Capítulo 10

Estimación y selección de modelos ARIMA

Rudolf E. Kalman (1930 -)

Matemático húngaro. Estudio en el MIT y fue profesor en Stanford y Florida. Inventor en 1961 con Richard Bucy del procedimiento recursivo de estimación del estado de un sistema dinámico que se conoce como filtro de Kalman. Este método es de uso habitual en ingeniería de control, tanto en el campo aeroespacial como en el industrial, y una de las herramientas más utilizadas para la estimación y previsión de series temporales. Desarrolló los fundamentos modernos de los sistemas dinámicos de control.

10.1 Introducción

En este capítulo supondremos inicialmente que se dispone de una serie estacionaria de T elementos $\{\omega_t\}$ $t = 1, ..., T$, para estimar los parámetros de un modelo ARMA concreto y estudiaremos la estimación del modelo mediante máxima verosimilitud. Utilizaremos la notación ω_t porque es frecuente que la serie a estimar sea una transformación de la serie original, z_t. Por ejemplo, con datos mensuales económicos es frecuente que $\omega_t = \nabla\nabla_{12} \log z_t$. El lector interesado en las aplicaciones puede prescindir de estas secciones sin pérdida de continuidad. La segunda parte del capítulo considera el caso en el que disponemos de varios modelos ARIMA estimados para una serie y se plantea el problema de decidir entre ellos seleccionando el más adecuado. Las ideas principales de seleccion de modelos son importantes y se utilizan mucho en el resto del libro.

Supondremos que no se dispone de información a priori sobre los valores de los parámetros del proceso y que, como suele ser habitual, disponemos de una muestra grande. En este caso la estimación Bayesiana y la estimación clásica por máxima verosimilitud conducen a los mismos resultados, aunque la interpretación de los resultados es distinta. Para simplificar la exposición presentamos unicamente la estimación MV, aunque el filtro de Kalman, la herramienta central de estimación de modelos ARMA, puede justificarse tanto desde el punto de vista clásico como del Bayesiano.

El estudio de la estimación comienza con el caso más simple: la estimación condicionada de procesos AR. El método es entonces similar a la estimación por mínimos cuadrados de un modelo de regresión. A continuación veremos la estimación exacta de procesos AR, que conduce a un problema de estimación no lineal en los parámetros, lo que require la utilización de algoritmos de optimización para problemas no lineales. A continuación, estudiaremos la estimación de modelos MA y ARMA, que es siempre no lineal y presenta dos fases principales. La primera es calcular cuanto vale la función de verosimilitud dado un valor de los parámetros. La segunda es encontrar un nuevo valor de los parámetros que haga más grande el valor de la función. La estimación consiste en iterar entre estas dos fases, hasta obtener el máximo de la función. Nos centraremos en cómo evaluar la función de verosimilitud mediante un algoritmo eficiente, el filtro de Kalman. En el apéndice 10.1 presentaremos brevemente los fundamentos de los métodos de optimización que se utilizan para obtener el máximo de la función.

Si hemos estimado varios modelos se presenta el problema de decidir entre ellos. Los criterios de ajuste no son útiles en este caso porque si aumentamos el número de parámetros siempre aumentará el ajuste del modelo. Tenemos que acudir entonces a criterios de selección de modelos, como los que se presentan en este capítulo.

10.2 La función de verosimilitud de un proceso ARMA

Supongamos que tenemos un proceso ARMA y se desea estimar los parámetros por máxima verosimilitud. Para ello debemos escribir la función de densidad conjunta y maximizarla respecto a los parámetros considerando a los datos como fijos. Para escribir la densidad conjunta de las T observaciones $\boldsymbol{\omega}_T = (\omega_1, ..., \omega_T)$, vamos a utilizar que la distribución conjunta de dos variables cualesquiera, $\mathbf{x}, \mathbf{y}$, puede siempre escribirse como el producto de dos distribuciones univariantes. La primera es la marginal de una de ellas y la segunda la distribución de la otra variable condicionada a los valores de la utilizada para la marginal. Es decir :

$$f(\mathbf{x}, \mathbf{y}) = f(\mathbf{x}) f(\mathbf{y}|\mathbf{x}) \tag{10.1}$$

