Estimación por Algoritmo EM para un Proceso Autoregresivo de Primer Orden Estrictamente Estacionario con Distribución Gamma$(a, 1)$

1 Introducción

El modelo AR(1) centrado (de media cero) usual se describe por la ecuación

$$Y_t = \rho Y_{t-1} + \epsilon_t$$

donde $\{\epsilon_t\}_t$ es un proceso de ruido blanco y $-1 < \rho < 1$ para garantizar la estacionariedad. La generalización a un proceso de media μ está dada por

$$(Y_t - \mu) = \rho(Y_{t-1} - \mu) + \epsilon_t$$

Cuando $\{\epsilon_t\}_t$ es Gaussiano, se tiene que $\{Y_t\}_t$ es Gaussiano y en particular, estrictamente estacionario. Sin embargo, en general resulta complicado establecer la distribución de ϵ_t oara garantizar que $\{Y_t\}_t$ sea estrictamente estacionario y tenga cierta distribución deseada $f_Y(y)$.

En [1] se propone una clase de modelos alternativos, que son estacionarios Markovianos de primer orden, donde existe una relación lineal entre los valores esperados de las observaciones, y donde es posible establecer un número de distribuciones marginales distintas a la Normal. Esto se lleva a cabo especificando la distribución marginal de Y_t y la densidad condicional de $(Y_t|Y_{t-1})$, requiriendo que se cumpla la relación

$$\mathbb{E}(Y_t|Y_{t-1}) = \rho Y_{t-1} + (1-\rho)\mu$$

donde $\mu = \int y f_Y(y)dy$ y f_Y es la densidad estacionaria deseada de Y_t. De manera similar al modelo estándar AR(1), la ecuación anterior implica que la función de autocorrelación del proceso, debe ser de la forma ρ^r. La demostración se encuentra en el Apéndice.

En [1] se muestra que es usualmente posible definir una densidad de transición $f_{Y_t|Y_{t-1}}(y_t|y_{t-1})$ para la cual se satisfagan las ecuaciones

$$f_Y(y) = \int f_{Y_t|Y_{t-1}}(y|z) f_Y(z) dz \qquad \text{y} \qquad \int y f_{Y_t|Y_{t-1}}(y|z)dy = \rho z + (1-\rho)\mu$$

En [1] este objetivo se logra introduciendo una variable latente X y considerando densidades de transición de la forma

$$f_{Y_t|Y_{t-1}}(y|z) = \int f_1(y|x) f_2(x|z) d\lambda(x)$$

donde $d\lambda$ representa ya sea la medida de conteo si X es discreta, o la medida de Lebesgue si X es continua.

El punto clave es notar que si se propone una densidad conjunta $f_{Y,X}(y,x)$ tal que $f_1(y|x) = f_{Y|X}(y|x)$ y $f_2(x|z) = f_{X|Y}(x|z)$ son las densidades condicionales y tal que $f_Y(y) = \int f_{Y,X}(y,x)d\lambda(x)$, entonces se sigue de forma inmediata que

$$\int f_{Y_t|Y_{t-1}}(y|z)f_Y(z)dz = \int \left(\int f_{Y|X}(y|x)f_{X|Y}(x|z)d\lambda(x) \right) f_Y(z)dz$$

$$= \int f_{Y|X}(y|x) \left(\int f_{X|Y}(x|z)f_Y(z)dz \right) d\lambda(x) = \int f_{Y|X}(y|x)f_X(x)d\lambda(x)$$

$$= \int f_{Y,X}(y,x)d\lambda(x) = f_Y(y)$$

Esta densidad de transición puede muestrearse simulando un proceso latente $\{X_t\}_t$, donde $Y_{t+1}|X_t \sim f_{Y|X}(\cdot|X_t)$ y $X_t|Y_t \sim f_{X|Y}(\cdot|Y_t)$.
Para obtener la relación lineal entre los valores esperados condicionales y las observaciones se requiere elegir $f_(Y,X)$ de manera que $f_{Y|X}(y|x) \propto f_{X|Y}(x|Y)f_Y(y)$ pertenezca a la misma familia de distribuciones que $f_Y(y)$.

En [1] se considera como ejemplo el caso en que se desea que $f_Y(y) \propto y^{a-1}e^{-y}$ con $a > 0$, es decir, se quiere que la distribución de proceso estacionario $\{Y_t\}_t$ sea Gamma$(a,1)$, cuya media y varianza son iguales a a.
Si se propone $f_{X|Y}(x|y) = \text{Poisson}(x|y\phi)$ con $\phi > 0$ y se fija $f_Y(y) = \text{Gamma}(a,1)$, entonces

$$f_{Y|X}(y|x) \propto f_{X|Y}(x|y)f_Y(y) \propto e^{-y\phi}(y\phi)^x y^{a-1}e^{-y} \propto y^{a+x-1}e^{-y(1+\phi)}$$

lo cual implica que $f_{Y|X}(y|x) = \text{Gamma}(a+x, 1+\phi)$, y claramente se satisface

$$\sum_{x=0}^{\infty} f_{Y,X}(y,x) = \sum_{x=0}^{\infty} f_{X|Y}(x|y)f_Y(y) = f_Y(y)$$

Así, se tiene $Y_{t+1}|X_t = x \sim \text{Gamma}(a+x, 1+\phi)$ y $X_t|Y_t = y \sim \text{Poisson}(y\phi)$. El cálculo de $\mathbb{E}(Y_t|Y_{t-1})$ se deriva a través de la siguiente expresión:

$$\mathbb{E}(\mathbb{E}(Y_t|X_{t-1})|Y_{t-1} = z) = \int \mathbb{E}(Y_t|X_{t-1} = x)f_{X|Y}(x|z)d\lambda(x) = \int \left(\int y f_{Y|X}(y|x)dy \right) f_{X|Y}(x|z)d\lambda(x)$$

$$= \int y \left(\int f_{Y|X}(y|x)f_{X|Y}(x|z)d\lambda(x) \right) dy = \int y f_{Y_t|Y_{t-1}}(y|z) = \mathbb{E}(Y_t|Y_{t-1} = z)$$

Así, se tiene que $\mathbb{E}(Y_t|Y_{t-1}) = \mathbb{E}(\mathbb{E}(Y_t|X_{t-1})|Y_{t-1})$ c.s., y como $Y_t|X_{t-1} \sim \text{Gamma}(a+X_{t-1}, 1+\phi)$ entonces

$$\mathbb{E}(Y_t|X_{t-1}) = \frac{a + X_{t-1}}{1+\phi} \qquad \text{c.s.}$$

Debido a la distribución condicional $X_{t-1}|Y_{t-1} \sim \text{Poisson}(Y_{t-1}\phi)$, se tiene que

$$\mathbb{E}(X_{t-1}|Y_{t-1}) = Y_{t-1}\phi \qquad \text{c.s.}$$

Por tanto, utilizando la identidad de esperanzas condicionales y las dos ecuaciones anteriores se llega a que

$$\mathbb{E}(Y_t|Y_{t-1}) = \frac{a + Y_{t-1}\phi}{1+\phi} = \rho Y_{t-1} + (1-\rho)\mu \qquad \text{c.s.}$$

con $\rho = \phi/(1+\phi) \in (0,1)$ y $\mu = a$. Así, se tiene un proceso $\{Y_t\}_t$ estrictamente estacionario con distribución Gamma$(a,1)$, y relación lineal requerida entre los valores esperados al tiempo t y las observación al tiempo $t-1$.

2 Estimación por Máxima Verosimilitud

Con la elección anterior para las densidades de transición $f_{X|Y}$ y $f_{Y|X}$, la función de transición del proceso $\{Y_t\}_t$ está dada por

$$f_{Y_t|Y_{t-1}}(y|z) = \sum_{x=0}^{\infty} \frac{e^{-z\phi}(z\phi)^x}{x!} \frac{(1+\phi)^{a+x} y^{a+x-1} e^{-y(1+\phi)}}{\Gamma(a+x)}$$

Debido al carácter Markoviano de $\{Y_t\}_t$, para toda $n \geq 1$ se puede expresar la densidad conjunta $f_{Y_0,\cdots,Y_n}$ como producto de las densidades de transición

$$f_{Y_0,\cdots,Y_n}(y_0,\cdots,y_n) = f_{Y_0}(y_0) \prod_{i=1}^{n} f_{Y_i|Y_{i-1}}(y_i|y_{i-1})$$

En un problema de estimación por máxima verosimilitud, dadas las observaciones $(y_0, y_1, \cdots, y_n)$ del proceso $\{Y_t\}_t$ en los tiempos $t = 0, 1, \cdots, n$, es de interés maximinzar la log-verosimilitud, dada por

$$l(a,\phi|y_0,\cdots,y_n) = \log(f_{Y_0}(Y_0|a)) + \sum_{i=1}^{n} \log(f_{Y_i|Y_{i-1}}(y_i|y_{i-1},a,\phi))$$

donde se usa la notación $f_{Y_i|Y_{i-1}}(y_i|y_{i-1},a,\phi)$ y $f_{Y_0}(y_0|a)$ para mostrar la dependencia de la distribución estacionaria y las transiciones, con los parámetros a y ϕ. Cada densidad de transición dentro de los logartimos está dada por la serie anterior, la cual no cuenta con una forma cerrada. Por esta razón, la log-verosimilitud que se quiere maximizar es intratable y es necesario recurrir a otras técnicas para estimar.

Un método que resulta natural utilizar debido a la construcción del proceso $\{Y_t\}_t$ a través del proceso latente $\{X_t\}_t$, es el Algoritmo EM. Este método se emplea en situaciones en las cuales maximizar la verosimilitud resulta complicado, pero donde el problema se simplifica al aumentar la muestra con datos latentes (no observados). En el contexto de este problema,

los datos latentes constan de los valores no observados $(x_0, x_1, \cdots, x_{n-1})$ del proceso $\{X_t\}_t$ en los tiempos $t = 0, 1, \cdots, n-1$.

Supóngase que se tiene un problema en el cual se quiere estimar un parámetro (o conjunto de parámetros) θ, y se cuenta con una observación del vector aleatorio $\mathbf{Z}$. La idea del algoritmo consiste en trabajar con la llamada log-verosimilitud completa $\log(f(\mathbf{Z}, \mathbf{T}|\theta))$, que consta de los datos observados $\mathbf{Z}$ y un conjunto de datos no observados o latentes $\mathbf{T}$, la cual tiene una expresión más sencilla que la log-verosimilitud observada $\log(f(\mathbf{Z}|\theta))$ (aquella que consta de los datos observados únicamente). Sin embargo, como no se cuenta con observaciones de las variables latentes $\mathbf{T}$ ni se conoce el valor de θ, se trabaja entonces maximizando iterativamente la esperanza condicional de la log-verosimilitud completa dado lo observado:

$$Q(\theta|\theta^{(t)}) = \mathbb{E}(\log(f(\mathbf{Z}, \mathbf{T}|\theta))|\mathbf{Z}, \theta^{(\mathbf{t})})$$

con $\theta^{(t+1)} = \text{argmax}_\theta Q(\theta|\theta^{(t)})$.

