_{t-k-1} \end{aligned}$$

Si $\{Y_t\}$ es estacionaria, entonces $\mathbb{E}(Y_t^2)$ es finita e independiente de t, por tanto

$$\mathbb{E}\left[\left(Y_t - \sum_{j=0}^{k} \phi^j \epsilon_{t-j}\right)^2\right] = \phi^{2k+2}\mathbb{E}\left[Y_{t-k-1}^2\right]$$

Y entonces

$$\lim_{k\to\infty} \mathbb{E}\left[\left(Y_t - \sum_{j=0}^{k} \phi^j \epsilon_{t-j}\right)^2\right] = \lim_{k\to\infty} \phi^{2k+2}\mathbb{E}\left[Y_{t-k-1}^2\right] = 0$$

Esto significa que Y_t es el límite en media cuadrática $\sum_{j=0}^{\infty} \epsilon_{t-j}$ y por tanto el proceso que se define mediante (2.14) es la única solución estacionaria de (2.13).

En caso de que $|\phi| > 1$, la serie en (2.14) no converge; pero se puede reescribir (2.13) en la forma $X_t = -\phi^{-1}\epsilon_{t+1} + \phi^{-1}X_{t+1}$.

Iterando se tiene que

$$\begin{aligned} X_t &= -\phi^{-1}\epsilon_{t+1} - \phi^{-2}\epsilon_{t+2} + \phi^{-2}X_{t+2} \\ &= \ldots = -\phi^{-1}\epsilon_{t+1} - \ldots - \phi^{-k-1}\epsilon_{t+k+1} + \phi^{-k-1}X_{t+k+1} \end{aligned}$$

Lo cual demuestra, con el mismo argumento que se usó antes, que

$$X_t = -\sum_{j=0}^{\infty} \phi^{-1}\epsilon_{t+j}$$

es la única solución estacionaria de (2.13). Esta solución no se debe confundir con la solución no-estacionaria $\{X_t\}$ de (2.13) que se obtiene cuando X_0 es cualquier variable aleatoria no correlacionada con $\{Z_t\}$.

La solución $X_t = -\sum_{j=0}^{\infty} \phi^{-1}\epsilon_{t+j}$ parece poco natural ya que depende de (y está correlacionado) de valores futuros de ϵ_s, a diferencia de la solución $X_t = \sum_{j=0}^{\infty} \phi^j \epsilon_{t-j}$ que tiene la propiedad de que X_t está no correlacionado con ϵ_s para $s > t$. Por estas razones, frecuentemente se restringe la atención $AR(1)$ cuando $|\phi| < 1$. Entonces se puede representar en términos de $\{\epsilon_s\}_{s=-\infty}^{t}$ y se dirá que $\{X_t\}$ es función causal (ó sólo causal ó independiente del futuro) de $\{\epsilon_t\}$, ó de forma más precisa, que $\{X_t\}$ es un proceso auto-regresivo causal. $\triangledown$

Observación 2.4.3.

Para un proceso $AR(1)$ con $|\phi| > 1$, si se define $W_t := X_t - \frac{1}{\phi}X_{t-1}$, se puede mostrar que $\{W_t\}$ es un ruido blanco con media cero y varianza σ_W^2 (que por supuesto es función de σ^2 y ϕ) y se puede reescribir

$$X_t = \frac{1}{\phi}X_{t-1} + W_t = \phi' X_{t-1} + W_t, \quad t \in \mathbb{Z}$$

Es decir, se puede expresar a un $AR(1)$, $\{X_t\}$ con $|\phi| > 1$, en la misma estructura $AR(1)$ con $\phi' := \phi^{-1}$ y $|\phi'| < 1$. Por tanto, desde un punto de vista de segundo orden, no se pierde riqueza en la estructura $AR(1)$ eliminando los procesos con $|\phi| > 1$. Si $\phi = \pm 1$, no existe solución estacionaria de (2.13) ▽

Observación 2.4.4.

Para un proceso $AR(1)$ con $|\phi| < 1$, la única solución estacionaria $X_t = \sum_{j=0}^{\infty} \phi^j \epsilon_{t-j}$, también se pudo encontrar de la siguiente forma: sean $\phi(B) = I - \phi(B)$ y $\pi(B) = \sum_{j=0}^{\infty} \phi^j B^j$, entonces $\psi(B) := \phi(B)\pi(B) = I$. Entonces, si se aplica π a ambos lados de $X_t - \phi X_{t-1} = \phi(B)(X_t) = \epsilon_t$ se obtiene

$$X_t = \pi(B)(\epsilon_t) = \sum_{j=0}^{\infty} \phi^j \epsilon_{t-j}$$

▽

Definición 2.4.2.

Se dice que la serie de tiempo $\{X_t\}$ es un proceso $ARMA(1,1)$ si es estacionaria y para todo $t \in \mathbb{Z}$

$$X_t - \phi X_{t-1} = \epsilon_t + \theta \epsilon_{t-1} \tag{2.15}$$

donde $\{\epsilon_t\}$ es un ruido blanco con varianza σ^2 y $\phi + \theta \neq 0$.

Notación 2.4.2.

Usando el operador de retraso B, (2.15) se puede reescribir como $\phi(B)(X_t) = \theta(B)(\epsilon_t)$, donde $\phi(B)$ y $\theta(B)$ son los filtros lineales

$$\phi(B) = 1 - \phi B \quad \text{y} \quad \theta(B) = 1 + \theta B$$

▽

Una primera pregunta que surge de manera natural es para qué conjunto de valores de ϕ y θ existe una solución estacionaria de (2.15). Supógase que $|\phi| < 1$ y sea $\chi(z)$ la serie de potencias de la función $\frac{1}{\phi(z)}$, es decir, $\chi(z) = \sum_{j=0}^{\infty} \phi^j z^j$ (la cual tiene coeficientes absolutamente sumables). Entonces $\chi(B)\phi(B) = 1$. Aplicando el filtro $\chi(B)$ a $\phi(B)(X_t) = \theta(B)(\epsilon_t)$ se tiene que $X_t = \chi(B)\theta(B)(\epsilon_t) = \psi(B)(\epsilon_t)$, donde

$$\psi(B) = \sum_{j=0}^{\infty} \psi_j B^j = (1 + \psi B + \psi^2 B^2 + \ldots)(1 + \theta B)$$

y por tanto, $\psi_0 = 1$ y para $j \in \mathbb{N}^+$, $\psi_j = (\theta + \phi)\phi^{j-1}$.

Por tanto, el proceso

$$X_t = \epsilon_t + (\theta + \phi) \sum_{j=1}^{\infty} \phi^{j-1} \epsilon_{t-j}$$

es la única solución estacionaria de (2.15).

Ahora, supóngase que $|\phi| > 1$. Nótese que

$$\frac{1}{\phi(z)} = -\sum_{j=1}^{\infty} \phi^{-j} z^{-j}$$

donde $\sum_{j=1}^{\infty} |-\phi^{-j}| = \frac{1/\phi}{1-1/\phi} = (\phi - 1)^{-1} < \infty$ (i.e. tiene coeficientes absolutamente sumables).

Se puede aplicar el mismo argumento que en el caso $|\phi| < 1$ para obtener la única solución estacionaria de (2.15): Sea $\chi(B) = -\sum_{j=1}^{\infty} \phi^{-j} B^{-j}$ y se aplica $\chi(B)$ a $\phi(B)(X_t) = \theta(B)(\epsilon_t)$ se tiene que $X_t = \chi(B)\theta(B)(\epsilon_t) = -\theta\phi^{-1}\epsilon_t - (\theta + \phi)\sum_{j=1}^{\infty} \phi^{-j-1}\epsilon_{t+j}$.

Si $\phi = -1$ ó $\phi = 1$, no existe solución estacionaria de (2.15). Por tanto, no hay un proceso $ARMA(1,1)$ con $\phi = -1$ ó $\phi = 1$ de acuerdo a la definición.

Se puede resumir sobre la existencia y naturaleza de las soluciones estacionarias del proceso $ARMA(1,1)$ de la siguiente forma:

- Existe una solución de la ecuación $ARMA(1,1)$ si y sólo si $\phi \neq \pm 1$.
- Si $|\phi| < 1$, entonces la única solución estacionaria está dada por

$$X_t = \epsilon_t + (\theta + \phi) \sum_{j=1}^{\infty} \phi^{j-1} \epsilon_{t-j}$$

  En este caso se dice que $\{X_t\}$ es causal ó función causal de $\{\epsilon_t\}$ ya que X_t se puede expresar en términos de valores pasados y presentes ϵ_s, $s \leq t$

- Si $|\phi| > 1$, entonces la única solución estacionaria está dada por

$$X_t = \chi(B)\theta(B)(\epsilon_t) = -\theta\phi^{-1}\epsilon_t - (\theta+\phi)\sum_{j=1}^{\infty}\phi^{-j-1}\epsilon_{t+j}$$

  La solución es no-causal ya que X_t es una función de ϵ_s, $s \geq t$.

Justo como la causalidad significa que X_t se puede expresar en términos de ϵ_t, $s \leq t$, el concepto de invertibilidad significa que ϵ_t se puede expresar en términos de X_s, $s \leq t$. Ahora, se mostrará que el proceso $ARMA(1,1)$ definido en (2.15) es invertible si $|\theta| < 1$. Para demostrar esto, considérese $\xi(z)$ la expansión en serie de potencias de $\frac{1}{\theta(z)}$, i.e. $\sum_{j=0}^{\infty}(-\theta)^j z^j$, que tiene coeficientes absolutamente sumables. Y nótese que $\xi(B)\theta(B) = 1$; aplicando $\xi(B)$ en ambos lados de $\phi(B)(X_t) = \theta(B)(\epsilon_t)$ se tiene que

$$\epsilon_t = \xi(B)\phi(B)(X_t) = \pi(B)(X_t)$$

donde

$$\pi(B) = \sum_{j=0}^{\infty}\pi_j B^j = (I - \theta B + (-\theta)^2 B^2 + \ldots)(I - \phi B)$$

Entonces,

$$\epsilon_t = X_t - (\phi+\theta)\sum_{j=1}^{\infty}(-\theta)^{j-1}X_{t-j}$$

Por tanto, el proceso $ARMA(1,1)$ es invertible ya que ϵ_t se puede expresar en términos de valores pasados y presentes del proceso X_s, $s \leq t$. Con un argumento similar al que se usó para demostrár la no-causalidad cuando $|\phi| > 1$ se puede demostrar que el proceso $ARMA(1,1)$ es no-invertible si $|\theta| > 1$ y en este caso

$$\epsilon_t = -\phi\theta^{-1}X_t + (\theta+\phi)\sum_{j=1}^{\infty}(-\theta)^{-j-1}X_{t+j}$$

Se puede resumir los resultados de la siguiente forma:

- Si $|\theta| < 1$, entonces el proceso $ARMA(1,1)$ es invertible y ϵ_t se puede expresar en términos de X_s, $s \leq t$ como $\epsilon_t = X_t - (\phi+\theta)\sum_{j=1}^{\infty}(-\theta)^{j-1}X_{t-j}$
- Si $|\theta| > 1$, entonces el proceso $ARMA(1,1)$ es no-invertible y ϵ_t se puede expresar en términos de X_s, $s \geq t$ como $\epsilon_t = -\phi\theta^{-1}X_t + (\theta+\phi)\sum_{j=1}^{\infty}(-\theta)^{-j-1}X_{t+j}$

Observación 2.4.5.

En los casos $|\theta| = 1$, el proceso $ARMA(1,1)$ es invertible en el sentido más general en el que ϵ_t es el límite (en media cuadrática) de combinaciones lineales finitas de X_s, $s \leq t$, aunque no se puede expresar explícitamente como una combinación lineal infinita de X_s, $s \leq t$. Cuando se ocupe el término invertible será en el sentido de que $\epsilon_t = \sum_{j=0}^{\infty} \pi_j X_{t-j}$ con $\sum_{j=0}^{\infty} |\pi_j| < \infty$ $\triangledown$

Observación 2.4.6.

Si el proceso $ARMA(1,1)$, $\{X_t\}$, es no-causal ó no-invertible con $|\theta| > 1$, se puede encontrar una nueva sucesión de ruido blanco W_t tal que $\{X_t\}$ es un proceso con estructura $ARMA(1,1)$ no-invertible pero causal con respecto a $\{W_t\}$. Por tanto, desde un perspectiva de segundo orden, no se pierde riqueza en la estructura $ARMA(1,1)$ si se restringe la atención a modelos $ARMA(1,1)$ causales e invertibles. Esto último es válido para modelos $ARMA$ de orden más alto. $\triangledown$

2.5. Algunos particularidades de modelos auto-regresivos

Los modelos auto-regresivos de series de tiempo son ya tradicionales en el análisis de datos en series de tiempo ya sea como parte de una gran cantidad de modelos o bien en sus formas generalizadas que se utilizan para modelos dependientes del tiempo no-estacionarios. Los conceptos y la estructura de modelos lineales auto-regresivos también forman una base para la apreciación de modelos no-lineales. En esta sección se recordarán algunas formas de dichos modelos (y más adelante se verán algunos asuntos relacionados con inferencia) y algunos otros temas relacionados con esta amplia gama de modelos.

2.5.1. Estructura de auto-regresiones

Considérese la serie de tiempo de cantidades igualmente espaciadas $\{X_t\}_{t=1}^{\infty}$ que proviene del modelo

$$X_t = \sum_{j=1}^{p} \phi_j X_{t-j} + \epsilon_t$$

donde $\{\epsilon_t\}_{t=1}^{\infty}$ es una sucesión de términos de error no-correlacionados y $\{\phi_j\}_{j=1}^{p}$ es un conjunto de parámetros constantes. Este modelo se define de forma secuencial: X_t se genera como una función de valores pasados, parámetros y errores aleatorios. Las variables aleatorias $\epsilon_1, \epsilon_2, \ldots$, se conocen como *términos de innovación* y se supondrá que son condicionalmente independientes a los valores pasados de la serie. Frecuentemente se supone que $\epsilon_t \sim N(0, \nu)$ y que son independientes. Este es el modelo autorregresivo estándar o $AR(p)$, donde $p \in \mathbb{N}^+$ es el orden de la autorregresión.

Los modelos AR se pueden ver desde un punto de vista empírico: se supone que los datos están relacionados a través del tiempo y la forma AR es una de las clases de modelos

empíricos más simples para estudiar las dependencias. Un motivación más formal es aquella que está basada en la teoría de procesos estocásticos estacionarios.

La definición secuencial del modelo y su naturaleza Markoviana dan como consecuencia que la densidad también tenga una estructura secuencial

$$p(x_{1:T}) = p(x_{1:p}) \prod_{t=p+1}^{T} p(x_t | x_{(t-p):(t-1)})$$

para cualquier $T > p$. Se debe notar que estas densidades están condicionadas a $(\phi_1, \phi_2, \ldots, \nu)$ pero no se incluyen en la expresión para no sobrecargar la notación, sin embargo se debe tener en mente y más aún cuando se esté haciendo un análisis desde la perspectiva bayesiana. Si los primeros p valores de la serie son conocidos y se consideran como constantes fijas y $T = n + p$ para algún $n \in \mathbb{N}^+$, entonces la densidad condicional de $\mathbf{X} = (X_T, X_T T - 1, \ldots, X_{p+1})'$ dados los primeros p valores está dada por

$$\begin{aligned} p(\mathbf{x}|x_{1:p}) &= \prod_{t=p+1}^{T} p(x_t | x_{(t-p):(t-1)}) \\ &= \prod_{p+1}^{T} g_{X_t}(x_t) = g_{\mathbf{X}_t}(\mathbf{x}) \end{aligned}$$

donde g_{X_t} es la densidad normal con media $\mathbf{f}_t'\boldsymbol{\phi}$ y varianza v con $\boldsymbol{\phi} = (\phi_1, \ldots, \phi_p)'$, $\mathbf{f}_t = (x_{t-1}, \ldots, x_{t-p})'$ y $g_{\mathbf{X}_t}$ es la densidad normal p-variada con vector de medias $\mathbf{F}'\boldsymbol{\phi}$ y matriz de varianzas-covarianzas vI_n con $\mathbf{F} = (\mathbf{f}_T, \ldots, \mathbf{f}_{p+1})$

2.5.1.1. Causalidad y estacionariedad en los procesos AR

Ya se dijo que un proceso $AR(p)$, $\{X_t\}$, es causal si se puede escribir como un proceso lineal dependiente del pasado, es decir si

$$X_t = \Psi(B)(\epsilon_t) = \sum_{j=0}^{\infty} \psi_j \epsilon_{t-j}$$

donde B es el operador de rezago, $B(\epsilon_t) = \epsilon_{t-1}$, $\psi_0 = 1$ y $\sum_{j=0} \infty |\psi_j| < \infty$.

$\{X_t\}$ es causal sólo cuando el polinomio característico auto-regresivo, definido como

$$\Upsilon(u) = 1 - \sum_{j=1}^{p} \phi_j u^j$$

tiene raíces con módulo mayor que 1. Es decir, $\{X_t\}$ es causal si $\Upsilon(u) = 0$ sólo cuando $|u| > 1$.

El polinomio característico auto-regresivo también se puede escribir como

$$\Upsilon(u) = \prod_{j=1}^{p}(1 - \alpha_j u)$$

así que las raíces de $\Upsilon(u)$ son los inversos multiplicativos de los α_j's. Los α_j's pueden ser reales ó ser parejas de complejos conjugados (dependiendo la paridad de p). En cualquiera de los casos, si $|\alpha_j| < 1$ para todo j, entonces el proceso es estacionario.

Ya se mencionó que en el caso $p = 1$ la condición $-1 < \phi_1 < 1$ implica estacionariedad y suponiendo innovaciones Gaussianas, la distribución estacionaria de X_t es $N(0, \frac{v}{1-\phi_1^2})$. Cuando $\phi_1 = 1$, el modelo se convierte en una caminata aleatoria no estacionaria. La distribución estacionaria bivariada de $(X_t, X_{t-1})'$ es normal con correlación $\rho_X(1) = \phi_1$ y la de $(X_t, X_{t-h})'$ es normal con correlación $\rho_X(1) = \phi_1^{|h|}$.

Cuando $p = 2$, $X_t = \phi_1 X_{t-1} + \phi_2 X_{t-2} + \epsilon_t$ y la condición de causalidad implica que los valores de los parámetros deben pertenecer a la región $-1 < \phi_2 < 1$, $\phi_1 < 1 - \phi_2$ y $\phi_1 > \phi_2 - 1$.

2.5.1.2. Representación espacial de los procesos $AR(p)$

La representación espacial de un modelo $AR(p)$ se puede aprovechar para estudiar la estructura matemática y la inferencia y análisis de datos. Una versión de esta representación de $X_t = \sum_{j=1}^{p} \phi_j X_{t-j} + \epsilon_t$ es

$$X_t = \mathbf{F}'\mathbf{Y}_t$$

$$\mathbf{Y}_t = \mathbf{G}\mathbf{Y}_{t-1} + \omega_t$$

Donde $\mathbf{Y}_t = (X_t, X_{t-1}, \ldots, X_{t-p+1})'$ es el vector de estados al tiempo t. La innovación al tiempo t aparece en el vector de error $\omega_t = (\epsilon_t, 0, \ldots, 0)'$. Adicionalmente, $\mathbf{F} = (1, 0, \ldots, 0)'$ y

$$\mathbf{G} = \begin{pmatrix} \phi_1 & \phi_2 & \phi_3 & \cdots & \phi_{p-1} & \phi_1 \\ 1 & 0 & 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 & 0 \\ \vdots & \vdots & \vdots & \ddots & 0 & \vdots \\ 0 & 0 & 0 & 0 & 1 & 0 \end{pmatrix}$$

El comportamiento esperado del futuro del proceso se puede describir mediante la *función de predicción* $d_t(h) := \mathbb{E}(X_{t+h}|x_{1:t})$ que es una función definida en $\mathbb{N}^+$ para un "origen" fijo $t \geq p$ condicionado a los últimos p valores observados de la serie en el estado actual del vector $\mathbf{Y}_t = (X_t, X_{t-1}, \ldots, X_{t-p+1})'$. En este caso se tiene que $d_t(h) = \mathbf{F}'\mathbf{G}^h\mathbf{Y}_t$. Esta función se puede analizar de forma más sencilla si $\mathbf{G}$ tiene eigenvalores diferentes (reales ó complejos). Estos eigenvalores son precisamente los inversos multiplicativos del polinomio auto-regresivo $\Upsilon(u)$, denotadas por α_j. Entoces, se puede escribir

$$d_t(h) = \sum_{j=1}^{p} c_{tj}\alpha_j^h$$

donde las $c'_{tj}s$ son constantes que dependen de $\boldsymbol{\phi}$ y el estado actual $\mathbf{Y}_t$. Cada coeficiente está dado por $c_{tj} = d_j e_{tj}$. Los valores d_j y e_{tj} son los elementos de los vectores de dimensión p, $\mathbf{d} = \mathbf{E}'\mathbf{F}$ y $\mathbf{e}_t = \mathbf{E}^{-1}\mathbf{Y}_t$, donde $\mathbf{E}$ es la eigenmatriz de $\mathbf{G}$, i.e. $\mathbf{E}$ es una matriz de $p \times p$ cuyas columnas son los eigenvectores en el orden correspondientes de los eigenvalores.

La forma de la función de predicción depende de la combinación de los eigenvalores reales y complejos de $\mathbf{G}$. Por ejemplo, supóngase que α_j es real y positivo; entonces su contribución a la función de predicción es $c_{tj}\alpha_j^h$. Si el proceso es tal que $|\alpha_i| < 1$ para todo i, esta función de h decrece exponencialmente hacia cero, monótonamente si $\alpha_j > 0$, y oscilará entre valores positivos o negativos si $\alpha < 0$. Si $|\alpha_j| > 1$, la función de predicción explota. La contribución relativa a la función de predicción se mide mediante la tasa de decaimiento y a la amplitud inicial c_{tj}, la última depende explícitamente del estado actual y por tanto, tienen diferente impacto en diferentes tiempos mientras el estado cambia en respuesta a la sucesión de innovaciones.

En el caso de eigenvalores complejos, el hecho de que $\mathbf{G}$ sea real implica que cualesquiera complejos aparecen en parejas (con sus conjugados). Supóngase que $\alpha_1 = r \cdot exp(i\omega)$ y $\alpha_2 = r \cdot exp(-i\omega)$ son conjuntados complejos con módulo r y argumento ω. Entonces los correspondientes factores complejos c_{t1} y c_{t1} son conjugados, $a_t \exp(\pm ib_t)$, y la contribución a $d_t(h)$ (que debe ser real) es

$$c_{t1}\alpha_1^h + c_{t2}\alpha_2^h = 2a_t r^h cos(\omega h + b_t)$$

Por tanto, ω determina la constante de frecuencia de una oscilación senosoidal en la función de predicción. En los casos en los que $r > 1$, entonces la variación senosoidal incrementa su amplitud cuando r^h se incrementa.

Generalmente la función de predicción es una combinación lineal de los términos que decrecen ó explotan y los factores que decrecen o explotan multiplicando componentes senosoidales de diferentes periodos.

2.5.1.3. Caracterización del proceso $AR(2)$

El caso $p = 2$ es importante en la práctica. El proceso es estacionario si $-1 < \phi_2 < 1$, $\phi_1 < 1 - \phi_2$ y $\phi_1 > \phi_2 - 1$. En este caso, el polinomio auto-regresivo cuadrático $\Upsilon(z)$ tiene raíces recíprocas α_1 y α_2 dentro del círculo unitario. Esto se resume de la siguiente forma

- Las dos raíces son reales cuando $\phi_1^2 + 4\phi_2 \geq 0$, y en este caso la función de predicción decrece exponencialmente.
- Las dos raíces son complejas (conjugadas) $r\exp(\pm i\omega$ cuando $\phi_1^2 + 4\phi_2 < 0$. Las raíces tienen módulo $r = \sqrt{-\phi_2}$ y el argumento está dado por $cos(\omega) = |\phi_1|/2r$. Y la función de predicción se comporta como un coseno exponencialmente amortiguado.

Y para la estacionariedad $-2 < \phi_1 < 2$; para raíces complejas se tiene la restricción adicional de que $-1 < \phi_2 < -\phi_1^2/4$. Por lo que en estos casos, el modelo $X_t = \phi_1 X_{t-1} + \phi_2 X_{t-2} + \epsilon_t$ representa un proceso cuasi-cíclico. Una varianza grande v induce un mayor grado de variación en esta dinámica, i.e. un proceso cuasi-cíclico. Si v es muy pequeña ó se vuele cero en algún punto, el proceso decaerá a cero en amplitud debido a este factor amortiguador. En la frontera de esta región $\phi_2 = -1$, el módulo es $r = 1$ y la función de predicción es senosoidal sin amortiguamiento; en este caso $\phi_1 = 2cos(\omega)$. Por tanto, para $|\phi_1| < 2$, el modelo $X_t = \phi_1 X_{t-1} + \phi_2 X_{t-2} + \epsilon_t$ tiene forma senosoidal con variación aleatoria en amplitud y fase; con una varianza de la innovación pequeña (ó cero) v, la forma senosoidal es constante y representa, esencialmente, una onda senosoidal fija (con sus respectivas amplitud y fase constantes).

2.5.1.4. Estructura de autocorrelación del proceso $AR(p)$

Ya se dijo también que la estructura de auto-correlación de un modelo $AR(p)$ está dada en términos de la solución a la ecuación en diferencias homogénea

$$\rho(h) - \phi_1\rho(h-1) - \ldots - \phi_p\rho(h-p) = 0, \quad h \geq p$$

En general, si $\alpha_1, \ldots, \alpha_r$ son los inversos de las raíces del polinomio auto-regresivo $\Upsilon(z)$, donde cada raíz tiene multiplicidad $m_1, m_2, \ldots, m_r$, respectivamente y $\sum_{i=1}^{r} m_i = p$, entonces la solución general de la ecución anterior es

$$\rho(h) = \alpha_1^h q_1(h) + \alpha_2^h q_2(h) + \ldots + \alpha_r^h q_r(h), \quad h \geq p$$

donde $q_j(h)$ es un polinomio de grado $m_j - 1$.

Por ejemplo, en el caso de un proceso $AR(2)$ se tienen los siguientes escenarios:

- El polinomio auto-regresivo tiene dos raíces reales diferentes, cada una con multiplicidad $m_1 = m_2 = 1$. Entonces, la función de autocorrelación tiene la forma

$$\rho(h) = a\alpha_1^h + b\alpha_2^h, \quad h \geq 2$$

donde a, b son constantes y α_1, α_2 son los inversos multiplicativos de las raíces de dichos polinomio. Bajo la condición de estacionariedad, esta función de autocorrelación decrece exponencialmente cuando h tiende a ∞ y su comportamiento es parecido al de la función de predicción (excepto quizá después de un re-escalamiento). Las constantes a, b se determinan especificando las dos condiciones iniciales $\rho(0) = 1$ y $\rho(-1) = \phi_1/(1 - \phi_2)$.

- El polinomio auto-regresivo tiene una raíz real con multiplicidad $m = 2$ y por tanto la función de autocorrelación tiene la forma

$$\rho(h) = (a + bh)\alpha_1^h, \quad h \geq 2$$

donde a, b son constantes y α_1 es el inverso e de la raíz. Bajo la condición de estacionariedad, esta función de autocorrelación también decrece exponencialmente cuando h tiende a ∞.

- El polinomio auto-regresivo tiene dos raíces complejas (conjugadas). En este caso, las raíces recíprocas se pueden escribir como $\alpha_1 = r \cdot exp(i\omega)$ y $\alpha_2 = r \cdot exp(-i\omega)$; por tanto, la función de autocorrelación tiene la forma

$$\rho(h) = ar^h cos(h\omega + b), \quad h \geq 2$$

donde a, b son constantes. Bajo la condición de estacionariedad las funciones de autocorrelación y de predicción tienen comportamiento de tipo coseno amortiguado.

2.6. Modelos $ARMA$

Definición 2.6.1. *(Proceso $ARMA(p, q)$)*

Se dice que $\{X_t\}$ es un proceso $ARMA(p, q)$ si $\{X_t\}$ es estacionario y para todo $t \in \mathbb{Z}$

$$X_t - \phi_1 X_{t-1} - \ldots - \phi_p X_{t-p} = \epsilon_t + \theta_1 \epsilon_{t-1} + \ldots + \theta_q \epsilon_{t-p} \tag{2.16}$$

donde $\{\epsilon_t\}$ es una ruido blanco con media cero y varianza σ^2 y los polinomios $1 - \phi_1 z - \ldots - \phi_p z^p$ y $1 + \theta_1 z + \ldots + \theta_q z^q$ no tienen factores mónicos en común.

Se dice que el proceso $\{X_t\}$ es un proceso $ARMA(p, q)$ con media μ si $\{X_t - \mu\}$ es un proceso $ARMA(p, q)$.

Es útil algunas veces expresar a la ecuación (2.16) en términos del operador de retraso

$$\phi(B)(X_t) = \theta(B)(\epsilon_t) \tag{2.17}$$

donde $\phi(z) = 1 - \phi_1 z - \ldots - \phi_p z^p$ y $\theta(z) = 1 + \theta_1 z + \ldots + \theta_q z^q$

Observación 2.6.1.

- Se dice que la serie de tiempo $\{X_t\}$ es un proceso auto-regresivo de orden p (ó $AR(p)$) si $\theta(z) = 1$
- Se dice que la serie de tiempo $\{X_t\}$ es un proceso de medias móviles de orden q (ó $MA(q)$) si $\phi(z) = 1$

Una parte importante de la definición del proceso $ARMA(p, q)$ es que $\{X_t\}$ sea estacionario. Cuando se estudio al proceso $ARMA(1, 1)$ se dijo que hay solución estacionaria única de $X_t = \phi_1 X_{t-1} + \epsilon_t$ si y sólo si $|\phi_1| \neq 1$. La última condición es equivalente a decir que el polinomio $\phi(z) = 1 - \phi_1(z)$ sea distinto de cero para $z = \pm 1$. Dicha condición es análoga para los procesos $ARMA(p, q)$ y establece que es polinomio $\phi(z) = 1 - \phi_1 z - \ldots - \phi_p z^p$ sea distinto de cero para todo $z \in \mathbb{C}$ tal que $|z| = 1$ (la región en $\mathbb{C}$ que define esta condición se conoce como círculo unitario). Si $\phi(z) \neq 0$ para todo z en el círculo unitario (i.e. $|z| = 1$), existe $\delta > 0$ tal que

$$\frac{1}{\phi(z)} = \sum_{j=-\infty}^{\infty} \chi_j z^j \quad \text{para } 1 - \delta < |z| < 1 + \delta$$

con $\sum_{j=-\infty}^{\infty} |\chi_j| < \infty$. Entonces, se puede definir $1/\phi(B)$ como el filtro lineal con coeficientes absolutamente sumables

$$\frac{1}{\phi(B)} = \sum_{j=-\infty}^{\infty}$$

Aplicando el operador $\chi(B) := \frac{1}{\phi(B)}$ a ambos lados de (2.17) se obtiene

$$X_t = \chi(B)\phi(B)(X_t) = \chi(B)\theta(B)(\epsilon_t) = \psi(B)(\epsilon_t) = \sum_{j=-\infty}^{\infty} \psi_j \epsilon_{t-j} \tag{2.18}$$

donde $\psi(z) = \chi(z)\theta(z) = \sum_{j=-\infty}^{\infty} \psi_j z^j$. Usando un argumento similar al que se hizo cuando se estudió el $ARMA(1, 1)$ se tiene que $\psi(B)(\epsilon_t)$ es la única solución estacionaria de (2.16).

Proposición 2.6.1. *(Existencia y unicidad de solución estacionaria $ARMA(p,q)$)*

Una solución estacionaria $\{X_t\}$ de la ecuación (2.16) existe (y es la única solución estacionaria) si y sólo si

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

Cuando se estudió la causalidad del proceso $ARMA(1,1)$, i.e. si X_t se puede expresar en términos ϵ_s, $s \leq t$, se dijo que $\{X_t\}$ es causal si y sólo si $|\phi_1| < 1$. Para el $ARMA(p,q)$ la condición análoga es que $\phi(z) \neq 0$ para $|z| \leq 1$, i.e. si el valor absoluto de los ceros del polinomio auto-regresivo es mayor que 1.

Se dice que el proceso $ARMA(p,q)$ es causal (ó función causal de $\{\epsilon_t\}$ si existen constantes ψ_j tales que $\sum_{j=0}^{\infty} |\psi_j| < \infty$ y para todo $t \in \mathbb{Z}$

$$X_t = \sum_{j=0}^{\infty} \psi_j \epsilon_{t-j} \tag{2.19}$$

La condición de causalidad es equivalente a

$$\phi(z) = 1 - \phi_1 z - \ldots - \phi_p z^p \neq 0 \;\; \text{para todo } z \in \mathbb{C} \text{ tal que } |z| \leq 1 \tag{2.20}$$

La demostración de la equivalencia entre causalidad y (2.20) se sigue de propiedades de las series de potencias. De la expresión (2.18), se puede observar que $\{X_t\}$ es causal si y sólo si $\chi(z) = \frac{1}{\phi(z)} = \sum_{j=0}^{\infty} \chi_j z^j$ (suponiendo que $\phi(z)$ y $\theta(z)$ no tienen factores mónicos en común); pero esto es precisamente equivalente a (2.20).

La sucesión $\{\psi_j\}$ en (2.19) está determinada por la relación $\psi(z) = \sum_{j=0}^{\infty} \psi_j z^j = \frac{\theta(z)}{\phi(z)}$, o equivalentemente por la identidad

$$(1 - \phi_1 z - \ldots - \phi_p z^p)(\psi_0 + \psi_1 z + \ldots) = 1 + \theta_1 z + \ldots + \theta_p z^p$$

Igualando los coeficientes de z^j, $j \in \mathbb{N}$

$$\begin{aligned} 1 &= \psi_0 \\ \theta_1 &= \psi_1 - \psi_0 \phi_1 \\ \theta_2 &= \psi_2 - \psi_1 \phi_1 - \psi_0 \phi_2 \\ &\ldots \end{aligned} \tag{2.21}$$

Ó equivalentemente

$$\psi_j - \sum_{k=1}^{p} \phi_k \psi_{j-k} = \theta_j, \quad j \in \mathbb{N} \tag{2.22}$$

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

La invertibilidad permite expresar a ϵ_t en términos de X_s, $s \leq t$ y tiene una caracterización similar en términos del polinomio de medias móviles.

Se dice que el proceso $ARMA(p,q)$ es invertible si existen constantes π_j tales que $\sum_{j=0}^{\infty} |\pi_j| < \infty$ y para todo $t \in \mathbb{Z}$

$$\epsilon_t = \sum_{j=0}^{\infty} \psi_j X_{t-j} \tag{2.23}$$

La condición de invertibilidad es equivalente a

$$\theta(z) = 1 + \theta_1 z + \ldots + \theta_q z^q \neq 0 \ \text{ para todo } z \in \mathbb{C} \text{ tal que } |z| \leq 1 \tag{2.24}$$

Intercambiando los polinomios auto-regresivo y de medias móviles en (2.22) se tiene que

$$\pi_j + \sum_{k=1}^{q} \theta_k \pi_{j-k} = -\phi_j, \quad j \in \mathbb{N} \tag{2.25}$$

donde $\phi_0 := -1$, para $j > p$ $\phi_j := 0$ y para $j < 0$ $\pi_j := 0$.

2.6.1. La función de autocovarianza de los proceso $ARMA(p,q)$

A continuación se obtendrá la función de autocovarianza del proceso $ARMA(p,q)$ causal definido por

$$\phi(B)(X_t) = \theta(B)(\epsilon_t) \tag{2.26}$$

donde $\{\epsilon_t\}$ es un ruido blanco con media cero y varianza σ^2, $\phi(z) = 1 - \phi_1 z - \ldots - \phi_p z^p$ y $\theta(z) = 1 + \theta_1 z + \ldots + \theta_q z^q$. La suposición de causalidad implica que

$$X_t = \sum_{j=0}^{\infty} \psi_j \epsilon_{t-j} \tag{2.27}$$

donde $\sum_{j=0}^{\infty} \psi_j z^j = \frac{\theta(z)}{\phi(z)}, \quad |z| \leq 1$ y ya se discutió cómo obtener la sucesión $\{\psi_j\}$.

- Método 1. De la proposición 2.4.1 y la representación (2.27) se tiene que

$$\gamma_X(h) = \mathbb{E}(X_{t+h}X_t) = \sigma^2 \sum_{j=0}^{\infty} \psi_j \psi_{j+|h|}$$

Por ejemplo, para el proceso $ARMA(1,1)$, $X_t - \phi X_{t-1} = \epsilon_t + \theta\epsilon_{t-1}$ con $|\phi| < 1$

$$\begin{aligned} \gamma_X(0) &= \sigma^2 \sum_{j=0}^{\infty} \psi_j^2 = \sigma^2 \left(1 + (\phi+\theta)^2 \sum_{j=0}^{\infty} \phi^{2j}\right) \\ &= \sigma^2 \left(1 + \frac{(\phi+\theta)^2}{1-\phi^2}\right) \end{aligned}$$

$$\begin{aligned} \gamma_X(1) &= \sigma^2 \sum_{j=0}^{\infty} \psi_j \psi_{j+1} = \sigma^2 \left(\phi + \theta + (\phi+\theta)^2 \phi \sum_{j=0}^{\infty} \phi^{2j}\right) \\ &= \sigma^2 \left(\phi + \theta + \frac{(\phi+\theta)^2\phi}{1-\phi^2}\right) \end{aligned}$$

y $\gamma_X(h) = \phi^{h-1}\gamma_X(1)$, $h \geq 2$. Por ejemplo, para el proceso $MA(q)$, $X_t = \epsilon_t + \theta_1\epsilon_{t-1} + \ldots + \theta_q\epsilon_{t-q}$

$$\gamma_X(h) = \begin{cases} \sigma^2 \sum_{j=0}^{q-|h|} \theta_j \theta_{j+|h|} & \text{si } |h| \leq q \\ 0 & \text{si } |h| > q \end{cases}$$

- Método 2. Si se multiplican ambos lados de

$$X_t - \phi_1 X_{t-1} - \ldots - \phi_p X_{t-p} = \epsilon_t + \theta_1\epsilon_{t-1} + \ldots + \theta_q\epsilon_{t-q}$$

por X_{t-k}, $k \in \mathbb{N}$ y se obtienen esperanzas de ambos lados se tiene que

$$\mathbb{E}(X_tX_{t-k}) - \phi_1\mathbb{E}(X_{t-1}X_{t-k}) - \ldots - \phi_p\mathbb{E}(X_{t-p}X_{t-k}) = \mathbb{E}(\epsilon_tX_{t-k}) + \theta_1\mathbb{E}(\epsilon_{t-1}X_{t-k}) + \ldots + \theta_q\mathbb{E}(\epsilon_{t-q}X_{t-k})$$

entonces,

$$\gamma_X(k) - \phi_1\gamma_X(k-1) - \ldots - \phi_p\gamma_X(k-p) = \sigma^2 \sum_{j=0}^{\infty} \theta_{k+j}\psi_j, \quad k \in \{0,1,\ldots,m-1\} \tag{2.28}$$

y

$$\gamma_X(k) - \phi_1\gamma_X(k-1) - \ldots - \phi_p\gamma_X(k-p) = 0, \quad k \in \{m, m+1, \ldots\} \tag{2.29}$$

donde $m := \text{máx}\{p, q+1\}$, $\psi_j := 0$ para $j < 0$, $\theta_0 := 1$ y $\theta_j := 0$ para $j \notin \{0,\ldots,q\}$

$$\gamma_X(h) = \alpha_1\xi_1^{-h} + \alpha_2\xi_2^{-h} + \ldots + \alpha_p\xi_p^{-h}, \quad h \geq m-p \tag{2.30}$$

- Método 3. Las autocovarianzas se pueden encontrar resolviendo $p+1$ ecuaciones de (2.28) y (2.29) para $\gamma_X(0),\ldots,\gamma_X(p)$ y se resulven las siguientes ecuaciones para encontrar recursivamente $\gamma_X(p+1), \gamma_X(p+2), \ldots$.

2.7. Estimación

Ahora, se definirán algunos estimadores de las funciones de autocovarianza, autocorrelación, covarianza cruzada y correlación cruzada a partir de datos en forma de serie de tiempo. Se supondrá que se tienen datos $x_{1:T}$. El estimador usual de la función de autocovarianza es la autocovarianza muestral que para $h > 0$ se define como

$$\hat{\gamma}_X(h) := \frac{1}{T}\sum_{t=1}^{T-h}(x_{t+h} - \bar{x})(x_t - \bar{x}), \quad \text{para } h \in \mathbb{N}$$

donde $\bar{x} = \frac{1}{T}\sum_{t=1}^{T} x_t$ es la media muestral. Con este estimador de la función de autocovarianza se puede obtener un estimador para la función de autocorrelación mediante

$$\hat{\rho}_X(h) = \frac{\hat{\gamma}_X(h)}{\hat{\gamma}_X(0)} \quad \text{para } h \in \mathbb{N}$$

2.7.1. Propiedades de la media y la función de autocovarianza muestrales

Un proceso $\{X_t\}$ estacionario está caracterizado, al menos en términos de segundo orden, por su media μ y su función de autocovarianza $\gamma(\cdot)$; por tanto, la estimación de μ, $\gamma(\cdot)$ y $\rho(\cdot)$ a partir de las observaciones $X_1, \ldots, X_T$ tiene un papel importante en problemas de inferencia y en particular en el problema de construir un modelo apropiado para los datos. A continuación, se estudiarán algunas propiedades de los estimadores muestrales $\bar{X}$, $\hat{\rho}$, respectivamente.

El estimador por momentos de la media μ de un proceso estacionario es la media muestral

$$\bar{X}_T = \frac{1}{T}\sum_{t=1}^{T} X_t$$

Que además es un estimador insesgado (de μ) pues

$$\mathbb{E}(\bar{X}_T) = \frac{1}{T}\sum_{t=1}^{T}\mathbb{E}(X_t) = \frac{1}{T}\mu T = \mu$$

El error cuadrático medio de $\bar{X}_T$ es

$$\begin{aligned} ECM_{\bar{X}_T}(\mu) &= Var(\bar{X}_T) = \frac{1}{T^2} Var(\sum_{t=1}^{T} X_t) \\ &= \frac{1}{T^2} \sum_{i=1}^{T} \sum_{j=1}^{T} Cov(X_i, X_j) = \frac{1}{T^2} \sum_{i-j=-T}^{T} (T - |i-j|)\gamma_X(i-j) \\ &= \frac{1}{T} \sum_{h=-T}^{T} \left(1 - \frac{|h|}{T}\right) \gamma_X(h) \end{aligned}$$

Nótese que si $\lim_{h\to\infty} \gamma_X(h) = 0$, entonces el lado derecho de (2.31) converge a cero, así que $\bar{X}_T$ converge en media cuadrática a μ. Si $\sum_{h=-\infty}^{\infty} |\sigma_X(h)| < \infty$, entonces

$$\lim_{T\to\infty} T \cdot ECM_{\bar{X}_T} = \sum_{|h|<\infty} \gamma_X(h)$$

Proposición 2.7.1.

Si $\{X_t\}$ es una serie de tiempo estacionaria con media μ y función de autocovarianza $\gamma_X(\cdot)$, entonces cuanto $T \to \infty$

$$Var(\bar{X}_T) \to 0 \quad \text{si } \gamma_X(T) \to 0$$

$$T \cdot Var(\bar{X}_T) \to \sum_{|h|<\infty} \gamma_X(h) \quad \text{si } \sum_{h=-\infty}^{\infty} |\sigma_X(h)| < \infty$$

2.7.2. Estimación ML, MAP y LS

Es posible obtener estimadores puntuales de los parámetros del modelo maximizando la función de verosimilitud o la distribución posterior completa (*full posterior distribution*). En la literatura se proponen muchos métodos y algoritmos para encontrar dichos estimados. A continuación, se dará un breve repaso de dichos métodos, ejemplificando con el $AR(1)$.

Se puede obtener un estimador de $\boldsymbol{\theta}$, $\hat{\boldsymbol{\theta}}$, maximizando la función de verosimilitud $p(x_{1:n}|\boldsymbol{\theta})$ con respecto a $\boldsymbol{\theta}$. En este caso se denota por $\hat{\boldsymbol{\theta}} = \boldsymbol{\theta}_{MV}$. Análogamente, si se maximiza la distribución posterior, $p(\boldsymbol{\theta}|x_{1:n})$ (en vez de la función de verosimilitud) se obtiene el estimado máximo aposteriori para $\boldsymbol{\theta}$ que se denota por $\hat{\boldsymbol{\theta}} = \boldsymbol{\theta}_{MAP}$.

Muchas veces, la función de verosimilitud y la distribución posterior son funciones de $\boldsymbol{\theta}$ con formas complicadas (por ejemplo no-lineales) y por tanto se necesitan algunos métodos numéricos, tal como el método de Newton-Raphson ó el método de scoring, para obtener $\boldsymbol{\theta}_{MV}$ y $\boldsymbol{\theta}_{MAP}$. En general, el algoritmo de Newton-Raphson se puede resumir de la siguiente

forma: sea $g(\boldsymbol{\theta})$ una función de $\boldsymbol{\theta} = (\theta_1, \ldots, \theta_k)'$ que se quiere maximizar y sea $\hat{\boldsymbol{\theta}}$ el máximo. En la m-ésima iteración se obtiene

$$\boldsymbol{\theta}^{(m)} = \boldsymbol{\theta}^{(m-1)} - [g''(\boldsymbol{\theta}^{(m-1)})]^{-1}[g'(\boldsymbol{\theta}^{(m-1)})] \tag{2.31}$$

donde $g'(\boldsymbol{\theta})$ y $g''(\boldsymbol{\theta})$ denotan las primera y segunda derivadas parciales de g, i.e. $g'(\boldsymbol{\theta})$ es un vector de k entradas dado por

$$g'(\boldsymbol{\theta}) = \left(\frac{\partial}{\partial\theta_1}g(\boldsymbol{\theta}), \ldots, \frac{\partial}{\partial\theta_k}g(\boldsymbol{\theta})\right)'$$

y $g''(\boldsymbol{\theta})$ es una matriz de $k \times k$ de las derivadas de segundo orden dada por $[\frac{\partial^2}{\partial\theta_i\partial\theta_j}g(\boldsymbol{\theta})]$ para $i, j \in \{1, \ldots, k\}$. Bajo ciertas condiciones, este algoritmo produce una sucesión $\boldsymbol{\theta}^{(1)}, \boldsymbol{\theta}^{(2)}, \ldots$ que convergerá a $\hat{\boldsymbol{\theta}}$. Es importante comenzar con un buen valor inicial $\boldsymbol{\theta}^{(0)}$ ya que este algoritmo no necesariamente converge para valores en las cuales $-g''(\cdot)$ no es positivo definida. Un método alternativo es el método de scoring, en éste se reemplaza $g''(\boldsymbol{\theta})$ en (2.31) por la matriz de valores esperados $\mathbb{E}(g''(\boldsymbol{\theta}))$.

En muchas aplicaciones prácticas, especialmente cuando se trabaja con modelos que involucran muchos parámetros, no es útil resumir las inferencias en términos de la moda posterior conjunta. En vez de esto, se puede hacer en términos de las modas posteriores marginales. Por ejemplo, supóngase que se puede particionar los parámetros del modelo en dos conjuntos $\boldsymbol{\theta}_1$ y $\boldsymbol{\theta}_2$ de tal forma que $\boldsymbol{\theta} = (\boldsymbol{\theta}_1', \boldsymbol{\theta}_2')'$ y supóngase que se está interesado en $p(\boldsymbol{\theta}_2|x_{1:n})$. El algoritmo EM (Esperanza-Maximización) propuesto por Dempster, Laird & Rubin (1977) es muy útil cuando se trabaja con modelos para los cuales $p(\boldsymbol{\theta}_2|x_{1:n})$ es difícil de maximizar directamente pero es relativamente facil trabajar con $p(\boldsymbol{\theta}_1|\boldsymbol{\theta}_2, x_{1:n})$ y $p(\boldsymbol{\theta}_2|\boldsymbol{\theta}_1, x_{1:n})$. El algoritmo EM se puede describir de la siguiente manera:

1. Empezar con algún valor inicial $\boldsymbol{\theta}_2^{(0)}$.
2. Para $m = 1, 2, \ldots$
   - Calcular $\mathbb{E}^{(m-1)}[\log(p(\boldsymbol{\theta}_2, \boldsymbol{\theta}_1|x_{1:n}))]$ dada por la expresión

     $$\int p(\boldsymbol{\theta}_2, \boldsymbol{\theta}_1|x_{1:n})p(\boldsymbol{\theta}_1|\boldsymbol{\theta}_2^{(m-1)}, x_{1:n})d\boldsymbol{\theta}_1 \tag{2.32}$$

     Este se conoce como el paso E.
   - Hacer $\boldsymbol{\theta}_2^{(m)}$ al valor que maximice (2.32). Este se conoce como el paso M.

En cada iteración, el algoritmo satisface que $p(\boldsymbol{\theta}_2^{(m)}|x_{1:n}) \geq p(\boldsymbol{\theta}_2^{(m-1)}|x_{1:n})$. No hay garantía de que el algoritmo EM converja a la moda; en el caso de distribuciones multimodales, el algoritmo puede converger a una moda local. En la literatura hay varias

alternativas para evitar tal "convergencia" hacia sólo una moda local, tales como implementar el algoritmo con varios puntos iniciales aleatorios ó usando métodos de simulated annealing. Algunas generalizaciones del algoritmo EM son el algoritmo ECM (esperanza-condicionamiento-maximización), el algoritmo ECME (una variante del ECM en el cual ó se maximiza la log-densidad posterior ó se maximiza la esperanza de la log-densidad posterior) y los algoritmos SEM (EM suplementado) (ver Gelman, Carlin, Stern & Rubin (2004)) ó algunas versiones estocásticas del algoritmo EM como el MCEM (Monte-Carlo EM, ver Wei & Tanner (1990)).

2.7.3. Mínimos cuadrados tradicionales

Las metodologías de máxima verosimilitud y Bayesianas para ajustar auto-regresiones lineales son dos de los métodos estándar del análisis de regresión. Por esta razón, ahora se hará un pequeño recuento de las principales ideas y resultados para el estudio de series de tiempo.

Un modelo lineal con una respuesta univariada y $p > 0$ variables de regresión (o bien predictores o covariados) tiene la forma $x_i = \mathbf{f}_i'\boldsymbol{\beta} + \epsilon_i, \quad i = 1, 2, \ldots$ donde x_i es la i-ésima observación de la variable de respuesta, y tiene valores correspondientes de los regresores en el vector de diseño $\mathbf{f}_i' = (f_{i1}, \ldots, f_{ip})$. Se supone que los vectores de diseño son conocidos y fijos antes de observar las correspondientes respuestas. También se supone que los términos de error ϵ_i son independientes y tienen distribución $N(0, \sigma^2)$. Generalmente se desea estimar el vector de parámetros de regresión $\boldsymbol{\beta} = (\beta_1, \ldots, \beta_p)'$ y la varianza del error. El modelo para un conjunto de n respuestas observadas $\mathbf{x} = (x_1, \ldots, x_n)'$ es

$$\mathbf{x} = \mathbf{F}'\boldsymbol{\beta} + \boldsymbol{\epsilon}$$

donde $\mathbf{F}$ es una matriz de $p \times n$ cuya i-ésima columna es $\mathbf{f}_i$. Además, $\boldsymbol{\epsilon} = (\epsilon_1, \ldots, \epsilon_n)'$ tiene distribución normal n-variada con media $\bar{0}$ y matriz de varianzas-covarianzas $\sigma^2\mathbf{I}_n$.

La distribución de muestreo se define como

$$p(x_{1:n}) = \prod_{i=1}^{n} N(x_i|\mathbf{f}_i'\boldsymbol{\beta}, \sigma^2) = \frac{1}{2\pi\sigma^2} \exp\left\{-\frac{1}{2\sigma^2} Q(\mathbf{x}, \boldsymbol{\beta})\right\}$$

donde $N(x_i|\mathbf{f}_i'\boldsymbol{\beta}, \sigma^2)$ denota la densidad normal con media $\mathbf{f}_i'\boldsymbol{\beta}$ y varianza σ^2 evaluada en x_i y $Q(\mathbf{x}, \boldsymbol{\beta}) = (\mathbf{x} - \mathbf{F}'\boldsymbol{\beta})'(\mathbf{x} - \mathbf{F}'\boldsymbol{\beta}) = \sum_{i=1}^{n}(x_i - \mathbf{f}_i'\boldsymbol{\beta})^2$. Y con esto se obtiene una función de verosimilitud para $(\boldsymbol{\beta}, \sigma^2)$. También se puede escribir $Q(\mathbf{x}, \boldsymbol{\beta})$ como

$$Q(\mathbf{x}, \boldsymbol{\beta}) = (\boldsymbol{\beta} - \hat{\boldsymbol{\beta}})'\mathbf{F}\mathbf{F}'(\boldsymbol{\beta} - \hat{\boldsymbol{\beta}}) + R$$

donde $\hat{\boldsymbol{\beta}} = (\mathbf{F}\mathbf{F}')^{-1}\mathbf{F}\mathbf{x}$ y $R = (\mathbf{x} - \mathbf{F}'\hat{\boldsymbol{\beta}})'(\mathbf{x} - \mathbf{F}'\hat{\boldsymbol{\beta}})$. Esto supone que $\mathbf{F}$ tiene rango completo p, y en otro caso se debe usar una transformación lineal apropiada en los vectores de diseño

para hacer que $\mathbf{F}$ tenga rango completo y para que se reduzca la dimensión del modelo. Aquí $\hat{\boldsymbol{\beta}}$ es el estimador máximo verosímil de $\boldsymbol{\beta}$ y R/n es el estimador máximo verosímil de σ^2; aunque un estimador muy común en la práctica de $sigma^2$ es $\frac{R}{n-p}$ (donde $n-$ se asocia a los grados de libertad)

2.7.4. Series de tiempo: Una perspectiva bayesiana

Se supondrá que se tiene n observaciones $x_{1:n}$ de un proceso de series de tiempo univariado $\{X_t\}_{t\in\mathcal{T}}$. Suponóngase que para cada X_t se tiene una distribución de probabilidad que se puede escribir como función de algún parámetro o conjunto de parámetros $\boldsymbol{\theta}$ y que dicha dependencia se puede describir en términos de una función de densidad de probabilidad $p(x_t|\boldsymbol{\theta})$. Como siempre en un entorno de inferencia si se piensa a $p(x_t|\boldsymbol{\theta})$ como una función de $\boldsymbol{\theta}$ en vez de una función de x_t se conocerá a esta función como función de verosimilitud. Mediante el teorema de Bayes se puede obtener la densidad posterior de $\boldsymbol{\theta}$ dado x_t, $p(\boldsymbol{\theta}|x_t)$ de la siguiente forma

$$p(\boldsymbol{\theta}|x_t) = \frac{p(\boldsymbol{\theta})p(x_t|\boldsymbol{\theta})}{p(x_t)}$$

donde $p(x_t) = \int p(\boldsymbol{\theta})p(x_t|\boldsymbol{\theta})d\boldsymbol{\theta}$ se conoce como función de densidad predictiva. La distribución *a priori*, $p(\boldsymbol{\theta})$, es una forma de incorporar creencias previas acerca de $\boldsymbol{\theta}$ y el teorema de Bayes permite actualizar dichas creencias después de observar los datos.

En el análisis de series de tiempo el teorema de Bayes también se puede utilizar de forma secuencial de la siguiente manera: antes de tomar datos, las creencias previas acerca de $\boldsymbol{\theta}$ se expresan de forma probabilista a través de $p(\boldsymbol{\theta})$. Supóngase que se toma una primera observación al tiempo $t = 1$, x_1, y se obtendrá $p(\boldsymbol{\theta}|x_1)$ utilizando el teorema de Bayes. Una vez que se obtiene x_2 se puede obtener $p(\boldsymbol{\theta}|x_{1:2})$ de nuevo mediante el teorema de Bayes a través de $p(\boldsymbol{\theta}|x_{1:2}) \propto p(\boldsymbol{\theta})p(x_{1:2}|\boldsymbol{\theta})$. Si X_1 y X_2 son condicionalmente independientes dado $\boldsymbol{\theta}$ se puede escribir $p(\boldsymbol{\theta}|x_{1:2}) \propto p(\boldsymbol{\theta}|x_1)p(x_2|\boldsymbol{\theta})$. En forma análoga, se puede obtener $p(\boldsymbol{\theta}|x_{1:n})$ secuencialmente si todas las observaciones son independientes. Sin embargo, en el análisis de series de tiempo las observaciones no son independientes. Una suposición frecuente es que cada observación al tiempo t depende sólo de $\boldsymbol{\theta}$ y de la observación tomada al tiempo $t-1$, i.e. una estructura Markoviana. En este caso se tiene que

$$p(\boldsymbol{\theta}|x_{1:n}) \propto p(\boldsymbol{\theta})p(x_1|\boldsymbol{\theta})\prod_{t=2}^{n} p(x_t|x_{t-1},\boldsymbol{\theta})$$

Sin embargo, hay algunos modelos más generales en los cuales X_t depende de un número arbitrario de observaciones pasadas.

Una clase importante de modelos de series de tiempo es aquella en al cual los parámetros también están indexados por el tiempo. En este caso, cada observación se relaciona con

el parámetro $\boldsymbol{\theta}_t$ que evoluciona a través del tiempo. Esta clase se conoce como modelos dinámicos lineales. En este tipo de modelos es necesario definir un proceso que defina la evolución en el tiempo de $\boldsymbol{\theta}_t$. Por ejemplo, considérese el modelo conocido como $TVAR(1)$ que se define como

$$\begin{cases} X_t = \phi_t X_{t-1} + \epsilon_t \\ \phi_t = \phi_{t-1} + \nu_t \end{cases}$$

donde ϵ_t y ν_t son independientes en el tiempo y mutuamente independientes y $\epsilon_t \sim N(0, \upsilon)$ y $\nu_t \sim N(0, w)$. En este modelo algunas distribuciones de interés son las distribuciones posteriores al tiempo t, $p(\phi_t|x_{1:t})$ y $p(\upsilon|x_{1:t})$; también $p(\phi_t|x_{1:n})$ y la distribución de predicción $p(x_{t+h}|x_{1:t})$.

2.7.5. Asignación del orden p en modelo auto-regresivos

Se puede hacen un análisis inferencial para diferentes valores p (el orden del modelo auto-regresivo) y analizarse las variaciones en la inferencia cuando el parámetro p cambia. Los valores grandes para p están limitados por el tamaño de la muestra y de hecho, sólo datos con ciertas características permitirán inferencias posteriores de referencia; se puede hacer inferencia en un gran número de parámetros (con un tamaño de muestra fijo)sólo con a prioris propias e informativas para dichos parámetros, en caso de no ser así, se tendrán problemas análogos a los de regresión de sobre-ajuste y colinelidad.

El simplemente incrementar secuencialmente p y analizando los residuales ajustados, los cambios en estimaciones de los parámetros posteriores, etc., puede ser interesante en la determinación del orden de un modelo auto-regresivo (aunque un poco inocente e heurístico). Se pueden calcular varias características numéricas que permitan evaluar de forma más objetiva estos cambios inferenciales. Dos de las más conocidas son el Criterio de Información de Akaike (AIC) y el Criterio de Información Bayesiano (BIC) (Akaike (1969), Akaike (1974), Schwarz (1978)). Entonces, el BIC y el AIC describirán de manera más formal una medida Bayesiana del ajuste del modelo. Cuando se está comparando modelos con diferentes números de parámetros, se hace con una muestra fija común; por tanto, se fija un orden máximo p^* y se comparan modelos con distintos órdenes $p \leq p^*$.

Para un orden de modelo seleccionado, p, la dependencia explícita sobre p se hace explícita escribiendo $\hat{\boldsymbol{\phi}}_p$ para los parámetros máximo verosímiles del AR y s_p^2 para la correspondiente varianza de las innovaciones, i.e. la suma residual de los cuadrados, divididos entre $n - p$ (con $n = T - p^*$). La medida AIC del ajuste del modelo se tomará como $2p + n \cdot log(s_p^2)$ y el BIC como $log(n)p + n \cdot log(s_p^2)$. Los valores de p que inducen AIC y BIC pequeños serán indicativos de que hay un buen ajuste de modelo cuando se trabaja en proceso (puramente) auto-regresivos (por supuesto, pueden ser malos modelos si se comparán con otras clases de modelos). Valores grandes de p tenderán a dar estimados

más pequeños de la varianza. El BIC tiende a seleccionar modelos más simples que el AIC.

En un análisis Bayesiano formal, el orden se considera como un parámetro desconocido y por tanto se puede "actualizar" cualquier a priori de p con el uso de las funciones de verosimilitud y la densidad marginal de los datos

$$p(x_{(p^+1:T)}|x_{1:p^*};p) = \int p(x_{(p^+1:T)}|\boldsymbol{\phi}_p, v, x_{1:p^*})p(\boldsymbol{\phi}_p, v)d\boldsymbol{\phi}_p dv$$

donde $p(\boldsymbol{\phi}_p, v)$ es una apriori bajo el modelo $AR(p)$ y recuérdese que la dimensión de $\boldsymbol{\phi}_p$ depende de p. Dadas aprioris propias $p(\boldsymbol{\phi}_p, v)$ en un rango de valores de $p \leq p^*$, se obtendrán medidas numéricas directas de ajuste relativo con respecto a estas densidades marginales que definen una función de verosimilitud para el orden del modelo. Sin emabargo, para llevar a cabo esto se requiere de una a priori propia $p(\boldsymbol{\phi}_p, v)$ que depende de la dimensión de p y esta dependencia es importante para determinar la función de verosilimitud resultante. El uso de a prioris "tradicionales" hace que este análisis no sea válido dada la "impropiedad".

2.7.6. Un poco de simulación

Una vez que se ha dicho como se puede abordar el estudio de las series de tiempo desde una pespectiva Bayesiana es totalmente pertinente mencionar algunas primeras técnicas de simulación útiles en el análisis de series de tiempo.

2.7.6.1. El algoritmo Metropolis-Hastings

Supóngase que la distribución posterior objetivo, $p(\boldsymbol{\theta}|x_{1:n})$, se puede calcular (excepto quizá por una constante de normalización). Mediante el algoritmo Metropolis-Hastings se obtiene una sucesión de simulaciones $\boldsymbol{\theta}^{(1)}, \boldsymbol{\theta}^{(2)}, \ldots$ cuya distribución converge a la distribución objetivo. Cada sucesión se puede considerar como la realización de una cadena de Markov cuya distribución estacionaria es $p(\boldsymbol{\theta}|x_{1:n})$. Este algoritmo se puede resumir de la siguiente manera:

1. Simular un punto inicial $\boldsymbol{\theta}^{(0)}$ tal que $p(\boldsymbol{\theta}^{(0)}|x_{1:n}) > 0$ de una distribución incial $p_0(\boldsymbol{\theta})$.

2. Para $m = 1, 2, \ldots$

   (i) Simular un candidato $\boldsymbol{\theta}^*$ de una distribución de saltos $J(\boldsymbol{\theta}^*|\boldsymbol{\theta}^{(m-1)})$. Si la distribución J es simétrica, i.e. si $J(\boldsymbol{\theta}_a|\boldsymbol{\theta}_b) = J(\boldsymbol{\theta}_b|\boldsymbol{\theta}_a)$ para cualesquiera $\boldsymbol{\theta}_a, \boldsymbol{\theta}_b$ y m, entonces este algoritmo de simulación que se conoce como algoritmo Metropolis. Si J_m no es simétrica, este algoritmo de simulación se conoce como algoritmo Metropolis-Hastings.

(ii) Calcúlese el cociente

$$r := \frac{p(\boldsymbol{\theta}^*|x_{1:n})/J(\boldsymbol{\theta}^*|\boldsymbol{\theta}^{(m-1)})}{p(\boldsymbol{\theta}^{(m-1)}|x_{1:n})/J(\boldsymbol{\theta}^{(m-1)}|\boldsymbol{\theta}^*)}$$

(iii) Hágase

$$\boldsymbol{\theta}^{(m)} = \begin{cases} \boldsymbol{\theta}^* & \text{con probabilidad igual a mín}\{r, 1\}; \\ \boldsymbol{\theta}^{(m-1)} & \text{en otro caso} \end{cases}$$

Una distribución de saltos ideal es una en la que es fácil simular y facilita el cálculo de r. La distribución de saltos $J(\cdot|\cdot)$ debe ser tal que en cada salto se mueva una distancia razonable en el espacio parametral y por tanto, la caminata aleatoria no sea muy lenta y también que no se rechacen demasiados saltos.

2.7.6.2. Gibbs sampling

Supóngase que $\boldsymbol{\theta}$ tiene k componentes, es decir, $\boldsymbol{\theta} = (\theta_1, \theta_2, \ldots, \theta_k)$. El Gibbs sampler se puede ver como un caso especial del algortimo Metropolis-Hastings para el cual la distribución de saltos en cada iteración m es una función $p(\theta_j^*|\theta_{-j}^{(m-1)}, x_{1:n})$, donde θ_{-j} denota un vector con todas las componentes de $\boldsymbol{\theta}$ excepto la componente θ_j. En otras palabras, para componente de θ se realiza un paso del tipo del algoritmo Metropolis para el cual la distribución de saltos está dada por

$$J_j(\boldsymbol{\theta}^*|\boldsymbol{\theta}^{(m-1)}) = \begin{cases} p(\theta_j^*|\theta_{-j}^{(m-1)}, x_{1:n}) & \text{si } \theta_{-j}^* = \theta_{-j}^{(m-1)}; \\ 0 & \text{en otro caso} \end{cases}$$

y por tanto, $r = 1$ y el salto siempre se acepta.

Si no es posible simular de $p(\theta_j^*|\theta_{-j}^{(m-1)}, x_{1:n})$ se puede utilizar una aproximación $g(\theta_j^*|\theta_{-j}^{(m-1)})$; sin embargo, en este caso será necesario calcular el cociente r del algoritmo Metropolis.

2.8. Predicción en series de tiempo estacionarias

Ahora, se considerará el problema de predecir valores X_{T+h}, $h > 0$, de una serie de tiempo estacionaria con media, μ, conocida y función de autocovarianza γ_X en términos de valores de $\{X_T, \ldots, X_1\}$ hasta el tiempo T. El objetivo es encontrar la combinación lineal de $1, X_T, X_{T-1}, \ldots, X_1$ que predice X_{T+h} con el mínimo error cuadrático medio. El mejor predictor lineal en términos de $1, X_T, X_{T-1}, \ldots, X_1$ se denotará por $P_T X_{T+h}$ y tiene la forma

$$P_T X_{T+h} = a_0 + a_1 X_T + \ldots + a_T X_1 \tag{2.33}$$

Entonces, se tienen que determinar los coeficientes $a_0, a_1, \ldots, a_T$ que minimice

$$\Upsilon(a_0, a_1, \ldots, a_T) = \mathbb{E}[(X_{T+h} - a_0 - a_1 X_T - \ldots - a_T X_1)^2] \tag{2.34}$$

Como Υ es una función cuadrática de $a_0, a_1, \ldots, a_T$ y está acotada por debajo por cero, existe $(a_0, a_1, \ldots, a_T)$ que minimiza Υ y dicho mínimo satisface

$$\frac{\partial}{\partial a_j}\Upsilon(a_0, a_1, \ldots, a_T) = 0, \quad j = 0, 1, \ldots, T \tag{2.35}$$

Ó equivalentemente,

$$\mathbb{E}\left[X_{T+h} - a_0 - \sum_{i=1}^{T} a_i X_{T+1-i}\right] = 0 \tag{2.36}$$

$$\mathbb{E}\left[\left(X_{T+h} - a_0 - \sum_{i=1}^{T} a_i X_{T+1-i}\right) X_{T+1-j}\right] = 0, \quad j = 1, 2, \ldots, T \tag{2.37}$$

Es decir, $a_0 = \mu(1 - \sum_{i=1}^{T} a_i)$ y

$$\Gamma_T \mathbf{a}_T = \boldsymbol{\gamma}_T(h) \tag{2.38}$$

donde $\mathbf{a}_T = (a_1, \ldots, a_T)'$, $\Gamma_T = [\gamma_X(i-j)]_{i,j=1}^{T}$ y

$$\boldsymbol{\gamma}_T(h) = (\gamma_X(h), \gamma_X(h+1), \ldots, \gamma_X(h+T-1))'$$

Por tanto,

$$P_T X_{T+h} = \mu + \sum_{i=1}^{T} a_i (X_{T+1-i} - \mu) \tag{2.39}$$

donde $\mathbf{a}_T$ satisface (2.38). De (2.39) se tiene que

$$\mathbb{E}(P_T X_{T+h}) = \mu + \sum_{i=1}^{T} a_i \mathbb{E}(X_{T+1-i} - \mu) = \mu$$

Entonces, $\mathbb{E}(P_T X_{T+h} - X_{T+h}) = 0$. Y también,

$$\begin{aligned}
\mathbb{E}[(X_{T+h} - P_T X_{T+h})^2] &= Var(P_T X_{T+h}) \\
&= \gamma_X(0) - 2\sum_{i=1}^{T} a_i \gamma_X(h+i-1) + \sum_{i=1}^{T}\sum_{j=1}^{T} a_i \gamma_X(i-j) a_j \\
&= \gamma_X(0) - \mathbf{a}_T' \boldsymbol{\gamma}_T(h)
\end{aligned}$$

Observación 2.8.1.

Para mostrar que las ecuaciones (2.36) y (2.37) determinan de manera única a $P_T X_{T+h}$ sean $a_0^{(1)}, a_1^{(1)}, \ldots, a_T^{(1)}$ y $a_0^{(2)}, a_1^{(2)}, \ldots, a_T^{(2)}$ dos soluciones y sea Z la diferencia entre los predictores correspondientes, i.e.

$$Z = a_0^{(1)} - a_0^{(2)} + \sum_{j=1}^{T} (a_j^{(1)} - a_j^{(2)}) X_{T+1-j}$$

Entonces,

$$Z^2 = Z\left(a_0^{(1)} - a_0^{(2)} + \sum_{j=1}^{T} (a_j^{(1)} - a_j^{(2)}) X_{T+1-j} \right)$$

pero por (2.36) y (2.37) $\mathbb{E}(Z) = 0$ y $\mathbb{E}(Z X_{T+1-j}) = 0$ para $j = 1, \ldots, T$. Por tanto $\mathbb{E}(Z^2) = 0$ y por tanto $Z = 0$ ∇.

Se puede resumir lo anterior en la siguiente proposición.

Proposición 2.8.1. *(Propiedades de $P_T X_{T+h}$)*

1. $P_T X_{T+h} = \mu + \sum_{i=1}^{T} a_i (X_{T+1-i} - \mu)$, donde $\mathbf{a}_T = (a_1, \ldots, a_T)'$ satisface (2.38)
2. $\mathbb{E}[(X_{T+h} - P_T X_{T+h})^2] = \gamma_X(0) - \mathbf{a}_T' \boldsymbol{\gamma}_T(h)$, dondo $\boldsymbol{\gamma}_T(h) = (\gamma_X(h), \gamma_X(h+1), \ldots, \gamma_X(h+T-1))'$
3. $\mathbb{E}(X_{T+h} - P_T X_{T+h}) = 0$
4. $\mathbb{E}[(X_{T+h} - P_T X_{T+h}) X_j] = 0$, $j = 1, \ldots, T$.

Ejemplo 2.8.1. *(Predicción a un paso del proceso $AR(1)$)*

Considérese la serie estacionaria $\{X_t\}_{t \in \mathbb{Z}}$ tal que

$$X_t = \phi X_{t-1} + \epsilon_t, \quad |\phi| < 1$$

donde $\{\epsilon_t\}$ es un ruido blanco con media cero y varianza σ^2. De (2.38) y (2.39) se tiene que el mejor predictor lineal de X_{T+1} en términos de $\{1, X_T, \ldots, X_1\}$ es $P_T X_{T+1} = \mathbf{a}_T' \mathbf{X}_T$ donde $\mathbf{X}_T = (X_T, \ldots, X_1)'$ y $\mathbf{a}$ satisface

$$\begin{pmatrix} 1 & \phi & \phi^2 & \cdots & \phi^{T-1} \\ \phi & 1 & \phi & \cdots & \phi^{T-2} \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ \phi^{T-1} & \phi^{T-2} & \phi^{T-3} & \cdots & 1 \end{pmatrix} \begin{pmatrix} a_1 \\ a_2 \\ \vdots \\ a_T \end{pmatrix} = \begin{pmatrix} \phi \\ \phi^2 \\ \vdots \\ \phi^T \end{pmatrix} \tag{2.40}$$

Nótese que una solución de (2.40) es $a_T = (\phi, 0, \ldots, 0)'$ y por tanto el mejor predictor lineal de X_{T+1} en términos de $\{X_1, \ldots, X_T\}$ es

$$P_T X_{T+1} = \mathbf{a}_T' \mathbf{X}_T = \phi X_T$$

cuyo error cuadrático medio está dado por

$$\mathbb{E}[(X_{T+1} - P_T X_{T+1})^2] = \gamma_X(0) - \mathbf{a}'\boldsymbol{\gamma}_T(1) = \frac{\sigma^2}{1-\phi^2} - \phi\gamma_X(1) = \sigma^2$$

∇

2.8.1. Predicción de variables aleatorias de segundo orden

Supónagase que Y y $W_T, \ldots, W_1$ son cualesquiera variables aleatorias con segundo momento finito. Supóngase que $\mu := \mathbb{E}(Y)$, $\mu_i := \mathbb{E}(W_i)$, $Var(Y)$, $Cov(Y, W_i)$ y $Cov(W_i, W_j)$ son conocidas. Defínanse los vectores $\mathbf{W} = (W_T, \ldots, W_1)'$, $\mu_W := (\mu_T, \ldots, \mu_1)'$,

$$\boldsymbol{\gamma} := Cov(Y, \mathbf{W}) = (Cov(Y, W_T), Cov(Y, W_{T-1}), \ldots, Cov(Y, W_1))'$$

y

$$\boldsymbol{\Gamma} := Cov(\mathbf{W}, \mathbf{W}) = [Cov(W_{T+1-i}, W_{T+1-j})]_{i,j=1}^{T}$$

Con argumentos análogos a los que se usaron para encontrar al mejor predictor lineal $P_T X_{T+h}$, se puede probar que el mejor predictor lineal (en el sentido de media cuadrática) de Y en términos de $1, W_T, \ldots, W_1$, $\wp(Y|\mathbf{W})$, está dado por

$$\wp(Y|\mathbf{W}) = \mu_y + \mathbf{a}'(\mathbf{W} - \mu_W) \tag{2.41}$$

donde $\mathbf{a} = (a_1, \ldots, a_T)'$ es cualquier solución de

$$\boldsymbol{\Gamma}\mathbf{a} = \boldsymbol{\gamma} \tag{2.42}$$

El error cuadrático medio del predictor es

$$\mathbb{E}[(Y - \wp(Y|\mathbf{W}))^2] = Var(Y) - \mathbf{a}'\boldsymbol{\gamma}$$

Ejemplo 2.8.2. *(Estimación de un valor faltante)*

Considérese la serie estacionaria $\{X_t\}_{t\in\mathbb{Z}}$ tal que

$$X_t = \phi X_{t-1} + \epsilon_t, \quad |\phi| < 1$$

donde $\{\epsilon_t\}$ es un ruido blanco con media cero y varianza σ^2. Supóngase que se observa la serie en los tiempos 1 y 3 y se desea usar estas observaciones para encontrar la combinación lineal de X_1 y X_3 que estime X_2 con el error cuadrático más pequeño. Se puede resolver este problema directamente a partir de (2.41) y (2.42) haciendo $Y = X_2$ y $\mathbf{W} = (X_1, X_3)'$. Es decir, con la expresión

$$\begin{pmatrix} 1 & \phi^2 \\ \phi^2 & 1 \end{pmatrix} \mathbf{a} = \begin{pmatrix} \phi \\ \phi \end{pmatrix}$$

con

$$\mathbf{a} = \frac{1}{1+\phi^2} \begin{pmatrix} \phi \\ \phi \end{pmatrix}$$

una solución. Por tanto, el mejor estimador de X_2 está dado por

$$\wp(X_2|\mathbf{W}) = \frac{\phi}{1+\phi^2}(X_1 + X_3)$$

con un error cuadrático medio de

$$\mathbb{E}[(X_2 - \wp(X_2|\mathbf{W}))^2] = \frac{\sigma^2}{1-\phi^2} - \mathbf{a}' \begin{pmatrix} \frac{\phi\sigma^2}{1-\phi^2} \\ \frac{\phi\sigma^2}{1-\phi^2} \end{pmatrix} = \frac{\sigma^2}{1+\phi^2}$$

∇

2.8.1.1. Propiedades del operador de predicion $\wp(\cdot|\mathbf{W})$

Para $\mathbf{W} = (W_T, \ldots, W_1)'$ y Y con segundo momento finito ya se vió cómo calcular el mejor predictor lineal $\wp(Y|\mathbf{W})$ en términos de $1, W_T, \ldots, W_1$ mediante $\wp(Y|\mathbf{W}) = \mu_y + \mathbf{a}'(\mathbf{W} - \mu_W)$. La función $\wp(\cdot|\mathbf{W})$ que transforma Y en $\wp(Y|\mathbf{W})$ se conoce como *operador de predicción*. Los operadores de predicción tienen muchas propiedades útiles que se pueden utilizar para simplificar el cálculo de los mejores predictores lineales. A continuación se enlistan algunas.

Proposición 2.8.2. *(Propiedades del operador de predicción $\wp(\cdot|\mathbf{W})$)*

Supóngase que $\mathbb{E}(U^2), \mathbb{E}(V^2) < \infty$, $\Xi = Cov(\mathbf{W}, \mathbf{W})$ y $\beta, \alpha_1, \ldots, \alpha_T$ constantes. Entonces,

1. $\wp(U|\mathbf{W}) = \mathbb{E}(U) + \mathbf{a}'(\mathbf{W} - \mathbb{E}(\mathbf{W}))$, donde $\Xi\mathbf{a} = Cov(U, \mathbf{W})$
2. $\mathbb{E}[(U - \wp(U|\mathbf{W}))\mathbf{W}]$ y $\mathbb{E}[U - \wp(U|\mathbf{W})] = 0$
3. $\mathbb{E}[(U - \wp(U|\mathbf{W}))^2] = Var(U) - \mathbf{a}'Cov(U, \mathbf{W})$
4. $\wp(\alpha_1 U + \alpha_2 V + \beta|\mathbf{W}) = \alpha_1\wp(U|\mathbf{W}) + \alpha_2\wp(V|\mathbf{W}) + \beta$
5. $\wp(\sum_{i=1}^{T} \alpha_i W_i + \beta|\mathbf{W}) = \sum_{i=1}^{T} \alpha_i W_i + \beta$
6. Si $Cov(U, \mathbf{W}) = 0$, entonces $\wp(U|\mathbf{W}) = \mathbb{E}(U)$
7. Si $\mathbf{V}$ es un vector aleatorio tal que las componentes de $\mathbb{E}(\mathbf{V}\mathbf{V}')$ son finitas, entonces $\wp(U|\mathbf{W}) = \wp(\wp(U|\mathbf{W}, \mathbf{V})|\mathbf{W})$

Ejemplo 2.8.3. *(Predicción a un paso de la serie $AR(p)$)*

Supóngase que $\{X_t\}_{t\in\mathbb{Z}}$ es una serie de tiempo estacionaria que satisface las ecuaciones

$$X_t = \phi_1 X_{t-1} + \ldots + \phi_p X_{t-p} + \epsilon_t, \quad t \in \mathbb{Z}$$

donde $\epsilon_t \sim WN(0, \sigma^2)$ y ϵ_t no está correlacionado con X_s para todo $s < t$. Entonces, si $T > p$ se puede aplicar el operador de predicción P_T en ambos lados de la ecuación anterior y usando las propiedades 4., 5. y 6. se tiene que

$$P_T(X_{T+1}) = \phi_1 X_T + \ldots + \phi_p X_{T+1-p}$$

∇

Ejemplo 2.8.4. *(Una serie $AR(1)$ con media diferente de 0)*

Se dice que una serie $\{Y_t\}$ es un proceso $AR(1)$ con media μ si $\{Y_t - \mu\}$ es un proceso $AR(1)$ con media 0. Si $Y_t = X_t + \mu$, entonces Y_t satisface la ecuación

$$Y_t - \mu = \phi(Y_{t-1} - \mu) + \epsilon_t$$

Si $P_T Y_{T+h}$ es el mejor predictor lineal de Y_{t+h} en términos de $\{1, Y_T, \ldots, Y_1\}$, entonces aplicando P_T en ambos lados de la ecuación anterior con $t = T+1, T+2, \ldots$ se obtiene

$$P_T Y_{T+h} - \mu = \phi(P_T Y_{T+h-1} - \mu), \quad h \in \mathbb{N}^+$$

Nótese que $P_T Y_T = Y_T$, entonces se puede resolver estas ecuaciones recursivamente para $P_T Y_{T+h}$, $h \in \mathbb{N}^+$ para obtener

$$P_T Y_{T+h} = \mu + \phi^h (Y_T - \mu)$$

El correspondiente error cuadrático medio es

$$\mathbb{E}[(Y_{T+h} - P_T Y_{T+h})^2] = \gamma(0)[1 - \mathbf{a}_T' \rho_T(h)]$$

donde $\gamma(0) = \frac{\sigma^2}{1-\phi^2}$ y $\rho(h) = \phi^h, h \geq 0$. Entonces, sustituyendo $a_T = (\phi^h, 0, \ldots, 0)'$ se tiene que

$$\mathbb{E}[(Y_{T+h} - P_T Y_{T+h})^2] = \sigma^2 \frac{1 - \phi^{2h}}{1 - \phi}$$

∇

Observación 2.8.2.

En general, si $\{Y_t\}$ es una serie de tiempo estacionaria con media μ y si $\{X_t\}$ es una serie definida mediante $X_t = Y_t - \mu$ (donde $\{X_t\}$ tiene media cero), entonces como el conjunto de todas las combinaciones lineales de $\{1, Y_T, \ldots, Y_1\}$ es el mismo que el conjunto de todas las combinaciones lineales de $\{1, X_T, \ldots, X_1\}$ el predictor lineal de cualquier variable aleatoria W en términos de $1, Y_T, \ldots, Y_1$ es el mismo que el predictor lineal en términos de $1, X_T, \ldots, X_1$. Denótese este predictor como P_T y aplicando P_T a la ecuación $Y_{T+h} = X_{T+h} + \mu$ se obtiene

$$P_T Y_{T+h} = \mu + P_T X_{T+h}$$

Por tanto, el mejor predictor lineal de Y_{T+h} se puede determinar si se encuentra el mejor predictor lineal de X_{T+h} y sumándole μ. Nótese que como $P_T X_{T+h} = \mu + \sum_{i=1}^{T} a_i (X_{T+1-i} - \mu)$ y $\mathbb{E}(X_t) = 0$, $P_T X_{T+h}$ es el mismo que el mejor predictor lineal de X_{T+h} sólo en términos de $X_T, \ldots, X_1$. ∇

2.8.1.2. El algoritmo de Durbin-Levinson

Por la observación anterior, se puede restringir la atención a series de tiempo estacionarias con media cero, haciendo lo correspondiente si se desea hacer predicciones de series de tiempo estacionarias con media distinta de cero. Si $\{X_t\}$ es una serie de tiempo con media cero y función de autocovarianza $\gamma(\cdot)$, entonces la expresión $\wp(Y|\mathbf{W}) = \mu_Y + \mathbf{a}'(\mathbf{W} - \mu_W)$ resuelve completamente el problema de determinar el mejor predictor lineal de $P_T X_{T+h}$ de X_T en términos de $\{X_T, \ldots, X_1\}$. Sin embargo, con esta metodología se tiene que resolver un sistema de T ecuaciones lineales y, como siempre, si T es grande hay un problema

numérico importante. En lo casos en los cuales el proceso se define mediante un sistema de ecuaciones lineales se puede aprovechar la linealidad de P_T. Para procesos estacionarios más generales resulta de utilidad si el predictor a un paso $P_T X_{T+1}$ basada en las T observaciones anteriores se puede usar para simplificar el cálculo $P_{T+1}X_{T+2}$ (el predictor a un paso basado en las $T+1$ observaciones previas). Se dice que son recursivos los algoritmos que utilizan esta idea. Dos ejemplos importantes de este tipo de algoritmos recursivos son el algoritmo de Durbin-Levinson y el algoritmo de innovación.

Se sabe de $\wp(Y|\mathbf{W}) = \mu_Y + \mathbf{a}'(\mathbf{W} - \mu_W)$ y de $\mathbf{\Gamma a} = \boldsymbol{\gamma}$ que si la matriz $\mathbf{\Gamma}_T$ es no-singular, entonces

$$P_T X_{T+1} = \boldsymbol{\phi}' \mathbf{X}_T = \phi_{T1} X_T + \ldots + \phi_{TT} X_1$$

donde $\boldsymbol{\phi}_T = \mathbf{\Gamma}_T^{-1} \boldsymbol{\gamma}_T$, $\boldsymbol{\gamma}_T = (\gamma(1), \ldots, \gamma(T))'$ y el correspondiente error cuadrático medio es

$$v_T := \mathbb{E}[(X_{T+1} - P_T X_{T+1})^2] = \gamma(0) - \boldsymbol{\phi}' \boldsymbol{\gamma}_T$$

Una condición útil para la no singularidad de todas las matrices de autocovarianza $\mathbf{\Gamma}_1, \mathbf{\Gamma}_2, \ldots$ es que $\gamma(0) > 0$ y $\lim_{h\to\infty} \gamma(h) = 0$.

Proposición 2.8.3. *(Algoritmo de Durbin-Levinson)*

Los coeficientes $\phi_{T,1}$ se pueden obtener recursivamente mediante

$$\phi_{TT} = \left[\gamma(n) - \sum_{j=1}^{T-1} \phi_{T-1,j} \gamma(T-j) \right] v_{T-1}^{-1} \tag{2.43}$$

$$\begin{pmatrix} \phi_{T,1} \\ \vdots \\ \phi_{T,T-1} \end{pmatrix} = \begin{pmatrix} \phi_{T-1,1} \\ \vdots \\ \phi_{T-1,T-1} \end{pmatrix} - \phi_{T,T} \begin{pmatrix} \phi_{T-1,T-1} \\ \vdots \\ \phi_{T-1,1} \end{pmatrix} \tag{2.44}$$

y $v_T = v_{T-1}[1 - \phi_{TT}^2]$ donde $\phi_{11} = \gamma(1)/\gamma(0)$ y $v_0 = \gamma(0)$.

Demostración: (por inducción sobre T)

La definición ϕ_{11} garantiza que la ecuación

$$R_T \boldsymbol{\phi}_T = \boldsymbol{\rho}_T$$

(donde $\rho_T = (\rho(1), \ldots, \rho(T))'$) se satisfaga para $T = 1$. Ahora se probará que ϕ_T definida recursivamente mediante (2.43) y (2.44) satisfacen el sistema de ecuaciones anterior.

Supóngase que esto se cumple para $T = k$. Particionándose R_{k+1} y definiendo

$$\rho_k^{(r)} := (\rho(k), \rho(k-1), \ldots, \rho(1))'$$

y

$$\phi_k^{(r)} := (\phi_{kk}, \phi_{k,k-1}, \ldots, \phi_{k1})'$$

se tiene que las recursiones implican

$$\begin{aligned} R_{k+1}\phi_{k+1} &= \begin{pmatrix} R_k & \rho_k^{(r)} \\ \rho_k^{(r)'} & 1 \end{pmatrix} \begin{pmatrix} \phi_k - \phi_{k+1,k+1}\phi_k^{(r)} \\ \phi_{k+1,k+1} \end{pmatrix} \\ &= \begin{pmatrix} \rho_k - \phi_{k+1,k+1}\rho_k^{(r)} + \phi_{k+1,k+1}\rho_k^{(r)} \\ \rho_k^{(r)'}\phi_k - \phi_{k+1,k+1}\rho_k^{(r)}\phi_k^{(r)} + \phi_{k+1,k+1} \end{pmatrix} = \rho_{k+1} \end{aligned}$$

Ahora se probará que los errores cuadráticos medios $v_T := \mathbb{E}[(X_{T+1} - \boldsymbol{\phi}_T'\mathbf{X}_T)^2]$ satisfacen que $v_0 = \gamma(0)$ y $v_T = v_{T-1}(1-\phi_{TT}^2)$. Como $P_0X_1 := \mathbb{E}(X_1) = 0$, entonces $v_T = \gamma(0)$. Como $\boldsymbol{\phi}_T'\mathbf{X}_T$ es el mejor predictor lineal de X_{T+1}, se puede escribir

$$v_T = \gamma(0) - \phi_T'\gamma_T = \gamma(0) - \phi_{T-1}'\gamma_{T-1} + \gamma_{TT}\phi_{T-1}^{(r)'}\gamma_{T-1} - \phi_{TT}\gamma(T)$$

Entonces,

$$v_T = v_{T-1} + \phi_{TT}(\phi_{T-1}^{(r)'}\gamma_{T-1} - \gamma(T))$$

y por el cálculo de ϕ_{TT} se tiene que

$$v_T = v_{T-1} + \phi_{TT}^2(\gamma(0) - \phi_{T-1}'\gamma_{T-1}) = v_{T-1}(1-\phi_{TT}^2)$$

□

2.8.1.3. El algoritmo de innovación

El algoritmo recursivo que se estudiará ahora se puede aplicar a todas las series con segundo momento finito, aún si éstas no son estacionarias; y en algunos casos especiales se puede simplificar considerablemente.