Esta expresión sigue siendo cierta si $\mathbf{x}$ e $\mathbf{y}$ son variables vectoriales. También si todas las funciones de densidad van condicionadas a otra variable $\mathbf{z}$, de manera que :

$$f(\mathbf{x}, \mathbf{y}|\mathbf{z}) = f(\mathbf{x}|\mathbf{z}) f(\mathbf{y}|\mathbf{x}, \mathbf{z}) \tag{10.2}$$

Consideremos la función de densidad conjunta de las T observaciones $\boldsymbol{\omega}_T$. Tomando $\mathbf{x}$ =ω_1 y $\mathbf{y}$ =$\omega_2, ..., \omega_T$ en (10.1), podemos escribir

$$f(\boldsymbol{\omega}_T) = f(\omega_1) f(\omega_2, ..., \omega_T|\omega_1)$$

y descomponiendo el segundo término, con (10.2) haciendo $\mathbf{z}$ =ω_1, $\mathbf{x}$ =ω_2 y $\mathbf{y}$ =ω_3,..., ω_T, resulta

$$f(\boldsymbol{\omega}_T) = f(\omega_1) f(\omega_2|\omega_1) f(\omega_3, ..., \omega_T|\omega_1, \omega_2)$$

y repitiendo este proceso, obtenemos finalmente

$$f(\boldsymbol{\omega}_T) = f(\omega_1) f(\omega_2|\omega_1) f(\omega_3|\omega_2, \omega_1) ... f(\omega_T|\omega_{T-1}, ..., \omega_1). \tag{10.3}$$

Esta expresión permite escribir la función de densidad conjunta como producto de distribuciones univariantes. La diferencia entre esta representación y la que se obtiene con datos independientes es que en lugar de tener el producto de las marginales de cada dato tenemos la marginal de la primera y el producto de las condicionadas de cada dato dados los anteriores.

La descomposición anterior10.3 permite escribir la verosimilitud de un modelo ARMA. Su suponemos normalidad, todas las distribuciones condicionadas serán normales. Llamaremos a sus parámetros:

$$E(\omega_t|\omega_{t-1}, ..., \omega_1) = \widehat{\omega}_{t|t-1}$$

$$Var(\omega_t|\omega_{t-1}, ..., \omega_1) = \sigma^2 v_{t|t-1}$$

donde $\sigma^2 = Var(a_t)$ se saca factor común para simplificar, ya que va a aparecer en el cálculo de todas las varianzas condiciondas. Entonces, la función de densidad conjunta de la muestra puede escribirse como:

$$f(\boldsymbol{\omega}_T) = \prod_{t=1}^{T} \sigma^{-2} v_{t|t-1}^{-1/2} (2\pi)^{-1/2} \exp\left\{ -\frac{1}{2\sigma^2} \sum_{1}^{T} \frac{(\omega_t - \widehat{\omega}_{t|t-1})^2}{v_{t|t-1}} \right\}$$

y llamando $e_t = \omega_t - \widehat{\omega}_{t|t-1}$ a la diferencia entre los valores observados y sus esperanzas condicionadas, que son los errores de predicción a un paso, tenemos que la función de verosimilitud exacta de un proceso ARMA es, llamando $\boldsymbol{\beta} = (\mu, \phi_1, ..., \theta_q, \sigma^2)$ al vector de parámetros :

$$L(\boldsymbol{\beta}) = -\frac{T}{2} \log \sigma^2 - \frac{1}{2} \sum_{t=1}^{T} \log v_{t|t-1} - \frac{1}{2\sigma^2} \sum_{t=1}^{T} \frac{e_t^2}{v_{t|t-1}} \tag{10.4}$$

donde tanto los coeficientes $v_{t|t-1}$ como los errores e_t dependen de los parámetros. Por tanto, evaluar la función de verosimilitud se reduce al problema de calcular los errores de predicción a un paso de cada observación dadas las anteriores y sus varianzas.