En [2] se prueba que cada maximización de Q produce un incremento en la log-verosimilitud observada, y se muestra que la sucesión $\{\theta^{(t)}\}_{t=0}^{\infty}$ converge a algún máximo local de $\log(f(\mathbf{Z}|\theta))$. Es por ello que se suele implementarse varias veces el algoritmo con distintos valores iniciales $\theta^{(0)}$.

El algoritmo EM se detalla a continución:

1. Inicialización: Dar un valor inicial $\theta^{(0)}$ para el parámetro.
2. Paso E: En el paso j, calcular $Q(\theta|\theta^{(j)}) = \mathbb{E}(\log(f(\mathbf{Z}, \mathbf{T}|\theta))|\mathbf{Z}, \theta^{(\mathbf{j})})$
3. Paso M: Determinar $\theta^{(j+1)} = \text{argmax}_\theta Q(\theta|\theta^{(j)})$
4. Iterar los pasos 2 y 3 hasta la convergencia.

Para el caso del problema de estimación sobre el proceso $\{Y_t\}_t$ dada la serie $(Y_0, Y_1, \cdots, Y_n)$, utilizado el algoritmo EM, es necesario entonces conocer lo siguiente:

- La log-verosimilitud completa

$$\log(f_{Y_0, X_0, \cdots, X_{n-1}, Y_n}(y_0, x_0, \cdots, x_{n-1}, y_n|a, \phi))$$

- Las densidades condicionales

$$f_{X_i|Y_0, \cdots, Y_n}(x_i|y_0, \cdots, y_n)$$

- Calcular la esperanza condicional

$$\mathbb{E}(\log(f_{Y_0, X_0, \cdots, X_{n-1}, Y_n}(Y_0, X_0, \cdots, X_{n-1}, Y_n|a, \phi))|Y_0, \cdots, Y_n)$$

Como consecuencia de la estructura condicional entre los procesos $\{Y_t\}_t$ y $\{X_t\}_t$ se tiene una forma sencilla para la verosimilitud completa

$$f_{Y_0,X_0,\cdots,X_{n-1},Y_n}(y_0,x_0,\cdots,x_{n-1},y_n|a,\phi) = f_{Y_0}(y_0|a)f_{X_0|Y_0}(x_0|y_0,\phi)f_{Y_1|X_0}(y_1|x_0,a,\phi)$$

$$f_{X_1|Y_1}(x_1|y_1,\phi)\cdots f_{Y_{n-1}|X_{n-1}}(y_{n-1}|x_{n-2},a,\phi)f_{X_{n-1}|Y_{n-1}}(x_{n-1}|y_{n-1},\phi)f_{Y_n|X_{n-1}}(y_n|x_{n-1},a,\phi)$$

$$= f_{Y_0}(y_0|a)\prod_{i=1}^{n} f_{Y_i|X_{i-1}}(y_i|x_{i-1},a,\phi)\prod_{i=0}^{n-1} f_{X_i|Y_i}(x_i|y_i,\phi)$$

Para simplificar la notación, se denotará a la log-verosimilitud completa por $l(a,\phi|\mathbf{Y},\mathbf{X})$, con $\mathbf{Y}=(Y_0,Y_1,\cdots,Y_n)$ y $\mathbf{X}=(X_0,X_1,\cdots X_{n-1})$. Así, se tiene entonces al tomar logaritmo en la ecuación anterior

$$l(a,\phi|\mathbf{Y},\mathbf{X}) = \log(f_{Y_0}(y_0|a)) + \sum_{i=1}^{n}\log(f_{Y_i|X_{i-1}}(y_i|x_{i-1},a,\phi)) + \sum_{i=0}^{n-1}\log(f_{X_i|Y_i}(x_i|y_i,\phi))$$

donde

$$f_{Y_0}(y_0|a) = \frac{y_0^{a-1}e^{-y_0}}{\Gamma(a)}, \quad f_{Y_i|X_{i-1}}(y_i|x_{i-1},a,\phi) = \frac{(1+\phi)^{a+x_{i-1}}y_i^{a+x_{i-1}-1}e^{-(1+\phi)y_i}}{\Gamma(a+x_{i-1})} \quad i=1,\cdots,n$$

$$\text{y} \qquad f_{X_i|Y_i}(x_i|y_i,\phi) = \frac{e^{-y_i\phi}(y_i\phi)^{x_i}}{x_i!} \quad i=0,\cdots n-1$$

Sustituyendo las funciones de transición en la log-verosimilitud completa y agrupando términos se llega a la siguiente expresión:

$$l(a,\phi|\mathbf{Y},\mathbf{X}) = \log(1+\phi)\left(na+\sum_{i=0}^{n-1}X_i\right) - \sum_{i=0}^{n-1}\log(\Gamma(a+X_i)) + (a-1)\sum_{i=0}^{n}\log(Y_i)$$

$$+\sum_{i=0}^{n-1}X_i\log(\phi Y_iY_{i+1}) - \sum_{i=0}^{n}Y_i - \phi\left(2\sum_{i=0}^{n}Y_i - Y_0 - Y_n\right) - \sum_{i=0}^{n-1}\log(X_i!) - \log(\Gamma(a))$$

Previamente al cálculo del valor esperado, se propone pre-estimar el parámetro a por medio del estimador

$$\hat{a} = \frac{\sum_{i=0}^{n}Y_i}{n+1}$$

y asignar este valor al parámetro a y tratarlo como conocido en el Algoritmo EM para estimar ϕ. Esta estimación es razonable ya que el proceso $\{Y_t\}_t$ es estacionario con densidad Gamma$(a,1)$ y se supone que Y_0 proviene de dicha distribución.
De no llevarse a cabo esta estimación previa, el problema de optimización se torna mucho más complicado y computacionalmente más costoso. Además, estimar a por un promedio temporal es adecuado dada la estacionariedad del proceso $\{Y_t\}_t$.

Tomando esperanza condicional respecto a $\mathbf{Y} = (Y_0, Y_1, \cdots, Y_n)$ y con ϕ' como parámetro para la distribución condicional $\mathbf{X}|\mathbf{Y}$, se obtiene

$$\mathbb{E}(l(a,\phi|\mathbf{Y},\mathbf{X})|\mathbf{Y},a,\phi') = \log(1+\phi)\left(na + \sum_{i=0}^{n-1}\mathbb{E}(X_i|\mathbf{Y},a,\phi')\right) - \sum_{i=0}^{n-1}\mathbb{E}(\log(\Gamma(a+X_i))|\mathbf{Y},a,\phi')$$

$$+(a-1)\sum_{i=0}^{n}\log(Y_i) + \sum_{i=0}^{n-1}\log(\phi Y_i Y_{i+1})\mathbb{E}(X_i|\mathbf{Y},a,\phi') - \sum_{i=0}^{n}Y_i - \phi\left(2\sum_{i=0}^{n}Y_i - Y_0 - Y_n\right)$$

$$-\sum_{i=0}^{n-1}\mathbb{E}(\log(X_i!)|\mathbf{Y},a,\phi') - \log(\Gamma(a))$$

Así, para calcular la esperanza condicional necesaria para llevar a cabo el Algoritmo EM, se requiere conocer las densidades condicionales

$$f_{X_i|\mathbf{Y}}(x_i|\mathbf{y},a,\phi') \qquad \forall \quad i = 0,1,\cdots,n-1$$

Sea entonces $i \in \{0,1,\cdots,n-1\}$ y obsérvese que

$$f_{X_i|\mathbf{Y}}(x_i|\mathbf{y},a,\phi') = \sum_{x_0=0}^{\infty}\cdots\sum_{x_{i-1}=0}^{\infty}\sum_{x_{i+1}=0}^{\infty}\cdots\sum_{x_{n-1}=0}^{\infty} f_{\mathbf{X}|\mathbf{Y}}(\mathbf{x}|\mathbf{y},a,\phi')$$

$$= \frac{1}{f_{\mathbf{Y}}(\mathbf{y}|a,\phi')}\sum_{x_0=0}^{\infty}\cdots\sum_{x_{i-1}=0}^{\infty}\sum_{x_{i+1}=0}^{\infty}\cdots\sum_{x_{n-1}=0}^{\infty} f_{\mathbf{Y},\mathbf{X}}(\mathbf{y},\mathbf{x}|a,\phi')$$

$$= \frac{f_{Y_{i+1}|X_i}(y_{i+1}|x_i,a,\phi')f_{X_i|Y_i}(x_i|y_i,\phi')}{f_{\mathbf{Y}}(\mathbf{y}|a,\phi')} \times \sum_{x_0=0}^{\infty} f_{Y_1|X_0}(y_1|x_0,a,\phi')f_{X_0|Y_0}(x_0|y_0,\phi') \times \cdots$$

$$\times \sum_{x_{i-1}=0}^{\infty} f_{Y_i|X_{i-1}}(y_i|x_{i-1},a,\phi')f_{X_{i-1}|Y_{i-1}}(x_{i-1}|y_{i-1},\phi')$$

$$\times \sum_{x_{i+1}=0}^{\infty} f_{Y_{i+2}|X_{i+1}}(y_{i+2}|x_{i+1},a,\phi')f_{X_{i+1}|Y_{i+1}}(x_{i+1}|y_{i+1},\phi') \times \cdots$$

$$\times \sum_{x_{n-1}=0}^{\infty} f_{Y_n|X_{n-1}}(y_n|x_{n-1},a,\phi')f_{X_{n-1}|Y_{n-1}}(x_{n-1}|y_{n-1},\phi') \times f_{Y_0}(y_0|a)$$

$$= \frac{f_{Y_{i+1}|X_i}(y_{i+1}|x_i,a,\phi')f_{X_i|Y_i}(x_i|y_i,\phi')}{f_{Y_0}(y_0|a)\prod_{j=1}^{n} f_{Y_j|Y_{j-1}}(y_j|y_{j-1},a,\phi')} \times \prod_{j=1,j\neq i+1}^{n} f_{Y_j|Y_{j-1}}(y_j|y_{j-1},a,\phi') \times f_{Y_0}(y_0|a)$$

