Universidad Nacional Autónoma
De México

Facultad de Ciencias

MODELOS PARA SERIES TEMPORALES VÍA MEZCLAS FINITAS

T E S I S

QUE PARA OBTENER EL TÍTULO DE:

ACTUARIO

P R E S E N T A :

ORTEGA IBAÑEZ OSCAR

Dr. RAMSÉS HUMBERTO MENA CHÁVEZ

2009

Universidad Nacional Autónoma de México
UNAM
Biblioteca Central
Dirección General de Bibliotecas de la UNAM

UNAM – Dirección General de Bibliotecas
Tesis Digitales
Restricciones de uso

DERECHOS RESERVADOS ©
PROHIBIDA SU REPRODUCCIÓN TOTAL O PARCIAL

Todo el material contenido en esta tesis esta protegido por la Ley Federal del Derecho de Autor (LFDA) de los Estados Unidos Mexicanos (México).

El uso de imágenes, fragmentos de videos, y demás material que sea objeto de protección de los derechos de autor, será exclusivamente para fines educativos e informativos y deberá citar la fuente donde la obtuvo mencionando el autor o autores. Cualquier uso distinto como el lucro, reproducción, edición o modificación, será perseguido y sancionado por el respectivo titular de los Derechos de Autor.

Hoja de Datos del Jurado

1. Datos del alumno
   Ortega
   Ibañez
   Oscar
   58 63 45 76
   Universidad Nacional Autonoma de Mexico
   Facultad de Ciencias
   Actuaría
   302094639

2. Datos del tutor
   Dr
   Ramsés Humberto
   Mena
   Chávez

3. Datos del sinodal 1
   Dr
   Alberto
   Contreras
   Cristan

4. Datos del sinodal 2
   Dra
   Silvia
   Ruiz
   Velasco Acosta

5. Datos del sinodal 3
   M en C
   Nelsón Omar
   Muriel
   Torrero

6. Datos del sinodal 4
   Mat
   Margarita Elvira
   Chávez
   Cano

7. Datos del trabajo escrito
   Modelos para series temporales vía modelos mezcla finitos
   Modelo GH-ARCH estacionario
   166 p
   2009

A los profesores que me ayudaron a hacer posible este trabajo, mi más sincero y profundo agradecimiento:

Dr. Ramsés Humberto Mena Chávez

Dra. Silvia Ruiz Velasco Acosta

M. en C. Nelsón Omar Muriel Torrero

Mat. Margarita Elvira Chávez Cano

Dr. Alberto Contreras Cristian

Índice general

NOTACIÓN

$\mathbb{N}$	Conjunto de los números naturales
$\mathbb{Z}$	Conjunto de los números enteros
$\mathbb{R}$	Conjunto de los números reales
$\mathbb{C}$	Conjunto de los números complejos
$\sum_{k=1}^{n} x_k$	Suma de los números $x_1, \ldots, x_n$
$\prod_{k=1}^{n} x_k$	Producto de los números $x_1, \ldots, x_n$
$\binom{m}{n}$	Combinaciones de m elementos de n en n
$\lvert x \rvert$	Valor absoluto del número real x
$g \circ f$	Composición de las funciones f y g
f^{-1}	Inversa de la función f
$\sin \phi$	Función seno
$\cos \phi$	Función coseno
$\tan \phi$	Función tangente
$\ln x$	Logaritmo natural de x
$\exp(x)$	Exponencial de x
$(\partial\, h(x))/(\partial x)$	Derivada parcial de la función $h(x)$
$\int h(x)\, dx$	Integral de la función $h(x)$
$\Gamma(a) = \int_0^\infty e^{-t}\, t^{x-1}\, dt$	Función Gamma
$x \cdot y$	Producto de los vectores x y y
$\lvert\lvert x \rvert\rvert$	Norma del vector x
x^T	Vector transpuesto
A^{-1}	Inversa de la matriz A
$\lvert A \rvert = \det(A)$	Determinante de la matriz A
$\mathbf{J}(A)$	Jacobiano de la matriz A
$\mathbf{H}(A)$	Hessiano de la matriz A
$(\Omega, \mathscr{F}, P)$	Espacio de probabilidad
Z	Variable aleatoria (continua o discreta)
$\mathbb{P}[Z \leq z] = F(z)$	Función de distribución de Z (caso continuo o discreto)
$\mathbb{P}[Z = z] = f(z)$	Función de densidad de Z (sólo en el caso discreto)

$\mu = \mathbb{E}[Z]$	Esperanza de Z
$\sigma^2 = \mathbb{V}[Z]$	Varianza de Z
$\mathbb{I}_{\{C\}}^{(x)}$	Función indicadora de la variable x sobre el conjunto C

Introducción

Una de las principales características que se desean en un modelo de series de tiempo es que su comportamiento sea estable con respecto al tiempo. Esto obedece a aspectos de interpretación, de parsimonia y en ocasiones para facilitar procedimientos de estimación (Tukey, 1967 y Box y Jenkins, 1976). Para presevar esta característica, en un principio, se consideró la construcción de modelos de series de tiempo cuya función de distribución es normal. Sin embargo, el uso de estos modelos se ha visto limitado debido a las complejas estructuras de dependencia observadas en distintas series. En respuesta a esto Lawrance y Lewis (1980) plantearon la construcción de modelos cuya distribución marginal es exponencial (EARMA(p,q)). Otros modelos con distribuciones marginales pertenecientes a otras familias se encuentran en Gaver y Lewis (1980) y Lawrance (1982) para mezclas de distribuciones exponenciales y para distribuciones gamma; McKenzie (1986, 1988) para distribuciones Poisson y binomial negativa; Joe(1996) para distribuciones infinitamente divisibles y cerradas bajo convolución.

En Pitt et. al.(2002) se introduce un método flexible para la construcción de modelos autorregresivos de orden uno estrictamente estacionarios con distribución marginal arbitraria, y mediante la generalización de este método (Raftery, 1985) en Mena y Walker (2007a) se presenta la construcción de modelos no lineales con distribución marginal no gaussiana. A partir de esta construcción Mena y Walker (2007b) establecen las condiciones bajo las cuales el modelo GH-ARCH es estrictamente estacionario. La manera en que se construye y estima este último modelo se considerará en el desarrollo del presente texto.

Con esto en mente, se parte de los conceptos básicos que permitan al lector potencial la fácil comprensión de los temas tratados subsecuentemente. Es así que el siguiente escrito se compone de tres partes. En la primera se describen los conceptos básicos sobre el análisis series de tiempo. Una vez que se ha estudiado la estructura básica de una serie de tiempo, se procede al análisis de las distintas representaciones de un proceso estacionario. Entre estas representaciones se destaca el uso de los modelos MA, AR, ARMA y ARCH. Posteriormente se examinarán

algunos de los métodos de estimación para los modelos ARMA(p,q) y ARCH(p), donde nos centraremos en el método de momentos y en el de máxima verosimilitud. Asimismo se presentarán los distintos criterios para la selección de un posible modelo. La finalidad de la segunda parte es presentar, de manera general, el uso de modelos de mezclas finitas para la construcción de distribuciones complejas. Partiendo de esta premisa, se define al modelo de mezclas finitas como la combinación lineal de distintas densidades. Además se examinan los distintos métodos para estimar los parámetros producidos en la mezcla, dando un especial énfasis al uso del algoritmo EM (Dempster, 1977). Por último se mencionan dos elementos básicos para la construcción del modelo GH-ARCH. El primero de estos es el uso del muestreo de Gibbs para la construcción de modelos de primer orden a través de variables latentes, y en algunos casos para la estimación de este modelo. El segundo corresponde a la descripción de la distribución hiperbólica generalizada, debido a que esta distribución es fundamental para la construcción de la probabilidad de transición del modelo GH-ARCH. En la tercera parte se presenta de manera detallada la construcción del modelo GH-ARCH estrictamente estacionario. La estimación de los parámetros producidos en este modelo se hace a través del algoritmo EM. Empero, debido a la forma que toma la probabilidad de transición, la maximización en términos análiticos en el Paso-M no es posible. Para solucionar esto último se empleo el algoritmo de Broyden-Fletcher-Goldfarb-Shanno (BFGS). Por último, para ilustrar el uso y estimación de este modelo, se analizarán las bases de datos que corresponden al precio diario de las acciones emitidas por JP MORGAN CHASE CO, CIT GROUP INC (DEL), AMER INTL GROUP NEW e INTL BUS MACHINE del 10 de julio de 2008 hasta el 10 de julio de 2009. Estos datos se eligieron debido a que su comportamiento implica una estructura de dependencia más compleja en relación con el uso de modelos lineales.

Conceptos Básicos

Capítulo 1

Conceptos Fundamentales

Considerando los conceptos expuestos en Brockwell et. al.(2002) y en William (1990), el presente capítulo tiene la finalidad de introducir algunas de las ideas básicas para el análisis de series de tiempo. En primera instancia, se define una serie de tiempo como un proceso estocástico. A partir de esta definición se estudian algunas de sus principales caraterísticas. Entre estas propiedades, se describen las condiciones bajo las cuales la serie es estacionaria en distribución, ya sea en un sentido débil o estricto; asimismo se define la media, varianza, covarianza, correlación y la función de autocorrelación parcial. La manera en que se estiman la media, varianza y autocorrelación se expone en la Sección 1.3. En la siguiente sección se plantea el modelo de series de tiempo como una mezcla de tendencias, ciclos y de componentes irregulares, además se estudian algunos de los principales métodos para estimar cada uno de estos componentes. Por último, teniendo en cuenta que algunas características de los modelos que se estudiarán en el Capítulo 2 dependen de la solución de las ecuaciones de diferencias lineales, en la Sección 1.5 se da un panorama general sobre la forma en que se obtienen las raíces de estas ecuaciones.

1.1. Proceso estocástico

Un proceso estocástico es un conjunto de variables aleatorias $\{Z_\tau\}_{\tau \in T}$ indexadas por un conjunto $\tau \in T$, que puede ser finito o infinito dependiendo del fenómeno a modelar. Por ejemplo, si

$$Z_n = Z_0 + \xi_1 + \cdots + \xi_n,$$

donde $\xi_1, \xi_2, \ldots$ son variables aleatorias con función de densidad común, y si Z_0 es una variable aleatoria independiente de las ξ_i, entonces Z_n es un proceso estocástico. Este proceso es conocido como la caminata aleatoria.

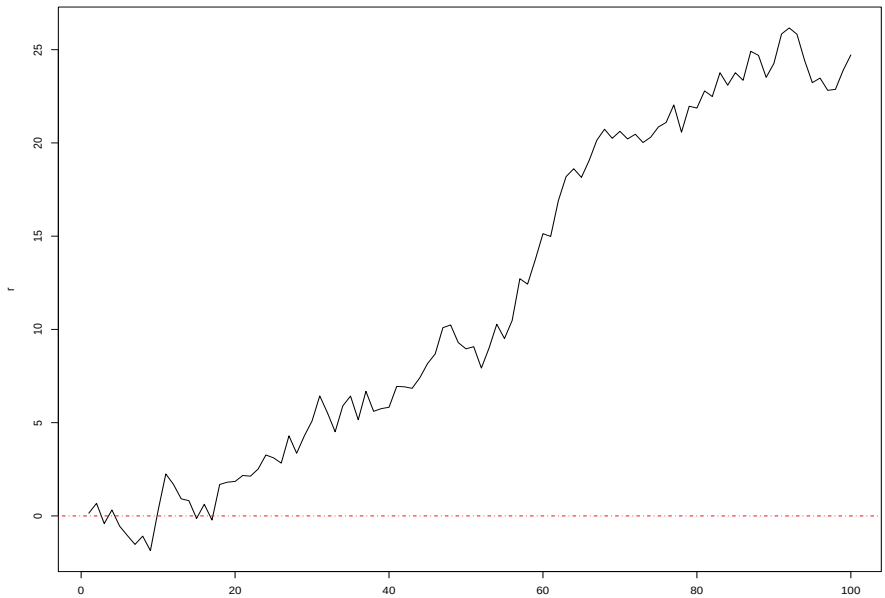

Figura 1.1: Ejemplo de una caminata aleatoria que consiste de 100 realizaciones, donde $Z_0 = 0$ y $\xi_i \sim \mathrm{N}(0, 1)$.

Definición 1. *Una serie de tiempo $\{Z_\tau\}_{\tau \in T}$ es una sucesión de observaciones tomadas en intervalos iguales de tiempo.*

Por lo tanto, una serie de tiempo es un caso particular de un proceso estocástico, donde el conjunto de subíndices corresponde al conjunto de los números enteros, es decir $T = \mathbb{Z}$. Para simplificar un poco la notación simplemente escribiremos $\{Z_\tau\}_{\tau \in T}$ como $\{Z_t\}$.

Por lo general, la manera en la que se estudian estos procesos es a patir de su función de distribución. De esta manera, si consideramos un conjunto finito de variables aleatorias $\{Z_{t_1}, Z_{t_2} \ldots, Z_{t_n}\}$ provenientes del proceso estocástico $\{Z_t : t = 0, \pm 1, \pm 2, \ldots\}$, la función de distribución de dimensión n se define mediante

$$F(Z_{t_1}, \ldots, Z_{t_n}) = \mathbb{P}[Z_{t_1} \leq z_{t_1}, \ldots, Z_{t_n} \leq z_{t_n}].$$

Definición 2. *Se dice que un proceso es estacionario de orden n si*

$$F(z_{t_1}, \ldots, z_{t_n}) = F(z_{t_1+k}, \ldots, z_{t_n+k}), \tag{1.1}$$

para $t_1, \ldots, t_n \in \mathbb{Z}$ *y* $k \in \mathbb{Z}$. *Además se dice que un proceso es estrictamente estacionario si* (1.1) *se cumple para* $n = 1, 2, \ldots$.

Sin embargo, debido a la compleja estructura de correlación que se presenta en algunas series, en términos prácticos es difícil considerar la definición 2. Por esta razón se plantea la siguiente definición:

Definición 3. *El proceso* $\{Z_t\}$ *es débilmente estacionario si sus primeros* $m < n$ *momentos existen y además no dependen del tiempo.*

Es así que a lo largo del presente texto se trabajará con procesos estrictamente estacionarios y que solo toman valores en el conjunto de números reales. Entonces para un determinado proceso $\{Z_t\}$, su media se define como

$$\mu_t = \mathbb{E}\left[Z_t\right],$$

su varianza

$$\sigma_t^2 = \mathbb{E}\left[(Z_t - \mu_t)^2\right],$$

la covarianza entre Z_{t_1} y Z_{t_2}

$$\gamma\left(t_1, t_2\right) = \mathbb{E}\left[(Z_{t_1} - \mu_{t_1})(Z_{t_2} - \mu_{t_2})\right],$$

y la correlación entre Z_{t_1} y Z_{t_2}

$$\rho\left(t_1, t_2\right) = \frac{\gamma\left(t_1, t_2\right)}{\sqrt{\sigma_{t_1}^2}\sqrt{\sigma_{t_2}^2}}.$$

Debido a que estamos considerando procesos estrictamente estacionarios, su media $\mu_t = \mu$ es constante, siempre que $\mathbb{E}\left[\,|Z_t|\,\right] < \infty$. Del mismo modo, si $\mathbb{E}\left[Z_t^2\right] < \infty$, entonces $\sigma_t^2 = \sigma^2$ para toda t. Además, dado que $F\left(z_{t_1}, z_{t_2}\right) = F\left(z_{t_1+k}, z_{t_2+k}\right)$ para cualesquiera números enteros t_1, t_2, k, se tiene que

$$\gamma\left(t_1, t_2\right) = \gamma\left(t_1 + k, t_2 + k\right)$$

y

$$\rho\left(t_1, t_2\right) = \rho\left(t_1 + k, t_2 + k\right).$$

Si $t_1 = t - k$ y $t_2 = t$ entonces

$$\gamma\left(t_1, t_2\right) = \gamma\left(t - k, t\right) = \gamma\left(t, t + k\right) = \gamma\left(k\right),$$

$$\rho\left(t_1, t_2\right) = \rho\left(t - k, t\right) = \rho\left(t, t + k\right) = \rho\left(k\right).$$

Por lo tanto, para un proceso estacionario con media y varianza constantes, se define la covarianza entre Z_t y Z_{t+k} como

$$\gamma(k) = \text{Cov}(Z_t, Z_{t+k}) = \mathbb{E}[(Z_t - \mu)(Z_{t+k} - \mu)],$$

en donde $\text{Cov}(Z_t, Z_s)$ solamente es una función de la diferencia entre los tiempos $|t-s|$. Si se define a la función de correlación entre Z_t y Z_{t+k} como

$$\rho(k) = \frac{\text{Cov}(Z_t, Z_{t+k})}{\sqrt{\sigma_t^2}\sqrt{\sigma_{t+k}^2}} = \frac{\gamma(k)}{\gamma(0)},$$

entonces para un proceso estrictamente estacionario con los dos primeros momentos finitos, la covarianza y la correlación entre Z_t y Z_{t+k} sólo dependerá de la diferencia del tiempo.

Es fácil ver que para un proceso estacionario la función de autocovarianza y la función de autocorrelación tienen las siguientes propiedades

1. $\gamma(0) = \sigma_t^2$; $\rho(0) = 1$,
2. $|\gamma(k)| \leq \gamma(0)$; $|\rho(0)| \leq 1$,
3. $\gamma(k) = \gamma(-k)$ y $\rho(k) = \rho(-k)$, $\forall\, k$,

es decir, $\gamma(k)$ y $\rho(k)$ son funciones pares y por lo tanto simétricas alrededor del 0. Esto se deduce del hecho de que la diferencia de tiempo entre Z_t y Z_{t+k} y Z_t y Z_{t-k} es la misma. Otra propiedad importante de $\gamma(k)$ y de $\rho(k)$ es que son no-negativas definidas en el sentido que

$$\sum_{i=1}^{n}\sum_{j=1}^{n} \alpha_i\, \alpha_j\, \gamma\,(|t_1 - t_j|) \geq 0 \tag{1.2}$$

y

$$\sum_{i=1}^{n}\sum_{j=1}^{n} \alpha_i\, \alpha_j\, \rho\,(|t_1 - t_j|) \geq 0 \tag{1.3}$$

para todo conjunto de puntos en el tiempo $t_1, \ldots, t_n$ y para cualesquiera números reales $\alpha_1, \ldots, \alpha_n$. El resultado (1.2) se deduce del hecho de que

$$0 \leq \sigma_t^2 = \sum_{i=1}^{n}\sum_{j=1}^{n} \alpha_i\, \alpha_j\, \text{Cov}(Z_{t_i}, Z_{t_j}) = \sum_{i=1}^{n}\sum_{j=1}^{n} \alpha_i\, \alpha_j\, \gamma\,(|t_1 - t_j|).$$

Se puede demostrar un resultado similar para $\rho(k)$ en (1.3) si se divide la desigualdad (1.2) entre $\gamma(0)$. Por lo tanto es importante saber que no cualquier función que cumpla las condiciones 1 a 3 puede ser una función de autocovarianza o de autocorrelación para un proceso. Una condición necesaria para que cualquier función sea de autocoavarianza o de autocorrelación es que sea no-negativa definida.

1.2. Función de autocorrelación parcial

Además de la correlación entre Z_t y Z_{t+k}, es deseable conocer la correlación entre Z_t y Z_{t+k} después de que la dependencia lineal entre las variables $Z_{t+1}, \ldots, Z_{t+k-1}$ se ha eliminado, es decir,

$$\text{Corr}\left(Z_t, Z_{t+k} | \, Z_{t+1}, \ldots, Z_{t+k-1}\right).$$

Nos referimos a esta función como la autocorrelación parcial. Entonces si consideremos un modelo de regresión donde la variable dependiente Z_{t+k}, de un proceso estacionario con media cero depende de k variables de retraso, $Z_{t+k-1}, Z_{t+k-2}, \ldots, Z_t$, es decir,

$$Z_{t+k} = \phi\left(k,1\right) Z_{t+k-1} + \phi\left(k,2\right) Z_{t+k-2} + \cdots + \phi\left(k,k\right) Z_t + \varepsilon_{t+k}, \tag{1.4}$$

donde $\phi_{k,i}$ denota el i-ésimo parámetro de regresión y ε_{t+k} es el error no correlacionado con Z_{t+k-j} para $j \geq 1$, y además tiene asociado una distribución normal. Multiplicando por Z_{t+k-j} en ambos lados de (1.4) y tomando esperanza, se tiene que

$$\gamma\left(j\right) = \phi\left(k,1\right)\gamma\left(j-1\right) + \phi\left(k,2\right)\gamma\left(j-2\right) + \cdots + \phi\left(k,k\right)\gamma\left(j-k\right),$$

entonces,

$$\rho\left(j\right) = \phi\left(k,1\right)\rho\left(j-1\right) + \phi\left(k,2\right)\rho\left(j-2\right) + \cdots + \phi\left(k,k\right)\rho\left(j-k\right).$$

Para $j = 1, 2, \ldots, k$, tenemos el siguiente sistema de ecuaciones:

$$\begin{aligned}
\rho\left(1\right) &= \phi\left(k,1\right)\rho\left(0\right) + \phi\left(k,2\right)\rho\left(1\right) + \cdots + \phi\left(k,k\right)\rho\left(k-1\right) \\
\rho\left(2\right) &= \phi\left(k,1\right)\rho\left(1\right) + \phi\left(k,2\right)\rho\left(1\right) + \cdots + \phi\left(k,k\right)\rho\left(k-2\right) \\
&\vdots \\
\rho\left(k\right) &= \phi\left(k,1\right)\rho\left(k-1\right) + \phi\left(k,2\right)\rho\left(k-2\right) + \cdots + \phi\left(k,k\right)\rho\left(0\right).
\end{aligned}$$

Usando la regla de Cramer sucesivamente para $k = 1, 2, \ldots$, tenemos que

$$\phi(k,k) = \frac{\begin{vmatrix} 1 & \rho(1) & \rho(2) & \cdots & \rho(k-2) & \rho(1) \\ \rho(1) & 1 & \rho(1) & \cdots & \rho(k-3) & \rho(2) \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ \rho(k-1) & \rho(k-2) & \rho(k-3) & \cdots & \rho(1) & \rho(k) \end{vmatrix}}{\begin{vmatrix} 1 & \rho(1) & \rho(2) & \cdots & \rho(k-2) & \rho(k-1) \\ \rho(1) & 1 & \rho(1) & \cdots & \rho(k-3) & \rho(k-2) \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ \rho(k-1) & \rho(k-2) & \rho(k-3) & \cdots & \rho(1) & 1 \end{vmatrix}} \tag{1.5}$$

donde $\phi(1,1) = \rho(1)$. Dado que, por definición $\rho(0) = \phi(0,0) = 1$ para cualquier proceso, cuando hablamos de la autocorrelación y de la autocorrelación parcial, nos referimos sólo a $\rho(k)$ y a $\phi(k,k)$ para $k \neq 0$.

1.3. Estimación de la media, autocovarianza y autocorrelación

A continuación, examinaremos las condiciones bajo las cuales podemos estimar la media, la autocovarianza y, por lo tanto, la autocorrelacion mediante el uso de la media muestral.

1.3.1. Media muestral

Con una sola realización, un estimador natural de la media $\mu = \mathbb{E}[Z_t]$ de un proceso estacionario es la media muestral

$$\bar{Z} = \frac{1}{n}\sum_{t=1}^{n} Z_t,$$

que es el valor promedio de n observaciones. Dado que $\bar{Z}$ es insesgado y consistente para μ, es decir

$$\mathbb{E}[\bar{Z}] = \mu,$$

y

$$\lim_{n\to\infty} \frac{1}{n}\sum_{t=1}^{n} Z_t = \mu. \tag{1.6}$$

entonces el estimador anterior es válido para la media. Además, se dice que el proceso es ergódico para la media si el resultado (1.6) se cumple. Una condición suficiente para que este resultado sea verdadero es que $\rho(k) \to 0$

cuando $k \to \infty$. Esto es así porque $\rho(k) \to 0$ cuando $k \to \infty$ implica que para toda $\varepsilon > 0$, podemos elegir una N de manera que $|\rho_k| < (1/4)\,\varepsilon$ para toda $k > N$, entonces

$$\begin{aligned}\left|\frac{1}{n}\sum_{k=-(n-1)}^{n-1}\rho(k)\right| &\leq \frac{2}{n}\sum_{k=0}^{n-1}|\rho(k)| \\ &\leq \frac{2}{n}\sum_{k=0}^{N}|\rho(k)| + \frac{2}{n}\sum_{k=N+1}^{n-1}|\rho(k)| \\ &\leq \frac{2}{n}\sum_{k=0}^{N}|\rho(k)| + \frac{\varepsilon}{2n}(n-N-1) \\ &\leq \varepsilon,\end{aligned}$$

donde elegimos a n lo bastante grande de modo que el primer término de la última desigualdad sea menor que $(1/2)\,\varepsilon$. Esto demuestra que si $\rho(k) \to 0$ cuando $k \to \infty$, entonces

$$\lim_{n\to\infty}\frac{1}{n}\sum_{k=-(n-1)}^{n-1}\rho(k) = 0.$$

1.3.2. Función de autocovarianza muestral

Un estimador de la función de autocovarianza es el siguiente,

$$\hat{\gamma}(k) = \frac{1}{n}\sum_{t=1}^{n-k}(Z_t - \bar{Z})(Z_{t+k} - \bar{Z}), \tag{1.7}$$

ó

$$\hat{\gamma}(k) = \frac{1}{n-k}\sum_{t=1}^{n-k}(Z_t - \bar{Z})(Z_{t+k} - \bar{Z}).$$

Ahora, considerando que

$$\begin{aligned}\sum_{t=1}^{n-k}(Z_t - \bar{Z})(Z_{t+k} - \bar{Z}) &= \sum_{t=1}^{n-k}\left[(Z_t - \mu) - (\bar{Z} - \mu)\right]\left[(Z_{t+k} - \mu) - (\bar{Z} - \mu)\right] \\ &= \sum_{t=1}^{n-k}(Z_t - \mu)(Z_{t+k} - \mu) - (\bar{Z} - \mu)\sum_{t=1}^{n-k}(Z_t - \mu) \\ &\quad - (\bar{Z} - \mu)\sum_{t=1}^{n-k}(Z_{t+k} - \mu) + (n-k)(\bar{Z} - \mu)^2 \\ &\approx \sum_{t=1}^{n-k}(Z_t - \mu)(Z_{t+k} - \mu) - (n-k)(\bar{Z} - \mu)^2,\end{aligned}$$

donde aproximamos los términos $\sum_{k=1}^{n-k}(Z_t-\mu)$ y $\sum_{k=1}^{n-k}(Z_{t+k}-\mu)$ por $(n-k)(\bar{Z}-\mu)$. Entonces,

$$\mathbb{E}\left[\hat{\gamma}\left(k\right)\right] \simeq \gamma\left(k\right) - \frac{k}{n}\gamma\left(k\right) - \left(\frac{n-k}{n}\right)\mathbb{V}\left[\bar{Z}\right].$$

Es evidente que este estimador es sesgado. A pesar de esto utilizaremos $\hat{\gamma}\left(k\right)$ como la función de autocovarianza muestral para estimar la función de autocovarianza $\gamma\left(k\right)$.

1.3.3. Función de autocorrelación muestral

Para algunas observaciones $Z_1, Z_2, \ldots, Z_n$ de la serie de tiempo $\{Z_t\}$, la ACF muestral se define como

$$\hat{\rho}\left(k\right) = \frac{\hat{\gamma}\left(k\right)}{\hat{\gamma}\left(0\right)} = \frac{\sum_{t=1}^{n-k}(Z_t-\bar{Z})(Z_{t+k}-\bar{Z})}{\sum_{t=1}^{n}(Z_t-\bar{Z})^2}, \qquad k = 1, 2, \ldots.$$

Es importante mencionar que algunas características de la serie, como la existencia de ciclos o tendendencias se pueden determinar a partir de la gráfica de la función de autocovarianza muestral.

Por ejemplo, en la gráfica 1.2 su comportamiento a través del tiempo no es estacionario, dado que para varios lags consecutivos el ACF es negativo o positivo, este cambio de signos es un claro indicio de que la serie original contiene tendencias. Sin embargo, en el correlograma 1.3 se puede observar la presencia de ciclos, esto se debe a la manera en que para distintos periodos de tiempo el comportamiento de los lags se repite. Los distintos métodos a través de los cuales estos dos componentes se eliminan, o por lo menos se aminoran, se estudian con detalle en la Sección 1.4.

Por otra parte, para un proceso Gaussiano estacionario, Bartlett (1946) demostró que para $k > 0$ y para $k + j > 0$,

$$\begin{aligned}\mathrm{Cov}\left(\hat{\rho}\left(k\right), \hat{\rho}\left(k+j\right)\right) \simeq \frac{1}{n}\sum_{i=-\infty}^{\infty}\big[&\rho\left(i\right)\rho\left(ij\right) + \rho\left(i+k+j\right)\rho\left(i-k\right) - 2\,\rho\left(k\right)\rho\left(i\right)\rho\left(i-k-j\right) \\ &- 2\,\rho\left(k+j\right)\rho\left(i\right)\rho\left(i-k\right) + 2\,\rho\left(k\right)\rho\left(k+j\right)\rho\left(i\right)^2\big].\end{aligned} \tag{1.8}$$

Para n grande, $\hat{\rho}\left(k\right)$ tiene aproximadamente una distribución normal con media $\rho\left(k\right)$ y varianza

$$\mathbb{V}\left[\hat{\rho}\left(k\right)\right] \simeq \frac{1}{n}\sum_{i=-\infty}^{\infty}\left[\rho\left(i\right)^2 + \rho\left(i+k\right)\rho\left(i-k\right) - 4\,\rho\left(k\right)\rho\left(i\right)\rho\left(i-k\right) + 2\,\rho\left(k\right)^2\rho\left(i\right)^2\right]. \tag{1.9}$$

Para un proceso en el que $\rho\left(k\right) = 0$ para $k > 0$, la aproximación de Bartlett para (1.9) es la siguiente

$$\mathbb{V}\left[\hat{\rho}\left(k\right)\right] \simeq \frac{1}{n}\left[1 + 2\,\rho\left(1\right)^2 + 2\,\rho\left(2\right)^2 + \cdots + 2\,\rho\left(m\right)^2\right].$$

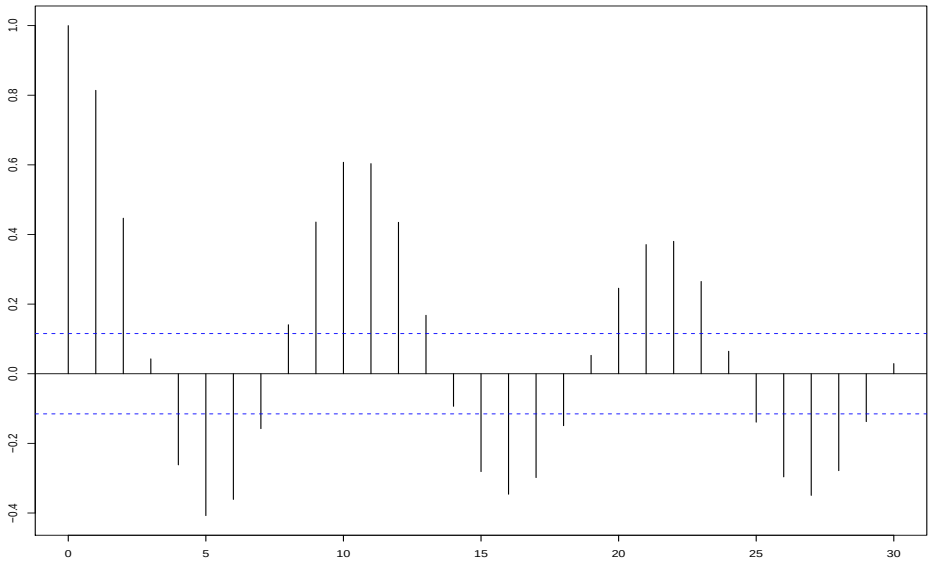

Figura 1.2: ACF muestral del número de manchas solares observadas de 1749 a 1997 (registro mensual).

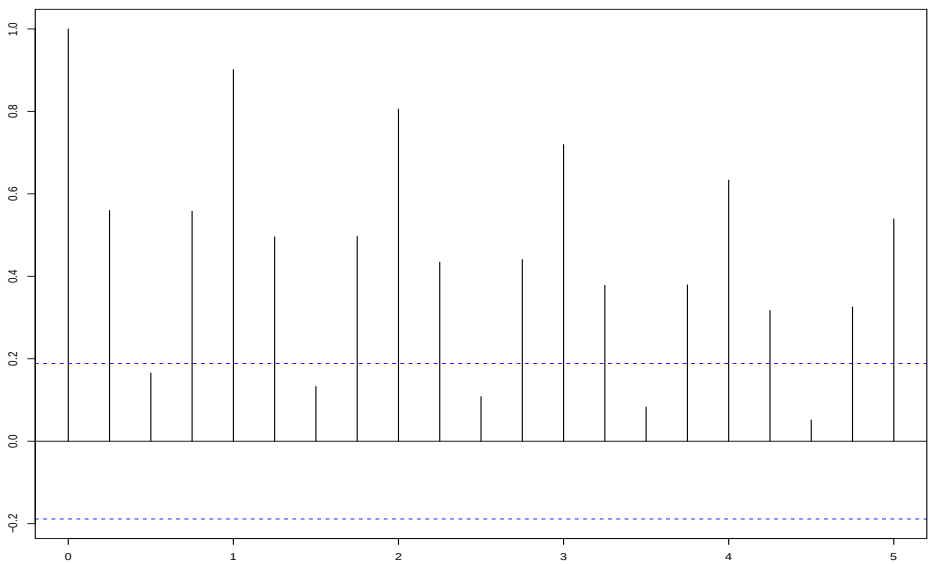

Figura 1.3: ACF muestral del consumo trimestal de gas de 1960 a 1986 (en millones).

En la práctica, se desconocen los valores reales de $\rho\,(i)$ $(i = 1, 2, \ldots, m)$ por lo que generalmente son reemplazados por sus respectivas estimaciones $\hat{\rho}\,(i)$, es por esto que se propone el siguiente error estándar de retraso para $\hat{\rho}\,(k)$:

$$S_{\hat{\rho}\,(k)} = \sqrt{\frac{1}{n}\,(\,1 + 2\,\rho_1^2 + 2\,\rho_2^2 + \cdots + 2\,\rho_m^2\,)}.$$

1.3.4. Función de autocorrelación parcial muestral

La PACF muestral se obtiene al sustituir $\rho\,(i)$ por $\hat{\rho}\,(i)$ en la Ecuación (1.8). En lugar de calcular los complicados determinantes cuando k es grande en (1.8), Durbin (1960) proporcionó un método recursivo para calcular $\phi\,(k, k)$ comenzando con un valor inicial $\hat{\phi}\,(1, 1) = \hat{\rho}\,(1)$, de la siguiente manera

$$\hat{\phi}\,(k+1, k+1) = \frac{\hat{\rho}\,(k+1) - \sum_{j=1}^{k} \hat{\phi}\,(k, j)\,\hat{\rho}\,(k+1-j)}{1 - \sum_{j=1}^{k} \hat{\phi}\,(k, j)\,\hat{\rho}\,(j)}$$

y

$$\hat{\phi}\,(k+i, j) = \hat{\phi}\,(k, j) - \hat{\phi}\,(k+1, k+1)\,\hat{\phi}\,(k, k+1-j), \quad \text{para} \quad j = 1, \ldots, k.$$

Bajo la hipótesis de que el proceso subyacente corresponde a un ruido blanco, la varianza de $\hat{\phi}\,(k, k)$ se puede aproximar mediante

$$\mathbb{V}\,[\hat{\phi}\,(k, k)] \simeq \frac{1}{n}.$$

Entonces, $\pm 2/\sqrt{n}$ se puede utilizar como los límites críticos de $\hat{\phi}\,(k, k)$ para poner a prueba la hipótesis de la existencia de un ruido blanco.

1.4. Modelo clásico de series de tiempo

Inicialmente, el modelo de series de tiempo se ha pensado como una mezcla de tendencias T_t, ciclos C_t y de componenetes irregulares e_t. Si suponemos que cada uno de estos componentes son independientes y aditivos, podemos escribir la serie de tiempo Z_t como

$$Z_t = T_t + C_t + e_t.$$

A continuación se muestran algunos métodos para estimar cada uno de estos componentes.

1.4.1. Método de promedios móviles

Este método se desarrolla sobre la hipótesis de que la suma anual de una serie posee poca variación. Entonces, considerando que $N_t = P_t + e_t$ es el componente no estacionario de la serie, una estimación de este factor se puede obtener utilizando un operador simétrico de promedios móviles, es decir,

$$\hat{N}_t = \sum_{i=-m}^{m} \lambda_i \, Z_{t-i},$$

donde $m \in \mathbb{Z}^+$ y las $\lambda_i's$ son constantes tales que $\lambda_i = \lambda_{-i}$ y $\sum_{i=-m}^{m} \lambda_i = 1$. Una estimación del componente de ciclos se deriva al substraer $\hat{N}_t$ de la serie original, es decir,

$$\hat{C}_t = Z_t - \hat{N}_t.$$

Las estimaciones anteriores se pueden obtener iterativamente a través de la repetición de los diversos operadores de promedios móviles. El procedimiento anterior de conoce como el ajuste estacionario. El caso más sencillo ocurre cuando $\lambda_i = \lambda_j \ \forall \ i \neq j$, es decir, todos los pesos son iguales, por lo tanto el operador de promedios móviles se reduce a

$$\hat{N}_t = \frac{1}{2m+1} \sum_{i=-m}^{m} Z_{t-i}.$$

Por ejemplo, en la gráfica de abajo se muestra la aplicación este filtro a la serie correspondiente al consumo de gas en Inglaterra, con $m = 2$, $m = 12$ y $m = 40$.

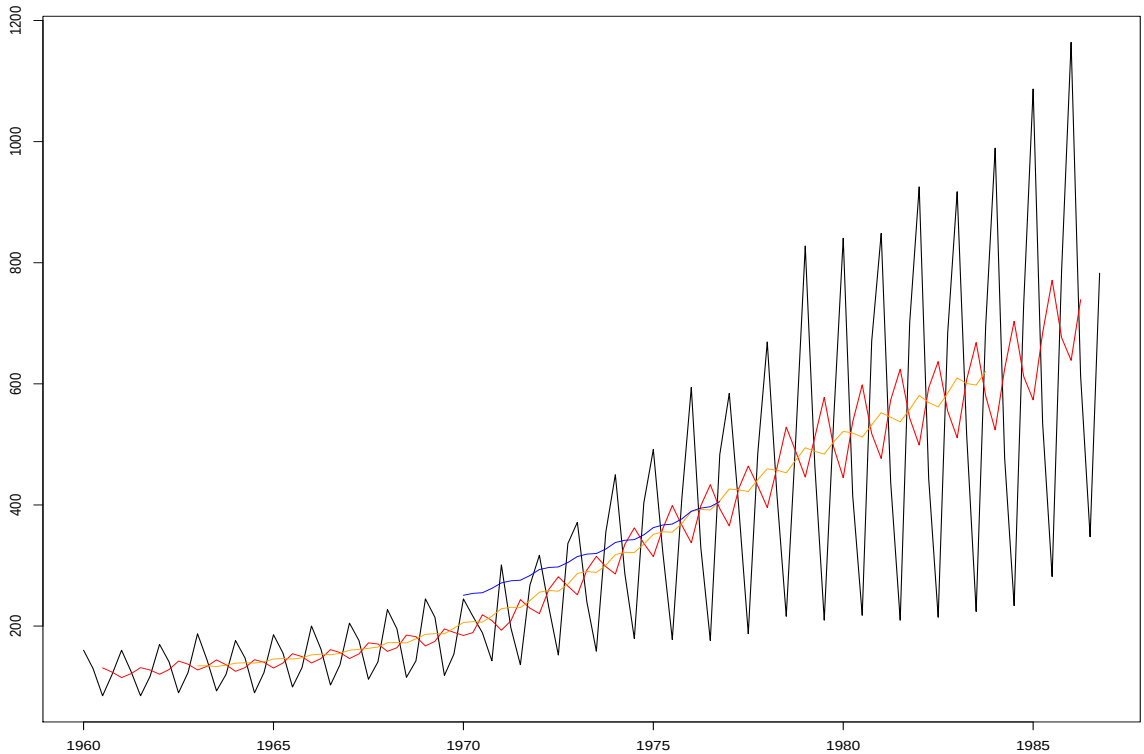

Figura 1.4: Introducción de distintos filtros: m=2, m=12 y m=40.

Por lo tanto, los pesos son los siguientes:

- m=2: $\lambda_i = \underbrace{\left(\frac{1}{5}, \ldots, \frac{1}{5}\right)}_{5 \text{ veces}}$,
- m=12: $\lambda_i = \underbrace{\left(\frac{1}{25}, \ldots, \frac{1}{25}\right)}_{25 \text{ veces}}$,
- m=40: $\lambda_i = \underbrace{\left(\frac{1}{81}, \ldots, \frac{1}{81}\right)}_{81 \text{ veces}}$.

Una posible interpretación de estos filtros es pensarlos como un promedio semanal ($m = 2$), mensual ($m = 12$) y trimestral ($m = 40$).

1.4.2. Método de regresión

En este método la serie de tiempo se escribe como un modelo de regresión:

$$\begin{aligned} Z_t &= T_t + C_t + e_t \\ &= \alpha_0 + \sum_{i=1}^{m} \alpha_i \, U_{it} + \sum_{j=1}^{k} \beta_j \, V_{jt} + e_t, \end{aligned} \tag{1.10}$$

donde $T_t = \alpha_0 + \sum_{i=1}^{m} \alpha_i \, U_{it}$ es la variable de tendencia y $C_t = \sum_{j=1}^{k} \beta_j \, V_{jt}$ corresponde a la variable de ciclos. De manera general, el componente de tendencias puede ser expresado como un polinomio de orden m con respecto al tiempo, es decir,

$$T_t = \alpha_0 + \sum_{i=1}^{m} \alpha_i \, t^i.$$

Análogamente, el componente de ciclos se puede expresar como una combinación lineal de variables indicadoras o como una combinación lineal de funciones seno y coseno de distintas frecuencias. Por ejemplo, la serie cíclica de periodo d se puede escribir como

$$C_t = \sum_{j=1}^{s-1} \beta_j \, D_{jt},$$

donde

$$D_{jt} = \begin{cases} 1, & \text{si } t = j, \\ 0, & \text{e.o.c.} \end{cases}$$

Notemos que cuando el ciclo del periodo es igual a s, necesitamos sólo $s-1$ variables dummy. En otras palabras, β_s se fija para que sea 0 de tal manera que el coeficiente β_j para $j \neq s$ represente el efecto cíclico del j-ésimo periodo en comparación con el período s. De manera alternativa, C_t se puede escribir de la siguente manera

$$C_t = \sum_{j=1}^{[s/2]} \left[\beta_j \sin\left(\frac{2\pi j}{s}\right) + \gamma_j \cos\left(\frac{2\pi j}{s}\right) \right],$$

donde $[s/2]$ es la parte entera de $s/2$. Por lo tanto, el modelo (1.10) se convierte en

$$Z_t = \alpha_0 + \sum_{i=1}^{m} \alpha_i \, t^i + \sum_{j=1}^{s-1} \beta_j \, D_{jt} + e_t \tag{1.11}$$

ó

$$Z_t = \alpha_0 + \sum_{i=1}^{m} \alpha_i \, t^i + \sum_{j=1}^{[s/2]} \left[\beta_j \sin\left(\frac{2\pi j}{s}\right) + \gamma_j \cos\left(\frac{2\pi j}{s}\right) \right] + e_t. \tag{1.12}$$

Por ejemplo, para la serie anterior, si

$$Z_t = \alpha_0 + \alpha_1\, t + \alpha_2\, t^2 + \beta_1 \sin(2\, z_j\, \pi) + \beta_2\, \sin(4\, z_j\, \pi) + \gamma_1\, \cos(2\, z_j\, \pi) + \gamma_2\, \cos(4\, z_j\, \pi) + e_t, \tag{1.13}$$

el modelo estimado en comparación con la serie original se muestra a continuación

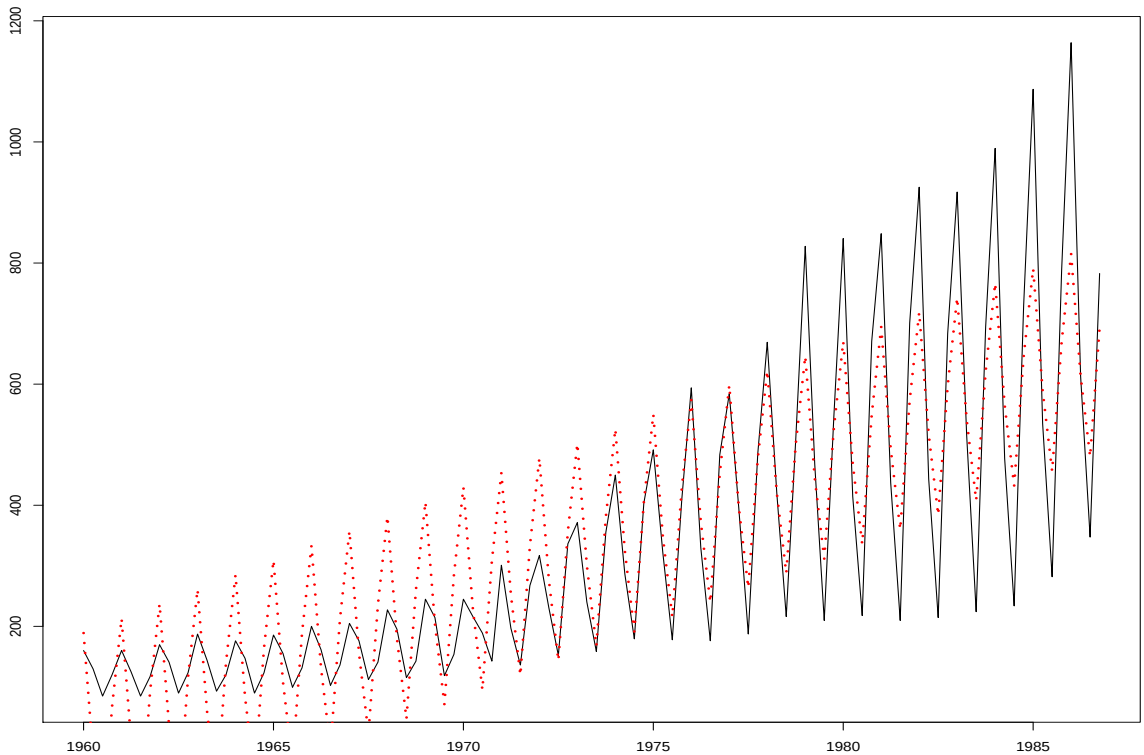

Figura 1.5: Estimación de la serie mediante un modelo de regresión de segundo grado, donde ($\cdots$) es el modelo estimado y (—) es el modelo observado.

Para un determinado conjunto de datos Z_t y para valores específicos de m y s, el método de regresión por mínimos cuadrados se puede utilizar para estimar α_i, β_j y γ_j. Las estimaciones de T_t, S_t y e_t para la

Ecuación (1.11) están dadas por

$$\hat{T}_t = \hat{\alpha}_0 + \sum_{i=1}^{m} \hat{\alpha}_i \, t^i,$$
$$\hat{C}_t = \sum_{j=1}^{s-1} \hat{\beta}_j \, D_{jt},$$
$$\hat{e}_t = Z_t - \hat{T}_t - \hat{C}_t.$$

De manera análoga, para la Ecuación (1.12), tenemos que,

$$\hat{T}_t = \hat{\alpha}_0 + \sum_{i=1}^{m} \hat{\alpha}_i \, t^i,$$

$$\hat{C}_t = \sum_{j=1}^{[s/2]} \left[\hat{\beta}_j \sin\left(\frac{2\pi j}{s}\right) + \hat{\gamma}_j \cos\left(\frac{2\pi j}{s}\right) \right]$$

$$\hat{e}_t = Z_t - \hat{T}_t - \hat{C}_t.$$

Para el modelo anteriormente planteado, la estimación de cada uno de sus parámetros se muestra en la siguiente tabla:

Parámetro	Estimación
α_0	3.371e+02
α_1	1.951e+03
α_2	3.868e+02
γ_1	1.730e+02
β_1	-3.396e+01
γ_2	4.186e+00
β_2	4.933e+12

Cuadro 1.1: Estimación de los parámetros para el modelo (1.13).

1.4.3. Método de diferencias

Definimos al operador de diferencias $\bigtriangledown$ como

$$\bigtriangledown^j (Z_t) = \bigtriangledown \left(\bigtriangledown^{j-1} (Z_t)\right), \quad \text{para} \quad j \geq 1 \quad \text{con} \quad \bigtriangledown (Z_t) = Z_t - Z_{t-1} \quad \text{y} \quad \bigtriangledown^0 (Z_t) = Z_t.$$

Por ejemplo,

$$\begin{aligned} \bigtriangledown^2 Z_t &= \bigtriangledown (\bigtriangledown Z_t) = (1-B)(1-B)\, Z_t = (1 - 2B + B^2)\, Z_t \\ &= Z_t - 2\, Z_{t-1} + Z_{t-2}, \end{aligned}$$

donde $B\, Z_t = Z_{t-1}$ se define como el operador de retraso. Entonces, si el operador $\bigtriangledown$ es aplicado a

$$T_t = \alpha_0 + \alpha_1\, t,$$

obtenemos la función constante

$$\bigtriangledown T_t = T_t - T_{t-1} = \alpha_0 + \alpha_1\, t - (\alpha_0 + \alpha\,(t-1)) = \alpha_1.$$

De la misma manera, cualquier polinomio con tendencias de grado m puede reducirse a una constante mediante la aplicación del operador $\bigtriangledown^m$. Por ejemplo, si

$$Z_t = T_t + e_t,$$

donde

$$T_t = \alpha_0 + \sum_{i=1}^{m} \alpha_i\, t^i$$

y e_t es un proceso estacionario con media cero, al aplicar $\bigtriangledown^m$ tenemos que

$$\bigtriangledown^m Z_t = m!\, \alpha_m + \bigtriangledown^m e_t,$$

es un proceso estacionario con media $m!\, \alpha_m$. Estas consideraciones sugieren la posibilidad de aplicar el operador $\bigtriangledown$ hasta encontrar una sucesión $\{\bigtriangledown^m z_t\}$, que se pueda modelar como la realización de un proceso estacionario. A menudo, en la práctica, el orden necesario de las diferencias m es bastante pequeño, con frecuencia es de uno o dos. Esto se basa en el hecho de que muchas de las funciones se pueden aproximar por un polinomio de grado razonablemente bajo.