La maximización de la función de verosimilitud exacta se efectúa con un algoritmo de optimización no lineal. Un algoritmo muy utilizado es el basado en el algoritmo de Gauss-Newton con una modificación debida Marquardt, que se describe en el apéndice 10.1

10.3 Procesos AR

10.3.1 El proceso AR(1)

Como ilustración consideremos el proceso AR(1), $\omega_t = \phi\omega_{t-1} + a_t$, de media cero. En este caso, según vimos en el capítulo :

$$E(\omega_1) = 0 \tag{10.5}$$

y

$$Var(\omega_1) = \frac{\sigma^2}{1-\phi^2} \tag{10.6}$$

Por tanto con la notación anterior $\widehat{\omega}_1 = 0$ y $v_1 = (1-\phi^2)^{-1}$. Para ω_2 tenemos que, al condicionar a ω_1, :

$$E(\omega_2|\omega_1) = \phi\omega_1$$

y

$$Var(\omega_2|\omega_1) = E\left[(\omega_2 - \phi\omega_1)^2\right] = E(a_2^2) = \sigma^2$$

y ahora $\widehat{\omega}_{2|1} = \phi\omega_1$ y $v_{2|1} = 1$.de la misma forma comprobamos que

$$E(\omega_t|\omega_{t-1}) = \widehat{\omega}_{t|t-1} = \phi\omega_{t-1}, \qquad t = 2, ..., T$$

y

$$Var(\omega_t|\omega_{t-1}) = \sigma^2 v_{t|t-1} = \sigma^2, \qquad t = 2, ..., T$$

En consecuencia la función de verosimilitud será:

$$f\left(\boldsymbol{\omega}_T\right) = f(\omega_1)\prod_{t=2}^{T}\sigma^{-2}\left(2\pi\right)^{-1/2}\exp\left\{-\frac{1}{2\sigma^2}\sum_{2}^{T}(\omega_t - \phi\omega_{t-1})^2\right\} \tag{10.7}$$

y tomando logaritmos y utilizando que $f(\omega_1)$ es normal con parámetros dados por (10.6) y (10.5), resulta:

$$\begin{aligned} L\left(\phi,\sigma^2|\boldsymbol{\omega}_T\right) &= \frac{-T}{2}\ln\sigma^2 + \frac{1}{2}\ln\left(1-\phi^2\right) - \\ &\quad \frac{\left(1-\phi^2\right)\left(\omega_1-\mu\right)^2}{2\sigma^2} - \frac{1}{2\sigma^2}\sum_{t=2}^{T}(\omega_t - \phi\omega_{t-1})^2 \end{aligned} \tag{10.8}$$

Para obtener el estimador de ϕ tendremos que derivar respecto a este parámetro e igualar a cero el resultado. Se obtiene una ecuación cúbica que tiene tres raices, y la que maximice la función de verosimilitud es el estimador MV.

Las expresiones (10.7) y (10.8) muestran que si prescindimos del primer término la verosimilitud es lineal en los parámetros. Si condicionamos a la primera observación tenemos que

$$f\left(\omega_2, ..., \omega_T|\omega_1\right) = \prod_{t=2}^{T}\sigma^{-2}\left(2\pi\right)^{-1/2}\exp\left\{-\frac{1}{2\sigma^2}\sum_{2}^{T}(\omega_t - \phi\omega_{t-1})^2\right\}$$