Y por tanto se tiene que

$$f_{X_i|\mathbf{Y}}(x_i|\mathbf{y},a,\phi') = \frac{f_{Y_{i+1}|X_i}(y_{i+1}|x_i,a,\phi')f_{X_i|Y_i}(x_i|y_i,\phi')}{f_{Y_{i+1}|Y_i}(y_{i+1}|y_i,a,\phi')} \qquad \forall i = 0,1,\cdots,n-1$$

y se cumple de forma inmediata que

$$\sum_{x=0}^{\infty} f_{X_i|\mathbf{Y}}(x|\mathbf{y},a,\phi') = \frac{\sum_{x=0}^{\infty} f_{Y_{i+1}|X_i}(y_{i+1}|x,a,\phi')f_{X_i|Y_i}(x|y_i,\phi')}{f_{Y_{i+1}|Y_i}(y_{i+1}|y_i,a,\phi')} = \frac{f_{Y_{i+1}|Y_i}(y_{i+1}|y_i,a,\phi')}{f_{Y_{i+1}|Y_i}(y_{i+1}|y_i,a,\phi')} = 1$$

y por tanto, se tiene que si g es una función Borel medible,

$$\mathbb{E}(g(X_i)|\mathbf{Y},a,\phi') = \frac{\sum_{x=0}^{\infty} g(x)f_{Y_{i+1}|X_i}(y_{i+1}|x,a,\phi')f_{X_i|Y_i}(x|y_i,\phi')}{f_{Y_{i+1}|Y_i}(y_{i+1}|y_i,a,\phi')}$$

Sustituyendo por las transiciones se concluye que

$$\mathbb{E}(g(X_i)|\mathbf{Y},a,\phi') = \frac{\sum_{x=0}^{\infty} g(x)\frac{(1+\phi')^{a+x}}{\Gamma(a+x)}y_{i+1}^{a+x-1}e^{-(1+\phi')y_{i+1}}\frac{e^{-y_i\phi'}(y_i\phi')^x}{x!}}{\sum_{x=0}^{\infty}\frac{(1+\phi')^{a+x}}{\Gamma(a+x)}y_{i+1}^{a+x-1}e^{-(1+\phi')y_{i+1}}\frac{e^{-y_i\phi'}(y_i\phi')^x}{x!}} \qquad \forall i = 0,1,\cdots,n-1$$

En el problema de estimación en cuestión, se desea calcular $\mathbb{E}(g(X_i)|\mathbf{Y},a,\phi')$ con $g(x) = x, \log(\Gamma(a+x))$ y $\log(x!)$, para $i = 0,1,\cdots,n-1$, y debido a que no existen expresiones cerradas para estos valores esperados, se requiere aproximarlos mediante algún método numérico.

Debido a que se cuenta con la relación de proporcionalidad,

$$f_{X_i|\mathbf{Y}}(x_i|\mathbf{y},a,\phi') \propto f_{Y_{i+1}|X_i}(y_{i+1}|x_i,a,\phi')f_{X_i|Y_i}(x_i|y_i,\phi')$$

resulta natural emplear el Algoritmo de Monte Carlo vía cadenas de Markov Metropolis-Hastings, en el cual se construye una cadena de Markov irreducible y aperiodica, cuya distribución estacionaria es alguna distribución objetivo que se desea integrar o de la cual se quieren obtener muestras. Este algoritmo es especialmente útil cuando no se puede evaluar directamente la constante de normalización de la distribución objetivo.

En [2] se discute a profundidad la teoría del método de Metropolis-Hastings y se discute el llamado Algoritmo Monte Carlo EM, en el cual se usa precisamente algún método de Monte Carlo para estimar las esperanzas condicionales involucradas en el Algoritmo EM.

Se presenta aquí un resumen del algoritmo Metropolis-Hastings con distribución objetivo $P(x)$:

1. Seleccionar una densidad condicional $Q(x|y)$ (llamada distribución propuesta) de la cual sea sencillo simular y con mismo soporte que $P(x)$, y un valor inicial $x^{(0)}$.

2. En el paso t: Generar $x' \sim Q(\cdot|x^{(t-1)})$
3. Calcular
$$\alpha = \min\left\{1, \frac{P(x')Q(x^{(t-1)}|x')}{P(x^{(t-1)})Q(x'|x^{(t-1)})}\right\}$$
4. Hacer
$$x^{(t)} = \begin{cases} x', \text{ con probabilidad } \alpha \\ x^{(t-1)}, \text{ con probabilidad } 1-\alpha \end{cases}$$
5. Repetir los pasos 2-4 hasta conseguir una cadena del tamaño deseado.

Una vez estimados los valores esperados para el cálculo de $\mathbb{E}(l(a,\phi|\mathbf{Y},\mathbf{X})|\mathbf{Y},a,\phi')$, se requiere optimizar como función de ϕ. Derivando la expresión obtenida para este valor esperado se obtiene

$$\frac{d}{d\phi}\mathbb{E}(l(a,\phi|\mathbf{Y},\mathbf{X})|\mathbf{Y},a,\phi') = \frac{na+\sum_{i=0}^{n-1}\mathbb{E}(X_i|\mathbf{Y},a,\phi')}{1+\phi}+\frac{\sum_{i=0}^{n-1}\mathbb{E}(X_i|\mathbf{Y},a,\phi')}{\phi}-\left(2\sum_{i=0}^{n}Y_i-Y_0-Y_n\right)$$

Igualando a cero y desarrollando se llega a a ecuación cuadrática

$$K_3\phi^2+\phi(K_3-K_1)-K_2=0$$

donde $K_1 = na+2\sum_{i=0}^{n-1}\mathbb{E}(X_i|\mathbf{Y},a,\phi')$, $K_2=\sum_{i=0}^{n-1}\mathbb{E}(X_i|\mathbf{Y},a,\phi')$ y $K_3 = 2\sum_{i=0}^{n}Y_i-Y_0-Y_n$, de manera que las raices pueden encontrarse utilizando la fórmula general

$$\phi = \frac{(K_1-K_3)\pm\sqrt{(K_1-K_3)^2+4K_3K_2}}{2K_3}$$

Como Y_t y X_t son positivas con probabilidad 1 para todo t, entonces se sigue que K_1, K_2 y K_3 son también positivas y por tanto se tienen siempre una raiz positiva y una negativa. En la construcción del proceso $\{Y_t\}_t$ se requiere que $\phi > 0$, por lo cual se elige la raiz positiva, dada por

$$\phi^* = \frac{(K_1-K_3)+\sqrt{(K_1-K_3)^2+4K_3K_2}}{2K_3}$$

Tomando la segunda derivada

$$\frac{d^2}{d\phi^2}\mathbb{E}(l(a,\phi|\mathbf{Y},\mathbf{X})|\mathbf{Y},a,\phi') = -\frac{na+\sum_{i=0}^{n-1}\mathbb{E}(X_i|\mathbf{Y},a,\phi')}{(1+\phi)^2}-\frac{\sum_{i=0}^{n-1}\mathbb{E}(X_i|\mathbf{Y},a,\phi')}{\phi^2}$$

se encuentra nuevamente que $\frac{d^2}{d\phi^2}\mathbb{E}(l(a,\phi|\mathbf{Y},\mathbf{X})|\mathbf{Y},a,\phi') < 0$ para todo ϕ, ya que Y_t y X_t son positivas con probabilidad 1. Esto implica que ϕ^* es punto máximo.

Una vez que se cuenta con los valores esperados (o en este caso estimaciones para los valores esperados) y una regla para optimizar, es posible probar el Algoritmo EM con datos simulados.

3 Simulación

Previo a la implementación del algoritmo, es necesario contar con una forma para simular valores del proceso $\{Y_t\}_t$ con parámetros (a, ϕ). El algoritmo de simulación es el siguiente:

1. Generar $y_0 \sim \text{Gamma}(a, 1)$
2. En el paso $n+1$ generar $x_n \sim \text{Poisson}(y_n\phi)$ y posteriormente generar $y_{n+1} \sim \text{Gamma}(a+x_n, 1+\phi)$.
3. Repetir el paso 2 hasta obtener el tamaño deseado de la serie.

En Fig.1 se muestran ejemplos de trayectorias del proceso $\{Y_t\}_t$ con 150 observaciones y distintos valores de a y ϕ. Debido a que la media y la varianza del proceso es a, es esperado el

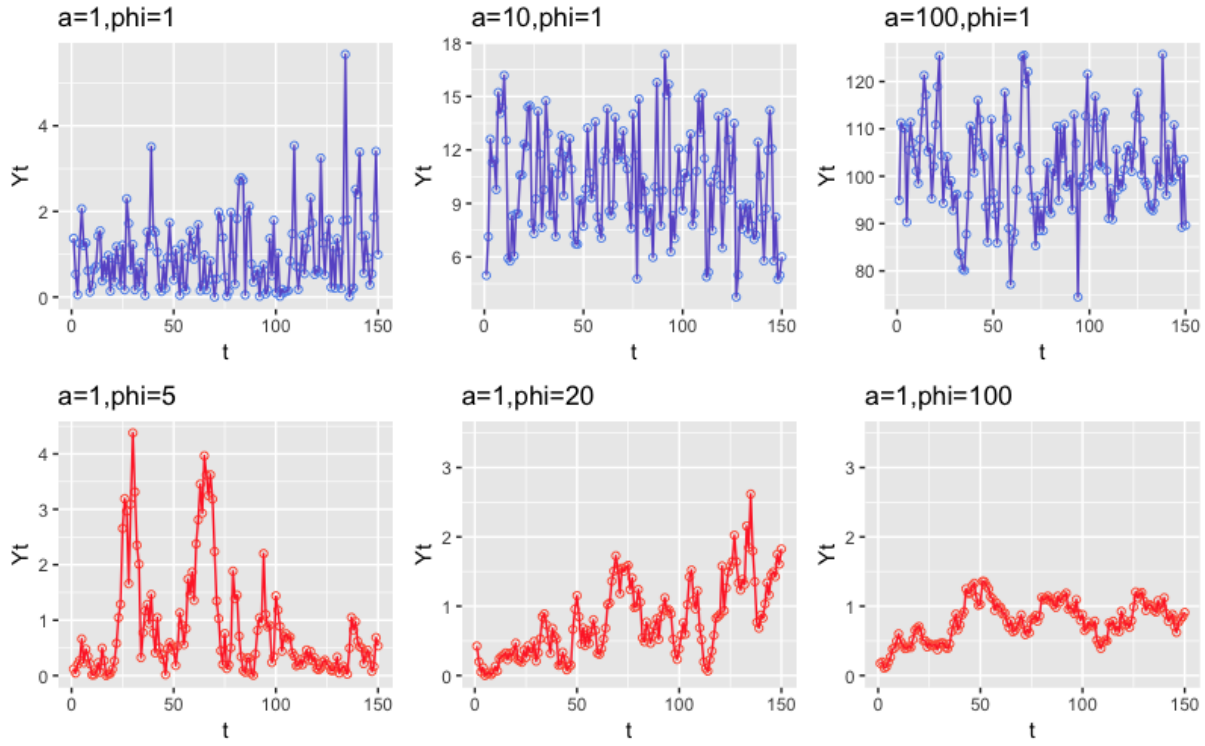

Figure 1: Realizaciones de 150 observaciones del proceso $\{Y_t\}_t$ con distintos valores para a y ϕ

comportamiento mostrado por las series al varias dicho parámetro al mantener fijo ϕ. Valores pequeños de a mantienen al proceso cerca del cero y restringen su variabilidad, mientras que valores más grandes de a lo alejan del origen e incrementan la amplitud debido al incremento de la varianza. Por otro lado, una inspección a los gráficos para distintos valores de ϕ con a

fijo, muestran que al incrementar el valor de dicho parámetro, observaciones consecutivas se agrupan más cercanamente una de otra en el tiempo, lo cual da evidencia de una correlación positiva alta, concordando con la función de autocorrelación que es de la forma $(\phi/(1+\phi))^r$.