Supóngase que $\{X_t\}$ es una serie de tiempo con media cero tal que $\mathbb{E}(|X_t|) < \infty$ y

$$\mathbb{E}(X_iX_j) = \kappa(i,j)$$

Y defínase

$$\hat{X}_T := \begin{cases} 0 & \text{si } T = 1 \\ P_{T-1}X_T & \text{si } T = 2, 3, \ldots \end{cases}$$

y

$$v_T := \mathbb{E}[(X_{T+1} - P_T X_{T+1})^2]$$

Ahora, se definen las innovaciones ó errores de predicción a un paso

$$U_T = X_T - \hat{X}_T$$

En términos vectoriales $\mathbf{U}_T = (U_1, \ldots, U_T)'$ y $\mathbf{X}_T = (X_1, \ldots, X_T)'$, la última ecuación se convierte en

$$\mathbf{U} = A_T \mathbf{X}_T$$

donde A_T tiene la forma

$$A_T = \begin{pmatrix} 1 & 0 & 0 & \cdots & 0 \\ a_{1,1} & 1 & 0 & \cdots & 0 \\ a_{2,2} & a_{2,1} & 1 & \cdots & 0 \\ \vdots & \vdots & \vdots & \ddots & 0 \\ a_{T-1,T-1} & a_{T-1,T-2} & a_{T-1,T-3} & \cdots & 1 \end{pmatrix} \tag{2.45}$$

A_T es no-singular con inversa, C_T, de la forma

$$C_T = \begin{pmatrix} 1 & 0 & 0 & \cdots & 0 \\ \theta_{1,1} & 1 & 0 & \cdots & 0 \\ \theta_{2,2} & \theta_{2,1} & 1 & \cdots & 0 \\ \vdots & \vdots & \vdots & \ddots & 0 \\ \theta_{T-1,T-1} & \theta_{T-1,T-2} & \theta_{T-1,T-3} & \cdots & 1 \end{pmatrix} \tag{2.46}$$

Por tanto, el vector de predictores a un paso, $\hat{\mathbf{X}}_T := (X_1, P_1 X_2, \ldots, P_{T-1} X_T)'$ se puede expresar como

$$\hat{\mathbf{X}}_T = \mathbf{X}_T - \mathbf{U}_T = C_T \mathbf{U}_T - \mathbf{U}_T = \boldsymbol{\Theta}_T (\mathbf{X}_T - \hat{\mathbf{X}}_T)$$

donde

$$\boldsymbol{\Theta}_T = \begin{pmatrix} 0 & 0 & 0 & \cdots & 0 \\ \theta_{1,1} & 0 & 0 & \cdots & 0 \\ \theta_{2,2} & \theta_{2,1} & 0 & \cdots & 0 \\ \vdots & \vdots & \vdots & \ddots & 0 \\ \theta_{T-1,T-1} & \theta_{T-1,T-2} & \theta_{T-1,T-3} & \cdots & 0 \end{pmatrix} \tag{2.47}$$

y $\mathbf{X}_T$ satisface $\mathbf{X}_T = C_T(\mathbf{X}_T - \hat{\mathbf{X}}_T)$. La ecuación $\hat{\mathbf{X}}_T = \boldsymbol{\Theta}_T(\mathbf{X}_T - \hat{\mathbf{X}}_T)$ se puede reescribir como

$$\hat{X}_{T+1} := \begin{cases} 0 & \text{si } T = 0 \\ \sum_{j=1}^{T} \theta_{Tj}(X_{T+1-j} - \hat{X}_{T+1-j}) & \text{si } T = 1, 2, 3, \ldots \end{cases}$$

por tanto, los predictores a un paso $\hat{X}_1, \hat{X}_2, \ldots$ se pueden calcular recursivamente una vez que se tienen los coeficientes θ_{ij}. El siguiente algorimo genera estos coeficientes y los errores cuadráticos medios $v_j = \mathbb{E}[(X_j - \hat{X}_j)^2]$ a partir de las covarianzas $\kappa(i, j)$.

Proposición 2.8.4. *(Algoritmo de innovación)*

Los coeficientes $\theta_{T1}, \ldots, \theta_{TT}$ se pueden obtener recursivamente mediante las ecuaciones $v_0 := \kappa(1, 1,)$,

$$\theta_{T,T-k} = v_k^{-1}[\kappa(T+1, k+1) \sum_{j=0}^{k-1} \theta_{k,k-j}\theta_{T,T-j}v_j], \quad 0 \leq k < T,$$

y

$$v_T = \kappa(T+1, k+1) - \sum_{j=0}^{T-1} \theta_{T,T-j}^2 v_j$$

A partir de v_0 se obtienen sucesivamente $\theta_{11}, v_1; \theta_{22}, \theta_{21}, v_2; \theta_{33}, \theta_{32}, \theta_{31}, v_3; \ldots$

Observación 2.8.3.

Mientras con el algoritmo de Durbin-Levinson se obtienen los coeficientes de $X_1, \ldots, X_T$ en la representación $\hat{X}_{T+1} = \sum_{j=1}^{T} \phi_{Tj} X_{T+1-j}$, con el algoritmo de innovación se obtienen los coeficientes de $(X_T - \hat{X}_T), \ldots, (X_1 - \hat{X}_1)$ en la expansión $\hat{X}_{T+1} = \sum_{j=1}^{T} \theta_{Tj}(X_{T+1-j} - \hat{X}_{T+1-j})$. La última expansión tiene ciertas ventajas que se obtienen del hecho de que las innovaciones no están correlacionadas. Una consecuencia directa de este algoritmo es que si $\theta_{T0} := 1$, entonces

$$X_{T+1} = X_{T+1} - \hat{X}_{T+1} + \hat{X}_{T+1} = \sum_{j=0}^{T} \theta_{Tj}(X_{T+1-j} - \hat{X}_{T+1-j}), \quad T = 0, 1, 2, \ldots$$

∇

Ejemplo 2.8.5. *(Predicción recursiva de un $MA(1)$)*

Si $\{X_t\}$ es la serie de tiempo definida por

$$X_t = \epsilon_t + \theta\epsilon_{t-1}, \quad \{\epsilon_t\} \sim WN(0, \sigma^2)$$

entonces $\kappa(i,j) = 0$ para $|i-j| > 1$, $\kappa(i,i) = \sigma^2(1+\theta^2)$ y $\kappa(i,i+1) = \theta\sigma^2$. Aplicando el algoritmo de innovación se obtiene

$$\theta_{Tj} = 0, \quad 2 \leq j \leq T$$

$$\theta_{T1} = v_{T-1}^{-1}\theta\sigma^2,$$

$$v_0 = (1+\theta^2)\sigma^2$$

y

$$v_T = [1+\theta^2 - v_{T-1}^{-1}\theta^2\sigma^2]\sigma^2$$

∇

Cálculo recursivo para predictores de h-pasos

Para la predicción de h-pasos se utiliza el resultado

$$P_T(X_{T+k} - P_{T+k-1}X_{T+k}) = 0, \quad k \in \mathbb{N}^+$$

Entonces,

$$\mathbb{E}[(X_{T+k} - P_{T+k-1}X_{T+k} - 0)X_{T+j-1}] = 0, \quad j = 1, \ldots, T$$

Por tanto,

$$\begin{aligned} P_T X_{T+h} &= P_T P_{T+h-1} X_{T+h} = P_T \hat{X}_{T+h} \\ &= P_T \left(\sum_{j=1}^{T+h-1} \theta_{T+h-1,j}(X_{T+h-j} - \hat{X}_{T+h-j}) \right) \end{aligned}$$

de nuevo usando el hecho de que $P_T(X_{T+k} - P_{T+k-1}X_{T+k}) = 0, \quad k \in \mathbb{N}^+$ y la linealidad de P_T se tiene que

$$P_T X_{T+h} = \sum_{j=1}^{T+h-1} \theta_{T+h-1,j}(X_{T+h-j} - \hat{X}_{T+h-j}) \tag{2.48}$$

donde los coeficientes θ_{Tj} se obtienen mediante el algoritmo de innovación. El error cuadrático medio se puede expresar como

$$\begin{aligned} \mathbb{E}[(X_{T+h} - P_T X_{T+h})^2] &= \mathbb{E}(X_{T+h}^2) - \mathbb{E}^2(P_T X_{T+h}) \\ &= \kappa(T+h, T+h) - \sum_{j=h}^{T+h-1} \theta_{T+h-1,j}^2 v_{T+h-j-1} \end{aligned}$$

2.8.1.4. Predicción de procesos estacionarios en términos de una infinidad de valores pasados

Cuando se tiene disponibles $X_m, \ldots, X_0, X_1, \ldots, X_T$ $(m < 0)$, es útil evaluar el mejor predictor lineal de X_{T+h} en términos de $1, X_m, \ldots, X_0, \ldots, X_m$. Este predictor se denotará por $P_{m,T}X_{T+h}$ se puede evaluar con los métodos que se estudiaron en las secciones anteriores. Si $|m|$ es grande, este predictor se puede aproximar por su límite en media cuadrática

$$\tilde{P}_T X_{T+h} = \lim_{m \to -\infty} P_{m,T} X_{T+h}$$

Se conocerá a $\tilde{P}_T$ como el operador de predicción basado en el pasado infinito, $\{X_t : -\infty < t \leq T\}$. Análogamente, se conocerá a P_T como el operador de predicción basado en el pasado finito, $\{X_1, \ldots, X_T\}$.

Para calcular $\tilde{P}_T X_{T+h}$ se debe considerar que como $P_T X_{T+h}$, en el mejor predictor lineal $\tilde{P}_T X_{T+h}$ cuando $\{X_t\}$ es un proceso estacionario con media cero y función de autocovarianza $\gamma(\cdot)$ está caracterizado por las ecuaciones

$$\mathbb{E}[(X_{T+h} - \tilde{P}_T X_{T+h}) X_{T+1-i}] = 0, \quad i = 1, 2, \ldots$$

Si se puede encontrar solución a este sistema de ecuaciones, éstas determinarán de manera única al predictor $\tilde{P}_T X_{T+h}$. Una forma (generalmente efectiva) de resolver este problema es supone que $\tilde{P}_T X_{T+h}$ se puede expresar de la forma

$$\tilde{P}_T X_{T+h} = \sum_{j=1}^{\infty} \alpha_j X_{T+1-j}$$

y por tanto el sistema de ecuaciones anterior se convierte en

$$\mathbb{E}\left[\left(X_{T+h} - \sum_{j=1}^{\infty} \alpha_j X_{T+1-j}\right) X_{T+1-i}\right] = 0, \quad i = 1, 2, \ldots$$

ó equivalentemente,

$$\sum_{j=1}^{\infty} \gamma(i-j)\alpha_j = \gamma(h+j-1), \quad i = 1, 2, \ldots$$

Este es un conjunto infinito de ecuaciones lineales con coeficientes desconocidos α_i que determina $\tilde{P}_T X_{T+h}$ siempre y cuando la serie converja.

Proposición 2.8.5. *(Propiedades de $\tilde{P}_T$)*

Supóngase que $\mathbb{E}(U), \mathbb{E}(V) < \infty$, a, b, c constantes y $\Gamma = Cov(\mathbf{W}, \mathbf{W})$. Entonces,

1. $\mathbb{E}[(U - \tilde{P}_T U)X_j] = 0,\ j \leq n.$
2. $\tilde{P}_T(aU + bV + c) = a\tilde{P}_T U + b\tilde{P}_T V + c.$
3. Si U es combinación lineal de X_j, $j \leq T$, entonces $\tilde{P}_T U = U$.
4. Si $Cov(U, X_j) = 0$ para todo $j \leq T$, entonces $\tilde{P}_T U = \mathbb{E}(U)$.

Estas propiedades se pueden utilizar para simplificar la obtención de $\tilde{P}_T X_{T+h}$ en el caso de que $\{X_t\}$ sea un proceso $ARMA$.

Ejemplo 2.8.6.

Considérese el proceso $\{X_t\}$ causal e invertible, $ARMA(1,1)$ definido mediante la expresión

$$X_t - \phi X_{t-1} = \epsilon_t + \theta\epsilon_{t-1}, \quad \{\epsilon_t\} \sim WN(0, \sigma^2)$$

Ya se estudió (dada su naturaleza causal e invertible) la existencia de las representaciones asociadas

$$X_{T+1} = \epsilon_{T+1} + (\theta + \phi)\sum_{j=1}^{\infty} \phi^{j-1}\epsilon_{T+1-j}$$

y

$$\epsilon_{T+1} = X_{T+1} - (\theta + \phi)\sum_{j=1}^{\infty}(-\theta)^{j-1}X_{T+1-j}$$

Aplicando el operador $\tilde{P}_T$ en la segunda ecuación (y utilizando las propiedades de $\tilde{P}_T$) se tiene que

$$\tilde{P}_T X_{T+1} = (\phi + \theta)\sum_{j=1}^{\infty}(-\theta)^{j-1}X_{T+1-j}$$

Ahora, aplicando el operador $\tilde{P}_T$ en la primera ecuación (y utilizando las propiedades de $\tilde{P}_T$) se tiene que

$$\tilde{P}_T X_{T+1} = (\theta + \phi)\sum_{j=1}^{\infty} \phi^{j-1}\epsilon_{T+1-j}$$

Por tanto,

$$X_{T+1} - \tilde{P}_T X_{T+1} = \epsilon_{T+1}$$

y entonces, el error medio cuadrático del predictor $\tilde{P}_T X_{T+1}$ está dado por $\mathbb{E}(Z_{T+1}^2) = \sigma^2$
▽

2.8.1.5. La descomposición de Wold

Considérese el proceso estacionario

$$X_t = Acos(ct) + Bsen(ct)$$

donde $c \in (0, \pi)$ es constante y A y B son variables aleatorias no correlacionadas con media cero y varianzas σ^2. Se puede probar (mediante el cálculo directo que)

$$X_t = (2cos(c))X_{t-1} - X_{t-2} = \tilde{P}_{t-1} X_t, \quad t \in \mathbb{Z}$$

Así que $X_t - \tilde{P}_{t-1} X_t = 0$ para todo $t \in \mathbb{Z}$. Los procesos con esta última propiedad se conocen como *deterministas*.

Proposición 2.8.6. *(Descomposición de Wold)*

Si $\{X_t\}$ es una serie de tiempo estacionaria no determinista, entonces

$$X_t = \sum_{j=0}^{\infty} \psi_j \epsilon_{t-j} + V_t$$

donde

1. $\psi_0 = 1$ y $\sum_{j=0}^{\infty} \psi_j^2 < \infty$
2. $\{\epsilon_t\} \sim WN(0, \sigma^2)$
3. Para cualesquiera $s, t \in \mathbb{Z}$, $Cov(\epsilon_t, V_t) = 0$
4. Para todo $t \in \mathbb{Z}$, $\epsilon_t = \tilde{P}_t \epsilon_t$
5. Para cualesquiera $s, t \in \mathbb{Z}$, $V_t = \tilde{P}_s \epsilon_t$
6. $\{V_t\}$ es determinista.

Observación 2.8.4.

En esta proposición $\tilde{P}_t Y$ es el mejor estimador lineal de Y en términos de combinaciones lineales ó límites de combinaciones lineales de $1, X_s$, $-\infty < s \leq t$. Las sucesiones $\{\epsilon_t\}, \{\psi_t\}$ y $\{V_t\}$ son únicas y se pueden escribir explícitamente como $\epsilon_t = X_t - \tilde{P}_{t-1} X_t$, $\psi_j = \frac{\mathbb{E}[X_t \epsilon_{t-j}]}{\mathbb{E}[\epsilon_t^2]}$ y $V_t = X_t - \sum_{j=0}^{\infty} \psi_j \epsilon_{t-j}$.

Ejemplo 2.8.7.

Si $X_t = U_t + Y$, donde $\{U_t\} \sim WN(0, \sigma^2)$, $\mathbb{E}(U_t Y) = 0$ para cualquier $t \in \mathbb{Z}$ y Y tiene media cero y esperanza τ^2, entonces $\tilde{P}_{t-1} X_t = Y$ ya que Y es el límite en media cuadrática cuando $s \to \infty$ de $\frac{1}{s}(X_{t-1} + \ldots + X_{t-s})$ y para cualesquiera $s \leq t-1$ se tiene que $\mathbb{E}[(X_t - Y) X_s] = 0$. Por tanto, las sucesiciones en la descomposición de Wold de $\{X_t\}$ están dadas por $\epsilon_t = U_t$, $\psi_0 = 1$, $V_t = Y$ y para $j \in \mathbb{N}^+$ $\psi_j = 0$. $\triangledown$

2.8.2. Predicción en procesos ARMA

El algortimo de innovación que se describió en las secciones anteriores proporciona un método recursivo para hacer pronóstico en procesos con media cero que no necesariamente sean estacionarios. Para el proceso ARMA causal

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

se puede simplificar la aplicación de dicho algoritmo considerablemente. La idea consiste en aplicar el algoritmo no al proceso (original ó de interés) sino al proceso transformado

$$W_t = \begin{cases} \sigma^{-1} X_t & t = 1, ..., m \\ \sigma^{-1} \phi(B)(X_t) & t > m \end{cases}$$

donde $m = \text{máx}\{p, q\}$.

Por convención se define $\theta_0 := 1$ y para $j > q$, $\theta_j := 0$. Sin pérdida de generalidad se puede suponer que $p, q \in \mathbb{N}^+$ escribiendo los correspondientes θ_i, ϕ_j igual a cero.

Las autocovarianzas $\kappa(i, j) = \mathbb{E}(W_i W_j)$ $(i, j \in \mathbb{N}^+)$ se obtienen de la forma

$$\kappa(i,j) = \begin{cases} \sigma^{-2} \gamma_X(i-j) & 1 \leq i, j \leq m \\ \sigma^{-2} \left[\gamma_X(i-j) - \sum_{r=1}^{p} \phi_r \gamma_X(r - |i-j|)\right] & \text{mín}\{i,j\} \leq m < \text{máx}\{i,j\} \leq 2m \\ \sum_{r=0}^{q} \theta_r \theta_{r+|i-j|} & \text{mín}\{i,j\} > m \\ 0 & \text{en otro caso} \end{cases}$$

Aplicando el algoritmo de innovación a $\{W_t\}$ se tiene que

$$\hat{W}_{T+1} = \begin{cases} \sum_{j=1}^{T} \theta_{Tj}(W_{T+1-j} - \hat{W}_{T+1-j}) & 1 \leq T < m \\ \sum_{j=1}^{q} \theta_{Tj}(W_{T+1-j} - \hat{W}_{T+1-j} & T > m \end{cases}$$

donde los coeficientes θ_{Tj} y los errores cuadráticos medios $\mathbb{E}[(W_{T+1} - \hat{W}_T + 1)^2]$ se encuentran recursivamente a partir del algoritmo de innovación y las autocovarianzas $\kappa(i,j)$. Una característica importante de los predictores $\hat{W}_{T+1}$ es la dismunución de θ_{Tj} cuando $T \geq m$ y $j > q$. Esta es una consecuencia de la construcción con el algoritmo de innovación y el hecho de que si $r > m$ y $|r - s| > q$, entonces $\kappa(r,s) = 0$.

Se puede ver a partir de la definición de W_t,

$$W_t = \begin{cases} \sigma^{-1} X_t & t = 1, ..., m \\ \sigma^{-1}\phi(B)(X_t) & t > m \end{cases}$$

que se puede escribir a X_T $(T \geq 1)$ como una combinación lineal de $W_j, j \in \{1, \ldots, T\}$ pero también a W_T $(T \geq 1)$ como una combinación lineal de $X_j, j \in \{1, \ldots, T\}$. Esto significa que el mejor predictor lineal de cualquier variable aleatoria Y en términos de $\{1, X_1, \ldots, X_T\}$ es el mismo que el mejor predictor lineal de Y en términos de $\{1, W_1, \ldots, W_T\}$. Se denotará a este predictor como $P_T Y$. En particular los predictores a un paso de W_{T+1} y X_{T+1} están dados por

$$\hat{W}_{T+1} = P_T W_{T+1}$$

y

$$\hat{X}_{T+1} = P_T X_{T+1}$$

Utilizando la linealidad de P_T y la definición de W_t se tiene que

$$\hat{W}_t = \begin{cases} \sigma^{-1}\hat{X}_t & t = 1, ..., m \\ \sigma^{-1}[\hat{X}_t - \phi_1 X_{t-1} - \ldots - \phi_p X_{t-p}] & t > m \end{cases}$$

y por tanto

$$X_t - \hat{X}_t = \sigma[W_t - \hat{W}_t], \quad t \in \mathbb{N}^+$$

Reemplazando $W_t - \hat{W}_t$ por $\sigma^{-1}(X_t - \hat{X}_t)$ en el cálculo de las autocovarianzas $\kappa(i,j)$ se obtiene que

$$\hat{X}_{T+1} = \begin{cases} \sum_{j=1}^{T} \theta_{Tj}(X_{T+1-j} - \hat{X}_{T+1-j}) & 1 \leq T < m \\ \phi_1 X_T + \ldots + \phi_p X_{T+1-p} + \sum_{j=1}^{q} \theta_{Tj}(X_{T+1-j} - \hat{X}_{T+1-j}) & T \geq m \end{cases}$$

y

$$\mathbb{E}[(X_{T+1} - \hat{X}_{T+1})^2] = \sigma^2 \mathbb{E}[(W_{T+1} - \hat{W}_{T+1})^2] = \sigma^2 r_T$$

donde θ_{Tj} y r_T se encuentran a partir del algoritmo de innovación la función de autocovarianzas $\kappa(i,j)$. La expresión anterior para $\hat{X}_{T+1}$ determina los predictores a un paso $\hat{X}_2, \hat{X}_3, \ldots$, recursivamente.

Observación 2.8.5.

Si $\{X_t\}$ es invertible, entonces

$$\lim_{T\to\infty} \mathbb{E}[(X_T - \hat{X}_T - \epsilon_T)^2] = 0,$$

para $j = 1, \ldots, q$

$$\lim_{T\to\infty} \theta_{Tj} = \theta_j$$

y

$$\lim_{T\to\infty} r_T = 1$$

∇

El cálculo algebraico de los coeficientes θ_{Tj} y r_T es muy complicado excepto en algunos modelos muy simples (tales como los que se analizarán a continuación). La mayoría de las recursiones se puede implementar numéricamente de forma muy directa.

Ejemplo 2.8.8. *(Predicción del proceso $AR(p)$)*

Aplicando la expresión que se obtuvo para $\hat{X}_{T+1}$ al proceso $ARMA(p,1)$ con $\theta_1 = 0$ se tiene que

$$\hat{X}_{T+1} = \phi_1 X_T + \ldots + \phi_T X_{T+1-p}, \quad T \geq p$$

∇

Ejemplo 2.8.9. *(Predicción del proceso $MA(q)$)*

Aplicando la expresión que se obtuvo para $\hat{X}_{T+1}$ al proceso $ARMA(1,q)$ con $\phi_1 = 0$ se tiene que

$$X_{T+1} = \sum_{j=1}^{\text{mín}\{T,q\}} \theta_{Tj}(X_{T+1-j} - \hat{X}_{T+1-j}), \quad T \in \mathbb{N}^+$$

donde los coeficientes θ_{Tj} se encuentran mediante el algoritmo de innovación y las autocovarianzas $\kappa(i,j)$. Como en este caso los procesos $\{X_t\}$ y $\{\sigma^{-1}W_t\}$ son idénticos, entonces la expresión de las covarianzas se reduce a

$$\kappa(i,j) = \sigma^{-2}\gamma_X(i-j) = \sum_{r=0}^{q-|i-j|} \theta_r \theta_{r+|i-j|}$$

∇

Ejemplo 2.8.10. *(Predicción del proceso $ARMA(1,1)$)*

Si

$$X_t - \phi X_{t-1} = \epsilon_t + \theta\epsilon_{t-1}, \quad \{\epsilon_t\} \sim WN(0, \sigma^2)$$

con $|\phi| < 1$, aplicando la expresión que se obtuvo para $\hat{X}_{T+1}$ a dicho proceso se obtiene

$$\hat{X}_{T+1} = \phi X_T + \theta_{T1}(X_T - \hat{X}_T)$$

Para calcular θ_{T1} se usará el hecho de que $\gamma_X(0) = \frac{1+2\theta\phi+\theta^2}{1-\phi^2}$ y por tanto para $i, j \in \mathbb{N}^+$ se obtienen las autocovarianzas

$$\kappa(i,j) = \begin{cases} (1+2\theta\phi+\theta^2)/(1-\phi^2) & i = j = 1 \\ 1+\theta^2 & i = j \geq 2 \\ \theta & |i-j| = 1, i \geq 1 \\ 0 & \text{en otro caso} \end{cases}$$

con estos valores de $\kappa(i,j)$, las recursiones del algoritmo de innovación se reducen a

$$r_0 = (1+2\theta\phi+\theta^2)/(1-\phi^2)$$

$$\theta_{T1} = \theta/r_{T-1}$$

$$r_T = 1 + \theta^2 - \frac{\theta^2}{r_{T-1}}$$

∇

2.8.2.1. Predicción h pasos hacia adelante de un proceso $ARMA(p,q)$

Como antes, $P_T Y$ denotará al mejor predictor lineal de Y en términos de $X_1, \ldots, X_T$ (que es el mismo que el mejor predictor lineal en términos de $W_1, \ldots, W_T$. Entonces, de (2.48) se tiene que

$$\begin{aligned} P_T W_{T+h} &= \sum_{j=h}^{T+h-1} \theta_{T+h-i,j}(W_{T+h-j} - \hat{W}_{T+h-j}) \\ &= \sigma^2 \sum_{j=h}^{T+h-1} \theta_{T+h-i,j}(X_{T+h-j} - \hat{X}_{T+h-j}) \end{aligned}$$

Utilizando esta deducción y aplicando el operador P_T en ambos lados de

$$W_t = \begin{cases} \sigma^{-1} X_t & t = 1, ..., m \\ \sigma^{-1}\phi(B)(X_t) & t > m \end{cases}$$

se concluye que el predictor h pasos hacia adelante $P_T X_{T+h}$ satisface

$$P_T X_{T+h} = \begin{cases} \sum_{j=h}^{T+h-1} \theta_{T+h-i,j}(X_{T+h-j} - \hat{X}_{T+h-j}) & 1 \leq h \leq m-T \\ \sum_{i=1}^{p} \phi_i P_T X_{T+h-i} + \sum_{j=h}^{T+h-1} \theta_{T+h-i,j}(X_{T+h-j} - \hat{X}_{T+h-j}) & h > m-T \end{cases}$$

Si $T > m = \text{máx}\{p, q\}$, para $h \in \mathbb{N}^+$

$$P_T X_{T+h} = \sum_{i=1}^{p} \phi_i P_T X_{T+h-i} + \sum_{j=h}^{T+h-1} \theta_{T+h-i,j}(X_{T+h-j} - \hat{X}_{T+h-j})$$

Para T fijo, una vez que los predictores $\hat{X}_1, \ldots, \hat{X}_T$ se obtuvieron, es directo el cálculo de los predictores $P_T X_{T+1}, P_T X_{T+2}, P_T X_{T+3}, \ldots$.

El error cuadrático medio de $P_T X_{T+h}$ se puede calcular mediante la expresión

$$\sigma_T^2(h) := \mathbb{E}[(X_{T+h} - P_T X_{T+h})^2] = \sum_{j=0}^{h-1} \left(\sum_{r=0}^{j} \chi_r \theta_{T+h-r,j-r} \right)^2 v_{T+h-j-1}$$

donde los coeficientes χ_j se calculan recursivamente mediante $\chi_0 = 1$ y

$$\chi_j = \sum_{k=1}^{\text{mín}\{p,j\}} \phi_k \chi_{j-k}, \quad j = 1, 2, \ldots$$

Aproximaciones en muestras grandes

Si además, se supone que el proceso $ARMA(p, q)$, $\phi(B)(X_t) = \theta(B)(\epsilon_t)$, es causal e invertible se tienen las representaciones

$$X_{T+h} = \sum_{j=0}^{\infty} \psi_j \epsilon_{T+h-j}$$

y

$$\epsilon_{T+h} = X_{T+h} + \sum_{j=1}^{\infty} \pi_j X_{T+h-j}$$

donde $\{\psi_j\}$ y $\{\pi_j\}$ son únicos.

Sea $\tilde{P}_T Y$ la aproximación con menor error cuadrático medio de Y que es combinación lineal ó límite de combinaciones lineales de $\{X_t\}$, $-\infty < t \leq T$ (ó equivalentemente de Z_t, $-\infty < t \leq T$). Aplicando P_T en ambos lados de las

$$X_{T+h} = \sum_{j=0}^{\infty} \psi_j \epsilon_{T+h-j}$$

y

$$\epsilon_{T+h} = X_{T+h} + \sum_{j=1}^{\infty} \pi_j X_{T+h-j}$$

se tiene que

$$\tilde{P}_T X_{T+h} = \sum_{j=h}^{\infty} \psi_j \epsilon_{T+h-j} \tag{2.49}$$

y

$$\tilde{P}_T X_{T+h} = -\sum_{j=1}^{\infty} \pi_j \tilde{P}_T X_{T+h-j} \tag{2.50}$$

Para $h = 1$, el j-ésimo término del lado derecho de (2.50) es simplemente X_{T+1-j}. Una vez que se evalúa $\tilde{P}_T X_{T+1}$, a partir de (2.50) se puede calcular $\tilde{P}_T X_{T+2}$. Los predictores $\tilde{P}_T X_{T+3}, \tilde{P}_T X_{T+4}, \ldots$ se pueden calcular sucesivamente de la misma forma. A partir de la ecuación (2.49) se tiene que el error de predicción está dada por

$$X_{T+h} - \tilde{P}_T X_{T+h} = \sum_{j=0}^{h-1} \psi_j \epsilon_{T+h-j}$$

y entonces el error cuadrático medio está dado por

$$\tilde{\sigma}^2(h) = \sigma^2 \sum_{j=0}^{h-1} \psi_j^2$$

Los predictores que se obtienen de esta manera se tiene la forma

$$\tilde{P}_T X_{T+h} = \sum_{j=0}^{\infty} c_j X_{T-j} \tag{2.51}$$

En la práctica, sólo se tienen observaciones $X_1, \ldots, X_T$ por tanto se tiene que truncar la serie (2.51) después de T términos. El predictor resultante es una aproximación útil de $P_T X_{T+h}$ si T es grande y los coeficientes c_j convergen a 0 cuando j crece. Además, $\tilde{\sigma}^2(h)$ es una aproximación de $\sigma_T^2(h)$ para T grande.

Si en el proceso $ARMA(p, q)$, $\{X_t\}$, el ruido blanco $\{\epsilon_t\}$ es un ruido blanco Gaussiano, entonces para todo $h \in \mathbb{N}^+$ el error de predicción $X_{T+h} - P_T X_{T+h}$ tiene distribución Gaussiana con media 0 y varianza $\sigma_T^2(h)$.

Si $c_{1-\alpha/2}$ es el cuantil al nivel $1-\alpha/2$ de la distribución normal estándar, entonces X_{T+h} está acotado por $P_T X_{T+h} \pm c_{1-\alpha/2}\sigma_T(h)$ con probabilidad $1 - \alpha$. Estas cotas se conocen como *cotas de predicción al nivel* $1 - \alpha$ *para* X_{T+h}.

Capítulo 3

Algunos modelos de series de tiempo para valores discretos

3.1. Introducción

Se entenderá por serie de tiempo entera a una serie de tiempo con espacio de estados discreto (generalemente $\mathbb{Z}$ ó algun conjunto identificado con $\mathbb{Z}$). El estudio de las series de tiempo enteras es una de las áreas de mayor interés en el análisis de series de tiempo; sin embargo es una de las menos desarrolladas. El hecho de que la serie tome exclusivamente valores enteros hace que muchos de los modelos tradicionales de dependencia sean poco factibles e incluso imposibles. En las últimas tres décadas se han desarrollado algunas metodologías para desarrollar una clase adecuada de modelos para estas situaciones. La idea de esta sección introductoria es revisar algunas de estas metodologías, casi de forma histórica.

Ejemplos de series de tiempo enteras hay muchos, algunas veces como conteos de eventos, objetos o individuos en intervalos consecutivos o en puntos del tiempo consecutivos. Algunos ejemplos de dichas series son:

1. El número de objetos defectuosos que se encuentran en muestras sucesivas en una línea de producción.
2. La sucesión de días con lluvia y sin lluvia en cierta ciudad.
3. El número de casos de alguna enfermedad en cierta zona en meses sucesivos.
4. El número de nacimientos y el número de métodos de parto de un hospital en meses sucesivos.
5. El número de accidentes de tráfico por día.

6. Las secuencias básicas de ADN.

7. El número de homicidios con arma de fuego en semanas sucesivas en cierta ciudad.

Para muchos de los modelos de series de tiempo enteras la simulación tiene un papel muy importante; generalmente se necesitará simular sucesiones de variables dependientes con una distribución marginal particular y estructura de correlación específica. A veces esto se hace simplemente por razones prácticas, por ejemplo en Phatarford & Mardia (1973) simularon el flujo de cierta presa de forma más realista. Algunas otras veces son más estadísticas, por ejemplo para evaluar los efectos de correlación serial, i.e. pruebas, estimadores, etc. desarrollados originalmente para datos independientes. A veces la motivación puede ser teórica y práctica. Por ejemplo, tales modelos son útiles para nuevas distribuciones discretas mutivariadas. En el trabajo de Phatarford & Mardia (1973) se usó una distribución bivariada propuesta por Edwards & Gurland (1961) en estudios de propensión a accidentes. Ambos pares de autores estuvieron menos interesados en la estructura subyacente del proceso generador involucrado que en sus distribuciones conjuntas. Estos procesos generadores son los procesos auto-regresivos Poisson y binomial (que se verán más adelante).

Además, en algunos casos los valores discretos son números relativamente grandes y tiene sentido aproximar estos mediante variables continuas. Sin embargo, esto no siempre es posible y es necesario encontrar modelos para series de tiempo de enteros relativamente pequeños. Adicionalmente, en muchas ocasiones, las series observadas presentan fuertes tendencias dependientes del tiempo o dependencia más general en covariables y generalmente, para propósitos prácticos, interesa la naturaleza de tales tendencias y dependencias. Muchos de los modelos en series de tiempo de conteo se analizaron inicialmente utilizando modelos lineales generalizados o algunas metodologías similares basadas en regresión y tratando cualquier correlación serial como un ruido que, en el mejor escenario, se puede explicar adecuadamente ajustando las tendencias.

Aunque ya es clara la necesidad de modelar y simular este tipo de datos, existen pocos modelos con este propósito. En la práctica, las cadenas de Markov representan una de las clases de modelos más usadas y que son adecuadas; sin embargo, tienden a estar sobreparametrizadas en la mayoría de los ejemplos prácticos (excepto quizá en los casos binarios o de dos estados). Con estos modelos basados en cadenas de Markov la estructura de correlación es limitada para las aplicaciones y no es fácil generelizar a modelos alternativos. Las cadenas de orden superior pueden ayudar con este tipo de problemas pero, de nuevo, con una gran cantidad de parámetros.

El gran interés en los modelos lineales simples de series de tiempo en el caso Gaussiano y sus correspondientes usos en la práctica, sistematizados y popularizados desde el trabajo de Box & Jenkins (1970), motivaron el desarrollo de los casos no-lineales y no-Gaussianos.

El resto de este capítulo consiste de varias secciones acerca de modelos ó clases de modelos en los que se estudiará en el orden en el que se propusieron (casi) históricamente. Se empezará con una breve sección de cadenas de Markov enfocándose en los métodos diseñados para simplificar el uso de cadenas de orden superior en series de tiempo que toman valores discretos. Después se dará una pequeña introducción de los proceso DARMA. Estos representan el primer intento real de definir una clase de modelos para series de tiempo discretas. Sus propiedades y estructura se basan en los procesos ARMA. Posteriormente, en la parte más extensa de este capítulo, se estudiarán modelos basados en adelgazamiento. Posteriormente se estudiarán modelos tipo regresión, éstos generalmente se basan en conceptos de modelos lineales generalizados que intentan incorporar tendencia y correlación serial adecuadamente. Finalmente, se dará una muy breve discusión con respecto a modelos espaciales y Bayesianos.

3.2. Modelos con cadenas de Markov

3.2.1. Cadenas de Markov saturadas

Ya que la propiedad de Markov es una suposición simple y matemáticamente manejable, ésta parece un buen primer intento para relajar la suposición de independencia. Es por esto que es bastante natural considerar cadenas de Markov en tiempo discreto con espacio de estados discreto como modelos de series de tiempo que toman valores discretos. Aunque más adelante algunos de los modelos serán cadenas de Markov con una estructura específica, ahora se analizarán los modelos que se conocen como modelos de *cadenas de Markov saturados* ó *completamente parametrizados*, esto significa que se considerarán cadenas de Markov cuyo espacio de estados tiene cardinalidad m y entonces se tendrán $m(m-1)$ probabilidades de transición.

A continuación se estudiarán las funciones de autocorrelación, autocorrelación parcial y estimación de las probabilidades de transición mediante máxima verosimilitud de cadenas de Markov estacionarias con espacio de estados $\{1, \ldots, m\}$.

Sea $\{X_n\}_{n\in\mathbb{N}}$ una cadena de Markov homogénea irreducible con espacio de estados $\{1, \ldots, m\}$ con matriz de probabilidades de transición $P = (p_{ij})$, donde como siempre

$$p_{ij} = \mathbb{P}(X_t = j | X_{t-1} = i)$$

En esta sección las cadenas que se considerarán son homogéneas (en el tiempo). Como la cadena es irreducible y tiene espacio de estados finito, existe una única distribución estacionaria, $\pi = (\pi_1, \ldots, \pi_m)$. Supóngase además que $\{X_t\}$ es estacionaria, i.e. para cualquier $t \in \mathbb{N}$, $X_t \sim \pi$.

Como $\{X_t\}$ es irreducible, entonces 1 es un eigenvalor de P y los eigenvectores correspondientes son únicos (excepto por constantes escalares). Entonces, dichos eigenvectores son múltiplos del vector columna $\mathbf{1} = (1, 1, \ldots, 1)'$ y π, respectivamente.

Sean V una matriz en cuya diagonal están los números ordenados $1, 2, \ldots, m$ y fuera de la diagonal las entradas son cero y $p_{ij}(k) := (P^k)_{ij}$. Se tienen los siguientes resultados para la media de X_t y la covarianza entre X_t y X_{t+k} para cualquier $k \in \mathbb{N}^+$

$$\mathbb{E}(X_t) = \sum_{i=1}^{m} i \cdot \pi_i = \pi v'$$

$$\begin{aligned} \mathbb{E}(X_t X_{t+k}) &= \sum_{i}^{m}\sum_{j}^{m} ij\mathbb{P}(X_{t+k} = j | X_t = i)\pi_i \\ &= \sum_{i,j} (i \cdot \pi_i) p_{ij}(k) \cdot j = \pi V P^k v' \end{aligned}$$

donde $v = (1, 2, \ldots, m)$. Finalmente,

$$Cov(X_t, X_{t+k}) = \pi V P^k v' - (\pi v')^2.$$

Aún si P no fuese diagonalizable, se puede simplificar la covarianza si se escribe a P en su forma canónica de Jordan. Así que será suficiente suponer que P se puede escribir como QD_1Q^{-1}, donde Q y D_1 son de la forma

$$Q = (\mathbf{1}', R), \quad Q^{-1} = \begin{pmatrix} \pi \\ W \end{pmatrix}$$

y

$$D_1 = \begin{pmatrix} 1 & \mathbf{0} \\ \mathbf{0}' & D_2 \end{pmatrix},$$

donde D_2 es semidiagonal con los eigenvalores de P que no son 1 sobre la diagonal, unos y ceros arriba de la diagonal y ceros en todas las demás entradas.

Por tanto,

$$P^k = QD_1^kQ^{-1} = 1'\pi + RD_2^kW,$$

y entonces

$$Cov(X_t, X_{t+k}) = \pi V(1'\pi + RD_2^kW)v' - (\pi v')^2 = (\pi VR)D_2^k(Wv').$$

Como esta expresión es válida para $k = 0$, se tiene que la función de autocorrelación de $\{X_t\}$ está dada por

$$\rho_k = \frac{Cov(X_t, X_{t+k})}{Var(X_t)} = \frac{(\pi VR)D_2^k(Wv')}{\pi VRWv'}.$$

Esta expresión involucra las k-ésimas potencias de los eigenvalores de P.

Si P es diagonalizable se tendrá una estructura más ordenada. Si los eigenvalores de P, además del 1, son $\lambda_2, \ldots, \lambda_m$ entonces la matriz D_1 se puede escribir como $diag(1, \lambda_2, \ldots, \lambda_m)$, las columnas de Q son los correspondientes eigenvectores derechos de P y los renglones de Q^{-1} son los correspondientes vectores izquierdos de P. Entonces, para $k \in \mathbb{N}$,

$$\begin{aligned} Cov(X_t, X_{t+k}) &= \pi VQD_1^kQ^{-1}v' - (\pi v')^2 \\ &= aD_1^kb' - a_1b_1 = \sum_{i=2}^{m} a_ib_iw_i^k \end{aligned}$$

donde $a = \pi VQ$ y $b' = Q^{-1}v'$. Por lo tanto, $Var(X_t) = \sum_{i=2}^m a_ib_i$ y para $k \in \mathbb{N}$

$$\rho_k = Corr(X_t, X_{t+k}) = \frac{\sum_{i=2}^m a_ib_i\lambda_i^k}{\sum_{i=2}^m a_ib_i}.$$

Esta es una combinación lineal de las k-ésimas potencias de los eigenvalores $\lambda_2, \ldots, \lambda_m$ (si estos eigenvalores son diferentes), análogo al caso de la función de autocorrelación del proceso $AR(m-1)$ Gaussiano. Sin embargo, ya no hay analogía si se considera a la función de autocorrelación parcial. Se tiene que en el caso $m = 2$ se tiene que $\rho_k = \rho_1^k$ para cualquier $k \in \mathbb{N}$ y esto también ocurre en otros casos, por ejemplo cuando todos los eigenvalores λ_i son iguales ó $a_ib_i = 0$ para todos excepto un valor de i. En el caso $m = 2$ también ocurre que ρ_1 es el otro eigenvalor de P (además del 1).

Antes de considerar la función de autocorrelación parcial de la cadena de Markov, se analizarán brevemente algunos resultados generales y se recordarán otros acerca de las autocorrelaciones parciales. Para cualquier proceso estacionario de segundo orden $\{X_t\}$, la autocorrelación parcial ϕ_{kk} es la correlación obtenida de los residuales que se obtienen al llevar a cabo una regresión de X_t y X_{t+k} sobre las $k-1$ variables intermedias. Recuérdese que esta autocorrelación parcial se obtenía mediante la expresión

$$\phi_{kk} = \frac{\begin{vmatrix} 1 & \rho_1 & \rho_2 & \cdots & \rho_{k-2} & \rho_1 \\ \rho_1 & 1 & \rho_1 & \cdots & \rho_{k-3} & \rho_2 \\ \vdots & \vdots & \vdots & \ddots & \vdots & \vdots \\ \rho_{k-1} & \rho_{k-2} & \rho_{k-3} & \cdots & \rho_1 & \rho_k \end{vmatrix}}{\begin{vmatrix} 1 & \rho_1 & \rho_2 & \cdots & \rho_{k-2} & \rho_{k-1} \\ \rho_1 & 1 & \rho_1 & \cdots & \rho_{k-3} & \rho_{k-2} \\ \vdots & \vdots & \vdots & \ddots & \vdots & \vdots \\ \rho_{k-1} & \rho_{k-2} & \rho_{k-3} & \cdots & \rho_1 & 1 \end{vmatrix}} \tag{3.1}$$

Observación 3.2.1.

Nótese que la ecuación (3.1) se cumple para todo proceso estacionario de segundo orden; sin embargo, no siempre es válida para procesos con una distribución específica (como la auto-regresiva Gaussiana ó la de promedios móviles). ▽

Considérese una cadena de Markov $\{X_t\}$ con espacio de estados $\{1, 2, \ldots, m\}$. Mediante la ecuación (3.1) se puede verificar que para $m = 2$ y cualquier otro caso en el que $\rho_k = \rho_1^k$ (para todo $k \in \mathbb{N}$), se tendrá que ϕ_{rr} es cero para todo $r > 1$, tal como ocurre con el $AR(1)$ Gaussiano. El siguiente ejemplo de una cadena con tres estados muestra un caso en el que ϕ_{33} puede ser diferente de cero. Este comportamiento es diferente al del proceso $AR(2)$ estándar, para el cual $\phi_{33} = 0, \quad r \geq 3$.

Ejemplo 3.2.1. *(Cadena de tres estados)*

Considérese una cadena de Markov con espacio de estados $\{1, 2, 3\}$ con matriz de probabilidades de transición

$$\begin{pmatrix} 0 & 1 & 0 \\ \frac{13}{16} & 0 & \frac{3}{16} \\ 1 & 0 & 0 \end{pmatrix}$$

La función de autocorrelacion es $\rho_k = \frac{75}{124}(-\frac{3}{4})^k + \frac{49}{124}(-\frac{1}{4})^k$ y las primeras tres autocorrelaciones son $\rho_1 = -\frac{137}{248}$, $\rho_2 = \frac{181}{496}$ y $\rho_3 = -\frac{1037}{3968}$; por tanto $\phi_{33} \neq 0$ ▽

Con el fin de estimar los $m(m-1)$ parámetros p_{ij} $(i \neq j)$ a partir de una realización $x_1, \ldots, x_T$ de la cadena $\{X_t\}$, se considerará primero la verosimilitud condicionada a la primera observación

$$\prod_{i=1}^{m} \prod_{j=1}^{m} p_{ij}^{f_{ij}}$$

donde f_{ij} es el número de transiciones del estado i al estado j (y por tanto $\sum_{ij} f_{ij} = T-1$). Como $p_{ii} = 1 - \sum_{k \neq i} p_{ik}$, derivando con respecto a p_{ij} al logaritmo de esta verosimilitud e igualando a cero se tiene que

$$\frac{f_{ii}}{1 - \sum_{k \neq i} p_{ik}} = \frac{f_{ij}}{p_{ij}}$$

El estimador intuitivamente plausible, $\hat{p}_{ij} = \frac{f_{ij}}{\sum_{k \neq i} f_{ik}}$, se puede ver como un estimador máximo verosimil condicional de p_{ij}. Nótese que la suposición de estacionariedad de la cadena de Markov no se utilizó para obtener $\hat{p}_{ij}$. De forma alternativa, se puede usar la verosimilitud no condicional, i.e. para una cadena de Markov estacionaria $\{X_t\}$ dicha verosimilitud es simplemente

$$\pi_{x_1} \frac{f_{ii}}{1 - \sum_{k \neq i} p_{ik}} = \frac{f_{ij}}{p_{ij}}$$

y se puede maximizar numéricamente (para encontrar estimadores de las probabilidades de transición p_{ij}), sujeto a restricciones sobre las sumas por renglones y de no negatividad. Algunos casos especiales como, por ejemplo, la cadena de dos estados se pueden obtener expresiones explícitas para los estimadores máximo verosímiles no condicionales (Bisgaard & Travis (1991)).

En una cadena de Markov saturada de m estados que tiene $m(m-1)$ parámetros (libres) no se puede estimar con confianza si m es grande y la serie observada es corta. Hay aplicaciones de cadenas de Markov que suponen esta estructura particular de la matriz de probabilidades de transición, por ejemplo el modelo de cadena de Markov de Albert (1994) para una serie de tiempo de las categorías de severidad de cierta enfermedad para periodos de recaída y remisión. En esta aplicación se utilizó una cadena de Markov de diez estados caracterizada sólo por ocho parámetros.

3.2.1.1. Un modelo de de cadena de Markov no homogénea para una serie de tiempo binaria

El siguiente modelo fue propuesto por Azzalini (1994), fue diseñado específicamente para trabajar con datos binarios y es un excelente ejemplo de cómo se puede utilizar una cadena de Markov para contruir un modelo más sofisticado. Considérese una cadena de Markov no-homogénea $\{X_t\}$ que toma los valores 0 y 1, con matriz de probabilidades de transición de X_{t-1} a X_t dada por

$$\begin{pmatrix} 1 - {}_t p_0 & {}_t p_0 \\ 1 - {}_t p_1 & {}_t p_1 \end{pmatrix}.$$

Es decir, para $j = 0, 1$

$$ {}_tp_j = \mathbb{P}(X_t = 1 | X_{t-1} = j). $$

Dada esta estructura de Markov, la distribución de este proceso se puede especificar en términos de sus medias marginales $\theta_t := \mathbb{E}(X_t)$ y sus momios

$$ \zeta := \frac{{}_tp_1/(1 - {}_t\, p_1)}{{}_tp_0/(1 - {}_t\, p_0)}, $$

que supondrán constantes en el tiempo. Sin embargo, Azzalini propone relajar esta suposición permitiendo que ζ dependa de covariables que dependen del tiempo, i.e si se permite que θ_t dependa de un vector x_t de covariados que dependen del tiempo

$$ g(\theta_t) = x_t \beta' $$

donde $g : (0,1) \to \mathbb{R}$ es una función liga (de *link*) adecuada (tal como logit o probit) y β es un vector de k parámetros. A partir de la relación

$$ \theta_t = (\theta_{t-1}) {}_tp_1 + (1 - \theta_{t-1}) {}_tp_0 $$

y de la definición de ζ se puede obtener una expresión de términos de ζ, θ_{t-1} y θ_t para ${}_tp_j$

$$ {}_tp_j = \frac{\delta - 1 + (\zeta - 1)(\theta_t - \theta_{t-1})}{2(\zeta - 1)(1 - \theta_{t-1})} + j \frac{1 - \delta + (\zeta - 1)(\theta_t - \theta_{t-1} - 2\theta_1 \theta_{t-1})}{2(\zeta - 1)\theta_{t-1}(1 - \theta_{t-1})} $$

para $t = 2, 3, \ldots, T$, donde

$$ \delta^2 = 1 + (\zeta - 1)[(\theta_t - \theta_{t-1})^2 \zeta - (\theta_t + \theta_{t-1})^2 + 2(\theta_t + \theta_{t-1})] $$

La fórmula anterior se cumple para $\zeta \neq 1$, pero si $\zeta = 1$ entonces ${}_tp_j := \theta_t$. Haciendo $\theta_1 = \mathbb{P}(X_1 = 1)$, se especifica completamente al proceso, i.e. $X_1, X_2, \ldots, X_T$ es una cadena de Markov binaria no-homogénea tal que para cualquier $t \in \mathbb{N}$, $\mathbb{E}(X_t) = \theta_t$ y todos los momios para (X_{t-1}, X_t) en todos los casos son iguales a ζ.

Para una serie observada $x_1, x_2, \ldots, x_T$, la función de log-verosimilitud para los parámetros β y $\kappa = Log(\zeta)$ está dada por

$$ l(\beta, \kappa) = \sum_{t=1}^{T} x_t Log\left(\frac{{}_tp_{x_{t-1}}}{1 - {}_t\, p_{x_{t-1}}} \right) + Log\left(1 - {}_t\, p_{x_{t-1}}\right) $$

donde ${}_tp_j$ es como en la ecuación anterior para $t > 1$ y ${}_tp_{x_0} =: \theta_1$.

La maximización de dicha log-verosimilutud se hace numéricamente (dada la complejidad inducida por las derivadas con respecto a los parámetros).

Azzalini describió cómo los modelos con covariables de este tipo tienen efecto sobre la distribución marginal de una observación dada (a diferencia de Zeger & Qaqish (1988) en los que las covariables tiene influencia en la distibución condicional dado el pasado). Él aplicó este modelo para estudiar los factores de riesgo coronario en niños en edad escolar.

3.2.2. Cadenas de Markov de orden superior

En aquellos casos en los que las observaciones de un proceso con espacio de estados finito parece no satisfacer la propiedad de Markov, una posibilidad es utilizar una cadena de Markov de orden superior, i.e. un proceso estacástico $\{X_t\}$ con la siguiente generalización de la propiedad de Markov para algún $l \geq 2$

$$\mathbb{P}(X_t = x_t | X_{t-1} = x_{t-1}, X_{t-2} = x_{t-2}, \ldots) = \mathbb{P}(X_t = x_t | X_{t-1} = x_{t-1}, X_{t-2} = x_{t-2}, \ldots, X_{t-l} = x_{t-l})$$

Aunque este modelo pareciera no tener una estructura de cadena de Markov "usual" i.e. no se está utilizado una cadena de Markov de primer orden, se puede redefinir el modelo de tal forma que se obtenga un proceso equivalente que sí tenga la propiedad de Markov en el sentido usual. Si se hace $Y_t := (X_{t-l+1}, X_{t-l+2}, \ldots, X_t)$, entonces $\{Y_t\}$ es una cadena de Markov (en el sentido usual ó de primer orden) con espacio de estados $M \times \cdots \times M = M^l$ (donde M es el espacio de estados de $\{X_t\}$). Aunque algunas propiedades son más complicadas de obtener, en esencia no hay una nueva teoría detrás del estudio de cadenas de orden superior. Por ejemplo, si $\{X_t\}$ tuviera distribución estacionaria, ésta se podría encontrar determinando la distribución estacionaria de $\{Y_t\}$ y marginalizando con respecto a cualquiera de las l componentes de Y_t.

Por supuesto, el uso de una cadena de Markov de orden superior (en vez de primer orden) incrementa el problema de sobreparametrización: una cadena de Markov de orden l sobre m estados implicará $m^l(m-1)$ probabilidades de transición independientes (aunque $\{Y_t\}$ es una cadena de Markov de m^l estados, el número de probabilidades de transición independientes es menor que $m^{2l} - m^l$ pues muchas de las entradas de la matriz de probabilidades de transición son cero). Pegram (1980) y Raftery (1985a, 1985b) propusieron una clase de modelos más asequible (en términos de aplicaciones) para cadenas de orden superior. Para $m = 2$, los modelos de Raftery son parecidos a los de Pegram, pero para $m > 2$ los de Raftery son más generales. Los modelos de Pegram tienen $m + l - 1$ parámetros y los de Raftery $m(m-1) + l - 1$.

Los modelos de Raftery se conocen como modelos distribución de transición mixta (MTD) y se definen de la siguiente forma.

Definición 3.2.1. *Modelo de distribución de transición mixta (MTD)*

Sea $\{X_t\}$ un proceso estocástico con espacio de estados $M = \{1, 2, \ldots, m\}$. Si

$$P(X_t = j_0 | X_{t-1} = j_1, \ldots, X_{t-l} = j_l) = \sum_{i=1}^{l} \kappa_i q_{j_i j_0} \tag{3.2}$$

donde $\sum_{i=1}^{l} \kappa_i = 1$ y $\Upsilon \in M_{m\times m}(\mathbb{R})$, $\Upsilon = (q_{jk})$ es una matriz estocástica tal que el lado derecho de (3.2) está acotado por cero y por uno. Entonces se dice que $\{X_t\}$ sigue un modelo MTD.

El último requisito de la definición anterior induce m^{l+1} parejas de restricciones lineales en los parámetros y asegura que las probabilidades de transición en (3.2) efectivamente sean probabilidades; además, la condición sobre la suma de los renglones de Υ asegura que la suma de estas probabilidades sobre j_0 es uno.

Nótese que la definición de Raftery no supone que los parámetros κ_i sean no negativos. De hecho, si se hace esta suposición el lado derecho de la ecuación (3.2) es una combinación lineal convexa de probabilidades y automáticamente estará acotada por cero y uno, haciendo a las restricciones un tanto redundantes.

La suposición de que los parámetros son no negativos simplifica los cálculos involucrados en la estimación de parámetros. Si no se puede hacer esta suposición esta estimación se puede complicar dado el gran número de restricciones no lineales sobre los parámetros. Sin embargo, Raftery & Tavaré (1994) y Schimert (1992) describieron cómo se puede reducir el número de restricciones y por tanto hacer más accesible al modelo en la práctica.

En el caso $l = 1$, este modelo es un modelo de cadena de Markov de primer orden con matriz de probabilidades de transición Υ. Otro caso interesante de modelo MTD es el que se obtiene cuando se toma

$$\Upsilon = \theta I + (1 - \theta)\mathbf{1}'\pi,$$

donde $\pi = (\pi_1, \ldots, \pi_m)$ es un vector con entradas positivas de tal manera que $\sum_{j=1}^{m} \pi_j = 1$. Con esto se obtiene el modelo de Pegram. Entonces, el vector π es la distribución estacionaria correspondiente a Υ (aunque Pegram lo estudió definiendo a π como la distribución límite de X_t).

Raftery demostró el siguiente teorema bajo la suposición de que las entradas de Υ son positivas.

Proposición 3.2.1.

Si π es la distribución estacionaria correspondiente a Υ, entonces $X_t = j$ tiene probabilidad límite π_j independientemente de condiciones iniciales. Es decir para cualesquiera $j, i_1, \ldots, i_l \in M$

$$\mathbb{P}(X_t = j | X_1 = i_1, \ldots, X_l = i_l) = \pi_j$$

Por tanto, suena razonable restringir la atención a modelos estacionarios $\{X_t\}$. Para tales modelos estacionarios Raftery demostró que la densidad conjunta de X_t y X_{t+k} satisface un sistema de ecuaciones parecido al sistema de ecuaciones de Yule-Walker. Es decir, si se define $P(k) = (p_{ij}(k))$ mediante

$$p_{ij}(k) = \mathbb{P}(X_t = i, X_{t+k} = j), \quad i, j \in M, k \in \mathbb{Z},$$

donde $P(0) = diag(\pi)$ en el caso $k = 0$. Entonces para $k \in \mathbb{N}$

$$P(k) = \sum_{g=1}^{l} \kappa_g P(k-g)\Upsilon \tag{3.3}$$

No siempre es posible encontrar una solución única pero Raftery dió condiciones suficientes sobre los parámetros para garantizar la unicidad (en los casos $l \geq 2$). A partir de (3.3), se obtiene un sistema de ecuaciones para las autocorrelaciones $\rho_k = Corr(X_t, X_{t+k})$, de nuevo, muy parecidas a las de Yule-Walker que también, sólo se pueden resolver en algunos casos particulares. Él consideró muy detalladamente el comportamiento de autocorrelacines de su modelo cuando $m = 3, l = 2$, $\pi = \frac{1}{3}\mathbf{1}$ y Υ tiene una estuctura especial que implica que las autocorrelaciones satisfacen exactamente las ecuaciones de Yule-Walker

$$\rho_1 = \phi_1 + \phi_2 \rho_1$$

$$\rho_2 = \phi_1 \rho_1 + \phi_2,$$

para algunos ϕ_1, ϕ_2 que no dependen ni de ρ_1 ni de ρ_2. El conjunto de correlaciones admisibles para este modelos tienen una forma muy complicada y está contenido en el conjunto correspondiente para el modelo $AR(2)$ Gaussiano.

Raftery (1985a) ajustó su modelo a tres conjuntos de datos relacionados con fuerza del viento, relaciones interpersonales y movilidad ocupacional. Los parámetros se estimaron directamente de la máximización numérica del logaritmo de la verosimilitud condicional sobre las primeras l observaciones (si l es grande). La selección del orden del modelo se hizo

mediante el criterio de información de Bayes (BIC) de Schwarz.

Algunas aplicaciones adicionales se presentaron en Raftery & Tavaré (1994) para secuencias de ADN, direcciones del viento y patrones en el canto de aves. En varios de los modelos ajustados se obtienen estimados negativos para algunos de los coeficientes κ_i.

Adke & Deshmukh (1988) generalizaron el teorema límite de Raftery para modelos MTD en cadenas de Markov de orden superior con espacios de estados numerable. De hecho, ellos también extendieron el resultado de modelos de cadenas de Markov de orden superior con espacio de estados arbitrario haciendo la suposición de que los coeficientes κ_i son positivos.

Los méritos de varias de las técnicas que se pueden aplicar a los modelos de Raftery en varias situaciones de muestreo se discuten en Li & Kwok (1990). Para una sola realización la metodología de estimación máximo verosímil condicional se compara con una metodología de minimización de tipo ji-cuadrada mediante simulaciones en las que $m = 3, l = 2, \pi = \frac{1}{3}\mathbf{1}$. Con los dos métodos se obtienen estimados de Υ comparables pero el método de minimización de ji-cuadrada da un mejor estimado de κ_1. Para macro datos, i.e. datos que conjuntan las observaciones en varias copias de un modelo de Raftery, se puede utilizar simulación y mínimos cuadrados no-lineales pero los resultados sugieren que en algunos casos los estimadores que se obtienen son inconsistentes. Finalmente, Li & Kwok consideran datos desagrupados de un "panel" de sujetos que se supone siguen modelos de Raftery pero con parámetros posiblemente diferentes entre los sujetos; con la evidencia que un ejercicio de simulación, Li & Kwok recomiendan una metodología Bayesiana para la estimación de los parámetros de cada sujeto.

Azzalini (1983) discutió acerca de la estimación máximo-verosímil del parámetro m (el orden) y utilizó una cadena de Markov binaria de segundo orden para estudiar la eficiencia de tal estimación (de orden cero ó de orden uno) en un caso donde se puede comparar con la estimación máximo verosímil exacta. Su conclusión es que el método funciona suficientemente bien cuando no se dispone de un estimador máximo verosímil. Azzalini & Bowman (1990) hicieron un ajuste mediante un modelo de cadena de Markov de segundo orden binaria para representar las longitudes de las erupciones sucesivas del geyser *Old Faithful*.

Una generalización de los modelos de Raftery se conocen como *modelos de Markov de orden infinito* y es una propuesta de Mehran (1989). Esto permite una infinidad de términos $\kappa_i q_{j,j_0}$ en vez de los l términos que aparecen en la ecuación (3.2), con $\sum_{i=1}^{\infty} \kappa_i = 1$ como antes. Sn embargo, en este caso los coeficientes κ_i están dados por alguna función paramétrica decreciente de i, por ejemplo, $\kappa_i = \varepsilon^{i-1}(1-\varepsilon)$ para algún $\varepsilon \in (0,1)$. Tales modelos son particularmente útiles cuando se tienen datos faltantes o esquemas de muestreo no consecutivo. Una sucesión finita se trata como una sucesión infinita en la cual faltan las observaciones desde cierto punto.

3.3. Modelos $DARMA$

En esta sección se estudiará un método simple para obtener una sucesión estacionaria de variables aleatorias dependientes que tengan una distribución marginal específica y una estructura de correlación (momentos conjuntos de segundo orden) también específica. Una ventaja de este modelo es que la especificación de los dos aspectos del modelo es independiente. Otra ventaja es que dicha sucesión se obtiene a partir de una transformación muy simple de una sucesión de variables aleatorias independientes.

A continuación, se definirán algunas cantidades que se usarán durante esta sección. Sea $\{Y_t\}$ una sucesión de variables aleatorias independientes que toman valores en un espacio discreto E cada una con distribución π. Sean $\{U_t\}$ y $\{V_t\}$ sucesiones independientes de variables aleatorias independientes idénticamente distribuidas tales que $U_t \sim Bnlli(\beta)$ y $V_t \sim Bnlli(\rho)$ con $\beta, \rho \in [0,1)$. Sea $\{S_t\}$ una sucesión de variables aleatorias independientes e idénticamente distribuidas tal que $sop(S_t) = \{1, 2, \ldots, M\}$ con $M \in \mathbb{N}$ y con función de densidad $f_S(\cdot)$.

El caso más general que se considerará es una sucesión de variables aleatorias definidas de acuerdo al siguiente modelo

$$X_t = U_t Y_{t-S_t} + (1 - U_t) A_{t-(M+1)} \quad t = 1, 2, \ldots \tag{3.4}$$

donde

$$A_t = V_t A_{t-1} + (1 - V_t) Y_t \tag{3.5}$$

para $t = -M, -M+1, \ldots$.

El modelo definido por (3.4) y (3.5) se conoce como modelo $DARMA(1, M+1)$ (proceso discreto mixto auto-regresivo de medias móviles con orden de auto-regresión 1 y orden de medias móviles $M+1$)

Si se empieza el proceso con $A_{-(M+1)}$ con distribución π, independiente de $\{Y_t\}_{t=-N}^{\infty}$, $\{U_t\}$, $\{V_t\}$ y $\{S_t\}$, entonces $\{X_t\}_{t=1}^{\infty}$ es una sucesión estacionaria de variables aleatorias dependientes que tiene distribución marginal π. En general, esta sucesión no será Markoviana, pero si $\beta = 0$ entonces sí lo será. La estructura de correlación se determina por los parámetros β, ρ y la distribución $f_S(\cdot)$. Algunos casos de distribuciones discretas de interés particular son las que se obtienen escogiendo a π como la distribución Poisson ó Geométrica.

En esta sección se estudiarán algunos casos particulares de procesos $DARMA(1, M+1)$, estos casos particulares ayudarán a que la nomenclatura sea más clara para procesos más generales. Nótese que si se selecciona $\beta = 1$ y $M = 0$ ó bien $\beta = 0$ y $\rho = 0$, entonces $\{X_t\}$

será una sucesión de variables aleatoras independientes todas con distribución π.

Otro nombre para el modelo definido en (3.4) y (3.5) es el modelo $DARMA$ hacia atrás (backward $DARMA$). El modelo hacia adelante (forward) se define de forma similar; sin embargo, dichos procesos son parecidos pero no equivalentes. Esto se debe a que en general $\{X_t\}$ no es reversible en el tiempo en el sentido de que $(X_1, \ldots, X_k)$ no tiene la misma distribución que $(X_{-k}, X_{-k+1}, \ldots, X_{-1})$. Las propiedades de ambos modelos se pueden obtener de forma análoga, aquí sólo se considerará el modelo hacia atrás.

Una motivación detrás de los modelos $DARMA$ es proporcionar un esquema para obtener modelos con los cuales se puedan analizar sucesiones estacionarias de variables aleatorias discretas dependientes con distribución y estructura de correlación específicas.

Otra motivación para el desarrollo de estos procesos es que sirven para estudiar procesos puntuales en los cuales los datos se dan en términos de conteos en intervalos de tiempo fijos en vez de los tiempos exactos de arribo. La mayoría de los modelos para procesos puntuales, además del proceso Poisson, se pueden describir muy fácilmente en términos de los tiempos llegada o los tiempos de entre llegadas, sin embargo es difícil obtener resultados acerca de la distribución conjunta de los conteos en los diferentes intervalos de tiempo fijos. Los modelos $DARMA$ se pueden utilizar en tales situaciones. Con estos modelos $DARMA$ también se puede modelar directamente procesos de conteo binarios en procesos puntuales en tiempo discreto.

3.3.1. Algunas propiedades preliminares del proceso $DARMA(1, M+1)$

Ahora se estudiarán algunas propiedades del proceso $DARMA(1, M+1)$. A menos que se especifique de otra forma se supondrá que $A_{-(M+1)}$ tiene distribución π y es independiente de $\{Y_t\}$, $\{U_t\}$, $\{V_t\}$ y $\{S_t\}$

3.3.1.1. La distribución marginal de X_t

A continuación se mostrará que X_t como se definió en (3.4) tiene distribución π para todo t. Nótese que de la expresión (3.5) se puede obtener recursivamente que A_t se puede "expandir" hacia atrás a su valor inicial $A_{-(M+1)}$ obteniendo $A_t = Y_{t-j}$ con probabilidad $\rho^j(1-\rho)$ para $j \in \{0, \ldots, M+t\}$ y $A_t = A_{-(M+1)}$ con probabilidad ρ^{M+t+1}, i.e. A_t es una mezcla de $Y_t, Y_{t-1}, \ldots, Y_{-M}$ y $A_{-(M+1)}$. Entonces, para $i \in E$

$$\mathbb{P}(A_t = i) = \sum_{j=0}^{M+t} \mathbb{P}(Y_{t-j} = i)\rho^j(1-\rho) + \mathbb{P}(A_{-(M+1)} = i)\rho^{M+t+1} = \pi(i), \;\; t = -N, -N+1, \ldots$$

Análogamente, para $i \in E, t \in \mathbb{N}^+$

$$\mathbb{P}(Y_{t-S_t}=i)=\sum_{j=0}^{M}\mathbb{P}(Y_{t-j}=i,S_t=j)=\pi(i)\sum_{j=0}^{M}\mathbb{P}(S_t=j)=\pi(i).$$

De (3.4) y (3.5) se sigue que para $i\in E, t\in\mathbb{N}^+$

$$\begin{aligned}\mathbb{P}(X_t=i) &= \beta\mathbb{P}(Y_{t-S_t}=i)+(1-\beta)\mathbb{P}(A_{t-(M+1)}=i)\\ &= \beta\pi(i)+(1-\beta)\pi(i)=\pi(i)\end{aligned}$$

$$\therefore \mathbb{P}(X_t=i)=\pi(i)$$

Por tanto, la distribución marginal de las X_t's es la misma que la de las Y_t's (que es π).

3.3.1.2. Propiedades de correlación

Aunque las X_t's tienen distribución estacionaria π, éstas no son independientes (como las Y_t que sí lo son). Esto se comprobará a través del cálculo de la covarianza entre X_t y X_{t+j}

$$\begin{aligned}Cov(X_{t+j},X_t) &= \beta^2 Cov(Y_{t+j-S_{t+j}},Y_{t-S_t})\\ &\quad+\beta(1-\beta)\left[Cov(Y_{t+j-S_{t+j}},A_{t-(M+1)})+Cov(Y_{t-S_t},A_{t-(M+1)+j})\right]\\ &\quad+(1-\beta)^2Cov(A_{t-(M+1)+j},A_{t-(M+1)})\end{aligned}\tag{3.6}$$

como S_{t+j} sólo toma valores en el conjunto $\{0,1,\ldots,M\}$, entonces para $j>M$,

$$Cov(Y_{t+j-S_{t+j}},Y_{t-S_t})=0$$

y para $j\in\{1,\ldots,M\}$

$$Cov(Y_{t+j-S_{t+j}},Y_{t-S_t})=Var(Y_1)\sum_{k=j}^{M}\mathbb{P}(S_{t+j}=k)\mathbb{P}(S_t=k-j).$$

Por tanto, para $j\in\{1,\ldots,M\}$

$$Corr(Y_{t+j-S_{t+j}},Y_{t-S_t})=\sum_{k=j}^{M}f_S(k)f_S(k-j)=\sum_{k=0}^{M-j}f_S(k)f_S(k+j).$$

Análogamente,

$$Cov(Y_{t-S_t}, A_{t-(M+1)+j}) = \sum_{k=0}^{M} \mathbb{E}(Y_{t-k}A_{t-(M+1)+j}|S_t = k) - \mathbb{E}(Y_{t-k}|S_t = k)\mathbb{E}(A_{t-(M+1)+j})$$

Por tanto, para $j \in \{1, \ldots, M\}$

$$\begin{aligned} Corr(Y_{t-S_t}, A_{t-(M+1)+j}) &= (1-\rho) \sum_{k=M+1-j}^{M} \rho^{k-(M+1-j)} f_S(k) \\ &= (1-\rho)\rho^{-(M+1-j)} \sum_{k=M+1-j}^{M} \rho^k f_S(k) \end{aligned}$$

y para $j > M$

$$Corr(Y_{t-S_t}, A_{t-(M+1)+j}) = \rho^{j-(M+1)} \sum_{k=0}^{M} \rho^k (1-\rho) f_S(k)$$

Ahora, considérese el segundo término de (3.6). Como S_{t+j} sólo toma valores en el conjunto $\{0, 1, \ldots, M\}$, para $j \in \mathbb{N}^+$ se tiene que

$$Corr(Y_{t+j-S_t}, A_{t-(M+1)}) = 0.$$

Por último, considérese el último término de (3.6). Se tiene que

$$\begin{aligned} &Cov(A_{t-(M+1)+j}, A_{t-(M+1)}) = \\ &\quad \rho Cov(A_{t-(M+1)+j-1}, A_{t-(M+1)}) + (1-\rho)Cov(Y_{t-(M+1)+j}, A_{t-(M+1)}) \end{aligned}$$

El segundo término es cero ya que $A_{t-(M+1)}$ es función sólo de $Y_{t-(M+1)}, Y_{t-(M+1)-1}, \ldots, Y_{-M}$ y $A-(M+1)$ y éstas son independientes de $Y_{t-(M+1)+j}$.

De nuevo, mediante un argumento inductivo se tiene que

$$Corr(A_{t-(M+1)+j}, A_{t-(M+1)}) = \rho^j \tag{3.7}$$

Con las expresiones (3.6) y (3.7) se obtendrá $\rho_\star(j)$, la correlación entre X_t y X_{t+j}.

Para $j \in \{1, \ldots, M\}$,

$$\rho_\star(j) = \beta^2 \sum_{k=0}^{M-j} f_S(k) f_S(k+j)$$
$$+\beta(1-\beta)\left[(1-\rho)\rho^{-(M+1-j)} \sum_{k=M+1-j}^{M} \rho^k \rho^k f_S(k)\right] + (1-\beta)^2 \rho^j$$

y para $j > M$

$$\rho_\star(j) = \beta(1-\beta)\rho^{j-(M+1)} \sum_{k=0}^{M} \rho^k (1-\rho) f_S(k) + (1-\beta)^2 \rho^j.$$

Nótese que $\rho_\star(j) \in [0,1]$ y para $j > N$, $\rho_\star(j)$ decrece geométricamente como función de j si $\rho \in (0,1)$. Ya que $\rho_\star(j)$ no depende de t, el proceso $DARMA(1, M+1)$ es estacionario de segundo orden.

3.3.1.3. Invarianza bajo transformaciones

Nótese que desde su definición X_t es una mezcla de las variables $Y_t, Y_{t-1}, \ldots, Y_{-M}$ y $A_{-(M+1)}$, i.e. es una selección aleatoria de una y sólo una de estas variables aleatorias. Por tanto, si se transforma cada una de las variables aleatorias $Y_t, Y_{t-1}, \ldots, Y_{-M}$ y $A_{-(M+1)}$ mediante la misma función, X_t se transformará de la misma forma y su distribución se obtendrá a partir de la transformación de $Y_t, Y_{t-1}, \ldots, Y_{-M}$ y $A_{-(M+1)}$.

Análogamente ocurrirá si se transforman las X_t's. Nótese que aplicar una transformación común individualmente a las X_t's no afectará el procedimiento de selección y por tanto, la estructura de correlación del proceso transformado será la misma que la del proceso no transformado. Esta invarianza marginal anterior ante transformaciones es importante para el análisis estadístico del proceso y también simplifica algunos cálculos (como se verá más adelante).

3.3.2. Propiedades de mezcla

Muchas características del proceso DARMA se siguen del hecho de que X_t es una mezcla de las variables aleatorias $A_{-M-1}, Y_{-M}, Y_{-M+1}, \ldots, Y_t$. Por ejemplo, si f es una función medible con codominio en $\mathbb{R}$, entonces $f(X_t)$ es también una mezcla de las variables aleatorias

$$f(A_{-M-1}), f(Y_{-M}), f(Y_{-M+1}), \ldots, f(Y_t).$$

De hecho, $\{f(X_t)\}$ es un proceso $DARMA$ con los mismos parámetros β, ρ y f_S que $\{X_t\}$ pero tendrá diferente distribución marginal.

Sea $\{X_t\}$ un proceso $DARMA(1, M+1)$. De (3.4) y (3.5) se tiene que la dependencia de X_t sobre $X_{t-1}, \ldots, X_1$ es a través de $Y_{t-1}, Y_{t-2}, \ldots, Y_{t-M}, A_{t-M-1}$. Primero se considerarán algunas propiedades asintóticas de $\{A_t\}$.

Por (3.4), $\{A_t\}$ es una cadena de Markov, i.e. A_{n+1} es condicionalmente independiente de

$$A_{-M-1}, A_{-M}, \ldots, A_{n-1}$$

dado A_n. Para $x \in \mathbb{R}$ y $B \in \mathscr{B}(\mathbb{R})$, sus probabilidades de transición están dadas por

$$P(x, B) = \mathbb{P}(A_n \in B | A_{n-1} = x) = \rho I_B(x) + (1-\rho)\pi(B).$$

Mediante un argumento inductivo se tiene que para $n \in \mathbb{N}^+$

$$P^n(x, B) = \rho^n I_B(x) + (1-\rho^n)\pi(B) \tag{3.8}$$

y por tanto

$$|P^n(x, B) - \pi(B)| = \rho^n |I_B(x) - \pi(B)| \leq \rho^n.$$

A continuación se probabará que el proceso $DARMA(1, M+1)$ es una ϕ-mezcla.

Proposición 3.3.1.

El proceso $DARMA(1, M+1)$ es una ϕ-mezcla con

$$\phi(n) := \begin{cases} 1 & \text{si } n \in \{1, 2, \ldots, M\}; \\ \rho^{n-M-1} & \text{si } n > M \end{cases} \tag{3.9}$$

Demostración:

Defínanse los eventos $B := f(X_1, \ldots, X_n) = b_*$ y $C := g(X_{n+m+M+1}, X_{n+m+M+2}, \ldots) = c_*$. Por (3.4), $X_{n+m+M+1}, X_{n+m+M+2}, \ldots$ es condicionalmente independiente de $X_1, \ldots, X_n$ dado A_n. Por tanto,

$$\begin{aligned} \mathbb{P}(B \cap C) - \mathbb{P}(B)\mathbb{P}(C) &= \mathbb{E}(f\mathbb{E}(g|A_n)) - \mathbb{E}(f)\mathbb{E}(g) \\ &= \mathbb{E}[f \sum_i \mathbb{E}(g|A_{n+m} = 1)(P^m(A_n, i) - \pi_i)] \\ &= \rho^m[\mathbb{E}(f\mathbb{E}(g|A_{n+m} = A_n)) - \mathbb{E}(f)\mathbb{E}(g)] \end{aligned} \tag{3.10}$$

donde $\pi_i := \mathbb{P}(Y_1 = i)$. Y de (3.8) y (3.10) se tiene que

$$|\mathbb{P}(B \cap C) - \mathbb{P}(B)\mathbb{P}(C)| \leq \rho^m \mathbb{P}(B)$$

□

3.3.3. Estimaciones del primer momento, percentiles y cuantiles

Ya se mostró que correlación entre X_t y X_{t+j}, $\rho_\star(j)$ para $j \in \{1, \ldots, M\}$, está dada por

$$\rho_\star(j) = \beta^2 \sum_{k=0}^{M-j} f_S(k) f_S(k+j) + \beta(1-\beta) \left[(1-\rho)\rho^{-(M+1-j)} \sum_{k=M+1-j}^{M} \rho^k \rho^k f_S(k) \right] + (1-\beta)^2 \rho^j$$

y para $j > M$

$$\rho_\star(j) = \beta(1-\beta)\rho^{j-(M+1)} \sum_{k=0}^{M} \rho^k (1-\rho) f_S(k) + (1-\beta)^2 \rho^j.$$

En particular, para $j \in \{1, 2, \ldots, M\}$

$$Corr(X_1, X_{1+j}) = (1-\beta)^2 \rho^j + \beta^2 \sum_{k=0}^{M-j} f_S(k) f_S(k+j) + \beta(1-\beta)(1-\rho)\rho^{-(M+1-j)} \cdot \sum_{k=M+1-j}^{M} \rho^k f_S(k) \tag{3.11}$$

y para $j \geq M+1$

$$Corr(X_1, X_{1+j}) = \rho^{j-M-1} \times \tag{3.12}$$
$$\rho^{j-M-1}[\beta(1-\beta)(1-\rho)(f_S(0) + \rho f_S(1) + \ldots + \rho^M f_S(M)) + (1-\beta)^2 \rho^{M+1}]$$

Nótese que la estructura de correlación del proceso se especifica independientemente de la distribución marginal, i.e. se puede seleccionar las Y_t's para dar X_t's con cualquier distribución marginal π.

En esta sección se obtendrán algunos resultados con respecto a la consistencia de estimadores asintóticos del primer momento, percentiles y cuantiles de la distribución marginal π. Se supondrá que los dos primeros momentos de la distribución π son finitos.

Sea $\zeta_t = \sum_{k=1}^{t} X_k$. Por la Ley Fuerte de Grandes Números, se satisface

$$\lim_{t\to\infty} \frac{\zeta_t}{t} = \mathbb{E}(X_1) \quad c.s. \tag{3.13}$$

Ahora se obtendrá una versión de Teorema de Límite Central. Nótese que

$$\sigma^2 := \lim_{t\to\infty} \frac{Var(\zeta_t)}{t} = c \cdot Var(X_1)$$

donde

$$c = 1 + 2\sum_{n=1}^{\infty} Corr(X_1, X_{1+n})$$

y con las ecuaciones (3.11) y (3.12) se tiene que

$$\sum_{n=1}^{\infty} Corr(X_1, X_{1+n}) =$$

$$\sum_{j=1}^{M} \left[(1-\beta)^2\rho^j + \beta^2 \sum_{k=0}^{M-j} F_S(k)F_S(k+j) + \beta(1-\beta)(1-\rho)\rho^{-(M+1-j)} \sum_{k=M+1-j}^{M} \rho^k F_S(k) \right]$$

$$+(1-\rho)^{-1}[\beta(1-\beta)(1-\rho)(F_S(0) + \rho F_S(1) + \ldots + \rho^M F_S(M)) + (1-\beta)^2\rho^{M+1}]$$

Nótese que σ^2 siempre es mayor que $Var(X_1)$, el valor para el caso de independencia, y se incrementa rápidamente cuando ρ se acerca a 1, i.e. cuando el efecto de auto-regresión se vuelve más grande. Una suposición natural es que la distribución marginal del proceso no es una distribución degenerada así que $Var(X_1)$ es estrictamente positiva, y ahora sí, por el Teorema del Límite Central se tiene que

$$\lim_{t\to\infty} \mathbb{P}\left(\frac{\zeta_t - t\mathbb{E}(X_1)}{\sqrt{t\sigma^2}} \leq x \right) = \Phi(x) \tag{3.14}$$

donde (como siempre) Φ es la función de distribución de una variable aleatoria normal estándar.

Las ecuaciones (3.13) y (3.14) muestran que el estimador $\frac{\zeta_t}{t}$ de $\mathbb{E}(X_1)$ es fuertemente consistente y tiene distribución asintótica normal con varianza σ^2, donde σ^2 es $c \cdot Var(X_1)$.

Con versiones adecuadas más restrictivas del Teorema del Límite Central se pueden encontrar estimadores para momentos de orden mayor como estimadores de coeficientes de correlación serial.

Ahora se intentará encontrar un estimador del i-ésimo percentil, i.e. $P(i) := \pi((-\infty, i])) = \mathbb{P}(Y_1 \leq i)$. Considérese

$$P_t(i) := \frac{1}{t} \sum_{k=1}^{t} I_{\{X_k \leq i\}}$$

Como $\{I_{\{X_k \leq i\}}\}$ se obtuvo a través de una transformación de $\{X_k\}$, entonces es un proceso $DARMA$ cuya correlación serial está dada por (3.11) y (3.12). De nuevo, con argumentos de Grandes Números se tiene que

$$\lim_{t \to \infty} P_t(i) = P(i) \quad c.s \tag{3.15}$$

y

$$\lim_{t \to \infty} \mathbb{P}\left(\frac{P_t(i) - P(i)}{\sqrt{\frac{c}{t} P(i)(1 - P(i))}} \leq x \right) = \Phi(x), \quad x \in \mathbb{R} \tag{3.16}$$

Por último se considerarán algunas propiedades asintóticas en la estimación de cuantiles de π.

Sea $a \in (0, 1)$, el a-cuantil de π es $\varsigma = \text{ínf}\{i : P(i) \geq a\}$. Sea Q_t el a-cuantil muestral de $X_1, \ldots, X_t$, i.e. Q_t es el valor más pequeño de $X_1, \ldots, X_t$ tal que al menos $\lfloor t \cdot a \rfloor$ son mayores que Q_t.

Como los eventos $\sup_{k \geq t}\{Q_k < \varsigma\}$ y $\text{ínf}_{k \geq t}\{P_k(\varsigma-) \geq a\}$ son equivalentes entonces

$$\limsup_{t \to \infty} \{Q_t < \varsigma\} = \liminf_{t \to \infty} \{P_t(\varsigma-) \geq a\}$$

Este último evento tiene probabilidad cero (por la definición de cuantil y (3.15)). Un argumento análogo lleva a

$$\mathbb{P}(\liminf_{t \to \infty} \{Q_t > \varsigma\}) = 0$$

por tanto

$$\lim_{t \to \infty} Q_t = \varsigma, \quad c.s.$$

Con esto se encontrará un intervalo de confianza para ς. Sean $X_{(1)} \leq X_{(2)} \leq, \ldots, \leq X_{(t)}$ las estadísticas de orden de $X_1, \ldots, X_t$.

Teorema 3.3.1.

$$\lim_{t\to\infty} \mathbb{P}\left(X_{\lfloor tP(\varsigma)-k(t)\rfloor} \leq \varsigma \leq X_{\lfloor tP(\varsigma)+k(t)\rfloor}\right) = \Phi(z) - \Phi(-z) \tag{3.17}$$

donde $k(t) = z\sqrt{tcP(\varsigma)(1-P(\varsigma))}$

Demostración:

El resultado es inmediato de (3.16) y del hecho de que

$$\mathbb{P}\left(X_{\lfloor tP(\varsigma)-k(t)\rfloor} \leq \varsigma \leq X_{\lfloor tP(\varsigma)+k(t)\rfloor}\right) = \mathbb{P}\left(\lfloor tP(\varsigma) - k(t)\rfloor \leq \sum_{k=1}^{t} I_{\{X_k\leq\varsigma\}} \leq \lfloor tP(\varsigma) + k(t)\rfloor\right)$$

□

En la práctica, para obtener un intervalo de confianza para ς usando la proposición anterior se necesita estimar $P(\varsigma)$ y c. La estimación de c se hará más adelante. Pero nótese que

$$|P_t(Q_t) - P(\varsigma)| \leq |P_t(Q_t) - P(Q_t)| + |P(Q_t) - P(\varsigma)|$$

Como $\lim_{t\to\infty} Q_t = \varsigma$ el término $|P(Q_t) - P(\varsigma)|$ converge a cero (cuando $t \to \infty$) y por el Teorema de Glivenko-Cantelli $|P_t(Q_t) - P(Q_t)|$ también converge a cero. Por tanto $P(Q_t)$ es un estimador consistente de $P(\varsigma)$.

3.3.4. Aplicación de la prueba χ^2 de bondad de ajuste

Ahora, se estudiarán algunos resultados acerca del comportamiento asintótico de la prueba de bondad de ajuste χ^2 para la distribución marginal de un proceso $DARMA(1, M+1)$ cuya hipótesis nula es que los datos provienen de una distribución marginal específica.

Primero se considerará el caso en el que π es una distribución discreta completamente especificada cuyo soporte es el conjunto $\{1, 2, \ldots, J\}$.

Sean $W_t(i) = I_{\{i\}}(X_t) - \pi_i$ para $i = 1, 2, \ldots, J$ y $\mathbf{W}_t = (W_t(1), W_t(2), \ldots, W_t(J))$. Como se está suponiendo que A_{-N-1} tiene distribución π, entonces $\{\mathbf{W}_t\}$ es una sucesión estacionaria tal que $\mathbb{E}(W_t(i)) = 0$ y $\mathbb{E}(W_t^2(i)) = \pi_i(1-\pi_i)$. Además, esta sucesión también es una ϕ-mezcla con la misma función que en la proposición (3.9). De nuevo por el Teorema del Límite Central $\frac{1}{t}\sum_{k=1}^{t} \mathbf{W}_k$ tiende a un vector aleatorio con distribución normal

multivariada con covarianzas $\sigma_{ii} = c\pi_i(1-\pi_i)$ y para $i \neq j$ $\sigma_{ij} = -c\pi_i\pi_j$, donde como se definió anteriormente

$$c = 1 + 2\sum_{n=1}^{\infty} Corr(X_1, X_{1+n})$$

Por tanto,

$$\frac{1}{\sqrt{t}}\sum_{k=1}^{t}\left(\frac{W_k(1)}{\sqrt{\pi_1}}, \ldots, \frac{W_k(1)}{\sqrt{\pi_1}}\right)$$

tiene distribución asintótica (cuando $t \to \infty$) normal multivariada con covarianzas $\sigma_{ii} = c(1-\pi_i)$ y para $i \neq j$, $\sigma_{ij} = -c\sqrt{\pi_i\pi_j}$. Por tanto, si

$$\wp_i(t) := \sum_{k=1}^{t} I_{\{i\}}(X_k)$$

se tiene el siguiente resultado.

Proposición 3.3.2.

Cuando $t \to \infty$ la distribución de

$$\sum_{i=1}^{J} \frac{(\wp_i(t) - t\pi_i)^2}{tc\pi_i}$$

tiende a la distribución $\chi^2_{(J-1)}$.

Supóngase que la distribución π (hipotética) depende de varios parámetros desconocidos $\alpha = (\alpha_1, \ldots, \alpha_m)$ $(m < J)$ que se estimarán a partir de datos muestrales. Sea $\hat{\alpha}$ el valor de α que maximiza la función

$$Ln(L(\alpha)) = \wp_1(t)Ln(\pi_1(\alpha)) + \ldots + \wp_J(t)Ln(\pi_J(\alpha))$$

Esta es la log-verosimilitud que se obtendría si se hubiese supuesto que $X_1, X_2, \ldots$ son independientes y todas tienen distribución π.

Proposición 3.3.3.

Supóngase que todos los puntos α en un conjunto compacto $K \subseteq \mathbb{R}^m$ y para todo $i \in \{1, \ldots, J\}$ satisfacen:

(i) $\sum_{i=1}^{J} \pi_i(\alpha) = 1$

(ii) Para todo $i \in \{1, 2, \ldots, J\}$, $\pi_i(\alpha) > d > 0$ (con d constante)

(iii) Para todo $i \in \{1, 2, \ldots, J\}$

$$\frac{\partial \pi_i(\alpha)}{\partial \alpha_j}, \ \frac{\partial^2 \pi_i(\alpha)}{\partial \alpha_i \alpha_j}$$

son continuas

(iv) La matriz de $J \times m$ cuya entrada (i, j) es $\frac{\partial \pi_i(\alpha)}{\partial \alpha_j}$, tiene rango m.

Entonces, si la verdadera distribución marginal subyacente es $\pi(\alpha^0)$ (con $\alpha^0 \in int(K)$) entonces

$$\hat{\alpha} \overset{t\to\infty}{\to} \alpha^0, \quad \text{en probabilidad}$$

y además la distribución de

$$\sum_{i=1}^{J} \frac{(\wp_i(t) - t\pi_i(\hat{\alpha}))^2}{tc\pi_i(\hat{\alpha})}$$

tiende a la distribución $\chi^2_{(J-m-1)}$.

Demostración:

Sea π_i^0 la probabilidad $\pi_i(\alpha^0)$ de que X_t sea igual a i, $i = 1, 2 \ldots, J$. Por la Ley Fuerte de los Grandes Números

$$\lim_{t\to\infty} \frac{Var(\wp_i(t))}{t} = c \cdot \pi_i(\alpha^0)(1 - \pi_i(\alpha^0)).$$

Por la desigualdad de Chebyshev, para $\lambda > 0$

$$\mathbb{P}(|\wp_i(t) - t\pi_i^0| \geq \lambda\sqrt{t}) \leq \frac{\pi_i^0(1 - \pi_i^0)(c - O(t))}{\lambda^2} \leq \frac{\pi_i^0(c - O(t))}{\lambda^2}.$$

Por tanto,

$$\mathbb{P}(|\wp_i(t) - t\pi_i^0\| \geq \lambda\sqrt{t} \text{ para algún } i) \leq \frac{c - O(t)}{\lambda^2},$$

y el sistema de ecuaciones

$$\frac{\partial}{\partial \alpha_j} L = 0, \quad j = 1, \ldots, m$$

tiene solución única $\hat{\alpha}$ y además $\hat{\alpha}$ converge en probabilidad a α^0 cuando $t \to \infty$ (esto significa que se tiene manera de estimar los parámetros en la distribución marginal pero no necesariamente es la forma más eficiente). □

Sean

$$\wp^I(i) = \frac{\wp_i(t) - t\pi_i^0}{\sqrt{t\pi_i^0}}$$

y

$$\wp^{II}(i) = \frac{\wp_i(t) - t\pi_i(\hat{\alpha})}{\sqrt{t\pi_i(\hat{\alpha})}}$$

$\wp^I = (\wp^I(1), \ldots, \wp^I(J))$ y $\wp^{II} = (\wp^{II}(1), \ldots, \wp^{II}(J))$. Sea B la matriz cuya entrada (i, j) es $\frac{\alpha^0}{\sqrt{\pi_i^0}} \frac{\partial \pi_i}{\partial \alpha_j}$ Por Cramér (1946) se tiene que $\wp^{II} = A\wp^I + R$, donde $A = I - B(B'B)^{-1}B'$ y R tiende a 0 en probabilidad (cuando $t \to \infty$). Con argumentos similares a los que se usaron en la Proposición anterior, la distribución de $\wp^I$ tiende a la distribución normal con media 0 y matriz de varianzas-covarianzas $c(I - \mathbf{pp}')$, donde $\mathbf{p}$ es una matriz columna cuya i-ésima entrada es $\sqrt{\pi_i^0}$. Entonces la distribución límite de $\wp^{II}$ también es normal multivariada con media 0 y matriz de varianzas-covarianzas

$$c[I - B(B'B)^{-1}B'](I - \mathbf{pp}')[I - B(B'B)^{-1}B'] = c[I - \mathbf{pp}' - B(B'B)^{-1}B'].$$

Las dos proposiciones anteriores establecen que aplicar la prueba de bondad de ajuste χ^2 con la distribución marginal de un proceso $DARMA$ es lo mismo que si se aplicara a variables aleatorias independientes excepto que aparece una constante c. Por (3.11) y (3.12) la constante c no depende de la distribución marginal.