Si definimos como verosimilitud condicionada a la resultante de esta función de densidad conjunta

$$L_C\left(\phi,\sigma^2|\omega_2,...,\omega_T\right) = \frac{-(T-1)}{2}\ln\sigma^2 - \frac{1}{2\sigma^2}\sum_{t=2}^{T}(\omega_t - \phi\omega_{t-1})^2$$

y el estimador del parámetro ϕ maximizando esta verosimilitud condicionada se obtiene minimizando la suma de cuadrados

$$\sum_{t=2}^{T}(\omega_t - \phi\omega_{t-1})^2$$

que conduce al estimador

$$\widehat{\phi} = \frac{\sum_{t=2}^{T}\omega_t\omega_{t-1}}{\sum_{t=2}^{T}\omega_{t-1}^2},$$

que es similar al de la pendiente en un modelo de regresión. El estimador MV condicionado de la varianza será:

$$\widehat{\sigma}^2 = \frac{\sum_{t=2}^{T}(\omega_t - \widehat{\phi}\omega_{t-1})^2}{T-1}$$

Hemos visto que si condicionamos al primer término y escribimos la verosimilitud de los elementos 2 al T tenemos un modelo lineal en los parámetros. La diferencia entre el estimador obtenido con la verosimilitud condicionada y la exacta será en general pequeña, y despreciable para muestras grandes.

10.3.2 Procesos AR(p)

Consideremos un proceso AR(p) general. La esperanza condicionada de ω_t, para $t = p+1,...,T$ dados los datos previos, $\omega_{t-1},...,\omega_1$ será, utilizando la ecuación del AR(p):

$$E\left[\omega_t|\omega_{t-1},...,\omega_1\right] = \mu + \phi_1(\omega_{t-1}-\mu) + ... + \phi_p(\omega_{t-p}-\mu)$$

y su varianza condicionada será:

$$Var\left(\omega_t|\omega_{t-1},...,\omega_1\right) = Var(a_t) = \sigma^2$$

En consecuencia, todas las distribuciones condicionadas para $t = p+1,...,T$ son normales, con media igual a la predicción de la observación dado el pasado y varianza σ^2, el error cuadrático medio de predicción a un paso cuando conocemos los parámetros. La función de verosimilitud condicionada se obtendrá a partir de la densidad conjunta de las observaciones $(\omega_{p+1,},...,\omega_T)$ condicionadas a las p primeras. Su expresión es:

$$L_C\left(\mu,\boldsymbol{\phi},\sigma^2\right) = -\frac{(T-p)}{2}\ln\sigma^2 - \frac{1}{2\sigma^2}\sum_{t=p+1}^{T}(\omega_t - \mu - \Sigma_{i=1}^{p}\phi_i(\omega_{t-i}-\mu))^2 \tag{10.9}$$

Maximizar esta verosimilitud respecto a μ y $\boldsymbol{\phi}$ equivale a minimizar la suma de cuadrados de los residuos o de errores de predicción a un paso:

$$S = \sum_{t=p+1}^{T}a_t^2 = \sum_{t=p+1}^{T}(\omega_t - \mu - \Sigma_{i=1}^{p}\phi_i(\omega_{t-i}-\mu))^2 \tag{10.10}$$

donde $a_t = (\omega_t - \mu - \Sigma_{i=1}^{p}\phi_i(\omega_{t-i}-\mu))$. Por tanto, maximizar la verosimilitud condicional equivale a mínimos cuadrados. El estimador de μ se obtiene derivando e igualando a cero:

$$\sum_{t=p+1}^{T}(\omega_t - \mu - \Sigma_{i=1}^{p}\phi_i(\omega_{t-i}-\mu)) = 0$$

y suponiendo que $\sum_{t=p+1}^{T} \omega_t \simeq \sum_{t=p+1}^{T} \omega_{t-i}$, lo que será aproximadamente cierto si T es grande, obtenemos que el estimador de la media es la media muestral de las observaciones consideradas:

$$\widehat{\mu} = \frac{\sum_{t=p+1}^{T} \omega_t}{T-p}.$$

Un mejor estimador de μ es $\overline{\omega} = \sum_{t=1}^{T} \omega_t/T$, la media muestral de todas las observaciones, que estudiamos en el Capítulo 3. Ambos estimadores son centrados, pero el calculado con toda la muestra tiene menor varianza (véase el ejercicio 10.1). Por tanto será el que utilicemos. Esto equivale a estimar inicialmente la media con todos los datos y después escribir la verosimilitud para las variables en desviaciones a la media.

