XXIX ESCUELA VENEZOLANA DE MATEMÁTICAS
EMALCA-VENEZUELA 2016

MODELOS DE MARKOV OCULTOS

Lisandro Fermín
Luis-Angel Rodríguez
Ricardo Ríos

MÉRIDA, VENEZUELA, 04 al 09 de septiembre de 2016

XXIX ESCUELA VENEZOLANA DE MATEMÁTICAS
EMALCA–VENEZUELA 2016

Modelos de Markov Ocultos

Lisandro Fermín

Universidad de Valparíso, Chile
lisandro.fermin@uv.cl

Ricardo Ríos

Universidad Central de Venezuela
ricardo.rios@ciens.ucv.ve

Luis-Angel Rodríguez

Universidad De Carabobo, Venezuela
larodri@uc.edu.ve

MÉRIDA, 4 AL 9 DE SEPTIEMBRE DE 2016

XXIX ESCUELA VENEZOLANA DE MATEMÁTICAS

La Escuela Venezolana de Matemáticas es una actividad de los postgrados en matemáticas de las instituciones siguientes: Centro de Estudios Avanzados del Instituto Venezolano de Investigaciones Científicas, Facultad de Ciencias de la Universidad Central de Venezuela, Facultad de Ciencias de la Universidad de Los Andes, Universidad Simón Bolívar, Universidad Centroccidental Lisandro Alvarado y Universidad de Oriente, y se realiza bajo el auspicio de la Asociación Matemática Venezolana. La XXIX Escuela Venezolana de Matemáticas recibió financiamiento de la Academia de Ciencias Físicas, Matemáticas y Naturales de Venezuela, el Instituto Venezolano de Investigaciones Científicas (Centro de Estudios Avanzados, Departamento de Matemáticas y Ediciones IVIC), la Universidad de los Andes (CEP, CDCHT, CODEPRE, Departamento de Matemáticas de la Facultad de Ciencias, Decanato de Ciencias y Vicerrectorado Administrativo), la Asociación Matemática Venezolana, la Unión Matemática de América Latina y el Caribe (UMALCA) y Centre International de Mathématiques Pures et Appliquées (CIMPA).

2010 Mathematics Subject Classification:

Instituto Venezolano de Investigaciones Científicas
Rif: G-20004206-0

Modelos de Markov Ocultos

Lisandro Fermín, Ricardo Ríos, Luis-Angel Rodríguez

Diseño y edición: Escuela Venezolana de Matemáticas
Depósito legal DC2016000283
ISBN 978-980-261-169-0
Caracas, Venezuela
2016

A Diego Fermín y Diomedes Bárcenas

II

Índice general

Capítulo 1

Introducción

En este libro estámos interesados en dar a conocer resultados recientes en el activo tema de los Procesos Aleatorios controlados por una Cadena de Markov oculta conocidos en la literatura anglosajona Switching Markov Processes.

Una cadena de Markov oculta es una sucesión de variables aleatorias $Y = \{Y_n\}_{n\geq 0}$ que se supone ocurre luego de la realización de una Cadena de Markov $X = \{X_n\}_{n\geq 1}$ que no es observada, llamada régimen, siendo que el modelaje estadístico se hace solo con la información aportada por el proceso Y.

En la literatura estándar de Cadenas de Markov ocultas las variables aleatorias Y_n condicionales a X_n son independientes, en nuestro caso podría existir dependencia a un paso. Esta caracterización nos permitirá tratar de manera unificada los tres ejemplos principales que estudiaremos: Cadenas de Markov ocultas con espacio de estado finito para las observaciones y los estados ocultos con espacio de estado finito, Cadenas de Markov con observaciones continuas y con espacio de estado finito y los procesos autorregresivos con régimen de Markov (AR-RM).

La dependencia que tiene Y_n de Y, X se puede hacer explícita porque se puede suponer sin perder generalidad que $Y_n = r(Y_{n-1}, X_n, e_n)$ para alguna función medible r y una sucesión de variables aleatorias $e = \{e_n\}_{n\geq 1}$ independientes de Y_0 y de X. A lo largo de estas notas

sólo consideramos como espacio de estados de la cadena de Markov X, al conjunto finito $\{1, \ldots, m\}$ y denotaremos por $P = [p_{ij}]$ su matriz de transición.

En general el proceso Y no es una cadena de Markov. Sin embargo el proceso conjunto $Z = (Y, X)$ si es una cadena de Markov. En la Figura 1.1 se observa un esquema del mecanismo de generación para un proceso AR-RM.

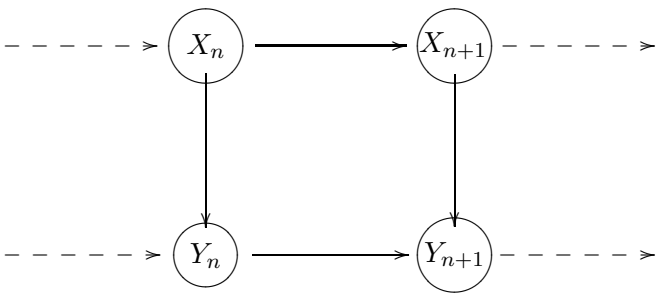

Figura 1.1: Esquema descriptivo de un modelo de Markov oculto

Cuando la función r sólo depende de los procesos X y e se obtiene la subclase de procesos conocidos como *Cadenas de Markov ocultas* (CMO).

Los modelos de CMO son usados en distintas áreas de las ciencias básicas y aplicadas, así como en la industria, las finanzas y la economía, desde el análisis de rutina hasta la resolución de problemas de alta envergadura: reconstrucción de imágenes, reconocimiento de patrones, tomografía, resolución de problemas inversos, etc. ver Cappe [11] y McDonald y Zucchini [33] para referencias más completas.

Los modelos de CMO fueron introducidos por Blackwell and Koopmans [50] como *funciones probabilísticas de una cadena de Markov*. Ellos se ocupan del siguiente problema probabilístico: para todos los procesos estacionarios $Y = \{Y_n\}_{n\geq 0}$ con valores en un conjunto discreto, se quiere caracterizar cuáles admiten la representación $Y_n = r(X_n)$ y estu-

diar sus propiedades. Heller [27] aborda este problema de una manera no constructiva. Utilizando algunos aspectos de la teoría de realización estocástica se pueden dar algoritmos que permitan, dado un proceso $Y = \{Y_n\}_{n\geq 0}$, que se puede representar como un modelo de CMO, construir una cadena de Markov $\{X_n\}_{n\geq 0}$ y una función r tal que el proceso $Y_n = r(X_n)$ o al menos Y_n y $r(X_n)$ tengan la misma distribución, ver Finesso [22].

Las primeras contribuciones relacionadas con la estimación por máxima verosimilitud de los modelos CMO (consistencia y normalidad asintótica) se deben a Baum, Petrie y sus colaboradores, quienes a mediados de los años sesenta del siglo XX desarrollaron sus propiedades en una serie de artículos, ver [7, 8] y sus referencias. Ellos proponen un algoritmo de cálculo numérico de la verosimilitud. Estos autores, introducen el algoritmo EM el cual se ha hecho muy popular con la aparición del trabajo de Dempster *et al.* [17]. Después, a mediados de los setenta, los modelos de CMO hacen una aparición fugaz aparición en la literatura estadística. En 1975 Baker [6] propone un CMO como modelo de reconocimiento automático de patrones y sigue siendo hoy día muy utilizado en este campo. Los aspectos computacionales del algoritmo de Baum y el reconocimiento de patrones están recogidos en Levison *et al.* [32].

La consistencia y la normalidad asintótica para el estimador de máxima verosimilitud (EMV) desarrolladas por Baum y Petrie [7] son extendidas al considerar espacios más generales para el proceso Y (por ejemplo $\mathbb{R}^d$). Leroux [31] establece la consistencia del EMV, mientras que la normalidad asintótica es establecida por Bickel *et al.* en [9, 10]. Las propiedades asintóticas del EMV cuando el espacio de estados de la cadena X es compacto son establecidas por Jensen y Petersen [52], Douc y Matias [34].

Regresando a nuestro planteamiento general $Y_n = r(Y_{n-1}, X_n, e_n)$, podemos suponer que la función r tiene una forma aditiva conveniente que permite escribir Y_n en la forma,

$$Y_n = r(Y_{n-1}, X_n) + e_n, \tag{1.1}$$

o de manera equivalente,

$$Y_n = r_{X_n}(Y_{n-1}) + e_n. \tag{1.2}$$

La ecuación (1.2) nos permite entender el modelo AR-RM como la combinación de m modelos de autorregresión seleccionados según la realización de la cadena de Markov X. En estas notas se puede presentar el modelo en las dos formas. Goldfeld and Quand [53] los introducen como una generalización de los modelos de regresión alternantes (switching en la literatura anglosajona) el cambio de régimen es aleatorio.

El interés en este tipo de procesos es amplio tanto teóricamente como en las aplicaciones. Los procesos AR-RM son usados en muchas áreas porque representan modelos heterogéneos no independientes. Hamilton [25] estudia, en el contexto econométrico, un proceso AR lineal con régimen de Markov para el análisis de la serie temporal del producto interno bruto de los Estados Unidos, con dos regímenes: uno de contracción y otro de expansión. Los modelos autorregresivos lineales con régimen de Markov también han sido usados en varios problemas surgidos de la ingeniería eléctrica: detección de fallas, control automático, ver Douc *et al.* [18].

Los aspectos probabilísticos de los AR-RM relacionados con la estabilidad del modelo son desarrollados en Yao y Attali [55]. La consistencia y la normalidad asintótica del estimador de máxima verosimilitud en el contexto de los procesos AR-RM son estudiadas por: Francq y Roussignol [51] (consistencia), Douc *et al.* [18] (consistencia y normalidad asintótica). Estos últimos las establecen tanto para el caso estacionario como para el caso no estacionario, utilizando una técnica de acoplamiento introducida por Bakry *et al.* [14] para el modelo CMO cuando el proceso Y toma valores en un espacio finito. Para la estimación semi paramétrica y no paramétrica referimos a [40, 30].

Nuestro objetivo en estas notas es introducir al lector en el estudio de series con cambios de régimen en el tiempo, de manera de caracterizar los distintos modos de la serie utilizando un modelo que permite tener buenas propiedades teóricas y versatilidad en la gama de algoritmos que se pueden implementar.

Capítulo 2

Series de Tiempo

En este capítulo, siempre con la intención de hacer estas notas autocontenidas, presentamos algunos aspectos básicos del extenso campo del análisis de series temporales.

Un modelo matemático para una medición temporal discretizada (proceso estocástico a tiempo discreto o serie de tiempo) puede ser formulado introduciendo una sucesión $\{Y_n\}_{n\in\mathbb{Z}}$ de variables aleatorias definidas en un espacio de probabilidad $(\Omega, \mathcal{F}, \mathbb{P})$.

En el enfoque empleado prescindimos de la formulación general de proceso estocástico y lo que supondremos es que conocemos la distribución conjunta de una muestra

$$(Y_0, Y_1, \ldots, Y_n)^T$$

de la sucesión $\{Y_n\}_{n\in\mathbb{Z}}$. En el documento esta distribución conjunta estará determinada por la densidad conjunta $p(y_0, y_1, \ldots, y_n)$ del vector $(Y_0, Y_1, \ldots, Y_n)^T$. Indicamos con $(\cdot)^T$ la operación transposición y puede aplicarse a vectores y matrices.

Definición 2.1. *La esperanza de la variable aleatoria $g(Y_0, Y_1, \ldots, Y_n)$ esta dada por*

$$\mathbb{E}g(Y_0, Y_1, \ldots, Y_n) = \int g(y_0, y_1, \ldots, y_n)p(y_0, y_1, \ldots, y_n)dy_1, \ldots, dy_n$$

donde $p(y)$ es la distribución marginal con respecto a $(Y_0, Y_1, \ldots, Y_n)^T$.

Consideramos $L^2_{\mathbb{R}}(\Omega, \mathcal{F}, P)$ el espacio de las variables aleatorias Y definidas sobre Ω y a valores en $\mathbb{R}$ tales que

$$\|Y\|_2 = (E[Y^T Y])^{1/2} < \infty,$$

donde Y es un vector columna e Y^T es su vector traspuesto. Este es un espacio de Hilbert con el producto escalar $\langle X, Y \rangle = E[X^T Y]$.

Definición 2.2. *Una serie de tiempo $\{Y_n\}_{n\in\mathbb{Z}}$ es de segundo orden si $Y_n \in L^2_{\mathbb{R}}(\Omega, \mathcal{F}, P)$, para todo $n \in \mathbb{Z}$.*

Definición 2.3. *Sea $\{Y_n\}_{n\in\mathbb{Z}}$ una serie de tiempo de segundo orden. Definimos su función de autocovarianza por*

$$\gamma_{jn} = cov(Y_n, Y_{n+j}) = \mathbb{E}(Y_n - \mathbb{E}(Y_n))(Y_{n+j} - \mathbb{E}(Y_{n+j})),$$

Observación: En esta notación $var(Y_n) = \gamma_{0n}$.

Las series temporales de segundo orden débilmente estacionarias se caracterizan por sus momentos de segundo orden.

Definición 2.4. *Una serie de tiempo $\{Y_n\}_{n\in\mathbb{Z}}$ de segundo orden se dice débilmente estacionaria si:*

1. *$\mathbb{E}(Y_n) = \mu$ para todo $n \in \mathbb{Z}$.*
2. *$\gamma_{jn} = \gamma_{j0}$ para todo $n \in \mathbb{Z}$. En este caso denotamos $\gamma_{j0} = \gamma_j$.*

Para una serie de tiempo de segundo orden débilmente estacionaria se define la función de autocorrelación por

$$\rho_j = \gamma_j / \gamma_0.$$

Como una consecuencia de la desigualdad de Cauchy-Schwarz se demuestra que $|\rho_j| < 1$ para todo $j \in \mathbb{Z}$.

Ejemplo 1: (Proceso ruido blanco) Sea $\{e_n\}_{n\in\mathbb{Z}}$ una sucesión de variables aleatorias no correlacionadas, centradas $\mathbb{E}(e_n) = 0$ y con $\mathbb{E}(e_n^2) = \sigma^2$. El proceso ruido blanco es débilmente estacionario. Cuando la sucesión

es indepediente el proceso se conoce como *ruido blanco fuerte*.

Ejemplo 2: Definimos $Y_n = \mu + e_n$ donde $\{e_n\}_{n\in\mathbb{Z}}$ es un ruido blanco. Entonces $\{Y_n\}_{n\in\mathbb{Z}}$ es una serie de tiempo débilmente estacionaria.

Ejemplo 3: Definimos $Y_n = n\mu + e_n$ donde $\{e_n\}_{n\in\mathbb{Z}}$ es un ruido blanco. Entonces $\{Y_n\}_{n\in\mathbb{Z}}$ no es débilmente estacionario.

2.1. Autocorrelación parcial

El coeficiente de autocorrelación parcial de una serie de tiempo de segundo orden débilmente estacionaria en el retardo k, denotado por ψ_{kk}, es la correlación simple entre Y_n y Y_{n-k} después de extraer la influencia de los retardos intermedios.

El cálculo de las autocorrelaciones parciales puede basarse en el modelo de regresión múltiple en desviaciones respecto a las medias poblacionales

$$Y_n = \psi_{1k}Y_{n-1} + \cdots + \psi_{kk}Y_{n-k} + u_n$$

Multiplicando por Y_{n-k} y tomando esperanza,

$$\mathbb{E}(Y_{n-k}Y_n) = \psi_{1k}\mathbb{E}(Y_{n-k}Y_{n-1}) + \cdots + \psi_{kk}\mathbb{E}(Y_{n-k}Y_{n-k}) + \mathbb{E}(Y_{n-k}u_n)$$

y si suponemos que el proceso está centrado, $\mathbb{E}(Y_n) = 0$ para todo $n \in \mathbb{Z}$, entonces

$$\gamma_n = \psi_{1k}\gamma_{n-1} + \cdots + \psi_{kk}\gamma_{n-k}$$

dividiendo por γ_0 se obtiene el llamado sistema de ecuaciones de Yule-Walker, que queda determinado por

$$\begin{pmatrix} \rho_1 \\ \rho_2 \\ \vdots \\ \rho_k \end{pmatrix} = \begin{pmatrix} \rho_0 & \rho_1 & \cdots \rho_{k-1} \\ \rho_1 & \rho_0 & \cdots \rho_{k-2} \\ \vdots & & \\ \rho_{k-1} & \rho_{k-2} & \cdots \rho_0 \end{pmatrix} \begin{pmatrix} \psi_{1k} \\ \psi_{2k} \\ \vdots \\ \psi_{kk} \end{pmatrix}$$

de donde obtenemos las autocorrelaciones parciales en términos de las autocorrelaciones simples $\rho_0, \rho_1, \ldots, \rho_k$. Aplicando la regla de Cramer

nos queda

$$\psi_{kk} = \frac{\begin{vmatrix} 1 & \rho_1 & \cdots \rho_1 \\ \rho_1 & 1 & \cdots \rho_2 \\ \vdots & & \\ \rho_{k-1} & \rho_{k-2} & \cdots \rho_k \end{vmatrix}}{\begin{vmatrix} 1 & \rho_1 & \cdots \rho_{k-1} \\ \rho_1 & 1 & \cdots \rho_{k-2} \\ \vdots & & \\ \rho_{k-1} & \rho_{k-1} & \cdots 1 \end{vmatrix}}$$

2.2. Procesos de Medias Móviles

Sea $\{e_n\}_{n\in\mathbb{Z}}$ un proceso ruido blanco. Un proceso Media Móvil de orden q, (MA(q), siglas en inglés), es definido por la siguiente ecuación

$$Y_n = \mu + e_n + \theta_1 e_{n-1} + \cdots + \theta_q e_{n-q},$$

donde $\mu, \theta_1, \ldots, \theta_q$ son parámetros reales constantes.
Veamos el proceso de media móvil de primer orden, MA(1). Consideraremos

$$Y_n = \mu + e_n + \theta e_{n-1}$$

donde θ y μ son constantes reales cualesquiera. Se tiene que

1. $\mathbb{E}(Y_n) = \mu$.

2. $\gamma_{0n} = (1+\theta^2)\sigma^2$, $\gamma_{1n} = \theta\sigma^2$ y $\gamma_{jn} = 0$ para $j > 1$.

De esta manera el proceso MA(1) es débilmente estacionario.
Para el proceso MA(1) se tiene que la funcion de autocorrelación es dada por

$$\rho_1 = \frac{\theta}{(1+\theta)^2},$$

y $\rho_j = 0$ para $j > 0$.

2.3. Procesos Autorregresivos

Un proceso autorregresivo de orden p (AR(p)) se define por la ecuación

$$Y_n = \phi_1 Y_{n-1} + \cdots + \phi_p Y_{n-p} + e_n$$

donde e_n es un ruido blanco y c, ϕ_j, $j = 1, \ldots, p$ son constantes reales. Presentaremos a continuación los casos $p = 1$ y $p = 2$.

Un proceso autorregresivo de orden 1 (AR(1)) se define por la ecuación

$$Y_n = c + \phi Y_{n-1} + e_n \tag{2.1}$$

donde e_n es un ruido blanco y c, ϕ son constantes reales. La ecuación (2.1) es equivalente a

$$Y_n = \phi B(Y_n) + w_n$$

donde B es el operador de retardo (backward shift) y $w_n = c + e_n$. Esto permite dar la solución débilmente estacionaria que define el AR(1), en efecto, $(I - \phi B)(Y_n) = w_n$ y aplicando la inversa del operador $(I - \phi B)^{-1}$ la cual existe bajo la condición $|\phi| < 1$,

$$Y_n = \sum_{j=0}^{\infty} \phi^j w_{n-j}, \tag{2.2}$$

podemos calcular la esperanza y la varianza de Y_n, las cuales están dadas por $\mathbb{E}(Y_n) = c/(1 - \phi)$ y

$$\gamma_0 = \frac{\sigma^2}{1 - \phi^2},$$

respectivamente, mientras que la función de autocovarianza es

$$\gamma_j = \frac{\phi^j \sigma^2}{1 - \phi^2}.$$

En consecuencia es fácil ver que la función de autocorrelación es $\rho_j = \phi^j$.

Observación: la representación de la solución de la ecuación (2.2) se conoce una media móvil infinita MA(∞).

Si $|\phi| > 1$ la serie (2.2) no converge en L^2. Sin embargo reescribiendo (2.1) en término de los valores futuros de Y_n se tiene la siguiente ecuación autoregresiva

$$Y_n = \phi^{-1}Y_{n+1} - \phi^{-1}w_{n+1}. \tag{2.3}$$

Como $|\phi^{-1}| < 1$, por el mismo argumento utilizado anteriormente para el caso $|\phi| < 1$, se tiene que,

$$Y_n = -\sum_{j=1}^{\infty} \phi^{-j} w_{n+j}, \tag{2.4}$$

siendo esta la única solución estacionaria de (2.1) o equivalentemente, de (2.3).

Luego, su función de covarianza es dada por

$$\gamma_j = \frac{\sigma^2 \phi^{-|j|}}{\phi^2 - 1}, \tag{2.5}$$

Si $|\phi| = 1$, se puede probar que no existe una solución estacionaria. Por consiguiente, no existe un $AR(1)$ para $|\phi| = 1$.

A continuación, presentamos en la Figura 2.1, dos realizaciones del proceso autoregresivo AR(1), y sus correspondientes correlogramas. Ambas trayectorias han sido simuladas considerando la misma realización del ruido blanco $\{e_n\}$, por lo que sólo se diferencian en el valor del parámetro autoregresivo.

Para la primera realización del proceso AR(1) identificado con el color azul, se utilizó el parámetro autoregresivo $\phi = 0{,}65$ y $\sigma = 1$, y para la segunda silmulación identificada con el color rojo, el parámetro autoregresivo $\phi = 0{,}98$ y $\sigma = 1$, ambos para los tiempos $n = 0, \ldots, 1000$. Se puede observar que la primera trayectoria presenta un comportamiento con poca variabilidad y de esta manera una dispersión más estable, mientras que para la segunda realización el comportamiento es más explosivo y con una variabilidad mucho más fuerte.

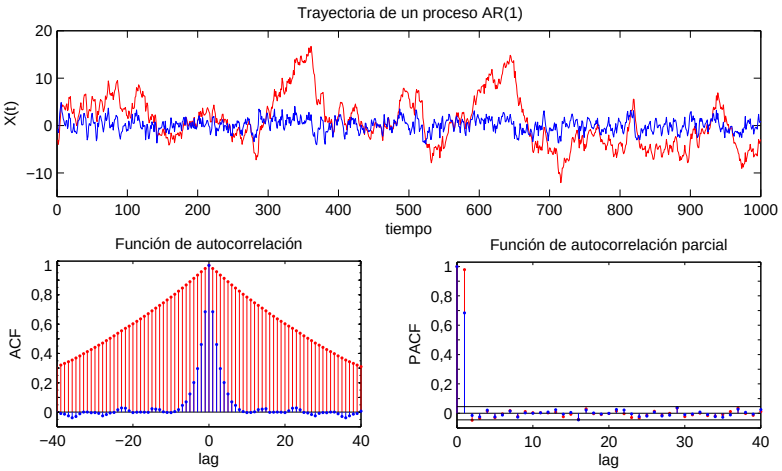

Figura 2.1: Simulaciónes del proceso AR(1)

También, se observa en la gráfica de la función de autocorrelación (ACF) que para ambos modelos se presenta un decrecimiento exponencial, característico de los modelos autoregresivos de coeficientes positivos, y finalmente a fin de distinguir entre las dos simulaciones para el modelo AR(1), presentamos la función de autocorrelación parcial (PACF); en el caso de un modelo AR(1), Y_n mantiene una relación directa con Y_{n-1}, por lo que el primer coeficiente de autocorrelación parcial es diferente de cero y el resto de los coeficientes iguales a cero.

Un proceso autorregresivo de orden 2, AR(2), se define por la ecuación,

$$Y_n = \phi_1 Y_{n-1} + \phi_2 Y_{n-2} + e_n$$

el estudio de este proceso lo establecemos considerando la representación equivalente del AR(2) como un AR(1) vectorial, definimos $Z_n = (Y_n, Y_{n-1})^T$, $E_n = (e_n, 0)^T$ y

$$A = \begin{pmatrix} \phi_1 & \phi_2 \\ 1 & 0 \end{pmatrix}$$

así

$$Z_n = AZ_{n-1} + E_n,$$

y la estacionaridad de este último depende del comportamiento del radio espectral $\rho(A)$ de la matriz A, es decir, si $\rho(A) < 1$ entonces

$$\lim_{n\to\infty} A^n = 0$$

y como $Z_n = A^n Z_0 + \sum_{k=0}^{n} A^k E_{n-k}$ cuando n tiende al infinito $Z_\infty \to \sum_{k=0}^{\infty} A^k E_k$ y la distribución de este proceso límite es la distribución estacionaria.

El radio espectral de A es el autovalor de mayor modulo, el cual en este caso esta dado por las raíces del polinómio $\lambda^2 - \phi_1\lambda - \phi_2$. En general para un proceso AR(p) se puede estudiar su estacionaridad si se estudia el polinomio característico asociado.

2.4. Proceso ARMA(p,q)

El proceso ARMA(p,q) (autorregresivo media Movil de orden p, q)

$$(I - \phi_1 B - \cdots - \phi_p B^p)Y_n = (I - \theta_1 B - \cdots - \theta_q B^q)e_n$$

El proceso será estacionario si las raíces de $1 - \phi_1\lambda - \cdots \phi_p\lambda^p$ están fuera del circulo unitario y es invertible si las de $1 - \theta_1\lambda - \cdots - \theta_q\lambda^q$ lo están.

2.5. Paseo al azar

Consideremos a manera introductoria la siguiente sucesión de variables aleatorias. Sea $S_n = X_1 + X_2 + \cdot + X_n$, con $S_0 = 0$ y X_i variables aleatorias Bernoulli independientes con valores $-1, 1$ y $\mathbb{P}(X_k = -1) = \mathbb{P}(X_k = 1) = 1/2$. Si consideramos las poligonales (n, S_n) que parten del origen, estas definen el paseo al azar o caminata aleatoria.

Observemos que el proceso S_n satisface la ecuación,

$$S_n = S_{n-1} + X_n$$

es decir, una caminata al azar se puede ver como un proceso autorregresivo con $\phi = 1$ y ruido X_n Bernoulli. En general llamaremos paseo al azar a un proceso AR(1) con $\phi = 1$. Este proceso no es estacionario

pero el proceso de sus incrementos $S_n - S_{n-1}$ si lo es.

Definiendo el operador $\nabla = I - B$, los incrementos del paseo al azar se escriben como $\nabla S_n = X_n$.

La varianza de S_n es igual a $\sigma^2 n$, su $cov(S_n, S_{n+h}) = \sigma^2 n$ y la función de autocorrelación

$$\frac{n}{\sqrt{n(n+k)}}.$$

si n es grande, los coeficientes de la función de autocorrelación serán próximos a uno y decrecerán muy lentamente en k.

2.6. ARIMA(p, d, q)

El paseo al azar se obtuvo admitiendo que la raíz de un AR(1) es unitaria, con lo que se convierte en no estacionario. Esta idea se puede generalizar para cualquier proceso ARMA permitiendo una o varias raíces unitarias en el operador AR. Se obtienen entonces procesos del tipo:

$$(I - \phi_1 B - \cdots - \phi_p B^p)(I - B)^d Y_n = (I - \theta_1 B - \cdots - \theta_q B^q) e_n$$

este es un proceso ARIMA(p, d, q). En esta notación p es el orden de la parte autorregresiva, d es el número de raíces unitarias y q el orden de la parte media móvil.

2.7. Metodología de Box y Jenkins

En lo que sigue presentamos una metodología utilizada para ajustar un modelo concreto a una serie de datos particular. Las tres etapas del modelado son las siguientes:

1. Identificación y selección del modelo: asegurarse de que las variables son estacionarias, la identificación de la estacionalidad de la serie (diferenciando en el tiempo si es necesario), y el uso de los gráficos de las funciones de autocorrelación y de autocorrelación parcial de la serie de tiempo para decidir cuál componente (si es el caso) se debe utilizar en el modelo, el promedio autorregresivo o una media móvil.

2. Estimación de parámetros usando algoritmos de cálculo para tener coeficientes que mejor se ajustan al modelo ARIMA seleccionado. Los métodos más comunes cuando se se supone al ruido blanco e gaussiano indepediente son estimación por máxima verosimilitud o mínimos cuadrados.

3. Comprobar el modelo mediante el ensayo, es decir si el modelo estimado se ajusta a las especificaciones de un proceso univariado estacionario. En particular, los residuos deben ser independientes, la media y la varianza constantes en el tiempo.

2.8. Ejercicios

1. Un proceso Y_n es estrictamente estacionario si es invariante por traslaciones, la distribución de $(X_{t_1}, \ldots, X_{t_n})$ es la misma que $(X_{t_1+h}, \ldots, X_{t_n+h})$ para todo $(t_1, \ldots, t_n)$ y todo h.

   a) Sea $\{Y_n\}$ un proceso estocástico con $Z_n = (Y_{2n-1}, Y_{2n}) \sim$ i.i.d $\mathcal{N}(0, \Sigma)$, con $\Sigma = \begin{pmatrix} 1 & \sigma \\ \sigma & 2 \end{pmatrix}$ demuestre que $Y_n \sim \mathcal{N}(0, 1)$ para todo n pero las distribución de (Y_1, Y_2) no es igual a la de (Y_2, Y_3) porque $(Y_2, Y_3) \sim \mathcal{N}(0, I)$.

   b) Demuestre que si $\{Y_n\}$ es un proceso estrictamente estacionario con $\mathbb{E}(Y_n^2)$ para todo n entonces $\{Y_n\}$ es un proceso débilmente estacionario.

2. Calcular la media, varianza y función de covarianza de un proceso MA(q). Concluya que el proceso MA(q) es débilmente estacionario.

3. Demuestre que si B es el operador de retardo entonces el operador B es acotado y el operador $I - \phi B$ es invertible si $|\phi| < 1$, utilice la norma del supremo.

4. Calcular la media, varianza y función de covarianza de un proceso AR(1). Concluya que el proceso AR(1) es un proceso débilmente estacionario.