1.5. Ecuaciones en diferencias lineales

La mayoria de los modelos constituídos por un número finito de parámetros hacen referencia tanto a una variable de salida, Z_t, como a una variable de entrada, a_t, en términos de ecuaciones en diferencias lineales. Por lo que a estos modelos se les conoce como *Modelos de Ecuaciones en Diferencias Lineales*. Las propiedades de estos modelos dependen de las características de las raíces de las ecuaciones en diferencias.

En general, una ecuación de diferencias lineales de orden n con coeficientes constantes está dada por

$$C_0\, Z_t + C_1\, Z_{t-1} + C_2\, Z_{t-2} + \cdots + C_n\, Z_{t-n} = e_t, \tag{1.14}$$

donde C_i $(i = 0, 1, \ldots, n)$ son constantes. Sin pérdida de generalidad, vamos a suponer que $C_0 = 1$. La función e_t en (1.14) es conocida como la función de forzamiento. La ecuación (1.14) se dice que es homogénea si $e_t = 0$, y no-homogénea (o completa) si $e_t \neq 0$. Utilizando el operador de retraso $C(B) = (1+C_1\, B+C_2\, B^2+\cdots+C_n\, B^n)$, podemos reescribir (1.14) como

$$C(B)\, Z_t = e_t.$$

Como una función de B, donde éste opera en el índice de tiempo t, $C(B) = 0$ es conocida como la ecuación auxiliar asociada a la ecuación lineal de diferencias dada. La solución de las ecuaciones de diferencias lineales se basa sobre todo en los siguientes lemas.

Lema 1. *Si $Z_t^{(1)}$ y $Z_t^{(2)}$ son dos soluciones de la ecuación homogénea $C(B)\, Z_t = 0$, entonces $b_1\, Z_t^{(1)} + b_2\, Z_t^{(2)}$ es también una solución para cualesquiera constantes b_1 y b_2*

Lema 2. *Si $Z_t^{(H)}$ es una solución de la ecuación homogénea $C(B)\, Z_t = 0$ y $C(B)\, Z_t = e_t$ entonces $Z_t^{(H)} + Z^{(P)}$ es una solución general de la ecuación completa.*

La solución párticular de una ecuación de diferencias dependerá de la forma de la función de forzamiento. Por lo general, el eventual comportamiento de un modelo de series de tiempo es a menudo regido por una ecuación homogénea de diferencias. Por lo que, es de vital importancia el tener en cuenta la solución general de una ecuación homogénea de diferencias.

Lema 3. *Si $Z_t = b\, t^j$ y si m es un entero no negativo, donde b es una constante cualquiera y $j \in \mathbb{Z}^+$ está fijo y $j < m$, entonces $(1 - B)^m\, Z_t = 0$.*

Demostración. Para $m = 1$, $Z_t = bt^0 = b$. Es evidente que $(1-B)\, Z_t = (1-B)\, b = b-b = 0$. Ahora supongamos

que $(1-B)^{m-1}\,Z_t = 0$. Entonces para $Z_t = b\,t^j$, $j < m$,

$$\begin{aligned}(1-B)^m\,Z_t &= (1-B)^{m-1}(1-B)\,b\,t^j \\ &= (1-B)^{m-1}\,b\{t^j - (t-1)^j\} \\ &= (1-B)^{m-1}\left\{-b\sum_{i=0}^{j-1}\binom{j}{i}(-1)^{j-i}\,t^i\right\}.\end{aligned}$$

Ahora, cada término en la última expresión que involucra a t contiene solamente potencias enteras menores a $m-1$. Entonces, por nuestra hipótesis de inducción, cada término es reducido a cero por el operador $(1-B)^{m-1}$. Por lo que el Lema está demostrado. □

Se desprende de los Lemas 1 y 3 que $(1-B)^m[\sum_{j=0}^{m-1} b_j t^j] = 0$.

Lema 4. *Si* $(1-RB)^m = 0$ *donde* $R \neq 0$, $m \in \mathbb{Z}^+$ *y* $Z_t = R^t\,t^j$ *donde* $j \in \mathbb{Z}^+$ *y* $j < m$. *Entonces* $(1-RB)^m\,Z_t = 0$.

Demostración. Primero, notemos que

$$\begin{aligned}(1-RB)\,Z_t &= (1-RB)\,R^t\,t^j \\ &= R^t\,t^j - R\times R^{t-1}\,(t-1)^j \\ &= R^t\,(1-B)\,t^j.\end{aligned}$$

El uso repetido del resultado anterior implica que

$$(1-RB)^m\,R^t\,t^j = R^t\,(1-B)^m\,t^j.$$

Por lo tanto, el resultado se sigue inmediatamente del Lema 3. □

Finalmente, tenemos el siguiente resultado.

Teorema 1. *Sea* $C(B)\,Z_t = 0$ *una ecuación homogénea de diferencias lineales donde* $C(B) = 1 + C_1\,B + C_2\,B^2 + \cdots + C_n\,B^n$. *Si* $C(B) = \prod_{i=1}^{N}(1-R_iB)^{m_i}$ *donde* $\sum_{i=1}^{N} m_i = n$ *y* $B_i = R_i^{-1}$, $i = 1, 2, \ldots, N$ *son raíces de multiplicidad* m_i *de* $C(B) = 0$, *entonces* $Z_t = \sum_{i=1}^{N}\sum_{j=0}^{m_i-1} R_i^t\,b_{ij}t^j$. *En párticular, si* $m_i = 1$ *para toda* i *y* R_i^{-1} *es distinto para* $i = 1, 2, \ldots, n$, *se tiene que* $Z_t = \sum_{i=1}^{n} b_i\,R_i^t$.

Demostración. El resultado se sigue inmediatamente de los Lemas 1, 3 y 4. □

Notemos que las raíces complejas de $C(B) = 0$ deben aparecer en pares. Es decir, si $(c + di)$ es una raíz, entonces su complejo conjugado $(c+di)^* = (c-di)$ es tambien una raíz. Pero como un número complejo siempre puede ser escrito en forma polar, entonces,

$$(c \pm di) = \alpha\,(\cos\phi \pm i\,\sin\phi),$$

donde

$$\alpha = (c^2 + d^2)^{1/2}$$

y

$$\phi = \tan^{-1}(d/c).$$

Dado que $(c \pm d\,i)^t = \alpha^t\,(\cos(\phi\,t) \pm i\,\sin(\phi\,t))$, para cada par de raíces complejas de multiplicidad m la solución de la ecuación homogénea de diferencias debe contener las secuencias $\alpha^t\,\cos(\phi\,t)$, $\alpha^t\,\sin(\phi\,t)$; $t\,\alpha^t\cos(\phi\,t)$, $t\,\alpha\sin(\phi\,t)$; $\ldots$; $t^{m-1}\,\alpha^t\cos(\phi\,t)$, $t^{m-1}\,\alpha^t\sin(\phi\,t)$.

Capítulo 2

Modelos Estacionarios

Una vez que se ha descrito la estructura básica de una serie de tiempo y a partir de las definiciones planteadas en William (1990), se procede al análisis de las distintas representaciones de un proceso estacionario. De esta manera, en la primera sección del presente Capítulo se estudian los modelos MA y AR considerando sus primeros dos momentos. Sin embargo, como el número de parámetros para estos dos modelos puede ser prohibitvamente grande, en la Sección 2.2 se introduce el modelo ARMA a partir de la mezcla de los modelos autorregresivos y de promedios moviles. Posteriormente revisaremos algunos de los distintos métodos de estimación para los parámetros del modelo ARMA(p,q). Principalmente nos centraremos en los métodos de momentos y el de máxima verosimilitud. En la Sección 2.4 se presentan los distintos criterios para la selección del modelo. La forma en la que se construye el error en media cuadrática para la predicción de valores futuros de un modelo ARMA se explica en la Sección 2.5.

Como el modelo ARCH es indispensable para la construcción del modelo GH-ARCH, en la última parte de este capítulo se proporciona, de manera general, la construcción de este modelo a partir del artículo publicado por Engle (1982). En esta sección se desarrollan las principales ideas sobre la construcción, condiciones de estacionariedad y estimación del modelo ARCH(p).

2.1. Representación de procesos MA y AR en series de tiempo

En el análisis de series de tiempo dos representaciones son especialmente útiles para expresar un proceso. Una de estás es escribir a Z_t como una combinación lineal de variables aleatorias no correlacionadas, es decir,

$$Z_t = \mu + a_t + \psi_1\, a_{t-1} + \psi_2\, a_{t-2} + \cdots = \sum_{j=0}^{\infty} \psi_j\, a_{t-j}, \tag{2.1}$$

donde $\psi_0 = 1$, $\{a_t\}$ es un ruido blanco con media cero y $\sum_{j=0}^{\infty} \psi_j^2 < \infty$. La Ecuación (2.1) se define de tal manera que

$$\mathbb{E}\left[\left(\dot{Z}_t - \sum_{j=0}^{n} \psi_j\, a_{t-j}\right)^2\right] \longrightarrow 0 \qquad \text{cuando } n \longrightarrow \infty,$$

donde $\dot{Z}_t = Z_t - \mu$. Al introducir el operador de retraso $B^j\, x_t = x_{t-j}$ podemos escribir (2.1) como,

$$\dot{Z}_t = \psi(B)\, a_t$$

donde $\psi(B) = \sum_{j=0}^{\infty} \psi_j\, B^j$. El proceso (2.1) es conocido como *Modelo de Promedios Moviles, MA*, en el análisis de series de tiempo. Wold (1938) demostró que un proceso estacionario no determinístico siempre se puede expresar como (2.1). Entonces esta representación es conocida en la literatura como la representación de Wold, y todo proceso que pueda ser representado de esta forma es conocido como proceso no determinístico.

Es fácil demostrar que para el proceso descrito en (2.1)

$$\begin{aligned} \mathbb{E}\left[Z_t\right] &= \mu, \\ \mathbb{V}\left[Z_t\right] &= \sigma_a^2 \sum_{j=0}^{\infty} \psi_j^2 \end{aligned} \tag{2.2}$$

y

$$\mathbb{E}\left[a_t Z_t\right] = \begin{cases} \sigma_a^2, & \text{para } j = 0, \\ 0, & \text{si } j > 0. \end{cases}$$

Entonces,

$$\gamma\left(k\right) = \mathbb{E}\left[\dot{Z}_t\, \dot{Z}_{t+k}\right] = \mathbb{E}\left[\sum_{i=0}^{\infty} \sum_{j=0}^{\infty} \psi_i\, \psi_j\, a_{t-i}\, a_{t+k-j}\right] = \sigma_a^2 \sum_{i=0}^{\infty} \psi_i\, \psi_{i+k} \tag{2.3}$$

y

$$\rho\left(k\right) = \frac{\sum_{i=0}^{\infty} \psi_i\, \psi_{i+k}}{\sum_{i=0}^{\infty} \psi_i^2}. \tag{2.4}$$

Evidentemente (2.3) y (2.4) son funciones que dependen de la diferencia del tiempo k. Sin embargo, dado que se trata de sumas infinitas, para que el proceso sea estacionario tenemos que demostrar que $\gamma(k)$ es finita para cada k. Ahora,

$$|\gamma(k)| = |\mathbb{E}[\dot{Z}_t\, \dot{Z}_{t+k}]| \leq [\mathbb{V}[Z_t]\, \mathbb{V}[Z_{t+k}]]^{1/2} = \sigma_a^2 \sum_{j=0}^{\infty} \psi_j^2.$$

Por lo tanto, una condición necesaria para que el proceso (2.1) sea estacionario es $\sum_{j=0}^{\infty} \psi_j^2 < \infty$.

Otra forma útil de escribir al proceso Z_t es mediante una representación *Autorregresiva, AR*, en la cual regresamos el valor de Z al tiempo t sobre su propio pasado más un error aleatorio, es decir,

$$\dot{Z}_t = \pi_1\, \dot{Z}_{t-1} + \pi_2\, \dot{Z}_{t-2} + \cdots + a_t$$

o equivalentemente,

$$\pi(B)\, \dot{Z}_t = a_t, \tag{2.5}$$

donde $\pi(B) = 1 - \sum_{j=1}^{\infty} \pi_j\, B^j$ y $1 + \sum_{j=1}^{\infty} |\pi_j| < \infty$. Box y Jenkins (1976) llaman a un proceso invertible si se puede escribir de la forma (2.5). Para un proceso lineal $Z_t = \psi(B) = a_t$ invertible que puede ser escrito en términos de un proceso AR, las raíces de $\psi(B) = 0$, como una función de B, deben quedar fuera del círculo unitario. Es decir, si β es una raíz de $\psi(B)$, entonces $|\beta| > 1$, donde $|\cdot|$ corresponde a una métrica Euclideana. Es fácil observar que no todos los procesos estacionarios son invertibles y que un proceso invertible no necesariamente es estacionario.

Por el resultado de Wold, para que el proceso (2.5) sea estacionario, este deberá poder ser reescrito en una representación MA, es decir,

$$\dot{Z}_t = \frac{1}{\pi(B)}\; a_t = \psi(B)\, a_t,$$

de tal manera que $\sum_{j=0}^{\infty} \psi_j^2 < \infty$ se satisfaga. Para lograr que esta condición se cumpla es necesario que todas las raíces de $\pi(B) = 0$ se encuentren fuera del círculo unitario, es decir, $|\delta| > 1$.

Aunque las representaciones de los procesos de promedios moviles y autorregresivos son útiles, no son los modelos que se utilizan al principio. Esto se debe a que contienen un número infinito de parámetros, imposibles de calcular a partir de un número finito de observaciones. Si sólo un número finito de pesos π son distintos de cero, $\pi_1 = \phi_1, \pi_2 = \phi_2, \ldots, \pi_p = \phi_p$ y $\pi_k = 0$ para $k > p$, entonces el proceso resultante es conocido como un proceso autoregresivo de orden p, entonces escribimos

$$\dot{Z}_t - \phi_1\, \dot{Z}_{t-1} - \cdots - \phi_p\, \dot{Z}_{t-p} = a_t. \tag{2.6}$$

Analogamente, en la representación de un proceso de promedios móviles, si sólo un número finito de pesos ψ son distintos de cero, $\psi_1 = -\theta_1, \psi_2 = -\theta_2, \ldots, \psi_q = -\theta_q$ y $\psi_k = 0$ para $k > q$, entonces el proceso resultante se conoce como proceso de promedios móviles de orden q, el cual se escribe de la siguiente manera

$$\dot{Z}_t = a_t - \theta_1 a_{t-1} - \cdots - \theta_q a_{t-q}. \tag{2.7}$$

Sin embargo, aunque nos limitemos a un orden finito en los procesos (2.6) y (2.7) el número de parámetros puede ser prohibitivamente grande. Una alternativa natural es la mezcla de los modelos autorregresivos y de promedios moviles

$$\dot{Z}_t - \phi_1 \dot{Z}_{t-1} - \cdots - \phi_p \dot{Z}_{t-p} = a_t - \theta_1 a_{t-1} - \cdots - \theta_q a_{t-q}.$$

Para un número fijo de observaciones, un gran número de parámetros en un modelo hace menos eficiente la estimación de los parámetros. En general se busca que el modelo tenga el menor número de parámetros posibles. Este es el principio de parsimonia en la construcción del modelo de Tukey (1967) y Box y Jenkins (1976).

2.2. Modelo ARMA(p,q)

Como se mencionó anteriomente, en la modelación de series de tiempo, a veces es necesario incluir términos autorregresivos y de promedios moviles. Esto condujo al planteamiento del modelo *Autorregresivo de Promedios Moviles, ARMA*:

$$\phi_p(B)\,\dot{Z}_t = \theta_q(B)\,a_t, \tag{2.8}$$

donde

$$\phi_p(B) = 1 - \phi_1 B - \cdots - \phi_p B^p,$$

y

$$\theta_q(B) = 1 - \theta_1 B - \cdots - \theta_q B^q.$$

Para que el proceso sea invertible, es necesario que las raíces de $\theta_q(B) = 0$ se encuentren fuera del círculo unitario. Para que sea estacionario, es necesario que las raíces de $\phi(B) = 0$ se encuentren fuera del círculo unitario. Además, suponemos que $\theta_q(B) = 0$ y $\phi(B) = 0$ no comparten raíces. De ahora en adelante, nos referimos a este proceso como un modelo $ARMA(p,q)$, en donde p y q denotan el orden del proceso autorregresivo y de promedios móviles respectivamente.

Un proceso ARMA invertible y estacionario puede ser escrito en una representación autoregresiva, es decir,

$$\pi(B)\dot{Z}_t = a_t,$$

donde

$$\pi(B) = \frac{\phi_p(B)}{\theta_q(B)} = (1 - \pi_1 B - \pi_2 B^2 - \cdots).$$

Este proceso también se puede escribir como un proceso de promedios móviles

$$\dot{Z}_t = \psi(B)\, a_t,$$

donde

$$\psi(B) = \frac{\theta_q(B)}{\phi_p(B)} = (1 + \psi_1 B + \psi_2 B^2 + \cdots).$$

2.2.1. ACF del proceso ARMA(p,q)

Para obtener la función de autocovarianza, reescribimos (2.8) como

$$\dot{Z}_t = \phi_1 \dot{Z}_{t-1} + \cdots + \phi_p \dot{Z}_{t-p} + a_t - \theta_1 a_{t-1} - \cdots - \theta_q a_{t-q}$$

y multiplicando por $\dot{Z}_{t-k}$ ambos lados de la expresión anterior se tiene que

$$\dot{Z}_{t-k}\dot{Z}_t = \phi_1 \dot{Z}_{t-k}\dot{Z}_{t-1} + \cdots + \phi_p \dot{Z}_{t-k}\dot{Z}_{t-p} + \dot{Z}_{t-k} a_t - \theta_1 \dot{Z}_{t-k} a_{t-1} - \cdots - \theta_q \dot{Z}_{t-k} a_{t-q}.$$

Ahora, tomamos el valor esperado para obtener

$$\gamma_k = \phi_1 \gamma_{k-1} + \cdots + \phi_p \gamma_{k-p} + \mathbb{E}[\dot{Z}_{t-k} a_t] - \theta_1 \mathbb{E}[\dot{Z}_{t-k} a_{t-1}] - \cdots - \theta_q \mathbb{E}[\dot{Z}_{t-k} a_{t-q}].$$

Dado que

$$\mathbb{E}[\dot{Z}_{t-k} a_{t-i}] = 0 \quad \text{para} \quad k > i,$$

entonces

$$\gamma_k = \phi_1 \gamma_{k-1} + \cdots + \phi_p \gamma_{k-p}, \qquad k \geq q+1,$$

por lo tanto,

$$\rho_k = \phi_1 \rho_{k-1} + \cdots + \phi_p \rho_{k-p}, \qquad k \geq q+1. \tag{2.9}$$

Además la expresión (2.9) satisface las ecuaciones homogéneas de diferencias de orden p. Por lo tanto la función de autocorrelación de un proceso ARMA(p,q) disminuye después del retraso q, al igual que la de un proceso AR(q). Sin embargo, las primeras q autocorrelaciones $\rho_q, \rho_{q-1}, \ldots, \rho_1$ dependen tanto de los parámetros autorregresivos como de los parámetros de promedios móviles en el modelo. Esta distinción es útil en la identificación del modelo.

2.3. Estimación de los parámetros

Una vez que se ha identificado un modelo tentativo, el siguiente paso es estimar los parámetros del modelo. Se estimarán los parámetros $\phi = (\phi_1, \phi_2, \ldots, \phi_p)^T$, $\mu = \mathbb{E}[Z_t]$, $\theta = (\theta_1, \theta_2, \ldots, \theta_q)$ y $\sigma_a^2 = \mathbb{E}[a_t^2]$ en el modelo

$$\dot{Z}_t = \phi_1 \dot{Z}_{t-1} + \phi_2 \dot{Z}_{t-2} + \cdots + \phi_p \dot{Z}_{t-p} + a_t - \theta_1 a_{t-1} - \cdots - \theta_q a_{t-q} \tag{2.10}$$

donde $t = 1, 2, \ldots, n$.

2.3.1. Método de momentos

Este método consiste en la sustitución de los momentos muestrales, como la media muestral, la varianza muestral, o la ACF muestral, con sus contrapartes teóricas para resolver las ecuaciones resultantes y obetener los estimadores correspondientes. Por ejemplo para un proceso AR(p)

$$\dot{Z}_t = \phi_1 \dot{Z}_{t-1} + \phi_2 \dot{Z}_{t-2} + \cdots + \phi_p \dot{Z}_{t-p} + a_t,$$

la media $\mu = \mathbb{E}[Z_t]$ es estimada por $\bar{Z}$. Para estimar ϕ, primero consideraremos que $\rho_k = \phi_1 \rho_{k-1} + \phi_2 \rho_{k-2} + \cdots + \phi_p \rho_{k-p}$ donde $k > 1$ para obtener el siguiente sistema de ecuaciones

$$\begin{aligned}
\rho_1 &= \phi_1 + \phi_2 \rho_1 + \phi_3 \rho_2 + \cdots + \phi_p \rho_{p-1} \\
\rho_2 &= \phi_1 \rho_1 + \phi_2 + \phi_3 \rho_1 + \cdots + \phi_p \rho_{p-2} \\
&\vdots \\
\rho_p &= \phi_1 \rho_{p-1} + \phi_2 \rho_{p-2} + \phi_3 \rho_{p-3} + \cdots + \phi_p
\end{aligned}$$

Entonces reemplazando ρ_k por $\hat{\rho}_k$, podemos obtener los estimadores para $\hat{\phi}_1, \hat{\phi}_2, \ldots, \hat{\phi}_p$ al resolver el siguiente sistema de ecuaciones

$$\begin{bmatrix} \hat{\phi}_1 \\ \hat{\phi}_2 \\ \vdots \\ \hat{\phi}_p \end{bmatrix} = \begin{bmatrix} 1 & \hat{\rho}_1 & \hat{\rho}_2 & \cdots & \hat{\rho}_{p-2} & \hat{\rho}_{p-1} \\ \hat{\rho}_1 & 1 & \hat{\rho}_1 & \cdots & \hat{\rho}_{p-3} & \hat{\rho}_{p-2} \\ \vdots & & & & & \vdots \\ \hat{\rho}_{p-1} & \hat{\rho}_{p-2} & \hat{\rho}_{p-3} & \cdots & \hat{\rho}_1 & 1 \end{bmatrix}^{-1} \begin{bmatrix} \hat{\rho}_1 \\ \hat{\rho}_2 \\ \vdots \\ \hat{\rho}_p \end{bmatrix}.$$

Los estimadores obtenidos son conocidos como los *Estimadores de Yule-Walker*. Por ejemplo, se simulo un proceso AR(3) mediante

$$Z_t = 0.6448\, Z_{t-1} - 0.0634\, Z_{t-2} - 0.2198\, Z_{t-3} + a_t, \tag{2.11}$$

donde $a_t \sim \mathrm{N}(0, 1)$. De esta manera se estimaron sus parámetros a través de la ecuaciones de Yule-Walker para verificar la validez de este procedimiento. Los resultados se muestran en la tabla 2.1.

Simulación	ϕ_1	ϕ_2	ϕ_3	ϕ_4	ϕ_5	ϕ_6
1	0.6843	-0.1890	-0.1778	-0.0558	0.1994	-0.1364
2	0.5988	-0.1158	-0.1839	-	-	-
3	0.6502	-0.0501	-0.2399	-	-	-
4	0.6678	-0.2131	-0.1503	-	-	-

Cuadro 2.1: Estimadores de Yule-Walker para el modelo 2.11.

Al obtener $\hat{\phi}_1, \hat{\phi}_2, \ldots, \hat{\phi}_p$, utilizamos el siguiente resultado

$$\begin{aligned}\gamma_0 =& \mathbb{E}\left[\dot{Z}_t\, \dot{Z}_t\right] = \mathbb{E}\left[\dot{Z}_t \left(\phi_1\, \dot{Z}_{t-1} + \phi_2\, \dot{Z}_{t-2} + \cdots + \phi_p\, \dot{Z}_{t-p} + a_t\right)\right] \\ =& \phi_1\, \gamma_1 + \phi_2\, \gamma_2 + \cdots + \phi_p\, \gamma_p + \sigma_a^2\end{aligned}$$

para conseguir el estimador por momentos de σ_a^2,

$$\hat{\sigma}_a^2 = \hat{\gamma}_0 \left(1 - \hat{\phi}_1\, \hat{\rho}_1 - \hat{\phi}_2\, \hat{\rho}_2 - \cdots - \hat{\phi}_p\, \hat{\rho}_p\right).$$

Entonces, para el modelo anterior, la estimación de su varianza es la siguiente,

Simulación	$\hat{\sigma}_a^2$
1	0.9505
2	1.1420
3	1.1520
4	0.9667

Cuadro 2.2: Estimación de la varianza para el modelo 2.11.

2.3.2. Método de máxima verosimilitud

La forma exacta y general de la función de verosimilitud para un modelo ARMA es muy complicada. Empero, para ilustrar su forma en un modelo de series de tiempo, consideremos un proceso AR(1)

$$\left(1 - \phi B\right) \dot{Z} = a_t \quad \text{ó} \quad \dot{Z}_t = \phi\, \dot{Z}_{t-1} + a_t.$$

Al reescribir este proceso en una rerpresentación de promedios móviles, se tiene que

$$\dot{Z}_t = \sum_{j=0}^{\infty} \phi^j \, a_{t-j}.$$

Es claro que las variables $\dot{Z}_t$ tienen una distribución $N\left(0, \sigma_a^2/(1-\phi^2)\right)$. Sin embargo, las variables $\dot{Z}_t$ están altamente correlacionadas. Para derivar la función de densidad conjunta de $(\dot{Z}_1, \dot{Z}_2, \ldots, \dot{Z}_n)$, teniendo en mente que se desea la función de verosimilitud, observemos que

$$\begin{aligned} e_1 &= \sum_{j=0}^{\infty} \phi^j \, a_{1-j} = \dot{Z}_1, \\ a_2 &= \dot{Z}_2 - \phi \, \dot{Z}_1, \\ a_3 &= \dot{Z}_3 - \phi \, \dot{Z}_2, \\ &\vdots \\ a_n &= \dot{Z}_n - \phi \, \dot{Z}_{n-1}. \end{aligned} \tag{2.12}$$

Es importante señalar que e_1 sigue una distribución $N\left(0, \sigma_a^2/(1-\phi^2)\right)$, a_t se distribuye $N\left(0, \sigma_a^2\right)$ y e_1 es independiente de a_t para $2 \leq t \leq n$. Entonces, la función de densidad conjunta de $(e_1, a_2, \ldots, a_n)$ será igual a

$$\mathbb{P}\left[e_1, a_2, \ldots, a_n\right] = \left[\frac{1-\phi^2}{2\,\pi\,\sigma_a^2}\right]^{1/2} \exp\left[-\frac{e_1^2\,(1-\phi^2)}{2\,\sigma_a^2}\right]^{(n-1)/2} \exp\left[-\frac{1}{2\,\sigma_a^2}\sum_{t=2}^{n} a_t^2\right].$$

Ahora, consideremos la siguiente transformación:

$$\begin{aligned} \dot{Z}_1 &= e_1, \\ \dot{Z}_2 &= \phi \, \dot{Z}_1 + a_2, \\ \dot{Z}_3 &= \phi \, \dot{Z}_2 + a_3, \\ &\vdots \\ \dot{Z}_n &= \phi \, \dot{Z}_{n-1} + a_n. \end{aligned}$$

El Jacobiano de esta transformación es igual a

$$\mathbf{J} = \begin{bmatrix} 1 & 0 & \cdot & \cdot & \cdot & \cdot & \cdot & 0 \\ -\phi & 1 & 0 & \cdot & \cdot & \cdot & \cdot & 0 \\ 0 & -\phi & 1 & 0 & \cdot & \cdot & \cdot & 0 \\ \vdots & \vdots & \vdots & \vdots & & & & \vdots \\ 0 & \cdot & \cdot & \cdot & \cdot & 0 & -\phi & 1 \end{bmatrix} = 1,$$

De lo anterior se deduce que

$$\mathbb{P}\left[\dot{Z}_1, \dot{Z}_2, \ldots, \dot{Z}_n\right] = \mathbb{P}\left[e_1, a_2, \ldots, a_n\right]$$

$$= \left[\frac{1-\phi^2}{2\,\pi\,\sigma_a^2}\right]^{1/2} \exp\left[-\frac{\dot{Z}_1^2\,(1-\phi^2)}{2\,\sigma_a^2}\right] \left[\frac{1-\phi^2}{2\,\pi\,\sigma_a^2}\right]^{(n+1)/2} \exp\left[-\frac{1}{2\,\sigma_a^2}\sum_{t=2}^{n}(\dot{Z}_t - \phi\,\dot{Z}_{t-1})^2\right].$$

Por lo tanto, la función exacta de log-verosimilitud está dada por

$$\ln L\left(\dot{Z}_1, \ldots, Z_n | \,\phi, \mu, \sigma_a^2\right) = -\frac{n}{2}\,\ln\left(2\,\pi\right) + \frac{1}{2}\,\ln(1-\phi^2) - \frac{n}{2}\,\ln\left(\sigma_a^2\right) - \frac{S\left(\phi,\mu\right)}{2\,\sigma_a^2},$$

donde

$$S\left(\phi,\mu\right) = (Z_1-\mu)^2(1-\phi^2) + \sum_{t=2}^{n}\left[(Z_t-\mu) - \phi\,(Z_{t-1}-\mu)\right]^2$$

es la suma de cuadrados solamente en función de ϕ y de μ. Por ejemplo, al calcular los parámteros para el modelo (2.11) se obtienen los siguientes resultados:

Simulación	ϕ_1	ϕ_2	ϕ_3	ϕ_4	ϕ_5	ϕ_6
1	0.6820	-0.1866	-0.1783	-0.0580	0.1996	-0.1359
2	0.6040	-0.1096	-0.1839	-	-	-
3	0.6537	-0.0481	-0.2511	-	-	-
4	0.6652	-0.2105	-0.1519	-	-	-

Cuadro 2.3: Estimadores por Máxima Verosimilitud para el modelo 2.11.

Sin embargo, la función de verosimilitud para un modelo ARMA se puede aproximar a partir de dos mecanismos:

- La estimación máximo verosímil condicional; y
- La estimación máximo verosímil no condicional.

En la estimación máximo verosímil condicional, considerando la hipótesis de que $\{Z_t\}$ es estacionario y $\{a_t\}$ es una serie de variables aleatorias i.i.d. $N\left(0, \sigma_a^2\right)$, para el modelo estacionario (2.10) la distribución conjunta de $\mathbf{a} = (a_1, a_2, \ldots, a_n)^T$ estará dada por

$$f(\mathbf{a}|\,\phi,\mu,\theta,\sigma_a^2) = \left(2\,\pi\,\sigma_a^2\right)^{-n/2} \exp\left\{-\frac{1}{2\,\sigma_a^2}\sum_{t=1}^{n} a_t^2\right\}.$$

Entonces al reescribir (2.10) como

$$a_t = \theta_1 a_{t-1} + \cdots + \theta_q a_{t-q} + \dot{Z}_t - \phi_1 \dot{Z}_{t-1} - \cdots - \phi_p \dot{Z}_{t-p},$$

podemos expresar la función de verosimilitud en función de los parámetros $(\phi, \mu, \theta, \sigma_a^2)$. Sea $\mathbf{Z} = (Z_1, Z_2, \ldots, Z_n)^T$ y se suponen las condiciones iniciales $\mathbf{Z}_* = (Z_{1-p}, \ldots, Z_{-1}, Z_0)^T$ y $\mathbf{a}_* = (a_{1-q}, \ldots, a_{-1}, a_0)^T$. Entonces la función de logverosímilitud condicional está dada por

$$\ln L_* (\phi, \mu, \theta, \sigma_a^2) = -\frac{n}{2} \ln (2\pi\sigma_a^2) - \frac{S_* (\phi, \mu, \theta)}{2\sigma_a^2} \tag{2.13}$$

donde

$$S_* (\phi, \mu, \theta) = \sum_{t=1}^{n} a_t^2 (\phi, \mu, \theta | \mathbf{Z}_*, \mathbf{a}_*, \mathbf{Z}) \tag{2.14}$$

es la función condicional de la suma de cuadrados. Por lo tanto, las cantidades $\hat{\phi}$, $\hat{\mu}$ y $\hat{\theta}$ que maximizan la Ecuación (2.13) se conocen como los estimadores máximo verosimiles condicionales.

Existen algunas alternativas para especificar las condiciones iniciales $\mathbf{Z}_*$ y $\mathbf{a}_*$. Como $\{Z_t\}$ es estacionario y $\{a_t\}$ es una serie de variables aleatorias i.i.d. $N(0, \sigma_a^2)$, entonces podemos reemplazar el valor desconocido de Z_t por la media muestral $\bar{Z}$ y el valor desconocido de a_t por su valor esperado. Para el modelo (2.10), también podemos suponer que $a_p = a_{p-1} = \cdots = a_{p-q+1} = 0$ y calcular a_t para $t \geq p+1$ utilizando (2.10). Es así que la suma de cuadrados condicional (2.14) se convierte en

$$S_* (\phi, \mu, \theta) = \sum_{t=1}^{n} a_t^2 (\phi, \mu, \theta | \mathbf{Z}), \tag{2.15}$$

Después de obtener los parámetros estimados, $\hat{\phi}$, $\hat{\mu}$ y $\hat{\theta}$, la estimación de σ_a^2 se calcula a partir de

$$\hat{\sigma}_a^2 = \frac{S_* (\hat{\phi}, \hat{\mu}, \hat{\theta})}{g.l.},$$

donde el número de grados de libertad, $g.l.$, es igual al número de términos utilizados en la suma de $S_* (\hat{\phi}, \hat{\mu}, \hat{\theta})$ menos el número de parámetros estimados. Si (2.15) se utiliza para calcular la suma de cuadrados, entonces,

$$g.l. = (n-p) - (p+q+1) = n - (2p+q+1).$$

Para poder hacer una mejora en la estimación, Box y Jenkins (1976) sugieren la siguiente función de logverosímilitud incondicional

$$\ln L (\phi, \mu, \theta, \sigma_a^2) = -\frac{n}{2} \ln (2\pi\sigma_a^2) - \frac{S (\phi, \mu, \theta)}{2\sigma_a^2} \tag{2.16}$$

donde $S(\phi, \mu, \theta)$ es la suma de cuadrados incondicional dada por

$$S(\phi, \mu, \theta) = \sum_{t=-\infty}^{n} \mathbb{E}^2 [a_t | \phi, \mu, \theta, \mathbf{Z}] \tag{2.17}$$

y $\mathbb{E}[a_t | \phi, \mu, \theta, \mathbf{Z}]$ es la esperanza condicional de a_t dados ϕ, μ, θ y $\mathbf{Z}$. Las cantidades $\hat{\phi}$, $\hat{\mu}$ y $\hat{\theta}$ que maximizan la Ecuación (2.16) se conocen como los estimadores máximo verosimiles incondicionales.

En la práctica, la suma (2.17) se aproxima mediante

$$S(\phi, \mu, \theta) = \sum_{t=-M}^{n} \mathbb{E}^2 [a_t | \phi, \mu, \theta, \mathbf{Z}],$$

donde M es un entero lo suficientemente grande de tal manera que

$$| \mathbb{E}[Z_t | \phi, \mu, \theta, \mathbf{Z}] - \mathbb{E}[Z_{t-1} | \phi, \mu, \theta, \mathbf{Z}] | < \varepsilon$$

para cualquier valor predeterminado ε donde $t \leq -(m+1)$. Esto implica que $\mathbb{E}[Z_t | \phi, \mu, \theta, \mathbf{Z}] \simeq \mu$ y por lo tanto $\mathbb{E}[a_t | \phi, \mu, \theta, \mathbf{Z}]$ es insignificante para $t \leq -(M+1)$.

La estimación de σ_a^2 se puede calcular de la siguiente manera

$$\hat{\sigma}_a^2 = \frac{S(\hat{\phi}, \hat{\mu}, \hat{\theta})}{n}.$$

2.4. Criterios para la selección del modelo

En el análisis de series de tiempo existen distintas formas para determinar el modelo adecuado. Para un determinado conjunto de datos, cuando existen múltiples modelos a considerar, el criterio de selección se basa en el resumen de residuales a partir de un modelo estimado. En esta sección se intoducirán algunos de los principales criterios para la selección del modelo básados en el análisis de residuales.

Criterio AIC de Akaike

Para evaluar la calidad del modelo estimado, Akaike (1973) introdujo el siguiente criterio de información:

$$AIC(M) = -2 \ln [\text{función de verosímilitud}] + 2M$$

donde M es el número de parámetros en el modelo. Por ejemplo, para un modelo ARMA y para n observaciones, se tiene que

$$\ln L = -\frac{n}{2} \ln (2\pi\sigma_a^2) = -\frac{1}{2\sigma_a^2} S(\phi, \mu, \theta). \tag{2.18}$$

Maximizando (2.18) con respecto a ϕ, μ, θ y σ_a^2, tenemos que,

$$\ln L = -\frac{n}{2}\,\ln\left(\hat{\sigma}_a^2\right) - \frac{n}{2}\,(1 + \ln\left(2\,\pi\right)). \tag{2.19}$$

Debido a que el segundo término en (2.19) es una constante, el criterio de AIC se reduce a

$$AIC(M) = n \ln\left(\hat{\sigma}_a^2\right) + 2\,M.$$

El orden óptimo del modelo es elegido por el valor M, que es una función de p y q, de tal manera que el $AIC(M)$ sea mínimo. Por ejemplo, el orden óptimo para la serie de datos Pigs[1], bajo un modelo AR(p), se obtiene cuando $p = 12$. Los resultados se muestran en la siguiente tabla:

Orden	AIC	Orden	AIC
1	49.125331	7	8.144714
2	6.218262	8	9.362332
3	7.788230	9	11.15654
4	4.302370	10	7.296035
5	4.864974	11	2.635909
6	6.602800	12	0.000000

Cuadro 2.4: Elección del orden de un modelo AR(p) a través del criterio AIC

Criterio BIC

Shibata (1976) ha demostrado que el criterio AIC tiende a sobrestimar el orden de la autorregresión. Por esta razón Akaike (1978,1979) ha desarrollado una extensión Bayesiana del AIC, conocida como el criterio BIC, el cual tiene la siguiente forma:

$$BIC\,(M) = n \ln\left(\hat{\sigma}_a^2\right) - (n - M)\,\ln\left(1 - \frac{M}{n}\right) + M \ln n + M_z\,\ln\left[\left(\frac{\hat{\sigma}_z^2}{\hat{\sigma}_a^2} - 1\right)/M\right],$$

donde $\hat{\sigma}_a^2$ es el estimador máximo verosimil de σ_a^2, M es el número de parámetros y $\hat{\sigma}_Z^2$ es la varianza muestral de la serie.

[1]Los datos corresponden al número total de cerdos obtenidos en un matadero en Victoria, EU. de enero de 1980 a agosto de 1995, ABS.

Criterio SBC de Schwartz

De forma similar al cirterio BIC de Akaike, Schwartz (1978) sugiere el siguiente criterio Bayesiano para la selección del modelo, el cual es conocido como cirterio SBC (Schwartz's Bayesian Criterion):

$$SBC\left(M\right) = n\ln\left(\hat{\sigma}_a^2\right) + M\ln n. \tag{2.20}$$

De nuevo en (2.20), $\hat{\sigma}_a^2$ es el estimador máximo verosimil de σ_a^2, M es el número de parámetros en el modelo y n es el número de observaciones que es equivalente a la cantidad de residuales que pueden calcularse a partir de la serie.

Criterio CAT de Parzen

En 1977 Parzen sugiere el siguiente criterio de selección del modelo, el cual se cononoce como CAT (criterio para las funciones autorregresivas de transferencia):

$$CAT\left(p\right) = \begin{cases} -\left(1+\frac{1}{n}\right), & p = 0, \\ \frac{1}{n}\sum_{j=1}^{p}\frac{1}{\hat{\sigma}_j^2} - \frac{1}{\hat{\sigma}_p^2}, & p = 1,2,3,\ldots \end{cases}$$

donde $\hat{\sigma}_j^2$ es el estimador insesgado de σ_a^2 cuando un modelo $AR\left(j\right)$ es estimado para la serie y n es el número de observaciones. El orden óptimo se elige de tal manera que el criterio $CAT\left(p\right)$ sea mínimo.

2.5. Predicción

2.5.1. Predicción de errores de media cuadrática para modelos ARMA

Uno de los objetivos más importantes en el análisis de series de tiempo es el de predecir valores futuros. La mayoria de los resultados sobre predicción se obtienen a partir de la teoría general de predicción lineal desarrollada por Kolmogorov (1939, 1941), Wiener (1949), Kalman (1960), Yaglom (1962), y Whittle (1983), entre otros.

Para derivar la predicción del error en media cuadrática, consideremos el modelo ARMA estacionario

$$\phi\left(B\right)Z_t = \theta\left(B\right)a_t.$$

Como el modelo es estacionario, entonces,

$$\begin{aligned} Z_t &= \psi\left(B\right)a_t \\ &= a_t + \psi_1\, a_{t-1} + \psi_2\, a_{t-2} + \cdots \end{aligned} \tag{2.21}$$

donde

$$\psi(B) = \sum_{j=0}^{\infty} \psi_j B^j = \frac{\theta(B)}{\psi(B)}$$

y $\psi_0 = 1$. Para $t = n + l$, tenemos que

$$Z_{n+l} = \sum_{j=0}^{\infty} \psi_j \, a_{n+l-j}. \tag{2.22}$$

Supongamos que en el tiempo $t = n$ se tienen las observaciones $Z_n, Z_{n-1}, Z_{n-2}, \ldots$ y deseamos pronosticar el valor futuro Z_{n+l} como una combinación lineal de las observaciones $Z_n, Z_{n-1}, Z_{n-2}, \ldots$. Dado que Z_t para $t = n, n-1, n-2, \ldots$ se puede escribir de la forma (2.21), la predicción del error de media cuadrática se puede expresar como

$$\hat{Z}_n(l) = \psi_l^* \, a_n + \psi_{l+1}^* \, a_{n-1} + \psi_{l+2}^* \, a_{n-2} + \cdots$$

donde ψ_j^* debe ser determinada. De esta manera, la predicción del error en media cuadrática será igual a

$$\mathbb{E}\left[(Z_{n+l} - \hat{Z}_n(l))^2\right] = \sigma_a^2 \sum_{j=0}^{l-1} \psi_j^2 + \sigma_a^2 \sum_{j=0}^{\infty} (\psi_{l+j} - \psi_{l+j}^*)^2,$$

la cual es más fácil de minimizar cuando $\psi_{l+j}^* = \psi_{l+j}$. Entonces,

$$\hat{Z}_n(l) = \psi_1 \, a_n + \psi_{l+1} \, a_{n-1} + \psi_{l+2} \, a_{n-2} + \cdots .$$

Pero utilizando (2.22) y que

$$\mathbb{E}\left[a_{n+j} | \, Z_n, Z_{n-1}, \ldots\right] = \begin{cases} 0, & j > 0, \\ a_{n+j}, & j \leq 0. \end{cases}$$

tenemos

$$\mathbb{E}\left[Z_{n+j} | \, Z_n, Z_{n-1}, \ldots\right] = \psi_1 \, a_n + \psi_{l+1} \, a_{n-1} + \psi_{l+2} \, a_{n-2} + \cdots .$$

Por lo tanto, la predicción del error en media cuadrática de Z_{n+l} estará dada por

$$\hat{Z}_n(l) = \mathbb{E}\left[Z_{n+j} | \, Z_n, Z_{n-1}, \ldots\right].$$

$\hat{Z}_n(l)$ es conocido como el paso de tamaño l de la predicción Z_{n+l} con origen en n. El error de predicción está dado por

$$e_n(l) = Z_{n+l} - \hat{Z}_n(l) = \sum_{j=0}^{l-1} \psi_j \, a_{n+l-j}. \tag{2.23}$$

Debido a que $\mathbb{E}\left[e_n(l) | \, Z_t, t \leq n\right] = 0$, la predicción es insesgada con varianza

$$\mathbb{V}\left[\hat{Z}_n(l)\right] = \mathbb{V}\left[e_n(l)\right] = \sigma_a^2 \sum_{j=0}^{l-1} \psi_j^2.$$

Por ejemplo, para un proceso normal, los límites de predicción del $(1-\alpha)\,100\,\%$ son

$$\hat{Z}_n \pm N_{\alpha/2}\left[1+\sum_{j=1}^{l-1}\psi_j^2\right]^{1/2}\sigma_a$$

donde $N_{\alpha/2}$ es el cuantil de una distribución normal estándar de tal manera que $\mathbb{P}\left[N > N_{\alpha/2}\right] = \alpha/2$.

Como se muestra en (2.23), la predicción del error $e_n(l)$ es una combinación lineal de las perturbaciones aleatorias futuras que se incorporan en el sistema después del tiempo n. Especificamente, la predicción del error a un paso es

$$e_n(1) = Z_{n+1} - \hat{Z}_n(1) = a_{n+1}.$$

Por lo que, las predicciones de los errores a un paso son independientes. Esto implica que $\hat{Z}_n(1)$ es sin duda el mejor predictor de Z_{n+1}. De lo contrario, si las predicciones de los errores a un paso son correlacionadas, entonces uno puede calcular la predicción $\hat{a}_{n+1}$ de a_{n+1} de los errores disponibles $a_n, a_{n-1}, a_{n-2}, \ldots$ y por lo tanto mejorar la predicción de Z_{n+1} simplemente utilizando $\hat{Z}_n(1) + \hat{a}_{n+1}$ como la predicción. Sin embrago, la predicción de los errores para intervalos de tiempo mayores es correlacionada. Es decir, considerando que

$$e_n(l) = Z_{n+l} - \hat{Z}_n(l) = a_{n+l} + \psi_1\, a_{n+l-1} + \cdots + \psi_{l-1}\, a_{n+1}$$

y

$$e_{n-j}(l) = Z_{n+l-j} - \hat{Z}_{n-j}(l) = a_{n+l-j} + \psi_1\, a_{n+l-j-1} + \cdots + \psi_{l-1}\, a_{n-j+1},$$

se realizan al mismo tiempo l pero cuyos origenes son n y $n-j$ para $j < l$.

2.6. Modelo ARCH

Los modelos econométricos tradicionales suponen que la varianza en un periodo de tiempo determinado es constante, lo cual se deriva de la forma en que se plantean. Es decir, se supone que son estacionarios, ya sea en un sentido débil o estricto. Empero, en economía, el comportamiento del momento actual responde a la expectativa generada sobre el valor del cambio producido en el momento precedente. Considerando lo anterior, Robert F. Engle (1982) determina un patrón de comportamiento estadístico para la varianza al introducir una nueva clase de procesos estocásticos, conocidos como *Modelos Autorregresivos Heterocedásticos: ARCH.* En respuesta al modelo ARCH y debido a un gran número de sofisticaciones se generó una gran familia de modelos, entre los cuales se encuentran los modelos GARCH, IGARCH, EARCH, TARCH, SWARCH, QS-ARCH, APARCH, FACTOR-ARCH, entre otros.

Engle (1982) cita tres situaciones que motivan y justifican la modelación de la heterocedasticidad condicional autorregresiva:

- La experiencia empírica nos lleva a contrastar períodos de amplia varianza de error seguidos de otros de varianza más pequeña. Es decir, el valor de la dispersión del error respecto a su media cambia en el tiempo.
- En segundo lugar, Engle (1982) expone la validez de estos modelos para determinar los criterios de mantenimiento o venta de activos financieros. Los agentes económicos deciden en función de la información proveniente del pasado respecto al valor medio de su rentabilidad y la volatilidad que ésta ha tenido. Con los modelos ARCH se tendrían en cuenta estos dos condicionantes.
- El modelo de regresión ARCH puede ser una aproximación a un sistema más complejo. Los modelos estructurales que admiten una especificación tipo ARCH infinito, se determinan a partir del cambio de los parámetros. Esto permite que este tipo de modelos sean capaces de contrastar la hipótesis de permanencia estructural.