En todos los resultados límite que se han analizado hasta el momento, la constante

$$c = 1 + 2\sum_{t=1}^{\infty} Corr(X_1, X_{t+1})$$

está presente. Gracias a esto, en muchas aplicaciones se debe estimar dicho valor de c. En general, la estimación de c es difícil si no se conoce el orden de dependencia del proceso. Sin embargo, si M es conocido, de las ecuaciones (3.11), (3.12) y la función de ϕ-mezcla se pueden estimar los primeros $M+2$ coeficientes de correlación serial para el proceso de la forma usual en la que se obtuvo un estimador consistente de c para el proceso $DARMA(1, M+1)$.

De forma más precisa, sea $\hat{\rho}(j)$ el estimador usual del coeficiente de correlación serial de orden j, después de $j = M$, por ejemplo en $j = M+k$, las correlaciones seriales se pueden escribir como $\rho(M+k) = \rho^{k-1}\zeta$ donde ζ es la constante de (3.12). Por tanto,

$$c = 1 + 2[\rho(1) + \ldots + \rho(M) + (1-\rho)^{-1}\zeta].$$

Para estimar ζ y ρ se puede tomar $\hat{\rho}(M+1)$ y $\hat{\rho}(M+2)$ que convergen casi seguramente a $\rho(M+1)$ y $\rho(M+2)$, respectivamente; y tomar a $\hat{\zeta} = \hat{\rho}(M+1)$ y $\hat{\rho} = \frac{\hat{\rho}(M+1)}{\hat{\rho}(M+1)}$ que convergen casi seguramente a ζ y ρ, respectivamente. Por tanto,

$$\hat{c} = 1 + 2[\hat{\rho}(1) + \ldots + \hat{\rho}(M) + (1-\hat{\rho})^{-1}\hat{\zeta}]$$

que también converge casi seguramente a c.

Como una característica práctica, para muestras pequeñas se tendrán problemas con este estimador pues $\hat{\rho}(M+2) \geq \hat{\rho}(M+1)$. Por tanto, es preferible suavizar el decreciemiento geométrico de las correlaciones seriales y estimar por ejemplo,

$$\hat{\rho} = \frac{[\hat{\rho}(M+2) + \hat{\rho}(M+3) + \hat{\rho}(M+4)]/3}{[\hat{\rho}(M+1) + \hat{\rho}(M+2) + \hat{\rho}(M+3)]/3}$$

que converge casi seguramente a p.

3.3.5. El proceso auto-regresivo $DAR(1)$

Ahora, se estudiarán algunas propiedades del proceso $DAR(1)$, $\{A_t\}$. Como antes, se supondrá que $A_{-(M+1)}$ tiene distribución π. Por lo analizado en las secciones anteriores, $\{A_t\}$ es una sucesión estacionaria de variables aleatorias con distribución marginal π y correlaciones

$$\rho_A(j) = Corr(A_t, A_{t+j}) = \rho^j, \quad j \in \mathbb{N}^+$$

Se sigue de (3.5) que $\{A_t\}$ es una cadena de Markov, i.e. para cualquier $i \in E$

$$\mathbb{P}(A_{t+1} = i | A_1 = a_1, \ldots, A_t = a_t) = \mathbb{P}(A_{t+1} = i | A_t = a_t)$$

Además, la matriz de probabilidades de transición está dada por

$$\mathbb{P}(A_{t+1} = i | A_t = k) = p^A_{k,i} = \begin{cases} (1-\rho)\pi(i) & \text{si } k \neq i; \\ \rho + (1-\rho)\pi(i) & \text{si } k = i \end{cases} \tag{3.18}$$

Nótese que el análisis se hace de forma opuesta al que es habitual en la teoría de cadenas de Markov ya que primero se especificó la distribución estacionara asociada con la cadena y se especificó la estructura de dependencia (Markoviana) mediante un solo parámetro ρ. Además, si se cambia ρ no se afecta π. Cuando $\rho = 0$, se tiene una sucesión estacionaria de variables aleatorias independientes e idénticamente distribuidas (con distribución π).

El hecho de que $\{A_n\}$ sea una cadena de Markov con una función de transición particularmente simple, hace que muchos cálculos se faciliten. Por ejemplo, en series de tiempo discretas, las rachas de valores dados de las variables aleatorias X_t son útiles para análisis estadístico. Algunas propiedades de estas rachas para el proceso $DAR(1)$ se pueden obtener fácilmente. Para un estado $i \in E$ (fijo) sea $T_i = \text{ínf}\{n \in \mathbb{N}^+ : A_n \neq i\} - 1$; T_i es la longitud de la racha de i's empezando al tiempo 1, donde la longitud puede ser $0, 1, \ldots$. Entonces, para $n \in \mathbb{N}^+$

$$\mathbb{P}(T_i \geq n) = \mathbb{P}(A_1 = A_2 = \ldots = A_n = 1) = \pi(i)(p_{i,i})^{n-1}$$

y $\mathbb{P}(T_i \geq 0) = 1 - \pi(i)$. Por tanto,

$$\mathbb{E}(T_i) = \frac{\pi(i)}{1 - p_{i,i}} = \frac{\pi(i)}{(1-\rho)(1-\pi(i))}$$

Si $\rho = 0$, entonces $A_t = Y_t$ para $t \in \mathbb{N}^+$ y $\mathbb{E}(T_i) = \frac{\pi(i)}{1-\pi(i)}$, como se esperaba, pues $\{A_t\}$ es una sucesión de variables aleatorias independientes en este caso. Nótese que para $\rho \in [0, 1]$ se tiene que

$$\mathbb{E}(T_i) \geq \frac{\pi(i)}{1 - \pi(i)}$$

es decir, la longitud esperada de una racha de i's para un proceso $DAR(1)$ siempre es mayor o igual que la longitud esperada de la racha para una sucesión de variables aleatorias independientes. Además, el incremento en la longitud esperada de rachas es uniforme para todos los estados. Esto es una consecuencia de que se esté trabajando con una cadena de Markov de un parámetro.

No es difícil calcular la función generadora de probabilidad de T_i, que está dada por

$$\begin{aligned} G_{T_i}(z) &= \sum_{n=0}^{\infty} z^n \mathbb{P}(T_i = n) \\ &= 1 - pi(i) + \pi(i) \sum_{n=1}^{\infty} z^n (p_{i,i})^{n-1}(1 - p_{i,i}) \\ &= 1 - pi(i) + \frac{z(1 - p_{i,i})\pi(i)}{1 - zp_{i,i}} \\ &= \frac{(1 - \pi(i))(1 - z\rho)}{1 - z\pi(i) - z\rho(1 - \phi(i)} \end{aligned}$$

para $z \in [0, 1]$. De nuevo, si $\rho = 0$, G_{T_i} se reduce a $\frac{1-\pi(i)}{1-z\pi(i)}$, que es precisamente la expresión para la función generadora de probabilidad de la longitud de una racha de i elementos para una sucesión de variables aleatorias independientes con distribución marginal π.

3.3.6. Proceso $DMA(M)$

Ahora, se considerará el proceso $DMA(M)$, $X_t = Y_{t-S_t}$. A diferencia del proceso $DAR(1)$, en general el proceso $DMA(M)$ no es Markoviano.

3.3.6.1. Propiedades de correlación

Por lo que se vió en la secciones anteriores, $\{X_n\}$ es una sucesión estacionaria de variables aleatorias con distribución marginal π y en general las correlaciones entre X_t y X_{t+j}, $\rho_\star(j)$, están dadas por

$$\begin{aligned} \rho_\star(j) = \beta^2 \sum_{k=0}^{M-j} f_S(k) f_S(k+j) \\ +\beta(1-\beta) \left[(1-\rho)\rho^{-(M+1-j)} \sum_{k=M+1-j}^{M} \rho^k \rho^k f_S(k) \right] + (1-\beta)^2 \rho^j \end{aligned}$$

para $j \in \{1, \ldots, M\}$ y para $j > M$

$$\rho_\star(j) = \beta(1-\beta)\rho^{j-(M+1)} \sum_{k=0}^{M} \rho^k (1-\rho) f_S(k) + (1-\beta)^2 \rho^j.$$

En particular para el proceso $DMA(M)$, para $j \in \{1, \ldots, M\}$

$$\rho_{\star M}(j) = Corr(X_t, X_{t+j}) = \sum_{k=0}^{M-j} f_S(k) f_S(k+j) = \sum_{v=j}^{M} f_S(v) f_S(v-j)$$

y para $j > M$ se tiene que $\rho_{\star M}(j) = 0$ y $\rho_{\star M}(0) = 1$. Obsérvese que cuando $M = 1$, el valor máximo de la correlación serial de primer orden es $\max_{f_S(0)} \rho_{\star M}(1) = \max_{f_S(0)}[f_S(0)(1 - f_S(0))] = 1/4$. De hecho, se puede mostrar que para cualquier $M \in \mathbb{N}^+$ la máxima correlación serial de primer orden, $\rho_{\star M}(1)$, que se puede alcanzar es $\frac{1}{4}$. También se puede maximizar la correlación en cualquier punto, por ejemplo el punto j, haciendo $f_S(j) = f_S(j) = \frac{1}{2}$; sin embargo, todas las otras correlaciones serán cero en este caso.

3.3.6.2. Distribución conjunta y reversibilidad en el tiempo

En lo que resta de esta pequeña sección se centrará la atención en el proceso $DMA(1)$, es decir, si $\alpha = \mathbb{P}(S_t = 0)$, entonces

$$X_t = \begin{cases} Y_t & \text{con probabilidad } \alpha; \\ Y_{t-1} & \text{con probabilidad } 1 - \alpha \end{cases} \tag{3.19}$$

En este caso $\rho_{\star M}(j)(1) = \alpha(1-\alpha)$ y para $j \geq 2$, $\rho_{\star M}(j)(1) = 0$.

No es complicado calcular la transformada de Laplace-Stieltjes de la distribución conjunta de variables en la sucesión $DMA(1)$. Sea $M_*(s) = \mathbb{E}(\exp\{-sY_1\})$. Nótese que sobre el conjunto $\{S_1 = r_1, \ldots, S_t = r_t\}$ (donde $r_i = 0$ ó $r_i = 1$, $i = 1, \ldots, t$)

$$\exp\left(-\sum_{k=1}^{t} s_k X_k\right) = \exp\left\{-s_1(r_1 Y_0 + (1-r_1)Y_1) - \ldots - s_t(r_t Y_{t-1} + (1-r_t)Y_t)\right\}$$

$$= \exp\left\{-s_1 r_1 Y_0 - s_t(1-r_t)Y_t - (s_1(1-r_1) + s_2 r_2)Y_1 - \ldots - (s_{t-1}(1-r_{t-1}) + s_t r_t)Y_{t-1}\right\}.$$

Ya que $\{Y_k\}$ y $\{S_k\}$ son sucesiones independientes de variables aleatorias independientes

$$\begin{aligned} \psi_t(s_1, \ldots, s_t) &= \mathbb{E}\left[\exp\left(-\sum_{k=1}^{t} s_k X_k\right)\right] \\ &= \sum_{\mathbf{r}} M_*(s_1 r_1) M_*(s_t(1-r_t)) \left(\prod_{k=1}^{t-1} M_*(s_k(1-r_k) + s_{k+1} r_{k+1})\right) \left(\prod_{i=1}^{t} \alpha(1-r_i) + (1-\alpha) r_i\right) \end{aligned}$$

donde la suma se toma sobre el conjunto de vectores de dimensión t, $\mathbf{r} = (r_1, \ldots, r_t)$ con $r_i = 0$ ó $r_i = 1$.

Uno de los objetivos por los que se desea calcular la distribución conjunta es verificar la reversibilidad en el tiempo del proceso.

Con el fin de demostrar la reversibilidad en el tiempo, se necesita probar que $\psi_t(s_1, \ldots, s_t) = \psi_t(s_t, \ldots, s_1)$ para cualesquiera $s_1, \ldots, s_t \geq 0$. De la expresión anterior para $\psi_t(\cdot, \ldots, \cdot)$ se puede ver que $\psi_t(s_t, \ldots, s_1)$ tiene la misma forma que $\psi_t(s_1, \ldots, s_t)$ excepto que r_i y $1 - r_i$ se intercambian para $i \in \{1, \ldots, t\}$. Por tanto, para mostrar reversibilidad en el tiempo basta probar que si $\psi_t(s_1, \ldots, s_t) = g(\alpha, 1-\alpha)$, entonces

$$g(\alpha, 1-\alpha) = g(1-\alpha, \alpha).$$

Para ejemplificar el argumento, considérese el coeficiente $g_\Gamma(\alpha, 1-\alpha)$ del término

$$\Gamma_* = M_*(s_1)M_*(s_2)M_*(s_3)M_*(s_4+s_5)M_*(s_6)M_*(s_7)M_*(s_8)M_*(s_9+s_{10})M_*(s_{11})M_*(s_{12})\cdots M_*(s_t)$$

en la expresión anterior para $\psi_t(\cdot, \ldots, \cdot)$. (Se está suponiendo que $t > 12$). El coeficiente $g_\Gamma(\alpha, 1-\alpha) = \mathbb{P}(D_1 \cap E_1 \cap D_2 \cap E_2 \cap D_3)$ donde

$$D_1 = \cup_{i=0}^{1}(S_1 = i, S_2 = i, S_3 = 1) \cup \left(\cup_{k=2}^{3}(S_1 = 1, \ldots, S_{k-1} = 1, S_k = 0, \ldots, S_3 = 0)\right)$$

$$E_1 = (S_4 = 0, S_5 = 1)$$

$$D_2 = \cup_{i=0}^{1}(S_6 = i, S_7 = i, S_8 = 1) \cup \left(\cup_{k=7}^{8}(S_6 = 1, \ldots, S_{k-1} = 1, S_k = 0, \ldots, S_8 = 0)\right)$$

$$E_2 = (S_9 = 0, S_{10} = 1)$$

$$D_1 = \cup_{i=0}^{1}(S_1 = i, S_2 = i, S_3 = 1) \cup \left(\cup_{k=12}^{t}(S_{11} = 1, \ldots, S_{k-1} = 1, S_k = 0, \ldots, S_t = 0)\right)$$

Ya que $\{S_i\}$ son independientes y cada evento en la unión de cada D_i es disjunta, entonces $\mathbb{P}(D_i)$ es simétrico en α y $1-\alpha$, $i = 1, 2, 3$. Además, $\mathbb{P}(E_i) = \alpha(1-\alpha)$ para $i = 1, 2$ y los eventos $\{D_i\}$ y $\{E_i\}$ son independientes. Entonces $g_\Gamma(\alpha, 1-\alpha) = g_\Gamma(1-\alpha, \alpha)$. Con argumentos similares a los que se utilizaron cuando se probó que cualquier término de ψ_t tiene un coeficiente que es simétrica en α y $1-\alpha$. Por tanto, $g(\alpha, 1-\alpha) = g(1-\alpha, \alpha)$ y entonces el proceso $DMA(1)$ es reversible en el tiempo.

3.3.6.3. Longitud de las rachas del proceso $DMA(1)$

A continuación se considerarán las rachas del proceso $DMA(1)$. Sean $i \in E$ (fijo) y $T_i := \text{ínf}\{n \in \mathbb{N}^+ : X_n \neq i\} - 1$ la longitud de la racha de i iniciando al tiempo 1, donde la longitud puede ser $0, 1, \ldots$. Primero se calculará $\mathbb{E}(T_i)$.

Sean $a_0 := 1$ y para $n \in \mathbb{N}^+$ $a_n := \mathbb{P}(X_1 = X_2 = \ldots = X_n)$. Entonces,

$$a_1 = \mathbb{P}(X_1 = i),$$

$$a_2 = \mathbb{P}(X_1 = X_2 = i) = \alpha\pi(i)a_1 + \alpha(1-\alpha)\pi(i)a_0 + (1-\alpha)^2\pi^2(i),$$

e inductivamente,

$$a_{n+1} = \mathbb{P}(X_1 = X_2 = \ldots = X_{n+1} = i)$$

$$= \alpha\pi(i)a_n + \alpha(1-\alpha)\pi(i)a_{n-1} + \alpha(1-\alpha)^2\pi^2(i)a_{n-2} + \ldots + \alpha(1-\alpha)^n\pi^n(i)a_0 + (1-\alpha)^{n+1}\pi^{n+1}(i)$$

Por tanto,

$$\mathbb{E}(T_i) = \sum_{n=1}^{\infty} a_n = \pi(i) + \alpha\pi(i)\sum_{n=1}^{\infty} a_n + \frac{\alpha(1-\alpha)\pi(i)}{1-(1-\alpha)\pi(i)}\sum_{n=1}^{\infty} a_n$$

$$+(1-\alpha)^2\pi^2(i)\frac{1}{1-(1-\alpha)\pi(i)} + \alpha(1-\alpha)\pi(i)\frac{a_0}{1-(1-\alpha)\pi(i)}$$

$$= \pi(i) + \frac{(1-\alpha)^2\pi^2(i) + \alpha(1-\alpha)\pi(i)}{1-(1-\alpha)\pi(i)} + \left(\alpha\pi(i) + \frac{\alpha(1-\alpha)\pi(i)}{1-(1-\alpha)\pi(i)}\right)\mathbb{E}(T_i)$$

Resolviendo la ecuación para $\mathbb{E}(T_i)$ se tiene que

$$\mathbb{E}(T_i) = \frac{p(i)}{1-p(i)} = \frac{\pi(i)[1+\alpha(1-\alpha)(1-\pi(i))]}{(1-\pi(i))(1-\alpha(1-\alpha)\pi(i))} \tag{3.20}$$

donde

$$p(i) = \pi(i)(1+\alpha(1-\alpha)) - \pi^2(i)\alpha(1-\alpha) = \pi(i) + \alpha(1-\alpha)\pi(i)(1-\pi(i))$$

Si $\alpha = 0$ ó $\alpha = 1$, X_t es una sucesión de variables aleatorias independientes y $\mathbb{E}(T_i) = \frac{\pi(i)}{1-\pi(i)}$. Obsérvese que $\mathbb{E}(T_i) \geq \frac{\pi(i)}{1-\pi(i)}$ para $\alpha \in [0,1]$, i.e. la longitud esperada de una racha de números i, para un proceso $DMA(1)$ es mayor que la longitud esperada de una racha de una sucesión de variables aleatorias independientes. Para una distribución dada π, el valor máximo que puede tomar $\mathbb{E}(T_i)$ ocurre cuando $\alpha = \frac{1}{2}$, en este caso

$$\mathbb{E}(T_i) = \frac{\pi(i)}{1-\pi(i)}\left(1 + \frac{1}{4-\pi(i)}\right).$$

Ahora se obtendrá la función generadora de probabilidad de T_i. Usando cálculos similares a los que se hicieron para obtener (3.20) se puede ver que

$$\sum_{n=1}^{\infty} z^n \mathbb{P}(T_i \geq n) = \sum_{n=1}^{\infty} z^n a_n = \frac{z\pi(i)[1 + z\alpha(1-\alpha)(1-\pi(i))]}{(1-z\pi(i))[1 + z\alpha(1-\alpha)(1-\pi(i))]}$$

Por tanto,

$$\sum_{n=0}^{\infty} z^n \mathbb{P}(T_i = n) = 1 - \frac{1-z}{z}\left(\sum_{n=1}^{\infty} z^n a_n\right) = \frac{(1-\pi(i))(1-z\pi(i)\alpha(1-\alpha))}{1 - z\pi(i) - z^2\pi(i)(1-\pi(i))\alpha(1-\alpha)} \quad (3.21)$$

Nótese que para $\alpha = 0$ ó $\alpha = 1$, se tiene que $G_T(z) = \frac{1-\pi(i)}{1-z\pi(i)}$ (como se esperaba). A partir de (3.21) se pueden obtener momentos de orden mayor de la longitud de las rachas.

3.3.7. El proceso $DARMA(1,1)$ binario

Ahora se analizará el proceso binario $DARMA(1,1)$ en el cual X_n sólo puede tomar los valores 0 ó 1, i.e.

$$X_t = \begin{cases} Y_t & \text{con probabilidad } \beta; \\ A_{t-1} & \text{con probabilidad } 1-\beta \end{cases} \quad (3.22)$$

$$A_t = \begin{cases} A_{t-1} & \text{con probabilidad } \rho; \\ Y_t & \text{con probabilidad } 1-\rho \end{cases} \quad (3.23)$$

donde $\{Y_t\}$ es una sucesión de variables aleatorias que pueden tomar los valores 0 ó 1 con distribución común π. En general, el proceso $DARMA(1,1)$ no es Markoviano.

Las series de tiempo binarias son de particular importancia para modelar el proceso de conteo diferencial en procesos puntuales a tiempo discreto.

Cuando $M = 0$ se tiene que la correlación entre X_t y X_{t+j} está dada por

$$\rho_\star(j) = Corr(X_t, X_{t+j}) = \rho^{j-1}[\beta(1-\beta)(1-\rho) + (1-\beta)^2\rho].$$

Ahora se considerarán algunas propiedades de la longitud de las rachas. Para $i \in \{0,1\}$ (fijo), sea $T_i := \text{ínf}\{n \in \mathbb{N} : X_n \leq i\} - 1$ como antes. Se calcularán $\mathbb{E}(T_i)$ y la función generadora de probabilidad de T_i. Nótese que aunque en general $\{X_t\}$ no es una cadena de Markov, $\{(A_t, X_t)\}$ sí lo es. Para $i, j, k, l \in \{0,1\}$

$$\mathbb{P}(A_{t+1} = j, X_{t+1} = k | A_t = i, X_t = l) =: Q_k(i,j)$$

no depende de l. Si Q_k es la matriz cuya entrada (i,j) es $Q_k(i,j)$, entonces se tiene que

$$Q_0 = \begin{pmatrix} \rho(1-\beta) + (1-\rho(1-\beta))\pi(0) & (1-\rho)(1-\beta)\pi(1) \\ \beta(1-\rho)\pi(0) & \beta\rho\pi(0) \end{pmatrix}$$

y

$$Q_1 = \begin{pmatrix} \beta\rho\pi(1) & \beta(1-\rho)\pi(1) \\ (1-\rho)(1-\beta)\pi(0) & \rho(1-\beta) + (1-\rho(1-\beta))\pi(1) \end{pmatrix}$$

Nótese que para $i \in \{0,1\}$, $\mathbb{P}(T_0 \geq n | A_0 = i) = Q_0^n(i,0) + Q_0^n(i,1) =: Q_0^n(i,E)$. Por tanto,

$$\mathbb{E}(T_0 | A_0 = 1) = \sum_{n=1}^{\infty} Q_0^n(i,E) =: R_0(i,E) - 1$$

donde $R_0(i,j) = \sum_{n=1}^{\infty} Q_0^n(i,j)$, con $Q_0^0 = I$ y $R_0(i,E) = R_0(i,0) + R_0(i,1)$. Se puede probar que

$$R_0 = (I - Q_0)^{-1}$$

$$(I - Q_0)^{-1} = \frac{1}{\Delta} \begin{pmatrix} 1 - \beta\rho\pi(0) & (1-\beta)(1-\rho)\pi(1) \\ \beta(1-\rho)\pi(0) & 1 - \rho(1-\beta) - (1-\rho(1-\beta))\pi(0) \end{pmatrix}$$

donde

$$\Delta = det(I - Q_0) = (1-\pi(0))[1 - \rho(1-\beta)(1-\beta\pi(0)) - \beta\pi(0)(1-\beta(1-\rho))]$$

$$\begin{aligned} \mathbb{E}(T_0) &= \frac{\pi(0)[1 - \beta\rho + \beta(1-\rho-\beta+2\beta\rho)(1-\pi(0))]}{(1-\pi(0))[1-\rho(1-\beta)(1-\beta\pi(0)) - \beta\pi(0)(1-\beta(1-\rho))]} \\ &= \frac{\pi(0)[Corr(X_t, X_{t+1}) + 1 - \rho + \beta\rho - \pi(0)(Corr(X_t, X_{t+1}) + 2\beta\rho - \rho)]}{(1-\pi(0))[1-\rho(1-\beta)(1-\beta\pi(0)) - \beta\pi(0)(1-\beta(1-\rho))]} \end{aligned}$$

y

$$\begin{aligned} \mathbb{E}(T_1) &= \frac{\pi(1)[1 - \beta\rho + \beta(1-\rho-\beta+2\beta\rho)(1-\pi(1))]}{(1-\pi(1))[1-\rho(1-\beta)(1-\beta\pi(1)) - \beta\pi(1)(1-\beta(1-\rho))]} \\ &= \frac{\pi(1)[Corr(X_t, X_{t+1}) + 1 - \rho + \beta\rho - \pi(1)(Corr(X_t, X_{t+1}) + 2\beta\rho - \rho)]}{(1-\pi(1))[1-\rho(1-\beta)(1-\beta\pi(1)) - \beta\pi(1)(1-\beta(1-\rho))]} \end{aligned}$$

Nótese que la longitud esperada de una racha de i para el proceso $DARMA(1,1)$ binario es mayor o igual que la longitud esperada de una racha de i para el caso independiente.

Ahora, se intentará calcular la función generadora de probabilidad de T_i,

$$G_{T_i}(z) = \sum_{n=0}^{\infty} z^n \mathbb{P}(T_i = n)$$

, para $i = 0, 1$. Nótese que

$$\mathbb{P}(T_0 = n | A_0 = i) = \sum_j Q_0^n(i,j)(1 - Q_0(j,E))$$

Por tanto,

$$\sum_{n=0}^{\infty} z^n \mathbb{P}(T_0 = n | A_0 = i) = \sum_j \sum_{n=0}^{\infty} Q_0^n(i,j)(1 - Q_0(j,E))$$

Sin embargo,

$$\sum_{n=0}^{\infty} z^n Q_0^n = (I - zQ_0)^{-1}$$

donde

$$(I - zQ_0)^{-1} = \frac{1}{\Delta_0(z)} \begin{pmatrix} 1 - z\beta\rho\pi(0) & z(1-\beta)(1-\rho)\pi(1) \\ z\beta(1-\rho)\pi(0) & 1 - z[\rho(1-\beta) + (1-\rho(1-\beta))\pi(0)] \end{pmatrix}$$

con

$$\Delta_0(z) = 1 - z\rho(1-\beta) + z\pi(0)[-\beta - 1 + \rho(1-\beta)]$$
$$+z^2\pi(0)(1-\pi(0))(-\beta(1-\beta) + 2\rho\beta(1-\beta)) + z^2\pi^2(0)\beta\rho$$

Es decir,

$$G_{T_0}(z) = \frac{1}{\Delta_0(z)}[1 - z\rho(1-\beta) + \pi(0)(1-\pi(0))(z\beta^2 - z\beta^2\rho - z\beta - z\beta^2\rho) + \pi(0)(z\rho - 1) - z\beta\rho\pi^2(1)]$$

y análogamente,

$$G_{T_1}(z) = \frac{1}{\Delta_1(z)}[1 - z\rho(1-\beta) + \pi(1)(1-\pi(1))(z\beta^2 - z\beta^2\rho - z\beta - z\beta^2\rho) + \pi(0)(z\rho - 1) - z\beta\rho\pi^2(1)]$$

Se pueden obtener momentos de orden mayor para las rachas a partir de funciones generadoras de probabilidad. Dichos momentos son importantes pues ayudan a determinar qué tan diferente es el proceso $DARMA(1,1)$ de otros modelos Markovianos, por ejemplo el modelo $DAR(1)$; la diferenciación que se considera aquí es qué tanto las rachas se alejan de la distribución geométrica.

3.4. Modelos de regresión

Una metodología importante es incorporar covariables. Dicha metodología fue propuesta por Zeger (1988) en un estudio de la incidencia de polio en Estados Unidos. Se utilizó una regresión Poisson para modelar la tendencia y la ciclicidad explícitamente pero reconociendo explícitamente la presencia de correlación serial en los residuales. La solución de Zeger (1988) consistió en modelar la autocorrelación mediante un proceso estacionario no observado $\{\epsilon_t\}$ tal que $\mathbb{E}(\epsilon_t) = 1$ y función de autocovarianza $Cov(\epsilon_t, \epsilon_{t-k}) = \rho_\epsilon(k)\sigma^2$ para $k \in \mathbb{N}$. Se supone que $X_t|\epsilon_t = y \sim Poisson(\mu_t y, \mu_t y)$, donde μ_t es una función determinista tal que $Ln(\mu_t) = \beta' Y_t$, donde Y_t es un vector de covariables al tiempo t. Por tanto,

$$\mathbb{E}(X_t) = \mu_t, \quad Var(X_t) = \mu_t + \sigma^2\mu_t^2$$

y

$$\rho_X(k) = \frac{\rho_\epsilon(k)}{\sqrt{\left(1 + \frac{1}{\mu_t\sigma^2}\right)\left(1 + \frac{1}{\mu_{t-k}\sigma^2}\right)}}$$

Zeger notó que hay una separación completa de la esperanza (marginal) μ_t, que es función de las covariables y de los parámetros desconocidos del proceso $\{\epsilon_t\}$ (que define el mecanismo con el que se genera la correlación serial).

La correlación en la serie que estudió Zeger (la de la incidencia de polio) es numéricamente pequeña, por tanto supuso que $\{\epsilon_t\}$ tenía la estructura de un $AR(1)$, i.e. $\rho_\epsilon(k) = \rho^k$.

Campbell (1994) extendió esta metodología a órdenes de dependencia más altos cuando trabajó en casos de síndrome de muerte súbita de bebés en relación con factores ambientales. Y Brännas & Johansson (1994, 1996) también trabajaron en esta metodología y la extendieron para el caso de datos de panel y particularmente Brännäs (1995a) trabajó el cuestiones de predicción y control.

El estudio de series de tiempo binarias mediante de cadenas de Markov cuyas probabilidades de transición son funciones de covariables lo empezó Cox (1970) y los ejemplificó e implementó Korn & Whittemore (1979) y Muenz & Rubinstein (1985). En particular, estos últimos modelaron las probabilidades de transición p_{00} y p_{10} mediante regresión logística con covariables y lo aplicaron a estudios longitudinales en ciencias médicas.

Estas ideas fueron generalizadas de una manera natural por Fahrmeir & Kaufmann (1987). Aunque el origen de sus modelos era para series de tiempo categóricas, éstos eran adecuados para series de tiempo multinomiales (ó binomiales). Las probabilidades multinomiales condicionales, i.e. condicionadas a la historia pasada del proceso, se modelaban desde una perspectiva usual de modelos lineales generalizados (por ejemplo con log-momios).

En un contexto similar, Zeger *et al* (1985) observaron que al modelar las distribuciones marginales de esta forma no siempre se obtienen expresiones simples (ó útiles) para las marginales y por tanto no se podían estudiar algunas propiedades (marginales) importantes. Ellos también estaban interesados en modelar datos longitudinales de tal forma que se suponía que hay un vector de covariables Y_i, asociado con el i-ésimo individuo, que no depende del tiempo. La serie binaria, $\{X_{it}\}$, para el i-ésimo sujeto se describe mediante una cadena de Markov con $\pi_i := \mathbb{P}(X_{it} = 1|Y_i)$ dado por $logit(\pi_i) = \beta' Y_i$, y todos los sujetos comparten una misma estructura de correlación $Corr(X_{i,t}, X_{i,t-1}|Y_i) = \rho$. Con pocas suposiciones con respecto a la verdadera estructura de dependecia subyacente se pueden obtener estimados de los parámetros de regresión así como sus errores estándar.

Liang & Zeger (1986) extendieron esta idea de modelar la estructura marginal a través de covariables (y tratando la dependencia como ruido) pero que éstos sí dependan del tiempo y con otras distribuciones de nuevo en un marco de modelos lineales generalizados. Liang & Zeger no buscaron las distribuciones marginales multivariadas de la sucesión de observaciones sino que usaron un modelo lineal generalizado para especificar la forma de la distribución marginal de $X_{i,t}$. De nuevo, con muy pocas suposiciones acerca de la forma de la dependencia del tiempo, se puede hacer inferencia utilizando ecuaciones estimadoras, obteniéndose estimadores consistentes de los parámetros de regresión y de sus varianzas. Este procedimiento está muy relacionado con quasi-verosimilutud. Liang & Zeger comentaron que si la naturaleza de la dependencia del tiempo es importante en el análisis de series, entonces puede ser más razonable modelar directamente a la distribución condicional.

Zeger & Qaqish (1988) presentaron varios modelos basados en distribuciones condicionales y éstos se discutieron en el contexto de datos longitudinales en el libro de Diggle *et al.* (1994). En el caso binario, el modelo es simplemente una extensión del modelo de regresión logística. Por tanto

$$logit(p_t) = \beta' Y_t + \sum_{i=1}^{q} \theta_i X_{t-i} \tag{3.24}$$

donde $p_t = P(X_t = 1|Y_t, X_{t-1}, X_{t-2}, \ldots)$. La dependencia sobre $\{X_{t-i}\}_{i=1}^{q}$ puede tener una forma más complicada y si todos los θ_i fuesen cero se obtendrían los resultados usuales de regresión logística, en caso de que no sean cero se pueden interpretar en términos de momios.

La extensión a datos de conteo más generales es un poco más problemática y en la literatura ya se han considerado algunas alternativas para el caso en el que, condicionado a las observaciones pasadas y al vector de covariables Y_t, X_t tiene distribución $Poisson(\mu_t)$. El análogo directo de (3.24) en el caso Possion y tomando $q = 1$ (por simplicidad) es

$$Ln(\mu_t) = \beta' Y_t + \theta X_{t-1}$$

Sin embargo, esta forma tiene aplicaciones limitadas ya que la media condicional crece exponencialmente conforme X_{t-1} crece (a menos que $\theta < 0$). Por tanto, en este caso sólo se permite asociación negativa. Y no sólo eso, la interpretación "obvia" del modelo se dificulta ya que está involucrada la expresion $\exp(\beta' Y_t)$ en la media cuando la observación anterior fue cero, y por tanto se preferirá (tendrá que) aproximar ésta por una media (no condicional). Una alternativa más práctica está dada por

$$Ln(\mu_t) = \beta' Y_t + \theta(Ln(X^*_{t-1}) - \beta' Y_{t-1})$$

donde $X^*_{t-1} := \text{máx}\{X_{t-1}, c\}$ y $c \in (0,1)$. En esta nueva propuesta se "permite" correlaciones tanto positivas como negativas, dependiendo del signo de θ. Otra alternativa que permite incorporar a la constante c a todo X_{t-1} y no simplemente valores de cero está dada, por ejemplo, por

$$Ln(\mu_t) = \beta' Y_t + \theta[Ln(X_{t-1} + c) - Ln(\exp(\beta' Y_{t-1}) + c)]$$

Estos modelos surgen de manera natural en un contexto de procesos de ramificación con dependencia en el tamaño (de la ramificación); y por supuesto, toda la teoría de procesos de ramificación se puede utilizar en estos casos.

Zeger & Qaqish (1988) se refieren a estos modelos como *modelos de regresion Markovianos* y son una clase útil de modelos ya que son flexibles, relativamente fáciles de entender y describen la dependencia del pasado explícitamente (de la misma forma que lo hace la incorporación de covariables).

Li (1994) intentó extender estos modelos de regresion Markovianos de tal forma que se pueda "imitar" la estructura de promedios móviles, incorporando valores pasados a la sucesión de las medias condicionales. Ignorando covariables, la forma de "promedios móviles de orden 1" para datos binarios tiene la forma

$$logit(\mu_t) = \mu + \theta(X_{t-1} - \mu_{t-1})$$

y bajo la suposición de distribución Poisson (para las X_t) se reduce a

$$Ln(\mu_t) = \mu + \theta Ln\left(\frac{X_{t-1}}{mu_{t-1}}\right)$$

Por supuesto, se puede extender estos modelos a órdenes mayores.

3.4.1. Modelos *parameter-driven*

El modelo de Zeger (1988) para los datos de polio pertenece a una clase de modelos que Cox (1981) describe como *parameter-driven* ya que, condicionado a un proceso parametral

no observado $\{\epsilon_t\}$, los datos son independientes con sus distribuciones marginales determinadas por el valor del parámetro actual. Se han desarrollado varios métodos de series de tiempo para datos de conteo con este principio. Por ejemplo, Keenan (1982) modela una serie de tiempo estacionaria binaria, $\{X_t\}$, suponiendo la existencia de un proceso estacionario subyacente $\{Z_t\}$ y una función de distribución tal que $\mathbb{P}(X_t = 1|Z_t = z) = F(z)$. En particular, él considera el caso en el que marginalmente X_t tiene distribución $N(0, \sigma^2)$ con una función de autocorrelación específica y F la función de distribución de una variable $N(0, \tau^2)$. En este trabajo se obtienen algunas propiedades de tales modelos, incluyendo distribuciones conjuntas, predicción y estimación del proceso subyacente.

En principio, es sencillo construir tales modelos. Los datos tienen una forma de distribución estándar condicionados al proceso parametral que genera la estructura de correlación. Algunas aplicaciones, tales como Zeger (1988) y Campbell (1994), en las que se trabajó con componentes deterministas y la estructura de correlación se estudió con una mezcla de modelos estándar. Otra clase de modelos *parameter-driven* usa una cadena de Markov como proceso parametral. Éstos se conocen como modelos *hidden-Markov* y se describen detalladamente en el libro de MacDonald & Zucchini (1997) incluyendo aplicaciones para estos modelos. Un modelo simple es el *Poisson-hidden-Markov*. Se supone que $\{S_t\}$ es una cadena de Markov homogénea e irreducible con espacio de estados denotado por $\{1, 2, \ldots, m\}$ y que, condicionado a $S_1 = s_1, S_2 = s_2, \ldots, S_n = s_n$, $X_1, X_2, \ldots, X_n$ son independientes y X_t tiene distribución Poisson con media λ_{s_t}. Por tanto, hay m distribuciones Poisson con medias $\{\lambda_k\}_{k=1}^m$ y el proceso $\{X_t\}$ "escoge" su distribución marginal actual de acuerdo al estado de la cadena de Markov al tiempo t. Hay m^2 parámetros en este modelo: las m medias y las $m(m-1)$ probabilidades de transición para la cadena de Markov. Generalmente el número de estados m es relativamente pequeño (a veces tan pequeño como 2).

Evidentemente, la forma de este modelo se puede ocupar cualquier otra distribución en vez de la Poisson. MacDonald & Zucchini (1997) detalla las estructuras de correlación y distribucional además de algunos aspectos de inferencia. El problema de selección del modelo, en particular el número de estados de la cadena de Markov, sigue siendo complicado y no se ha resuelto satisfactoriamente (aunque parece que el uso de criterios de información, como el BIC, tienen buenas expectativas). También se puede hacer extensiones a cadenas de Markov de orden más alto, posiblemente usando la mezcla de distribución de transición como el de Raftery & Tavaré (1994), procesos multivariados y procesos en los cuales los parámetros son función del estado de la cadena y de covariables; por ejemplo, los coeficientes de regresión se determinan a partir del estado de la cadena. Esto último también se puede modelar permitiendo que las probabilidades de transición de la cadena de Markov sean funciones de covariables.

3.5. Modelos espaciales y Bayesianos

En la estructura lineal Gaussiana, los modelos espaciales son modelos *parameter-driven* en los cuales un vector de estados θ_t (no observado) evoluciona linealmente a través del tiempo y sólo se observa una función lineal de éste, generalmente un ruido. Por tanto, hay dos ecuaciones: una para la observación y otra para la evolución (del vector de estados)

$$X_t = F_t\theta_t + \nu_t \tag{3.25}$$

$$\theta_t = G_t\theta_{t-1} + \omega_t \tag{3.26}$$

donde $\{\nu_t\}$ y $\{\omega_t\}$ son sucesiones independientes de variables aleatorias independientes con media cero y matrices de varianzas-covarianzas V_t y W_t, respectivamente. Es común suponer que $\{F_t\}$, $\{G_t\}$, $\{V_t\}$ y $\{W_t\}$ son conocidas. Hay dificultades (como siempre) al extender una estructura de este estilo para datos exclusivamente discretos.

La metodología "tradicional" es la que proponen Durbin & Koopman (2000) y consiste en reemplazar (3.25) por la relación

$$p(x_t|\theta_t, x_{t-1}^*, y_t^*) = p(x_t|F_t\theta_t) \tag{3.27}$$

donde $x_{t-1}^* = \{x_{t-1}, x_{t-2}, \ldots, x_1\}$ y $y_t^* = \{y_t, x_{t-1}, \ldots, y_1\}$ son posibles covariados y F_t puede ser función de éstos. Es común suponer que el vector de estados θ_t aún satisface (3.26), pero ahora el proceso de innovación $\{\omega_t\}$ puede ser no-Gaussiano. En algunos modelos no se especifica la distribución (de la innovación), sólo se cuenta con los primeros dos momentos. Esta descripción es muy general y se puede usar en la mayoría de los casos de datos no-Gaussianos.

Los intentos más exitosos de esta generalización son los que se hicieron desde una perspectiva Bayesiana y a continuación se hará una muy breve descripción de éstos. Los principales promotores de la predicción desde la perspectiva Bayesiana son West & Harrison (1989) quienes presentan a detalle su metología en su libro (1989). En el caso Gaussiano, su modelo es lineal (como antes) y lo llaman modelo lineal dinámico. Ellos extienden este modelo a distribuciones no-Gaussianas pero sí de la familia exponencial y desarrollan lo que ellos llamaron modelo lineal dinámico generalizado (GLM). Como el nombre sugiere, se usa la metodología básica lineal general y, en principio, se puede modelar series binomiales, Poisson, binomial negativa y otras discretas. Por tanto, el parámetro natural η_t (de la parametrización natural de la familia exponencial) de la distribución se modela a través de una función liga (de *link*) $g(\cdot)$ como una función lineal del vector de estados θ_t, i.e. $g(\eta_t) = F_t\theta_t$; y θ_t evoluciona en el tiempo de acuerdo a la ecuacion (3.26). El análisis de este modelo es Bayesiano, pero no completamente, en el sentido de que la distribución a priori de η_t es la conjugada a priori para la familia exponencial pero el parámetro se obtiene aproximando las distribuciones correspondientes por el comportamiento de sus primeros dos

momentos. El trabajo se presentó originalmente en West, Harrison & Migon (1985), el cual tiene una discusión y algunos ejemplos.

Esta construcción, usando (3.27) para la familia exponencial y (3.26) para el predictor lineal η_t es muy atractivo en el contexto de series de tiempo discretas y constituye el modelo básico para otras metodologías en esta área. A veces se conoce como el GLM dinámico (modelo general lineal dinámico). Las principales diferencias entre distintas metodologías tienden a ser más computacionales y de detalles de la estimación.

Por ejemplo, Singh & Roberts (1992) definieron (3.26) sólo en términos de los primeros dos momentos y hacen estimación a través de un algoritmo iterativo de mínimos cuadrados ponderados filtrado. Kashiwagi & Yanagimoto (1992), basados en Kitigawa (1987), construyeron modelos espaciales no-Gaussianos en los cuales se obtienen las densidades no-Gaussianas en cada paso del filtro de Kalman (la mayoría de las veces numéricamente). Estos modelos tienen una estructura muy relacionada con los modelos lineales generalizados. Por ejemplo, en el caso Poisson se supone que la serie $\{X_t\}$ es Poisson de media λ_t y que el parámetro natural $Ln(\lambda_t)$ sigue una caminata aleatoria Gaussiana, i.e. $\nabla Ln(\lambda_t) \sim N(0, \sigma^2)$, donde $\nabla Ln(\lambda_t) := Ln(\lambda_t) - Ln(\lambda_{t-1})$.

Fahrmeir (1992) utilizó el modelo básico de (3.27) y (3.26), donde $\{\omega_t\}$ en (3.26) tiene distribución normal. Él se refiere a su versión como GLM dinámico y hace la estimación mediante la maximización de las densidades posteriores usando una generalización del filtro de Kalman extendido y suavizamiento.

Chan & Ledolter (1995) retomaron los datos de polio de Zeger (1988) y los modelaron como un GLM dinámico en el cual el proceso latente, $\{\theta_t\}$, es un $AR(1)$. Explícitamente, condicionado a los covariados de tendencia y ciclicidad, Y_t, $(X_t|Y_t, \theta_t) \sim Poisson(a_t\mu_t)$, donde $a_t = e^{\beta' Y_t}$ y $Ln(\mu_t) = \theta_t = \rho\theta_{t-1} + \omega_t$, donde ω_t es Gaussiano. Su propuesta consiste en considerar al proceso no-observado $\{\theta_t\}$ como datos faltantes y usar una variación del algoritmo EM para maximizar la verosimilitud. En su versión del algoritmo, el paso E se realiza indirectamente usando simulación Monte-Carlo. Esta metodología es directa y la predicción y el suavizamiento son claros. Ellos mencionaron que se necesitaban diagnósticos para esta clase de modelos.

El uso de técnicas de simulación, en particular las de cadenas de Markov Monte-Carlo y el Gibbs sampler, en esta clase de modelos la consideró Carlin *et al.* (1992) y posteriormente Carter & Kohn (1994) y Frühwirth-Schnatter (1994). Algunos ejemplos interesantes de estás técnicas en la época fueron los de Cargnoni *et al.* (1997) en un estudio de datos longitudinales que consistian de colecciones de vectores multinomiales que describian el número de estudiantes que permanecían en el sistema educativo italiano.

En el artículo de Lee & Nelder (1996) se discute acerca de los GLM jerárquicos, los GLM jerárquicos dinámicos y su relación con el filtro de Kalman y el suavizamiento. Una vez más, el modelo básico es el (3.27) para distribuciones de la familia exponencial y se supone (3.26). En esta formulación, el filtro se obtiene mediante un algoritmo de maximización para la verosimilitud jerárquica.

El artículo de Jørgensen *et al.* (1999) desarrolla el GLM dinámico incluyendo covariados en dos diferentes formas, distinguiendo entre dos tipos de covariables y reemplazando (3.26) por un proceso de Markov Gamma. Condicionando a los covariados $\{Y_t\}$, el proceso de obseraciones (3.27) es similar al que estudiaron Chan & Ledolter (1995), i.e. $(X_t|Y_t,\theta_t) \sim Poisson(a_t\theta_t)$, donde $a_t = e^{\beta' Y_t}$, pero (3.26) se reemplaza por $\theta_t|\theta_{t-1} \sim Gamma(\frac{\theta_{t-1}}{\sigma^2}, b_t\sigma^2)$. Por tanto, (3.26) se sigue cumpliendo ya que $\theta_t = b_t\theta_{t-1} + \omega_t$ donde ω_t es no-Gaussiano. Además, los autores distinguieron entre dos tipos de covariables: aquellos de corto plazo que aparecen en $\{Y_t\}$ y aquellos de largo plazo que aparecen en $\{Z_t\}$ y $Ln(b_t) = \alpha'\nabla(Z_t)$, donde $\nabla(Z_t) = Z_t - Z_{t-1}$. En la aplicación a los datos de polio, $\{Y_t\}$ contiene covariados de ciclicidad y $\{Z_t\}$ de tendencia. Una ecuación con la que se obtienen estimadores insesgados es la que se obtiene a partir del algoritmo EM en el cual el paso E se reemplaza por un suavizamiento de Kalman.

Durbin & Koopman (2000) utilizaron el mismo GLM dinámico (3.26) y (3.27) pero la inferencia estaba basada en el filtro de Kalman, suavizamiento, simulación (pero no MCMC) y muestreo de importancia (Gaussiano). Usando esta metodología, los autores mostraron cómo estimar todas las cantidades que usualmente se necesitan ya sea en un entorno clásico (frecuentista) o en un entorno Bayesiano. El hecho de no hacerlo mediante técnicas MCMC significó que no hay preocupaciones con cuestiones de convergencia y todas las cantidades que se requieren se calcularon fácilmente. La densidad de importancia necesaria para la simulación se construye a partir de modelos lineales Gaussianos que aproximan a la no-Gaussiana en una vecindad de la moda condicional del vector de estados dados los datos.

La descripción que se hará en lo que queda de esta pequeña sección intenta reflejar con precisión los diferentes enfoques para la inferencia sobre lo que es, en muchos sentidos, el mismo modelo. Se intenta hacer notar el importante papel de el auxilio computacional para la solución de problema que se pueden presentar en la inferecia y predicción de este tipo de modelos, y también la relativa complejidad de las soluciones.

Sin embargo, en un caso, hay soluciones que son relativamente directas e intuitivas. Es el caso más simple, pero el más usado: es el dado por las ecuaciones (3.25) y (3.26) con $F_t = 1$ y $G_t = 1$, i.e. la serie tiene una media θ_t que sigue una caminata aleatoria. Esta es una de las formas estándar de modelar una serie cuyo único parámetro es la media y ésta cambia a través del tiempo. Cuando se supone que el proceso de ruido es estacionario, la forma de la solución suele ser un simple promedio móvil con ponderación exponencial

cuya constante de suavizamiento es una función del cociente de las varianzas de los ruidos. Smith (1979) extendió este modelo a las distribuciones de la familia exponencial usando un método diferente al de West *et al.* (1985). Smith (1979) observó que la ecuación (3.25) se puede reescribir en este caso como $\mathbb{E}(X_t|\theta_t) = \theta_t$ y que aunque esto se puede extender fácilmente a otras distribuciones para este modelo, no hay una manera simple de generizar la forma de caminata aleatoria del mecanismo de evolución (3.26). El problema básicamente es que no se sabe como generalizar la "revisión" de la distribución de $(\theta_{t-1}|X^{t-1})$ a la de $(\theta_t|X^{t-1})$. Su solución requiere que la densidad de la segunda distribución sea proporcional a la densidad de la primera mediante un potencia fraccional de ω. Esto asegura que la "revisión" tenga propiedades similares al caso Gaussiano lineal, i.e. que las medias seran las mismas y las varianzas crezcan y que las predicciones del proceso sean simplemente un promedio móvil con ponderación exponencial. Él presentó detalles de la solución para algunas distribuciones, incluyendo la Poisson, la Binomial y la Binomial Negativa.

Harvey & Fernandes (1989) generalizan esta metodología (y Harvey la comenta a detalle en su libro en 1989). En el caso Poisson $X_t|\mu_t$ se distribuye Poisson con media λ_t y se supone además que μ_t tiene una apriori conjugada para la Poisson, i.e. la distribución Gamma. Se usará la parametrización de la densidad Gamma de la forma

$$f(x) = \frac{b^a}{\Gamma(a)} x^{a-1} e^{-bx} I_{(0,\infty)}(x)$$

$(\mu_{t-1}|X^{t-1}) \sim Gamma(a_{t-1}, b_{t-1})$ y según Smith (1979) $(\mu_t|X^{t-1}) \sim Gamma(a_{t|t-1}, b_{t|t-1})$, donde para $\omega \in [0,1]$

$$a_{t|t-1} := \omega a_{t-1} \quad \text{y} \quad b_{t|t-1} := \omega b_{t-1} \tag{3.28}$$

Con la observación de la siguiente observacion, X_t, la distribución posterior de μ_t condicionada a X^t es una $Gamma(a_t, b_t)$ donde $a_t := a_{t|t-1} + X_t$ y $b_t := b_{t|t-1} + 1$. Por tanto, se tiene que $\hat{\mu}_t = \mathbb{E}(\mu_t|X^t) = \frac{a_t}{b_t}$ y usando la relación entre (a_t, b_t) y (a_{t-1}, b_{t-1}), i.e.

$$a_t = \omega a_{t-1} + X_t \quad \text{y} \quad b_t = \omega b_{t-1} + 1 \tag{3.29}$$

se tiene que

$$\hat{\mu}_t = \hat{\mu}_{t-1} + k_t(X_t - \hat{\mu}_{t-1}) \tag{3.30}$$

$$k_t = \frac{k_{t-1}}{k_{t-1} + \omega} \tag{3.31}$$

Nótese que $k_t \overset{t\to\infty}{\to} (1-\omega)$ y en (3.30) $\hat{\mu}_t$ se obtiene por suavizamiento exponencial. De hecho, las ecuaciones (3.30) y (3.31) son las ecuaciones recursivas que se usan para encontrar el estimador de la media cuando se usa mínimos cuadrados descontados (con factor de

descuento ω).

Harvey (1989) estima el factor de "descuento" ω a partir de datos históricos maximizando la verosimilitud (que está dada en términos del producto de las densidades condicionales $p(x_t|X^{t-1})$). Estas probabilidades definen la distribución predictiva, que en este caso (Poisson con la Gamma aprori para su media) es binomial negativa, i.e. $(X_t|X^{t-1}) \sim BinNeg(a_{t|t-1}, (1+b_{t|t-1})^{-1})$.

También notó que aunque el mecanismo estocástico subyacente a la transición de μ_{t-1} a μ_t al tiempo $t-1$ se define sólo implícitamente mediante el comportamiento de los parámetros de la distribución Gamma, es posible especificar tal mecanismo explícitamente. De hecho, se puede escribir $\mu_t = \frac{\mu_{t-1}}{\omega\beta_t}$, donde β_t tiene distribución $Beta(\omega a_{t-1}, (1-\omega)a_{t-1})$. La parametrización de la distribución $Beta(\alpha, \beta)$ que se está considerando es

$$f(x) = \frac{1}{B(\alpha,\beta)} x^{\alpha-1}(1-x)^{\beta-1} I_{(0,1)}(x)$$

Este proceso refleja el hecho de que predecir μ_t de μ_{t-1} al tiempo $t-1$ sin ninguna nueva información, su ubicación no cambia pero la incertidumbre acerca de éste se incrementa.

La distribución binomial se trata análogamente. Si X_t tiene distribución $Bin(n, p_t)$ y $p_t|X^t \sim Beta(a_t, b_t)$, entones las ecuaciones (3.28) y (3.29) se convierten en

$$a_t = \omega a_{t-1} + X_t \quad \text{y} \quad b_t = \omega b_{t-1} + (n - X_t)$$

$\hat{\mu}_t = \mathbb{E}(np_t|X^t)$ y las ecuaciones (3.30) y (3.31) se siguen cumpliendo. La distribución predictiva en este caso es $Beta - Binomial$, i.e. $X_t|X^{t-1}$ tiene función de densidad de probabilidad

$$p(x) = \binom{n}{x} \frac{B(a+x, b+n-x)}{B(a,b)}$$

donde $a = a_{t|t-1}$ y $b = b_{t|t-1}$.

Como en el caso Poisson, es posible definir el mecanismo estocástico que especifica la transición de p_{t-1} a p_t al tiempo $t-1$. En este caso, esta dado por

$$p_t = \frac{B_a p_{t-1}}{B_a p_{t-1} + B_b(1-p_{t-1})}$$

donde B_a tiene distribución $Beta(\omega a_{t-1}, (1-\omega)a_{t-1})$, B_b tiene distribución $Beta(\omega b_{t-1}, (1-\omega)b_{t-1})$ y B_a y B_b son independientes.

El caso binario se trata como binomial con $n = 1$ y Harvey (1989) extiende este procedimiento al caso multinomial con una distribución Dirichlet como cojugada apriori.

El caso binomial negativo se puede estudiar casi exactamente como el caso binomial. La conjugada a priori para p_t en $BinNeg(r, p_t)$ es la distribución $Beta$ y si se reparametriza de tal forma que $(p_t|X^t) \sim Beta(a_t, b_t + 1)$, entonces las ecuaciones (3.28) y (3.29) toman la forma

$$a_t = \omega a_{t-1} + X_t \quad \text{y} \quad b_t = \omega b_{t-1} + r$$

$\hat{\mu}_t = \mathbb{E}(rp_t|X^t)$ y las ecuaciones (3.30) y (3.31) se siguen cumpliendo. La distribución predictiva en este caso es $Beta - Pascal$, i.e. $X_t|X^{t-1}$ tiene función de densidad de probabilidad

$$p(x) = \binom{r+x-1}{x} \frac{B(a+x, b+1+r)}{B(a, b+1)}$$

donde $a = a_{t|t-1}$ y $b = b_{t|t-1}$. La forma de la apriori aquí es un poco diferente a la que utilizó Harvey (1989) pero esta garantiza la existencia de la media de la apriori y hace que el análisis sea análogo al de los otros casos (Poisson y Binomial). De nuevo, es posible definir el mecanismo estocástico que especifica la transición de p_{t-1} a p_t al tiempo $t-1$. Sin embargo, se considerará la transición de los momios $r_t := \frac{p_t}{1-p_t}$, de r_{t-1} a r_t al tiempo $t-1$. Este está dado por

$$r_t = \frac{B_a r_{t-1}}{B_{b1}}$$

donde B_a tiene distribución $Beta(\omega a_{t-1}, (1-\omega)a_{t-1})$, B_{b1} tiene distribución $Beta(\omega b + 1, (1-\omega)b)$, B_a y B_{b1} son independientes y $a = a_{t-1}$ y $b = b_{t-1}$. Nótese que r_t tiene distribución Beta de segunda clase con parámetros $(a, b+1)$, i.e. su densidad está dada por

$$f(x) = \frac{x^{a-1}}{(1+x)^{a+b+1} B(a, b+1)} I_{(0,\infty)}(x)$$

Capítulo 4

Modelos auto-regresivos para series enteras

4.1. Introducción

La idea general de este capítulo es estudiar las series de tiempo de (pequeños) conteos, mediante algunos modelos auto-regresivos (ó simplemente regresivos). Generalmente tales series consisten de enteros no-negativos (pequeños), lo cual hace naturalmente que la mayoría de los modelos para datos continuos sea inapropiada.

Ya se han desarrollado algunas metodologías estadísticas explícitas para datos con este tipo de características discretas.

Cox (1981) propuso dividir estos modelos en dos grandes categorías:

(i) *Observation-driven*

(ii) *Parameter-driven*

Los del tipo (ii) estudian la relación del proceso latente con las observaciones. Los del tipo (i) estudian la relación entre observaciones presentes y pasadas.

Estas ideas se han trabajado en:

- Mac Donald & Zucchini (1997): *Hidden Markov and other models for discrete-valued time series.*
- Kedem & Fokianos (2002): *Regression models for time series analysis.*
- McKenzie (2003): *Discrete variate time series.*

En el capítulo anterior ya se analizaron algunas de las técnicas más importantes para el análisis de series de tiempo enteras. Ahora, desde esta perspectiva de Cox (1981), se trabajará en la clase *observation-driven* conocida como procesos autorregresivos con valores

enteros (INAR), procesos de medias móviles con valores enteros (INMA) que se introdujeron por McKenzie (1985) y Al-Osh & Alzaid (1988) y se analizarán algunas generalizaciones-de técnicas basadas en adelgzamiento. Básicamente se estudirán con más detalle algunos modelos que se basan en adelgazamiento binomial.

4.2. Modelos de series de tiempo enteras basados en adelgazamiento binomial

Una clase de modelos bastante amplia es aquella que se basa en la idea de "adelgazamiento". Estos modelos se han estudiado por un conjunto relativamente pequeño de personas: McKenzie (1985a,b; 1986; 1987; 1988a,b), Al-Osh & Alzaid (1987; 1988; 1991), Alzaid & Al-Osh (1988; 1990; 1993), Du & Li (1991) y Al-Osh & Aly (1992). Dion, Gauthier & Latour (1995) demostraron que estos modelos son simplemente funcionales de procesos de ramificación multitipo con inmigración (y de hecho en modelos sencillos se puede dar esta interpretación desde la definición del proceso). Esto permitió unificar y extender muchos de los resultados que se tienen para modelos basados en adelgazamiento. En particular, se usan resultados acerca de estimación en procesos de ramificación para obtener los correspondientes para modelos basados en adelgazamiento.

Definición 4.2.1. *(Operador de adelgazamiento binomial)*

Sea X una variable aleatoria tal que $X \geq 0$ *c.s.* y $sop(X) \subseteq \mathbb{N}^+$ y $\alpha \in [0,1]$

$$\alpha \circ X = \sum_{i=1}^{X} Y_i$$

donde $\{Y_k\}_{k\in\mathbb{Z}}$ es una sucesión de variables aleatorias independientes e idénticamente distribuidas con distribución $Bnlli(\alpha)$ e independientes de X.

Observación 4.2.1.

Nótese que $\alpha \circ X | X = x \sim Bin(x, \alpha)$ ▽

Definición 4.2.2. *(Función generadora de probabilidad y función generadora de probabilidad alterna)*

Sea X una variable aleatoria tal que $sop(X) \subseteq \mathbb{Z}$.

(i) Se define la función generadora de probabilidad de X, G_X, como

$$G_X(r) := \mathbb{E}(r^X)$$

(ii) Se define la función generadora de probabilidad alterna de X, G_X^A, como

$$G_X^A(r) = G_X(1-r) = \mathbb{E}((1-r)^X)$$

Lema 4.2.1.

Si X es una variable aleatoria tal que $X \sim Bnlli(p)$, entonces su función generadora de probabilidad está dada por

$$G_X(s) = 1 - p + ps$$

Demostración:

$$\begin{aligned} G_X(s) &= \mathbb{E}(s^X) = s^0 \cdot \mathbb{P}(X=0) + s^1 \cdot \mathbb{P}(X=1) \\ &= (1-p) + ps = 1 - p + ps \end{aligned}$$

□

Lema 4.2.2. *(Algunas propiedades del operador de adelgazamiento binomial)*

Sea X una variable aleatoria tal que $X \geq 0$ *c.s.* y $sop(X) \subseteq \mathbb{N}^+$, $\alpha \in [0,1]$ y $\alpha \circ X = \sum_{i=1}^{X} Y_i$ entonces

(i) $G_{\alpha\circ X}(s) = G_X(1-\alpha+\alpha s)$

(ii) $\mathbb{E}(\alpha \circ X) = \alpha\mathbb{E}(X)$

(iii) $Var(\alpha \circ X) = \alpha^2 Var(X) + \alpha(1-\alpha)\mathbb{E}(X)$

Demostración:

(i)

$$\begin{aligned} G_{\alpha\circ X}(s) &= \mathbb{E}(s^{Y_1+\ldots+Y_X}) = \mathbb{E}\left(\prod_{j=1}^{X} s^{Y_j}\right) = \mathbb{E}\left[\mathbb{E}\left(\prod_{j=1}^{X} s^{Y_j}|X\right)\right] \\ &= \mathbb{E}\left[\prod_{j=1}^{X}\mathbb{E}\left(s^{Y_j}|X\right)\right] = \mathbb{E}\left[\left(\mathbb{E}(s^{Y_1})\right)^X\right] \text{ por la independencia de las } Y_i\text{'s y } X \\ &= G_X(\mathbb{E}(s^{Y_1})) = G_X(G_{Y_1}(s)) = G_X(1-\alpha+\alpha s) \text{ pues } Y_1 \sim Bnlli(\alpha) \end{aligned}$$

$$\therefore G_{\alpha\circ X}(s) = G_X(1-\alpha+\alpha s)$$

(ii)

$$\begin{aligned} \mathbb{E}(\alpha\circ X) &= \mathbb{E}\left[\sum_{j=1}^{X} Y_j\right] = \mathbb{E}\left[\mathbb{E}\left(\sum_{j=1}^{X} Y_j|X\right)\right] \\ &= \mathbb{E}[X\mathbb{E}(Y_1)] = \mathbb{E}(Y_1)\mathbb{E}(X) = \alpha\mathbb{E}(X) \end{aligned}$$

$$\therefore \mathbb{E}(\alpha\circ X) = \alpha\mathbb{E}(X)$$

(iii)

$$\begin{aligned}(\alpha \circ X)^2 &= \left(\sum_{j=1}^{X} Y_j\right)^2 = \left(\sum_{j=1}^{X} Y_j\right)\left(\sum_{i=1}^{X} Y_i\right) \\ &= Y_1^2 + Y_1Y_2 + Y_1Y_3 + \ldots + Y_1Y_X + Y_2Y_1 + Y_2^2 + \ldots + Y_2Y_X + \ldots \\ &+ Y_XY_1 + Y_XY_2 + \ldots + Y_X^2 \\ &= : A_1 + A_2 + \ldots + A_X\end{aligned}$$

donde $A_j := \sum_{i=1}^{X} Y_iY_j$ para $j \in \{1, \ldots, X\}$.

Nótese que para todo $j \in \{1, \ldots, X\}$ $\mathbb{E}(A_j|X) = \mathbb{E}(A_1|X)$ y además

$$\begin{aligned}\mathbb{E}(A_1|X) &= \mathbb{E}\left(\sum_{i=1}^{X} Y_iY_j|X\right) = \mathbb{E}(Y_1^2) + \mathbb{E}\left(\sum_{i=2}^{X} Y_iY_j|X\right) \\ &= \mathbb{E}(Y_1^2) + \sum_{i=2}^{X}\mathbb{E}(Y_1Y_i) = \mathbb{E}(Y_1^2) + \sum_{i=2}^{X}\mathbb{E}(Y_1)\mathbb{E}(Y_i) \\ &= [\mathbb{E}^2(Y_1) + Var(Y_1)] + \sum_{i=2}^{X}\alpha\cdot\alpha = \alpha^2 + \alpha(1-\alpha) + (X-1)\alpha^2 \\ &= \alpha + \alpha^2(X-1)\end{aligned}$$

$$\therefore \mathbb{E}(A_1|X) = \alpha + \alpha^2(X-1)$$

Por tanto,

$$\begin{aligned}\mathbb{E}[(\alpha \circ X)^2] &= \mathbb{E}[\mathbb{E}((\alpha \circ X)^2|X)] = \mathbb{E}[\mathbb{E}(A_1 + A_2 + \ldots A_X|X)] \\ &= \mathbb{E}\left[\mathbb{E}\left(\sum_{j=1}^{X} A_j|X\right)\right] = \mathbb{E}\left[\sum_{j=1}^{X}\mathbb{E}\left(A_j|X\right)\right] \\ &= \mathbb{E}\left[\sum_{j=1}^{X}\mathbb{E}\left(A_1|X\right)\right] = \mathbb{E}\left[X(\alpha + \alpha^2(X-1))\right] \\ &= \mathbb{E}\left[\alpha X + \alpha^2X^2 - \alpha^2X\right] = \mathbb{E}\left[\alpha^2X^2 + \alpha(1-\alpha)X\right] \\ &= \alpha^2\mathbb{E}(X^2) + \alpha(1-\alpha)\mathbb{E}(X)\end{aligned}$$

$$\therefore \mathbb{E}[(\alpha \circ X)^2] = \alpha^2 \mathbb{E}(X^2) + \alpha(1-\alpha)\mathbb{E}(X)$$

Finalmente,

$$\begin{aligned} Var(\alpha \circ X) &= \mathbb{E}[(\alpha \circ X)^2] - \mathbb{E}^2[\alpha \circ X] \\ &= \alpha^2 \mathbb{E}(X^2) + \alpha(1-\alpha)\mathbb{E}(X) - (\alpha \mathbb{E}(X))^2 \\ &= \alpha^2 \mathbb{E}(X^2) + \alpha(1-\alpha)\mathbb{E}(X) - \alpha^2 \mathbb{E}^2(X) \\ &= \alpha^2(\mathbb{E}(X^2) - \mathbb{E}^2(X)) + \alpha(1-\alpha)\mathbb{E}(X) = \alpha^2 Var(X) + \alpha(1-\alpha)\mathbb{E}(X) \end{aligned}$$

$$\therefore Var(\alpha \circ X) = \alpha^2 Var(X) + \alpha(1-\alpha)\mathbb{E}(X)$$

□

4.2.1. Modelos con marginal geométrica

Proposición 4.2.1.

Considérese el proceso $\{X_t\}$ definido mediante

$$X_t = \alpha \circ X_{t-1} + Z_t \tag{4.1}$$

donde Z_t (la inovación) es independiente de X_{t-1}, $\alpha \in [0,1]$. Entonces

$$\therefore G^A_{Z_t}(r) = \frac{G^A_{X_t}(r)}{G^A_{X_t}(\alpha r)}$$

Demostración:

De la ecuación (4.3) y la independencia entre $\alpha \circ X_{t-1}$ y Z_t se tiene que

$$G^A_{X_t}(z) = G^A_{\alpha \circ X_{t-1}}(z) G^A_{Z_t}(z) \tag{4.2}$$

$$\begin{aligned} G^A_{Z_t}(r) &= \mathbb{E}[(1-r)^{Z_t}] = G_{Z_t}(1-r) = \frac{G_{X_t}(1-r)}{G_{X_t}(1-\alpha+\alpha(1-r))} \\ &= \frac{G_{X_t}(1-r)}{G_{X_t}(1-\alpha r)} = \frac{G^A_{X_t}(r)}{G^A_{X_t}(\alpha r)} \end{aligned}$$

$$\therefore G^A_{Z_t}(r) = \frac{G^A_{X_t}(r)}{G^A_{X_t}(\alpha r)}$$

□

Por tanto, si una sucesión de variables aleatorias independientes e idénticamente distribuidas $\{Z_t\}$ con función generadora de probabilidad alterna G^A_Z se utiliza en $X_t = \alpha \circ X_{t-1} + Z_t$ y X_0 tiene una función generadora de probabilidad alterna $G^A_{X_0}$, entonces se tendrá una sucesión de variables aleatorias dependientes cuya distibución marginal estacionaria también tiene la función generadora de probabilidad alterna G^A_X.

La forma de la ecuación $X_t = \alpha \circ X_{t-1} + Z_t$ sugiere que hay una relación lineal en los parámetros y de hecho, muchas de sus propiedades son parecidas a las del modelo $AR(1)$ (que sí es lineal).

Supóngase que

$$X_t = \alpha \circ X_{t-1} + Z_t \tag{4.3}$$

donde Z_t (la inovación) es independiente de X_{t-1}, $\alpha \in [0,1]$ y X_{t-1} tiene distribución $Geo(1-\theta)$ i.e. para $k \in \mathbb{N}$

$$\mathbb{P}(X_{t-1} = k) = (1-\theta)\theta^k$$

La función generadora de probabilidad alterna de X_{t-1}, $G^A_{X_{t-1}}$, está dada por

$$\begin{aligned} G^A_{X_{t-1}}(z) &= \mathbb{E}((1-z)^{X_{t-1}}) = \sum_{k=0}^{\infty}(1-z)^k \mathbb{P}(X_{t-1}=k) \\ &= \sum_{k=0}^{\infty}(1-z)^k(1-\theta)\theta^k = (1-\theta)\sum_{k=0}^{\infty}[\theta(1-z)]^k \\ &= (1-\theta)\frac{1}{1-\theta(1-z)} = \frac{1-\theta}{\theta}\frac{1}{\frac{1}{\theta}-1+z} \\ &= \lambda\frac{1}{\frac{1-\theta}{\theta}+z} = \frac{\lambda}{\lambda+z} \end{aligned}$$

donde $\lambda := \frac{1-\theta}{\theta}$.

Y la función generadora de probabilidad alterna de $\alpha \circ X_{t-1}$ está dada por

$$\begin{aligned} G^A_{\alpha\circ X_{t-1}}(z) &= G_{\alpha\circ X_{t-1}}(1-z) = G_{X_{t-1}}(1-\alpha+\alpha(1-z)) \\ &= G_{X_{t-1}}(1-\alpha z) = G^A_{X_{t-1}}(\alpha z) = \frac{\lambda}{\lambda+\alpha z} \end{aligned}$$

De la ecuación (4.3) y la independencia entre $\alpha \circ X_{t-1}$ y Z_t se tiene que

$$G^A_{X_t}(z) = G^A_{\alpha \circ X_{t-1}}(z) G^A_{Z_t}(z) \tag{4.4}$$

es decir,

$$G^A_{X_t}(z) = \frac{\lambda}{\lambda + \alpha z} G^A_{Z_t}(z)$$

Entonces, la condición para que X_t tenga la misma distibución que X_{t-1} es que $G^A_{X_t}(z) \equiv G^A_{X_{t-1}}(z) = \frac{\lambda}{\lambda + z}$.

Es decir, la condición para que X_t tenga la misma distibución que X_{t-1} es que Z_t tenga una función generadora de probabilidad alterna

$$G^A_{Z_t}(z) \equiv \frac{\lambda/(\lambda + z)}{\lambda/(\lambda + \alpha z)} = \alpha + (1 - \alpha)\frac{\lambda}{\lambda + z}$$

i.e. Z_t es cero (con probabilidad α) ó tiene distribución $Geo(1 - \theta)$. Equivalentemente, Z_t se puede describir como el producto de variables aleatorias geométricas (independientes) y Bernoulli. Si Z_t satisface este requisito y X_0 tiene la distribución $Geo(1 - \theta)$, entonces X_t también tendrá distribución geométrica para cualquier $t \in \mathbb{N}$.

La estructura de correlación de la sucesión estacionaria $\{X_t\}$ es simple: la autocorrelación de orden k, ρ_k es simplemente α^k (como en el caso del $AR(1)$ usual $X_t = \alpha X_{t-1} + \epsilon_t$). Este hecho se demostrará más adelante en el marco de un análisis de las propiedades teóricas de este tipo de procesos.

Cualquier proceso aleatorio $\{X_t\}$ que satisfaga la ecuación (4.3) es una cadena de Markov. El operador de adelgazamiento "$\circ$" es un análogo discreto muy natural a la multiplicación escalar (veáse Apéndice B). Además, el modelo $X_t = \alpha \circ X_{t-1} + Z_t$ tiene una interpretación útil pues se puede considerar a X_t como el número de sobrevivientes de aquellos presentes al tiempo $t - 1$ (cada uno con probabilidad de supervivencia α) más Z_t el número de individuos que se incorporan entre $t - 1$ y t.

El modelo anterior es el proceso auto-regresivo geométrico de orden uno de McKenzie (1985b) estudiado también por Alzaid & Al-Osh (1988). Este se ha modificado y generalizado en varias direcciones por los mismos autores. Estos modelos modelos se diseñaron para tener una distribución marginal dada (por ejmplo Poisson, Binomial o Binomial Negativa) y una estructura de dependencia (auto-regresiva, de medias móviles ó $ARMA$). Sin embargo, es necesario hacer enfásis en que los términos "auto-regresivo" y "medias móviles" con frecuencia se usan vagamente en este contexto, simplemente para indicar la similaridad en su forma con la de ARMA estándar. Por ejemplo, Alzaid & Al-Osh (1990) definen su proceso con valores enteros de orden p donde, de hecho, su función de autocorrelación es

parecida a a del $ARMA(p, p-1)$ estándar.

McKenzie (1986) definió los modelos de medias móviles y auto-regresivos con marginal geométrica (justo como el que se describió en esta sección) como análogos de los correspondientes modelos $ARMA$ exponenciales de Lawrence & Lewis (1980) obteniéndolos a partir de reemplazar la multiplicación escalar con el adelgazamiento y la distribución exponencial por la distribución geométrica. El modelo más general de este tipo que estudió McKenzie es un $ARMA(p, q)$ geométrico. Sin embargo, sólo se analizará con detalle el proceso $ARMA(1, 1)$ geométrico. Sean $\{M_t\}, \{U_t\}$ y $\{V_t\}$ sucesiones independientes de variables aleatorias independientes e idénticamente distribuidas tales que M_t tiene distribución geométrica con media λ^{-1}, $U_t \sim Bnlli(1-\alpha)$ y $V_t \sim Bnlli(1-\beta)$. También supóngase que W_0 tiene distribución geométrica con media λ^{-1} y que

$$X_t = \beta \circ M_t + V_t W_{t-1}$$

y

$$W_t = \alpha \circ W_{t-1} + U_t M_t$$

donde se supone que en todos las operaciones de adelgazamiento se realizan de forma independiente de todas las componentes del proceso que se está considerando, incluyendo otras operaciones de adelgazamiento. W_t y X_t tienen distribución geométrica con media λ^{-1} y $\{X_t\}$ es un proceso $ARMA(1, 1)$ geométrico. La autocorrelación de orden k del proceso $\{X_t\}$ está dada por

$$\rho_k = (1-\beta)[(1-\alpha)\beta + \alpha(1-\beta)]\alpha^{k-1}$$

Nótese que $\{W_t\}$ es un proceso $AR(1)$ geométrico y en los casos $\beta = 0$ ó $\alpha = 0$, $\{X_t\}$ es un $AR(1)$ geométrico ó un $MA(1)$ geométrico, respectivamente.

Se han descrito algunos otros modelos con marginal geométrica, por ejemplo McKenzie (1986) trabajó con el análogo geométrico del proceso $NEAR(1)$ de Lawrence & Lewis (1981), éste se define mediante

$$X_t = (\beta U_t) \circ X_{t-1} + (1 - V_t + (1-\alpha)\beta V_t) \circ M_t$$

donde $\{M_t\}, \{U_t\}$ y $\{V_t\}$ son sucesiones independientes de variables aleatorias independientes e idénticamente distribuidas tales que M_t tiene distribución geométrica, $U_t \sim Bnlli(\alpha)$ y $V_t \sim Bnlli(\frac{\alpha\beta}{1-(1-\alpha)\beta})$. El caso en el que $\alpha = 1$ se reduce al $AR(1)$ geométrico que ya se describió. Mckenzie (1985b) también dió un análogo geométrico del modelo $NEAR(2)$ de Lawrence & Lewis (1985).

4.2.2. Modelos con marginal binomial negativa

La distribución geométrica es un caso particular de la distribución binomial negativa, y McKenzie (1986) también consideró la construcción de un proceso $AR(1)$ binomial negativo. Sólo con fines didácticos, se considerará la parametrización de una variable aleatoria X con distribución binomial negativa con parámetro de forma β y parámetro de escala λ satisface

$$\mathbb{P}(X=k)=\binom{\beta+k-1}{k}\left(\frac{\lambda}{1+\lambda}\right)^{\beta}\left(\frac{1}{1+\lambda}\right)^{k},\quad k\in\mathbb{N}=\{0,1,\ldots\}$$

los parámetros β,λ son positivos. Nótese que en particular el parámetro de forma β no necesariamente es entero. La función generadora de probabilidad alterna de dicha distribución binomial negativa está dada por $\left(\frac{\lambda}{\lambda+z}\right)^{\beta}$, que es justo la transformada de Laplace de la densidad gamma con parámetros de forma y escala, β y λ i.e. para $x>0$

$$\frac{\lambda^{\beta}x^{\beta-1}e^{-\lambda x}}{\Gamma(\beta)}$$

Esto sugiere que para definir un proceso $AR(1)$ binomial negativo, lo único que se tiene que hacer es reemplazar la multiplicación escalar por el operador de adelgazamiento y la distribución gamma por la distribución binomial negativa en el proceso $AR(1)$ Gamma (Gaver & Lewis (1980)). El modelo resultante es

$$X_t=\alpha\circ X_{t-1}+Z_t$$

con $\alpha\in(0,1)$, X_t con distribución binomial negativa con parámetro de forma β y parámetro de escala λ para cualquier $t\in\mathbb{N}$, y la innovación,Z_t, tiene función generadora de probabilidad alterna

$$\left(\frac{\lambda+\alpha z}{\lambda+z}\right)^{\beta}=\left(\alpha+(1-\alpha)\frac{\lambda}{\lambda+z}\right)^{\beta}$$

La construcción de una variable aleatoria que tiene una función generadora de probabilidad alterna para un β general (no necesariamente entero) es bastante ingenua. McKenzie (1987) describió tal construcción y está basada en un proceso *shot-noise* y es esencialmente el mismo que el desarrollado por Lawrence (1992) para resolver el problema correspondiente para el proceso Gamma.

La complejidad del proceso de innovación de McKenzie consiste en definir un tipo diferente de proceso binomial negativo (McKenzie (1986)). Esto es análogo al proceso $AR(1)$ gamma-beta descrito por Lewis (1985) que es una auto-regresión con coeficiente aleatorio con marginal Gamma. El correspondiente $AR(1)$ binomial negativo se define como

$$X_t=A_t\circ X_{t-1}+M_t$$

Donde A_t tiene distribución Beta con parámetros α y $\beta - \alpha$, M_t tiene distribución binomial negativa con parámetros $\beta - \alpha$ y λ de forma y escala, respectivamente ($0 < \alpha < \beta$, $\lambda > 0$). Y además A_t, X_{t-1} y M_t son mutuamente independientes. Si X_0 tiene distribución binomial negativa con parámetros β y λ (de forma y escala, respectivamente), entonces X_t tiene la misma distribución que X_0 y para cualquier $k \in \mathbb{N}$

$$\rho_k = (\alpha/\beta)^k$$

Nótese que al tiempo t hay tres elementos aleatorios: las variables aleatorias A_t y M_t (no observadas) y el de la operación de adelgazamiento.

El caso especial en el que $\beta = 1$ de este proceso binomial negativo es el proceso $AR(1)$ geométrico diferente al que se analizó en la sección anterior. La distribución de este nuevo proceso $AR(1)$ geométrico no se distribuye geométricamente ni es el producto de geométricas y Bernoulli. La distribución de la innovación es binomial negativa con parámetros $1 - \alpha$ y λ, donde $\alpha \in (0, 1)$ y $\lambda > 0$.

4.2.3. Modelos con marginal Poisson

En este trabajo, los modelos con marginal Poisson se considerarán a detalle: se harán cálculos de correlaciones explícitos, se obtendrán y analizarán algunos de los estimadores clásicos para este modelo e incluso se realizarán algunas simulaciones particulares. Sin embargo, por completez, se dará un breve resumen de sus principales propiedades y alternativas en esta sección y más adelante (durante todo el capítulo) se retomará como ejemplo/aplicación de algunos resultados obtenidos en modelos auto-regresivos.

Un proceso Poisson $AR(1)$ se discute en McKenzie(1985b; 1988b), Al-Osh& Alzaid (1987) y Alzaid & Al-Osh (1988), donde se trata a éste como un caso especial de su modero $INAR(1)$ ("auto-regresivo de orden uno con valores en los enteros"). Aunque también McKenzie (1988b) hizo generalizaciones con un proceso Poisson $ARMA$ y un proceso Poisson-múltiple $AR(1)$. Al-Osh & Alzaid (1988) analizaron varias propiedades del caso Poisson de sus modelos $INMA(1)$ e $INMA(q)$ (modelos de medias móviles de valores enteros que se estudiarán un poco más a detalle en secciones posteriores). Finalmente (antes de una teoría más elaborada) Alzaid & Al-Osh (1990) estudiaron propiedades del $INAR(p)$ Poisson y lo relacionaron con el $AR(1)$ Poisson múltiple de McKenzie.

El proceso $AR(1)$ Poisson de McKenzie es una solución estacionaria $\{X_t\}$ de la ecuación

$$X_t = \alpha \circ X_{t-1} + Z_t$$

con la hipótesis de que Z_t y X_{t-1} son independientes y $\{Z_t\}$ es una sucesión de variables aletorias independientes e idénticamente distribuidas.

Suponga que

$$\mathbb{P}(X_t = x) = e^{-\lambda}\frac{\lambda^x}{x!}I_{0,1,\ldots}(x)$$

Nótese que

$$\begin{aligned} G_{X_t}(r) &= \mathbb{E}[r^{X_t}] = \sum_{x=0}^{\infty} r^x e^{-\lambda}\frac{\lambda^x}{x!} = e^{-\lambda}\sum_{x=0}^{\infty}\frac{(r\lambda)^x}{x!} \\ &= e^{-\lambda}e^{r\lambda} = e^{-\lambda(1-r)} \end{aligned}$$

$$\therefore G_{X_t}(r) = e^{-\lambda(1-r)}$$

De aquí que

$$\begin{aligned} G_{Z_t}(r) &= \frac{G_{X_t}(r)}{G_{X_t}(1-\alpha+\alpha r)} = \frac{e^{-\lambda(1-r)}}{e^{-\lambda(1-(1-\alpha+\alpha r))}} \\ &= \frac{e^{-\lambda(1-r)}}{e^{-\lambda(\alpha-\alpha r)}} = \frac{e^{-\lambda(1-r)}}{e^{-\lambda\alpha(1-r)}} \\ &= exp[(1-r)(-\lambda+\lambda\alpha)] = e^{-\lambda(1-\alpha)(1-r)} \end{aligned}$$

$$\therefore G_{Z_t}(r) = e^{-\lambda(1-\alpha)(1-r)}$$

$$\therefore Z_t \sim Poisson(-\lambda(1-\alpha))$$

Una diferencia entre el $AR(1)$ geométrico y $AR(1)$ Poisson es que en el caso Poisson las distribuciones marginal y de la inovación son Poisson. De forma más precisa, $X_t \sim Poisson(\lambda)$ si y sólo si $Z_t \sim Poisson((1-\alpha)\lambda)$. El papel de la distribución Poisson en la ecuación anterior es análogo al del la distribución normal en el $AR(1)$ Gaussiano $(X_t = \alpha X_{t-1} + \varepsilon_t)$.

Otras propiedades interesantes del proceso $AR(1)$ Poisson son que su función de autocorrelación es α^k y que es una cadena de Markov reversible con probabilidades de transición

$$\mathbb{P}(X_t = j|X_{t-1} = i) = \sum_{k=0}^{j}\binom{i}{k}\alpha^k(1-\alpha)^{i-k}e^{-\lambda(1-\alpha)}\frac{(\lambda(1-\alpha))^{j-k}}{(j-k)!}$$

Aquí, λ es la media de X_t y se usa la convención de que $\binom{i}{k} = 0$ para $k > i$. La regresión de X_t sobre X_{t-1} es lineal (y viceversa, por la reversibilidad) pero la varianza de X_t dado X_{t-1} no es constante con respecto a X_{t-1}. Esta última propiedad es una en la que el AR(1) Poisson es diferente de su contraparte el $AR(1)$ Gaussiano.

Al proceso de medias móviles de orden 1 Poisson, lo definieron McKenzie (1988b) y Al-Osh & Alzaid (1988) y es un proceso $\{X_t\}$ que satisface

$$X_t = Z_t + \beta \circ Z_{t-1}$$

para $0 \leq \beta \leq 1$ y $\{Z_t\}$ es una sucesión de variables aleatorias independientes e idénticamente distribuidas como Poisson. Si la media de Z_t es $\frac{\lambda}{1+\beta}$, entonces la de X_t es λ. La autocorrelación, ρ_k, es cero para $k \geq 2$, y $\rho_1 = \frac{\beta}{1+\beta}$. Algo que es notable es que la distribución conjunta de X_t y X_{t-1} es de la misma forma que en el $AR(1)$ Poisson que se describió anteriormente: para ambos, la función generadora de probabilidad conjunta alterna es

$$\mathbb{E}\left[(1-u)^{X_{t-1}}(1-v)^{X_t}\right] = \exp\{-\lambda(u+v-\rho_1 uv)\}$$

El proceso de promedios móviles Poisson de orden q es una extensión natural del de orden 1

$$X_t = Z_t + \sum_{i=1}^{q} \beta_i \circ Z_{t-i}$$

donde $0 \leq \beta_i \leq 1$ para cualquier i, $\beta := \sum_{i=0}^{q} \circ X_{t-i}$ y $\{Z_t\}$ es una sucesión de variables aleatorias independientes e idénticamente distribuidas con media λ/β. Por convención $\beta_0 = 1$. La distribución de X_t es $Poisson(\lambda)$ y la función de autocorrelación es

$$\rho_k = \begin{cases} \sum_{i=0}^{q-k} \frac{\beta_i \beta_{i+k}}{\beta} & \text{si } k = 1, ..., q; \\ 0 & \text{si } k = q+1, q+2, \ldots \end{cases}$$

El proceso $ARMA(1,q)$, es un proceso $\{X_t\}$ en el cual

$$X_t = Z_{t-q} + \sum_{k=1}^{q} \beta_k \circ W_{t+1-k}$$

y

$$Z_t = \alpha \circ Z_{t-1} + W_t$$

La sucesión $\{W_t\}$ se toma como una sucesión de variables aleatorias independientes e idénticamente distribuidas $Poisson((1-\alpha)\lambda)$ y Z_0 es una variable aleatoria independiente $Poisson(\lambda)$. Entonces se tiene que $\{Z_t\}$ es un $AR(1)$ Poisson con media λ y $\{X_t\}$ es una sucesión estacionaria de variables aleatorias $Poisson((1+(1-\alpha)b)\lambda)$, donde b se define como $\sum_{k=1}^{q} \beta_k$. Si $\alpha = 0$, entonces $\{X_t\}$ es un $MA(q)$ Poisson; si $\beta_k = 0$ para todo k, entonces $\{X_t\}$ es un $AR(1)$ Poisson. McKenzie (1988b) demostró que la función de autocorrelación de proceso $ARMA(1,q)$ Poisson general satisface que $\rho_k = \alpha^{k-q}\rho_q$ para $k \geq q$ y

$$\rho_q = \frac{\alpha^q + (1-\alpha)\sum_{i=1}^{q} \beta_i \alpha^{i-1}}{1 + (1-\alpha)b}$$

McKenzie (1988b) demostró algunos resultados con respecto a la distribución de n observaciones consecutivas de un proceso $ARMA(1,q)$ Poisson, $\{X_t\}$, y utilizó dichos resultados para obtener conclusiones acerca de la reversibilidad del proceso. La distribución conjunta de estas n observaciones consecutivas es la distribución Poisson multivariada de Teicher (1954). Un resultado interesante es que $Cov(X_t^2, X_{t-k})$ y $Cov(X_t, X_{t-k}^2)$ son iguales aunque el proceso puede no ser reversible. Los procesos $AR(1), MA(1), MA(2)$ Poisson en general son reversibles pero para $q \geq 3$ el proceso $MA(q)$ puede no ser reversible.