5. Sean los procesos $\{V_n\}$ y $\{U_n\}$ definidos por $V_n = e_1 + \phi e_2 + \cdots + \phi^{n-1} e_n$ y $U_n = \phi^{n-1} e_1 + \phi^{n-2} e_2 + \cdots + e_n$ respectivamente. Sea $\{Y_n\}$ el proceso definido por la ecuación (2.1) (con $c = 0$). Demuestre que U_n y V_n tienen la misma distribución y que si $|\phi| < 1$, $U_n - Y_n \to 0$ c.s. y $V_n \to \sum_{n \geq 0} \phi^{n-1} e_n$ c.s. entonces $Y_n \to \sum_n \phi^{n-1} e_n$ en distribución.

6. Demuestre que la función de autocorrelación parcial de un proceso AR(1) se anula para retardos k mayor que 1.

7. Demuestre que la función de autocorrelación parcial de un proceso MA(1) es

$$\psi_{kk} = \frac{-\theta^k (1 - \theta^2)}{1 - \theta^{2(k+1)}}.$$

8. Suponga que $Y_n = \phi_1 Y_{n-1} + u_n$ y $u_n = \phi_2 u_{n-1} + a_n$ donde a_t es un proceso ruido blanco. Demuestre que Y_n es un proceso AR(2).

9. Suponga que $Y_n = \phi_1 Y_{n-1} + u_n$ y $u_n = a_n - \theta_2 a_{n-1}$ donde a_t es un proceso ruido blanco. Demuestre que Y_n es un proceso AR(1,1).

10. Explique detallamente las propiedades de esperanza, varianza, covarianza y autocorrelaciones parciales de un proceso ARMA(1,1).

11. Realice simulaciones de MA(1), MA(2) y AR(1). Compare las funciones de autocorrelación muestral con las autocorrelaciones teóricas.

Capítulo 3

Cadenas de Markov

En este capítulo introducimos ideas básicas relacionadas con Cadenas de Markov a tiempo discreto definidas sobre un espacio de probabilidad $(\Omega, \mathcal{F}, \mathbb{P})$ a valores en un espacio medible $(E, \mathcal{E})$, el conjunto E denota el espacio de estados y $\mathcal{E}$ una σ-álgebra. El capítulo sigue de cerca el enfoque de Prieto [39] y Duflo [20].

3.1. Núcleos de Markov

Un *núcleo de transición* sobre el espacio medible $(E, \mathcal{E})$, es una función $Q : E \times \mathcal{E} \to [0, 1]$ tal que,

1. Para cada $x \in E$, $Q(x, \cdot)$ es una medida de probabilidad sobre $(E, \mathcal{E})$.

2. Para cada $\Gamma \in \mathcal{E}$, $Q(\cdot, \Gamma)$ es una variable aleatoria.

Una *filtración* sobre un espacio medible $(\Omega, \mathcal{F})$ es una familia creciente $\{\mathcal{F}_n\}_{n\geq 0}$ de sub-σ-álgebras de $\mathcal{F}$, ($\mathcal{F}_n \subseteq \mathcal{F}_{n+1}$, para cada $n \geq 0$).

Dada una sucesión de variables aleatorias (v.a) $\{X_n\}_{n\geq 0}$, la *filtración natural* se define como la menor σ-álgebra, tal que $X_0, ..., X_n$ son Borel-medibles, esta se denota como $\mathcal{F}_n = \sigma(X_0, ..., X_n)$.

Una sucesión de v.a $\{X_n\}_{n\geq 0}$ es adaptada a la filtración $\{\mathcal{F}_n\}_{n\geq 0}$ si para cada $n \in \mathbb{N}$, la variable aleatoria X_n es $\mathcal{F}_n$-medible ($X_n \in \mathcal{F}_n$).

Una sucesión de v.a $X = \{X_n\}_{n\geq 0}$ sobre el espacio de probabilidad $(\Omega, \mathcal{F}, \mathbb{P})$ adaptadas a una filtración $\{\mathcal{F}_n\}_{n\geq 0}$, con valores en el espacio medible $(E, \mathcal{E})$ es una *Cadena de Markov homogénea con núcleo de transición* Q si y sólo si para cada $n \in \mathbb{N}$,

$$\mathbb{P}(X_{n+1} \in A|\mathcal{F}_n) = Q(X_n, A), \tag{3.1}$$

para cada $A \in \mathcal{E}$. La distribución de X_{n+1} condicional a $\mathcal{F}_n$ es la medida aleatoria $Q(X_n, \cdot)$.

La distribución de la v.a X_0 se llama ley inicial. Asociaremos a Q la transición en n pasos Q^n definida recursivamente por:

$$\begin{aligned} Q^0(x,\cdot) &= \delta_x(\cdot), \\ Q^1(x,\cdot) &= Q(x,\cdot), \\ &\vdots \\ Q^n(x,\cdot) &= \int Q(y,\cdot)Q^{n-1}(x,dy). \end{aligned}$$

donde δ_x es la delta de Dirac, esto es,

$$\delta_x(A) = \begin{cases} 1 & \text{si} \quad x \in A, \\ 0 & \text{si no.} \end{cases}$$

Si $B \in \mathcal{E}$ y ν es una medida sobre $\mathbb{R}$,

$$\nu Q^n(B) = \int Q^n(x, B)\nu(dx). \tag{3.2}$$

Además, si g es una función $\mathcal{E}$-medible y acotada en E, denotaremos,

$$Q^n(g(x)) = \int g(y)Q^n(x, dy). \tag{3.3}$$

A continuación presentamos algunos ejemplos de Cadenas de Markov, considerando siempre la filtración natural.

Ejemplo 3.1. ***Caso numerable.***
Sea E un conjunto numerable y $q : E \times E \to [0,1]$ una función tal que, para todo $i \in E$, $\sum_{j\in E} q(i,j) = 1$.

Sea $\{X_n\}_{n\geq 0}$ una sucesión de variables aleatorias con valores en E tal que, para todo n y todo vector $(i_0, ..., i_n)$ de E^{n+1} se tiene:

$$\mathbb{P}(X_n = i_n | X_{n-1} = i_{n-1}, ..., X_0 = i_0) = q(i_n, i_{n-1}),$$

entonces $\{X_n\}_{n\geq 0}$ define una Cadena de Markov con núcleo $Q(i,A) = \sum_{j\in A} q(i,j)$ y espacio de estados $(E, \mathcal{P}(E))$, donde $\mathcal{P}(E)$ denota el conjunto de las partes de E.

Cuando E tiene cardinal finito, $E = \{1, ..., m\}$, denotamos por q la matriz de transición en un paso, con entradas $q(i,j)$ y q^n es la matriz de transición en n pasos, con entradas $q^n(i,j)$. Se puede probar usando la propiedad de probabilidad total que:

$$q^{n+l}(i,j) = \sum_{h=1}^{m} q^n(i,h) q^l(h,j),$$

esta propiedad se conoce comoEcuación de Chapman-Kolmogorov.

Ejemplo 3.2. ***Caso continuo con densidad.***
Sea $E \subseteq \mathbb{R}^d$ y $q : E \times E \longrightarrow \mathbb{R}^+$ una función de densidad positiva en la segunda variable, es decir, $q(x,y) > 0$, para todo $y \in \mathbb{R}$ y $\int_{\mathbb{R}^d} q(x,y)dy = 1$. Entonces $Q(x,A) = \int_A q(x,y)dy$ define un núcleo de transición sobre $(E, \mathcal{E})$ con $\mathcal{E} = \mathcal{B}(E)$ la σ-álgebra de Borel. Llamaremos a la función q la densidad de transición.

Suponga que la densidad de transición es una gaussiana con varianza 1, es decir, $q(x,y) = e^{-\frac{1}{2}(y-x)^2}/\sqrt{2\pi}$, entonces,

$$Q(x,A) = \frac{1}{\sqrt{2\pi}} \int_A e^{-\frac{1}{2}(y-x)^2} dy. \tag{3.4}$$

Ahora, calculemos Q^2 (la segunda iteración de Q),

$$\begin{aligned} Q^2(x,A) &= \int Q(y,A) Q(x,dy), \\ &= \int \int_A q(y,z) q(x,y) dz\, dy, \end{aligned}$$

usando el teorema de Fubini, se tiene,

$$\begin{aligned} Q^2(x,A) &= \int_A \int q(y,z)q(x,y)dy\ dz, \\ &= \int_A \int \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}(z-y)^2} \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}(y-x)^2} dy\ dz, \\ &= \frac{1}{2\pi} \int_A \int e^{-\frac{1}{2}[(z-y)^2+(y-x)^2]} dy\ dz, \end{aligned}$$

completando cuadrados, obtenemos,

$$\begin{aligned} Q^2(x,A) &= \frac{1}{2\pi} \int_A \int e^{-\frac{1}{2}[(z-y+y-x)^2-2(z-y)(y-x)]} dy\ dz, \\ &= \frac{1}{2\pi} \int_A \int e^{-\frac{1}{2}(z-x)^2+yx-y^2-zx+zy} dy\ dz, \\ &= \frac{1}{2\pi} \int_A \int e^{-\frac{1}{2}(z-x)^2} e^{yx-y^2-zx+zy} dy\ dz, \\ &= \frac{1}{2\pi} \int_A e^{-\frac{1}{2}(z-x)^2} e^{-zx} \int e^{y(x+z)-y^2} dy\ dz, \\ &= \frac{1}{2\pi} \int_A e^{-\frac{1}{2}(z-x)^2-zx} \sqrt{\pi} e^{\frac{1}{4}(z+x)^2} dz, \\ &= \frac{1}{2\sqrt{\pi}} \int_A e^{-\frac{1}{4}(z-x)^2} dz. \end{aligned}$$

Note que la función de densidad correspondiente es una normal de media cero y varianza 2.

■

Ejemplo 3.3. ***Modelo iterativo.***
Existe una manera iterativa de construir Cadenas de Markov. Sea $(G, \mathcal{G})$ un espacio medible, $\{e_n\}_{n\geq 0}$ una sucesión de variables aleatorias (v.a) independientes idénticamente distribuidas (i.i.d) sobre G. Sea $E \times G$ dotado de la σ-álgebra producto $\mathcal{E} \bigotimes \mathcal{G}$. Consideremos una función medible $F : E \times G \to E$ entonces el proceso definido por la ecuación,

$$X_{n+1} = F(X_n, e_{n+1}), \tag{3.5}$$

lo llamamos modelo iterativo. Si el estado inicial X_0 es independiente de $\{e_n\}_{n\geq 0}$ y definimos

$$Q(x,A) := \mathbb{P}(F(x,e) \in A),$$

entonces Q es un núcleo de transición y $\{X_n\}_{n\geq 0}$ definida por la ecuación (3.5) es una Cadena de Markov adaptada a la filtración natural $\mathcal{F}_n = \sigma(X_0, ..., X_n)$. En efecto, para todo $n > 0$

$$\begin{aligned}\mathbb{P}(X_{n+1} \in A|\mathcal{F}_n) &= \mathbb{P}(F(X_n, e_{n+1}) \in A|\mathcal{F}_n)\\ &= \mathbb{P}(F(X_n, e) \in A|X_n)\\ &= \mathbb{P}(F(x, e) \in A|X_n = x)\\ &= Q(X_n, A).\end{aligned}$$

■

Una cadena de Markov en general se puede escribir como un modelo iterativo. El siguiente lema demuestra que en particular toda cadena de Markov con valores en un espacio de estados finito es un modelo iterativo.

Lema 3.4. *Toda cadena de Markov sobre un conjunto finito de estados E es un modelo iterativo.*

Demostración: Supongamos que $\{X_n\}_{n\geq 1}$ es una cadena de Markov sobre el conjunto de estados $E = \{1, 2\}$ y con matriz de transición

$$P = \begin{pmatrix} 1-p & p \\ q & 1-q \end{pmatrix}.$$

Sea $\{U_n\}_{n\geq 1}$ una sucesión independiente de variables aleatorias uniformes sobre $[0, 1]$ y definamos la función $F : E \times [0, 1] \to E$ por

$$F(1,u) = \begin{cases} 1 & \text{si } p \leq u \leq 1 \\ \\ 2 & \text{si } 0 \leq u < p \end{cases} \qquad F(2,u) = \begin{cases} 1 & \text{si } 0 \leq u < q \\ \\ 2 & \text{si } q \leq u \leq 1 \end{cases}$$

por lo tanto la cadena de Markov $\{\zeta_n\}_{n\geq 1}$ definida como sistema iterativo por

$$\zeta_n = F(\zeta_{n-1}, U_n)$$

y $\{X_n\}_{n\geq 1}$ tienen la misma matriz de transición, en efecto

$$\mathbb{P}(\zeta_n = i|\zeta_{n-1} = 1) = \begin{cases} \mathbb{P}(F(1, U_n) = 1) = 1 - p & \text{si } i = 1 \\ \mathbb{P}(F(1, U_n) = 2)) = p & \text{si } i = 2 \end{cases}$$

y

$$\mathbb{P}(\zeta_n = i|\zeta_{n-1} = 2) = \begin{cases} \mathbb{P}(F(2, U_n) = 1) = q & \text{si } i = 1 \\ \mathbb{P}(F(2, U_n) = 2) = 1 - q & \text{si } i = 2 \end{cases}$$

por lo tanto se obtiene el lema.

■

Un caso particular del modelo iterativo, son los *procesos autorregresivos*, el cual desarrollamos en el siguiente ejemplo.

Ejemplo 3.5. ***Procesos Autorregresivos.***
Recordemos que un proceso autorregresivo de primer orden en $\mathbb{R}$*, se define por,*

$$X_{n+1} = \rho X_n + e_{n+1}, \tag{3.6}$$

donde ρ *es un número real y* $\{e_n\}_{n\geq 0}$ *es una sucesión de v.a. i.i.d. Retomando el ejemplo del modelo iterativo (3.5), en este caso, tenemos que* $F(X_n, e_{n+1}) = \rho X_n + e_{n+1}$.

Si Γ *es la medida de probabilidad de la sucesión de ruidos* $\{e_n\}_{n\geq 0}$*, el núcleo de transición de la Cadena de Markov* $\{X_n\}_{n\geq 0}$ *está dado por,*

$$\begin{aligned} Q(x, A) &= \mathbb{P}(F(x, e) \in A), \\ &= \mathbb{P}(\rho x + e \in A), \\ &= \mathbb{P}(e \in A - \rho x), \\ &= \int_{A-\rho x} \Gamma(z)dz. \end{aligned}$$

Luego, si $z \in A - \rho x$*, entonces* z *es de la forma* $z = u - \rho x$*, donde* $u \in A$*. Por lo tanto,* $u = z + \rho x$*, entonces* $Q(x, A) = \int_A \Gamma(u - \rho x)du$*. De ahí se tiene que,*

$$Q(x, A) = \Gamma(A - \rho x). \tag{3.7}$$

Luego,

$$\begin{aligned} Q^n(x,A) &= \int Q(y,A)Q^{n-1}(x,dy) \\ &= \int\int_A \Gamma(u-\rho y)du\Gamma^{n-1}(y-\rho^{n-1}x)dy \\ &= \int_A\int \Gamma(u-\rho y)\Gamma^{n-1}(y-\rho^{n-1}x)dy\ du \\ &= \int_A\int \Gamma(u-\rho^n x-\rho v)\Gamma^{n-1}(v)dv\ du \\ &= \int_A \Gamma^n(u-\rho^n x)du \\ &= \Gamma^n(A-\rho^n x), \end{aligned}$$

donde $\Gamma^n(u) = \int \Gamma(u-\rho v)\Gamma^{n-1}(v)dv$ *para todo* $n \geq 2$.

■

3.2. Medida Invariante

En esta sección introduciremos el concepto de medida invariante, ϕ- irreductibilidad y algunos otros tópicos relacionados con esta medida, que serán importantes a lo largo de este trabajo.

Definición 3.6. *Sea* $(E,\mathcal{E})$ *un espacio medible y* Q *un núcleo de transición de* $(E,\mathcal{E})$ *en si mismo. Una medida* π *positiva se dice invariante con respecto a* Q *si* $\pi Q = \pi$*; ella es excesiva si* $\pi Q \leq \pi$.

Ejemplo 3.7. ***Procesos Autorregresivos (Continuación).***
Para ilustrar la definición anterior calcularemos la medida invariante del ***AR(1)*** *en su forma general y con ruido gaussiano, luego usaremos este resultado para obtenerla como en el ejemplo 3.5.*

Sea,

$$X_{n+1} = \rho X_n + \beta + \gamma^{1/2}e_{n+1}, \tag{3.8}$$

con $|\rho| < 1$. *Calculemos la transformada de Fourier en ambos lados para* $n = 0$,

$$X_1 = \rho X_0 + \beta + \gamma^{1/2}e_1,$$

si e_1 *es independiente de* X_0*, entonces,*

$$\mathbb{E}(e^{itX_1}) = \mathbb{E}(e^{\rho itX_0})\mathbb{E}(e^{it\gamma^{1/2}e_1})e^{it\beta},$$

como $e_1 \sim N(0,1)$*,*

$$\mathbb{E}(e^{itX_1}) = \mathbb{E}(e^{\rho itX_0})e^{it\beta}e^{-\gamma\frac{t^2}{2}},$$

como queremos que la medida sea invariante si tomamos $\psi(t) = \mathbb{E}(e^{itX_1})$ *tenemos,*

$$\begin{aligned}
\psi(t) &= \psi(\rho t)e^{it\beta}e^{-\gamma\frac{t^2}{2}}, \\
\psi(t) &= \psi(\rho t)e^{\frac{it\beta}{1-\rho}}e^{\frac{-\rho it\beta}{1-\rho}}e^{\frac{\gamma t^2\rho^2}{2(1-\rho^2)}}e^{\frac{-\gamma t^2}{2(1-\rho^2)}}, \\
\frac{\psi(t)}{e^{\frac{it\beta}{1-\rho}}e^{\frac{-\gamma t^2}{2(1-\rho^2)}}} &= \frac{\psi(\rho t)}{e^{\frac{-\rho it\beta}{1-\rho}}e^{\frac{\gamma t^2\rho^2}{2(1-\rho^2)}}},
\end{aligned}$$

Sea, $g(t) = \frac{\psi(t)}{e^{\frac{it\beta}{1-\rho}}e^{\frac{-\gamma t^2}{2(1-\rho^2)}}}$*, entonces,*

$$g(t) = g(\rho t),$$

como g *es continua y derivable con primera derivada* g' *continua, entonces*

$$g'(t) = \rho g'(\rho t),$$

ahora, elegimos $f(t) = g'(t)$*, luego,*

$$\begin{aligned}
f(t) &= \rho f(\rho t), \text{ por inducción} \\
f(t) &= \rho^n f(\rho^n t),
\end{aligned}$$

tomando el límite cuando $n \to \infty$*, nos queda,*

$$\begin{aligned}
f(t) &= \lim_{n\to\infty} \rho^n f(\rho^n t), \text{por la continuidad de} f \\
f(t) &= \lim_{n\to\infty} \rho^n f(\lim_{n\to\infty} \rho^n t), \text{luego, } |\rho| < 1, \\
f(t) &= \lim_{n\to\infty} \rho^n f(0), \\
f(t) &= 0f(0) = 0,
\end{aligned}$$

luego, como $f(t) = 0$ *para todo* t, *entonces* $g'(t) = 0$, *por lo tanto,* $g(t) = c$, *donde* c *es una constante.*

Pero $g(0) = \frac{\psi(0)}{e^0 e^0} = \psi(0) = 1$, *así,* $c = 1$.
Por otra parte,

$$1 = g(t) = \frac{\psi(t)}{e^{\frac{it\beta}{1-\rho}} e^{\frac{-\gamma t^2}{2(1-\rho^2)}}},$$

de donde,

$$\begin{aligned} \psi(t) &= e^{\frac{it\beta}{1-\rho}} e^{\frac{-\gamma t^2}{2(1-\rho^2)}}, \\ \psi(t) &= e^{\frac{it\beta}{1-\rho} - \frac{-\gamma t^2}{2(1-\rho^2)}}. \end{aligned}$$

Concluimos que la medida invariante es $N\left(\frac{\beta}{1-\rho}, \frac{\gamma}{1-\rho^2}\right)$.

En el ejemplo 3.5, $\beta = 0$ *y* $\gamma = 1$, *se tiene que la medida invariante es* $N\left(0, \frac{1}{1-\rho^2}\right)$. ■

Cuando se hace referencia a Cadenas de Markov con núcleo de transición sobre espacios de estados generales, es importante definir el primer tiempo de entrada y el primer tiempo de retorno a un conjunto del espacio de estado.

Sea $\{X_n\}_{n\geq 0}$ una Cadena de Markov con núcleo de transición Q sobre $(E, \mathcal{E})$, para cualquier conjunto $A \in \mathcal{E}$, se define,

$$\begin{aligned} \sigma_A &= \text{ínf}\{n \geq 0 : X_n \in A\}, \\ \tau_A &= \text{ínf}\{n \geq 1 : X_n \in A\}, \end{aligned}$$

donde σ_A y τ_A son el primer tiempo de entrada y el primer tiempo de retorno al conjunto A, respectivamente.

Otras definiciones importantes para desarrollar la teoría sobre espacios de estados generales son,

Definición 3.8. *Un conjunto* $A \in \mathcal{E}$ *se dice que es accesible para el núcleo de transición* Q *(ó* Q*-accesible), si* $\mathbb{P}_x(\tau_A < \infty) > 0$ *para todo* $x \in E$.

Definición 3.9. *Una Cadena de Markov $\{X_n\}_{n\geq 0}$, con núcleo de transición Q, se dice ϕ-irreductible (ϕ es una medida) si para todo $x \in E$ y para todo $A \in \mathcal{E}$, con $\phi(A) > 0$, existe un $n \in \mathbb{N}$ tal que $Q^n(x, A) > 0$. En otras palabras, $\{X_n\}_{n\geq 0}$, es ϕ-irreductible si algún conjunto con ϕ-medida positiva es accesible a cualquier punto en el espacio de estado. Esta medida es llamada medida de irreductibilidad de Q.*

En general, hay muchas medidas de irreductibilidad; dos medidas de irreductibilidad no son necesariamente equivalentes. Sin embargo, se puede demostrar que existe una *medida de irreductibilidad máxima* ψ, tal que cualquier medida de irreductibilidad ϕ es absolutamente continua con respecto a ψ, ver [47].

Teorema 3.10. *Sea Q un núcleo de transición ϕ-irreductible sobre $(E, \mathcal{E})$. Entonces, existe una medida de irreductibilidad ψ tal que todas las medidas de irreductibilidad son absolutamente continuas con respecto a ψ y para todo $A \in \mathcal{E}$,*

$$\mathbb{P}_x(\tau_A < \infty) > 0 \Leftrightarrow \psi(A) > 0 \qquad \text{para todo } x \in E.$$

Vea la demostración de este teorema en [12] Cappé, Moulines y Rydén (2005).

El teorema anterior implica que si ψ es una medida de irreductibilidad, entonces A es accesible si y sólo si $\psi(A) > 0$.

Ejemplo 3.11. ***Procesos Autorregresivos (Continuación).*** *Retomando el modelo autorregresivo del ejemplo 3.5, decimos que es ϕ-irreductible siempre que la medida de los ruidos tengan una densidad positiva con respecto a la medida de Lebesgue λ. Si tomamos $\phi = \lambda$, la definición del núcleo de transición (3.7) se tiene,*

$$\begin{aligned} Q(x, A) &= \Gamma(A - \rho x), \\ &= \mathbb{P}(e_1 \in A - \rho x). \end{aligned}$$

Sea γ la densidad positiva del ruido respecto a la medida de Lebesgue λ y supongamos $\lambda(A) > 0$. Si $\mathbb{P}(e_1 \in A - \rho x) = 0$ podemos afirmar que,

$$\int_A \gamma(u - \rho x)du = \int \mathbb{1}_A(u)\gamma(u - \rho x)du = 0.$$

Como $\gamma > 0$, se tiene que $\mathbb{1}_A(u)\gamma(u - \rho x)du = 0$ c.s-λ, en consecuencia $\lambda(A) = 0$ y se obtiene una contradicción. Por lo tanto,

$$Q(x, A) > 0.$$

En consecuencia la cadena es ϕ-irreductible en un sólo paso. ■

Una Cadena de Markov se dice *ergódica*, si las iteraciones del núcleo de transición convergen a la medida invariante.

Ahora bien, retomando el ejemplo 3.2, demostraremos que si $Q(x, \cdot)$ admite densidad con respecto a la medida de Lebesgue, es decir, que $Q(x, A) = \int_A q(x, y)dy$ y π es invariante para Q entonces π admite densidad con respecto a la medida de Lebesgue, para ello veamos la siguiente proposición.

Proposición 3.12. *Sea $E = \mathbb{R}^d$, existe una función boreliana q sobre $\mathbb{R}^{2d}$ tal que $Q(x, dy) = q(x, y)dy$, entonces toda medida invariante π tiene una función de densidad h que satisface*

$$\int h(x)q(x, y)dx = h(y). \tag{3.9}$$

En efecto, para toda función boreliana positiva g

$$\int g(y)\pi(dy) = \int \pi(dx) \int q(x, y)g(y)dy = \int \left(\int \pi(dx)q(x, y) \right) g(y)dy.$$

Una Cadena de Markov con núcleo de transición Q satisface la *condición de reversibilidad*, si existe una función f tal que,

$$Q(y, x)f(y) = Q(x, y)f(x), \tag{3.10}$$

para todo (x, y).

Teorema 3.13. *Suponga que una Cadena de Markov con núcleo de transición Q satisface la condición de reversibilidad, con f una función de densidad de probabilidad. Entonces, la densidad f es la densidad invariante de la cadena.*

Sea Q como en el ejemplo 3.2, para algún conjunto medible B, usando la ecuación (3.2), se tiene,

$$\pi Q(B) = \int Q(x, B)\pi(dx),$$

luego, como Q y π tienen densidad respecto a la medida de Lebesgue,

$$\begin{aligned} \int Q(x, B)\pi(dx) &= \int \int_B q(x, y)dy f(x)dx, \\ &= \int \int_B q(x, y)f(x)dy\, dx, \end{aligned}$$

usando la condición de reversibilidad (3.10),

$$\int \int_B q(x, y)f(x)dy\, dx = \int \int_B q(y, x)f(y)dy\, dx,$$

dado que q es una función de densidad sabemos que $\int q(y, x)dx = 1$. Además q y f son positivas, aplicando el teorema de Fubini tenemos que,

$$\begin{aligned} \int \int_B q(y, x)f(y)dy\, dx &= \int_B \left(\int q(y, x)f(y)dx \right) dy, \\ &= \int_B f(y) \left(\int q(y, x)dx \right) dy, \\ &= \int_B f(y)dy = \pi(B), \end{aligned}$$

En la siguiente proposición se dan algunas propiedades de la medida invariante.

Proposición 3.14. *Son válidas las siguientes proposiciones:*

1. *Una probabilidad excesiva es invariante.*
2. *Sean π y ν dos medidas invariantes; la parte absolutamente continua y la parte singular de ν con respecto a π son invariantes.*
3. *Si la cadena es ϕ-irreductible, entonces tiene una única medida invariante.*

Veamos la prueba de cada una de las proposiciones.

1. Sea π una probabilidad excesiva. Para todo $B \in \mathcal{E}$

$$\pi Q(B) \leq \pi(B), \quad 1 - \pi Q(B) = \pi Q(B^c) \leq \pi(B^c) = 1 - \pi(B).$$

entonces $\pi Q = \pi$.

2. Sean π y ν dos medidas invariantes. Por la descomposición de Lebesgue:
$\nu = h\pi + 1\!\!1_N \nu$, donde $h\pi$ es absolutamente continua respecto a π y $1\!\!1_N \nu$ es la parte singular

$$0 = \pi(N) = \int Q(x, N) d\pi(x).$$

Además, para todo boreliano B,

$$\begin{aligned} \int Q(x, B) h(x) d\pi(x) &= \int Q(x, B \cap N^c) h(x) d\pi(x) \\ &\leq \nu(B \cap N^c) = (h\pi)(B). \end{aligned}$$

La medida $h\pi$ se dice *acotada* y *excesiva*, entonces es invariante. Para la diferencia, se sigue lo mismo sobre $1\!\!1_N \nu$.

3. Sobre la hipótesis de irreductibilidad, ϕ es absolutamente continua con respecto a todas las probabilidades invariantes; de acuerdo a la parte (2.) estas son equivalentes. Sean ν y π dos probabilidades invariantes, $\nu = h\pi$.

Sea $a > 0$, la medida $\nu_a = \text{ínf}(h, a)\pi$ es excesiva ya que:

$$\nu_a Q \leq a\pi Q = a\pi \quad y \quad \nu_a Q \leq (h\pi)Q = \nu = h\pi.$$

Así, $\nu_a \leq \nu$, por lo tanto, ν_a es invariante, es nulo o equivalente a π.

Ahora bien, $(a\nu - \nu_a) = (a - h) + \pi$ es también invariante, nula o equivalente a π. Para cada a, $\pi(h < a)$ es 0 ó 1, luego $h = ctte.$, π-c.s y $\nu = \pi$.

Una cadena de Markov $\{X_n\}_{n\geq 0}$ es Harris recurrente, si cada vez que $\phi(A) > 0$, se tiene $\mathbb{P}_{X_0}(\sum_{n=1}^{\infty} \mathbb{1}_A(X_n) = \infty) = 1$ para cualquier valor inicial X_0 de la cadena de Markov. En este caso, existe una única (salvo constante multiplicativa) medida invariante. Si además esta medida es finita, el proceso se llama Harris recurrente positivo.

Una cadena de Markov con núcleo Q es felleriana si, para toda función g continua y acotada, la función Qg definida por

$$Qg(x) := \int g(y)Q(x, dy)$$

es continua. Una cadena de Markov es fuertemente felleriana si para toda función medible y acotada g la función Qg es continua. En particular el modelo iterativo es felleriano si F es una función continua y es fuertemente felleriano cuando la sucesión $\{u_n\}_{n\geq 0}$ tiene densidad con respecto a la medida de Lebesgue, ver Duflo [20], pág 19.