Es así que, en primera instancia, Engle (1982) considera un modelo autorregresivo de primer orden

$$\dot{Z}_t = \phi_1 \, \dot{Z}_{t-1} + a_t.$$

El enfoque estándar de heterocedasticidad es el de introducir una variable exógena, x_t, de tal manera que x_t predice la variación. Con una media igual a cero, un primer modelo a plantear es el siguiente

$$\dot{Z}_t = a_t \, x_{t-1},$$

donde $\mathbb{V}[a_t] = \sigma_a^2$. La varianza de $\dot{Z}_t$ simplemente es igual a $\sigma_a^2 \, x_{t-1}^2$ y, por lo tanto, el intervalo del pronóstico depende de la evolución de una variable exógena. Sin embargo esta representación estándar parece ser insuficiente. Un segundo modelo que permite a la varianza condicional depender de la última realización de la serie, es el modelo bilineal descrito por Granger y Andersen (1978). Un caso simple de este modelo es el siguiente

$$\dot{Z}_t = a_t \, \dot{Z}_{t-1}.$$

En este nuevo modelo, la variaza condicional es igual a $\sigma_a^2 \, \dot{Z}_{t-1}^2$. Sin embargo la varianza incondicional es cero o infinito, lo que hace que este planteamiento sea poco atractivo. Teniendo en cuenta las limitaciones de los modelos anterores, Engle (1982) plantea el siguiente modelo

$$\dot{Z}_t = a_t \, h_t^{1/2}, \tag{2.24}$$

donde $h_t = \alpha_0 + \alpha_1 \dot{Z}_{t-1}^2$ y $\mathbb{V}[a_t] = 1$. Este modelo es conocido como proceso Autorregresivo Heterocedástico de primer orden, ARCH(1). Si consideramos la hipótesis de normalidad y si denotamos al conjunto de la información pasada como ψ_{t-1}, entonces

$$\dot{Z}_t | \psi_{t-1} \sim N(0, h_t), \quad h_t = \alpha_0 + \alpha_1 \dot{Z}_{t-1}^2. \tag{2.25}$$

Además la varianza del modelo puede ser expresada de manera más general como

$$h_t = h(Z_{t-1}, Z_{t-2}, \ldots, Z_{t-p}, \alpha)$$

donde p es el orden del proceso ARCH y α es el vector de parámetros desconocidos.

Sin embargo, es posible plantear algunas variantes para la expresión de la varianza del modelo, las cuales pueden ser más apropiadas para determinadas aplicaciones. Por ejemplo, Engle (1982) considera las siguientes dos alternativas:

- $h_t = \exp(\alpha_0 + \alpha_1 Z_{t-1}^2)$,
- $h_t = \alpha_0 + \alpha_1 |Z_{t-1}|$.

Como Engle (1982) menciona, estas dos formas ofrecen un contraste interesante. La forma exponencial tiene la ventaja de que la varianza es positiva para todos los valores de alfa, pero no es difícil demostrar que los datos generados a partir de este modelo tienen varianza infinita para cualquier valor de $\alpha_1 \neq 0$. La forma del valor absoluto requiere que ambos parámetros sean positivos, pero puede demostrarse que la varianza es finita para cualquier valor de los parámetros.

2.6.1. Condiciones de estacionariedad para el modelo ARCH

Engle (1982) plantea, mediante la existencia de los momentos del proceso, las condiciones para que el proceso (2.24) sea debilmente estacionario. El siguiente Teorema caracteriza los momentos del proceso ARCH.

Teorema 2. *Para un entero r, los momentos de orden 2r de un proceso lineal ARCH de primer orden con* $\alpha_0 > 0$, $\alpha_1 \geq 0$ *existen si, y solo si,*

$$\alpha_1^r \prod_{j=1}^{r} (2j-1) < 1.$$

Por simetría los momentos impares son igules a cero.

Para calcular los momentos de un proceso lineal ARCH de primer orden la demostración del Teorema 2 proporciona una expresión para estos momentos. Para un proceso ARCH de orden p las condiciones para que el proceso sea estacionario se generalizan a partir del Teorema 2 mediante el siguiente teorema,

Teorema 3. *El proceso lineal ARCH de orden p, con* $\alpha_0 > 0$, $\alpha_1, \ldots, \alpha_p \geq 0$, *tiene una covarianza estacionaria si, y sólo si, la ecuación característica asociada tiene todas sus raíces fuera del círculo unitario. La varianza estacionaria está dada por*

$$\mathbb{E}\left[Z_t^2\right] = \frac{\alpha_0}{1 - \sum_{j=1}^{p} \alpha_j}.$$

2.6.2. Modelo de regresión ARCH

Como menciona Engle (1982), una interpretación alternativa para el modelo ARCH, es considerar que los errores en un modelo de regresión lineal siguen un proceso ARCH. Si el orden del modelo es p, entonces las condiciones para este proceso son

$$\begin{aligned} \dot{Z}_t |\, \psi_{t-1} &\sim N\left(x_t \beta, h_t\right), \\ h_t &= \alpha_0 + \alpha_1\, \epsilon_{t-1}^2 + \cdots + \alpha_p\, \epsilon_{t-p}^2, \\ \epsilon_t &= \dot{Z}_t - x_t\, \beta, \\ l &= \frac{1}{T} \sum_{t=1}^{T} l_t, \\ l_t &= -\frac{1}{2} \log h_t - \frac{1}{2}\, \epsilon_t^2 / h_t, \end{aligned} \tag{2.26}$$

donde x_t incluye un retraso dependiente y variables hexógenas. Esta función de verosimilitud será maximizada con respecto a los parámetros desconocidos α y β.

Bajo las condiciones en (2.26), el estimador por mínimos cuadrados de β es consistente cuando x y ϵ no están correlacionados. Si las $x's$ son constantes fijas, entonces el error estándar por mínimos cuadrados será correcto; sin embargo, si existen variables de retraso en x_t y si los errores estándar se calculan convencionalmente, entonces estos no serán consistentes, ya que los cuadrados de los errores estarán correlacionados con los cuadrados de las $x's$. Esta es una extensión del argumento de White (1980) sobre heterocedasticidad y además sugiere el uso de una forma alternativa para la matriz de covarianza, de tal manera que proporciona una estimación consistente de los errores estándar por mínimos cuadrados.

Si los regresores no incluyen variables de retraso dependientes, el proceso es estacionario y si consideramos a Z como un vector de variables dependientes de dimensión $T \times 1$ y a x como una matriz de variables independientes de dimensión $T \times K$, entonces

$$\begin{aligned} \mathbb{E}\left[Z|x\right] &= x\, \beta, \\ \mathbb{V}\left[Z|x\right] &= \sigma_a^2\, l, \end{aligned}$$

además las hipótesis de Gauss-Markov se satisfacen. La estimación ordinaria por mínimos cuadrados resulta ser el mejor estimador lineal insesgado para el modelo (2.26); además, la varianza estimada es consistente e insesgada. Sin embargo, la estimación por máxima verosimilitud es asintóticamente superior. Esto se debe a que la estimación por mínimos cuadrados no alcanza el límite de Cramér y Rao. Los estimadores máximo verosímiles para β y para α se obtienen a partir de la solución de las condiciones de primer orden y de la expresión analítica de la matriz de información. La expresión para la matriz de información que Engle (1982) proporciona es la siguiente:

$$\begin{aligned}\mathtt{g}_{\beta\beta} &= \frac{1}{T}\sum_t \mathbb{E}\left[\mathbb{E}\left[\frac{\partial^2 l_t}{\partial\beta\,\partial\beta'}|\,\psi_{t-1}\right]\right] \\ &= \frac{1}{T}\sum_t \mathbb{E}\left[\frac{x_t' x_t}{h_t} + \frac{1}{2\,h_t^2}\frac{\partial h_t}{\partial\beta}\frac{\partial h_t}{\partial\beta'}\right].\end{aligned}$$

Entonces el estimador consistente para un modelo de regresión ARCH de orden p estará dado por

$$\hat{\mathtt{g}}_{\beta\beta} = \frac{1}{T}\sum\left[\frac{x_t' x_t}{h_t} + 2\sum_j \alpha_j^2\,\frac{\epsilon_{t-j}^2}{h_t^2}\,x_{t-j}'\,x_{t-j}\right]. \tag{2.27}$$

Al agrupar $x_t'\,x_t$, en (2.27) la ecuación se puede reescribir como

$$\begin{aligned}\hat{\mathtt{g}}_{\beta\beta} &= \frac{1}{T}\sum_t x_t'\,x_t\left[h_t^{-1} + 2\,\epsilon_t^2\sum_{j=1}^{p}\alpha_j^2\,h_{t+j}^{-2}\right] \\ &\equiv \frac{1}{T}\sum_t x_t'\,x_t\,r_t^2.\end{aligned}$$

De manera similar, los elementos de la diagonal inferior de la matriz de información se pueden expresar de la siguiente forma:

$$\hat{\mathtt{g}}_{\alpha\,\beta} = \frac{1}{T}\sum_t \mathbb{E}\left[\frac{1}{2\,h_t^2}\frac{\partial h_t}{\partial\alpha}\frac{\partial h_t}{\partial\beta}\right].$$

Antes de presentar el siguiente resultado, definiremos algunos conceptos. Si ξ_t es un vector de variables aleatorias de dimensión $p\times 1$ del espacio muestral Ξ, cuyos elementos son $\xi_t^{'} = (\xi_{t-1},\ldots,xi_{t-p})$. Para caulquier ξ_t, si ξ_t^* es idéntico excepto por el m-ésimo elemento que ha sido multiplicado por -1 y si m se encuentra entre 1 y p, entonces

Definición 5. *El proceso ARCH definido por* (2.25) *es simétrico si*

- $h\,(\xi_t) = h\,(\xi_t^*)$ *para toda* m *y para* $\xi\in\Xi$,
- $\partial h\,(\xi_t)/\partial\alpha_j = \partial h\,(\xi_t^*)/\partial\alpha_j$ *para toda* m, i *y para* $\xi\in\Xi$,

- $\partial h\left(\xi_t\right)/\partial\xi_{t-m} = -\partial h\left(\xi_t^*\right)/\partial\xi_{t-m}$ *para toda* m *y para* $\xi \in \Xi$.

Otra caracterización de los modelos ARCH generales está en términos de las condiciones de regularidad.

Definición 6. *El modelo ARCH definido por* (2.25) *es regular si*

- $\min h\left(\xi_t\right) \geq \delta$ *para algun* $\delta > 0$ *y para* $\xi \in \Xi$,
- $\mathbb{E}\left[\,|\,\partial h(\xi_t)/\partial\alpha_i\,|\,|\,\partial h(\xi_t)/\partial\xi_{t-m}\,|\,|\,\psi_{t-m-1}\,|\,\right]$ *existe para toda* i, m, t.

De esta manera podemos asegurar el siguiente resultado.

Teorema 4. *El modelo lineal ARCH de orden p satisface las condiciones de regularidad si* $\alpha_0 > 0$ *y* $\alpha_1, \ldots, \alpha_p \geq 0$.

Este último Teorema determina las condiciones bajo las cuales los elementos de la diagonal inferior de la matriz de información son iguales a cero.

Teorema 5. *Si un modelo de regresión ARCH es simétrico y regular, entonces* $\hat{\mathfrak{g}}_{\alpha\,\beta} = 0$.

La demostración de ambos Teoremas puede consultarse en Engle (1982). Las implicaciones para el Teorema 5 son de gran envergadura, ya que la estimación de α y β puede realizarse por separado sin pérdida de eficiencia asintótica. Además sus varianzas pueden calcularse por separado.

2.6.3. Estimación del modelo de regresión ARCH

Dado que la estimación de α y de β se puede hacer de manera separada y considerando que el cálculo de cualquiera se puede estimar con base en el cálculo del otro, el procedimiento recomendado es estimar inicialmente β por el método de mínimos cuadrados y obtener los residuales. A partir de estos residuales, se puede construir un estimador eficiente para α. La estimación para los parámetros del modelo se basan en el algoritmo de puntaje (*scoring*). Cada paso del vector de parámetros ϕ produce la estimación ϕ^{i+1} basada en ϕ^i de acuerdo a

$$\phi^{i+1} = \phi^i + [\hat{\mathfrak{g}}^i_{\phi\,\phi}]^{-1}\frac{1}{T}\sum_t \frac{\partial\, l_t^i}{\partial\phi},$$

donde $\hat{\mathfrak{g}}^i$ y $\partial\, l_t^i/\partial\phi$ se evalúan en ϕ^i. La ventaja de este algoritmo es, en parte, que requiere sólo las primeras derivadas de la función de verosimilitud y, en parte, que utiliza las propiedades estadísticas de los problemas para adaptar el algoritmo para su aplicación.

En el modelo lineal de orden p, el valor de α se obtiene mediante la siguiente iteración

$$\alpha^{i+1} = \alpha^i + (\tilde{z}'\tilde{z})^{-1}\tilde{z}'\, f^i$$

donde

$$\begin{aligned}
\tilde{z}_t &= (1, e_{t-1}^2, \ldots, e_{t-p}^2)/h_t^i, \\
\tilde{z}_t' &= (\tilde{z}_1', \ldots, \tilde{z}_T'), \\
f_t^i &= (e_t^2 - h_t^i)/h_t^i, \\
f^{i'} &= (f_1^i, \ldots, f_T^i).
\end{aligned}$$

En estas expresiones, e_t es el residual de la i-ésima iteración , h_t^i es la varianza condicional estimada y α^i es la estimación del vector de parámetros desconocidos en la iteración i. Cada paso es fácilmente construido a partir de una regresión por mínimos cuadrados en las variables transformadas. La matriz de varianzas y covarianzas de los parámetros es consistentemente estimada por $2\,(\tilde{z}'\,\tilde{z})^{-1}$. Esta expresión difiere ligeramente de $\hat{\sigma}^2\,(\tilde{z}'\,\tilde{z})^{-1}$ calculado para la regresión auxiliar. Asintóticamente, $\hat{\sigma}^2 = 2$, siempre que las hipótesis distribucionales sean correctas.

Los parámetros en α deben estar definidos positivamente y deben cumplir algunas condiciones de estacionariedad. Estas restricciones podrían ser impuestas a través de funciones de penalidad. Este último enfoque se utiliza en Engle (1982), aunque tal vez una reformulación del modelo podría emplear elementos al cuadrado para imponer restricciones de no negatividad:

$$h_t = \alpha_0^2 + \alpha_1\,\epsilon_{t-1} + \cdots + \alpha_p^2\,\epsilon_{t-p}^2.$$

La convergencia para cada iteración puede formularse de distintas maneras. Un criterio simple es del gradiente en torno a la inversa del Hessiano. Para un vector de parámetros, ϕ, se tiene que

$$\theta = \frac{\partial l'}{\partial \phi}\left(\frac{\partial^2 l}{\partial\phi\,\partial\phi'}\right)^{-1}\frac{\partial l}{\partial\phi}. \tag{2.28}$$

Usar θ como un criterio de convergencia suena atractivo, ya que ofrece una normalización natural y se puede interpretar como el término remanente en una expansión de series de Taylor. En todo caso, sustituyendo el gradiente y la matriz de información estimada en (2.28), $\theta = R^2$ en la regresión auxiliar.

Para una estimación determinada de α, el algoritmo de scoring se puede utilizar para mejorar la estimación de β. El algoritmo de scoring para β es

$$\beta^{i+1} = \beta^i + [\hat{\mathbf{g}}_{\beta\,\beta}]^{-1}\frac{\partial l^i}{\partial\beta}. \tag{2.29}$$

Si se define $\tilde{x}_t = x_t\, r_t$ y $\tilde{e}_t = e_t\, s_t/r_t$ con $\tilde{x}$ y $\tilde{e}$ como la correspondiente matriz y vector, (2.29) se puede reescribir para estimar ϵ_t en la i-ésima iteración mediante

$$\beta^{i+1} = \beta^i + (\tilde{x}'\,\tilde{x})^{-1}\tilde{x}'\,\tilde{e}.$$

Entonces, se calculan iterativamente los valores para β para poder obtener la matriz de varianzas covarianzas final de la estimación máximo verosimil de β.

Bajo las condiciones del teorema de Crowder (1976) para martingalas, se puede establecer que los estimadores máximo verosimiles $\hat{\alpha}$ y $\hat{\beta}$ tienen una distribución asintótica normal

$$\sqrt{T\,(\hat{\alpha} - \alpha)} \xrightarrow{D} N\,(0, \mathsf{g}^{-1}_{\alpha\,\alpha}),$$
$$\sqrt{T\,(\hat{\beta} - \beta)} \xrightarrow{D} N\,(0, \mathsf{g}^{-1}_{\beta\,\beta}).$$

Preliminares

Capítulo 3

Modelos de mezclas finitas

Un modelo de mezclas finitas es capaz de modelar de manera completa distribuciones complejas. Puede, por lo tanto, manejar situaciones en donde alguna familia paramétrica no satisface un modelo debido a las variaciones locales en los datos observados. De hecho, el uso de modelos de mezclas finitas se remonta a finales del siglo XIX en donde Newcomb (1886) presenta una aplicación de mezclas normales para datos discrepantes (*outliers*). Además, como toda distribución continua puede ser aproximada por una mezcla finita de densidades normales con varianza común, o matriz de covarainzas en el caso multivariado, los modelos de mezclas finitas proveen un conveniente marco semiparamétrico en donde un modelo distribucional desconocido toma forma. Para demostrar esto último, Marron y Wand (1992) presentaron algunos ejemplos de mezclas de distribuciones normales univariadas correspondientes a distintas combinaciones de sus componentes[1]. La idea es que cualquier densidad puede ser aproximada arbritrariamente por una mezcla de densidades normales, lo cual se ilustra en las gráficas de la Figura 3.1.

Entonces, a partir de lo expuesto por Titterington (1985) y McLachlan (2000), la finalidad de este capítulo es presentar la teoría básica sobre modelos de mezclas finitas. Es así que se define a la función de densidad de una mezcla como una combinación lineal de distintas densidades. Asimismo, en las secciones 3.2 y 3.3 se generaliza este concepto al introducir variables categóricas y continuas en una misma mezcla y al suponer modelos lineales generales como los componentes de la misma. Por último en la Sección 3.4 se presentan las condiciones necesarias para que distintas mezclas sean identificables.

[1] Estos componentes se listan en el Apéndice B.

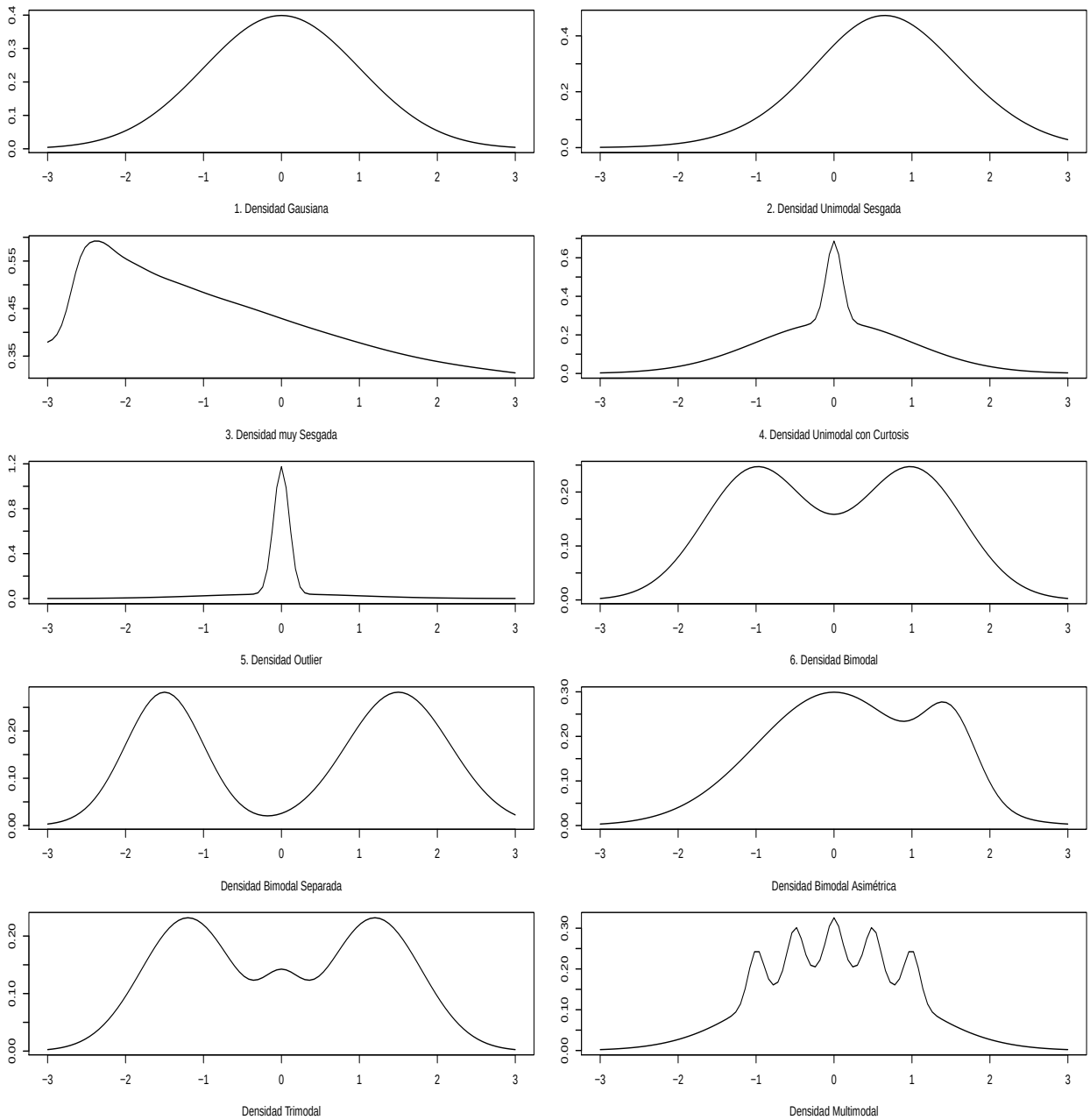

Figura 3.1: Mezcla de ditribuciones normales univariadas. De Marron y Wand (1992)

3.1. Definiciones y conceptos básicos

Supongamos que tenemos una variable (o vector) aleatoria X que toma valores en un espacio muestral Ω y cuya distribución puede ser representada por una función de densidad de la siguiente forma

$$f_M(x) = \pi_1 f_1(x) + \cdots + \pi_k f_k(x), \quad x \in \mathscr{X}; \tag{3.1}$$

donde

$$\pi_j > 0, \quad j = 1, \ldots, k; \quad \pi_1 + \cdots + \pi_k = 1$$

y

$$f_j(\cdot) \geq 0, \quad \int_{\mathscr{X}} f_j(x)\, dx = 1, \quad j = 1, \ldots, k.$$

Entonces, se dice que X es una mezcla finita de densidades y que $f_M(\cdot)$ es la *función de densidad de una mezcla finita*. Los parámetros $\pi_1, \ldots, \pi_k$ se conocen como los pesos de la mezcla y $f_1(\cdot), \ldots, f_k(\cdot)$ son las densidades de composición de la mezcla. En muchas situaciones $f_1(\cdot), \ldots, f_k(\cdot)$ se pueden especificar mediante una forma paramétrica, y el lado derecho de (3.1) tendrá una representación más explicita

$$\pi_1 f_1(x|\, \theta_1) + \cdots + \pi_k f_k(x|\, \theta_k), \tag{3.2}$$

en donde θ_j denota los distintos parámetros de $f_j(\cdot)$. Se va a denotar a la colección de todos los parámetros producidos en las densidades de la mezcla por $\dot{\theta}$ y a la colección completa de todos los distintos parámetros producidos en el modelo por $\dot{\Theta}$. Por ejemplo, un modelo mezcla frecuentemente utilizado es el que consiste de dos componentes normales homocedásticos,

$$f_M(x) = \pi\, \phi(x|\, \mu_1, \sigma^2) + (1 - \pi)\, \phi(x|\, \mu_2, \sigma^2), \tag{3.3}$$

donde

$$\phi(x|\, \mu_k, \sigma^2) = (2\,\pi)^{-\frac{1}{2}}\, \sigma^{-1} \exp\left\{ -\frac{1}{2\,\sigma^2}(x - \mu_k)^2 \right\}$$

denota a la función de densidad normal univariada con media μ_j y varianza σ^2. En este caso $\pi_1 = \pi$, $\pi_2 = 1 - \pi$, $\theta_1 = (\mu_1, \sigma)$, $\theta_2 = (\mu_2, \sigma)$, $\dot{\theta} = (\mu_1, \mu_2, \sigma)$ y $\dot{\Theta} = (\pi, \mu_1, \mu_2, \sigma)$.

En los modelos de mezclas finitas no existe el requisito de que las densidades en (3.2) pertenezcan a la misma familia parámetrica. Por lo tanto, la función de densidad de una mezcla finita tendrá la siguiente forma

$$f_M(x|\, \dot{\Theta}) = \sum_{j=1}^{k} \pi_j\, f(x|\, \dot{\theta}_j), \tag{3.4}$$

donde $f\left(\cdot|\,\dot{\theta}\right)$ denota a un miembro genérico de la familia parámetrica. En el caso de un modelo mezcla finito definido por la expresión (3.4) cada uno de los $\dot{\theta}_j$ para $j = 1, \ldots, k$ es un elemento del mismo espacio paramétrco, denotado por Θ. Si $G_\pi\left(\cdot\right)$ denota una medida de probabilidad sobre Θ definida por $\dot{\pi}$, entonces (3.4) puede ser escrita de manera formal como

$$f_M\left(x|\,\dot{\Theta}\right) = \int_\Theta f\left(x|\,\dot{\theta}\right) d\, G_\pi\left(\dot{\theta}\right). \tag{3.5}$$

Sin embargo, centraremos nuestra atención al caso en que $G_\pi(\cdot)$ define una medida finita y discreta sobre θ. La forma de la expresión (3.5) sugiere una obvia generalización de la mezcla, debido a que $G_\pi(\cdot)$ es una medida más general sobre Θ. Cabe señalar que en la formulación de este modelo el número de componentes es fijo. Pero de hecho en muchas aplicaciones, el valor de k es desconocido y tiene que ser inferido de los datos disponibles, junto con las proporciones y parámetros de la mezcla.

Una interpretación de los modelos de mezclas finitas, es considerar la introducción de Z como una variable aleatoria categórica que toma los valores $1, \ldots, k$ con probabilidades $\pi_1, \ldots, \pi_k$ respectivamente. Si la densidad condicional de X dada $Z = j$ es $f_j\left(x\right)$ para $j = 1, \ldots, k$, entonces la densidad marginal de X estará dada por

$$f\left(x|\dot{\theta}_j\right) = f_j\left(x\right).$$

En este contexto la variable Z puede ser pensada como una variable que etiqueta a la variable X. Entonces será conveniente trabajar con k componentes etiquetados por el vector Z, en lugar de una única variable categórica Z, donde el j-ésimo elemento de Z está definido como cero o uno, en función de que el componente de origen de X en la mezcla sea igual a i o no. Por lo tanto Z sigue una distribución multinomial que consiste de k categorías con probabilidades $\pi_1, \ldots, \pi_k$, es decir

$$\mathbb{P}\left[Z_j = z_j\right] = \pi_1^{x_{1j}}\, \pi_2^{x_{2j}} \cdots \pi_k^{x_{kj}}; \tag{3.6}$$

y escribiremos

$$Z_j \sim \mathrm{Mult}_k\left(1, \dot{\pi}\right), \tag{3.7}$$

donde $\dot{\pi} = (\pi_1, \ldots, \pi_k)$.

En la interpretación del modelo de mezclas finitas (3.1) se presenta la situación en la que el modelo se aplica de manera directa. Es decir, cuando X se extrae de una población G, la cual consiste de k grupos, $G_1, \ldots, G_k$ en proporciones $\pi_1, \ldots, \pi_k$. Si la densidad de X en el grupo G_j está dada por $f_j\left(x\right)$ para $j = 1, \ldots, k$, entonces la densidad de X tendrá la forma de la mezcla (3.1). En esta situación, los k componentes de la mezcla se podrán identificar físicamente con los k grupos existentes.

En la mayoria de los ejemplos presentados en la literatura estadística donde la población es una mezcla de k distintos grupos, se tiene el conocimiento de una distribución a priori, por lo menos en un sentido físico. Sin embargo, en muchos de los ejemplos que manejan el uso de modelos mezcla, los componenetes no pueden ser identificados a través de la existencia de grupos externos. En algunas situaciones, los componentes son introducidos dentro del modelo de mezclas finitas para permitir mayor flexibilidad en la modelación de una población heterogénea que aparentemente es incapaz de ser modelada por un único componente. En el extremo de este ejercicio, podemos obtener la estimación no paramétrica del kernel de una densidad si se encaja una mezcla con $k = n$ componentes en proporciones iguales $1/n$, donde n es el tamaño de la muestra observada. Por ejemplo, si $x_1, \ldots, x_n$ denota una muestra univariada de tamaño n, entonces podemos obtener la estimación del kernel de la densidad de X de la siguiente manera

$$\hat{f}(x) = \frac{1}{nh} \sum_{i=1}^{n} k\left(\frac{x - x_i}{h}\right), \tag{3.8}$$

si en (3.1) fijamos $k = n$ y $\pi_i = 1/n$, entonces

$$f_M(x) = \frac{1}{h} k\left(\frac{x - x_i}{h}\right)$$

para alguna función kernel $k(\cdot)$ y para algún parámetro h.

Para ilustrar algunas de las formas adoptadas por una mezcla de densidades univariadas consideremos la mezcla (3.3). Si los dos componentes normales se encuentran lo suficientemente separados, entonces se esperaría que la densidad de la mezcla sea una densidad bimodal. Para demostrar esto, se ha graficado la densidad de la mezcla para distintos valores de Δ en los casos donde $\mu_1 = 0$, $\mu_2 = \Delta$, $\sigma^2 = 1$ y cuyos pesos son iguales ($\pi = 1 - \pi = 0.5$).

Se puede observar que conforme Δ aumenta, la forma de $f_M(x)$ cambia de unimodal a bimodal. El umbral para este cambio es cuando $\Delta = 2$ donde, de manera más general, se tiene que

$$\Delta = \frac{\mid \mu_1 - \mu_2 \mid}{\sigma}$$

es la distancia de Mahalanobis entre componentes homocedásticos de una mezcla de densidades normales. Esta gráfica muestra la forma en donde la resolución gráfica de una mezcla en sus componentes puede ser una simple tarea cuando los componentes están ampliamente separados ($\Delta = 3$ y $\Delta = 4$), pero que puede ser todo un reto cuando los componentes están demasiado juntos ($\Delta = 1$).

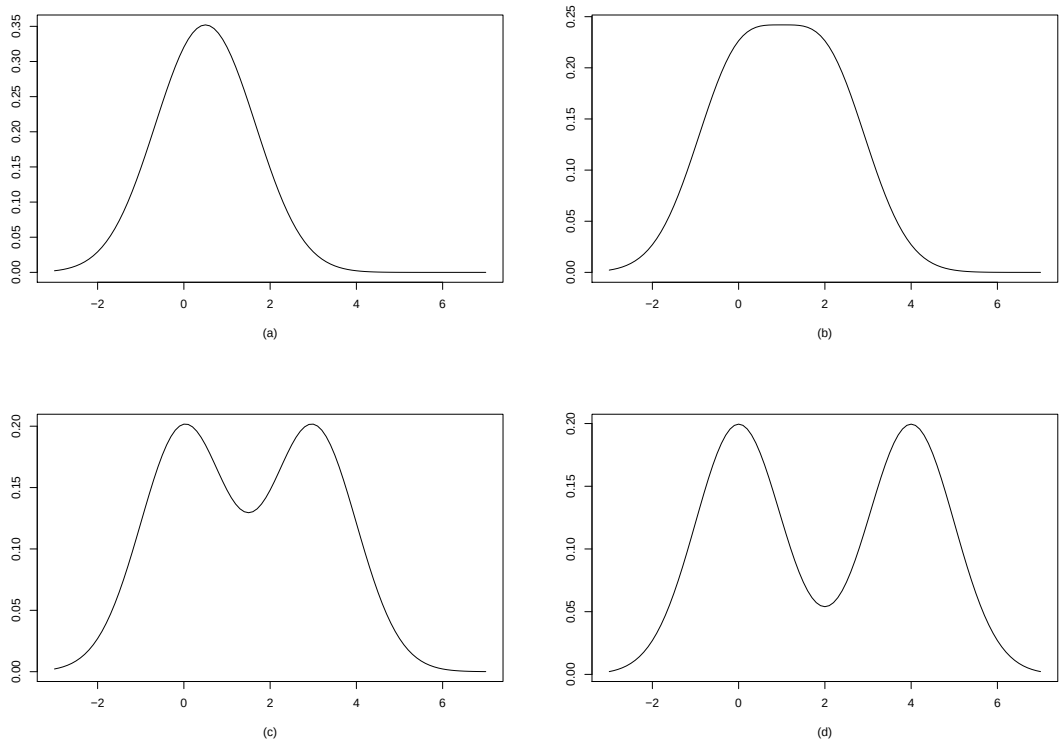

Figura 3.2: Gráfica de una mezcla de dos distribuciones normales univariadas en proporciones iguales con varianza común $\sigma^2 = 1$ y medias $\mu_1 = 0$ y $\mu_2 = \Delta$, en este caso (a) $\Delta = 1$; (b) $\Delta = 2$; (c) $\Delta = 3$ y (d) $\Delta = 4$.

Si las medias de las dos densidades del modelo (3.3) están lo suficiente cercanas, el traslape entre las dos densidades se hace más evidente y la distinción entre los componentes se dificulta más si estos no están representados en proporciones iguales.

Para demostrar esto, consideremos el modelo anteriormente planteado, pero cuyos pesos son iguales a $\pi = 0.75$ y $1-\pi = 0.25$. En este caso, se puede ver que la forma de la mezcla cambia de sesgada a bimodal para $\Delta = 4$. La forma de la mezcla para $\Delta = 3$ demuestra bitangencialidad. Esta ocurre cuando dos puntos distintos x_1 y x_2 comparten la misma tangente. Por lo que la bitangencialidad está implícita, pero no implica, bimodalidad. De manera informal, la bimodalidad implica una joroba extra en la curva, pero la bitangencialidad es simplemente un golpe extra de unimodalidad.

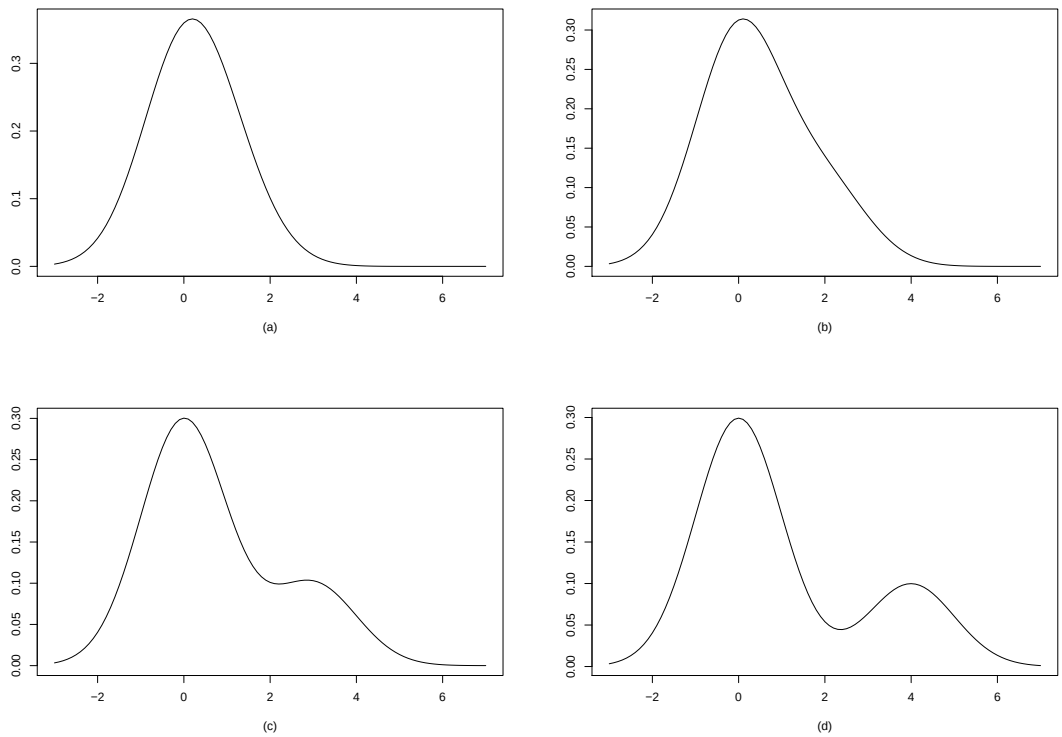

Figura 3.3: Gráfica de una mezcla de dos distribuciones normales univariadas en proporciones 0.75 y 0.25 con varianza común $\sigma^2 = 1$ y medias $\mu_1 = 0$ y $\mu_2 = \Delta$, en este caso (a) $\Delta = 1$; (b) $\Delta = 2$; (c) $\Delta = 3$ y (d) $\Delta = 4$.

3.2. Mezclas continuas y variables categóricas

Consideraremos el modelo mezcla

$$f\left(x_i|\,\dot{\Theta}\right) = \sum_{j=1}^{k} \pi_j\, f_j\left(x_i|\,\theta_j\right) \tag{3.9}$$

donde algunas de las variables son categóricas. La forma más sencilla de modelar las densidades de las variables mixtas es considerar que las variables categóricas son independientes una de otra y asumir que las funciones que corresponden a las variables continuas se han adoptado para, por ejemplo, una distribución normal multivariada. Aunque esto parece una manera burda de proceder, es común que en la práctica ocurra como una forma de agrupación mixta de las características de los datos.

Podemos refinar este enfoque adoptando el modelo de localización. Supongamos que p_1 de las p variables características en X_i son categóricas, donde la q-ésima variable toma m_q distintos valores para $q = 1, \ldots, p_1$. Entonces existen

$$w = \prod_{q=1}^{p_1} w_q$$

distintos modelos de estas p_1 variables categóricas. Con el modelo de localización, las p_1 variables categóricas se sustituirán por una sola variable aleatoria multinomial $X_i^{(1)}$ con m celdas. Es decir, $(X_i^{(1)})_s = 1$ si las realizaciones de las p_1 variables categóricas en X_i corresponde al s-ésimo modelo. Cualquier asociación entre las variables categóricas originales se convierte en relaciones entre las resultantes celdas de probabilidad multinomial. El modelo de localización asume que la distribución de las $p - p_1$ variables continuas es normal, con media $\mu_{j\,s}$ y matriz de varianzas y covarianzas Σ_j , que es la misma para todas las s celdas.

En la práctica, el número de parámetros con el enfoque del modelo de localización puede ser grande si la distribución multinomial que sustituye a las variables categóricas tiene demasiadas celdas. Supongamos que las primeras dos variables características son variables binarias que, sin pérdida de generalidad, toman los valores uno o cero. Entonces definimos la variable $x_{v\,i} = (x_i)_v$ para $v = 1, \ldots, p$. Además considermos que $X_{1\,i}$ es independiente de todas las otras variables características, pero que $X_{2\,i}$ no es independiente de las restantes $p - 2$ variables continuas. Entonces la partición del vector característico x_i será

$$x_i = (x_{1\,i}, x_{2\,i}, x_i^{(2)^T})^T, \tag{3.10}$$

donde $x_i^{(2)}$ es el vector que contiene las $p-2$ variables continuas. Entonces la j-ésima densidad de X_j se modela como

$$f_j\,(x_i) = \left\{ \prod_{v=1}^{2} f_B\,(x_{vi} |\, \theta_{v\,j}) \right\} \phi\,(x_i^{(2)} |\, \mu_{j\,s}, \sigma_j), \tag{3.11}$$

donde

$$f_B\,(w) = \theta^w\,(1-\theta)^{1-w}$$

denota la función de densidad para una variable aleatoria binaria que toma los valores cero y uno con probabilidades θ y $1-\theta$ respectivamente, y $\mu_{j\,s}$, $s = 1, 2$ es la media de $X_i^{(2)}$ correspondiente a los dos distintos valores de la variable binaria $X_{2\,i}$.

Si el ajuste preliminar sugiere que no es razonable tomar las dos variables binarias $X_{1\,i}$ y $X_{2\,i}$ independientes, entonces serán sustituidas por una única variable multinomial, $X_i^{(2)}$ con $s = 4$ celdas. Por lo tanto, la distribución condicional del subvector de variables continuas $X_i^{(2)}$ se asume que tiene un distribución normal con matriz de varianzas y covarianzas Σ_j y media $\mu_{j\,s}$ que corresponden a $(X_i^{(1)})_s = 1$ para $s = 1, \ldots, 4$.

En una situación general que involucra variables categóricas y continuas, se divide la función de vectores en tantos subvectores como sea posible de tal manera que su distribución sea independiente. El caso extremo sería considerar que todas las p variables características son independientes y que incluya la estructura de correlación cuando sea necesario. La estructura de correlación en un modelo se debe ampliar para incluir asociaciones locales adicionales entre las variables categóricas, continuas o ambas. Esto se considera formalmente en términos del cambio en la log-verosimilitud. Pero las consideraciones informales incluyen la inspección de las gráficas de dispersión. En contraste con el uso del estadístico de prueba $-2\log\lambda$ para el número de componentes, su nula distribución asintótica con los grados de libertad iguales a las diferencias entre los dos modelos, no deben dar resultados engañosos en esta función cuando los modelos están anidados.

3.3. Mezclas de modelos lineales generalizados

El *modelo lineal generalizado* (GLM) originalmente fue propuesto por Nelder y Wedderburn (1972). Este modelo propone que el logaritmo de la densidad (univariada) de la variable X_i tiene la siguiente forma

$$\log f\left(x_i;\theta_i,\kappa\right) = m_i\,\kappa^{-1}\{\theta_i\,x_i - b\left(\theta_i\right)\} + c\left(x_i;\kappa\right), \tag{3.12}$$

donde θ_i es el parámetro natural o canónico, κ es el parámetro de dispersión y m_i es el peso a priori. La media y varianza de X_i estarán dadas por

$$\mathbb{E}\left[X_j\right] = \mu_i = b'\left(\theta_i\right)$$

y

$$\mathbb{V}\left[X_i\right] = \kappa b''\left(\theta_j\right).$$

respectivamente, donde $b'\left(\cdot\right)$ y $b''\left(\cdot\right)$ denotan la primera y segunda derivada con respecto a θ_i. En un GLM, se asume que

$$\begin{aligned}\eta_i &= g\left(\mu_i\right)\\ &= y_i^T\,\beta,\end{aligned}$$

donde y_i es un vector de covariables o de variables exploratorias en la j-ésima respuesta x_j, β es un vector de parámetros desconicidos y $g(\cdot)$ es una función monótona, conocida como la función liga. Entonces, para una mezcla con k distribuciones en proporciones $\pi_1,\ldots,\pi_k$, se tiene que la densidad de la j-ésima variable de respuesta X_i está dada por

$$f_M\left(x_i|\,\dot{\Theta}\right) = \sum_{j=1}^{k}\pi_j\,f\left(x_i|\,\theta_{i\,j},\kappa_j\right), \tag{3.13}$$

donde para un parámetro de dispersión κ_j, se obtiene que

$$\log f\left(x_i;\theta_{i\,j},\kappa_j\right)=m_i\,\kappa_j^{-1}\left\{\theta_{i\,j}\,x_i-b_i\left(\theta_{i\,j}\right)\right\}+c_j\left(x_i;\kappa_j\right)\quad\text{para}\quad j=1\ldots,k \tag{3.14}$$

Para el i-ésimo componente del GLM, consideramos a μ_j como la media de X_i, $h_j\left(\mu_{i\,j}\right)$ como la función de enlace y $\eta_j=h_j\left(\mu_{i\,j}\right)=\beta_j^T\,y_i$ como el predictor lineal $(j=1,\ldots,k)$.

Los componentes de la mezcla serán los mismos que en el GLM, de modo que la log-densidad para el i-ésimo componente se puede escribir de la siguiente forma

$$\log f\left(x_i;\theta_{i\,j},\kappa_j\right)=\kappa_j^{-1}\{\theta_{i\,j}\,x_i-b_i\left(\theta_{i\,j}\right)\}+c_j\left(x_i;\kappa_j\right),\qquad j=1,\ldots,k. \tag{3.15}$$

Las proporciones de la mezcla pueden ser modeladas como función de un vector de covariables asociado con la respuesta. Este vector de covariables puede o no tener algunos elementos en común con el vector de covariables y, de los cuales los componetentes de las medias de la mezcla puedan depender. Sin pérdida de generalidad, denotaremos ambos vectores de covariables por y , como covariables irrelevantes en las formas de regresión para medias canónicas y cuyas proporciones de la mezcla pueden tener sus coeficientes iguales a cero.

Un modelo común para expresar el i-ésimo peso de la mezcla como una función de y, es el modelo logístico. Bajo este modelo, podemos hacer una correspondencia entre la i-ésima observación con el vector de covariables y_i

$$\begin{aligned}\pi_{i\,j}&=\pi_j(y_i;\alpha)\\&=\exp(\omega_j^T\,y_i)/\{1+\sum_{h=1}^{k-1}\exp(\omega_h^T\,y_i)\}\qquad j=1,\ldots,k,\end{aligned} \tag{3.16}$$

donde $\omega_k=0$ y $\alpha=(\omega_1^T,\ldots,\omega_{k-1}^T)^T$ contiene los coeficientes de la regresión logística. El primer elemento de y_i usualmente es tomado como uno, de modo que el primer elemento de cada ω_j sea una intersección. Si ahora $\dot{\Theta}$ es el vector de parámetros desconocidos, dado por

$$\dot{\Theta}=(\alpha^T,\beta^T)^T,$$

donde β contiene los elementos conocidos de $\beta_1,\ldots,\beta_k$ a priori que son distintos. Como las proporciones de la mezcla son modeladas para depender de alguna o de todas las covariables, esto implica que puede haber problemas de identificabilidad con algunos de los parámetros en α y β. En párticular con los términos de intersección de α y los elementos de β.

3.4. Identificabilidad

En general, una familia paramétrica de distribuciones $F(x, \dot{\Theta})$ es identificable si distintos valores del parámetro $\dot{\Theta}$ determinan distintos miembros de la familia de las distribuciones

$$\{F\,(x, \dot{\Theta}) : \dot{\Theta} \in \Omega\},$$

donde Ω es el espacio de parámetros especificado; es decir,

$$F\,(x, \dot{\Theta}) = F\,(x, \dot{\Theta}^*), \tag{3.17}$$

si y solo si

$$\dot{\Theta} = \dot{\Theta}^*. \tag{3.18}$$

Entonces, consideremos a

$$\mathscr{C} = \{F\,(x, \theta_j),\, \theta_j \in \Omega,\, x \in \mathbb{R}^d\}$$

como la clase de distribuciones d-dimensionales de donde se forman las mezclas. Por lo tanto, identificamos a la clase de mezclas finitas de $\mathscr{C}$ con la clase apropiada de funciones de distribución

$$\mathscr{H} = \left\{ H\,(x) : H\,(x) = \sum_{j=1}^{k} \pi_j\, F\,(x, \theta_j)\ \pi_j > 0,\ \sum_{j=1}^{k} \pi_j = 1,\ F\,(\cdot, \theta_j) \in \mathscr{C},\ \forall j, k = 1, 2, \ldots, x \in \mathbb{R}^d \right\}.$$

Podemos abreviar $F\,(x, \theta_j)$ por $F_j\,(x)$, y en ocasiones, denotar a la mezcla por

$$H = \sum_{j=1}^{k} \pi_j\, F_j.$$

En todas la expresiones, $F_1, \ldots, F_k$, se asumirán que son distintos miembros de $\mathscr{C}$.

Definición 7. *(Identificabilidad) Supongamos que H, H^* son dos miembros de $\mathscr{H}$, por lo que*

$$H = \sum_{j=1}^{k} \pi_j\, F_j, \qquad H^* = \sum_{j=1}^{k^*} \pi_j^*\, F_j^*,$$

entonces $H \equiv H^*$ si, y sólo si, $k = k^*$ y podemos ordenar la sumas de tal manera que $\pi_j = \pi_j^*$, $F_j = F_j^*$, $j = 1, \ldots, k$. Entonces $\mathscr{H}$ es identificable.

En pocas palabras, lo que la Definición 7 nos dice, es que $\mathscr{H}$ es identificable, si y sólo si, todos los miembros de $\mathscr{H}$ son distintos. Notemos que definiciones equivalentes de $\mathscr{H}$ o de identificabilidad pueden ser escritas en términos de funciones de densidad, siempre y cuando estas existan.

Algunas veces en la práctica, en particular con datos univariados, puede existir un orden natural de los componentes de acuerdo al tamaño de sus medias. Esta falta de identificabilidad no es motivo de preocupación tanto en el desarrllo como en la estimación de los modelos de mezclas finitas por máxima verosimilitud, y en especial en el algoritmo EM.

Como señala Crawford (1994) la no identificabilidad debida a la sobrevaloración, es decir, el ajuste de demasiados componentes en el modelo, es muy problemático. Por ejemplo, el modelar una mezcla de $k-1$ componentes incorrectamente por una mezcla de k componentes se puede dar de las siguientes dos maneras:

- Uno de los pesos de los k componentes de la mezcla puede ser igual a cero.
- Dos de las densidades en los k componentes de la mezcla se pueden tomar como la misma.

Un enfoque distinto sobre el problema de la identificabilidad es el que utiliza una función de identificabilidad (Kadane, 1974). Esta es esencialmente la misma que el enfoque de Redner (1981) el cual utiliza el cociente del espacio topológico $\tilde{\Omega}$ obtenido por un mapeo equivalente a los valores de $\dot{\Theta}$ en un único punto.

3.4.1. Un teorema sobre identificabilidad

Definimos a $< \mathscr{C} >$ como la expansión de $\mathscr{C}$ sobre el conjunto de los números reales, es decir, es la clase de todas las combinaciones lineales de $\mathscr{C}$.