Para obtener el estimador de $\boldsymbol{\phi}$, sutituyendo μ por $\overline{\omega}$ en (10.10) y llamando $\mathbf{x}_t' = (\omega_{t-1} - \overline{\omega}, ..., \omega_{t-p} - \overline{\omega})$, se obtiene el estimado habitual de minimos cuadrados en modelos de regresión

$$\widehat{\boldsymbol{\phi}} = \left(\sum_{t=p+1}^{T} \mathbf{x}_t \mathbf{x}_t' \right)^{-1} \left(\sum_{t=p+1}^{T} \mathbf{x}_t (\omega_t - \overline{\omega}) \right), \tag{10.11}$$

Esta expresión para muestras grandes será aproximadamente:

$$\widehat{\boldsymbol{\phi}} = \widehat{\Gamma}_p^{-1} \widehat{\boldsymbol{\gamma}}_p \tag{10.12}$$

donde

$$\widehat{\Gamma}_p = \begin{vmatrix} \widehat{\gamma}_0 & \cdots & \widehat{\gamma}_{p-1} \\ \cdots & \widehat{\gamma}_0 & \cdots \\ \widehat{\gamma}_{p-1} & \cdots & \widehat{\gamma}_0 \end{vmatrix}, \qquad \widehat{\boldsymbol{\gamma}}_p = \begin{vmatrix} \widehat{\gamma}_1 \\ \cdots \\ \widehat{\gamma}_p \end{vmatrix}$$

que son las ecuaciones de Yule-Walker. Sin embargo en pequeñas muestras ambos estimadores son diferentes y tanto más diferentes cuanto mayor sea el orden del proceso. Los estimadores de mínimos cuadrados, o condicionales, consideran la muestra de $T-p$ observaciones y todos los coeficientes de (10.11) se calculan como cocientes de sumas de $T-p$ términos en el numerador y en el denominador. Los estimadores de Yule-Walker utilizan distintos términos en cada sumando: $\widehat{\gamma}_0$ se calcula con los T datos pero $\widehat{\gamma}_k$ con $T-k$. Este desequilibrio introduce sesgos importantes y puede demostrarse (Tjostheim y Paulsen, 1983) que los estimadores de mínimos cuadrados son más precisos que los de Yule-Walker.

La estimación por máxima verosimilitud exacta requiere calcular las esperanzas y varianzas condicionadas para las p primeras observaciones. Esto puede hacerse de manera similar a como se hizo para un AR(1) pero veremos más adelante un procedimiento general para obtenerlas.

10.4 Estimación de modelos MA y ARMA

10.4.1 Estimación Condicional

La estimación de modelos MA y mixtos es más complicada que la de los AR por dos razones. En primer lugar la función de verosimilitud, tanto la condicional como la exacta, es siempre no lineal en los parámetros. En segundo, el procedimiento de condicionar a ciertos valores iniciales, que lleva a resultados simples en los AR, es más complicado para procesos MA y ARMA, haciendo el cálculo de las esperanzas y varianzas condicionadas más difícil. Para ilustrar estas dificultades, consideremos el caso de un MA(l):

$$\omega_t = a_t - \theta a_{t-1}$$

con esperanza marginal cero. La esperanza de ω_t condicionada a sus valores previos ya no es inmediata, como en los AR, y para obtenerla tenemos que expresar ω_t en función de los valores anteriores. Comenzando con $t=2$, como $\omega_2 = a_2 - \theta a_1$, y $a_1 = \omega_1 + \theta a_0$, tenemos que

$$\omega_2 = -\theta\omega_1 + a_2 - \theta^2 a_0$$

de donde deducimos que la esperanza de la distribución condicionada es:

$$E(\omega_2|\omega_1) = -\theta\omega_1$$

y la varianza

$$var(\omega_2|\omega_1) = \sigma^2(1+\theta^4)$$

Procediendo de esta forma para $t = 3, 4, ..,$ se obtiene que

$$\omega_t = -\theta\omega_{t-1} - \theta^2\omega_{t-2} - ... - \theta^{t-1}\omega_1 + a_t - \theta^t a_0$$

que conduce a

$$E(\omega_t|\omega_{t-1}, ..., \omega_1) = -\theta\omega_{t-1} - \theta^2\omega_{t-2} - ... - \theta^{t-1}\omega_1$$

y

$$var(\omega_t|\omega_{t-1}, ..., \omega_1) = \sigma^2(1+\theta^{2t})$$

Estas expresiones son no lineales en los parámetros y pesadas de calcular por este método de sustitución en procesos MA(q) generales.

Un enfoque alternativo es condicionar también en las primeras innovaciones no observadas. Observemos que para cada valor de los parámetros θ, la expresión

$$a_t = \omega_t + \theta a_{t-1} \tag{10.13}$$

permite calcular recursivamente las perturbaciones a_t, condicionadas a un valor inicial a_0. Tomando $a_0 = 0$ podemos calcular todas las restantes perturbaciones a partir de los ω_t. Entonces

$$E(\omega_t|\omega_{t-1}, ..., \omega_1, a_0) = -\theta a_{t-1}$$

y

$$var(\omega_t|\omega_{t-1}, ..., \omega_1, a_0) = E\left[(\omega_t + \theta a_{t-1})^2\right] = E\left[{a_t}^2\right] = \sigma^2$$

que conduce a la verosimilitud condicionada

$$L_C\left(\theta|\omega_1, a_0\right) = \frac{-(T-1)}{2}\ln\sigma^2 - \frac{1}{2\sigma^2}\sum_{t=2}^{T} a_t^2\left(\theta|a_0\right)$$

La maximización de esta función se realiza mediante un algortimo no lineal que se describe en el apéndice 10.1

La estimación condicionada de modelos ARMA(p, q) se realiza siguiendo los mismos principios. Llamando $r = \max(p, q)$ y $\boldsymbol{\beta} = (\mu, \phi_1, ..., \theta_q, \sigma^2)$ al vector de parámetros, la función de verosimilitud condicional es:

$$L_C\left(\boldsymbol{\beta}|\mathbf{a}_0, \boldsymbol{\omega}_p\right) = \frac{-(T-r)}{2}\ln\sigma^2 - \frac{1}{2\sigma^2}\sum_{T=r+1}^{T} a_t^2\left(\boldsymbol{\beta}|\mathbf{a}_0, \boldsymbol{\omega}_p\right) \tag{10.14}$$

donde $\mathbf{a}_0$ y $\boldsymbol{\omega}_p$ son los vectores iniciales a los que condicionamos la estimación. Los residuos se calculan recursivamente mediante:

$$\widehat{a}_t = \omega_t - c - \phi_1\omega_{t-1} - ... - \phi_p\omega_{t-p} + \theta_1\widehat{a}_{t-1} + ... + \theta_q\widehat{a}_{t-q} \quad t = r+1, ..., T \tag{10.15}$$

donde $c = \mu(1 - \phi_1 - ... - \phi_p)$ y se supone que los primeros r residuos son cero. La maximización de (10.14) requiere un valor inicial de los parámetros que puede obtenerse con el algoritmo de Hannan y Rissanen que se presenta a continuación.

El Algoritmo de Hannan y Rissanen

Este algoritmo proporciona estimadores iniciales para un proceso ARMA(p,q). Si se itera, puede utilizarse para obtener estimadores de procesos ARMA utilizando sólo regresiones. El algortimo funciona en dos etapas.