3.3. El método de estabilidad de Lyapunov

Una función de Lyapunov es una función continua $V : \mathbb{R}^d \to \mathbb{R}^+$ tal que el $lim_{\|x\|\to\infty} V(x) = \infty$. En particular en lo que sigue elegiremos $V(x) = \|x\|^q$ con $q > 1$.

Lema 3.15 (Estabilización de orden p)**.** *Supongamos $\{Z_n\}_{n>-p}$ y $\{e_n\}_{n\geq 0}$ son dos sucesiones de números reales positivos, tal que para p constantes positivas $\alpha_1, \ldots, \alpha_p$,*

$$Z_n \leq \alpha_1 Z_{n-1} + \cdots + \alpha_p Z_{n-p} + e_n.$$

Supongamos que $\alpha_1 + \cdots + \alpha_p < 1$. Entonces existe $0 < \alpha < 1$ tal que

$$Z_n \leq const.(\alpha^n \|\overrightarrow{Z}_0\| + e_n + \alpha e_{n-1} \cdots \alpha^{n-1} e_1)$$

donde $\overrightarrow{Z} = (Z_1, \ldots, Z_p)$. En consecuencia

$$\sup_{k\leq n} Z_k = O\left(\sup_{k\leq n} e_k\right) \text{ y para } q \geq 1 \ \sum_{k=1}^{n} Z_k^q = O\left(1 + \sum_{k=1}^{n} e_k^q\right)$$

En consecuencia se obtiene el siguiente criterio de estabilidad:

Proposición 3.16. *Sea $\{X_n\}$ una sucesión de variables aleatorias en $\mathbb{R}^d$. Existen constantes positivas $\alpha_1, \ldots, \alpha_p$, tales que*

$$V(X_n) \leq \alpha_1 V(X_{n-1}) + \cdots \alpha_p V(X_{n-p})$$

con

- $\alpha_1 + \cdots + \alpha_p < 1$.
- *Sea $\{e_n\}$ una sucesión de variables aleatorias positivas, independientes e identicamentes distribuidas con momentos de orden $q \geq 1$. Entonces*
  - *Casi seguramente* $\sup_{k \leq n} V(X_k) = O\left(\sup_{k \leq n} e_k\right) = O(n^{1/q})$ *y* $\sum_{k=0}^{n} V(X_k) = O(n)$.
  - *Si $V(X_0)^q$ es integrable entonces* $\sup_n \mathbb{E}(X_n)^q < \infty$.

La demostración de estos resultados puede ser consultada en Duflo [20].

3.4. Ejercicios

1. Supongamos que $\{e_n\}$ es un ruido blanco gaussiano con valores en $\mathbb{R}^d$ suponemos que tiene media 0 y matriz de covarianza Γ con traza $tr\Gamma$.

   a) Demuestre para todo $a > 0$,

   $$\frac{1}{a}\exp(-a^2/2)(1-a^{-2}) \leq \int_a^\infty \exp(-u^2/2)du \leq \frac{1}{a}\exp(-a^2/2).$$

   b) Para $d = 1$, un ruido blanco centrado con varianza 1, demuestre que c.s

   $$\lim\sup |e_n|/2(\ln n)^{1/2} = 1.$$

   c) En general demuestre que,

   $$\lim\sup \|e_n\|/2(\ln n)^{1/2} \leq (tr\Gamma)^{1/2}$$

   deduzca que c.s $\sup_{k\leq n} \|e_n\| = O((\ln n)^{1/2})$.

2. Autorregresivos con umbrales. Consideremos el proceso $\{Y_n\}$ definido por

   $$Y_n = \begin{cases} aY_{n-1} + e_n \text{ Si } Y_{n-1} \leq s \\ bY_{n-1} + e_n \text{ Si } Y_{n-1} > s \end{cases}$$

   $\{e_n\}$ es un ruido blanco gaussiano con media 0 y varianza 1. Con densidad denotada por p y función de distribución Φ. Suponemos Y_0 y $\{e_n\}$ independientes.

   a) Para $|a| < 1$, $|b| < 1$, demuestre que este modelo es estable y demuestre que c.s $\sup_{k\leq n} Y_k = O((\ln n)^{1/2})$, (Ayuda: utilice el ejercicio anterior).

   b) Supongamos que $s = 0$, $a < 1$, $b < 1$ y $ab < 1$. Examinando los casos, $a < 0$ y $b > 0$, $a > 0$ y $b < 0$ por último a y b negativo, demuestre la estabilidad del modelo.

   c) Suponemos que el modelo es estable y tiene medida invariante μ. Demuestre que μ tiene una densidad h con respecto a la medida de Lebesgue y esta es la única solución de la ecuación

   $$h(y) = \int_{-\infty}^{s} p(y-at)h(t)dt + \int_s^\infty p(y-bt)h(t)dt$$

   Ayuda: utilice el método de estabilidad de Lyapunov.

Capítulo 4

Modelos de Markov Ocultos

En este capítulo desarrollamos la parte central del curso. Damos las principales propiedades y ejemplos de interés para ahondar en posteriores investigaciones.

Formalmente definimos un modelo de Markov oculto como una cadena de Markov bivariada $\{(X_k, Y_k)\}$ con valores en $(E \times F, \mathcal{E} \otimes \mathcal{F})$ para la cual

- $\{X_k\}$ es una cadena de Markov homogenea con valores en E (no observada) y núcleo de transición P.
- El núcleo de transición Q de la cadena conjunta $\{(X_k, Y_k)\}$ se factoriza como

$$Q((x, y), A \times B) = \int_B \int_A P(x, dx')G((x, y), dy')$$

donde G es un núcleo de emisión.

Nuestra noción de modelo de Markov oculto extiende la definición clásica. Nuestro objetivo es incluir los procesos autorregresivos con régimen de Markov permitiéndonos dar un tratamiento unificado. Observemos que si G no depende de y se obtiene la definición usual de modelo de

Markov oculto.

Nos concentraremos en tres ejemplos, las cadenas de Markov ocultas con espacio de estado finito para las observaciones y los estados ocultos con espacio de estado finito, cadenas de Markov con observaciones continuas y con espacio de estado finito y los procesos autorregresivos con régimen de Markov.

4.1. Cadenas de Markov ocultas finitas

Consideremos un proceso estocástico $\{Y_k\}_{k\geq 0}$ con valores en un conjunto finito $\{1,\ldots,r\}$. Este proceso es generado en dos etapas: Primero se muestrea un valor i de una cadena de Markov homegenea $\{X_n\}$ con valores en $\{1,\ldots,m\}$ y matriz de transición $P=\{p_{ij}\}$. Entonces para cada valor $i\in\{1,\ldots,m\}$ se muestrea un valor $s\in\{1,\ldots,r\}$ de la distribución de probabilidad definida por $g_{is}=\mathbb{P}(Y_n=s|X_n=i)$. La cadena $\{X_n\}$ no es observada, razón por la cual un tal proceso recibe el nombre de modelo de cadena de Markov oculta. Estos procesos están determinados por: la distribución de X_0, denotada por γ, la probabilidad de transición P, la matriz de emisiones $G=\{g_{il}\}$ de dimensiones $m\times r$.

Definimos un modelo de Markov oculto como una cadena de Markov bivariada $\{(X_k,Y_k)\}$ con valores en $\{1,\ldots,m\}\times\{1,\ldots,r\}$ para la cual

- $\{X_k\}$ es una cadena de Markov homogenea con espacio de estado $\{1,\ldots,m\}$ (no observada).
- Las entradas de la matriz de transición Q de la cadena conjunta $\{(X_k,Y_k)\}$ se factorizan como $Q_{(i,l)(j,s))}=p_{ij}g_{js}$.

Teorema 4.1. *Sea $\{(Y_k,X_k)\}$ un modelo de cadena de Markov oculta. Condicional a $\{X_k\}$ la variables aleatorias Y_k son independientes.*

Una aplicación de este modelo es el alineamiento de secuencias de ADN. El uso de modelos ocultos de Markov ha tenido un éxito notable para abordar este problema. El enfoque más simple es el siguiente. El parámetro k de tiempo representa la posición a lo largo de la cadena de ADN. La señal X_k es un proceso Markov en $E=\{0,1\}$: el k-ésimo par de la

base está en una región de codificación si $X_k = 1$, y en una región no codificada en caso contrario. El proceso de observación Y_k tiene un espacio de estados de cuatro letras $F = \{\mathrm{A}, \mathrm{C}, \mathrm{G}, \mathrm{T}\}$, de modo que Y_k representa el tipo de par de bases de orden k. El núcleo de transición y de emisión P y G se estiman a partir de los datos de secuencia. Una vez hecho esto, podemos ejecutar un procedimiento de estimación inversa para determinar que regiones de una secuencia de ADN están codificadas o no. Este enfoque es más bien ingenuo, sin embargo, es sorprendentemente bueno. Las regiones codificadas y no codificadas se caracterizan por frecuencias relativas para cada uno de los pares de bases. El enfoque se puede mejorar por la elección de un modelo oculto de Markov subyacente más sofisticado.

4.2. Cadenas de Markov ocultas gaussianas

Consideramos un modelo de Markov oculto definido por la ecuación

$$Y_n = \mu_{X_n} + \sigma_{X_n} e_n$$

donde $\{X_n\}_{n\geq 0}$ es una cadena de Markov con valores en el espacio de estado $\{1, \ldots, m\}$ y $\{e_n\}$ es una sucesión de v.a $\mathcal{N}(0,1)$ independientes de $\{X_n\}$ con los vectores $(\mu_1, \ldots, \mu_m) \in \mathbb{R}$ y $(\sigma_1^2, \ldots, \sigma_m^2) \in \mathbb{R}^+$.

Este modelo ha sido utilizado en diversas aplicaciones, a continuación describimos su uso en modelación de canales iónicos.

Una célula, por ejemplo del cuerpo humano, necesita intercambiar varios tipos de iones (sodio, potasio, etc.) con su entorno, esto lo usa para su metabolismo y para propósitos de comunicación química. La membrana de la célula es impermeable a tales iones pero contiene los llamados canales iónicos, cada uno adaptado para dejar pasar a través de ellos cada tipo de ion. Cada canal es en realidad una molécula, una proteína que puede adoptar diferentes configuraciones o estados. Cuando el canal permite el flujo de iones está abierto, en otro caso está cerrado. Un flujo de iones es un intercambio de cargas eléctricas del orden de los picoamperes. En otras palabras, cada estado del canal está caracterizado por el nivel de conductancia. Estos niveles corresponden a canales totalmente abiertos, totalmente cerrados o entre estos. Dicha actividad puede ser

medida con pruebas muy complejas que arrojan series que cambian entre los distintos niveles. De aquí la motivación de caracterizar la dinámica que estas series describen.

Modelar la actividad del canal se basa en dos pasos. Primero, se elige un modelo de Markov el cual se específica por el número de estados y su conectividad, la elección de un modelo esta típicamente motivada por otros datos como el número de subunidades y el número de conexiones entre sitios en el canal de proteína. En segundo lugar, se estiman los parámetros del modelo. Los parámetros del modelo de Markov a tiempo continuo son: las tasas de transición entre los diferentes estados, las probabilidades del estado inicial, y el nivel actual de cada estado.

En algunos casos en las pruebas realizadas no se percibe directamente si el canal esta abierto o cerrado, es allí donde la corriente observada se considera como la suma de dos componentes: una señal de ruido, que es la salida de tiempo discreto de un modelo oculto de Markov homogéneo de primer orden con estados finitos y un ruido gaussiano que corresponde al aparato de grabación. Para detalles acerca de la estimación del caso particular de esta aplicación ver Rosales [44] y sus referencias.

4.3. Procesos AR con régimen de Markov

Consideramos un proceso autorregresivo con régimen de Markov lineal AR-RM se define por

$$Y_n = \rho_{X_n} Y_{n-1} + b_{X_n} + \sigma_{X_n} e_n \tag{4.1}$$

donde $\{X_n\}_{n\geq 0}$ es una cadena de Markov con espacio de estado $\{1,\ldots,m\}$ y $\{e_n\}$ es una sucesión de v.a $\mathcal{N}(0,1)$ independientes de $\{X_n\}$ y de Y_0, con $\rho_1,\ldots,\rho_m$ tales que $|\rho_i|<1$, $b_1,\ldots,b_m$ contantes reales y $\sigma_1,\ldots,\sigma_m$ contantes positivas.

En el siguiente teorema se demuestra la existencia de una solución estacionaria para el proceso AR-RM.

Teorema 4.2. *Supongamos que* $\mathbb{E}_\varrho(\log(\rho_X)) < 0$*, donde* ϱ *es la medida invariante de la cadena* X*. Sin perdidad de generalidad suponemos que*

$(\sigma_1, \ldots, \sigma_m) = (1, \ldots, 1)$. *Sea ϑ la distribución de la variable*

$$Y_\infty = \sum_{n=0}^{\infty} \rho_{X_n} \cdots \rho_{X_1} (b_{X_{k+1}} + e_{n+1}).$$

Para un proceso AR-RM lineal se satisfacen las siguientes proposiciones:

- *El proceso $\{Y_n\}$ converge en distribución a ϑ.*
- *Propiedad de olvido. Para cualquier distribución de la variable Y_0, el proceso $\{Y_n\}$ converge en distribución a ϑ.*

Demostración: Tenemos que iterando la ecuación (4.1)

$$Y_n = \rho_{X_n} \cdots \rho_{X_1} Y_0 + b_{X_n} + e_n + \sum_{k=1}^{n-1} \rho_{X_n} \cdots \rho_{X_{k+1}} (b_{X_k} + e_k),$$

del Teorema Ergódico aplicado a la cadena $\{X_n\}_{n \geq 1}$, se tiene que

$$\lim_{n \to \infty} \frac{1}{n} \sum_{k=1}^{n} \log(\rho_{X_k}) \to \mathbb{E}_\varrho(\log(\rho_X)) \text{ c.s.}$$

donde $X \sim \varrho$, la ecuación anterior es equivalente a

$$\lim_{n \to \infty} \rho_1^{\frac{n_1}{n}} \cdots \rho_m^{\frac{n_m}{n}} \to \rho_1^{\varrho_1} \cdots \rho_m^{\varrho_m}.$$

Además $\rho_{X_n} \cdots \rho_{X_1} = (\rho_1^{\frac{n_1}{n}} \cdots \rho_m^{\frac{n_m}{n}})^n$ y por hipótesis $\rho_1^{\varrho_1} \cdots \rho_m^{\varrho_m} < 1$ entonces del teorema de convergencia dominada se obtiene

$$\lim_{n \to \infty} \rho_{X_n} \cdots \rho_{X_1} = 0, \text{ c.s.}$$

Por otra parte, como consecuencia del Teorema Ergódico para $\{X_k\}_{k \geq 1}$ y la LFGN para las variables aleatorias e_k tenemos $b_{X_k} + e_k = O(k)$ y así

$$\sum_{k=n}^{\infty} \rho_{X_k} \cdots \rho_{X_1} (b_{X_{k+1}} + e_{k+1}) \leq C \sum_{k=n}^{\infty} (\rho_1^{\varrho_1} \cdots \rho_m^{\varrho_m})^k |k+1| \to 0$$

entonces $\sum_{k=n}^{\infty} \rho_{X_k} \cdots \rho_{X_1}(b_{X_{k+1}} + e_{k+1})$ es la cola de una serie sumable y si para $k = 0$ definimos $\rho_{X_k} \cdots \rho_{X_1} = 1$ entonces

$$Y_\infty = \sum_{k=0}^{\infty} \rho_{X_k} \cdots \rho_{X_1}(b_{X_{k+1}} + e_{k+1})$$

y esta es una solución para el modelo AR-RM.

Esta solución es estacionaria, en efecto sea

$$Y_0 = \sum_{k=0}^{\infty} \rho_{X_k} \cdots \rho_{X_1}(b_{X_{k+1}} + e_{k+1})$$

y denotemos por ϑ su distribución. Si sustituimos Y_0 en el modelo, entonces valen las siguientes igualdades en sentido de distribución

$$\begin{aligned} Y_n &= \rho_{X_n} \cdots \rho_{X_1} \sum_{k=0}^{\infty} \rho_{X_k} \cdots \rho_{X_1}(b_{X_{k+1}} + e_{k+1}) \\ &\quad + b_{X_n} + e_n + \sum_{k=1}^{n-1} \rho_{X_n} \cdots \rho_{X_{k+1}}(b_{X_k} + e_k) \\ &= \sum_{k=0}^{\infty} \rho_{X_{k+n}} \cdots \rho_{X_1}(b_{X_{n+k+1}} + e_{k+1}) + \sum_{k=1}^{n-1} \rho_{X_n} \cdots \rho_{X_{k+1}}(b_{X_k} + e_k). \end{aligned}$$

Luego, para cada $n \in \mathbb{N}$, Y_n se distribuye ϑ.

Ahora demostramos que el proceso tiene olvido de la ley inicial. Elegimos Y_0' independiente de Y_0 y de la cadena de Markov $\{X_n\}_{n\geq 1}$ entonces

$$|Y_n - Y_n'| = |\rho_{X_n} \cdots \rho_{X_1}||Y_0 - Y_0'|$$

de donde $|Y_n - Y_n'| \to 0$. ∎

Ejemplo. Supongamos que $\theta_i = (0, \rho_i)^t$ para todo $i = 1, \ldots, m$ y que la variable aleatoria e_n se distribuyen $\mathcal{N}(0, 1)$ entonces la solución estacionaria toma la forma:

$$\begin{aligned} Y_\infty = \sum_{n=0}^{\infty} \rho_{X_n} \cdots \rho_{X_1} e_{n+1} &= \sum_{n=0}^{\infty} \rho_1^{n_1} \cdots \rho_m^{n_m} e_{n+1} \\ &\approx \sum_{n=0}^{\infty} (\rho_1^{\varrho_1} \cdots \rho_m^{\varrho_m})^n e_{n+1}, \end{aligned}$$

y la distribución estacionaria es $\vartheta \sim \mathcal{N}(0, \frac{1}{1-d^2})$ con $d = \rho_1^{\varrho_1} \cdots \rho_m^{\varrho_m}$.

4.4. Procesos AR no lineales con régimen de Markov

Consideramos un proceso autorregresivo con régimen de Markov no lineal (ARN-RM) definido por

$$Y_n = r_{X_n}(Y_{n-1}) + \sigma_{X_n} e_n \tag{4.2}$$

donde $\{X_n\}_{n\geq 1}$ es una cadena de Markov con espacio de estados finito $\{1, \ldots, m\}$ y $\{e_n\}$ es una sucesión de v.a $\mathcal{N}(0,1)$ independientes de $\{X_n\}$ y la v.a Y_0. Se denota por $P = [p_{ij}]$ la matriz de transición cuyos elementos son $p_{ij} = \mathbb{P}(X_n = j | X_{n-1} = i)$.

Suponemos que un proceso ARN-RM verifica las siguientes condiciones:

- La cadena de Markov $\{X_n\}_{n\geq 1}$ es recurrente y positiva. Su distribución invariante se denota por, $\varrho = (\varrho_1, \ldots, \varrho_m)$.
- La sucesión de v.a $\{e_n\}_{n\geq 1}$ es independiente e identicamente distribuida.
- Y_0, la cadena de Markov $\{X_n\}_{n\geq 1}$ y la sucesión $\{e_n\}_{n\geq 1}$ son v.a. mutuamente independientes.

Demostraremos que el proceso conjunto $\{Z_n\}_{n\geq 1}$ definido por $Z_n = (X_n, Y_n)$ es un proceso de Markov, viendo que puede ser escrito como un modelo iterativo.

Lema 4.3. *El proceso conjunto $\{Z_n\}_{n\geq 1}$ definido por $Z_n = (X_n, Y_n)$ es una cadena de Markov con espacio de estados $\{1, \ldots, m\} \times \mathbb{R}$. Además,*

- *Esta cadena es felleriana si las funciones de regresión r_i son continuas, para $i = 1, \ldots, m$.*
- *Si la v.a. e_1 admite una densidad Φ con respecto a la medida de Lebesgue la cadena de Markov $\{Z_n\}_{n\geq 1}$ es fuertemente felleriana.*
- *Si la densidad Φ de e_1 es estrictamente positiva entonces la cadena de Markov $\{Z_n\}_{n\geq 1}$ es φ-irreductible.*

Demostración: El lema 3.4 implica que

$$X_n = F(X_{n-1}, u_n).$$

para una sucesión de variables aleatorias independientes $\{u_n\}_{n\geq 1}$ las cuales son a su vez independientes de $\{e_n\}_{n\geq 1}$ y donde $F : E\times[0,1] \to E$ es una función medible. Así, según el lema 3.4, $\{Z_n\}_{n\geq 1}$ es un modelo iterativo markoviano de la forma

$$Z_n = \begin{pmatrix} X_n \\ Y_n \end{pmatrix} = \begin{pmatrix} F(X_{n-1}, u_n) \\ r_{F(X_{n-1},u_n)}(Y_{n-1}) + e_n \end{pmatrix}.$$

Denotemos por Q el núcleo de transición del proceso conjunto. Demostraremos que Q es φ-irreductible, donde φ es la medida definida por $\mu_c\lambda$ (contar $\otimes$ Lebesgue) sobre $\mathcal{P}\{1,\ldots,m\}\times\mathcal{B}(\mathbb{R})$, donde $\mathcal{P}(\{1,\ldots,m\})$ es la familia de las partes de $\{1,\ldots,m\}$ y $\mathcal{B}(\mathbb{R})$ es la familia de bolerianos de $\mathbb{R}$. Sean $A \in \mathcal{P}(\{1,\ldots,m\})$ y $B \in \mathcal{B}(\mathbb{R})$. Supongamos que $\mu_c(A)\lambda(B) > 0$, entonces

$$\begin{aligned} Q(Z, A\times B) &:= \mathbb{P}\left(x \in A,\ r_x(y) + e_1 \in B\right) \\ &= \delta_x(A)\mathbb{P}\left(e_1 \in B - r_x(y)\right). \end{aligned}$$

Supongamos que $\delta_x(A)\mathbb{P}\left(e_1 \in B - r_x(y)\right) = 0$ entonces $\mathbb{P}(e_1 \in B - r_x(y)) = 0$, ya que $\delta_x(A) > 0$. Luego, $\mathbb{P}(e \in B - r_x(y)) = 0$ y, si suponemos que la densidad del ruido es estrictamente positiva, podemos afirmar que

$$\int \mathbb{1}_B(x)\Phi(x)dx = 0$$

y como $\Phi > 0$; entonces, $\Phi(x)\mathbb{1}_B(x) = 0$, c.s-$x$, en consecuencia $\lambda(B) = 0$ y se obtiene una contradicción porque $\lambda(B) > 0$. Por lo tanto

$$Q(Z, A\times B) > 0$$

en consecuencia, la cadena es φ-irreductible. ■

Yao y Attali en [55] demuestran que con la hipótesis de sublinealidad

$$r_i(y) \leq \rho_i|y| + b_i$$

para $i = 1, \ldots, m$, la condición de estabilidad

$$\mathbb{E}_{\varrho}(\log \rho) = \sum_{i=1}^{m} \log(\rho_i)\varrho_i < 0$$

y la función de Lyapounov $V(x, y) = \|y\|^{\frac{1}{p}} + 1$ sobre $\{1, \ldots, m\} \times \mathbb{R}$, se satisface la desigualdad

$$Q^p V(x, y) \leq \rho_p V(x, y) + \beta_p + 1 - \rho_p \tag{4.3}$$

para constantes positivas, ρ_p, β_p con $\rho_p < 1$. Bajo la suposición de que la cadena $\{Z_n\}_{n\geq 0}$ es felleriana existe una una medida invariante. Esta medida invariante será única cuando la cadena de Markov $\{Z_n\}_{n\geq 0}$ sea por ejemplo φ-irreductible, 1.IV.19 de Duflo [20]. Entonces bajo las hipótesis del lema 1, queda garantizada la existencia de una única medida invariante.

La desigualdad (4.3) permite demostrar que la cadena $\{Z_n\}_{n\geq 0}$ es V-uniformemente ergódica, es decir

$$\|Q^n - \vartheta\|_V \to 0 \quad n \to \infty$$

donde ϑ es la única medida invariante del proceso conjunto $\{Z_n\}_{n\geq 0}$ y por marginalización se obtiene la medida invariante del proceso $\{Y_n\}_{n\geq 0}$.

Teorema 4.4. *Consideremos un proceso ARN-RM definido por la ecuación (4.2), supongamos que las funciones de regresión r_i en cada régimen son sublineales y que se satisface la condición de estabilidad. Entonces*

1. *Existe una única solución estacionaria geométricamente ergódica.*
2. *Si además $\mathbb{E}(|e_1|^s) < \infty$ y la matriz $Q_s = \left(\rho_j^s\, p_{ij}\right)_{i,j=1\ldots m}$ tiene radio espectral estrictamente menor que 1, entonces $\mathbb{E}(|Y_k|^s) < \infty$.*

La demostración de la parte 1 es la discusión anterior al enunciado del teorema. Para la existencia de momentos referimos a Yao y Attali [55]. Para demostrar la irreductibilidad es fundamental suponer que existe una densidad positiva para la sucesión de innovaciones, que en el caso

cuando la distribución es discreta deja de ser cierto. Por ello en Yao y Attali se sustituye la condición de sublinealidad por una condición de Lipschitz de las funciones r_i y se demuestra directamente, sin pasar por la cadena conjunta $\{Z_n\}_{n\geq 0}$, que el proceso $\{Y_n\}_{n\geq 0}$ es estable.

En el caso cuando e_1 tiene distribución discreta se puede demostrar que la cadena de Markov conjunta Z admite una medida invariante. En efecto, suponiendo que las funciones de regresión r_i para $i = 1, \ldots, m$ son continuas, la cadena de Markov $\{Z_n\}_{n\geq 0}$ es felleriana y si adicionalmente las funciones de regresión son sublineales, entonces se satisface una condición de contracción del tipo (4.3). Esto es suficiente para garantizar existencia de la medida invariante pero no la unicidad.

Una técnica que permite demostrar unicidad para cadenas fellerianas que satisfacen una condición de deriva y que no son irreductibles es verificar que se satisface una condición de ser alcanzable, es decir, existe $z \in \{1, \ldots, m\} \times \mathbb{R}$ tal que $\sum_{k=1}^{\infty} Q^k(z, A \times B) > 0$ para todos los conjuntos abiertos $A \times B$ que contienen a z. En este caso, queda como problema abierto demostrar que bajo innovaciones con distribución discreta se satisface la condición de ser alcanzable.

También se puede debilitar la hipótesis de continuidad de las funciones r_i. Attali en [5] introduce la noción de cadenas quasi-fellerianas la cual es más débil que la fellerianidad y con esta demuestra el siguiente teorema.

Teorema 4.5. *ARN-RM definido por la ecuación (4.2). Para el cual:*

1. *Existe la densidad Φ del proceso $\{e_n\}_{n\geq 1}$.*
2. *Las funciones r_i son Riemann integrables.*
3. *Para cada $(i, y) \in \{1, \ldots, m\} \times \mathbb{R}$, la sucesión*

$$\left\{ 1/n \sum_{k=1}^{n} Q^k((i, y), di \times dy) \right\}_{n\geq 1}$$

es tensa.

Entonces $\{Z_n\}_{n\geq 1}$ es una cadena de Markov Harris positiva.

Para la demostración y la definición de quasi-fellerianidad ver Attali [5].

4.4.1. Existencia de la distribución finito dimensional del proceso conjunto

Comenzamos denotando por $V_{1:n}$ el vector aleatorio $(V_1, \ldots, V_n)^t$ y $v_{1:n} = (v_1, \ldots, v_n)^t$ cualquier realización. Ahora veamos la existencia de la densidad conjunta de las variables $Y_{0:n}, X_{1:n}$, para el proceso autorregresivo con régimen de Markov, definido por la ecuación (4.2).

Lema 4.6. *Para el proceso AR-RM definido en (4.2), el vector* $(Y_{0:n}, X_{1:n})$ *admite densidad de probabilidad*

$$\begin{aligned} & p(Y_{0:n} = y_{0:n}, X_{1:n} = x_{1:n}) \\ & \quad = \Phi(y_n - r_{x_n}(y_{n-1})) \cdots \Phi(y_1 - r_{x_1}(y_0)) p_{x_{n-1}x_n} \cdots p_{x_1x_2} \mu_{x_1} p_{Y_0}(y_0) \end{aligned}$$

con respecto a la medida producto $\lambda \otimes \mu_c$*, donde* λ *y* μ_c *denotan las medidas de Lebesgue y de contar, respectivamente.*

Demostración: Definimos el cambio de variables

$$T(Y_1, \ldots, Y_n) = (e_1, \ldots, e_n)$$

donde $e_k = Y_k - r_{X_k}(Y_{k-1})$, para $k = 1, \ldots, n$. Así, por el teorema del cambio de variables, como la matriz jacobiana T es triangular su determinante es igual a 1, por lo tanto para cualquier función medible h,

$$\begin{aligned} & \mathbb{E}(h(Y_{1:n}, X_{1:n}, Y_0)) \\ & \quad = \mathbb{E}(h(T^{-1}(e_{1:n}), X_{1:n}, Y_0)) \\ & \quad = \int \sum_{i_{1:n}} h(T^{-1}(u_{1:n}), i_{1:n}, y_0)) p(e_{1:n} = u_{1:n}, X_{1:n} = i_{1:n}, Y_0 = y_0) du_{1:n} dy_0 \end{aligned}$$

utilizando la independencia conjunta,

$$p(e_{1:n} = u_{1:n}, X_{1:n} = i_{1:n}, Y_0 = y_0) = p(e_{1:n} = u_{1:n}) p(X_{1:n} = i_{1:n}) p_{Y_0}(y_0)$$

como las densidades de Y_0 y e_1 existen,

$$\begin{aligned} & \mathbb{E}(h(Y_{1:n}, X_{1:n}, Y_0)) \\ & \quad = \int \sum_{i_{1:n}} h(T^{-1}(u_{1:n}), i_{1:n}, y_0) \prod_{k=1}^{n} \Phi(u_k) \prod_{k=2}^{n} p_{i_{k-1}i_k} \mu_{i_1} p_{Y_0}(y_0) du_{1:n} dy_0. \end{aligned}$$

Por lo tanto las v.a $(Y_1, \ldots, Y_n, X_{1:n}, Y_0)$ admiten densidad conjunta con respecto a la medida de Lebesgue producto y la medida de contar producto. ■

La ventaja de este resultado es que cualquier otra distribución marginal de interés se obtiene por integración de la densidad conjunta obtenida.