Teorema 6. *(Yakowitz y Spragins, 1968) Una condición necesaria y suficiente para que $\mathscr{H}$ sea identificable, es que $\mathscr{C}$ sea un conjunto de combinaciones lineales independiente sobre la linea de los números reales, $\mathbb{R}$.*

Demostración. $\Longrightarrow$] Supongamos que $\mathscr{C}$ no es una combinación lineal sobre $\mathbb{R}$, entonces, para alguna $k > 0$, existe una combinación lineal nula de distintos miembros de $\mathscr{C}$ tal que, para alguna m, $0 < m < k$,

$$\sum_{j=1}^{k} \eta_j F_j = 0,$$

con $\eta_j < 0$ para $j \leq m$, $\eta_j > 0$ para $j > m$. Entonces

$$\sum_{j=1}^{m} |\eta_j| F_j \equiv \sum_{j=m+1}^{k} |\eta_j| F_j \tag{3.19}$$

y, dado que $\{F_j\}$ es el conjunto de todas las funciones de distribución, es decir $\sum_{j=1}^{k} F_j = 1$, entonces se tiene que

$$\sum_{j=1}^{m} |\eta_j| = \sum_{j=m+1}^{k} |\eta_j| = b, \quad b \in \mathbb{R} - \{0\}.$$

Si definimos $\pi_j = |\pi_j|/b$ para $j = 1, \ldots, k$, entonces la ecuación (3.19) puede expresarse como

$$\sum_{j=1}^{m} \pi_j F_j \equiv \sum_{j=m+1}^{k} \pi_j F_j$$

y por lo tanto tenemos dos distintas representaciones para la misma mezcla finita, lo que implica que $\mathscr{H}$ es no identificable.

$\Longleftarrow$] Si $\mathscr{C}$ es linealmente independiente sobre $\mathbb{R}$ entonces forma una base sobre $< \mathscr{C} >$. Entonces todo miembro de $< \mathscr{C} >$ tiene una única representación como una combinación lineal de miembros de $\mathscr{C}$. Dado que $\mathscr{H} \subset < \mathscr{C} >$, la identificabilidad de $\mathscr{H}$ se sigue de manera inmediata. □

Corolario 1. *$\mathscr{H}$ es identificable si y sólo si la imagen de $\mathscr{C}$ bajo cualquier isomorfismo sobre $< \mathscr{C} >$ es linealmente independiente en la imagen de $\mathscr{C}$.*

El corolario del Teorema 6 tiende a ser más fácil de aplicar directamente que el teorema en sí mismo. En particular, nos permite trabajar en términos de funciones generadoras, que son a menudo más convenientes de manejar matemáticamente que las correspondientes funciones de distribución.

Capítulo 4

Estimación

Podemos argumentar que la estimación de los modelos de mezclas finitas comenzó con el trabajo de Pearson (1894). En este artículo se expone la estimación de los parámetros de una mezcla de dos distribuciones normales heterocedásticas por el método de momentos. En el caso especial donde se tienen $k = 2$ componentes normales con matriz de covarianza común, Lindsay y Basak (1993) derivaron un sistema de ecuaciones de momentos cuya única solución da una estimación de $\dot{\Theta}$. Recientemente, Craigmile y Titterington (1998) consideraron el método de momentos así como el de máxima verosimilitud para mezclas de distribuciones uniformes. En un desarrollo reciente, DasGupta (1999) presento un algoritmo para la estimación de componentes normales cuyas matrices de covarianza son iguales.

Por lo tanto, a partir de lo planteado por Titterington (1985) y McLachlan (2000), se examinarán algunos de los principales métodos para la estimación de los parámetros producidos en un modelo de mezclas finitas. Entre los distintos métodos que se examinarán, se encuentran los métodos gráficos, en donde su análisis se basa en la función de densidad y de distribución empirica; el método de momentos; el método de máxima verosimilitud; el algoritmo EM para modelos mezcla finitos; métodos bayesianos; estimación por mínima distancia; estimaciones basadas en transformaciones y la descomposición numérica de mezclas. Sin embargo el uso de estos métodos depende en gran medida de la forma que adoptan las densidades de los distintos componentes de la mezcla. Por esta razón se da un especial enfásis al algoritmo EM (Dempster, 1977). Esto se debe a su tratabilidad y a que las estimaciones obtenidas son asintóticamente mejores en comparación con los demás métodos descritos en este capítulo.

4.1. Métodos gráficos

Una gran variedad de procedimientos exploratorios basados en gráficas han sido desarrollados para hacer frente a la estimación de los modelos de mezclas finitas. Los objetivos de estos procedimientos generalmente son de dos tipos:

- Para indicar si los datos provienen, o no, de una mezcla de densidades; y
- Para proveer una estimación cruda de los parámetros del modelo de mezclas finitas.

La mayoria de las publicaciones hacen referencia a datos univariados y gran parte de estas solo se ocupan de mezclas cuyas densidades tiene asociada una distribución normal o lognormal. Como se vera más adelante, la mayoria de estos métodos gráficos son intentos para obtener estimaciones crudas de los parámetros de la mezcla, que por lo general y en un principio, son la única forma de un análisis estadístico.

Existen dos tipos principales de gráficas para el análisis de datos univariados. Estos dependen de que la función que se represente sea de distribución o de densidad. En particular, las formas de las gráficas incluyen el histograma y la función de distribución empírica.

4.1.1. Métodos basados en funciones de densidad

Uno de los primeros intentos para determinar el número de componentes de una mezcla era a partir del número de modas que se encuentran en un histograma. Pero el uso de esta gráfica se ve limitado debido a que el número de modas que se contabilicen va a depender de la partición del histograma. Considerando que el número de modas es uno de los principales factores para determinar el número de componentes de la mezcla, Tanner (1962) propone un método en el cual caracteriza a las modas como un máximo local (primera derivada igual a cero y segunda derivada negativa) y las antimodas como un mínimo local (primera derivada igual a cero y segunda derivada positiva). Un enfoque menos crudo, que se basa en el uso de histogramas, es el que describe Bhattacharya (1967). El método se compone de dos partes:

- El logaritmo de una densidad normal es concava en la variable, es decir la derivada es lineal, con pendiente negativa.
- Cuando hay una gran cantidad de datos y la agrupación impuesta por el histograma es muy fina, las alturas del histograma son aproximadamente proporcionales a la densidad.

Entonces, la gráfica de las primeras diferencias de los logaritmos de las frecuencias del histograma de una mezcla de componentes normales bien separados, deberá mostrar una secuencia de pendientes negativas (gráfica de líneas) las cuales corresponderán a cada uno de los componentes. Aunque la gráfica obtenida no será lo suficientemente clara para determinar los componentes de la mezcla, deberá existir evidencia, al menos cruda, de la existencia del número de componentes. Evidentemente, las posiciones y orientaciones de las líneas contienen información que puede ser utilizada para proporcionar estimaciones crudas de los parámetros. Bajo la hipótesis de que el conjunto de datos proviene de una distribución $N(\mu, \sigma^2)$ y que el tamaño del intervalo del histograma es h, Bhattacharya (1967) derivo el siguiente procedimiento para estimar μ y σ^2 a partir de una gráfica de lineas,

$$\hat{\mu} = \lambda + \frac{h}{2}$$
$$\hat{\sigma}^2 = \left[\frac{d \cdot h \cdot \cot(\theta)}{b}\right] - \frac{h^2}{12}$$

donde d y b son escalas relativas sobre los ejes X e Y, λ es la intersección sobre el eje X y θ es el ángulo entre la línea y la dirección negativa del eje de las X. Para la estimación de los pesos de la mezcla, Bhattacharya (1967) sugirio varios métodos basados en el ajuste de mínimos cuadrados de las frecuencias esperadas. Bhattacharya (1967) hace algunas variaciones sobre la matriz de mínimos cuadrados, pero parece poco probable que la mejora obtenida sobre el método de Teaner (1962) sea significativa. En particular, dado que se trata de un método gráfico, las estimaciones sobre la media y la varianza no son fiables.

El traslapo de los componentes predispone claramente las estimaciones. En el método de Bhattacharya, como en muchos otros, es posible restar las frecuencias que puedan tener su origen en los componentes exteriores. Entonces, partiendo de esto, se pueden obtener estimaciones sesgadas. De este modo, el grado de solapamiento se evalúa para estimar las frecuencias en la región de superposición de donde proceden los componentes. Ahora podemos restar de las frecuencias observadas y tambien pueden ser contadas en las estimaciones de los pesos. La sustracción sucesiva informal de los componentes después del ajuste cuadrático a los logaritmos de las frecuencias se describe por Buchanan-Wollaston y Hodgson (1928). Si tres de las frecuencias son utilizadas, una de las cuadráticas se ajusta perfectamente.

4.1.2. Métodos basados en la función de distribución

La alternativa para dibujar una función de densidad es dibujar la función de distribución empírica y ver si existe evidencia de una posible mezcla en la gráfica. Al investigar la posibilidad de una mezcla de distribuciones

normales, es natural el uso de papel de probabilidad normal, lo que conduce a un gráfica normal cuantil-cuantil (Q-Q plot). Este gráfico puede ser descrito como una gráfica de la estimación de $F_M^{-1}(p)$ contra $\Phi^{-1}(p)$, $0 < p < 1$, donde $F_M(\cdot)$ es la función de distribución de la mezcla y $\Phi(\cdot)$ es la función de distribución de una variable aleatoria normal estandar. Algunas desviaciones de linealidad son características de ciertos tipos de mezcla, aunque tiene que haber una buena cantidad de separación para que el patrón sea claro. Dada la falta de fiabilidad de las estimaciones en métodos que se basan solamente en gráficos Q-Q, es probable que se tengan que adecuar a ojo los puntos de inflexión del gráfico para encontrar las estimaciones crudas del modelo.

La superposición de las distribuciones es la causa de uno de los principales problemas en la estimación de los parámetros de la mezcla. Esto se debe a que las líneas de estimación de los parámetros que se obtienen son bastante difíciles de elegir, adémas de ser no lineales, estos no podrán ser los mismos debido a la falta de normalidad y a la influencia de la separación de las gráficas. Empero, los métodos se pueden diseñar para contrarrestar esto. Generalmente al dibujar gráficas es mejor que estas se adapten a los puntos alejados de las zonas de superposición y, en un grado inferior, lejos de las colas, donde es probable que existan pocos datos. Un sesgo potencial en los parámetros de las estimaciones obtenidas utilizando el método anterior se encontrará en las estimaciones de los pesos de la mezcla. Este se basa en las estimaciones de los puntos de inflexión de la función de distribución acumulativa, es decir, mínimos locales de la densidad. El sesgo puede, en principio, ser removido cuando el punto de corte se traslada a x_0, donde para el caso de dos componentes se tiene que

$$\pi_1 \int_{-\infty}^{x_0} f_1(x)\, dx = \pi_2 \int_{x_0}^{\infty} f_2(x)\, dx.$$

Si $f_1(\cdot)$ y $f_2(\cdot)$ tienen una distribución nomal, está ecuación puede ser escrita de la siguiente manera

$$\pi_1 \Phi\left(\frac{x_0 - \mu_1}{\sigma_1}\right) = \pi_2 \left[1 - \Phi\left(\frac{x_0 - \mu_2}{\sigma_2}\right)\right]. \tag{4.1}$$

En la práctica podemos usar las estimaciones iniciales para los parámetros en (4.1) para obtener un valor para el punto de corte, x_0. Esto puede hacerse en cualquiera de los extremos y, en caso necesario, el procedimiento itera para mejorar las estimaciones. Si $\pi_1 = \pi_2$, x_0 se puede encontrar explícitamente en términos de los otros parámetros, para

$$\frac{x_0 - \mu_1}{\sigma_1} = -\frac{x_0 - \mu_2}{\sigma_2},$$

esto es

$$x_0 = \frac{\mu_2\, \sigma_1 + \mu_1\, \sigma_2}{\sigma_1 + \sigma_2}.$$

Este enfoque fue examinado por Brown (1978), que señala que este procedimiento se vuelve tedioso cuando el número de componentes es mayor al que se expone arriba y que puede ser inestable si las estimaciones iniciales

son pobres.

Como una alternativa de las gráficas Q-Q podemos considerar al gráfico P-P, que fue examinado en un contexto de mezclas por Fowlkes (1979). Entonces se graficará

$$\Phi\left(\frac{x_{(i)}-\bar{x}}{s}\right)-p_i$$

contra

$$\frac{x_{(i)}-\bar{x}}{s}, \quad i=1,\ldots,n$$

donde $x_{(1)} \leq \cdots \leq x_{(p)}$ representa una muestra ordenada, $\bar{X}$ y s representan la media y la desviación estandar muestral respectivamente y

$$p_i=\frac{2i-1}{2\,n}, \quad i=1,\ldots,n.$$

Esto da como resultado la gráfica muestral $\Phi - P$ versus Q (Fowlkes, 1979), la cual se puede ver como

$$\Phi\left(\frac{x-\mu}{\sigma}\right)-F\left(x\right) \quad \text{contra} \quad \frac{x-\mu}{\sigma}.$$

Con datos normales, la gráfica es una línea recta horizontal. La evidencia empírica presentada por Fowlkes (1979) sugiere que esta gráfica es al menos tan útil como la gráfica Q-Q para la detección de mezclas.

El uso de gráficas sobre la base de la descomposición en mezclas de Weibull, sobre la misma base que en el caso normal, es examinado por Kao (1959). Para datos exponenciales, la gráfica log(1- función de distribución empírica) versus x deberá ser lineal. Esto lleva también a un análisis gráfico y a técnicas de estimación apropiadas. Las distribuciones uniformes automaticamente producen gráficos Q-Q lineales. Un conjunto de datos de una mezcla de componentes uniformes debe conducir, por lo tanto, a una dispersión lineal compuesta por variables. El cambio de los puntos indica el final del espacio del componente muestral y la pendiente de la dispersión dará lugar a la estimacion de los pesos de la mezcla. El ajuste sistemático de la forma lineal involucra la metodología del ajuste de gráficas de probabilidad y del cambio del punto de inflección.

4.2. Método de momentos

Supongamos que tenemos un conjunto de n observaciones independientes de una población cuyo modelo de probabilidad depende de r parámetros desconocidos. Supongamos además que $\mu\left(\dot{\Theta}\right)$ denota un vector de r momentos funcionalmente independientes y que m denota el conjunto correspondiente a la muestra de momentos. El método de momentos es el estimador $\hat{\Theta}$ que satisface

$$\mu\left(\hat{\Theta}\right)=m. \tag{4.2}$$

En general, existe un gran número de problemas con los estimadores de momentos. Pero para utilizar este método se deberán tener las siguientes consideraciones,

- La solución explicita de (4.2) puede no ser fácil o incluso posible.
- La solución de (4.2) puede no ser única y no puede encontrarse automáticamente en una región factible de $\mathbb{R}^r$.
- Aunque la consistencia de $\mu(\hat{\Theta})$ y, por consiguiente en algunos casos típicos, la consistencia de $\hat{\Theta}$ por lo general sigue la ley de los grandes números, $\hat{\Theta}$ no puede ser asintóticamente eficiente.
- El cálculo exacto de $\mathrm{Cov}(\hat{\Theta})$ normalmente no es posible. Sin embargo, una expansión de Taylor en el argumento puede ser utilizada a menudo para mostrar que, aproximadamente, y para muestras grandes,

$$\mu\left(\dot{\Theta}_0\right) + D\left(\dot{\Theta}_0\right)\left(\hat{\Theta} - \dot{\Theta}_0\right) = m, \tag{4.3}$$

donde D es la matriz cuadrada de derivadas de los elementos en μ y $\dot{\Theta}_0$ es el valor real. Entonces, aproximadadmente,

$$\begin{aligned} \mathrm{Cov}\left(\hat{\Theta}\right) &= D\left(\dot{\Theta}_0\right)^{-1} \mathrm{Cov}_{\dot{\Theta}_0}\left(m\right)\left[D\left(\dot{\Theta}_0\right)^T\right]^{-1} \\ &\approx D\left(\hat{\Theta}\right)^{-1} \mathrm{Cov}_{\hat{\Theta}}(m)\left[D\left(\hat{\Theta}\right)^T\right]^{-1}. \end{aligned} \tag{4.4}$$

Todos estos problemas ocurren frecuentemente con los estimadores de momentos. Sin embargo, existe una larga historia de aplicaciones de tales métodos, en parte debido a que el primer punto normalmente no ocurre en la práctica y en parte por el problema en los cálculos relacionados con métodos alternativos como la máxima verosimilitud esto, por supuesto, antes de la llegada de las computadoras.

4.3. Método de máxima verosimilitud

Supongamos que se tienen n observaciones independientes $X_1 = x_1, \ldots, X_n = x_n$ de una mezcla, entonces la función de verosimilitud asociada a esta muestra es la siguiente

$$L_0\left(\dot{\Theta}\right) = \prod_{i=1}^{n} f\left(x_i|\,\dot{\Theta}\right) = \prod_{i=1}^{n} \sum_{j=1}^{k} \pi_j\, f\left(x_i|\,\dot{\theta}_j\right). \tag{4.5}$$

La maximización de $L_0\left(\dot{\Theta}\right)$ con respecto a $\dot{\Theta}$, para los datos $X_1 = x_1, \ldots, X_n = x_n$, proporciona la estimación máximo verosimil de $\dot{\Theta}$. Equivalentemente, y muy usualmente, la cantidad que se maximiza es la

log-verosimilitud

$$\mathscr{L}_0\left(\dot{\Theta}\right) = \log L_0\left(\dot{\Theta}\right) = \log \prod_{i=1}^{n} \sum_{j=1}^{k} \pi_j \, f\left(x_i | \, \dot{\theta}_j\right) = \sum_{i=1}^{n} \log \left\{ \sum_{j=i}^{k} \pi_j \, f_j(x_i | \, \dot{\theta}_j) \right\}. \tag{4.6}$$

El cálculo directo del estimador máximo verosimil (4.6) requiere de la solución de la ecuación de verosimilitud,

$$\frac{\partial \, \log L_0\left(\dot{\Theta}\right)}{\partial \, \dot{\Theta}} = 0. \tag{4.7}$$

Se puede manipular de tal forma que el MLE de $\dot{\Theta}$ satisfaga

$$\pi_j = \frac{1}{n} \sum_{i=1}^{n} \tau_j\left(x^{(n)}; \hat{\Theta}\right) \qquad j = 1, \ldots, k \tag{4.8}$$

y

$$\sum_{j=1}^{k} \sum_{i=1}^{n} \tau_j\left(x^{(n)}; \hat{\Theta}\right) \frac{\partial}{\partial \dot{\theta}} \log f_j\left(x^{(n)}; \hat{\theta}_j\right) = 0, \tag{4.9}$$

en donde

$$\tau_j\left(x_i; \dot{\Theta}\right) = \frac{\pi_j f_j\left(x_i; \theta_i\right)}{\sum_{h=1}^{k} \pi_h \, f_h\left(x_i; \theta_h\right)} \tag{4.10}$$

es la probabilidad posterior de $x^{(n)}$ correspondiente al i-ésimo componente de la mezcla.

Para modelos paramétricos simples, el enfoque máximo verosimil es muy popular, en parte debido a la existencia de una teoría asintótica atractiva y porque las estimaciones son a menudo fáciles de calcular. Para modelos de mezclas finitas, sin embargo, vamos a descubrir que la teoría asintótica y algunos aspectos sobre su cálculo no son siempre tan sencillos. Para ejemplificar este método, consideremos el siguiente ejemplo:

Ejemplo 1. *Mezcla de dos densidades conocidas*

Supongamos que

$$\mathscr{L}_0\left(\dot{\Theta}\right) = \mathscr{L}_0\left(\pi\right) = \sum_{i=1}^{n} \log\left[\pi \, f_1\left(x_i\right) + (1 - \pi) \, f_2\left(x_i\right)\right] = \sum_{i=1}^{n} \log\left[\pi \left(f_{i1} - f_{i2}\right) + f_{i2}\right],$$

donde $f_{ij} = f_j\left(x_i\right)$ para $j = 1, 2, \ldots, n$. Si además escribimos $p_i = \pi \, f_1\left(x_i\right) + (1 - \pi) \, f_2\left(x_i\right)$, entonces la ecuación de verosimilitud será igual a

$$0 = \frac{\partial \mathscr{L}_0}{\partial \pi} = \sum_{i=1}^{n} \frac{f_{i1} - f_{i2}}{p_i}. \tag{4.11}$$

Existen dos características preocupantes acerca de la solución de (4.11). La primera es que (4.11) es equivalente a una ecuación polinómica de grado hasta (n-1) en π. Sin embargo, existe a lo más una raíz real debido a la concavidad de $\mathscr{L}_0$:

$$\frac{\partial^2 \mathscr{L}_0}{\partial \pi^2} = -\sum_{i=1}^{n} \left[\frac{f_{i1} - f_{i2}}{p_i}\right]^2 < 0.$$

El segundo problema surge porque la solución para (4.11) no puede satisfacer que $0 \leq \hat{\pi} \leq 1$, entonces la estimación máximo verosimil de π estará dada por

- $\hat{\pi}$ si $0 \leq \hat{\pi} \leq 1$;
- 0 si $\partial \mathscr{L}_0 / \partial \pi|_{\pi=0} < 0$;
- 1 si $\partial \mathscr{L}_0 / \partial \pi|_{\pi=1} > 0$.

Si bien es tranquilizante que (4.11) sólo tiene una raíz real, no es posible obtener una solución explícita, por lo que se tendrán que utilizar métodos numéricos, como por ejemplo el método de Raphson o el método de puntajes. Un tercer procedimiento iterativo puede ser utilizado mediante la sustitución de f_{i2} en términos de p_i y f_{i1} en (4.11), entonces reagrupando la ecuación resultante, se tiene que

$$\pi = \frac{1}{n} \sum_{i=1}^{n} \pi \frac{f_{1i}}{p_i} = \frac{1}{n} \sum_{i=1}^{n} w_{i1}(\pi), \tag{4.12}$$

donde w_{i1} claramente está entre 0 y 1. Esto sugiere el siguiente procedimiento de aproximaciones sucesivas:

$$\pi^{(m+1)} = \frac{1}{n} \sum_{i=1}^{n} w_{i1}(\pi^{(m)}), \qquad m = 0, 1, \ldots \tag{4.13}$$

4.4. Algoritmo EM para modelos de mezclas finitas

Hasselblad(1966, 1969), Wolfe(1965, 1967, 1970) y Day (1969) observaron que en algunos casos especiales, las ecuaciones (4.8) y (4.9) sugieren un cálculo iterativo de la solución. Para un valor inicial, $\dot{\Theta}_0$ de $\dot{\Theta}$, un nuevo estimador $\dot{\Theta}^{(1)}$ puede ser calculado para $\dot{\Theta}$; lo que a su vez puede ser sustituido para producir una nueva actualización $\dot{\Theta}^{(2)}$ y así sucesivamente, hasta la convergencia. La solución de la ecuación de verosimilitud se puede identificar como una aplicación directa del algoritmo EM de Dempster (1977). Este procedimiento iterativo consiste de dos pasos, el Paso-E (por esperenza) y el Paso-M (por maximización).

El algoritmo EM es particularmente útil para problemas de estimación donde existen observaciones pérdidas. Supongamos que el conjunto de datos observados consiste del vector

$$x = (x_1, \ldots, x_n)^T,$$

que se considera como incompleto, ya que los vectores $z_1, \ldots, z_n$ asociados al componente de la etiqueta no están disponibles. En este marco, donde cada x_i se conceptualiza como un derivado de los componentes del modelo de mezclas finitas (3.1), z_i es un vector de dimensión k con $z_{ij} = (z_i)_j = 0$ ó 1, esto en función de que x_i surja o no de el k-ésimo componente de la mezcla. Por lo tanto, el vector de datos completo será el siguiente

$$X_c = (x, z)^T, \tag{4.14}$$

donde

$$z = (z_1, \ldots, z_n)^T. \tag{4.15}$$

Los vectores de etiqueta $z_1 \ldots, z_n$ corresponden a las realizaciones de los vectores aleatorios $Z_1, \ldots, Z_n$. Es conveniente suponer que se distribuyen de acuerdo a (3.7). Esta suposición significa que la distribución de los datos completos del vector X_c implica la apropiada para el vector de datos incompletos x. Entonces la log-verosimiltud del conjunto de los datos completos de $\dot{\Theta}$, está dada por

$$\mathscr{L}_c\left(\dot{\Theta}\right) = \sum_{j=1}^{k} \sum_{i=1}^{n} z_{ij}\{\log \pi_j + \log f_j\left(x_i; \theta_j\right)\}. \tag{4.16}$$

La adición del conjunto de datos no observados al problema es manejado por el Paso-E.

4.4.1. Paso-E

El Paso-E requiere el cálculo de la esperanza condicional de $\mathscr{L}_c(\dot{\Theta})$ dado x, usando $\dot{\Theta}^{(0)}$ para $\dot{\Theta}$, es decir

$$Q\left(\dot{\Theta}; \dot{\Theta}^{(0)}\right) = \mathbb{E}_{\dot{\Theta}^{(0)}}\left\{\mathscr{L}_c\left(\dot{\Theta}\right) | \, x\right\}. \tag{4.17}$$

El operador esperanza tiene el subíndice $\dot{\Theta}^{(0)}$ para dejar en claro que esta esperenza se efectuará utilizando $\dot{\Theta}^{(0)}$ para $\dot{\Theta}$. De ello se deduce que en la iteración $m+1$, el Paso-E requiere el cálculo de $Q\left(\dot{\Theta}; \dot{\Theta}^{(m)}\right)$, donde $\dot{\Theta}^{(m)}$ es el valor de $\dot{\Theta}$ después de la m-ésima iteración. Como la log-verosimilitud de los datos completos es lineal en los datos no observados, el Paso-E en la iteración $m+1$ requiere el cálculo de la actual esperanza condicional de Z_{ij} dados los datos observados x, donde Z_{ij} es la variable aleatoria correspondiente a z_{ij}. Entonces

$$\mathbb{E}_{\dot{\Theta}^{(m)}} = \mathbb{P}_{\dot{\Theta}^{(m)}}\left\{Z_{ij} = 1 | \, x\right\} = \tau_j\left(x_i; \dot{\Theta}^{(m)}\right), \tag{4.18}$$

donde, por (4.10),

$$\tau_j\left(x_i; \dot{\Theta}^{(m)}\right) = \frac{\pi_j^{(m)} f_j(x_i; \theta_j^{(m)})}{f\left(x_i; \theta_j^{(m)}\right)} = \frac{\pi_j^{(m)} f_j(x_i; \theta_j^{(m)})}{\sum_{h=1}^{k} \pi_h^{(m)} f_h\left(x_i; \theta_h^{(m)}\right)}, \tag{4.19}$$

para $i = 1, \ldots, n$; $j = 1 \ldots, k$. La cantidad $\tau_j\left(x_i; \dot{\Theta}^{(m)}\right)$ es la probabilidad posterior del i-ésimo miembro de la muestra con valores perteneciente al j-ésimo componente de la mezcla. Usando (4.18), se puede asumir que la esperenza condicional de (4.16) dado x es igual a

$$Q\left(\dot{\Theta}; \dot{\Theta}^{(m)}\right) = \sum_{j=1}^{k} \sum_{i=1}^{n} \tau_j\left(x_i; \dot{\Theta}^{(m)}\right) \{\log \pi_j + \log f_j\left(x_i; \theta_j\right)\}. \tag{4.20}$$

4.4.2. Paso-M

El Paso-M en la iteración $m+1$ requiere de la maximización global de $Q\left(\dot{\Theta}; \dot{\Theta}^{(k)}\right)$ con respecto a $\dot{\Theta}$ sobre el espacio parametral Ω. Para el modelo de mezclas finitas, las estimaciones actualizadas de los pesos de la mezcla son calculados independientemente de la estimación actualizada del vector de parámetros $\dot{\theta}$. Si los z_{ij} son observados, entonces el MLE del conjunto de datos completos de π_j estará dado por

$$\hat{\pi}_j = \frac{1}{n} \sum_{i=1}^{n} z_{ij} \qquad j = 1, \ldots, k. \tag{4.21}$$

Como el Paso-E simplemente involucra el reemplazar cada z_{ij} con su actual esperanza condicional $\tau_j\left(x_i; \dot{\Theta}^{(m)}\right)$ en la log-verosimilitud de los datos completos, el estimador actual de π_j esta dado por

$$\pi_j^{(m+1)} = \frac{1}{n} \sum_{i=1}^{n} \tau_j\left(x_i; \dot{\Theta}^{(m)}\right) \qquad j = 1, \ldots, k. \tag{4.22}$$

Así, en la estimación de π_j de la iteración $m+1$, existe una contribución de cada observación x_i igual a la probabilidad posterior del j-ésimo componente del modelo de mezclas finitas. En cuanto a la actualización de $\dot{\theta}$ en el Paso-M de la iteración $m+1$, se puede ver que a partir de $\dot{\theta}^{(m+1)}$ se obtiene como una apropiada raíz de

$$\sum_{j=I}^{k} \sum_{i=1}^{n} \tau_j\left(x_i; \dot{\Theta}^{(m)}\right) \frac{\partial}{\partial \dot{\theta}} \log f_j\left(x_i; \theta_j\right) = 0. \tag{4.23}$$

Una característica del algoritmo EM es que la solución de (4.23) a menudo existe en forma cerrada. El Paso-E y el Paso-M se alternan repetidamente hasta que la diferencia

$$L\left(\dot{\Theta}^{(m+1)}\right) - L\left(\dot{\Theta}^{(m)}\right)$$

es una cantidad arbitrariamente pequeña.

Ejemplo 2. *Mezcla de dos distribuciones normales homocedásticas.*

Si considermos la mezcla

$$f_M\left(x_i\right) = \pi\,\phi\left(x_i|\,\mu_1,\sigma^2\right) + \left(1-\pi\right)\phi\left(x_i|\,\mu_2,\sigma^2\right), \tag{4.24}$$

donde

$$\phi\left(x_i|\,\mu_j,\sigma^2\right) = \frac{1}{\sqrt{2\,\pi\,\sigma^2}}\,\exp\left\{\frac{1}{2\,\sigma^2}\left(x_i-\mu_j\right)^2\right\}$$

denota a la función de densidad normal univariada con media μ_j y varianza σ^2 para $j=1,2$. La log-verosimilitud de los datos completos esta dada por

$$\mathscr{L}_c\left(\dot{\Theta}\right) = \sum_{i=1}^{n}\left(1-z_i\right)\log\,\phi\left(x_i|\,\mu_1,\sigma^2\right) + z_i\,\phi(x_i|\,\mu_2,\sigma^2). \tag{4.25}$$

Al derivar (4.25) con respecto a $\dot{\Theta}$ obtenemos los estimadores máximo verisimiles para π,μ_1,μ_2 y σ^2:

$$\hat{\pi} = \frac{1}{n}\sum_{i=1}^{n} z_i, \tag{4.26}$$

$$\hat{\mu}_1 = \frac{\sum_{i=1}^{n} z_i\,x_i}{\sum_{i=1}^{n} z_i}, \tag{4.27}$$

$$\hat{\mu}_2 = \frac{\sum_{i=1}\left(1-z_i\right)x_i}{n-\sum_{i=1}^{n} z_i}, \tag{4.28}$$

$$\hat{\sigma}^2 = \frac{1}{n}\left[\sum_{i=1}^{n} z_i\left(x_i-\mu_1\right)^2 + \sum_{i=1}^{n}\left(1-z_i\right)\left(x_i-\mu_2\right)^2\right]. \tag{4.29}$$

Como se explico anteriormente, el Paso-E involucra el cálculo de

$$\mathbb{E}_{\dot{\Theta}^{(m)}}\left[\mathscr{L}_c\,|x_i\right] = \frac{\pi_j^{(m)}\,f_j\left(x_i;\theta_j^{(m)}\right)}{f\left(x_i;\theta_j^{(m)}\right)} = \frac{\pi_j^{(m)}\,f_j\left(x_i;\theta_j^{(m)}\right)}{\sum_{h=1}^{k}\pi_h^{(m)}\,f_h\left(x_i;\theta_h^{(m)}\right)} = \tau_i^{(m)}.$$

Y el Paso-M en la iteración $m+1$ requiere de la maximización global de

$$Q\left(\dot{\Theta};\dot{\Theta}^{(m)}\right) = \sum_{j=1}^{k}\sum_{i=1}^{n}\tau_j\left(x_i;\dot{\Theta}^{(m)}\right)\{\log\pi_j + \log f_j\left(x_i;\theta_j\right)\}.$$

con respecto a $\dot{\Theta}$ sobre el espacio parametral Ω. Es decir, en este paso se va a sustituir $\tau_i^{(m)}$ por z_i en (4.26) y hasta (4.29), entonces

$$\pi_j^{(m+1)} = \frac{1}{n}\sum_{i=1}^{n} \tau_j\left(x_i; \dot{\Theta}^{(m)}\right),$$

$$\mu_1^{(m+1)} = \frac{\sum_{i=1}^{n} \tau_j\left(x_i; \dot{\Theta}^{(m)}\right) x_i}{\sum_{i=1}^{n} \tau_j\left(x_i; \dot{\Theta}^{(m)}\right)},$$

$$\mu_2^{(m+1)} = \frac{\sum_{i=1} \left(1 - \tau_j\left(x_i; \dot{\Theta}^{(m)}\right)\right) x_i}{n - \sum_{i=1}^{n} \tau_j\left(x_i; \dot{\Theta}^{(m)}\right)},$$

$$\sigma^{2^{(m+1)}} = \frac{1}{n}\left[\sum_{i=1}^{n} \tau_j\left(x_i; \dot{\Theta}^{(m)}\right)(x_i - \mu_1)^2 + \sum_{i=1}^{n}\left(1 - \tau_j\left(x_i; \dot{\Theta}^{(m)}\right)\right)(x_i - \mu_2)^2\right],$$

para $j = 1, 2$.

Entonces se simulo la mezcla (4.24) con los siguientes parámetros:

$$\dot{\Theta} = (1/2, 2, 1, 1).$$

A partir de los resultados anteriores se estimo $\dot{\Theta}$ a través de algoritmo EM. Los resultados se muestran en la tabla 4.1.

Iteración	π	μ_1	μ_2	σ^2
10	0.4950106	1.705152	1.226815	0.6104789
25	0.5225931	1.692305	1.213243	0.6451429
45	0.5287654	1.689444	1.210178	0.6529800
60	0.5294312	1.689136	1.209846	0.6538489
100	0.5295984	1.689059	1.209763	0.6540402
150	0.5296008	1.689057	1.209762	0.6540429

Cuadro 4.1: Estimadores basados en el algoritmo EM.

En esta tabla se puede observar que los valores obtenidos difieren poco de los valores reales, sin embargo, la varianza sigue muy distante del valor real.

4.4.3. Valores iniciales para el algoritmo EM

Por lo general, en la práctica se tiene que especificar un valor para $\dot{\Theta}^{(0)}$. Además si se elige incorrectamente a $\dot{\Theta}^{(0)}$, la convergencia del algoritmo EM podría ser demasiado lenta. De hecho, en algunos casos en donde la verosimilitud se encuentra en el límite del espacio muestral, la secuencia de estimaciones $\{\dot{\Theta}^{(m)}\}$ generadas por el algoritmo EM divergen si $\dot{\Theta}^{(0)}$ se elige demasiado cerca del límite. Otro problema que se presenta en los modelos de mezclas finitas es que la ecuación de verosimilitud tenga raíces múltiples que corresponden a distintos máximos locales. Por lo que el algoritmo EM deberá aplicarse a partir de una amplia gama de valores, que son el resultado de la búsqueda de todos los máximos locales.

Un enfoque alternativo es el de realizar el primer Paso-E especificando un valor $\tau_i^{(0)}$ para $\tau(x_i; \dot{\Theta})$ para $i = 1, \ldots, n$, en donde

$$\tau\left(x_i; \dot{\Theta}\right) = \left(\tau_1\left(x_i; \dot{\Theta}\right), \ldots, \tau_k\left(x_i; \dot{\Theta}\right)\right)^T$$

es el vector que contiene las k probabilidades posteriores de los componentes pertenecientes a la mezcla para x_i. Esto último suele ser llevado a cabo al establecer $\tau_i^{(0)} = z_i^{(0)}$ para $i = 1, \ldots, n$, donde

$$z^{(0)} = (z_1^{(0)}, \ldots, z_n^{(0)})^T$$

define una partición inicial de los datos en k grupos.

Para datos de dimensión mayor, un valor inicial $z^{(0)}$ se puede estimar a través de la utilización de algunos algoritmos de agrupación (clustering), tales como el algoritmo de k-medias o un procedimiento jerárquico, si n no es demasiado grande.

4.4.4. Comenzando con valores aleatorios

Otra forma de especificar un valor inicial de los datos es dividirlos de manera aleatoria en k grupos correspondientes a los k componentes del modelo de mezclas finitas. Es decir, para cada observación x_i se genera aleatoriamente un número entero entre 1 y k. Si este entero aleatorio es igual a h, entonces nos fijamos si el i-ésimo elemento de $z_i^{(0)}$ es igual a uno para $j = h$ e igual a cero para $j \neq h,\ j = 1, \ldots, k$.

Con inicios aleatorios, el efecto del teorema del límite central permite tener a los parámetros inicialmente similares, al menos en muestras grandes. Una forma de reducir este efecto es primero seleccionando una pequeña

submuestra aleatoria de los datos, para que después sea asignada aleatoriamente a los k componentes. Entonces el primer Paso-M se realiza en base de la submuestra. La submuestra tiene que ser lo suficientemente grande como para asegurar que el primer Paso-M sea capaz de producir una estimación no degenerada del vector de parámetros. Por ejemplo, en la estimación de una mezcla de componentes normales de dimensión p con matrices de covarianza no restringidas, tiene que haber por lo menos $p+1$ observaciones asignadas a cada uno de los componentes. Esto con el fin de garantizar estimaciones no singulares de las matrices de varianzas y covarainzas de cada uno de los componentes del modelo.

Un método alternativo para especificar el inicio aleatorio, al menos en el contexto de k componentes normales con medias μ_j y matrices de varianzas y covarianzas Σ_j, es generar aleatoriamente e independientemente $\mu_j^{(0)}$ mediante

$$\mu_1^{(0)}, \ldots, \mu_k^{(0)} \overset{i.i.d.}{\sim} \mathrm{N}\left(\bar{x}, \mathbf{V}\right), \tag{4.30}$$

donde $\bar{x}$ es la media muestral y

$$\mathbf{V} = \frac{1}{n} \sum_{i=1}^{n} \left(x_i - \bar{x}\right)\left(x_i - \bar{x}\right)^T \tag{4.31}$$

es la matriz de varianzas y covarianzas de los datos observados. Con este método, existe más variación entre los valores iniciales $\mu_j^{(0)}$ para las medias de los componentes μ_j que con una partición aleatoria de los datos en k grupos. Además es menos exigente a la hora de hacer los cálculos. Las matrices de varianzas y covarianzas de los componentes Σ_j y los pesos π_j se pueden especificar de la siguiente manera

$$\Sigma_j^{(0)} = \mathbf{V} \quad \text{y} \quad \pi_j^{(0)} = \frac{1}{k}, \qquad j = i, \ldots, k. \tag{4.32}$$

Como se ilustra en McLachlan y Basford (1988), un factor clave en la estimación de un modelo de mezclas finitas es la precisión de la estimación del vector de proporciones de la mezcla. Para mezclas univariadas Fowlkes (1979) sugirió determinar el punto de inflexión en una gráfica $Q-Q$ para estimar la proporción de las poblaciones. El resto de los parámetros pueden ser estimados a partir de la muestra particionada en grupos de acuerdo con la estimación de las proporciones de la mezcla.

4.4.5. Tasa de convergencia del algoritmo EM

El algoritmo EM define un mapeo $\dot{\Theta} \rightarrow M\left(\dot{\Theta}\right)$, del espacio parametral de $\dot{\Theta}$ de tal forma que cada iteración $\dot{\Theta}^{(m)} \rightarrow \dot{\Theta}^{(m+1)}$ esta definida por

$$\dot{\Theta}^{(m+1)} = M\left(\dot{\Theta}^{(m)}\right), \qquad m = 0, 1, 2, \ldots.$$

Si $\dot{\Theta}^{(m)}$ converge a algún punto $\dot{\Theta}^*$ y $M(\dot{\Theta})$ es continua, entonces $\dot{\Theta}^*$ es un punto fijo del algoritmo, es decir, $\dot{\Theta}^*$ debe satisfacer

$$\dot{\Theta}^* = M(\dot{\Theta}^*). \tag{4.33}$$

A partir de una expansión en series de Taylor de $\dot{\Theta}^{(m+1)}$ sobre el punto $\dot{\Theta}^{(m)} = \dot{\Theta}^*$, tenemos una aproximación de $\dot{\Theta}^*$ de modo que

$$\dot{\Theta}^{(m+1)} - \dot{\Theta}^* \approx J(\dot{\Theta}^*)(\dot{\Theta}^{(m)} - \dot{\Theta}^*), \tag{4.34}$$

donde $J(\dot{\Theta})$ es la matriz Jacobiana de dimensión $d \times d$ para $M(\dot{\Theta}) = (M_1(\dot{\Theta}), \ldots, M_d(\dot{\Theta}))^T$, cuyo elemento perteneciente al renglón r y a la columna s será igual a

$$J_{rs}(\dot{\Theta}) = \frac{\partial M_r(\dot{\Theta})}{\partial \dot{\Theta}_s},$$

en donde $\dot{\Theta}_s = (\dot{\Theta})_s$. Entonces, en una aproximación de $\dot{\Theta}^*$, el algoritmo EM es esencialmente una iteración lineal, con una matriz tasa $J(\dot{\Theta}^*)$. Por está razón, $J(\dot{\Theta}^*)$ es a menudo citada como la matriz de la tasa de convergencia o, simplemente, la tasa de convergencia.

Para un vector $\dot{\Theta}$, una medida de la tasa real observada de concergencia es la tasa global de convergencia, la cual se define como

$$r = \lim_{k \to \infty} \frac{||\dot{\Theta}^{(m+1)} - \dot{\Theta}^*||}{||\dot{\Theta}^{(m)} - \dot{\Theta}^*||},$$

donde $||\cdot||$ es cualquier norma perteneciente a un espacio d-dimensional Euclideano. Es bien sabido que, en determinadas condiciones de regularidad,

$$r = \lambda_{\text{máx}} \equiv \text{el mayor de los eigenvalores de } J(\dot{\Theta}^*).$$

En la práctica, r comunmente se determina mediante

$$r = \lim_{k \to \infty} \frac{||\dot{\Theta}^{(m+1)} - \dot{\Theta}^{(m)}||}{||\dot{\Theta}^{(m)} - \dot{\Theta}^{(m-1)}||}. \tag{4.35}$$

4.4.6. Matriz de convergencia en términos de matrices de información

Supongamos que $\{\dot{\Theta}^{(m)}\}$ es una secuencia del algoritmo EM, para la cual

$$\frac{\partial Q(\dot{\Theta}; \dot{\Theta}^{(m)})}{\partial \dot{\Theta}} = 0 \tag{4.36}$$

se satisface para $\dot{\Theta} = \dot{\Theta}^{(m+1)}$. Dempster (1977) demostro que si $\dot{\Theta}^{(m)}$ converge a un punto $\dot{\Theta}^*$, entonces

$$J(\dot{\Theta}^*) = \mathcal{I}_c^{-1}(\dot{\Theta}^*; x)\mathcal{I}_m(\dot{\Theta}^*; x), \tag{4.37}$$

donde $\mathcal{I}_c^{-1}(\dot{\Theta}^*; x)$ es la esperanza condicional de la matriz de los datos completos definida por

$$\mathcal{I}_c = \mathbb{E}_{\dot{\Theta}}\{\mathbf{I}_c(\dot{\Theta}; \mathbf{X_c})|\, x\}, \tag{4.38}$$

donde

$$\mathbf{I}_c(\dot{\Theta}; \mathbf{X_c}) = -\frac{\partial^2 \log L_c(\dot{\Theta})}{\partial \dot{\Theta}\, \partial \dot{\Theta}^T}$$

y

$$\mathcal{I}_m(\dot{\Theta}; x) = -\mathbb{E}_{\dot{\Theta}}\{\partial^2 \log k(\mathbf{X_c}|\, x; \dot{\Theta})/\partial \dot{\Theta}\, \partial \dot{\Theta}^T|\, x\} \tag{4.39}$$

es la esperanza de la matriz de información para $\dot{\Theta}$ basada en x_c.

Por lo tanto la tasa de convergencia del algoritmo EM estará dada por el mayor de los eigenvalores de la matriz de información $\mathcal{I}_c^{-1}(\dot{\Theta}^*; x)\mathcal{I}_m(\dot{\Theta}^*; x)$. Es decir, mide la proporción de información sobre $\dot{\Theta}$ que no esta disponible por no considerar el conjunto de datos perdidos. Cuanto mayor es la proporción de información que falta, más lento es el índice de convergencia.

La tasa de convergencia del algoritmo EM se puede expresar, de manera equivalente, en términos del eigenvalor más pequeño de

$$\mathcal{I}_c^{-1}(\dot{\Theta}^*; x)\mathcal{I}_m(\dot{\Theta}^*; x).$$

Esto es porque podemos expresar $J(\dot{\Theta}^*)$ de la siguiente forma

$$J(\dot{\Theta}^*) = \mathbf{I}_d - \mathcal{I}_c^{-1}(\dot{\Theta}^*; x)\mathcal{I}_m(\dot{\Theta}^*; x), \tag{4.40}$$

donde $\mathbf{I}_d$ denota a la matriz identidad de dimensión $d \times d$.

Por último cabe señalar que, Windham y Cutler (1992) basaron una prueba sobre el número de componentes en un modelo de mezclas finitas en el eigenvalor más pequeño de $\mathcal{I}_c^{-1}(\hat{\Theta}^*; x)\mathcal{I}_m(\hat{\Theta}^*; x)$. Su motivación es que, heurísticamente, un gran valor de este pequeño eigenvalor sugiere un buen agrupamiento de los datos, mientras que un pequeño valor sugiere lo contrario.

4.5. Algoritmo EM incremental (IEM)

El algoritmo EM Incremental fue propuesto por Hinton y Neal (1998) para mejorar la tasa de convergencia del algoritmo EM. En el algoritmo IEM, se llava a cabo un Paso-E parcial antes de que se lleve a cabo el siguiente Paso-M. Es decir, supongamos que los datos observados $x_1, \ldots, x_n$ se encuentrán divididos en B bloques. Si r es igual a la parte entera de n/B, entonces cada bloque contiene r observaciones, aparte de, por ejemplo, el

B-ésimo bloque que tendrá más de r cuando n no es un múltiplo de B. El algoritmo IEM procede mediante la aplicación del Paso-E en un sólo bloque de observaciones antes de realizar el Paso-M. De esta manera, cada punto de los datos x_j es visitado después de B parciales Pasos-E y B Pasos-M se han llevado a cabo.

Entonces, si $\dot{\Theta}^{(m)}$ denota el valor de $\dot{\Theta}$ después de la m-ésima actualización y si $\dot{\Theta}^{(m+b/B)}$ denota el valor de $\dot{\Theta}$ después de la b-ésima iteración en la actualización $(k+1)$. En el contexto de un modelo de mezclas finitas con k componentes, el algoritmo IEM es implementado en la itreación $b+1$ de la actualización $(k+1)$ de la siguiente manera:

Paso-E: Para $j = 1, \ldots, k$, reemplazar la variable indicadora z_{ij} en la log-verosimilitud de los datos completos por $\tau_j\left(x_i; \dot{\Theta}^{(m+b/B)}\right)$ para las x_i en el bloque $b+1$, $b = 0, \ldots, B-1$

El uso de un Paso-E parcial plantea dos puntos que no se encuentran en el algoritmo EM estándar. Uno de ellos es la forma de evaluar la convergencia cuando se utiliza el algoritmo IEM con la aplicación del Paso-E a través de bloques de datos. Después de la iteración $b+1$ en la actualización $m+1$, la log-verosimilitud se puede aproximar mediante

$$\log L\left(\dot{\Theta}^{(m+(b+1)/B)}\right) \approx \log L\left(\dot{\Theta}^{(m+b/B)}\right) + \sum_{i \in S_b} \left\{\log f\left(x_i; \dot{\Theta}^{(m+(b+1)/B)}\right) - \log f\left(\dot{\Theta}^{(m+b/B)}\right)\right\}. \quad (4.41)$$

El segundo punto está relacionado con la actualización inicial a través de los datos. Los primeros bloques pueden contener algunas observaciones con una alta probabilidad de pertenencia a algún componente de la mezcla. Esto que puede dar lugar a que la estimación del componente de la mezcla sea insignificante. Como consecuencia, las observaciones en los bloques subsecuentes pueden tener prácticamente una estimación nula en sus probabilidades posteriores. Este problema puede evitarse mediante la ejecución del algoritmo EM estándar para las primeras exploraciones o por lo menos esperar hasta que el Paso-E se realize durante varios bloques antes de realizar el primer Paso-M. Por supuesto que esto debe considerarse como excesivamente conservador en la mayoría de las aplicaciones con bases de datos muy grandes.

En relación con el tiempo necesario para realizar el algoritmo IEM para una actualización, los B Pasos-E parciales se demoran más para ejecutarse que un Paso-E completo en el algoritmo EM estándar. La elección del número de bloques a fin de optimizar el tiempo de convergencia del algoritmo IEM es un problema interesante. McLachlan y Ng (2000) sugiere el usar $B \approx n^{2/5}$ como una simple guia, sim embargo, cuando las matrices de covarianza son diagonales McLachlan y Ng (2000a) sugieren la modificación de esta guía para $B \approx n^{1/3}$. La elección óptima dependerá del número de parámetros desconocidos.

Una característica a resaltar dentro del algoritmo IEM es que el tiempo de convergencia comienza a aumentar conforme aumenta el número de bloques. Esto se debe al tiempo adicional del cálculo necesario para realizar los M pasos adicionales y además tener que invertir las matrices de covarianza en la actualización de las probabilidades posteriores de los componentes en cada actualización del conjunto de datos. En particular, uno debe evitar tener que invertir la matriz de covarianzas después de cada actualización de la probabilidad posterior de una única observación. McLachlan y Ng (2000) han modificado estas fórmulas donde el peso de una sola observación no cambia de 1 a 0, es decir, se suprime pero no es más que actualizado a otro valor entre 0 y 1.