Etapa 1: Obtenemos una estimación inicial de los residuos del modelo ajustando una AR largo de orden $k > p + q$. Sean $\widehat{\pi}_i$ los coeficientes estimados utilizando (10.11). Los residuos se calculan mediante

$$\widehat{a}_t = \omega_t - \widehat{c} - \sum_{i=1}^{k} \widehat{\pi}_i \omega_{t-i}$$

Etapa 2: Con los residuos estimados en la primera etapa, se estima la regresión

$$\omega_t = c + \phi_1 \omega_{t-1} + ... + \phi_p \omega_{t-p} - \theta_1 \widehat{a}_{t-1} - ... - \theta_q \widehat{a}_{t-q} + u_t \tag{10.16}$$

donde las variables $\widehat{a}_{t-j}$ se construyen a partir del vector de innovaciones $\widehat{a}_t$ calculadas en la etapa anterior. La estimación de esta regresión proporciona los estimadores iniciales.

El algoritmo de Hannan Rissanen puede utilizarse para obtener estimadores de modelos ARMA iterando las dos etapas anteriores que sólo requieren regresiones. En efecto, con los parámetros estimados en la etapa 2 podemos calcular nuevos residuos y repetir la estimación de (10.16) hasta obtener convergencia. En estas condiciones obtenemos estimdores próximos a los MV. El algoritmo puede modificarse para mejorar sus propiedades asintóticas (véase Koreisha y Pukkila, 1990) Cuando se utiliza para obtener estimadores iniciales sólo se realiza el ciclo una vez, mediante las etapas 1 y 2 descritas.

10.4.2 Estimación MV Exacta

La estimación exacta requiere calcular los parámetros de las distribuciones de cada observación condicionada a las precedentes, $\widehat{\omega}_{t|t-1}$ y $\sigma^2 v_{t|t-1}$, y viene dada por (10.4). En la sección siguiente veremos un método general para calcularla.

Es importante resaltar que la descomposición de la función de densidad conjunta como producto de condicionadas es un caso particular de una descomposición que se utiliza en muchas aplicaciones estadísticas y que se conoce como factorización de Cholesky. Observemos que la expresión general de la función de densidad conjunta de $\boldsymbol{\omega}_T$ es, en logaritmos:

$$\text{Log} f(\boldsymbol{\omega}_T) = -\frac{T}{2} \log \sigma^2 - \frac{1}{2} \log |\mathbf{M}_T| - \frac{1}{2\sigma^2} (\boldsymbol{\omega}_T - \boldsymbol{\mu})' \mathbf{M}_T^{-1} (\boldsymbol{\omega}_T - \boldsymbol{\mu}) \tag{10.17}$$

Si comparamos (10.17) y (10.4) vemos que en (10.17) se ha sustituido la forma cuadrática $(\boldsymbol{\omega}_T - \boldsymbol{\mu})' \mathbf{M}_T^{-1} (\boldsymbol{\omega}_T - \boldsymbol{\mu})$ por la expresión más simple $\mathbf{e}'\mathbf{D}\mathbf{e}$, donde $\mathbf{D}$ es una matriz diagonal con términos $v_{t|t-1}^{-1}$ y $\mathbf{e} = (e_1, ..., e_T)'$ es el vector de errores de predicción. Además, en lugar del determinante de $\log |\mathbf{M}_T|$ aparece en (10.4) el término $\sum_{t=1}^{T} \log v_{t|t-1}$. Estos cambios pueden interpretarse como la realización de una transformacion

$$\mathbf{e} = \mathbf{L}(\boldsymbol{\omega}_T - \boldsymbol{\mu})$$

donde la matriz $\mathbf{L}$ es triangular inferior y tiene unos en la diagonal de manera que e_t sea sólo función de los valores actuales y previos, $(\omega_t, ..., \omega_1)$ y no de los posteriores $(\omega_{t+1}, ..., \omega_T)$. De esta manera:

$$\mathbf{e}'\mathbf{D}\mathbf{e} = (\boldsymbol{\omega}_T - \boldsymbol{\mu})' \mathbf{L}' \mathbf{D} \mathbf{L} (\boldsymbol{\omega}_T - \boldsymbol{\mu}) = (\boldsymbol{\omega}_T - \boldsymbol{\mu})' \mathbf{M}_T^{-1} (\boldsymbol{\omega}_T - \boldsymbol{\mu})$$