4.4.2. Propiedades de dependencia

Un proceso $Y = \{Y_k\}_{k\in\mathbb{Z}}$ es fuertemente α-mezclante (mixing en inglés), si

$$\alpha_n := \sup\{|\mathbb{P}(A \cap B) - \mathbb{P}(A)\mathbb{P}(B)| : A \in \mathcal{M}^0_{-\infty}, B \in \mathcal{M}^\infty_n\} \to 0, \quad (4.4)$$

cuando $n \to \infty$ y denotamos por $\mathcal{M}^b_a$, con $a, b \in \overline{\mathbb{Z}}$, la σ-algebra generada por $\{Y_k\}_{k=a:b}$.

Es absolutamente regular, si

$$\beta_n := \mathbb{E}\left(ess\sup\{\mathbb{P}(B|\mathcal{M}^0_{-\infty}) - \mathbb{P}(B) : B \in \mathcal{M}^\infty_n\}\right) \to 0, \quad \text{con } n \to \infty. \quad (4.5)$$

Los valores α_n son llamados coeficientes fuertemente mezclantes, y los valores β_n son los coeficientes absolutamente regulares. Ver Doukhan (1994) para ejemplos y propiedades bajo condiciones de dependencia, [19]. En general, tenemos la desigualdad $2\alpha_n \leq \beta_n \leq 1$.

Los coeficientes α-mezclantes se pueden escribir como:

$$\alpha_n := \sup\{|\mathrm{cov}(\phi, \xi)| : 0 \leq \phi, \xi \leq 1, \phi \in \mathcal{M}^0_{-\infty}, \xi \in \mathcal{M}^\infty_n\}. \quad (4.6)$$

En el caso de procesos de Markov estrictamente estacionarios X, con espacio de estados $(E, \mathcal{E})$, núcleo de transición Q y medida de probabilidad invariante ϱ; los coeficientes β toman la forma siguiente, see Doukhan ([19], sección 2.4):

$$\beta_n := \mathbb{E}\left(\sup\{|Q^{(n)}(X, B) - \varrho(B)| : B \in \mathcal{E}\}\right). \quad (4.7)$$

Proposición 4.7. *El proceso ARN-RM definido por la ecuación (4.2) y bajo las condiciones del lema 1, es estrictamente estacionario, α-mezclante y sus coeficientes decrecen geométricamente.*

Demostración: por teorema 1 en [19, Section 2.4], la ergodicidad geométrica implica la propiedad β-mezclante para Z. Más aun, exite $0 < \zeta < 1$ y $c > 0$ tal que el coficiente β-mezclante satisface

$$\beta_n(Z) \leq c\,\zeta^n.$$

Así, de la desigualdad $2\alpha_n(Z) \leq \beta_n(Z)$, el proceso Z es también α-mezclante.

Además, el proceso Y se obtiene desde Z como $Y_n = \pi(Z_n)$, donde π es la función de proyección. Como, la proyección π es una función continua tenemos $\mathcal{M}_a^b(Y) \subset \mathcal{M}_a^b(Z)$ para todo a, b. Entonces, la expresión (4.4) para los coeficientes α-mezclante implica que

$$\alpha_n(Y) \leq \alpha_n(Z) \leq \frac{1}{2}\beta_n(Z) \leq \frac{c}{2}\zeta^n.$$

Por lo tanto, Y es α-mezclante y su coeficiente $\alpha_n(Y)$ decrece geométricamente. ■

Ejemplo: proceso AR-RM que no es fuertemente α-mezclante. Si consideramos el caso de un proceso AR-RM lineal con $\theta_i = (0, \rho_i)^t$ para todo $i = 1, \ldots, m$, que la variable aleatoria e_1 se distribuye Bernoulli con parámetro q y que $Y_0 = 0$ entonces el proceso definido por

$$Y_n = \sum_{k=0}^{n-1} \rho_{X_k} \cdots \rho_{X_1} e_{k+1},$$

con la convención de que $\rho_{X_k} \cdots \rho_{X_1} = 1$ para $k = 0$, no es fuertemente α-mezclante.

En efecto se puede probar (ver D. Andrews [3]) que si $0 < \rho_i \leq 1/2$ entonces existe un conjunto $A \in \mathcal{M}_{-\infty}^0$ con $\mathbb{P}(A) > 0$ y existen conjuntos $B_s \in \mathcal{M}_n^\infty$ con $\mathbb{P}(B_s) \leq c$ para $s \in \mathbb{N}$, y alguna constante $c < 1$ tal que $\mathbb{P}(B_s|A) = 1$, para todo s y esto implica que

$$\alpha_s(Y) \geq \mathbb{P}(A \cap B_s) - \mathbb{P}(A)\mathbb{P}(B_s) = \mathbb{P}(A)(\mathbb{P}(B_s|A) - \mathbb{P}(B_s)) \leq \mathbb{P}(A)(1-c)$$

y por lo tanto $\alpha_s(Y)$ no decrece a 0 cuando $s \to \infty$, por lo tanto el proceso Y no es fuertemente α-mezclante.

4.5. Ejercicios

1. Demuestre el teorema 4.1.

2. Autorregresivos con régimen de Markov. Sea $\{X_n\}$ una cadena de Markov homogénea con espacio de estados $E = \{1, 2\}$. Consideremos el proceso $\{Y_n\}$ definido por

$$Y_n = \begin{cases} aY_{n-1} + e_n \text{ Si } X_n = 1 \\ bY_{n-1} + e_n \text{ Si } X_n = 2 \end{cases}$$

   $\{e_n\}$ es un ruido blanco gaussiano con media 0 y varianza 1. Con densidad denotada por ϕ y función de distribución Φ. Suponemos Y_0 y $\{e_n\}$ independientes.

   a) Para $|a| < 1$, $|b| < 1$, demuestre que este modelo es estable y demuestre que c.s $\sup_{k \leq n} Y_k = O((\ln n)^{1/2})$.

   b) Puede debilitarse la condición de estabilidad del apartado anterior.

   c) Suponemos que el modelo es estable y tiene medida invariante μ. Demuestre que μ tiene una densidad h con respecto a la medida de Lebesgue y de la forma de la misma.

3. ¿ Qué se puede decir de la estabilidad de un proceso AR-RM cuando el ruido tiene distribución soportada en un conjunto discreto?

Capítulo 5

Modelos con datos incompletos

En este capítulo introducimos algunos algoritmos numéricos para obtener el estimador de máxima verosimilitud para modelos de Markov Ocultos.

Si consideremos un modelo de Markov oculto $\{Y_k\}_{k\geq 0}$ dado un conjunto de observaciones $y_1, \ldots, y_n$, la función de verosimilitud se define por

$$\mathbb{P}_\theta(Y_1 = y_1, \ldots, Y_n = y_n)$$

la cual denotamos por L y el estimador de máxima verosimilitud (EMV) esta dado por

$$\hat{\theta}_n = \arg\max_\theta \log \mathbb{P}_\theta(Y_1 = y_1, \ldots, Y_n = y_n) \tag{5.1}$$

esto significa que el parámetro θ que estimamos es aquel que hace más probables (verosímiles) los datos observados.

Consideremos la función de verosimilitud para un modelo de Markov oculto se escribe como

$$\begin{aligned} \mathbb{P}_\theta(Y_{1:n} = y_{1:n}) &= \sum_{x_{1:n}} \mathbb{P}_\theta(Y_{1:n}, X_{1:n} = x_{1:n}) \\ &= \int_{x_{1:n}} \mathbb{P}_\theta(Y_{1:n} = y_{1:n}, X_{1:n} = x_{1:n}) \mu_c(dx_{1:n}) \end{aligned} \tag{5.2}$$

donde μ_c es la medida de contar producto.

Para modelos parcialmente observados como las cadenas de Markov ocultas su compleja estructura impide tener estimadores en forma de expresiones cerradas, en efecto se observa de la ecuación (5.2) que obtener el argumento máximo de (5.1) implica un número elevado de operaciones, este se reduce si utilizamos el algoritmo Esperanza-Maximización (EM) el cual permite realizar la estimación de manera eficiente para modelos de Markov ocultos. Este algoritmo es introducido en forma particular para modelos de CMO Baum *et al.* [8] en su versión de algoritmo forward-backward, el cual es una forma temprana del algoritmo EM. El algoritmo EM propuesto en su forma general por Dempster *et al.* [17] maximiza la función log-verosimilitud en problemas con datos incompletos.

5.1. Algoritmo EM

El algoritmo EM es un método recursivo que permite cambiar la maximización de la función de verosimilitud observada por un problema de maximización de un funcional de la función de verosimilitud completa.

Para definir este funcional introducimos la distribución condicional

$$\mathbb{P}_\theta(X_{1:n} = x_{1:n}|Y_{1:n} = y_{1:n}) = \frac{\mathbb{P}_\theta(Y_{1:n} = y_{1:n}, X_{1:n} = x_{1:n})}{\mathbb{P}_\theta(Y_{1:n} = y_{1:n})}$$

y consideramos la función auxiliar

$$\mathcal{A}(\theta, \theta') = \mathbb{E}_{\theta'}(\log \mathbb{P}_\theta(Y_{1:n} = y_{1:n}, X_{1:n} = x_{1:n})|Y_{1:n} = y_{1:n}) \qquad (5.3)$$

El siguiente teorema demuestra que encontrar los máximos de la función $\mathbb{P}_\theta(Y_{1:n} = y_{1:n})$ es equivalente a los máximos de la función $\mathcal{A}(\theta, \theta')$.

Teorema 5.1. *Si* $\mathcal{A}(\theta, \theta') \geq \mathcal{A}(\theta, \theta)$ *entonces* $\mathbb{P}_{\theta'}(Y_{1:n} = y_{1:n}) \geq \mathbb{P}_\theta(Y_{1:n} = y_{1:n})$

Demostración: Consideremos

$$\begin{aligned}
&\frac{\mathbb{P}_{\theta'}(Y_{1:n}=y_{1:n})}{\mathbb{P}_{\theta}(Y_{1:n}=y_{1:n})}\\
&\quad=\frac{\int_{x_{1:n}}\mathbb{P}_{\theta'}(Y_{1:n}=y_{1:n},X_{1:n}=x_{1:n})\mu_c(dx_{1:n})}{\mathbb{P}_{\theta}(Y_{1:n}=y_{1:n})}\\
&\quad=\int_{x_{1:n}}\frac{\mathbb{P}_{\theta'}(Y_{1:n}=y_{1:n},X_{1:n}=x_{1:n})}{\mathbb{P}_{\theta}(Y_{1:n}=y_{1:n},X_{1:n}=x_{1:n})}\mathbb{P}_{\theta}(X_{1:n}=x_n|Y_{1:n}=y_{1:n})\mu_c(dx_{1:n})
\end{aligned}$$

Como consecuencia de la desigualdad de Jensen

$$\begin{aligned}
&\log\left[\int_{x_{1:n}}\frac{\mathbb{P}_{\theta'}(Y_{1:n}=y_{1:n},X_{1:n}=x_{1:n})}{\mathbb{P}_{\theta}(Y_{1:n}=y_{1:n},X_{1:n}=x_{1:n})}\mathbb{P}_{\theta}(X_{1:n}=x_n|Y_{1:n}=y_{1:n})\mu_c(dx_{1:n})\right]\\
&\quad\geq\int_{x_{1:n}}\log\left[\frac{\mathbb{P}_{\theta'}(Y_{1:n}=y_{1:n},X_{1:n}=x_{1:n})}{\mathbb{P}_{\theta}(Y_{1:n}=y_{1:n},X_{1:n}=x_{1:n})}\right]\mathbb{P}_{\theta}(X_{1:n}=x_n|Y_{1:n}=y_{1:n})\mu_c(dx_{1:n})\\
&\quad=\quad\mathcal{A}(\theta,\theta')-\mathcal{A}(\theta,\theta)
\end{aligned}$$

en consecuencia

$$\log\left(\frac{\mathbb{P}_{\theta'}(Y_{1:n}=y_{1:n})}{\mathbb{P}_{\theta}(Y_{1:n}=y_{1:n})}\right)\geq\mathcal{A}(\theta,\theta')-\mathcal{A}(\theta,\theta)$$

por hipótesis $\mathcal{A}(\theta,\theta')-\mathcal{A}(\theta,\theta)\geq 0$ de donde $\mathbb{P}_{\theta'}(Y_{1:n}=y_{1:n})\geq\mathbb{P}_{\theta}(Y_{1:n}=y_{1:n})$ ■

Dempster *et al.* [17] notan que

$$\mathcal{A}(\theta,\theta')=\log\mathbb{P}_{\theta}(Y_{1:n}=y_{1:n})-\mathcal{H}(\theta,\theta'),$$

esto dice que la verosimilitud observada y la cantidad auxiliar del algoritmo EM difieren en la cantidad

$$\begin{aligned}
&\mathcal{H}(\theta,\theta')\\
&\quad=\quad-\int_{x_{1:n}}\log\mathbb{P}_{\theta}(X_{1:n}=x_n|Y_{1:n}=y_{1:n})\mathbb{P}_{\theta'}(X_{1:n}=x_n|Y_{1:n}=y_{1:n})\mu_c(dx_{1:n})
\end{aligned}$$

que se reconoce como la entropía de la función de probabilidad $\mathbb{P}_{\theta}(X_{1:n}=x_n|Y_{1:n}=y_{1:n})$ y cuyo incremento

$$\begin{aligned}
&\mathcal{H}(\theta,\theta')-\mathcal{H}(\theta',\theta')\\
&\quad=\int_{x_{1:n}}\log\left[\frac{\mathbb{P}_{\theta}(X_{1:n}=x_n|Y_{1:n}=y_{1:n})}{\mathbb{P}_{\theta'}(X_{1:n}=x_n|Y_{1:n}=y_{1:n})}\right]\mathbb{P}_{\theta'}(X_{1:n}=x_n|Y_{1:n}=y_{1:n})\mu_c(dx_{1:n})
\end{aligned}$$

es la distancia de Kullback-Leibler o entropía relativa.

A continuación describimos los pasos del algoritmo EM:

1. Iniciar con $\theta_0 \in \Theta$.
2. Paso E: Calcular $\mathcal{A}(\theta, \theta^{(t)})$.
3. Paso M: $\theta^{(t+1)} = \arg\max_\theta \mathcal{A}(\theta, \theta^{(t)})$.
4. Repetir hasta que $\mathbb{P}_{\theta^{(t+1)}}(Y_{1:n} = y_{1:n}) - \mathbb{P}_{\theta^{(t)}}(Y_{1:n} = y_{1:n}) < Tol$.

Ejemplo 5.2. *Modelos de Markov oculto finitos*
En este caso escribimos la función de verosimilitud en la forma

$$\prod_{k=1}^{n} \left(\prod_{i=1}^{m} \gamma_i^{\mathbb{1}_i(X_k)} \prod_{i,j=1}^{m} p_{ij}^{\mathbb{1}_{(i,j)}(X_{k-1},X_k)} \prod_{i=1}^{m} q_{iy_k}^{\mathbb{1}_i(X_k)} \right)$$

donde $\mathbb{1}_i(X_k)$ indica si la cadena visitó o no el estado i en el tiempo k y $\mathbb{1}_{(i,j)}(X_{k-1}, X_k)$ indica si hubo una transición de i a j desde el tiempo $k-1$ hasta el tiempo k.

Al aplicar el logaritmo y la esperanza condicional dado los datos en esta nueva expresión de la verosimilitud completa se obtiene la expresión siguiente para la función auxiliar $\mathcal{A}$,

$$\begin{aligned} &\sum_{k=1}^{n}\sum_{i}^{m} \mathbb{E}_{\theta'}(\mathbb{1}_i(X_k)|Y_{1:n} = y_{1:n}) \log(\gamma_i) \\ &\quad + \sum_{k=1}^{n}\sum_{i,j=1}^{m} \mathbb{E}_{\theta'}(\mathbb{1}_{(i,j)}(X_{k-1}, X_k)|Y_{1:n} = y_{1:n}) \log(p_{ij}) \\ &\quad + \sum_{k=1}^{n}\sum_{i=1}^{m} \mathbb{E}_{\theta'}(\mathbb{1}_i(X_k)|Y_{1:n} = y_{1:n}) \log(q_{iy_k}) \end{aligned} \tag{5.4}$$

Observemos que

$$\mathbb{E}_{\theta'}(\mathbb{1}_{(i,j)}(X_{k-1}, X_k)|Y_{1:n} = y_{1:n}) = \mathbb{P}_{\theta'}(X_{k-1} = i, X_k = j|Y_{1:n} = y_{1:n})$$

y

$$\mathbb{E}_{\theta'}(\mathbb{1}_i(X_k)|Y_{1:n} = y_{1:n}) = \mathbb{P}_{\theta'}(X_k = i|Y_{1:n} = y_{1:n})$$

Al derivar la función auxiliar para encontrar las expresiones para el paso M se obtiene

$$\hat{\gamma}_i^{(t)} = \mathbb{P}(X_1 = i | Y_1^n = y_{1:n}) \tag{5.5}$$

$$\hat{p}_{ij}^{(t)} = \frac{\sum_{k=2}^n \mathbb{P}(X_{k-1} = i, X_k = j | Y_1^n = y_{1:n})}{\sum_{k=2}^n \mathbb{P}(X_{k-1} = i | Y_1^n = y_{1:n})} \tag{5.6}$$

$$\hat{q}_{il}^{(t)} = \frac{\sum_{k=1}^n \mathbb{1}_l(y_k) \mathbb{P}(X_k = j | Y_1^n = y_{1:n})}{\sum_{k=1}^n \mathbb{P}(X_k = j | Y_1^n = y_{1:n})} \tag{5.7}$$

Las cantidades anteriores se evaluan utilizando las recursiones de Baum y Welch.

Ejemplo 5.3. *Modelo MS-AR*
En este caso la función de verosimilitud esta dada por

$$\prod_{k=1}^n \prod_{i,j=1}^m p_{ij}^{\mathbb{1}_{i,j}(x_k, x_{k+1})} \prod_{k=1}^n \prod_{i=1}^m \left[\frac{1}{\sqrt{2\pi\sigma_i^2}} \exp\left(-\frac{(y_k - \rho_i y_{k-1} - b_i)^2}{\sigma_i^2} \right) \right]^{\mathbb{1}_i(x_k)} \tag{5.8}$$

Para describir el paso $t+1$ de este algoritmo, consideramos

$$\begin{aligned}
&\mathcal{A}(\theta, \theta^{(t)}) \\
&= \mathbb{E}(\log P_\theta(Y_{0:n}, X_{1:n} = x_{1:n}) | Y_{0:n} = y_{0:n}, \theta^{(t)}) \\
&= \sum_{n=1}^{N-1} \sum_{i,j=1}^m \mathbb{E}(\mathbb{1}_{i,j}(X_n, X_{n+1}) | Y_{0:n} = y_{0:n}, \theta^{(t)}) \log(p_{ij}) \\
&- \sum_{n=1}^{N-1} \sum_{i=1}^m \mathbb{E}(\mathbb{1}_i(X_n) | Y_{0:n} = y_{0:n}, \theta^{(t)}) \left[\frac{\log(2\pi\sigma_i^2)}{2} + \frac{(y_k - \rho_i y_{k-1} - b_i)^2}{\sigma_i^2} \right]
\end{aligned}$$

El algoritmo EM se desarrolla en dos pasos, en el paso E se evalúa la función $\mathcal{A}(\theta, \theta^{(t)})$ y en el paso M calculamos

$$\theta^{(t+1)} = \underset{\theta \in \Theta}{argmax}\, \mathcal{A}(\theta, \theta^{(t)}).$$

Como e_1 tiene distribucuión gaussiana el modelo pertenece a la familia exponencial, por lo que satisface las hipótesis que garantizan la convergencia del algoritmo EM.

5.1.1. Recursiones de Baum y Welch

Definimos las cantidades auxiliares

$$\alpha_k(i) = \mathbb{P}_\theta(Y_1^k = y_{1:n}, X_k = i)$$

y

$$\beta_k(i) = \mathbb{P}_\theta(Y_{k+1}^n = y_{k+1}^n | X_k = i)$$

Estas cantidades se relacionan con la función de verosimilitud observada por la siguiente formula

$$\mathbb{P}_\theta(Y_{1:n}) = \sum_{i=1}^{m} \alpha_k(i)\beta_k(i) \tag{5.9}$$

Las cantidades α_k y β_k se calculan mediante las recursiones

$$\alpha_{k+1}(j) = \left(\sum_{i=1}^{m} \alpha_k(i)p_{ij}\right) q_{jy_{k+1}} \tag{5.10}$$

$$\beta_k(i) = \sum_{j=1}^{m} q_{jy_{k+1}}\beta_{k+1}(j)p_{ij} \tag{5.11}$$

donde para cada $i \in \{1, \ldots, m\}$, $\alpha_1(i) = \gamma_i q_{iy_1}$ y $\beta_n(i) = 1$.

Las ecuaciones (5.5),(5.6) y (5.7) se evaluan a partir de las cantidades,

$$\mathbb{P}(X_k = i | Y_1^n = y_{1:n}) = \frac{\alpha_k(i)\beta_k(i)}{\mathbb{P}_\theta(Y_{1:n} = y_{1:n})} \tag{5.12}$$

y

$$\mathbb{P}(X_{k-1} = i, X_k = j | Y_1^n = y_{1:n}) = \frac{\alpha_k(i)p_{ij}q_{jy_{k+1}}\beta_{k+1}(j)}{\mathbb{P}_\theta(Y_{1:n} = y_{1:n})} \tag{5.13}$$

5.2. Monte Carlo EM

Recordemos que el algoritmo EM evalua la función auxliar $\mathcal{A}(\theta, \theta')$ definida en (5.3). En esta sección se considera la evaluación de esta cantidad por una aproximación numérica de tipo Monte Carlo,

$$\hat{\mathcal{A}}(\theta, \theta') = \frac{1}{s}\sum_{j=1}^{s} \log \mathbb{P}_\theta(Y_{1:n} = y_{1:n}, X^j)$$

donde $X^1, \ldots, X^s$ son muestras de la densidad condicional $\mathbb{P}_\theta(X_{1:n}|Y_{1:n} = y_{1:n})$. El algoritmo EM es modificado reemplazando $\mathcal{A}$ por $\hat{\mathcal{A}}$ en el paso E. A continuación describimos los pasos del algoritmo MCEM:

1. Iniciar con $\theta_0 \in \Theta$.
2. Paso S: Simular $X^1, \ldots, X^s$
3. Paso I: integración Monte Carlo para evaluar $\mathcal{A}(\theta, \theta^{(t)})$.
4. Paso M: $\theta^{(t+1)} = \arg \max_\theta \hat{\mathcal{A}}(\theta, \theta^{(t)})$.
5. Repetir hasta que $\mathbb{P}_{\theta^{(t+1)}}(Y_{1:n} = y_{1:n}) - \mathbb{P}_{\theta^{(t)}}(Y_{1:n} = y_{1:n}) > Tol$.

La simulación bajo la distribución condicional $\mathbb{P}_\theta(X_{1:n}|Y_{1:n} = y_{1:n})$ no siempre es simple, por ejemplo cuando esta distribución no se puede descomponer como el producto de marginales de dimensión menor es conveniente recurrir a métodos MCMC.

5.3. El algoritmo SAEM

Como en la sección anterior dividiremos el paso E del algoritmo EM en dos pasos: un paso de simulación y un paso de aproximación estocástica que sustituye el paso de integración del algoritmo MCEM. El algoritmo propuesto es el siguiente

1. Iniciar con $\theta_0 \in \Theta$.
2. Paso S: Simular $X^1, \ldots, X^s$
3. Paso ES: Actualizar $\mathcal{A}(\theta, \theta^{(t)})$ por la cantidad

$$\begin{aligned}&\mathcal{A}(\theta, \theta^{(t)}) \\ &\quad = \mathcal{A}(\theta, \theta^{(t-1)}) + \gamma_t \left[\frac{1}{s} \sum_{j=1}^{s} \log \mathbb{P}_\theta(Y_{1:n} = y_{1:n}, X^j) - \mathcal{A}(\theta, \theta^{(t-1)}) \right]\end{aligned}$$

   donde γ_t es una sucesión positiva.
4. Paso M: $\theta^{(t+1)} = \arg \max_\theta \hat{\mathcal{A}}(\theta, \theta^{(t)})$.
5. Repetir hasta que $\mathbb{P}_{\theta^{(t+1)}}(Y_{1:n} = y_{1:n}) - \mathbb{P}_{\theta^{(t)}}(Y_{1:n} = y_{1:n}) > Tol$.

5.3.1. Paso ES (Carter y Kohn)

En esta sección describimos el método de simulación que usamos en el algoritmo SAEM. Para muestrear la distribución condicional

$$p_\theta(x_{1:n}|y_{0:n}) = \lambda_{x_1} p(y_1|y_0, x_1) \dots a_{x_{n-1}x_n} p(y_n|y_{n-1}, x_n)/p_\theta(y_{1:n}|y_0),$$

para todo $x_{1:n} \in \{1, \dots, m\}^N$. Carter y Kohn en [13] proponen un método de muestreo que es una versión estocástica del algoritmo forward-backward propuesto por Baum *et al.* [8]. Esto se tiene observando que $p_\theta(x_{1:n}|y_{0:n})$ admite la descomposición,

$$p_\theta(x_{1:n}|y_{0:n}) = p_\theta(x_n|y_{0:n}) \prod_{k=1}^{n-1} p_\theta(x_k|x_{k+1}, y_{0:n}).$$

Dado X_{k+1} conocido, $p_\theta(X_k|X_{k+1}, y_{0:n})$ es una distribución discreta, lo cual nos sugiere la siguiente estrategia de muestreo. Para $k = 2, \dots, n$, $i \in \{1, \dots, m\}$, calculamos recursivamente el filtro óptimo $p(X_k|y_{0:k}, \theta)$ como

$$p(X_n = i|y_{0:k}, \theta) \propto p_\theta(y_k|y_{k-1}, X_k = i) \sum_{i=1}^{m} p_{ij} p(X_{k-1} = j|y_{1:k}, \theta).$$

Entonces, muestreamos X_n de $p(X_n|y_{0:n}, \theta)$ y para $k = n-1, \dots, 1$, X_k se muestrea de

$$p(X_k = i|X_{k+1} = x_{k+1}, y_{0:k}, \theta) = \frac{a_{ix_{k+1}} p(X_k = i|y_{0:k}, \theta)}{\sum_{l=1}^{m} a_{il} p(X_k = l|y_{0:k}, \theta)}.$$

Este procedimiento genera una cadena de Markov $\{x_{1:n}^{(t)}\}_{t\geq 1}$ ergódica en el espacio de estados finito $\{1, \dots, m\}^N$, tal que $p(x_{1:n}|y_{0:n}, \theta)$ es su distribución estacionaria. La ergodicidad sigue demostrando irreductibilidad y aperiodicidad, para esto observamos que el núcleo de transición Q de la cadena simulada es positivo,

$$Q\left(x_{1:n}^{(t)}|x_{1:n}^{(t-1)}, \theta\right) \propto \quad p\left(x_{1:n}^{(t)}|y_{0:n}, \theta\right) \prod_{n=1}^{N-1} p\left(x_{1:n}^{(t)}|x_{1:n}^{(t)}, y_{0:n}, \theta\right) > 0.$$

En este caso por teoremas clásicos de cadenas de Markov finitas (Kemeny y Snell [28]) se satisface que,

$$\left\| Q\left(x_{1:n}^{(t+1)}, x_{1:n}^{(t)}, \theta\right) - p(X_{1:n}|y_{0:n}, \theta) \right\| \leq C\rho^{t-1}, \tag{5.14}$$

con $C = card(\{1,\ldots,m\}^N)$, $\rho = (1 - 2K_x^*)$ y $K^* = \text{ínf}\, K(x'|x,\theta)$, para $x, x' \in \{1,\ldots,m\}^N$.

Ejemplo 5.4. *Cadenas de Markov ocultas y MS-AR gaussianos*

La verosimilitud completa del modelo (5.8), pertenece a la familia de distribuciones exponecial. En este caso, el paso EA se sustituye por una aproximación de tipo Robins-Monro (ver Duflo [20]) para estadísticos suficientes $S(X_{1:n})$ de la cadena de Markov oculta, definidos por

$$\hat{S}^{(t)} = \hat{S}^{(t-1)} + \gamma_t(S(x_1^{n,(t)}) - \hat{S}^{(t-1)}). \tag{5.15}$$

En nuestro caso $S = (S_1, S_2, S_3)$, donde:

- $S_1(X_{1:n}) = [\mathbb{1}_i(X_k)]_{1\le i\le m, 1\le k\le n}$.
- $S_2(X_{1:n}) = (n_1(X_{1:n}),\ldots,n_m(X_{1:n}))$.
- $S_3(X_{1:n}) = [n_{ij}(X_{1:n})]_{1\le i,j\le m}$.