4.5.1. Actualización de los bloques para estadísticos suficientes

Si alguna de las distribuciones de los componentes pertenece a la familia exponencial, entonces es más fácil trabajar en términos de esperanzas condicionales que corresponden a estadísticos suficientes. Por lo tanto, para el bloque b y para el valor actual $\dot{\Theta}^{(m)}$ de $\dot{\Theta}$, si

$$\begin{aligned} T_{j\,1,b}^{(m)} &= \sum_{i \in S_b} \tau_j\,(x_i; \dot{\Theta}^{(m)}), \\ T_{j\,2,b}^{(m)} &= \sum_{i \in S_b} \tau_j\,(x_i; \dot{\Theta}^{(m)}) x_i, \\ T_{j\,3,b}^{(m)} &= \sum_{i \in S_b} \tau_j\,(x_i; \dot{\Theta}^{(m)}) x_i\, x_i^T. \end{aligned} \tag{4.42}$$

para $j = 1, \ldots, k$, donde $S_b \in \{1, \ldots, n\}$ contiene los subíndices de las x_i que pertenecen al b-ésimo bloque. La esperanza actual del estadístico suficiente podrá ser expresada sobre los B bloques en los términos de las ecuaciones (4.42) para obtener

$$T_{j\,q}^{(m)} = \sum_{b=1}^{B} T_{i\,q,b}^{(m)}, \qquad i = 1, \ldots, k;\ q = 1, 2, 3. \tag{4.43}$$

En las expresiones (4.42), el exponente m denota una iteración y no necesariamente alguna actualización del algoritmo IEM. Además, en el Paso-M de la iteración $b+1$ en la actualización $(k+1)$ del algoritmo IEM, las estimaciones de π_j, μ_j y Σ_j se actualizan de la siguiente manera:

$$\pi_j^{(m+(b+1)/B)} = T_{j\,1}^{(m+b/B)}/n, \tag{4.44}$$

$$\mu_j^{(m+(b+1)/B)} = T_{j\,2}^{(m+b/B)}/T_{j\,1}^{(m+b/B)}, \tag{4.45}$$

y

$$\sigma_j = \left\{ T_{j\,3}^{(m+b/B)} - T_{j\,1}^{(m+b/B)^{-1}} \, T_{j\,2}^{(m+b/B)} \, T_{j\,2}^{(m+b/B)^T} \right\} / T_{j\,1}^{(m+b/B)}, \tag{4.46}$$

para $j = 1, \ldots, k$. Las esperanzas condicionales $T_{i\,q}^{(m+b/B)}$ de los estadísticos suficientes del lado derecho de (4.44) a (4.46) pueden expresarse en términos de sus valores en la iteración anterior, usando el siguiente resultado

$$T_{j\,q}^{(m+b/B)} = T_{j\,q}^{(m+(b-1)/B)} - T_{j\,q,b+1}^{(m-1+b/B)} + T_{i\,q,b+1}^{(m+b/B)}, \qquad j = 1, \ldots, k, \; q = 1, 2, 3. \tag{4.47}$$

Es decir, en el Paso-E parcial y en la iteración $b+1$ de la actualización $m+1$, sólo los términos del bloque $b+1$ tienen que ser calculados, dado que el primer y segundo término del lado derecho de (4.47) están disponibles a partir de la iteración y actualización anterior, respectivamente.

4.5.2. Fórmulas eficientes de actualización

Supongamos, sin pérdida de generalidad, que el i-ésimo bloque consiste de la i-ésima observación x_i. Entonces para actualizar las probabilidades posteriores de

$$\tau_j \left(x_i; \dot{\Theta}^{(m+(i-1)/n)} \right) \quad \text{a} \quad \tau_j \left(x_i; \dot{\Theta}^{(m+i/n)} \right)$$

en la iteración $i+1$ de la actualización $m+1$ del algoritmo IEM, tenemos que actualizar los valores π_j, μ_j y Σ_j^{-1} para $j = 1, \ldots, k$. Por conveniencia, escribiremos $\tau_j \left(x_i; \dot{\Theta}^{(m+(j-1)/n)} \right)$, $\pi_j^{(m+(j-1)/n)}$, $\mu_j^{(m+(j-1)/n)}$ y $\Sigma_j^{(m+(j-1)/n)}$ como τ_{ij}, π_j, μ_j y Σ_j respectivamente. Las cantidades correspondientes a $\dot{\Theta}^{(m+(j-1)/n)}$ reemplazadas por $\dot{\Theta}^{(m+j)/n)}$ serán denotadas como τ_{ij}^*, π_j^*, μ_j^* y Σ_j^* respectivamente. McLachlan y Ng (2000) demostrarón que cuando $\dot{\Theta}^{(m+(j-1)/n)}$ es actualizado por $\dot{\Theta}^{(m+j)/n)}$ en $\tau_j(x_i; \dot{\Theta})$, entonces π_j, μ_j, Σ_j^{-1} y $|\Sigma_j|$ se pueden actualizar de la siguiente manera:

$$\pi_j^* = \frac{n\,\pi_j - \tau_{i\,j} + \tau_{i\,j}^*}{n}, \tag{4.48}$$

$$\mu_j^* = \mu_j - \frac{(\tau_{i\,j} - \tau_{i\,j}^*)\,(x_i - \mu_j)}{n\,\pi_j^*}, \tag{4.49}$$

$$\Sigma_j^{*-1} = \frac{\pi_j^*}{\pi_j} \left[\Sigma_j^{-1} + \frac{(\tau_{i\,j} - \tau_{i\,j}^*)\,\Sigma_j^{-1}\,(x_i - \mu_j)\,(x_i - \mu_j)^T\,\Sigma_j^{-1}}{n\,\pi_j^* - (\tau_{i\,j} - \tau_{i\,j}^*)\,(x_i - \mu_j)^T} \, \Sigma_j^{-1}(x_i - \mu_j) \right] \tag{4.50}$$

y

$$|\Sigma_j^*| = \left(\frac{\pi_j}{\pi_j^*} \right)^p \, |\Sigma_j| \left[1 - \frac{\tau_{i\,j} - \tau_{i\,j}^*}{n\,\pi_j^*} \, (x_i - \mu_j)^T \, \Sigma_j^{-1} \, (x_i - \mu_j) \right], \tag{4.51}$$

para $j = 1, \ldots, k$.

El uso de (4.48) a (4.51) reduce considerablemente la cantidad de tiempo del cálculo en la actualización de las k probabilidades posteriores de los componentes para x_i. En particular, el uso de las últimas dos expresiones evita tener que calcular directamente la inversa y los determinates de las matrices de covarianza. Lamentablemente, no es posible generalizar estas fórmulas para el caso de bloques que constan de más de una observación.

4.6. Algoritmo EM para cadenas de Markov

Al considerar el modelo (3.7), la función de densidad de los vectores $x_1, \ldots, x_n$ se asume como condicionalmente independiente de $z_1, \ldots, z_n$, es decir,

$$f\left(x_1, \ldots, x_n | \, z_1, \ldots, z_n; \xi\right) = \prod_{i=1}^{n} f\left(x_i | \, z_i; \xi\right),$$

donde

$$f\left(x_i | \, z_i; \xi\right) = \prod_{j=1}^{k} f_j\left(x_i; \theta_j\right)^{z_i \, j}$$

y ξ denota los parámetros desconocidos en $\theta_1, \ldots, \theta_k$.

El *Modelo Oculto de Markov, HMM,* debilita la hipótesis de independencia en X_i mediante la adopción de observaciones sucesivas que se correlacionan a través de su componente de origen. Con este enfoque, la hipótesis de independencia (3.7) en el vector de variables indicadoras z_i se debilita. Por lo general, un modelo Markoviano estacionario se formula para la distribución de los vectores ocultos $Z_1, \ldots, Z_n$. En una dimensión, este modelo Markoviano es una cadena de Markov y en dos o más dimensiones es un campo aleatorio de Markov. La distribución condicional del vector observado X_i se formula para que dependa sólamente del valor del componente de origen Z_i y que sea condicionalmente independiente. Las cadenas de Markov ocultas proporcionan un modelo realista cuando las observaciones x_i aparecen secuencialmente en el tiempo y tienden a agruparse o alternarse entre los posibles componentes (subpoblaciones). La estimación de los parámetros en los modelos ocultos de Markov por lo general se basan en la estimación máximo verosimil en métodos Bayesianos. Sin embargo, el método de momentos se dificulta al grado de ser intratable.

Si la dependencia entre los vectores de variables indicadoras Z_i se especifica por una cadena de Markov estacionaria con matriz de transición $\mathbf{A} = ((\pi_{h\,i}))$, $h, i = 1, \ldots, k$. Entonces

$$\mathbb{P}\left[Z_{j,i+1} = 1 | \, Z_{i\,h}\right] = \pi_{j\,h} \qquad j, h = 1, \ldots, k$$

para cada i, $i = 1, \ldots, n-1$. La distribución inicial de la cadena de Markov se define por π_{0j}. Por lo tanto, si β contiene las probabilidades iniciales π_{0j} y las probabilidades de transición π_{ij}, entonces podemos escribir la distribución de Z como

$$\mathbb{P}[z;\beta] = \mathbb{P}[z_1;\beta] \prod_{i=2}^{n} \mathbb{P}[z_i|\, z_{i-1};\beta],$$

donde

$$\mathbb{P}[z_1;\beta] = \prod_{j=1}^{k} \pi_{0\,j}^{z_{j\,1}}$$

y

$$\mathbb{P}[z_i|\, z_{i-1};\beta] = \prod_{h=1}^{k}\prod_{j=1}^{k} \pi_{h\,j}^{z_{h,i-1}z_{i\,j}}.$$

Si consideramos a X_i como una variable discreta, donde

$$f_j(x_i) = \mathbb{P}[X_i = x_i|\, Z_{i\,j} = 1], \qquad j = 1, \ldots, k;\ i = 1, \ldots, n,$$

es decir, es la probabilidad de que $X_i = x_i$ dado que es un miembro del j-ésimo componente de la cadena. El vector de parámetros desconocidos $\dot{\Theta}$ consistirá de β y de las probabilidades de los componentes para los distintos valores asumidos por la variable aleatoria X_i. Entonces, el vector de datos completos se conformará por los datos observados y los datos no observados, $z = (z_1^T, \ldots, z_n^T)^T$. Por lo tanto, la log-verosimilitud de los datos completos estará dada por

$$\mathscr{L}_C(\dot{\Theta}) = \log \mathbb{P}[z] + \log f(x|\, z) = \log \mathbb{P}[z] + \sum_{j=1}^{k}\sum_{i=1}^{n} \log f_j(x_i;\theta_j), \tag{4.52}$$

donde

$$\log \mathbb{P}[z] = \sum_{j=1}^{k} z_{j\,1} \log \pi_j + \sum_{h=1}^{k}\sum_{j=1}^{k}\sum_{i=1}^{n-1} z_{h,i}z_{j,i+1} \log \pi_{h\,j}.$$

4.6.1. Paso-E

El Paso-E requiere del cálculo de la esperenza condicional de (4.52) dado el dato observado x. Al tomar esta esperanza se tiene que en la iteración $m+1$

$$Q(\dot{\Theta}, \dot{\Theta}^{(m)}) = \sum_{j=1} \tau_{j\,1}^{(m)} \log \pi_{0\,j} + \sum_{h=1}^{k}\sum_{j=1}^{k}\sum_{i=1}^{n-1} \tau_{h\,i\,j}^{(m)} \log \pi_{h\,j} + \sum_{j=1}^{k}\sum_{i=1}^{n} \tau_{i\,j}^{(m)} \log f_j(x_i;\theta_j), \tag{4.53}$$

donde $\tau_{h\,i\,j}^{(m)}$ y $\tau_{i\,j}^{(m)}$ denotan los valores actuales de las probabilidades condicionales definidas como

$$\tau_{h\,i\,j} = \mathbb{P}[Z_{h\,i} = 1, Z_{j,i+1} = 1|\, x], \qquad i = 1, \ldots, n-1, \tag{4.54}$$

y

$$\tau_{i\,j} = \mathbb{P}\left[Z_{i\,j} = 1 | \, x\right], \qquad i = 1, \ldots, n.$$

De (4.54), tenemos que

$$\tau_{i\,j} = \sum_{h=1}^{k} \tau_{h\,j,i-1}, \qquad i = 2, \ldots, n$$

y

$$\tau_{j\,1} = \frac{\pi_{0\,j} f_j(x_1)}{\sum_{h=1}^{k} \pi_{0\,h} f_h\left(x_1\right)}.$$

Las probabilidades posteriores $\tau_{h\,i\,j}$ y $\tau_{i\,j}$ se pueden expresar en términos de las siguientes probabilidades:

$$a_{i\,j} = \mathbb{P}\left[X_1 = x_1, \ldots, X_i = x_i, Z_{i\,j} = 1\right] \qquad i = 1, \ldots, n,$$

y

$$b_{i\,j} = \mathbb{P}\left[X_{i+1} = x_{i+1}, \ldots, X_n = x_n | \, Z_{i\,j} = 1\right] \qquad i = n-1, n-2, \ldots, 1.$$

Rabiner (1989) se refiere a $a_{i\,j}$ como la probabilidad de retraso y a $b_{i\,j}$ como la probabilidad de adelanto. De ello se deduce que τ_{hij} puede expresarse mediante

$$\tau_{h\,i\,j} = \frac{a_{h\,i}\, \pi_{h\,i}\, f_j\left(x_{i+1}\right) b_{j,i+1}}{\sum_{h=1}^{k} \sum_{j=1}^{k} a_{h\,j}\, \pi_{hj}\, f_j\left(x_{i+1}\right) b_{j,i+1}}, \tag{4.55}$$

puesto que es el numerador es $\mathbb{P}\left[Z_{hi} = 1, Z_{j,i+1} = 1, Y = y\right]$ y el denominador es $\mathbb{P}[Y = y]$. Los valores de $a_{ij}^{(m)}$ se calculan por una recusión forward según se indica en la iteración $m+1$, es decir,

Inicio:

$$a_{j\,1}^{(m)} = \pi_{0\,j}^{(m)}\, f_j^{(m)}(x_1) \qquad j = 1, \ldots, k.$$

Inducción:

$$a_{j,i+1}^{(m)} = f_j^{(m)}\left(x_{i+1}\right) \sum_{h=1}^{k} a_{h\,i}^{(m)} \pi_{h\,j}^{(m)} \qquad i = 1, \ldots, n-1.$$

Final:

$$\mathbb{P}_{\dot{\Theta}^{(m)}}\left[X_1 = x_1, \ldots, X_n = x_n\right] = \sum_{j=1}^{k} a_{j\,n}^{(m)},$$

donde $\mathbb{P}_{\dot{\Theta}^{(m)}}[\cdot]$ denota a la probabilidad con respecto a $\dot{\Theta}$ reemplazada por $\dot{\Theta}^{(m)}$. Los valores para $b_{hi}^{(m)}$ se calculan por una recursión hacia atrás según se indica en la iteración $m+1$, es decir,

Inicio:

$$b_{h\,n}^{(m)} = 1 \qquad h = 1, \ldots, k.$$

Inducción:

$$b_{h\,i}^{(m)} = \sum_{j=1}^{k} \pi_{h\,j}^{(m)} f_j^{(m)}(x_{i+1})\, b_{j,i+1}^{(m)} \qquad i = n-1, \ldots, 1;\ h = 1, \ldots, k.$$

El cálculo final en el Paso-E consiste en conectar estos valores y los valores actuales de los parámetros en la ecuación (4.55) de la siguiente manera:

$$\tau_{h\,i\,j}^{(m)} = \frac{a_{h\,i}^{(m)} \pi_{h\,j}^{(m)} f_j^{(m)}(x_i+1)\, b_{j,i+1}^{(m)}}{\sum_{h=1}^{k} \sum_{j=1}^{k} a_{h\,i}^{k} \pi_{h\,j}^{(m)} f_j^{(m)}(x_{i+1})\, b_{j,i+1}^{(m)}} \qquad i = 1, \ldots, n-1.$$

4.6.2. Paso-M

El Paso-M consiste en encontrar las estimaciones actualizadas de los parámetros de la función (4.53). Se trata de una combinación de los estimadores máximo verosimiles para los parámetros de una distribución multinomial y las probabilidades de transición de una cadena de Markov. La actualización de los parámetros se calculan de la siguiente manera:

$$\pi_{0\,j}^{(m+1)} = \tau_{j\,1}^{(m)},$$

$$\pi_{h\,j}^{(m+1)} = \sum_{i=1}^{n-1} \tau_{h\,i\,j}^{(m)} \Big/ \sum_{i=1}^{(n-1)} \tau_{h\,i}^{(m)},$$

$$f_j^{(m+1)}(x_i) = \sum_{l=1}^{n-1} \tau_{i\,l}^{(m)} \delta(x_l - x_i) \Big/ \sum_{l=1}^{n-1} \tau_{j\,l}^{(m)},$$

donde $\delta(u-v)$ es uno si $u = v$ y cero en otro caso.

Recientemente, Dunmur y Titterington (1998) estudiaron la influencia de las condiciones iniciales en la estimación máximo verosimil para una cadena de Markov binaria y homogénea. Llegaron a la conclusión de que el MLE depende de las estimaciones iniciales. En algunos casos, el algoritmo no converge a una solución razonable; pero para un punto fijo el algoritmo anterior se utiliza en el Paso-E.

Sin embargo, como se explica en Leroux y Puterman (1992), el algoritmo hacia atrás-hacia adelante para el cálculo de $a_{i\,j}$ y de $b_{i\,j}$ es numéricamente inestable en muchas situaciones. Esto se debe a que $a_{i\,j}$ converge rápidamente a cero o diverge a infinito conforme i aumenta. En general, Leroux y Puterman (1992) sugieren determinar para cada i el valor de r para el cual $10^{-r}\sum_j a_{i\,j}$ se encuentre entre 0.1 y 1.0 y multiplicando $a_{i\,j}$ por 10^{-r}. Entonces $a_{j,i+1}$ es calculada. Un procedimiento similar se aplica para encontrar el valor de $b_{i\,j}$. De está manera podemos obtener el valor de $\tau_{i\,j}$ y el valor de $\tau_{h\,i\,j}$.

4.7. Métodos bayesianos

El uso del método Bayesiano ha sido limitado hasta la aparición del artículo publicado por Smith (1990). Este artículo se centra en el gran potencial del muestreo de Gibbs en una gran variedad de problemas estadísticos. En particular, se observo que casi todos los cálculos Bayesianos podrían estimarse a través del muestreo de Gibbs. Considerando un modelo con vector de parámetros desconocidos $\dot{\Theta} \in \bar{\Theta}$, el paradigma de la inferencia Bayesiana es muy fácil de describir. Si $L\,(\dot{\Theta})$ denota la verosimilitud de $\dot{\Theta}$ dada la muestra $x^{(n)} = (x_1, \ldots, x_n)$, el teorema de Bayes establece el mecanismo por el que las creencias acerca de la información a priori de $\dot{\Theta}$ sobre los datos observados $x^{(n)}$ se pueden expresar como una densidad $f(\dot{\Theta})$, que actualiza las creencias acerca de la información de $\dot{\Theta}$ sobre los datos observados $x^{(n)}$, denotada por $f(\dot{\Theta}|\, x^{(n)})$, es decir,

$$f\,(\dot{\Theta}|\, x^{(n)}) = \frac{L\,(\dot{\Theta})\, f(\dot{\Theta})}{\int_{\Theta} L(\dot{\Theta})\, f(\dot{\Theta})\, d\,\dot{\Theta}}. \tag{4.56}$$

La aplicación del paradigma Bayesiano para modelos de mezclas finitas no será en absoluto sencillo, a menos que $\dot{\Theta}$ consista solo de uno o dos parámetros desconocidos. En este último caso, las gráficas o las densidades marginales posteriores pueden ser fácilmente obtenidas a partir de (4.56) cuyos cálculos se fácilitan simplemente evaluando $f\,(\dot{\Theta}|\, x^{(n)})$ sobre un conjunto conveniente de puntos. Cuando $\dot{\Theta}$ consiste de tres o más parámetros desconocidos, nos encontramos con el problema general de llevar a cabo una eficiente integración numérica en varias dimensiones, con el fin de obtener el denominador de (4.56), así como la inferencia marginal en funciones de dimensión menor de $\dot{\Theta}$. No vamos a entrar aquí en detalles, ya que una vez que se encuentra una estrategia general de integración numérica, un modelo de mezclas finitas es un caso especial de un problema de inferencia Bayesiana multiparamétrica.

Una situación en la cual el progreso analítico es posible, ocurre cuando la información de la muestra sobre $\dot{\Theta}$ se puede pensar como la información a priori, tal como se especifica en $p\,(\dot{\Theta})$. Para fines prácticos, $p\,(\dot{\Theta}|\, x^{(n)})$ puede verse como una versión normalizada de $L\,(\dot{\Theta})$ de modo que la teoría asintótica Bayesiana está estrechamente

relacionada con el enfoque de máxima verosimilitud. Conforme a las condiciones de regularidad, la distribución posterior asintótica de $\dot{\Theta}$ es $N(\hat{\Theta}, \hat{\Sigma})$, donde $\hat{\Theta}$ es el estimador máximo verosimil y

$$(\hat{\Sigma}^{-1})_{ij} = - \left. \frac{\partial^2 L(\dot{\Theta})}{\partial \theta_i \, \partial \theta_j} \right|_{\dot{\Theta}=\hat{\Theta}}$$

es la matriz inversa estimada. De ello se desprende que algunos de los algoritmos discutidos en la Sección 4.3 son igualmente útiles para un enfoque Bayesiano.

Otra situación en la que el enfoque analítico es limitado surge solo cuando los pesos de la mezcla son desconocidos, por lo que se les asigna una distribución a priori Dirichlet, que es proporcional a

$$\prod_{j=1}^{k} \pi_j^{\alpha_j - 1}, \quad \alpha_j > 0, \quad j = 1, \ldots, k. \tag{4.57}$$

Si los datos consisten de n observaciones independientes $x_1, \ldots, x_n$ de la mezcla

$$f_M(x|\,\pi) = \pi_1 \, f_1(x) + \cdots + \pi_k \, f_k(x),$$

entonces

$$L(\dot{\Theta}) = \sum_{i=1}^{n} \left[\sum_{j=1}^{k} \pi_j \, f_j(x_i) \right] = \sum_{r_1 + \cdots + r_k = n} C(x^{(n)}; r_1, \ldots, r_k) \prod_{j=1}^{k} \pi_j^{r_j}, \tag{4.58}$$

donde $C(x^{(n)}; r_1, \ldots, r_k)$ es una función fácilmente idetificable de $\{f_j(x_i), i = 1, \ldots, n, \, j = 1, \ldots, k\}$. Dado que la densidad posterior de π es proporcional al producto de (4.57) y (4.58), se deduce que $f_M(\pi|\,x^{(n)})$ tiene la forma de una mezcla de k^n densidades de Dirichlet. Estas últimas corresponden a las k^n posibles densidades posteriores que se derivan de la expreción (4.57) asumiendo una identificación particular de las n observaciones individuales con las k posibles distribuciones. Si $J_1, \ldots, J_{k^n}$ denota las posibles identificaciones, entonces

$$f_M(\pi|\,x^{(n)}) = \sum_{s=1}^{k^n} f_M(\pi|\,x^{(n)}, J_s) \, f_M(J_s|\,x^{(n)}),$$

de modo que los pesos en la mezcla posterior reflejan la relativa credibilidad en la identificación de las observaciones con las fuentes.

Cuando solo se desconoce a π, los factores de Dirichlet $f_M(\pi|\,x^{(n)}, J_s)$ en $f_M(\pi|\,x^{(n)})$ son independientes de $x^{(n)}$. Como resultado de ello, el número de distintas densidades de Dirichlet en $f_M(\pi|\,x^{(n)})$ es mucho menor que k^n. Para el caso más simple ($k = 2$) el número de de términos en $f_M(\pi|\,x^{(n)})$ se puede reducir de 2^n a $(n+1)$, cada uno correspondiente a un componente beta.

4.8. Estimación por mínima distancia

4.8.1. Estimación por mínima distancia basada en funciones de distribución

Supongamos que $F(\cdot|\,\dot{\Theta})$ es la función de distribución de interés. Si

$$\delta(G, F)$$

es una medida de la distancia entre dos funciones de distribución F y G entonces un estimador de minima distancia para $\dot{\Theta}$ es el valor de $\dot{\Theta}$ que minimiza

$$\delta[F_n(\cdot), F(\cdot|\,\dot{\Theta})].$$

Esta distancia será denotada por $\delta(\dot{\Theta})$, generalmente $\dot{\Theta}$ será un punto estacionario de $\delta(\dot{\Theta})$ el cual satisfacerá

$$\mathbf{D}_{\hat{\Theta}}\,\delta(\hat{\Theta}) = \mathbf{0}, \tag{4.59}$$

Esto es, por supuesto, una reminiscencia de la solución de máxima verosimilitud y vamos a ser capaces de obtener la máxima verosimilitud como un caso especial del enfoque de la estimación por mínima distancia. Como en la Sección 4.3, la clave para obtener las propiedades asíntoticas de $\hat{\Theta}$ será a través de una expansión lineal de Taylor de (4.59) sobre el valor real $\dot{\Theta}_0$, sujeto a las habituales condiciones de regularidad. Aproximadamente, para $\hat{\Theta}$ cerca de $\dot{\Theta}_0$,

$$\mathbf{D}_{\dot{\Theta}_0}\,\delta(\dot{\Theta}_0) + \mathbf{D}^2_{\dot{\Theta}_0}\,\delta(\dot{\Theta}_0)(\hat{\Theta} - \dot{\Theta}_0) = 0. \tag{4.60}$$

A menudo esto dará lugar al siguiente resultado asíntotico

$$\hat{\Theta} \to N(\dot{\Theta}_0, \mathbf{V}(\dot{\Theta}_0)),$$

en distribución, cuando $n \to \infty$, donde

$$\mathbf{V}(\dot{\Theta}_0) = \left\{\mathbb{E}\left[\mathbf{D}^2_{\dot{\Theta}_0}\,\delta(\dot{\Theta}_0)\right]\right\}^{-1} \mathrm{Cov}\left[\mathbf{D}^2_{\dot{\Theta}_0}\delta(\dot{\Theta}_0)\right]\left\{\mathbb{E}\left[\mathbf{D}^2_{\dot{\Theta}_0}\,\delta(\dot{\Theta}_0)\right]\right\}^{-1}.$$

Esto último motiva a la invocación de los algoritmos de Newton-Raphson o el Método de puntaje para el cálculo de $\hat{\Theta}$ cuando los métodos numéricos son obligatorios. Aunque gran parte de nuestra discusión introductoria será en términos de distribuciones continuas univariadas sobre la línea real, las distribuciones discretas y multivariadas también pueden ser tratadas. En este último caso, F representa una medida de probabilidad discreta.

Es evidente que existe una amplia gama de estimadores que pueden ser obtenidos por este método, dependiendo de la elección de la medida de distancia $\delta(\cdot,\cdot)$. La palabra distancia se utilizarán a pesar de que $\delta(\cdot,\cdot)$ puede o no satisfacer las propiedades formales de una métrica. La desigualdad del triángulo no es importante en el contexto actual. No es de vital importancia que $\delta(\cdot,\cdot)$ sea simétrica en sus dos argumentos. Exigimos, sin embargo, que

$$\delta(G,F) \geq \delta(G,G), \qquad \text{para toda F, G,}$$

en donde la igualdad se da si y sólo si $F(x) = G(x)$. Usualmente $\delta(G,G) = 0$ para toda G. Bajo estas circunstacias, la consistencia del estimador está asegurada siempre que mantenga las condiciones de regularidad y

- La familia $F(\cdot|\dot{\Theta})$ sea identificable;
- $\{F_n(\cdot)\}$ es consistente respecto a la función de distribución real.

Esto último es generalmente cierto cuando $F_n(\cdot)$ es la función de distribución empírica. El Cuadro 4.2 muestra algunas de las medidas de distancia que se han sugerido.

Cabe destacar algunos puntos importantes sobre las metricas expuestas en el Cuadro 4.2 son las siguientes:

- Aunque algunas de las medidas son métricas, esto no es cierto para todos los casos, por ejemplo, las medidas de Kullback-Leibler (KL), Levy (L), Ji-cuadrada (C), Ji-cuadrada modificado (MC), y la norma L_2 promediada no son métricas.
- Algunas de las medidas pueden considerarse como casos especiales de las demás: por ejemplo, δ_C es un caso especial de δ_{WL}.
- $\delta_C(F,G) = \delta_{MC}(G,F)$
- $\delta_{KL}\{F_n(\cdot), F(\cdot \mid \dot{\Theta})\}$ es equivalente al criterio de máxima verosimilitud, en cuanto a la estimación de $\dot{\Theta}$ se refiera.
- En la medida ponderada, las funciones con pesos no negativos $w(\cdot)$ son usadas.
- Cuando se considera un espacio muestral discreto (p,q) es el conjunto de probabilidad asociado con las probabilidades (F,G). En estos casos, vamos a escribir, por ejemplo, $\delta(p,q)$. El vector r denotará la cantidad correspondiente para $F_n(\cdot)$; es decir el vector de frecuencias relativas.

	Espacio Muestral Continuo	Espacio Muestral Discreto
Norma L_2 con		
funciones de distribución		
$\delta_{LA}(F,G)$	$\int [F(x)-G(x)]^2\ dx$	$\sum_{i=1} \left(\sum_{j=1}^{i} p_j - \sum_{j=1}^{i} q_j\right)^2$
Norma L_2 con densidades		
$\delta_{LB}(F,G)$	$\int [f(x)-g(x)]^2\ dx$	$\sum_{i=1} (p_i - q_i)^2$
Norma L_2 ponderada		
$\delta_{WLA}(F,G)$	$\int [F(x)-G(x)]^2 w(x)\ dx$	$\sum_{i=1} \left[\sum_{j=1}^{i} p_j - \sum_{j=1}^{i} q_j\right]^2 w_i$
$\delta_{WLB}(F,G)$	$\int [f(x)-g(x)]^2 w(x)\ dx$	$\sum_{i=1} (p_i - q_i)^2 w_i$
Norma L_2 promediada		
$\delta_{ALA}(F,G)$	$\int [F(x)-G(x)]^2\ dF(x)$	$\sum_{i=1} \left[\sum_{j=1}^{i} (p_j - q_j)\right]^2 p_i$
$\delta_{ALB}(F,G)$	$\int [f(x)-g(x)]^2\ dF(x)$	$\sum_{i=1} (p_i - q_i)^2 p_i$
Ji-cuadrada		
$\delta_{C}(F,G)$	$\int \frac{[f(x)-g(x)]^2}{f(x)}\ dx$	$\sum_{i=1} (p_i - q_i)^2 / p_i$
$\delta_{MC}(F,G)$	$\int \frac{[f(x)-g(x)]^2}{g(x)}\ dx$	$\sum_{i=1} (p_i - q_i)^2 / q_i$
Norma superior		
$\delta_{S}(F,G)$	$\sup_x \mid F(x) - G(x) \mid$	$\sup_x \left\| \sum_{j=1}^{i} (p_j - q_j) \right\|$
Distancia de Wolfowitz		
$\delta_{W}(F,G)$	$\int \lvert F(x) - G(x) \rvert\ dx$	$\sum_i \left\lvert \sum_{j=1}^{i} (p_j - q_j) \right\rvert$
Distancia de Hellinger		
$\delta_{H}(F,G)$	$\int [\sqrt{f(x)} - \sqrt{g(x)}]^2\ dx$	$\sum_{i=1} \left(\sqrt{p_i} - \sqrt{q_i}\right)^2$
Distancia de Levy		
$\delta_{L}(F,G)$	$\inf_x \{\varepsilon : F(x-\varepsilon) - \varepsilon \leq G(x) \leq F(x+\varepsilon) + \varepsilon\}$	
Kullback-Leibler		
$\delta_{KL}(F,G)$	$\int \log[dF(x)/dG(x)] dF(x)$	$\sum_{i=1} \log(p_i/q_i)$

Cuadro 4.2: Tomado de D. M. Titterington, A. F. M. Smith y U. E. Makov (1985)

- En el caso discreto de las medidas tales como δ_{LA}, δ_{WLA} y δ_S, basadas en funciones de distribución hay un orden implícito de los puntos en el espacio muestral. En muchas aplicaciones, esto será bastante natural, sobre todo si los datos consisten de datos univariantes agrupados de una muestra en el espacio continuo. Si no existe este orden natural, puede que no tenga sentido utilizar estas medidas.

- Los problemas surgen con $\delta_{kl}(p, q)$ si, por ejemplo, $p_i \neq 0$ pero $q_i = 0$. Lo mismo es cierto para δ_C y δ_{MC}.

- Formalmente las versiones de densidades basadas en medidas de distancias en ejemplos con espacios muestrales continuos pueden ser no válidas debido a la falta de diferenciabilidad de $F_n(\cdot)$.

La elección de la métrica es fundamental para la estiamación de los parámetros del modelo. Considreando la estrecha relación con la estimación por máxima verosimilitud, δ_{KL} es claramente la opción preferida. Para problemas en donde es de interés el comportamiento en las colas de la distribución, δ_C y δ_{MC} son favorecidas. Desde un punto de vista asíntotico, tienden a tener las mismas características que δ_{KL} en la medida en que produce estimadores consistententes y asintoticamente normales. En términos prácticos, una muestra pequeña, el comportamiento del estimador y el grado de dificultad de los cálculos asociados, pueden ser las consideraciones más importantes en la elección de la medida. Dado que el comportamiento de una muestra pequeña variará de aplicación a aplicación, la viabilidad del cálculo es tal vez la cuestión dominante.

4.8.2. Estimación de los pesos de la mezcla basada en distancias cuadráticas

Algunas medidas de distancia llevan a criterios cuadráticos similares. Por ejemplo, $\delta_{WLB}(F_n, F)$ puede ser escrita de la siguiente forma

$$\delta_{WLB}(F_n, F) = \int \frac{[\, d\, F_n(x) - d\, F(x\, \dot{\Theta})\,]^2}{d\, H(x)}.$$

En la práctica el espacio muestral es finito, posiblemente como resultado de la agrupación. Esto por supuesto, antes de evaluar las frecuencias relativas y los pesos. El poder obtener resultados explícitos podría llevar a preferir el criterio ji-cuadrado modificado $\delta_{MC}(\mathbf{p}(\pi), \mathbf{r})$ sobre el criterio ji-cuadrado $\delta_C(\mathbf{p}(\pi), \mathbf{r})$, particularmente en vista de la equivalencia asintótica de los dos estimadores resultantes.

Ahora supongamos que el espacio muestral $\mathscr{X}$ es continuo. Entonces, para toda n,

$$\sup_x |F(x|\, \pi) - F_n(x)|$$

es alcanzado para uno de los puntos del conjunto de datos $x_1, \ldots, x_n$. Supongamos, sin pérdida de generalidad, que $x_1 \leq \cdots \leq x_n$. Ahora nuestro objetivo es minimizar π_0 sujeto a

$$\left|\sum_{j=1}^{k}\pi_j F_j(x_i) - \frac{i}{n}\right| \leq \pi_0, \quad \left|\sum_{j=1}^{k}\pi_j F_j(x_i) - \frac{i-1}{n}\right| \leq \pi_0,$$

donde $\pi_1 + \cdots + \pi_k = 1$, $\pi_0, \pi_1, \ldots, \pi_k \geq 0$ para $i = 1, \ldots, n$

Dado que las expresiones anteriores pueden ser escritas como desigualdades ordinarias, los pesos, $\pi_0, \pi_1, \ldots, \pi_k$, son lineales bajo una interpretación de programación lineal. Deely y Kruse (1968) generalizaron esto último al desarrollar un método para la estimación de una mezcla de distribuciones. El proposito es estimar solo los pesos de la mezcla sin ningún interés en las distribuciones de la mezcla y sin la necesidad de especificar el modelo paramétrico.

4.9. Estimadores basados en transformaciones

En esta sección, utilizaremos la medida de distancia entre las transformaciones de las funciones de distribución empírica y teórica. Supongamos que, para una variable auxiliar $t \in \mathscr{T}$,

$$G(t|\,\dot{\Theta}) = \mathbb{E}[g(t,X)] = \int g(t,x)\,dF(x|\,\dot{\Theta}),$$

siempre existe. Si el espacio muestral es discreto, podemos interpretar a la integral como una suma y si la variable x es multivariada, t tendrá que ser un vector. Si además $\mathbb{E}[g^2(t,X)]$ es finita para toda $t \in \mathscr{T}$, si $X_1, \ldots, X_n$ representa una muestra aleatoria de $F(\cdot|\,\dot{\Theta})$ y si definimos

$$\bar{g}_n(t) = \frac{1}{n}\sum_{i=1}^{n} g(t,X_i),$$

entonces, por la ley de los grandes números,

$$\bar{g}_n(t) \to G(t|\,\dot{\Theta}),$$

en un sentido propio cuando $n \to \infty$ para cada $t \in \mathscr{T}$. Puede incluso ser posible garantizar la convergencia uniforme. Una forma natural para la estimación de $\dot{\Theta}$, es la minimización de algunas medidas de distancia entre $\bar{g}_n(t)$ y $G(t|\,\dot{\Theta})$,

$$\delta[\,\bar{g}_n(t), G(t|\,\dot{\Theta})\,].$$

Si este criterio es denotado por $\delta(\dot{\Theta})$, entonces las ecuaciones (4.59) y (4.60) pueden ser utilizadas para obtener el estimador $\hat{\Theta}$. Los estimadores basados en la función de distribución corresponden al caso especial en donde

$g(t, X)$ es la función indicadora

$$g(t,x) = \begin{cases} 1, & \text{si } x < t, \\ 0, & \text{en otro caso.} \end{cases}$$

El rango de posibles elecciones para δ es grande, sin embargo restringimos nuestra atención a medidas de distancia cuadráticas. Por lo tanto nos centraremos en la norma ponderada L_2

$$\delta\left\{\bar{g}_n(\cdot), G(\cdot|\,\dot{\Theta})\right\} = \int_{\mathscr{T}} |G(t|\,\dot{\Theta}) - \bar{g}_n(t)|^2\, dW(t), \tag{4.61}$$

donde $W(\cdot)$ es una medida ponderada sobre $\mathscr{T}$. Con $\mathscr{X} = \mathscr{T} = \mathbb{R}$ y $g(t,x)$ definida por la función indicadora y por (4.61) se elige la métrica $\delta_{WLA}[\,F_n(\cdot), F(\cdot|\,\dot{\Theta})\,]$. Por ejemplo, para datos univariados, algunos candidatos obvios para $G(t|\,\dot{\Theta})$ son los siguientes:

- *Función característica* (tansformación de Fourier)(Paulson, Holcomb, y Leitch, 1975; Thornton y Paulson, 1977; Heathcote, 1977; Bryant y Paulson, 1983):

$$G_C(t|\,\dot{\Theta}) = \mathbb{E}[e^{itX}]; \quad \bar{g}_n(t) = \frac{1}{n}\sum_{r=1}^{n} \exp\{i\,t\,x_r\}, \quad t \in \mathbb{R}, \quad \text{donde } i = \sqrt{-1}.$$

- *Función generadora de momentos* (Transformada de Laplace)(Quandt y Ramsey, 1978):

$$G_M(t|\,\dot{\Theta}) = \mathbb{E}[e^{tX}]; \quad \bar{g}_n(t) = \frac{1}{n}\sum_{r=1}^{n} \exp\{t\,x_r\}.$$

  Donde G_M existe solamente para cierto rango de t.

- *Función generadora de probabilidades* (datos discretos):

$$G_P(t|\,\dot{\Theta}) = \mathbb{E}[t^X]; \quad \bar{g}_n(t) = \frac{1}{n}\sum_{r=1}^{n} t^{x_r}.$$

La elección de la función de pesos $W(\cdot)$ es totalmente abierta. Por lo general la práctica y el sentido común son fundamentales para la elección de esta función. Los estimadores resultantes serán funciones de $W(\cdot)$ y, en principio, es posible definir un óptimo que conduce, en cierto sentido, a estimadores con matriz de covarianza asintotica mínima. Pero la complicada dependencia en $W(\cdot)$ sugiere que esta solución ideal no es práctica.

4.10. Descomposición numérica de mezclas

El objetivo de esta sección será el de descomponer en un número de componentes unimodales la curva de la densidad. Generalmente se asume que los componentes son funciones de densidad simétricas. En la práctica a

menudo la forma que toman corresponde a densidades normales o Cauchy. El principio usado para generar los componentes es el minimizar

$$\delta\left[F_0\left(\cdot\right), F\left(\cdot|\,\dot{\Theta}\right)\right] = \int \left[f_0\left(x\right) - p\left(x|\,\dot{\Theta}\right)\right]^2 d\,W\left(x\right),$$

donde $f_0\left(\cdot\right)$ denota la curva de referencia. Entonces $F_0\left(\cdot\right)$ no es una función de distribución empírica como tal. La función de peso es generalmente tomada como un punto de medida cuyo soporte es finito, de modo que $\dot{\Theta}$ se elige para minimizar

$$\sum_{i=1}^{n} \left[f_0\left(x_i\right) - p\left(x_i|\,\dot{\Theta}\right)\right]^2, \tag{4.62}$$

para algunos $x_1, \ldots, x_n$. Cabe destacar que $x_1 \ldots, x_n$ ya no son puntos de referencia, sino que simplemente representan una cuadrícula de valores en los que se basa el análisis de mínimos cuadrados. Aunque la mayoria de los artículos sobre este método se basan en (4.62) con una cuadrícula de puntos igualmente espaciados en un espacio muestral univariado, es obvio que existe un margen más sutil para la red de los puntos elegidos, no de manera uniforme ponderada por medio de mínimos cuadrados o la ampliación de espacios muestrales multivariados. Esta última extensión se facilita por el uso de las funciones de densidad en (4.62) en comparación con el uso de funciones de distribución. Un enfoque alternativo sería el de suavisar la función de distribución empírica o el histograma. Esto nos daría como resultado que $f_0\left(\cdot\right)$ sea una densidad con una curva suave. Esto podría lograrse mediante el uso de la estiamción del kernel de la densidad o por métodos de Boneva, Kendall, y Stefanov (1971), sobre la base de splines, y Van Ryzin (1973).

La minimización de (4.62) con respecto a $\dot{\Theta}$ requerirá de métodos númericos, por lo que el uso de software es indispensable para este método. Los parámetros se pueden ajustar a fin de mejorar la adecuación, sobre la base del error cuadrado integrado. Como en las secciones 4.8.1 y 4.9, la minimización explícita es posible si una medida cuadrática de ajuste se utiliza y sólo los pesos de la mezcla son desconocidos.

Capítulo 5

Muestreo de Gibbs

El muestreo de Gibbs es fundamental en la construcción y estimación del modelo GH-ARCH estacionario. Entonces, en base a lo descrito en Casella y George (1992), la finalidad de este capítulo es dar un panorama general sobre esta técnica.

En primer instancia se definen los elementos básicos para la construcción de una cadena de Gibbs en el caso univariado. A partir de esta construcción se plantea el caso para dos y más variables. Posteriormente se establecen los mecanismos mediante los cuales se pueden conocer algunas propiedades de la densidad deseada, como el cálculo de su media y varianza. El estudio de la convergencia en el muestreo de Gibbs es de especial interés si se desea que la muestra obtenida converja a la distribución deseada. Es así que en la última sección se presentan algunas estrategias para la detección de tal convergencia en una secuencia de Gibbs.

5.1. Una interpretación heurística

El muestreo de Gibbs es una técnica que permite generar distribuciones marginales sin que se tenga que calcular directamente su densidad conjunta. Supongamos que tenemos una función de densidad conjunta $f(x, y_1, \ldots, y_p)$ en donde es de interés obtener las características de la función de densidad marginal

$$f(x) = \int \ldots \int f(x, y_1, \ldots, y_p)\, dy_1 \ \ldots \ dy_p \tag{5.1}$$

tales como la media o la varianza. Lo más natural es calcular $f(x)$ y a partir de este resultado, obtener las características deseadas. Sin embargo, muchas de las veces resulta muy difícil el calcular cada una de las integrales de la expresión (5.1), ya sea de manera analítica o númerica. En tales casos el muestreo de Gibbs

provee una alternativa para obtener $f(x)$. En lugar de calcular o aproximar directamente $f(x)$, nos permite generar una muestra $X_1, \ldots, X_m \sim f(x)$ sin que sea necesario conocer a $f(x)$. Simulando una muestra lo suficientemente grande, la media, la varianza, o muchas otras características de $f(x)$ pueden calcularse para el grado deseado de precisión. Es importante darse cuenta de que el resultado final de todos los calculos, aunque basados en simulaciones, son las características de la población que se desean. Por ejemplo, para calcular la media de $f(x)$, podríamos utilizar $\frac{1}{n}\sum_{i=1}^{n} X_i$ y el hecho de que

$$\lim_{n\to\infty} \frac{1}{n} \sum_{i=1}^{n} X_i = \int_{-\infty}^{\infty} x\, f(x)\, dx = \mathbb{E}[X]. \tag{5.2}$$

Así, tomando n lo suficientemente grande, una característica de la población, incluso la propia densidad, se puede obtener para cualquier grado de exactitud.

Para entender el funcionamiento del muestreo de Gibbs, primero exploraremos el caso de dos variables. Comenzaremos con un par de variables aleatorias (X, Y). El muestreo de Gibbs genera una muestra de $f(x)$ a través de las distribuciones condicionales $f(x|\, y)$ y $f(y|\, x)$. Esto se hace para generar una *Secuencia de Gibbs* de variables aleatorias

$$Y_0', X_0', Y_1', X_1', Y_2', X_2', \ldots Y_k', X_k'. \tag{5.3}$$

El valor inicial $Y_0' = y_0'$ se especifica, y el resto de la expresión (5.3) se obtiene iterativamente por la generación alternada de los valores

$$\begin{aligned} X_j' &\sim f(x|\, Y_j' = y_j') \\ Y_{j+1}' &\sim f(y|\, X_j' = x_j'). \end{aligned} \tag{5.4}$$

Entonces, nos referimos a (5.4) como el *Muestreo de Gibbs*. De hecho, es razonable que bajo estas condiciones, la distribución de X_k' converga a $f(x)$, la verdadera distribución marginal de X, cuando $k \to \infty$. Así, para k lo suficientemente grande, la observación final en (5.3) es efectivamente un punto de la muestra de $f(x)$. La convergencia, en distribución, de la secuencia de Gibbs puede ser explotada en una gran variedad de formas para obtener una aproximación de los valores de $f(x)$. Si k es lo suficientemente grande, entonces se obtiene una muestra i.i.d. de $f(x)$.

El promedio de las densidades condicionales $f(x|\, Y_k' = y_k')$ será una aproximación cerrada para $f(x)$ y se puede estimar $f(x)$ con

$$\hat{f}(x) = \frac{1}{m} \sum_{i=1}^{m} f(x|\, y_i), \tag{5.5}$$

donde $y_1, \cdots, y_m$ es la secuencia de los valores de cada realización de las obsevaciones finales de Y para cada secuencia de Gibbs. La teoría detras de la Ecuación (5.5) es que el valor esperado de la densidad condicional es igual a

$$\mathbb{E}\left[f(x|\, y)\right] = \int f\left(x|\, y\right) f\left(y\right) dy = f\left(x\right), \tag{5.6}$$

un cálculo similar al de la Ecuación (5.5), dado que $y_i, \cdots, y_m$ aproxima una muestra de $f\left(y\right)$. Además esta ilustra un aspecto importante sobre el uso del muestreo de Gibbs para evaluar las características de $f\left(x\right)$. Las cantidades $f\left(x|\, y_1\right), \cdots f\left(x|\, y_m\right)$, calculadas usando los valores simulados $y_1, \cdots y_m$, conllevan más información sobre $f\left(x\right)$ que $x_1 \cdots x_m$ por si sola y darán mejores estimaciones. Por ejemplo, un estimador de la media de $f\left(x\right)$ es

$$\frac{1}{m}\sum_{i=1}^{m} x_i,$$

pero un mejor estimador es

$$\frac{1}{m}\sum_{i=1}^{m} E\left[X|\, y_i\right].$$

5.2. El caso bivariado

Supongamos que, para dos variables aleatorias X e Y, necesitamos conocer las densidades condicionales $f_{X|\,Y}\left(x|\, y\right)$ y $f_{Y|\,X}\left(y|\, x\right)$. Podriamos determinar de manera directa la densidad marginal de X y entonces la densidad conjunta de X e Y, mediante el siguiente argumento. Por definición,

$$f_X\left(x\right) = \int f_{X\,Y}\left(x, y\right) dy,$$

donde $f_{X\,Y}\left(x, y\right)$ es la función de densidad conjunta, que es desconocida. Ahora, usando el hecho de que $f_{X\,Y}\left(x, y\right) = f_{X|\,Y}\left(x|\, y\right) f_Y\left(y\right)$, entonces,

$$f_X\left(x\right) = \int f_{X|\,Y}(x|\, y)\, f_Y\left(y\right) dy$$

y si de manera analoga se sustituye a $f_Y\left(y\right)$, entonces,

$$\begin{aligned} f_X\left(x\right) &= \int f_{X|\,Y}\left(x|\, y\right) \int f_{Y|\,X}\left(y|\, t\right) f_X\left(t\right) dt \\ &= \int \left[\int f_{X|\,Y}(x|\, y)\, f_{Y|\,X}\left(y|\, t\right)\right] f_X\left(t\right) dt \\ &= \int h\left(x, t\right) f_X\left(t\right) dt, \end{aligned} \tag{5.7}$$

donde $h(x,t) = \left[\int f_{X|Y}(x|y) f_{Y|X}(y|t)\right]$. La Ecuación (5.7) define un punto fijo para la solución de la integral, donde $f_X(x)$ es una solución. Además, la ecuación (5.7) es el límite para el sistema iterativo de Gibbs, ilustrando como el muestreo de densidades condicionales produce una distribución marginal.