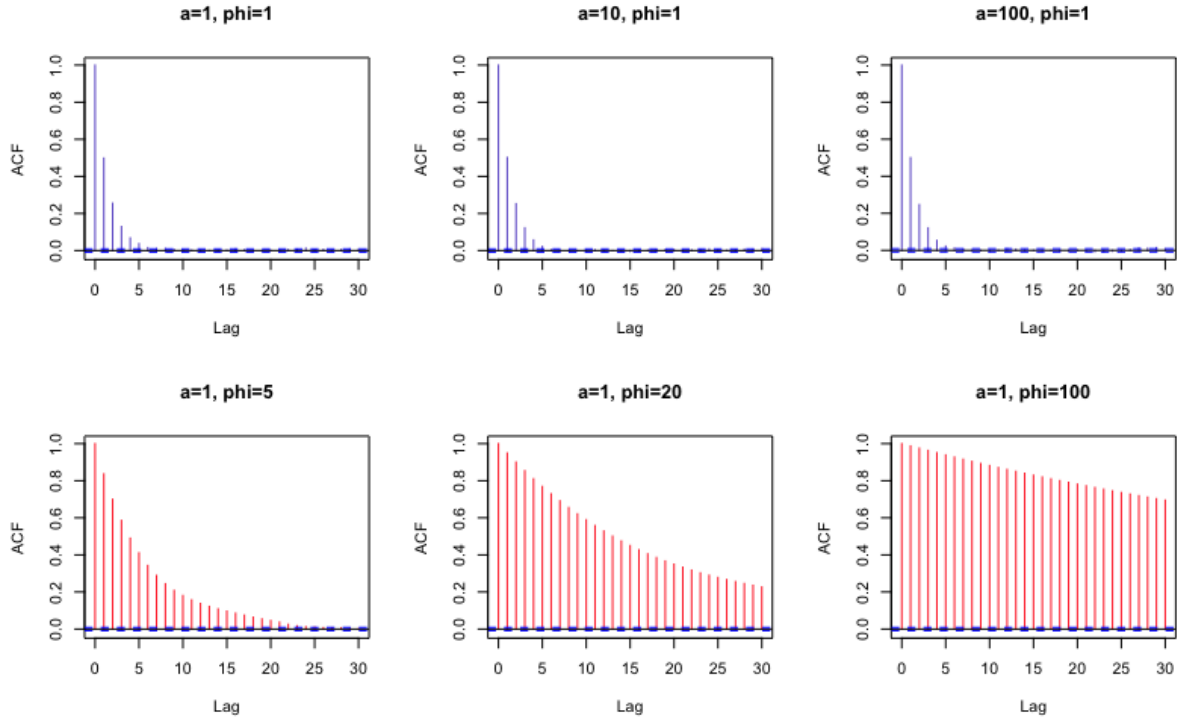

Figure 2: Correlograma del proceso $\{Y_t\}_t$ con distintos valores para a y ϕ

En la Fig.3 se incluyen los correlogramas para las series con los parámetros anteriores correspondientes (estimando la autocorrelación con series de 50,000 observaciones para tener estimados precisos para lags grandes), y se observa claramente la fuerte dependencia de la autocorrelación con el parámetro ϕ, así como el nulo impacto de a sobre la misma.

De forma similar a como se calculó $\mathbb{E}(Y_t|Y_{t-1})$, se tiene que

$$\mathbb{E}(Y_t^2|Y_{t-1}) = \mathbb{E}(\mathbb{E}(Y_t^2|X_{t-1})|Y_{t-1}) = \mathbb{E}\left(\frac{a+X_{t-1}}{(1+\phi)^2} + \frac{(a+X_{t-1})^2}{(1+\phi)^2}|Y_{t-1}\right)$$

$$= \frac{a+Y_{t-1}\phi}{(1+\phi)^2} + \frac{a^2+2aY_{t-1}\phi+Y_{t-1}\phi+(Y_{t-1}\phi)^2}{(1+\phi)^2}$$

Entonces

$$\mathrm{Var}(Y_t|Y_{t-1}) = \mathbb{E}(Y_t^2|Y_{t-1}) - \mathbb{E}^2(Y_t|Y_{t-1}) = \frac{a+Y_{t-1}\phi}{(1+\phi)^2} + \frac{a^2+2aY_{t-1}\phi+Y_{t-1}\phi+(Y_{t-1}\phi)^2}{(1+\phi)^2}$$

$$-\frac{(a+Y_{t-1}\phi)^2}{(1+\phi)^2} = \frac{a+2\phi Y_{t-1}}{(1+\phi)^2} \xrightarrow{\phi\to\infty} 0$$

de forma que al crecer ϕ, condicionalmente a Y_{t-1}, Y_t se encuentra cerca de $\mathbb{E}(Y_t|Y_{t-1})$, donde

$$\mathbb{E}(Y_t|Y_{t-1}) = \frac{a+Y_{t-1}\phi}{1+\phi} \xrightarrow{\phi\to\infty} Y_{t-1}$$

lo cual muestra que en efecto, para valores crecientes de ϕ, observaciones cercanas en el tiempo son cada vez más similares entre ellas.

Recordemos que dentro del Algoritmo EM, se utilizará el Algoritmo MCMC Metropolis-Hastings para estimar las esperanzas condicionales involucradas. En la implementación de dicho algoritmo, es crucial la elección de la distribución propuesta $Q(x|y))$, que debe ser sencilla de muestrear y debe proponer valores dentro de las regiones de alta densidad de la distribución objetivo, de lo contrario habrá una tasa de aceptación muy baja, y como consecuencia se tendrá una convergencia extremadamente lenta a la distribución estacionaria, y autocorrelación persistente. [Robert y Casella] Por esta razón tiene sentido establecer una Q que proponga valores en donde el proceso latente $\{X_t\}_t$ tiende a tomar valores. La Ley de la Esperanza Iterada implica que

$$\mathbb{E}(X_t) = \mathbb{E}(\mathbb{E}(X_t|Y_t))$$

y como $X_t|Y_t \sim \text{Poisson}(Y_t\phi)$ entonces

$$\mathbb{E}(X_t) = \mathbb{E}(Y_t\phi) = a\phi \qquad \forall t$$

ya que $\mathbb{E}(Y_t) = a$ para todo t. Similarmente, por la Ley de la Varianza Total, se tiene que para todo t

$$\text{Var}(X_t) = \mathbb{E}(\text{Var}(X_t|Y_t)) + \text{Var}(\mathbb{E}(X_t|Y_t)) = \mathbb{E}(Y_t\phi) + \text{Var}(Y_t\phi) = a\phi + a\phi^2 = a\phi(1+\phi)$$

Si bien no existe una regla estándar para elegir la distribución condicional Q, se recomienda elegirla de manera que $Q(\cdot|y)$ sea independiente de y, y que sea una buena aproximación a la distribución objetivo. Las esperanzas que quieren estimarse se toman respecto a las distribuciones condicionales $f_{X_i|\mathbf{Y}}$, y por la Ley de la Varianza Total tenemos que

$$\text{Var}(X_t) \geq \mathbb{E}(\text{Var}(X_t|\mathbf{Y}))$$

de forma que en promedio, las varianzas condicionales son menores o iguales a la varianza total o no condicional. Por esta razón, se propone tomar a Q como una distribución Poisson de parámetro λ cercano a $\mathbb{E}(X_t) = a\phi$. Una elección de parámetro que mostró funcionar y no posee una forma complicada, es

$$\lambda = \phi \times \frac{a+Y_t+Y_{t+1}}{3}$$

es decir, para estimar $\mathbb{E}(X_t|\mathbf{Y}, a, \phi)$ utilizando MCMC, se propone Q tal que

$$Q(x|z) = \text{Poisson}\left(x|\phi \times \frac{a + Y_t + Y_{t+1}}{3}\right) = \frac{e^{-\phi \times \frac{a+Y_t+Y_{t+1}}{3}}(\phi \times \frac{a+Y_t+Y_{t+1}}{3})^x}{x!}$$

y es independiente de z. Esta distribución condicional posee una varianza $\phi \min\{a, Y_t\}$ menor a la varianza del proceso $\{X_t\}_t$, sin embargo la desigualdad anterior motiva esta elección. Además, la elección de varianzas demasiado grandes para la distribución condicional Q puede provocar tasas de aceptación bajas en el algoritmo MCMC, y por tanto convergencia lenta a la distribución estacionaria.

Además de que Q podría no resultar conveniente para ciertas combinaciones de valore de a y phi, el problema principal con esta distribución condicional Q es el hecho de que en el problema de estimación, ϕ es desconocido y es justamente el parámetro que se quiere estimar. Para esto se propone dar una primera estimación de ϕ con base en el correlograma, y utilizar este valor de ϕ para Q y también como valor inicial en el algoritmo EM.
El método propuesto es utilizando mínimos cuadrados sobre la función de autocorrelación. Supóngase que se tienen lo estimados de las autocorrelaciones $C_1, C_2, \cdots C_n$ donde C_j denota la autocorrelación estimada con un lag igual a j. Como la autocorrelación teórica para un lag j está dada por ρ^j con $\rho = \phi/(1 + \phi)$, entonces se propone encontrar el valor de ρ que minimice la suma de cuadrados

$$L(\rho) = \frac{1}{n}\sum_{j=1}^{n}(C_j - \rho^j)^2 = \frac{1}{n}\sum_{i=0}^{n}(C_j - \rho^j)^2$$

Derivando e igualando a cero y recordando que $\rho \in (0, 1)$ se tiene

$$\sum_{j=0}^{n} C_j = \sum_{j=0}^{n} \rho^j = \frac{1 - \rho^{n+1}}{1 - \rho}$$

de manera que la única raíz contenida en el intervalo $(0, 1)$ del polinomio

$$\rho^{n+1} + (1 - \rho)\sum_{j=0}^{n} C_j - 1 = 0$$

es utilizada para obtener

$$\phi = \frac{\rho}{1 - \rho}$$

el cual es un valor razonable para la pre-estimación de ϕ para ser usada en la distribución propuesta Q y como valor inicial en el Algoritmo EM. Posteriormente, una vez que el Algoritmo EM vaya arrojando valores actualizados de ϕ, estos pueden proponerse para ser utilizados en la distribución Q en cada nueva iteración del algoritmo.