El paso de maximización, cuando $\rho_i = 0$, está dado por,

$$\begin{aligned}
\widehat{a}_{ij}^{(t+1)} &= \frac{S_3^{(t+1)}[i,j]}{S_2^{(t+1)}(i)} \\
\widehat{b}_i^{(t+1)} &= \frac{\sum_{k=1}^n S_1^{(t+1)}[i,k] y_n}{S_2^{(t+1)}(i)} \\
\widehat{\sigma^2}_i^{(t+1)} &= \frac{1}{n}\sum_{k=1}^n S_1^{(t+1)}[i,k]\left(y_n - b_i^{(t+1)}\right)^2,
\end{aligned}$$

y para $\rho_i \neq 0$,

$$\begin{aligned}
\widehat{a}_{ij}^{(t+1)} &= \frac{S_3^{(t+1)}[i,j]}{S_2^{(t+1)}(i)} \\
\widehat{\rho}_i^{(t+1)} &= \frac{\sum_{k=1}^{n-1} S_1^{(t+1)}[i,n] y_k y_{k-1} - \sum_{k=1}^{n-1} S_1^{(t+1)}[i,k] y_k \sum_{k=1}^{N} S_1^{(t+1)}[i,k] y_{k-1}}{\sum_{k=1}^{n-1} S_1^{(t+1)}[i,k] y_{k-1}^2 - \left(\sum_{k=1}^{n-1} S_1^{(t+1)}[i,k] y_k \right)^2} \\
\widehat{b}_i^{(t+1)} &= \sum_{k=1}^{n-1} S_1^{(t+1)}[i,k] y_k - \widehat{\rho}_i \sum_{k=1}^{n} S_1^{(t+1)}[i,k] y_{k-1} \\
\widehat{\sigma^2}_i^{(t+1)} &= \frac{1}{n} \sum_{k=1}^{n} S_1^{(t+1)}[i,k] \left(y_n - \rho_i y_{n-1} - b_i^{(t+1)} \right)^2
\end{aligned}$$

Al considerar fijas las observaciones $y_{1:n}$ *las expresiones anteriores definen de forma explícita, en cada uno de los dos casos de estudio, la aplicación* $\widehat{\theta} = \theta(S)$ *entre los estadísticos suficientes y el espacio de parámetros. Esta es necesaria para el estudio de convergencia del algoritmo SAEM.*

5.3.2. Ejemplos numéricos

Ilustramos el comportamiento de los métodos de estimación considerando algunos datos simulados, estos fueron extraidos de Rodríguez [42]. Trabajamos con un CMO y dos AR-RM. Estimamos el número de estados utilizando el Critero de Información Bayesiano (BIC), este método considera la estimación por máxima verosimilitud penalizada (MVP), utilizando como función de penalidad $pen = \frac{\log(N)}{2} dim(\Psi_m)$. Con respecto a la consistencia de métodos de verosimilitud penalizada el lector interesado puede consultar R. Ríos y L. A. Rodríguez [41].

Para evaluar la función de verosimilitud en cualquier parámetro ψ se calcula

$$p(y_{1:N}|y_0\psi) = \sum_{i=1}^{m} \iota_N(i),$$

donde $\iota_n(i) = p(y_{1:n}, X_n = i)$ se evalúia recursivamente con la siguiente fórmula forward de Baum,

$$\iota_n(j) = \sum_{i=1}^{m} \iota_{n-1}(i) p_{ij} p(y_n | y_{n-1}, X_n = i)$$

ver D. Le Nhu *et al.* [37].

En la siguiente secciones describimos los ejemplos y mostramos los resultados.

5.3.3. HMMs

En la simulación del modelo CMO tenemos los siguiente parámetros: $dim(\Psi_m) = m^2 + 1$ $N = 500$, $m = 3$, $\sigma^2 = 1{,}5$, $\theta = (-2,\ 1,\ 4)$,

$$P = \begin{pmatrix} 0.9 & 0.05 & 0.05 \\ 0.05 & 0.9 & 0.05 \\ 0.05 & 0.05 & 0.9 \end{pmatrix},$$

la serie observada es graficada en la Figura 5.1.

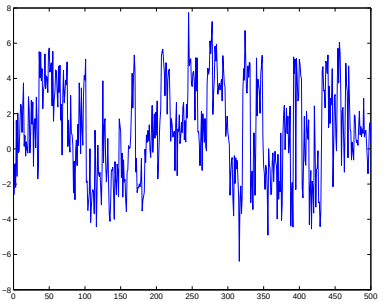

Figura 5.1: Serie observada $y_1, \ldots, y_{500}$ en el caso CMO.

El Cuadro 5.1 contiene los valores al evaluar el máximo de la verosimilitud penalizada para $m = 2, \ldots, 7$, observemos que $\hat{m} = 3$.

m	$-l(\psi)$	pen	$-l(\psi)+pen$
2	802.32	15.53	817.85
3	419.09	31.07	450.16
4	417.70	52.82	470.52
5	464.70	80.78	545.48
6	445.89	114.97	560.86
7	436.26	155.36	591.62

Cuadro 5.1: Evaluación de MVP

En este caso $\hat{\psi}$ lo estimamos utilizando SAEM, donde los valores obtenidos son, $\hat{\sigma^2} = 1{,}49$, $\hat{\theta} = (-1{,}98,\ 4{,}09,\ 0{,}91)$,

$$\hat{P} = \begin{pmatrix} 0{,}8650 & 0{,}0274 & 0{,}1076 \\ 0{,}0404 & 0{,}8943 & 0{,}0653 \\ 0{,}0658 & 0{,}0648 & 0{,}8694 \end{pmatrix},$$

en la Figura 5.2 graficamos la sucesión $\{\psi^{(t)}\}$, $t = 1, \ldots, 4000$ y observamos la convergencia de los estimados.

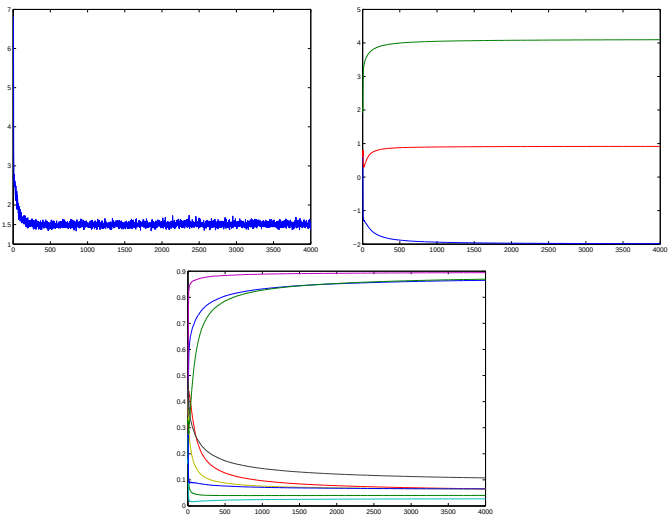

Figura 5.2: Convergencia de los estimados, σ^2, θ y P.

5.3.4. AR-RM

En el primer proceso AR-RM que simulamos elegimos los parámetros: $dim(\Psi_m) = m(m+1)+1$, $N = 500$, $m = 2$, $\sigma^2 = 1.5$,

$$\theta = \begin{pmatrix} 1 & -1 \\ -0.5 & 0.5 \end{pmatrix}, \quad P = \begin{pmatrix} 0.9 & 0.1 \\ 0.1 & 0.9 \end{pmatrix},$$

la serie observada es graficada en la Figura 5.3.

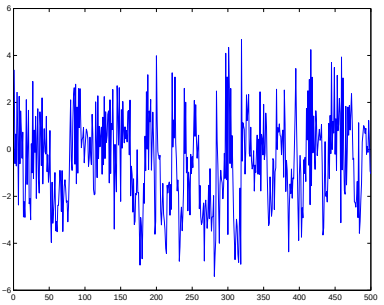

Figura 5.3: Serie observada $y_1, \ldots, y_{500}$ para el proceso AR-RM

El Cuadro 5.2 contiene los valores para el MVP para $m = 2, \ldots, 6$, observemos que $\hat{m} = 2$. En este caso $\hat{\psi}$ fue estimado utilizando SAEM, cuyos valores son, $\hat{\sigma^2} = 1{,}42$,

$$\hat{\theta} = \begin{pmatrix} 1.07 & -0.96 \\ -0.5 & 0.5 \end{pmatrix} \quad \hat{P} = \begin{pmatrix} 0{,}8650 & 0{,}1350 \\ 0{,}1130 & 0{,}8870 \end{pmatrix},$$

en la Figura 5.4 graficamos la sucesión $\{\psi^{(t)}\}$, $t = 1, \ldots, 1000$ y observamos la convergencia de los estimados.

m	$-l(\psi)$	pen	$-l(\psi)+pen$
2	351.14	18.64	369.78
3	346.64	37.28	383.92
4	355.10	64.14	417.24
5	354.52	93.21	447.73
6	384.50	130.50	515.00

Cuadro 5.2: Evaluación de MVP

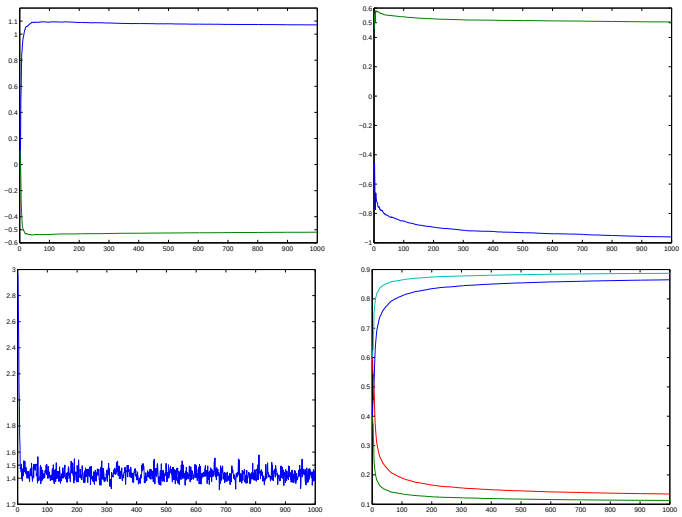

Figura 5.4: Convergencia de los estimados, θ_1, θ_2, σ^2, y P.

En la segunda simulación del AR-RM elegimos los parámetros: $N = 500$, $m = 2$, $\sigma^2 = 1.5$,

$$\theta = \begin{pmatrix} 1 & -2 \\ -0.7 & 1.08 \end{pmatrix} \quad P = \begin{pmatrix} 0.9 & 0.1 \\ 0.1 & 0.9 \end{pmatrix}$$

la gráfica de la serie se observa en la Figura 5.5. En este caso observemos que una de las pendientes de las rectas de regresión tiene coeficiente $\rho_i > 1$ que en el caso de un proceso autorregresivo de orden 1 implicaría la inestabilidad del mismo. Este ejemplo muestra la versatilidad de los AR-

RM porque permiten modelar series temporales que son heterogéneas y volátiles por trozos.

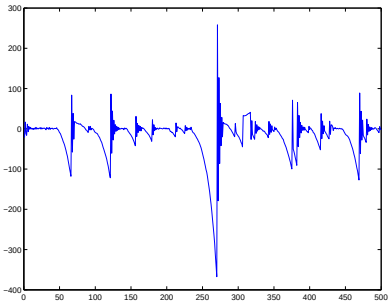

Figura 5.5: Serie observada $y_1, \ldots, y_{500}$ para el AR-RM.

Para este ejemplo $m = 2$ es fijo y $\hat{\psi}$ es estimado usando SAEM, los valores son, $\hat{\sigma^2} = 1{,}42$,

$$\hat{\theta} = \begin{pmatrix} 0.85 & -2.01 \\ -0{,}69 & 1{,}08 \end{pmatrix} \quad \hat{P} = \begin{pmatrix} 0{,}9093 & 0{,}0907 \\ 0{,}019 & 0{,}9181 \end{pmatrix}.$$

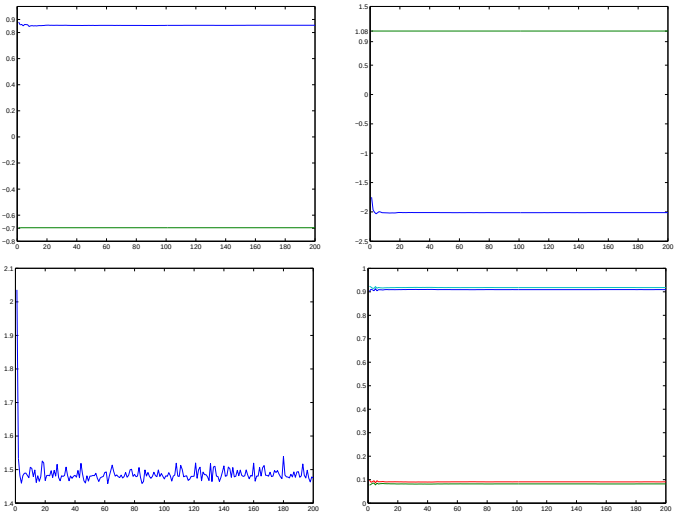

Figura 5.6: Convergencia de los estimados, θ_1, θ_2, σ^2, and P.

En la Figura 5.6 graficamos la sucesión $\{\psi^{(t)}\}$, $t = 1, \ldots, 1000$ y observamos la convergencia de los estimados.

5.4. Identidad de Fisher y Louis

Comenzaremos con propiedades generales de la matriz de información. Sea V una variable aleatoria con densidad g. Definimos el logaritmo de la función de verosimilitud como $l(\theta) = \log g_\theta(x)$, la función de score por $\nabla_\theta l(\theta)$ y la matriz de información por $I(\theta) = \mathbb{E}(\nabla^2_\theta l(\theta))$. Tenemos:

- La esperanza del score es cero. En efecto,

$$\mathbb{E}(\nabla_\theta l(\theta)) = \int \nabla_\theta \log g_\theta(x) g_\theta(x) \mu(dx) = \nabla_\theta \int g_\theta(x) \mu(dx) = 0.$$

- La matriz de información es la varianza del score, veámoslo

$$\nabla^2_\theta \log g_\theta(x) = \frac{\nabla^2_\theta g_\theta(x)}{g_\theta(x)} - \frac{\nabla_\theta g_\theta(x) \nabla_\theta g_\theta(x)^T}{g_\theta(x)^2}$$

  por lo tanto

$$\nabla^2_\theta \log g_\theta(x) = \frac{\nabla^2_\theta g_\theta(x)}{g_\theta(x)} - \nabla_\theta l(\theta) \nabla_\theta l(\theta)^T$$

  y como $\mathbb{E}(\frac{\nabla^2_\theta g_\theta(x)}{g_\theta(x)}) = 0$, se tiene la identidad de la matriz de información

$$I(\theta) = \mathbb{E}(\nabla^2_\theta l(\theta)) = -\mathbb{E}(\nabla_\theta l(\theta) \nabla_\theta l(\theta)^T).$$

Ahora si regresamos a un modelo de Markov oculto obtendremos identidades que relacionan la verosimilitud observado con la verosimilitud completa. Definimos $l(\theta) = \log \int \mathbb{P}_\theta(X = x, Y = y) \mu_c(dx)$ entonces

$$\begin{aligned} \nabla l(\theta) &= \frac{\int \nabla \mathbb{P}_\theta(X=x, Y=y) \mu_c(dx)}{\int \mathbb{P}_\theta(X=x, Y=y) \mu_c(dx)} && (5.16) \\ &= \int \frac{\nabla \mathbb{P}_\theta(X=x, Y=y)}{\mathbb{P}_\theta(X=x, Y=y)} \frac{\mathbb{P}_\theta(X=x, Y=y)}{\mathbb{P}_\theta(Y=y)} \mu_c(dx) && (5.17) \\ &= \int \nabla \log \mathbb{P}_\theta(X=x, Y=y) \mathbb{P}_\theta(X=x|Y=y) \mu_c(dx) && (5.18) \end{aligned}$$

entonces se obtiene la identidad de Fisher,

$$\nabla l(\theta) = \mathbb{E}_\theta(\nabla \log \mathbb{P}_\theta(X, Y)|Y = y).$$

Calculemos la matriz de información de la función de verosimilitud observada, para esto consideremos la información de Fisher de la distribución condicional $p_\theta(X|Y)$,

$$\begin{aligned} \mathbb{E}(\nabla^2 \log \mathbb{P}_\theta(X|Y)) &= \mathbb{E}\nabla^2(\log \mathbb{P}_\theta(X, Y|Y = y) - l(\theta)) \\ &= \mathbb{E}(\nabla^2 \log \mathbb{P}_\theta(X, Y)|Y = y) - \nabla^2 l(\theta), \end{aligned}$$

y al despejar

$$-\nabla^2 l(\theta) = -\mathbb{E}(\nabla^2 \log \mathbb{P}_\theta(X, Y)|Y = y) + \mathbb{E}(\nabla^2 \log \mathbb{P}_\theta(X|Y))|Y = y) \tag{5.19}$$

la identidad anterior es conocida como identidad de Louis y esta puede ser reescrita observando de la identidad de la matriz de información que

$$\mathbb{E}(\nabla^2 \log \mathbb{P}_\theta(X|Y))|Y = y) = -\mathbb{E}(\nabla_\theta \log \mathbb{P}_\theta(X|Y) \nabla_\theta \log \mathbb{P}_\theta(X|Y)^T|Y = y)$$

5.5. Ejercicios

1. Maximizando la función auxiliar $\mathcal{A}$ en la forma de la ecuación (5.4) encuentre las fórmulas (5.5), (5.6) y (5.7). Ayuda: Utilice multiplicadores de Lagrange.

2. Demuestre la fórmula de la ecuación (5.9).

3. Demuestre la validez de las recursiones de Baum y Welch dadas por (5.10) y (5.11).

4. Demuestre las expresiones (5.12) y (5.13).

5. Simule un modelo de Markov oculto en un computador. Utilice las recursiones de Baum y Welch para poner en marcha la estimación utilizando los datos simulados.

Capítulo 6

Convergencia del EMV

En este capítulo estudiamos algunos aspectos relacionados con la convergencia del estimador de máxima verosimilitud en modelos de Markov Ocultos.

Recordemos que el estimador por máxima verosimilitud se define como

$$\hat{\theta}_n = \arg\max_{\theta} l_n(\theta),$$

donde $l_n(\theta) = \log L(\theta)$ y $L(\theta) = \mathbb{P}_\theta(Y_1 = y_1, \ldots, Y_n = y_n)$. Decimos que el estimador de máxima verosimilitud es consistente si $\hat{\theta}_n \to \theta_0$ cuando $n \to \infty$ c.s. Para demostrar la consistencia del estimador de máxima verosimilitud siguiendo el enfoque de Wald (1949) es suficiente:

1. Demostrar que existe una función determinística $l(\theta)$ tal que

   $$\lim_{n\to\infty} v(n) l_n(\theta) = l(\theta) \text{ c.s.}$$

   donde $v(n)$ es una sucesión de normalización que no depende de θ.

2. Dar condiciones para que $l(\theta)$ tenga un único máximo en $\theta = \theta_0$.

3. Concluimos que $\hat{\theta}_n = \arg\max_\theta l_n(\theta) \to \arg\max_\theta l(\theta) = \theta_0$, cuando $n \to \infty$.

En el caso de procesos autorregresivos con régimen de Markov, para obtener una ley fuerte de grandes números para el proceso de verosimilitud

se utilizan técnicas de procesos de Markov no homogeneos, básicamente en dos direcciones, en la primera se construye una cadena de Markov extendida (ver Rynkiewicz [46], Krishnamurthy [29]) la cual satisface un teorema ergódico para luego marginalizar y en la segunda se utilizan técnicas de aproximación al proceso de verosimilitud por procesos estacionarios y desigualdades de minorización de ciertos núcleos de transición (ver Douc *et. al.* [18]).

6.1. Consistencia del EMV

En este trabajo siguiendo el enfoque de prueba de consistencia de Ramón van Handel, [48] capítulo 7, unido a la aplicación de la propiedad α-mezclante (ver sección (4.4.2)) se puede obtener la siguiente prueba de la consistencia del estimador de máxima verosimilitud. Verifiquemos que el paso 3, para demostrar consistencia se satisface bajo convergencia uniforme.

Lema 6.1. *Supongamos que el espacio de parámetros Θ es compacto. Sea $l_n : \Theta \to \mathbb{R}$ una sucesión de funciones continuas que converge uniformemente a $l : \Theta \to \mathbb{R}$. Entonces*

$$\hat{\theta}_n = \arg\max_{\theta} l_n(\theta) \to \arg\max_{\theta} l(\theta)$$

Demostración: Como una función continua sobre un compacto alcanza su máximo existe $\theta_n \in \arg\max_\theta l_n(\theta)$ para todo n. Por otra parte se satisfacen las siguientes desigualdades

$$\begin{aligned} 0 &\leq \sup_{\theta\in\Theta} l(\theta) - l(\theta_n) = \sup_{\theta\in\Theta}(l(\theta) - l_n(\theta) + l_n(\theta)) - l(\theta_n) \\ &\leq \sup_{\theta\in\Theta}(l(\theta) - l_n(\theta)) + \sup_{\theta\in\Theta}(l_n(\theta) - l(\theta_n)) \\ &\leq \sup_{\theta\in\Theta}(l(\theta) - l_n(\theta)) + (l_n(\theta_n) - l(\theta_n)) \leq 2\sup_{\theta\in\Theta}(l(\theta) - l_n(\theta)) \to 0, \end{aligned}$$

cuando $n \to \infty$. Entonces

$$\lim_{n\to\infty} l(\theta_n) = \sup_{\theta\in\Theta} l(\theta). \tag{6.1}$$

Supongamos que para la sucesión $\{\theta_n\}$ sus puntos límites no pertenecen al conjunto $\{\tilde{\theta} : \ l(\tilde{\theta}) = \max_{\theta\in\Theta} l(\theta)\}$. Por compacidad de Θ existe una

subsucesión $\{\theta'_n\} \subset \{\theta_n\}$ que converge a θ' y tal que $\theta'_n \notin \{\tilde{\theta} : \ l(\tilde{\theta}) = \text{máx}_{\theta\in\Theta} l(\theta)\}$. Pero $l(\theta)$ es una función continua porque es límite uniforme de funciones continuas, así $l(\theta'_n) \to l(\theta') < \sup_{\theta\in\Theta} l(\theta)$ cuando $n \to \infty$, de acuerdo a (6.1) esto es una contradicción. ■

Definimos $l_n(\theta) = n^{-1} \log p_\theta(Y_{1:n}|Y_0)$, demostraremos que $l_n(\theta) \to l(\theta)$, cuando $n \to \infty$, para cada $\theta \in \Theta$, para esto será necesario expresar $p_\theta(Y_{1:n}|Y_0)$ en términos del filtro de predicción $\mathbb{P}(X_k|Y_{0:n})$.

Lema 6.2. *Sea* $\delta = \text{ínf}_{i,j=1:m} p_{ij}$. *Definimos para* $0 < l < k$

$$D_{k,l} = \log \int \int p(Y_k|Y_{0:k-1}, x_k) a_{x_{k-1},x_k} \mathbb{P}(x_{k-1}|Y_{l:k-1}) \mu_c(dx_k) \mu_c(dx_{k-1}),$$

Entonces $|D_{k,l} - D_{k,0}| \leq 2\delta^{-1}(1-\delta)^{k-1-l}$.

Demostración: Ver Rodríguez [43]. ■

Proposición 6.3. *Supongamos que el espacio de parámetros* Θ *es compacto y* $l_n(\theta) = n^{-1} \log p_\theta(Y_{1:n}|Y_0)$ *entonces* $l_n(\theta)$ *es continua y* $l(\theta) = \lim_{n\to\infty} l_n(\theta)$ *existe c.s para cada* $\theta \in \Theta$.

Demostración: Ver van Handel [48] . ■

Teorema 6.4. *Supongamos que* Θ *es un conjunto compacto. Además*

1. $\theta = \theta^*$ *si y sólo si* $\mathbb{P}_\theta = \mathbb{P}_{\theta^*}$.
2. *Para todo* $i, j \in \{1, \ldots, m\}$ *y todo* $y, y' \in \mathbb{R} \times \mathbb{R}$ *las funciones* $\theta \to p_{ij}$ *y* $p_\theta(Y_1 = y|Y_0 = y', X_1 = i)$ *son continuas.*
3. *Se satisface la siguiente condición de Lipschitz*

$$|D_k^\theta - D_k^{\theta'}| \leq K\|\theta - \theta'\|.$$

Entonces el estimador de máxima verosimilitud $\hat{\theta}_n$ *es consistente.*

Demostración: Para la consistencia es suficiente demostrar, suponiendo la condición de Lipschitz 3. que la convergencia en la proposición 6.3 es uniforme en θ. Como las funciones $l_n(\theta)$ son continuas su límite $l(\theta)$ lo es. La hipótesis $|D_k^\theta - D_k^{\theta'}| \leq c\|\theta - \theta'\|$ implica que $l_n(\theta)$ es Lipschizt

por lo tanto $l(\theta)$ también.

Como Θ es compacto, se puede cubrir por un número finito de abiertos de radio ε para cualquier $\varepsilon > 0$. Existe $\{\theta_1, \ldots, \theta_l\} \subset \Theta$ tal que cada $\theta \in \Theta$ está a distancia ε de algún punto en $\{\theta_1, \ldots, \theta_l\} \subset \Theta$. Por desigualdad triangular se tiene,

$$|l_n(\theta) - l(\theta)| \leq |l_n(\theta_k) - l_n(\theta)| + |l_n(\theta_k) - l(\theta_k)| + |l(\theta_k) - l(\theta)|$$

Por la condición de Lipschitz 3. tenemos que

$$|l_n(\theta_k) - l_n(\theta)| \leq K|\theta_k - \theta|$$

al tomar esperanza se obtiene la misma cota para $|l(\theta_k) - l(\theta)|$, así

$$\sup_{\theta \in \Theta} |l_n(\theta) - l(\theta)| \leq 2K\varepsilon + \max_{k=1:l} |l_n(\theta_k) - l(\theta_k)|$$

Como $\varepsilon > 0$ es arbitrario y desde proposición 6.3 se tiene que $l_n(\theta_k) \to l(\theta_k)$ puntualmente cuando $n \to \infty$, entonces $l_n \to l$ uniformemente c.s.

Como el estimador de máxima verosimilitud está definido por $\hat{\theta}_n = \arg\max_\theta$ y demostramos que $l_n \to l$ uniformemente, por lo tanto se sigue del Lema 4 que

$$\hat{\theta}_n \to \theta^* = \arg\max_\theta l(\theta),$$

y este valor es único en virtud de la suposición de identificabilidad. ■

6.2. Normalidad asintótica del EMV

Para demostrar la normalidad asintótica del estimador de máxima verosimilitud, la idea se basa en que el gradiente de la función se anula en su máximo y el desarrollo de Taylor siguiente permite escribir,

$$0 = \nabla_\theta l_n(\hat{\theta}_n) = \nabla_\theta l_n(\theta^*) + \nabla^2_\theta l_n(\theta^*)(\hat{\theta}_n - \theta^*) + R_n$$

al despejar en la expresión anterior y normalizar por $\sqrt{n}$, se tiene

$$\sqrt{n}(\hat{\theta}_n - \theta^*) = -(\nabla^2_\theta l_n(\theta^*))^{-1}(\nabla_\theta l_n(\theta^*) + R_n)\sqrt{n}.$$

Estos argumentos nos permiten enunciar el próximo teorema.

Teorema 6.5. *Bajo las hipótesis del teorema 6.4. Suponemos que la matriz de información asintótica de Fisher* $J(\theta^*) = var(\nabla_\theta l(\theta^*))$ *es no singular y* θ^* *pertenece al interior de* Θ*.Entonces cuando* $n \to \infty$*,*

- $\sqrt{n}\nabla_\theta l_n(\theta^*) \to N(0, J(\theta*))$*, en distribución.*
- $-(\nabla^2_\theta l_n(\theta^*)) \to J(\theta^*)$*, c.s.*
- $\sqrt{n}R_n \to 0$*, c.s.*

lo que permite concluir que $\sqrt{n}(\hat{\theta}_n - \theta^*) \to \mathcal{N}(0, J(\theta^*)^{-1})$*, en distribución.*

Para una demostración de este resultado referimos a Douc *et. al.* [18]. La demostración es muy general y abarca a procesos autorregresivos controlados por cadenas de Markov con espacio de estados compactos no necesariamente finitos. Como menciona Ailliot en [1] las hipótesis de este resultado podrían ser debilitadas en el caso de AR-RM, las integrales en Douc *et. al.* son sumas finitas que permitirían intercambios de límites y derivadas.

Una prueba de este resultado para procesos autorregresivos no lineales con régimen de Markov usando la propiedad α-mezclante en dada en Rodríguez [43].

6.3. Caso lineal y gaussiano

En lo sucesivo nos concentraremos en el caso lineal y gaussiano. Consideramos como función de verosimilitud para el conjunto de observaciones $Y_{0:n}$ y el parámetro $\theta = (\psi, \sigma^2, P)$ a la distribución condicional $p_\theta(Y_{1:n}|Y_0)$.

Recordemos la notación: $n_i = \sum_{k=1}^{n} \mathbb{1}_i(x_k)$, para cada $1 \leq i \leq m$, es el número de visitas de una realización de la cadena de Markov $\{X_n\}_{n\geq 1}$ al estado i en los primeros n pasos. $n_{ij} = \sum_{k=1}^{n-1} \mathbb{1}_{i,j}(x_{k-1}, x_k)$ es el número de transiciones de i a j en n pasos.

En virtud de la regla de la probabilidad total, la función de verosimilitud del modelo se representa como

$$p_\theta(Y_{1:n}|Y_0) = \sum_{x_{1:n}} p_\theta(Y_{1:n}, x_{1:n}|y_0) = \sum_{x_{1:n}} p_{\theta,\sigma^2}(Y_{1:n}|Y_0, x_{1:n}) p_P(x_{1:n}) \tag{6.2}$$

donde

$$p_{\theta,\sigma^2}(Y_{1:n}|Y_0, X_{1:n}) = \prod_{k=1}^{n}\prod_{i=1}^{m}\left[\frac{\exp\left(-\frac{(Y_k-\rho_i Y_{k-1}-b_i)^2}{\sigma_i^2}\right)}{\sqrt{2\pi\sigma_i^2}}\right]^{\mathbb{1}_i(X_k)}.$$

y

$$p_P(x_{1:n}) = \prod_{k=1}^{n}\prod_{i,j=1}^{m} p_{ij}^{\mathbb{1}_{i,j}(x_k,x_{k+1})}.$$

Demostraremos que las hipótesis del teorema 6.4 para la consistencia son ciertas en este caso particular. Viendo directamete que el proceso de verosimilitud normalizado por $v(n) = 1/n$ es equicontinuo. Este resultado es demostrado para cadenas de Markov ocultas con espacio de estados y observaciones discretas en Finesso [22] y es extendido a AR-RM lineales gaussianos en Ríos y L. A. Rodríguez [41], la prueba del siguiente teorema es parte de este trabajo.

Teorema 6.6. *El conjunto de funciones* $f_n(\theta) = \frac{1}{n}\log p_\theta(Y_{1:n}|Y_0)$ *es una sucesión equicontinua c.s-*$\mathbb{P}_{\theta_0}$.