Para definir un proceso $AR(1)$ Poisson múltiple, McKenzie (1988b) primero definió $\alpha \circ Y$ para una variable aleatoria Y cuyo soporte es $\mathbb{N} = \{0, 1, \ldots\}$ y α un vector de probabilidades cuya suma no es mayor que uno. Esto se hace especificando que, condicionado sobre $Y = y$, $\alpha \circ Y$ tiene distribución multinomial con parámetros y (las categorías) y α. Por tanto, la operación $\circ$ ahora se conocerá como *adelgazamiento multinomial*. Para $A \in M_{p\times p}(\mathbb{R})$, $A = (\alpha_1\ \alpha_2\ \cdots\ \alpha_p)$ y un vector Y entonces se define

$$A \circ Y = \sum_{i=1}^{p} \alpha_i \circ Y_i$$

en donde cada operación de adelgazamiento multinomial se lleva a cabo de manera independiente. El proceso $AR(1)$ se define como la solución estacionaria, $\{X_t\}$, de la ecuación

$$X_t = A \circ X_{t-1} + E_t$$

con $\{E_t\}$ una sucesión de vectores aleatorios (de dimensión p) independientes e idénticamente distribuidos y cada vector X_t consiste de variables aleatorias Poisson independientes. McKenzie estudió resultados generales con respecto a la distribución de la innovación y la estructura de correlación de tales procesos. Uno de los resultado más destacables es el que establece que la j-ésima componente de X_t tiene función del autocorrelación $\{\rho_j(k)\}$ que satisface

$$\rho_j(k) = \begin{cases} (A^k)_{ij} & \text{si } k = 1, ..., p; \\ \sum_{i=1}^{p} \phi_i \rho_j(k-i) & \text{si } k = p+1, p+2, \ldots \end{cases}$$

para ciertas constantes $\phi_1, \ldots, \phi_p$. Nótese que por este resultado, se puede afirmar que cada componente tiene la misma estructura que el proceso $ARMA(p, p-1)$ estándar y puede disminuir el orden para matrices A con alguna estructura particular más manejable (como diagonal o triangular).

McKenzie también sugiere algunas extensiones de los modelos Poisson en las siguientes direcciones: distribuciones marginales Poisson compuesta, correlación negativa (que se excluye en todos los modelos basados en adelgazamiento definidos hasta el momento) y la

existencia de tendencias y ciclos (como en el análisis de series de tiempo estándar).

Alzaid & Al-Osh (1990) describieron el caso Poisson (que se estudiará detalladamente en las siguientes secciones) de su proceso general $INAR(p)$ y demostraron que para $p = 2$ hay un proceso Poisson $AR(1)$ multivariado incrustado del tipo analizado por McKenzie, i.e. con independencia de las componentes. Este es el proceso $\{\mathbf{X}_t\}$ defindo por el vector

$$\mathbf{X}_t = \begin{pmatrix} X_t \\ \alpha_2 \circ X_{t-1} \end{pmatrix}$$

donde α_2 es el "coeficiente" de X_{t-2} de la ecuación

$$X_t = \alpha_1 \circ X_{t-1} + \alpha_2 \circ X_{t-2} + Z_t$$

Ya que el proceso incrustado tiene una estructura más simple, se puede usar para obtener propiedades de $\{X_t\}$ mediante la distribución conjunta de X_{t-1} y X_t y el hecho de que $\{X_t\}$ es reversible.

4.2.4. Modelos con marginal binomial

McKenzie (1985b) propuso un proceso $AR(1)$ binomial, $\{X_t\}$, que satisface

$$X_t = \alpha \circ X_{t-1} + \beta \circ (m - X_{t-1}),$$

con $X_t \sim Bin(m, \theta)$ para cualquier $t \in \mathbb{Z}$, $\alpha \in (0, 1)$ y a menos que éste sea mayor que 1, $\beta = \frac{\alpha\theta}{1-\theta}$ (se puede hacer una modificación en el caso en el que $\frac{\alpha\theta}{1-\theta} > 1$). La función de autocorrelación está dada por $\rho_k = (\alpha - \beta)^k$. La construcción usual (la dada por la ecuación $X_t = \alpha \circ X_{t-1} + Z_t$) no es posible ya que la distribución binomial no es auto-descomponible discretoa (a diferencia de las distribuciones Poisson, geométrica y binomial negativa) en el sentido de Steutel van Harn (1979) (Véase apéndice B).

Al-Osh & Alzaid (1991) construyeron la clase de los modelos $ARMA$ binomiales de una forma un poco diferente. Esta construcción se basa en lo que se conoce como adelgazamiento hipergeométrico. Para definir este modelo se necesitan los siguientes preliminares:

Sea $S \sim Bin(m, p)$ y sea $T(S)$ una variable aleatoria tal que, condicionado a S, tiene distribución hipergeométrica con parámetros m, s, M, i.e.

$$\mathbb{P}(T(S) = k | S = s) = \frac{\binom{s}{k}\binom{m-s}{M-k}}{\binom{m}{M}}$$

para $k \in \{0, 1, \ldots, \min m, M\}$. De aquí que si $S \sim Bin(m, p)$ y $T(S) entonces \sim Bin(M, p)$ y $T(S)$ y $S - T(S)$ son independientes.

El proceso $AR(1)$ Binomial de Al-Osh & Alzaid, X, se define mediante la expresión

$$X_t = T(X_{t-1}) + Z_t$$

donde Z_t es independiente de X_{t-1} y de $T(X_{t-1})$ y tiene distribución $Bin(m-M,p)$. Por tanto, si $X_{t-1} \sim Bin(m,p)$, entonces X_t también tiene distribución $Bin(m,p)$. Si este proceso $AR(1)$ binomial es estacionario, entonces su función de autocorrelación está dada por

$$\rho_k = (M/m)^k, \quad k \in \mathbb{N}$$

y es una cadena de Markov reversible. La regresión de X_t sobre X_{t-1} es lineal, pero la correspondiente varianza condicional no es constante.

De forma análoga, los autores definen un proceso estacionario $MA(q)$, $\{X_t\}$, que satisface

$$X_t = Z_t + \sum_{i=1}^{q} T_i(Z_{t-i})$$

donde $\{Z_t\}$ es una sucesión de variables aleatorias independientes idénticamente distribuidas como $Bin(m-n,p)$, la distribución de $T_i(Z_{t-i})$ dado $Z_{t-i} = z$ es $Hipergeo(m-n, z, m_i)$ con $m = m_1 + \ldots + m_q$ y las diferentes operaciones de adelgazamiento hipergeométrico se realizan de forma independiente. La función de autocorrelación de dicho proceso está dada por

$$\rho_k = \begin{cases} \frac{1}{m(m-n)} \sum_{i=0}^{q-k} m_i m_{i+k} & \text{si } k = 1, ..., q; \\ 0 & \text{si } k > q \end{cases}$$

con $m_0 := 1$. Es decir, el proceso tiene la misma propiedad de truncamiento en la autocorrelación que el de las medias móviles estándar. Sin embargo, como en casi todos los modelos basados en adelgazamiento, las correlaciones se restringen a ser no negativas.

El correspondiente modelo estacionario $ARMA(1,q)$ se define mediante

$$X_t = Y_{t-q} + \sum_{i=1}^{q} T_i(Z_{t+1-i})$$

y

$$Y_t = T(Y_{t-1}) + Z_t$$

Esta estructura es análoga a la construcción del $ARMA(1,q)$ Poisson de McKenzie. Las funciones de autocorrelación también son muy parecidas: para $k \in \{1, 2 \ldots, q\}$ las expresiones para ρ_k tienen forma similar, y en ambos casos existe alguna $c \in \mathbb{R}$ tal que para $k \geq q$,

$\rho_k = c^{k-q}\rho_q$. La distribución conjunta de X_{t-1} y X_t se puede obtener de manera sencilla para el proceso $ARMA(1,1)$ binomial y tiene la misma simetría que el $AR(1)$ binomial. Por tanto, la regresión de X_t sobre X_{t-1} es lineal (y viceversa) también en este caso. La conclusión de Al-Osh & Alzaid (a partir de la simetría de la distribución conjunta de X_t y X_{t-1}) de que $\{X_t\}$ es reversible en el caso $ARMA(1,1)$ binomial aún no parece justificada.

Al-Osh & Alzaid también definieron un $AR(1)$ binomial multivariado, y lo hicieron de forma análoga al de la definición del $AR(1)$ Poisson multivariado de MacKenzie. Este modelo comparte con el modelo de McKenzie el hecho de que tiene la misma estructura de correlación que el $ARMA(p, p-1)$ estándar.

4.2.5. Modelos que no se basan en algunas suposición distribucional explícita

Algunos de los resultados que se mencionaron anteriormente en el contexto de una distribución marginal particular son válidos en general. Por ejemplo, la propiedad $\rho_k = \alpha^k$ de los modelos $AR(1)$ geométrico y Poisson es una propiedad de cualquier proceso $INAR(1)$ de Al-Osh & Alzaid (1987) (que se estudiarán con detalle en la siguiente sección).

El proceso $INAR(p)$ de Al-Osh & Alzaid se define mediante

$$X_t = \sum_{i=1}^{p} \alpha_1 \circ X_{t-i} + Z_t$$

donde, como antes, $\{Z_t\}$ es una sucesión de variables aleatorias independientes idénticamente distribuidas tal que $sop(Z_t) \subseteq \mathbb{N}$, $\sum_{i=1}^{p} \alpha_i < 1$ y la distribución condicional del vector aleatorio $(\alpha_1 \circ X_t, \alpha_2 \circ X_t, \ldots, \alpha_p \circ X_t)$ dada $X_t = x$ es multinomial con parámetros x y $\alpha = (\alpha_1 \ldots, \alpha_p)$, independiente de la historia del proceso, i.e. independiente de X_{t-k} y todas las operaciones de adelgazamiento para $k \in \mathbb{N}$. Esta estructura particular (que no es la misma que la del $INAR(p)$ de Du & Li) implica que la estructuras de correlación es similar a la de un proceso $ARMA(p, p-1)$ estándar, no a la de un $AR(p)$. El proceso $INAR(p)$ de Du & Li sí implica una estructura de correlación similar a la de un $AR(p)$ estándar. En ambos casos, la condición para la existencia de una distribución estacionaria o límite es que las raíces del polinomio

$$z^p - \alpha_1 z^{p-1} - \ldots - \alpha_{p-1} z - \alpha_p$$

estén dentro del círculo unitario.

4.2.6. Modelos no estacionarios

En la práctica, la mayoría de las series de tiempo, discretas ó no, no son estacionarias. Típicamente hay tendencias de largo plazo y tendencias periódicas que reflejen la ciclicidad;

y por supuesto, éstas se deben modelar.

Considérese el modelo $X_t = \alpha \circ X_{t-1} + Z_t$, si se aplica el operador esperanza de ambos lados se obtiene que

$$\mu_t = \alpha\mu_{t-1} + \omega_t \tag{4.5}$$

Donde $\mu_t := \mathbb{E}(X_t)$ y $\omega_t := \mathbb{E}(W_t)$. En el caso estacionario, se tiene que $\omega_t = (1-\alpha)\mu_{t-1}$ pero, en general, la ecuación (4.5) se cumplirá sin importar la forma de la sucesión de medias (siempre y cuando sean positivas). Por tanto, se le puede asociar a $\{X_t\}$ una media dependiente del tiempo μ_t, simplemente manipulando las medias del proceso de innovación $\{\omega_t\}$. Por ejemplo, supóngase que se quiere que $\{X_t\}$ tenga una tendencia lineal de la forma $a + bt$, entonces haciendo $\omega_t = \alpha b + (1-\alpha)(a+bt)$ en (4.5) se asegura que $\{\mu_t\}$ tenga la forma correcta. Esta idea se analiza en McKenzie (1985a) en donde se consideran diferentes formas para la ciclicidad y los cambios de nivel. Parece ser una propuesta bastante razonable para resolver el problema de modelar tendencias deterministas en modelos basados en adelgazamiento.

Por otro lado, la tendencia que se modela y la estructura de correlación no son independientes, y esto puede ser un serio problema en la práctica. En el ejemplo anterior, las medias de innovación $\{\omega_t\}$ debe ser positiva. Es fácil probar que esto implica que

$$\alpha < 1 + \frac{b}{a+bt}.$$

Sin embargo, b puede ser negativo siempre y cuando $a + bt$ sea positivo y por tanto, α tendría una cota positiva estrictamente menor que 1. De hecho, de forma mas general en (4.5) se requiere que

$$\alpha < \frac{\mu_t}{\mu_{t-1}}, \quad \text{para todo } t,$$

y si $\{\mu_t\}$ es decreciente puede ser complicado encontrar un modelo de este tipo.

Este problema puede ocurrir en algún tipo de tendencia determinista y en todos los modelos que se han discutido. Por ejemplo, en el modelo binario dado por

$$X_t = A_t X_{t-1} + B_t(1 - X_{t-1})$$

(con $\{A_t\}$ y $\{B_t\}$ son sucesiones independientes de variables aleatorias independientes idénticamente distribuidas como $Bnlli(\alpha)$ y $Bnlli(\beta)$, respectivamente) no hay proceso de innovación explícito. Sin embargo, se puede incorporar dependencia del tiempo en la probabilidad θ_t (la media del proceso) permitiendo que $\{A_t\}$ ó $\{B_t\}$ sean dependientes del tiempo. En este caso se obtendría

$$\theta_t = \alpha_t\theta_{t-1} + \beta_t(1 - \theta_{t-1})$$

donde, en el caso estacionario $\alpha_t = \alpha$ y $\beta_t = \beta = \frac{(1-\alpha)\theta}{1-\theta}$. Obsérvese que $X_t = A_t X_{t-1} + B_t(1 - Xt - 1)$ es una cadena de Markov de dos estados con $\mathbb{P}(X_t = 1|X_{t-1} = 1) = \alpha_t$ y $\mathbb{P}(X_t = 1|X_{t-1} = 0) = \beta_t$. Como en el caso estacionario, β es una función de α y la correlación (i.e. está relacionada con α de una forma muy simple), entonces parece natural inducir la tendencia determinista haciendo que β_t sea dependiente del tiempo y que α permanezca constante. Sin embargo, incluso bajo todas estas suposiciones la correlación y la tendencia son inter-dependientes. Para asegurar que β_t sea una probabilidad α debe satisfacer

$$1 - \frac{1-\theta_t}{\theta_{t-1}} < \alpha < \frac{\theta_t}{\theta_{t-1}}$$

donde, de nuevo, habrá problemas si $\{\theta_t\}$ es decreciente.

En algunos casos, al menos este problema se puede evitar. Por ejemplo, Azzalini (1994) propuso modelar datos binarios mediante una cadena de Markov. Esencialmente tiene la misma estructura $X_t = A_t X_{t-1} + B_t(1 - Xt - 1)$ y $\{\theta\}$ satisface $\theta_t = \alpha_t\theta_{t-1} + \beta_t(1 - \theta_{t-1})$. En este trabajo se relaciona α_t y β_t definiendo $\psi_t := \frac{\alpha_t(1-\beta_t)}{(1-\alpha_t)\beta_t}$ y que éste sea constante o función de los covariados y $\{\theta_t\}$ también puede ser función de covariados que dependen del tiempo. Sin embargo, en esta formulación parece que la forma de las componentes deterministas no restringe el rango de valores válidos para los parámetros del modelo. Adicionalmente, él propone algunas extensiones para medidas repetidas en datos binarios e incluso trabaja con algunos aspectos de datos faltantes. Él afirma que no hay problema si la depencia serial no se modela exactamente siempre y cuando no afecte la estimación de los parámetros de regresión. Esto parece alentador ya que frecuentemente se tiene interés en la significancia de estos parámetros. Además, sus simulaciones sugieren que los errores estándar estimados pueden ser robustos contra la falta de especificación de la estructura de independencia.

Ha habido varios intentos para hacer que los modelos estándar basados en adelgazamiento sean más atractivos para aplicaciones económicas al considerar covariados. Por ejemplo, Brännäs (1995) trabajó con una serie de conteo de datos suecos usando un Poisson $INAR(1)$ en el cual los parámetros son funciones de covariados. Por tanto, el modelo toma la forma $X_t = \alpha_t \circ X_{t-1} + Z_t$, donde $Ln(\alpha_t) - Ln(1 - \alpha_t) = \beta' Y_t$, Z_t se distribuye Poisson con media $\lambda_t = e^{\gamma' Y_t}$ y Y_t es un vector de covariados al tiempo t de dimensión p. En dicho trabajo se discute acerca de estimación y predicción.

4.3. El modelo auto-regresivo $INAR(p)$ de Al-Osh & Alzaid

La idea de esta sección es definir y estudiar algunas de las propiedades de una extensión del proceso $INAR(1)$ con el fin de tener modelos más flexibles para procesos de conteo. La forma "inocente" en la que se hace es reemplazando en el $AR(p)$ estándar, la multiplicación escalar por operaciones de adelgazamieno "$\circ$" como en el proceso $INAR(1)$. Sin embargo,

en esta propuesta de Al-Osh & Alzaid las analogías entre el $AR(p)$ y el $INAR(p)$ se limitan a esta "sustitución". De hecho, dadas las relaciones de dependencia que existen en este $INAR(p)$, la autocorrelación de este proceso es más parecida a la del $ARMA(p, p-1)$ estándar. Desafortunademente este proceso no cumple la propiedad de Markov, sin embargo un acomplamiento de este proceso sí la cumple.

El modelo $INAR(1)$, $X_t = \alpha \circ X_{t-1} + Z_t$, es apropiado para modelar datos de procesos de ramificación y de conteo. Sin embargo, las realizaciones de algún proceso de conteo $\{X_t\}$ pueden no sólo atribuirse al pasado inmediato X_{t-1} sino también a realizaciones pasadas del proceso $\{X_{t-j}\}_{j=2}^{p}$, para alguna constante p. Por tanto, para modelar tales procesos se necesita extender la estructura $INAR(1)$ considerando una forma análoga del proceso $AR(p)$ reemplazando la operación escalar "$\alpha_i X_{t-i}$" por la operación "$\alpha_i \circ X_i$" para $i \in \{1, 2, \ldots, p\}$. Sin embargo, al hacer este reemplazo de operaciones se deben hacer algunas suposiciones adicionales con el fin de que la estructura de dependencia del proceso esté bien definida.

Definición 4.3.1. *(Proceso $INAR(p)$)*

Se dice que el proceso $\{X_t\}$ es un proceso $INAR(p)$ si admite la representación

$$X_t = \sum_{i=1}^{p} \alpha_i \circ X_{t-i} + Z_t, \quad t \in \mathbb{Z} \tag{4.6}$$

donde $\{Z_t\}$ es una sucesión de variables aleatorias no negativas tal que $sop(Z_t) \subseteq \mathbb{N}$ con media μ_Z y varianza σ_Z^2, ambos finitos, y $\{a_i\}_{i=1}^{p}$ son constantes no negativas tales que $\sum_{i=1}^{p} \alpha_i < 1$. La distribución condicional del vector $(\alpha_1 \circ X_t, \alpha_2 \circ X_t, \ldots, \alpha_p \circ X_t)$ dado $X_t = x_t$ es multinomial con parámetros $(\alpha_1, \alpha_2, \ldots, \alpha_p, x_t)$ y es independiente de la historia pasada de proceso, i.e. dado $X_t = x_t$ la variable aleatoria $\alpha_i \circ X_t$ es independiente de X_{t-k} y sus supervivientes $\alpha_j \circ X_{t-k}$ para $i, j = 1, 2, \ldots, p$ y $k > 0$.

El proceso $INAR(p)$ definido en (4.6) se puede ver como un caso especial de un proceso de ramificación con inmigración; pues considérese una población animal en la cual la hembra se puede reproducir a lo más una vez durante su vida reproductiva (que se puede dividir en p periodos disjuntos). La probabilidad de que dicha hembra tenga descendecia durante el i-ésimo periodo es α_i. Es claro que el tamaño de la t-ésima generación, X_t, es simplemente el total de la descendencia de las últimas p generaciones más el proceso de inmigración Z_t que entra al sistema durante el intervalo de tiempo $(t-1, t]$. Nótese que si $\sum_{i=1}^{p} \alpha_i \geq 1$ el tamaño de la generación X_t se incrementará monótonamente y se tendrá un proceso explosivo. La interpretación previa es simplemente una generalización del proceso de ramificación $INAR(1)$ en el que sólo la última generación se puede reproducir.

Se debe notar que aunque la forma del modelo $INAR(p)$ es muy parecida a la del proceso auto-regresivo de orden p, la dependencia del tiempo definida por el operador "$\circ$" hace que este proceso sea diferente (por supuesto en la relación de dependencia) que el proceso $AR(p)$ estándar. Para fijar estas ideas se comenzará con el $INAR(2)$, es decir con el modelo

$$X_t = \alpha_1 \circ X_{t-1} + \alpha_2 \circ X_{t-2} + Z_t$$

En el proceso $AR(2)$ estándar X_t es una combinación lineal (con la multiplicación usual) de X_{t-1} y X_{t-2} sin necesariamente considerar toda la estructura estocástica previa. Sin embargo, en el proceso $INAR(2)$ no ocurre esto $\alpha_1 \circ X_{t-2}$ y $\alpha_2 \circ X_{t-2}$ que son componentes de X_{t-1} y X_t, respectivamente, están conectados de una forma más intrincada que en el caso del $AR(2)$ estándar en el cual sólo la presencia de X_{t-2} los conecta. Aquí las operaciones "$\alpha_1\circ$" y "$\alpha_2\circ$" sobre X_{t-2} son dependientes y se realizan en diferentes puntos del tiempo. La estructura es más clara cuando se piensa en términos de simulación. Al tiempo t, se observa X_t y se tiene disponible X_{t-1} y $V_{t-1} := \alpha_2 \circ X_{t-1}$, habiendo usado $\alpha_1 \circ X_{t-1}$ para obtener X_t. El primer paso es formar $U_t := \alpha_1 \circ X_t$ y $V_t := \alpha_2 \circ X_t$ y entonces $X_{t+1} = U_t + V_t + Z_{t+1}$ y V_t está disponible para usarse en la obtención de X_{t+2}.

La estructura de dependencia mutua entre los componentes de X_t (i.e. $\alpha_i \circ X_t$, $i = 1, 2, \ldots, p$) que aparecen en diferentes tiempos induce una estructura de promedios móviles en el proceso. De hecho, más adelante se verá que el comportamiento de la función de autocorrelación de proceso $INAR(p)$ es muy parecida al proceso $ARMA(p, p-1)$.

4.3.1. Distribución límite de X_t

Ahora, se discutirá acerca de la distribución límite del proceso $\{X_t\}$ definido por

$$X_t = \sum_{j=1}^{p} \alpha_j \circ X_{t-j} + Z_t, \quad t \in \mathbb{Z}$$

Defínase $\{w_j\}_{j=0}^{\infty}$ una sucesión de pesos tal que

$$w_0 = 1$$

$$w_j = \sum_{i=1}^{\min\{j,p\}} \alpha_i w_{j-i}$$

La sucesión $\{w_j\}_{j=0}^{\infty}$ forma los coeficientes de la representación usual de promedios móviles para el proceso $INAR(p)$. Por tanto,

$$\left(\sum_{k=0}^{\infty} w_k z^k\right)\left(1 - \sum_{k=1}^{p} \alpha_k z^k\right) = 1$$

Sean G_{X_t} y G_{Z_t} las funciones generadoras de probabilidad de X_t y Z_t, respectivamente. Entonces

$$\begin{aligned} G_{X_t}(s) &= \mathbb{E}\left[s^{X_t}\right] = \mathbb{E}\left[s^{\sum_{j=1}^{p}\alpha_j\circ X_{t-j}+Z_t}\right] \\ &= \mathbb{E}\left[(1-w_1+w_1 s)^{X_{t-1}} s^{\sum_{j=2}^{p}\alpha_j\circ X_{t-j}}\right] G_{Z_t}(1-w_0+w_0 s) \end{aligned}$$

Usando este argumento de forma iterativa se tiene

$$G_{X_t}(s) = \prod_{j=0}^{k-1} G_{Z_{t-j}}(1-w_j+w_j s)\mathbb{E}\left[\prod_{l=1}^{p}(1-w_{k-l}+w_{k-l}s)^{\sum_{i=l}^{p}\alpha_i\circ X_{t-k+l-i}}\right] \quad \text{para } k \geq p \tag{4.7}$$

Para encontrar la distribución límite de $\{X_t\}$ se necesita el siguiente lema.

Lema 4.3.1.

Considérese los pesos $\{w_j\}$ como antes. Si las raíces del polinomio

$$z^p - \alpha_1 z^{p-1} - \ldots - \alpha_p \quad \text{con } \alpha_p \neq 0$$

están dentro del círculo unitario, entonces existe $\lambda \in (0,1)$ tal que $0 \leq w_j \leq c\lambda^j$, $j = 0,1,2,\ldots$ para alguna constante c.

Teorema 4.3.1. *(Distribución límite I)*

Sea $\{X_t\}_{t=0}^{\infty}$ un proceso $INAR(p)$ con parámetros $\alpha_1, \alpha_2, \ldots, \alpha_p$ con $\alpha_p > 0$. Si las raíces del polinomio $z^p - \alpha_1 z^{p-1} - \ldots - \alpha_p$ están dentro del círculo unitario y

$$\sum_{j=0}^{\infty}(j+1)^{-1}\sum_{k=j+1}^{\infty}\mathbb{P}(Z_1 = k) < \infty$$

Entonces el proceso $\{X_t\}_{t\in\mathbb{N}}$ tiene una distribución límite, cuya función generadora de probabilidad está dada por

$$G(s) = \prod_{j=0}^{\infty} G_Z(1-w_j+w_j s) \tag{4.8}$$

Demostración:

De (4.7) se tiene que

$$G_{X_{t+p}}(s) = \prod_{j=0}^{t-1} G(1 - w_j + w_j s)\mathbb{E}\left[\prod_{k=1}^{p}(1 - w_{t-k} + w_{t-k}s)^{\sum_{i=1}^{p}\alpha_i \circ X_{p+k-i}}\right]$$

pero por el Lema 4.3.1 y los Teoremas de Convergencia Acotada se tiene que

$$\lim_{t\to\infty} \mathbb{E}\left[\prod_{k=1}^{p}(1 - w_{t-k} + w_{t-k}s)^{\sum_{i=1}^{p}\alpha_i \circ X_{p+k-i}}\right] = 1$$

Entonces, es suficiente demostrar que $\prod_{j=0}^{t-1} G_Z(1 - w_j + w_j s)$ es convergente. Sin embargo, $\prod_{j=0}^{t-1} G_Z(1 - w_j + w_j s) < \infty$ si y sólo si $\sum_{j=1}^{t}(1 - G_Z(1 - w_j + w_j s)) < \infty$. Como G_Z es no decreciente, de nuevo por el Lema 4.3.1 se tiene que

$$\begin{aligned} 1 - G_Z(1 - w_j + w_j s) &\leq 1 - G_Z(1 - c\lambda^j(1 - s)) \\ &\leq 1 - G_Z(1 - \lambda^j) \text{ para } 1 - 1/c \leq s < 1 \end{aligned}$$

Ahora, usando un argumento de Heathcote (1966) se tiene que $\prod_{j=0}^{t-1} G_Z(1-w_j+w_j s) < \infty$ si y sólo si $\sum_{j=0}^{\infty}(j+1)^{-1}\sum_{k=j+1}^{\infty}\mathbb{P}(Z_1 = k) < \infty$ (que se está suponiendo por hipótesis). □

El siguiente teorema menciona que la distribución límite del proceso $\{X_t\}$ con función generadora de probabilidad dada por (4.8) también es válida si el proceso es estacionario y empezó desde el pasado infinito.

Teorema 4.3.2. *(Distribución límite II)*

Sea $\{X_t\}_{t\in\mathbb{Z}}$ un proceso $INAR(p)$ estacionario con parámetros $\alpha_1, \alpha_2, \ldots, \alpha_p$ con $\alpha_p > 0$. Si las raíces del polinomio $z^p - \alpha_1 z^{p-1} - \ldots - \alpha_p$ están dentro del círculo unitario y

$$\sum_{j=0}^{\infty}(j+1)^{-1}\sum_{k=j+1}^{\infty}\mathbb{P}(Z_1 = k) < \infty$$

Entonces el proceso $\{X_t\}_{t\in\mathbb{Z}}$ tiene una distribución límite, cuya función generadora de probabilidad está dada por

$$G(s) = \prod_{j=0}^{\infty} G_Z(1 - w_j + w_j s)$$

Observación 4.3.1.

El Teorema 4.3.1 es equivalente a decir que la distribución marginal de X_t es idéntica a la que sea obtiene de un proceso de medias móviles de valores enteros ($INMA$) de orden infinito con parámetros $\{w_j\}_{j=0}^{\infty}$ y sucesión de innovación $\{Z_t\}$, i.e.

$$X_t \overset{d}{=} \sum_{j=0}^{\infty} w_j \circ Z_{t-j},\ t \in \mathbb{Z}$$

Vale la pena mencionar que a diferencia del proceso de medias móviles estándar, el proceso de ruido de innovación $\{Z_t\}$ satisface que dado $Z_t = z_t$ las variables $w_j \circ Z_t$ y $w_i \circ Z_t$ son dependientes. ▽

4.3.2. Propiedades de correlación

Para estudiar las propiedades de correlación del proceso $INAR(p)$ se hará la suposición de que la variable $\alpha_i \circ X_t | X_t = x_t$ es independiente de la historia pasada del proceso $\{X_{t-k}\}$ y de $\alpha_j \circ X_{t-k}$ para $i, j = 1, 2, \ldots, p$ y $k \geq 1$.

Entonces de la expresión

$$X_t = \sum_{i=1}^{p} \alpha_i \circ X_{t-i} + Z_t, \quad \text{para } t \in \mathbb{Z}$$

se tiene que

$$\begin{aligned}\mathbb{E}(X_t) &= \mathbb{E}\left(\sum_{i=1}^{p}\alpha_i \circ X_{t-i}\right) + \mathbb{E}(Z_t) \\ &= \sum_{i=1}^{p}\alpha_i \mathbb{E}(X_{t-i}) + \mu_Z\end{aligned}$$

donde $\mu_Z := \mathbb{E}(Z_t)$.

Si además se supone que $\{X_t\}$ es débilmente estacionario entonces

$$\mu_X := \mathbb{E}(X_t) = \frac{\mu_Z}{1-\alpha_1-\alpha_2-\ldots-\alpha_p}$$

Ahora, se obtendrá la función de autocovarianza, $\gamma_X(k)$:

$$\begin{aligned}\gamma_X(k) = Cov(X_{t-k}, X_t) &= Cov\left(X_{t-k}, \sum_{i=1}^{p}\alpha_i \circ X_{t-i} + Z_t\right) \\ &= \sum_{i=1}^{p} Cov\left(X_{t-k}, \alpha_i \circ X_{t-i}\right) + Cov(X_{t-k}, Z_t) \\ &= \sum_{i=1}^{p} Cov\left(X_{t-k}, \alpha_i \circ X_{t-i}\right) + I_{\{k\}}(0)\sigma_Z^2\end{aligned}$$

Sin embargo, nótese que

$$Cov\left(X_{t-l}, \alpha \circ X_t\right) = \alpha Cov(X_{t-l}, X_t)$$

Dada esta propiedad, defínase

$$\gamma(l, \alpha_i) := Cov\left(X_{t-l}, \alpha \circ X_t\right)$$

Entonces, se cumple que

$$\gamma(l, \alpha) = \alpha\gamma_X(l), \quad l \in \mathbb{N}$$

De aquí que, usando un argumento de que $(\alpha_1 \circ X|X = x, \alpha_2 \circ X|X = x, \ldots, \alpha_p \circ X|X = x)$ tiene distribución multinomial con parámetros $(\alpha_1, \alpha_2, \ldots, \alpha_p, x)$, se tiene que

$$Cov(\alpha_j \circ X_{t-k}, \alpha_i \circ X_{t-l}) = \begin{cases}\alpha_j\gamma(k-l, \alpha_i) & \text{si } k < l \\ \alpha_j\alpha_i\gamma_X(0) + \alpha_j(I_{\{i\}}(j) - \alpha_i)\mu_x & \text{si } k = l \\ \alpha_i\gamma(l-k, \alpha_j) & \text{si } k > l\end{cases} \tag{4.9}$$

Se puede determinar $\gamma(-l, \alpha)$ utilizando recursivamente (4.9). Comenzando con $l = 1$ se tiene que

$$\begin{aligned}\gamma(-l, \alpha_i) &= \alpha_1\alpha_i\gamma_X(0) + \alpha_1(I_{\{1\}}(i) - \alpha_i)\mu_X + \alpha_i\sum_{j=2}^{p}\gamma(\alpha_j, j-1) \\ &= \alpha_i\gamma_X(-1) + \alpha_1(I_{\{1\}}(i) - \alpha_i)\mu_X\end{aligned}$$

Defínase $\nu(l, \alpha_i) := \gamma(l, \alpha_i) - \alpha_i\gamma_X(l)$. Como $\gamma(l, \alpha) = \alpha\gamma_X(l)$, entonces para $l \in \mathbb{N}$, $\nu(l, \alpha_i) = 0$. Y por la ecuación

$$\nu(-1, \alpha_i) = \alpha_1(I_{\{1\}}(i) - \alpha_i)\mu_X$$

Por ejemplo,

$$\begin{aligned}\nu(-2, \alpha_i) &= \sum_{j=1}^{p} Cov(\alpha_j \circ X_{t-j}, \alpha_i \circ X_{t-2}) - \alpha_i\gamma_X(-2) \\ &= \alpha_1\gamma(-1, \alpha_i) + \alpha_2\alpha_i\gamma_X(0) + \alpha_2(I_{\{2\}}(i) - \alpha_i)\mu_X \\ &\quad +\alpha_i\sum_{j=3}^{p} Cov(\alpha_j \circ X_{t-j}, X_{t-2}) - \alpha_i\gamma(-2) \\ &= \alpha_1\nu(-1, \alpha_i) + \alpha_2(I_{\{2\}}(i) - \alpha_i)\mu_X\end{aligned}$$

Haciendo esto varias veces, se obtiene

$$\nu(-l, \alpha_i) = \sum_{j=1}^{l-1}\alpha_j\nu(j-l, \alpha_i) + \alpha_i(I_{\{l\}}(i) - \alpha_l)\mu_x$$

Esto significa que $\nu(-l, \alpha_i)$ es una función lineal de la media del proceso μ_x. A partir de la definición de $\nu(l, \alpha_i)$ se puede ver que $\gamma(k-i, \alpha_i)$ satisface

$$\gamma(k-i, \alpha_i) = \alpha_i\gamma_X(k-i) + \nu(k-i, \alpha_i)$$

donde para $k \geq i$, $\nu(k-i, \alpha_i) = 0$.

Además,

$$\gamma_X(k) = \sum_{i=1}^{p}\alpha_i\gamma_X(k-i) + \sum_{i=k+1}^{p}\nu(k-i, \alpha_i) + I_{\{0\}}(k)\sigma_Z^2$$

Se puede ver a partir de la expresión anterior que la autocovarianza del proceso $INAR(p)$ es de la misma forma que la del proceso $ARMA(p, p-1)$ Gaussiano. Este comportamiento de $\gamma_X(k)$ se debe a la estructura de dependencia mutua entre las componentes de X_t, i.e. $\alpha_j \circ X_{t-i}$ para $i = 1, 2, \ldots, p$ aparece diferentes veces en lo que se ha discutido hasta el momento.

4.3.3. Representación espacial del proceso $INAR(p)$

En esta búsqueda de analogías entre los procesos $INAR(p)$ y $AR(p)$, ahora se dará una representación espacial del modelo $INAR(p)$ (pues el $AR(p)$ estándar sí admite una representación espacial). Y con esta representación espacial se encontrará la distribución conjunta del proceso y se comparará este proceso con el $AR(1)$ Poisson múltiple de McKenzie (1988). Como antes, se será más conciso con el proceso $INAR(2)$ y se enunciarán algunas propiedades para estructuras de auto-regresión con orden más alto.

El vector espacial que se considerará para este proceso es $\mathbf{X}_t^* := (X_t, \alpha_2 \circ X_{t-1})'$. Este vector es más fácil de manipular que el vector $\mathbf{X}_t := (X_t, X_{t-1})'$. La distribución conjunta y las correspondientes distribuciones marginales del proceso $\{X_t\}$ se pueden obtener a partir de la distribución conjunta de $\{\mathbf{X}_t^*\}$.

El proceso $\{\mathbf{X}_t^*\}$ es Markoviano ya que

$$\mathbb{E}(s^{X_t} r^{\alpha_2 \circ X_{t-1}} | \mathbf{X}_{t-1}^*, \mathbf{X}_{t-2}^*, \ldots) G_{Z_t}(1 - \alpha_1 - \alpha_2 + \alpha_1 s + \alpha_2 r)^{X_{t-1}} r^{\alpha_2 \circ X_{t-1}}$$

Por tanto, es suficiente considerar la distribución conjunta de $\mathbf{X}_t^*$ y $\mathbf{X}_{t-1}^*$ para determinar la distribución conjunta del proceso. Como se tiene la hipótesis de que dado $X_{t-1} = x$, $(\alpha_1 \circ X_{t-1}, \alpha_2 \circ X_{t-1})'$ tiene distribución multinomial y es independiente de X_{t-2} y sus supervivientes, la función generadora de probabilidad conjunta de $\mathbf{X}_t^*$ y $\mathbf{X}_{t-1}^*$ está dada por

$$\begin{aligned} G_{\mathbf{X}_t^*, \mathbf{X}_{t-1}^*}(\mathbf{s}, \mathbf{r}) &= \mathbb{E}(s_1^{X_t} s_2^{\alpha_2 \circ X_{t-1}} r_1^{X_{t-1}} r_2^{\alpha_2 \circ X_{t-2}}) \\ &= G_{\mathbf{Z}}(\mathbf{s}) G_{\mathbf{X}_{t-1}^*}(r_1(1 - \alpha_1 - \alpha_2 + \alpha_1 s_1 + \alpha_2 s_2), s_1 r_2) \end{aligned}$$

donde $\mathbf{s} = (s_1, s_2)'$, $\mathbf{r} = (r_1, r_2)'$ y $G_{\mathbf{Z}}$ es la función generadora de momentos de $\mathbf{Z}_t := (z_t, 0)'$.

Esto significa que la distribución conjunta del proceso $\{\mathbf{X}_t^*\}$ está determinada por las distribuciones marginales del proceso mismo y la del proceso $\{\mathbf{Z}_t\}$.

A continuación, se obtendrán las distribuciones marginal y límite del proceso $\{\mathbf{X}_t^*\}$.

$$\begin{aligned} G_{\mathbf{X}_t^*}(\mathbf{s}) &= \mathbb{E}(s_1^{X_t} s_2^{\alpha_2 \circ X_{t-1}}) \\ &= \mathbb{E}(s_1^{\alpha_1 \circ X_{t-1} + \alpha_2 X_{t-2} + Z_t} s_2^{\alpha_2 \circ X_{t-1}}) \\ &= G_{\mathbf{Z}}(\mathbf{s}) G_{\mathbf{X}_{t-1}^*}(1 - \alpha_1 - \alpha_2 + \alpha_1 s_1 + \alpha_2 s_2, s_1) \\ &= G_{\mathbf{Z}}(\mathbf{s}) G_{\mathbf{X}_{t-1}^*}(\mathbf{1} - \mathbf{A}'(\mathbf{1} - \mathbf{s})) \end{aligned}$$

donde $\mathbf{1} := (1,1)'$ y

$$\mathbf{A} = \begin{pmatrix} \alpha_1 & 1 \\ \alpha_2 & 0 \end{pmatrix}$$

La forma de $G_{\mathbf{X}_t^*}(\mathbf{s})$ sugiere que el proceso $\mathbf{X}_t{}^*$ se puede ver como una proceso de ramificación multi-tipo con inmigración. Steutel & van Harn (1986) consideraron una función generadora de probabilidad con esta forma para una matriz general para definir el concepto de auto-descomponibilidad discreta multivariada y McKenzie (1988) también consideró una descomposición como esta para definir un proceso vectorial auto-regresivo de orden uno.

La metodología anterior se puede extender a process $INAR$ de orden mayor. Para el proceso $INAR(p)$ general el correpondiente vector de estados está dado por

$$\mathbf{X}_t = \left(X_t, \sum_{i=2}^{p} \alpha_i \circ X_{t+1-i}, \sum_{i=3}^{p} \alpha_i \circ X_{t+2-i}, , \alpha_p \circ X_{t-1} \right)'$$

que es de la misma forma que el vector de estados del $AR(p)$ Gaussiano (Harvey & Phillips (1979)).

De nuevo, con un procedimiento iterativo se tiene que

$$G_{\mathbf{X}_t^*}(\mathbf{s}) = G_{\mathbf{X}_1^*}(\mathbf{1} - (\mathbf{A}')^t(\mathbf{1} - \mathbf{s})) \prod_{i=1}^{t-1} G_{\mathbf{Z}_t}(\mathbf{1} - (\mathbf{A}')^i(\mathbf{1} - \mathbf{s})) G_{\mathbf{Z}_t}(\mathbf{s})$$

Por tanto, las distribuciones marginales (y por tanto la distribución conjunta) de $\{\mathbf{X}_t^*\}$ se pueden determinar mediante la distribución inicial de $\mathbf{X}_1^*$ y la distribución de la sucesión $\{\mathbf{Z}_t\}$. Sin embargo, si el proceso es estacionario la distribución de $\mathbf{X}_t^*$ será la de la distribución límite. A continuación se discutirá acerca de la distribución límite de $\{\mathbf{X}_t^*\}$.

Encontrando directamente las potencias de al matriz $\mathbf{A}'$ se puede ver que para $n \in \mathbb{N}^+$, $\mathbf{A}^n$ tiene la forma

$$\mathbf{A}^n = \begin{pmatrix} w_n & w_{n-1} \\ \alpha_2 w_{n-1} & \alpha_2 w_{n-2} \end{pmatrix}$$

donde los pesos w_j se definen como los pesos de la representación en promedios móviles para el $INAR(p)$ y $w_j = 0$ para $j < 0$. Por el lema (4.3.1) se tiene que

$$\lim_{n\to\infty} w_n = 0$$

Y por tanto,

$$\lim_{n\to\infty} A^n = 0$$

Y con un argumento similar al que se usó en la demostración del Teorema (4.3.1) la distribución límite de X_n tiene distribución límite tiene función generadora de probabilidad

$$G^*(\mathbf{s}) = \prod_{i=0}^{\infty} G_Z(1 - w_i(1 - s_1) - \alpha_2 w_{i-1}(1 - s_2)) \tag{4.10}$$

Por tanto, dada la distribución de la sucesión de innovación $\{Z_t\}$ se puede determinar $G^*(\mathbf{s})$

4.3.4. $INAR(p)$ Poisson

Ahora, se utilizá la representación

$$G_{\mathbf{X}_t^*}(\mathbf{s}) = G_{\mathbf{Z}}(\mathbf{s}) G_{\mathbf{X}_{t-1}^*}(\mathbf{1} - \mathbf{A}'(\mathbf{1} - \mathbf{s}))$$

en el caso en el que $\{Z_t\}$ es una sucesión de variables aleatorias independientes con distribución Poisson.

La distribución Poisson es quizá la distribución más común para modelar procesos de conteo. En el proceso $INAR(1)$, la distribución Poisson tiene un papel similar al de la distribución Gaussiana en el $AR(1)$. Específicamente tiene la propiedad de que si la sucesión de innovación $\{Z_t\}$ tiene distribución Poisson y la distribución inicial también es Poisson, entonces la distribución marginal de $\{X_t\}$ es Poisson. Ahora, se considerará una extensión del proceso $INAR(1)$ Poisson a una estructura de auto-regresión de orden mayor y se comparará este proceso con el Poisson múltiple definido por McKenzie (1988).

4.3.4.1. El proceso Poisson múltiple incrustado

Supóngase que la sucesión de innovación $\{Z_t\}$ en 4.6 tiene distribución $Poisson(\lambda)$ y que el proceso $INAR(p)$ es estacionario. Por la expresión (4.10), en el proceso $INAR(2)$ se cumple que

$$G^*(\mathbf{s}) = \exp\{-\lambda\varpi((1 - s_1) + \alpha_2(1 - s_2))\}$$

donde

$$\varpi := \sum_{t=0}^{\infty} w_t = \frac{1}{1-\alpha_1-\alpha_2} < \infty$$

i.e. X_t y $\alpha_2 \circ X_{t-1}$ son variables independientes tales que $X_t \sim Poisson(\varpi\lambda)$ y $\alpha_2 \circ X_{t-1} \sim Poisson(\lambda\varpi\alpha_2)$.

Definición 4.3.2. *(Poisson múltiple)*

Se dice que un proceso p-variado, $\{\mathbf{Y}_t\}$ es un Poisson múltiple si las componentes de $\mathbf{Y}_t$ son independientes y todas tienen distribución Poisson.

Por tanto $\{\mathbf{X}_t^*\}$ es un Poisson múltiple auto-regresivo de orden uno (McKenzie (1988)).

Utilizando la ecuación

$$G_{\mathbf{X}_t^*,\mathbf{X}_{t-1}^*}(\mathbf{s},\mathbf{r}) = G_{\mathbf{Z}}(\mathbf{s})G_{\mathbf{X}_{t-1}^*}(r_1(1-\alpha_1-\alpha_2+\alpha_1 s_1+\alpha_2 s_2), s_1 r_2)$$

se puede ver que la función generadora de probabilidad conjunta de $\mathbf{X}_t^*$ y $\mathbf{X}_{t-1}^*$ está dada por

$$G_{\mathbf{X}_t^*,\mathbf{X}_{t-1}^*}(\mathbf{s},\mathbf{r}) = \exp\left\{-\boldsymbol{\theta}'(1-\mathbf{s}) - \boldsymbol{\theta}'(1-\mathbf{r}) + (1-\mathbf{s})'\Gamma_{\mathbf{X}_t^*}(1)'(1-\mathbf{r})\right\}$$

donde $\boldsymbol{\theta}' = (\lambda\varpi, \lambda\alpha_2\varpi)$ y

$$\Gamma_{\mathbf{X}_t^*}(1) = \begin{pmatrix} \lambda\varpi\alpha_1 & \lambda\varpi\alpha_2 \\ \lambda\varpi\alpha_2 & 0 \end{pmatrix}$$

es la matriz de autocovarianzas de orden 1 del proceso $\{\mathbf{X}_t^*\}$.

Como el proceso $\{\mathbf{X}_t^*\}$ es de Markov y la función generadora de probabilidad de $\mathbf{X}_t^*$ y $\mathbf{X}_{t-1}^*$ es "simétrica" en $\mathbf{s}$ y $\mathbf{r}$, entonces el proceso $\{\mathbf{X}_t^*\}$ es reversible (McKenzie (1988)).

4.3.4.2. El proceso $INAR(2)$ Poisson

Como se mencionó anteriormente, la función generadora de probabilidad conjunta de un proceso $INAR(p)$ se puede obtener de que el correspondiente proceso de ramificación multi-tipo $\mathbf{X}_t^*$. Específicamente, la función generadora de probabilidad de $\mathbf{X}_t^* = (X_t, X_{t-1})'$, $G_{\mathbf{X}_t^*}(\cdot)$, en el proceso $INAR(2)$ se puede obtener la expresión

$$G_{\mathbf{X}_t}(s_1,s_2) = \exp\{-\lambda\varpi(1-\alpha_1)[(1-s_1)+(1-s_2)] - \lambda\varpi\alpha_1(1-s_1 s_2)\}$$

Y de forma más general, la distribución conjunta de $X_1, \ldots, X_t$ es una distribución marginal de una distribución conjunta del correspondiente proceso "incrustado" $\mathbf{X}_1^*, \mathbf{X}_2^*, \ldots, \mathbf{X}_t^*$. Por la reversibilidad en el tiempo del proceso $\{\mathbf{X}_t^*\}$ se puede concluir que el proceso $\{X_t\}$ es reversible en el tiempo. De forma más precisa

$$\begin{aligned} G_{X_1,\ldots,X_t}(s_1,\ldots,s_t) &= G_{\mathbf{X}_1^*,\ldots,\mathbf{X}_t^*}(\mathbf{s}_1,\ldots,\mathbf{s}_t) \\ &= G_{\mathbf{X}_1^*,\ldots,\mathbf{X}_t^*}(\mathbf{s}_t,\ldots,\mathbf{s}_1) \\ &= G_{X_1,\ldots,X_t}(s_t,\ldots,s_1) \end{aligned}$$

donde $\mathbf{s}_i := (s_i, 1)'$, $i \in \{1, 2, \ldots, t\}$. Como resultado de esta propiedad se tiene que

$$\mathbb{E}(X_t | X_{t-1} = x_1, X_{t-2} = x_2) = \mathbb{E}(X_{t-2} | X_{t-1} = x_1, X_t = x_2)$$

Estas esperanzas condicionales se puede encontrar utilizando la función generadora de probabilidad conjunta $G_{X_{t-2},X_{t-1},X_t}(\cdot,\cdot,\cdot)$ obteniéndose

$$\begin{aligned} \mathbb{E}(X_t | X_{t-1} = x, X_{t-2} = z) = \lambda \Bigg[& 1 + (1-\alpha_1)\varpi \frac{p(x-1,z)}{p(x,z)} \\ & + \alpha_2 \varpi \frac{p(x,z-1)}{p(x,z)} + \alpha_1^2 \varpi \frac{p(x-1,z-1)}{p(x,z)} \Bigg] \end{aligned}$$

donde $p(\cdot,\cdot)$ es la función de probabilidad conjunta de las variables X_{t-1} y X_t. La forma de esta ecuación sugiere que la regresión en el proceso $INAR(2)$ en general puede no ser lineal.

4.4. El modelo auto-regresivo $INAR(p)$ de Du & Li

En esta sección se dará una versión diferente del proceso $INAR(p)$ a la dada por Al-Osh y Alzaid. Esta versión de modelo $INAR(p)$ es una propuesta de Du & Li (1991) y está construida de tal forma que este modelo $INAR(p)$ tenga la misma estructura de correlación que el $AR(p)$ estándar. La diferencia fundamental entre ambas versiones de $INAR(p)$ es la forma en la que se lleva a cabo la operación de adelgazamiento "$\circ$".

El proceso $INAR(1)$ está dado por

$$X_t = \alpha \circ X_{t-1} + Z_t, \quad t \in \mathbb{Z} \tag{4.11}$$

donde $\{Z_t\}$ es una sucesión de variables aleatorias idénticamente distribuidas tales que $sop(Z_t) \subseteq \mathbb{N}$ e independiente de todas las series de conteo (que definen $\circ$) en (4.11). Aquí

se está suponiendo que $\{Z_t\}$ son independientes e idénticamente distibuidas con el fin de hacer a X_{t-1} independiente de las series de conteo, justo esta es la diferencia "más fuerte" que la de Al-Osh y Alzaid.

Por supuesto, la extensión natural de (4.11) es

$$X_t = \alpha_1 \circ X_{t-1} + \ldots + \alpha_p \circ X_{t-1p} + Z_t, \quad t \in \mathbb{Z} \tag{4.12}$$

donde $\{Z_t\}$ es una sucesión de variables aleatorias idénticamente distribuidas tales que $sop(Z_t) \subseteq \mathbb{N}$ e independiente de todas las series de conteo (que definen $\circ$), todas las series de conteo son mutuamente independientes y $0 \leq \alpha_1, \ldots, \alpha_p \leq 1$.

4.4.1. Teorema de existencia

A continuación, se intententará analizar las condiciones sobre $\alpha_1, \ldots, \alpha_p$ para que exista una solución estacionaria de (4.12) exista.

Teorema 4.4.1. *(Teorema de existencia)*

Sea $\{Z_t\}$ una sucesión de variables aleatorias no-negativas tal que $sop(Z_t) \subseteq \mathbb{Z}$, $\mathbb{E}(Z_t) = \mu_z$, $Var(Z_t) = \sigma_z^2$ y $\alpha_i \in [0,1]$, $i = 1, 2, \ldots, p$. Si las raíces del polinomio

$$z^p - \alpha_1 z^{p-1} - \ldots - \alpha_{p-1} z - \alpha_p$$

están dentro del círculo unitario, entonces existe un único proceso de series de tiempo estacionario no negativo y con valores en los enteros $\{X_t\}$ tal que

$$X_t = \alpha_1 \circ X_{t-1} + \ldots + \alpha_p \circ X_{t-p} + Z_t \tag{4.13}$$

$$Cov(X_s, Z_t) = 0, \quad s < t$$

Para probar este teorema de existencia se necesitan los siguientes lemas.

Lema 4.4.1. *(Algunas propiedades de $\circ$)*

(a) $\mathbb{E}(\alpha \circ X) = \alpha\mathbb{E}(X)$

(b) $\mathbb{E}[(\alpha \circ X)^2] = \alpha(1-\alpha)\mathbb{E}(X) + \alpha^2\mathbb{E}(X^2)$

(c) $\mathbb{E}[(\alpha \circ X - \alpha \circ Y)^2] = \alpha(1-\alpha)\mathbb{E}(|X-Y|) + \alpha^2\mathbb{E}[(X-Y)^2]$ donde las series de conteo en $\alpha \circ X$ y $\alpha \circ Y$ son las mismas

(d) $\mathbb{E}[(\alpha \circ X)(\beta \circ Y)] = \alpha\beta\mathbb{E}(XY)$ donde

$$\alpha \circ X = \sum_{j=1}^{X} Y_j \quad \text{y} \quad \beta \circ Y = \sum_{j=1}^{Y} Y_j^*,$$

$\{Y_j\}$ es independientes de $\{Y_j^*\}$ y (X, Y) es independiente de $\{Y_j\}$ y $\{Y_j^*\}$

Notación 4.4.1. *(Operador $\circ$ matricial)*

Sean

$$A = \begin{pmatrix} \alpha_1 & \alpha_2 & \cdots & \alpha_{p-1} & \alpha_p \\ 1 & 0 & \cdots & 0 & 0 \\ 0 & 0 & \ddots & 0 & 0 \\ 0 & 0 & \cdots & 1 & 0 \end{pmatrix}$$

y

$$\mathbf{X} = \begin{pmatrix} X_1 \\ X_2 \\ \vdots \\ X_p \end{pmatrix}$$

Se usa la notación

$$A \circ \mathbf{X} = \begin{pmatrix} \sum_{j=1}^{p} \alpha_j \circ X_j, X_1, \ldots, X_{p-1} \end{pmatrix}$$

∇

Lema 4.4.2. *(Propiedades de $\circ$ matricial)*

(a) $\mathbb{E}(A \circ X) = A\mathbb{E}(X)$

(b) $\mathbb{E}[(A \circ X)(A \circ X)'] = A\mathbb{E}(XX')A' + C$, donde $C \in M_{p\times p}(\mathbb{R})$ es de la forma

$$C = \begin{bmatrix} \sum_{j=1}^{p} \alpha_j(1-\alpha_j)\mathbb{E}(X_j) & \mathbf{0} \\ \mathbf{0} & \mathbf{0} \end{bmatrix}$$

Notación 4.4.2.

Sean $A \in M_{m\times n}(\mathbb{R})$, $A = (a_{ij})$ y $X, Y \in M_{n\times 1}(\mathbb{R})$.

(i) Si para cualesquiera $i \in \{1, \ldots, m\}, j \in \{1, \ldots, n\}$ $a_{ij} \geq 0$, entonces se escribirá $A \succeq 0$

(ii) Si $X - Y \succeq 0$, se escribirá $X \succeq Y$

$\triangledown$

Lema 4.4.3.

(a) (Propiedad de monotonía) Si $X \succeq Y$, $A \succeq 0$, entonces $AX \succeq AY$

(b) Si $X_n \xrightarrow{L^2} X$ y $X_n \xrightarrow{c.s.} Y$, entonces $X = Y$, *c.s.*

(c) Si $X_n \xrightarrow{L^2} X$ entonces $\alpha \circ X_n \xrightarrow{L^2} \alpha \circ X$ donde la serie de conteo en $\alpha \circ X_n$ y $\alpha \circ X$ es la misma.

También se necesitará el concepto conocido como *cuadrado de Kronecker* de una matriz (pues se utilizará en la demostración del Teorema de Existencia). Como ejemplo se considerará una matriz de 3×3. De acuerdo a Bellman (1974), el cuadrado de Kronecker de la matriz $A \in M_{3\times 3}(\mathbb{R})$, $A_{[2]}$, se define como

$$A_{[2]} = \begin{pmatrix} a_{11}^2 & a_{12}^2 & a_{13}^2 & 2a_{11}a_{12} & 2a_{11}a_{13} & 2a_{12}a_{13} \\ a_{21}^2 & a_{22}^2 & a_{23}^2 & 2a_{21}a_{22} & 2a_{21}a_{23} & 2a_{22}a_{23} \\ a_{31}^2 & a_{32}^2 & a_{33}^2 & 2a_{31}a_{32} & 2a_{31}a_{33} & 2a_{32}a_{33} \\ a_{11}a_{21} & a_{11}a_{22} & a_{13}a_{23} & a_{11}a_{22} + a_{12}a_{21} & a_{11}a_{23} + a_{13}a_{21} & a_{12}a_{23} + a_{13}a_{22} \\ a_{11}a_{31} & a_{12}a_{32} & a_{13}a_{33} & a_{11}a_{32} + a_{12}a_{31} & a_{11}a_{33} + a_{13}a_{31} & a_{12}a_{33} + a_{13}a_{32} \\ a_{21}a_{31} & a_{22}a_{32} & a_{23}a_{33} & a_{21}a_{32} + a_{22}a_{31} & a_{21}a_{33} + a_{23}a_{31} & a_{22}a_{33} + a_{23}a_{32} \end{pmatrix}$$

Se puede definir de la misma manera $A_{[2]}$ para una matriz $A \in M_{p\times p}(\mathbb{R})$. Nótese que para $A \in M_{p\times p}(\mathbb{R})$, $A_{[2]}$ es una matriz cuadrada de $p+\frac{p(p-1)}{2}$ renglones y $p+\frac{p(p-1)}{2}$ columnas. El siguiente teorema exhibe la relación entre los eigenvalores de A y los eigenvalores de $A_{[2]}$.

Lema 4.4.4.

Si $\lambda_1, \lambda_2, \ldots, \lambda_p$ son los eigenvalores de una matriz A, entonces $\lambda_1, \lambda_2, \ldots, \lambda_p, \lambda_1\lambda_2, \ldots, \lambda_1\lambda_p, \lambda_2\lambda_3, \ldots, \lambda_{p-1}\lambda_p$ son los eigenvalores de $A_{[2]}$

Ahora sí, se probará el antes mencionado Teorema de Existencia

Teorema 4.4.2. *(Teorema de existencia)*

Sea $\{Z_t\}$ una sucesión de variables aleatorias no-negativas tal que $sop(Z_t) \subseteq \mathbb{Z}$, $\mathbb{E}(Z_t) = \mu_z$, $Var(Z_t) = \sigma_z^2$ y $\alpha_i \in [0,1]$, $i = 1, 2, \ldots, p$. Si las raíces del polinomio

$$z^p - \alpha_1 z^{p-1} - \ldots - \alpha_{p-1} z - \alpha_p$$

están dentro del círculo unitario, entonces existe un único proceso de series de tiempo estacionario no negativo y con valores en los enteros $\{X_t\}$ tal que

$$X_t = \alpha_1 \circ X_{t-1} + \ldots + \alpha_p \circ X_{t-p} + Z_t$$

$$Cov(X_s, Z_t) = 0, \quad s < t$$

Demostración:

Defínase

$$S_{n,t} = \begin{cases} 0 & \text{si } n < 0; \\ Z_t & \text{si } n = 0; \\ \alpha_1 \circ S_{n-1,t-1} + \ldots + \alpha_p \circ S_{n-p,t-p} & \text{si } n > 0 \end{cases} \tag{4.14}$$

y

$$\mathbf{S}_{n,t} = \begin{pmatrix} S_{n,t} \\ S_{n-1,t-1} \\ \vdots \\ S_{n-p+1,t-p+1} \end{pmatrix}$$

entonces (4.14) se puede escribir como

$$\mathbf{S}_{n,t} = A \circ \mathbf{S}_{n-1,t-1} + \mathbf{Z}_t \tag{4.15}$$

donde

$$A = \begin{pmatrix} \alpha_1 & \alpha_2 & \cdots & \alpha_{p-1} & \alpha_p \\ 1 & 0 & \cdots & 0 & 0 \\ 0 & 0 & \ddots & 0 & 0 \\ 0 & 0 & \cdots & 1 & 0 \end{pmatrix}$$

$$\mathbf{Z}_t = \begin{pmatrix} Z_t \\ 0 \\ \vdots \\ 0 \\ 0 \end{pmatrix}$$

Sean $\boldsymbol{\mu} := \mathbb{E}(\mathbf{Z}_t)$, $\Sigma_1 := \mathbb{E}(\mathbf{Z}_t\mathbf{Z}_t')$. Sea $L^2(\Omega, \mathscr{F}, \mathbb{P}) = \{X : \mathbb{E}(X^2) < \infty\}$. Defínase el producto escalar en $L^2(\Omega, \mathscr{F}, \mathbb{P})$ usual $\langle X, Y\rangle := \mathbb{E}(XY)$. Es bien sabido que con este producto interno $L^2(\Omega, \mathscr{F}, \mathbb{P})$ es un espacio de Hilbert.

Afirmación 1: $S_{n,t} \in L^2(\Omega, \mathscr{F}, \mathbb{P})$

Prueba:

De (4.15) y el lema 4.4.2 se tiene que

$$\begin{aligned} \boldsymbol{\mu}_n &:= \mathbb{E}(\mathbf{S}_{n,t}) = A\mathbb{E}(\mathbf{S}_{n-1,t-1}) + \boldsymbol{\mu} \\ &= \boldsymbol{\mu} + A\boldsymbol{\mu} + \ldots + A^n\boldsymbol{\mu} \end{aligned}$$

i.e. $\boldsymbol{\mu}_n = \mathbb{E}(\mathbf{S}_{n,t})$ no depende de t. Ademas,

$$\mathbb{E}(\mathbf{S}_{n,t}\mathbf{S}_{n,t}') = \Sigma_1 + \boldsymbol{\mu}\boldsymbol{\mu}_{n-1}'A' + A\boldsymbol{\mu}_{n-1}\boldsymbol{\mu}' + B_{n-1} + A\mathbb{E}(\mathbf{S}_{n-1,t-1}\mathbf{S}_{n-1,t-1}') \tag{4.16}$$

donde $B_{n-1} \in M_{p\times p}(\mathbb{R})$

$$B_{n-1} = \begin{bmatrix} \sum_{i=1}^{p} \alpha_i(1-\alpha_i)\mathbb{E}(S_{n-i,t-i}) & \mathbf{0} \\ \mathbf{0} & \mathbf{0} \end{bmatrix}$$

Repitiendo este cálculo n veces se tiene que $\mathbb{E}(\mathbf{S}_{n,t}\mathbf{S}'_{n,t}) < \infty$.

$$\therefore S_{n,t} \in L^2(\Omega, \mathscr{F}, \mathbb{P})$$

Ahora se probará la existencia del proceso.

Nótese que

$$S_{n,t} - S_{n-1,t} = \sum_{t=1}^{p}(\alpha_i \circ S_{n-i,t-i} - \alpha_i \circ S_{n-i-1,t-i})$$

Entonces,

$$\mathbb{E}[(S_{n,t} - S_{n-1,t})^2]$$

$$\begin{aligned}
= & \sum_{t=1}^{p}(\alpha_i \circ \mathbb{E}[S_{n-i,t-i} - \alpha_i \circ S_{n-i-1,t-i})^2] \\
& +2 \sum_{1\le i\le j\le p} \mathbb{E}[(\alpha_i \circ S_{n-i,t-i} - \alpha_i \circ S_{n-i-1,t-i})(\alpha_j \circ S_{n-j,t-j} - \alpha_j \circ S_{n-j-1,t-i})] \\
= & \sum_{t=1}^{p} \alpha_i^2 \mathbb{E}[(S_{n-i,t-i} - S_{n-i-1,t-i})^2] + \sum_{t=1}^{p} \alpha_i(1-\alpha_i)\mathbb{E}[|S_{n-i,t-i} - S_{n-i-1,t-i}|] \\
& +2 \sum_{1\le i\le j\le p} \alpha_i\alpha_j |\mathbb{E}[(S_{n-i,t-i} - S_{n-i-1,t-i})(S_{n-j,t-j} - S_{n-j-1,t-j})] - \mathbb{E}[|S_{n-i,t-i} - S_{n-i-1,t-i}|]| \\
\le & \sum_{t=1}^{p} \alpha_j \mathbb{E}[|S_{n-j,t-j} - S_{n-j-1,t-j}|]
\end{aligned}$$

Y también para $j \in \{1, 2, \ldots, p-1\}$

$$\mathbb{E}[(S_{n,t} - S_{n-1,t})(S_{n-j,t-j} - S_{n-j-1,t-j})]$$

$$\begin{aligned}
= & \sum_{k=1}^{p} \mathbb{E}[(\alpha_k \circ S_{n-k,t-k} - \alpha_k \circ S_{n-k-1,t-k})(S_{n-j,t-j} - S_{n-j-1,t-j})] \\
= & \sum_{k=1}^{p} \alpha_k \mathbb{E}[(S_{n-k,t-k} - S_{n-k-1,t-k})(S_{n-j,t-j} - S_{n-j-1,t-j})] \\
= & \alpha_j \mathbb{E}[(S_{n-j,t-j} - S_{n-j-1,t-j})^2] + 2\sum_{k\ne j} \alpha_k \mathbb{E}[(S_{n-k,t-k} - S_{n-k-1,t-k})(S_{n-j,t-j} - S_{n-j-1,t-j})]
\end{aligned}$$

Defínase el vector columna de $2p + \frac{p(p-1)}{2}$ renglones

$$\mathbf{X}(n,t) := \begin{pmatrix} \mathbb{E}[(S_{n,t} - S_{n-1,t})^2] \\ \vdots \\ \mathbb{E}[(S_{n,t} - S_{n-1,t})^2] \\ \mathbb{E}[|S_{n,t} - S_{n-1,t}|] \\ \vdots \\ \mathbb{E}[|S_{n-p+1,t-p+1} - S_{n-p,t-p+1}|] \\ \mathbb{E}[(S_{n,t} - S_{n-1,t})(S_{n-1,t-1} - S_{n-2,t-1})] \\ \vdots \\ \mathbb{E}[(S_{n-p+2,t-p+2} - S_{n-p+1,t-p+2})(S_{n-p+1,t-p+1} - S_{n-p,t-p+1})] \end{pmatrix}$$

Se puede demostrar que

$$\mathbf{X}(n,t) \preceq D\mathbf{X}(n-1,t-1)$$

donde

$$D = \begin{bmatrix} A_1 & C & A_2 \\ 0 & A & 0 \\ A_3 & 0 & A_4 \end{bmatrix}$$

con

$$C = \begin{pmatrix} \alpha_1(1-\alpha_1) & \cdots & \alpha_p(1-\alpha_p) \\ 0 & \cdots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \cdots & 0 \end{pmatrix},$$

$$A = \begin{pmatrix} \alpha_1 & \alpha_2 & \cdots & \alpha_{p-1} & \alpha_p \\ 1 & 0 & \cdots & 0 & 0 \\ 0 & 0 & \ddots & 0 & 0 \\ 0 & 0 & \cdots & 1 & 0 \end{pmatrix},$$

$$A_1 = \begin{pmatrix} \alpha_1^2 & \alpha_2^2 & \cdots & \alpha_{p-1}^2 & \alpha_p^2 \\ 1 & 0 & \cdots & 0 & 0 \\ 0 & 0 & \ddots & 0 & 0 \\ 0 & 0 & \cdots & 1 & 0 \end{pmatrix},$$

$$A_2 = \begin{bmatrix} 2\alpha_1\alpha_2 \ldots 2\alpha_1\alpha_p 2\alpha_2\alpha_3 \ldots 2\alpha_{p-1}\alpha_p \\ \mathbf{0} \end{bmatrix},$$

$$A_3 = \begin{bmatrix} \alpha_1 & 0 & \cdots & 0 & 0 \\ 0 & \alpha_2 & \cdots & 0 & 0 \\ \vdots & \vdots & \ddots & \vdots & \vdots \\ 0 & 0 & 0 & \alpha_{p-1} & 0 \\ & & \mathbf{0} & & \end{bmatrix},$$

$$A_4 = \begin{pmatrix} \alpha_2 & \alpha_3 & \cdots & \alpha_{p-1} & \alpha_p & 0 & 0 & \cdots & 0 & 0 & 0 & 0 & \cdots & 0 & 0 & \cdots & 0 & 0 & 0 \\ \alpha_1 & 0 & \cdots & 0 & 0 & \alpha_3 & \alpha_4 & \cdots & \alpha_{p-1} & \alpha_p & 0 & 0 & \cdots & 0 & 0 & \cdots & 0 & 0 & 0 \\ 0 & \alpha_1 & \cdots & 0 & 0 & \alpha_2 & 0 & \cdots & 0 & 0 & \alpha_4 & \alpha_5 & \cdots & \alpha_{p-1} & \alpha_p & \cdots & 0 & 0 & 0 \\ \vdots & \vdots & \ddots & \vdots & \vdots & \vdots & \vdots & \ddots & \vdots & \vdots & \vdots & \vdots & \ddots & \vdots & \vdots & \ddots & \vdots & \vdots & \vdots \\ 0 & 0 & \cdots & \alpha_1 & 0 & 0 & 0 & \cdots & \alpha_2 & 0 & 0 & 0 & \cdots & \alpha_3 & 0 & \cdots & \alpha_{p-2} & 0 & \alpha_p \\ 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ \vdots & \vdots & \ddots & \vdots & \vdots & \vdots & \vdots & \ddots & \vdots & \vdots & \vdots & \vdots & \ddots & \vdots & \vdots & \ddots & \vdots & \vdots & \vdots \\ 0 & 0 & \cdots & 1 & 0 & 0 & 0 & \cdots & 0 & 0 & 0 & 0 & \cdots & 0 & 0 & \cdots & 0 & 0 & 0 \\ 0 & 0 & \cdots & 0 & 0 & 1 & 0 & \cdots & 0 & 0 & 0 & 0 & \cdots & 0 & 0 & \cdots & 0 & 0 & 0 \\ 0 & 0 & \cdots & 0 & 0 & 0 & 1 & \cdots & 0 & 0 & 0 & 0 & \cdots & 0 & 0 & \cdots & 0 & 0 & 0 \\ \vdots & \vdots & \ddots & \vdots & \vdots & \vdots & \vdots & \ddots & \vdots & \vdots & \vdots & \vdots & \ddots & \vdots & \vdots & \ddots & \vdots & \vdots & \vdots \\ 0 & 0 & \cdots & 0 & 0 & 0 & 0 & \cdots & 1 & 0 & 0 & 0 & \cdots & 0 & 0 & \cdots & 0 & 0 & 0 \\ \vdots & \vdots & \ddots & \vdots & \vdots & \vdots & \vdots & \ddots & \vdots & \vdots & \vdots & \vdots & \ddots & \vdots & \vdots & \ddots & \vdots & \vdots & \vdots \\ 0 & 0 & \cdots & 0 & 0 & 0 & 0 & \cdots & 0 & 0 & 0 & 0 & \cdots & 0 & 0 & \cdots & 0 & 0 & 0 \end{pmatrix}$$

Los eigenvalores de D son la unión de los eigenvalores de A y

$$B := \begin{bmatrix} A_1 & A_2 \\ A_3 & A_4 \end{bmatrix}$$

El polinomio característico de A es $z^p - \alpha_1 z^{p-1} - \ldots - \alpha_{p-1} z - \alpha_p$ así que los eigenvalores de A están dentro del círculo unitario por hipótesis. Después de una larga aritmética matricial se puede ver que B es una forma condensada del cuadrado de Kronecker de A, i.e. $B = A_{[2]}$. Por el lema 4.4.4 los eigenvalores de B están dentro del círculo unitario (pues producto de números dentro del círculo unitario vuelve a estar dentro del círculo unitario). Por tanto, los eigenvalores de D están dentro del círculo unitario.

Por el lema 4.4.3 se tiene que

$$\mathbf{X}(n,t) \preceq D\mathbf{X}(n-1,t-1) \preceq D^2\mathbf{X}(n-2,t-2) \preceq \ldots \preceq D^n\mathbf{X}(0,t-n)$$

y

$$\begin{aligned}\mathbf{X}(0,t-n) &= (\mathbb{E}(Z_{t-n}^2),0,\ldots,0,\mathbb{E}(Z_{t-n}),0,\ldots,0)' \\ &= (\sigma^2+\mu^2,0,\ldots,0,\mu,0,\ldots,0)'\end{aligned}$$

Sea $\lambda = \text{máx}\{|v| : v \text{ es eigenvalor de } D\}$. Para n suficientemente grande se tiene que

$$\mathbf{X}(n,t) \leq M\lambda^n(1,1,\ldots,1)'$$

donde M es una constante. Por tanto,

$$\mathbb{E}[(S_{n,t}-S_{n-1,t})^2] \leq M\lambda^n,$$

i.e.

$$\|S_{n,t}-S_{n-1,t}\| \leq \sqrt{M}\lambda^{n/2}$$

Y además

$$\begin{aligned}\|S_{n+k,t}-S_{n,t}\| &\leq \sum_{j=1}^{k}\|S_{n+j,t}-S_{n+j-1,t}\| \leq \sum_{j=1}^{k} M^{1/2}\lambda^{(n+j)/2} \\ &\leq M_1\lambda^{(n+1)/2} \overset{n\to\infty}{\to} 0\end{aligned}$$

Entonces, $S_{n,t}$ converge en L^2 (para t fija). Sea $X_t = \lim_{n\to\infty} S_{n,t}$. Por la definición de $S_{n,t}$ se tiene que

$$X_t = \sum_{j=1}^{p}\alpha_j \circ X_{t-j} + Z_t$$

Aplicando varias veces la ecuación (4.15) se tiene que

$$\mathbb{E}(\mathbf{S}_{n,t}\mathbf{S}'_{n-k,t-k}) = \boldsymbol{\mu}\boldsymbol{\mu}'_{n-k} + A\mathbb{E}(\mathbf{S}_{n-1,t-1}\mathbf{S}'_{n-k,t-k})$$

repitiendo este cálculo k veces, ya se mencionó que $\mathbb{E}(\mathbf{S}_{n,t}\mathbf{S}'_{n-k,t-k})$ no depende de t. Y como

$$\mathbb{E}(\mathbf{X}_t\mathbf{X}'_{t-k}) = \lim_{n\to\infty}\mathbb{E}(\mathbf{S}_{n,t}\mathbf{S}'_{n-k,t-k})$$

Entonces, $\mathbb{E}(\mathbf{X}_t\mathbf{X}'_{t-k})$ no depende de t.

$$\therefore \{X_t\} \text{ es estacionario}$$

También nótese que

$$\mathbb{E}(|S_{n,t}-S_{n-1,t}|) \leq \|S_{n,t}-S_{n-1,t}\| \leq M^{1/2}\lambda^{n/2} \quad (\lambda \in (0,1))$$

así que

$$\sum \mathbb{E}(|S_{n,t} - S_{n-1,t}|) < \infty$$

$$\therefore \sum S_{n,t} - S_{n-1,t} < \infty \ c.s.$$

Sea $Y_t = \sum S_{n,t} - S_{n-1,t} = \lim_{n\to\infty} S_{n,t}$. Entonces por el lema 4.4.3 se tiene que $X_t = Y_t$ *c.s.*, es no negativo y toma valores en los enteros, por tanto es X_t.

Para probar la unicidad supóngase que existe otra serie de tiempo estacionaria que satisface (4.13). Entonces

$$X_t - Y_t = \sum(\alpha_i \circ X_{t-i} - \alpha_i \circ Y_{t-i})$$

y con argumentos dados similares a los que se hicieron previemente se tiene que

$$\|X_t - Y_t\| \leq M\lambda^n \overset{n\to\infty}{\to} 0$$

Por tanto, $\|X_t - Y_t\| = 0$, i.e. $X_t = Y_t$.

Finalmente,

Como para $k > 0$, $Cov(S_{n-k,t-k}, Z_t) = 0$ y $Cov(X_{t-k}, Z_t) = 0$, entonces $Cov(X_{t-k}, Z_t) = \lim_{n\to\infty} Cov(S_{n-k,t-k}, Z_t) = 0$

$$\therefore Cov(X_s, Z_t) = 0, \quad s < t$$

□

4.4.2. Ergodicidad y momentos

Sean $\{Y(t)\}$ todas las series de conteo $\alpha_1 \circ X_{t-1} + \ldots + \alpha_p \circ X_{t-p} + \epsilon_t$ en (4.13). Evidentemente, $\{Y(t)\}$ es una serie de tiempo de dimensión p. De (4.13) se tiene que

$$\sigma(X_t, X_{t-1}, \ldots) \subset \sigma(Z_t, Y(t), Z_{t-1}, Y(t-1), \ldots)$$

y por tanto

$$\cap_{t=0}^{-\infty}\sigma(X_t, X_{t-1}, \ldots) \subset \cap_{t=0}^{-\infty}\sigma(Z_t, Y(t), Z_{t-1}, Y(t-1), \ldots) \tag{4.17}$$

el lado derecho de (4.17) es la cola de la σ-álgebra independiente de la sucesión aleatoria $\{Z_t, Y(t)\}$; la probabilidad de un evento en ésta es 0 ó 1, por tanto, la cola de la σ-álgebra de $\{X_t\}$ tiene sólo conjuntos medibles con probabilidad 0 ó 1, por Wang (1982) $\{X_t\}$ es ergódica.

Sea $\hat{\gamma}_k = \frac{1}{T}\sum_{t=1}^{T-k}(X_t - \bar{X})(X_{t+k} - \bar{X})$ la autocovarianza muestral de $\{X_t\}$ y $\hat{\rho}_k = \frac{\hat{\gamma}_k}{\hat{\gamma}_0}$ el coeficiente de autocorrelación muestral. Gracias a las propiedades de ergodicidad y estacionariedad de $\{X_t\}$, $\hat{\gamma}_k$ y $\hat{\rho}_k$ son fuertemente consistentes. Entonces, se tiene el siguiente teorema.

Teorema 4.4.3.

$\hat{\gamma}_k$ y $\hat{\rho}_k$ son fuertemente consistentes

Sea

$$\mathbf{X}_t = \begin{pmatrix} X_t \\ X_{t-1} \\ \vdots \\ X_{t-p+1} \end{pmatrix}$$

Entonces, (4.13) se puede escribir como

$$\mathbf{X}_t = A \circ \mathbf{X}_{t-1} + \mathbf{Z}_t$$

donde

$$A = \begin{pmatrix} \alpha_1 & \alpha_2 & \cdots & \alpha_{p-1} & \alpha_p \\ 1 & 0 & \cdots & 0 & 0 \\ 0 & 0 & \ddots & 0 & 0 \\ 0 & 0 & \cdots & 1 & 0 \end{pmatrix}$$

$$\mathbf{Z}_t = \begin{pmatrix} Z_t \\ 0 \\ \vdots \\ 0 \\ 0 \end{pmatrix}$$

Si

$$\boldsymbol{\gamma}_k := \mathbb{E}[(\mathbf{X}_t - \mathbb{E}(\mathbf{X}_t))(\mathbf{X}_{t-k} - \mathbb{E}(\mathbf{X}_{t-k}))']$$

entonces, para $k \in \mathbb{N}_+$

$$\begin{aligned} \boldsymbol{\gamma}_k &= \mathbb{E}(\mathbf{X}_t\mathbf{X}'_{t-k}) - \mathbb{E}(\mathbf{X}_t)\mathbb{E}(\mathbf{X}'_{t-k}) \\ &= A\mathbb{E}(\mathbf{X}_{t-1}\mathbf{X}'_{t-k}) + \mathbb{E}(\mathbf{Z}_t)\mathbb{E}(\mathbf{X}'_{t-k}) - \mathbb{E}(\mathbf{X}_t)\mathbb{E}(\mathbf{X}'_{t-k}) \\ &= A\boldsymbol{\gamma}_{k-1} + (A - I)\mathbb{E}(\mathbf{X}_t)\mathbb{E}(\mathbf{X}'_t) + \mathbb{E}(\mathbf{Z}_t)\mathbb{E}(\mathbf{X}'_t) \\ &= A\boldsymbol{\gamma}_{k-1} \end{aligned}$$

Y como $\mathbb{E}(\mathbf{X}_t) = A\mathbb{E}(\mathbf{X}_{t-1}) + \mathbb{E}(\mathbf{Z}_t)$, entonces

$$\gamma_X(k) = \alpha_1\gamma_X(k-1) + \alpha_2\gamma_X(k-2) + \ldots + \alpha_p\gamma_X(k) \tag{4.18}$$

ó equivalentemente

$$\rho_X(k) = \alpha_1\rho_X(k-1) + \alpha_2\rho_X(k-2) + \ldots + \alpha_p\rho_X(k) \tag{4.19}$$

donde $\gamma_X(k) = \mathbb{E}(X_tX_{t-k})$. Las ecuaciones (4.18) y (4.19) implican que las estructuras de correlación de los modelos $INAR(p)$ y $AR(p)$ son las mismas. De lo observado anteriormente se tiene que el modelo $INAR(p)$ es similar al modelo $AR(p)$ no sólo en su dominio estacionario sino también en su estructura de correlación. Por tanto, se pueden aplicar las metodologías de estimación desarrolladas para el modelo $AR(p)$ en el modelo $INAR(p)$.

4.4.3. Estimación de parámetros

4.4.3.1. Estimación Yule-Walker

Ya se dijo que el proceso $INAR(p)$ satisface

$$\rho_X(k) = \alpha_1\rho_X(k-1) + \alpha_2\rho_X(k-2) + \ldots + \alpha_p\rho_X(k)$$

Entonces, para $k = 1, 2, \ldots, p$ se tienen las ecuaciones de Yule-Walker

$$\boldsymbol{\Upsilon}\boldsymbol{\alpha} = \boldsymbol{\rho} \tag{4.20}$$

donde

$$\boldsymbol{\Upsilon} = [\rho_X(|i-j|)]_{p\times p}$$

,

$$\boldsymbol{\alpha} = (\alpha_1, \alpha_2, \ldots, \alpha_p)'$$

y

$$\boldsymbol{\rho} = (\rho_X(1), \rho_X(2), \ldots, \rho_X(p))'$$

Reemplazando $\rho_X(i)$ por $\hat{\rho}_X(i)$ en $\boldsymbol{\Upsilon}\boldsymbol{\alpha} = \boldsymbol{\rho}$ se obtiene el llamado estimador de Yule-Walker $\hat{\alpha}$ para α. Dicho estimador satisface

$$\hat{\boldsymbol{\Upsilon}}\hat{\boldsymbol{\alpha}} = \hat{\boldsymbol{\rho}} \tag{4.21}$$

Entonces, el estimador para $\hat{\mu}$ está dado por

$$\hat{\mu} = (1 - \hat{\alpha}_1 - \cdots - \hat{\alpha}_p)\bar{X} \tag{4.22}$$

y el estimador para σ^2 está dado por

$$\hat{\sigma}^2 = \frac{1}{T-p} \sum_{t=p+1}^{T} (\hat{Z}_t - \tilde{Z}_N)^2 \tag{4.23}$$

donde

$$\hat{Z}_t = X_t - \hat{\alpha}_1 X_{t-1} - \cdots - \hat{\alpha}_p X_{t-p}$$

$$\tilde{Z}_N = \frac{1}{T-p} \sum_{t=p+1}^{T} \hat{Z}_t$$

Teorema 4.4.4. *(Consistencia de los estimadores)*

$\hat{\alpha}_1, \ldots, \hat{\alpha}_p, \hat{\mu}$ y $\hat{\sigma}^2$ son fuertemente consistentes.

4.4.3.2. Estimación por mínimos cuadrados condicionales

Sean

$$\mathscr{F}_t := \sigma(X_1, \ldots, X_t)$$

$$\boldsymbol{\beta} = \begin{pmatrix} \mu \\ \alpha_1 \\ \vdots \\ \alpha_p \end{pmatrix} = \begin{pmatrix} \beta_1 \\ \beta_2 \\ \vdots \\ \beta_{p+1} \end{pmatrix}$$

$$\begin{aligned} g(\boldsymbol{\beta}, \mathscr{F}_t) &:= \mathbb{E}(X_t | \mathscr{F}_t) && (4.24) \\ &= \mu + \alpha_1 X_{t-1} + \ldots + \alpha_p X_{t-p} \quad \text{para } t > p && (4.25) \end{aligned}$$

$$Q_T(\boldsymbol{\beta}) = \sum_{t=p+1}^{T} (X_t - g(\boldsymbol{\beta}, \mathscr{F}_t))^2 \tag{4.26}$$

Se tomara a $\hat{\boldsymbol{\beta}}$ que minimice $Q_T(\boldsymbol{\beta})$ como estimador de $\boldsymbol{\beta}$, i.e.

$$Q_T(\hat{\boldsymbol{\beta}}) = \min_{\boldsymbol{\beta}} \{Q_T(\boldsymbol{\beta})\}$$

$\hat{\boldsymbol{\beta}}$ se conoce como el estimador por mínimos cuadrados condicionales. Como es costumbre en estimación por mínimos cuadrados, $\hat{\boldsymbol{\beta}}$ se puede encontrar resolviendo las ecuaciones

$$\frac{\partial}{\partial \mu} Q_T = 0$$

$$\frac{\partial}{\partial \alpha_k} Q_T = 0, \quad k = 1, 2, \ldots, p$$

Obteniéndose $\hat{\boldsymbol{\beta}} = (\hat{\mu}^*, \hat{\boldsymbol{\alpha}'}^*)'$, donde

$$\hat{\mu}^* = \bar{X}_T^{(0)} - \sum_{j=1}^{p} \hat{\alpha}_j^* \bar{X}_T^{(j)} \tag{4.27}$$

$$\hat{\boldsymbol{\Upsilon}}^* \hat{\boldsymbol{\alpha}}^* = \hat{\boldsymbol{\rho}}^* \tag{4.28}$$

con

$$\bar{X}_T^{(j)} = \frac{1}{T-p} \sum_{t=p+1}^{T} X_{t-j}$$

$$\hat{\gamma}_{k-j}^* = \frac{1}{T-p} \sum_{t=p+1}^{T} (X_{t-j} - \bar{X}_T^{(j)})(X_{t-1} - \bar{X}_T^{(k)})$$

$$\hat{\rho}_{k-j}^* = \frac{\hat{\gamma}_{k-j}^*}{\hat{\gamma}_0^*}$$

$$\hat{\boldsymbol{\Upsilon}}^* \in M_{p \times p}(\mathbb{R}), \ \hat{\boldsymbol{\Upsilon}}^* = [\hat{\rho}_{|i-j|}^*]$$

$$\boldsymbol{\rho}^* = \begin{pmatrix} \hat{\rho}_1^* \\ \hat{\rho}_2^* \\ \vdots \\ \hat{\rho}_p^* \end{pmatrix}$$

Es fácil ver de las ecuaciones (4.21), (4.22), (4.27) y (4.28) que cuando T es suficientemente grande, $\hat{\boldsymbol{\Upsilon}}^* - \hat{\boldsymbol{\Upsilon}}$, $\hat{\boldsymbol{\rho}}^* - \hat{\boldsymbol{\rho}}$ y $\bar{X}_T^{(j)} - \bar{X}$ son cercanos a cero y por tanto, $\hat{\boldsymbol{\alpha}}*$ y $\hat{\boldsymbol{\alpha}}$ son muy parecidos, así como $\hat{\mu}^*$ y $\hat{\mu}$. Es decir, los estimados de mínimos cuadrados condiocionales son muy parecidos a los dados en (4.21) y (4.22).

También se fácil ver que la función g definida en (4.24), $\frac{\partial}{\partial \beta_i} g$ y $\frac{\partial^2}{\partial \beta_i \partial \beta_j} g$ satisfacen todas las condiciones de regularidad propuestas por Klimko & Nelson (1978) y por lo tanto, las estimaciones por mínimos cuadrados condicionales son fuertemente consistentes.

Defínase para $t \geq p+1$,

$$g_t := g(\boldsymbol{\beta}, \mathscr{F}_t)$$

$$U_p(\boldsymbol{\beta}) = X_{p+1} - \alpha_1 X_p - \ldots - \alpha_p X_1 - \mu$$

De nuevo, por Klimko & Nelson (1978), los estimadores por mínimos cuadrados condicionales se distribuyen asintóticamente normal con varianza $V^{-1}WV^{-1}$, donde $W, V \in M_{(p+1)\times(p+1)}(\mathbb{R})$

$$W = \left[\mathbb{E}\left(U_p^2(\boldsymbol{\beta})\frac{\partial g_{p+1}}{\partial \beta_i}\frac{\partial g_{p+1}}{\partial \beta_j}\right)\right]$$

$$V = \left[\mathbb{E}\left(\frac{\partial g_{p+1}}{\partial \beta_i}\frac{\partial g_{p+1}}{\partial \beta_j}\right)\right]$$

Para $i, j \in 2, 3, \ldots,$

$$V_{ij} = \mathbb{E}\left(\frac{\partial g_{p+1}}{\partial \beta_i}\frac{\partial g_{p+1}}{\partial \beta_j}\right) = \mathbb{E}(X_{t-i}X_{t-j}) = \gamma_{i-j} + \mu_x^2$$

donde $\mu_x := \mathbb{E}(X_t)$ y $V_{i,j} = \mathbb{E}(X_{t-j}) = \mu_x$, por tanto

$$V = \begin{bmatrix} 1 & \mu_x \mathbf{1}' \\ \mu_x \mathbf{1} & \Gamma + \mu_x \mathbf{1}\mathbf{1}' \end{bmatrix}$$

donde, $\mathbf{1} \in M_{p\times 1}(\mathbb{R})$, $\mathbf{1} = (1, 1, \ldots, 1)'$ y $\Gamma \in M_{p\times p}(\mathbb{R})$ $\Gamma := [\gamma'_{i-j}]$. La inversa de V es de la forma

$$V^{-1} = \begin{bmatrix} 1 + \mu_x^2 \mathbf{1}'\Gamma^{-1}\mathbf{1} & -\mu_x \mathbf{1}'\Gamma^{-1} \\ -\mu_x \Gamma^{-1}\mathbf{1} & \Gamma^{-1} \end{bmatrix}$$

Y para $i, j \in \{2, \ldots, p+1\}$,

$$W_{ij} := \mathbb{E}\left(U_p^2(\boldsymbol{\beta})\frac{\partial g_{p+1}}{\partial \beta_i}\frac{\partial g_{p+1}}{\partial \beta_j}\right)$$

que mediante un argumento condicional se puede escribir de la forma

$$W_{ij} = \sum_{k=1}^{p} \alpha_k(1-\alpha_k)\mathbb{E}(X_{p-k+1}X_{p+2-i}X_{p+2-j}) + \sigma^2\mathbb{E}(X_{p+2-i}X_{p+2-j}) \quad (4.29)$$

$$= \sum_{k=1}^{p} \alpha_k(1-\alpha_k)\mathbb{E}(X_{p-k+1}X_{p+2-i}X_{p+2-j}) + \sigma^2 V_{ij} \quad (4.30)$$

Análogamente,

$$W_{1j} := \mathbb{E}\left(U_p^2(\boldsymbol{\beta})\frac{\partial g_{p+1}}{\partial \beta_1}\frac{\partial g_{p+1}}{\partial \beta_j}\right) \tag{4.31}$$

$$= \sum_{k=1}^{p}\alpha_k(1-\alpha_k)\mathbb{E}(X_{p-k+1}X_{p+1-j}) + \sigma^2 V_{1j} \tag{4.32}$$

De (4.29) y (4.31) se tiene que

$$W_{ij} = \sum_{k=1}^{p}\alpha_k(1-\alpha_k)\mathbb{E}(X_{p-k+1}X_{p+2-i}X_{p+2-j}) + \sigma^2 V_{ij}$$

para $i, j \in \{1, 2, \ldots, p+1\}$ con la convención de que $X_{p+1} := 1$. Entonces, se tiene el siguiente teorema.