Aunque la distribución conjunta de X e Y determina todas las distribuciones condicionales y marginales, no siempre se da el caso en el que un conjunto de distribuciones condicionales apropiadas puedan determinar una distribución marginal apropiada y, por lo tanto, una adecuada distribución conjunta.

5.3. Más de dos variables

Como el número de variables, por lo general, tiende a aumentar, la relación entre las distribuciones condicionales, marginales y las distribuciones conjuntas se vuelve más compleja. Esto significa que existen muchas maneras para crear una única solución, como en (5.7) y es posible utilizar diferentes tipos de distribuciones condicionales para calcular la distribución marginal de interés. Tales metodologías son parte de las técnicas generales de *Sustitución por Muestreo*. Aquí ilustramos simplemente dos versiones de esta técnica.

En el caso de dos variables, todos los algoritmos para el muestreo de sustitución son iguales. El caso de tres variables, es lo suficientemente complejo como para ilustrar las diferencias entre los distintos algoritmos, pero lo suficientemente simple para que nos permita describirlo en detalle. La generalización a los casos de más de tres variables es bastante sencillo.

Supongamos que queremos calcular la distribución marginal $f_X(x)$ en un problema que depende de las variables aleatorias X, Y y Z. Una solución única, como en (5.7), se puede derivar si consideramos el par (Y,Z) como una sola variable aleatoria, entonces

$$f_X(x) = \int \left[\int \int f_{X|YZ}(x|y,z)\, f_{YZ|X}(y,x|t)\, dy\, dz\right] f_X(t)\, dt, \tag{5.8}$$

es análogo a (5.7). Los ciclos entre $f_{X|YX}$ y $f_{YZ|X}$, una vez más, dan lugar a una secuencia de variables aleatorias que convergen en distribución a $f_X(x)$. Esta es la idea detrás del *Algoritmo del Aumento de Datos* de Tanner y de Wong (1987). Por un muestreo iterativo de $f_{X|YX}$ y $f_{YZ|X}$, podemos obtener sucesivamente una mejor aproximación a $f_X(x)$. En cambio, el muestreo de Gibbs muestrea iterativamente para $f_{X|YZ}$, $f_{Y|XZ}$ y $f_{Z|XY}$. Es decir, la j-ésima iteración estará dada por

$$\begin{aligned} X'_j &\sim f(x|\, Y'_j = y'_j, Z'_j = z'_j) \\ Y'_{j+1} &\sim f(y|\, X'_j x ='_j, Z'_j = z'_j) \\ Z'_{j+1} &\sim f(z|\, X'_j = x'_j, Y'_{j+1} = y'_{j+1}). \end{aligned} \tag{5.9}$$

Por lo que el sistema iterativo (5.9) producirá una secuencia de Gibbs

$$Y_0', Z_0', X_0', Y_1', Z_1', X_1', Y_2', Z_2', X_2', \ldots, \tag{5.10}$$

con la propiedad que, para k lo suficientemente grande, $X_k' = x_k'$es efectivamente un punto muestral de $f(x)$. Aunque no es inmediatamente evidente, la iteración en (5.9) también solucionará la ecuación (5.8). De hecho, una característica que define el muestreo de Gibbs es que siempre utiliza el conjunto completo de condicionales univariadas para definir la iteración. Besag (1974) establece que este conjunto es suficiente para determinar la distribución conjunta y cualquier otra marginal, por lo que puede ser utilizado para solucionar (5.8).

5.4. Detectando la convergencia

El muestreo de Gibbs genera una cadena de Markov de variables aleatorias, las cuales convergen a la distribución de interés, $f(x)$. Muchos de los enfoques para extraer información de la secuencia de Gibbs explotan está propiedad mediante la selección de k, tomando X_j' para $j \geq k$ como una muestra de $f(x)$. Entonces el problema se reduce en elegir el valor de k.

Una estrategia general para tal elección de k, es para controlar la convergencia de la secuencia de Gibbs. Por ejemplo, Gelfand y Smith (1990) y Gelfand, Hills, Racine-Poor, y Smith (1990) sugieren el monitoreo de las densides estimadas para m secuencias de Gibbs independientes y la elección de k para que sea el punto de partida para que las densidades sean iguales bajo la prueba de "felt-tip pen". Tanner (1991) sugiere la supervisión de una secuencia de pesos que miden la discrepancia entre la muestra y la distribución deseada. Desafortunadamente, tales enfoques de supervisión no son inexpugnables, como se muestra en Gelman y Rubin (1991). Una alternativa puede ser la de elegir k basandonos en consideraciones teóricas, como en Banfield y Raftery (1990).

Una alternativa natural para muestrear el k-ésimo valor para un número considerable de repeticiones independientes de una secuencia de Gibbs, es generar una secuencia de Gibbs y extraer todas la r-ésimas observaciones. Para r lo suficientemente grande, tendremos una aproximación de una muestra i.i.d. de $f(x)$. Una ventaja de este enfoque es que reduce la dependencia de los valores iniciales, pero una posible desventaja es que la secuencia de Gibbs puede permanecer en un pequeño subconjunto de la muestra por un largo espacio de tiempo. Para bases de datos grandes, en donde el problema a la hora de simular se vuelve complicado debido a las límitaciones computacionales, un enfoque que nos permite explotar la secuencia de Gibbs es utilizar todas las realizaciones de X_j' para $j \leq k$. Aunque los datos resultantes serán dependientes, seguirá siendo el caso en

el que la distribución empírica de X'_j converge a $f(x)$. Notemos, que desde este punto de vista se puede ver que la eficiencia de la toma de muestras de Gibbs está determinada por el ritmo de convergencia. Intuitivamente, este tipo de convergencia será mejor cuando X'_j se mueva rápidamente a través del espacio muestral, una característica que puede ser pensada como una mezcla.

Capítulo 6

Distribución hiperbólica generalizada

La distribución hiperbólica generalizada fue introducida por Barndorff-Nielsen (1977) para modelar distintos fenómenos considerando distribuciones más generales. Su uso es fundamental para la construcción de la probabilidad de transición en el modelo GH-ARCH estacionario. Entonces, con base en lo planteado en Mena y Walker (2007b), la finalidad de este capítulo es exhibir las principales características de esta distribución. Inicialmente se partirá de la distribución Gaussiana inversa generalizada, despúes, mediante una mezcla de medias y varianzas de distribuciones Normales (Barndorff-Nielsen, 1977), se definirá a la distribución GH. Además se mencionarán algunas características de esta distribución, como la forma en que se obtienen sus momentos y el comportamiento asintótico de sus colas.

6.1. Definiciones y conceptos básicos

Para definir la distribución hiperbólica generalizada, primero mencionaremos algunas características de la distribición *Gaussiana Inversa Generalizada, GIG*.

Definición 8. *Decimos que la variable aleatoria X tiene una distribución Gaussiana Inversa Generalizada si su función de densidad se escribe de la siguiente manera,*

$$GIG\left(x;\lambda,\delta,\gamma\right)=\frac{\left(\frac{\gamma}{\delta}\right)^{\lambda/2}}{2\,K_{\lambda}\left(\sqrt{\delta\,\gamma}\right)}\,x^{\lambda-1}\,\exp\left\{-\frac{1}{2}\left(\delta\,x^{-1}+\gamma\,x\right)\right\},\quad x>0, \tag{6.1}$$

donde $\lambda \in \mathbb{R}$, $(\delta, \gamma) \in \Theta_\lambda$,

$$\Theta_\lambda = \begin{cases} \delta \geq 0,\, \gamma > 0, & si\ \lambda > 0, \\ \delta > 0,\, \gamma > 0, & si\ \lambda = 0, \\ \delta > 0,\, \gamma \geq 0, & si\ \lambda < 0. \end{cases}$$

y

$$K_\lambda(x) = \frac{1}{2} \int_0^\infty u^{\lambda-1} \exp\left\{-\frac{1}{2} x\,(u^{-1} + u)\right\} du, \qquad x > 0$$

es la función de Bessel de tercer orden con índice λ.

Los casos en donde $\delta = 0$ y $\gamma = 0$ serán interpretados como casos límites. Usando la expansión asisntotica 1 que se encuentra en el Apéndice C podemos observar que si $\lambda > 0$ y $\delta \downarrow 0$, la densidad (6.1) se reduce a

$$\mathrm{GIG}(x; \lambda, 0, \gamma) = \frac{(\gamma/2)^\lambda}{\Gamma(\lambda)} x^{\lambda-1} \exp\{-(\gamma/2)x\} = \mathrm{Ga}\left(x; \lambda, \frac{\gamma}{2}\right),$$

que es la distribución Gamma. Análogamente, cuando $\lambda < 0$, $\gamma \downarrow 0$ y usando la expansión asintotica 2 que se encuentra en el Apéndice C,

$$\mathrm{GIG}(x; \lambda, \delta, 0) = \frac{(2/\delta)^\lambda}{\Gamma(-\lambda)} x^{\lambda-1} \exp\{-(\delta/2)x\} = \mathrm{Iga}\left(x; -\lambda, \frac{\delta}{2}\right),$$

caracteriza la distribución gama inversa.

Barndorff-Nielsen (1977) construyo la distribución hiperbólica generalizada como una mezcla de medias y varianzas de distribuciones Normales. Más precisamente, se dice que la variable aleatoria X tiene una distribución hiperbólica generalizada $\mathrm{GH}(x; \lambda, \alpha, \beta, \delta, \mu)$ si

$$X | Y = y \sim \mathrm{N}(\mu + \beta\, y, y),$$

donde Y es una variable aleatoria con distribución $\mathrm{GIG}(\lambda, \delta, \sqrt{\alpha^2 - \beta^2})$ y $\mathrm{N}(\mu + \beta\, y, y)$ denota una distribución normal con media $\mu_t = \mu + \beta\, y$ y varianza $\sigma_t^2 = y$. A partir de esto, se puede verificar que su función de densidad está dada por

$$\mathrm{GH}(x; \lambda, \alpha, \beta, \delta, \mu) = \int_0^\infty \mathrm{N}(\mu + \beta\, y, y)\, \mathrm{GIG}(\lambda, \delta, \sqrt{\alpha^2 - \beta^2})\, dy. \tag{6.2}$$

Con esto, la distribución t de Student se puede ver como un caso particular de la distribución GH, a saber,

$$\mathrm{St}(x; \mu, \beta^2, \upsilon) = \int_0^\infty \mathrm{N}(x; \mu, y^{-1})\, \mathrm{Ga}(y; \upsilon/2, \upsilon\beta^2/2)\, dy, \tag{6.3}$$

donde $\mathrm{St}(\mu, \sigma^2, \upsilon)$ denota una distribución t de Student no centrada con parámetro de localización μ, parámetro de dispersión σ y υ grados de libertad. Muchos otros casos están incluidos en distribuciones GIG y por lo tanto en distribuciones GH.

La solución de (6.2) establece de manera análitica a la densidad GH. Por lo tanto, considerando los conceptos anteriores, podemos definir de manera consisa a la distribución GH:

Definición 9. *Se dice que la variable aleatoria X tiene una distribución hiperbólica generalizada si su función de densidad es*

$$GH(x;\lambda,\alpha,\beta,\delta,\mu) = a\,(\lambda,\alpha,\beta,\delta)\,\{\delta^2+(x-\mu)^2\}^{(\lambda-1/2)/2}\,K_{\lambda-1/2}\left(\alpha\sqrt{\delta^2+(x-\mu^2}\right)\exp\{\beta(x-\mu)\} \qquad (6.4)$$

con

$$a\,(\lambda,\alpha,\beta,\delta) = \frac{(\alpha^2-\beta^2)^{\lambda/2}}{\sqrt{2\,\pi}\,\alpha^{\lambda-\frac{1}{2}}\,K_\lambda\,(\delta\sqrt{\alpha^2-\beta^2})},$$

donde $x \in \mathbb{R}$ y K_ν es la función de Bessel de tercer orden con índice ν.

Como se puede observar, la función de densidad depende de cinco parámetros: $\alpha > 0$ como un parámetro de forma, β con $0 \leq |\beta| < \alpha$ que determina la asimetría, $\mu \in \mathbb{R}$ la localización, $\delta > 0$ como un factor de escala y el parámetro $\lambda \in \mathbb{R}$ que caracteriza a ciertas sub-clases y se relaciona con la cantidad de masa en las colas.

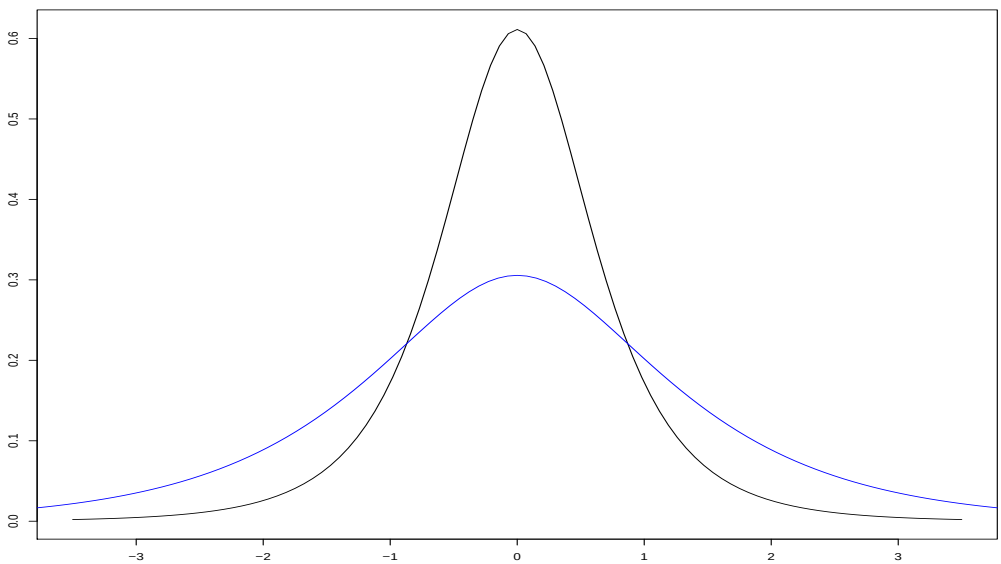

Figura 6.1: Gráficas de la distribución GH

Por ejemplo, en la gráfica de arriba se muestra el comportamiento de la distribución GH al considerar $\alpha = 1$, $\beta = 0$, $\delta = 1$, $\mu = 0$ con $\lambda = -1$ y $\lambda = 1$.

6.2. Propiedades básicas de la distribución GH

En esta sección nos centraremos en el comportamiento de las colas de la distribución GH y en los momentos que caracterizan a la distribución.

Proposición 1. *(Colas de la distribución GH) Si* $X \sim GH(x;\lambda,\alpha,\beta,\delta,\mu)$*, entonces*

$$GH(x;\lambda,\alpha,\beta,\delta,\mu) \sim c\,|x|^{\lambda-1}\exp\{\beta\,x-\alpha\,|x|\} \quad \textit{cuando} \quad x\to\pm\infty,$$

donde c *es una constante. Además, si* $\beta>0$ *la cola derecha decae si*

$$GH(x;\lambda,\alpha,\beta,\delta,\mu) \sim c\,x^{\lambda-1}\exp\{x\,(\beta-\alpha)\} \quad \textit{cuando} \quad x\to+\infty,$$

y si $\beta<0$ *la cola izquierda decae si*

$$GH(x;\lambda,\alpha,\beta,\delta,\mu) \sim c\,|x|^{\lambda-1}\exp\{x\,(\alpha+\beta)\} \quad \textit{cuando} \quad x\to-\infty.$$

Esta proposición demuestra que las colas de la distribución GH es el producto de una función potencia y una función exponencial.

Proposición 2. *(Momentos) Si* $X \sim GH(\lambda,\alpha,\beta,\delta,\mu)$ *entonces el* n*-ésimo momento no centrado se caracteriza a través de la siguiente fórmula de recursión*

$$M_{GH}^{(n)} = \sum_{i=0}^{n}\binom{n}{i}\mu^{n-i}\,M_{GIG}\,(i,\beta,\omega,\eta), \quad n=1,2,\ldots$$

donde $M_{GIG}\,(0,\beta,\omega,\eta)=1$*,* y

$$M_{GIG}\,(n,\beta,\omega,\eta) = \begin{cases} \sum_{i=1}^{r}\frac{(2r-1)!\,\beta^{2\,i-1}}{(2\,i-1)!\,(r-i)!\,2^{r-i}}\,\frac{K_{\lambda+r+i-1}\,(\omega)\,\eta^{r+i-1}}{K_{\lambda}\,(\omega)} & n=2\,r-1, \\ \sum_{i=0}^{r}\frac{(2r)!\,\beta^{2\,i}}{(2\,i)!\,(r-i)!\,2^{r-i}}\,\frac{K_{\lambda+r+i}\,(\omega)\,\eta^{r+i}}{K_{\lambda}\,(\omega)} & n=2\,r, \end{cases}$$

con $\omega=\delta\,\sqrt{\alpha^2-\beta^2}$ y $\eta=\delta\sqrt{\alpha^2-\beta^2}$.

Para detalles de la demostración véase Mena y Walker (2007b).

Modelo GH-ARCH estacionario

Capítulo 7

Modelos estacionarios a través de variables latentes

Una de las principales características que se desean en un modelo estocástico es que su comportamiento sea estable, por ejemplo, con respecto al tiempo. Esta razón obedece a aspectos de interpretación y de estimación. Entre los principales métodos para medir tal característica se encuentran la estacionariedad, fuerte y débil, la reversivilidad, la ergodicidad, la recurrencia, entre otros. La manera en que se construyen estos modelos se considerará en el desarrollo del presente capítulo. En la primer parte nos centraremos en el estudio de modelos estrictamente estacionarios de orden uno, principalmente en la construcción planteada por Pitt et al. (2002). De esta manera se procederá a estudiar modelos de orden mayor a uno a partir de la generalización del modelo de Pitt et al. (2002) mediante la introducción de modelos de mezclas finitas (Raftery, 1985). En la segunda parte, con base en lo planteado por Mena y Walker (2007b), se estudiarán las condiciones bajo las cuales el proceso GH-ARCH es estrictamente estacionario. Posteriormente se estimarán los parámetros producidos en el modelo GH-ARCH(p) a partir del algoritmo EM (Dempster, 1977). Por último, para ilustrar lo anterior, se analizarán las bases de datos que corresponde al precio diario de las acciones emitidas por JP MORGAN CHASE CO, CIT GROUP INC (DEL), AMER INTL GROUP NEW y TINTL BUS MACHINE del 10 de julio de 2008 hasta el 10 de julio de 2009 en donde se estiman los parámetros de la densidad estacionaria y se establece el orden óptimo del modelo GH-ARC(p).

7.1. Modelos de primer orden a través de variables latentes

Pitt et al. (2002) introducen un método flexible para la construcción de modelos autorregresivos de orden uno estrictamente estacionarios con distribución marginal arbitraria. Este procedimiento, de manera general, es el siguiente: Considerando que se desea un modelo estacionario AR(1) con distribución marginal $f_X(x)$, se introduce una densidad condicional y se define la densidad de transición del proceso $\{X_t\}$ mediante

$$p(x_{t-1}, x) = \int f_{X|Y}(x|y) f_{Y|X}(y|x_{t-1})\, d\lambda(y), \tag{7.1}$$

donde $\lambda(\cdot)$ representa una medida discreta, si y es discreta, o la medida de Lebesge si y es continua. La manera en que se especifica la forma de $f_{Y|X}(y|x_{t-1})$ es mediante la introducción de un proceso latente $\{Y_t\}$ a través del siguiente mecanismo

$$\{Y_{t+1}|X_t = x\} \sim f_{Y|X}(\cdot|x), \qquad \{X_{t+1}|Y_{t+1} = y\} \sim f_{X|Y}(\cdot|y),$$

para $t = 1, 2, \ldots$. Al aplicar el teorema de Bayes se obtiene la forma de $f_{X|Y}(x|y)$, es decir,

$$f_{X|Y}(x|y) = \frac{f_{Y|X}(y|x) f_X(x)}{\int f_{Y|X}(y|x) f_X(x)\, d\lambda(x)} \propto f_{Y|X}(y|x) f_X(x).$$

Los procesos construidos $\{X_t\}$ y $\{Y_t\}$ heredan las propiedades de una cadena de Markov generada a través de un muestreo de Gibbs, la única diferencia radica en que las cadenas generadas siempre son estacionarias.

Es fácil ver que la densidad de transición (7.1) define un proceso estacionario y reversible con densidad marginal $f_X(\cdot)$. Además $f_X(\cdot)$ constituye una densidad invariante para la transición (7.1), es decir,

$$f_X(x) = \int p(x_{t-1}, x) f_X(x_{t-1})\, d\lambda(x_{t-1}). \tag{7.2}$$

Por lo tanto, la dependencia en el modelo estará impuesta por la elección de $f_{Y|X}(y|x)$. Dado que esta imposición puede tomar muchas formas, Pitt et al. (2002) restringen esta imposición mediante la existencia de una relación lineal con respecto a la media, es decir,

$$\mathbb{E}[X_t|X_{t-1}] = \rho X_{t-1} + (1-\rho)\mu \tag{7.3}$$

donde $\mu = \int x f_X(x)\, dy$, $0 < \rho < 1$ y $f_X(x)$ es la densidad estacionaria de $\{X_t\}$. Grundwald et al. (2000) señalaron que la mayoría de los modelos estacionarios no Gaussianos conocidos en la literatura se limitan a la simple relación de dependencia (7.3).

Una de las ventajas del enfoque de Pitt et al. (2002) es que, para el caso de un AR(1), la Ecuación (7.1) se puede manipular (logrando expresiones de fácil manejo). Si la integral se puede calcular de manera analítica,

podemos entonces operar directamente con la densidad de transición. Si la integral en (7.1) es difícil de calcular o manipular, la representación de la integral a través de la variable latente es útil.

Para ejemplificar lo anterior, Pitt et al. (2002) construyeron un modelo AR(1) a partir del proceso $\{X_t\}$ con distribución estacionaria

$$\mathrm{Ga}\,(x; a, 1) \propto x^{a-1}\, e^{-x} \quad \mathbb{I}(x > 0), \quad a > 0.$$

Entonces, si se define $f_{Y|\,X}\,(y|\,x)$ como una distribución Poisson,

$$\mathrm{Po}\,(y; x\,\phi) \propto (x\,\phi)^y / y!\, \exp\{-x\,\phi\} \quad \mathbb{I}_{\{y=0,1,2,\cdots\}},$$

por el teorema de Bayes, la conjugación entre estas dos distribuciones conduce a que

$$\begin{aligned} f_{X|\,Y}\,(x|\,y) &\propto f_{Y|\,X}\,(y|\,x)\, f_X\,(x) \\ &\propto x^{a-1}\, \exp\{-x\}\, (x\,\phi)^y\, \exp\{-x\,\phi\} \\ &\propto x^{a+y-1}\, \exp\{-x\,(1+\phi)\} \\ &\propto \mathrm{Ga}\,(x; a+y, 1+\phi). \end{aligned}$$

En este ejemplo, Pitt et al.(2002) demuestran que la dependencia lineal (7.3) se satisface, debido a que,

$$\begin{aligned} \mathbb{E}\,[X_t|\, X_{t-1} = x] &= \mathbb{E}\,[\mathbb{E}\,[X_t|\, Y]|\, X_{t-1} = x] \\ &= \frac{a + \phi\, x}{1+\phi} \\ &= \frac{\phi\, x}{1+\phi} + \left(\frac{1+\phi-\phi}{1+\phi}\right) a \\ &= \frac{\phi}{1+\phi}\, x + \left(1 - \frac{\phi}{1+\phi}\right) a, \end{aligned}$$

por lo tanto, se puede ver que $\rho = \phi/(1+\phi)$ y $\mu = a$ donde ρ se restringe en el intervalo $(0, 1)$. El caso general donde $b \neq 1$ se desarrolla en Mena y Walker (2007a), en donde,

$$\begin{aligned} \mathbb{E}\,[X_t|\, X_{t-1} = x] &= \frac{1-\rho}{b}\, \{a + \mathbb{E}\,[Y|\, X_{t-1} = x]\} \\ &= \rho x + (1-\rho)\, \frac{a}{b}, \end{aligned}$$

entonces, es fácil ver que $ACF(h) = \rho^h$.

Partiendo de la Ecuación (7.1) y considerando a $\lambda\,(\cdot)$ como la medida de Lebesge, Mena y Walker (2007a) proporcionaron, de forma analítica, la expresión de la densidad de transición en un paso para el modelo AR(1),

es decir,

$$p\left(x_{t-1}, x_t\right) = \sum_{y=0}^{\infty} \mathrm{Ga}\left(x_t; y+a, b+\phi\right)\mathrm{Po}\left(y; x_{t-1}\,\phi\right) \tag{7.4}$$

$$= \frac{\exp\{-(b/(1-\rho))[x_t + \rho\, x_{t-1}]\}}{(1-\rho)\,\rho^{(a-1)/2}} \sqrt{\frac{x_t}{x_{t-1}}}^{a-1} I_{a-1}\left(\frac{2\,b\,\rho^{1/2}\sqrt{x_t\, x_{t-1}}}{1-\rho}\right),$$

donde $x, x_{t-1} > 0$ e $I_v\left(\cdot\right)$ denota la función de Bessel modificada de primer orden con índice ν.

La forma en que se construye $f_{Y\,|\,X}\left(y|\,x\right)$ dada $f_X\left(x\right)$ a partir de modelos más generales se expone en Pitt et al. (2002). Para esto suguieren el uso de los siguientes modelos:

- La familia exponencial de convoluciones cerradas i.d.,

$$f\left(y;\theta,\tau\right) = c\left(y;\tau\right)\,\exp\{y\,\theta - \tau\, M(\theta)\}, \quad \tau > 0,$$

- La introducción de una nueva familia de densidades, y

$$f_Y\left(y\right) = c\,\exp\{a\,g(y) - b\,h(y)\}\,g'(y),$$

  donde $h'/g' = y$ y a, b y c son constantes.

En Mena y Walker (2005) la elección de tal distribución puede hacerse mediante un enfoque Bayesiano no parámetrico. Esta idea parte del hecho de que la densidad $f_{Y\,|\,X}\left(y|\,x\right)$ puede ser vista como la distribución posterior correspondiente a la distribución inicial en la variable Y. Entonces $p\left(x_{t-1}, x_t\right)$ se convierte en la distribución posterior predictiva.

7.2. Modelos de mezclas finitas estrictamente estacionarios

La construcción de modelos estrictamente estacionarios de orden mayor se establece a partir de generalizar el mecanismo de Pitt et al. (2002). Esta técnica fue introducida por Raftery (1985). En su artículo, considera una función de densidad construida por una mezcla discreta, es decir,

$$f\left(x|\,x_{t-1}, \cdots, x_{t-p}\right) = \sum_{k=1}^{p} w_k\, p_k\left(x_{t-k}, x\right), \tag{7.5}$$

para toda t, donde $w_k \geq 0$ y $\sum_{k=1}^{p} w_k = 1$. La manera en que se determina la dependencia en este modelo es el valor de retraso x_{t-k}.

No obstante, las distintas elecciones de la forma paramétrica supuesta para $p(\cdot,\cdot)$, en el caso no Gaussiano, representa una limitante para el uso potencial de estos modelos debido a la relación entre las condiciones de estacionariedad y la disponibilidad de métodos para su posible estimación. Además, cuando el soporte del proceso no se encuentra en el conjunto de los números rales, no siempre es claro qué clase de densidad de transisión debe de usarse. Una de las principales ventajas del enfoque de Pitt et al. (2002) es que aborda ambos temas para procesos estrictamente estacionarios y también proporciona una forma fácil de construir formas paramétricas para las distribuciones de transición. Asimismo, la representación de (7.1) para la densidad de trancisión en la construcción de Pitt. et al. (2002) no sólo nos proporciona los medios para la construcción de modelos más complejos con respecto a la dependencia de sus estructuras, sino también para estudiar momentos de orden mayor.

Considerando lo anterior, Mena y Walker (2007a) establecen las condiciones bajo las cuales un modelo de mezclas finitas es estrictamente estacionario, es decir,

Teorema 7. *Un modelo de mezclas finitas $\{X_t\}_{t\in\mathbb{N}}$ es estrictamente estacionario con densidad marginal f_X cuándo $X_1 \sim F_X$ y para todo $t \geq 2$*

$$f\left(x_t \middle| x^{[t,t_p]}\right) = \sum_{k=1}^{t_p-1} w_k\, p\left(x_{t-k}, x_t\right) + \left(1 - \sum_{k=1}^{t_p-1} w_k\right) p\left(x_{t-t_p}, x_t\right) \tag{7.6}$$

donde $t_p := (t-1) \wedge p$, $i \wedge p := min\{i,p\}$, $\sum_{k=1}^{p-1} w_k \leq 1$, $X^{[t,i]} := (X_{t-1}, \ldots, X_{t-i})$ son los valores aleatorios retrasados con un desfase en i y que tiene su origen en el tiempo t, $x^{[t,i]}$ son los puntos observados y la densidad de transición $p(x_k,\cdot)$ estará dada por la Ecuación (7.1).

El Teorema 7 utiliza la estacionariedad estricta como una característica constructiva en lugar de una propiedad que está en función de algunos valores de los parámetros. Este también abarca densidades condicionales en función de valores menores a p, proporcionando una caracterización completa de todas las distribuciones finito dimensionables de $\{X_t\}_{t\in\mathbb{N}}$. En este enfoque, la densidad de transición requerida para la construcción del modelo de mezclas finitas está dada de tal manera que garantiza que las distribuciones marginales $f_X(\cdot)$ sean invariantes. Entonces, la construcción de la transición a través de un proceso latente, es la siguiente,

$$f\left(x_t \middle| x^{[t,p]}\right) = \sum_{k=1}^{p} w_k \int f_{X|Y}\left(x \middle| y_t\right) f_{Y|X}\left(y_t \middle| x_{t-k}\right) d\lambda(y_t) = \int f_{X|Y}\left(x \middle| y_t\right) f\left(y_t \middle| x^{[t,p]}\right) d\lambda(y_t), \tag{7.7}$$

donde

$$f\left(y_t|\, x^{[t,p]}\right) = \sum_{k=1}^{p} w_k\, f_{Y|\,X}\left(y_t|\, x_{t-k}\right). \tag{7.8}$$

Esta representación nos permite estudiar todas las propiedades de dependencia del modelo de mezclas finitas estacionario, incluso en los casos en donde la densidad de transición no se conoce explícitamente, lo que permite considerar modelos con complejas estructuras de dependencia.

En Mena y Walker (2007a) se considera una generalización de la propiedad (7.3) para modelos de orden p, es decir,

$$\mathbb{E}\left[X_t|\, X^{[t,p]}\right] = \sum_{k=1}^{p} w_k \left\{\rho\, X_{t-k} + (1-\rho)\,\mu\right\} = \rho\, \left(\sum_{k=1}^{p} w_k\, X_{t-k}\right) + (1-\rho)\,\mu. \tag{7.9}$$

Entonces, de manera general, para modelos de mezclas finitas que satisfacen la Ecuación (7.9), la ecuación de diferencias de orden p, estará dada por,

$$r(h) = \rho \sum_{k=1}^{p} w_k\, r(h-k), \qquad h \geq p.$$

7.3. Modelo estacionario GH-ARCH

Generalmente, las herramientas para el análisis de la dependencia en series de tiempo se basan en los momentos de segundo orden. Sin embargo, los momentos de orden mayor pueden ser cruciales en el análisis de la dependencia, este es el caso de los modelos ARCH (Engle, 1982).

Teniendo en cuenta que el modelo ARCH sólo considera el comportamiento estadístico para la varianza a través de un periodo de tiempo específico, algunas características para la serie bajo estudio $\{X_t\}$, como la agrupación de la volatilidad, lepkurtosis y la asimetría no están consideradas en este modelo. En respuesta a esto, Barndorff-Nielsen (1997) propone el siguiente modelo ARCH(p) capaz de incluir todas estás características:

$$\begin{aligned} X_{t+p}|\, Y_{t+p} &\sim \mathrm{N}(\mu + \beta\, Y_{t+p}, Y_{t+p}) \quad \text{para} \quad t = 1, 2, \ldots && \text{(densidad observada)} \\ Y_{t+p}|\, X^{(t,p-1)} &\sim \mathrm{GIG}(1/2, r(X^{(t,p-1)}; \theta), \alpha^2) && \text{(densidad del espacio de estados),} \end{aligned}$$

para $p = 1, 2, \ldots, \mu$, $\alpha > 0$, $0 \leq |\beta| < \alpha$ y $X^{(t,p-1)} = \{X_t, X_{t+1}, \ldots, X_{t+p-1}\}$. Dentro de este contexto, Barndorff-Nielsen (1997) propuso la siguiente especificación para $r(\cdot; \theta)$,

$$r\left(X^{(t,i)}; \theta\right) = \left(\varepsilon + \sum_{j=0}^{i} \rho_j\, X_{t+j}^2\right)^k,$$

donde $\varepsilon, \rho_j, k > 0, j \geq 0$. Otras especificaciones de la función $r(\cdot;\theta)$ con argumentos econométricos se estudian en Andersson (2001) y Jensen y Lunde (2001). Sin embargo, ninguno de estos modelos de tipo ARCH llevan siempre a una estricta estacionariedad. Teniendo en cuenta lo anterior, Mena y Walker (2007b) establecen las condiciones bajo las cuales el modelo de Barndorff-Nielsen (1997) es estrictamente estacionario.

7.3.1. Modelo estacionario GH-ARCH(1)

Supongamos que se desea construir un modelo estacionario con densidad marginal

$$f_X(x) = \mathrm{St}(x; 0, \beta^2, \nu).$$

Si

$$f_{Y|X}(y|x) = \mathrm{N}(x; 0, y^{-1}) \quad \text{y} \quad f_Y(y) = \mathrm{Ga}(y; \nu/2, \nu\beta^2/2),$$

entonces, la conjugación entre estas dos distribuciones implica que

$$\begin{aligned}
f_{X|Y}(x|y) &\propto \mathrm{N}(x; 0, y^{-1})\,\mathrm{Ga}(y; \nu/2, \nu\beta^2/2) \\
&\propto \frac{1}{\sqrt{2\pi y^{-1}}} \exp\left\{-\frac{1}{2y^{-1}}(x-0)^2\right\} \frac{\left(\frac{\nu\beta^2}{2}\right)^{\nu/2}}{\Gamma\left(\frac{\nu}{2}\right)} y^{\nu/2-1} \exp\left\{-(\nu\beta^2/2)\,y\right\} \\
&\propto \sqrt{y} \exp\left\{-\frac{y}{2}x^2\right\} y^{(\nu/2)-1} \exp\left\{-(\nu\beta^2/2)y\right\} \\
&\propto y^{\frac{\nu+1}{2}} \exp\left\{-y\left(\frac{x^2+\nu\beta^2}{2}\right)\right\} \\
&\propto \mathrm{Ga}\left(y; \frac{\nu+1}{2}, \frac{x^2+\nu\beta^2}{2}\right) \\
&\therefore f_{X|Y}(x|y) = \mathrm{Ga}\left(y; \frac{\nu+1}{2}, \frac{x^2+\nu\beta^2}{2}\right).
\end{aligned}$$

Con estos componentes es posible construir un modelo Markoviano $\{X_t\}$ con densidad de transición

$$p\left(x_t, x_{t+1}\right) = \int_0^\infty \mathrm{N}\left(x_{t+1}; 0, y^{-1}\right) \mathrm{Ga}\left(y; \frac{\nu+1}{2}, \frac{x^2+\nu\,\beta^2}{2}\right) dy$$

$$= \int_0^\infty \frac{y^{1/2}}{\sqrt{2\,\pi}} \exp\left\{-\frac{y}{2}\,x_{t+1}^2\right\} \frac{\left(\frac{x_t^2+\nu\,\beta^2}{2}\right)^{\nu/2}}{\Gamma\left(\frac{\nu+1}{2}\right)} y^{((\nu+1)/2)-1} \exp\left\{-\left(\frac{x_t^2+\nu\,\beta^2}{2}\right) y\right\} dy$$

$$= \frac{\left(\frac{x_t^2+\nu\,\beta^2}{2}\right)^{\nu/2}}{\sqrt{2\,\pi}\,\Gamma\left(\frac{\nu+1}{2}\right)} \int_0^\infty y^{1/2}\, y^{((\nu+1)/2)-1} \exp\left\{-\frac{y}{2}\,x_{t+1}^2\right\} \exp\left\{-\left(\frac{x_t^2+\nu\,\beta^2}{2}\right) y\right\} dy$$

$$= \frac{\left(\frac{x_t^2+\nu\beta^2}{2}\right)^{\nu/2}}{\sqrt{2\,\pi}\,\Gamma\left(\frac{\nu+1}{2}\right)} \underbrace{\int_0^\infty y^{(\nu/2+1)-1} \exp\left\{-y\left(\frac{x_{t+1}^2+(x_t^2+\nu\,\beta^2)}{2}\right)\right\} dy}_{\mathrm{Ga}\left(y; \frac{\nu}{2}+1, \frac{x_{t+1}^2+x_t^2+\nu\,\beta^2}{2}\right)}$$

entonces,

$$p\left(x_t, x_{t+1}\right) = \frac{\left(\frac{x_t^2+\nu\,\beta^2}{2}\right)^{\nu/2}}{\sqrt{2\,\pi}\,\Gamma\left(\frac{\nu+1}{2}\right)} \frac{\Gamma\left(\frac{\nu}{2}+1\right)}{\left(\frac{x_{t+1}^2+x_t^2+\nu\,\beta^2}{2}\right)^{\nu/2+1}} = \frac{\Gamma\left(\frac{\nu}{2}+1\right)}{\Gamma\left(\frac{\nu+1}{2}\right)} \frac{1}{\sqrt{2\pi}} \frac{\left(\frac{x_t^2+\nu\,\beta^2}{2}\right)^{(\nu+1)/2}}{\left(\frac{x_{t+1}^2+x_t^2+\nu\,\beta^2}{2}\right)^{(\nu/2)+1}}$$

$$= \frac{\Gamma\left(\frac{\nu}{2}+1\right)}{\Gamma\left(\frac{\nu+1}{2}\right)} \frac{1}{\sqrt{2\,\pi}} \frac{2}{\sqrt{2}} \left(\frac{x_t^2+\nu\beta^2}{x_{t+1}^2+x_t^2+\nu\,\beta^2}\right)^{\nu/2} \frac{\sqrt{x_t^2+\nu\,\beta^2}}{x_{t+1}^2+x_t^2+\nu\,\beta^2}$$

$$= \frac{\Gamma\left(\frac{\nu}{2}+1\right)}{\Gamma\left(\frac{\nu+1}{2}\right)} \frac{1}{\sqrt{\pi}} \left(\frac{x_t^2+\nu\,\beta^2}{x_{t+1}^2+x_t^2+\nu\,\beta^2}\right)^{\nu/2} \frac{x_t^2+\nu\,\beta^2}{\sqrt{x_t^2+\nu\beta^2}} \frac{1}{x_{t+1}^2+x_t^2+\nu\,\beta^2}$$

$$= \frac{\Gamma\left(\frac{\nu}{2}+1\right)}{\Gamma\left(\frac{\nu+1}{2}\right)} \frac{1}{\sqrt{\pi}} \left(\frac{x_t^2+\nu\beta^2}{x_{t+1}^2+x_t^2+\nu\beta^2}\right)^{\nu/2} \left(\frac{1}{x_t^2+\nu\,\beta^2}\right)^{1/2} \left(\frac{x_t^2+\nu\,\beta^2}{x_{t+1}^2+x_t^2+\nu\,\beta^2}\right)$$

$$= \frac{\Gamma\left(\frac{\nu}{2}+1\right)}{\Gamma\left(\frac{\nu+1}{2}\right)} \frac{1}{\sqrt{\pi}} \left(\frac{x_t^2+\nu\,\beta^2}{x_{t+1}^2+x_t^2+\nu\,\beta^2}\right)^{(\nu/2)+1} \left(\frac{1}{x_t^2+\nu\,\beta^2}\right)^{1/2}$$

$$= \frac{\Gamma\left(\frac{\nu}{2}+1\right)}{\Gamma\left(\frac{\nu+1}{2}\right)} \frac{1}{\sqrt{\pi}} \left(\frac{x_{t+1}^2 + x_t^2 + \nu\,\beta^2}{x_t^2 + \nu\,\beta^2}\right)^{-((\nu/2)+1)} \left(\frac{1}{x_t^2 + \nu\,\beta^2}\right)^{1/2}$$

$$= \frac{\Gamma\left(\frac{\nu}{2}+1\right)}{\Gamma\left(\frac{\nu+1}{2}\right)\sqrt{x_t^2 + \nu\,\beta^2}\,\sqrt{\pi}} \left(1 + \frac{x_{t+1}^2}{x_t^2 + \nu\,\beta^2}\right)^{-((\nu/2)+1)}$$

$$= \frac{\Gamma\left(\frac{(\nu+1)+1}{2}\right)}{\Gamma\left(\frac{\nu+1}{2}\right)\sqrt{\frac{x_t^2+\nu\,\beta^2}{\nu+1}}\,\sqrt{(\nu+1)\,\pi}} \left(1 + \frac{x_{t+1}^2}{(\nu+1)\left(\frac{x_t^2+\nu\,\beta^2}{\nu+1}\right)}\right)^{-\frac{(\nu+1)+1}{2}}$$

$$= \mathrm{St}\left(x_{t+1}; 0, \frac{x_t^2 + \nu\,\beta^2}{\nu+1}, \nu+1\right).$$

Es fácil ver que este modelo puede reescribirse como

$$X_{t+1} = \sqrt{\frac{x_t^2 + \nu\,\beta^2}{\nu+1}}\,\varepsilon_{t+1}, \qquad \varepsilon \sim \mathrm{St}\,(0,1,\nu+1). \tag{7.10}$$

que puede ser reconocido como el modelo ARCH(1) estacionario con inovaciones Student t propuesto por Pitt y Walker (2005) o como la versión estacionaria del modelo t-ARCH(l) de Bollerslev (1987).

Considerando el modelo anterior, si

$$f_{X|\,Y}\,(\cdot|\,y) = \mathrm{N}\,(\mu + \beta\,y, y) \quad \text{y} \quad f_Y\,(\cdot) = \mathrm{GIG}\,(\lambda, \delta^2, \alpha^2 - \beta^2), \tag{7.11}$$

donde $\mu, \beta, \lambda \in \mathbb{R}$ y $0 \leq |\beta| < \alpha$, entonces, al aplicar el teorema de Bayes, se tiene que,

$$f_{Y|\,X}\,(y|\,x) \propto \mathrm{N}\,(x; \mu + \beta\,y, y)\,\mathrm{GIG}\,(y; \lambda, \delta^2, \alpha^2 - \beta^2)$$

$$\propto \frac{1}{\sqrt{2\,\pi\,y}}\,\exp\left\{-\frac{1}{2\,y}\,(x - (\mu + \beta\,y))^2\right\}\,y^{\lambda-1}\exp\left\{-\frac{1}{2}\,(\delta^2\,y^{-1} + (\alpha^2 - \beta^2)\,y)\right\}$$

$$\propto y^{(\lambda-1/2)-1}\exp\left\{-\frac{1}{2}\,\left[y^{-1}(x - (\mu + \beta\,y))^2 + \delta^2\,y^{-1} + (\alpha^2 - \beta^2)\,y\right]\right\}$$

$$\propto y^{(\lambda-1/2)-1}\exp\left\{-\frac{1}{2}\left[y^{-1}((x-\mu)^2-2\,\beta y\,(x-\mu)+\beta^2\,y^2)+\delta^2 y^{-1}+(\alpha^2-\beta^2)\,y\right]\right\}$$

$$\propto y^{(\lambda-1/2)-1}\exp\left\{-\frac{1}{2}\left[y^{-1}(x-\mu)^2-2\,\beta(x-\mu)+\beta^2\,y+\delta^2\,y^{-1}+\alpha^2\,y-\beta^2\,y\right]\right\}$$

$$\propto y^{(\lambda-1/2)-1}\exp\left\{-\frac{1}{2}\left[y^{-1}((x-\mu)+\delta^2)+\alpha^2\,y\right]\right\}\exp\left\{\beta\,(x-\mu)\right\}$$

$$\propto y^{(\lambda-1/2)-1}\exp\left\{-\frac{1}{2}\left[y^{-1}((x-\mu)+\delta^2)+\alpha^2\,y\right]\right\}$$

$$\propto \text{GIG}\left(y;\lambda-\frac{1}{2},(x-\mu)^2+\delta^2,\alpha^2\right)$$

$$\therefore f_{Y|\,X}=\text{GIG}\left(y;\lambda-\frac{1}{2},(x-\mu)^2+\delta^2,\alpha^2\right).$$

Por lo tanto la probabilidad de transición de un solo paso estará dada por

$$p\,(x_t,x_{t+1})=\int_0^\infty \text{N}\,(x_{t+1};\mu+\beta\,y)\;\text{GIG}\left(y;\lambda-\frac{1}{2},(x_{t+1}-\mu)^2+\delta^2,\alpha^2\right)\,dy$$

$$=\int_0^\infty \frac{1}{\sqrt{2\,\pi y}}\exp\left\{-\frac{1}{2\,y}\,(x_{t+1}-(\mu+\beta\,y))^2\right\}\frac{\left(\frac{\alpha^2}{(x_{t+1}-\mu)^2+\delta^2}\right)^{(\lambda-1/2)/2}}{2\,K_{\lambda-1/2}\,(\sqrt{((x_{t+1}-\mu)^2+\delta^2)\,\alpha^2})}\,y^{(\lambda-1/2)-1}$$

$$\times\exp\left\{-\frac{1}{2}\,(y^{-1}\,((x_{t+1}-\mu)^2+\delta^2)+\alpha^2\,y\right\}\,dy,$$

sea

$$\mathscr{E}:=\frac{\left(\frac{\alpha^2}{(x_{t+1}-\mu)^2+\delta^2}\right)^{(\lambda-1/2)/2}}{2\,\sqrt{2\,\pi}\,K_{\lambda-1/2}\,(\sqrt{((x_{t+1}-\mu)^2+\delta^2)\,\alpha^2})},$$

entonces,

$$p\,(x_t, x_{t+1}) = \mathscr{E} \int_0^\infty y^{(\lambda-1)-1} \exp\left\{-\frac{1}{2}\,[\,y^{-1}\,((x_{t+1}-\mu)-\beta y)^2 + y^{-1}\,((x_{t+1}-\mu)^2+\delta^2) + \alpha^2\,y\,]\right\} dy$$

$$= \mathscr{E} \int_0^\infty y^{(\lambda-1)-1} \exp\left\{-\frac{1}{2}\,[\,2\,y^{-1}\,(x_{t+1}-\mu)^2 - 2\,\beta\,(x_{t+1}-\mu) + \beta^2\,y + \delta^2\,y^{-1} + \alpha^2\,y\,]\right\} dy$$

$$= \mathscr{E} \int_0^\infty y^{(\lambda-1)-1} \exp\left\{-\frac{1}{2}\,[\,y^{-1}\,(2\,(x_{t+1}+\delta^2) + (\alpha^2+\beta^2)\,y - 2\,\beta\,(x_{t+1}-\mu)\,]\right\} dy$$

$$= \mathscr{E} \int_0^\infty y^{(\lambda-1)-1} \exp\left\{-\frac{1}{2}\,[\,y^{-1}\,(2\,(x_{t+1}+\delta^2) + (\alpha^2+\beta^2)\,y\,]\right\} \exp\{\beta\,(x_{t+1}-\mu)\}\,dy$$

$$= \mathscr{E}\, \exp\{\beta\,(x_{t+1}-\mu)\} \underbrace{\int_0^\infty y^{(\lambda-1)-1} \exp\left\{-\frac{1}{2}\,[\,y^{-1}\,(2\,(x_{t+1}+\delta^2) + (\alpha^2+\beta^2)\,y\,]\right\})\}\,dy}_{\mathrm{GIG}\,(\lambda-1, 2\,(x_{t+1}-\mu)^2+\delta^2, \alpha^2+\beta^2)}$$

$$= \mathscr{E}\, \exp\{\beta\,(x_{t+1}-\mu)\}\, \frac{2\,K_{\lambda-1}\,(\sqrt{2\,(x_{t+1}-\mu)^2+\delta^2}\,\sqrt{\alpha^2+\beta^2})}{\left(\frac{\alpha^2+\beta^2}{2\,(x_{t+1}-\mu)^2+\delta^2}\right)^{(\lambda-1)/2}},$$

por lo tanto, considerando la definición de $\mathscr{E}$, tenemos que,

$$p\,(x_t, x_{t+1}) = \frac{\left(\frac{\alpha^2}{(x_{t+1}-\mu)^2+\delta^2}\right)^{(\lambda-1/2)/2}}{2\,\sqrt{2\pi}\,K_{\lambda-1/2}\,(\sqrt{((x_{t+1}-\mu)^2+\delta^2)\,\alpha^2})}\, \frac{2\,K_{\lambda-1}\,(\sqrt{2\,(x_{t+1}-\mu)^2+\delta^2}\,\sqrt{\alpha^2+\beta^2})}{\left(\frac{\alpha^2+\beta^2}{2\,(x_{t+1}-\mu)^2+\delta^2}\right)^{(\lambda-1)/2}}$$

$$\times \exp\{\beta\,(x_{t+1}-\mu)\},$$

$$= \frac{\left(\frac{\alpha^2}{(x_{t+1}-\mu)^2+\delta^2}\right)^{(\lambda-1/2)/2}}{\left(\frac{\alpha^2+\beta^2}{2\,(x_{t+1}-\mu)^2+\delta^2}\right)^{(\lambda-1)/2}}\, \frac{2\,K_{\lambda-1}\,(\sqrt{2\,(x_{t+1}-\mu)^2+\delta^2}\,\sqrt{\alpha^2+\beta^2})}{2\,\sqrt{2\,\pi}\,K_{\lambda-1/2}\,(\sqrt{((x_{t+1}-\mu)^2+\delta^2)\,\alpha^2})}\, \exp\{\beta\,(x_{t+1}-\mu)\}$$

$$= \left(\frac{\alpha^2}{(x_{t+1}-\mu)^2+\delta^2}\right)^{\lambda/2-1/4} \left(\frac{\alpha^2+\beta^2}{2\,(x_{t+1}-\mu)^2+\delta^2}\right)^{-\lambda/2+1/2} \frac{K_{\lambda-1}\left(\sqrt{2\,(x_{t+1}-\mu)^2+\delta^2}\,\sqrt{\alpha^2+\beta^2}\right)}{\sqrt{2\,\pi}\,K_{\lambda-1/2}\left(\sqrt{((x_{t+1}-\mu)^2+\delta^2)\,\alpha^2}\right)}$$

$$\times \exp\{\beta\,(x_{t+1}-\mu)\}.$$

Por otra parte si $\rho^2 = (x_{t+1}-\mu)^2$, entonces,

$$\left(\frac{\alpha^2}{\rho^2+\delta^2}\right)^{\lambda/2-1/4} \left(\frac{\alpha^2+\beta^2}{2\,\rho^2+\delta^2}\right)^{-\lambda/2+1/2} = \left(\frac{2\,\rho^2+\delta^2}{\alpha^2+\beta^2}\right)^{\lambda/2} \left(\frac{\alpha^2+\beta^2}{2\,\rho^2+\delta^2}\right)^{1/2} \left(\frac{\alpha^2}{\rho^2+\delta^2}\right)^{\lambda/2} \left(\frac{\rho^2+\delta^2}{\alpha^2}\right)^{1/4}$$

$$= \left(\frac{2\,\rho^2+\delta^2}{\alpha^2+\beta^2}\right)^{(\lambda-1)/2} \left(\frac{\alpha^2}{\rho^2+\delta^2}\right)^{(\lambda-1/2)/2}$$

$$= \frac{(\,2\,\rho^2+\delta^2\,)^{(\lambda-1)/2}\,(\alpha^2)^{(\lambda-1/2)/2}}{(\,\alpha^2+\beta^2\,)^{(\lambda-1)/2}\,(\,\rho^2+\delta^2\,)^{(\lambda-1/2)/2}}$$

$$= \frac{(\,2\,\rho^2+\delta^2\,)^{(\lambda-1)/2}\,(\alpha^2)^{(\lambda-1/2)/2}}{\left(\sqrt{\alpha^2+\beta^2}\right)^{\lambda-1}\,\left(\sqrt{\rho^2+\delta^2}\right)^{\lambda-1/2}}.$$

Por lo tanto,

$$p\,(x_t, x_{t+1}) = a\,(\,\lambda - 1/2, \sqrt{\alpha^2+\beta^2}, \beta, \sqrt{(x_{t+1}-\mu)^2+\delta^2}\,)\,\{(x_{t+1}-\mu)^2+\delta^2+(\,x_{t+1}-\mu\,)^2\}^{(\lambda-1)/2} \quad (7.12)$$

$$\times\, K_{(\lambda-1/2)-1/2}(\sqrt{\alpha^2+\beta^2}\sqrt{(x_{t+1}-\mu)^2+\delta^2+(x_{t+1}-\mu)^2}\,)\,\exp\{\beta\,(x_{t+1}-\mu)\},$$

donde

$$a(\lambda,\alpha,\beta,\delta) = \frac{(\,\alpha^2-\beta^2\,)^{\lambda/2}}{\sqrt{2\,\pi}\,\alpha^{\lambda-\frac{1}{2}}\,\delta^{\lambda}\,K_{\lambda}\left(\delta\,\sqrt{\alpha^2-\beta^2}\right)},$$

es decir,

$$p\,(x_t, x_{t+1}) = \mathrm{GH}\left(x_{t+1}; \lambda-\frac{1}{2}, \sqrt{\beta^2+\alpha^2}, \beta, \sqrt{(x_t-\mu)^2+\delta^2}, \mu\right). \quad (7.13)$$

Entonces, podemos asegurar que un modelo estrictamente estacionario hipérbolico generalizado ARCH(1) es un proceso de Markov $\{X_t\}$ con distribución de transición (7.13) y distribución marginal $\mathrm{GH}\,(\lambda,\alpha,\beta,\delta,\mu)$ para toda $t \geq 1$.