En cada iteración del Algoritmo EM, para una serie de observaciones $\mathbf{Y} = (Y_0, Y_1, \cdots, Y_n)$ deben ser estimadas $\mathbb{E}(X_0|\mathbf{Y}, a, \phi)$, $\mathbb{E}(X_1|\mathbf{Y}, a, \phi)$, $\cdots$, $\mathbb{E}(X_{n-1}|\mathbf{Y}, a, \phi)$, cada una por MCMC Metropolis-Hastings, es decir, en cada iteración se generan n cadenas de Markov y con cada una de ellas se estima cada esperanza condicional correspondiente.
Con la finalidad de ver el algoritmo MCMC en acción, se simula una serie de 100 observaciones $(t = 0, 1, \cdots, 99)$ de $\{Y_t\}_t$ $(a = 10, \phi = 5)$, y se toma el caso particular de ejecutar el Algoritmo MCMC para estimar $\mathbb{E}(X_{25}|\mathbf{Y}, a, \phi)$, $\mathbb{E}(X_{50}|\mathbf{Y}, a, \phi)$ y $\mathbb{E}(X_{75}|\mathbf{Y}, a, \phi)$. Para cada caso, se simulan 3 cadenas, a partir de las cuales se generan traceplots de 200 pasos, gráficos de la media acumulada con 20,000 pasos, correlogramas, histogramas y pruebas de bondad de ajuste Kolmogorov-Smirnov (versión discreta) para diagnosticar la convergencia a la distribución estacionaria.

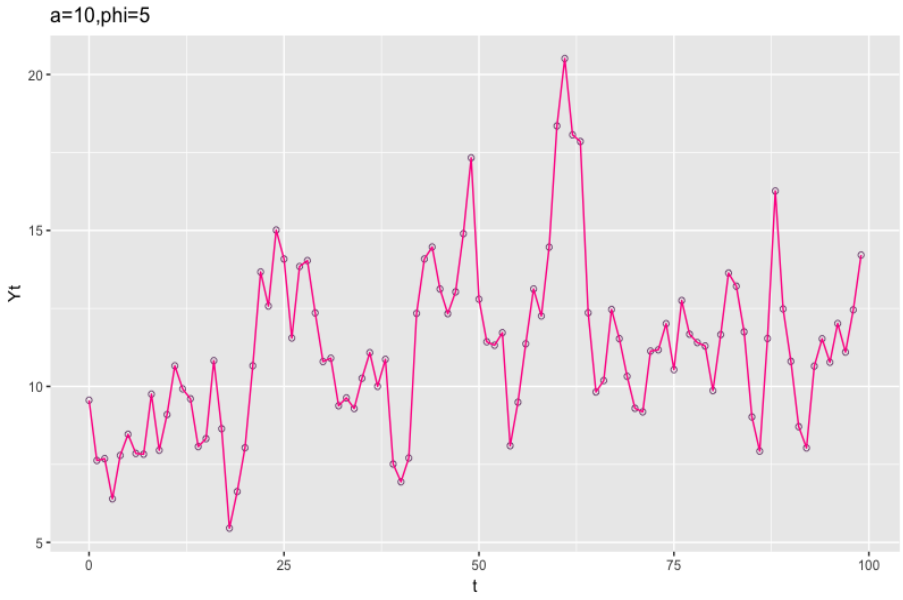

Figure 3: Serie de 100 observaciones del proceso $\{Y_t\}_t$

- Para estimar $\mathbb{E}(X_{25}|\mathbf{Y}, a, \phi)$ son utilizados únicamente los valores observados de Y_{25} y Y_{26}, que son 14.08463 y 11.55213 respectivamente. Para las tres cadenas se proponen 3 valores iniciales usando la distribución Q, y se implementa el algoritmo Metrópolis-Hastings para cada uno de los valores iniciales.

  En la parte superior de la Fig.4 se muestran los traceplots de las 3 cadenas generadas (trayectorias) únicamente de 200 pasos para poder observar el comportamiento de las cadenas con facilidad. En la parte inferior pueden apreciarse las medias acumuladas.

  Posteriormente se incluyen correlogramas para cada una de ellas.

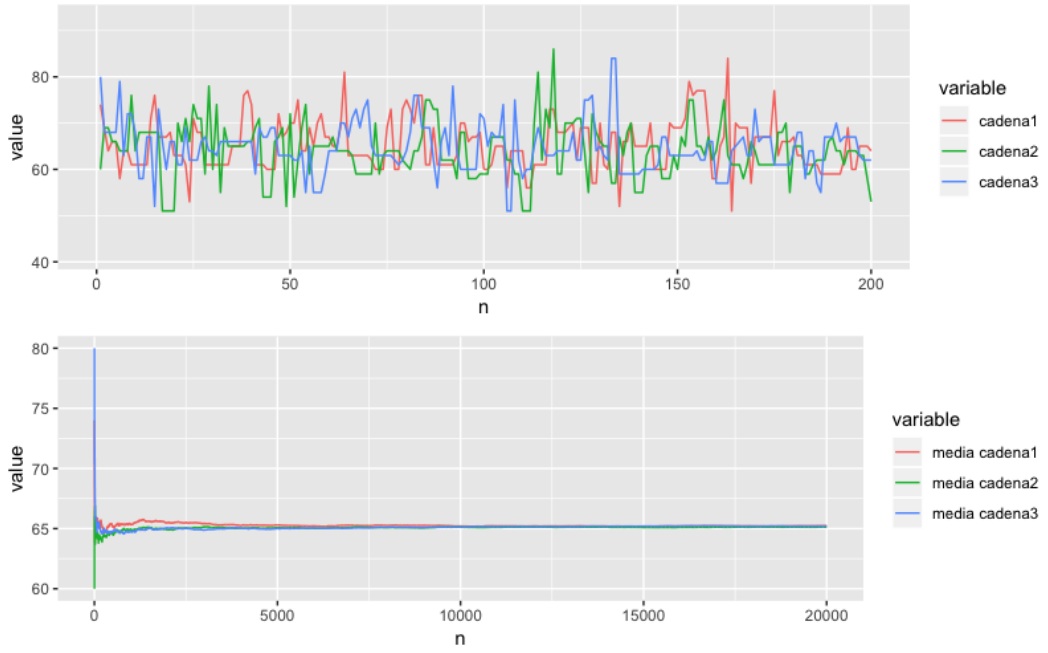

Figure 4: Traceplots y medias acumuladas de las 3 cadenas simuladas.

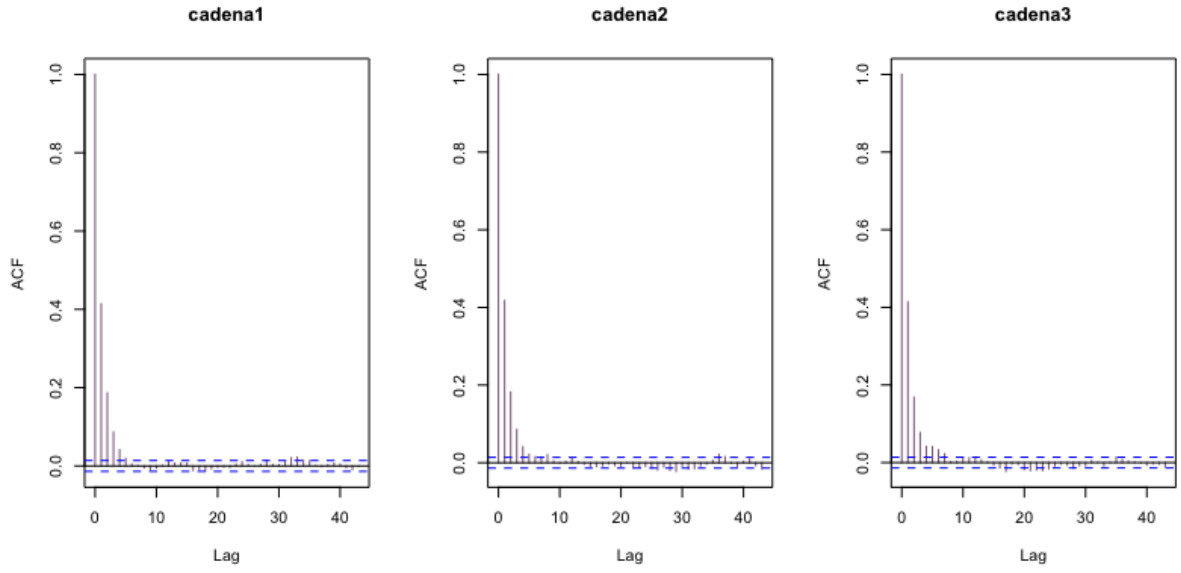

Figure 5: Correlograma estimado para cada una de las 3 cadenas simuladas.

Tanto los traceplots, medias acumuladas y autocorrelaciones estimadas dan evidencia

de un comportamiento similar de las cadenas. Esto sugiere la convergencia a la distribución estacionaria. Para probar estacionariedad, primero se descartan las primeras 1,000 observaciones de las cadenas (burn-in o periodo de calentamiento) y se lleva a cabo un thinning de las trayectorias, es decir, se toman submuestras dando saltos de tamaño k, donde k es algún lag al cual se observa correlación estimada dentro de las bandas punteadas. Para este caso un valor de k aceptable parecer ser 7. La finalidad del thinning es contar con muestras con observaciones no correlacionadas. Si bien el Teorema Ergódico justifica que el thinning no es necesario para estimar valores esperados, en necesario tener muestras aproximadamente independientes para aplicar pruebas de estacionariedad.

Finalmente, el valor de

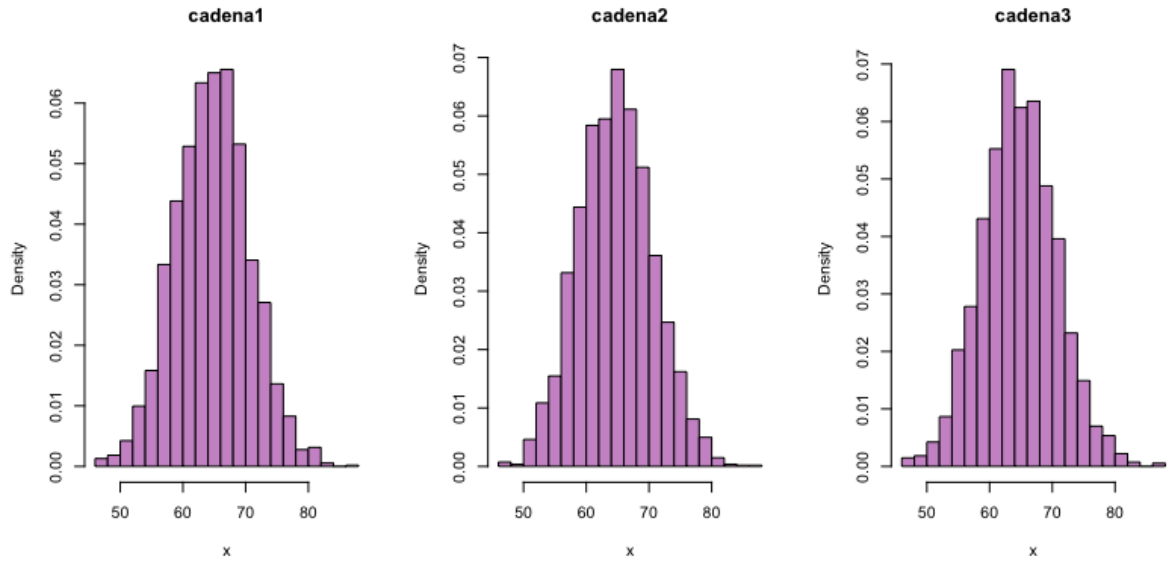

Figure 6: Histograma de frecuencias para cada una de las 3 cadenas simuladas posterior al procedimiento de thinning.