Con este resultado se satisfacen las hipótesis del teorema 6.4 garantizando la consistencia del estimador de máxima verosimilitud para el caso lineal y gaussiano. Suponiendo además que la matriz de información asintótica de Fisher $J(\theta^*) = \mathrm{var}(\nabla_\theta l(\theta^*))$ es no singular y θ^* pertenece al interior de Θ, entonces se obtiene desde el teorema 6.5 la gaussianidad del estimador EMV.

A continuación estudiamos el comportamiento del cociente de verosimilitud (CV) para probar la hipótesis nula de identificar un modelo CMO contra la alternativa de un proceso AR-RM. Para la prueba de hipótesis de un modelo de CMO contra un proceso AR-RM seguimos las ideas de

Giudici *et al.* [24], demostramos que la teoría asintótica del CV es válida en este caso. Consideramos la prueba

$$H_0 : \ \rho = 0$$

contra

$$H_1 : \ \rho \neq 0.$$

Teorema 6.7. $2(l(\hat{\rho}) - l(0)) \to \chi^2_1$, *bajo* $\mathbb{P}_0$.

El teorema garantiza que podemos emplear la prueba CV para rechazar H_0 si:

$$-2(l(\hat{\rho}) - l(0)) \geq \chi^2_{1,q}$$

donde $\chi^2_{1,q}$ es el q-cuantil de la distribución χ^2_1.

6.4. Estimación del orden

Para un modelo de Markov oculto, el cardinal del conjunto de estados $E = \{1, \ldots, m\}$ se conoce como el orden del modelo. Un procedimiento que podríamos utilizar para estimar el mismo es el siguiente: para cada $m = 1, \cdots, M$ utilizamos el algoritmo EM o algún otro procedimiento que nos permita calcular el estimador de máxima verosimilitud. Obteniendo un modelo candidato $\theta^{(m)}$ para cada uno de estos cálculamos $l(\theta^{(m)})$. Lo que uno espera es que exista un m' tal que el valor $l(\theta^{(m')})$ sea máximo y elegir este con el estimador de m.

Sin embargo, este no funciona, para un modelo de Markov oculto de orden m siempre es posible duplicando un estado obtener un modelo de orden $m + 1$ que represente al de orden m. Esta es una situación típica cuando se tienen modelos estadísticos anidados. El estimador de máxima verosimilitud de m es una función creciente. En particular, un estimador del orden no existe.

Una manera de solventar esta situación es definir un estimador de m como un estimador de máxima verosimilitud penalizado, es decir

$$\hat{m}_n = arg\,\text{máx} \ \ l(\theta^{(m)}) - pen(n, m)$$

donde $pen(n, m)$ es una función estrictamente creciente en m para cada n. Si denotamos por m^* el valor verdadero del orden para $m > m^*$ de la función $l(\theta^{(m)})$ debería estabilizarse pero como la función de penalización sigue creciendo por lo tanto $l(\theta^{(m)}) - pen(n, m)$ tiene un valor máximo alrededor de $m \approx m^*$. El problema es entonces elegir la función de penalización tal que el estimador del orden sea consistente $\hat{m}_n \to m^*$, c.s.

En lo que sigue demostramos que el número de estados no es sobreestimado. En efecto,

$$\mathbb{P}(\hat{m}_n > m*) \leq \sum_{m>m^*} \mathbb{P}(\hat{m}_n = m)$$

de la definición del estimador tenemos

$$\begin{aligned} \mathbb{P}(\hat{m}_n = m) &\leq \mathbb{P}(l(\theta^{(m)}) - pen(n, m) \geq l(\theta^*) - pen(n, m^*) \\ &\leq \mathbb{P}(l(\theta^{(m)}) - l(\theta^*) \geq pen(n, m) - pen(n, m^*)) \end{aligned}$$

elegiendo $pen(n, m) = \varsigma(n)\varsigma(m)$ donde $\varsigma(n) \to 0$ y $\varsigma(m)$ es una función creciente

$$\mathbb{P}(\hat{m}_n = m) \leq \mathbb{P}\left(\frac{l(\theta^{(m)}) - l(\theta^*)}{\varsigma(n)} \geq \varsigma(m) - \varsigma(m^*)\right)$$

y si suponemos $\frac{l(\theta^{(m)}) - l(\theta^*)}{\varsigma(n)} \to 0$ como $m > m^*$ y $\varsigma(m)$ creciente entonces

$$\mathbb{P}\left(\frac{l(\theta^{(m)}) - l(\theta^*)}{\varsigma(n)} \geq \varsigma(m) - \varsigma(m^*)\right) \to 0$$

cuando $n \to \infty$ entonces c.s $\hat{m}_n \leq m^*$, es decir el estimador penalizado no sobreestima el verdadero número de estados.

Para demostrar la subestimación del estimador y concluir que $\hat{m}_n \to m^*$ procedemos como sigue

$$\mathbb{P}(\hat{m}_n < m^*) \leq \sum_{m=1}^{m^*-1} \mathbb{P}(\hat{m}_n = m)$$

de la definición del estimador tenemos que

$$\begin{aligned}\mathbb{P}(\hat{m}_n = m) &\leq \mathbb{P}(l(\theta^{(m)}) - pen(n,m) \geq l(\theta^*) - pen(n,m^*) \\ &\leq \mathbb{P}\left(\frac{l(\theta^{(m)}) - l(\theta^*)}{n} \geq \frac{pen(n,m) - pen(n,m^*)}{n}\right)\end{aligned}$$

si uniforme en θ

$$\lim_{n\to\infty} \frac{l(\theta^{(m)}) - l(\theta^*)}{n} = -\gamma$$

para una constante real $\gamma > 0$ y bajo la hipótesis de que $\varsigma(n) \to 0$ se tiene que

$$\mathbb{P}\left(\frac{l(\theta^{(m)}) - l(\theta^*)}{n} \geq \frac{pen(n,m) - pen(n,m^*)}{n}\right) \to 0$$

concluimos que c.s $\hat{m}_n \geq m^*$.

En realidad no hemos resuelto el problema porque lo realmente díficil es dar las velocidades correctas de las funciones de penalización, las cuales están fuera del alcance de nuestra presentación introductoría. Para referencias un poco más detalladas de esta temática consultar Dacunha-Castelle [15] y sus referencias y la consistencia de un estimador penalizado para el orden en procesos AR-RM gaussianos en Ríos y Rodríguez [41].

6.5. Extensiones

Para cadenas de Markov ocultas se han considerado otros tipos de contrastes diferentes del EMV por ejemplo Mevel [35] considera el siguiente contraste

$$\tilde{S}_N(\theta) = \frac{1}{N}\sum_{n=1}^{N}(Y_n - \mathbb{E}_\theta(Y_n|Y_{0:n-1}))^2. \tag{6.3}$$

El estimador por mínimos cuadrados condicional (MCC) se define como

$$\tilde{\theta} = \underset{\theta\in\Theta}{\operatorname{argmin}} \tilde{S}_N(\theta). \tag{6.4}$$

Mevel demuestra la consistencia débil y la normalidad asintótica del estimador MCC. Para procesos ARN-RM en Ríos y Rodríguez [41] demuestran su consistencia y normalidad asintótica. Este contraste tiene la limitación que es necesario conocer la esperanza condicional $\mathbb{E}_\theta(Y_n|Y_{0:n-1})$, como la esperanza condicional del contraste depende de $\{Y_n\}$, θ, $\{X_n\}$ y de la función de densidad Φ y como ésta es desconocida, $\mathbb{E}(Y_n|Y_{0:n-1})$ también lo es, por lo que el estimador $\tilde{\theta}$ no puede ser obtenido por minimización de $\tilde{S}_N(\theta)$. Pero se puede reemplazar en la ecuación (6.3) la esperanza condicional por un estimador no paramétrico basado en la muestra $y_0, \ldots, y_N$ y estimar θ_* minimizando este nuevo contraste.

El criterio de Mínimos Cuadrados Condicional Modificado (MCCM) se define entonces por

$$S_N(\theta) = \frac{1}{N}\sum_{n=1}^{N}(Y_n - \hat{\mathbb{E}}(Y_n|Y_{0:n-1}))^2 \tag{6.5}$$

donde $\hat{\mathbb{E}}(Y_n|Y_{0:n-1})$ es un estimador no paramétrico de $\mathbb{E}(Y_n|Y_{0:n-1})$ basado en $y_0, \ldots, y_N$. El estimador $\hat{\theta}$ de θ_* es

$$\hat{\theta} = \underset{\theta\in\Theta}{\operatorname{argmin}} S_N(\theta). \tag{6.6}$$

En Ríos y Rodríguez [41] se demuestra la consistencia en probabilidad de este estimador semiparamétrico y se calcula su velocidad de convergencia.

6.6. Ejercicios

1. Consideremos las funciones $f(x) = e^{-x^2}$ y

$$f_n(x) = e^{-x^2} + 2e^{-(nx-n+\sqrt{n})^2},$$

para $x \in [-1, 1]$. Demuestre que:

   a) $f_n(x) \to f(x)$ cuando $n \to \infty$ para cada $x \in [-1, 1]$.

   b) $arg\,\text{máx}_x\, f_n(x) \to 1$ cuando $n \to \infty$ y $arg\,\text{máx}_x\, f(x) = 0$.

   c) Concluya que la sucesión $\{f_n\}$ converge puntualmente sin embargo el máximo de f_n no converge al máximo de f.

2. Demuestre el teorema 6.7.

3. Para una sucesión $\{Y_n\}$ independiente, identicamente distribuida con densidad común p_θ utilizando el esquema de Wald y las hipótesis que considere conveniente demuestre la consistencia del estimador de máxima verosimilitud.

Capítulo 7

Estimación no paramétrica

En este capítulo seguimos un enfoque de estimación no paramétrico aplicado a procesos autorregresivos con régimen de Markov. En particular nos centramos en estimadores de tipo núcleos de convolución. Comenzamos introduciendo la función de distribución empírica.

La aproximación de un modelo por la función de distribución empírica se basa en la construcción de un estimado a la función de distribución por una muestra $(X_1, \ldots, X_n)$ de variables aleatorias i.i.d. con distribución común $F(t)$. La función de distribución empírica se define por:

$$\hat{F}_n(t) = \frac{\{\text{número de elementos en la muestra} \leq t\}}{n} = \frac{1}{n}\sum_{i=1}^{n} \mathbb{1}_{\{x_i \leq t\}}$$

recordemos que $\mathbb{1}_A$ denota el indicador del evento A. Para un t fijo, el indicador $\mathbb{1}_{\{x_i \leq t\}}$ es una variable aleatoria Bernoulli con parametro $p = F(t)$, de aquí que $n\hat{F}_n(t)$ es una variable aleatoria binomial con media $nF(t)$ y varianza $nF(t)(1 - F(t))$. Esto implica que $\hat{F}_n(t)$ es un estimador insesgado de $F(t)$.

Algunas propiedades de la función de distribución:

- Como consecuencia de la ley fuerte de grandes números, el estimador $\hat{F}_n(t)$ converge a $F(t)$ cuando $n \to \infty$ casi seguramente, para cada valor de t:

$$\hat{F}_n(t) \xrightarrow{a.s.} F(t),$$

- Existe un resultado más fuerte, el teorema de Glivenko-Cantelli, el cual asegura la convergencia uniforme en t, cuando $n \to \infty$,:

$$\|\hat{F}_n - F\|_\infty \equiv \sup_{t\in\mathbb{R}} \left|\hat{F}_n(t) - F(t)\right| \xrightarrow{a.s.} 0.$$

- La distribución asintótica puede ser caracterizada de distintas maneras. Entre estas, el teorema del límite central del estimador puntual, establece que $\hat{F}_n(t)$ tiene distribución normal con tasa de convergencia estándar $\sqrt{n}$, cuando $n \to \infty$:

$$\sqrt{n}\big(\hat{F}_n(t) - F(t)\big) \xrightarrow{d} \mathcal{N}\Big(0, F(t)\big(1 - F(t)\big)\Big).$$

Existen resultados más precisos como el Teorema de Donsker y la ley del logaritmo iterado pero no los enunciaremos en esta presentación introductoria.

Bajo la suposición de la existencia de la función de densidad, es decir $f(t) = F'(t)$ estamos interesados en construir estimadores no paramétricos de la función de densidad $f(t)$. Como f es la derivada de F podemos aproximarla por derivadas numéricas de $\hat{F}_n$, a pesar de que $\hat{F}_n$ no es una función continua, esta idea de aproximación nos llevará a construir un estimador con buenas propiedades.

Observemos que:

$$f(t) \approx \frac{F(t+h) - F(t)}{h} \approx \frac{\hat{F}_n(t+h) - \hat{F}_n(t)}{h}$$

pero

$$\frac{\hat{F}_n(t+h) - \hat{F}_n(t)}{h} = \frac{1}{hn}\sum_{k=1}^{n} \mathbb{1}_{\{X_i \le t+h\}} - \mathbb{1}_{\{X_i \le t\}} = \frac{1}{hn}\sum_{k=1}^{n} \mathbb{1}_{(0,1]}\left(\frac{t - X_k}{h}\right)$$

de donde

$$\hat{f}_n(t) = \frac{1}{hn}\sum_{k=1}^{n} \mathbb{1}_{(0,1]}\left(\frac{t - X_k}{h}\right) \tag{7.1}$$

Para ver la convergencia puntual de este estimador observamos que como $\hat{f}_n(t)$ no es un estimador insesgado debemos considerar un término de

sesgo y otro de varianza, en efecto la desigualdad triangular nos permite escribir

$$|\hat{f}_n(t) - f(t)| \leq |f(t) - \mathbb{E}(\hat{f}_n(t))| + |\hat{f}_n(t) - \mathbb{E}(\hat{f}_n(t))|$$

y entonces la convergencia se reduce a controlar el término de sesgo $|f(t) - \mathbb{E}(\hat{f}_n(t))|$ que es una cantidad determinística y la varianza de la cantidad aleatoria $|\hat{f}_n(t) - \mathbb{E}(\hat{f}_n(t))|$.

Para el término de sesgo se tiene que

$$\mathbb{E}(\hat{f}_n(t)) = \mathbb{E}\left(\frac{\hat{F}_n(t+h) - \hat{F}_n(t)}{h}\right) = \frac{F(t+h) - F(t)}{h}$$

de donde

$$|f(t) - \mathbb{E}(\hat{f}_n(t))| = \left| f(t) - \frac{F(t+h) - F(t)}{h} \right| \to 0$$

cuando $h \to 0$. Mientras que para la cantidad aleatoria se tiene por la desigualdad de Chebyshev,

$$\mathbb{P}(|\hat{f}_n(t) - \mathbb{E}(\hat{f}_n(t))| > \varepsilon) \leq \frac{\mathrm{Var}(\hat{f}_n(t))}{\varepsilon^2}$$

y

$$\begin{aligned} \mathrm{Var}(\hat{f}_n(t)) &= \frac{1}{nh}\left(\frac{F(t+h) - F(t)}{h}(1 - (F(t+h) - F(t))\right) \\ &= \frac{f(t)}{nh} + O\left(\frac{1}{nh}\right), \end{aligned}$$

este último término tiende a cero cuando $nh \to \infty$ y $n \to \infty$.

Si en el estimador definido en (7.1) definimos $K(t) = \mathbb{1}_{(0,1]}(t)$ obtenemos

$$\hat{f}(t) = \frac{1}{nh}\sum_{k=1}^{n} K\left(\frac{t - X_k}{h}\right).$$

Podemos elegir una clase de funciones K con buenas propiedades, así obtenemos la clase de estimadores de la densidad por núcleos de convolución. Cuando el núcleo esta definido por $K(t) = \mathbb{1}_{(1/2,1/2]}(t)$ se conoce

como el *núcleo ingenuo* y es introducido por Rosenblatt en [45].

Si consideramos una muestra bidimensional $(Y_1, X_1), \ldots, (Y_n, X_n)$ para la cual existe una versión de la regresión de Y sobre X,

$$r(x) = \mathbb{E}(Y|X = x),$$

el estimador de la función de densidad de una distribución de probabilidad puede ser utilizado para construir un estimador de la función r. En efecto, la esperanza condicional considerada se expresa como

$$\mathbb{E}(Y|X = x) = \int y\mathbb{P}(Y = y|X = x)dy \tag{7.2}$$

y

$$\mathbb{P}(Y = y|X = x) = \frac{f(y, x)}{f(x)}$$

donde $f(y, x)$ es la densidad conjunta del vector (Y, X) y $f(x)$ es la densidad marginal de X, es decir $f(x) = \int f(y, x)dy$. Si se propone como estimador de la densidad conjunta,

$$\hat{f}(y, x) = \frac{1}{nh^2} \sum_{k=1}^{n} K\left(\frac{y - Y_k}{h}\right) K\left(\frac{x - X_k}{h}\right)$$

y $\hat{f}(x) = \frac{1}{nh} \sum_{k=1}^{n} K\left(\frac{x-X_k}{h}\right)$ el estimador de f. Al sustituir en la ecuación (7.2), se obtiene

$$\hat{r}(x) = \frac{\sum_{k=1}^{n} Y_k K\left(\frac{x-X_k}{h}\right)}{\sum_{k=1}^{n} K\left(\frac{x-X_k}{h}\right)},$$

este estimador de la función de regresión r fue introducido de manera independiente por Naradaya [36] y Watson [49] en 1964, y es conocido en la literatura como el estimador de Nadaraya-Watson.

Para cadenas de Markov ocultas los estimadores tipo núcleos de convolución son introducidos por Harel y Puri [26] estudiando su consistencia. En lo que sigue para procesos autorregresivos no lineales con régimen de Markov se introducen estimadores de tipo núcleo.

Supongamos una muestra de los datos completos $\{Y_k, X_k\}_{k=1:n}$ de un proceso ARN-RM definido por la ecuación (4.2), la cantidad de interés es la función de autorregresión $r(y) = \mathbb{E}(Y_1|Y_0 = y)$, esta puede ser escrita como

$$r(y) = \sum_{i=1}^{m} \mathbb{E}(Y_1|Y_0 = y, X_1 = i)\mathbb{P}(X_1 = i),$$

de aquí que sea suficiente estimar las funciones de autorregresión en cada régimen

$$r_i(y) = \mathbb{E}(Y_1|Y_0 = y, X_1 = i), \tag{7.3}$$

para $i = 1, \ldots, m$ y $y \in \mathbb{R}$.

Sea

$$g_i(y) := r_i(y)f_i(y), \tag{7.4}$$
$$f_i(y) := \mu_i p(Y_0 = y). \tag{7.5}$$

El estimador de Nadaraya-Watson de r_i se define por

$$\hat{r}_i(y) = \begin{cases} \hat{g}_i(y)/\hat{f}_i(y) & \text{if } \hat{f}_i(y) \neq 0, \\ 0 & \text{en otro caso} \end{cases} \tag{7.6}$$

con

$$\hat{g}_i(y) := \frac{1}{nh} \sum_{k=0}^{n-1} Y_{k+1} K_h\left(y - Y_k\right) \mathbb{1}_i(X_{k+1}), \tag{7.7}$$

$$\hat{f}_i(y) := \frac{1}{nh} \sum_{k=0}^{n-1} K_h\left(y - Y_k\right) \mathbb{1}_i(X_{k+1}), \tag{7.8}$$

y $K_h(y) = K(y/h)$.

La convergencia del estimador cociente $\hat{r}_i = \hat{g}_i(y)/\hat{f}_i(y)$ que se ha propuesto se obtiene aplicando un método utilizado por G. Collomb, ver [21], el cual estudia simultaneamente la convergencia de $\hat{g}_i(y)$ and $\hat{f}_i(y)$, cuando $n \to \infty$.

El estimador de Nadaraya-Watson $\hat{r}(y) = (\hat{r}_1(y), \ldots, \hat{r}_m(y))$, para cada y, se puede obtener como la solución de un problema de mínimos cuadrados ponderados; en nuestro caso esto consiste en encontrar el mínimo del potencial U definido por

$$U(y, Y_{1:n}, X_{1:n}, \theta) = \frac{1}{nh} \sum_{k=1}^{n} \sum_{i=1}^{m} K_h(y - Y_k) \mathbb{1}_i(X_{k+1})(Y_{k+1} - \theta_i)^2, \quad (7.9)$$

con respecto a $\theta = (\theta_1, \ldots, \theta_m)$ en un conjunto convexo abierto Θ de $\mathbb{R}^m$. Así, el estimador de regresión esta dado por

$$\hat{r}(y) = \underset{\theta \in \Theta \subset \mathbb{R}^m}{\operatorname{argmin}} \; U(y, Y_{1:n}, X_{1:n}, \theta).$$

En el caso de datos parcialmente observados, es decir, cuando no se observa $\{X_k\}_{k \geq 1}$, no se puede obtener una expresión explicita para la solución $\hat{r}(y)$. Por esta razón, consideramos un algoritmo recursivo que aproxima la solución. Nuestro enfoque aproxima el estimador $\hat{r}(y)$ por un algoritmo recursivo similar al de Robbins-Monro, [12, 20, 54]. Este involucra dos pasos: primero un paso Monte-Carlo que restaura los datos no observados $\{X_n\}_{n \geq 1}$, y un segundo paso una aproximación de Robbins-Monro con el proposito de minimizar el potencial U.

Aquí utilizamos algunas notaciones que utilizaremos.

- Para cada $1 \leq i \leq m$, $n_i = \sum_{k=1}^{n} \mathbb{1}_i(X_k)$ es el número de visitas de la cadena de Markov $\{X_k\}_{\geq 1}$ al estado i en los primeros n pasos, y $n_{ij}(X_{1:n}) = \sum_{k=1}^{n-1} \mathbb{1}_{i,j}(X_{k-1}, X_k)$ es el número de transiciones desde i a j en los primeros n pasos.

- $\psi^t = (\theta^t, P^t)$ es un vector que contiene las funciones estimadas $\theta^t = (\theta_1^t, \ldots, \theta_m^t)$ y las matriz de transición estimada P^t, en la t-esima iteración del algoritmo de Robbins-Monro.

Restauración-estimación (Robbins-Monro)

Para cada y fijo

Paso 0. Se inicia con una realización $X_{1:n}^0 = X_1^0, \ldots, X_n^0$. Cálculamos las funciones de regresión estimadas $\hat{r}^0(y) = (\hat{r}_1^0(y), \ldots, \hat{r}_m^0(y))$ de

la ecuación (7.6) en términos de los datos observados $Y_{1:n}$ y la realización $X^0_{1:n}$. Cálculamos la matriz de transición $P^0 = (p^0_{ij})_{i,j=1:m}$, by $p^0_{ij} = n_{ij}(X^0_{1:n})/n_i(X^0_{1:n})$ para $i, j = 1, \ldots, m$, la medida inicial $\varrho^0 = (\varrho^0_{1:m})$ por $\varrho^0_i = n_i(X^0_{1:n})/n$, para $i = 1, \ldots, m$. Defininos $\theta^0 = \hat{r}^0(y)$.

Para $t \geq 1$,

Paso R. Restaurar los datos no observados por una muestra $X^t_{1:n}$ de la distribución condicional $p(X_{1:n}|Y_{0:n}, \psi^{t-1})$.

Paso E. Actualizamos la estimación $\psi^t = (\theta^t, P^t)$ por

$$\theta^t = \theta^{t-1} - \gamma_t \nabla_\theta U\left(y, Y_{1:n}, X^t_{1:n}, \theta^{t-1}\right), \tag{7.10}$$

donde $\nabla_\theta U\left(y, Y_{1:n}, X^t_{1:n}, \theta^{t-1}\right) = \nabla_\theta U\left(y, Y_{1:n}, X^t_{1:n}, \theta\right)\big|_{\theta=\theta^{t-1}}$, $P^t = (p^t_{ij})_{i,j=1:m}$ dado por $p^t_{ij} = n_{ij}(X^t_{1:n})/n_i(X^t_{1:n})$, y $\varrho^t = (\varrho^t_i)_{i=1:m}$ se define por $\varrho^t_i = n_i(X^t_{1:n})/n$.

Paso A. Reducimos la varianza asintótica del algoritmo utilizando los promedios $\bar{\theta}^t = \sum_{k=1}^t \theta^k/t$ en lugar de θ^t, el cual se cálcula recursivamente por $\bar{\theta}^0 = \theta^0$, y

$$\bar{\theta}^t = \bar{\theta}^{t-1} + \frac{1}{t}\left(\theta^t - \bar{\theta}^{t-1}\right). \tag{7.11}$$

El siguiente resultado nos permite escribir el algoritmo propuesto como un algoritmo de gradiente estocástico. Sea,

$$\mathbb{E}_{\psi'}\left(U(y, Y_{1:n}, X^t_{1:n}, \theta)|\mathcal{F}_{t-1}\right) = u(y, Y_{1:n}, \theta),$$

con $\mathbb{E}_{\psi'}(\cdot) = \mathbb{E}(\cdot|Y_{0:n}, \psi')$ y $\psi' = (\theta', P') \in \mathcal{F}_{t-1}$ la σ-álgebra generada por $\{X^s_{1:n}\}_{s=1:(t-1)}$. Esta esperanza condicional es una esperanza con respecto a la función de distribución condicional $p(X^t_1 : n|Y_{0:n}, \psi')$.

Lema 7.1. *Para cada $\theta \in \Theta$ tenemos,*

$$u(y, Y_{1:n}, \theta) = \frac{1}{nh}\sum_{k=1}^{n}\sum_{i=1}^{m} K_h(y - Y_k)\mathbb{P}(X_{k+1} = i|Y_{0:n}, \theta')(Y_{k+1} - \theta_i)^2 \tag{7.12}$$

y $\mathbb{E}_{\theta'}\left(\nabla_\theta U(y, Y_{1:n}, X^t_{1:n}, \theta)|\mathcal{F}_{t-1}\right) = \nabla_\theta u(y, Y_{1:n}, \theta)$.

Demostración: Aplicando esperanza en (7.9), sigue que (7.12) es cierta. Para la segunda parte, utilizamos el hecho de que el potencial U es absolutamente integrable con respecto a la medida $\mathbb{P}(X_{1:n}^t = x|Y_{0:n}, \theta')\mu_c(dx)$ con $\mu_c(dx)$ la medida de contar definida sobre el conjunto $\{1, \ldots, m\}^n$. De acuerdo al teorema de convergencia dominada

$$\begin{aligned}\mathbb{E}_{\theta'}\left(\nabla_\theta U(y, Y_{1:n}, X_{1:n}^t, \theta)|\mathcal{F}_{t-1})\right) & \\ &= \int \nabla_\theta U(y, Y_{1:n}, x, \theta)\mathbb{P}(X_{1:n}^t = x|Y_{0:n}, \theta')\mu_c(dx) \\ &= \nabla_\theta \int U(y, Y_{1:n}, x, \theta)\mathbb{P}(X_{1:n}^t = x|Y_{0:n}, \theta')\mu_c(dx) \\ &= \nabla_\theta u(y, Y_{1:n}, \theta).\end{aligned}$$

■

Por lo tanto, el algoritmo de Restauración-Estimación es un algoritmo de gradiente estocástico que minimiza $u(y, Y_{1:n}, \theta)$ y que puede ser escrito como

$$\theta^t = \theta^{t-1} + \gamma_t\left(-\nabla_\theta u(y, Y_{1:n}, \theta^{t-1}) + \varsigma_t\right), \tag{7.13}$$

donde

$$\begin{aligned}\varsigma_t &= -\nabla_\theta U\left(y, Y_{1:n}, X_{1:n}^t, \theta^{t-1}\right) + \mathbb{E}_{\theta'}\left(\nabla_\theta U\left(y, Y_{1:n}, X_{1:n}^t, \theta^{t-1}\right)|\mathcal{F}_{t-1}\right) \\ &= -\nabla_\theta U\left(y, Y_{1:n}, X_{1:n}^t, \theta^{t-1}\right) + \nabla_\theta u(y, Y_{1:n}, \theta^{t-1}).\end{aligned}$$

Así, el algoritmo de gradiente estocástico es obtenido por perturbación del sistema gradiente

$$\dot{\theta} = -\nabla_\theta u(y, Y_{1:n}, \theta).$$

Para demostrar los resultados de convergencia introducimos las condiciones que nos permitarán obtener los resultados asintóticos.

7.1. Hipótesis generales

Sea $K : \mathbb{R} \to \mathbb{R}$, un núcleo positivo, simétrico, con soporte compacto tal que $\int K(t)dt = 1$. Suponemos que el núcleo K y la densidad Φ son funciones acotodas, i.e

B1 $\|K\|_\infty < \infty$.

B2 $\|\Phi\|_\infty < \infty$.

Bajo la condición B1, el núcleo K es de orden 2, i.e. $\int tK(t)dt = 0$ y $0 < \left|\int t^2 K(t)dt\right| < \infty$.

Denotamos por C un subconjunto compacto de $\mathbb{R}$, suponemos las siguientes condiciones de regularidad:

R1 Existe una constante $c, \beta > 0$, tal que

$$\forall y, y' \in \mathrm{C},\ |K(y) - K(y')| < c|y - y'|^\beta.$$

R2 La función de densidad de la variable aleatoria Y_0, Φ, y r_i tienen segundas derivadas continuas en el interior de C.

R3 Para todo $k \in \mathbb{N}$, las funciones

$$r_{i,k}(t,s) = \mathbb{E}(|Y_1 Y_{k+1}|\,|Y_0 = t, Y_k = s, X_1 = i, X_{k+1} = i)$$

son continuas.

Definimos $g_{2,i}(y) := f_i(y) r_{i,0}(y,y) = f_i(y)\mathbb{E}(Y_1^2 | Y_0 = y, X_1 = i)$, la cual es continua como consecuencia de la condición R3.

La sucesión $\{h_n\}_{n\geq 1}$ de números reales satisface la siguiente condición

S1 Para todo $n \geq 0$, $h_n > 0$, $\lim_{n\to\infty} h_n = 0$ y $\lim_{n\to\infty} nh_n = \infty$.

Finalmente, suponemos las siguientes condiciones de momentos:

M1 $\mathbb{E}(\exp(|Y_0|)) < \infty$ y $\mathbb{E}(\exp(|e_1|)) < \infty$.

M2 $\mathbb{E}(|Y_0|^s) < \infty$ y $\mathbb{E}(|e_1|^s) < \infty$, para algún $s > 2$.