Recientemente, Mena y Walker (2007b) establecieron una expresión analítica para la correlación de un modelo estacionario GH-ARCH(1), la cual está dada por

$$\operatorname{Corr}(X_{t+1}, X_t) = \frac{\operatorname{Cov}(X_{t+1}, X_t)\, K_\lambda(\omega)}{\eta\, K_{\lambda+1}(\omega) + \operatorname{Cov}(X_{t+1}, X_t)\, K_\lambda(\omega)},$$

donde

$$\operatorname{Cov}(X_{t+1}, X_t) = \beta^2 \eta^2 \left\{ \frac{K_{\lambda+2}(\omega)}{K_\lambda(\omega)} - \left(\frac{K_{\lambda+1}(\omega)}{K_\lambda(\omega)} \right)^2 \right\}$$

y $\omega = \delta\sqrt{\alpha^2 - \beta^2}$ y $\eta = \delta/\sqrt{\alpha^2 - \beta^2}$.

7.3.2. Modelo estacionario GH-ARCH(p)

Consideremos una estructura de dependencia con un desface mayor a 1, donde la distribución $(p+1)$-dimensional esta dada por

$$p(X^{(t,p)}) = q(X_t) \prod_{i=1}^{p} p(X_{t+i} | X^{(t,i-1)}), \tag{7.14}$$

donde $X^{(t,i)} := (X_t, \ldots, X_{t+i})$ para toda $t \in \mathbb{N}$. Al igual que antes, con el fin de mantener la estricta estacionariedad de la secuencia $\{X_t\}$ con distribución conjunta dada por la Ecuación (7.14), tenemos que imponer nuevas condiciones para el mecanismo de actualización $p(X_{t+i}|X^{(t,i-1)})$. Esto es posible mediante el siguiente muestreo de Gibbs

$$\begin{aligned} F_{Y|X^i}(\cdot \,|\, x^{(t,i-1)}) &\sim \{ Y_{t+i} | X^{(t,i-1)} = x^{(t,i-1)}, Y^{(t,i-1)} = y^{(t,i-1)} \} \\ F_{X|Y}(\cdot \,|\, y_{t+i}) &\sim \{ X_{t+i} | Y_{t+i} = y, X^{(t,i-1)} = x^{(t,i-1)}, Y^{(t,i-1)} = y^{(t,i-1)} \}, \end{aligned} \tag{7.15}$$

donde $i = 1, \ldots, p$, X^i denota un vector aleatorio de dimensión i y $x^{(t,i-1)}$ un vector de dimensión i que denota los valores del espacio de tiempo correspondiente a $X^{(t,i-1)}$. Debido a la independencia condicional subyacente y bajo el conocimiento de F_Y, la densidad correspondiente a $F_{Y|X^i}$ se puede obtener de la siguiente manera

$$f_{Y|X^i}(y_{i+1} | x^{(t,i-1)}) \propto q_y(y_{t+i}) \prod_{j=1}^{i} f_{X|Y}(x_{t+j-1} | y_{t+j}). \tag{7.16}$$

Por lo tanto si suponemos que $X_t \sim f_X(x)$ y utilizamos la estructura de independencia condicional, sólo es necesario determinar la forma de $F_{Y|X^i}$ para $i = 1, \ldots, p$ a fin de construir de un proceso estacionario con distribución marginal F_X. Al igual que en el modelo de Pitt et al. (2002), la especificación de las formas funcionales de $F_{Y|X^i}$ es bastante abierta. Por lo tanto, la densidad transición en un paso asociada con este

modelo está dada por,

$$p\,(\,x_{t+i}|\,x^{(t,i-1)}\,)=\int f_{X|\,Y}\,(\,x_{t+i}|\,y\,)\,f_{Y|\,X^i}\,(\,y|\,x^{(t,i-1)}\,)\,\lambda(dy). \tag{7.17}$$

De manera general, en Mena y Walker (2007b) la forma en la que se asegura que el proceso GH-ARCH(p) es estrictamente estacionario es mediante la elección de $r(\cdot,\cdot)$. Entonces para construir la distribución condicional (7.17) se debe suponer (7.11). Se puede ver que

$$f_{X^p|\,Y}\,(\,x^p\,|\,y\,)=\mathrm{N}_p\,(\,x^p;M+y\,\Delta\,B,y\,\Delta\,), \tag{7.18}$$

donde $\Delta\in\mathbb{R}^{p\times p}$ es una matriz positiva definida con $\det(\Delta)=1$. La especificación de f_Y está dada por

$$f_Y\,(\cdot)=\mathrm{GIG}\,(\,\lambda,\delta^2,\alpha^2-B^T\Delta\,B\,), \tag{7.19}$$

donde $\lambda\in\mathbb{R}$, $\delta>0$, $\Delta\in\mathbb{R}^{p\times p}$ y $\alpha^2>B^T\,\delta\,B$. Entonces a partir de las Ecuaciones (7.18) y (7.19) y al aplicar el Teorema de Bayes, tenemos que

$$f_{Y|\,X^p}\propto\mathrm{N}_p\,(\,x^p;M+y\,\Delta\,B,y\,\Delta)\,\mathrm{GIG}\,(\lambda,\delta^2,\alpha^2-B^T\,\Delta\,B)$$

$$\propto\frac{y^{\lambda-1}}{|y\,\Delta|^{1/2}}\,\exp\left\{-\frac{1}{2}\,[\,(x^p-(M+y\,\Delta\,B))^T(y\,\Delta)^{-1}(x^p-(M+y\,\Delta\,B))\,]\right\}$$

$$\times\,\exp\left\{-\frac{1}{2}\,(\,\delta^2\,y^{-1}+(\alpha^2-B^T\,\Delta\,B)\,y\,)\right\}$$

$$\propto y^{(\lambda-p/2)-1}\,\exp\left\{-\frac{1}{2}\,[\,(x^p-(M+y\,\Delta\,B)\,)^Ty^{-1}\Delta^{-1}(x^p-(M+y\Delta B))+\delta^2y^{-1}+(\alpha^2-B^T\,\Delta\,B)\,y\,]\right\}$$

$$\propto y^{(\lambda-p/2)-1}\,\exp\left\{-\frac{1}{2}\,[\,((x^p-M)-y\Delta B))^Ty^{-1}\Delta^{-1}\,((x^p-M)-y\Delta B))+\delta^2y^{-1}+\alpha^2\,y-B^T\,\Delta\,B\,y\,]\right\}$$

$$\propto y^{(\lambda-p/2)-1}\,\exp\left\{-\frac{1}{2}\,[\,((x^p-M)^T-y\,\Delta\,B^T))\,y^{-1}\Delta^{-1}((x^p-M)-y\,\Delta\,B)\,)+\delta^2\,y^{-1}+\alpha^2\,y-B^T\,\Delta\,B\,y\,]\right\}$$

$$\propto y^{(\lambda-p/2)-1}\,\exp\left\{-\frac{1}{2}\,[\,y^{-1}((x^p-M)^T\,\Delta^{-1}\,(x^p-M)+\delta^2)+B^T\,\Delta\,B\,y+\alpha^2\,y-B^T\,\Delta\,B\,y\,]\right\}$$

$$\propto y^{(\lambda-p/2)-1}\,\exp\left\{-\frac{1}{2}\,[\,y^{-1}\,((x^p-M)^T\,\Delta^{-1}\,(x^p-M)+\delta^2)+\alpha^2\,y\,]\right\}$$

$$\propto \text{GIG}\left(y;\lambda-\frac{p}{2},(x^p-M)^T\,\Delta^{-1}\,(x^p-M)+\delta^2,\alpha^2\right)$$

$$\therefore f_{Y|\,X^p}=\text{GIG}\left(y;\lambda-\frac{p}{2},(x^p-M)^T\,\Delta^{-1}(x^p-M)+\delta^2,\alpha^2\right).$$

Mena y Walker (2007b) construyen la probabilidad de transición a partir del mecanismo de actualización (7.15) de la siguiente manera

$$p\,(x_{t+i}|x^{(t,i-1)})=\int_{\mathbb{R}_+}\text{N}\,(\,x_{t+i};\mu_{t+i}+y\,\beta_{t+i}\,)\;\text{GIG}\,\left(\,y;\lambda-\frac{i}{2},r^2_{(t,i-1)},\alpha^2\,\right)\,dy$$

$$=\int_0^\infty\frac{1}{\sqrt{2\,\pi\,y}}\,\exp\left\{-\frac{1}{2\,y}(\,x_{t+i}-(\mu_{t+i}+y\,\beta_{t+i})\,)^2\right\}\,\frac{\left(\frac{\alpha^2}{(x^{(t,i-1)}-M)^T\,\Delta^{-1}\,(x^{(t,i-1)}-M)+\delta^2}\right)^{(\lambda-1/2)/2}}{2\,K_{\lambda-1/2}\,(\sqrt{((x^{(t,i-1)}-M)^T\,\Delta^{-1}\,(x^{(t,i-1)}-M)+\delta^2)\alpha^2})}$$

$$\times\exp\left\{-\frac{1}{2}\,[\,y^{-1}\,((x^{(t,i-1)}-M)^T\,\Delta^{-1}\,(x^{(t,i-1)}-M)+\delta^2)+\alpha^2\,y\,]\right\}$$

$$=\frac{\left(\frac{\alpha^2}{(x^{(t,i-1)}-M)^T\,\Delta^{-1}\,(x^{(t,i-1)}-M)+\delta^2}\right)^{(\lambda-1/2)/2}}{2\,\sqrt{2\,\pi}\,K_{\lambda-1/2}(\sqrt{((x^{(t,i-1)}-M)^T\,\Delta^{-1}\,(x^{(t,i-1)}-M)+\delta^2)\,\alpha^2})}\,\int_0^\infty y^{(\lambda-i/2-1/2)-1}$$

$$\times\exp\left\{-\frac{1}{2\,y}(x_{t+i}-(\mu_{t+i}+y\,\beta_{t+i}))^2\right\}\,\exp\left\{-\frac{1}{2}\,[\,y^{-1}\,((x^{(t,i-1)}-M)^T\,\Delta^{-1}\,(x^{(t,i-1)}-M)+\delta^2)+\alpha^2\,y\,]\right\}\,dy$$

sea

$$\mathscr{F}:=\frac{\left(\frac{\alpha^2}{(\,x^{(t,i-1)}-M\,)^T\,\Delta^{-1}\,(\,x^{(t,i-1)}-M\,)+\delta^2}\right)^{(\lambda-1/2)/2}}{2\,\sqrt{2\,\pi}\,K_{\lambda-1/2}\,\left(\,\sqrt{(\,(x^{(t,i-1)}-M)^T\,\Delta^{-1}\,(x^{(t,i-1)}-M)+\delta^2\,)\,\alpha^2}\,\right)},$$

entonces,

$$= \mathscr{F} \int_0^\infty y^{(\lambda - i/2 - 1/2)-1} \exp\left\{ -\frac{1}{2\,y} \left(x_{t+i} - (\mu_{t+i} + y\,\beta_{t+i}) \right)^2 \right\}$$

$$\times \exp\left\{ -\frac{1}{2} \left[y^{-1} \left((x^{(t,i-1)} - M)^T \, \Delta^{-1} \, (x^{(t,i-1)} - M) + \delta^2 \right) + \alpha^2 \, y \right] \right\} dy$$

$$= \mathscr{F} \int_0^\infty y^{(\lambda - i/2 - 1/2)-1}$$

$$\times \exp\left\{ -\frac{1}{2} \left[y^{-1} \left(x_{t+i} - (\mu_{t+i} + y\beta_{t+i}) \right)^2 + y^{-1} \left((x^{(t,i-1)} - M)^T \, \Delta^{-1} (x^{(t,i-1)} - M) + \delta^2 \right) + \alpha^2 \, y \right] \right\} dy$$

$$= \mathscr{F} \int_0^\infty y^{(\lambda - i/2 - 1/2)-1}$$

$$\times \exp\left\{ -\frac{1}{2} \left[y^{-1} \left[\left((x_{t+i} - \mu_{t+i}) - y\beta_{t+i} \right)^2 + \left((x^{(t,i-1)} - M)^T \Delta^{-1} (x^{(t,i-1)} - M) + \delta^2 \right) \right] + \alpha^2 \, y \right] \right\} dy,$$

si definimos

$$r_{(t,i-1)} := \sqrt{(x^{(t,i-1)} - M)^T \, \Delta^{-1} \, (x^{(t,i-1)} - M) + \delta^2},$$

entonces,

$$= \mathscr{F} \int_0^\infty y^{(\lambda - i/2 - 1/2)-1} \exp\left\{ -\frac{1}{2} \left[y^{-1} \left[((x_{t+i} - \mu_{t+i}) - y\beta_{t+i})^2 + r^2_{(t,i-1)} \right] + \alpha^2 \, y \right] \right\} dy$$

$$= \mathscr{F} \int_0^\infty y^{(\lambda - i/2 - 1/2)-1} \exp\left\{ -\frac{1}{2} \left[y^{-1} \left[(x_{t+i} - \mu_{t+i})^2 - 2\,y\,\beta_{t+i} \, (x_{t+i} - \mu_{t+i}) + y^2 \, \beta^2_{t+i} + r^2_{(t,i-1)} \right] + \alpha^2 \, y \right] \right\} dy$$

$$= \mathscr{F} \int_0^\infty y^{(\lambda - i/2 - 1/2)-1} \exp\left\{ -\frac{1}{2} \left[y^{-1} \, (x_{t+i} - \mu_{t+i})^2 - 2\,\beta_{t+i} \, (x_{t+i} - \mu_{t+i}) + y\,\beta^2_{t+i} + y^{-1} \, r^2_{(t,i-1)} + \alpha^2 \, y \right] \right\} dy$$

$$= \mathscr{F} \int_0^\infty y^{(\lambda - i/2 - 1/2)-1} \exp\left\{ -\frac{1}{2} \left[y^{-1} \, ((x_{t+i} - \mu_{t+i})^2 + r^2_{(t,i-1)}) - 2\,\beta_{t+i} \, (x_{t+i} - \mu_{t+i}) + (\alpha^2 + \beta^2_{t+i}) \, y \right] \right\} dy$$

$$= \mathscr{F} \int_0^\infty y^{(\lambda-i/2-1/2)-1} \exp\left\{-\frac{1}{2}\left[y^{-1}\left((x_{t+i}-\mu_{t+i})^2 + r^2_{(t,i-1)}\right) + \left(\alpha^2+\beta^2_{t+i}\right)y\right]\right\} \exp\left\{\beta_{t+i}\left(x_{t+i}-\mu_{t+i}\right)\right]\} \, dy$$

$$= \mathscr{F} \exp\left\{\beta_{t+i}\left(x_{t+i}-\mu_{t+i}\right)\right\} \underbrace{\int_0^\infty y^{(\lambda-i/2-1/2)-1} \exp\left\{-\frac{1}{2}\left[y^{-1}\left((x_{t+i}-\mu_{t+i})^2 + r^2_{(t,i-1)}\right) + (\alpha^2+\beta^2_{t+i})y\right]\right\} dy}_{\text{GIG}\left(y;\lambda-\frac{1+i}{2},(x_{t+i}-\mu_{t+i})+r^2_{(t,i-1)},\alpha^2+\beta^2_{t+i}\right)}$$

$$= \mathscr{F} \exp\left\{\beta_{t+i}\left(x_{t+i}-\mu_{t+i}\right)\right\} \frac{2\,K_{\lambda-i/2-1/2}\left(\sqrt{\left((x_{t+i}-\mu_{t+i})^2 + r^2_{(t,i-1)}\right)\left(\alpha^2+\beta^2_{t+i}\right)}\right)}{\left(\frac{\alpha^2+\beta^2_{t+i}}{(x_{t+i}-\mu_{t+i})^2+r^2_{(t,i-1)}}\right)^{(\lambda-i/2-1/2)/2}},$$

por lo tanto, considerando la definición de $\mathscr{F}$ y la de $r^2_{(t,i-1)}$, se tiene que

$$p\left(x_{t+i}|\, x^{(t,i-1)}\right) = \frac{\left(\frac{\alpha^2}{r^2_{(t,i-1)}}\right)^{(\lambda-1/2)/2}}{2\sqrt{2\,\pi}\,K_{\lambda-1/2}\left(\sqrt{r^2_{(t,i-1)}\,\alpha^2}\right)} \frac{2\,K_{\lambda-i/2-1/2}\left(\sqrt{\left((x_{t+i}-\mu_{t+i})^2 + r^2_{(t,i-1)}\right)\left(\alpha^2+\beta^2_{t+i}\right)}\right)}{\left(\frac{\alpha^2+\beta^2_{t+i}}{(x_{t+i}-\mu_{t+i})^2+r^2_{(t,i-1)}}\right)^{(\lambda-i/2-1/2)/2}}$$

$$\times \exp\left\{\beta_{t+i}\left(x_{t+i}-\mu_{t+i}\right)\right\},$$

esto implica que,

$$p\left(x_{t+i}|\, x^{(t,i-1)}\right) = \frac{\left(\frac{\alpha^2}{r^2_{(t,i-1)}}\right)^{(\lambda-i/2)/2}}{\left(\frac{\alpha^2+\beta^2_{t+i}}{(x_{t+i}-\mu_{t+i})^2+r^2_{(t,i-1)}}\right)^{(\lambda-i/2-1/2)/2}} \frac{K_{\lambda-i/2-1/2}\left(\sqrt{\left((x_{t+i}-\mu_{t+i})^2 + r^2_{(t,i-1)}\right)\left(\alpha^2+\beta^2_{t+i}\right)}\right)}{\sqrt{2\,\pi}\,K_{\lambda-i/2}\left(\sqrt{r^2_{(t,i-1)}\,\alpha^2}\right)}$$

$$\times \exp\left\{\beta_{t+i}\left(x_{t+i}-\mu_{t+i}\right)\right\}$$

$$= \left(\frac{\alpha^2}{r^2_{(t,i-1)}}\right)^{\lambda/2-i/4} \left(\frac{\alpha^2+\beta^2_{t+i}}{(x_{t+i}-\mu_{t+i})^2 + r^2_{(t,i-1)}}\right)^{-\lambda/2+i/4+1/4} \exp\left\{\beta_{t+i}\left(x_{t+i}-\mu_{t+i}\right)\right\}$$

$$\times \frac{K_{\lambda-i/2-1/2}\left(\sqrt{\left((x_{t+i}-\mu_{t+i})^2 + r^2_{(t,i-1)}\right)\left(\alpha^2+\beta^2_{t+i}\right)}\right)}{\sqrt{2\,\pi}\,K_{\lambda-i/2}\left(\sqrt{r^2_{(t,i-1)}\,\alpha^2}\right)},$$

por otra parte si $\rho^2 = (x_{t+i} - \mu_{t+i})^2$ entonces,

$$\left(\frac{\alpha^2}{r^2_{(t,i-1)}}\right)^{\lambda/2-i/4}\left(\frac{\alpha^2+\beta^2_{t+i}}{\rho^2+r^2_{(t,i-1)}}\right)^{-\lambda/2+i/4+1/4} = \left(\frac{\alpha^2}{r^2_{(t,i-1)}}\right)^{(\lambda-i/2)/2}\left(\frac{1}{\alpha^2+\beta^2_{t+i}}\right)^{(\lambda-i/2-1/2)/2}$$

$$\times \left\{\rho^2 + r^2_{(t,i-1)}\right\}^{(\lambda-i/2-1/2)/2}$$

$$= \frac{(\alpha^2)^{(\lambda-i/2)/2}}{(r_{(t,i-1)})^{\lambda-i/2}} \frac{1}{\left(\sqrt{\alpha^2+\beta^2_{t+i}}\right)^{(\lambda-i/2)-1/2}}$$

$$\times \left\{\rho^2 + r^2_{(t,i-1)}\right\}^{(\lambda-i/2-1/2)/2}$$

$$= \frac{(\alpha^2)^{(\lambda-i/2)/2}}{\left(\sqrt{\alpha^2+\beta^2_{t+i}}\right)^{(\lambda-i/2)-1/2}\ (r_{(t,i-1)})^{\lambda-i/2}}$$

$$\times \left\{\rho^2 + r^2_{(t,i-1)}\right\}^{(\lambda-i/2-1/2)/2}$$

Por lo tanto,

$$p\left(x_{t+i}|\, x^{(t,i-1)}\right) = a\left(\lambda - i/2, \sqrt{\alpha^2+\beta^2_{t+i}}, \beta_{t+i}, r_{(t,i-1)}, \mu_{t+i}\right) \tag{7.20}$$

$$\times \left\{(x_{t+i}-\mu_{t+i})^2 + (x^{(t,i-1)}-M)^T\,\Delta^{-1}\,(x^{(t,i-1)}-M) + \delta^2\right\}^{(\lambda-i/2-1/2)/2}$$

$$\times K_{(\lambda-i/2)-1/2}\left(\sqrt{\left((x_{t+i}-\mu_{t+i})^2 + r^2_{(t,i-1)}\right)\left(\alpha^2+\beta^2_{t+i}\right)}\right)\exp\left\{\beta_{t+i}\,(x_{t+i}-\mu_{t+i})\right\},$$

es decir,

$$X_{t+i}|\, X^{(t,i-1)} \sim \mathrm{GH}\left(x_{t+i};\, \lambda - \frac{i}{2},\, \sqrt{\alpha^2+\beta^2_{t+i}},\, \beta_{t+i},\, r_{(t,i-1)},\, \mu_{t+i}\right)$$

para $i = 1, \ldots, p$; si $\Delta = I$, $B = (\beta, \beta, \ldots, \beta)^T$, $\beta_{t+i} = \beta$, $M = (\mu, \mu, \ldots, \mu)^T$ y $\mu_{t+i} = \mu$, para $i = 1, \ldots, p$.

El modelo resultante se conoce como *Modelo Hiperbólico Generalizado Estacionario ARCH(p)*, o de manera resumida como *GH-ARCH(p)*. Este modelo también se puede ver como la versión estacionaria del modelo propuesto por Barndorff-Nielsen (1997) donde

$$r\left(X^{(t,i)};\delta,\mu\right)=\left(\delta^2+\sum_{j=0}^{i}(X_{t+j}-\mu)^2\right)^{1/2}.$$

En este caso, el modelo GH-ARCH(p) no tiene una representación estocástica como en (7.10) ya que considera el comportamiento de los valores $x_1,\ldots,x_p$. Sin embargo, después de considerar estos valores, el modelo puede ser escrito en la siguiente forma

$$X_{t+p}=\mu+\sqrt{\delta^2+\sum_{j=1}^{p}(X_{t+p-j}-\mu)^2}\,\varepsilon_{t+p},\quad \varepsilon_{t+p}\sim \mathrm{GH}\left(\lambda-\frac{p}{2},\sqrt{\alpha^2+\beta^2},\beta,1,0\right). \tag{7.21}$$

No obstante, para cualquier elección de λ su manejo, tanto analítico como numérico, se dificulta.

7.4. Estimación del modelo estacionario GH-ARCH(p)

Dado que la estimación máximo verosímil de los parámetros en el modelo GH-ARCH(p) no es posible, al menos de manera analítica, será necesario el uso de métodos numéricos. En otras palabras, utilizando la descomposición (7.16), la estimación se puede obtener a través del algoritmo EM (Dempster, 1977).

Dada una muestra $x^{(n)}=(x_1,\cdots,x_n),\ n>p$, la log-verosimilitud para los datos aumentados está dada por

$$\mathscr{L}_c\left(\dot{\Theta}\right)=\log L_{x,z}\left(\dot{\Theta}\right)=\sum_{k=1}^{p}\sum_{t=1}^{n} z_{k\,t}\left[\log w_k+\log p\left(x_{t-k},x_t;\dot{\theta}\right)\right], \tag{7.22}$$

En este caso, el Paso-E implica el cálculo de la esperanza de (7.22) con respecto a la distribución condicional de $Z|\,X$, lo que se reduce a

$$\tau_k\left(x_t,\dot{\Theta}^{(m)}\right):=\mathbb{E}\left[Z_{k\,t}|\,x^{(m)}\right]=\frac{w_k^{(m)}\,p\left(x_{t-k},x_t;\theta^{(m)}\right)}{f\left(x_t|\,x^{[t,p]};\theta^{(m)}\right)}.$$

El Paso-M implica máximizar

$$\dot{\Theta}^{(m+1)}=\max_{\dot{\Theta}} Q\left(\dot{\Theta}|\,\dot{\Theta}^{(m)}\right) \tag{7.23}$$

donde

$$Q\left(\dot{\Theta}|\,\dot{\Theta}^{(m)}\right)=\sum_{k=1}^{p}\sum_{t=1}^{n}\tau_k\left(x_t,\dot{\Theta}^{(m)}\right)\left[\log w_k+\log p\left(x_t-k;\dot{\theta}\right)\right]. \tag{7.24}$$

Los pesos se actualizan a través de

$$w_k^{(m+1)} = \frac{\sum_{i=1}^{n} \tau_k\,(\,x_t, \dot{\Theta}^{(m)}\,)}{n-k} \quad \text{para} \quad k = 1, \ldots, p. \tag{7.25}$$

Hasta el momento no hemos utilizado el hecho de que el componente de la transición en un paso también puede ser descompuesto a través de un vector latente $Y^{(n)} = (Y_2, \ldots, Y_n)$. En estos casos la log-verosimilitud de los datos completos esta dada por

$$\log L_{x,y,z}\,(\dot{\Theta}) = \log f_X\,(x_1; \dot{\theta}) + \sum_{k=1}^{p} \sum_{t=1}^{n} Z_{k\,t}\{\,\log w_k + \log\,[\,f_{X|\,Y}\,(x_t|\,y_t; \dot{\theta})\,f_{Y|\,X}(y_t|\,x_{t-k}; \dot{\theta})\,]\,\}. \tag{7.26}$$

Por lo tanto el algoritmo EM implica calcular la esperanza de (7.26) con respecto a $Y|\,X$. Para el modelo GH-ARCH(p), se tiene que

$$l_{\mathbf{x},\mathbf{y}}(\dot{\Theta}) = \log\,(\,\mathrm{GH}\,(x_1, \lambda, \alpha, \beta, \delta, \mu)) + \sum_{i=1}^{T-1} \log\,(\mathrm{N}\,(x_{i+1}; \mu + \beta y_{i+1}, y_{i+1})) \tag{7.27}$$

$$+ \sum_{i=1}^{p-1} \log\left(\mathrm{GIG}\left(y_{i+1}; \lambda - \frac{i}{2}, r^2_{(1,i-1)}, \alpha^2\right)\right)$$

$$+ \sum_{i=1}^{T-p} \log\left(\mathrm{GIG}\left(y_{i+p}; \lambda - \frac{i}{2}, r^2_{(i,p-1)}, \alpha^2\right)\right),$$

donde $l_{\mathbf{x},\mathbf{y}}(\dot{\Theta}) = \log L_{\mathbf{x},\mathbf{y}}(\dot{\Theta})$. Por lo general, este procedimiento requiere el cálculo del gradiente $\nabla\, l_\theta$, donde $l_\theta = \log L_{\mathbf{x}}(\theta)$. Para detalles sobre su cálculo consúltese Mena y Walker (2007b).

La principal dificultad con el algoritmo EM es obtener la esperanza, dado que la distribución $F^{\theta_{(j)}}_{\mathbf{Y}|\,\mathbf{X}}$ no tiene una forma simple. Sin embargo, en este caso, podemos construir la distribución de $F^{\theta_{(j)}}_{\mathbf{Y}|\,\mathbf{X}}$ teniendo en cuenta que

$$f\,(y_{t+i} \mid x_{t+i}, x^{(t,i-1)}) \propto f^{\theta}_{X|Y}\,(x_{t+i} \mid y_{t+i}) f_{Y|X}(y_{t+i} \mid x^{(t,i-1)}),$$

donde $x^{(t,i-1)} = (x_t, x_{t+i}, \ldots, x_{t+i-1})$. En este caso, podemos ver que

$$f\,(y_{t+i}|x_{t+i}, x^{(t,i-1)}) = \mathrm{GIG}\,\left(y_{t+i}; \lambda - \frac{i+1}{2}, r^2_{(t,i)}, \alpha^2 + \beta^2\right),$$

para $i, \ldots, T-1$. Por lo tanto, para un modelo GH-ARCH(p) estacionario podemos representar al vector

aleatorio $\{\mathbf{Y}|\,\mathbf{X}\}$ considerando las siguientes variables aleatorias condicionalmente independientes

$$Y_t \sim \text{GIG}\,(\lambda, \delta^2, \alpha^2 - \beta^2), \tag{7.28}$$

$$Y_{i+1}|\,x^{(1,i)} \sim \text{GIG}\,\left(\lambda - \frac{i+1}{2}, r^2_{(1,i)}, \alpha^2 + \beta^2\right), \quad \text{para} \quad i = 1, \ldots, p-1,$$

$$Y_{i+p}|\,x^{(i,p)} \sim \text{GIG}\,\left(\lambda - \frac{i+1}{2}, r^2_{(i,p)}, \alpha^2 + \beta^2\right), \quad \text{para} \quad i = 1, \ldots, T-p.$$

Utilizando está descomposición, es posible calcular la densidad $f^{\theta_{(j)}}_{\mathbf{Y}|\mathbf{X}}$ para la distribución requerida simplemente multiplicando las densidades correspondientes a las variables anteriores.

En Walker (1996), se demostro que el Paso-M involucrado en el algoritmo EM se puede simplificar de la siguiente manera; si $\theta^{-i}_{(j)} = (\theta^1_{(j)}, \ldots, \theta^{i-1}_{(j)}, \theta^{i+1}_{(j)}, \ldots, \theta^d_{(j)})$, entonces el Paso-M se puede simplificar al resolver

$$\frac{\partial\, Q\,(\theta^i|\,\theta^{-i}_{(j)}, \mathbf{x})}{\partial\, \theta^i} = \mathbb{E}_{\theta_{(j)}}\left[\frac{\partial\, l^{aug}_{\mathbf{x},\mathbf{y}}\,(\theta^i)}{\partial\, \theta^i}\right]\Bigg|_{\theta^i = \theta^i_{(j+1)}} = 0, \tag{7.29}$$

para $i = 1, \ldots, d$ y donde la esperanza $\mathbb{E}_{\theta_{(j)}}[\cdot]$ se toma con respecto a $F^{\theta_{(j)}}_{\mathbf{Y}|\,\mathbf{X}}$. La esperanza en (7.29) se puede obtener a través de una simulación de Monte Carlo de $\{\mathbf{Y}|\,\mathbf{X}\}$. En este caso, tal esperanza tiene una forma analítica si se utiliza la descomposición (7.28).

En Mena y Walker (2007b) se propone una forma alternativa para estimar los parámetros del modelo GH-ARCH(p) estacionario. El procedimiento es el siguiente: Dado un conjunto inicial de valores $\theta_0 = (\lambda_0, \alpha_0, \beta_0, \delta_0, \mu_0)$, la primera actualización (primera iteración del algoritmo EM) está dada al resolver individualmente

$$\frac{\partial\, l_{\mathbf{x}}\,(\,\lambda, \alpha_0, \beta_0, \delta_0, \mu_0\,)}{\partial\, \lambda}\Bigg|_{\lambda=\lambda_1} = 0, \qquad \frac{\partial\, l_{\mathbf{x}}\,(\,\lambda_1, \alpha, \beta_0, \delta_0, \mu_0\,)}{\partial\, \alpha}\Bigg|_{\alpha=\alpha_1} = 0 \tag{7.30}$$

$$\frac{\partial\, l_{\mathbf{x}}\,(\,\lambda_1, \alpha_1, \beta, \delta_0, \mu_0\,)}{\partial\, \beta}\Bigg|_{\beta=\beta_1} = 0, \qquad \frac{\partial\, l_{\mathbf{x}}\,(\,\lambda_1, \alpha_1, \beta, \delta, \mu_0\,)}{\partial\, \delta}\Bigg|_{\delta=\delta_1} = 0$$

$$\frac{\partial\, l_{\mathbf{x}}\,(\,\lambda_1, \alpha_1, \beta_1, \delta_1, \mu\,)}{\partial\, \mu}\Bigg|_{\mu=\mu_1} = 0,$$

Por lo tanto, utilizamos θ_1 para obtener θ_2 y así sucesivamente, hasta la convergencia. Es decir, de manera general se está resolviendo

$$\mathbb{E}_{\theta_{(j)}}\left[\nabla l_{\mathbf{x},\mathbf{y}}(\theta)\right] = \nabla l_{\mathbf{x}}(\theta). \tag{7.31}$$

Es deseable que la Ecuación (7.30) tenga una solución analítica. En general, para todo el dominio del conjunto de parámetros, esto no esposible. Sin embargo, la solución numérica de las ecuaciones anteriores es más sencilla, en términos de tiempo y de eficiencia, en contraste con el procedimiento de maximización para la verosimilitud completa. Notemos que en general el enfoque MLE anterior no garantiza la convergencia al valor óptimo. Sin embargo, en este caso, Mena y Walker (2007b) utilizan el algoritmo EM para garantizar tal convergencia.

Ejemplo 3. *Estimación de un modelo GH-ARCH(p).*

Para ilustrar el mecanismo de estimación anteriormente descrito consideremos las siguientes bases de datos financieras:

Nombre	**Símbolo**	**Desde**	**Hasta**	**QLB(30)**
JP MORGAN CHASE CO	JPM	10-Julio-08	10-Julio-09	259.987
CIT GROUP INC (DEL)	CIT	10-Julio-08	10-Julio-09	255.461
AMER INTL GROUP NEW	AIG	10-Julio-08	10-Julio-09	265.700
INTL BUS MACHINE	IBM	10-Julio-08	10-Julio-09	465.730

Cuadro 7.1: Bases de datos a considerar en el análisis. El QLB para la prueba de Ljung-Box se realizo con 30 lags. El valor critico al 5 % es de 43.773.

Las series de la Tabla 7.1 se muestran en la Figura 7.1 y el ACF de los log-rendimientos al cuadrado de cada una de estas series se expone en la Figura 7.2. Debido a que en cada una de estas gráficas se observa una tendencia creciente y decreciente en distintos periodos de tiempo y considerando el resultado del estadístico QLB para la prueba de Ljung-Box podemos concluir que el proceso $\{Z_t\}$ depende del tiempo, es decir no es estacionario.

Siendo, como se ha explicado, la estacionariedad una propiedad necesaria para el análisis estadístico, se propone considera los log-rendimientos, que es una caso especial de la transformación de Box-Jenkins (cuando $d = 0$). Es decir, si $\{Z_t\}$ es el proceso original, entonces,

$$Y_t = \ln\left(\frac{Z_t}{Z_{t-1}}\right) = \ln(Z_t) - \ln(Z_{t-1}) \quad \text{con} \quad t = 1, \ldots, 252. \tag{7.32}$$

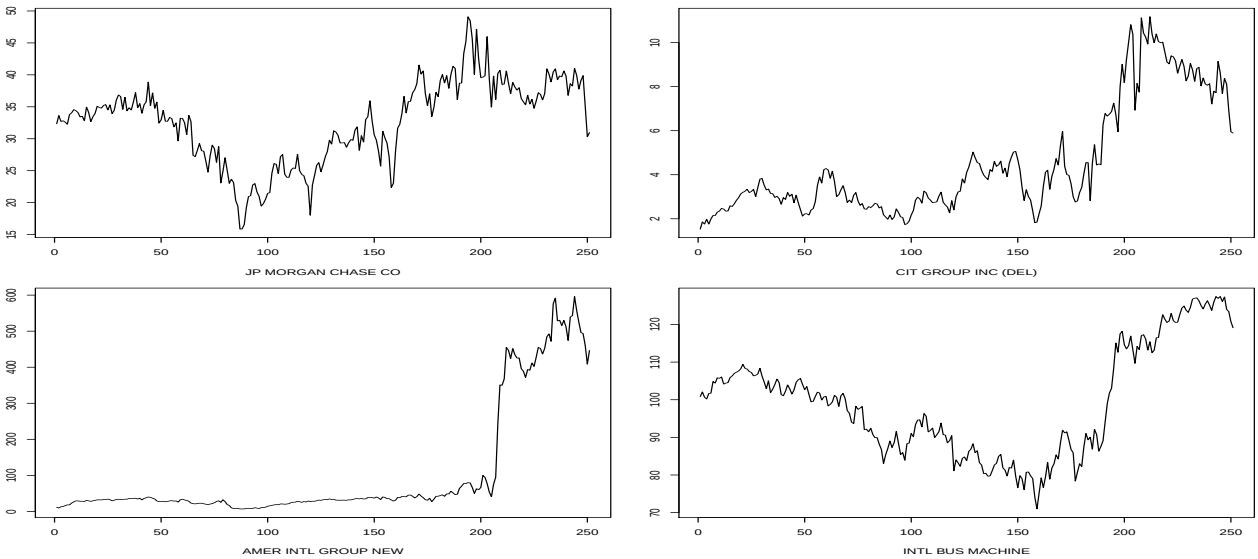

Figura 7.1: Precio diario de las acciones emitidas por JP MORGAN CHASE CO, CIT GROUP INC (DEL), AMER INTL GROUP NEW e INTL BUS MACHINE del 10 de julio de 2008 hasta el 10 de julio de 2009.

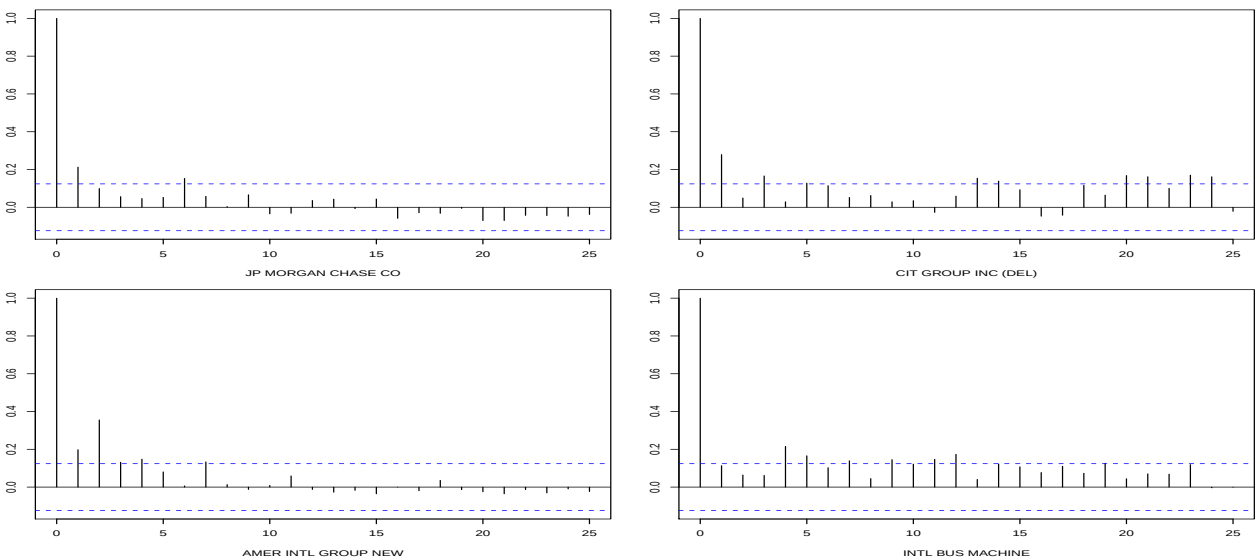

Figura 7.2: ACF correspondiente al cuadrado de los log-rendimientos para las series de la Figura 7.1.

Una de las ventajas de esta modificación es que se pueden observar los cambios relativos en las variables y comparar directamente con otras variables cuyos valores pueden ser muy diferentes de los valores originales. Además esta transformación es frecuentemente utilizada en procesos con tendencia exponencial creciente y/o decreciente.

Entonces comenzamos modelando la serie $\{Y_t\}$ a partir de un modelo GH-ARCH(p)[1]. Para estimar los parámetros de la probabilidad de transición se utilizará el algoritmo EM descrito en la Sección 7.4. El Paso-E se calculará a partir de la log-verosimilitud aumentada (7.27) de los datos transformados $\{Y_t\}$. Empero, dado que la maximización en términos analíticos de la expresión obtenida en el Paso-E resulta muy complicada, el Paso-M se realizará a través del algoritmo de Broyden-Fletcher-Goldfarb-Shanno (BFGS). En Press et. al (2007) este algoritmo se describe de la siguiente manera: Consideramos al sistema coordenado $\mathbf{X}$ con $\mathbf{P}$ el punto de origen del sistema. Como toda función f puede ser aproximada mediante una serie de Taylor, entonces

$$\begin{aligned} f(\mathbf{x}) &= f(\mathbf{P}) + \sum_i \frac{\partial f}{\partial x_i} + \frac{1}{2} \sum_{i,j} \frac{\partial^2 f}{\partial x_i - \partial x_j} x_i x_j + \cdots \\ &\approx c - \mathbf{b} \cdot \mathbf{x} + \frac{1}{2} \mathbf{x} \cdot \mathbf{A} \cdot \mathbf{x}, \end{aligned} \tag{7.33}$$

donde

$$c \equiv f(\mathbf{P}), \qquad \mathbf{b} \equiv -\nabla f|_{\mathbf{P}} \qquad \text{y} \qquad [\mathbf{A}]_{i\,j} \equiv \left. \frac{\partial^2 f}{\partial x_i \, \partial x_j} \right|_{\mathbf{P}}.$$

A la matriz $\mathbf{A}$, cuyos componentes son las segundas derivadas parciales de la función $f(x)$, se le conoce como el *Hessiano* de la función $f(\mathbf{x})$ evaluada en $\mathbf{P}$. El cálculo del gradiente en la aproximación (7.33) se hace de la siguiente manera

$$\nabla f = \mathbf{A} \cdot x - \mathbf{b}.$$

Sin embargo, en la mayoria de los casos no se tiene ninguna información sobre los parámetros de $\mathbf{A}$ y $\mathbf{b}$. Por lo que, la idea básica del algoritmo BFGS es estimar iterativamente a la matriz $\mathbf{A}^{-1}$ mediante una secuencia de matrices $\mathbf{H}_i$ tales que

$$\lim_{i \to \infty} \mathbf{H}_i = \mathbf{A}^{-1}.$$

Para localizar un cero en la función gradiente consideremos la posibilidad de encontrar un máximo a través del método de Newton, cerca del punto actual x_i mediante la aproximación de la función por un polinomio de segundo orden,

$$f(x) = f(x_i) + (x - x_i) \cdot \nabla f(x_i) + \frac{1}{2} (x - x_i) \cdot \mathbf{A} \cdot (x - x_i)$$

[1] El algoritmo para determinar el orden óptimo del modelo se muestra en el Apéndice E.

donde

$$\nabla f(x) = \nabla f(x_i) + A \cdot (x - x_i).$$

Es decir, nos interesa conocer $\nabla f(x) = 0$ para determinar el próximo punto de la iteración:

$$x - x_i = -\mathbf{A}^{-1} \cdot \nabla f(x_i). \tag{7.34}$$

La razón por la que este método se considera como quasi-Newton es que se utiliza una estimación de la matriz $\mathbf{A}^{-1}$ en comparación con el método de Newton que emplea al Hessiano de f. La idea detrás de los métodos quasi-Newton es comenzar con una aproximación de la matriz $\mathbf{A}$ que sea simétrica y positiva definida (usualmente esta corresponde a la matriz identidad) y construir la aproximación de las $\mathbf{H}_i's$ de tal manera que la matriz $\mathbf{H}_i$ siga siendo positiva definida y simétrica. Entonces restando la ecuación (7.34) en x_{i+1} para la misma ecuación en x_i se tiene que

$$x_{i+1} - x_i = \mathbf{A}^{-1} \cdot (\nabla f_{i+1} - \nabla f_i) \tag{7.35}$$

donde $\nabla f_j = \nabla f(x_j)$. Habiendo hecho el paso de x_i a x_{i+1} la nueva aproximación $\mathbf{H}_{i+1}$ debe satisfacer (7.35), es decir,

$$x_{i+1} - x_i = \mathbf{H}_{i+1} \cdot (\nabla f_{i+1} - \nabla f_i).$$

El algoritmo BFGS consiste en la actualización de la matriz $\mathbf{H}_i$ mediante la siguiente formula

$$\begin{aligned}\mathbf{H}_{i+1} = \mathbf{H}_i &+ \frac{(\mathbf{x}_{i+1} - \mathbf{x}_i) \otimes (\mathbf{x}_{i+1} - \mathbf{x}_i)}{(\mathbf{x}_{i+1} - \mathbf{x}_i) \cdot (\nabla f_{i+1} - \nabla f_i)} - \frac{[\mathbf{H}_i \cdot (\nabla f_{i+1} - \nabla f_i)] \otimes [\mathbf{H}_i \cdot (\nabla f_{i+1} - \nabla f_i)]}{(\nabla f_{i+1} - \nabla f_i) \cdot \mathbf{H}_i \cdot (\nabla f_{i+1} - \nabla f_i)} \\ &+ [(\nabla f_{i+1} - \nabla f_i) \cdot \mathbf{H}_i \cdot (\nabla f_{i+1} - \nabla f_i)]\, \mathbf{u} \otimes \mathbf{u},\end{aligned}$$

donde $\otimes$ denota el producto externo o directo entre dos vectores y

$$\mathbf{u} \equiv \frac{(\mathbf{x}_{i+1} - \mathbf{x}_i)}{(\mathbf{x}_{i+1} - \mathbf{x}_i) \cdot (\nabla f_{i+1} - \nabla f_i)} - \frac{\mathbf{H}_i \cdot (\nabla f_{i+1} - \nabla f_i)}{(\nabla f_{i+1} - \nabla f_i) \cdot \mathbf{H}_i \cdot (\nabla f_{i+1} - \nabla f_i)}.$$

De esta manera, en la Tabla 7.2 se muestra la estimación de los parámetros λ, α, β, δ y μ para la probabilidad de transición (7.20). Además se presenta el criterio de Akaike descrito en la Sección 2.4. El orden óptimo para cada serie de datos se muestra en negritas. Por último, en la Figura 7.3 se exhibe el histograma de los log-rendimientos de las series de datos de la Tabla 7.1 junto con la distribución estacionaria ajustada para el precio diario de las acciones emitidas por JP MORGAN CHASE CO, CIT GROUP INC (DEL), AMER INTL GROUP NEW e INTL BUS MACHINE desde el 10 de julio de 2008 hasta el 10 de julio de 2009.