Teorema 4.4.5. *(Consistencia del estimador por mínimos cuadrados condicionales)* $\hat{\boldsymbol{\beta}} = (\hat{\mu}^*, \hat{\boldsymbol{\alpha}}'^*)'$ es fuertemente consistente. Si $\mathbb{E}(Z_t^3) < \infty$, entonces $\sqrt{T-p}(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta})$ tiene distribución normal $(p+1)$-variada con vector de medias $\mathbf{0}$ y matriz de varianzas-covarianzas $V^{-1}WV^{-1}$

4.4.4. Predicción

De nuevo, considérese $\mathscr{F}_T = \sigma(X_1, \ldots, X_T)$ (la σ-álgebra generada por $X_1, \ldots, X_T$). Entonces, el predictor de mínima varianza, $X_T(1)$, de X_{T+1} es

$$\hat{X}_T(1) = \mathbb{E}(X_{T+1}|\mathscr{F}_T) = \alpha_1 X_T + \ldots + \alpha_p X_{T+1-p} + \mu$$

Análogamente se tiene que

$$\hat{X}_T(m) = \mathbb{E}(X_{T+m}|\mathscr{F}_T) = \sum_{j=1}^{p}\mathbb{E}(\alpha_j \circ X_{T+m-j}|\mathscr{F}_T) + \mu$$

Pero por la propiedad de torre de la esperanza condicional se tiene que

$$\begin{aligned}\mathbb{E}(\alpha_j \circ X_{T+m-j}|\mathscr{F}_T) &= \mathbb{E}[\mathbb{E}(\alpha_j \circ X_{T+m-j}|\mathscr{F}_T)|\mathscr{F}_T] \\ &= \alpha_j\mathbb{E}(X_{T+m-j}|\mathscr{F}_T) = \alpha_j\hat{X}_T(m-j)\end{aligned}$$

Por tanto,

$$\hat{X}_T(m) = \sum \alpha_j \hat{X}_T(m-j) + \mu$$

4.5. Cinco modelos de series de tiempo enteras basados en adelgazamiento: AR y MA discretos

A continuación, se discutirán cinco modelos observation-driven básicos de series de tiempo:

- $INAR(1)$
- $INMA(1)$
- Primera variante de $INAR(2)$
- Segunda variante de $INAR(2)$
- $INMA(2)$

4.5.1. El proceso $INAR(1)$

Ya en este capítulo y el anterior se definió el proceso $INAR(1)$, $\{X_t\}_{t\in\mathbb{Z}}$ mediante la ecuación en diferencias

$$X_t = \alpha \circ X_{t-1} + Z_t \tag{4.33}$$

donde $\alpha \in [0,1)$ (fijo) y $\{Z_t\}_{t\in\mathbb{Z}}$ es una sucesión de variables aleatorias discretas no-negativas independientes e idénticamente distribuidas tal que $\mathbb{E}(|Z_t|^2) < \infty$. También se supondrá que para todo $t \in \mathbb{Z}$ X_{t-1} y Z_t son independientes.

Como el proceso generado por la ecuación anterior es estacionario, cualquier efecto de condiciones iniciales no es importante.

Este proceso se parece mucho al $AR(1)$ Gaussiano, pero no en su carácter de modelo lineal debido a la operación "$\circ$" en lugar de la operación de multiplicación escalar en el modelo continuo.

El objetivo de la operación $\circ$ es asegurar que la serie tome valores enteros. Este operador ya se ha definido y se conoce como a*delgazamiento binomial* ú *operador de adelgazamiento* y se define como

$$\alpha \circ X_t := Y_{1,t-1} + Y_{2,t-1} + \ldots + Y_{X_{t-1},t-1} = \sum_{j=1}^{X_{t-1}} Y_{j,t-1} \tag{4.34}$$

donde $\{Y_{k,t-1}\}_{k\in\mathbb{Z}}$ son variables aleatorias independientes e idénticamente distribuidas con distribución $Bnlli(a)$, i.e.

$$\mathbb{P}(Y_{k,t-1}=1)=\alpha \quad \text{y} \quad \mathbb{P}(Y_{k,t-1}=0)=1-\alpha$$

Es importante notar que operaciones de adelgazamiento subsecuentes se realizan de forma independiente de cada otra, y de hecho, es una de las pocas oportunidades que se tiene de hacer explícita esta independencia con la notación $\{Y_{k,t}\}$ y que el adelgazamiento es una operación aleatoria con una distribución de probabilidad asociada, i.e. $\alpha\circ X|X=x\sim Bin(x,\alpha)$.

Aunque esta no es la manera formal de definir del concepto de adelgazamiento este es bien conocido en la literatura probabilística en partícular en la de procesos de ramificación de Bienaymé-Galton-Watson y también en la teoría de sumas detenidas.

El proceso $INAR(1)$ es parte de una clase de modelos, un poco más amplia, conocida como *modelos lineales condicionales autorregresivos de primer orden*, $CLAR(1)$, que se han estudiado en los trabajos de Grunwald et. al (2000). Y una de las "ventajas" de considerar esta clase $CLAR(1)$ es que implícitamente se está considerando $AR(1)$ Gaussiano (será muy claro despúes de que se dé la definición).

Definición 4.5.1. *(Procesos $CLAR(1)$)*

Para cualquier proceso $\{X_t\}_{t\in\mathbb{Z}}$, la estructura $CLAR(1)$ se define mediante

$$\eta(X_{t-1}):=\mathbb{E}(X_t|X_{t-1})=aX_{t-1}+\kappa$$

Proposición 4.5.1. *(Esperanza y varianza de los procesos $CLAR(1)$)*

Si $\{X_t\}_{t\in\mathbb{Z}}$ es un proceso con estructura $CLAR(1)$, entonces

(i) $\mathbb{E}(X_t)=\frac{\kappa}{1-a}$

(ii) $Var(X_t)=\frac{\kappa}{1-a}$

Demostración:

$$\begin{aligned}
\mathbb{E}(X_t) &= \mathbb{E}[\mathbb{E}(X_t|X_{t-1})] = \mathbb{E}(aX_{t-1}+\kappa) = a\mathbb{E}(X_{t-1})+\kappa \\
&= a\mathbb{E}[\mathbb{E}(X_{t-1}|X_{t-2})]+\kappa = a\mathbb{E}(aX_{t-2}+\kappa)+\kappa = a^2\mathbb{E}(X_{t-2})+a\kappa+\kappa \\
&= a^2\mathbb{E}[\mathbb{E}(X_{t-2}|X_{t-3})]+a\kappa+\kappa = a^2\mathbb{E}(aX_{t-3}+\kappa)+a\kappa+\kappa \\
&= a^3\mathbb{E}(X_{t-3})+a^2\kappa+a\kappa+\kappa = ... = a^k\mathbb{E}(X_{t-k})+\kappa\sum_{j=0}^{k-1}a^j = ... \\
&= \lim_{k\to 0} a^k\mathbb{E}(X_{t-k})+\kappa\sum_{j=0}^{\infty}a^j = 0+\kappa\frac{1}{1-a} = \frac{\kappa}{1-a}
\end{aligned}$$

□

Observación 4.5.1.

Nótese que esta proposición sugiere, sólo en términos de momentos, una similitud de los procesos $CLAR(1)$ y variables aleatorias independientes idénticamente distribuidas como $Poisson(\frac{\kappa}{1-a})$

▽

Alternativamente, una estructura estacionaria autorregresiva de primer orden se puede definir mediante una función de autocorrelación decreciente exponencialmente

$$\rho(k) = Corr(X_t, X_{t-k}) = a^{|k|}, \quad k \in \mathbb{Z} \tag{4.35}$$

Grunwald et. al.(2000) demostraron que bajo ciertas condiciones las ecuación (4.5.1) es implicada por una estructura $CLAR(1)$, pero la implicación inversa no necesariamente es cierta.

Como el proceso $INAR(1)$ tiene una estructura $CLAR(1)$ entonces su función de autocorrelación está dada por

$$\rho(k) = Corr(X_t, X_{t-k}) = a^{|k|}, \quad k \in \mathbb{Z}$$

de donde se puede notar que la correlación siempre es positiva.

Definición 4.5.2. *(Proceso de ramificación Bienaymé-Galton-Watson con inmigración)*

Un proceso de ramificación Bienaymé-Galton-Watson con inmigración se define mediante

$$X_t = \sum_{j=1}^{X_{t-1}} Y_{j,t-1} + W_t, \quad t \in \mathbb{Z}$$

donde $Y_{j,t-1}$ y W_t denotan un lattice independiente de variables aleatorias (lattice iid).

Nótese que si $\mathbb{E}(Y_{j,t-1}) < 1$ entonces el proceso $INAR(1)$ es equivalente a un proceso Bienaymé-Galton-Watson (directo de la definición).

Hasta el momento no se han hecho supuestos distribucionales en el proceso de serie de tiempo $INAR(1)$, excepto que $\{Z_t\}$ sean discretas. Ahora, se harán algunas suposiciones en este sentido. Una suposición natural para Z_t es que tiene distribución $Poisson(\lambda)$.

Al-Osh & Alzaid (1987) suponen que $Z_t \sim Poisson(\lambda)$ y se supondrá esto hasta que se especifique de otra forma.

Proposición 4.5.2. *(Innovaciones Poisson)*

Si $Z_t \sim Poisson(\lambda)$, entonces $X_t \sim Poisson(\frac{\lambda}{1-a})$

Notación 4.5.1.

A partir de este momento al proceso $INAR(1)$ tal que $W_t \sim Poisson(\lambda)$ se le denotará mediante $PoINAR(1)$ ▽

4.5.2. El proceso $INAR(2)$ primera variante

Si existe algún tipo de dependencia de orden mayor en los datos, no es adecuado el modelo $INAR(1)$ (que es Markoviano). A continuación, se puntualizará acerca de dos estructuras auto-regresivas de orden dos: la de Al-Osh & Alzaid y la de Du & Li. Dichas generalizaciones no necesariamente son directas y se presentan algunos inconvenientes que no ocurren en el caso $p = 1$. Ambas propuestas dependen de mecanismos de adelgazamiento (diferentes). El proceso $INAR(2)$ se define de la forma relativamente obvia

$$X_t = a_1 \circ X_{t-1} + a_2 \circ X_{t-2} + Z_t \tag{4.36}$$

Una de las propuestas es la de Alzaid & Al-Osh (1990) y es, en cierto sentido, una extensión directa del proceso $INAR(1)$. El otro modelo es que le se propone en los trabajos Du & Li (1991) y es un análogo más parecido al modelo auto-regresivo Gaussiano de orden alto.

En la especificación de Al-Osh & Alzaid (1990) del proceso $INAR(2)$, las operaciones de adelgazamiento se realizan de la siguiente forma: en cualquier periodo de tiempo, la población se "adelgaza" de acuerdo a dos operaciones $a_1\circ$ y $a_2\circ$, con los sobrevivientes, respectivamente, considerando partes de la población uno ó periodos atrás. El adelgazamiento cualquier tiempo t es independiente de la historia del pasada del proceso. Para asegurar la estacionariedad del proceso se tiene que cumplir que $a_1 + a_2 < 1$. Bajo innovaciones Poisson, el vector $(a_1 \circ X_{t-1}, a_2 \circ X_{t-1})$ dado $X_{t-1} = x$ tiene una distribución trinomial con parametros (a_1, a_2, x). Esta es una extensión natural multivariada de la suposición de que $\alpha \circ X_{t-1}|X_{t-1} = x \sim Bin(x, a)$ en el caso $PoINAR(1)$. Esta construcción tiene dos consecuencias importantes:

- El proceso sigue teniendo una interpretación física.
- Se induce una estructura de medios móviles de tal forma que el la función de autocorrelación sea similar a la del $ARMA(2,1)$ Gaussiano.

Esta especificación se denotará como $PoINAR(2) - AA$ ya que se tendrá una distribución marginal para X_t Poisson. En este caso se cumple que

$$\mathbb{E}(X_t) = Var(X_t) = \frac{\lambda}{1 - a_1 - a_2}$$

En este caso, la esperanza condicional de X_t dado X_{t-1} no es lineal en X_{t-1},

$$\mathbb{E}(X_t|\mathscr{F}_{t-1}) = \lambda\left[1 + \left(a_1 + \frac{a_1 a_2}{1 - a_1 - a_2}\right)\frac{p(y-1,z)}{p(y,z)} + \frac{a_2}{1 - a_1 - a_2}\frac{p(y,z-1)}{p(y,z)} + \frac{a_1^2}{1 - a_1 - a_2}\frac{p(y-1,z-1)}{p(y,z)}\right]$$

donde $\mathscr{F}_{t-1} = \{X_{t-1}, X_{t-2}, \ldots\}$ y

$$\begin{aligned} p(y,z) &= \mathbb{P}(X_{t-1} = y, X_{t-2} = z) \\ &= \exp\left(\lambda\frac{a_1 - 2}{1 - a_1 - a_2}\right)\sum_{i=0}^{\text{mín}\{y,z\}} \frac{(\lambda\frac{1-a_1}{1-a_1-a_2})^{y+z-2i}(\frac{\lambda a_1}{1-a_1-a_2})^i}{(y-i)!(z-i)!i!} \end{aligned}$$

i.e. esta especificación no es miembro de la clase $CLAR$. La función de autocorrelación satisface la ecuacion en diferencias de segundo orden.

$$\rho(k) = a_1 + \rho(k-1) + a_2\rho(k-2), \quad k \in 2, 3, \ldots \tag{4.37}$$

donde $\rho(0) = 1$ y $\rho(1) = a_1$.

Como en el caso del proceso $ARMA$ Gaussiano, el espacio parametral se puede particionar en dos secciones: una donde la función de autocorrelación decaiga exponencialmente y otra dos oscile antes de irse a cero (para $k \in \{2, 3, \ldots\}$). Para procesos en los que $a_2 < a_1 - a_1^2$, la función de autocorrelación decae exponencialmente a cero, si $a_2 > a_1 - a_1^2$ el comportamiento es oscilatorio. Y si $a_2 = a_1 - a_1^2$, las autocorrelaciones de primer y segundo orden son iguales.

4.5.3. El proceso $INAR(2)$ segunda variante

En la especificación del proceso $INAR(2)$ de Du & Li (1991), que se denotará como $INAR(2) - DL$ tiene la especificación

$$X_t = a_1 \circ X_{t-1} + a_2 \circ X_{t-2} + Z_t$$

La principal diferencia es cómo se llevan a cabo las operaciones de adelgazamiento $a_1 \circ X_{t-1}$ y $a_2 \circ X_{t-2}$. En esta especificación, cada conteo en un periodo dado t se adelgaza dos veces: una vez por $a_1\circ$ al tiempo $t + 1$; y subsecuentemente por $a_2\circ$ al tiempo $t + 2$.

El proceso se puede interpretar como un caso especial proceso de ramificación multitipo con inmigración.

En este caso también se cumple que

$$\mathbb{E}(X_t) = \frac{\lambda}{1 - a_1 - a_2}$$

y también se cumple que para $a_1, a_2 \in [0, 1)$, el proceso $INAR(2) - DL$ es estacionario siempre que $a_1 + a_2 < 1$. También tiene propiedades de correlación idénticas a las del modelo $AR(2)$ Gaussiano. Esta quizá es la diferencia más importante con el proceso $INAR(2) - AA$.

Una diferencia adicional entre las variantes AA y DL es el hecho de que en la especificación $INAR(2) - AA$ la esperanza condicional de X_t dado X_{t-1} (la regresión de X_t sobre X_{t-1}) no es lineal. Pero en el $INAR(2) - DL$ se tiene

$$\mathbb{E}(X_t | \mathscr{F}_{t-1}) = a_1 X_{t-1} + a_2 X_{t-2} + \lambda,$$

donde $\mathscr{F}_t = X_{t-1}, X_{t-2}, \ldots$. Esta condición define lo que se conoce como la familia $CLAR(2)$.

Y también, la distribución marginal de X_t no es Poisson, aunque la innovación sea Poisson.

4.5.4. El proceso $INMA(1)$

Los modelos $INAR(1)$ e $INAR(2)$ (ambas variantes) suponen que el proceso $\{X_t\}$ se puede representar satisfactoriamente en términos de un número finito de valores pasados del proceso mismo. En la literatura de series de tiempo con espacio de estados continuo, un teorema importante que motiva el uso de las clases de modelos AR y MA es el Teorema de descomposición de Wold para series de tiempo estacionarias (Véase capítulo 2). Un origen inocente de los modelos auto-regresivos es el de la representación mediante promedios móviles de orden infinito.

Supóngase que se propone un modelo $INAR(p)$, $p \in \{1, 2\}$, para un conjunto de datos (donde, en principio se supondrá que el $INAR(2)$ es el propuesto por Al-Osh & Alzaid). En la sección anterior se demostró que X_t tiene la misma distribución que

$$\sum_{i=1}^{\infty} c_j \circ Z_{t-j} + Z_t$$

donde $c_j = a^j$ si $p = 1$ y $c_1 = a_1$, $c_j = a_1 c_{j-1} + a_2 c_{j-2}, j = 2, 3, \ldots$ si $p = 2$. Con esto se obtiene un representación de medias móviles de orden infinito (en términos del adelgazamiento binomial) para el ampliamente estudiado proceso $INAR$.

Un tipo diferente de estructura de dependencia se puede obtener a partir de los procesos $INAR$ truncando la representación $INMA(\infty)$. La representación más simple es usar un proceso de medias móviles de orden 1 ($INMA(1)$).

Definición 4.5.3. *(Proceso $INMA(1)$)*

Se dice que un proceso de series de tiempo $\{X_t\}_{t\in\mathbb{Z}}$ es un proceso entero de medias móviles de orden 1, $INMA(1)$, si

$$X_t = b \circ Z_{t-1} + Z_t, \quad t \in \mathbb{Z} \tag{4.38}$$

donde $b \in [0, 1]$ y $\{Z_t\}_{t\in\mathbb{Z}}$ es una sucesión de variables aleatorias discretas no-negativas independientes e idénticamente distribuidas tal que $\mathbb{E}(|Z_t|^2) < \infty$.

Observación 4.5.2.

Aquí el operador de adelgazamiento $b \circ Z_{t-1}$ tiene la forma

$$b \circ Z_t := Y_{1,t-1} + Y_{2,t-1} + \ldots + Y_{Z_{t-1},t-1} = \sum_{j=1}^{Z_{t-1}} Y_{j,t-1}$$

donde $\{Y_{k,t-1}\}_{k\in\mathbb{Z}}$ son variables aleatorias independientes e idénticamente distribuidas con distribución $Bnlli(b)$, i.e.

$$\mathbb{P}(Y_{k,t-1} = 1) = b \;\text{ y }\; \mathbb{P}(Y_{k,t-1} = 0) = 1 - b$$

∇

Este modelo corresponde a un sistema de colas a tiempo continuo $M/D/\infty$ que se observa en puntos equidistantes de tiempo. Incluso se le puede dar una interpretación física a este proceso $INMA(1)$ considerando a X_t como el número de partículas en un cierto espacio (bien definido) al tiempo t. Este número se forma de las particulas que entraron durante $(t-1,t]$ y los sobrevivientes (representados por el operador de adelgazamiento) que entraron durante $(t-2,t-1]$. Cada partícula tiene la misma probabilidad de sobrevivir b. En comparación con el proceso $INAR(1)$, el adelgazamiento sólo actua en los inmigrantes al tiempo $t-1$, NO entre todas las particulas presentes en el espacio. Todas las partículas tienen una vida máxima de dos periodos y esto asegura que sólo las observaciones adyacentes a X_t estén correlacionadas. El proceso resultante $(\{X_t\}_{t\in\mathbb{Z}})$ no es ni Markoviano ni un proceso de Bienaymé-Galton-Watson con inmigración.

Nótese que $Var(X_t) = Var(b \circ Z_{t-1}) + Var(Z_t)$ (pues $b \circ Z_{t-1}$ es independiente de Z_t) Pero

$$\begin{aligned} Var(b \circ Z_{t-1}) &= Var(\mathbb{E}(b \circ Z_{t-1}|Z_{t-1})) + \mathbb{E}(Var(b \circ Z_{t-1}|Z_{t-1})) \\ &= Var\left(\mathbb{E}\left(\sum_{j=1}^{Z_{t-1}} Y_{j,t-1}|Z_{t-1}\right)\right) + \mathbb{E}\left(Var\left(\sum_{j=1}^{Z_{t-1}} Y_{j,t-1}|Z_{t-1}\right)\right) \\ &= Var(Z_{t-1}b) + +\mathbb{E}(Z_{t-1}b(1-b)) = b^2\sigma_Z^2 + b(1-b)\mu_Z \end{aligned}$$

Entonces,

$$Var(X_t) = b^2\sigma_Z^2 + b(1-b)\mu_Z + Var(Z_t) = b^2\sigma_Z^2 + b(1-b)\mu_Z + \sigma_Z^2 = (1+b^2)\sigma_Z^2 + b(1-b)\mu_Z$$

Y también

$$\begin{aligned} Cov(X_t, X_{t+1}) &= Cov(b \circ Z_{t-1} + Z_t, b \circ Z_t + Z_{t+1}) \\ &= Cov(b \circ Z_{t-1}, b \circ Z_t) + Cov(b \circ Z_{t-1}, Z_{t+1}) + Cov(Z_t, b \circ Z_t) + Cov(Z_t, Z_{t+1}) \\ &= Cov(b \circ Z_{t-1}, b \circ Z_t) + Cov(Z_t, b \circ Z_t) \end{aligned}$$

Entonces la estructura de dependencia se resume mediante su función de autocorrelación

$$\rho_X(k) = \begin{cases} \frac{b\sigma_Z^2}{b(1-b)\mu_Z+(1+b^2)\sigma_Z^2} & \text{si } |k| = 1 \\ 0 & \text{si } |k| \neq 1 \end{cases}$$

que es análoga al proceso $MA(1)$ Gaussiano. Se puede probar que si $b \in [0,1]$, entonces $\rho(1) \in [0, 0.5]$. De nuevo, un primer candidato para la distribución marginal del proceso $INMA(1)$ es la distribución Poisson.

Proposición 4.5.3. *(Innovaciones Poisson)*

Si $Z_t \sim Poisson(\lambda)$, entonces $X_t \sim Poisson(\lambda(1+b))$ y dicho proceso se conoce como $PoINMA(1)$

Es decir, bajo inovaciones Poisson, la media y la varianza de X_t está dada por

$$\mathbb{E}(X_t) = \lambda(1+b) = Var(X_t)$$

$\rho_X(1) = \frac{b}{1+b}$ y para $k \in 2, 3, \ldots$ $\rho_X(k) = 0$.

Observación 4.5.3.

Nótese que

$$\begin{aligned} \mathbb{E}(X_t|X_{t-1} = x) &= \mathbb{E}\left(\sum_{i=1}^{X_{t-1}} Y_{i,t-1}|X_{t-1} = x\right) + \mathbb{E}(Z_t|X_{t-1} = x) \\ &= xb + \mathbb{E}(Z_t) \text{ pues } X_{t-1} \text{ es independiente de } Y_{i,t-1} \text{ y } Z_t \end{aligned}$$

$$\therefore \mathbb{E}(X_t|X_{t-1}) = bX_{t-1} + \mathbb{E}(Z_t)$$

Y análogamente, se puede probar que $Var(X_t|X_{t-1}) = bX_{t-1} + \mathbb{E}(Z_t)$.

Es decir, la esperanza condicional de X_t (regresión de X_t sobre X_{t-1}) es función lineal de X_{t-1} y por tanto el modelo $INMA(1)$ es un miembro de la familia $CLAR(1)$.

4.5.5. El proceso $INMA(2)$

Una extensión natural del proceso $INMA(1)$, es el modelo $INMA(2)$.

Definición 4.5.4. *(Proceso $INMA(2)$)*

Se dice que un proceso de series de tiempo $\{X_t\}_{t\in\mathbb{Z}}$ es un proceso entero de medias móviles de orden 2, $INMA(2)$, si

$$X_t = b_1 \circ Z_{t-1} + b_2 \circ Z_{t-2} + Z_t, \quad t \in \mathbb{Z} \tag{4.39}$$

donde $b_1, b_2 \in [0,1]$, $\{Z_t\}_{t\in\mathbb{Z}}$ es una sucesión de variables aleatorias discretas no-negativas independientes e idénticamente distribuidas tal que $\mathbb{E}(|Z_t|^2) < \infty$ y las operaciones $b_1 \circ Z_{t-1}$ y $b_2 \circ Z_{t-2}$ se realizan independietemente.

La estructura de dependencia se resume mediante su función de autocorrelación

$$\rho_X(k) = \begin{cases} \frac{\sum_{j=1}^{2-k} b_j b_{j+k}}{1+b_1+b_2} & \text{si } |k| = 1,2 \\ 0 & \text{si } |k| > 2 \end{cases}$$

con $b_0 := 1$. Nótese que también tiene la propiedad de "corte" o anulación análoga a la del $MA(2)$ Gaussiano.

Con innovaciones Poisson, la distribición marginal de modelo Poisson resultante $PoINMA(2)$ es $X_t \sim Poisson(\lambda(1 + b_1 + b_2))$. También se puede dar una interpretación de proceso $INMA(2)$ como una generalización de los sistemas $M/D/\infty$.

Al-Osh & Alzaid (1988) propuesieron otra version del $INMA(2)$ conservando la suposición de independencia entre las operaciones de adelgazamiento de la definición anterior pero estableciendo un estructura de independencia entre los adelgazamientos de los términos $b_1 \circ Z_t$ y $b_2 \circ X_t$, pero con estructura de esperanza y varianza condicionales como las anteriores, resultando en un proceso con un ACF ligeralemente diferente (Brännas & Hall (2001)). Sin embargo, esta especificación no ofrece ventajas significactivas para las aplicaciones.

4.5.6. Estimación de parámetros de interés

A continuación se discutirá acerca de un conjunto de parámetros analizados por Jung *et al.* (2005) para la media de la innovación y el parámetro de dependencia en un modelo $INAR(1)$. Dichos estimadores se obtuvieron por los métodos de momentos, de mínimos cuadrados y por máxima verosimilitud. Según Jung *et al.* (2005), los estimadores de Yule-Walker y los de mínimos cuadrados condicionales funcionan mejor (aún en cuando haya sobredisperción). En muestras pequeñas es recomedable incorporar algún factor de corrección de sesgo, aunque en muestras mayores a 500 su efecto es insignificante. Debido a

la dificultad analítica de la función de verosimilitud (distribuciones condicionales analíticamente complicadas), particularmente en modelos de orden grande, se centrará más la atención en los estimadores por momentos que se basan en la función de autocorrelación parcial muestral.

4.5.6.1. Estimadores por momentos

Para el modelo $PoINAR(1)$, Jung *et al.* (2005) trabajaron en el estimador de Yule-Walker para a, que se denotará por $\hat{a}_{YW}$, es simplemente la autocorrelación muestral de primer orden y el correspondiente para λ es $\hat{\lambda}_{YW} = (1-\hat{a}_{YW})\bar{X}$. También se puede utilizar un método que se conoce como el *método de momentos generalizado* propuesto, para series de tiempo, por Harris (1999) y Brännäs (1994). En principio, se pueden utilizar momentos no condicionales ó condicionales para estimación por momentos generalizada, aunque los momentos condicionales cumplen las restricciones de forma más satisfactorias que sus contrapartes no condicionales.

Para las dos especificaciones del $INAR(2)$, se puede obtener los estimadores de Yule-Walker a partir de las autocorrelaciones muestrales de primer y de segundo orden

$$\hat{\rho}(k) = \frac{(T-k)^{-1}\sum_{t=k+1}^{T}(X_t-\bar{X})(X_{t-k}-\bar{X})}{T^{-1}\sum_{t=1}^{T}(X_t-\bar{X})^2}, \quad k=1,2.$$

Para el modelo $PoINAR(2)-AA$, a partir de la ecuación $\rho(k) = a_1\rho(k-1) + a_2\rho(k-2), \quad k \geq 2$, se obtiene que los estimadores de Yule-Walker para a_1 y a_2 están dados por

$$\hat{a}_{1|YWAA} = \hat{\rho}(1)$$

y

$$\hat{a}_{2|YWAA} = \hat{\rho}(2) - \hat{\rho}^2(1)$$

El estimador de Yule-Walker para λ se puede obtener a partir de la media no condicional del $PoINAR(2)-AA$ y éste está dado por

$$\hat{\lambda}_{YWAA} = (1-\hat{a}_{1|YWAA}-\hat{a}_{2|YWAA})\bar{X}$$

Los correspondientes estimadores de Yule-Walker para el modelo $INAR(2)-DL$ se pueden obtener recordando que las estructura de correlación de esta variante de $INAR(2)$ es parecida a la del $AR(2)$ Gaussiano. Los estimadores de Yule-Walker para esta variante son

$$\hat{a}_{1|YWDL} = \hat{\rho}(1)\left[\frac{1-\hat{\rho}(2)}{1-\hat{\rho}^2(1)}\right]$$

y

$$\hat{a}_{2|YWDL} = \frac{\hat{\rho}(2) - \hat{\rho}^2(1)}{1 - \hat{\rho}^2(1)} = 1 - \frac{1 - \hat{\rho}(2)}{1 - \hat{\rho}^2(1)}$$

El estimador de Yule-Walker para λ está dado por

$$\hat{\lambda}_{YWDL} = (1 - \hat{a}_{1|YWDL} - \hat{a}_{2|YWDL})\bar{X}$$

Du & Li (1991) estudiaron propiedades asintóticas de estos estimadores, entre ellas su matriz de varianzas-covarianzas asintóticas.

Los estimadores de Yule-Walker para los parámetros del proceso $PoINMA(1)$ se pueden obtener directamente (aritméticamente) de los momentos no condicionales

$$\hat{b}_{YW} = \frac{\hat{\rho}(1)}{1 - \hat{\rho}(1)}$$

y

$$\hat{\lambda}_{YW} = \frac{\bar{X}}{1 + \hat{b}_{YW}}$$

Y lo mismo sucede para los estimadores de Yule-Walker para los parámetros del proceso $PoINMA(2)$

$$\hat{b}_{1|YW} = \frac{\hat{\rho}(1) - 1 + \sqrt{(1 - \hat{\rho}(1))^2 + 4\hat{\rho}(1)\hat{\rho}(2)}}{2\hat{\rho}(1)},$$

$$\hat{b}_{2|YW} = \frac{\hat{\rho}(1) + 2\hat{\rho}(2) - 1 + \sqrt{(1 - \hat{\rho}(1))^2 + 4\hat{\rho}(1)\hat{\rho}(2)}}{2(1 - \hat{\rho}(2))}$$

y

$$\hat{\lambda}_{YW} = \frac{\bar{X}}{1 + \hat{b}_{1|YW} + \hat{b}_{2|YW}}$$

4.5.6.2. Otros estimadores

Los estimadores por mínimos cuadrados condicionales de Klimko & Nelson (1978) para los parámetros a y λ en el modelo $PoINAR(1)$ tienen la siguiente forma

$$\hat{a}_{CLS} = \frac{(T-1)\sum_{t=2}^{T} X_t X_{t-1} - \left(\sum_{t=2}^{T} X_t\right)\left(\sum_{t=2}^{T} X_{t-1}\right)}{(T-1)\sum_{t=2}^{T} X_{t-1}^2 - \left(\sum_{t=2}^{T} X_{t-1}\right)^2} \tag{4.40}$$

y

$$\hat{\lambda}_{CLS} = \frac{\sum_{t=2}^{T} X_t - \hat{a}_{CLS} \sum_{t=2}^{T} X_{t-1}}{T-1} \tag{4.41}$$

Como el modelo $INMA(1)$ también pertenece a la clase $CLAR(1)$, un estimador para b también se puede basar en el lado derecho de la ecuación (4.40) obteniéndose

$$b_{CLS} = \frac{\hat{a}_{CLS}}{1 - \hat{a}_{CLS}}.$$

La media de innovación también se estima mediante la ecuación (4.41). En el modelo $INAR(2) - DL$ también se pueden obtener estimadores mediante metodologías de mínimos cuadrados utilizando el hecho de que dicho modelo pertenece a la clase $CLAR$ pero no se analizará aquí.

La estimación máximo verosímil para el modelo $PoINAR(1)$ se puede hacer directamente (como en Hamilton (1994)), sin embargo se prefiere el método de estimación máximo verosímil condicional en el cual se considera que X_0 está dada. Bajo la suposición de que X_t tiene distribución Poisson, Al-Osh & Alzaid (1987) escribieron la verosimilitud condicional para el modelo $PoINAR(1)$ y ésta es de la forma

$$\ell_c(a, \lambda | X_1, \ldots, X_T) = -(T-1)\lambda + \sum_{t=2}^{T} Ln(C(X_{t-1}, X_t)) \tag{4.42}$$

donde

$$C(x_{t-1}, x_t) = \sum_{k=0}^{m} \frac{a^k (1-a)^{x_{t-1}-k} \lambda^{x_t - k}}{k!(x_t - k)!(x_{t-1} - k)!} \tag{4.43}$$

y $m = \text{mín}\{x_t, x_{t-1}\}$.

Para obtener estimadores máximo verosímiles se utilizan métodos numéricos. Suena razonable que este método se comporté con estabilidad en modelos de pocos parámetros aún cuando se tengan muestras de tamaño pequeño y por supuesto se complica computacionalmente en dimensiones mayores.

La estimación máximo verosímil es factible en el contexto del modelo $PoINMA(1)$ esto se puede ver partir del hecho de que la densidad condicional $p(x_t|x_{t-1})$ en el $PoINAR(1)$ y en el $PoINMA(1)$ es la misma (McKenzie (1988)). Por tanto, utilizando la relación $a = \frac{b}{1+b}$, la función de verosímilitud del modelo $PoINMA(1)$ es de la misma forma que la de las ecuaciones (4.42) y (4.43) y por tanto, se puede obtener el estimador máximo verosímil de b utilizando argumentos de invarianza de los estimadores máximo verosímiles sobre a.

La estimación en modelos $INAR$ e $INMA$ con órdenes de correlación y medias móviles (respectivamente) grandes no es tan directa ya que sea deben calcular funciones de distribución condicionales con forma funcional complicada incluso con innovaciones Poisson, y por supuesto se complican aún más las rutinas de máximización numérica.

4.5.7. Observaciones acerca de la estimación de parámetros

Una vez que se han obtenido estimadores para los parámetros, es evidente la necesidad de obtener estimaciones adecuadas de su precisión. Los errores estándar asintóticos para los estimadores de Yule-Walker en el modelo $INAR(1)$ se pueden obtener mediante los métodos de Heyde & Seneta (1972), específicamente aquellas expresiones estructuralmente equivalente al modelo de Bienaymé-Galton-Watson con inmigración. El método no se puede generalizar fácilmente a estimadores Yule-Walker para dimensiones más grandes (como por ejemplo el $INAR(2)$). Los errores asintóticos estándar para los estimadores CLS (de mínimos cuadrados condicionales) en el $INAR(1)$ se basan en resultados de Klimko & Nelson (1978). Los estimadores Yule-Walker y CLS en este modelo tienen el mismo comportamiento asintótico. En el caso de estimadores máximo verosímiles, las varianzas de los estimadores se pueden obtener de la forma usual i.e. mediante el Hessiano ó los resultados asintóticos de la matriz de información. En el caso del $PoINMA(1)$, la varianza asintótico de b_{ML} se puede obtener directamente de $Var(\hat{a}_{MV})$ mediante la ecuación $Var(\hat{a}_{MV}) = Var(\hat{b}_{MV})(1 + \hat{b}_{MV})^2$. Las precisiones para los estimadores de momentos generalizados se pueden obtener a partir de la matriz de varianzas-covarianzas asintótica (Newey & West (1987)). Una advertencia para el uso de estos errores estándar asintóticos convencionales (que se pueden usar en estimadores máximo verosímiles basadas en la suposición Poisson) es que son muy sensibles ante la presencia de sobredispersión . Esto se puede evitar utilizando métodos de bootstrap, particularmente útiles en modelos de segundo orden.

En la siguiente tabla se muestra algunos de los estimadores para a λ y su respectivo error asintótico estádar en la serie de tiempo Fürth para el modelo $PoINAR(1)$.

Parámetro de adelgazamiento		Media de las innovaciones	
Parámetro	Error asintótico	Parámetro	Error asintótico
$\hat{a}_{YW} = 0.666$	0.037	$\lambda_{YW} = 0.532$	0.062
$\hat{a}_{CLS} = 0.665$	0.037	$\lambda_{CLS} = 0.540$	0.062
$\hat{a}_{MVC} = 0.688$	0.023	$\lambda_{MVC} = 0.532$	0.041
$\hat{a}_{MV} = 0.686$	0.023	$\lambda_{MV} = 0.494$	0.041
$\hat{a}_{YW} = 0.707$	0.024	$\lambda_{YW} = 0.488$	0.044

Estos resultados para los estimadores de Yule-Walker son idénticos a los que obtuvieron Mills & Seneta (1989). Dichos autores afirman que hay un nivel ineficiente para el ajuste

en los datos. Si se revisa la autocorrelación residual muestral y las autocorrelaciones parciales muestrales de los residuales se observa presencia de dependencia después de ajustar el modelo $INAR(1)$, confirmando así, la afirmación.

Cuando se intenta hacer un ajuste de un modelo $INMA(1)$, se presentan discrepancias inmediatamente ya que la autocorrelación de primer orden de los datos Fürth superan en un medio a su contraparte teórica; por tanto, no hay un parámetro factible para b y ya no se considerará para este conjunto de datos (el cálculo del estimador CLS de b es 1.985, hecho que violenta la definición de adelgazamiento binomial). Lo mismo ocurre cuando se considera el modelo $INMA(2)$, obténiéndose estimadores de Yule-Walker para b_1 de 1.008, de 0.961 para b_2 y de 0.536 para λ; de nuevo, el estimado para b_1 violenta la definición de operador binomial, por tanto no se considerará este modelo.

Con la especificación del $INAR(2)$ de Alzaid & Al-Osh (1990) se tiene que $\hat{a}_{1|YWAA} = 0.664$, $\hat{a}_{2|YWAA} = -0.119$ y $\lambda_{YWAA} = 0.723$. Los correspondientes estimados para el modelo $INAR(2) - DL$ son $\hat{a}_{1|YWAA} = 0.808$, $\hat{a}_{2|YWAA} = -0.214$ y $\lambda_{YWAA} = 0.646$. Estos resultados exhiben que ninguna de las variantes $INAR(2)$ es satisfactoria para este conjunto de datos, esto se debe a que ambas estimaciones de a_2 son negativas. Estos resultados exhiben una restricción para usar las especificaciones de las dos variantes del modelo $INAR(2)$. De las expresiones $\hat{a}_{2|YWAA} = \hat{\rho}(2) - \hat{\rho}^2(1)$ y $\hat{a}_{2|YWDL} = \frac{\hat{\rho}(2)-\hat{\rho}^2(1)}{1-\hat{\rho}^2(1)}$, es evidente que para que el estimador de a_2 haga sentido debe ocurrir que $\hat{\rho}(2) > \hat{\rho}^2(1)$ y no sólo eso, la suma de los parámetros de adelgazamiento debe ser menor que uno. En el caso de la serie de tiempo Fürth se tiene que $\hat{\rho}(2) - \hat{\rho}^2(1) = -0.119$.

4.5.8. Pruebas para independencia serial

En la literatura existen varias pruebas paramétricas y no-paramétricas para probar la presencia de dependecia serial en una muestra de conteo $(x_1, \ldots, x_t)$ indexada con el tiempo, es decir, una serie de tiempo de conteo. Muchas pruebas no-paramétricas fallan al tomar en cuenta la naturaleza discreta de los datos y pueden ser ineficientes debido a la presencia de múltiples empates (recuéredese que se tiene en mente series de tiempo enteras de baja frecuencia). La idea es centrar la atención en pruebas diseñadas para series de tiempo de conteo, contextualizándolas en la mayoría de los casos en modelos basados en adelgazamiento binomial.

La primera prueba que se considerará es la ya conocida: prueba de rachas (runs). En esta prueba se tiene que "dicotomizar" la serie original según algún criterio. Es muy común (y de hecho algunos autores lo recomiendan) utilizar la mediana para hacer esta división; se eliminarán aquellas observaciones de la muestra que sean idénticas a la mediana muestral. Sin embargo, en modelos estacionarios de series de tiempo entera de baja frecuencia, es común observar la mediana y por tanto se desecharán muchas observaciones afectando

considerablemente la potencia de la prueba, y por tanto se preferirá usar la media muestral para hacer la dicotomización (Gibbons & Chakraborti (1992)) ya que incluso con datos enteros, ésta difícilmente será un número entero y por tanto no se perderá información muestral en este proceso de dicotomización.

Si la hipótesis nula H_0 es no hay dependencia serial, entonces bajo la hipótesis nula la distribución del número de rachas se puede obtener con argumentos de combinatoria. Para una observación de la serie de conteo $x_1, \ldots, x_T$, el estadístico de prueba, que se denota por Z, tiene la forma

$$Z = \frac{R - 1 - (2T_1(T - T_1))/T}{\sqrt{2T_1(T - T_1)[2T_1(T - T_1) - T]/T^2(T - 1)}}$$

donde R es el número de rachas y T_1 es número de observaciones que se retiraron por coincidir con la variable de criterio de dicotomización. Los niveles de significancia convencionales sólo se pueden alcanzar a través de un diseño de prueba aleatorizada.

Wald & Wolfowitz (1940) probaron que el estadística de prueba (que se denotó convenientemente por Z) converge en distribución a una variable aleatoria normal estándar bajo la hipótesis nula de independencia de la serie de conteo (observable). Ya que la distribución del número de rachas es discreta se recomienda utilizar un factor de corrección para esta aproximación, se denotará al correspondiente estadístico de prueba como Z_{cc} y dicha estadística es asintónticamente equivalente a Z.

El diseño de prueba que se trabajará aquí es de una cola (one-sided). Esto esta motivado por el hecho de que bajo la hipótesis alternativa los procesos $INAR(1)$, $INAR(2)$ e $INMA(1)$ tienen autocorrelación positiva. Por tanto, bajo H_0 se espera que el número de rachas sea pequeño. En la práctica, no se usará la clase de modelos $INARMA$ si la autocovarianza muestral de primer orden es negativa. Por tanto, se rechaza la hipótesis nula (hay independencia serial) si $Z < z_\alpha$ (o alternativamente si $Z_{cc} < z_\alpha$), donde z_α es el cuantil relevante (dependiendo del nive de confianza deseado) de la distribución normal estándar.

Otra metodología para probar la presecia de dependencia serial en series de tiempo de conteo es la que se obtiene mediante el estadístico de Freeland (1998). Dicho estadística de prueba se denota por S y se define como

$$S := \frac{1}{\sqrt{T}\bar{x}} \sum_{t=2}^{T} (x_{t-1} - \bar{x})(x_t - \bar{x}) \tag{4.44}$$

donde

$$\bar{x} = \frac{1}{T} \sum_{t=1}^{T} x_t$$

Si la hipótesis nula es que $\{X_t\}$ son variables aleatorias indepenientes idénticamente distribuidas como $Poisson(\lambda)$, Freeland (1998) demostró que S converge en distribución a una variable aleatoria con distribución normal estándar. Se puede obtener una versión modificada del estadístico de Freeland utilizando el hecho de que la media y varianza de una variable aleatoria con distribución Poisson son iguales. La estadística modificada es

$$S^* := \frac{\sqrt{T}\sum_{t=2}^{T}(x_{t-1}-\bar{x})(x_t-\bar{x})}{\sum_{t=1}^{T}(x_t-\bar{x})^2} \tag{4.45}$$

S^* es asintóticamente equivalente a S ya que bajo la hipótesis nula de que $\{X_t\}$ son independientes y todas tienen distribución $Poisson(\lambda)$, la media y varianza muestrales son asintóticamente iguales. Parece razonable que esta estadística trabajará mejor en aquellas situaciones en las que la distribución Poisson es muy restrictiva. Esto es muy cómún en situaciones prácticas en las que hay mucha sobredispersión en los datos. De nuevo, sólo se están considerando pruebas de una cola y se rechazará H_0 si se obtienen valores grandes de S ó S^*.

Una tercera clase de pruebas considera dos estadísticos de pruebas diseñados por Venkatamaran (1982) y Mills & Seneta (1989) para medir bondad de ajuste en un proceso de ramificación simple. Bajo la hìpótesis nula de independencia e idéntica dstribución de $\{X_t\}$, la estadística de prueba de Venkatamaran (1982) está dada por

$$Q_{acf}(k) = \frac{\sum_{i=1}^{k}\hat{\varrho}_{i+1}^2[\sum_{t=1}^{T}(x_t-\bar{x})^2]^2}{\sum_{t=i+2}^{T}(x_t-\bar{x})^2(x_{t-i-1}-\bar{x})^2} \tag{4.46}$$

donde

$$\hat{\varrho}_k = \frac{\sum_{t=k+1}^{T}(x_t-\bar{x})(x_{t-k}-\bar{x})}{\sum_{t=1}^{T}(x_t-\bar{x})^2} \tag{4.47}$$

y $k \in \mathbb{N}^+$. Bajo la hipótesis nula de que $\{X_1,\ldots,X_T\}$ son variables aleatorias indepenientes idénticamente distribuidas como $Poisson(\lambda)$ se puede mostrar que $Q_{acf}(k)$ converge en distribución a una variable aleatoria con distribución $\chi^2_{(k)}$ cuando $T\to\infty$ (para $k\in\mathbb{N}^+$ fija). La línea de razonamiento de este resultado para $k=1$ es como sigue: se necesita la distribución límite de la autocorrelación muestral $\hat{\varrho}_2$ y como los sumandos en el numerador de $\hat{\varrho}_2$ no son independientes pero sí forman diferencia martigala con respecto a alguna filtración adecuada, se necesita una Teorema Central del Límite para martingalas, y si se prueba que $\sqrt{T}\hat{\varrho}_2$ converge en distribución a una variable aleatoria normal estándar y $T\hat{\varrho}_2^2$ converge en distribución a una variable aletoria ji-cuadrada con un grado de libertad y usando argumentos de aditividad de la distribución ji-cuadrada se puede extender el resultado para $k=0,1,\ldots$. Por la Ley Débil de Grandes Números se tiene que

$$plim\left[\frac{1}{T}\sum_{t=1}^{T}(X_t-\bar{X})^2\right]^2=\lambda^2$$

Y el denominador de Q_{acf} está formado por sumandos dependientes, pero utilizando alguna Ley Débil de Grandes Números apropiada se tiene que

$$plim\left[\frac{1}{T}\sum_{t=3}^{T}(X_t-\bar{X})^2(X_{t-2}-\bar{X})^2\right]=\lambda^2$$

esto establece que la estadística $Q_{acf}(1)$ es asintóticamente equivalente a una basada en $T\hat{\varrho}_2^2$.

La segunda estadística es la propuesta por Mills & Seneta (1989) también para el caso de independencia e idéntica distribución está dada por

$$Q_{pacf}(k):=\frac{\sum_{i=1}^{k}\hat{\phi}_{i+1}^2[\sum_{t=1}^{T}(x_t-\bar{x})^2]^2}{\sum_{t=i+2}^{T}(x_t-\bar{x})^2(x_{t-i-1}-\bar{x})^2}$$

donde $\hat{\phi}_k$ es la autocorrelación parcial muestral de orden k (con $k\in\mathbb{N}^+$). Con argumentos análogos a la caso de $Q_{acf}(k)$, se puede probar que $Q_{pacf}(k)$ se distribute asintóticamente como una variable aleatoria ji-cuadrada con k grados de libertad.

Nótese que la estadística $Q_{acf}(k)$ está basada en las autocorrelaciones muestrales de orden $k+1$ y la $Q_{pacf}(k)$ también está basada en las autocorrelaciones parciales muestrales de orden $k+1$, i.e. ambas estadísticas ignoran las autocorrelaciones muestrales de primer orden. Esta es una observación directa de la construcción de dichas estadísticas de prueba (más intuitivamente que como se obtuvieron en Cameron & Triverdi (1998)). De forma análoga a los modelos AR y MA para series de tiempo continuas, estas dos estadísticas ayudan a distinguir entre las estructuras $INAR(1)$ e $INMA(1)$ para datos de conteo. Además, la forma de las funciones de autocorrelación y autocorrelación parcial en modelos basados en adelgazamiento para conteos tiene analogías genéricas con sus contrapartes continuas; por ejemplo, la función de autocorrelación de un modelo $INMA(q)$ también tiene la propiedad de corte despúes de un rezago q como en la determinación de modelos en el caso Gaussiano.

Se recomienda utilizar diferentes pruebas ya que éstas pueden ser muy sensibles ante diferentes estructuras de dependencia en los datos. La siguiente tabla presenta la varias de estas estadísticas para los datos Fürth. Dichas estadísticas fueron obtenidas con una significancia del 1 %.

Estadístico	Valor
Z	-10.71
S	14.12
S^*	14.93
$Q_{acf}(1)$	40.02
$Q_{acf}(5)$	72.41
$Q_{pacf}(1)$	17.51
$Q_{pacf}(5)$	24.77

Las estadísticas $Q_{acf}(1)$ y $Q_{pacf}(1)$ dan evidencia de presencia de correlación serial en los datos, por tanto se puede inferir que ni el $PoINAR(1)$ ni el $PoINMA(1)$ se ajustan adecuadamente a los datos. Estas conclusiones coinciden con la evidencia de Mill & Seneta (1989), que encontraban ajuste pobre a su modelo $BGWI$ (que es equivalente al modelo $INAR(1)$). Los siguientes candidatos naturales son los modelos $INAR(2)$ e $INMA(2)$.

4.6. Una pequeña generalización del modelo de Du & Li: $GINAR(p)$

En este punto, se estudiará una pequeña generalización del modelo $INAR(p)$ de Du & Li. Dicha generalización causal y estacionaria se conoce como proceso $GINAR(p)$, auto-regresivo con valores enteros generalizado (*generalized integer-valued autoregressive*). Primero se demostará que dicho proceso existe y se dará un criterio muy simple para verificar que dicho proceso sea causal y estacionario (bastará que la suma de los coeficientes de auto-regresión sea menor que uno).

Ya se dijo que el proceso $INAR(p)$ de Du & Li (1991) tiene una estructura de correlación igual a la del $AR(p)$ estándar. Ahora se analizará la riqueza que tiene esta estructura mediante las ecuaciones de Yule-Walker.

Una extensión natural, de hecho es la propuesta de Latour (1998), del proceso $X_t = \sum_{j=1}^{p} \alpha_p \circ X_{t-j} + Z_t$ con "$\circ$" el operador binomial (i.e. se tiene sumas de variables aleatorias $Bnlli(\alpha_j)$ como operador de van Harn & Steutel) es que en la operación $\alpha \circ X = \sum_{j=1}^{X} Y_j$ en vez de que las Y_j's sean variables aleatorias $Bnlli$, éstas sean variables aleatorias independientes con media α_i y varianza β_i.

Observación 4.6.1.

No se cambiará la notación para el operador de van Harn & Steutel, sólo se debe considerar que en esta secciones las operaciones se están haciendo con este operador generalizado. $\triangledown$

Considérese un proceso auto-regresivo, no necesariamente con valores enteros, que satisfaga

$$X_t = \sum_{k=1}^{p} \alpha_k X_{t-k} + e_t, \quad t \in \mathbb{Z} \tag{4.48}$$

donde $\{e_t\}_{t\in\mathbb{Z}}$ es un ruido blanco con varianza σ^2. Sean $\gamma(k)$, $k = 0, 1, , p$ las primeras $p+1$ autocovarianzas de este proceso. Las siguientes ecuaciones se conocen como *ecuaciones de Yule-Walker* y proporcionan un nexo entre los parámetros y las autocovarianzas del modelo

$$\gamma(h) = \begin{cases} \sum_{i=1}^{p} \alpha_i \gamma(i) + \sigma^2 & \text{si } h = 0; \\ \sum_{i=1}^{p} \alpha_i \gamma(h-i) & \text{si } h = 1, 2, \ldots, p \end{cases} \tag{4.49}$$

En algunos casos, dadas las autocovarianzas se pueden encontrar los parámetros $\alpha_1, \ldots, \alpha_p$ y σ^2 del modelo (4.48). Si se supone que $\gamma(0) > 0$, entonces los coeficientes de la matriz de las últimas p ecuaciones del sistema (4.49) es positiva definida. Por tanto, la solución de este subsistema es única y σ^2 se puede encontrar sustituyendo $h = 0$ en (4.49).

En algunos otros casos, dados los valores de $\alpha_1, \ldots, \alpha_p$ y σ^2 se podría querer encontrar una solución $\gamma(k),\ k = 0, 1, \ldots, p$ de (4.49) y se puede afirmar que es única. En la literatura clásica, pareciera que los autores no analizan esto. El siguiente lema es una criterio simple para la existencia de una única solución.

Lema 4.6.1.

Sean $\alpha_1, \ldots, \alpha_p \in [0, 1]$ tales que $\sum_{i=1}^{p} \alpha_i < 1$. El sistema de ecuaciones (4.49) con incógnitas $\gamma(k),\ k = 0, 1, \ldots, p$ tiene solución única.

Demostración:

Sea $\mathbf{C} \in M_{(p+1)\times(p+1)}(\mathbb{R})$ tal que la entrada (i, j) es el coeficiente de $\gamma(i)$ en el lado derecho de la j-ésima columna del sistema (4.49), $i = 1, \ldots, p$. Sea $\boldsymbol{\gamma}_p = (\gamma(0), \ldots, \gamma(p))'$ y sea $\mathbf{c} \in M_{(p+1)\times 1}(\mathbb{R})$ cuya primera entrada es 1 y las demás son iguales a cero. Con esta notación se tiene que

$$(\mathbf{I} - \mathbf{C})\boldsymbol{\gamma}_p = \sigma^2 \mathbf{c} \tag{4.50}$$

el total en cada renglón de $\mathbf{C}$ es $\sum_{i=1}^{p} \alpha_i$ y la matriz $\mathbf{C}^* := \left(\sum_{i=1}^{p} \alpha_i\right)^{-1} \mathbf{C}$ es una matriz estocástica. Por tanto, el eigenvalor, λ^*, más grande de $\mathbf{C}^*$ es 1 y los otros eigenvalores tienen

valor absoluto menor o igual que 1. También un resultado de álgebra lineal es que si λ es el eigenvalor más grande de $\mathbf{C}$ entonces $\lambda = (\sum_{i=1}^{p} \alpha_i)\,\lambda^*$, y como $\sum_{i=1}^{p} \alpha_i < 1$

$$\sum_{k=0}^{\infty} \mathbf{C}^k = (\mathbf{I} - \mathbf{C})^{-1}$$

Entonces, la solución de la ecuación (4.50) es

$$\boldsymbol{\gamma}_p = \sigma^2 (\mathbf{I} - \mathbf{C})^{-1} \mathbf{c}$$

□

Lema 4.6.2.

Sea $\alpha(z) := 1 - \alpha_1 z - \alpha_2 z^2 - \ldots - \alpha_p z^p$ y $\varpi(z) := z^p - \alpha_1 z^{p-1} - \ldots - \alpha_{p-1} z - \alpha_p$ con $\sum_{k=1}^{p} |\alpha_k| < 1$. Las raíces de $\alpha(z)$ están fuera del círculo unitario y las de $\varpi(z)$ están dentro del círculo unitario

Demostración:

Sea $\mathcal{C} = \{z \in \mathbb{C} : |z| < 1\}$ el círculo unitario. Defínanse las funciones $f(z) := 1$ y $g(z) := -(\alpha_1 z + \ldots + \alpha_p z^p)$. Nótese que f y g son analíticas sobre $\mathcal{C}$. Por tanto (sobre $\mathcal{C}$)

$$|g| \leq \sum_{k=1}^{p} |\alpha_k z^k| \leq \sum_{k=1}^{p} |\alpha_k| < 1 = |f|$$

Por el Teorema de Rouché, $f(z)$ y $f(z) + g(z) = 1 - \alpha_1 z - \ldots - \alpha_p z^p = \alpha(z)$ tienen el mismo número de ceros en $\mathcal{C}$. Pero f no tiene ceros en $\mathcal{C}$ (de hecho no tiene ceros en ningún lado). Por tanto, las raíces de $\alpha(z)$ están fuera del círculo unitario. Con un argumento similar se prueba que las raíces de $\varpi(z)$ están dentro del círculo unitario. □

4.6.1. Existencia del proceso $GINAR(p)$

Antes que todo se dirá qué se entenderá por que exista el proceso $GINAR(p)$. Considérese un proceso estacionario $\{X_t\}_{t\in\mathbb{Z}}$ tal que $sop(X_t) \subseteq \mathbb{N}^+$. El proceso $\{X_t\}_{t\in\mathbb{Z}}$ es un proceso $GINAR(p)$ si y sólo si existe una sucesión de variables aleatorias independientes idénticamente distribuidas, $\{Z_t\}_{t\in\mathbb{Z}}$, tales que $sop(Z_t) \subseteq \mathbb{N}^+$ con media $\mu_Z > 0$ y varianza $\sigma_Z^2 > 0$ y p operadores generalizados de Steutel & van Harn, $\alpha_j\circ$, $j = 1, 2, \ldots, p$ tales que

$$X_t = \sum_{j=1}^{p} \alpha_j \circ X_{t-j} + Z_t, \quad t \in \mathbb{Z} \tag{4.51}$$

En la ecuación (4.51) parece ser más apropiado escribir $\alpha_j^{(t)}\circ$ en vez de $\alpha_j\circ$ para indicar explícitamente que existe una sucesión de variables aleatorias $\{Y_k^{(t,j)}\}_{k\in\mathbb{Z}}$ para cada $t\in\mathbb{Z}$. Se supondrá que $\mathbb{E}(Y_k^{(t,j)})=\alpha_j$, $\alpha_1,\ldots,\alpha_p\geq 0$, $\alpha_p\neq 0$ y que $Var(Y_k^{(t,j)})=\beta_j$.

Por el momento también supóngase que $\{X_t\}_{t\in\mathbb{Z}}$ tiene media constante (i.e. un prerequisito será estacionaredad débil). Tomando esperanzas en ambos lados de (4.51) se tiene que

$$\mathbb{E}(X_t)=\mu_Z\left(1-\sum_{j=1}^{p}\alpha_j\right)^{-1}$$

como $\alpha_1,\alpha_2,\ldots,\alpha_{p-1}\in[0,1]$, $\alpha_p,\mu_Z>0$ y $\mathbb{E}(X_t)>0$ se tiene que $\sum_{j=1}^{p}\alpha_k<1$.

A continuación se demostrará un lema necesario para la demostración de la existencia del proceso $GINAR(p)$.

Lema 4.6.3.

Sea $\mathbf{A}\in M_{r\times r}(\mathbb{R}^+)$ tal que todos los eigenvalores de $\mathbf{A}$ están dentro del círculo unitario. Defínase la sucesión $\{\mathbf{S}(n)\}_{n=0}^{\infty}$

$$\mathbf{S}(n)=\sum_{k=0}^{n}\mathbf{1}'\mathbf{A}^{n-k}\mathbf{1}\mathbf{A}^k(\mathbf{A}^k)'$$

donde $\mathbf{1}\in M_{r\times 1}(\mathbb{R}^+)$, $\mathbf{1}=(1,\ldots,1)'$. Entonces,

$$\lim_{n\to\infty}\mathbf{S}(n)=\mathbf{0}$$

Demostración:

Nótese que $\mathbf{S}_1(n)=\sum_{k=0}^{n}\mathbf{1}'\mathbf{A}^k\mathbf{1}$ converge a $\mathbf{1}'(\mathbf{I}-\mathbf{A})^{-1}\mathbf{1}$. Entonces para $a_k:=\mathbf{1}'\mathbf{A}^k\mathbf{1}$, entonces $\sum_{k=0}^{n}a_k$ converge. Esto que la serie de potencias

$$\mathbf{Y}_1(x):=\sum_{k=0}^{n}a_kx^k$$

converge absolutamente y uniformemente (para $|x|\leq 1$). La serie

$$\mathbf{S}_2(n) := \sum_{k=0}^{n} \mathbf{A}^k (\mathbf{A}^k)'$$

también es convergente, pues es una sucesión creciente acotada por

$$(\mathbf{I} - \mathbf{A})^{-1}[(\mathbf{I} - \mathbf{A})^{-1}]'$$

Así que si se define $b_{i,j}^{(k)} := \left(\mathbf{A}^k(\mathbf{A}^k)'\right)_{i,j}$, entonces la serie de potencias

$$\mathbf{Y}_2(x) := \sum_{k=0}^{n} b_{i,j}^{(k)} x^k$$

converge absolutamente y uniformemente (para $|x| \leq 1$).

Por lo anterior, la serie $\mathbf{Y} := \mathbf{Y}_1\mathbf{Y}_2$ converge absolutamente y uniformemente (para $|x| \leq 1$). Pero

$$\mathbf{Y}(x) := \sum_{n=0}^{\infty} \left(\sum_{k=0}^{n} a_{n-k} b_{i,j}^{(k)} \right) x^k$$

Por tanto, $\sum_{k=0}^{n} a_{n-k} b_{i,j}^{(k)} \to 0$

□

Proposición 4.6.1. *(Existencia del proceso GINAR(p))*

Dados p operadores de Steutel & van Harn $\alpha_1\circ, \ldots, \alpha_p\circ$ tales que $\sum_{k=1}^{p} \alpha_k < 1$ y una sucesión de variables aleatorias independientes idénticamente distribuidas $\{Z_t\}_{t\in\mathbb{Z}}$ tales que $sop(Z_t) \subseteq \mathbb{N}$, entonces existe un proceso estacionario, $\{X_t\}_{t\in\mathbb{Z}}$ con $sop(X_t) \subseteq \mathbb{N}^+$ tal que

$$X_t = \sum_{i=1}^{p} \alpha_i \circ X_{t-i} + Z_t, \quad t \in \mathbb{Z}$$

y si $s < t$, entonces $Cov(X_s, Z_t) = 0$

Demostración:

Primero se definirá un sucesión de variables aleatorias $\{X_t^{(n)}\}_{n\in\mathbb{N}}$ que se aproxima a X_t. A saber, defínase

$$X_t^{(n)} := \begin{cases} 0 & \text{si } n < 0; \\ Z_t & \text{si } n = 0; \\ \sum_{i=1}^{p} \alpha_i^{(t)} \circ X_{t-i}^{(n-i)} & \text{si } n > 0 \end{cases} \tag{4.52}$$

En esta notación, el símbolo $\alpha_i^{(t)}\circ$ significa que la sucesión de conteo $\{Y_k^{(t,i)}\}_{k\in\mathbb{N}}$ es la que se usa en la operación $\alpha_i\circ$ para $t \in \mathbb{Z}$ fija.

Ahora se demostrará que las variables $X_t^{(n)} \in L^2(\Omega, \mathscr{F}, \mathbb{P})$. Entonces, para $t \in \mathbb{Z}$ fija se demostrará que $\{X_t^{(n)}\}_{t\in\mathbb{Z}}$ es una sucesión de Cauchy. Como $L^2(\Omega, \mathscr{F}, \mathbb{P})$ es un espacio de Hilbert, entonces $X_t := \lim_{n\to\infty} X_t^{(n)}$ existe como un elemento de $L^2(\Omega, \mathscr{F}, \mathbb{P})$.

Se probará que $X_t^{(n)}$ tiene primer momento finito y se hará mediante inducción fuerte.

Primero,

$$\mathbb{E}(X_t^{(0)}) = \mathbb{E}(Z_t) = \mu_Z < \infty$$

Ahora, supóngase que para $k = 0, 1, \ldots, n-1$, $\mathbb{E}(X_t^{(k)}) < \infty$. Ahora,

$$\begin{aligned} \mathbb{E}(X_t^{(n)}) &= \sum_{i=1}^{p} \mathbb{E}(\alpha_i^{(t)} \circ X_{t-i}^{(n-i)}) + \mathbb{E}(Z_t) \\ &= \sum_{i=1}^{p} \alpha_i \mathbb{E}(X_{t-i}^{(n-i)}) + \mu_Z < \infty \quad \text{por la hipótesis de inducción} \end{aligned} \tag{4.53}$$

De la definición de $X_t^{(n)}$ y del hecho de que $\{Z_t\}_{t\in\mathbb{Z}}$ es una sucesión de variables aleatorias independientes, se puede notar que $\mathbb{E}(X_t^{(n)})$ depende de n pero no de t; por tanto, se denotará como $\mu^{(n)}$ al valor de $\mathbb{E}(X_t^{(n)})$. La ecuación (4.53) implica que

$$\mu^{(n)} = \sum_{i=1}^{p} \alpha_i \mu^{(n-i)} + \mu_Z$$

Para la varianza, primero nótese que

$$Var(X_t^{(0)}) = Var(Z_t) = \sigma_Z^2 < \infty$$

y para $k \in \{1, 2, \ldots, p\}$ considérese

$$Cov(X_t^{(n)}, X_{t-k}^{(n-k)}) = \mathbb{E}(X_t^{(n)} X_{t-k}^{(n-k)}) - \mu^{(n)}\mu^{(n-k)}$$

Pero

$$\mathbb{E}(X_t^{(n)} X_{t-k}^{(n-k)}) = \alpha_k \mathbb{E}((X_{t-k}^{(t-k)})^2) + \sum_{i=1, i\neq k}^{p} \alpha_i \mathbb{E}(X_{t-i}^{(n-i)} X_{t-k}^{(n-k)}) + \mu^{(n-k)} \mu_Z$$

En términos de covarianzas se tiene que

$$Cov(X_t^{(n)}, X_{t-k}^{(n-k)}) = \sum_{i=1}^{p} Cov(X_{t-i}^{(n-i)}, X_{t-k}^{(n-k)}) \tag{4.54}$$

Es importante notar que sólo las covarianzas de las variables $X_{t-l}^{(n-l)}$, $l \in \{1, 2, \ldots, p\}$ en el lado derecho de la ecuación (4.54). Ahora sí se calculará $Var(X_t^{(n)})$.

$$Var(X_t^{(n)}) = \sigma_Z^2 + \sum_{i=1}^{p} \beta_i \mathbb{E}(X_{t-i}^{(n-i)}) + \sum_{i=1}^{p} \sum_{k=1}^{p} \alpha_i \alpha_k Cov(X_{t-i}^{(n-i)}, X_{t-k}^{(n-k)})$$

estas dos últimas ecuaciones muestran que la varianza de $X_t^{(n)}$ se puede expresar en términos de segundos momentos centrados de las variables $X_{t-l}^{(n-l)}$, $l = 1, \ldots, p$. Con este argumento recursivo aplicado n veces, se verifica que $Var(X_t^{(n)}) < \infty$ y depende de n pero no de t, i.e. $X_t^{(n)}$ pertenece a $L^2(\Omega, \mathscr{F}, \mathbb{P})$.

El siguiente paso es demostrar que esto es una sucesión de Cauchy. Defínase

$$U(n,t,k) := |X_t^{(n)} - X_t^{(n-k)}| \quad l(n,t,k) := \text{mín}\{X_t^{(n)} - X_t^{(n-k)}\}$$

Nótese que $U(n,t,k) \geq 0$ y además

$$\begin{aligned}
U(n,t,k) &\leq \sum_{i=1}^{p} \left| \alpha_i^{(t)} \circ X_{t-i}^{(n-i)} - \alpha_i^{(t)} \circ X_{t-i}^{(n-i-k)} \right| \\
&= \sum_{i=1}^{p} \left| \sum_{j=1}^{X_{t-i}^{(n-i)}} Y_j^{t,i} - \sum_{j=1}^{X_{t-i}^{(n-i-k)}} Y_j^{t,i} \right| \\
&= \sum_{i=1}^{p} \sum_{j=1}^{U(n-i,t-i,k)} Y_{l(n-i,t-i,k)+j}^{(t,i)} \\
&= \sum_{i=1}^{p} \alpha_i \circ U(n-i, t-i, k)
\end{aligned}$$

Ahora, sea

$$\mathbf{A} = \begin{pmatrix} \alpha_1 & \alpha_2 & \cdots & \alpha_{p-1} & \alpha_p \\ 1 & 0 & \cdots & 0 & 0 \\ 0 & 1 & \cdots & 0 & 0 \\ \vdots & \vdots & \ddots & \vdots & \vdots \\ 0 & 0 & \cdots & 1 & 0 \end{pmatrix}$$

y

$$\mathbf{A} \circ \begin{pmatrix} X_1 \\ \vdots \\ X_p \end{pmatrix} = \begin{pmatrix} \sum_{i=1}^{p} \alpha_i \circ X_i \\ X_1 \\ \vdots \\ X_{p-1} \end{pmatrix}$$

Considérese el vector

$$\mathbf{U}_{t,k}^{(n)} = \begin{pmatrix} U(n,t,k) \\ U(n-1,t-1,k) \\ \vdots \\ U(n-p+1,t-p+1,k) \end{pmatrix}$$

donde

$$\mathbf{B} := \begin{pmatrix} \beta_1 & \beta_2 & \cdots & \beta_p \\ 0 & 0 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & 0 \end{pmatrix}$$

Aplicando esta recurrencia n veces se tiene que

$$\mathbb{E}(\mathbf{U}_{t,k}^{(n)}\mathbf{U}_{t,k}^{(n)'}) \leq \sum_{j=0}^{n-1} \mathbf{A}^j diag(\mathbf{B}\mathbf{A}^{n-j+1}\boldsymbol{\mu}')(\mathbf{A}^j)'$$
$$+\mathbf{A}^n \begin{pmatrix} \sigma_Z^2 + \mu_Z^2 & \mathbf{0}'_{p-1} \\ \mathbf{0}_{p-1} & \mathbf{0}_{p-1\times p-1} \end{pmatrix} \mathbf{A}^{n'}$$

Todos los eigenvalores de $\mathbf{A}$ están dentro del círculo unitario por tanto el último término de esta expresión tiende a $\mathbf{0}$. El primer término

$$\sum_{j=0}^{n-1} \mathbf{A}^j diag(\mathbf{B}\mathbf{A}^{n-j+1}\boldsymbol{\mu}')(\mathbf{A}^j)'$$

está acotado por

$$\max_{1\leq j\leq r}\{\beta_j\}\mu_Z\sum_{j=0}^{n-1}\mathbf{1}'\mathbf{A}^{n-j+1}\mathbf{1}\mathbf{A}^j\mathbf{A}^{j'}$$

por el Lema (4.6.3), la sucesión

$$\mathbf{S}(n)=\sum_{j=0}^{n}\mathbf{1}'\mathbf{A}^{n-j}\mathbf{1}\mathbf{A}^j\mathbf{A}^{j'}$$

converge a la matriz $\mathbf{0}$.

□

El proceso $INAR(1)$ con distribución marginal Poisson basada en el operador binomial es un caso particular de un proceso de ramificación simple con inmigración. A continuación se analizará otro ejemplo de un proceso $GINAR(1)$.

Ejemplo 4.6.1.

Sean G_{X_t}, G_Z y G_Y la funciones generadoras de probabilidad de X_t, Z_t y $Y_k^{(t)}$, respectivamente. Para un proceso estrictamente estacionario $INAR(1)$ se tiene que

$$G_{X_t}(s)=G_{X_t}(G_Y(s))G_Z(s) \tag{4.55}$$

Supóngase que en la sucesión $\{Z_t\}_{t\in\mathbb{Z}}$ es tal que $X_t\sim Poisson(\lambda)$ i.e. $G_Z=exp(-\lambda(1-s))$. Considérese una sucesión de variables aleatorias independientes con función de probabilidad dada por

$$p_Y(y)=\begin{cases}\frac{1-e^{-\lambda/m}}{\beta} & \text{si } y=0;\\ \frac{e^{-\lambda/m}(\lambda/m)^y}{\beta y!}(\frac{m\beta}{\lambda}y-1) & \text{si } y\geq 1\end{cases}$$

donde $m\in\mathbb{N}^+$ y $\lambda\in(0,1)$ tal que $\beta\geq\lambda/m$. La función generadora de probabilidad de esta variable está dada por

$$G_Y(s)=\frac{1-(1-\beta s)\exp\{-\lambda(1-s)/m\}}{\beta}$$

Por tanto, el valor esperado de esta variable está dada por $\alpha=1-\frac{\lambda(1-\beta)}{m\beta}$. Con elecciones adecuadas de m,λ y β se puede tener $\alpha<1$. Con estas distribuciones, el proceso $GINAR(1)$

$$X_t=\alpha\circ X_{t-1}+Z_t,\quad t\in\mathbb{Z}$$

es tal que la distribución marginal de X_t es binomial negativa con parámetros m y β y función generadora de probabilidad dada por

$$G_{X_t}(s) = \left(\frac{1-\beta}{1-\beta s}\right)^{-m}$$

ya que G_{X_t}, G_Z y G_Y satisfacen la ecuación (4.55). ▽

Sean $\alpha_1, \alpha_2, \ldots, \alpha_p \geq 0$ tales que $\sum_{i=1}^{p} \alpha_i < 1$ y $\beta_1, \beta_2, \ldots, \beta_p, \sigma^2 \in \mathbb{R}^+$. Defínase Y_i de tal forma que

$$Y_t - \mu_Y = \sum_{k=1}^{p} \alpha_k (Y_{t-k} - \mu_Y) + \tilde{e}_t \tag{4.56}$$

donde $\{\tilde{e}_t\}_{t\in\mathbb{Z}}$ es un ruido blanco Gaussiano con varianza

$$\sigma^2 := \mu_y \sum_{i=1}^{p} \beta_i + \sigma_Z^2$$

Por el Lema (4.6.2), el operador auto-regresivo $\alpha(B) = 1 - \alpha_1 B - \ldots - \alpha_p B^p$ tiene todas sus raíces fuera del círculo unitario y por tanto, $\{Y_t\}_{t\in\mathbb{Z}}$ es un proceso $AR(p)$ causal y estacionario.

El comportamiento de la función de autocovarianza asociada con un proceso estacionario causal caracteriza el proceso en cierto sentido. Recuérdese el siguiente resultado de teoría clásica de series de tiempo.

Proposición 4.6.2.

Supóngase que $\{X_t\}$ y $\{Y_t\}$ son procesos estacionarios con media cero ambos con función de autocovarianza $\gamma(\cdot)$. Si $\{Y_t\}$ es un proceso $ARMA(p,q)$ entonces $\{X_t\}$ es también un proceso $ARMA(p,q)$

Y el siguiente corolario es importante en el análisis de series de tiempo enteras.

Corolario 4.6.1.

Si $\{X_t\}_{t\in\mathbb{Z}}$ es un proceso $GINAR(p)$ que satisface la ecuación (4.51). Entonces, $\{X_t\}_{t\in\mathbb{Z}}$ es un proceso $AR(p)$ que se puede escribir como

$$X_t - \mu_X = \sum_{k=1}^{p} \alpha_k (X_{t-k} - \mu_X) + e_t \tag{4.57}$$

donde $\{e_t\}_{t\in\mathbb{Z}}$ es un proceso de ruido blanco de varianza $\sigma^2 := \mu_X \sum_{i=1}^{p} \beta_i + \sigma_Z^2$ y β_i es la varianza de las variables involucradas en el i-ésimo operador de Steutel & van Harn, $i = 1, 2, , p$

Demostración:

Como $\{X_t\}_{t\in\mathbb{Z}}$ es un proceso $GINAR(p)$ se tiene que $\alpha_1, \ldots, \alpha_{p-1} \in [0, 1)$, $\alpha_p \in (0, 1)$ y $\sum_{k=1}^{p} \alpha_k < 1$.

Sea $\sigma^2 = \mu_X \sum_{i=1}^{p} \beta_i + \sigma_Z^2$. El proceso Y_t que satisface (4.56) es un proceso estacionario $AR(p)$ causal Gaussiano. Por los resultados de Gauthier & Latour (1994), la función del autocovarianza del proceso $GINAR(p)$ $\{X_t\}_{t\in\mathbb{Z}}$ con $\sigma^2 = \mu_X \sum_{i=1}^{p} \beta_i + \sigma_Z^2$ satisface (4.49). Por el Lema (4.6.1), la solución de este sistema es única. Por tanto, la función de autocovarianza del proceso $\{X_t\}_{t\in\mathbb{Z}}$ es idéntica a la función de autocovarianza del proceso $AR(p)$, $\{Y_t\}_{t\in\mathbb{Z}}$.

Por la proposición (4.6.2), $\{X_t\}_{t\in\mathbb{Z}}$ es un proceso $AR(p)$, que se puede escribir como en la ecuación

$$X_t - \mu_X = \sum_{k=1}^{p} \alpha_k (X_{t-k} - \mu_X) + e_t$$

Como $\mu_Z = (1 - \sum_{i=1}^{p} \alpha_i)\mu_X$. La ecuación (4.57) también se puede escribir como

$$X_t = \sum_{k=1}^{p} \alpha_k X_{t-k} + \mu_Z + e_t$$

Otra forma de mostrar que un $GINAR(p)$ es un $AR(p)$ se basa en el siguiente argumento: Sea $\{X_t\}_{t\in\mathbb{Z}}$ un proceso $GINAR(p)$ y sea $e_t := \mathbb{E}(X_t|\sigma(X_{t-1}, X_{t-2}, \ldots))$. Se puede probar que $\{e_t\}$ es una diferencia martingala. Por tanto, satisface la ecuación de recurrencia

$$X_t - \mu_X = \sum_{k=1}^{p} \alpha_k (X_{t-k} - \mu_X) + e_t$$

Evidentemente, esta última ecuación establece que $\{X_t\}_{t\in\mathbb{Z}}$ es un proceso $AR(p)$

□

4.7. Generalización a operadores generales de Steutel & van Harn

La idea de esta sección final es estudiar una familia de procesos auto-regresivos que toma valores enteros de orden p ($INAR(p)$) mediante el uso de la multiplicación generalizada de Steutel & van Harn (1982).

Se obtendrán varias propiedades distribucionales y de regresión (esperanza condicional) para estos modelos. También se analizará un conjunto de procesos estacionarios $INAR(p)$ con marginales específicas y algunas generalizaciones de éstos tal como se hizo en la primera sección de este capítulo.

Ya se han estudiado en este capítulo y en el anterior algunos modelos para datos en series de tiempo discretas. Varios autores han desarrollado estos y más modelos de series de tiempo para datos de conteo y han hecho algunas aplicaciones en ajustes para dichs datos y simulación de sucesiones de variables dependientes con distribución Poisson y Binomial Negativa (entre otros). Estos modelos se basan en el operador de adelgazamiento binomial "$\circ$" que se define de la siguiente forma: si X es una variable aleatoria tal que $sop(X) \subseteq \mathbb{N}^+$ y $\alpha \in (0,1)$, entonces

$$\alpha \circ X := \sum_{i=1}^{X} X_i \tag{4.58}$$

donde $\{X_i\}$ es una sucesión de variables aleatorias independientes e idénticamente distribuidas e independientes de X tales que $X_i \sim Bnlli(\alpha)$. La operación "$\circ$" incorpora la naturaleza discreta de las variables y actúa como el análogo estándar de la multiplicación en los modelos de Box-Jenkins.

Ahora se intentará generalizar estos conceptos a operadores de van Harn & Steutel no necesriamente el binomial. van Harn et al. (1982) definieron a multiplicación generalizada $\odot_G$ como una extensión del adelgazamiento binomial y se utilizó para definir los conceptos de auto-descomponibilidad y estabilidad discretos. Posteriormente, van Harn & Steutel (1993) utilizaron $\odot_G$ para definir y resolver ecuaciones de estabilidad que involucran procesos a tiempo continuo con valores discretos (en $\mathbb{N}^+$) con incrementos independientes y estacionarios.

En esta pequeña sección introductoria se recordarán algunas definiciones y resultados para poder definir adecuadamente a procesos que generalizarán los adelgazamientos bino-

mial de Al-Osh & Alzaid.

Definición 4.7.1. *(Función generadora de probabilidad)*

Sea $\{p_n\}_{n=0}^{\infty}$ una distribución de probabilidad sobre $\mathbb{N}$. Se define la función generadora de probabilidad de dicha distribución, G, como

$$G(z) = \sum_{n=0}^{\infty} p_n z^n, \quad |z| \leq 1$$

Sea $\mathcal{G} := \{G_r : r \geq 0\}$ un semigrupo de composición continua de funciones generadoras de probabilidad tales que $G_r \neq 1$ y $\delta_G := -Ln(G_1'(1)) > 0$. Es decir para cualquier $z \in \mathbb{R}$ tal que $|z| \leq 1$

$$G_s \circ G_r(z) = G_{s+r}(z), \quad s, r \geq 0 \tag{4.59}$$

$$\lim_{r \downarrow 0} G_r(z) = z, \quad \lim_{r \to \infty} G_r(z) = 1 \tag{4.60}$$

Definición 4.7.2. *(Generador infinitesimal de $\mathcal{G}$)*

Se define el generador infinitesimal del semigrupo $\mathcal{G}$, U, como

$$U(z) = \lim_{r \downarrow 0} \frac{G_r(z) - z}{r}, \quad |z| \leq 1$$

y satistace que para todo $0 \leq z < 1$, $U(z) > 0$.

Se puede demostra que existen una constante $a > 0$ y una distribución de probabilidad $\{h_n\}_{n=0}^{\infty}$ sobre $\mathbb{N}^+$ con función generadora de probabilidad $H(\cdot)$ tal que $h_1 := 0$,

$$H'(1) = \sum_{n=1}^{\infty} n h_n \leq 1$$

$$U(z) = a(H(z) - z), \quad -1 \leq z \leq 1 \tag{4.61}$$

Defínase la función $\mathcal{A} : z \in [0, 1) \to \mathbb{R}$ como

$$\mathcal{A}(z) = exp\left\{ - \int_0^z \frac{dx}{U(x)} \right\} \tag{4.62}$$

Las funciones $U(\cdot)$ y $\mathcal{A}(\cdot)$ satisfacen

$$U(G_r(z)) = U(z)G_r'(z), \quad \mathcal{A}(G_r(z)) = e^{-r}\mathcal{A}(z), \quad r \in \mathbb{R}^+, z \in [0,1) \tag{4.63}$$

Además

$$\delta_G = a(1 - H'(1)) = -U'(1), \quad G_r'(1) = e^{-\delta_r r}, \quad r \in \mathbb{R}^+ \tag{4.64}$$

Defínase la función $\mathcal{B} : z \in [0,1) \to \mathbb{R}$ como

$$\mathcal{B}(z) = \lim_{r\to\infty} \frac{G_r(z) - G_r(0)}{1 - G_r(0)}$$

y se puede probar que es una función generadora de probabilidad que satisface $\mathcal{B}(0) = 0$ y toma la forma

$$\mathcal{B}(z) = 1 - \mathcal{A}^{\delta_G}(z) \tag{4.65}$$

Definición 4.7.3. *(Multiplicación generalizada $\odot_G$)*

Sea X una variable aleatoria tal que $sop(X) \subseteq \mathbb{N}^+$ y $\eta \in (0,1)$. Se define multiplicación generalizada, $\odot_G$, mediante

$$\eta \odot_G X := \sum_{i=1}^{X} Y_i \tag{4.66}$$

donde $\{Y_i\}_{i=1}^{\infty}$ es una sucesión de variables aleatorias independientes, independientes de X con función generadora de probabilidad (común) G_r, donde $r := -Ln(\eta)$

4.7.1. Procesos $\mathcal{G} - INAR(1)$

Definición 4.7.4. *(Proceso $\mathcal{G} - INAR(1)$)*

Se dice que una sucesión de variables aleatorias $\{X_t\}_{t\in\mathbb{Z}}$ es un proceso $\mathcal{G} - INAR(1)$ si para cualquier $t \in \mathbb{Z}$

$$X_t = \eta \odot_G X_{t-1} + Z_t \tag{4.67}$$

donde $\eta \in (0,1)$ y $\{Z_t\}_{t\in\mathbb{Z}}$ es una sucesión de variables aleatorias independientes tal que $sop(Z_t) \subseteq \mathbb{N}^+$ y Z_t es independiente de las $\{Y_i\}$ que definen la operación $\odot_G$. A $\{Z_t\}_{t\in\mathbb{Z}}$ se le conoce (como antes) como sucesión de innovación.

La multiplicación generalizada $\eta \odot_G X_{t-1}$ en (4.67) se realiza de forma independiente para cada $t \in \mathbb{Z}$. De manera más formal, se supondrá la existencia de un arreglo $\{Y_{i,t} : i \in \mathbb{N}^+, t \in \mathbb{Z}\}$ de variables aleatorias independientes e idénticamente distribuidas con $sop(Y_{i,t}) \subseteq \mathbb{N}^+$ e independientes de $\{Z_t\}_{t\in\mathbb{Z}}$. Cada $Y_{i,t}$ tiene función generadora de probabilidad $G_r(z)$ donde $r = -Ln(\eta)$ y

$$\eta \odot X_{t-1} = \sum_{i=1}^{X_{t-1}} Y_{i,t-1} \tag{4.68}$$

Estas suposiciones hacen que el modelo (4.67) sea una cadena de Markov.

Las funciones generadoras de probabilidad de X_t y Z del proceso $\mathcal{G} - INAR(1)$ (4.67) satisfacen la ecuación

$$G_{X_t}(z) = G_{X_{t-1}}(G_r(z))G_Z(z),\ r = -Ln(\eta),\ t \in \mathbb{Z} \tag{4.69}$$

Usando (4.69) repetidas veces y el hecho de que $\mathcal{G}$ es un semigrupo entonces un proceso $\mathcal{G} - INAR(1)$, $\{X_t\}_{t\in\mathbb{Z}}$, admite la siguiente representación para todo $k \in \mathbb{N}^+$

$$X_t \overset{d}{=} \eta^k \odot_G X_{t-k} + \sum_{i=1}^{k-1} \eta^i \odot_G Z_{t-i},\quad t \in \mathbb{Z}$$

En la siguiente proposición se mencionan algunas propiedades de un proceso $\mathcal{G}-INAR(1)$.

Proposición 4.7.1. *(Algunas propiedades del proceso $\mathcal{G} - INAR(1)$)*

Supóngase que $\sum_{n=2}^{\infty} n(n-1)h_n < \infty$. Sea $\{X_t\}_{t\in\mathbb{Z}}$ un proceso $\mathcal{G} - INAR(1)$ (para alguna $\eta \in (0,1)$) tal que para cualquier $t \in \mathbb{Z}$, $\mathbb{E}(X_t^2) < \infty$, $\mu_Z := \mathbb{E}(Z_0) < \infty$ y $\sigma_Z^2 = Var(Z_0) < \infty$. Entonces

(i) La regresión de X_t sobre X_{t-1} es lineal, i.e.

$$\mathbb{E}(X_t|X_{t-1}) = \eta^{\delta_G} X_{t-1} + \mu_Z, \quad t \in \mathbb{Z}$$

(ii) La varianza condicional de X_t dado X_{t-1} también es lineal, i.e.

$$Var(X_t|X_{t-1}) = AX_{t-1} + \sigma_Z^2, \quad t \in \mathbb{Z}$$

donde

$$A = (1 - U''(1)/U'(1))\eta^{\delta_G}(1 - \eta^{\delta_G})$$

(iii) Para cualquier $t \in \mathbb{Z}$ y $k \in \mathbb{N}^+$, la covarianza $\gamma_t(k) := Cov(X_{t-k}, X_t)$ está dada por

$$\gamma_t(k) = \eta^{k\delta_G} Var(X_{t-k})$$

(iv) Para cualquier $t \in \mathbb{Z}$ y $k \in \mathbb{N}^+$

$$\mathbb{E}(X_t) = \eta^{k\delta_G}\mathbb{E}(X_{t-k}) + \mu_Z \sum_{i=0}^{k-1} \eta^{i\delta_G}$$

$$Var(X_t) = \eta^{2k\delta_G} Var(X_{t-k}) + A\sum_{i=1}^{k} \eta^{2(i-1)\delta_G}\mathbb{E}(X_{t-i}) + \sigma_Z^2 \sum_{i=1}^{k} \eta^{2(i-1)\delta_G}$$

donde A es como en el inciso (ii).

Demostración:

Primero nótese que por (4.61) y $\sum_{n=2}^{\infty} n(n-1)h_n < \infty$ implica que $U''(1)$ existe. Condicionando en (4.68) y por (4.64) se tiene que $\mathbb{E}(\eta \odot_G X_{t-1}|X_{t-1}) = \eta^{\delta_G} X_{t-1}$. Derivando dos veces con respecto a z la expresión $U(G_r(z)) = G_r'(z)U(z)$ (con $r = -Ln(\eta)$) y haciendo tender z a 1 se tiene que $G_r''(1) = \eta^{\delta_G}(\eta^{\delta_G} - 1)\frac{U''(1)}{U'(1)}$. De nuevo por (4.68) y (4.64) se tiene que

$$\mathbb{E}((\eta \odot_G X_{t-1})^2|X_{t-1}) = Var(Y_{1,t-1})X + \eta^{\delta_G} X^2$$

Y nótese que

$$Var(Y_{1,t-1}) = G_t''(1) + G_r'(1) - (G_r'(1))^2 = \eta^{\delta_G}(1-\eta^{\delta_G})\left(1 - \frac{U''(1)}{U'(1)}\right)$$

□

El siguiente resultado garantiza la existencia de un proceso estrictamente estacionario $\mathcal{G} - INAR(1)$.

Proposición 4.7.2.

Dado $\eta \in (0,1)$ y una sucesión de variables aleatorias independientes e idénticamente distribuidas $\{Z_t\}_{t\in\mathbb{Z}}$ tal que $sop(Z_t) \subseteq \mathbb{N}^+$y $\mathbb{E}(Z_t^2) < \infty$. Entonces existe un proceso estacionario $G - INAR(1)$, $\{X_t\}_{t\in\mathbb{Z}}$, tal que

$$X_t = \eta \odot_G X_{t-1} + Z_t$$

tal que si $s < t$ entonces $Cov(X_s, Z_t) = 0$

Ahora, se estudiará la relación entre la auto-descomponibilidad y los procesos estacionarios $\mathcal{G} - INAR(1)$.

Definición 4.7.5.

Se dice que una distribución de probabilidad sobre $\mathbb{N}^+$ es $\mathcal{G}$-auto-descomponible si para cualquier $r > 0$ existe una función generadora de probabilidad $G_r(\cdot)$ tal que

$$G(z) = G(G_r(z))G_r(z), \quad -1 \leq z \leq 1 \tag{4.70}$$

Cualquier distribución $\mathcal{G}$-auto-descomponible puede surgir como la distribución marginal de un proceso estacionario $\mathcal{G} - INAR(1)$. De forma más precisa, se tiene el siguiente resultado.