Una vez aplicado el thinning, se generan histogramas y también se comparan por pares las trayectorias de las cadenas utilizando la versión discreta de la prueba de Kolmogorov-Smirnov de dos muestras para probar que ambas muestras provienen de la misma distribución. Los resultados se muestran en la Tabla 1.

La prueba de KS no es significativa (nivel 0.05 de significancia) en ninguno de los tres casos, es decir, no se cuenta con evidencia para rechazar la hipótesis de igualdad de distribución, lo cual sugiere fuertemente la convergencia a la distribución estacionaria, y por tanto refuerza la idea de poder utilizar una de las cadenas generadas por este

Prueba KS para dos muestras	p-value
cadena 1 vs cadena 2	0.9996
cadena 1 vs cadena 3	0.9917
cadena 2 vs cadena 3	0.9978

Table 1: Tabla de resultados de la prueba de KS aplicada a las trayectorias por pares posterior al thinning.

medio para estimar el valor esperado de interés.

Finalmente, utilizando la primera cadena, tomamos la media de todas las observaciones, descartando las primeras 1,000, y estimamos

$$\mathbb{E}(X_{25}|\mathbf{Y}, a, \phi) \approx 65.23499$$

- Para estimar $\mathbb{E}(X_{50}|\mathbf{Y}, a, \phi)$ son utilizados únicamente los valores observados de Y_{50} y Y_{51}, que son 12.79582 y 11.42963 respectivamente. Se lleva a cabo el mismo análisis que en el caso anterior.

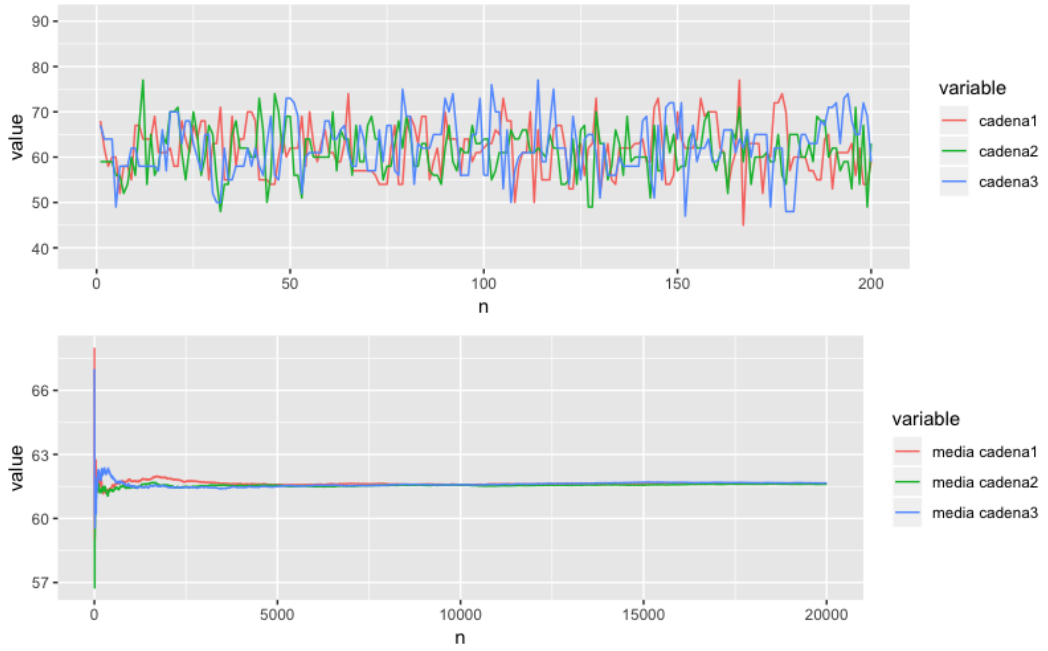

Figure 7: Traceplots y medias acumuladas de las 3 cadenas simuladas.

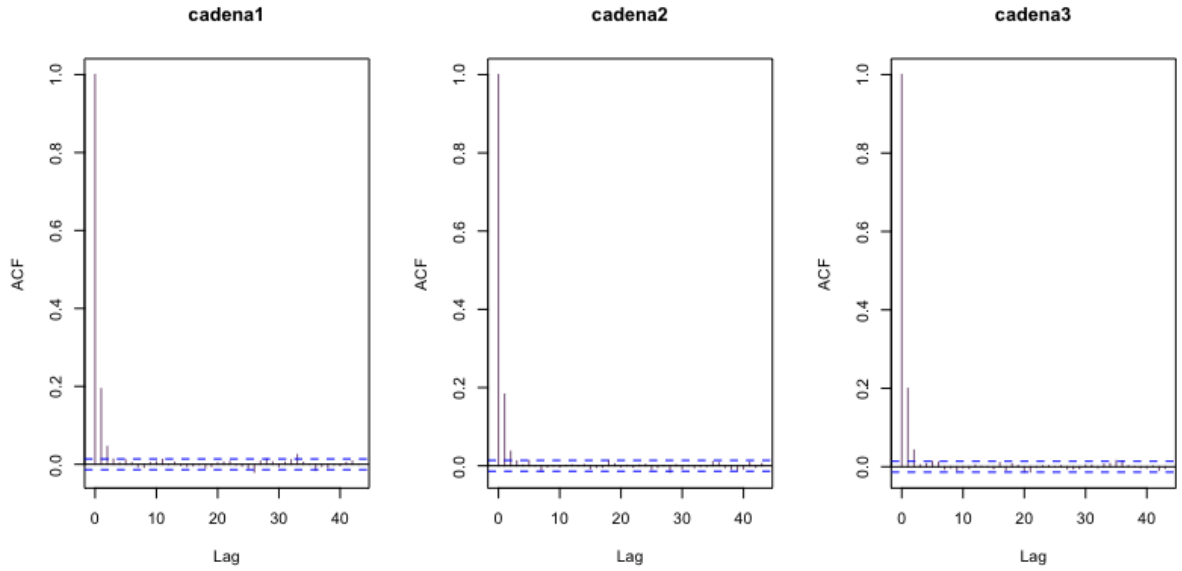

Figure 8: Correlograma estimado para cada una de las 3 cadenas simuladas.

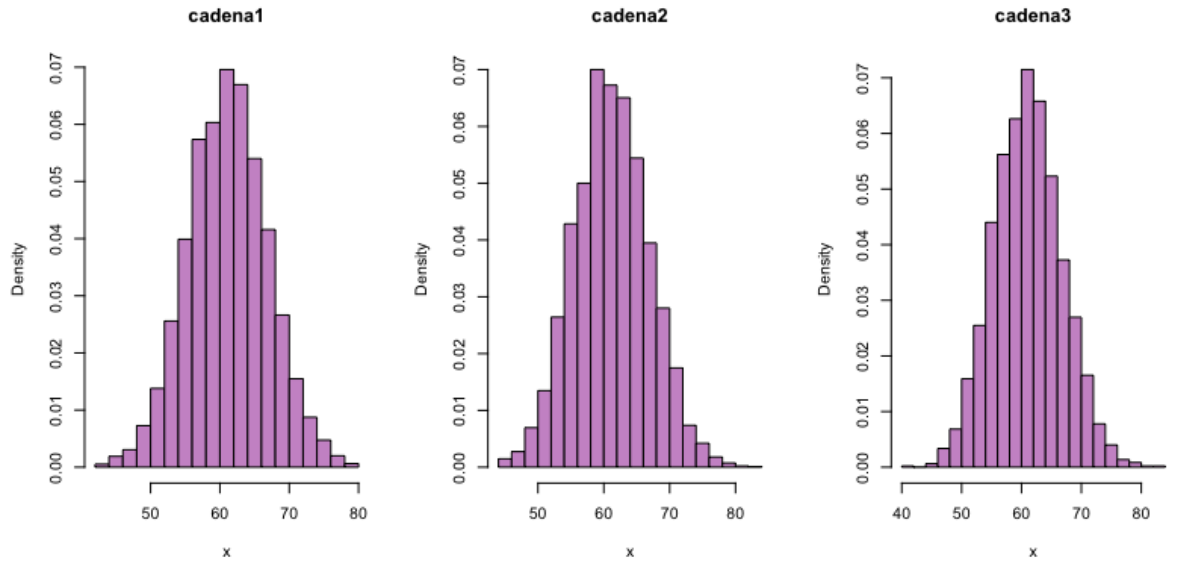

Figure 9: Histograma de frecuencias para cada una de las 3 cadenas simuladas posterior al procedimiento de thinning.

Nótese que en este caso, la autocorrelación se anula para lags más pequeños, lo cual

puede dar evidencia sobre tasas de aceptación del algoritmo MCMC más altas. En esta ocasión, el thinning se hace con un salto de tamaño 4 y un burn-in de 1,000 nuevamente. De manera similar al caso anterior, se presentan los resultados de las prueba de estacionariedad en una tabla.

Prueba KS para dos muestras	p-value
cadena 1 vs cadena 2	0.9384
cadena 1 vs cadena 3	0.6117
cadena 2 vs cadena 3	0.4629

Table 2: Tabla de resultados de la prueba de KS aplicada a las trayectorias por pares posterior al thinning.

Nuevamente, las pruebas son no significativas, por lo cual no se rechaza la estacionariedad. Finalmente se estima el valor esperado deseado utilizando la primera cadena simulada.

$$\mathbb{E}(X_{50}|\mathbf{Y}, a, \phi) \approx 61.62881$$

- Para estimar $\mathbb{E}(X_{75}|\mathbf{Y}, a, \phi)$ son utilizados únicamente los valores observados de Y_{75} y Y_{76}, que son 10.53689 y 12.75379 respectivamente. Se lleva a cabo el mismo análisis que en los casos anteriores.