Orden	λ	α	β	δ	μ	$loglik$	AIC
JP MORGAN CHASE CO							
1	-0.4528	16.668	-1.0239	0.076982	0.0058976	300.700	-591.400
3	1.1299	22.832	-0.4574	0.022985	0.0032471	307.141	-604.281
6	1.0551	23.973	-0.5336	0.040359	0.0036055	308.926	**-607.852**
7	1.2850	25.608	-0.7604	0.039033	0.0046982	308.520	-607.039
9	2.1804	31.003	-0.6478	0.030817	0.0041501	305.839	-601.679
CIT GROUP INC (DEL)							
1	-0.89077	-7.620	0.54147	0.134500	0.000465	184.860	-359.720
5	0.95364	13.831	1.06760	0.070084	-0.006565	191.583	-373.166
9	0.70409	13.547	0.51355	0.094568	0.000736	191.848	**-373.696**
13	2.37270	19.393	0.25413	0.063937	0.004289	187.982	-365.964
17	0.74450	19.016	0.38371	0.185410	0.002539	183.526	-357.053
AMER INTL GROUP NEW							
1	-0.67494	3.1378	0.31469	0.078970	0.0097949	209.346	-408.692
2	-0.54126	4.0320	0.27281	0.078325	0.0107090	214.459	-418.917
3	-0.74768	3.0066	0.32428	0.089928	0.0096230	218.967	**-427.935**
4	-0.50400	4.1617	0.29220	0.086327	0.0102560	217.195	-424.391
5	-0.32062	4.8273	0.34915	0.082638	0.0092220	216.897	-423.794
INTL BUS MACHINE							
1	0.6172	-19.078	3.1824	0.057781	-0.018965	287.762	-565.524
2	1.2441	21.162	4.6290	0.025038	-0.027620	292.426	**-574.852**
3	1.9826	25.804	4.8208	0.017119	-0.028317	292.126	-574.252
4	2.9859	33.099	7.1754	0.045384	-0.041421	286.738	-563.476
5	6.1031	45.119	7.8076	0.013389	-0.045767	281.725	-553.451

Cuadro 7.2: Resultados de la estimación a través del algoritmo EM para un modelo GH-ARCH(p). El orden óptimo de cada modelo se presenta en negritas. El tiempo total, para la estimación y la simulación, nunca superó los 5 segundos. El error absoluto utilizado para la convergencia se estableció en 10^{-6}.

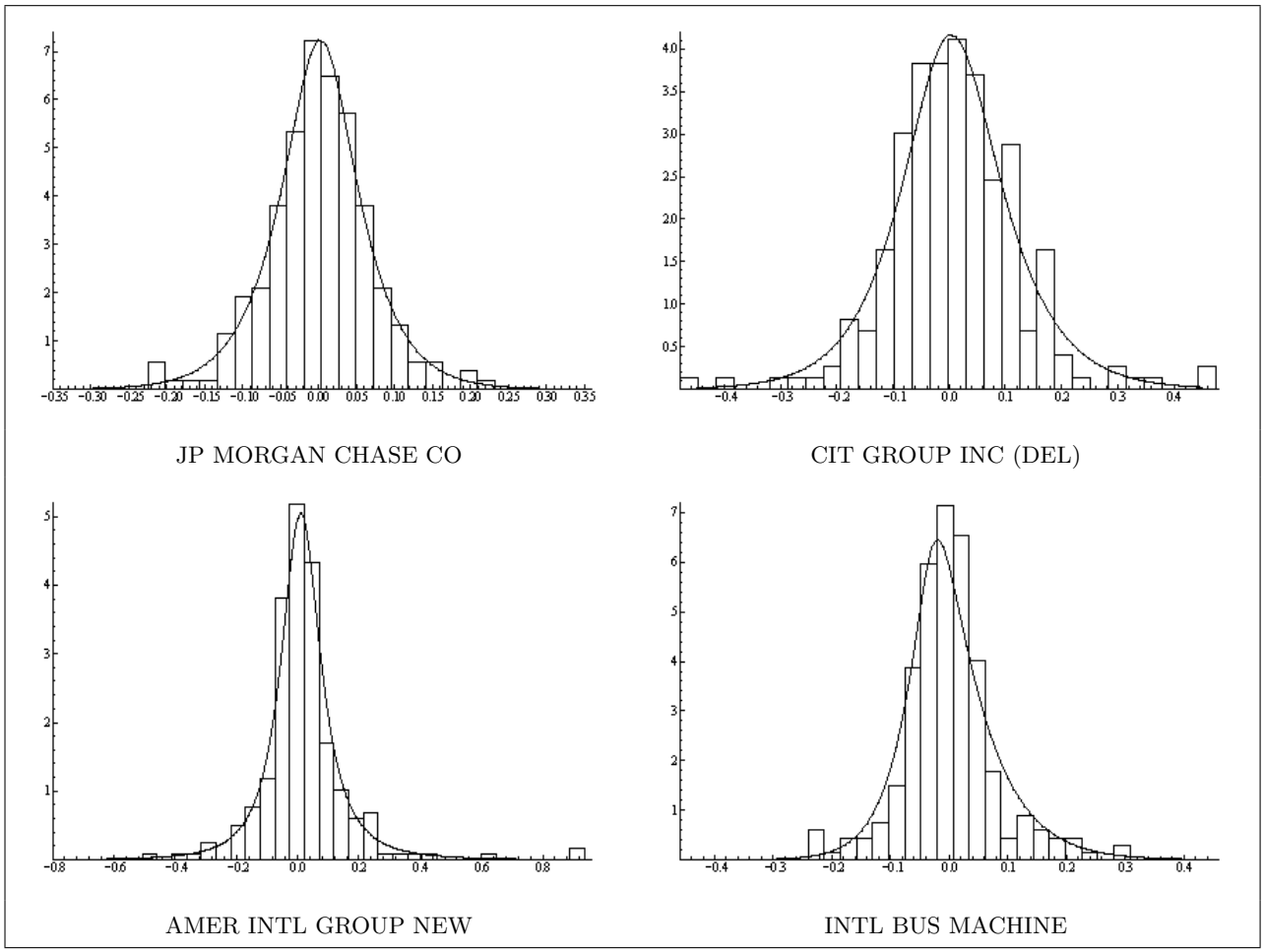

Figura 7.3: Comparación entre el histograma de las serie de datos de la Tabla 7.1 junto con la distribución estacionaria ajustada a partir de los valores que se muestran en la Tabla 7.2.

Por lo tanto se concluye que para el precio de las acciones emitidas diaramente del 10 de julio de 2008 hasta el 10 de julio de 2009 por JP MORGAN CHASE CO, CIT GROUP INC (DEL), AMER INTL GROUP NEW e INTL BUS MACHINE se les va a ajustar un modelo GH-ARCH(6), GH-ARCH(9), GH-ARCH(3) y GH-ARCH(2) respectivamente cuya probabilidad de transición estará dado por la Ecuación (7.20). Además, se puede observar que en la Figura 7.3 las distribuciones ajustadas tiene una alta concentración alrededor de la media, es decir se captura satisfactoriamente la lepcurtosis, lo cual es un buen referente para la elección de dicho modelo.

Conclusiones

La construcción de los modelos de series de tiempo vía modelos de mezclas finitas implica la formulación de distintas técnicas estadísticas. Por ejemplo, la introducción de variables latentes a partir de un muestreo de Gibbs, el uso de modelos de mezclas finitas para la generalización del modelo anterior y el empleo de modelos que consideran una estructura de dependencia más compleja como, por ejemplo, la agrupación de la volatilidad. El estudio de esta característica dio como resultado el modelo ARCH planteado por Engle (1982).

Empero, algunas propiedades, como la estacionariedad, la asimétria o la existencia de colas pesadas en la distribución de los datos no se considerán en este modelo. Por esta razón se han planteado distintos modelos que permitan abordar dichas propiedades. En el caso específico del modelo GH-ARCH(p) estrictamente estacionario planteado en el presente trabajo cada uno de estos enfoques se conjugan para dar origen a un modelo que permite ajustar series cuya estructura de dependencia es no lineal y en donde la variabilidad de los datos juega un papel fundamental.

Otra de las virtudes del modelo GH-ARCH estrictamente estacionario es el conocimiento analítico de su probabilidad de transición. Esto permite, además de conocer sus momentos, plantear el uso de cualquier estadístico de forma cerrada sin el uso de aproximaciones. Sin embargo, la forma de esta distribución tiene sus limitantes. Al estimar los parámetros o lo pesos de una mezcla, el cálculo de la esperanza condicional o la máximización en el Paso-M se complica al punto en donde se tienen que utilizar métodos númericos. No obstante, en comparación con el modelo planteado por Jensen y Lunde (2001), en donde se plantea un método de estimación más eficiente, el modelo GH-ARCH es estrictamente estacionario. Empero, como se vio anteriormente, los parámetros obtenidos son consistentes con el modelo planteado. A pesar de estas limitantes, todavia no existe un modelo que sea capaz de modelar de manera completa cualquier serie de tiempo, por lo que el uso de estos modelos se ha popularizado sobre todo en aplicaciones financieras.

Apéndice A

Distribuciones Comunes

[1] **A.1. Distribución Normal**, $\mathrm{N}_p(\theta, \Sigma)$

$$f(x|\theta\Sigma) = (\det\Sigma)^{-1/2}(2\pi)^{-p/2} \exp\left\{-\frac{1}{2}(x-\theta)'\Sigma^{-1}(x-\theta)\right\},$$

donde $\theta \in \mathbb{R}^p$ y Σ es una matriz simétrica y definida positiva de dimensión $p \times p$.

$$\mathbb{E}_{\theta,\Sigma}[x] = \theta,$$
$$\mathbb{E}[(x-\theta)(x-\theta)'] = \Sigma.$$

Cuando Σ no esta definida, la distribución $\mathrm{N}_p(\theta, \Sigma)$ no tiene definida una densidad con respecto a la medida de Lebesgue en $\mathbb{R}^p$. Para $p = 1$, la distribución *Log-normal* se define como la distribución de e^x cuando $x \sim \mathrm{N}(\theta, \sigma^2)$.

A.2. Distribución Gamma, $\mathrm{Ga}(\alpha, \beta)$

$$f(x|\alpha, \beta) = \frac{\beta^\alpha}{\Gamma(\alpha)}\ x^{\alpha-1} \exp\{-\beta x\}\ \mathbb{I}_{[0,\infty)}(x),$$

donde $\alpha, \beta > 0$.

$$\mathbb{E}_{\alpha,\beta}[x] = \frac{\alpha}{\beta},$$
$$\mathbb{V}_{\alpha,\beta}[x] = \frac{\alpha}{\beta^2}.$$

[1] Tomado de: Christian P. Robert, The Bayesian Choice, a decision-theoretict motivation, Springer-Verlag, 1994, p 381-385.

Los casos particulares de la distribución gamma son, la *distribución Erlang*, $\mathrm{Ga}(\alpha, 1)$, la *distribución exponencial*, $\mathrm{Ga}(1, \beta)$ (denotada como $\exp(\beta)$), y la *distribución Ji-cuadrada*, $\mathrm{Ga}(\nu/2, 1/2)$, (denotada como χ^2_ν).

A.3. Distribución Beta, $\mathrm{Be}(\alpha, \beta)$

$$f(x|\alpha, \beta) = \frac{x^{\alpha-1}}{B(\alpha, \beta)}(1-x)^{\beta-1}\ \mathbb{I}_{[0,1]}(x),$$

donde

$$B(\alpha, \beta) = \frac{\Gamma(\alpha)\Gamma(\beta)}{\Gamma(\alpha+\beta)}, \qquad \alpha, \beta > 0.$$

$$\mathbb{E}_{\alpha,\beta}[x] = \frac{\alpha}{\alpha+\beta},$$
$$\mathbb{V}_{\alpha,\beta}[x] = \frac{\alpha\beta}{(\alpha+\beta)^2(\alpha+\beta+1)}.$$

La distribución Beta se pude obtener como la distribución de $y_1/(y_1+y_2)$ donde $y_1 \sim \mathrm{Ga}(\alpha, 1)$ y $y_2 \sim \mathrm{Ga}(1, \beta)$.

A.4. Distribución Sudent-t, $\mathrm{St}_p(\nu, \theta, \Sigma)$

$$f(x|\nu, \theta, \Sigma) = \frac{\Gamma((\nu+p)/2)\Gamma(\nu/2)}{(\det\,\Sigma)^{1/2}(\nu\pi)^{p/2}}\left[1 + \frac{(x-\theta)'\Sigma^{-1}(x-\theta)}{\nu}\right]^{-(\nu+p)/2}$$

donde $\nu > 0$, $\theta \in \mathbb{R}^p$, y Σ es una matriz simétrica definida positiva de dimensión $p \times p$.

$$\mathbb{E}_{\nu,\theta\Sigma}[x] = \theta \qquad (\nu > 0),$$
$$\mathbb{E}_{\theta,\Sigma}[(x-\theta)(x-\theta)'] = \frac{\nu\Sigma}{\nu-2} \qquad (\nu > 2).$$

Cuando $p = 1$, un caso particular de la distribución Student-t de la *distribución Cauchy*, $\mathcal{C}(\theta, \sigma^2)$, la cual corresponde al caso en donde $\nu = 1$. La distribución Student-t puede derivarse de la distribución de x/z cuando $x \sim \mathrm{N}_p(\theta, \Sigma)$ y $\nu z^2 \sim \chi^2_\nu$.

A.5. Distribución F de Fisher, $\mathrm{F}(\nu, \rho)$

$$f(x|\nu, \rho) = \frac{\Gamma((\nu+\rho)/2)\nu^{\rho/2}\rho^{\nu/2}}{\Gamma(\nu/2)\Gamma(\rho/2)}\ \frac{x^{(\nu-2)/2}}{(\nu+\rho x)^{(\nu+\rho/2}}\ \mathbb{I}^{(x)}_{[0,\infty)},$$

donde $\nu, \rho > 0$.

$$\mathbb{E}_{\nu,\rho}[x] = \frac{\rho}{\rho - 2},$$
$$\mathbb{V}_{\nu,\rho}[x] = \frac{2\rho^2(\nu + \rho - 2)}{\nu(\rho - 4)(\rho - 2)^2}, \qquad \rho > 4.$$

La distribución $F(p, q)$ es también la distribución de $(x - \theta)'\Sigma^{-1}(x - \theta)/p$ cuando $x \sim \mathrm{St}_p(q, \theta, \Sigma)$. Sin embargo, si $x \sim F(\nu, \rho)$, $\rho x/(\nu + \rho x) \sim \mathrm{Be}(\nu, \rho)$.

A.6. Distribución Gamma Inversa, $\mathrm{Iga}(\alpha, \beta)$

$$f(x|\alpha, \beta) = \frac{\beta^\alpha}{\Gamma(\alpha)} \frac{\exp\{-\beta/x\}}{x^{\alpha+1}} \mathbb{I}^{(x)}_{[\,0,\infty)},$$

donde $\alpha, \beta > 0$.

$$\mathbb{E}_{\alpha,\beta}[x] = \frac{\alpha}{\beta},$$
$$\mathbb{V}_{\alpha,\beta}[x] = \frac{\alpha}{\beta^2}.$$

Esta distribución se obtiene a partir de la distribución de x^{-1} cuando $x \sim \mathrm{Ga}(\alpha, \beta)$.

A.7. Distribución Ji-cuadrada no Centrada, $\chi^2_\nu(\lambda)$

$$f(x|\lambda) = \frac{1}{2}(x/\lambda)^{(p-2)/4} I_{(p-2)/2}(\sqrt{\lambda x}) \exp\{-(\lambda + x)/2\},$$

donde $\lambda \geq 0$.

$$\mathbb{E}_\lambda[x] = p + \lambda,$$
$$\mathbb{V}_\lambda[x] = 3p + 4\lambda.$$

Esta distribución se puede derivar como la distribución de $x_1^2 + \cdots + x_p^2$ donde $x_i \sim \mathrm{N}(\theta_i, 1)$ y $\theta_1^2 + \cdots + \theta_p^2 = \lambda$.

A.8. Distribución Dirichlet, $D_k(\alpha_1, \ldots, \alpha_k)$

$$f(x|\alpha_1, \ldots, \alpha_k) = \frac{\Gamma(\alpha_0)}{\Gamma(\alpha_1) \cdots \Gamma(\alpha_k)} x_1^{\alpha_1 - 1} \cdots x_k^{\alpha_k - 1} \mathbb{I}_{\{\sum x_i = 1\}},$$

donde $\alpha_1, \ldots, \alpha_k > 0$ y $\alpha_0 = \alpha_1 + \cdots + \alpha_k$.

$$\begin{aligned}
\mathbb{E}_\alpha[x_i] &= \frac{\alpha_i}{\alpha_0}, \\
\mathbb{V}_[x_i] &= \frac{(\alpha_0 - \alpha_i)\alpha_i}{\alpha_0^2(\alpha_0 + 1)}, \\
\mathrm{Cov}[x_i, x_j] &= -\frac{\alpha_i \alpha_j}{\alpha_0^2(\alpha_0 + 1)}, \quad \text{para} \quad i \neq j.
\end{aligned}$$

Como un caso particular, notemos que $(x, 1-x) \sim \mathrm{D}_2(\alpha_1, \alpha_2)$ es equivalente a que $x \sim \mathrm{Be}(\alpha_1, \alpha_2)$.

A.9. Distribución Pareto, $\mathrm{Pa}(\alpha, x_0)$

$$f(x|\alpha, x_0) = \alpha \frac{x_0^\alpha}{x^{\alpha+1}} \; \mathbb{I}^{(x)}_{[x_0, \infty)},$$

donde $\alpha > 0$ y $x_0 > 0$.

$$\begin{aligned}
\mathbb{E}_{\alpha, x_0}[x] &= \frac{\alpha x_0}{\alpha - 1}, \qquad \alpha > 1, \\
\mathbb{V}_{\alpha, x_0}[x] &= \frac{\alpha x_0^2}{(\alpha - 1)^2(\alpha - 2)} \qquad \alpha > 2.
\end{aligned}$$

A.10. Distribución Binomial, $\mathrm{Bi}(n, p)$

$$f(x|p) = \binom{n}{x} p^x (1-p)^{n-x} \; \mathbb{I}^{(x)}_{\{0, \ldots, n\}},$$

donde $0 \leq p \leq 1$.

$$\begin{aligned}
\mathbb{E}_p[x] &= np, \\
\mathbb{V}[x] &= np(1-p).
\end{aligned}$$

A.11. Distribución Multinomial, $\mathrm{M}_k(n; p_1, \ldots, p_k)$

$$f(x_1, \ldots, x_k | p_1, \ldots, p_k) = \binom{n}{x_1 \cdots x_k} \prod_{i=1}^{k} p_i^{x_i} \; \mathbb{I}_{\sum x_i = n},$$

donde $p_i \geq 0$ $(1 \leq i \leq k)$ y $\sum_i p_i = 1$.

$$\begin{aligned}
\mathbb{E}_p[x_i] &= np_i, \\
\mathbb{V}[x_i] &= np_i(1-p_i), \\
\mathrm{Cov}[x_i, x_j] &= -np_i p_j \quad \text{para} \quad i \neq j.
\end{aligned}$$

Notemos que si $x \sim \mathrm{M}_k(n; p_1, ldots, p_k)$, $x_i \sim \mathrm{Bi}(n, p_i)$, y entonces la distribución binomial corresponde al caso en donde $(x, n-x) \sim \mathrm{M}_2(n; p, 1-p)$.

A.12. Distribución Poisson, $\mathrm{Poi}(\lambda)$

$$f(x|\lambda) = \exp\{-\lambda\}\frac{\lambda^x}{x!}\ \mathbb{I}_{\mathbb{N}}^{(x)},$$

donde $\lambda > 0$.

$$\mathbb{E}_\lambda[x] = \lambda,$$
$$\mathbb{V}_\lambda[x] = \lambda.$$

A.13. Distribución Binomial Negativa, $\mathrm{BN}(n, p)$

$$f(x|p) = \binom{x}{n+x+1}\ p^x(1-p)^x\ \mathbb{I}_{\mathbb{N}}^{(x)},$$

donde $0 \leq p \leq 1$.

$$\mathbb{E}_p[x] = n\frac{1-p}{p},$$
$$\mathbb{V}_p[x] = n\frac{1-p}{p^2}.$$

A.14. Distribución Hipergeométrica, $\mathrm{Hyp}(N; n; p)$

$$f(x|p) = \frac{\binom{pn}{x}\binom{(1-p)N}{n-x}}{\binom{N}{n}}\ \mathbb{I}_{\{n-(1-p)N,\ldots,pN\}}^{(x)}\ \mathbb{I}_{\{0,1,\ldots,n\}}^{(x)},$$

donde $0 \leq p \leq 1$, $n < N$ y $pN \in \mathbb{N}$.

$$\mathbb{E}_{N,n,p}[x] = np,$$
$$\mathbb{V}_{N,n,p}[x] = \frac{(N-n)np(1-p)}{N-1}.$$

Apéndice B

Mezcla de distribuciones normales

Densidad	$f(x)$
1. Gausiana	$N(0,1)$
2. Unimodal Sesgada	$\frac{1}{5}N(0,1) + \frac{1}{5}N(\frac{1}{2},(\frac{2}{3})^2) + \frac{3}{5}N(\frac{13}{15},(\frac{5}{9})^2)$
3. Muy Sesgada	$\sum_{i=0}^{7} \frac{1}{8}N\left(3\{(\frac{2}{3})^i - 1\},(\frac{2}{3})^{2i}\right)$
4. Unimodal con Curtosis	$\frac{2}{3}N(0,1) + \frac{1}{3}N(0,(\frac{1}{10})^2)$
5. Outlier	$\frac{1}{10}N(0,1) + \frac{9}{10}N(0,(\frac{1}{10})^2)$
6. Bimodal	$\frac{1}{2}N(-1,(\frac{2}{3})^2) + \frac{1}{2}N(1,(\frac{2}{3})^2)$
7. Bimodal Separada	$\frac{1}{2}N(-\frac{3}{2},(\frac{2}{3})^2) + \frac{1}{2}N(\frac{3}{2},(\frac{2}{3})^2)$
8. Bimodal Asimétrica	$\frac{3}{4}N(0,1) + \frac{1}{4}N(\frac{3}{2},(\frac{1}{2})^2)$
9. Trimodal	$\frac{9}{20}N(-\frac{6}{5},(\frac{3}{5})^2) + \frac{9}{20}N(\frac{6}{5},(\frac{3}{5})^2) + \frac{1}{10}N(0,(\frac{1}{4})^2)$
10. Multimodal	$\frac{1}{2}N(0,1) + \sum_{i=0}^{4} \frac{1}{10}N(\frac{i}{2} - 1,(\frac{1}{10})^2)$

Cuadro B.1: Tomado de Geoffrey McLachlan y David Peel (2000, Capítulo 1)

Apéndice C

Funciones de Bessel Modificadas

Propiedades de las Funciones de Bessel Modificadas	
Propiedades	Expansión asintotica cuando $x \downarrow 0$
1. $K_{-\nu}(x) = K_\nu(x)$	1. $K_\nu(x) \sim (1/2)\Gamma(\nu)((x/2))^{-\nu}, \quad \nu > 0$
2. $K_{1/2}(x) = \sqrt{\pi/2x}e^{-x}$	2. $K_\nu(x) \sim (1/2)\Gamma(-\nu)((x/2))^{\nu}, \quad \nu < 0$
3. $K_{\nu+\epsilon}(x) > K_\nu(x), \quad \nu, \epsilon, x > 0$	3. $K_0(x) \sim -\log(x)$
4. $K_{\nu+1}(x) = (2\nu/x)K_\nu(x) + K_{\nu-1}(x)$	
Representación en forma de integral	Expansión asintotica cuando $x \uparrow \infty$
$K_\nu = (1/2)\int_0^\infty y^{\nu-1}\exp(-(x-2)(y+y^{-1}))\ dy$	$K_\nu(x) \sim \sqrt{(\pi/2x)}e^{-x}$
Derivadas $\partial/\partial x$	
1. $K_0'(x) = -K_1(x)$	
2. $K_\nu'(x) = -(1/2)(K_{\nu+1}(x) + K_{\nu-1}(x))$	
3. $K_\nu'(x) = (\nu/x)K_\nu(x) - K_{\nu+1}(x)$	
4. $(\log K_\nu(x))' = \nu/x - R_\nu(x)$,	
donde $R_\nu(x) := \frac{K_{nu+1}(x)}{K_\nu(x)}, \quad x > 0$	

Cuadro C.1: Tomado de Abramowitz y Stegun (1992) (Capítulo 9) y Eberlein y Hammerstein (2004)

Apéndice D

Algoritmo en Ox para estimar los parámetros del modelo GH-ARCH(p).

A continuación se muestra la rutina[1] que se utilizo en el ejemplo 3 del Capítulo 7 para estimar el orden del modelo y los parámetros de la probabilidad de transición:

```
/*****************************************************************************************
*  Este programa estima los parámetros de un modelo GH-ARCH(p) estacionario para un       *
*  conjunto de datos. Los detalles del procedimiento se explican dentro del desarrollo de  *
*  las funciones.                                                                          *
*                                                                                          *
*  Para simular cualquier base de datos solo habra que incluir la base de datos a modelar  *
*  dentro del "main(){aqui}"del programa.                                                  *
*****************************************************************************************/

/*********************************  Inclusión de librerias  *********************************/

# include <oxstd.h>
# include <oxprob.h>
# include <oxdraw.h>
# include <oxfloat.h>
# import <maximize>

/*********************************  Declaración de variables globales  *********************/

const decl pi = M_PI;
```

[1]Este algoritmo es una adaptación del original realizado por el Dr. Ramsés Mena Chávez en su artículo: Mena, R. H. and Walker, S. G. (2007b). On the stationary version of the generalized hyperbolic ARCH model. Annals of the Institute of Statistical Mathematics, 59, 325-348.

```
decl y;        //Datos usados para la estimación del modelo GH-ARCH.
decl q;        //Lag.

/*********************************  Declaración de las funciones  ************************/

CtnsHist(const lugar, Y, const Nb);
GH(const x,const L,const a,const b,const d,const m);
TK(const p,const q, const t);
LW(const p);
LW2(const p, const AdFunc, const avScore, const amHess);
rangigK(const r, const c, const p, const l, const z);
ranghK(const r, const c, const p, const l, const z);
SIMGS(const p, const y0, const T,const sty,const vol,const z2, const w);
QLB(const w, const mlag);
est(const p,const P);
ESTSIM(const namedata, const M, const T, const Ns, const p, const lag, const w,
       const forecast,const sty);

/*********************************  Programa principal  **********************************/

main()
{
     ESTSIM("JPM.txt",230,20,20,<0,1,0,0.5,0>,6,0,1,1);
     ESTSIM("AIG.txt",230,20,20,<0,1,0,0.5,0>,3,0,1,1);
     ESTSIM("CIT.txt",230,20,20,<0,1,0,0.5,0>,9,0,1,1);
     ESTSIM("IBM.txt",230,20,20,<0,1,0,0.5,0>,2,0,1,1);
}

/*--------------------------------  Funciones y procedimientos  ----------------------------*/

/*********************************  Procedimiento Principal  *******************************/

/*-------------------------------------------------------------.
|  Dado un conjunto de datos y algunos parámetros, este        |
|  procedimiento reconoce cualquier conjunto de datos como un  |
|  conjunto de datos historicos, y además predice algun punto  |
|  o conjunto de datos que se desee.                           |
|  Es decir, si queremos comparar nuestras simulaciones con    |
|  datos reales durante el periodo de simulación, tenemos que  |
|  proporcionar los datos correspondientes a ese periodo de    |
|  tiempo. En otro caso, los datos no se compararán con el     |
|  periodo predicho.                                           |
|                                                              |
|  De esta manera, este procedimiento necesita las siguientes  |
|  variables:                                                  |
|                                                              |
|  in: namedata (nombre del archivo)                           |
|  M (Longitud del periodo historico para el análisis)         |
|  T (Periodo de predicción)                                   |
```

```
|  Ns (Número de simulaciones utilizadas en los estimadores MC)|
|  p (Valores iniciales para los parámetros)                   |
|  lag (lag del modelo GH-ARCH)                                 |
|  w (periodo de calentamiento utilizado en SIMGS)              |
|  frct (1/0 indica si el conjunto de datos esta en el periodo de|
|       predicción)                                             |
|  sty (1/0 indica si la hipotesis estacionaria es utilizada    |
|       o no en el modelo GH-ARCH)                              |
|  out: prints, plots etc.                                      |
.--------------------------------------------------------------*/
ESTSIM(const namedata, const M, const T, const Ns, const p, const lag, const w,
       const forecast,const sty);
{
decl i,tim;                                    //Contadores de bucles.
decl S,SM,Yt,YT,ST,YTreal,STE,STpathest,H,P,r;  //Variables de inicio.
decl Sdata,STdev,STIL,STUL,v,v1,vpathest,Y2t,Ct,K,ir;
tim=timer();
q=lag;
S=loadmat(namedata); r=rows(S);                // Datos de la serie, e.g. stock prices. (r,1).
 if(frct==1)
 {
  SM=S[r-M-T-1:r-T-1];                         // Datos historicos. (M+1,1).
  ST=S[r-T-1:r-1];                             // Datos historicos correspondientes al periodo
                                                  de predicción. (T+1,1).
  YTreal=log(ST[1:T])-log(ST[0:T-1]);          // Log-returns de ST. (T,1).
 }
 else { SM=S[r-M-1:r-1]; }
  Yt= log(SM[1:M])-log(SM[0:M-1]);             // Log-returns de SM. (M,1).
  y=Yt;                                        // Leyendo los datos historicos para la
                                                  variable global en la estimación.
  est(p,&P);                                   // Corriendo el procedimiento de estimación.

    YT=zeros(T,Ns);                            // Matriz inicial para asignar las
                                                  simulaciones Ns. (T,Ns).

    v=zeros(T,Ns);
    for(i=0;i<=Ns-1;i++)                       // Generando las simulaciones requeridas.
                                                  Estas son a partir de cuatro esquemas:
     {
     YT[][i]=SIMGS(P, y[M-q:M-1], T, sty,&v1,&Y2t,w); //hipótesis estacionaria, y simulando
                                                  directamente para la transción

     v[][i]=v1;
     }                                         // densidad (GH) SIMGH(..,1 ó 0), así como con
                                                  o sin hipótesis estacionaria.
    H=cumulate(YT);                            // Sumas de los log-returns para cualquier tiempo
```

```
                                                    1<t<T y para toda simulación. (T,Ns).
    STE=SM[M].*exp(H);                          // Observación simulada en cualquier momento 1<t<T
                                                    y para toda simulación. (T,Ns).
    STpathest=meanr(STE);                       // Promedio de las Ns simulaciones. (T,1).
    STIL=STpathest-STdev;
    STUL=STpathest+STdev;
    vpathest=meanr(v);

    Ct=zeros(T,1); ir=0.05; K=SM[M]+10;
    for(i=0;i<=T-1;i++)
     {
      Ct[i]=(exp(-ir*(i+1))/Ns)*sumr(max(((STE[i][]./STpathest[i]).*SM[M].*exp(ir*(i+1))-K)',0));
     }
/********************************  Mostrando la información y gráficando  ******************/

println("\n----------------------" ,namedata,"----------------------"\n );
println("\n----------------------"(orden=" ,q," simulaciones=" ,Ns,")---------------");
print("LogLik",LW(P));
print(\n"Valores de los parámetros para la distribución GH:\n","%12.8f",P',"\n);
print("QLB(30)=",QLB(y,30),"\n\n","AIC=",-2*LW(P)+10,"\n");
decl rrx, rry;
rrx=range(-0.8,-0.8,0.01); rry=GH(rrx,-0.92917429,1.34778502,1.27381378,0.09896981,-0.03939853);
DrawXMatrix(0, rry,{"TRUE"}, rrx,"x",0,1);
CtnsHist(0, Yt,30);
SetDraw(SET_ LEGENDHIDE,1);
SaveDrawWindow("Ajuste.eps"); CloseDrawWindow();
}
/********************************  Maximización  ****************************************/

/*-----------------------------------------------------------------.
|  Rutina para calcular el MLE.                                    |
|  in: p (valores iniciales para los prámetros);                   |
|      Pe (nombre de la variable asignada a las estimaciones)      |
|  out: Pe;                                                        |
.-----------------------------------------------------------------*/
est(const p,const Pe)
{
    decl lw, P;
    P=p';
     MaxControl(-1,50);                         //MaxControlEps(1e-2,5e-4);
     MaxBFGS(LW2, &P, &lw, 0, 1);
    Pe[0]=P;
    return 1;
}
/********************************  Procedimiento de simulación  ************************/
/*-----------------------------------------------------------------.
```

```
|  Procedimiento para generar los valores y1....yT bajo el     |
|  modelo GH-ARCH. Esta parte se hara de dos formas, se        |
|  simulará a partir de una representación NORMAL-GIG o        |
|  directamente a través de una distribución GH, en ambos casos|
|  la simulación comenzara utilizando los valores anteriores o |
|  las hipótesis estacionarias. Los valore latentes (vol), se  |
|  simulan implicitamente a través del primer procedimiento.   |
|  in: p (parámetros); y0 (valore iniciales past-lagged );     |
|      T (Tiempo);                                             |
|      sty (1/0 indica si los valores estacionarios pueden ser |
|           considerados);                                     |
|      vol (variable que asigna la variable latente v.)        |
|      z2  (variable que asigna y1...YT bajo el segundo        |
|           procedimiento);                                    |
|      w  (periodo de calentamiento, integer>0)                |
|  out: y1...yT; vol; z2;                                      |
.--------------------------------------------------------------*/
SIMGS(const p, const y0, const T,const sty,const vol,const z2, const w)
{
    decl t,x,z,zgh;
    x=zeros(T+w,1);
    z=zeros(T+w+1,1);
    zgh=zeros(T+w+1,1);
    if (sty == 1) {z[0]=rangh(1,1,p[0],p[3],sqrt(p[1]^2-p[2]^2), p[2])+p[4]; zgh[0]=z[0];)
                  for(t=1;t<=q-1;t++) { z[t]=ranghK(1,1,p,t,z[0:t-1]); zgh[t]=z[t];} }
    else { for(t=0;t<=q-1;t++){ z[t]=y0[t]; zgh[t]=z[t]; } }

    for (t=q; t<=T+w; t++)
    {
    x[t-1]=rangigK(1,1,p,q,z[t-q:t-1]);
    z[t]=p[4]+p[2]*x[t-1]+sqrt(x[t-1])*rann(1,1);
    zgh[t]=ranghK(1,1,p,q,zgh[t-q:t-1]);
    }
    vol[0]=x[w:T+w-1];
    z2[0]=zgh[1+w:T+w];
    return z[1+w:T+w];
    }
/*******************************  Probabilidad de transición  ***************************/
/*--------------------------------------------------------------.
|  Número aleatorio para la probabilidad de transición GH:     |
|  Esta función utiliza una función de Ox para generar un      |
|  número GH.                                                  |
|  in: r (renglones); c (columnas); p (parametros); l (lag);   |
|      z (valores pasados);                                    |
|  out: número aleatorio (escalar o vector)                    |
```

```
.-------------------------------------------------------------*/
ranghK(const r, const c, const p, const l, const z)
{
return rangh(r,c,p[0]-l/2,sqrt(sumc((z-p[4]).^2)+p[3]^2),sqrt(p[1]^2+(l-1)*p[2]^2),p[2])+p[4];
}
/*******************************  Probabilidad de transición gig  ************************/
/*-------------------------------------------------------------.
|  Número aleatorio para la distribución GIG:                  |
|  Esta función utiliza una función de Ox para generar un      |
|  número GH.                                                  |
|  in: r (renglones); c (columnas); p (parametros); l (lag);   |
|      z (valores pasados);                                    |
|  out: número aleatorio (escalar o vector)                    |
.-------------------------------------------------------------*/
rangigK(const r, const c, const p, const l, const z)
{
return rangig(r,c,p[0]-l/2,sqrt(sumc((z-p[4]).^2)+p[3]^2),sqrt(p[1]^2+(l-1)*p[2]^2),p[2]);
}
/*  Función máximo verosimil para el modelo GH-ARCH (función modificad para el uso del MaxBFGS)  */
/*-------------------------------------------------------------.
|  Igual que el GH(), solo que se modifica para su uso en la   |
|  función MaxBFGS de Ox. El procedimiento de estimación puede |
|  ser modificado utilizando el gradiente (avScore) y/o el     |
|  Hessiano (amHess).                                          |
.-------------------------------------------------------------*/
LW2(const p, const AdFunc, const avScore, const amHess)
{
AdFunc[0]=LW(p);
return 1;
}
/*******************************  Función máximo verosimil GH-ARCH  **********************/
/*-------------------------------------------------------------.
|  Calcula la verosimilitud del modelo GH-ARCH.                |
|  in: p=[L,a,b,d,m]'; (vector con parámetros)                 |
|  out: valor de la verosimilitud (escalar)                    |
|  use: GH(), TK(), global data z.ªnd global lag "q".          |
.-------------------------------------------------------------*/
LW(const p)
{
decl t,lw,rt;
rt=rows(y);
lw=log(GH(y[0],p[0],p[1],p[2],p[3],p[4]));
for(t=1; t<=q-1; t++) { lw=lw+log(TK(p,t,t)); }
for(t=q; t<=rt-1; t++) { lw=lw+log(TK(p,q,t)); }
```

```
return lw;
}
/********************************  Kernel de Transición  *********************************/
/*-----------------------------------------------------------------.
|  Calculo de la transición "densidad condicional"para el          |
|  modelo GH-ARCH.                                                 |
|  in: p=[L,a,b,d,m]'; l=lag; t= tiempo; uso de la variable        |
|  global y;                                                       |
|  out: valor de la densidad (escalar)                             |
.-----------------------------------------------------------------*/
TK(const p, const l, const t)
{
return GH(y[t],p[0]-l/2,sqrt(p[1]^2+l*p[2]^2),p[2],sqrt(sumc((y[t-l:t-1]-p[4]).^2)+p[3]^2),p[4]);
}
/********************************  Densidad Hiperbólica Generalizada  **********************/
/*-----------------------------------------------------------------.
|  Calcula la densidad de la distribución GH en x.                 |
|  in : x; L;a;b;d;m; (parámetros individuales)                    |
|  out: valor de la densidad (escalar)                             |
.-----------------------------------------------------------------*/
GH(const x, const L, const a, const b, const d, const m)
{
return densgh(x-m,L,d,sqrt(a^2-b^2),b);
}
/********************************  Histograma modificado  **********************************/
/*-----------------------------------------------------------------.
|  Muestra el histograma de los log-rendimientos.                  |
|  in: Y (data); Nb (num de barras);                               |
|  out: Histograma                                                 |
.-----------------------------------------------------------------*/
CtnsHist(const lugar, Y ,const Nb)
{
decl MIN,MAX,Bw,H0,H,I;
Y=sortc(Y); MIN=minc(Y); MAX=maxc(Y); I=(MAX-MIN)/Nb; Bw=(range(MIN,MAX,I));
H0=sumc(Y.>=Bw .&& Y.<Bw+I); H0[Nb-1]=H0[Nb-1]+1; H0=H0[0:Nb-1]./(I*rows(Y));
DrawHistogram(lugar, H0, MIN, (MAX-MIN)/Nb,2, 0);
}
/*************************************************************************************************/
```

Bibliografía

[1] Andersson, J. (2001). *On the normal inverse Gaussian stochastic volatility model.* Journal of Business Economic Statistics 19(1), 44-54.

[2] Barndorff-Nielsen, O. E. (1977). *Exponentially decreasing distributions for the logarithm of particle size.* Proceedings of the royal Society. London A. 353, 401-419.

[3] Barndorff-Nielsen, O. E. (1997). *Normal inverse Gaussian distributions and stochastic volatility modeling.* Scandinavian Journal of Statistics 24(1), 1-13.

[4] Bartlett, M. S. (1978). *An Introduction to Stochastic Processes.* 3rd ed. Cambridge.

[5] Besag, J. (1974). *Spatial interaction and the statistical analysis of life systems.* Journal of the Royal Statistical Society, Ser. B, 36, 192-236.

[6] Bhattacharya, C. G. (1967). *A simpled method for resolution of a distribution into its Gaussian components.* Biometrics 23, 115-135.

[7] Bollerslev, T. (1987). *A conditionally heterosedastic time series model for speculative prices and rates of return.* The Review of Economics and Statistics 69, 542-47.

[8] Box, G. E. P. and Jenkins, G. M. (1976). *Time Series Analysis, Forecasting and Control.* 2nd ed. San Francisco.

[9] Brockwell P. and Davis R. (2002). *Introduction to Time Series and Forecasting.* Springer.

[10] Crowder, M. J. (1976). *Maximum likelihood estimation for dependent observations.* Journal of the Royal Statistical Society, Series B 38, 45-53.

[11] DasGupta, S. (1999). *Learning mixtures of Gaussians*. Technical Report No. UCB/CSD 99-1047. Berkley: Computer Science Division, University of California.

[12] Day, N. E. (1969). *Estimating the components of a mixture of two distributions*. Biometrika 56, 463-474.

[13] Deely, J. J., and Kruse, R. L. (1968). *Construction of sequences estimating the mixture distribution*. Annals of Mathematical Statistics 39, 286-288.

[14] Dempster, A. P., Laird, N. M., and Rubin, D. B. (1977). *Maximum likelihood from incomplete data via the EM algorithm (with discission)*. Journal of the Royal Statistical Society B 39, 1-38.

[15] Dunmur, A. P., and Titterington, D. M. (1988). *The influence of initial conditions on maximum likelihood estimation of the parameters of a binary hidden Markov model*. Statistics & Probability Letters 40, 67-73.

[16] Engle, R. F. (1982). *Autoregressive conditional heteroscedasticity with estimates of the variance of united kingdom inflation*. Econometrica, 50(4), 987-1008.

[17] Fowlkes, E. B. (1979). *Some methods for studying mixture of two normal (lognormal) distributions*. Journal of the American Statistical Association 74, 561-575.

[18] Gaver, D. P., Lewis, P. A. W. (1980). *First-order autoregressive gamma sequences and point processes*. Advances in Applied Probability 12, 727-745.

[19] Gelfand, A. E., and Smith, A. F. M. (1990). *Sampling-based approaches to calculating marginal densities*. Journal of the American Statistical Association 85, 398-409.

[20] Gelfand, A. E., Hills, S. E., Racine-Poon, A., and Smith, A. F. M. (1990). *Illustration of bayesian inference in normal data models using Gibbs sampling*. Journal of the American Statistical Association 85, 972-985.

[21] Gelman, A., and Rubin, D. (1991). *An overview and approach to inference from iterative simulation*. Technical Report, University of California-Berkley, Dept. of Statistics.

[22] Geoffrey McLachlan y David Peel. (2000). *Finite Mixture Models*. Jhon Wiley y Sons, Inc.

[23] George Casella y Edward I. George. (1992). *Explaining the Gibbs Sampler*. The American Statistican Vol. 46, No. 3.

[24] Granger, C. W. J., and A. Andersen. (1978). *An Introduction to Bilinear Time-Series Models*. Göttingen: Vandenhoeck and Ruprecht.

[25] Hasselblad, V. (1966). *Estimation of parameters for a mixture of normal distributions.* Technometrics 8, 431-444.

[26] Hasselblad, V. (1969). *Estimation of finite mixtures of distributions from the exponential family.* Journal of the American Statistical Association 64, 1459-1471.

[27] Jensen, M. B., Lunde, A. (2001). *The NIG-S&ARCH model: a fat tailed, stochastic, and autoregressive conditional heterosedastic volatility model.* Econometrics Journal 4, 319-342.

[28] Joe, H. (1996). *Time series models with univariate margins in the convolution-closed infinitely divisible class.* Journal of Applied Probability 33, 664-77.

[29] Kadane, J. B. (1974). *The role of identification in Bayesian theory.* In studies in Bayesian Econometrics and Statistics, S. Fienberg and A. Zellner (Eds.). New York.

[30] Lawrance, A. J., Lewis, P. A. W. (1980). *The exponential autoregressive-moving average $EARMA(p\,,\,q)$ process.* Journal of the Royal Statistical Society. Series B. Statistical Methodology 42, 150-61.

[31] Lawrwnce, A. J. (1982). *The innovation distribution of a gamma distributed autoregressive process.* Scandinavian Journal of Statistics 9, 234-36.

[32] Leroux, B. G., and Puterman, M. L. (1992). *Maximum-penalized-likelihood estimation for independent and Markov-dependent mixture models.* Biometrics 48, 545-558.

[33] Lindsay, B. G. and Basak, P. (1993). *Multivariate normal mixtures: a fast, consistent method of moments.* Journal of the American Statistical Association 88, 468-476.

[34] Marron, J. S. and Wand, M. P. (1992). *Exact mean integrated squared error.* Annals of Statistics 20, 712-736.

[35] McKenzie, E. (1986). *Autoregressive moving-average processes with negative-binomial and geometric marginal distributions.* Advances in Applied Probability 18, 679-705.

[36] McKenzie, E. (1986). *Some ARMA models for dependent sequences of Poisson counts.* Advances in Applied Probability 20, 822-835.

[37] McLahlan, G. J. and Basford, K. E. (1988). *Mixture models: Inference and Applications to Clustering.* New York: Marcel Dekker.

[38] McLahlan, G. J. and Ng. S. K. (2000a). *A sparse version of the incremental EM algorithm for large databases*. Technical report. Brisbane: Department of Mathematics, University of Queensland.

[39] McLahlan, G. J. and Ng. S. K. (2000b). *A comparison of some information criteria for the number of components in a mixture model*. Technical report. Brisbane: Department of Mathematics, University of Queensland.

[40] Mena, R. H. and Walker, S.G. (2007a). *Stationary Mixture Transition Distribution (MTD) models via predictive distributions*. Journal of Statistical Planning and Inference, 137, 3103-3112.

[41] Mena, R. H. and Walker, S. G. (2007b). *On the stationary version of the generalized hyperbolic ARCH model*. Annals of the Institute of Statistical Mathematics, 59, 325-348.

[42] Mena, R. H. and Walker, S. G. (2005). *Stationary autoregressive models via a Bayesian nonparametric approach*. Journal of Time Series Analysis, 26, 789-805.

[43] Neal, R. M., and Hinton, G. E. (1998). *A view of the EM algorithm that justifies incremental, sparse, and others variants*. In Learning in Graphical Models, M. I. Jordan (Ed.). Dordrecht: Kluwer, pp. 355-368.

[44] Nelder, J. A. and Wedderburn, R. W. M. (1972). *Generalized linear models*. Journal of the Royal Statistical Society A 135, 370-384.

[45] Newcomb, S. (1886). *A generalized theory of the combination of observations so as to obtain the best result*. American Journal of Mathematics 8, 343-366.

[46] Pearson, K. (1894). *Contributions to the theory of mathematical evolution*. Philosophical Transactions of the Royal Society of London A 185, 71-110.

[47] Pitt, M.K., Chatfield, C., Walker, S.G. (2002). *Constructing first order autoregressive models via latent processes*. Scandinavian Journal of Statistics 29, 657-663.

[48] Press, H. William, Teukolsky, Saul A., Vetterling, William T. and Flannery Brian P. (2007). *Numerical recipies in C*. Cambridge University Press.

[49] Rabiner, L. R. (1989). *A tutorial on hidden Markov models and select applications in speech recognition*. Proceedings of the IEEE 77, 257-286.

[50] Raftery, A.E. (1985). *A model for high order Markov chains*. J. Roy. Statist. Soc. B 47, 528-539.

[51] Raftery, A. E., and Banfield, J. D. (1990). *Stopping the Gibbs sampler, the use of morphology, and other issues in spatial statistics*. Technical Report, University of Washington, Dept. of Statistics.

[52] Redner, R. A. (1981). *Note on the consistency of the maximum likelihood estimate for nonidentificable distributions*. Annlas of Statistics 9, 225-228.

[53] Tanner, M. A. (1991). *Tools for statistical inference*. New York: Springer-Verlag.

[54] Titterington, D. M., Smith, A. F. M., and Makov, U. E. (1985). *Statical Analysis of Finite Mixture Distributions*. Jhon Wiley y Sons, Inc.

[55] White, H. (1980). *A heterocedasticity consistent covarianze matrix estimator and a direct Test for heterocedasticity*. Econometrica 48, 817-838.

[56] William W. S. Wei, (1990). *Time Series Analysis, Univariate and Multivariate Methods*. Addison-Wesley Publishing Company.

[57] Windham, M. P., and Cutler, A. (1992). *Information ratios for validating mixture analyses*. Journal of the American Statistical Association 87, 1188-1192.

[58] Wolfe, J. H. (1965). *A computer program for the computation of maximum likelihood analysis of types*. Research Memo. SRM 65-12. San Diego: U.S. Naval Personnel Research Activity.

[59] Wolfe, J. H. (1967). *NORMIX: Computational methods for estimating the parameters of multivariate normal mixtures of distributions*. Research Memo. SRM 68-2. San Diego: U.S. Naval Personnel Research Activity.

[60] Wolfe, J. H. (1970). *Pattern clustering by multivariate mixture analysis*. Multivariate Behavioral Research 5, 329-350.