Proposición 4.7.3.

Sea G la función generadora de probabilidad de una distribución $\mathcal{G}$-auto-descomponible. Para cualquier $\eta \in (0,1)$ existe un proceso estacionario $\mathcal{G}-INAR(1)$, $\{X_t\}_{t\in\mathbb{Z}}$, cuya distribución marginal tiene función generadora de probabilidad G

Demostración:

Utilizando (4.70) para cada $\eta \in (0,1)$ se puede construir un proceso $\mathcal{G} - INAR(1)$, $\{X_t\}_{t=0}^{\infty}$, de la forma (4.67) cuya sucesión de innovación $\{Z_t\}_{t=0}^{\infty}$ tenga función generadora de probabilidad (común) $G_r(\cdot)$ (con $r = -Ln(\eta)$) tal que X_0 tenga función generadora de probabilidad $G(\cdot)$. Se sigue de (4.69) y (4.70) que las X_t's tienen distribución idéntica, lo cual implica que $\{X_t\}_{t=0}^{\infty}$ es estacionario (ya que es una cadena de Markov). □

Ahora se dará una representación de "promedio móviles" para el proceso $\mathcal{G} - INAR(1)$.

Proposición 4.7.4.

Sea $\{X_t\}_{t\in\mathbb{Z}}$ un proceso estacionario $\mathcal{G} - INAR(1)$. Entonces, $\{X_t\}_{t\in\mathbb{Z}}$ admite una representación de promedios móviles de orden infinito (para algún $\eta \in (0,1)$), i.e.

$$X_t \stackrel{d}{=} \sum_{i=1}^{\infty} \eta^i \odot_G Z_{t-i}, \quad t \in \mathbb{Z}$$

donde la convergencia de la serie es en sentido débil.

Con esta representación, es fácil encontrar la media, la varianza y la función de autocorrelación del proceso $\mathcal{G} - INAR(1)$.

Corolario 4.7.1.

Supóngase que $\sum_{n=2}^{\infty} n(n-1)h_n < \infty$. Sea $\{X_t\}_{t\in\mathbb{Z}}$ un proceso $\mathcal{G} - INAR(1)$ (para algún $\eta \in (0,1)$) tal que para cualquier $t \in \mathbb{Z}$, $\mathbb{E}(X_t^2) < \infty$, $\mu_Z := \mathbb{E}(Z_0) < \infty$ y $\sigma_Z^2 = Var(Z_0) < \infty$. Entonces

(i) Para cualquier $t \in \mathbb{Z}$,

$$\mathbb{E}(X_t) = \frac{\mu_Z}{1-\eta^{\delta_G}}$$

$$Var(X_n) = \frac{(1 - U''(1)/U'(1))\eta^{\delta_G}\mu_z + \sigma_Z^2}{1-\eta^{2\delta_G}}$$

(ii) Para cualquier $k \in \mathbb{N}$ y $t \in \mathbb{Z}$, el coeficiente de correlación de X_{t-k} y X_t es

$$\rho(k) = \eta^{k\delta_G}$$

Nótese que la función de autocorrelación del proceso $\mathcal{G}-INAR(1)$, $\rho(k)=\eta^{k\delta_G}$, tiene la misma forma que la del proceso $AR(1)$ estándar. Su decaimiento es exponencial (conforme k crece). Sin embargo, a diferencia del caso $AR(1)$, $\rho(\cdot)$ siempre es positiva.

Observación 4.7.1.

Supóngase que $\{X_t\}_{t=0}^{\infty}$ es un proceso $\mathcal{G}-INAR(1)$. Si las funciones generadoras de probabilidad $G_Z(\cdot)$ y $G_r(\cdot)$ $(r=-Ln(\eta))$ satisfacen

$$\int_0^1 \frac{1-G_Z(x)}{G_r(x)-x}dx<\infty,\quad r=-Ln(\eta)$$

entonces por Foster & Williamson (1971), $\{X_t\}_{t=0}^{\infty}$ tiene una distribución límite. Si se da a X_0 dicha distribución entonces se obtiene la estacionariedad (pues $\{X_t\}_{t=0}^{\infty}$ es una cadena de Markov). ▽

4.7.2. Soluciones estacionarias del proceso $\mathcal{G}-INAR(1)$

En esta parte, se estudiarán varias soluciones del proceso $\mathcal{G}-INAR(1)$.

Definición 4.7.6.

Sea X una variable aleatoria tal que $sop(X)\subseteq\mathbb{N}^+$. Se dice que X tiene una distribución $\mathcal{G}$-estable con exponente ϱ si existe una sucesión de variables aleatorias independientes e idénticamente distribuidas $\{X_i\}_{i=0}^{\infty}$ tal que para todo $i\in\mathbb{N}$ $X_i\overset{d}{=}X$ y para cualquier $t\in\mathbb{N}$

$$X\overset{d}{=}t^{-1/\varrho}\odot_G\sum_{i=1}^{t}X_i$$

Observación 4.7.2.

Las distribuciones $\mathcal{G}$-estables son $\mathcal{G}$-auto-descomponibles y existen sólo cuando $\varrho\in(0,\delta_G]$. Además, la función generadora de probabilidad, $G(z)$, de una distribución $\mathcal{G}$-estable con exponente $\varrho\in(0,\delta_G]$ admite la representación canónica

$$G(z)=exp\{-\lambda\mathcal{A}(z)^{\varrho}\} \tag{4.71}$$

para algún $\lambda>0$ y $\mathcal{A}(z)$ definida en (4.62).

Por la Proposición (4.7.3) para cada $\eta\in(0,1)$ existe un proceso estacionario $\mathcal{G}-INAR(1)$, $\{X_t\}_{t\in\mathbb{Z}}$ con una distribución marginal G-estable con exponente $\varrho\in(0,\delta_G]$.

La distribución marginal de la sucesión de innovación $\{Z_t\}_{t\in\mathbb{Z}}$ se obtiene resolviendo la ecuación (4.69) en términos de G_Z y utilizando (4.63) también es $\mathcal{G}$-estable con exponente ϱ y tiene función generadora de probabilidad dada por

$$G_Z(z) = exp\{-\lambda(1-\eta^{\varrho})A(z)^{\varrho}\}$$

Además, se puede demostrar que los procesos estacionarios $\mathcal{G}-INAR(1)$ cuya marginal es $\mathcal{G}$-estable con media finita surgen sólo en el caso que $\varrho = \delta_G$ y $\mathcal{B}'(1) < \infty$ (donde $\mathcal{B}(z)$ se definió (4.65)). El proceso tiene varianza finita si además $\mathcal{B}''(1) < \infty$.

Haciendo $\eta = e^{-r}$, la función generadora de probabilidad, G, de la distribución marginal de un proceso estacionario $G-INAR(1)$ $\mathcal{G}$-estable satisface que para cualquier $r > 0$ existe $c(r) \in (0,1)$ tal que

$$G(G_r(z)) = G(z)^{c(r)}, \quad z \in [0,1] \tag{4.72}$$

De hecho, esta propiedad caracteriza a tales procesos.

Proposición 4.7.5.

Sea G una función generadora de probabilidad tal que $G(z) \neq 0$ para todo $z \in [0,1]$. Entonces G es $\mathcal{G}$-estable con algún exponente $\varrho \in (0, \delta_G)$ si y sólo si para cualquier $r > 0$ existe $c(r)$ tal que (4.72) se cumple. La función $c(r)$ necesariamente es de la forma $c(r) = e^{-r\varrho}$

Demostración:

Sólo se necesita probar ($\Rightarrow$). Sea $\psi(z) := Ln(G(z))$ de (4.72) se tiene que para cualquier $r > 0$ existe $c(r) \in (0,1)$ tal que

$$\psi(G_r(z)) = c(r)\psi(z) \tag{4.73}$$

Sea $\psi_1(z) := \frac{\psi(z)}{\psi(0)}$. Nótese que $c(r) = \frac{\psi(G_r(0))}{\psi(0)}$ y entonces (4.73) se convierte en

$$\psi_1(G_r(z)) = \psi_1(G_r(0))\psi_1(z) \tag{4.74}$$

derivando (4.74) con respecto a r se tiene que

$$\frac{\partial}{\partial r}G_r(z)\psi_1'(G_r(z)) = \frac{\partial}{\partial r}G_r(0)\psi_1'(G_r(0))\psi_1(z)$$

usando el hecho de que $\frac{\partial}{\partial r}G_r(z) = U(G_r(z))$ y haciendo $r \downarrow 0$, de (4.60) se tiene que

$$\frac{\psi_1'(z)}{\psi_1(z)} = \frac{U(0)}{U(z)}\psi_1'(0)$$

cuya solución es $\psi_1 = \mathcal{A}(z)^{\varrho}$ donde $\varrho = -\psi_1'(0)U(0) > 0$. Por tanto, $G(z)$ tiene la forma (4.71). Como $G(z)$ es una función generadora de probabilidad $\varrho \leq \delta_G$ debe satisfacer que $\varrho \leq \delta$. La forma de $c(r)$ se sigue de la unicidad. □

A continuación se estudiará un proceso estacionario $\mathcal{G} - INAR(1)$ con una distribución marginal $\mathcal{G}$-geométrica estable.

Definición 4.7.7. *(Distribución $\mathcal{G}$-geométrica estable)*

Se dice que una variable aleatoria X tal que $sop(X) \subseteq \mathbb{N}^+$ tiene una distribución $\mathcal{G}$-geométrica estable si para cualquier $p \in (0,1)$ existe $\vartheta(p)$ tal que

$$X \stackrel{d}{=} \vartheta(p) \odot_G \sum_{i=1}^{N_p} X_i$$

donde $\{X_i\}_{i=0}^{\infty}$ es una sucesión de variables aleatorias independientes e idénticamente distribuidas tal que para todo $i \in \mathbb{N}$ $X_i \stackrel{d}{=} X$, $N_p \sim Geo(p)$ (independiente de$\{X_i\}_{i=0}^{\infty}$).

Observación 4.7.3.

Las distribuciones $\mathcal{G}$-geométricas estables son $\mathcal{G}$-auto-descomponibles y sus funciones generadoras de probabilidad admiten la representación canónica

$$G(z) = (1 + \lambda \mathcal{A}(z)^{\varrho})^{-1} \tag{4.75}$$

para $\varrho \in (0, \delta_G]$ y $\lambda > 0$. Se conocerá a las distribuciones con función generadora de probabilidad (4.75) como distribuciones G-geométricas estables con exponente ϱ.

Por la Proposición (4.7.3) para cada $\eta \in (0,1)$ existe un proceso estacionario $\mathcal{G} - INAR(1)$, $\{X_t\}_{t\in\mathbb{Z}}$ con una distribución marginal $\mathcal{G}$-geométrica estable con función generadora de probabilidad dada por (4.75). La distribución marginal de la sucesión de innovación $\{Z_t\}_{t\in\mathbb{Z}}$ se obtiene resolviendo la ecuación (4.69) en términos de G_Z y utilizando (4.63) obteniéndose la función generadora de probabilidad dada por

$$G_Z(z) = \eta^{\varrho} + (1 - \eta^{\varrho})(1 + \lambda \mathcal{A}(z)^{\varrho})^{-1}, \quad \varrho \in (0, \delta_G], \lambda > 0$$

Esto implica que un proceso estacionario $\mathcal{G} - INAR(1)$ $\{X_t\}_{t\in\mathbb{Z}}$ con distribución $\mathcal{G}$-geométrica estable se puede escribir como

$$X_t = \eta \odot_G X_{t-1} + D_t E_t, \quad t \in \mathbb{Z}$$

donde $\{D_t\}_{t\in\mathbb{Z}}$ y $\{E_t\}_{t\in\mathbb{Z}}$ son sucesiones independientes de variables aleatorias independientes tales que $D_t \sim Bnlli(1-\eta^{\varrho})$ y E_t tiene la misma distribución que X_t. Además, un proceso estacionario $\mathcal{G}-INAR(1)$ con marginal $\mathcal{G}$-geométrica estable tiene media finita sólo si $\varrho = \delta_G$ y $\mathcal{B}'(1) < \infty$. Éste tiene varianza finita si $\mathcal{B}''(1) < \infty$.

Haciendo $\eta = e^{-r}$, la función generadora de probabilidad, G, de la distribución marginal de un proceso estacionario $\mathcal{G}-INAR(1)$ $\mathcal{G}$-geométrica estable satisface que para cualquier $r > 0$ existe $c(r) \in (0,1)$ tal que

$$G(z) = G(G_r(z))[c(r) + (1-c(r))G(z)] \tag{4.76}$$

Ahora se mostrará la afirmación recíproca.

Proposición 4.7.6.

Sea $G(z)$ la función generadora de probabilidad de una distribución no-degenerada sobre $\mathbb{Z}^+$. Entonces $G(z)$ es $\mathcal{G}$-geométrica estable con algún exponente $\varrho \in (0, \delta_G]$ si y sólo si para cualquier $r > 0$ existe $c(r)$ tal que (4.76) se cumple. La función $c(r)$ necesariamente es de la forma $c(r) = e^{-r\varrho}$

Demostración:

Sólo se necesita probar el ($\Rightarrow$). Reescribiendo $G(z) = (1+\psi(z))^{-1}$ se sigue de (4.76) que para cualquier $r > 0$ existe $c(r) \in (0,1)$ tal que $\psi(G_r(z)) = c(r)\psi(z)$. Utilizando el mismo argumento que se hizo en la demostración de la proposición anterior se tiene que $\psi(z) = \lambda A^{\varrho}(z)$ para algún $\varrho \in (0, \delta_G]$. La forma de $c(r)$ se sigue de la unicidad. □

Ahora, se definirá la distribución discreta de Linnik y se construirá el correspondiente proceso estacionario $\mathcal{G}-INAR(1)$.

Definición 4.7.8. *(Distribución de Linnik)*

Sea X variable aleatoria tal que $sop(X) \subseteq \mathbb{N}^+$. Se dice que X tiene distribución $\mathcal{G}$-compuesta de Linnik si su función generadora de probabilidad tiene la forma

$$G(z) = (1+\lambda\mathcal{A}(z)^{\varrho})^{-b} \tag{4.77}$$

para algún $\varrho \in (0, \delta_G]$, $\lambda > 0$ y $b > 0$

Steutel & van Harn (1993) demostraron que las distribuciones $\mathcal{G}$-compuestas de Linnik son G-auto-descomponibles y surgen como soluciones de ecuaciones de estabilidad para procesos con soporte en $\mathbb{N}^+$ e incrementos independientes y estacionarios.

De nuevo por la auto-descomponibilidad, para cada $\eta \in (0,1)$ existe un proceso estacionario $\mathcal{G} - INAR(1)$, $\{X_t\}_{t\in\mathbb{Z}}$, con distribución marginal $\mathcal{G}$-compuesta de Linnik cuya sucesión de innovación $\{Z_t\}_{t\in\mathbb{Z}}$ tiene función generadora de probabilidad dada por

$$G_Z(z) = \left(\frac{1+\lambda\eta^{\varrho}\mathcal{A}^{\varrho}}{1+\lambda\mathcal{A}^{\varrho}}\right)^{b}$$

Con un argumento de aditividad de funciones generadoras de probabilidad se puede mostrar que $\{Z_t\}_{t\in\mathbb{Z}}$ tiene la representación

$$Z \stackrel{d}{=} \sum_{i=1}^{N} (\eta^{U_i}) \odot_G W_i$$

donde $\{W_i\}_{i=0}^{\infty}$ es una sucesión de variables aleatorias independientes e idénticamente distribuidas todas con distribución $\mathcal{G}$-geométrica estable (todas con generadora de probabilidad $G(z) = (1+\lambda A^{\varrho}(z))^{-1}$), $\{U_i\}$ son variables aleatorias independientes todas con distribución $Unif(0,1)$ y $N \sim Poisson(-b\varrho Ln(\eta))$ y todas estas variables son independientes. Esto permite una interpretación de shot-noise para el proceso análoga a la del proceso $AR(p)$ dada por Lawrence(1982).

Un proceso shot-noise se define mediante

$$X(t) = \sum_{m=N(-\infty)}^{N(t)} \eta^{t-\tau_m} \odot W_m \tag{4.78}$$

donde $\{W_m\}_{m=0}^{\infty}$ son variables aleatorias independientes e idénticamente distribuidas tales que $sop(W_m) \subseteq \mathbb{N}^+$ y $\{N(t)\}_{t\geq 0}$ es un proceso Poisson con tiempos de ocurrencia τ_m. Si todas las W_m's tienen la misma función generadora de probabilidad $G(z) = (1+\lambda\mathcal{A}^{\varrho}(z))^{-1}$ y $N(t)$ tiene tasa $-b\varrho Ln(\eta)$, entonces $X(t)$ de (4.78) muestreado en $t \in \mathbb{Z}$ proporciona otra representación del proceso estacionario $\mathcal{G} - INAR(1)$ con función generadora de probabilidad $G(z) = (1+\lambda\mathcal{A}^{\varrho}(z))^{-b}$.

Un proceso estacionario $\mathcal{G} - INAR(1)$ con distribución marginal $\mathcal{G}$-compuesta de Linnik tiene media finita sólo si $\varrho\delta_G$ y $\mathcal{B}'(1) < \infty$ y tendrá varianza finita si además $\mathcal{B}''(1) < \infty$.

Ejemplo 4.7.1.

En van Harn et al. (1982) se mencionan algunos ejemplos interesantes de semigrupos de composición continua de funciones generadoras de probabilidad con los cuales se puede

generar un proceso $\mathcal{G} - INAR(1)$. Considérese la familia de semigrupos $\{\mathcal{G}^{(\theta)} : \theta \in [0,1)\}$ cuyas funciones generadoras de probabilidad están dadas por

$$G_r^{(\theta)}(z) = 1 - \frac{\bar{\theta} e^{-\bar{\theta} r}(1-z)}{\bar{\theta} + \theta(1 - e^{-\bar{\theta} r})(1-z)}, \ t \geq 0, |z| \leq 1, \bar{\theta} := 1 - \theta \tag{4.79}$$

En este caso se tiene que $\delta_{G^{(\theta)}} = \bar{\theta}$, $U_{G^{(\theta)}}(z) = (1-z)(1-\theta z)$ y $\mathcal{A}_{G^{(\theta)}}(z) = \left(\frac{1-z}{1-\theta z}\right)^{1/\bar{\theta}}$. Nótese que para $\theta = 0$, $\mathcal{G}^{(\theta)}$ corresponde al semigrupo estándar con $G_r^{(0)}(z) = 1 - e^{-r} + ze^{-r}$ y la multiplicación $\odot_{G^{(0)}}$ se convierte en el operador de adelgazamiento binomial.

El proceso $AR(1)$-Poisson de Mckenzie (1988) es el proceso estacionario $\mathcal{G}^{(0)} - INAR(1)$ con una marginal $G^{(0)}$-estable. De forma más general, el proceso $INAR(1)$-Poisson-geométrico de Aly & Bouzar (1994) surge como el proceso estacionario $G^{(\theta)} - INAR(1)$ con una marginal $G^{(\theta)}$-estable. El proceso $\mathcal{G} - INAR(1)$-Mittag-Leffler de Pillai & Jayakumar (1995) (y en particular el $INAR(1)$ geometrico de McKenzie (1986)) es el proceso estacionario $\mathcal{G}^{(0)} - INAR(1)$ con una marginal $G^{(0)}$-geométrica estable. El proceso $INAR(1)$-Linnik de Aly & Bouzar (2000) (y en particular el $INAR(1)$ binomial negativo de McKenzie (1986)) es le proceso estacionario $\mathcal{G}^{(0)} - INAR(1)$ con una marginal $\mathcal{G}^{(0)}$-Linnik.

Finalmente, Zhu & Joe (2003) utilizaron una versión reparametrizada del semigrupo $\mathcal{G}^{(\theta)}$ para construir un proceso de Markov a tiempo continuo $\{X_t\}_{t \geq 0}$ (con espacio de estados en $\mathbb{N}^+$) a través de la ecuación

$$X_t = e^{-\mu(t-s)} \odot_{G^{(\theta)}} X(s) + \epsilon(s,t), \quad s < t$$

donde $\mu > 0$ y $\epsilon(s,t)$ tiene soporte en $\mathbb{N}^+$ independiste de X_s. ∇

4.7.3. Reversibilidad en el tiempo del proceso $\mathcal{G} - INAR(1)$

Se dice que un proceso estocástico $\{X_t\}_{t \in \mathbb{Z}}$ es reversible en el tiempo si para cualesquiera $t \in \mathbb{Z}, k \in \mathbb{N}$ $(X_t, X_{t+1}, \ldots, X_t + k)$ y $(X_{t+k}, X_{t+k-1}, \ldots, X_1)$ tienen la misma distribución conjunta.

Sea $\{X_t\}_{t \in \mathbb{Z}}$ un proceso $\mathcal{G} - INAR(1)$. Por la propiedad de Markov, $\{X_t\}_{t \in \mathbb{Z}}$ es reversible en el tiempo si y sólo si para cualquier $t \in \mathbb{Z}$ (X_{t-1}, X_t) y (X_t, X_{t-1}) tienen la misma distribución conjunta. En términos de la función generadora de probabilidad conjunta, $G_{X_{t-1}, X_t}(z_1, z_2)$, de (X_{t-1}, X_t) que se define mediante

$$G_{X_{t-1}, X_t}(z_1, z_2) = \mathbb{E}(z_1^{X_{t-1}} z_2^{X_t}), \quad |z_1|, |z_2| < 1$$

$\{X_t\}_{t \in \mathbb{Z}}$ es reversible en el tiempo si y sólo si $G_{X_{t-1}, X_t}(z_1, z_2) = G_{X_{t-1}, X_t}(z_2, z_1)$ para cualesquiera $t \in \mathbb{Z}$ y $|z_1|, |z_2| < 1$.

Por (4.67) y un argumento condicional se puede demostrar que

$$G_{X_{t-1},X_t}(z_1,z_2) = G_Z(z_2)G_{X_{t-1}}(z_1G_r(z_2)), \quad r = -Ln(\eta), \eta \in (0,1)$$

Por las propiedades que se enunciaron del proceso $\mathcal{G}-INAR(1)$, un proceso $\{X_t\}_{t\in\mathbb{Z}}$ $\mathcal{G}-INAR(1)$ reversible en el tiempo (tal que $\mathbb{E}(X_t), \mathbb{E}(Z_t) < \infty$) tiene la propiedad que se conoce como de regresión lineal hacia atrás, i.e. existe $c > 0$ y $d \geq 0$ tal que para cualquier $t \in \mathbb{Z}$

$$\mathbb{E}(X_{t-1}|X_t) = d + cX_t \tag{4.80}$$

A continuación se darán las condiciones para que un proceso estacionario $\mathcal{G}-INAR(1)$, con media y varianza finita, tenga la propiedad de regresión lineal hacia atrás.

Proposición 4.7.7.

Supóngase que la distribución $\{h_n\}_{n=0}^{\infty}$ satisface

$$\sum_{n=2}^{\infty} h_n nLn(n) < \infty$$

Sea $\{X_t\}_{t\in\mathbb{Z}}$ un proceso estacionario $\mathcal{G}-INAR(1)$ con media y varianza finitas con la propiedad de regresión lineal hacia atrás (4.80). Entonces la función generadora de probabilidad $G(\cdot)$ de la distribución marginal de $\{X_t\}_{t\in\mathbb{Z}}$ tiene la forma

$$G(z) = \exp\left(-\lambda \int_z^1 \frac{\mathcal{B}(x)}{x} dx\right)$$

donde $\lambda > 0$ y $\mathcal{B}(z)$ es la función generadora de probabilidad de (4.65).

Demostración:

Sea $t \in \mathbb{N}^+$ y $G_{X_{t-1},X_t}(z_1,z_2)$ la función generadora de probabilidad de (X_{t-1}, X_t). Por (4.69) y (?) se tiene que

$$G_{X_{t-1},X_t}(z_1,z_2) = \frac{G(z_1G_r(z_2))G(z_2)}{G(G_r(z_2))}, \quad r = -Ln(\eta), \eta \in (0,1)$$

derivando $G_{X_{t-1},X_t}(z_1,z_2)$ con respecto a z_1 y haciendo $z_1 = 1$ y $z_2 = z$ se tiene que para cualquier $t \in \mathbb{Z}$

$$\mathbb{E}(X_{t-1}z^{X_t}) = \frac{G_r(z)G(z)G'(G_r(z))}{G(G_r(z))}$$

y por (4.80) se tiene que para algunas $c > 0$ y $d \geq 0$, por el Teorema de Esperanza Iterada

$$\mathbb{E}(X_{t-1}z^{X_t}) = \mathbb{E}\left[z^{X_t}\mathbb{E}(X_{t-1}|X_t)\right] = cz\mathbb{E}(X_t z^{X_t-1}) + d\mathbb{E}(z^{X_t})$$

para cualquier $t \in \mathbb{N}$. Sea $Q(z) := \frac{zG'(z)}{G(z)}$. Nótese que $\mathbb{E}(X_{t-1}z^{X_t}) = G'(x)$. Por (?) y por (?) se tiene que

$$cQ(z) + d = Q(G_r(z))$$

y entonces

$$cQ'(z) = G_r'(z)Q'(G_r(z))$$

y nótese que $Q'(1) = Var(X_t) \neq 0$; por tanto $c = G_r'(1) = e^{-\delta_G r}$ y

$$Q'(z) = e^{\delta_G r}G_r'(z)Q'(G_r(z))$$

con un argumento inductivo se tiene que

$$Q'(z) = e^{t\delta_G r}\prod_{j=0}^{t-1} G_r'(G_{jr}(z))Q'(G_{tr}(z))$$

y por la propiedad de semigrupo, junto con (4.63) se tiene que

$$G_r'(G_{jr}(z)) = \frac{U(G_{(j+1)r}(z))}{U(G_{jr}(z))}$$

para $j \in 0, \ldots, t$. Por tanto,

$$Q'(z) = e^{t\delta_G r}\frac{U(G_{tr}(z))}{U(z)}Q'(G_{tr}(z)) \tag{4.81}$$

De nuevo, por las propiedades de semigrupo se tiene que

$$\lim_{t\to\infty} G_{tr}(z) = 1, \qquad \lim_{t\to\infty}\frac{U(G_{rt}(z))}{G_{tr}(z)-1} = U'(1) = -\delta_G, \text{ y}$$

$$\lim_{t\to\infty}\frac{G_{tr}(z)-1}{G_{tr}(0)-1} = 1 - \mathcal{B}(z)$$

Además, como $\sum_{n=2}^{\infty} h_n nLn(n) < \infty$, entonces

$$\lim_{t\to\infty} e^{t\delta_G r}(G_{tr}(0)-1) = -1$$

haciendo $t \to \infty$ en (4.81) se tiene que

$$Q'(z) = \delta_G Q'(1) \frac{1 - \mathcal{B}(z)}{U(z)}$$

De las definiciones de $\mathcal{A}$ y $\mathcal{B}$ se puede ver que $\frac{1}{U(z)} = -\frac{\mathcal{A}'(z)}{\mathcal{A}(z)}$ y $1 - \mathcal{B}(z) = [\mathcal{A}(z)]^{\delta_G}$. Por tanto, $Q(0) = 0$

i.e.

$$Q(z) = \int_0^z Q'(x)dx = Q'(1)(1 - [\mathcal{A}(z)]^{\delta_G}) = Q'(1)\mathcal{B}(z)$$

lo cual implica que

$$\frac{G'(z)}{G(z)} = Q'(1)\frac{\mathcal{B}(z)}{z}$$

ó bien

$$Ln(G(z)) = -Q'(1) \int_z^1 \frac{\mathcal{B}(x)}{x} dx$$

□

Observación 4.7.4.

Se puede intercambiar la propiedad de regresión hacia atrás en esta proposición por la propiedad de reversibilidad en el tiempo (que es más fuerte) ▽

Observación 4.7.5.

Para la familia de semigrupos $\{\mathcal{G}^{(\theta)} : \theta \in [0, 1)\}$ de (4.79), la condición $\sum_{n=2}^{\infty} h_n n Ln(n) < \infty$ se satisface (pues $h_n = 0$ para $n \geq 3$). En este caso, la función generadora de probabilidad, $G(z)$, de esta proposición se convierte en

$$G(z) = \begin{cases} e^{-\lambda(1-z)} & \text{si } \theta = 0 \ (\lambda > 0) \\ \left(\frac{1-\theta}{1-\theta z}\right)^r & \text{si } 0 < \theta < 1 \ (r > 0) \end{cases}$$

La distribución Poisson (o bien la distribución binomial negativa con probabilidad de éxito θ) es la única distribución que surge como la distribución marginal de un proceso $\mathcal{G}^{(0)} - INAR(1)$ estacionario (o bien de un $\mathcal{G}^{(\theta)} - INAR(1)$) con media y varianza finitos y con la propiedad de regresión lineal hacia atrás.

4.7.4. Un proceso $\mathcal{G} - INAR(p)$

Lawrence & Lewis (1980) introdujeron el proceso auto-regresivo mixto de orden p, $AR(p)$

$$X_t = \sum_{i=1}^{p} I_{(\xi_t=i)}\eta_i X_{t-i} + \epsilon_t \tag{4.82}$$

Donde $\{\xi_t\}$ y $\{\epsilon_t\}$ son sucesiones independientes de variables aleatorias independientes e idénticamente distribuidas tales que $0 < \eta_i < 1$, $\mathbb{P}(\xi_t = i) = c_i,\ i = 1, 2, \ldots, p$ y $\sum_{i=1}^{p} c_i = 1$. Los autores obtuvieron la distribución de la variable de innovación ϵ_t para el proceso estacionario $AR(p)$ con una marginal exponencial. Pillai & Jayakumar (1994) fueron más allá y obtuvieron la distribución de ϵ_t para un proceso estacionario $AR(p)$ con marginal Mittag-Leffler sobre $\mathbb{R}^+$. Utilizando el operador binomial de Steutel & van Harn, Jayakumar (1995) hizo una definición análoga a (4.82) discreta y construyo el proceso discreto $INAR(p)$ Mittag-Leffler.

En esta sección se estudiará un proceso generalizado $INAR(p)$ usando el operador $\odot_G$. En particular, se obtendrá la distribución marginal del proceso estacionario $INAR(p)$ con una marginal $\mathcal{G}$-geométrica estable.

Definición 4.7.9.

Se dice que una sucesión de variables aleatorias $\{X_t\}_{t\in\mathbb{Z}}$ tal que $sop(X_t) \subseteq \mathbb{N}^+$ es un proceso $\mathcal{G} - INAR(p)$ si para cualquier $t \in \mathbb{Z}$

$$X_t = \sum_{i=1}^{p} I_{\{\xi_t=i\}}\eta_i \odot_G X_{t-i} + Z_t \tag{4.83}$$

donde $\{\xi_t\}_{t\in\mathbb{Z}}$ y $\{Z_t\}_{t\in\mathbb{Z}}$ son sucesiones independientes de variables aleatorias independientes idénticamente distribuidas tales que $sop(\xi_t), sop(Z_t) \subseteq \mathbb{N}^+$, $\eta_1, \ldots, \eta_p \in (0, 1)$, $\mathbb{P}(\xi_t = i) = c_i,\ i = 1, 2, \ldots, p$ y $\sum_{i=1}^{p} c_i = 1$

La multiplicación generalizada $\eta_i \odot_G X_{t-i}$ en (4.83) se realiza de forma independiente para cada i. De manera más formal, se supone la existencia de p arreglos independientes $\{Y_{i,t}^{(j)} : i \in \mathbb{N}, t \in \mathbb{Z}\}$, $j = 1, 2, \ldots, p$ de variables aleatorias independientes e idénticamente distribuidas, independientes de $\{\xi_t\}_{t\in\mathbb{Z}}$, independientes de $\{Z_t\}_{t\in\mathbb{Z}}$ tales que para cada $j \in \{1, 2, \ldots, p\}$, el arreglo tiene la misma función generadora de probabilidad G_{r_j}, $r_j = -Ln(\eta_j)$ y

$$\eta_j \odot X_{t-j} = \sum_{i=1}^{X_{t-j}} Y_{i,t-j}^{(j)}$$

En términos de funciones generadoras de probabilidad se sigue que

$$G_{X_t}(z) = \left(\sum_{i=1}^{p} c_i G_{X_{t-j}}(G_{r_i}(z)) \right) G_Z(z), \quad t_i = -Ln(\eta_i)$$

La estructura de auto-correlación del proceso estacionario $\mathcal{G} - INAR(p)$ está dada en la siguiente proposición y en su corolario.

Proposición 4.7.8.

Supóngase que $\sum_{n=2}^{\infty} n(n-1)h_n < \infty$. Sea $\{X_t\}_{t\in\mathbb{Z}}$ un proceso estacionario $\mathcal{G} - INAR(p)$ tal que $\mathbb{E}(X_0), Var(X_0) < \infty$, $\mu_Z = \mathbb{E}(Z_0) < \infty$ y $Var(Z_0) < \infty$. Sean $\{(c_i, \eta_i)\}_{i=1}^{p}$ como en (4.83). Entonces,

(i) Para cualquier $t \in \mathbb{Z}$,

$$\mathbb{E}(X_t) = \mu_Z \left(1 - \sum_{i=1}^{p} c_i \eta_i^{\delta_G} \right)^{-1}$$

(ii) La función de autocovarianza, $\gamma(k) = Cov(X_{t-k}, X_t)$, está dada por

$$\gamma(k) = \begin{cases} \sum_{i=1}^{p} c_i \eta_i^{\delta_G} \gamma(i) + \Xi & \text{si } k = 0; \\ \sum_{i=1}^{p} c_i \eta_i^{\delta_G} \gamma(k-i) & \text{si } k \in \mathbb{N}^+ \end{cases} \tag{4.84}$$

donde

$$\Xi = \sigma_Z^2 + \left(\frac{1 - U''(1)}{U'(1)} \right) \mathbb{E}(X_0) \sum_{i=1}^{p} \eta_i^{\delta_G}(1 - \eta_i^{\delta_G}) + \mathbb{E}(X_0^2) \sum_{i=1}^{p} (c_i - c_i^2) \eta_i^{2\delta_G}$$

Demostración:

La estacionariedad implica que $\gamma(k = \gamma(-k)$. Utilizando (4.83), (4.64) y el teorema de esperanza iterada se tiene que

$$\mathbb{E}[X_{t-k}(\eta_i \odot_G X_{t-k})] = \eta_i^{\delta_G} \mathbb{E}(X_{t-k} X_{t-i}), \quad i = 1, 2, \ldots, p$$

de donde se obtiene que

$$\mathbb{E}(X_t X_{t-k}) = \sum_{i=1}^{p} c_i \eta_i^{\delta_G} \mathbb{E}(X_{t-k} X_{t-i}) + \mathbb{E}(Z_t)\mathbb{E}(X_{t-k})$$

Con esta última ecuación y la expresión para $\mathbb{E}(X_t)$ se tiene la expresión para $\gamma(k)$ y como $U''(1)$ existe (pues $\sum_{n=2}^{\infty} n(n-1)h_n < \infty$) se tiene que

$$\gamma(0) = \sum_{i=1}^{p} c_i^2 \eta_i^{2\delta_G} + 2 \sum_{1 \leq i < j \leq p} c_i c_j \eta_i^{\delta_G} \eta_j^{\delta_G} \gamma(j-i) + \Xi$$

□

La sucesión $\{\gamma(k)\}_{k \in \mathbb{Z}}$ está completamente determinada por (4.84) una vez que se encuentran los $\gamma(k)$. Los $\gamma(k)$ son solución del sistema lineal (4.84) restringido a $k = 0, 1, \ldots, p$. Como $\sum_{i=1}^{p} c_i \eta_i^{\delta_G} < 1$, este sistema tiene una única solución dada por

$$\mathbf{\Gamma}_p = \Xi(\mathbf{I} - \mathbf{C})^{-1}\mathbf{c}$$

donde $\mathbf{\Gamma}_p = (\gamma(0), \gamma(1), \ldots, \gamma(p))$. $\mathbf{I} \in M_{p+1 \times p+1}(\mathbb{R})$ es la matriz identidad y $\mathbf{C} \in M_{p+1 \times p+1}(\mathbb{R})$ es la matriz cuya entrada (i, j) es igual al coeficiente de $\gamma(i)$ del lado derecho de la j-ésima ecuación de (4.84) y $i, j = 0, 1, \ldots, p$, $\mathbf{c} = (1, 0, \ldots, 0)$.

Corolario 4.7.2.

Supóngase que $\sum_{n=2}^{\infty} n(n-1)h_n < \infty$. Sea $\{X_t\}_{t \in \mathbb{Z}}$ un proceso estacionario $\mathcal{G} - INAR(p)$ tal que $\mathbb{E}(X_0), Var(X_0) < \infty$, $\mu_Z = \mathbb{E}(Z_0) < \infty$ y $Var(Z_0) < \infty$. Sean $\{(c_i, \eta_i)\}_{i=1}^{p}$ como en (4.83). Entonces para $k \in \mathbb{N}^+$ la función de autocorrelación está dada por

$$\rho_k = \sum_{i=1}^{p} c_i \eta_i^{\delta_G} \rho_{k-i}$$

donde $\rho_0 = 1$

Ahora, se determinará la distribución marginal de la sucesión de innovación de un proceso estacionario $\mathcal{G} - INAR(p)$ con una distribución marginal $\mathcal{G}$-geométrica estable (cuya función generadora de probabilidad está dada por (**??**)). Antes de hacer esto se enunuciará un lema.

Lema 4.7.1.

Para $i = 1, 2$, sea $f_\star^{(i)}(r) := \sum_{j=1}^{n_i} c_{ij} \frac{a_{ij}}{a_{ij}+r}$, $r \geq 0$, $a_{ij} > 0$, $c_{ij} \geq 0$ y $\sum_{j=1}^{n_i} c_{ij} = 1$. Si $\text{máx}_{1\leq i\leq n_1} a_{1i} < \text{mín}_{1\leq i\leq n_2} a_{2i}$, entonces existen $c'_k \geq 0$ y $d_k > 0$ tales que $\sum_{k=0}^{n_1+n_2-1} c'_k = 1$ y d_k estrictamente creciente (con los d_k's siendo los a_{1k}'s en orden creciente para $k = 1, 2, \ldots, n_1$) y

$$\frac{f_\star^{(1)}(r)}{f_\star^{(2)}(r)} = c'_0 + \sum_{k=0}^{t_1+t_2-1} c'_k \frac{d_k}{d_k + r}$$

Demostración:

Nótese que $f_\star^{(i)}$ es la transformada de Laplace-Stieltjes de una mezcla de distribuciones exponenciales. Agrupando y re-ordenado si es necesario se supondrá sin pérdida de generalidad que $0 < a_{ij} < \ldots < a_{it}$, $i = 1, 2$. Por un resultado de Feller (1971) se tiene que

$$f_\star^{(i)}(r) = C_i \frac{b_{i1}+r}{a_{i1}+r} \cdots \frac{b_{i,t_i-1}+r}{a_{i,t_i-1}+r} \frac{1}{a_{i,t_i}+r}$$

donde $0 < a_{i1} < b_{i1} < a_{i2} < b_{i2} < \ldots < b_{i,n_i-1} < a_{it_i}$ y $C_i > 0$.

El hecho de que $a_{1t_1} < a_{21}$ y $\frac{f_\star^{(1)}(0)}{f_\star^{(2)}(0)} = 1$ implican que $c'_k \geq 0$ y $\sum_{k=0}^{t_1+t_2-1} c'_k = 1$.

Proposición 4.7.9.

Si X_t es un proceso estacionario $\mathcal{G} - INAR(p)$ con marginal $\mathcal{G}$-geométrica estable (con función generadora de probabilidad dada en (**??**), entonces

$$Z_t \stackrel{d}{=} \sum_{j=0}^{p} I_{(V_t=j)} \beta_j \odot_G E_t$$

donde $\{V_t\}_{t\in\mathbb{Z}}$ y $\{E_t\}_{t\in\mathbb{Z}}$ son sucesiones independientes de variables aleatorias independientes tales que $\mathbb{P}(V_t = j) = c'_j$, $j = 0, 1, \ldots, p$, $\sum_{j=0}^{p} c'_j = 1$, E_t tiene la misma distribución que X_t, $\beta_0 = 0$, $\beta_1 = 1$ y $\beta_2, \ldots, \beta_p \in (0, 1)$

Demostración:

Usando (4.63), (**??**) y (**??**) se tiene que

$$G_Z(z) = \frac{(1+\lambda\mathcal{A}(z)^\varrho)^{-1}}{\sum_{i=1}^p c_i a_i (a_i+\lambda\mathcal{A}(z)^\varrho)^{-1}}$$

con $\lambda > 0$, $\varrho \in (0, \delta_G]$ y $a_i = \eta_i^{-\varrho} > 1$, $i = 1, 2, \ldots, p$. Aplicando el Lema anterior a esta generadora con $t_1 = 1, t_2 = p, a_{11} = 1, a_{2j} = a_j$ y $r = \lambda\mathcal{A}(z)^\varrho$ se tiene que

$$G_Z = c_0' + c_1'(1+\lambda\mathcal{A}(z)^\varrho)^{-1} + \sum_{j=2}^{p} c_j'(1+\lambda\beta_j^\varrho\mathcal{A}(z)^\varrho)^{-1} \tag{4.85}$$

□

Observación 4.7.6.

En la demostración de esta proposición se exhibe que en el caso de un proceso estacionario $\mathcal{G}-INAR(p)$ con marginal $\mathcal{G}$-geométrica estable, la solución (4.85) de (**??**) existe para el rango completo de p, $\{r_i\}$ y $\{c_i\}$. El recíproco de esta afirmación también es cierta. Es decir, si G es la función generadora de probabilidad de una distribución no degenerada tal que (4.85) es la solución de (**??**) (donde $G_{X_k}(z) = G(z)$ para cualquier k) para el rango completo de p, $\{r_i\}$ y $\{c_i\}$, entonces G necesariamente $\mathcal{G}$-geométrica estable. Esta es una consecuencia directa de la proposición (**??**) ya que cuando $p = 1$, las hipótesis se convierten en

$$G(z) = G(G_r(z))[c(r) + (1-c(r))G(z)]$$

▽

Pueden surgir mezclas finitas de distribuciones $\mathcal{G}$-geométricas estables como marginales de procesos estacionarios $\mathcal{G}-INAR(p)$. Tales distribuciones tienen función generadora de probabilidad de la forma

$$G(z) = \sum_{j=1}^{m} q_j(1+\lambda_j\mathcal{A}^\varrho(z))^{-1} \tag{4.86}$$

donde $\varrho \in (0, \delta_G]$, $\lambda_j > 0, q_j \geq 0$ y $\sum_{j=1}^m q_j = 1$. Resolviendo (**??**) para G_Z con G_X como en (4.86) se tiene que

$$G_Z(z) = \frac{\sum_{j=1}^m q_j b_j (b_j+\mathcal{A}^\varrho(z))^{-1}}{\sum_{i=1}^p \sum_{j=1}^m c_i q_j b_{ij} (b_{ij}+\mathcal{A}^\varrho(z))^{-1}}$$

donde $b_j = 1/\lambda_j$ y $b_{ij} = b_j\eta_i^{-\varrho}$. Si

$$\frac{\text{máx}_{1\leq i\leq m}\,\lambda_i}{\text{mín}_{1\leq i\leq m}\,\lambda_i} < \min_{1\leq i\leq p}\,\eta_i^{-\varrho}$$

entonces se puede aplicar directamente el lema anterior (con $t_1 = m, t_2 = p$, las a_{1k}'s son las b_k's, las a_{2k}'s son las b_{ij}'s y $r = \mathcal{A}^\varrho(z)$) demostrando así que G_Z es la función generadora

de probabilidad de una mezcla de distribuciones $\mathcal{G}$-geométricas estables y la distribución degenerada en 0.

Se puede utilizar una familia de semigrupos $\{\mathcal{G}^{(\theta)} : \theta \in [0,1)\}$ de (4.79) para construir procesos estacionarios $\mathcal{G} - INAR(p)$ con marginales $\mathcal{G}$-geometricas estables (o mezclas finitas de las mismas). Un caso particular es el proceso de Mittag-Leffler $INAR(p)$ de Jayakumar (1995) es un proeso $\mathcal{G}^{(0)} - INAR(p)$ con una marginal $\mathcal{G}$-geométrica-estable. Como se acaba de mencionar, pueden surgir mezclas finitas de distribuciones discretas Mittag-Leffler como la marginal de un proceso $\mathcal{G}^{(0)} - INAR(p)$.

Finalmente, se debe mencionar que se puede construir un proceso $\mathcal{G} - INAR(p)$ más general usando la definición de Latour (1998) del proceso $INAR(p)$. En este caso, (4.83) se convierte en

$$X_t = \sum_{i=1}^{p} \eta_i \odot_G X_{t-i} + Z_t \tag{4.87}$$

donde $\eta_1, \eta_2, \ldots, \eta_p \in (0,1)$ y $\{Z_t\}_{t\in\mathbb{Z}}$ es una sucesión de variables aleatorias tal que $sop(Z_t) \subseteq \mathbb{N}^+$ con media μ_Z y varianza σ_Z^2 finitas. Por Latour (1998) un proceso estacionario $INAR(p)$ con media finita del tipo (4.87) existe si y sólo si

$$\sum_{i=1}^{p} \eta_i^{\delta_G} < 1$$

La Proposición (??) y el Corolario (??) siguen siendo válidos haciendo unas pequeñas modificaciones, obteniéndose

$$\mathbb{E}(X_t) = \mu_Z \left(1 - \sum_{i=1}^{p} \eta_i^{\delta_G}\right)^{-1},$$

y función de autocovarianza

$$\gamma(k) = \begin{cases} \sum_{i=1}^{p} \eta_i^{\delta_G} \gamma(i) + B_1 & \text{si } k = 0; \\ \sum_{i=1}^{p} \eta_i^{\delta_G} \gamma(k-i) & \text{si } k \in \mathbb{N}^+ \end{cases}$$

donde

$$B_1 = \sigma_Z^2 + \left(1 - \frac{U''(1)}{U'(1)}\right) \mathbb{E}(X_0) \sum_{i=1}^{p} \eta_i^{\delta_G}(1 - \eta_i^{\delta_G})$$

y función de autocorrelación

$$\rho_k = \sum_{i=1}^{p} \eta_i^{\delta_G} \rho_{k-i}$$

Capítulo 5

MCMC para procesos ARMA de valores enteros

5.1. Introducción

En los capítulos anteriores ya se dieron algunas definiciones y ejemplos de modelos de series de tiempo enteras desde diferentes perspectivas. También se mencionó que algunas metodologías estudiadas en series de tiempo como los modelos ARMA (modelo autoregresivo y de medias móviles univariados) son inadecuadas ya que suponen que los datos pueden tomar cualquier valor de $\mathbb{R}$ y particulamente pierden sentido cuando se trata de datos de conteo con baja frecuencia en los cuales aproximaciones continuas de procesos discretos son bastante inapropiados.

Si bien los procesos INARMA han sido muy estudiados los últimos 30 años, son mucho menos utilizados y entendidos que los procesos ARMA. En particular, la inferencia acerca de los parámetros del proceso INARMA ha sido bastante limitada. Casi se ha restringido a los procesos $INAR(1)$ desde las perspectivas clásicas y bayesianas.

5.2. Repaso del proceso ARMA con valores enteros

Aunque ya se dió un gran catálogo de modelos de series de tiempo enteras, se repasará el modelo ARMA de valores enteros que se utilizará con el fin de restringir la atención de simulación a éste y unificar la notación que se utilizará.

Definición 5.2.1. *(Proceso de serie de tiempo $INARMA$)*

Se dice que una serie de tiempo $\{X_t\}_{t\in\mathbb{Z}}$ sigue un proceso ARMA con valores enteros con parámetros p y q ($INARMA(p,q)$) si es un proceso de medias móviles autorregresivos ($ARMA$) con parámetros p, q y para cualquier $t \in \mathbb{Z}$, $sop(X_t) \subseteq \mathbb{Z}$ que satisface la siguiente ecuación en diferencias

$$X_t = \sum_{i=1}^{p} \alpha_i \circ X_{t-i} + \sum_{j=1}^{q} \beta_j \circ Z_{t-j} + Z_t,\ t \in \mathbb{Z}$$

para algunos operadores generalizados de Steutel & van Harn $\alpha_1, \ldots, \alpha_p$ y $\beta_1, \ldots, \beta_q$; y $\{Z_t\}_{t\in\mathbb{Z}}$ son idenpendientes e idénticamente distribuidas de acuerdo con una variable aletoria no negativa que tome valores en los enteros arbitraria, Z (pero especificada) tal que $\mathbb{E}[Z^2] < \infty$. También se supondrá que todos los operadores son independientes.

Si bien los procesos $INAR(p)$ se definieron (y han sido analizados) para cualquier operador generalizado de Steutel & van Harn, por ahora se restringirá la atención a operadores binomiales.

Se utiliza el operador binomial ya que es una elección natural de un operador generalizado de Steutel & van Harn en muchas situaciones. Sin embargo, la mayoría de los procedimientos y resultados presentados se pueden extender para operadores de Steutel & Van Harn de manera muy sencilla.

Definición 5.2.2. *(Operador binomial)*

Considérese una variable aleatoria $W \geq 0$ *c.s* tal que $sop(W) \subseteq \mathbb{Z}$. Se define el operador binomial, $\gamma\circ$, para la variable aleatoria W como

$$\gamma \circ W = \begin{cases} \varsigma & \text{si } W > 0; \\ 0 & \text{si } W = 0 \end{cases}$$

donde $\varsigma \sim Bin(W, \gamma)$ con $\gamma \in [0, 1)$.

Para asegurar que el proceso $INARMA(p,q)$ sea estacionario de segundo orden se requiere que $\boldsymbol{\alpha} = (\alpha_1, \alpha_2, \ldots, \alpha_p)$ sea tal que las raíces del polinomio de grado p

$$x^p - \alpha_1 x^{p-1} - \alpha_2 x^{p-2} - \cdots - \alpha_{p-1} x - \alpha_p$$

estén dentro del círculo unitario. Sin embargo, se utilizará el criterio de que $\sum_{i=1}^{p} \alpha_i < 1$ para asegurar que el proceso sea estacionario. Dicho criterio es más fuerte pero es más fácil de verificar. Nótese que los requisitos para garantizar la estacionariedad en los procesos $INARMA(p,q)$ son análogos para garantizar la estacionariedad en los procesos $ARMA(p,q)$. Se añadirá una restricción para $\boldsymbol{\beta} = (\beta_1, \beta_2, \ldots, \beta_q)$ y esta es $\sum_{j=1}^{q} \beta_j < 1$. Para procesos $ARMA(p,q)$ univariados esta condición es suficiente para que la serie de tiempo sea invertible. Sin embargo, para procesos $INARMA(p,q)$ no se sabe si dicha condición también garantiza la invertibilidad.

Una extensión natural para los procesos $INARMA(p,q)$ es la incorporación de un término de ruido D_t. Si $\{W_t\}_{t\in\mathbb{Z}}$ denota un proceso INARMA(p,q) con ruido, entonces

$$W_t = X_t + D_t$$

donde $\{X_t\}_{t\in\mathbb{Z}}$ es un proceso $INARMA(p,q)$ y $\{D_t\}_{t\in\mathbb{Z}}$ es una sucesión de variables aleatorias independientes e ideénticamente distribuidas no negativas tal que $sop(D_t) \subseteq \mathbb{Z}$. Además se supondrá que los procesos $\{X_t\}_{t\in\mathbb{Z}}$ y $\{D_t\}_{t\in\mathbb{Z}}$ son independientes.

Una suposición particular en esta sección de implementación es que Z_t y D_t tienen distribución Poisson; sin embargo, se pudieron considerar algunas distribuciones discretas alternativas (como las que se discutieron en el capítulo anterior ó algunas opciones autodescomponibles discretas).

5.2.1. Naturaleza de los datos que se estudiarán

Evidentemente las datos en series de tiempo entera son sucesiones de números que forman una serie de conteo. Se supondrá que se observa parte de la serie de tiempo y que se cuenta con todas las observaciones para la sección de tiempo observada. Por supuesto la metodología que se desarrollará se puede generalizar para incorporar datos faltantes (al considerar los datos faltantes como parámetros adicionales que se deben incorporar al modelo). Para el proceso $INARMA(p,q)$ se supondrá que los datos, i.e. la serie de tiempo observada es $\mathbf{x} = (x_{1-r}, x_{2-r}, \ldots, x_n)$ donde $r = \text{máx}\{p,q\}$ para algún $n \in \mathbb{N}^+$. Para un proceso $INARMA(p,q)$ con ruido aditivo los datos están dados por $\mathbf{w} = (w_{1-r}, w_{2-r}, \ldots, w_n)$ para algún $n \in \mathbb{N}^+$. Sin pérdida de generalidad se supondrá que $r < n$.

5.3. Verosimilitud y algoritmos MCMC

5.3.1. Verosimilitud del proceso $INARMA(p,q)$

A partir de ahora, se supondrá que $Z \sim Poisson(\lambda)$ y además que los parámetros $p > 0$ y $q > 0$ del proceso $INARMA(p,q)$ son conocidos (el procedimiento se simplifica considerablemente si $p = 0$ ó $q = 0$)

Para una serie de tiempo observada

$$\mathbf{x} = (x_{1-r}, x_{2-r}, ..., x_n) \text{ con } r = \text{máx}(p,q)$$

se tiene interés en hacer inferencia acerca de los parámetros $\boldsymbol{\alpha}$, $\boldsymbol{\beta}$ y λ que se suponen desconocidos. La inferencia de la distribución predictiva m pasos adelante (*m-step ahead*), $\mathbf{X}_m^{pred} = (X_{n+1}, X_{n+2}, \ldots, X_{n+m})$, es trivial pues dados los parámetros, predecir es muy fácil i.e. una vez estimados los parámetros, la predicción es muy sencilla.

Con el fin de facilitar la inferencia para el proceso $INARMA(p,q)$ es necesario aumentar la información como sigue:

Para $t \in \mathbb{Z}$, sean

$$Y_{t,i} := \alpha_i \circ X_{t-i} \text{ con } i \in \{1, \ldots, p\}$$

$$V_{t,j} := \beta_j \circ Z_{t-j} \text{ con } j \in \{1, \ldots, q\}$$

entonces, para $t \in \mathbb{Z}$ se tiene que

$$Z_t = X_t - \sum_{i=1}^{p} Y_{t,i} - \sum_{j=1}^{q} V_{t,j}$$

Para $t \in \mathbb{Z}$, defínase $\mathbf{Y}_t = (Y_{t,1}, Y_{t,2}, \ldots, Y_{t,p})$, $\mathbf{Y} = (\mathbf{Y}_1, \mathbf{Y}_2, \ldots, \mathbf{Y}_n)$, $\mathbf{V}_t = (V_{t,1}, V_{t,2}, \ldots, V_{t,q})$, $\mathbf{V} = (\mathbf{V}_1, \mathbf{V}_2, \ldots, \mathbf{V}_n)$ y $\mathbf{Z}_t = (Z_1, Z_2, \ldots, Z_t)$. Para $t \in \mathbb{N}^+$ $\mathbf{y}_t = (y_{t,1}, y_{t,2}, \ldots, y_{t,p})$, $\mathbf{y} = (\mathbf{y}_1, \mathbf{y}_2, \ldots, \mathbf{y}_n)$, $\mathbf{v}_t = (v_{t,1}, v_{t,2}, \ldots, v_{t,q})$ y $\mathbf{v} = (\mathbf{v}_1, \mathbf{v}_2, \ldots, \mathbf{v}_n)$. Para $q \in \mathbb{N}^+$ sea $\mathbf{z}_{IN} = (z_{1-q}, z_{2-q}, \ldots, z_0)$, que representa los valores iniciales de Z y para $t \in \mathbb{N}^+$ sea $\mathbf{z}_t = (z_1, z_2, \ldots, z_t)$ donde $\mathbf{z}_0$ corresponde a ninguna observación.

Nótese que dado $(\mathbf{z}_{IN}, \mathbf{z}_{t-1}, \mathbf{x}_{t-1}, \boldsymbol{\alpha}, \boldsymbol{\beta}, \lambda)$ cada una de las componentes de $(\mathbf{V}_t, \mathbf{Y}_t, \mathbf{Z}_t)$ es indepeniente, donde $\mathbf{x}_s = (x_{1-r}, x_{2-r}, \ldots, x_s)$ (para $s \geq 0$). Por tanto, para $t \in \mathbb{N}^+$

$$f_t(\mathbf{v}_t, \mathbf{y}_t, z_t|\mathbf{z}_{IN}, \mathbf{z}_{t-1}, \mathbf{x}_{t-1}, \boldsymbol{\alpha}, \boldsymbol{\beta}, \lambda) = f_t(z_t|\mathbf{z}_{IN}, \mathbf{z}_{t-1}, \mathbf{x}_{t-1}, \boldsymbol{\alpha}, \boldsymbol{\beta}, \lambda)$$
$$\times \left(\prod_{j=1}^{q} f_t(v_{t,j}|\mathbf{z}_{IN}, \mathbf{z}_{t-1}, \mathbf{x}_{t-1}, \boldsymbol{\alpha}, \boldsymbol{\beta}, \lambda)\right) \left(\prod_{i=1}^{p} f_t(y_{t,i}|\mathbf{z}_{IN}, \mathbf{z}_{t-1}, \mathbf{x}_{t-1}, \boldsymbol{\alpha}, \boldsymbol{\beta}, \lambda)\right)$$
$$= \frac{\lambda^{z_t}}{z_t!} e^{-\lambda} \left(\prod_{i=1}^{p} \binom{x_{t-i}}{y_{t,i}} \alpha_i^{y_{t,i}} (1-\alpha_i)^{x_{t-i}-y_{t,i}}\right) \left(\prod_{j=1}^{q} \binom{z_{t-j}}{v_{t,j}} \beta_j^{v_{t,j}} (1-\beta_j)^{z_{t-j}-v_{t,i}}\right)$$

pero por el teorema de Bayes

$$f_t(\mathbf{v}_t, \mathbf{y}_t, z_t|x_t, \mathbf{z}_{IN}, \mathbf{z}_{t-1}, \mathbf{x}_{t-1}, \boldsymbol{\alpha}, \boldsymbol{\beta}, \lambda)$$
$$= \frac{f_t(x_t|\mathbf{v}_t, \mathbf{y}_t, z_t, \mathbf{z}_{IN}, \mathbf{z}_{t-1}, \mathbf{x}_{t-1}, \boldsymbol{\alpha}, \boldsymbol{\beta}, \lambda)}{f_t(x_t|\mathbf{z}_{IN}, \mathbf{z}_{t-1}, \mathbf{x}_{t-1}, \boldsymbol{\alpha}, \boldsymbol{\beta}, \lambda)} f_t(\mathbf{v}_t, \mathbf{y}_t, z_t|\mathbf{z}_{IN}, \mathbf{z}_{t-1}, \mathbf{x}_{t-1}, \boldsymbol{\alpha}, \boldsymbol{\beta}, \lambda)$$

lo cual implica que

$$f_t(\mathbf{v}_t, \mathbf{y}_t, z_t|\mathbf{x}, \mathbf{z}_{IN}, \mathbf{z}_{t-1}, \boldsymbol{\alpha}, \boldsymbol{\beta}, \lambda)$$
$$\propto \frac{\lambda^{z_t}}{z_t!} e^{-\lambda} \left(\prod_{i=1}^{p} \binom{x_{t-i}}{y_{t,i}} \alpha_i^{y_{t,i}} (1-\alpha_i)^{x_{t-i}-y_{t,i}}\right) \left(\prod_{j=1}^{q} \binom{z_{t-j}}{v_{t,j}} \beta_j^{v_{t,j}} (1-\beta_j)^{z_{t-j}-v_{t,i}}\right)$$

sujeto a las restricciones

$$z_t + \sum_{i=1}^{p} y_{t,i} + \sum_{j=1}^{q} v_{t,j} = x_t \quad v_{t,j} \leq z_{t-j} \ (j \in \{1, \ldots, q\}) \text{ y } y_{t,i} \leq x_{t-i} \ (i \in \{1, \ldots, p\}),$$

en otro caso $f_t(\mathbf{v}_t, \mathbf{y}_t, z_t|\mathbf{x}, \mathbf{z}_{IN}, \mathbf{z}_{t-1}, \boldsymbol{\alpha}, \boldsymbol{\beta}, \lambda) = 0$,

Y por lo tanto,

$$f(\mathbf{v}, \mathbf{y}_, \mathbf{z}_n|\mathbf{x}, \mathbf{z}_{IN}, \boldsymbol{\alpha}, \boldsymbol{\beta}, \lambda) = \prod_{t=1}^{n} f_t(\mathbf{v}_t, \mathbf{y}_t, z_t|\mathbf{x}, \mathbf{z}_{IN}, \mathbf{z}_{t-1}, \boldsymbol{\alpha}, \boldsymbol{\beta}, \lambda)$$

Sin embargo, los parámetros $\boldsymbol{\alpha}, \boldsymbol{\beta}, \lambda$ y z_{IN} son desconocidos. Por tanto, desde una perspectiva bayesiana, se le asignarán densidades *apriori* a cada uno de estos parámetros:

$$\pi(\boldsymbol{\alpha}) \propto p!, \ \alpha_i \in [0, 1), \ i \in \{1, \ldots, p\}, \ \sum_{i=1}^{p} \alpha_i < 1$$

$$\pi(\boldsymbol{\beta}) \propto q!,\ \beta_j \in [0,1),\ j \in \{1,\ldots,q\},\ \sum_{j=1}^{q} \beta_j < 1$$

$$\lambda \sim Gamma(A_\lambda, B_\lambda) \text{ para algunos } A_\lambda, B_\lambda > 0$$

Es decir, se le asignó *aprioris* no informativas para $\boldsymbol{\alpha}$ y $\boldsymbol{\beta}$ sobre el espacio parametral permisible.

Para $1-r \leq t \leq 1$, se tomará $z_t|\lambda, \mathbf{x} \sim Poisson(\lambda)$ sujeto a que $z_t \leq x_t$ no sólo por que la apriori de Z_t es Poisson con media λ sino que también por definición $z_t \leq x_t$.

Por tanto, la verosimilitud de $(\mathbf{V}, \mathbf{Y}, \mathbf{Z}_n, \mathbf{Z}_{IN}, \boldsymbol{\alpha}, \boldsymbol{\beta}, \lambda)$ dados los datos $\mathbf{x}$ está dada por

$$f(\mathbf{v}, \mathbf{y}, \mathbf{z}_n, \mathbf{z}_{IN}, \boldsymbol{\alpha}, \boldsymbol{\beta}, \lambda|\mathbf{x}) = f(\mathbf{v}, \mathbf{y}, \mathbf{z}_n|\mathbf{z}_{IN}, \boldsymbol{\alpha}, \boldsymbol{\beta}, \lambda, \mathbf{x})\pi(\boldsymbol{\alpha})\pi(\boldsymbol{\beta})\pi(\mathbf{z}_{IN}|\lambda, \mathbf{x})\pi(\lambda)$$

Y por tanto,

$$f(\mathbf{v}, \mathbf{y}, \mathbf{z}_n, \mathbf{z}_{IN}, \boldsymbol{\alpha}, \boldsymbol{\beta}, \lambda|\mathbf{x}) \propto \prod_{t=1}^{n} [\frac{\lambda^{z_t}}{z_t!} e^{-\lambda} \left(\prod_{i=1}^{p} \binom{x_{t-i}}{y_{t,i}} \alpha_i^{y_{t,i}} (1-\alpha_i)^{x_{t-i}-y_{t,i}} \right)$$
$$\times \left(\prod_{j=1}^{q} \binom{z_{t-j}}{v_{t,j}} \beta_j^{v_{t,j}} (1-\beta_j)^{z_{t-j}-v_{t,j}} \right)] \lambda^{A_\lambda} e^{-\lambda B_\lambda} \prod_{i=1}^{r} \frac{\lambda^{z_{1-i}}}{z_{i-1}!} e^{-\lambda}$$

De la ecuación anterior se tiene que las distribuciones posteriores para α_i $(i \in \{1,\ldots,p\})$, β_j $(j \in \{1,\ldots,q\})$ y λ están dadas por

$$\alpha_i|\mathbf{v}, \mathbf{y}, \mathbf{z}_n, \mathbf{z}_{IN}, \boldsymbol{\alpha}_{i-}, \boldsymbol{\beta}, \lambda, \mathbf{x} \sim Beta\left(1 + \sum_{t=1}^{n} y_{t,i}, 1 + \sum_{t=1}^{n} (x_{t-i} - y_{t,i})\right)$$

donde $\boldsymbol{\alpha}_{i-} = (\alpha_1, \ldots, \alpha_{i-1}, \alpha_{i+1}, \ldots, \alpha_p)$, sujeto a la restricción $\sum \alpha_k < 1$

$$\beta_j|\mathbf{v}, \mathbf{y}, \mathbf{z}_n, \mathbf{z}_{IN}, \boldsymbol{\alpha}, \boldsymbol{\beta}_{j-}, \lambda, \mathbf{x} \sim Beta\left(1 + \sum_{t=1}^{n} v_{t,j}, 1 + \sum_{t=1}^{n} (z_{t-j} - v_{t,j})\right)$$

donde $\boldsymbol{\beta}_{j-} = (\beta_1, \ldots, \beta_{j-1}, \beta_{j+1}, \ldots, \beta_q)$, sujeto a la restricción $\sum \beta_l < 1$, y finalmente

$$\lambda|\mathbf{v}, \mathbf{y}, \mathbf{z}_n, \mathbf{z}_{IN}, \boldsymbol{\alpha}, \boldsymbol{\beta}, \mathbf{x} \sim Gamma\left(A_\lambda + \sum_{t=1-q}^{n} z_t, B_\lambda + n + q\right)$$

5.3.2. Algoritmo MCMC

A continuación, se describirá un algoritmo MCMC para hacer la estimación de los parámetros

(Paso 1:) Actualizar $\boldsymbol{\alpha}$

Se simularán α_i utilizando el siguiente muestreo de Gibbs/rechazo:

(i) Simular de una valor propuesto para α_i de una distribución

$$Beta\left(1+\sum_{t=1}^{n} y_{t,i}, 1+\sum_{t=1}^{n}(x_{t-i}-y_{t,i})\right)$$

(ii) Si el α_i simulado satisface $\sum \alpha_k < 1$, entonces se acepta y actualiza el valor de α_i. En otro caso, se regresa al paso (i).

(Paso 2:) Actualizar $\boldsymbol{\beta}$

Se simularán β_j utilizando el siguiente muestreo de Gibbs/rechazo:

(i) Simular de una valor propuesto para β_j de una distribución

$$Beta\left(1+\sum_{t=1}^{n} v_{t,j}, 1+\sum_{t=1}^{n}(z_{t-j}-v_{t,j})\right)$$

(ii) Si el β_j simulado satisface $\sum \beta_l < 1$, entonces se acepta y actualiza el valor de β_j. En otro caso, se regresa al paso (i).

(Paso 3:) Actualizar λ

Simular λ de su distribución condicional

$$Gamma\left(A_\lambda+\sum_{t=1-q}^{n} z_t, B_\lambda+n+q\right)$$

(Paso 4:) Actualizar $(\mathbf{v}, \mathbf{y}, \mathbf{z}_n)$

En un tiempo dado, un conjunto de componentes $(\mathbf{v}_t, \mathbf{y}_t, z_t)$ ó es fijo ó es aleatorio. Se utilizará el siguiente algoritmo de Metropolis-Hastings con instrumental independiente:

(i) Para $i \in \{1,\ldots,p\}$, simular $y'_{t,i}$ de la distribución $Bin(x_{t-i}, \alpha_i)$

(ii) Para $j \in \{1,\ldots,q\}$, simular $v'_{t,j}$ de la distribución $Bin(z_{t-j}, \beta_j)$

(iii) Si $x_t < \sum_{j=1}^{q} v'_{t,j} + \sum_{i=1}^{p} y'_{t,i}$, entonces repetir los pasos (i) y (ii)

(iv) Hágase $z'_t = x_t - (\sum_{j=1}^{q} v'_{t,j} + \sum_{i=1}^{p} y'_{t,i})$.
La distribución (instrumental) de $(\mathbf{v}'_t, \mathbf{y}'_t, z'_t)$ es independiente de $(\mathbf{v}_t, \mathbf{y}_t, z_t)$ y depende sólo de $(\mathbf{x}, \mathbf{z}_{IN}, \mathbf{z}_{t-1}, \boldsymbol{\alpha}, \boldsymbol{\beta})$. Sea $q_t(\mathbf{v}'_t, \mathbf{y}'_t, z'_t|\mathbf{x}, \mathbf{z}_{IN}, \mathbf{z}_{t-1}, \boldsymbol{\alpha}, \boldsymbol{\beta})$ la densidad condicional instrumental para $(\mathbf{v}'_t, \mathbf{y}'_t, z'_t)$ dado $(\mathbf{x}, \mathbf{z}_{IN}, \mathbf{z}_{t-1}, \boldsymbol{\alpha}, \boldsymbol{\beta})$.

(v) Calcular la probabilidad de aceptación

$$A = \text{mín}\left\{\frac{f(\mathbf{v}', \mathbf{y}', \mathbf{z}'_n|\mathbf{x}, \mathbf{z}_{IN}, \boldsymbol{\alpha}, \boldsymbol{\beta}, \lambda) q_t(\mathbf{v}_t, \mathbf{y}_t, z_t|\mathbf{x}, \mathbf{z}_{IN}, \mathbf{z}_{t-1}, \boldsymbol{\alpha}, \boldsymbol{\beta})}{f(\mathbf{v}, \mathbf{y}, \mathbf{z}_n|\mathbf{x}, \mathbf{z}_{IN}, \boldsymbol{\alpha}, \boldsymbol{\beta}, \lambda) q_t(\mathbf{v}'_t, \mathbf{y}'_t, z'_t|\mathbf{x}, \mathbf{z}_{IN}, \mathbf{z}_{t-1}, \boldsymbol{\alpha}, \boldsymbol{\beta})}, 1\right\}$$

Nótese que para $s \neq t$, $(\mathbf{v}'_s, \mathbf{y}'_s, z'_s) = (\mathbf{v}_s, \mathbf{y}_s, z_s)$. Por tanto, en la distribución instrumental que se escogió en los pasos (i) y (ii) la mayoría de los términos de las probabilidad A se cancelan y por tanto

$$A = \text{mín}\left\{\frac{z_t!}{z'_t!}\lambda^{z'_t - z_t} B_t, 1\right\}$$

donde $B_n = 1$ para $t = n$ y para $t \in \{1, \ldots, n-1\}$ se tiene que

$$B_t = \begin{cases} \prod_{j=1}^{q \wedge (n-t)} \frac{z'_t!(z_t - v_{t+j,j})!}{z_t!(z'_t - v'_{t+j,j})!}(1-\beta_j)^{z'_t - z_t} & \text{si } z'_t \geq v_{t+j,j} \ (1 \leq j \leq \text{mín}\{q, n-t\}); \\ 0 & \text{en otro caso} \end{cases}$$

Por tanto, si se acepta un movimiento entonces hágase $(\mathbf{v}_t, \mathbf{y}_t, z_t)$ igual $(\mathbf{v}'_t, \mathbf{y}'_t, z'_t)$, en otro caso $(\mathbf{v}_t, \mathbf{y}_t, z_t)$ permanece sin cambios.

(Paso 5:) Actualizar $\mathbf{z}_{IN}$

En un tiempo dado, una componente es fija ó es aleatoria. Para actualizar z_t se utilizará el siguiente algoritmo, (para $t \in \{1-q, \ldots, 0\}$

(i) Simular z'_t de la distribución $Poisson(\lambda)$

(ii) Si $z'_t > x_t$, repítase el paso (i); Por tanto, la probabilidad instrumental es

$$q(z'_t|\lambda, \mathbf{x}) = \frac{\pi(z'_t|\lambda)}{F_\lambda(x_t)}$$

donde $F_\lambda(\cdot)$ es la función de distribución de una variable aleatoria con distribucion Poisson con media λ.

(iii) Calcular la probabilidad de aceptación

$$\begin{aligned} A &= \text{mín}\left\{\frac{f(\mathbf{v}, \mathbf{y}, \mathbf{z}_n|\mathbf{x}, \mathbf{z}'_{IN}, \boldsymbol{\alpha}, \boldsymbol{\beta}, \lambda)\pi(z'_t|\lambda)q(z_t|\lambda, \mathbf{x})}{f(\mathbf{v}, \mathbf{y}, \mathbf{z}_n|\mathbf{x}, \mathbf{z}_{IN}, \boldsymbol{\alpha}, \boldsymbol{\beta}, \lambda)\pi(z_t|\lambda)q(z'_t|\lambda, \mathbf{x})}, 1\right\} \\ &= \begin{cases} \text{mín}\left\{\prod_{j=1-t}^{q} \frac{z'_t!(z_t - v_{t+j,j})!}{z_t!(z'_t - v'_{t+j,j})!}(1-\beta_j)^{z'_t - z_t}, 1\right\} & \text{si } z'_t \geq v_{t+j,j} \ (1-t \leq j \leq q); \\ 0 & \text{en otro caso} \end{cases} \end{aligned}$$

Es decir, se describió un algoritmo MCMC para obtener muestras de las distribuciones posteriores de los parámetros del modelo $(\boldsymbol{\alpha}, \boldsymbol{\beta}, \lambda)$ y los datos aumentados $(\mathbf{v}, \mathbf{y}, \mathbf{z}_n, \mathbf{z}_{IN})$. Generalmente se tiene mayor atención en la distribución posterior de los parámetros del modelo $\pi(\boldsymbol{\alpha}, \boldsymbol{\beta}, \lambda|\mathbf{x})$ (que se puede obtener fácilmente con el algorimo que se describió).

En el análisis de series de tiempo generalemte se tiene interés en las propiedades predictivas del modelo. Desde una perspectiva Bayesiana, se tiene interés en la distribución predictiva m-pasos hacia adelante, $\mathbf{X}_m^{pred}$, i.e. en $\pi(\mathbf{x}_m^{pred}|\mathbf{x})$, donde $\mathbf{x}_m^{pred} = (x_{n+1}, x_{n+2}, \ldots, x_{n+m})$. Se puede ver que

$$\begin{aligned}\pi(\mathbf{x}_m^{pred}|\mathbf{x}) &= \int_0^1 \int_0^1 \int_0^\infty \sum_{\mathbf{k}} \sum_{\mathbf{l}} \pi(\mathbf{x}_m^{pred}|\boldsymbol{\alpha}, \boldsymbol{\beta}, \lambda, \mathbf{z}_{IN} = \mathbf{k}, \mathbf{z}_n = \mathbf{l}, \mathbf{x}) \\ &\quad \times \pi(\boldsymbol{\alpha}, \boldsymbol{\beta}, \lambda, \mathbf{z}_{IN} = \mathbf{k}, \mathbf{z}_n = \mathbf{l}|\mathbf{x}) d\lambda d\boldsymbol{\beta} d\boldsymbol{\alpha}\end{aligned}$$

Por tanto, para obtener muestras de $\mathbf{X}_m^{pred}$, el algoritmo MCMC se puede extender de la siguiente forma:

(Paso 6:) Actualizar $\mathbf{X}_m^{pred}$

Se actualizará una componente a la vez, desde x_{n+1} hasta x_{n+m}. Entonces para $k \in \{1, \ldots, m\}$, simúlese z_{n+k} de la distribución $Poisson(\lambda)$, $y_{n+k,i}$ de la distribución $Bin(x_{n+k-i}, \alpha_i)$ $(i \in \{1, \ldots, p\})$ y $v_{n+k,j}$ de la distribución $Bin(z_{n+k-j}, \beta_j)$ $(j \in \{1, \ldots, q\})$. Finalmente, hágase

$$x_{n+k} = z_{n+k} + \sum_{i=1}^{p} y_{n+k,i} + \sum_{j=1}^{q} v_{n+k,j}$$

Por tanto, se obtuvo una observación (z_{n+k}, x_{n+k}) de la distribución

$$\pi(z_{n+k}, x_{n+k}|\boldsymbol{\alpha}, \boldsymbol{\beta}, \lambda, \mathbf{z}_n, \mathbf{z}_{IN}, \mathbf{x}, \mathbf{x}_{k-1}^{pred})$$

Entonces, con el paso 6 de este algortimo se puede obtener una muestra de la distribución posterior predictiva de $\mathbf{X}_m^{pred}$, $\pi(\mathbf{x}_m^{pred}|\mathbf{x})$.

5.3.3. El caso del proceso $INAR(p)$ con ruido aditivo

Ahora, se analizará un proceso $INAR(p)$ con ruido Poisson, $D_t \sim Poisson(\mu)$. La principal diferencia entre el proceso $INAR(p)$ (estándar ó sin ruido) es que $\mathbf{x} = (x_{1-p}, x_{2-p}, \ldots, x_n)$ del proceso $INAR$ NO se observa, por tanto se aumentarán los datos para incluir $\mathbf{x}$. Nótese

que dado $\mathbf{x}$, se obtiene $\mathbf{d} = (d_{1-p}, d_{2-p}, \ldots, d_n)$ pues $w_t = x_t + d_t$.

Supóngase que $\boldsymbol{\alpha}$, λ, μ y $\mathbf{x}$ son conocidos. Entonces, para $t \in \mathbb{N}^+$

$$f_t(d_t, \mathbf{y}_t, z_t|\mathbf{w}, \mathbf{x}, \boldsymbol{\alpha}, \lambda, \mu) \propto \frac{\lambda^{z_t}}{z_t!} e^{-\lambda} \frac{\mu^{d_t}}{d_t!} e^{-\mu} \prod_{i=1}^{p} \binom{x_{t-i}}{y_{t,i}} \alpha_i^{y_{t,i}} (1-\alpha_i)^{x_{t-i}-y_{t,i}}$$

sujeto a las restricciones

$$d_t = w_t - x_t, \quad \sum_{i=1}^{p} y_{t,i} + z_t = x_t, \quad y_{t,i} \leq x_{t-i} \ (i \in \{1, \ldots, p\}).$$

y en otro caso

$f_t(d_t, \mathbf{y}_t, z_t|\mathbf{w}, \mathbf{x}, \boldsymbol{\alpha}, \lambda, \mu) = 0$. Por tanto,

$$f(\mathbf{d}, \mathbf{y}, \mathbf{z}_n|\mathbf{w}, \mathbf{x}, \boldsymbol{\alpha}, \lambda, \mu) = \prod_{t=1}^{n} f_t(d_t, \mathbf{y}_t, z_t|\mathbf{w}, \mathbf{x}, \boldsymbol{\alpha}, \lambda, \mu)$$

donde $\pi(\boldsymbol{\alpha}) \propto 1; 0 \leq \alpha_i < 1 \quad (i \in \{1, \ldots, p\}), \quad \sum_{i=1}^{p} \alpha_i < 1$ y $\pi(\lambda)$ es la densidad $Gamma(A_\lambda, B_\lambda)$ para algunos $A_\lambda, B_\lambda > 0$. También convenientemente (por ser conjugada) se dira que $\pi(\mu)$ es la densidad $Gamma(A_\mu, B_\mu)$ para algunos $A_\mu, B_\mu > 0$, por tanto la distribución posterior $\pi(\mu|\mathbf{w}, \mathbf{d}, \mathbf{x}, \mathbf{y}, \mathbf{z}_n, \boldsymbol{\alpha}, \lambda)$ es

$$Gamma\left(A_\mu + \sum_{t=1-p}^{n} d_t, B_\mu + n + p\right)$$

El algoritmo MCMC descrito en la sección 5.3.2 se puede adaptar fácilmente para incluir el término de ruido. Los parámetros $(\boldsymbol{\alpha}, \boldsymbol{\lambda}, \mathbf{y}, \mathbf{z}_n)$ se actualizan exactamente como en la sección 5.3.2 usando la información aumentada $\mathbf{x}$ (y con $q = 0$). Por lo tanto, sólo se requiere el procedimiento para actualizar los parámetros $(\mathbf{x}, \mathbf{d}, \mu)$. Sean $\mathbf{x}_n = (x_1, x_2, \ldots, x_n)$, $\mathbf{d}_n = (d_1, d_2, \ldots, d_n)$, $\mathbf{x}_{IN} = (x_{1-p}, x_{2-}, \ldots, x_0)$ y $\mathbf{d}_{IN} = (d_{1-p}, d_{2-}, \ldots, d_0)$. Por tanto, $\mathbf{x} = (\mathbf{x}_{IN}, \mathbf{x}_n)$ y $\mathbf{d} = (\mathbf{d}_{IN}, \mathbf{d}_n)$. El siguiente algoritmo se debe agregar después del quinto paso del algoritmo MCMC de la sección 5.3.2.

(Paso 1:) Actualizar $(\mathbf{x}_n, \mathbf{d}_n, \mathbf{z}_n)$

Se actualizará una terna (x_t, d_t, z_t) a la vez, ya sea en un orden fijo ó aleatorio. Se utilizará el siguiente algoritmo de Metropolis-Hastings:

(i) Simúlese d_t' de la distribución $Poisson(\mu)$ y hágase $x_t' = w_t - d_t'$.

(ii) Si $x_t' < \sum_{i=1}^{p} y_{t,i}$, entonces repítase el paso (i).

(iii) Hágase $z'_t = x'_t - \sum_{i=1}^{p} y_{t,i}$. Nótese que la distribución propuesta para (x'_t, d'_t, z'_t) depende sólo de μ; por tanto se denotará por $q(x'_t, d'_t, z'_t|\mu)$ para la densidad propuesta.

(iv) Calcúlese la probabilidad de aceptación, A, de movimiento:

$$A = \text{mín}\left\{\frac{f(\mathbf{d}', \mathbf{y}, \mathbf{z}'_n|\mathbf{w}, \mathbf{x}', \boldsymbol{\alpha}, \lambda, \mu)q(d_t, z_t, x_t|\mu)}{f(\mathbf{d}, \mathbf{y}, \mathbf{z}_n|\mathbf{w}, \mathbf{x}, \boldsymbol{\alpha}, \lambda, \mu)q(d'_t, z'_t, x'_t|\mu)}, 1\right\}$$

En la expresión anterior la mayoría de los terminos del numerados y denominador se cancelan, obteniéndose

$$A = \text{mín}\{1, \frac{z_t!}{z'_t!}\lambda^{z'_t - z_t}B_t\}$$

donde $B_n = 1$ y para $t \in \{1, \ldots, n-1\}$,

$$B_t = \begin{cases} \text{mín}\left\{\prod_{i=1}^{\text{mín}\{p, n-t\}} \frac{x'_t!(x_t - y_{t+i,i})!}{x_t!(x'_t - y_{t+i,i})!}(1-\alpha_i)^{x'_t - x_t}, 1\right\} & \text{si } x'_t \geq y_{t+i,i} \ (1 \leq i \leq \text{mín}\{p, n-t\}); \\ 0 & \text{en otro caso} \end{cases}$$

Por tanto, si se acepta el movimiento hágase (x_t, d_t, z_t) igual a (x'_t, d'_t, z'_t) y en otro caso dejar a (x_t, d_t, z_t) sin cambios.

(Paso 2:) Actualizar $(\mathbf{x}_{IN}, \mathbf{d}_{IN})$

Se actualizará una pareja (x_t, d_t) a la vez, ya sea en un orden fijo ó aleatorio. Se utilizará el siguiente algoritmo:

(i) Simúlese d'_t de la distribución $Poisson(\mu)$ y hágase $x'_t = w_t - d'_t$.

(ii) Si $x'_t < 0$, entonces repítase el paso (i). Sea $q(d'_t, x'_t|\mu)$ la distribución propuesta.

(iii) Calcúlese la probabilidad de aceptación, A, de movimiento:

$$A = \text{mín}\left\{\frac{f(\mathbf{d}', \mathbf{y}, \mathbf{z}_n|\mathbf{w}, \mathbf{x}', \boldsymbol{\alpha}, \lambda, \mu)q(d_t, x_t|\mu)}{f(\mathbf{d}, \mathbf{y}, \mathbf{z}_n|\mathbf{w}, \mathbf{x}, \boldsymbol{\alpha}, \lambda, \mu)q(d'_t, x'_t|\mu)}, 1\right\}$$

Ésta se reduce a

$$A = \begin{cases} \text{mín}\left\{\prod_{i=1-t}^{p} \frac{x'_t!(x_t - y_{t+i,i})!}{x_t!(x'_t - y_{t+i,i})!}(1-\alpha_i)^{x'_t - x_t}, 1\right\} & \text{si } x'_t \geq y_{t+i,i} \ (1 - t \leq i \leq p); \\ 0 & \text{en otro caso} \end{cases}$$

Por tanto, si se acepta el movimiento hágase (x_t, d_t) igual a (x'_t, d'_t) y en otro caso dejar a (x_t, d_t) sin cambios.

(Paso 3:) Actualizar μ

Simular μ de la distribución

$$Gamma\left(A_\mu + \sum_{t=1-p}^{n} d_t, B_\mu + n + p\right)$$

Entonces, se obtendrán muestras de las disribuciones posteriores (predictivas) de $\mathbf{W}_m^{pred} = (W_{n+1}, W_{n+1}, \ldots, W_{n+m})$, $\mathbf{X}_m^{pred}$ ó de ambas ya es muy sencillo y consiste en agregar el siguiente paso del algoritmo MCMC anterior.

(Paso 4:) Actualizar $(\mathbf{x}_m^{pred}, \mathbf{w}_m^{pred})$

Se actualizará una pareja de componentes a la vez, desde (x_{n+1}, w_{n+1}) hasta (x_{n+m}, w_{n+m}). Entonces para $k \in \{1, \ldots, m\}$, simúlese z_{n+k} de la distribución $Poisson(\lambda)$, $y_{n+k,i}$ de la distribución $Bin(x_{n+k-i}, \alpha_i)$ $(i \in \{1, \ldots, p\})$ y d_{n+k} de la distribución $Poisson(\mu)$. Finalmente, hágase

$$(x_{n+k}, w_{n+k}) = \left(z_{n+k} + \sum_{i=1}^{p} y_{n+k,i}, z_{n+k} + \sum_{i=1}^{p} y_{n+k,i} + d_{n+k}\right)$$

5.4. Algoritmos para determinar el orden

En general, los algoritmos MCMC son una herramienta útil cuando se estudian procesos $INARMA$ ya que la inferencia utiliza *data augmentation*; pues la verosimilitud para los parámetros del modelo es poco manejable, pero seleccionando cuidadosamente información extra la verosimilitud se vuelve más tratable y se hace un poco más directo el análisis inferencial.

Se estudiarán dos metodologías que consideran a p y a q como parámetros del modelo. La primera, es una extensión del algoritmo de la sección 5.3.2. Ya que el número de parámetros en el modelo depende de p y q, es necesario desarrollar un algoritmo MCMC que tenga movimientos entre diferentes espacios parametrales. Esto se hace mediante un algoritmo MCMC de saltos reversibles ($RJMCMC$) (este algoritmo se define a detalle en Green(1995)). El principal inconveniente que se presenta en el uso de los algoritmos $RJMCMC$ con movimientos entre diferentes espacios parametrales es que la probabilidad de tales movimientos puede ser muy baja. Se evitará este tipo de problemas analizando un algoritmo $RJMCMC$ muy eficiente, aprovechando la estructura del modelo $INARMA(p, q)$. En particular, se identificará un parámetro particular que se puede estimar directamente de los datos y utilizar este parámetro para construir movimientos eficientes entre espacios

parametrales.

La segunda metodología utiliza *data augmentation* en un contexto de estimación máximo verosímil a través de un algoritmo EM. Este algoritmo EM se puede utilizar para p y q fijos y comparando la verosimilitud correspondiente mediante algún criterio de selección de modelos como el BIC (tal como se hace en el caso del $ARMA$ estándar). Desafortunadamente, dicho algoritmo EM sólo se puede construir para procesos $INAR(p)$ (con la especificación de Du &Li), aunque una posible solución para el caso general de $INARMA(p,q)$ es utilizar un algoritmo EM de Monte-Carlo. Sin embargo en este caso no se podrá calcular la verosimilitud correspondiente, luego entonces, una comparación de modelo a través del criterio de información NO es posible. Se discutirá cómo el algoritmo $RJMCMC$ se puede utilizar para determinar con cierto grado de precisión un parámetro inicial para el algoritmo EM.

5.4.1. Algoritmo MCMC con saltos reversibles

Ahora, el algoritmo que se describió en la sección 5.3.2 se utilizará como sub-algoritmo para los movimientos dentro de un modelo (para p, q fijos). Ahora se discutirá acerca del paso de cambio de orden. En cada iteración, se propone incrementar ó dismininuir el orden por 1 unidad al orden del AR, donde cada movimiento (ya sea descendente o ascendente) se hace con probabilidad 0.5. Esto es sujeto a las restricciones de que p esté entre $p = 0$ y p_{max}. Se hará lo mismo con el orden MA con restricciones entre $q = 0$ y q_{max}.

Antes de estudiar un algoritmo de determinación del orden se considerará el modelo $INARMA(p,q)$ y se describirán las principales características de este algoritmo $RJMCMC$. El $RJMCMC$ es una herramienta muy útil; sin embargo, frecuentemente hace gran cantidad de movimientos de cambio de modelo ineficientes, i.e. la probabilidad de aceptar movimientos entre diferentes modelos (órdenes) puede ser extremadamente baja. Brooks *et al.* (2003) consideran que no hay una forma genérica de construir algoritmos $RJMCMC$ eficientes en este sentido. Sin embargo, un buen entendimiento del modelo puede hace que se desarrollen algoritmos $RJMCMC$ eficientes.

Nótese que si la serie de tiempo es estacionaria entonces

$$\mathbb{E}(X_t) = \sum_{i=1}^{p} \mathbb{E}(X_{t-i}) + \sum_{j=1}^{q} \mathbb{E}(Z_{t-j}) + \mathbb{E}(Z_{t-j})$$

donde $\mathbb{E}(X_t) = \mathbb{E}(X_s)$ para cualesquiera $s, t \in \mathbb{Z}$. Además si $Z \sim Poisson(\lambda)$ entonces

$$\mathbb{E}(X_t) = \lambda \frac{1 + \sum_{j=1}^{q} \beta_j}{1 - \sum_{i=1}^{p} \alpha_i} =: \kappa$$

El punto clave que se aprovachará es que κ se puede estimar de forma adecuada a partir de los datos mediante $\hat{\kappa} = \frac{1}{n}\sum_{i=1}^{n} x_i$. Por tanto, aunque la información respecto a $(\boldsymbol{\alpha}, p)$, $(\boldsymbol{\beta}, q)$ y λ no está disponible directamente, los datos son extremadamente informativos con respecto a κ. Además, con el algoritmo que se estudió en la sección 5.3.2 se obtienen observaciones independientes e idénticamente distribuidas de la distribución posterior de κ. Entonces, una sugerencia natural cuando se construye el paso de "cambio" de orden se querrá mantener fija a κ. Esto se puede implementar muy fácilmente siempre que no se esté proponiendo un movimiento de un orden p (q) del AR (del MA) ya sea hacia ó desde $p = 0$ $(q = 0)$. Es decir, el procedimiento es muy simple cuando se proponen movimientos de orden en el mismo modelo, ya sea en la clase $INAR(p)$, $INMA(q)$ o $INARMA(p, q)$. Los movimientos entre diferentes clases de modelos es un poco más delicada.