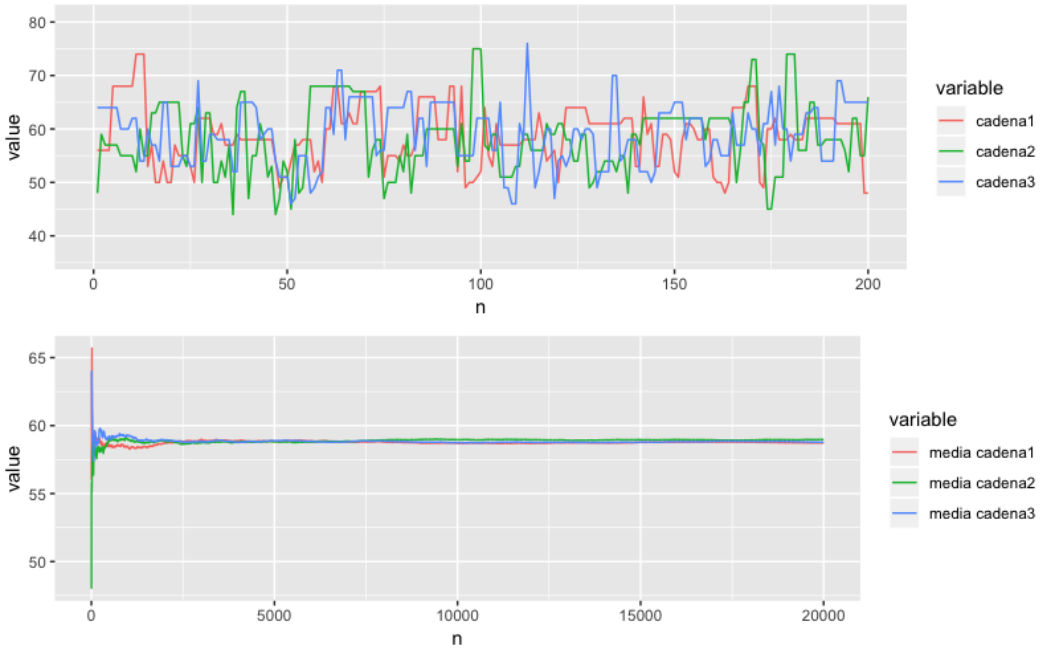

Figure 10: Traceplots y medias acumuladas de las 3 cadenas simuladas.

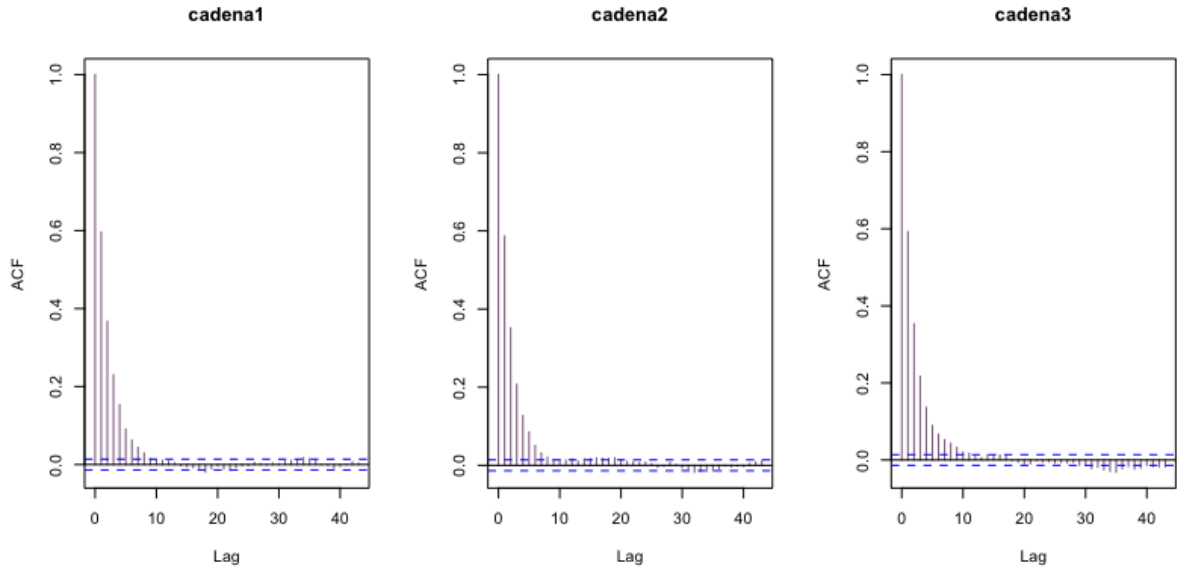

Figure 11: Correlograma estimado para cada una de las 3 cadenas simuladas.

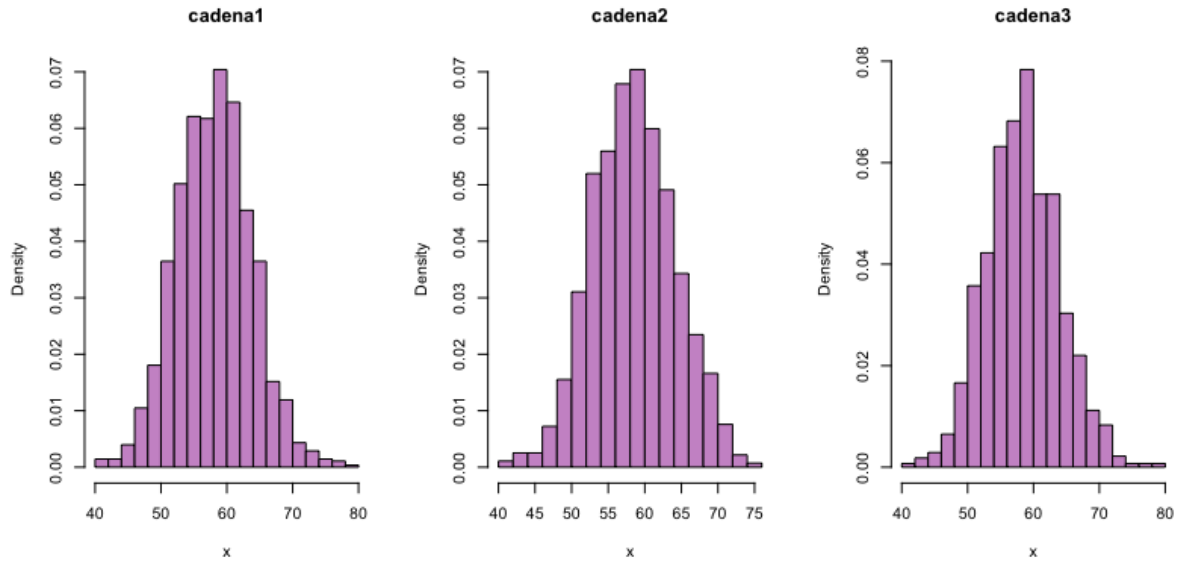

Figure 12: Histograma de frecuencias para cada una de las 3 cadenas simuladas posterior al procedimiento de thinning.

Nótese que en este caso, la autocorrelación demora más en situarse dentro de las bandas,

lo cual generalmente habla tasas de aceptación del algoritmo MCMC más bajas. En esta ocasión, el thinning se hace con un salto de tamaño 13. De manera similar al caso anterior, se presentan los resultados de las prueba de estacionariedad en una tabla.

Prueba KS para dos muestras	p-value
cadena 1 vs cadena 2	0.4025
cadena 1 vs cadena 3	0.4296
cadena 2 vs cadena 3	0.9397

Table 3: Tabla de resultados de la prueba de KS aplicada a las trayectorias por pares posterior al thinning.

En este caso, el burn-in elegido fue de 2,000, ya que de tomarse igual a 1,000, una de las comparaciones con la prueba KS es significativa, lo cual implica que se requiere esperar más pasos para que las cadenas converjan a la distribución estacionaria. Esta convergencia más lenta es consecuencia de las tasas de aceptación más bajas del algoritmo MCMC. Es este caso, esto no representa un problema y se pueden mejorar las estimaciones de los valores esperados tomando cadenas más largas.

Finalmente se estima el valor esperado deseado utilizando la primera cadena simulada.

$$\mathbb{E}(X_{75}|\mathbf{Y}, a, \phi) \approx 58.73728$$

A través de la inspección al proceso de estimación de valores esperados por algoritmo MCMC es posible notar que características tales como las tasas de aceptación del algoritmo, la autocorrelación de las cadenas y el tiempo de convergencia a la distribución estacionaria dependen de la observación X_i sobre la cual quieran estimarse los valores esperados condicionales. Es entonces una buena idea tomar un mismo periodo de burn-in largo para asegurar que todas las cadenas para distintos X_i's hayan convergido a la distribución límite.

Una vez revisado el proceso de estimación de las esperanzas condicionales, es momento de implementar el Algoritmo EM para estimar, con base en datos dados.

4 Resultados

Se obtuvieron seis series de tiempo distintas, compuestas de diferentes números de observaciones del proceso $\{Y_t\}_t$ con distintas combinaciones de parámetros. Se implementó el algoritmo EM aquí discutido con estimación de esperanzas condicionales utilizando MCMC, Se eligió tomar cadenas de 10,000 pasos, con un burn-in de 2,000 pasos para la estimación por MCMC. En el caso del algoritmo EM, se fijó un número máximo de 10,000 iteraciones, y debido al ruido inducido por la estimación por MCMC, se calculó un promedio móvil basado

en las últimas 100 iteraciones del algoritmo EM, para suavizar el ruido e identificar cuándo el algoritmo EM ha convergido. El criterio de paro seleccionado consiste en detener el algoritmo una vez que los promedios móvil cambian de signo, indicando que el algoritmo EM ha comenzado a oscilar en un pequeño intervalo de valores. El valor estimado que se selecciona corresponde al último promedio móvil calculado en el algoritmo. De tener un cálculo exacto de los valores esperados, este criterio de paro y el uso de promedios móviles sería innecesario, y podría pararse una vez que la sucesión de valores para ϕ obtenida del EM, cambie en una fracción menor a una tolerancia elegida.

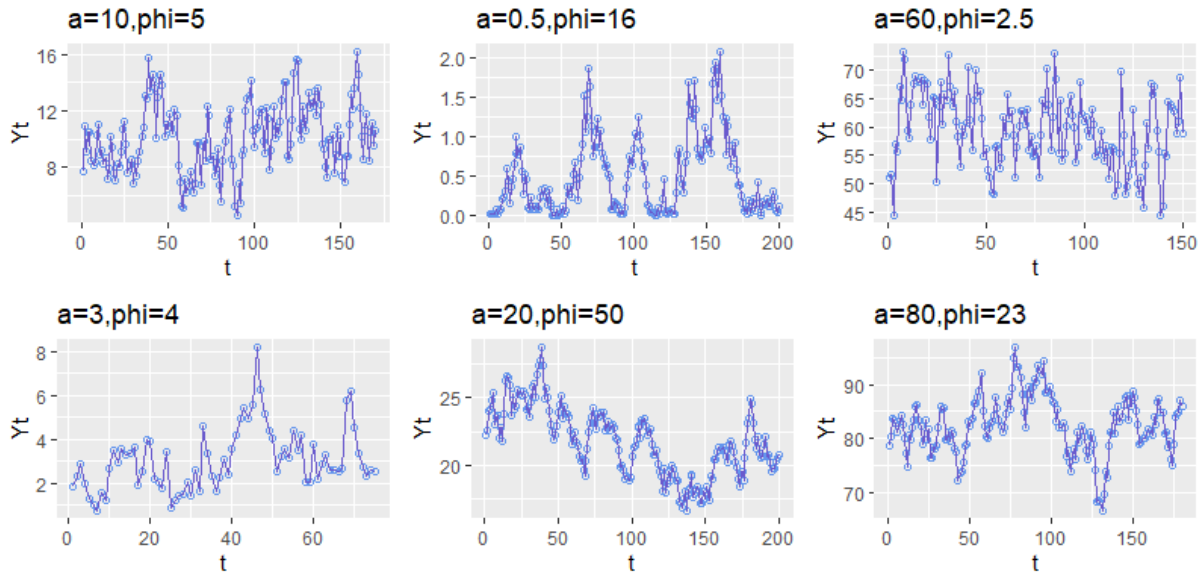

Figure 13: Seis series de tiempo con distinto número de observaciones y diferentes valores para a y ϕ

En la tabla siguiente se muestran los resultados del proceso de estimación, redondeados a 4 cifras decimales. Se incluyen los valores reales de los parámetros, el valor inicial para ϕ, los parámetros estimados, y el número de iteraciones requeridas para la convergencia del algoritmo.