Observación 7.2. *Obervemos que M1 implica M2, y M2 implica E5, la cual es una condición suficiente para la estabilidad del modelo ARN-RM.*

Como Y_0 y e_1 son independientes, la condición M1 implica

$$\mathbb{E}(\exp(|Y_1|)) \leq c\mathbb{E}(\exp(|Y_0|))\mathbb{E}(\exp(|e_1|)).$$

Mas aún, E3 y M2 implican que $\mathbb{E}(|Y_1|^s) < \infty$. *Esta condición también se puede obtener a partir de M1.*

La condición M1 se supone para garantizar la convergencia uniforme c.s. sobre subconjuntos compactos y la condición M2 para garantizar la convergencia puntual c.s. del estimador.

7.2. Identificabilidad

Demostramos la identificabilidad del proceso ARN-RM, siguiendo el reciente trabajo de [16] para la estimación no paramétrica de modelos de cadenas de Markov ocultas. Suponemos las siguientes condiciones:

I1 La matriz de transición $P = (p_{ij})_{i,j=1:m}$ tiene rango máximo.

I2 Las funciones $r_1, \ldots, r_m$ son diferentes c.s; es decir, si $i \neq j$ entonces $r_i(y') \neq r_j(y')$ para casi todo y'.

I3 La función de probabilidad Φ es tal que las funciones

$$\Phi(y - r_1(y')), \ldots, \Phi(y - r_m(y'))$$

son linealmente independiente; i.e.

$$\sum_{i=1}^{m} \alpha_i \Phi(y - r_i(y')) = 0, \text{ for all } y, y' \Longleftrightarrow \alpha_1 = \ldots = \alpha_m = 0.$$

I4 La función de probabilidad Φ es tal que $\Phi(y - \tilde{r}_{\tilde{k}}(y')) = \Phi(y - r_k(y'))$ para todo y si y solo si $\tilde{r}_{\tilde{k}}(y') = r_k(y')$.

Denotamos por $p^{(3)}_{P,r}$ la función de densidad de probabilidad conjunta de Y_0, Y_1, Y_2, Y_3. Observemos que si la cadena de Markov X es irreductible existe una única medida invariante μ, como consecuancia del lema 4.6 $p^{(3)}_{P,r}$ es bien definida por

$$\begin{aligned} p^{(3)}_{P,r} &= p(Y_0 = y_0) \sum_{i=1}^{m} \left(\sum_{j=1}^{m} \mu_j a_{ji} \Phi(y_1 - r_j(y_0)) \right) \otimes \Phi(y_2 - r_i(y_1)) \\ &\quad \otimes \left(\sum_{j=1}^{m} p_{ij} \Phi(y_3 - r_j(y_2)) \right), \end{aligned} \tag{7.14}$$

donde $p(Y_0 = y_0)$ es la densidad de probabilidad de Y_0, $\mu = (\mu_1, \ldots, \mu_m)$ es la distribución estacionaria de P la cual es la distribución de X_1. En el caso donde la cadena de Markov no es irreductible, no hay unicidad de la medida invariante y la distribución de X_1 tiene que ser especificada.

Proposición 7.3. *Supongamos que m es conocido, supogamos las condiciones I1-I4, P y r son identificables a partir de $p^{(3)}_{P,r}$, salvo permutaciones de los estados ocultos.*

Demostración: La demostración de esta proposición sigue de la misma idea dada en [16]. Se demuestra que, si $\tilde{P}$ es una $m \times m$ probabilidad de transición, y si las funciones de regresión $\tilde{r} = (\tilde{r}_1, \ldots, \tilde{r}_m)$ son tales que $p^{(3)}_{\tilde{P},\tilde{r}} = p^{(3)}_{P,r}$, entonces existe una permutación τ de los índices $\{1, ..., m\}$ tal que, para todo $i, j = 1, ..., m$, $\tilde{p}_{ij} = p_{\tau(i)\tau(j)}$ and $\tilde{r}_i = r_{\tau(i)}$.

De las condiciones I1 y I3, las funciones $\left(\sum_{j=1}^m \mu_j p_{ji} \Phi(y - r_j(y'))\right)_{i=1:m}$ son linealmente independiente, y también las funciones

$$\left(\sum_{j=1}^m p_{ij} \Phi(y - r_j(y'))\right)_{i=1:m}.$$

Entonces, de acuerdo a [2, Teorema 8] existe una permutación τ de los índices $\{1, \ldots, m\}$ tal que, para todo $i = 1, \ldots, m$:

$$\begin{aligned}
\sum_{j=1}^m \tilde{\mu}_j \tilde{p}_{ji} \Phi(y_1 - \tilde{r}_j(y_0)) &= \sum_{j=1}^m \mu_j p_{j\tau(i)} \Phi(y_1 - r_j(y_0)) \\
\Phi(y_2 - \tilde{r}_i(y_1)) &= \Phi(y_2 - r_{\tau(i)}(y_1)) \\
\sum_{j=1}^m \tilde{p}_{ij} \Phi(y_3 - \tilde{r}_j(y_2)) &= \sum_{j=1}^m p_{\tau(i)j} \Phi(y_3 - r_j(y_2)).
\end{aligned}$$

Ahora, usando la conmutatividad de la suma, se obtiene

$$\begin{aligned}
\sum_{j=1}^m \tilde{\mu}_j \tilde{p}_{ji} \Phi(y_1 - r_{\tau(j)}(y_0)) &= \sum_{j=1}^m \mu_{\tau(j)} p_{\tau(j)\tau(i)} \Phi(y_1 - r_{\tau(j)}(y_0)) \\
\sum_{j=1}^m \tilde{p}_{ij} \Phi(y_3 - r_{\tau(j)}(y_2)) &= \sum_{j=1}^m p_{\tau(i)\tau(j)} \Phi(y_3 - r_{\tau(j)}(y_2)).
\end{aligned}$$

Entonces, a partir de las condiciones I3 y I4, $\tilde{p}_{ij} = p_{\tau(i)\tau(j)}$, $\tilde{\mu}_j \tilde{p}_{ji} = \mu_{\tau(j)} p_{\tau(j)\tau(i)}$, y $\tilde{r}_i = r_{\tau(i)}$ c.s. ■

Observación 7.4. *La condición I2 implica la identificabilidad de las funciones de regresión r_i para casi todo y'. Sin embargo, la continuidad dada por la condición E2 asegura la identificabilidad para todo y'.*

El corolario siguiente da la identificabilidad en el caso donde la innovación $\{e_n\}$ es un ruido blanco gaussiano.

Corolario 7.5. *Supongamos que m es conocido y Φ la densidad de una distribución gaussiana con media cero y varianza σ^2. Suponemos las condiciones I1-I2, entonces P y r son identificables a partir de $p^{(3)}_{P,r}$, salvo permutaciones de los índices de los estados ocultos.*

Demostración: Dado que $\Phi(y) = \frac{1}{\sqrt{2\pi}\sigma} e^{-y^2/2\sigma}$, se puede verificar que la condición I4 se satisface. Para demostrar la condición I3, suponemos para todo y, y'

$$\sum_{i=1}^{m} \alpha_i \Phi(y - r_i(y')) = 0.$$

Entonces, para todo y', t, tenemos

$$\sum_{i=1}^{m} \alpha_i \int_{-\infty}^{+\infty} e^{ty} \Phi(y - r_i(y'))dy = \sum_{i=1}^{m} \alpha_i e^{r_i(y')t + \sigma^2 t^2/2} = 0.$$

Así,

$$\sum_{i=1}^{m} \alpha_i e^{r_i(y')t} = \sum_{k=0}^{\infty} \sum_{i=1}^{m} \alpha_i \frac{r_i^k(y')t^k}{k!} = 0 \Longrightarrow \sum_{i=1}^{m} \alpha_i r_i^k(y') = 0, \forall k \geq 0.$$

Consideremos la matriz de Vandermonde de orden $m \times m$

$$V = \begin{pmatrix} 1 & r_1(y') & \dots & r_1^{m-1}(y') \\ 1 & r_2(y') & \dots & r_2^{m-1}(y') \\ \vdots & \vdots & \ddots & \vdots \\ 1 & r_m(y') & \dots & r_m^{m-1}(y') \end{pmatrix}.$$

El determinante de la matriz de Vandermonde V se expresa como:

$$det(V) = \prod_{1 \leq i < j \leq m} (r_i(y') - r_j(y')).$$

A partir de la condición I2, $det(V) \neq 0$ para casi todo y'. Entonces el sistema de ecuaciones

$$\sum_{i=1}^{m} \alpha_i r_i^k(y') = 0, \quad \text{para} \quad k = 0, \ldots, m-1.$$

tiene una única solución, $\alpha_1 = \ldots = \alpha_m = 0$.

La identificabilidad se obtiene aplicando la proposición 7.3. ■

7.3. Consistencia para datos completamente observados

En lo que sigue, se establece la convergencia uniforme sobre compactos del estimador de Nadaraya-Watson $\hat{r}_i$ definido en (7.6). Para esto, se demostraran tres lemas técnicos. Para las demostraciones de estos lemas consultar Fermín *et. al* [30].

El primer lema permite tratar de forma unificada el comportamiento asintóticos de las variancias y covariancias de $\hat{f}_i$ y una versión truncada de $\hat{g}_i$. Los otros dos lemas dan una cota asintótica del sesgo y la varianza en términos de las funciones de regresión r_i.

Definimos $A_{n,h} \approx B_h$ por $\lim_{h\to 0} \lim_{n\to\infty} A_{n,h} = \lim_{h\to 0} B_h$, i.e. para n grande y h suficientemente pequeño $A_{n,h}$ es aproximadamente igual a B_h. Analogamente, definimos $A_{n,h} \preceq B_h$ por $\lim_{h\to 0} \lim_{n\to\infty} A_{n,h} \leq \lim_{h\to 0} B_h$, en particular escribimos $A_{n,h} \preceq B$ para denotar que B es una cota de la sucesión $A_{n,h}$, para n grande y h suficientemente pequeño.

Lema 7.6. *Supongamos que el modelo ARN-RM satisface las condiciones E1-E2, E5-E6, D1, B1-B2, S1 y R3 en el conjunto compacto C. Sea $\{M_n\}_{n\geq 1}$ una sucesión decreciente de números positivos que tiende a infinito. Sea*

$$T_{k,n} = aK_h(y - Y_k)\mathbb{1}_i(X_{k+1}) + bY_{k+1}\mathbb{1}_{\{|Y_{k+1}|\leq M_n\}}K_h(y - Y_k)\mathbb{1}_i(X_{k+1}).$$

Entonces, las siguiente proposiciones son ciertas para todo $y \in C$:

i) $\mathrm{var}(T_{0,n}) \approx h(a^2 f_i(y) + 2abg_i(y) + b^2 g_{2,i}(y)) \int K^2(z)dz + o(h^2)$.

ii) $\mathrm{cov}(T_{0,n}, T_{k,n}) \preceq h^2(a^2 c_1 + 2abc_2(y) + b^2 c_3(y)) + o(h^3)$.

iii) $\mathrm{cov}(T_{0,n}, T_{k,n}) \leq (a^2 + 2abM_n + b^2 M_n^2) 4\|K\|_\infty^2 \alpha_k$, *for any* $n > 0$.

Lema 7.7. *Supongamos que el modelo ARN-RM satisface las condiciones E1-E2, E5-E6, D1, B1-B2, S1, M2 y R2-R3 en el conjunto compacto* C. *Sea* $\{M_n\}_{n\geq 1}$ *una sucesión decreciente de números positivos que tiende a infinito, y* $\delta > 1$. *Entonces las siguientes desigualdades asintóticas son ciertas, para todo* $y \in C$*:*

i) $\mathbb{P}(|\hat{g}_i(y) - \mathbb{E}\hat{g}_i(y)| \geq \epsilon) \preceq 4\left(1 + \frac{\epsilon^2 nh}{16\delta}\right)^{-\delta/2} + c_1 \frac{16 M_n \zeta^{u_n}}{\epsilon h} + c_2 \frac{M_n^{(2-s)}}{\epsilon^2 h^2}$,

ii) $\mathbb{P}(|\hat{f}_i(y) - \mathbb{E}\hat{f}_i(y)| \geq \epsilon) \preceq 4\left(1 + \frac{\epsilon^2 nh}{16\delta}\right)^{-\delta/2} + c_3 \frac{16\zeta^{u_n}}{\epsilon h}$.

Lema 7.8. *Supongamos que el modelo ARN-RM satisface las condiciones E1, E6, D1, B1 y R2 en el conjunto compacto* C. *Entonces las siguientes proposiciones son válidas.*

i) $\sup_{y\in \mathrm{C}} |\mathbb{E}\hat{g}_i(y) - g_i(y)| = O(h^2)$.

ii) $\sup_{y\in \mathrm{C}} |\mathbb{E}\hat{f}_i(y) - f_i(y)| = O(h^2)$.

Observación 7.9. *El lema 7.6 es un resultado preliminar para probar el lema 7.7.*

Teorema 7.10. *Supongamos que el modelo ARN-RM* (4.2) *satisface las condiciones E1-E4, E6-E7, D1, B1-B2, S1 y R1-R3 en el compacto* C. *Entonces,*

i) Si $nh/\log n \to \infty$ *y la condición M2 se satisface, entonces para todo* $y \in C$

$$|\hat{r}_i(y) - r_i(y)| \to 0 \quad c.s.$$

ii) Si $nh/\log n \to \infty$ *y la condición M1 se satisface, entonces*

$$\sup_{y\in \mathrm{C}} |\hat{r}_i(y) - r_i(y)| \to 0 \quad c.s.$$

Demostración: Comenzamos con la siguiente desigualdad triangular en el conjunto de positividad de $f_i(y)$,

$$|\hat{r}_i(y) - r_i(y)| \leq |\hat{g}_i(y) - g_i(y)|\frac{1}{|\hat{f}_i(y)|} + |\hat{f}_i(y) - f_i(y)|\left|\frac{r_i(y)}{\hat{f}_i(y)}\right|, \quad (7.15)$$

lo cual implica la desigualdad

$$\begin{aligned}\sup_{y\in \mathrm{C}}|\hat{r}_i(y) - r_i(y)| \quad &\leq \quad \sup_{y\in \mathrm{C}}|\hat{g}_i(y) - g_i(y)|\frac{1}{\text{ínf}_{y\in \mathrm{C}}\,|\hat{f}_i(y)|} \\ &\quad + \sup_{y\in \mathrm{C}}|\hat{f}_i(y) - f_i(y)|\frac{\sup_{y\in \mathrm{C}}|r_i(y)|}{\text{ínf}_{y\in \mathrm{C}}\,|\hat{f}_i(y)|}.\end{aligned} \quad (7.16)$$

De acuerdo con la descomposición sesgo-varianza, la demostración se obtiene a través de los lemas 7.7 y 7.8, garantizando la positividad estricta de $\text{ínf}_{y\in \mathrm{C}}\,|\hat{f}_i(y)|$.

Así, aplicando el lema 7.7 con $\epsilon = \epsilon_0\sqrt{\frac{\log n}{nh}}$, $M_n = n^{\gamma}$, $u_n = (h\log n)^{-1}$, y δ suficientemente grande de manera que $log(n) = o(r)$, tenemos

$$\begin{aligned}&\mathbb{P}\left(|\hat{g}_i(y) - \mathbb{E}\hat{g}_i(y)| \geq \epsilon_0\sqrt{\frac{\log n}{nh}}\right) \\ &\qquad \preceq 4\exp\left(-\frac{\epsilon_0^2\log(n)}{32}\right) + c_1\frac{16n^{\gamma+\frac{1}{2}}\zeta^{u_n}}{\epsilon_0(\log n)^{\frac{1}{2}}h^{\frac{1}{2}}} + c_2\frac{n^{1+\gamma(2-s)}}{\epsilon_0^2\log(n)h} \\ &\qquad \preceq n^{-\frac{\epsilon_0^2}{32}} + c_1\frac{16n^{\gamma+\frac{1}{2}}\zeta^{u_n}}{\epsilon_0(\log n)^{\frac{1}{2}}h^{\frac{1}{2}}} + c_2\frac{n^{1+\gamma(2-s)}}{\epsilon_0^2\log(n)h}.\end{aligned}$$

Para $\vartheta = (s-2)\gamma - d - 2 > 0$ y $\frac{\epsilon_0^2}{32} = (s-2)\gamma - d - 1$ entonces,

$$\mathbb{P}\left(|\hat{g}_i(y) - \mathbb{E}\hat{g}_i(y)| \geq \epsilon_0\sqrt{\frac{\log n}{nh}}\right) \preceq cn^{-(1+\vartheta)}. \quad (7.17)$$

Del lema de Borel-Cantelli, la convergencia casi segura de $|\hat{g}_i(y) - \mathbb{E}\hat{g}_i(y)|$ a 0 es demostrada. Procedemos analogamente para obtener la convergencia casi segura de $|\hat{f}_i(y) - \mathbb{E}\hat{f}_i(y)| \to 0$.

De acuerdo al lema 7.8 tenemos

$$\begin{aligned}\inf_{y\in \mathrm{C}}|\hat{f}_i(y)| &\geq \inf_{y\in \mathrm{C}}|f_i(y)| - \sup_{y\in \mathrm{C}}|\hat{f}_i(y)-\mathbb{E}\hat{f}_i(y)| - \sup_{y\in \mathrm{C}}|\mathbb{E}\hat{f}_i(y)-f(y)| \\ &\geq \frac{1}{2}\inf_{y\in \mathrm{C}}|f_i(y)| > 0.\end{aligned}$$

Así, por los lemas previos 7.7 y 7.8 y la desigualdad (7.15) tenemos la convergencia puntual de $|\hat{r}_i(y) - r_i(y)|$.

Para demostrar la convergencia uniforme sobre un compacto C, necesitamos ver la desigualdad asintótica de tipo (7.17) para el término $\sup_{y\in \mathrm{C}}|\hat{g}_i(y)-\mathbb{E}\hat{g}_i(y)|$, y de forma análoga para $\sup_{y\in \mathrm{C}}|\hat{f}_i(y)-\mathbb{E}\hat{f}_i(y)|$, para la desigualdad (7.16). Para esto procedemos utilizando un esquema de truncamiento como en [4], supondremos la condición M1.

Sea $\Delta_k = Y_{k+1}K_h(y-Y_k)\mathbb{I}_i(X_{k+1})$ y la variable truncada $\tilde{\Delta}_k = \Delta_k \mathbb{I}_{\{|Y_{k+1}|\leq M_n\}}$. Entonces, definimos el estimador de núcleo truncado g_i por

$$\tilde{g}_i(y) = \frac{1}{nh}\sum_{k=0}^{n-1}\tilde{\Delta}_k.$$

Como $\|K\|_\infty < \infty$, tomando $M_n = M_0 \log n$, entonces tenemos

$$\mathbb{P}\left(\sup_{y\in \mathrm{C}}|\hat{g}_i(y)-\tilde{g}_i(y)|>0\right) \leq n\mathbb{P}(|Y_1|>M_0\log n) \leq \mathbb{E}(\exp(|Y_1|))n^{1-M_0}$$

y en virtud de la desigualdad de Cauchy-Schwarz y la condición R3,

$$\begin{aligned}&\sup_{y\in \mathrm{C}}\mathbb{E}\left(|\hat{g}_i(y)-\tilde{g}_i(y)|\right) \\ &\qquad\leq \frac{1}{h}\sup_{y\in \mathrm{C}}\mathbb{E}\left(|Y_1|\mathbb{I}_{\{|Y_1|>M_0\log(n)\}}K_h\left(y-Y_0\right)\mathbb{I}_i(X_1)\right) \\ &\qquad\leq \frac{1}{h}n^{-M_0/2}\mathbb{E}(\exp(|Y_1|))^{1/2}\sup_{y\in \mathrm{C}}\mathbb{E}\left(|Y_1|^2K_h^2\left(y-Y_0\right)\mathbb{I}_i(X_1)\right)^{1/2} \\ &\qquad\leq c_4\frac{n^{-M_0/2}}{h^{1/2}}.\end{aligned}$$

Ahora, reducimos los cálculos a un argumento de Chaining, ver [21, pags. 32 y 78], para el caso de estimadores de núcleos con variables acotadas.

Sea C un conjunto cubierto por un número finito ν_n de intervalos B_k con diametro $2L_n$ y centro en t_k. Entonces,

$$\begin{aligned}&\sup_{y\in \mathrm{C}} |\tilde{g}_i(y) - \mathbb{E}\tilde{g}_i(y)| \\ &\quad \leq \max_{k=1,\ldots,\nu_n} |\tilde{g}_i(t_k) - \mathbb{E}\tilde{g}_i(t_k)| + \sup_{y\in \mathrm{C}} |\tilde{g}_i(t_k) - \tilde{g}_i(y)| + \sup_{y\in \mathrm{C}} |\mathbb{E}\tilde{g}_i(y) - \mathbb{E}\tilde{g}_i(t_k)|\end{aligned}$$

Ahora, consideremos el lado derecho de la ecuación anterior. Primero, del lema 7.7

$$\begin{aligned}&\mathbb{P}\left(\max_{k=1,\ldots,\nu_n} |\tilde{g}_i(t_k) - \mathbb{E}\tilde{g}_i(t_k)| > \frac{\varepsilon_0}{2}\sqrt{\frac{\log n}{nh}}\right) \\ &\qquad \leq \sum_{k=1}^{\nu_n} \mathbb{P}\left(|\tilde{g}_i(t_k) - \mathbb{E}\tilde{g}_i(t_k)| > \frac{\varepsilon_0}{2}\sqrt{\frac{\log n}{nh}}\right) \\ &\qquad \preceq \nu_n \left(4n^{-\frac{\epsilon_0^2}{128}} + c_1 \frac{32 n^{1/2} M_n \zeta^{u_n}}{\epsilon_0 (\log n)^{1/2} h^{1/2}}\right).\end{aligned}$$

Para el segundo y tercer término, utilizamos la siguiente desigualdad obtenida de la condición R1,

$$\begin{aligned}|\tilde{g}_i(t_k) - \tilde{g}_i(y)| &\leq M_n \frac{1}{nh} \sum_{k=1}^{n} |K_h(t_k - Y_k) - K_h(y - Y_k)| \\ &\leq c\frac{M_n}{h^{1+\beta}} |y - t_k|^{\beta} \leq c\frac{M_n L_n^{\beta}}{h^{1+\beta}},\end{aligned}$$

para algunas constantes $c, \beta > 0$.

Por lo tanto,

$$\begin{aligned}&\mathbb{P}\left(\sup_{y\in \mathrm{C}} |\hat{g}_i(y) - \mathbb{E}\hat{g}_i(y)| \geq \epsilon_0 \sqrt{\frac{\log n}{nh}}\right) \\ &\quad \preceq \nu_n \left(4n^{-\frac{\epsilon_0^2}{128}} + c_1 \frac{32 n^{1/2} M_n \zeta^{u_n}}{\epsilon_0 (\log n)^{1/2} h^{1/2}}\right) + 2\mathbb{P}\left(\frac{M_n L_n^{\beta}}{h^{1+\beta}} \geq \frac{\epsilon_0}{2}\sqrt{\frac{\log n}{nh}}\right) \\ &\qquad + c_4 \frac{n^{-M_0/2}}{h^{1/2}}.\end{aligned}$$

Haciendo $L_n^{\beta} = n^{-\frac{1}{2}} h^{\frac{1}{2}+\beta} M_n^{-1}$ y $\nu_n = c_5/L_n$ se obtiene

$$\mathbb{P}\left(\sup_{y\in \mathrm{C}} |\hat{g}_i(y) - \mathbb{E}\hat{g}_i(y)| \geq \epsilon_0 \sqrt{\frac{\log n}{nh}}\right) \preceq \frac{c_5}{L_n}\left(4n^{-\frac{\epsilon_0^2}{128}} + c_1 \frac{32 n^{1/2} M_n \zeta^{u_n}}{\epsilon_0 (\log n)^{1/2} h^{1/2}}\right) + c_4 \frac{n^{-M_0/2}}{h^{1/2}}.$$

Tomando $u_n = (h \log n)^{-1}$, $\vartheta = \frac{\epsilon_0^2}{128} + \frac{1+d}{2\beta} + d - 1 > 0$, y $M_0 = 2(\vartheta+1)+d$, tenemos

$$\mathbb{P}\left(\sup_{y\in \mathrm{C}} |\hat{g}_i(y) - \mathbb{E}\hat{g}_i(y)| \geq \epsilon_0 \sqrt{\frac{\log n}{nh}}\right) \preceq c n^{-(1+\vartheta)}. \tag{7.18}$$

Así, en virtud del lema de Borel-Cantelli se verifica la convergencia casi segura del término $\sup_{y\in \mathrm{C}} |\hat{g}_i(y) - \mathbb{E}\hat{g}_i(y)|$.

La convergencia uniforme sobre compactos de $\hat{r}_i$ sigue de la misma forma como la convergencia puntual casi segura. ■

Observación 7.11. *Notemos que la demostración de la convergencia c.s, la probabilidad del término en* (7.17) *es sumable si* $\vartheta = (s-2)\gamma - d - 2 > 0$. *Esto es válido sólo si* $s > 2$, *y la restricción impuesta en la condición M2 es satisfecha.*
Desde (7.17) *and* (7.18) *se demuestra que la tasa de convergencia en el teorema 7.10 es* $\sqrt{\frac{\log n}{nh}}$.

7.4. Consistencia para el caso de datos parcialmente observados

En esta sección presentamos la demostración de la consistencia algoritmo de tipo Robbins-Monro para la estimación del modelo ARN-RM model en el caso cuando los datos son parcialmente observados. En lo que sigue detallamos cada paso del algoritmo.

Paso 0: Algoritmo SAEM

Para inicializar utilizamos el algoritmo SAEM descrito en la sección 5.3. Recordemos que en esta se supone linealidad de las funciones de regresión y gaussianidad del ruido.

Paso R: Filtro de Carter y Kohn

En el paso R utilizamos el método de Carter y Kohn descrito con detalle en la sección 5.3 apartado 5.3.1.

Paso E: Estimación

En cada iteracción de este algoritmo, evaluamos $\nabla_\theta U(y, Y_{1:n}, X_{1:n}, \theta)$ el gradiente del potencial. Para cada $1 \leq i \leq m$, calculamos las componentes

$$\frac{\partial U}{\partial \theta_i}(y, Y_{1:n}, X_{1:n}, \theta) = \hat{g}_i(y, Y_{1:n}, X_{1:n}) - \theta_i \hat{f}_i(y, Y_{1:n}, X_{1:n}),$$

y procedemos a actualizar esta cantidad. Esta manera de calcular las componentes tiene la ventaja de no evaluar el cociente directamente $\hat{r}_i$, evitando el problema de los ceros de $\hat{f}_i$.

Paso A: Promedio o Agregación

Para reducir la varianza asintótica de los parámetros estimados $\{\theta^t\}$, promediamos utilizando la técnica introducida por [38]. La idea es utilizar los promedios $\{\bar{\theta}^t\}$ definidos por $\bar{\theta}^t = 1/t \sum_{k=1}^{t} \theta^k$ en vez de $\{\theta^t\}$, este promedio se calcula recursivamente por medio de la ecuación (7.11).

El análisis de convergencia de las aproximaciones de Robbins-Monro son estudiados en [20] en el caso general. En este capítulo utilizamos un enfoque como el introducido en [12, pag. 431], para la convergencia del algoritmo de gradiente estocástico para la función de verosimilitud en modelos de Markov ocultos, considerando que en nuestro caso particular $u(\theta)$ es una funcióm continuamente diferenciable del parámetro θ, el resultado siguiente es obtenido para cada y fijo.