5.4.1.1. Sólo movimientos auto-regresivos

Se empezará describiendo el algoritmo para el orden AR ya sea creciente o decreciente, donde ni p ni p' (el nuevo orden propuesto) son 0. Con el fin de hacer transiciones entre los subespacios parametrales que satisfagan la condición de balance detallado en los algoritmos con saltos reversibles (ver capítulo 1), se utilizará una mapeo invertible determinístico que cumpla con la suposición de "*dimension matching*" de Green (1995). Este mapeo de separación-amalgamiento se describe a continuación:

Considérese el movimiento de p a $p' = p+1$. Una forma de mantener fija a κ es seleccionar $\boldsymbol{\alpha}'$ de tal forma que

$$\sum_{i=1}^{p'} \alpha'_i = \sum_{i=1}^{p} \alpha_i$$

con $\boldsymbol{\beta}' = \boldsymbol{\beta}$ y $\lambda = \lambda'$. Esto se puede implementar considerando $U \sim Unif[0, 1]$ y seleccionando aleatoriamente (de manera uniforme) K de entre $\{1, 2, \ldots, p\}$. Entonces, para $i \in \{1, 2, \ldots, p\}/K$, hágase $\alpha'_i = \alpha_i$. Sea $\alpha_K = U\alpha_K$ y $\alpha'_{p+1} = (1-U)\alpha_K$. Por tanto, se está "dividiendo" el K-ésimo término AR y esto se debe hacer acompañado de una "división" similar en los correspondietes términos de datos aumentados, i.e. para $1 \leq t \leq n$ considérese $S_t \sim Bin(y_{t,K}, U)$, y hágase $y'_{t,K} = S_t$, $y'_{t,p+1} = y_{t,K} - S_t$ y $\mathbf{S} = (S_1, S_2, \ldots, S_n)$. Todos los otros datos aumentados permanecen fijos.

Para el movimiento inverso, i.e. de $p+1$ a $p' = p$. Selecciónese aleatoriamente (de forma uniforme) K de entre $\{1, \ldots, p\}$. Hágase $\alpha'_K = \alpha_k + \alpha_{p+1}$ y para $1 \leq t \leq n$, hágase $y'_{t,K} = y_{t,K} + y_{t,p+1}$. Es decir, se está amalgamando el $(p+1)$-ésimo término AR en el K-ésimo término AR. Por tanto, haciendo $\mathbf{w} = (\mathbf{y}, \mathbf{v}, \mathbf{z})$ (los datos aumentados) y $\theta = (\boldsymbol{\alpha}, p, \boldsymbol{\beta}, q, \lambda)$ (los parámetros), se tiene que la probabilidad de aceptación de movimiento a incrementar orden es

$$A = \text{mín}\left\{1, \frac{f(\theta', \mathbf{w}'|\mathbf{x})}{f(\theta, \mathbf{w}|\mathbf{x})} \frac{q(\theta', \mathbf{w}' \to \theta, \mathbf{w})}{q(\theta, \mathbf{w} \to \theta', \mathbf{w}')} |J|\right\}$$

donde J denota el Jacobiano de la transformación de $(\theta, \mathbf{w}, U, K, S)$ a $(\theta', \mathbf{w}', K)$.

Entonces,

$$\frac{f(\theta', \mathbf{w}'|\mathbf{x})}{f(\theta, \mathbf{w}|\mathbf{x})} = \frac{\pi(p')p'}{\pi(p)} \times \prod_{t=1}^{n} \frac{\binom{x_{t-K}}{y'_{t,K}} (\alpha'_K)^{y_{t,K'}} (1-\alpha'_K)^{x_{t-K}-y_{t,K'}} \binom{x_{t-p'}}{y'_{t,p'}} (\alpha'_{p'})^{y'_{t,p'}} (1-\alpha'_{p'})^{x_{t-p'}-y'_{t,p'}}}{\binom{x_{t-K}}{y_{t,K}} (\alpha_K)^{y_{t,K}} (1-\alpha_K)^{x_{t-K}-y_{t,K}}}$$

Para el movimiento de $(\theta, \mathbf{w})$ a $(\theta', \mathbf{w}')$ se tiene que

$$q(\theta, \mathbf{w} \to \theta', \mathbf{w}') = \frac{1}{2} \cdot \frac{1}{p} \cdot 1 \cdot \prod_{t=1}^{n} \binom{y'_{t,K}}{y_{t,K}} U^{y_{t,K}} (1-U)^{y_{t,K}-y'_{t,K}}$$

donde los términos corresponden al incrementar el orden, K, U y $\mathbf{S}$, respectivamente. El movimiento inverso sólo depende de la probabilidad de decrecimiento del orden y la selección de K y está dada por

$$q(\theta', \mathbf{w}' \to \theta, \mathbf{w}) = \frac{1}{2} \cdot \frac{1}{p}$$

Por último, el Jacobiano se puede factorizar en $n+1$ partes correspondientes al mapeo en $(\boldsymbol{\alpha}, U) \mapsto \boldsymbol{\alpha}'$ y los mapeos $(y_{t,K}, S_t) \mapsto (y'_{t,K}, y'_{t,p+1})$ $(t \in \{1, \ldots, n\})$, obteniéndose $|J| = a_K| \cdot 1^n = a_K$.

Claramente el movimiento decreciente de p a $p' = p - 1$ es simplemente el inverso del anterior.

5.4.1.2. Sólo movimientos entre medias móviles

Para el paso de cambio en el orden MA, el procedimiento es esencialmente idéntico al que se dió, excepto que se tienen que combinar y separar los términos MA. En particular, cuando se cambia el orden se asegura que $\boldsymbol{\beta}$ sea tal que

$$\sum_{j=1}^{q'} \beta'_j = \sum_{j=1}^{q} \beta_j$$

con $\boldsymbol{\alpha} = \boldsymbol{\alpha}'$ y $\lambda = \lambda$. Para el movimiento de incremento de orden de q a $q' = q + 1$, se selecionará aleatoriamente (de manera uniforme) K de entre $\{1, \ldots, q\}$ y $U \sim Unif[0, 1]$, dividiendo el K-ésimo término en $\beta'_K = U\beta_K$ y $\beta'_{q+1} = (1-U)\beta_K$. Entonces, la "división" de los datos aumentados correspondientes se hace considerando $S_t \sim Bin(v_{t,K}, U)$ $(1 \leq t \leq n)$ y haciendo $v'_{t,K} = S_t$, $v'_{t,q+1} = v_{t,K} - S_t$ y $S = (S_1, S_2, \ldots, S_n)$. El movimiento de orden decreciente de q simplemente considera el amalgamiento del q-ésimo término con uno de los órdenes bajos de los términos MA (seleccionado aleatoriamiente) y combinando los coeficientes β y los términos de datos aumentados de la forma natural.

5.4.1.3. Movimientos auto-regresivos y de medias móviles

Por último, se considerará el algoritmo para moverse entre el $INARMA(p, q)$ y los submodelos $INAR(p)$ e $INMA(q)$. De nuevo se mantendrá a κ fija pero no se puede aplicar el algortimo de separación/amalgamiento que se analizó antes. Se considerará el movimiento de un $INAR(p)$ a un $INARMA(p, 1)$. En este caso $\boldsymbol{\alpha}$ será fija pero λ variará. Sea $U \sim Unif[0, 1]$ y hágase $\beta'_1 = U$ y $\lambda' = \frac{\lambda}{1+U}$. La actualización de los términos aumentados es secuencial con $S_t \sim Bin(min\{z_t, z'_{t-1}\}, \frac{U}{1+U})$, $v'_{t,1} = S_t$ y $z'_t = z_t - S_t$ $(1 \leq t \leq n)$. En el movimiento inverso natural de un $INARMA(p, 1)$ a un $INAR(p)$ se hace con $\lambda' = \lambda(1+\beta_1)$ y combinando los términos de datos aumentados mediante $z'_t = z_t + v_{t,1}$ $(1 \leq t \leq n)$.

El movimiento entre $INMA(q)$ a $INARMA(1, q)$ es un poco más intrincado. En este caso, $\boldsymbol{\beta}$ permanece fijo y λ y $(\mathbf{v}, \mathbf{z})$ varían. Sea $U \sim Unif[0, 1]$ y hágase $\alpha'_1 = U$ y $\lambda' = \lambda(1 - U)$. Para $1 \leq t \leq n$ y æ $\in \{0, 1, \ldots, q\}$ hágase $S_{t,j} \sim Bin(v_{t,j}, U)$ con la convención de que $v_{t,0} := z_t$. También hágase $y'_{t,1} = \sum_{j=0}^{q} S_{t,j}$ y $v'_{t,j} = v_{t,j} - S_{t,j}$ $(j \in \{0, 1, \ldots, q\})$. Por tanto, considerando el hecho de que si λ decrece entonces v y z lo hacen (para un β fijo). Para el movimiento inverso de $INARMA(1, q)$ a $INMA(q)$, hágase $\lambda' = \frac{\lambda}{1-\alpha_1}$ manteniendo a $\boldsymbol{\beta}$ fijo. Nótese que $\mathbf{S}_t = (S_{t,0}, S_{t,1}, \ldots, S_{t,q}) \sim Multinom(y_{t-1}, \gamma_0, \gamma_1, \ldots, \gamma_q)$ donde $\boldsymbol{\gamma} = (\gamma_0, \gamma_1, \ldots, \gamma_q)$ y $\gamma_k = \frac{\beta_k}{\sum_{j=0}^{q} \beta_j}$ $(k \in \{0, 1, \ldots, q\})$ con la convención de que $\beta_0 := 1$. Finalmente, para $t \in \{1, \ldots, n\}$ hágase $v'_{t,j} = v_{t,j} + S_{t,j}$ $(j \in \{0, 1, \ldots, q\})$.

5.4.2. Algoritmo EM

Si bien las técnicas MCMC no son exclusivas de la perspectiva Bayesiana, éstas tienen un papel muy importante y útil en el paradigma Bayesiano (de hecho, así se usó en la sección 5.4.1). La idea de esta sección es discutir un poco acerca de la estimación máximo verosímil desde una perspectiva Bayesiana con un critero clásico de selección de modelo: el criterio de información. Como la verosimilitud (dados los datos $\mathbf{x}$) tiene una forma analítica difícil de trabajar, se preferirá trabajar utilizando un algoritmo de aumentación de *data augmentation*. La aumentación natural es $\mathbf{w} = (\mathbf{v}, \mathbf{y}, \mathbf{z})$ como la que se utilizó (casi implícitamente) en la sección 5.3.2 con la que se obtuvo una verosimilitud proporcional a

$$f(\mathbf{v},\mathbf{y},\mathbf{z},\boldsymbol{\alpha},\boldsymbol{\beta},\lambda|\mathbf{x}) \propto \pi(p)p!\pi(q)q!\prod_{t=1}^{n}[\frac{\lambda^{z_t}}{z_t!}e^{-\lambda}\left(\prod_{i=1}^{p}\binom{x_{t-i}}{y_{t,i}}\alpha_i^{y_{t,i}}(1-\alpha_i)^{x_{t-i}-y_{t,i}}\right)$$
$$\times\left(\prod_{j=1}^{q}\binom{z_{t-j}}{v_{t,j}}\beta_j^{v_{t,j}}(1-\beta_j)^{z_{t-j}-v_{t,j}}\right)]\lambda^{A_\lambda}e^{-\lambda B_\lambda}\prod_{i=1}^{r}\frac{\lambda^{z_{1-i}}}{z_{i-1}!}e^{-\lambda}$$

con $A_\lambda = 0 = B_\lambda$. Entonces, el estimador máximo verosímil para $\theta = (\boldsymbol{\alpha},\boldsymbol{\beta},\lambda)$ tiene la forma

$$\hat{\alpha}_i = \frac{\sum_{t=1}^{n} y_{t,i}}{\sum_{t=1}^{n} x_{t-i}}, \quad i \in \{1,\ldots,p\}$$

$$\hat{\beta}_j = \frac{\sum_{t=1}^{n} v_{t,j}}{\sum_{t=1}^{n} z_{t-j}}, \quad j \in \{1,\ldots,q\}$$

$$\hat{\lambda} = \frac{1}{n}\sum_{t=1}^{n} z_t$$

Sin embargo, el algoritmo EM se basa en su habilidad para calcular los valores esperados de $\mathbf{w}$ dado $\mathbf{x}$ y θ. Para el modelo $INARMA(p,q)$, la dependecia de $(\mathbf{v}_t,\mathbf{y}_t,z_t)$ con $\mathbf{z}_{t-1}$ hace que el paso de esperanza sea poco factible. Se puede utilizar un algortimo EM Monte-Carlo, pero este es muy lento debido a la lenta convergencia de los parámetros y la alta tasa de rechazo que se obtiene en el paso E.

Para el modelo $INAR(p)$ se puede calcular los valores esperados de $\mathbf{y}$ dados $\mathbf{x}$ y θ. Por otra parte, la verosimilitud de θ se obtiene como un subproducto del algoritmo EM. El algoritmo EM sólo se puede usar en el $INAR(p)$ con un orden fijo p. Sin embargo, el cálculo de la verosimilitud permite comparar diferentes órdenes de p utilizando algunos criterios de selección de modelo, en particular el criterio de información Bayesiana (BIC). Esto se hará más adelante para comparar diferentes modelos $INAR(p)$. Se puede implementar fácilmente el algoritmo EM para $p = 1$ y dicho algoritmo es muy eficiente en el paso E y converge rápidamene a los parámetros máximo verosímiles. Sin embargo, dicha eficiencia disminuye cuando p se incrementa. Esto se debe en parte al hecho de que el paso E requiere cálculo de probabilidad para todas las posibilidades de datos faltantes (cuya cantidad se incrementa en orden exponencial como función de p), y por supuesto, también hay una convergencia más lenta a los parámetros máximo verosímiles (que están altamente correlacionados). Además, cuando p se incrementa la precisión computacional se vuelve un tema delicado ya que para $t \in \{1,\ldots,n\}$ cada posibilidad individual para el dato faltante $(\mathbf{y}_t,z_t)$ tiene probabilidad muy pequeña. Esto hace evidente al ver, en ocasiones, una ligera disminución en el cálculo del logaritmo de la verosimilitud, a pesar de que el algoritmo es correcto.

5.5. Implementaciones: Simulaciones y datos reales

5.5.1. Resultados de la implementación

Con el fin de analizar las debilidades y fortalezas de los algortimos MCMC que se describieron en las secciones anteriores, se realizarán las implementaciones de dichos algoritmos. En el cuadro 5.1 se establecen los modelos de los cuales se harán las simulaciones.

Modelo	n	$\boldsymbol{\alpha}$	$\boldsymbol{\beta}$	λ	Ruido
AR(3)	400	(0.2,0.3,0.15)	-	2	-
MA(3)	400	-	(0.2,0.3,0.15)	2	-
ARMA(2,2)	400	(0.4,0.2)	(0.4,0.2)	2	-
AR(3)+Ruido	400	(0.2,0.3,0.15)	-	2	$Poisson(1)$

Cuadro 5.1: *Una selección de modelos para la simulación*

En cada uno de los modelos, se simularon 10,000 de valores con un periodo de calentamento de 1,000 iteraciones. El periodo de calentamiento es esencial para asegurar que los estimados de la distribución posterior no se vean afectados por la elección de los valores iniciales de la cadena de Markov. Además se escogieron aprioris $Gamma(1,1)$ para λ y μ (para ambas es la misma).

En el cápítulo 1, y de hecho es la máxima de los algoritmos MCMC, se dijo que los métodos de cadenas de Markov Monte-Carlo generan observaciones dependientes de la distribución posterior de interés. Por tanto, es importante que se genere una muestra suficientemente grande para no "perder información" en comparación con observaciones independientes e idénticamente distribuidas de la distribución posterior (debido a que se tiene dependencia en las observaciones, i.e. formalmente no se tienen muestras aleatorias, sino realizaciones de procesos estocásticos con estructura más intrincada).

En la literatura hay algunos métodos formales para analizar el comportamiento de los algoritmos MCMC, como por ejemplo con la función de autocorrelación integrada ó la varianza de las medias agrupadas (Geyer (1992)). Sin embargo, por el momento sólo será suficiente una inspección visual de la serie de tiempo de la realización generada para detectar que la cadena de Markov se simula de su distribución estacionaria (en este caso, la distribución posterior). Esta exploración visual de la serie también permite determinar el grado de "*mixing*" de la cadena de Markov, donde un mejor *mixing* corresponde a menor dependencia entre observaciones sucesivas de la distribución posterior y por tanto un menor varianza en el error de la simulación (error Monte-Carlo). En las figuras 5.1 a 5.4 se muestran la gráficas de series de tiempo de las primeras 10,000 observaciones para λ para cada uno de los cuatro modelos que se consideran.

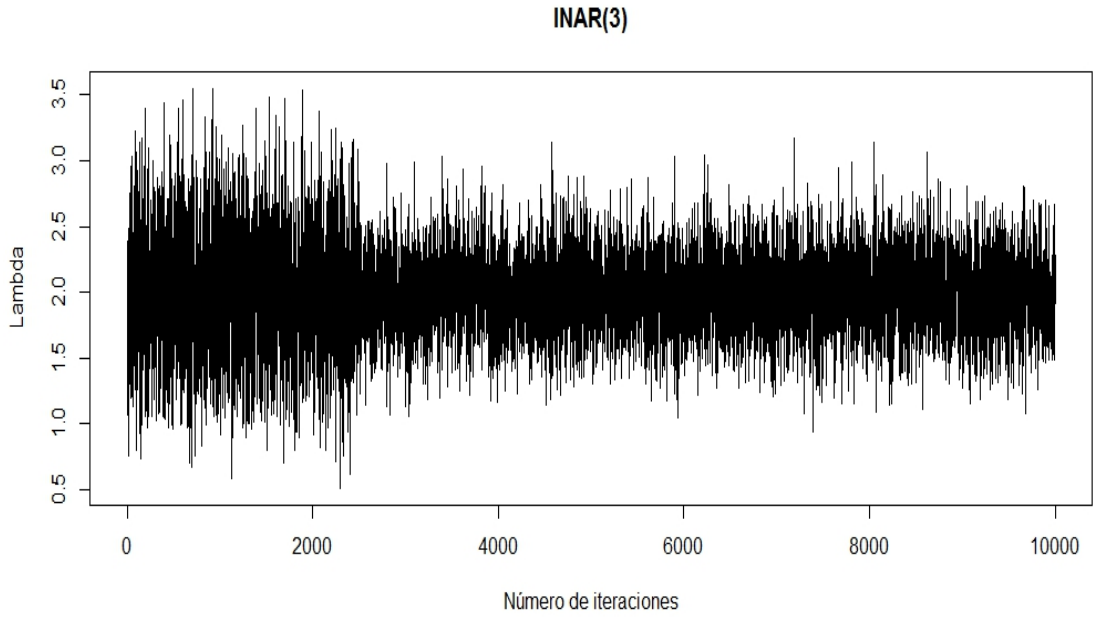


Figura 5.1: *Gráfico tipo serie de tiempo para las primeras 10,000 observaciones de* λ *bajo supuesto* $INAR(3)$

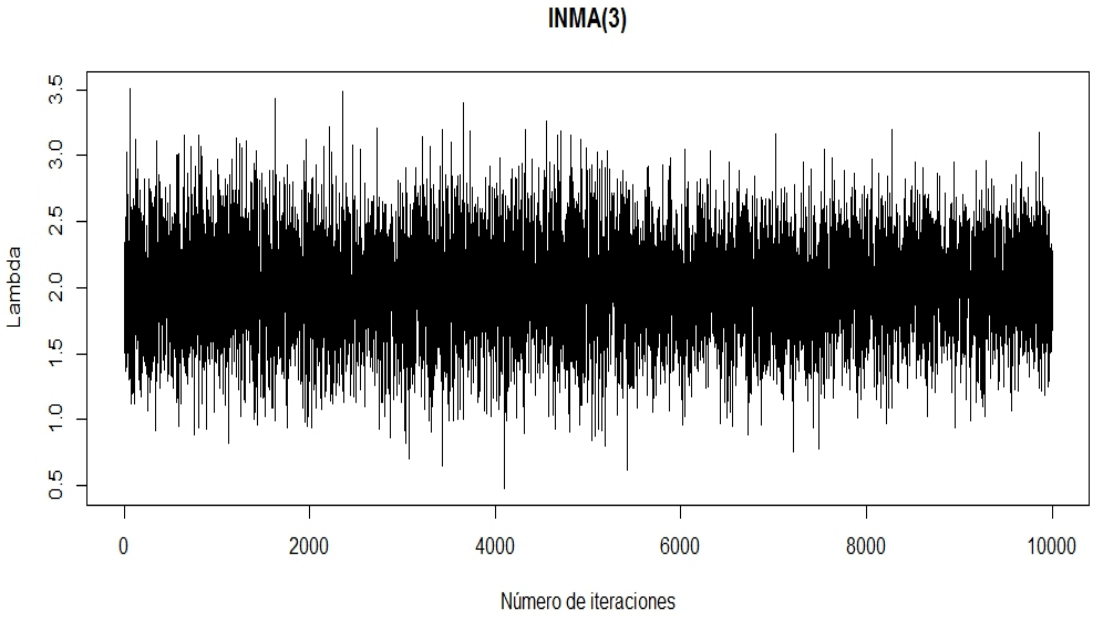


Figura 5.2: *Gráfico tipo serie de tiempo para las primeras 10,000 observaciones de* λ *bajo supuesto* $INMA(3)$

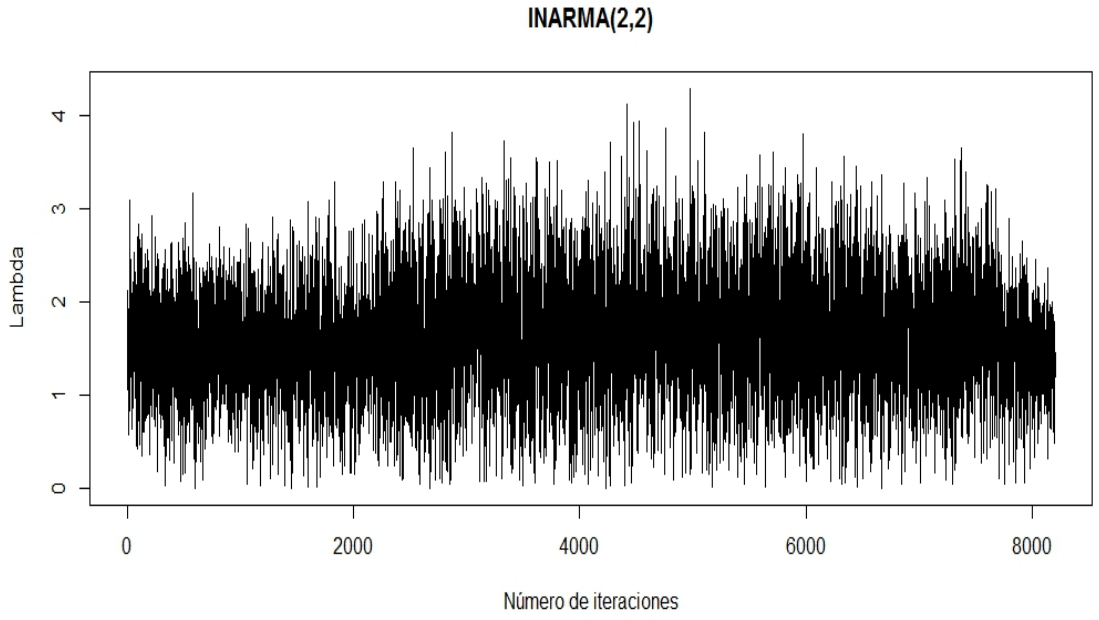


Figura 5.3: *Gráfico tipo serie de tiempo para las primeras 10,000 observaciones de* λ *bajo supuesto* $INARMA(2,2)$

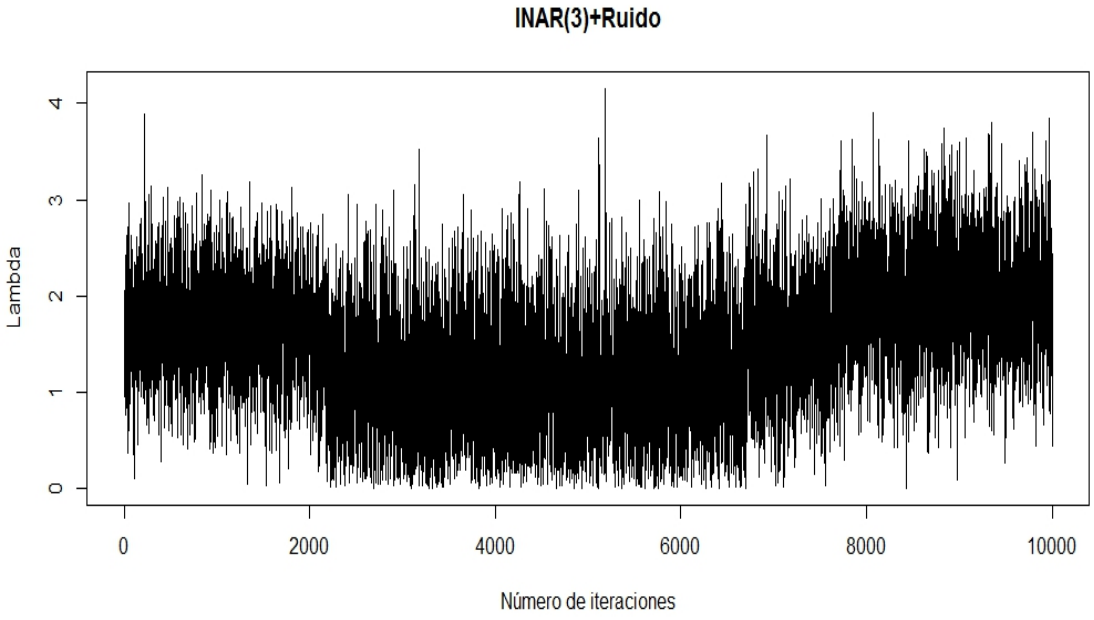


Figura 5.4: *Gráfico tipo serie de tiempo para las primeras 10,000 observaciones de λ bajo supuesto $INAR(3)$ + ruido aditivo*

El *mixing* de la cadena de Markov con respecto a λ es una manera visual de verificar el comportamiento de los algortimos MCMC. Los algortimos se comportan adecuadamente para estructuras $INAR$ e $INMA$ pero no lo suficiente en las estructuras $INARMA$ e $INAR$ con ruido aditivo. El desempeño de este algoritmo MCMC se relaciona fuertemente con la información contenida en los datos con respecto a los diferentes parámetros. Es decir, mientras más informativos sean los datos con respecto a los parámetros del modelo subyacente, se comportará mejor el algoritmo MCMC. En este caso, el algoritmo MCMC se comporta muy bien para el modelo $INAR(p)$ ya que los datos $(x_{1-p}, x_{2-p}, \ldots, x_n)$ es muy informativo con respecto a los parámetros del modelo. Sin embargo, dicho modelo pareciera comportarse mejor en el proceo $INMA(q)$ aún cuando el proceso subyacente $(z_{1-q}, z_{2-q}, \ldots, z_n)$ no sea observado (y necesite simularse). El cuadro 5.2 muestra algunas estimaciones de la media y desviación estándar de las distribuciones posteriores (obtenidos a partir de simulaciones del MCMC).

Para el $INARMA(2,2)$, los datos son más informativos para la parte auto-regresiva que para la de medias móviles. Esto ocurre generalmente en este tipo de estructura mixta. Este hecho lo exhiben las varianzas considerablemente grandes.

La presencia de un parámetro de ruido también afecta al algortimo. Aunque el ruido y el proceso son independientes, las distribuciones posteriores conjuntas para los parámetros de ambos procesos NO son independientes. Por tanto, mientras mayor sea el ruido, será más difícil detectar al proceso $INAR(p)$ subyacente, luego entonces, una mala estimación de los parámetros.

Modelo	$\boldsymbol{\alpha}$	$\boldsymbol{\beta}$	λ	μ
AR(3)				
Media	(0.2176,0.3175,0.1105)	-	1.946	-
Desv. Est.	(0.2096,0.3180,0.1064)	-	1.926	-
MA(3)				
Media	-	(0.1696,0.3180,0.1494)	2.101	-
Desv. Est.	-	(0.0533,0.0619,0.0577)	0.137	-
ARMA(2,2)				
Media	(0.3478,0.3122)	(0.5906,0.1903)	1.574	-
Desv. Est.	(0.0638,0.0512)	(0.199,0.1288)	0.3016	-
AR(3) + ruido				
Media	(0.1866,0.3732,0.1284)	-	1.396	1.789
Desv. Est.	(0.0578,0.0773,0.0626)	-	0.387	0.756

Cuadro 5.2: *Media y desviación estándar estimadas de la distribución posterior de los parámetros*

Ahora, se hará la implementación del algoritmos que se analizaron en las secciones 5.3.2 y 5.4.1. Los modelos a partir de los cuales se hicieron las simulaciones están en el cuadro 5.3.

Modelo	n	$\boldsymbol{\alpha}$	$\boldsymbol{\beta}$	λ
AR(3)	400	(0.4,0.1,0.3)	-	2
MA(3)	400	-	(0.4,0.1,0.3)	2
ARMA(2,1)	400	(0.4,0.2)	(0.4,0.2)	2

Cuadro 5.3: *Una selección de modelos para la simulación*

En cada modelo, se implementó un algoritmo RJMCMC para obtener una muestra de tamaño 100,000 con un periodo de calentamiento de 10,000 iteraciones con $p_{max} = q_{max} = 10$. La apriori que se eligió para $\pi(\lambda)$ es una $Gamma(1,1)$. Para cada modelo, se generó una trayectoria de longitud 400. El análisis está basado en las primeras n observaciones con $n = 100, 200, 400$.

La elección de las aprioris para p y q es importante. Si p ó q se incrementan en una unidad, el número de parámetros se incrementará en $n + 1$ correspondiente al nue-

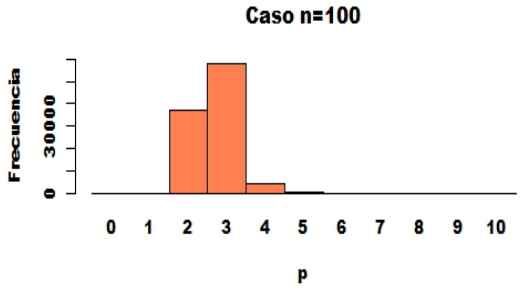
vo parámetro AR ó MA y los nuevos n valores aumentados. Entonces, una penalización BIC del orden del modelo tendrá buenos resultados en los datos simulados. Es decir, si se hace $\pi(p) \propto n^{-p/2}$ y $\pi(q) \propto n^{-q/2}$, tales apropir generarán resultados consistentes cuando $n(\leq 400)$ se incrementa. Algunas aprioris alternativas son $\pi(p), \pi(q)$ como $Poisson(\tau)$ ó $Geo(\tau)$ pero no son tan efectivas según el BIC.

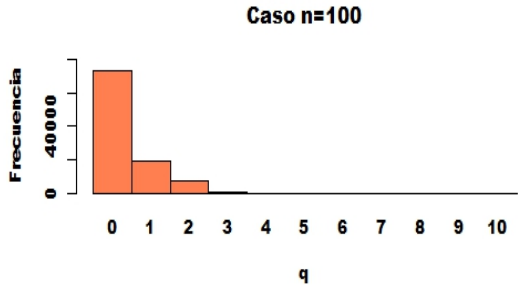
Modelo	Probabilidad
INAR(3)	0.902
INARMA(3,2)	0.049
INARMA(3,1)	0.039

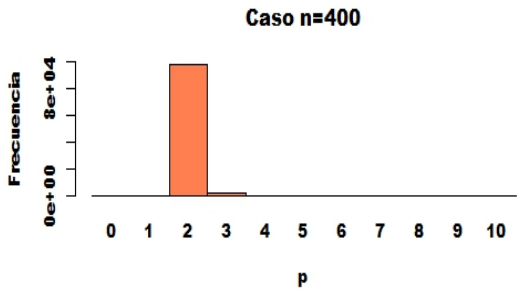
Cuadro 5.4: *Probabilidades del modelo cuando los datos viene de un* $INAR(3)$*,* $n = 400$

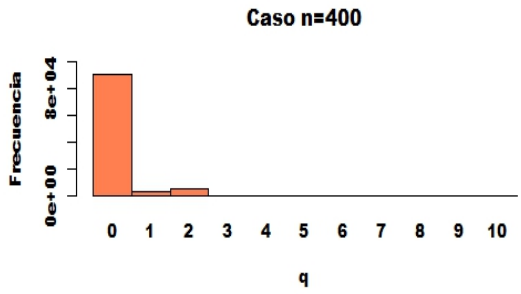
Modelo	$\boldsymbol{\alpha}$	$\boldsymbol{\beta}$	λ
AR(3)			
Media	(0.4563,0.1358,0.2299)	-	1.327
Desv. Estándar	(0.0428,0.0455,0.0513)	-	0.3417
MA(3)			
	-	(0.3993,0.1819,0.3632)	2.003
	-	(0.0694,0.0745,0.0728)	0.1098
ARMA(2,1)			
	(0.428,0.165)	(0.618)	1.717
	(0.0634,0.0473)	(0.2112)	0.2334

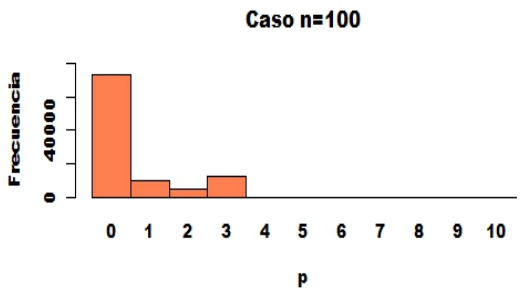
Cuadro 5.5: *Media y desviación estándar estimadas de la distribución posterior de los parámetros,* $n = 400$

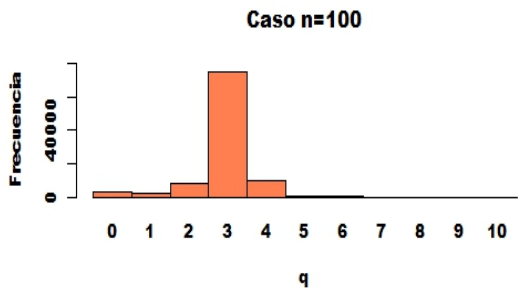

Caso n=100
Frecuencia
0
30000
0 1 2 3 4 5 6 7 8 9 10
p

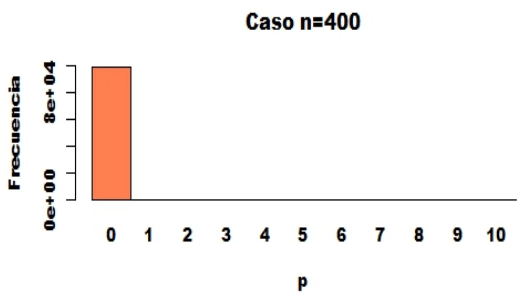

Caso n=100
Frecuencia
0
40000
0 1 2 3 4 5 6 7 8 9 10
q

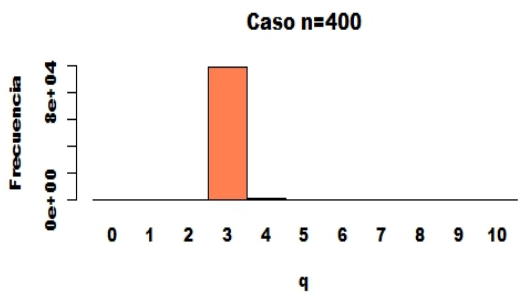

Caso n=400
Frecuencia
0e+00
8e+04
0 1 2 3 4 5 6 7 8 9 10
p

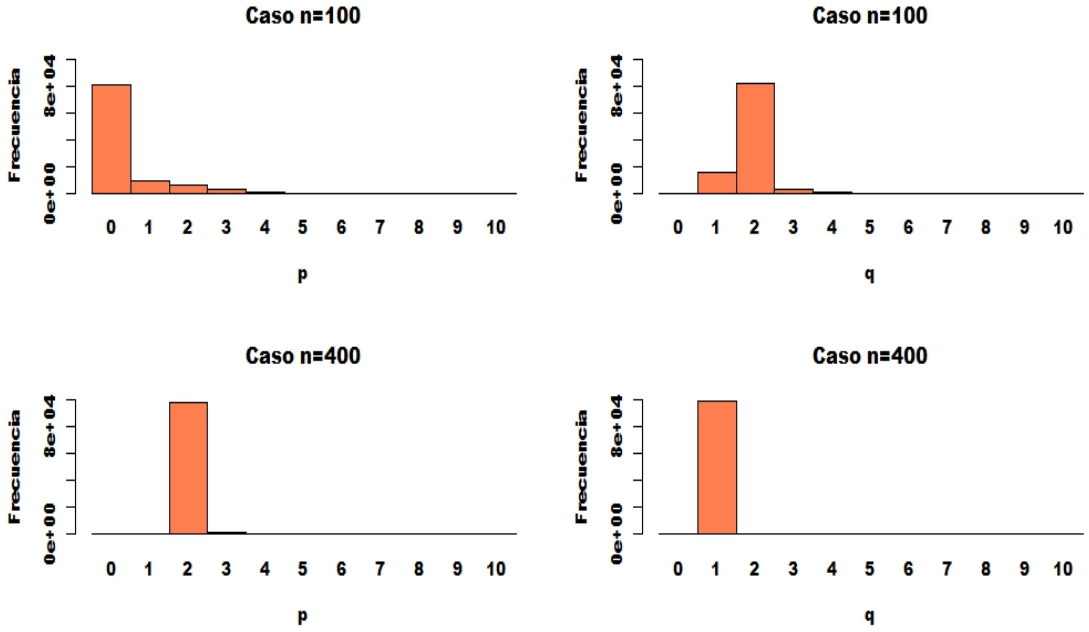

Caso n=400
Frecuencia
0e+00
8e+04
0 1 2 3 4 5 6 7 8 9 10
q


Caso n=100
Frecuencia
0
40000
0 1 2 3 4 5 6 7 8 9 10
p


Caso n=100
Frecuencia
0
40000
0 1 2 3 4 5 6 7 8 9 10
q


Caso n=400
Frecuencia
0e+00
8e+04
0 1 2 3 4 5 6 7 8 9 10
p


Caso n=400
Frecuencia
0e+00
8e+04
0 1 2 3 4 5 6 7 8 9 10
q



5.5.2. Datos reales

Uno de los objetivos principales para estudiar los algoritmos MCMC en las secciones 5.3.2 y 5.4.1 fue para implementar esta metodología en datos reales. En este sección se aplicarán los resultados que se obtuvieron en las secciones anteriores a dos conjuntos de datos: Partículas de oro de Westgren y score de pacientes esquizofrénicos. Aunque sólo se implementarán el $RJMCMCM$ para determinar el orden a la primera; esto se debe a que la segunda tiene un punto de corte, haciendo complicado el paso de cambio de modelo.

5.5.2.1. Partículas de oro de Westgren

Este conjunto de datos consiste de 380 conteos de partículas de oro en una solución en una solución de puntos equidistantes en el tiempo, aunque en este trabajo sólo se dispone de 375. Los datos se publicaron por primera vez en Westgren (1916) y la han analizado Jung & Tremayne (2006) en la estimación por momentos del modelo $INAR(2)$. Jung & Tremayne (2006) afirmaron que el $INAR(2)$ es un modelo adecuado para estos datos e hicieron su análisis en las primeras 370 observaciones y se usaron las 10 observaciones restantes con fines de predicción. Y de hecho, con el $RJMCMC$ se verá que el $INAR(2)$ es el modelo $INAR$ más adecuado para este conjunto de datos. En la figura 5.5 se muestra el conjunto de datos de partículas de oro de Westgren.

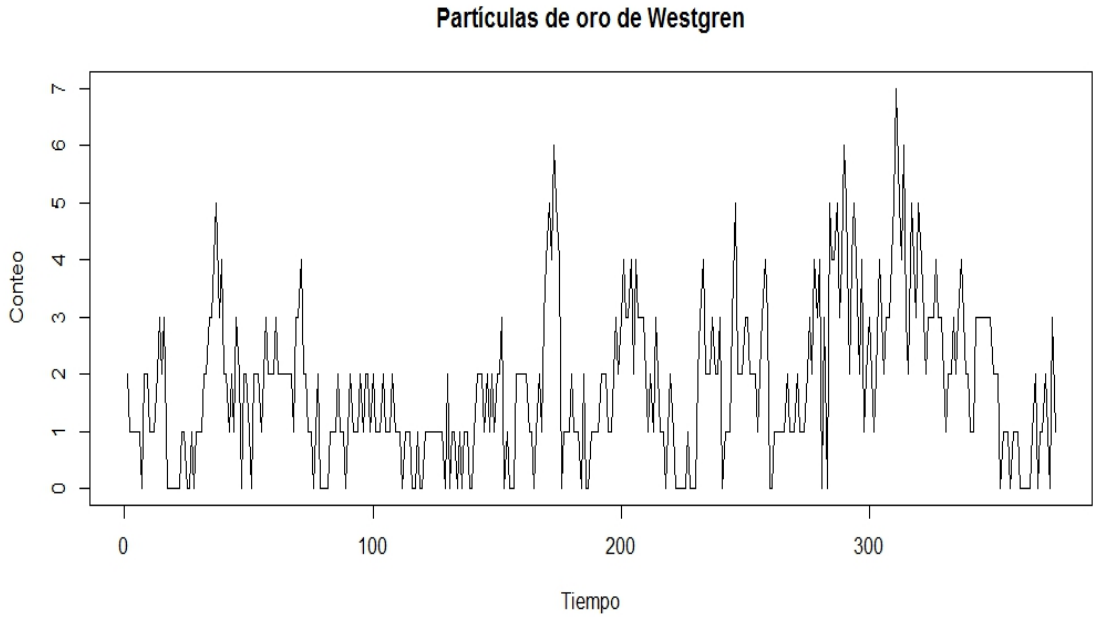


Figura 5.5: *Serie de tiempo de partículas de oro de Westgren*

La idea es comparar esta metodología con la de Jung & Tremayne (2006). En particular, se ajustará un modelo $INAR(2)$ para las primeras 365 observaciones y se hará estimación de los parámetros del $INAR(2)$. Se obtendrá la distribución conjunta predictiva para las siguientes 10 observaciones. En el capítulo 4 se consideraronn dos modelos $INAR(2)$: el $INAR(2)$ de Alzaid & Al-Osh (1990) y el $INAR(2)$ de Du & Li (1991). Se implementará el modelo de Du & Li, sin embargo esta metodología MCMC se puede adaptar fácilmente al modelo de Al-Osh & Alzaid.

Con el algoritmo MCMC se obtendrán 100,000 simulaciones de la distribución posterior del parámetro con un periodo de calentamiento de 10,000 iteraciones. Estas muestras se usarán para simular 100,000 realizaciones de los datos puntuales 376, 377, y 385. Las medias posteriores estimadas y las desviaciones estándar para los parámetros del modelo están en el Cuadro 5.6. Estos resultados son comparables con los parámetros que estimaron Jung & Tremayne (2006).

Parámetro	α_1	α_1	λ
Media	0.45869	0.18277	0.53341
Desviación estándar	0.0451	0.05581	0.0699

Cuadro 5.6: *Estimación de los parámetros de un INAR(2) para los datos de partículas de oro de Westgren*

Para la distribución conjunta predictiva de los últimos 10 puntos, se obtuvieron resultados idénticos a los de Jung & Tremayne (2006) para los valores medianos (que se predijeron). Y en la figura 5.6 se muestran los histogramas de ls densidades posteriores de los puntos 376, 377 y 385, además del histograma empírico de la serie de partículas de oro de Westgren. Nótese que las predicciones parecen bastante acertadas, dado el histograma de los datos completos.

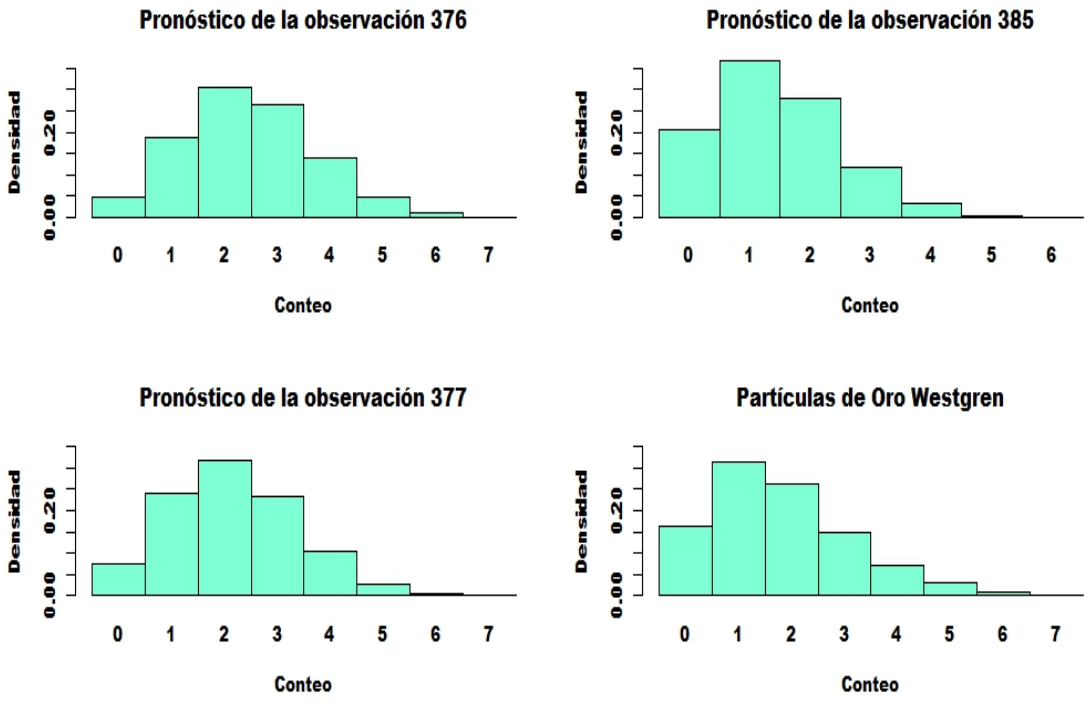
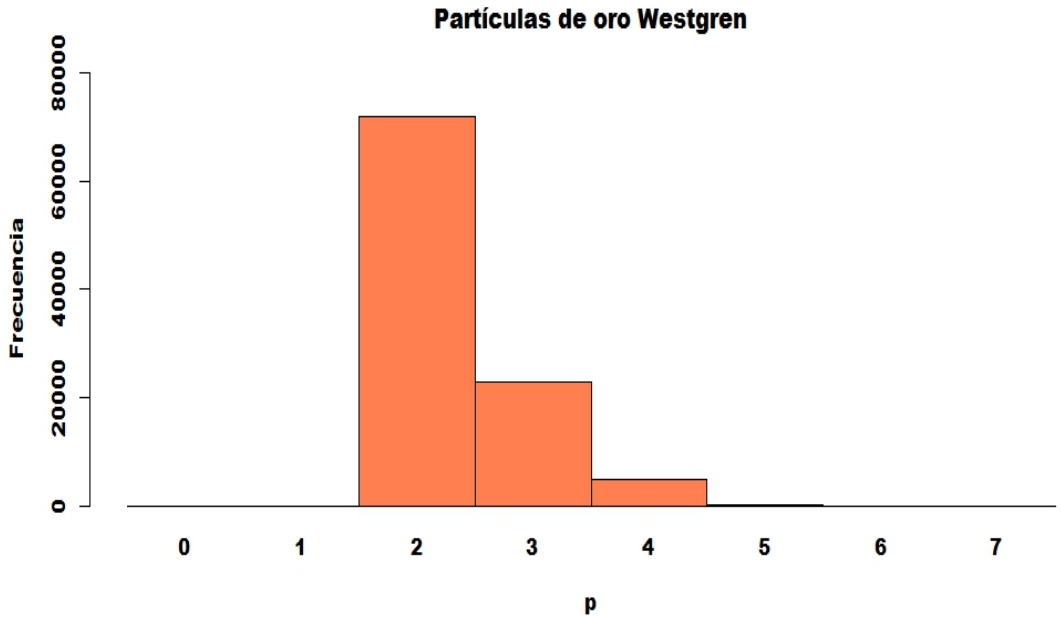
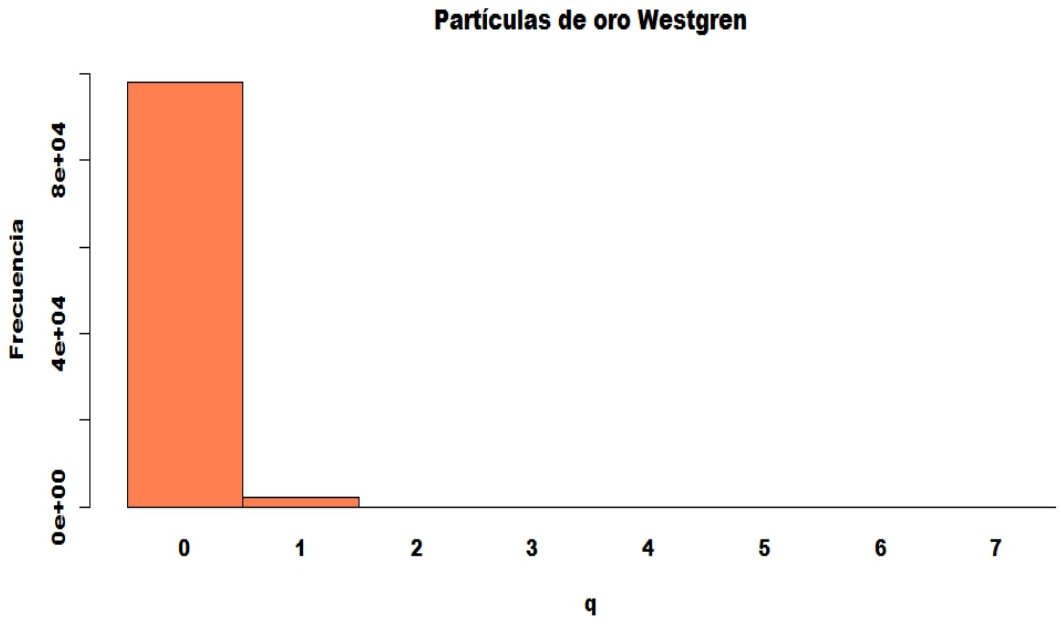


Figura 5.6: *Histogramas de las distribuciones predictivas de los puntos 376, 377 y 385, además de la distribución empírica de la serie de partículas de oro de Westgren.*

Para este conjunto de datos se implementará el algoritmo $RJMCMC$ para determinar el orden. Se simularán 100,000 observaciones de las distribuciones de p y q (de cada una)

con un periodo de calentamiento de 10,000. Se utilizarán a *prioris* para p, q y λ como $\pi(p) \propto n^{-p/2}$, $\pi(p) \propto n^{-q/2}$ y $\lambda \sim Gamma(1, 1)$.



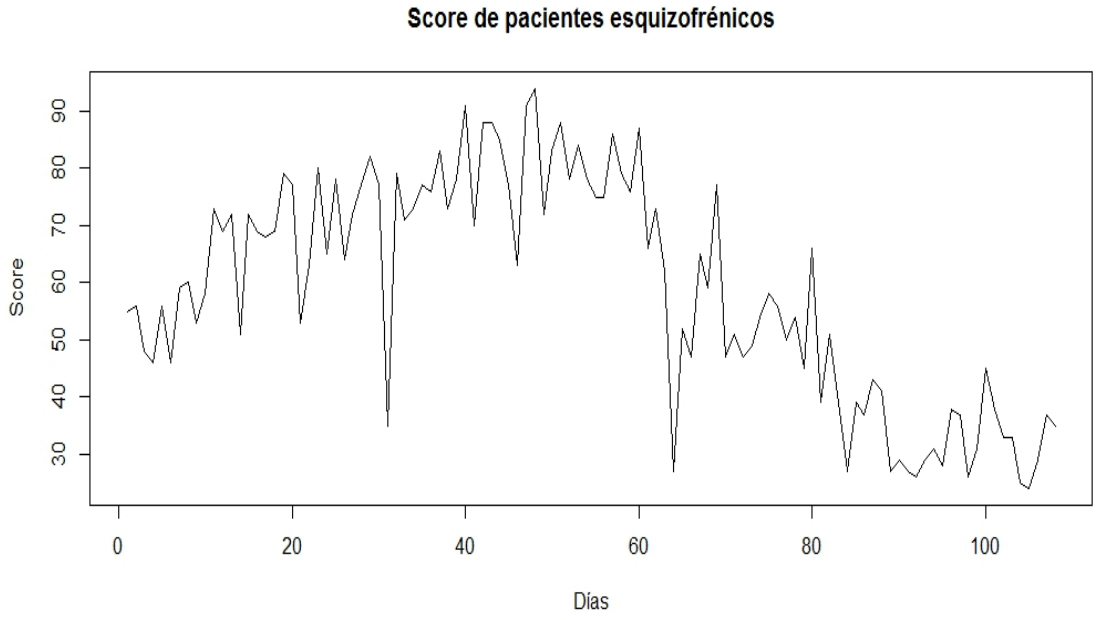


Modelo	α_1		α_2		α_3		λ	
	Media	Desv. Est.	Media	Desv. Est.	Media	Desv. Est.	Media	Desv. Est.
INAR(2)	0.471	0.0499	0.182	0.0526	-		0.563	0.0711
INAR(3)	0.449	0.0482	0.154	0.0601	0.114	0.0476	0.466	0.0728

Cuadro 5.7: *Media y desviación estándar para los estimadores de los parámetros en los datos de partículas de oro Westgren*

5.5.2.2. Score de pacientes esquizofrénicos

El segundo conjunto de datos consiste de las observaciones diarias del puntaje que obtuvo un paciente esquizofrénico en una prueba de velocidad perceptiva (véase McCleary & Hay (1980)). El conjunto de datos consiste de 120 calificaciones diarias consecutivas. Sin embargo, a partir del día 61 los pacientes comenzaron a recibir un poderoso tranquilizante (chloropromazine).



En los datos, se puede observar claramente una disminución en el puntaje después del día 61. Se propone modelar este puntaje mediante un proceso $INAR(1)$ con diferentes parámetros con diferentes parámetros (α_1, λ) para antes y después de recibir el tranquilizante. Por tanto, se considerará el modelo

$$X_t = \begin{cases} \alpha_1^A \circ X_{t-1} & \text{si } t \leq 60 \\ \alpha_1^D \circ X_{t-1} & \text{si } t \geq 61 \end{cases}$$

donde $\{Z_t\}_{t=1}^{120}$ son variables aleatorias independientes, $\{Z_t\}_{t=1}^{60}$ tienen distribución $Poisson(\lambda^A)$ y $\{Z_t\}_{t=61}^{120}$ tienen distribución $Poisson(\lambda^D)$ (los superíndices A y D representan antes y después del tratamiento).

No es difícil adaptar la verosimilitud (y por tanto el algoritmo MCMC) para incorporar un punto de cambio, y más porque dicho punto de cambio es conocido. Por supuesto, se puede modificar la metodología al caso en el que la posición y el número del punto de cambio es desconocido.

Se implementó el algoritmo y se obtuvieron 10,000 simulaciones de la distribución posterior de (α_1, λ) con un periodo de calentamiento de 1,000 simulaciones. La idea de este análisis es determinar los efectos del tranquilizante en los parámetros; aunque también hubiése sido factible obtener la distribución predictiva para encontrar alguna banda de con-

fianza para observaciones futuras.

Con las simulaciones de α^A y α^D se obtienes densidades posteriores similares sugiriendo que el parámetro α_1 no se ve afectado por la aplicación del tranquilizante y que los datos se deben modelar con $\alpha^D = \alpha^A = \alpha$. Por tanto, se repitió el ejercicio de simulación para un mismo α_1 para el caso antes de tratamiento y después de tratamiento. En la figura 5.7 se muestran las densidades posteriores para los parámetros para las dos implementaciones.

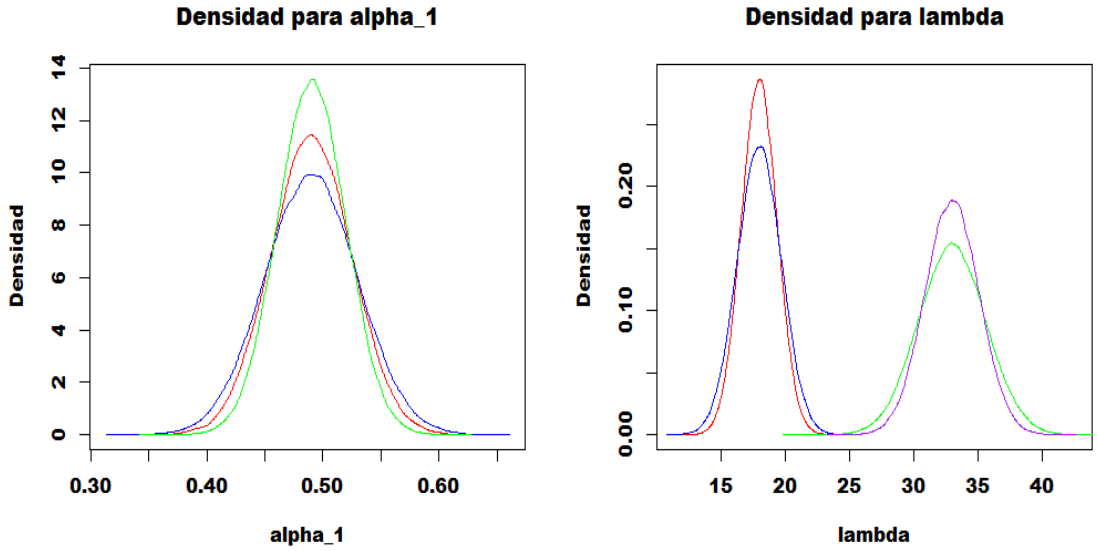


Figura 5.7: *Densidades posteriores para los parámetros del INAR(1) antes y después de tratamiento*

Con las dos implementaciones del algoritmo MCMC se obtienen resultados similares para las medias posteriores de $\alpha_1, \lambda^A, \lambda^D$. Como se supuso el mismo α_1, la reducción en el puntaje se deberá a modificaciones en el tamaño del parámetro λ. Cuando se supone la misma α_1, las medias posteriores de λ^A y λ^D son 33.97 y 19.43, respectivamente. Sin embargo, si se supone $\alpha_1^D \neq \alpha_1^A$, las medias posteriores de λ^A y λ^D son 34.62 y 19.51, respectivamente.

Nótese que la distribución posterior para α_1 ($\alpha_1 = \alpha_1^D = \alpha_1^A$) es más compacta que las distribuciones posteriores individuales de α_1^D y α_1^A (cuando $\alpha_1^D \neq \alpha_1^A$). Esto se debe a que la inferencia para α_1 se basa en un serie de tiempo de longitud 120 y las inferencias para α_1^D, α_1^A se basan en series de longitud 60. Esto también tiene efecto en que las distribuciones posteriores para λ_1^A y λ_1^D (en el caso de α_1 común) sean leptokurticas.

Dado este punto de cambio, se no se implementó en este conjunto de datos el algoritmo para determinar el orden.

Conclusiones

En este trabajo se revisaron muchas de las ideas más conocidas en series de tiempo que toman valores en un espacio a lo más numerable. Se explicó porqué algunas de las técnicas estándar en el análisis de series de tiempo no son adecuadas y las aproximaciones asintóticas se tienen que considerar de forma muy cuidadosa y la mayoría de las veces son inadecuadas.

Se aprovecharon las ventajas de las técnicas de simulación actuales para resolver algunos problemas de inferencia y predicción asociados a este tipo de datos discretos. Se dedicó todo el capítulo 1 para estudiar las principales técnicas de simulación principalmente los algoritmos Metropolis-Hastings y Gibbs *sampler*. Se estudiaron las principales condiciones para la convergencia de estos algoritmos y algunas propuestas para utilizar algoritmos que tengan partes Gibbs-sampler pero también Metropolis-Hastings, para poder definir un algorirmo MCMC con saltos reversibles.

El primer catálogo de modelos para series de tiempo con valores que toman valores en un conjunto a lo más numerable se hace en el mediante teoría de cadenas de Markov con espacio de estados numerables (principalmente los modelos de Raftery) y se terminó con los modelos $DARMA$ (así como DAR y DMA) de Lewis & Jacobs.

El capítulo más amplio es el capítulo 4, en éste se discutió extensamente acerca de las propuestas para modelar series de tiempo con valores discretos intentando extender (o redefinir) la clase de modelos $ARMA$. Se hicieron analogías con las definiciones y proposiciones del capítulo 2. Sin embargo, la principal clase que se analizó es la de los modelos $INAR$. Se dieron algunos resultados para la existencia de soluciones estacionarias para replicar las estructuras de correlación de los procesos AR. El factor clave para hacer estas extensiones es lo que se conoce como operador de adelgazamiento binomial. Se hicieron dos propuestas, definiendo dos operaciones de adelgazamiento diferente; se generalizó a operadores de van Harn & Steutel para finalmente llevarlo a semigrupos de operadores de adelgazamiento, exhibiendo que los conceptos de autodescomponibilidad y estabilidad son muy importantes.

Particularmente se profundizó en los casos $INAR$ e $INMA$ cuando se tienen procesos de innovación con distribución Poisson. Se hicieron algunos diagnósticos sencillos, junto con un ejercicio de simulación para calcular las potencias de algunas pruebas para determinar independencia serial.

En el último capítulo se utilizaron las técnicas de simulación del capítulo 1 y el marco teórico de los capítulos 3 y 4 para analizar, definir principales propiedades, inferencia básica y predicción en los modelos $INARMA$. Además se implementaron algoritmos para hacer estimación, predicción y determinación el orden de la estructura de correlación en datos reales y simulados.

Apéndices

Apéndice A

Un poco de teoría de cadenas de Markov

Es importante mencionar que en este texto no se considerarán modelos de cadenas de Markov a tiempo continuo (que también se conocen como procesos de Markov), i.e. no se considerarán procesos estocásticos $\{X_t\}_{t\in\mathcal{T}}$ con $\mathcal{T}$ un intervalo de $\mathbb{R}$ ó incluso todo $\mathbb{R}$. Gracias a la naturaleza de la simulación, sólo se considerarán procesos estocásticos a tiempo discreto, i.e. $\{X_t\}_{t\in\mathcal{T}}$ tal que $\mathcal{T}\subseteq\mathbb{N}^+$

A.1. Definiciones y primeros resultados

Una cadena de Markov es una sucesión de variables aleatorias. La indexación con $\mathbb{N}$ se puede pensar como variables aleatorias que evolucionan a través del tiempo en donde la probabilidad de un movimiento depende del conjunto particular en el que se encuentra la cadena.

Aunque esta es una manera muy intutiva de definir el concepto de proceso estocástico, un forma un poco más "elegante" desde el punto de vista matemático es definir a la cadena en términos de una función que determine estos movimientos; dicha función se conoce como kernel de transición.

Definición A.1.1. *(Kernel de transición)*

Sea $\mathcal{X}\subseteq\mathbb{R}^k$ y $\mathcal{B}(\mathcal{X})$ la σ-álgebra de Borel generada por $\mathcal{X}$. Un kernel de transición es una función $K:\mathcal{X}\to\mathcal{B}(\mathcal{X})$ tal que

(i) Para todo $x\in\mathcal{X}$, $K(x,\cdot)$ es una medida de probabilidad.

(ii) Para todo $A\in\mathcal{B}(\mathcal{X})$, $K(\cdot,A)$ es medible.

Observación A.1.1.

Cuando $\mathcal{X}$ es discreto, el kernel de transición es simplemente una matriz (de trancisión), K, cuyos entradas son

$$P_{xy} := \mathbb{P}(X_n = y | X_{n-1} = x), \qquad x, y \in \mathcal{X}$$

En el caso continuo, el kernel denota la densidad condicional $K(x, z)$ de la transición $K(x, \cdot)$, es decir

$$\mathbb{P}(X \in A|x) = \int_A K(x, z)dz$$

$\triangledown$

Generalmente la cadena se define en $\mathbb{N}$ en vez de en $\mathbb{Z}$, por tanto, la distribución de X_0 (el estado inicial de la cadena) juega un papel muy importante. En el caso discreto cuando el kernel K es una matriz de transición, dada una distribución inicial $\nu = (\omega_1, \omega_2, \ldots)$, la distribución marginal de X_1 se obtiene a partir de la expresión

$$\nu_1 = \nu K \tag{A.1}$$

y multiplicando sucesivamente se tiene que $X_n \sim \nu_n = \nu K^n$. Análogamente, en el caso continuo, si ν denota la distribución inicial de la cadena, i.e. $X_0 \sim \nu$, entonces P_ν denotará la distribución de probabilidad de $\{X_n\}$.

Notación A.1.1.

Cuando X_0 es fija, en particular para ν igual a la masa de Dirac δ_{x_0} se usa la notación P_{x_0} $\triangledown$

Definición A.1.2. *(Cadena de Markov)*

Dado un kernel de transición, una sucesión de variables aleatorias $\{X_t\}_{t=0}^{\infty}$ es una cadena de Markov si para cualquier $t \in \mathbb{N}^+$ la distribución condicional de X_t dado que $X_{t-1} = x_{t-1}, X_{t-2} = x_{t-2}, \ldots, X_0 = x_0$ es la misma que distribución de X_t dado que $X_{t-1} = x_{t-1}$; es decir,

$$\mathbb{P}(X_{k+1} \in A|x_0, x_1, \ldots, x_k) = \mathbb{P}(X_{k+1} \in A|x_k) = \int_A K(x_k, dx)$$

La cadena es homogénea (en el tiempo) si la distibución de $(X_{t_1}, \cdots, X_{t_k})$ dado $X_{t_0} = x_{t_0}$ es la misma distribución que la de $(X_{t_1-t_0}, \cdots, X_{t_k-t_0})$ dado $X_0 = x_{t_0}$ para cualquier $k \in \mathbb{N}^+$ y cualesquiera $t_0 \leq t_1 \leq \ldots \leq t_k$.

Considérese una cadena de Markov, si la distribución de X_0 se conoce, la construcción de dicha cadena está completamente determinada por su kernel de transición, que básicamente es la distribución de X_n dado que $X_{n-1} = x_{n-1}$.

Generalmente se estudian cadenas de Markov homogéneas y durante el desarrollo de este trabajo se supondrá esto (a menos que se especifique de otra forma); sin embargo, una mala implementación de los algoritmos MCMC puede generar cadenas de Markov no-homogéneas en las cuales no se pueden ocupar los criterios de convergencia que se desarrollarán.

En general, el hecho de que el kernel K determine las propiedades de la cadena $\{X_n\}$ se puede obtener de las siguientes relaciones

$$\mathbb{P}_x(X_1 \in A) = K(x, A_1)$$

$$\mathbb{P}_x((X_1, X_2) \in A_1 \times A_2) = \int_{A_1} K(y_1, A_2) K(x, dy)$$

$$\dots$$

$$\mathbb{P}_x((X_1, \dots, X_n) \in A_1 \times \cdots \times A_n) = \int_{A_1} \cdots \int_{A_{n-1}} K(y_{n-1}, A_n) K(x, dy_1) \cdots K(y_{n-2}, dy_{n-1})$$

En particular, la relación $\mathbb{P}_x(X_1 \in A) = K(x, A_1)$ establece que $K(x_n, dx_{n+1})$ es una versión de la distribución condicional de X_{n+1} dado X_n. Sin embargo, con la definición de cadena de Markov que se ha dado, i.e. primero especificando el kernel, no es necesario considerar las diferentes versiones de las probabilidades condicionales. En este sentido se puede decir que la construcción de cadenas de Markov a través de kernel es matemáticamente más "formal". Además en el estudio de técnicas MCMC los objetos de interés son estás distribuciones condicionales y es importante no preocuparse por sus diferentes versiones. Las propiedades de una cadena de Markov que se considerarán en este apéndice son independientes de la versión de probabilidad condicional seleccionada.

Notación A.1.2.

Sea $K^1 := K(x, A)$, el kernel para n transiciones está dado por

$$K^n(x, A) := \int_{\mathcal{X}} K^{n-1}(y, A) K(x, dy)$$

∇

Con ll siguiente resultado se obtienen fórmula de tipo convolución para K^{m+n} y se conoce como ecuaciones de Chapman-Kolmogorov.

Lema A.1.1. *(Ecuaciones de Chapman-Kolmogorov)*

Para cualesquiera $(m, n) \in \mathbb{N}^2$, $x \in \mathcal{X}$ y $A \in \mathcal{B}(\mathcal{X})$

$$K^{m+n}(x, A) = \int_{\mathcal{X}} K^n(y, A) K^m(x, dy)$$

En términos vagos, las ecuaciones de Chapman-Kolmogorov establecen que para ir de x a A en $m+n$ pasos, se debe pasar por algún y en el n-ésimo paso. En el caso discreto, el Lema anterior se puede interpretar simplemente como un producto de matrices y se sigue directamente de (A.1). En el caso general, se necesita considerar a K como un operador sobre el éxito de las funciones integrables, i.e. se necesita definir

$$Kh(x) := \int h(y) K(x, dy), \quad h \in \mathcal{L}_1(\lambda)$$

donde λ representa la medida dominante del modelo. Entonces K^n es la n-ésima composición de P, i.e. $K^n = K \circ K^{n-1}$.

Definición A.1.3. *(Resolvante)*

Un resolvante asociado con un Kernel P es un kernel de la forma

$$K_\varepsilon(x, A) = (1 - \varepsilon) \sum_{i=0}^{\infty} \varepsilon^i K^i(x, A)$$

con $\varepsilon \in (0, 1)$. La cadena con el kernel K_ε se conoce como la K_ε-cadena.

Dada una distribución inicial ν, se puede asociar una cadena $\{X_n^\varepsilon\}$ con el kernel K_ε que formalmente corresponde a una subcadena de la cadena original $\{X_n\}$, donde los índices de la cadena se generan de una distribución geométrica con parámetro $1 - \varepsilon$. Por tanto, K_ε efectivamente es un kernel y más adelante se verá que la cadena de Markov $\{X_n\}$ tiene ciertas propiedades de regularidad.

Si $\mathbb{E}_\nu(\cdot)$ denota la esperanza asociada con la distribución P_ν, entonces la propiedad de Markov débil se puede escribir de la siguiente forma.

Proposición A.1.1. *(Propiedad de Markov débil)*

Para toda distribución inicial ν y para cada muestra $(X_0, X_1, \ldots, X_n)$

$$\mathbb{E}_\nu(h(X_{n+1}, X_{n+2}, \ldots)|x_0, \ldots, x_n) = \mathbb{E}_{x_n}(h(X_1, X_2, \ldots)) \tag{A.2}$$

Nótese que si h es la función indicadora, entonces esta definición es exactamente la misma que la definición de cadena de Markov. Sin embargo, (A.2) se puede generalizar a otras clases de funciones (por tanto se le llama "débil") y es particularmente útil junto con el concepto de tiempos de paro se utiliza para asegurar la convergencia de las métodos MCMC que se estudian en este trabajo.

Definición A.1.4. *(Tiempo de paro, reglas de paro y número de visitas)*

Sea $A \in \mathcal{B}(\mathcal{X})$. El primer n para el cual la cadena entra al conjunto A se define como

$$\tau_A := \text{ínf}\{n \in \mathbb{N}^+ : X_n \in A\} \tag{A.3}$$

y se conoce como tiempo de paro en A; por convención $\tau_A = \infty$ si para todo $n \in \mathbb{N}^+$ $X_n \notin A$.
Una función $\zeta(x_1, x_2, \ldots)$ es una regla de paro si el conjunto $\{\zeta = n\}$ es medible con respecto a la σ-álgebra $\sigma(X_1, \ldots, X_n)$.
Asociado con el conjunto A se define

$$\eta_A := \sum_{n=1}^{\infty} I_A(X_n)$$

y se conoce como el número de visitas á A.

Generalmente, se tiene interés en las reglas de paro de la forma (A.3) que reflejan el hecho de que τ_A toma el valor n cuando ninguno de los valores de $X_1, \ldots, X_{n-1}$ están en el conjunto A dado pero en el n-ésimo paso sí se alcanza. El siguiente resultado se conoce como Propiedad de Markov fuerte y se prueba utilizando la propiedad débil condicionando sobre $\{\zeta = n\}$.

Proposición A.1.2. *(Propiedad de Markov fuerte)*

Para toda distribución inicial ν y para cada tiempo de paro $\zeta < \infty c.s$

$$\mathbb{E}_\nu(h(X_{\zeta+1}, X_{\zeta+2}, \ldots)|x_1, \ldots, x_\zeta) = \mathbb{E}_{x_\zeta}(h(X_1, X_2, \ldots))$$

A.2. Irreductibilidad, átomos y conjuntos "pequeños"

A.2.1. Irreductibilidad

Una primera medida de la sensitividad de la cadena de Markov ante condiciones iniciales, x_0 ó ν es la irreductibilidad. Esta propiedad es muy importante en los algoritmos de cadenas de Markov Monte-Carlo ya que garantiza la convergencia sin tener que analizar a detalle el operador de transición.

Definición A.2.1. *(Irreductibilidad - caso discreto)*

En el caso discreto, una cadena de Markov es irreducible si todos los estados se comunican, i.e. si para todo $x, y \in \mathcal{X}$

$$\mathbb{P}_x(\tau_y < \infty) > 0$$

donde τ_y es el primer tiempo en el que se visita y.

Observación A.2.1.

En muchos casos $\mathbb{P}_x(\tau_y < \infty)$ es uniformemente igual a cero, por tanto, se tiene que definir una medida auxiliar φ sobre $\mathcal{B}(\mathcal{X})$ para definir de manera más formal el concepto de irreductibilidad. ▽

Definición A.2.2. *(φ-irreductibilidad)*

Sea φ una medida sobre $\mathcal{B}(\mathcal{X})$. Sea $\{X_n\}$ una cadena de Markov con kernel de transición $K(x, y)$. Se dice que $\{X_n\}$ es φ-irreducible si para todo $A \in \mathcal{B}(\mathcal{X})$, tal que $\varphi(A) > 0$, existe $n \in \mathbb{N}$ tal que $K^n(x, A) > 0$ para todo $x \in \mathcal{X}$ ó bien si $\mathbb{P}_x(\tau_A < \infty) > 0$. Se dice que $\{X_n\}$ es φ-irreducible fuertemente si para todo $A \in \mathcal{B}(\mathcal{X})$, tal que $\varphi(A) > 0$, $K(x, A) > 0$ para todo $x \in \mathcal{X}$.

Teorema A.2.1.

La cadena $\{X_n\}$ es φ-irreducible si y sólo si para cualesquiera $x \in \mathcal{X}$ y $A \in \mathcal{B}(\mathcal{X})$, tal que $\varphi(A) > 0$, se cumple alguna de las siguientes propiedades:

(i) Existe $n \in \mathbb{N}^+$ tal que $K^n(x, A) > 0$

(ii) $\mathbb{E}_x(\eta_A) > 0$

(iii) Para todo $\varepsilon \in (0, 1)$ $K_\varepsilon(x, A) > 0$

Se usa la cadena K_ε para construir un kernel estrictamente positivo en caso de que la cadena sea φ-irreducible.

La medida φ en la definición de irreductibilidad no es fundamental en el sentido de que la irreductibilidad es una propiedad intrínseca de $\{X_n\}$.

Teorema A.2.2.

Si $\{X_n\}$ es una cadena de Markov φ-irreducible, entonces existe una medida de probabilidad ψ sobre $\mathcal{B}(\mathcal{X})$ tal que

(i) $\{X_n\}$ es ψ-irreducible

(ii) Si existe una medida ψ_* tal que $\{X_n\}$ es ψ_*-irreducible, entonces ψ domina a ψ_* (i.e. $\psi_* \prec\prec \psi$)

(iii) Si $\psi(A) = 0$, entonces $\psi(\{y \in \mathcal{X} : \mathbb{P}_y(\tau_A < \infty) > 0\}) = 0$

(iv) La medida ψ es equivalente a

$$\psi_0(A) = \int_{\mathcal{X}} K_{0.5}(x, A)\varphi(dx), \quad \forall A \in \mathcal{B}(\mathcal{X})$$

i.e. $\psi \prec\prec \psi_0$ y $\psi_0 \prec\prec \psi$

A través de este resultado se puede "encontrar" (construir) una medida de irreductibilidad maximal ψ mediante la medida candidato φ.

A.2.2. Átomos y conjuntos "pequeños"

En el caso discreto, el kernel de transición necesariamente es atómico en el sentido usual, i.e. existen puntos en el espacio de estados con masa (de probabilidad) positiva. La extensión de esta noción al caso general la introdujo Nummelin (1978) y es suficiente para controlar la cadena casi de forma tan precisa como en el caso discreto.

Definición A.2.3. *(Átomo)*

Se dice que la cadena de Markov $\{X_n\}$ tiene un átomo $\gamma \in \mathcal{B}(\mathcal{X})$ si existe una medida no nula ϖ tal que para cualesquiera $x \in \gamma$, $A \in \mathcal{B}(\mathcal{X})$

$$K(x, A) = \varpi(A)$$

Si $\{X_n\}$ es ψ-irreducible, entonces se dice que el átomo es accesible si $\psi(\gamma) > 0$

Aunque esta se cumple trivialmente para todo posible valor de X_n en el caso discreto, en el caso continuo está noción generalmente es muy fuerte pues implica que el kernel de transición es constante en un conjunto de medida positiva. Una generalización más poderosa es la que se conoce como *condición de minorización*. La condición de minorización establece que existe un conjunto $C \in \mathcal{B}(\mathcal{X})$, una constante $\epsilon > 0$ y una medida de probabilidad ϖ tal que para cualesquiera $x \in C$, $A \in \mathcal{B}(\mathcal{X})$

$$K(x, A) \geq \epsilon \varpi(A) \tag{A.4}$$

Por tanto, la medida de probabilidad ν es una componente constante del kernel de transición sobre C.

La condicion de minorización (A.4) motiva la siguiente definición.

Definición A.2.4. *(Conjunto pequeño)*

Un conjunto C es pequeño si existe $m \in \mathbb{N}^+$ y una medida diferente de cero ϖ_m tales que para cualesquiera $x \in C$, $A \in \mathcal{B}(\mathcal{X})$

$$K^m \geq \varpi_m(A) \tag{A.5}$$

Una condición suficiente para que C sea pequeño es que para $m = 1$ la K_ε-cadena cumpla (A.5). El siguiente resultado establece un nexo entre conjuntos pequeños e irreductibilidad.

Teorema A.2.3.

Sea $\{X_n\}$ una cadena ψ-irreducible. Para cualquier conjunto $A \in \mathcal{B}(\mathcal{X})$ tal que $\psi(A) > 0$ existen $m \in \mathbb{N}^+$ y un conjunto pequeño $C \subseteq A$ tal que $\varpi_m(C) > 0$. Además $\mathcal{X}$ se puede particionar en conjuntos pequeños.

La partición de $\mathcal{X}$ en conjuntos pequeños se basa en un conjunto pequeño arbitrario C y la sucesión

$$C_{nm} = \{y : K^n(y, C) > 1/m\}$$

Con la condición (A.4) es más fácil exhibir a los conjuntos pequeños que a los átomos. Y no sólo eso, Meyn & Tweedie (1993) probaron que para cadenas de Markov suficientemente regulares (en un sentido topológico), todo compacto es pequeño. Sin embargo, los átomos son un caso especial de conjuntos pequeños con propiedades de estabilidad más fuertes ya que la probabilidad de transición es invariante sobre γ.

Si la condición de minorización se cumple para $\{X_n\}$ existen dos formas de obtener una cadena de Markov $\{\tilde{X}_n\}$ que tenga muchas de las propiedades de $\{X_n\}$ y tenga un átomo γ. El primer método se conoce como *separación de Nummelin* y construye una cadena con dos copias de $\{X_n\}$ (Nummelin (1978)).

Un segundo método fue desarrollado casi a la par por Athreya & Ney (1978) y utiliza un tiempo de paro para obtener un átomo.

Definición A.2.5. *(Tiempo de renovación)*

Un tiempo de renovación (ó tiempo de regeneración) es una regla de paro τ con la propiedad de que $(X_\tau, X_{\tau+1}, \ldots)$ es independiente de $(X_{\tau-1}, X_{\tau-2}, \ldots)$

Si se cumple (A.4) y si la probabilidad $\mathbb{P}_x(\tau_C < \infty)$ de regreso a C en un tiempo finito es idénticamente igual a 1 sobre $\mathcal{X}$, Athreya & Ney (1978) modificaron el kernel del transición cuando $X_n \in C$ simulando X_{n+1} como

$$X_{n+1} \sim \begin{cases} \varpi & \text{con probabilidad } \epsilon \\ \frac{K(X_n,\cdot) - \epsilon\varpi(\cdot)}{1-\epsilon} & \text{con probabilidad } 1-\epsilon \end{cases} \tag{A.6}$$

i.e. simuando X_{n+1} de ϖ con probabilidad ϵ cada vez que X_n esté en C. Esta modificación no cambia la distribución marginal de X_{n+1} dado x_n ya que para cualquier $A \in \mathcal{B}(\mathcal{X})$

$$\epsilon\varpi(A) + (1-\epsilon)\frac{K(x_n, A) - \epsilon\varpi(A)}{1-\epsilon} = K(x_n, A)$$

obteniendo tiempos de renovación para cada j tal que $X_j \in C$ y $X_{j+1} \sim \varpi$.

Ahora, claramente se ve cómo los tiempos de renovación resultan en cadenas independientes. Cuando $X_{j+1} \sim \varpi$, este evento es totalmente independiente de cualquier historia pasada pues el estado actual de la cadena no tiene efecto sobre la medida ϖ. Nótese también el papel clave de la condición de minorización. Ésta permite crear una cadena con la misma distribución marginal que la de la cadena original. Para $j > 0$, sea

$$\tau_j := \text{ínf}\{n > \tau_{j-1} : X_n \in C \text{ y } X_{n+1} \sim \varpi\}$$

la sucesión de tiempos de renovación con $\tau_0 := 0$. Athreya & Ney (1978) definieron el concepto de *cadena aumentada* (también conocida como *cadena que separa*) $\tilde{X}_n := (X_n, \tilde{\omega}_n)$ con $\tilde{\omega}_n = 1$ cuando $X_n \in C$ y X_{n+1} se genera de ϖ. Es fácil demostrar que el conjunto $\gamma \times \{1\}$ es un átomo de la cadena $\{\tilde{X}_n\}$ y la subcadena resultante $\{X_n\}$ sigue siendo de Markov con kernel de transición $K(x_n, \cdot)$.

La noción de conjunto pequeño sólo es útil en casos finitos y discretos cuando las probabilidades individuales de los estados son demasiado pequeñas para permitir una tasa razonable de renovación. En estos casos, los conjuntos pequeños se forman de colecciones de estados y ϖ se define como un mínimo de ciertas probabilidades. En otro caso, los conjuntos pequeños que se reducen a un sólo valor también son átomos.

A.2.3. Ciclos y aperiodicidad

El comportamiento de $\{X_n\}$ a veces se puede restringir de forma determinista en los movimiento de X_n a X_{n+1}. A continuación, se formalizarán dichas restricciones. Las cadenas generadas por métodos de cadenas de Markov Monte-Carlo no presentan este comportamiento y por tanto no se tendrán los inconvenientes correspondientes.

Definición A.2.6. *(Periodo)*

En el caso discreto, se define el periodo de una estado $\omega \in \mathcal{X}$ como

$$d(\omega) := mcd\{m \in \mathbb{N}^+ : K^m(\omega, \omega) > 0\}$$

Observación A.2.2.

El periodo es constante en todos los estados que se comunican con ω (i.e. es una propiedad de clase). En el caso de una cadena de Markov irreducible con espacio de estados $\mathcal{X}$ finito, la matriz de transición se puede escribir (mediante un posible reordenamiento de los estados) como una matriz en bloques

$$\begin{pmatrix} 0 & D_1 & 0 & \cdots & 0 \\ 0 & 0 & D_2 & \cdots & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ D_d & 0 & 0 & \cdots & 0 \end{pmatrix}$$

donde los bloques D_i son matrices estocásticas. Esta representación ejemplifica la visita forzada de un grupo de estados a otro con un regreso al grupo inicial cada d-ésimo paso. Si la cadena es irreducible (i.e. todos los estados se comunican) hay un único periodo. Una cadena irreducible es *aperiódica* si su periodo es 1. La extensión al caso general requiere la existencia de un conjunto pequeño. ▽

Definición A.2.7. *(Ciclo)*

Una cadena $\{X_n\}$ ψ-irreducible tiene un ciclo de longitud d si existen un conjunto pequeño C, un entero asociado M y una distribución de probabilidad ϖ_M tales que d es el maximo común divisor de

$$\{m \in \mathbb{N}^+ : \exists \delta_m > 0 \text{ tal que } C \text{ es pequeño para } \varpi_m \geq \delta \varpi_M\}$$

En general, Se puede hacer una descomposición en matrices estocásticas como la enterior. Se puede demostrar que el número d es independiente del conjunto pequeño C y que este número intrínsecamente caracteriza a la cadena $\{X_n\}$. Entonces, el periodo de $\{X_n\}$ se define como el entero más grande que satisface la Definición (A.2.7) y $\{X_n\}$ es aperiódica si $d = 1$. Si existen un conjunto pequeño A y una medida minorizante ϖ_1 tales que $\varpi_1(A)$ (si es posible ir de A á A en un solo paso se dice que la cadena es *fuertemente aperiódica*). Nótese que se puede utilizar la K_ε-cadena para transformar una cadena aperiódica en una cadena fuertemente aperiódica.

En espacio de estados discreto, si un estado $x \in \mathcal{X}$ satisface $P_{xx} > 0$, entonces la cadena es aperiódica; aunque esta no es unà condición necesaria.

Cuando la cadena tiene espacio de estados continuo y el kernel de transición tienen un componente que es absolutamente continuo con respecto a la medida de Lebesgue, con densidad $f(\cdot|x_n)$, una condición suficiente para la aperiodicidad es que $f(\cdot|x_n)$ sea positiva en una vecindad de x_n. Entonces, la cadena puede permanecer en esta vecindad

por unnnúmero arbitrario de veces antes de visitar cualquier conjunto A.

A.3. Transitoriedad y recurrencia

A.3.1. Clasificación de cadenas irreducibles

Desde un punto de vista algorítmico, una cadena de Markov debe tener buenas propiedades de estabilidad que garanticen una aproximación aceptable de modelo simulado. De hecho, la irreductibilidad garantiza que todo conjunto A será visitado por la cadena $\{X_n\}$, pero esta propiedad no es suficientemente fuerte para garantizar que las trayectorias de $\{X_n\}$ pasarán por A un número suficiente de veces.

Definición A.3.1. *(Estado transitorio)*

En un espacio de estados finito $\mathcal{X}$, se dice que un estado $\omega \in \mathcal{X}$ es transitorio si el número promedio de visitas a ω, $\mathbb{E}_\omega(\eta_\omega)$, es finito. Y se dirá que $\omega \in \mathcal{X}$ es recurrente si $\mathbb{E}_\omega(\eta_\omega) = \infty$.

En cadenas irreducibles, las propiedades de recurrencia y transitoriedad son propiedades globales (i.e. de clase) más que de estados particulares (hecho que se puede probar fácilmente con las ecuaciones de Champan-Kolmogorov). Por tanto si η_A es el número promedio de visitas a A, para cualquier $(x, y) \in \mathcal{X}^2$ o bien ocurre $\mathbb{E}_x(\eta_y) < \infty$ en el caso transitorio o bien ocurre $\mathbb{E}_x(\eta_y) = \infty$ en el caso recurrente. Entonces la cadena se llamará transitoria o recurrente (dependiendo que propiedad satisfaga) y en el caso irreducible necesariamente se cumple una de las dos.

El estudio del caso general (i.e. no necesarimente en espacio de estados discreto) se basa en cadenas con átomos; la extensión a cadenas generales (con conjuntos pequeños) se conoce como *separación* y fue propuesto por Athreya & Ney (1978). A continuación se extenderán los conceptos de recurrencia y transietoriedad.

Definición A.3.2.

(i) Se dice A es recurrente si para todo $x \in A$, $\mathbb{E}_x(\eta_A) = \infty$

(ii) Se dice que A es uniformente transitorio si existe una constante M tal que para todo $x \in A$, $\mathbb{E}_x(\eta_A) < M$

(iii) Se dice que A es transitorio si existe una cubierta de $\mathcal{X}$ de conjuntos uniformente transitorios, i.e. una colección numerable de conjuntos uniformemente transitorios B_i tales que

$$A = \bigcup_i B_i$$

Teorema A.3.1.

Sea $\{X_n\}$ una cadena de Markov ψ-irreducible con un átomo accesible γ

(i) Si γ es recurrente, entonces cualquier $A \in \mathcal{B}(\mathcal{X})$ tal que $\psi(A) > 0$ es recurrente.

(ii) Si γ es transitorio, entonces $\mathcal{X}$ es transitorio.

La propiedad (i) es más relevante en el análisis de técnicas de cadenas de Markov Monte-Carlo y se puede probar a partir de las ecuaciones de Chapman-Kolmogorov. La propiedad (ii) es más difícil de probar y se usa el hecho de que $\mathbb{P}_\gamma(\tau_\gamma < \infty) < 1$ para un conjunto transitorio cuando $\mathbb{E}_x(\eta_A)$ se descompone condicionalmente sobre las última visita a γ (Meyn & Tweedie (1993)).

Definición A.3.3. *(Cadena recurrente y cadena transitoria)*

Se dice que una cadena de Markov, $\{X_n\}$, es recurrente si

(i) Existe una medida ψ tal que $\{X_n\}$ es ψ-irreducible, y

(ii) Para todo $A \in \mathcal{B}(\mathcal{X})$ tal que $\psi(A) > 0$, $\mathbb{E}_x(\eta_A) = \infty$

Se dice que una cadena de Markov, $\{X_n\}$, es transitoria si es ψ-irreducible y $\mathcal{X}$ es transitorio.

La clasificación que se hizo en el Teorema anterior se puede extender fácilmente a ca-

denas fuertemente aperiódicas ya que deben cumplir la condición de minorización (A.5) y por tanto se pueden descomponer como (A.4) cuando la cadena $\{X_n\}$ y su versión separada $\{\tilde{X}_n\}$ son ambas recurrentes ó ambas transitorias.