(a, ϕ)	Observaciones	ϕ Inicial	$\hat{a}$	$\hat{\phi}$	Iteraciones
(10,5)	170	2.3891	10.0932	4.8716	1,584
(0.5,16)	200	9.1717	0.4978	15.9344	341
(60,2.5)	150	1.2704	59.5786	2.0806	3,281
(3,4)	75	2.5614	3.0626	3.7678	512
(20,50)	200	14.7068	21.7613	47.5695	8,199
(80,23)	180	7.1418	82.5368	20.2428	13,997

Table 4: Tabla de resultados de estimación.

Los valores iniciales propuestos para ϕ se obtuvieron a partir de las primeras cinco autocorrelaciones (incluyendo la primera, cuyo valor siempre es 1). Es posible notar en la tabla que en general los valores iniciales de ϕ no son tan cercanos a los valores reales. Esto se debe a que para series de pocas observaciones, las estimaciones de la autocorrelación no son tan buenas. Sin embargo, los valores de ϕ iniciales obtenidos por este método, son suficientemente cercanos al valor real, de manera que el Algoritmo MCMC no se atasca debido a parámetros inadecuados de la distribución condicional propuesta Q. Al incrementarse el número de observaciones, la autocorrelación será mejor estimada, y por tanto los valores iniciales de ϕ serán más cercanos al valor real.

Es bien sabido que el Algoritmo EM, si bien es una poderosa herramienta para estimar por máxima verosimilitud, resulta ser un algoritmo que en general converge muy lentamente, por lo cual se suele requerir un número elevado de iteraciones y puede resultar computacionalmente costoso.

En adición al gran número de iteraciones que se requieren para converger, mientras mas observaciones contenga la serie de tiempo analizada, mayor número de cadenas de Markov deben ser construidas para estimar las esperanzas condicionales. Para una serie con $n+1$ observaciones $\mathbf{y} = (y_0, y_1, \ldots, y_n)$,se requiere estimar los n valores esperados

$$\mathbb{E}(X_i|\mathbf{Y} = \mathbf{y})$$

en cada iteración. Por tanto, a menos que se cuente con un gran poder de cómputo, analizar series con un gran número de observaciones puede resultar ineficiente, a pesar de que más observaciones pueden brindar mejores estimados. Si se cuenta con una serie muy larga, es una buena idea estimar la autocorrelación empleando toda la serie, a partir de ella dar un valor inicial ϕ cercano al valor real, y estimar ϕ por medio del Algoritmo EM empleando un segmento más corto de la serie.

Los valores estimados para a y ϕ son en general buenos. Sin embargo, para una serie con un valor de ϕ muy alto, la autocorrelación será muy grande y por ello, estimar a utilizando el promedio ergódico, puede dar un mal estimado si no se cuenta con muchas observaciones. Una modificación propuesta al método para tratar de mejorar la estimación consiste en estimar tanto a como ϕ en el Algoritmo EM, y utilizar

$$\hat{a} = \frac{\sum_{i=0}^{n} Y_i}{n+1}$$

únicamente como valor inicial. El hecho de no haber incluido la estimación de a dentro del Algoritmo EM en este trabajo, se debe a que la regla de optimización para ϕ es sencilla y está dada en términos de una raiz de un polinomio de segundo grado. Al incluir a a dentro del proceso de estimación, en cada paso el Algoritmo EM se debe maximizar conjuntamente

$$\mathbb{E}(l(a, \phi|\mathbf{Y}, \mathbf{X})|\mathbf{Y}, a', \phi')$$

como función de a y ϕ. Además de que la optimización no tendrá una expresión cerrada, no se cuenta con una forma explícita para

$$\mathbb{E}(\log(\Gamma(a+X_i))|\mathbf{Y}, a', \phi'), \qquad i = 0, \ldots, n-1$$

por lo cual se tendrán que estimar también por MCMC. Así, en cada iteración del algoritmo EM empleando una serie de $n+1$ observaciones, se estimarán las n esperanzas condicionales $\mathbb{E}(X_i|\mathbf{Y}, a', \phi')$, las n esperanzas condicionales para $\mathbb{E}(\log(\Gamma(a+X_i))|\mathbf{Y}, a', \phi')$, y posteriormente se llevará a cabo la optimización mediante algún método numérico iterativo que a su vez requerirá la re-estimación sucesiva de los valores esperados $\mathbb{E}(\log(\Gamma(a+X_i))|\mathbf{Y}, a', \phi')$, elevando significativamente el costo computacional del problema de estimación.

5 Apéndice

En esta sección se prueba por inducción el resultado siguiente:

Para el proceso $\{Y_t\}_t$ estacionario Markov de primer orden, que satisface la ecuación

$$\mathbb{E}(Y_{t+1}|Y_t) = \rho Y_t + (1-\rho)\mu$$

con μ la media del proceso, se tiene la siguiente forma para la autocorrelación:

$$\text{Cor}(Y_s, Y_t) = \rho^{|t-s|}$$

Demostración

Basta con mostrar que

$$\text{Cor}(Y_{t+n}, Y_t) = \rho^n \qquad \forall n \geq 0$$

lo cual se llevará a cabo por inducción.

Para $n = 0$ es inmediato que

$$\text{Cor}(Y_t, Y_t) = \frac{\text{Cov}(Y_t, Y_t)}{\sqrt{\text{Var}(Y_t)\text{Var}(Y_t)}} = \frac{\text{Var}(Y_t)}{\text{Var}(Y_t)} = 1$$

Para $n = 1$, nótese que por la Ley de la Esperanza Iterada y usando que Y_t es $\sigma(Y_t)$-medible, se tiene que

$$\mathbb{E}(Y_{t+1}Y_t) = \mathbb{E}(\mathbb{E}(Y_{t+1}Y_t|Y_t)) = \mathbb{E}(Y_t\mathbb{E}(Y_{t+1}|Y_t)) = \mathbb{E}(Y_t(\rho Y_t + (1-\rho)\mu)) = \rho\mathbb{E}(Y_t^2) + (1-\rho)\mu\mathbb{E}(Y_t)$$

Sea σ^2 la varianza del proceso. Se tiene entonces

$$\mathbb{E}(Y_{t+1}Y_t) = \rho(\sigma^2 + \mu^2) + (1-\rho)\mu^2 = \rho\sigma^2 + \mu^2$$

$$\implies \quad \text{Cov}(Y_{t+1}, Y_t) = \mathbb{E}(Y_{t+1}Y_t) - \mathbb{E}(Y_{t+1})\mathbb{E}(Y_t) = \rho\sigma^2$$

Por tanto se llega a que

$$\text{Cor}(Y_{t+1}, Y_t) = \frac{\text{Cov}(Y_{t+1}, Y_t)}{\sqrt{\text{Var}(Y_{t+1})\text{Var}(Y_t)}} = \frac{\rho\sigma^2}{\text{Var}(Y_t)} = \rho$$

Ahora supóngase que para $n = k$ se cumple la ecuación

$$\text{Cor}(Y_{t+k}, Y_t) = \rho^k$$

Como $\{Y_t\}_t$ es Markov de primer orden, se tiene que Y_{t+k+1} y Y_t son condicionalmente independientes, dado Y_{t+k}, es decir

$$\mathbb{E}(Y_{t+k+1}Y_t|Y_{t+k}) = \mathbb{E}(Y_{t+k+1}|Y_{t+k})\mathbb{E}(Y_t|Y_{t+k})$$

Con este resultado, la Ley de la Esperanza Iterada y sabiendo que Y_{t+k} es $\sigma(Y_{t+k})$-medible, se tiene que

$$\mathbb{E}(Y_{t+k+1}Y_t) = \mathbb{E}(\mathbb{E}(Y_{t+k+1}Y_t|Y_{t+k})) = \mathbb{E}(\mathbb{E}(Y_{t+k+1}|Y_{t+k})\mathbb{E}(Y_t|Y_{t+k}))$$

$$= \mathbb{E}((\rho Y_{t+k} + (1-\rho)\mu)\mathbb{E}(Y_t|Y_{t+k})) = \rho\mathbb{E}(Y_{t+k}\mathbb{E}(Y_t|Y_{t+k})) + (1-\rho)\mu\mathbb{E}(\mathbb{E}(Y_t|Y_{t+k}))$$

$$= \rho\mathbb{E}(\mathbb{E}(Y_{t+k}Y_t|Y_{t+k})) + (1-\rho)\mu\mathbb{E}(Y_t) = \rho\mathbb{E}(Y_{t+k}Y_t) + (1-\rho)\mu^2$$

Entonces

$$\text{Cov}(Y_{t+k+1}, Y_t) = \mathbb{E}(Y_{t+k+1}Y_t) - \mathbb{E}(Y_{t+k+1})\mathbb{E}(Y_t) = \mathbb{E}(Y_{t+k+1}Y_t) - \mu^2 = \rho\mathbb{E}(Y_{t+k}Y_t) - \rho\mu^2$$

$$\rho(\mathbb{E}(Y_{t+k}Y_t) - \mu^2 + \mu^2) - \rho\mu^2 = \rho\text{Cov}(Y_{t+k}, Y_t)$$

Dividiendo por σ^2 y usando la hipótesis de inducción se llega a

$$\text{Cor}(Y_{t+k+1}, Y_t) = \rho\text{Cor}(Y_{t+k}, Y_t) = \rho\rho^k = \rho^{k+1}$$

y por tanto se concluye que

$$\text{Cor}(Y_{t+n}, Y_t) = \rho^n \qquad \forall n \geq 0$$

References

[1] Micheal K. Pitt, Chris Chatfield, and Stephen G. Walker. *Constructing First Order Stationary Autoregressive Models via Latent Processes*. Scandinavian Journal of Statistics, 29: 657–663, 2002

[2] Trevor Hastie, Robert Tibshirani, and Jerome Friedman. *The Elements of Statistical Learning*. 3rd ed., Springer, 2009.

[3] Christian Robert and George Casella. *Monte Carlo Statistical Methods*. 2nd ed., Springer, 1999.