Teorema 7.12. *Supongamos la condición B1, que $\{\gamma_t\}$ es una sucesión de números positivos tal que*

$$\sum_t \gamma_t = \infty, \qquad \sum_t \gamma_t^2 < \infty,$$

y que la clausura del conjunto $\{\bar{\theta}^t\}$ es un subconjunto compacto de Θ. Entonces, casi seguramente, la sucesión $\{\bar{\theta}^t\}$ satisface

$$\lim_{t \to \infty} \nabla_\theta u(y, Y_{1:n}, \bar{\theta}^t) = 0.$$

Mas aún, $\lim_{t\to\infty}\bar{\theta}^t = \theta^*$ *y* $\nabla_\theta u(y, Y_{1:n}, \theta^*) = 0$, *c.s.*

Demostración: Sea $M^t = \sum_{s=0}^{t}\gamma_s\varsigma_s$. La sucesión $\{M^t\}$ es una $\mathcal{F}_t$ martingala, en efecto

$$\begin{aligned}\mathbb{E}(M^t|\mathcal{F}_{t-1}) &= \mathbb{E}(\gamma_t\varsigma_t + M^{t-1}|\mathcal{F}_{t-1})\\ &= \mathbb{E}(\gamma_t\varsigma_t|\mathcal{F}_{t-1}) + \mathbb{E}(M^{t-1}|\mathcal{F}_{t-1})\\ &= M^{t-1}.\end{aligned}$$

Además satisface $\sum_{t=1}^{\infty}\mathbb{E}(\|M^t - M^{t-1}\|^2|\mathcal{F}_{t-1}) < \infty$. Sigue entonces,

$$\mathbb{E}(\|M^t - M^{t-1}\|^2|\mathcal{F}_{t-1}) = \gamma_t^2\mathbb{E}(\|\varsigma_t\|^2|\mathcal{F}_{t-1})$$

y

$$\begin{aligned}\|\varsigma_t\|^2 &= \sum_{i=1}^{m}\left(\frac{\partial U}{\partial\theta_i}(y, Y_{1:n}, X^t, \theta^{t-1}) - \frac{\partial u}{\partial\theta_i}(y, Y_{1:n}, \theta^{t-1})\right)^2\\ &= \frac{4}{n^2h^2}\sum_{i=1}^{m}\left(\sum_{k=1}^{n}(Y_{k+1} - \theta_i^{t-1})K_h(y - Y_k)B_i^t(k)\right)^2,\end{aligned}$$

donde $B_i^t(k) = \mathbb{I}_i(X_{k+1}^t) - \mathbb{E}\left(\mathbb{I}_i(X_{k+1}^t|\mathcal{F}_{t-1})\right)$ son variables aleatorias centradas Bernoulli. Entonces, $\mathbb{E}(\|\varsigma_t\|^2|\mathcal{F}_{t-1})$ es

$$\frac{4}{n^2h^2}\sum_{i=1}^{m}\sum_{k,k'=1}^{n}(Y_{k+1} - \theta_i^{t-1})(Y_{k'+1} - \theta_i^{t-1})K_h(y - Y_k)K_h\left(y - Y_{k'}\right)\chi_i^t(k, k'),$$

con $\chi_i^t(k, k') = \mathrm{cov}(\mathbb{I}_i(X_{k+1}^t), \mathbb{I}_i(X_{k'+1}^t)|\mathcal{F}_{t-1})$. Así, por la desigualdad de Cauchy-Schwarz's tenemos

$$\chi_i^t(k, k') \le \sqrt{\mathrm{var}(B_i^t(k)|\mathcal{F}_{t-1})}\sqrt{\mathrm{var}(B_i^t(k')|\mathcal{F}_{t-1})} \le 1/4,$$

y

$$\begin{aligned}\mathbb{E}(\|\varsigma_t\|^2|\mathcal{F}_{t-1}) &\le \frac{1}{n^2h^2}\sum_{i=1}^{m}\left(\sum_{k=1}^{n}(Y_{k+1} - \theta_i^{t-1})K_h(y - Y_k)\right)^2 \quad (7.19)\\ &= \|\Psi(\theta^{t-1})\|^2,\end{aligned}$$

donde $\Psi(\theta) = (\Psi_1(\theta), \ldots, \Psi_m(\theta))$ y

$$\Psi_i(\theta) = \frac{1}{nh} \sum_{k=1}^{n} (Y_{k+1} - \theta_i) K_h(y - Y_k).$$

De la propiedad de compacidad $\|\Psi(\theta)\|^2$ es finita, por lo tanto

$$\mathbb{E}(\|M^t - M^{t-1}\|^2 | \mathcal{F}_{t-1}) \leq \|\Psi(\theta^{t-1})\|^2 \sum_{t=1}^{\infty} \gamma_t^2 < \infty.$$

En acuerdo al lema de Borel-Cantelli [12, Lemma 11.2.9], la sucesión $\{M^t\}$ tiene un limite finito c.s. y a partir del teorema [12, Theorem 11.3.2] la sucesión $\{\theta^t\}$ satisface

$$\lim_{t \to \infty} \nabla_\theta u(y, Y_{1:n}, \theta^t) = 0.$$

Usando que la función $\nabla_\theta u$ es continua demostramos que $\theta^* = \lim_{t\to\infty} \theta^t$ satisface $\nabla_\theta u(y, Y_{1:n}, \theta^*) = 0$, así por teorema de Cesàro, $\lim_{t\to\infty} \bar{\theta}^t = \theta^*$. ■

Vimos en la subsección 5.3.1 que la sucesión $\{X_{1:n}^t\}_{t\in\mathbb{N}}$ es una cadena de Markov ergódica con distribución invariante dada por la distribución $p(X_{1:n} = x_{1:n} | Y_{0:n}, \psi^*)$, la tasa de convergencia se obtiene desde la ecuación (5.14). Más aún, esta distribución invariante satisface

$$\begin{aligned} & p(X_{1:n} = x_{1:n} | Y_{0:n}, \psi^*) \\ &= \frac{\varrho_{x_1}^* p(Y_1 | Y_0, X_1 = x_1, \psi^*) \ldots p_{x_{n-1}x_n}^* p(Y_n | Y_{n-1}, X_n = x_n, \psi^*)}{p(Y_{1:n} | Y_0, \psi^*)}, \end{aligned}$$

para todo $x_{1:n} \in \{1, \ldots, m\}^N$. Donde $\psi^* = (\theta^*, P^*)$, con $\theta^* = \lim_{t\to\infty} \bar{\theta}^t$ es el límite en el teorema 7.12, $P^* = \lim_{t\to\infty} P^t$ es la probabilidad de transición de la cadena límite $X = (X_k)$; es decir $P^* = (p_{ij}^*(Y_{0:n}))_{i,j=1:}$ dada por

$$p_{ij}^*(Y_{0:n}) = p(X_{k+1} = j | X_k = i, Y_{0:n}, \psi^*),$$

y $\varrho* = (\varrho_i^*)_{i=1:m}$ con $\varrho^* = \lim_{t\to\infty} \varrho^t$; es decir,

$$\varrho_i^*(Y_{0:n}) = p(X_k = i | Y_{0:n}, \psi^*).$$

Como, para todo $t \geq 0$ tenemos $\varrho^t P^t = \varrho^t$, entonces deducimos que $\varrho^* P^* = \varrho^*$. Ahora, tomando $n \to \infty$, es fácil verificar que $P^*(Y_{0:n}) \to P$ y $\varrho^*(Y_{0:n}) \to \mu$, cuando $n \to \infty$ donde P es la matriz de transición y μ una medida invariante para la cadena de Markov X.

Por otro lado, el punto crítico $\theta^* \in \Theta \subset \mathbb{R}^m$ del gradiente de u, tiene i-esima componente θ_i^* dada por

$$\begin{aligned}
\theta_i^*(y, Y_{0:n}) &= \frac{\sum_{k=0}^{n-1} Y_{k+1} K_h(y - Y_k) \mathbb{P}(X_k = i | Y_{0:n}, \psi^*)}{\sum_{k=0}^{n-1} K_h(y - Y_k) \mathbb{P}(X_k = i | Y_{0:n}, \psi^*)} \\
&= \frac{\mathbb{E}\left[\sum_{k=0}^{n-1} Y_{k+1} K_h(y - Y_k) \mathbb{1}_i(X_k = i) | Y_{0:n}, \psi^*\right]}{\mathbb{E}\left[\sum_{k=0}^{n-1} K_h(y - Y_k) \mathbb{1}_i(X_k = i) | Y_{0:n}, \psi^*\right]} \\
&= \frac{\mathbb{E}\left[\hat{g}_i(y) | Y_{0:n}, \psi^*\right]}{\mathbb{E}\left[\hat{f}_i(y) | Y_{0:n}, \psi^*\right]}.
\end{aligned}$$

El teorema 7.10 implica que, bajo las condiciones E1-E4, E6-E7, D1, B1-B2, S1, M2 y R1-R3 en el conjunto compacto C, si $nh/\log n \to \infty$ entonces para todo $y \in C$

$$\hat{g}_i(y) \to g_i(y), \quad \hat{f}_i(y) \to f_i(y) = \mu_i p(Y_0 = y) \quad \text{c.s.}$$

Esto implica que $\mathbb{E}[\hat{g}_i(y) | Y_{0:n}, \psi^*] \to g_i(y)$, $\mathbb{E}[\hat{f}_i(y) | Y_{0:n}, \psi^*] \to f_i(y)$ c.s. Por lo tanto, $\theta_i^*(y, Y_{0:n}) \to r_i(y)$, *c.s.*, cuando $n \to \infty$. Como una consecuencia tenemos

$$\lim_{n \to \infty} \lim_{t \to \infty} \theta_i^t(y, Y_{0:n}) = r_i(y), \ a.s.$$

Observación 7.13. *Observemos que* $\int f_i(y) dy = \mu_i$. *Así, si el conjunto compacto* C *es tal que* $\mathbb{P}(Y_0 \in C) = 1$ *entonces* $\int_C \hat{f}_i(y) dy \to \mu_i$, *cuando* $n \to \infty$.

7.5. Ejemplos numéricos

En esta sección ilustramos el desempeño del algoritmo desarrollado en la sección previa y se lo aplicamos a datos simulados.

7.5.1. Ejemplo 1

En este primer ejemplo, utilizamos un modelo ARN-RM con $m = 2$ y funciones de autorregresión

$$r_1(y) = 0{,}7y + 2e^{(-10y^2)}, \quad r_2(y) = \frac{2}{1 + e^{10y}} - 1,$$

donde r_1 es una función bump r_2 es un función logística.

Estas funciones fueron consideradas por [23]. Sea Φ la densidad de una distribución gaussiana con media cero y varianza $\sigma^2 = 0{,}4$. La matriz de transición esta dada por

$$P = \begin{pmatrix} 0{,}98 & 0{,}02 \\ 0{,}02 & 0{,}98 \end{pmatrix}.$$

Utilizando una implementación en el programa Matlab de los algoritmos descritos, generamos una muestra de longitud $n = 1000$. Para cada k, simulamos X_k y a partir de esta generamos Y_k. La serie de datos simulados es graficada en la figura 7.1 (izquierda).

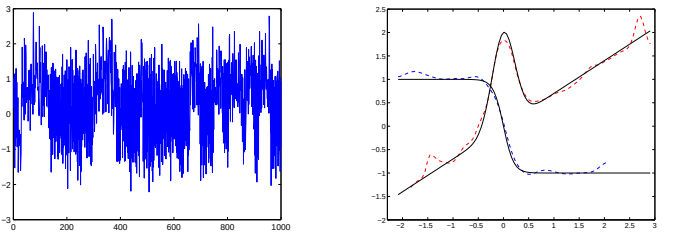

Figura 7.1: Datos simulados $Y_{1:n}$ (izquierda). Funciones estimados con el método de núcleo. $(X_{1:n}, Y_{1:n})$ (derecha).

Para la función de regresión r_i, utilizamos la densidad gaussiana estándar como núcleo K, aunque este no tiene soporte compacto. Como parámetro de ventana elegimos $h = (n/\log(n))^{1/5}$.

Suponemos que observamos los datos completos $\{Y_k, X_k\}_{k=1:1000}$, la figura 7.1 (derecha) muestra el comportamiento de r_1 y r_2 (línea continua)

y sus respectivos estimados por núcleos (línea punteada).

Ahora, suponiendo que sólo observamos $\{Y_k\}_{k=1:1000}$, entonces implementamos el algoritmo de Restauración-Estimación. En el paso 0 del algoritmo estimamos la cadena de Markov $X^0_{1:n}$ utilizando el algoritmo SAEM para el modelo MS-AR,

$$Y_n = \rho_{X_n} Y_{n-1} + b_{X_n} + \sigma_{X_n} e_n.$$

los parámetros obtenidos utilizando esta implementación son,

$$\hat{P} = \begin{pmatrix} 0{,}983 & 0{,}017 \\ 0{,}017 & 0{,}983 \end{pmatrix},$$

y la funciones lineales estimadas son de la forma $\hat{r}_1(y) = -0{,}8239y - 0{,}0218$ y $\hat{r}_2(y) = 0{,}2943y + 0{,}6334$.

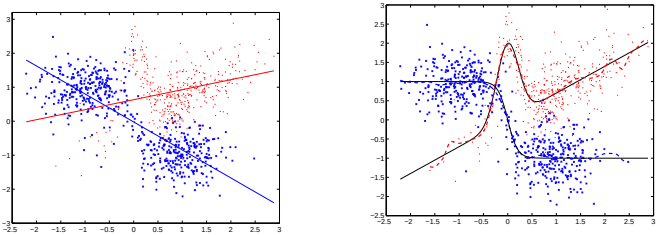

Figura 7.2: Estimados con SAEM y gráfico scatter de los datos simulados (izquierda). Estimador no paramétrico obtenido con el procedimiento de Robbins-Monro (derecha).

La figura 7.2 (izquierda) muestra el gráfico Y_k contra Y_{k-1} y el ajuste líneal.

Implementamos el algoritmo de Robbins-Monro con $t = 1 : T$ iteraciones y con tamaño de paso definido por

$$\gamma_t = \begin{cases} 1 & t \leq T_1 \\ (t - T_1)^{-1} & t \geq T_1 + 1 \end{cases}.$$

En la figura 7.2 (derecha) mostramos el gráfico de Y_k contra Y_{k-1}, r_1 y r_2 (línea solida) y los estimados obtenidos por el procedimiento de Robbins-Monro (línea punteada) para la última iteración.

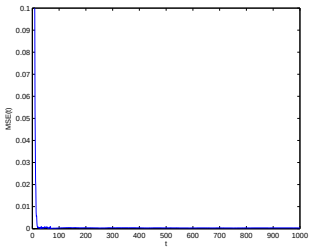

Figura 7.3: Error cuadrático medio de P^t para $t = 1 : T$.

En la figura 7.3 se muestra el error cuadrático medio de $\hat{P}^t$, podemos apreciar que la convergencia se alcanza rapidamente a una velocidad exponencial.

7.5.2. Ejemplo 2

En este ejemplo consideramos un modelo ARN-RM con $m = 3$. Las funciones autorregresivas son

$$r_1(y) = 0{,}7y + 2e^{(-10y^2)}, \;\; r_2(y) = \frac{2}{1+e^{10y}} - 1, \;\text{ and }\; r_3(y) = -2cos(y) - 1.$$

Las funciones r_1, r_2 son las mismas consideradas en el ejemplo 7.5.1. Sea Φ una densidad gaussiana con media cero y varianza $\sigma^2 = 0{,}4$, la matriz de transición esta dada por,

$$P = \begin{pmatrix} 0{,}98 & 0{,}01 & 0{,}01 \\ 0{,}01 & 0{,}98 & 0{,}01 \\ 0{,}01 & 0{,}01 & 0{,}98 \end{pmatrix}.$$

Simulamos una trayectoria de (Y, X) de tamaño $n = 3000$. Los datos simulados son graficados en la figura 7.4.

Implementamos el algoritmo de Robbins-Monro para los datos Y considerando que X es no observado. La densidad gaussiana es considerada como núcleo K, el parámetro de ventana $h = (n/\log(n))^{1/5}$, el paso $\gamma_t = t^{-0{,}6}$. En el paso 0 los estimados para la cadena de Markov $X^0_{1:n}$ son considerados como variables uniformes. Debido a la complejidad de las funciones de regresión en este ejemplo, la estimación de un modelo

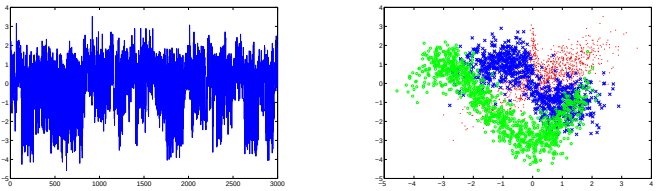

Figura 7.4: Datos simulados $Y_{1:n}$ (izquierda). Gráfico scatter de $(Y_k, Y_k - 1)$, los puntos están etiquetados con respecto al estado real de X_k: puntos rojos para $X_k = 1$, azul con x para $X_k = 2$ y verde con círculo para $X_k = 3$ (derecha).

MS-AR no es un buen estimado inicial. Para $T = 1000$, se obtiene el siguiente resultado. La matriz de transición estimada es

$$\hat{P} = \begin{pmatrix} 0{,}9665 & 0{,}0244 & 0{,}0091 \\ 0{,}0161 & 0{,}9338 & 0{,}0500 \\ 0{,}0230 & 0{,}0312 & 0{,}9458 \end{pmatrix}.$$

La figura 7.5 muestra la gráfica de Y_k contra Y_{k-1} y las funciones de regresión estimadas por el procedimiento de Robbins-Monro.

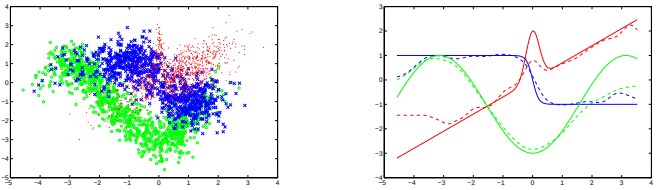

Figura 7.5: Gráfico scatter de $(Y_k, Y_k - 1)$, los puntos son etiquetados con respecto al estado estimado de X_k: puntos rojos para $X_k = 1$, azul con x para $X_k = 2$ y verde con círculo $X_k = 3$ (derecha). Estimación no paramétrica, las funciones reales son mostradas con una línea solida y los estimados por una punteada (izquierda).

En la figura 7.6, se muestra el error cuadrático medio para la matriz estimada P^t.

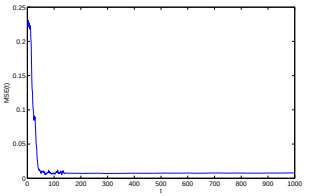

Figura 7.6: Error cuadrático medio de P^t para $t = 1 : T$.

7.5.3. Ejemplo 3

En este ejemplo se considera $m = 4$ estados. Las funciones de autorregresión son

$$r_1(y) = 0{,}7y + 2e^{(-10y^2)}, \quad r_2(y) = \frac{2}{1+e^{10y}} - 1$$

$r_3(y) = -2cos(y) - 1, \quad r_4(y) = (0{,}4y + 2{,}5)\mathbb{1}_{\{y<0\}} + (-0{,}4y + 2{,}5)\mathbb{1}_{\{y\geq 0\}}$.
Se elige Φ una densidad gaussiana con media cero y varianza $\sigma^2 = 0{,}25$, y la matriz de transición dada por

$$P = \begin{pmatrix} 0{,}9000 & 0{,}0500 & 0 & 0 \\ 0{,}0500 & 0{,}9000 & 0{,}0500 & 0 \\ 0 & 0{,}0500 & 0{,}9000 & 0{,}0500 \\ 0 & 0 & 0{,}1000 & 0{,}9000 \end{pmatrix}.$$

Simulamos una trayectoria de (Y, X) de tamaño de $n = 3000$. La serie de datos simulados $\{Y_k\}_{k=1:n}$ son gráficados en la figura 7.7.

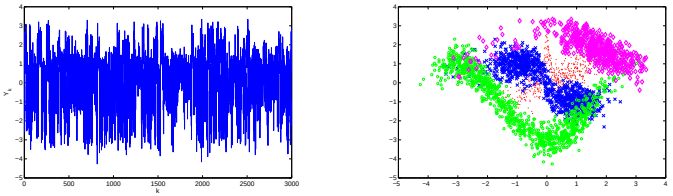

Figura 7.7: Datos simulados $Y_{1:n}$ (izquierda). Gráfico scatter de $(Y_k, Y_k - 1)$, los puntos son etiquetados con respecto al estado real de X_k: puntos rojos para $X_k = 1$, azul con x para $X_k = 2$, verde con círculo para $X_k = 3$, y mangenta con diamantes para $X_k = 4$ (derecha).

Para $T = 1000$, se obtienen los resultados siguientes. La matriz de transición estimada es

$$\hat{P} = \begin{pmatrix} 0{,}7655 & 0{,}0151 & 0{,}1439 & 0{,}0755 \\ 0{,}0406 & 0{,}7627 & 0{,}1128 & 0{,}0839 \\ 0{,}1449 & 0{,}0705 & 0{,}7558 & 0{,}0288 \\ 0{,}0656 & 0{,}0933 & 0{,}1109 & 0{,}7302 \end{pmatrix}.$$

La figura 7.8 muestra el gráfico de Y_k contra Y_{k-1} y los estimados de las funciones de regresión obtenidos por el procedimiento de Robbins-Monro.

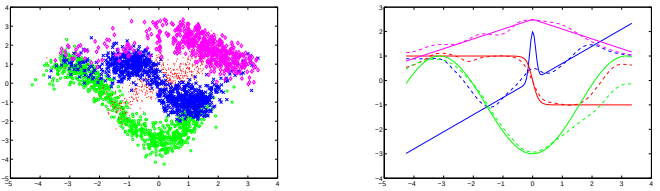

Figura 7.8: Gráfico scatter de $(Y_k, Y_k - 1)$, los puntos son etiquetados con respecto al estado real de X_k: puntos rojos para $X_k = 1$, azul con x para $X_k = 2$, verde con círculo para $X_k = 3$, y mangenta con diamantes para $X_k = 4$ (derecha). Estimación no paramétrica, las funciones reales se muestran con una línea solida y los estimados con una punteada. (izquierda).

La figura 7.9 muestra el error cuadrático medio los estimados P^t.

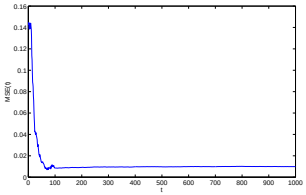

Figura 7.9: Error cuadrático medio de P^t para $t = 1 : T$.

Observamos que el comportamiento del algoritmo es bueno para los primeros dos ejemplos, es decir los casos $m = 2$ y $m = 3$ respectivamente.

Mas aún, la convergencia de la matriz de transición P^t es rapidamente alcanzada. Sin embargo, cuando el número de estados $m \geq 4$, surgen varios problemas al realizar la estimación no paramétrica. Observamos en este caso que el algoritmo tiene la dificultad de identificar los estados de la cadena de Markov oculta en los puntos donde se intersectan las funciones de regresión; es decir una perdida numérica de identificabilidad ocurre en estos puntos debido al tamaño del paso de discretización y el numéro de datos disponibles. Una mala clasificación de los datos ocurre cuando la varianza es grande con respecto al rango de las funciones de regresión. Finalmente, cuando m es grande, el algoritmo es más sensible a la elección de los parámetros de partida y el tamaño de la ventana h. En conclusión las buenas propiedades del algoritmo se pierden al incrementar el número de estados. Esto es porque el número de parámetros del modelo se incrementa y así la complejidad del modelo. Esto es la llamada maldición de la dimensionalidad.

Bibliografía

[1] P. Ailliot. Some theoretical results on Markov-switching autoregressive models with gamma innovations. *C. R. Acad. Sci. Paris, Ser. I*, 343:271–274, 2006.

[2] E.S. Allman, C. Matias, and J.A. Rhodes. Identifiability of parameters in latent structure models with many observed variables. *Ann. Statist.*, 37:3099–3132, 2009.

[3] D. Andrews. Non-strong mixing autoregressive procesess. *J. Appl Prob.*, 21:930–934, 1984.

[4] P. Ango-Nze, P. Buhlmann, and P. Doukhan. Weak dependence beyond mixing and asymptotics for nonparametric regression. *Annals of Statistics*, 30:397–430, 2002.

[5] J-G Attali. Ergodicity of a certain class of non Feller models : applications to ARCH and Markov switching models. *ESAIM: PS*, 8:76–86, 2004.

[6] J. K. Baker. *Stochastic Modeling for automatic Speech Understanding Speech Recognition.* Readings in speech recognition, 1990.

[7] L. E. Baum and T. Petrie. Statistical inference for the probabilistic functions of finite Markov chains. *Ann. Math. Stat.*, 37:1554–1563, 1966.

[8] L. E. Baum, T. Petrie, G. Soules, and N. Weiss. A maximization tecnique occuring in the statistical analysis of a probabilistic functions of Markov chains. *Ann. Math. Stat.*, 41:164–171, 1970.

[9] P. J. Bickel and Y. Ritov. Inference in hidden Markov models I: local asymptotic normality in the stationary case. *Bernoulli*, 2:199–228, 1996.

[10] P. J. Bickel, Y. Ritov, and T. Rydén. Asymptotic normality of maximum-likelihood estimator for general hidden Markov models. *The Annals of Statistics*, 26(4):1614–1635, 1998.

[11] O. Cappe. Ten years of HHMs. Preprint (online). Available: http://www-sig.enst.fr/˜cappe, 2001.

[12] O. Cappé, E. Moulines, and T. Rydén. *Inference in Hidden Markov Models* . Springer-Verlag, 2005.

[13] C. K Carter and R. Kohn. On Gibbs sampling for state space model. *Biometrika*, 81:541–553, 1994.

[14] X. Milhaud y P. Vandekerkhove D. Bakry. Statistique de chaînes de Markov cachées à espaces d'états fini. Le cas non stationnarie. *C.R. Acad. Sci. Paris*, 325-I:203–206, 1997.

[15] D. Dacunha-Castelle. Orden de un Modelo Estadístico: Estimación y Tests. *Boletín de la Asociación Matemática Venezolana*, V(1):9–27, 1998.

[16] Y. DeCastro, E. Gassiat, and C. Lacour. Minimax adaptive estimation of non-parametric hidden markov models. arXiv:1501.04787 [math.ST], 2015.

[17] A. Dempster, N. Laird, and D. Rubin. Maximum-likelihood from incomplete data via the EM algorithm. *J. Roy. Statist. Soc. Ser. B*, 39:1–38, 1977.

[18] R. Douc, E. Moulines, and T. Rydén. Asymptotic properties of the maximum likelihood estimator in autoregressive models with Markov regime. *Ann. Statist.*, 32:2254–2304, 2004.

[19] P. Doukhan. *Mixing: Propierties and Examples.*, volume 85. Lecture Notes in Statist., 1994.

[20] M. Duflo. Algorithmes Stochastiques. Springer-Verlag, Berlin, 1996.

[21] F. Ferraty, N. Antón, and P. Vieu. *Regresión No paramétrica: Desde la Dimensión Uno hasta la dimensión Infinita*. Servicio editorial de Universidad del País Vasco, 2001.

[22] L. Finesso. *Estimation of the order a finite Markov chain*. PhD thesis, University of Maryland, 1990.

[23] J. Franke, J. P. Stockis, J. Tadjuidje, and W.K. Li. Mixtures of nonparametric autoregressions. *Journal of Nonparametric Statistics*, 23(2):287–303, 2011.

[24] P. Giudici, T. Ryden, and P. Vandekerkhove. Likelihood-ratio test fir hidden markov models. *Biometrics*, pages 742–751, 2000.

[25] J.D. Hamilton. A new approach to the economic analysis of non stationary time series and the business cycle. *Econometrica*, pages 357–384, 1989.

[26] M. Harel and M. Puri. U-statistiques conditionnells universellement consistantes pour des modèles de Markov cachés. *S. R. Acad. Sci. Paris, Série I*, 333:953–956, 2001.

[27] A. Heller. On Stochastic Process Derived form Markov Chains. *Ann. Math. Stat.*, 36:1286–1291, 1965.

[28] J. G. Kemeny and J. L. Snell. *Finite Markov Chains*. Van Nostrand, Princenton, New Jersey, 1960.

[29] V. Krishnamurthy and G. G. Yin. Recursive Algorithms for estimation of hidden Markov Models with markov regime. *IEEE Trans. Information theory*, 48(2):458–476, 2002.

[30] Fermín L., Ríos, and L. A. Rodríguez. A robbins monro algorithm for nonparametric estimation of functional ar process with markov-switching: consistency. arXiv:1407.3747v5, 2014.

[31] B. G. Leroux. Maximum-likelihood estimation hidden Markov models. *Stoch. Proc. and their Appl*, 40:127–143, 1992.

[32] C. Liu and P. Narayan. An Introduction to the Application of the Theory of Probabilistic of a Markov Process to Automatic Speech Recognition . *Bell. Syst. Tech. J.*, 62:1035–1074, 1983.

[33] I.L. MacDonald and W. Zucchini. *Hidden Markov and Other Models for discrete-valued Time Series*. Chapman and Hall, 1997.

[34] C. Matias. *Estimation dans des modèles à variables cachées*. PhD thesis, Universite París XI, Orsay, París, 2001.

[35] L. Mevel. Statistique asymtotique pour les modeles Markov cachés. PhD thesis, Universite Rennes I, 1997.

[36] E.A. Nadaraya. On Estimating regression. *Theory probab. An*, 9(1):141–142, 1964.

[37] D. Le Nhu, B. G. Leroux, and M. L. Puterman. Exact Likelihood Evaluation in a Markov Mixture Model for Time Series Series of Seizure Counts. *Biometrics*, 48:317–323, 1992.

[38] B. T. Polyak and A. B. Juditsky. Acceleration of stochastic approximation by averaging. *SIAM J. Control Optim.*, 30:838–855, 1992.

[39] R. Prieto. Convergencia de Cadenas de Markov en espacio de estados generales. Universidad Central de Venezuela, Tesis Licenciatura, 2008.

[40] R. Ríos and L. A. Rodríguez. Estimación semiparamétrica en procesos autorregresivos con régimen de markov. *Divulgaciones Matemáticas*, 16(1):155–171, 2008.

[41] R. Ríos and L. A. Rodríguez. Penalized estimate of the number of states in gaussian linear ar with markov regime. *Electronic Journal of Statistics*, pages 1111–1128, 2008.

[42] L. A. Rodríguez. Algunas propiedades de procesos autorregresivos lineales y no lineales con régimen de markov. *Boletín de la Asociación Matemática Venezolana*, XXII(1):15–44, 2015.

[43] L. A. Rodríguez. Asymptotic properties of the maximum likelihood estimator for nonlinear AR processes with markov-switching. arXiv:1605.09457, 2016.

[44] R. Rosales. MCMC for hidden Markov models incorporating aggregation of states and filtering. *Bulletin of Mathematical Biology*, 66(5):1173–1199, 2004.

[45] M. Rosenblatt. Remarks on some non parametric estimates of a density function. *Ann. Math. Statist.*, 27(3):832–837, 1956.

[46] J. Rynkiewicz. Models hybrides integrant des reseaux de neurones artificiels a des modeles de chaines de Markov cachee: application a la prediction de series temporelles. PhD thesis, Universite París I, 2000.

[47] Meyn S.P. and Tweedie R.L. Markov Chain and Stochastic Stability. Springer - Verlag, London, 1993.

[48] R. v. Handel. *Hidden Markov Models.* Lecture notes: *https://www.princeton.edu/ rvan/*, 2008.

[49] G. S. Watson. Smoth regression analisys. *Sankhya*, 26(4):359–372, 1964.

[50] D. Blackwell y L. Koopmans. On the Identifability Problem for Functions of Finite Markov Chains. *Ann. Math. Stat.*, 28:1010–1015, 1957.

[51] C. Francq y M. Roussignol. Ergodicity of autoregressive process with markov-switching and consistency of the maximum likelihood estimator. *Statistics*, 32:151–173, 1998.

[52] J. L. Jensen y N. V. Petersen. Asymptotic normality of the maximum likelihood estimator in state models. *Ann. Statist*, 27:514–535, 1999.

[53] S. M. Goldfeld y R. E. Quandt. A Markov model for switching regressions. *Journal of Econometrics*, 1:3–16, 1973.

[54] J. Yao. On Recursive Estimation in Incomplete Data Models. *Statistics*, 34:27–51, 2000.

[55] J. Yao and J. G. Attali. On stability of nonlinear AR process with Markov switching. *Adv. Applied Probab*, 1999.

La Asociación Matemática Venezolana fue fundada en 1990 como una organización civil sin fines de lucro cuya finalidad es trabajar por el desarrollo de las matemáticas en Venezuela.

Asociación Matemática Venezolana
Apartado 47.898, Caracas 1041-A, Venezuela

Gobierno Bolivariano de Venezuela
Ministerio del Poder Popular para Educación Universitaria, Ciencia y Tecnología
1816-2016 PATRIOTAS UNIDOS SABEMOS VENCER