La generalización a una cadena irreducible arbitraria se sigue de las propiedades de la K_ε-cadena correspondiente (que es fuertemente aperiódica) mediante la relación

$$\sum_{n=0}^{\infty} K^n = \frac{1-\varepsilon}{\varepsilon} \sum_{n=0}^{\infty} K_\varepsilon^n$$

ya que

$$\mathbb{E}_x(\eta_A) = \sum_{n=0}^{\infty} K^n(x, A) = \frac{\varepsilon}{1-\varepsilon} = \sum_{n=0}^{\infty} K_\varepsilon^n(x, A)$$

Por tanto, se obtiene el siguiente resultado de clasificación.

Teorema A.3.2.

Una cadena ψ-irreducible es o bien recurrente o bien transitoria.

A.3.2. Criterio para recurrencia

Los resultados anteriores establecían una clara dicotomía entre transitoriedad y recurrencia para cadenas de Markov irreducibles. Sin embargo, dadas las definiciones de cadena de Markov recurrente, es útil tener un criterio simple para recurrencia. Naturalmente, la primera propuesta que surge como analogía con cadenas de Markov con espacio de estados discreto es mediante conjuntos pequeños.

Proposición A.3.1.

Una cadena ψ-irreducible $\{X_n\}$ es recurrente si existe un conjunto pequeño C tal que $\psi(C) > 0$ y para cualquier $x \in C$ $\mathbb{P}_x(\tau_C < \infty) = 1$

Demostración:

Primero se probará que el conjunto C es recurrente. Sea $x \in C$. Considérese $u_n := K^n(x, C)$ y $f_n := \mathbb{P}_x(X_n \in C, X_{n-1} \notin C, \ldots, X_1 \notin C,)$, que es la probabilidad de visitar por primera vez C en el $n$ésimo instante. También defínase

$$\tilde{U}(s) := 1 + \sum_{n=1}^{\infty} u_n s^n, \quad Q(S) := \sum_{n=1}^{\infty} f_n s^n$$

La ecuación

$$u_n = f_n + f_{n-1}u_1 + \ldots + f_1 u_{n-1} \tag{A.7}$$

describe la relación entre la probabilidad de una visita a C al tiempo n y las probabilidades de primera visita a C. Esto implica que

$$\tilde{U}(s) = \frac{1}{1 - Q(s)}$$

Sin embargo, nótese que $\tilde{U}(s) = \mathbb{E}_x(\eta_C) = \infty$ y $Q(1) = \mathbb{P}_x(\tau_C < \infty) = 1$. La ecuación (A.7) también es válida para la cadena $\{\tilde{X}_n\}$ ya que una visita a $C \times \{0\}$ asegura la independencia de la renovación. Ya que $\mathbb{E}_x(\eta_C)$ asociada con $\{X_n\}$ es más grande que $\mathbb{E}_{\tilde{x}}(\eta_{C\times\{0\}})$ asociada con $\{\tilde{X}_n\}$ y $\mathbb{P}_x(\tau_C < \infty)$ de $\{X_n\}$ es igual a $\mathbb{P}_{\tilde{x}}(\tau_{C\times\{0\}} < \infty)$ de $\{\tilde{X}_n\}$, la recurrencia se puede extender de $\{\tilde{X}_n\}$ a $\{X_n\}$. La recurrencia de $\{\tilde{X}_n\}$ se sigue del teorema anterior pues $C \times \{0\}$ es un átomo recurrente para $\{\tilde{X}_n\}$. □

A.3.3. Harris-recurrencia

Es posible hacer más fuertes las propiedades de estabilidad de una cadena $\{X_n\}$ pidiendo no sólo un número promedio infinito de visitas a cada conjunto pequeño sino también un número infinito de visitas para cada trayectoria de la cadena de Markov. Recuérdese que η_A es el número de veces que la cadena pasa por A, y se está considerando $\mathbb{P}_x(\eta_A = \infty) = 1$ como la probabilidad de visitar A un número infinito de veces.

Definición A.3.4. *(Harris-recurrencia)*

Se dice que un conjunto A es Harris-recurrente si para cualquier $x \in A$, $\mathbb{P}_x(\eta_A = \infty) = 1$. La cadena $\{X_n\}$ es Harris-recurrente si existe una medida ψ tal que $\{X_n\}$ es ψ-irreducible y para todo conjunto A tal que $\psi(A) > 0$ A es Harris-recurrente.

Recuérdese que la recurrencia equivalía a que $\mathbb{E}_x(\eta_\gamma) = \infty$ y es una condición más débil que $\mathbb{P}_x(\eta_A = \infty) = 1$. La siguiente proposición expresa a la Harris-recurrencia como una condición sobre $\mathbb{P}_x(\tau_A < \infty)$.

Proposición A.3.2.

Si para cualquier $A \in \mathcal{B}(\mathcal{X})$, $\mathbb{P}_x(\tau_A < \infty) = 1$ para todo $x \in A$, entonces para cualquier $x \in \mathcal{X}$, $\mathbb{P}_x(\eta_A = \infty) = 1$ y $\{X_n\}$ es Harris-recurrente.

Demostración:

Sea $B \in \mathcal{B}(\mathcal{X})$. El número promedio de visitas a B antes de visitar á A es

$$U_A(x, B) := \sum_{n=1}^{\infty} \mathbb{P}_x(X_n \in B, \tau_A \geq n)$$

Entonces, $U_A(x, A) = \mathbb{P}_x(\tau_A < \infty)$ ya que si $B \subseteq A$,

$$\mathbb{P}_x(X_n \in B, \tau_A \geq n) = \mathbb{P}_x(X_n \in B, \tau = n) = \mathbb{P}_x(\tau_B = n).$$

Análogamente, si $\tau_A(k)$ (para $k \in \mathbb{N}^+$) es el tiempo de la de k-ésima visita a A, entonces $\tau_A(k)$ satisface que para todo $x \in A$

$$\mathbb{P}_x(\tau_A(2) < \infty) = \int_A \mathbb{P}_y(\tau_A < \infty) U_A(x, dy) = 1$$

e inductivamente

$$\mathbb{P}_x(\tau_A(k+1) < \infty) = \int_A \mathbb{P}_y(\tau_A(k) < \infty) U_A(x, dy) = 1$$

Como $\mathbb{P}_x(\eta_A \geq k) = \mathbb{P}_x(\tau_A(k) < \infty)$ y

$$\mathbb{P}_x(\eta_A = \infty) = \lim_{k \to \infty} \mathbb{P}_x(\eta_A \geq k)$$

se tiene que $\mathbb{P}_x(\eta_A = \infty) = 1$ para cualquier $x \in A$. □

Nótese que la propiedad de Harris-recurrencia sólo es necesaria cuando $\mathcal{X}$ es no numerable. Si $\mathcal{X}$ es finito o numerable, se puede probar que $\mathbb{E}_x(\eta_x) = \infty$ si y sólo si para todo $x \in \mathcal{X}$ $\mathbb{P}_x(\tau_x < \infty) = 1$ con un argumento similar al que se usó en la prueba del Teorema (A.3.1). En el caso general, se puede demostrar que si $\{X_n\}$ es Harris-recurrente, entonces para todo $x \in \mathbb{X}$, $B \in \mathcal{B}(\mathcal{X})$ tal que $\psi(B) > 0$, $\mathbb{P}_x(\eta_B = \infty) = 1$. Esta propiedad proporciona una condición suficiente para Harris-recurrencia, generalizando la Proposición (A.3.1).

Teorema A.3.3.

Si $\{X_n\}$ es una cadena de Markov ψ-irreducible con un conjunto pequeño C tal que para todo $x \in \mathcal{X}$ $\mathbb{P}_x(\tau_C < \infty) = 1$, entonces $\{X_n\}$ es Harris recurrente.

Compárese este Teorema con la Proposición (A.3.1), en donde $\mathbb{P}_x(\tau_C < \infty) = 1$ sólo para $x \in C$. Este Teorema también permite reemplazar la recurrencia por Harris recurrencia en el siguiente Teorema.

Teorema A.3.4.

Sea $\{X_n\}$ una cadena de Markov ψ-irreducible. Si existen un conjunto pequeño C y una función V tales que

$$C_V(n) := \{x : V(x) \leq n\}$$

es un conjunto pequeño, para cualquier n; entonces las cadena es recurrente si sobre C^c se tiene que $\Delta V(x) \leq 0$

Tierney (1994) y Chan & Geyer (1994) analizaron el papel de la Harris recurrencia en algoritmos MCMC y observaron que la mayoría de los algoritmos cumplían la Harris-recurrencia.

A.4. Medidas invariantes

La cadena $\{X_n\}$ tendrá más estabilidad si la distribución de X_n no depende de n. De manera más formal, esto requiere de la existencia de una distribución de probabilidad π tal que si $X_n \sim \pi$, entonces $X_{n+1} \sim \pi$ y los métodos de cadenas de Markov Monte-Carlo se basan en el hecho de esta existencia, que define un tipo particular de convergencia conocida como recurrencia positiva. Las cadenas de Markov que se construyen mediante métodos de cadenas de Markov Monte-Carlo tienen buenas propiedades de estabilidad (excepto en algunos casos muy patológicos). Por tanto, se tiene que discutir ampliamente los conceptos de medidas invariantes y recurrencia positiva.

Definición A.4.1. *(Medida invariante)*

Sea π una medida σ-finita. Se dice que π es invariante para el kernel de transición $K(\cdot,\cdot)$ (y para la cadena asociada) si para todo $B \in \mathcal{B}(\mathcal{X})$

$$\pi(B) = \int_{\mathcal{X}} K(x,B)\pi(dx)$$

Observación A.4.1.

Cuando existe una medida de probabilidad invariante para una cadena ψ-irreducible (y por tanto recurrente por el Teorema (A.3.2)) se dice que la cadena es *positiva*. Las cadenas recurrentes en las que no existe una medida de probabilidad invariante finita se conocen como *recurrentes nulas*. ▽

La distribución invariante también se conoce como *estacionaria* si π es una medida de probabilidad, ya que $X_0 \sim \pi$ implica que para todo $n \in \mathbb{N}^+$, $X_n \sim \pi$, por tanto la cadena es estacionaria en distribución. Cuando π no es finita es más difícil interpretar en términos del comportamiento de la cadena. Se puede demostrar que si la cadena es irreducible y tiene una medida invariante σ-finita, ésta medida es única excepto quizá por una constante multiplicativa. El nexo entre la positividad y la recurrencia se enuncia en la siguiente proposición y formaliza el hecho intuitivo de que la existencia de una medida invariante evita que la masa de probabilidad se "escape a infinito".

Proposición A.4.1.

Si la cadena $\{X_n\}$ es positiva, entonces es recurrente.

Demostración:

Si $\{X_n\}$ es transitoria, existe una cubierta de $\mathcal{X}$ por conjuntos transitoriamente uniformes, A_j, con las cotas

$$\mathbb{E}_x(\eta_{A_j}) \leq M_j$$

para cualesquiera $x \in A_j$, $j \in \mathbb{N}$. Entonces por la invarianza de π

$$\pi(A_j) = \int K(x,A_j)\pi(dx) = \int K^n(x,A_j)\pi(dx)$$

Entonces, para cuaquier $k \in \mathbb{N}$

$$k\pi(A_j) = \sum_{n=0}^{k} \int K^n(x, A_j)\pi(dx) \leq \int \mathbb{E}_x(\eta_{A_j})\pi(dx) \leq M_j,$$

ya que por la definición de número de visitas á A se sigue que $\sum_{n=0}^{k} K^n(x, A_j) \leq \mathbb{E}_x(\eta_{A_j})$. Haciendo k tender a ∞ se tiene que para cualquier $j \in \mathbb{N}$ $\pi(A_j) = 0$ y por tanto no se puede obtener una medida de probabilidad invariante □

Por tanto, ahora se puede hablar de *cadenas positivas* o *cadenas Harris-positivas* sin la denominación de recurrente y Harris-recurrente. La proposición anterior es útil cuando se puede probar la positividad de $\{X_n\}$, sin embargo, las cadenas generadas por métodos de cadenas de Markov Monte-Carlo por construcción tienen distribución invariante.

Un resultado clásico de cadenas de Markov irreducibles con espacio de estados finitos es que cuando existe la distribución estacionaria, ésta está dada por

$$\pi_x = [\mathbb{E}_x(\tau_x)]^{-1}$$

con la interpretación de $\mathbb{E}_x(\tau_x)$ como el número promedio de pasos entre dos visitas a x (este resultado se conoce como Teorema de Kac). También se tiene que $[\mathbb{E}_x(\tau_x)]^{-1}$ es el eigenvector asociado con el eigenvalor 1 de la matriz de transición P. Ahora se enunciará un resultado más general cuando $\{X_n\}$ tiene un átomo.

Teorema A.4.1.

Sea $\{X_n\}$ ψ-irreducible con átomo γ. La cadena es positiva si y sólo si $\mathbb{E}_\gamma(\tau_\gamma) < \infty$. En este caso, la distribución invariante para $\{X_n\}$, π, satisface

$$\pi(\gamma) = [\mathbb{E}_\gamma(\tau_\gamma)]^{-1}$$

Observación A.4.2.

La notación $\mathbb{E}_\gamma(\cdot)$ es válida en este caso pues el kernel de transición es el mismo para todo $x \in \gamma$ (por la definición de átomo). Este teorema establece que la positividad es un propiedad de estabilidad más fuerte que la recurrencia ya que

$$\mathbb{P}_\gamma(\tau_\gamma = \infty) = 0$$

▽

Demostración (del Teorema):

Si $\mathbb{E}_\gamma(\tau_\gamma) < \infty$, entonces $\mathbb{P}_\gamma(\tau_\gamma < \infty) = 1$, y por la proposición (A.3.1) $\{X_n\}$ es recurrente.

Considérese una medida, π, definida como

$$\pi(A) = \sum_{n=1}^{\infty} \mathbb{P}_\gamma(X_n \in A, \tau_\gamma \geq n)$$

$$\begin{aligned} \int K(x, A_j)\pi(dx) &= \pi(\gamma)K(\gamma, A) + \int_{\gamma^c} \sum_{n=1}^{\infty} K(x_n, A)\mathbb{P}_\gamma(\tau_\gamma \geq n, dx_n) \\ &= K(\gamma, A) + \sum_{n=2}^{\infty} \mathbb{P}_\gamma(X_n \in A, \tau_\gamma \geq n) = \pi(A) \end{aligned}$$

Esta también es finita pues

$$\begin{aligned} \pi(A) &= \sum_{n=1}^{\infty} \mathbb{P}_\gamma(\tau_\gamma \geq n) = \sum_{n=1}^{\infty} \sum_{m=n}^{\infty} \mathbb{P}_\gamma(\tau_\gamma = m) \\ &= \sum_{m=1}^{\infty} m\mathbb{P}_\gamma(\tau_\gamma = m) = \mathbb{E}_\gamma(\tau_\gamma) < \infty \end{aligned}$$

Como π es invariante cuando $\{X_n\}$ es positiva, la unicidad de la distribución invariante implica que $\pi(\mathcal{X}) < \infty$ y por tanto $\mathbb{E}_\gamma(\tau_\gamma) < \infty$. Renormalizando π a $\pi/\pi(\mathcal{X})$ se tiene que $\pi(\gamma) = [\mathbb{E}_\gamma(\tau_\gamma)]^{-1}$. □

Desde una perspectiva más clásica, se puede tratar el caso general mediante la separación de $\{X_n\}$ en $\{\tilde{X}_n\}$ (que tiene un átomo) y la medida invariante de $\{\tilde{X}_n\}$ induce una medida invariante para $\{X_n\}$ mediante marginalización.

Teorema A.4.2.

Si $\{X_n\}$ es una cadena recurrente, entonces existe una medida invariante σ-finita única, excepto quizá por múltiplos escalares.

A.5. Ergodicidad y convergencia

A.5.1. Ergodicidad

Considerar a la cadena de Markov $\{X_n\}$ desde una perspectiva temporal es natural, por tanto es importante considerar el comportamiento límite de X_n. La existencia y unicidad de una distribución invariante π hace que ésta sea un candidato natural para ser la distribución límite. Ahora se estudiarán condiciones suficientes sobre $\{X_n\}$ para que X_n se distribuya asintóticamente como π. Los siguientes teoremas son resultados de convergencia de cadenas de Markov fundamentales y motivan la implementación de métodos de cadenas de Markov Monte-Carlo. Sin embargo, dichs resultados son difíciles de establecer y la mayoría de las pruebas se harán para el caso de espacio de estados discreto, la extensión al caso general se puede revisar en Meyn & Tweedie (1993).

Hay muchas condiciones que se pueden establecer sobre la convergencia de P^n (la distribución de X_n) hacia π. Quizá una de las más importantes es el de la *ergodicidad*, i.e. la independencia de las condiciones iniciales.

Definición A.5.1. *(Átomo ergódico)*

Sea $\{X_n\}$ una cadena Harris-positiva con distribución invariante π. Se dice que un átomo γ es ergódico si

$$\lim_{n\to\infty} |K^n(\gamma,\gamma) - \pi(\gamma)| = 0$$

Observación A.5.1.

En el caso numerable, la existencia de un átomo ergódico es suficiente para establecer la convergencia, de acuerdo con la norma de variación total

$$\|\nu_1 - \nu_2\|_{TV} := \sup_A |\nu_1(A) - \nu_2(A)|$$

▽

Proposición A.5.1.

Sea $\{X_n\}$ una cadena Harris-positiva sobre $\mathcal{X}$, numerable. Si existe un átomo ergódico $\gamma \subset \mathcal{X}$, entonces para cualquier $x \in \mathcal{X}$

$$\lim_{n\to\infty} \|K^n(x,\cdot) - \pi\|_{TV} = 0$$

Demostración:

El primer paso sigue una fórmula de descomposición que se conoce como "primera entrada y última salida":

$$\begin{aligned} K^n(x,y) &= \mathbb{P}_x(X_n = y, \tau_\gamma \geq n) && \text{(A.8)} \\ &+ \sum_{j=1}^{n-1} \left(\sum_{k=1}^{j} \mathbb{P}_x(X_k \in \gamma, \tau_\gamma \geq k) K^{j-k}(\gamma,\gamma) \right) \mathbb{P}_\gamma(X_{n-j} = y, \tau_\gamma \geq n-j) && \text{(A.9)} \end{aligned}$$

que relaciona $K^n(x,y)$ con la última visita de γ. Esto muestra que la influencia del valor inicial x ya que $\mathbb{P}_x(X_n = y, \tau_\gamma \geq n)$ converge a 0. La expresión

$$\pi(A) = \sum_{n=1}^{\infty} \mathbb{P}_\gamma(X_n \in A, \tau_\gamma \geq n)$$

de la medida invariante implica que

$$\pi(y) = \pi(\gamma) \sum_{j=1}^{\infty} \mathbb{P}_\gamma(X_j \in A, \tau_\gamma \geq j)$$

y con estás dos expresiones se obtiene que

$$\begin{aligned} \|K^n(x,\cdot) - \pi\|_{TV} &= \sum_y |K^m(x,y) - \pi(y)| \leq \sum_y \mathbb{P}_x(X_n = y, \tau_\gamma \geq n) \\ &+ \sum_y \sum_{j=1}^{n-1} \left| \sum_{k=1}^{j} \mathbb{P}_x(X_k \in \gamma, \tau_\gamma = k) K^{j-k}(\gamma,\gamma) - \pi(\gamma) \right| \mathbb{P}_\gamma(X_{n-j} = y, \tau_\gamma \geq n-j) \\ &+ \sum_y \pi(\gamma) \sum_{j=n-1}^{\infty} \mathbb{P}_\gamma(X_j = y, \tau_\gamma \geq j) \end{aligned}$$

El segundo paso de la prueba es demostrar que cada término en la descomposición se va a cero cuando n tiende a ∞. El primer término $\mathbb{P}_x(\tau_\gamma \geq n)$ tiende a cero ya que la cadena es Harris-recurrente. El tercer sumando es el residuo de la serie convergente

$$\sum_y \pi(\gamma) \sum_{j=n-1}^{\infty} \mathbb{P}_\gamma(X_j = y, \tau_\gamma \geq j) = \sum_y \pi(y)$$

El término de en medio es la suma sobre las y's de la convolución de dos sucesiones $a_n = |\sum_{k=1}^{n} \mathbb{P}_x(X_k \in \gamma, \tau_\gamma = k) K^{n-k}(\gamma,\gamma) - \pi(\gamma)|$ y $b_n = \mathbb{P}_\gamma(X_n = y, \tau_\gamma \geq n)$. La sucesión $\{a_n\}$ converge a cero ya que el átomo γ es ergódico. □

La descomposición (A.8) es muy constructiva ya que muestra el papel del átomo γ como generador de un proceso de renovación. Ahora se estudiará una extensión que permite trabajar con el caso general utilizando técnicas de acoplamiento.

El principio de acoplamiento utiliza dos cadenas $\{X_n\}$ y $\{X'_n\}$ con el mismo kernel, el evento de "acoplamiento" ocurre cuando dichas cadenas se encuentran en γ, i.e. al primer tiempo n_0 tal que $X_{n_0} \in \gamma$ y $X'_{n_0} \in \gamma$. Después de este instante, las propiedades de $\{X_n\}$ y $\{X'_n\}$ son idénticas y si una de las dos cadenas es estacionaria, no hay dependencia de las condiciones iniciales de ambas cadenas. Por tanto, si se puede mostrar que el tiempo de acoplamiento (i.e. el tiempo que pasa para que ambas cadenas se encuentren) es finito casi para todo punto inicial, entonces la ergodicidad de la cadena se cumple.

Para un átomo recurrente γ sobre un espacio numerable $\mathcal{X}$ sea $\tau_\gamma(k)$ la k-ésima visita a γ (k=1,2,...) y sea $p = (p(1), p(2), \ldots)$ la distribución del *tiempo de excursión*,

$$S_k = \tau_\gamma(k+1) - \tau_\gamma(k)$$

entre dos visitas a γ. Si $q = (q(0), q(1), \ldots)$ es la distribución de $\tau_\gamma(1)$ (que depende de las condiciones inciales x_0 ó ν), entonces la distribución de $\tau_\gamma(n+1)$ está dada por la convolución $q \star p^{n\star}$ (i.e. la distribución de la suma de n variables aleatorias independientes e idénticamente distribuidas de acuerdo a p y una variable que se distribuye de acuerdo a q) ya que

$$\tau_\gamma(n+1) = S_n + \ldots + S_1 + \tau_{\gamma(1)}$$

Por tanto, se considerarán dos sucesiones $\{S_i\}$ y $\{S'_i\}$ tales que $S_1, S_2, \ldots$ y $S'_1, S'_2, \ldots$ son independientes e idénticamente distribuidas de p con $S_0 \sim q_0$ y $S'_0 \sim r$. También defínase

$$Z_q(n) := \sum_{j=0}^{n} I_{S_1+\ldots+S_j=n}, \quad Z_r(n) := \sum_{j=0}^{n} I_{S'_1+\ldots+S'_j=n}$$

que corresponden a los eventos de que las cadenas $\{X_n\}$ y $\{X'_n\}$ visiten γ al tiempo n. Entonces, el tiempo de acoplamiento está dado por

$$T_{qr} = \text{mín}\{j : Z_q(j) = Z_r(j) = 1\}$$

que satisface el siguiente lema.

Lema A.5.1.

Si el tiempo medio de excursión satisface

$$m_p = \sum_{n=0}^{\infty} np(n) < \infty$$

y si p es aperiódico (el máximo común divisor del soporte de p es 1), entonces el tiempo de acoplamiento T_{pq} es finito c.s., i.e. para todo q,

$$\mathbb{P}(T_{pq} < \infty) = 1$$

Si p es aperiódica con media finita m_p, entonces Z_p satisface

$$\lim_{n\to\infty} |\mathbb{P}(Z_q(n) = 1) - m_q^{-1}| = 0$$

Por tanto, la probabilidad de visitar γ al tiempo n es asintóticamente independiente de la distribución inicial.

Teorema A.5.1.

Para una cadena de Markov positivo recurrente y aperiódica sobre un espacio de estados numerable y para cada estado inicial x se cumple

$$\lim_{n\to\infty} \|K^n(x,\cdot) - \pi\|_{TV} = 0$$

Demostración:

Como $\{X_n\}$ es positiva recurrente, $\mathbb{E}_\gamma(\tau_\gamma)$ es finita. Por tanto, m_p es finita, $\lim_{n\to\infty} |\mathbb{P}(Z_q(n) = 1) - m_q^{-1}| = 0$ y todo átomo es ergódico. Entonces, por la Proposición (A.5.1)

$$\lim_{n\to\infty} \|K^n(x,\cdot) - \pi\|_{TV} = 0$$

□

Para espacios de estados generales, se necesita la Harris-recurrencia para asegurar la convergencia de K^n a π.

Observación A.5.2.

Otra caracterización de la Harris-recurrencia es la convergencia de $\|K_x^n - \pi\|_{TV}$ a 0 para todo valor x en vez de para casi todo valor x. ▽

Teorema A.5.2.

Si $\{X_n\}$ es Harris-positiva y aperiódica, entonces

$$\lim_{n\to\infty} \left\| \int K^n(x,\cdot)\nu(dx) - \pi \right\|_{TV} = 0$$

para toda distribución inicial ν.

Teorema A.5.3.

Si π es una distribución invariante para P, entonces

$$\left\| \int K^n(x,\cdot)\nu(dx) - \pi \right\|_{TV} = 0$$

es decreciente como función de n

Demostración:

Primero, una definición equivalente de la norma de variación total es

$$\|\nu\|_{TV} = \frac{1}{2} \sup_{|h|<1} \left| \int h(x)\nu(dx) \right|$$

Entonces,

$$\begin{aligned} 2\left\|K^{n+1}(x,\cdot)\nu(dx) - \pi\right\|_{TV} &= \sup_{|h|<1} \left| \int h(x)K^{n+1}(x,dy)\nu(dx) - \int h(y)\pi(dy) \right| \\ &= \sup_{|h|<1} \left| \int h(y) \int K^n(x,dw)K(w,dy)\nu(dx) - \int h(y) \int K(w,dy)\pi(dw) \right| \end{aligned}$$

por definición $K^{n+1}(x,dy) = \int K^n(x,dw)K(w,dy)$ y por la invarianza de π, $\pi(dy) = \int K(w,dy)\pi(dw)$. Reacomodando términos se tiene que

$$
\begin{aligned}
2\left\|K^{n+1}(x,\cdot)\nu(dx)-\pi\right\|_{TV} &= \sup_{|h|<1}\left|\int\left(\int h(y)K(w,dy)\right)K^n(x,dw)\nu(dx)-\int\left(\int h(y)K(w,dy)\right)\pi(dw)\right| \\
&\leq \sup_{|h|<1}\left|\int h(w)K^n(x,dw)\nu(dx)-\int h(w)\pi(dw)\right|
\end{aligned}
$$

donde la desigualdad se sigue del hecho de que la cantidad entre paréntesis es una función con norma menor que 1. □

Observación A.5.3.

La equivalencia

$$\|\nu\|_{TV}=\frac{1}{2}\sup_{|h|<1}\left|\int h(x)\nu(dx)\right|$$

implica la convergencia

$$\lim_{n\to\infty}|\mathbb{E}_\nu(h(X_n))\mathbb{E}^\pi(X)|=0$$

para cualquier función acotada h. ▽

Teorema A.5.4.

Sea $\{X_n\}$ positiva, recurrente y aperiódica.

(i) Si $\mathbb{E}^\pi(h(X))=\infty$, entonces para cada $x\in\mathcal{X}$

$$\lim_{n\to\infty}\mathbb{E}_x(|h(X_n)|)=\infty$$

(ii) Si $\int|h(x)|\pi(dx)<\infty$, entonces

$$\lim_{n\to\infty}\sup_{|m(x)|\leq|h(x)|}|\mathbb{E}_y(m(X_n))-\mathbb{E}^\pi(m(X))|=0$$

sobre conjuntos pequeños C tales que

$$\sup_{y\in C}\mathbb{E}_y\left[\sum_{t=0}^{\tau_C-1}h(X_t)\right]<\infty$$

A.5.2. Convergencia geométrica

La convergencia del límite del inciso (i) del Teorema anterior, de la esperanza de $h(x)$ al tiempo n a la esperanza de $h(x)$ bajo la distribución estacionaria π asegura de cierta manera el "buen comportamiento" de la cadena $\{X_n\}$ sin importar el valor inicial X_0 (o su distribución). Una descripción más precisa de las propiedades de convergencia que involucran el estudio de la velocidad de convergencia de K^n a π. Una evaluación de esta velocidad es importante para los algoritmos de cadenas de Markov Monte-Carlo en el sentido de que se estudian las reglas de paro de estos algoritmos. También se requiere una velocidad de convergencia mínima para que se pueda aplicar el Teorema del Límite Central.

Para estudiar la velocidad de convergencia de manera más precisa, se tiene que definir una extensión de la norma de variación total, que se denota por $\|\cdot\|_h$, que permite una cota superior diferente de uno sobre las funciones. La generalización se define como

$$\|\nu\|_h = \sup_{|g|\leq h} \left| \int g(x)\nu(dx) \right|$$

Definición A.5.2.

Una cadena $\{X_n\}$ es geométricamente h-ergódica, con $h \geq 1$ sobre $\mathcal{X}$, si $\{X_n\}$ es Harris-positiva, con distribución estacionaria π tal que $\mathbb{E}^\pi(h) < \infty$ y existe $r_h > 1$ tal que para todo $x \in \mathcal{X}$

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

El caso $h = 1$ corresponde a la *ergodicidad geométrica* de $\{X_n\}$

La h-ergodicidad geométrica significa que $\|K^n(x,\cdot) - \pi\|_h$ es decreciente al menos a velocidad geométrica ya que

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

implica que

$$\|K^n(x,\cdot) - \pi\|_h < \frac{M}{r_h^n}$$

con

$$M = \sum_{n=1}^{\infty} r_h^n \|K^n(x,\cdot) - \pi\|_h$$

Si $\{X_n\}$ tiene un átomo, entonces para cualquier número real $r > 1$,

$$\mathbb{E}_x\left[\sum_{n=1}^{\tau_\gamma} h(X_n)r^n\right] < \infty, \quad \sum_{n=1}^{\infty} |\mathbb{P}_\gamma(X_n \in \gamma) - \pi(\gamma)|r^n < \infty$$

La serie asociada con $|\mathbb{P}_\gamma(X_n \in \gamma) - \pi(\gamma)|r^n$ converge fuera del círculo unitarioi la serie de potencias asociadas con $|\mathbb{P}_\gamma(\tau_\gamma = n)$ converge para valores de $|r|$ estrictamente mayores que uno. Este resultado se conoce como Teorema de Kendall.

Definición A.5.3. *(Átomo geométricamente ergódico y átomo de Kendall)*

Se dice que un átomo accesible γ es geométricamente ergódico si existe $r > 1$ tal que

$$\sum_{n=1}^{\infty} |K^n(\gamma, \gamma) - \pi(\gamma)|r^n$$

y se dice que γ es un átomo de Kendall si existe $\kappa > 1$ tal que

$$\mathbb{E}_\gamma(\kappa^{\tau_\gamma}) < \infty$$

Teorema A.5.5.

Sea $\{X_n\}$ es ψ-irreducible, con distribución invariante π, si existe un átomo geométricamente ergódico γ, entonces existen $r > 1$, $\kappa > 1$ y $R < \infty$ tal que para casi todo $x \in \mathcal{X}$

$$\sum_{n=1}^{\infty} r^n \|K^n(x, \cdot) - \pi\|_{TV} < R\mathbb{E}_x(\kappa^{\tau_\gamma}) < \infty$$

A.5.3. Ergodicidad uniforme

La propiedad de ergodicidad uniforme es más fuerte que la ergodicidad geométrica en el sentido de que la tasa de convergencia geométrica debe ser uniforme sobre todo el espacio. Esto se analizará más adelante en las aplicaciones del Teorema del Límite Central.

Definición A.5.4. *(Ergodicidad uniforme)*

Se dice que la cadena $\{X_n\}$ es uniformemente ergódica si

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

La ergodicidad se puede establecer mediante alguna de las siguientes propiedades equivalentes:

Teorema A.5.6.

Las siguientes afimaciones son equivalentes:

(i) $\{X_n\}$ es uniformemente ergódica

(ii) Existen $R < \infty$ y $r < 1$ tales que

$$\forall x \in \mathcal{X} \quad \|K^n(x,\cdot) - \pi\|_{TV} < \frac{R}{r^n}$$

(iii) $\{X_n\}$ es aperiódica y $\mathcal{X}$ es un conjunto pequeño

(iv) $\{X_n\}$ es aperiódica y existen un conjunto pequeño C y $\kappa > 1$ tales que

$$\sup_{x\in\mathcal{X}} \mathbb{E}_x(\kappa^{\tau_C}) < \infty$$

Observación A.5.4.

Si $\mathcal{X}$ es pequeño, existe una distribución de probabilidad, φ, sobre $\mathcal{X}$ y constantes $\varepsilon < 1$, $\delta > 0$ y n tal que si $\varphi(A) > \varepsilon$, entonces

$$\inf_{x\in\mathcal{X}} K^n(x,A) > \delta$$

Esta propiedad a veces se conoce como *condición de Doeblin*. Este requisito demuestra la fortaleza de la ergodicidad uniforme y sugiere algunos problemas con respecto a su verificación. En el caso finito, la ergodicidad uniforme se puede obtener del hecho de que X sea pequeño ya que

$$\mathbb{P}(X_{n+1} = y | X_n = x) \geq \inf_z p_{zy} = \rho_y, \quad x, y \in \mathcal{X}$$

lleva a la elección de una medida de minorización ϖ como

$$\varpi(y) = \frac{\rho_y}{\sum_{z\in\mathcal{X}} \rho_z}$$

siempre que $\rho_y > 0$ para algún $y \in \mathcal{X}$. Si $\{X_n\}$ es recurrente y aperiódica, esta condición de positividad la cumple la cadena $\{Y_m\} = \{X_{nd}\}$ para d suficientemente grande. $\triangledown$

A.6. Teoremas límite

Aunque ya se han motivado algunos resultados para justificar algunas propiedades de los algoritmos de cadenas de Markov Monte-Carlo, en lo que resta de este apéndice se analizará lo esencial para el procesamiento de dichos algoritmos. De hecho, los diferentes resultados de convergencia (básicamente los de ergodicidad) que se estudiaron con anterioridad sólo involucran a la medida de probabilidad P_x^n (bajo diferentes normas) que sólo considera a la cadena $\{X_n\}$ en el instante n, i.e. se analizan las propiedades probabilísticas del comportamiento promedio de la cadena en un instante fijo. Tales propiedades justifican algunos métodos de simulación pero son de menor importancia para el control de la convergencia de una simulación dada, donde las propiedades de la realización $\{x_n\}$ de la cadena son las únicas características que interesan. Meyn & Tweedie (1993)llaman a este tipo de características propiedades de trayectorias muestrales.

Se debe considerar la diferencia entre el análisis probabilístico (que describe el comportamiento promedio de las muestras) y la inferencia estadística (que se debe razonar por indcutivamente a partir e la muestra observada). Aunque las propiedades probabilísticas se pueden justificar o refutar, estas no contradicen el hecho de que el análisis estadístico se debe hacer condicionado a la muestra observada. Desde una perspectiva de cadenas de Markov con un análisis condicional se pueden aprovechar las propiedades de convergencia de P_x^n a π sólo para verificar la convergencia, a las cantidades de interés, de funciones de la trayectoria observada de la cadena. El hecho de que $\|P_x^n - \pi\|$ sea cercano a 0 ó incluso converja geométricamente rápido a 0 con velocidad ρ^n $(0 < \rho < 1)$ no da información directa acerca de la única observación disponible de P_x^n, X_n.

Los problemas de aplicar directamente los teoremas clásicos de convergencia (Leyes de Grandes Números, Ley del logaritmo iterado, Teorema Central del Límite) de la muestra $(X_1, \ldots, X_n)$ generalmente se deben a la estructura de dependencia Markoviana entre las observaciones X_i y la no estacionariedad de la sucesión. Sólo si $X_0 \sim \pi$, la cadena será estacionaria. Ya que es equivalente integrar sobre las condiciones iniciales y eliminar la necesidad de análisis condicional. Esta condición especialmente rara en métodos de cadenas de Markov Monte-Carlo.

Entonces se supondrá que la cadena empieza de un punto X_0 cuya distribución no es la distribución estacionaria de la cadena y por tanto se trabajará directamente con cadenas no estacionarias. Se empezará con algunos resultados de convergencia análogos a las Leyes de Grandes Números que se en la literatura de simulación se conocen como *Teoremas ergódicos*.

A.6.1. Teoremas ergódicos

Dadas $X_1, \ldots, X_n$ observaciones de una cadena de Markov, ahora se estudiará el comportamiento de las sumas parciales

$$S_n(h) = \frac{1}{n} \sum_{i=1}^{n} h(X_i)$$

cuando $n \to \infty$ volviendo al caso iid a través de la renovación cuando $\{X_n\}$ tiene un átomo. Primero se considerará el concepto de función harmónica, que se relaciona con la ergodicidad para las cadenas de Markov Harris-recurrentes.

Definición A.6.1. *(Función harmónica)*

Se dice que una función h es harmónica para la cadena $\{X_n\}$ si

$$\mathbb{E}[h(X_{n+1})|x_n] = h(x_n)$$

Observación A.6.1.

Por supuesto que la función h en esta definición debe ser medible, para que la operación de esperanza condicional haga sentido. ▽

Estas funciones son invariantes para el kernel de transición (en el sentido funcional no en el probabilístico) y caracterizan la Harris-recurrencia como sigue.

Proposición A.6.1.

Para una cadena de Markov positiva, si las únicas funciones acotadas harmónicas son las funciones constantes, entonces la cadena es Harris-recurrente.

Demostración:

La probabilidad de un número infinito de regresos $Q(x, A) = \mathbb{P}_x(\eta_A = \infty)$, como una función de x, $h(x)$, es una función harmónica ya que

$$\mathbb{E}_y[h(X_1)] = \mathbb{E}_y[P_{X_1}(\eta_A = \infty)] = \mathbb{P}_y(\eta_A = \infty)$$

por lo tanto $Q(x, A)$ es constante en x.

La función $Q(x, A)$ describe un evento cola, un evento cuya ocurrencia no depende de $X_1, \ldots, X_m$ para cualquier m finito. Dichos eventos generalmente satisfen una ley 0-1, i.e. sus probabilidad de ocurrencia son 0 ó 1. Sin embargo, la mayoría de las veces estas leyes se establecen en el caso independiente y su estudio para cadenas de Markov está fuera de los alcances de este trabajo, así que sólo se dirá que $Q(x, A)$ cumple una ley 0-1.

Si π es una medida invariante y $\pi(A) > 0$, entonces no es posible que $Q(x, A) = 0$. Para ver esto, se hará por contradicción. Supóngase que $Q(x, A) = 0$, entonces la cadena vista á A sólo un número finito de veces y el promedio

$$\frac{1}{N} \sum_{i=1}^{N} I_A(X_i)$$

no converge a $\pi(A)$ (contradiciendo a la Ley Fuerte de Grandes Números). Entonces, para cualquier x, $Q(x, A) = 1$. Por tanto, la cadena es Harris-recurrente. □

Esta proposición se puede interpretar como una propiedad de continuidad de la transición funcional $Kh(x) = \mathbb{E}_x[h(X_1)]$ de la siguienye forma: Por inducción, una función harmónica h satisface $h(x) = \mathbb{E}_x[h(X_n)]$ y por el Teorema (A.5.4), $h(x)$ es casi seguramente igual a $\mathbb{E}^\pi[h(X)]$, i.e. es constante casi en todos lados. Para cadenas Harris-recurrentes, la prposición anterior establece que $h(x)$ es constante en TODOS lados.

La proposición anterior es muy útil cuando se quiere determinar la Harris-recurrencia de algunos métodos de cadenas de Markov Monte-Carlo. También, el comportamiento de las funciones harmónicas acotadas caracteriza la Harris-recurrencia.

Lema A.6.1.

Para cadenas de Markov Harris-recurrentes, las constantes son las únicas funciones harmónicas acotadas.

Una consecuencia de este lema es que si $\{X_n\}$ es Harris-positiva con distribución estacionaria π y si $S_n(h)$ converge ν_0-casi seguramente a

$$\int_{\mathcal{X}} h(x)\pi(dx)$$

para una distribución inicial ν_0, esta convergencia ocurrirá para toda distribución inicial ν. De hecho, la convergencia

$$\mathbb{P}_x(S_N(h)) \to \mathbb{E}^\pi(h)$$

es harmónica. De nuevo, esto muestra que la Harris-recurrencia es un tipo superior de estabilidad en el sentido de que la convergencia casi segura se reemplaza por convergencia puntual.

El resultado principal de esta sección se conoce como *Ley Fuerte de Grandes Números para Cadenas de Markov* ó bien *Teorema Ergódico*, el cual garantiza la convergencia de $S_n(h)$.

Teorema A.6.1. *(Teorema ergódico)*

Si $\{X_n\}$ tiene una medida invariante σ-finita, π, entonces las siguientes afirmaciones son equivalentes:

(i) Si $f, g \in L^1(\pi)$ tal que $\int g(x)d\pi(x) \neq 0$, entonces

$$\lim_{n\to\infty} \frac{S_n(f)}{S_n(g)} = \frac{\int f(x)d\pi(x)}{\int g(x)d\pi(x)}$$

(ii) La cadena de Markov $\{X_n\}$ es Harris-recurrente

Demostración:

Si (i) se cumple, tómese a f como la función indicadora en un conjunto A con medida finita y g una función arbitraria con integral finita y positiva. Si $\pi(A) > 0$

$$\mathbb{P}_x(X \in A \text{ infinitas veces}) = 1$$

para cada $x \in \mathcal{X}$, que establece la Harris-recurrencia.

Si (ii) se cumple, sólo se necesita considerar al caso atómico mediante un argumento de separación. Sea γ un átomo y $\tau_\gamma(k)$ el tiempo de $(k+1)$-ésima visita a γ. Si l_N es el número de visitas a γ al tiempo N, se tiene las siguientes cotas

$$\begin{aligned}\sum_{j=0}^{l_N-1}\sum_{n=\tau_\gamma(j)+1}^{\tau_\gamma(j+1)} f(x_n) &\leq \sum_{k=1}^{N} f(x_k) \\ &\leq \sum_{j=0}^{l_N}\sum_{n=\tau_\gamma(j)+1}^{\tau_\gamma(j+1)} f(x_n) + \sum_{k=1}^{\tau_\gamma(0)} f(x_k)\end{aligned}$$

Los bloques

$$S_j(f) = \sum_{n=\tau_\gamma(j)+1}^{\tau_\gamma(j+1)} f(x_n)$$

son independientes e idénticamente distribuidos. Por tanto,

$$\frac{\sum_{i=1}^{n} f(x_i)}{\sum_{i=1}^{n} g(x_i)} \leq \frac{l_N}{l_N-1}\frac{[\sum_{j=0}^{l_n} S_j(f) + \sum_{k=1}^{\tau_\gamma} f(x_k)]/l_N}{\sum_{j=0}^{l_N-1} S_j(g)/(l_N-1)}$$

por tanto, el teorema se cumple aplicando la Ley Fuerte de Grandes Números (para variables aleatotias independientes e idénticamente distribuidas). □

Observación A.6.2.

Nótese que en este Teorema, π no necesita ser una medida de probabilidad, por tanto se puede tener algún tipo de estabilidad fuerte incluso si la cadena es recurrente nula. ▽

Cuando se trabaja con algoritmos de métodos de cadenas de Markov Monte-Carlo, generalmente este resultado se ocupa cuando se considera un análisis Bayesiano con medidas posteriores impropias.

A.6.2. Teoremas del límite central

Existe un camino general de las Leyes de los Grandes Números al Teorema del Límite Central. Además, la demostración del Teorema ergódico sugiere que es una extensión directa del Teorema del Límite Central para variables aleatorias independientes e idénticamente distribuidas. Desafortunadamente, este no es el caso debido a las condiciones de varianza finita que explícitamente involucran al átomo de la cadena que separa. Por esto, se estudiarán condiciones alternativas para que se pueda aplicar el Teorema del Límite Central.

Proposición A.6.2.

Si $\{X_n\}$ es Harris-positiva con un átomo γ tal que

(i) $\mathbb{E}_\gamma(\tau_\gamma^2) < \infty$

(ii) $\mathbb{E}_\gamma\left[\left(\sum_{n=1}^{\tau_\gamma} |h(X_n)|\right)^2\right] < \infty$

(iii) $\varrho_h^2 = \pi(\gamma)\mathbb{E}_\gamma\left[\left(\sum_{n=1}^{\tau_\gamma}(h(X_n) - \mathbb{E}^\pi(h))\right)^2\right] > 0$

Entonces el Teorema Central del Límite se cumple, es decir

$$\frac{1}{\sqrt{N}}\left(\sum_{n=1}^{N}(h(X_n) - \mathbb{E}^\pi(h))\right) \xrightarrow{L} Z$$

donde $Z \sim N(0, \varrho_h^2)$

Demostración:

Usando la misma notación que en la demostración del Teorema ergódico, si $\bar{h} := h - \mathbb{E}^\pi(h)$ se tiene que

$$\frac{1}{\sqrt{l_N}}\sum_{i=1}^{l_N} S_i(\bar{h}) \xrightarrow{L} N\left(0, \mathbb{E}_\gamma^2\left[\sum_{n=1}^{\tau_\gamma}\bar{h}(X_n)\right]\right)$$

según el Teorema del Límite Central para las variables independientes, $S_i(\bar{f})$ cuando N/l_N convergen a $\mathbb{E}_\gamma(S_0(1)) = [\pi(\gamma)]^{-1}$ casi seguramente. Ya que

$$\left|\sum_{i=1}^{l_N-1} S_i(\bar{h}) - \sum_{k=1}^{N}\bar{h}(X_k)\right| \leq S_{l_N}(|\bar{h}|)$$

y

$$\lim_{n\to\infty}\frac{1}{n}\sum_{j=1}^{n} S_i^2(|\bar{h}|) = \mathbb{E}_\gamma[S_0^2(|\bar{h}|)]$$

entonces

$$\limsup_{N\to\infty}\frac{S_{l_N}(|\bar{h}|)}{\sqrt{N}} = 0$$

□

Este resultado establece que una extensión del Teorema del Límite Central al caso no atómico puede ser más delicado con sólo las suposiciones del Teorema ergódico. De hecho, se tiene que extender estas suposiciones a la cadena que separa, $\{\tilde{X}_n\}$.

Un caso sencillo es cuando la cadena es reversible.

Definición A.6.2. *(Cadena reversible)*

Se dice que una cadena de Markov estacionaria, $\{X_n\}$, es reversible si la distribución de $X_{n+1}|X_{n+2} = x$ es la misma que la de $X_{n+1}|X_n = x$

Por tanto, para una cadena reversible, la dirección de tiempo no importa en la dinámica de la cadena. Con las suposición de reversbilidad, el Teorema del Límite Central se cumple debido a la positividad estricta de ϱ_g. Esto lo establecieron Kipnis & Varadhan (1986).

Teorema A.6.2.

Si $\{X_n\}$ es aperiódica, irreducible, reversible con distribución invariante π, entonces el Teorema Central del Límite se aplica cuando

$$0 < \varrho_g^2 = \mathbb{E}_\pi(\bar{g}^2(X_0)) + 2\sum_{k=1}^{\infty} \mathbb{E}_\pi(\bar{g}(X_0)\bar{g}(X_k)) < \infty$$

Es importante notar que, en general, la reversibilidad es una suposición muy restrictiva; aunque en algunos métodos de cadenas de Markov Monte-Carlo se puede inducir esta reversibilidad agregando pasos adicionales de simulación.

Apéndice B

Un centavo de autodescomponibilidad

B.1. Introducción

Definición B.1.1. *(Autodescomponibilidad)*

Se dice que una distribución de probabilidad en $\mathbb{R}$ se autodescompone si su función característica, φ, satisface

$$\varphi(t) = \varphi(\varrho t)\varphi_\varrho(t), \quad t \in \mathbb{R}, \varrho \in (0,1)$$

donde φ_ϱ es una función característica. Lo cual implica para la correspondiente a variable aleatoria que

$$X = \varrho Y + X_\varrho, \quad \varrho \in (0,1)$$

donde Y y X_ϱ son independientes y X y Y tienen idéntica distribución.

Observación B.1.1.

Como en la definición $\varrho \in (0,1)$, ninguna variable aleatorias con soporte contenido en $\mathbb{Z}$ (excepto la variable aleatoria $X = 0$) puede ser autodescomponible. De hecho, se sabe que todas las distribuciones autodescomponibles son absolutamente continuas. ∇

En este anexo se estudiarán algunos análogos de autodescomponibilidad y estabilidad para distribuciones sobre $\mathbb{N}$.

Las distribuciones que se autodescomponen discretamente y las distribuciones discretas estables comparten algunas propiedades básicas con sus contrapartes continuas. Por ejemplo, las distribuciones que se autodescomponen discretamente son unimodales y las

distribuciones discretas estables son muy "parecidas" a sus análogas continuas en $(0, \infty)$.

A continuación se enunciará un lema que se necesitará para definir correctamente estos conceptos análogos discretos.

Lema B.1.1.

(i) Si G es una función generadora de probabilidad entonces

$$lim_{x\uparrow 1}(1-x)G'(x) = 0$$

(ii) Sea G una función generadora de probabilidad de la distribución $\{p_i\}_{i=0}^{\infty}$ tal que $p_0 \in (0,1)$. G es infinitamente divisible si y sólo si tiene la forma

$$G(z) = \exp\{\lambda\xi(z) - 1\}$$

donde $\lambda > 0$ y ξ es una función generadora de probabilidad (única) tal que $\xi(0) = 0$. Equivalentemente, G es infinitamente divisible si y sólo si

$$G(z) = \exp\{-\int_z^1 R(u)du\},$$

donde $R(u) = \sum_{n=0}^{\infty} r_n u^n$ con $r_n \geq 0$ y $\sum_{k=0}^{n} \frac{r_n}{n+1} < \infty$ i.e. si y sólo si $\{p_i\}_{i=0}^{\infty}$ satisface

$$(n+1)p_{n+1} = \sum_{k=0}^{n} p_k r_{n-k}, \quad n \in \mathbb{N}$$

con $r_0 \geq 0$.

Demostración:

Por el Teorema del valor medio, para $x \in [0,1)$ se tiene que $1 - G(x) = (1-x)G'(y)$ para algún $y \in (x,1)$. Como G' es no decreciente, entonces

$$(1-x)G'(x) \leq (1-x)G'(y) = 1 - G(x) \xrightarrow{x\uparrow 1} 0$$

□

B.2. Autodescomponibilidad discreta

Definición B.2.1. *(Autodescomponibilidad discreta)*

Se dice que una distribución sobre $\mathbb{N}$ se autodescompone discretamente si su funcíón generadora de probabilidad, G, satisface

$$G(z) = G(1 - \varrho - \varrho z)G_{\varrho}(z), \quad |z| < 1, \varrho \in (0, 1)$$

donde G_{ϱ} es una función generadora de probabilidad.

La expresión anterior se puede escribir en términos de variables aleatorias como

$$X = \varrho \circ Y + X_{\varrho}$$

donde $\varrho \circ Y$ y X_{ϱ} son independientes y X y Y son idénticamente distribuidas; y $\varrho \circ Y$ se define (en distribución) mediante su función generadora de probabilidad $G(1 - \varrho + \varrho z)$ ó como

$$\varrho \circ Y = \sum_{i=1}^{Y} W_i$$

donde $\{W_i\}$ es una sucesión de variables aleatorias independientes e idénticamente distribuidas $W_i \sim Bnlli(\varrho)$, independiente de Y.

Nótese que $\varrho \circ Y \in$ toma valores en $\mathbb{N}$, $1 \circ Y = Y$, $0 \circ Y = 0$ y $\mathbb{E}(\varrho \circ Y) = \mathbb{E}[\mathbb{E}(\varrho \circ Y)|Y] = \mathbb{E}(\varrho Y) = \varrho \mathbb{E}(Y)$ pues $\varrho \circ Y | Y = y \sim Bin(y, \varrho)$.

A continuación se establecerá la forma canónica de las funciones generadoras de probabilidad que se autodescomponen discretamente.

Teorema B.2.1.

Una función generadora de probabilidad se autodescompone discretamente si y sólo si tiene la forma

$$G(z) = \exp\left\{-\lambda \int_z^1 \frac{1 - \xi(u)}{1 - u} du\right\},$$

donde $\lambda > 0$ y ξ es una función generadora de probabilidad (única) tal que $\xi(0) = 0$. O equivalentemente, G se autodescompone discretamente si y sólo si es infinitamente divisible y tiene una medida canónica r_n (como en el Lema anterior) que es no creciente.

Demostración:

($\Rightarrow$)

Supóngase que G se autodescompone discretamente. Entonces, para $r > 0$ y $r(1-\varrho_n)^{-1} \in \mathbb{N}^+$ defínase

$$Q_{r,n}(z) := (G_{\varrho_n}(z))^{r/(1-\varrho_n}$$

Nótese qu $Q_{r,n}$ es una función generadora de probabilidad. Como

$$G(1-\varrho+\varrho z) = G(z) + (1-\varrho)(1-z)G'(z) + o(1-\varrho), \quad \text{cuando } \varrho \uparrow 1$$

Entonces para ϱ_n tal que $\varrho_n \uparrow 1$ cuando $n \to \infty$ se tiene que

$$Q_r(z) := \lim_{n\to\infty} Q_{r,n}(z) = exp\{-r(1-z)G'(z)/G(z)\}$$

Como $Q_r(z) \to 1$ cuando $z \uparrow 1$ y por el Teorema de continuidad para las funciones generadoras de probabilidad, para cualquier $r > 0$, Q_r es una función generadora de probabilidad. Por tanto, $Q := Q_1$ es infinitamente divisible y aplicando el lema anterior a Q se tiene que

$$R(z) := \frac{G'(z)}{G(z)} = -\frac{Log(Q(z))}{1-z} = \lambda\frac{1-\xi(z)}{1-z}$$

Concluyendo que

$$G(z) = \exp\left\{-\lambda\int_z^1 \frac{1-\xi(u)}{1-u}du\right\}$$

Comparando $G(z) = \exp\{-\int_z^1 R(u)du\}$ y $G(z) = \exp\left\{-\lambda\int_z^1 \frac{1-\xi(u)}{1-u}du\right\}$ se puede ver que G es infinitamente divisible con

$$r_n = \lambda\left(1 - \sum_{j=1}^{n} g_j\right) = \lambda\sum_{j=n+1}^{\infty} g_j$$

que es no creciente.

($\Leftarrow$)

Supóngase que G es de la forma $G(z) = \exp\left\{-\lambda\int_z^1 \frac{1-\xi(u)}{1-u}du\right\}$. Entonces G satisface $G(z) = G(1-\varrho-\varrho z)G_\varrho(z)$ con

$$G_\varrho(z) = exp\{-\int_z^{1-\varrho(1-z)} R(u)du\}$$

i.e. con

$$R_\varrho(z) := \frac{G'_\varrho(z)}{G_\varrho(z)} = R(z) - \varrho R(1-\varrho(1-z))$$

con coeficientes

$$r_n - \varrho \sum_{k=n}^{\infty} \binom{k}{n} \varrho^n (1-\varrho)^{k-n} r_k \geq r_n [1 - \varrho^{n+1} \sum_{j=0}^{\infty} \binom{n+j}{j} (1-\varrho)^j] = 0$$

donde se está usando el hecho e que $\{r_n\}$ es no creciente. Por tanto G_ϱ es infinitamente divisible. □

La unimodalidad de una distribución que se autodescompone discretamente es un corolario del siguiente teorema.

Teorema B.2.2.

Sean $\{p_n\}_{n=0}^{\infty}, \{r_n\}_{n=0}^{\infty} \subseteq \mathbb{R}$ sucesiones tales que para todo $n \in \mathbb{N}^+$ $p_n \geq 0$, $p_0 > 0$, $\{r_n\}_{n=0}^{\infty}$ es no-creciente y además

$$(n+1)p_{n+1} = \sum_{k=0}^{n} p_k r_{n-k}, \quad n \in \mathbb{N}$$

Entonces, $\{p_n\}_{n=0}^{\infty}$ es unimodal, i.e. $p_n - p_{n-1}$ cambia de signo a los más una vez (donde $p_{-1} := 0$). Y además, $\{p_n\}_{n=0}^{\infty}$ es no-creciente si y sólo si $r_0 \leq 1$.

Demostración:

Sean $d_n := p_n - p_{n-1}$ y $\lambda_n := r_n - r_{n-1}$

$$(n+1)d_{n+1} = (r_0 - 1)p_n - \sum_{j=0}^{n-1} \lambda_j p_{n-j-1}, \quad n \in \mathbb{N}$$

Notese que para $n \in \mathbb{N}^+$ $d_n \leq 0 \Leftrightarrow r_0 \leq 1$. Sea $r_0 > 1$ y supóngase que

$$d_1 > 0, d_2 \geq 0, \ldots, d_{n_1} \geq 0, d_{n_1+1} < 0, \ldots, d_{n_1+m} =: d_{n_2} \leq 0, d_{n_2+1} > 0$$

Si para $j > n$ se hace $p_{n-j} = 0$,

$$p_{n_1-j} \leq p_{n_2-j}, \quad j = m+1, m+2, \ldots$$

$$p_{n_1-j} \leq p_{n_1}, \quad j = 1, 2, \ldots, m.$$

Y por tanto,

$$(n_1+1)d_{n_1+1} = (r_0-1)p_{n_1} - \sum_{j=0}^{n_1-1} \lambda_j p_{n_1-j-1} < 0$$

$$(n_2+1)d_{n_2+1} = (r_0-1)p_{n_2} - \sum_{j=0}^{n_2-1} \lambda_j p_{n_2-j-1} > 0$$

pero como $\sum_{j=0}^{m-1} \lambda_j p_{n_2} \leq \sum_{j=0}^{m-1} \lambda_j p_{n_2-j-1}$ entonces se tiene que $r_0 = r_n + \sum_{j=0}^{n-1} \lambda_j$ y

$$(r_m - 1)p_{n_2} > \sum_{j=1}^{n_1-1} \lambda_j p_{n_2-j-1}$$

Es decir,

$$\begin{aligned} \sum_{j=0}^{n_1-1} \lambda_j p_{n_1-j-1} &\leq \sum_{j=0}^{m-1} \lambda_j p_{n_1} + \sum_{j=m}^{n_1-1} \lambda_j p_{n_2-j-1} \\ &< p_{n_1}(r_0 - r_m) + p_{n_2}(r_m - 1) < p_n(r_0 - 1 \end{aligned}$$

lo cual es una contradicción. □

Corolario B.2.1.

Sea $\{p_n\}_{n=0}^{\infty}$ una distribución discreta autodescomponible discreta. Entonces,

(i) $\{p_n\}_{n=0}^{\infty}$ es unimodal

(ii) $\{p_n\}_{n=0}^{\infty}$ es no-creciente si y sólo si $r_0 = \frac{p_1}{p_0} \leq 1$

Equivalentemente, una distribución sobre $\mathbb{N}$ infinitamente divisible (con $p_0 > 0$) es unimodal si $\{r_n\}_{n=0}^{\infty}$ es no-creciente, es no-creciente si y sólo si $r_0 \leq 1$

Observación B.2.1.

En el teorema anterior no se está suponiendo que $r_0, r_1, \ldots$ son no-negativos, es decir, esta es una condición suficiente para la unimodalidad más general que en las distribuciones infinitamente divisibles. Sin embargo, si $\{p_n\}_{n=0}^{\infty}$ es no negativa y no-creciente entonces $\{r_n\}_{n=1}^{\infty}$ es no negativa. ▽

B.3. Estabilidad discreta

El conjunto de distribuciones sobre $\mathbb{R}$ que son estables (estrictamente) con exponente γ es el subconjunto de las distribuciones autodescomponibles que satisface

$$(s+t)^{1/\gamma} \stackrel{d}{=} s^{1/\gamma}X_1 + t^{1/\gamma}X_2, \quad s,t>0$$

donde X_1 y X_2 son independientes y con la misma distribución que X. Se puede reescribir la ecuación anterior como

$$X = \alpha X_1 + (1-\varrho^\gamma)^{1/\gamma}X_2, \quad \varrho \in (0,1)$$

Reemplazando ϱX_1 por $\varrho \circ X_1$ y $(1-\varrho^\gamma)^{1/\gamma}X_2$ por $(1-\varrho^\gamma)^{1/\gamma}\circ X_2$ se obtiene la definición de estabilidad para el caso discreto. En términos de la función generadora de probabilidad se tiene que

$$G(z) = G(1-\varrho(1-z))G(1-(1-\varrho^\gamma)^{1/\gamma}(1-z)), \quad |z|<1, \varrho\in(0,1)$$

Definición B.3.1.

Se dice que una función generadora de probabilidad G tal que $G(0)\in(0,1)$ es estable discreta (estrictamente) con coeficiente $\gamma>0$ si satisface

$$G(z) = G(1-\varrho(1-z))G(1-(1-\varrho^\gamma)^{1/\gamma}(1-z)), \quad |z|<1, \varrho\in(0,1)$$

A partir de la definición se puede ver que

$$\frac{1-G(1-(1-\varrho^\gamma)^{1/\gamma}(1-z))}{(1-\varrho)(1-z)} = \frac{G(1-\varrho(1-z))-G(z)}{(1-\varrho)(1-z)G(1-\varrho(1-z))} \stackrel{z\uparrow 1}{\longrightarrow} \frac{G'(z)}{G(z)}$$

Haciendo $u=(1-\varrho^\gamma)^{1/\gamma}$, lo anterior significa que

$$\frac{1-G(1-u(1-z))}{(u(1-z))^\gamma} \stackrel{u\downarrow 0}{\longrightarrow} \frac{(1-z)^{1-\gamma}}{\gamma}\frac{G'(z)}{G(z)}$$

y con $z=0$,

$$\frac{1-G(1-u)}{u^\gamma} \stackrel{u\downarrow 0}{\longrightarrow} \frac{p_1}{\gamma p_0}$$

Y por tanto según las últimas dos expresiones

$$\frac{G'(z)}{G(z)} = \frac{p_1}{p_0}(1-z)^{\gamma-1} \tag{B.1}$$

Como $G'(1) > 0$ (posiblemente infinito), de la expresión anterior se tiene que $\gamma \in (0, 1]$. Integrando (B.1) se tiene que

$$G(z) = G_\gamma^\lambda(z) := exp\{-\lambda(1-z)^\gamma\}, \quad |z| \leq 1, \lambda > 0$$

Entonces G definida de esta manera satisface que

$$G(z) = G(1-\varrho(1-z))G(1-(1-\varrho^\gamma)^{1/\gamma}(1-z)), \quad |z| < 1, \varrho \in (0,1)$$

□

Teorema B.3.1.

Las funciones generadoras de probabilidad discretas estables existen sólo para $\gamma \in (0, 1]$ y todas las funciones generadoras de probabilidad con exponente γ están dadas por

$$G(z) = G_\gamma^\lambda(z) := \exp\{-\lambda(1-z)^\gamma\}, \quad |z| < 1, \lambda > 0$$

Observación B.3.1.

Las funciones generadoras de probabilidad estables discretas son muy parecidas a las transformadas de Laplace $e^{-\lambda\tau^\gamma}$ de las distribuciones estables sobre $(0, \infty)$. ▽

Corolario B.3.1.

La distribución Poisson es estable discreta con exponente uno.

Corolario B.3.2.

Una distribución estable discreta se autodescompone discretamente y por tanto es unimodal.

Observación B.3.2.

Si se define una función generadora de probabilidad en el dominio de atracción (discreta) de una función generadora de probabilidad G_γ, existe ϱ_n tal que

$$\lim_{n\to\infty}[G(1-\varrho_n+\varrho_n z)]^n = G_\gamma(z)$$

Entonces, todas las distribuciones con primer momento finito son atraídas por la distribución Poisson tomando $\varrho_n = 1/n$. Se puede desarrollar una teoría general de atracción. Sin embargo, como para $\gamma \in (0,1)$ se tiene $G_\gamma(1-y) = exp(-y^\gamma)$ entonces

$$G^n(1-\varrho_n y) = \mathbb{E}^n[exp\{Log(1-\varrho_n y)\}] \overset{n\to\infty}{\sim} \mathbb{E}^n[exp(-\varrho_n y X)]$$

Una variablea aleatoria X con soporte en $\mathbb{N}$ está en el dominio de atracción de G_γ^λ si y sólo si está en el dominio de atracción de $exp(-\lambda y^\gamma)$. ▽

Ahora, se considerará la ecuación

$$G(z) = G(1-\varrho+\varrho z)G_\varrho(z), \quad |z|<1, \varrho\in(0,1)$$

primero considerando (un análogo más formal de autodescomponibilidad)

$$G(z) = \frac{G(\varrho z)}{G(\varrho)}G_\varrho(z), \quad |z|<1, \varrho\in(0,1)$$

Se pueden tratar estas dos últimas ecuaciones como equivalente según el siguiente teorema.

Teorema B.3.2.

Una función generadora de probabilidad, G tal que $G(0) > 0$ satisface

$$G(z) = \frac{G(\varrho z)}{G(\varrho)}G_\varrho(z), \quad |z|<1, \varrho\in(0,1)$$

si y sólo si es infinitamente divisible si y sólo si sigue una distribución Poisson compuesta.

Definiendo $\varrho * X$ (en distribución) mediante su función generadora de probabilidad $1-\varrho+\varrho G(z)$ ó por

$$\varrho * X = \sum_{j=1}^{N} X_j$$

con N como antes, se puede considerar la ecuación $X = \varrho * Y + X_\varrho$ ó en términos de su función generadora de probabilidad

$$G(z) = (1-\varrho+\varrho G(z))G_\varrho(z), \quad |z|<1, \varrho\in(0,1)$$

para obtener el siguiente resultado.

Teorema B.3.3.

Una función generadora de probabilidad, G tal que $G(0) > 0$ satisface

$$G(z) = [1 - \varrho + \varrho G(z)]G_{\varrho}(z), \quad |z| < 1, \varrho \in (0, 1)$$

si y sólo si es un geométrico compuesto.

Bibliografía

Bibliografía

Adke, S. & Deshmukh, S. (1988). *Limit distribution of a high order Markov chain.* J. R. Statist. Soc. B 50, 105-108.

Albert, P. (1991). *A two-state Markov mixture model for a time series of epileptic seizure counts.* Biometrics 47, 1371-1381.

Al-Osh, M. & Aly, E. (1992). *First order autoregressive time serues with negative binomial and geometric marginals.* Communications in Statistics: Theory and Methods, 21, 2483-2492.

Al-Osh, M. & Alzaid, A. (1987). *First-order integer-valued autoregressive ($INAR(1)$) process.* Journal of Time Series Analysis 8, 261-275.

Al-Osh, M. & Alzaid, A. (1988). *Integer-valued moving average (INMA) process.* Statistical Papers 29, 281-300.

Al-Osh, M. & Alzaid, A. (1991). *Binomial autoregressive moving average models.* Stochastic Models, 7(2): 261-282.

Aly, E. & Bouzar, N. (1994). *Explicit stationary distributions for some Galton-Watson processes with immigration.* Comm. Statist. Stochastic Models 10, no. 2, 499-517.

Aly, E. & Bouzar, N. (2000). *On geometric infinite divisibility and stability.* Ann. Inst. Statist. Math. 52, no. 4, 790-799.

Alzaid, A. & Al-Osh, M. (1988). *First-order integer-valued autoregressive ($INAR(1)$) process: distributional and regression properties.* Statist. Neerl. 42, 53-61.

Alzaid, A. & Al-Osh, M. (1990). *An integer-valued pth-order autoregressive structure ($INAR(p)$) process.* Journal of Applied Probability 27, 314-323.

Alzaid, A & Al-Osh, M. (1993). *Some Autoregressive Moving Average Processes with Generalized Poisson Marginal Distributions.* Ann. Inst. Statist. Math., 45(20): 223-232.

Athreya, K. & Ney, P. (1978). *A new approach to the limit theory of recurrent Markov chains.* Trans. Amer. Math. Soc. 245, 493-501.

Azzalini, A. (1983). *Maximum likelihood estimation of order "m" for stationary stochastic processes.* Biometrika, vol. 70; p. 381-387.

Azzalini, A. (1994). *Logistic regression and other discrete data models for serially correlated observations.* Journal of the Italian Statistical Soc, Vol. 3; p. 169-179.

Azzalini, A. (1994). *Logistic regression for autocorrelated data with application to repeated measures.* Biometrika, vol. 81; p. 767-775.

Azzalini, A. & Bowman A. (1990). *A look at some data on the Old Faithful Geyser.* J. of the Royal Statistical Soc. Series C-Appl. Stat., Vol. 39; p. 357-365.

Bellman, R. (1974). *Introduction to Matrix Analyisis.* 2nd. Edition. Bombay: Tata McGraw-Hill.

Besag, G. (1974). *Spatial interaction and the statistical analysis of lattice systems (with discussion).* J. Royal Statist. Soc. Series B, 36, 192-326.

Besag, J., Green, E., Higdon, D. & Mengersen, K.L. (1995). *Bayesian computation and stochastic systems (with discussion).* Statistical Science 10, 3-66.

Bisgaard, S. & Travis, L. (1991). *Existence and uniqueness of the solution of the likelihood equations for binary Markov chains.* Statist. Prob. Letters 12, 29-35

Box, G. & Jenkins, G. (1970). *Time Series Analysis: Forecasting and Control.* Holden-Day. San Francisco.

Brännäs, K. (1994). *Estimation and testing in integer-valued AR(1) models.* University of Umea Discussion Paper No. 335.

Brännäs. K. (1995a). *Prediction and control of a time series count data model.* International Journal of Forecasting 11. 263-270.

Brännäs. K. (1995b). *Explanatory variables in the AR*(1) *count data model.* Umea Economic Studies No. 381.

Brännäs, K. & Hall, A. (2001). *Estimation in integer-valued moving average models.* Appl. Stochastic Models Bus. Ind., 17: 277-291.

Brännäs, K. & Johansson, P. (1994). *Time Series Count Data Regression.* Commun. Statist. Theory Meth. 23, 2907-2925.

Brännäs, K. & Johansson, P. (1996). *Panel data for regression of counts.* Statist. Papers 37, 191-213.

Brooks, S., Giudici, P. & Roberts, G. (2003). *Efficient construction of reversible jump Markov Chain Monte Carlo proposal distributions.* J. R. Statist. Soc. (B) 65, 3-55.

Cameron, A. & Trivedi, P. (1998). *Regression Analysis of Count Data.* Cambridge University Press, Cambridge.

Campbell, M. (1994). *Time series regression for counts: an investigation of the relationship between suden infant death syndrome and enviromental temperature.* J. R. Statist. Soc. (A) 157, 191-208.

Cargnoni, C., Müller, P. & West, M. (1997). *Bayesian Forecasting of Multinomial Time Series through Conditionally Gaussian Dynamic Models.* J. Amer. Statist. Ass. 92, 640-647.

Carlin, B., Polson, N. & Stoffer, D. (1992). *A Monte Carlo Approach to Nonnormal and Nonlinear State Space Modelling.* J. Amer. Statist. Ass. 87, 493-500.

Carter, C. & Kohn, R. (1994). *On Gibbs Sampling for State Space Models.* Biometrika 81, 541-553.

Casella, G. (1996). *Statistical thoery and Monte Carlo algorithms (with discussion).* TEST 5, 249-344.

Casella, G. & George, E. (1992). *Explaining the Gibbs Sampler.* The American Statistician, 46, 167-174.

Chan, K. & Geyer, C. (1994). *Discussion of "Markov chains for exploring posterior distribution".* Ann. Statist. 22, 1747-1758.

Chan, K. & Ledolter, J. (1995). *Monte Carlo EM Estimation for Time Series Models involving Counts.* J. Amer. Statist. Ass. 90, 242-252.

Chib, S. & Greenberg, E. (1995). *Understanding the Metropolis-Hastings Algorithm.* The American Statistician, 49, 327-335.

Cox, D. (1970). *The Anaylsis of Binary Data.* Methuen, London.

Cox, D. R. (1981). *Statistical analysis of time series: Some recent developments.* Scandinavian Journal of Statistics 8, 93-115.

Damien, P., Wakefield, J. & Walker, S. (1999). *Gibbs sampling for Bayesian non-conjugate and hierarchical models by using auxiliary variables.* J. Royal Statist. Soc. Series B, 61(2):331-344.

Dempster, A., Laird, N. & Rubin, D. (1977). *Maximum Likelihood from Incomplete Data via the EM Algorithm.* Journal of the Royal Statistical Society. Series B (Methodological), Vol. 39, No. 1., pp. 1-38.

Diebolt, J. & Robert, C. (1993). *Discussion of "Bayesian computations via the Gibbs sampler" by A.F.M. Smith and G.O. Roberts.* J. Royal Statist. Soc. Series B 55, 71-72.

Diebolt, J. & Robert, C. (1994). *Estimation of finite mixture distributions by Bayesian sampling.* J. Royal Statist. Soc. Series B 56, 363-375.

Diggle, P., Liang, K-Y. & Zeger, S. (1994). *Anaylsis of Longitudinal Data.* Oxford University Press, Oxford.

Dion, J., Gauthier, G. & Latour, A. (1995). *Branching Processes with Immigration and Integer-valued Time Series.* SERDICA 21, 123-136.

Du, J. & Li, Y. (1991). *The integer-valued autoregressive ($INAR(p)$) model.* Journal of Time Series Analysis 12, 129-142.

Durbin, J. & Koopman, S. (2000). *Time series analysis of non-Gaussian observations based on state space models from both classical and Bayesian perspectives (with discussion).* J. R. Statist. Soc. (B) 62, 3-56.

Edwards, C. & Gurland, J. (1961). *A class of distributions applicable to accidents.* J. American Statist. Assoc., 56, 503-517.

Enciso-Mora, V., Neal, P. & Subba Rao, T. (2009). *Efficient Order Selection Algorithms for Integer-Valued ARMA Processes.* Journal of Time Series Analysis, 30(1): 1-18.

Fahrmeir, L. (1992). *Posterior Mode Estimation by Extended Kalman Filtering for Multivariate Dynamic Generalized Linear Models.* J. Amer. Statist. Ass. 87, 501-509.

Fahrmeir, L. & Kaufmann,H. (1987). *Regression Models for Non-Stationary Categorical Time Series.* Journal of Time Series Anal. 8, 147-160.

Feller, W. (1971). *An Introduction to Probability Theory and Its Applications. Vol. II.*, John Wiley & Sons, New York.

Foster, J. & Williamson, J. (1971). *Limit Theorems for the Galton-Watson process with time-dependent immigration.* Z. Wahrsch. Verw. Gebiete 20, 227-235.

Freeland, K. (1998). *Statistical Analysis of discrete time series with application to the analysis of worker's compensation claims data.* Ph. D. Thesis, University of British Columbia.

Freeland, R. & McCabe, B. (2004). *Analysis of low count time series data by Poisson autoregression.* Journal of Time Series Analysis, 25(5): 701-722.

Freeland, R. & McCabe, B. (2004). *Forecasting discrete valued low count time series.* International Journal of Forecasting, 20: 427-434.

Freeland, R. & McCabe, B. (2005). *Asymptotic properties of CLS estimators in the Poisson AR(1) model.* Statistics & Probability Letters, 73:147-153.

Frühwirth-Schnatter, S. (1994). *Data Augmentation and Dynamic Linear Models.* J. Time Series Analysis 15, 183-202.

Fuller, W. (1976) *Introduction to Statistical Time Series.* Wiley, Nwe York.

Gauthier, G. & Latour, A. (1994) *Convergence forte des estimateurs des paramèters d'un processus GENAR(p).* Ann. Sci. Math. Québec 18, 49-71.

Gaver, D. & Lewis, P. (1980). *Firsr order autoregressive gamma sequences and point processes.* Adv. Appl. Prob. 12, 724-745.

Gelfand, A. & Smith, A. (1990). *Sampling based approaches to calculating marginal densities.* J. American Statist. Assoc., 85:398-409.

Gelman, A., Carlin, J., Stern, H., & Rubin, D. (2004). *Bayesian Data Analysis*, Second Edition, London: Chapman & Hall.

Geman, S. & Geman, D. (1984). *Stochastic relaxation, Gibbs distributions and the Bayesian restoration of images.* IEEE Trans. Pattern Anal. Mach. Intell. 6:721-742.

Geyer, C. (1992). *Practical Markov Chain Monte Carlo (with discussion).* Statistical Science 7, 473-511.

Gibbons, J. & Chakraborti, S. (1992). *Nonparametric Statistical Inference.* 3rd. Edition. New York Marcel Dekker.

Gilks, W., Best, N. & Tan, K. (1995). *Adaptive rejection Metropolis sampling within Gibbs sampling.* Applied Statist. (Ser. C), 44:455-472.

Green, P. (1995). *Reversible jump MCMC computation and Bayesian model determination.* Biometrika, 82(4):711-732.

Grunwald, G., Hyndman, R., Tedesco, L. & Tweedie, R. (2000). *Non-Gaussian conditional linear AR(1) models.* Australian and New Zealand Journal of Statistics. 42, 479-495.

Hamilton, J. D. (1994). *Time series analysis.* Princeton University Press.

Harris, D. (1999). *GMM estimation of time series models.* In Mátyás L., editor *Generalized method of moments estimation.* Cambridge University Press. 149-170.

Harvey, A. (1989). *Forecasting, structural time series models and the Kalman filter.* Cambridge University Press, Cambridge.

Harvey, A. & Fernandes, C. (1989). *Time series models for count or qualitative observations.* J. Bus. Econ. Statist. 7, 407-422.

Harvey, A. & Phillips, G. (1979). *Maximum likelihood estimation of regression models with autoregressive-moving average disturbance.* Biometrika 66, 49-58.

Hastings, W. (1970). *Monte Carlo Sampling methods using Markov chains and their application.* Biometrika, 57:97-109.

Heathcote, C. (1966) *A branching process allowing immigration.* J. R. Statist. Soc. B. 28, 213-217.

Heyde, C. & Seneta, E. (1972). *Estimation theory for growth and immigration rates in a multiplicative process.* Journal of Applied Probability 9, 235-256.

Hills, S. & Smith, A. (1992). *Parametrization issues in Bayesian inference.* In Bernardo, J., Berger, J., David, A. & Smith, A., editors, Bayesian Statistics 4, pages 641-649. Oxford University Press, Oxford.

Hills, S. & Smith, A. (1993). *Diagnostic plots for improved parametrization in Bayesian inference.* Biometrika 80, 61-74.

Hobert, J.P. & Casella, G. (1996). *The effect of improper priors on Gibbs sampling in hierarchical linear models.* J. Amer Statist. Assoc. 91, 1461-1473.

Hobert, J.P. & Casella, G. (1998). *Functional compatibility, Markov chains, and Gibbs sampling with improper posteriors.* J. Comput. Graph. Statist. 7, 42-60.

Jayakumar, K. (1995). *An integer valued autoregressive (INAR(p)) time series model.* Bulletin of the International Statistical Institute, Proceedings of the 50th Sesion Book 3, pp. 1365-1374.

Jørgensen, B., Lundbye-Christensen, S., Song, P. & Sun, L. (1999). *A state-space model for multivariate longitudinal count data.* Biometrika 86, 169-181.

Jung, R. & Tremayne A. (2003). *Testing for serial dependence in time series models of counts.* Journal of Time Series Analysis 24, 65-84.

Jung, R., Ronning, G. & Tremayne A. (2005). *Estimation in conditional first order autoregression with discrete support.* Statistical Papers 46, 195-224.

Kashiwagi, N. & Yanagimoto, T. (1992). *Smoothing serial count data through a state-space model.* Biometrics 48, 1187-1194.

Kedem, B. & Fokianos, K. (2002). *Regression models for time series analysis.* John Wiley & Sons.

Keenan, D. (1992). *A time series analysis of binary data.* J. Amer. Statist. Ass. 77, 816-821.

Kipnis, C. & Varadhan, S. (1986). *Central Limit Theorem for additive functionals of reversible Markov processes and applications to simple exclusions.* Comm. Math. Phys. 104, 1-19.

Kitagawa, G. (1987). *Non-Gaussian state space modelling for non-stationary time series (with discussion).* J. Amer. Statist. Ass. 82, 1032-1063.

Klimko, L. A. & Nelson, P.I. (1978). *On conditional least squares estimation for stochastic processes.* The Annals of Statistics 6, 629-642.

Korn, E. & Whittemore, A. (1979). *Methods for analysing panel studies of acute health effects of air pollution.* Biometrics 35, 795-902.

Latour, A. (1997). *The Multivariate $GINAR(p)$ Process.* Advances in Applied Probability, 29(1): 228-248.

Latour, A. (1998). *Existence and stochastic structure of a non-negative integer-valued autoregressive process.* Journal of Time Series Analysis, 19(4): 439-455.

Lawrence, A. (1982). *The innovation distribution of a gamma distributed autoregressive process.* Scand. J. Statist. 9, 234-236.

Lawrence, A. & Lewis, P. (1980). *The Exponential Autoregressive-Moving Average EARMA(p,q) Process.* J. R. Statist. Soc. (B) 42, 150-161.

Lawrence, A. & Lewis, P. (1981). *A new autoregressive time series model in exponential variables (NEAR(1)).* Adv. Appl. Prob. 13, 826-845.

Lawrence, A. & Lewis, P. (1985). *Modelling and residual analysis of nonlinear autoregressive time series in exponential variables. Average EARMA(p,q) Process.* J. R. Statist. Soc. (B) 47, 165-183.

Lee, Y. & Nelder, J. (1996). *Hierarchical Generalized Linear Models (with discussion).* J. R. Statist. Soc. (B) 58, 619-678.

Levine, R. (1996). *Post-processing random variables.* Technical report, Biometrics Unit, Cornell Univ. PhD. Thesis.

Lewis, P. (1985). *Some simple models for continuous variate time series.* Water Resour. Bull. 21, 635-644.

Li, W. (1994). *Time series models based on Generalized Linear Models: Some Further Results.* Biometrics 50, 506-511.

Li, W. & Kwok, M. (1990). *Some results on the estimation of a high order Markov chain.* Commun. Statist-Simul. Comp. 19, 363-380.

Liang, K-Y., & Zeger, S. (1986) *Longitudinal data analysis using generalized linear models.* Biometrika 73, 13-22.

Liu, J. (1995). *Metropolized Gibbs sampler.* Technical report, Dept. of Statistics, Stanford Univ., CA.

Liu, J. & Chen, R. (1995). *Blind deconvolution via sequential imputations.* J. American Statist. Assoc. 90. 567-576.

Liu, C., Liu, J. & Rubin, D. (1994). *A variational control variable for assesing the convergence of the Gibbs sampler.* Amer. Statist. Assoc. 74-78. Statistical Computing Section.

Liu, C. & Rubin, D. (1992). *The ECME algorithm: a simple extension of EM and ECM with faster monotonous convergence.* Biometrika 81, 633-648.

MacDonald, I. & Zucchini, W. (1997). *Hidden Markov and Other Models for Discrete-valued Time Series.* Chapman & Hall, London.

McKenzie, E. (1985). *Some simple models for discrete variate time series.* Water Resources Bulletin 21. 645-50.

McKenzie, E. (1985a). *Some simple models for discrete variate time series.* Water Resources Bulletin 21. 645-50.

McKenzie, E. (1985b). *Contribution to the Discussion of 'Modelling and Residual Analysis of Nonlinear Autoregressive Time-Series' by Lawrence, A. & Lewis, P.*. J. R. Statist. Soc. (B) 47, 187-188.

McKenzie, E., (1986). *Autoregressive moving-average processes with negative binomial and geometric marginal distributions.* Adv. in Appl. Probab. 18, no. 3. 679-705.

McKenzie, E. (1987). *Innovation distributions for gamma and negative binomial autoregressions.* Scand. J. Statist. 14, 79-85.

McKenzie, E. (1988). *Some ARMA models for dependent sequences of Poisson counts.* Advances in Applied Probability 20, 822-835.

McKenzie, E. (1988a). *The distributional structure of finite moving average processes.* J. Appl. Prob. 25, 313-321.

McKenzie, E. (1988b). *Some ARMA models for dependent sequences of Poisson counts.* Advances in Applied Probability 20, 822-835.

McKenzie, E. (2003). *Discrete variate time series.* Handbook of Statistics (Shanbhag, D. & Rap, C. - editors), Volume 21, Elsevier Science, 573-606.

Mehran, F. (1989). *Analysis of discrete longitudinal data: infinite-lag Markov models.* Statistical Data Analysis and Inference. Ed. Y. Dodge. Elsevier Science Publishers, Amsterdam, 533-541.

Mengersen, K. & Tweedie, R. (1996). *Rates of convergence of the Hastings and Metropolis algorithms.* Ann. Statist., 24:101-121.

Metropolis, N., Rosenbluth, A., Rosenbluth, M., Teller, A. & Teller, E. (1953). *Equations of state calculations by fast computing machines.* J. Chem. Phys., 21:1087-1092.

Meyn, S. & Tweedie, R. (1993). *Markov Chains and Stochastic Stability.* Spronver-Verlag, New York.

Mills, T. M. & Seneta, E. (1989). *Goodness-of-fit for a branching process with immigration.* Stochastic Processes and their Applications 33, 151-161.

Muenz, L. & Rubinstein, L. (1985). *Markov Models for Covariate Dependence of a Binary Sequence.* Biometrics 41, 91-101.

Müller, P. (1991). *A generic approach to posterior integration and Gibbs sampling.* Technical report, Purdue Univ. West Lafayette, Indiana.

Müller, P. (1993). *Alternatives to the Gibbs sampling scheme.* Technical report, Institute of Statistics and Decision Sciences, Duke Univ.

Natarajan, R. & McCulloch. C. (1998). *Gibbs sampling with diffuse proper priors: A valid approach to data-driven inference?.* J. Comput. Graph. Statist. 7, 267-277.

Neal, P. & Rao, S. (2007). *MCMC for integer-valued ARMA processes. Journal of Time Series Analysis*, 28(1): 92-10.

Newey, W. & West, K. (1987). *A simple, positive semi-definite, heteroskedasticity and autocorrelation consistent covariance matrix.* Econometrica, 55. 703-708.

Pegram, G. (1980). *An autoregressive model for multilag Markov chains.* J. Appl. Prob. 17, 350-362.

Peskum, P. (1973). *Optimum Monte Carlo sampling using Markov Chains.* Biometrika 60, 607-612.

Phatarford, R. & Mardia, K. (1973) *Some results for dams with Markovian inputs.* J. Appl. Probab. 10, 166-180.

Pillai, R. & Jayakumar, K. (1994) *Specialised class L property and stationary autoregressive process.* Statit. Probab. Lett. 19, 1, 51-56.

Raftery, A. (1985a). *A model for high-order Markov chains.* J. Roy. Statist. Soc. B. 47, 528-539.

Raftery, A. (1985b). *A new model for discrete-valued time series: autocorrelations and extensions.* Rassegna di Metodi Statistici ed Applicazioni 3-4,149-162.

Raftery, A. & Tavaré, S. (1994). *Estimation and modelling repeated patterns in high order Markov chains with the mixture transition distribution model.* Appl. Statist. 43, 179-199.

Roberts, G. & Rosenthal, J. (1998). *Markov chain Monte-Carlo: Some practical implications of theoretical results* (with discussion). Canadian J. Statist. 26, 5-32.

Roberts, G. & Sahu, S. (1997). *Updating schemes, covariance structure, blocking and parametrisation for the Gibbs sampler.* J. Roy. Statist. Soc. Ser. B. 59, 291-318.

Roberts, G. & Tweedie, R. (1996). *Geometric convergence and central limit theorems for multidimensional Hastings and Metropolis algorithms.* Biometrika, 83:95-110.

Schimert, J. (1992). *A high order hidden Markov model.* Ph. D. dissertation, University of Washington.

Singh, A. & Roberts, G. (1992). *State Space Modelling of Cross-Classified Series of Counts.* Int. Statist. Rev. 60, 321-335.

Smith, J. (1979). *A generalization of the Bayesian steady forecasting model.* J. R. Statist. Soc. (B) 41, 375-387.

Steutel, F. & van Harn, K. (1979). *Discrete analogues of self-decomposability and stability.* Ann. Probab. 7, no. 5, 893-899.

Steutel, F. & van Harn, K. (1986). *Discrete operator self-decomposability and queueing networks.* Stochastic Models, 2(2): 161-169.

Steutel, F. & van Harn, K. (1993). *Stability equations for processes with stationary independent increments using branching processes and Poisson mixtures.* Stoch. Models, 19, no. 2, 235-254.

Tanner, M. & Wong, W. (1987). *The calculation of posterior distributions by data augmentation.* J. American Statist. Assoc., 82:528-550.

Teicher, H. (1954). *On the multivariate Poisson distribution.* Skand. Aktuariettidskr. 37, 1-9.

Tierney, L. (1994). *Markov Chains for Exploring Posterior Distributions.* (with discussion). Annals of Statistics, 22, 1701-1762.

Tierney, L. & Mira, A. (1999). *Some adaptive Monte Carlo Methods for Bayesian inference.* Statistics in Medicine, 18, 2507-2515.

van Harn, K., Steutel, F.W. & Vervaar,W. (1982). *Self-decomposable discrete distributions and branching processes.* Z. Wahrsch. Verw. Gebiete 61, no. 1, 97-118.

Venkatamaran, K. (1982). *A time series approach to the study of the simple subcritical Galton-Watson process with immigration.* Advances in Applied Probability.

Wakefield, J., Smith, A. Racine-Poon, A. & Gelfand, A. (1991). *Bayesian analysis of linear and non-linear population models using the Gibbs sampler.* Applied Statistics (Ser C) 43, 201-22.

Wald, A. & Wolfowitz, J. (1940). *On a test whether two samples are from the sample population.* Annals of Mathematicas Statistics 11, 147-162.

Wang, Z. (1982). *Stochastic Process.* Beijing: Scientific Press.

Wei, G. & Tanner, M. (1990). *A Monte Carlo Implementation of the EM Algorithm and the Poor Man's Data Augmentation Algorithms.* J. American Statist. Ass. Vol. 85, No. 411. 699-704.

West, M. & Harrison, P. (1989). *Bayesian Forecasting and Dynamic Models.* Springer-Verlag, New York.

West, M., Harrison, P. & Migon, H. (1985). *Dynamic generalized linear models and Bayesian forecasting (with discussion).* J. Amer. Statist. Ass. 80, 73-97.

Zeger, S. (1988). *A regression model for time series of counts.* Biometrika 75, 621-629.

Zeger, S., Liang, K-Y & Self, S. (1985). *The analysis of binary longitudinal data with time-independient covariates.* Biometrika 72, 31-38.

Zeger, S. & Qaqish, B. (1988). *Markov regression models for time series: a quasilikelihood approach.* Biometrics 44, 1019-1031.

Zhu, R. & Joe, H. (2003). *A new type of discrete self-decomposability and its application to continous-time Markov processes for modelling count data in time series.* Stoch. Models 19, no. 2, 235-254.