y hemos descompuesto la matriz $\mathbf{M}_T^{-1}$ como producto de dos matrices triangulares $\mathbf{T}'\mathbf{T}$ donde $\mathbf{T} = \mathbf{D}^{1/2}\mathbf{L}$ y donde $\mathbf{D}^{1/2}$ es la matriz diagonal que contiene los términos $v_{t|t-1}^{-1/2}$. De la misma forma, el determinante de $\mathbf{M}_T$ es el producto de los términos diagonales de $\mathbf{D}$, ya que el determinante de la matriz triangular con unos en la diagonal, $\mathbf{L}$, es la unidad.

La clave de esta descomposición es la factorización de la matriz definida positiva $\mathbf{M}_T^{-1}$ como producto de dos matrices triangulares, y se conoce como factorización de Cholesky. En la sección siguiente veremos un algoritmo recursivo muy potente para realizar esta descomposición: el filtro de Kalman.

10.5 El filtro de Kalman

El filtro de Kalman es un procedimiento recursivo muy rápido computacionalmente que tiene muchas aplicaciones en series temporales. En particular, permite evaluar rápidamente la función de verosimilitud de cualquier modelo ARMA calculando los errores de prediccion a un paso y sus varianzas. El filtro fue inicialmente diseñado para resolver un problemas más general: la estimación de sistemas de control en el espacio de los estados, como se explica a continuación. El filtro tiene una inmediata interpretación Bayesiana, por lo que se ha utilizado mucho en la formulación de series temporales mediante este enfoque.

10.5.1 Modelos en el espacio de los estados

Supongamos que observamos un sistema que puede representarse mediante una ecuación de observación:

$$\mathbf{z}_t = \mathbf{H}_t \boldsymbol{\alpha}_t + \boldsymbol{\epsilon}_t \tag{10.18}$$

donde $\mathbf{z}_t$ es un vector de observaciones de dimensiones $k \times 1$, $\mathbf{H}_t$ es una matriz $k \times p$ que suponemos conocida para todo t, $\boldsymbol{\alpha}_t$ un vector de variables de estado de dimension $p \times 1$ que no se observa y $\boldsymbol{\epsilon}_t$ es un proceso de ruido blanco que suponemos tiene distribución $N(\mathbf{0}, \mathbf{V}_t)$. Además, la descripción del sistema incluye una ecuación que describe la evolución dinámica de las variables de estado, $\boldsymbol{\alpha}_t$, llamada ecuación de estado

$$\boldsymbol{\alpha}_t = \boldsymbol{\Omega}_t \boldsymbol{\alpha}_{t-1} + \mathbf{u}_t \tag{10.19}$$

donde $\boldsymbol{\Omega}_t$ es una matriz conocida de dimensión $p \times p$ y $\mathbf{u}_t$ otro proceso de ruido blanco, independiente del anterior, que tiene disribución $N_p(\mathbf{0}, \mathbf{R}_t)$. Suponemos que las matrices del sistema $\mathbf{H}_t$ y $\boldsymbol{\Omega}_t$ se conocen para todos los instantes del sistema.

Por ejemplo, supongamos que observamos la posición en el espacio de un satélite en cada instante t, de manera que, para cada t, tenemos un vector, $\mathbf{z}_t$, de dimensión tres (las coordenadas de la posición del satélite en el espacio). Suponemos que la posición del satélite depende de un conjunto de variables de estado, $\boldsymbol{\alpha}_t$, que no son observables, pero que están relacionadas con nuestras observaciones a través de la matriz $\mathbf{H}_t$, que es conocida. Por ejemplo, si la posición del satélite depende de su velocidad y su aceleración, el vector de estado tiene dimensión dos. Suponemos que existe en cada instante un error de observación, representado por el vector $\boldsymbol{\epsilon}_t$. Por otro lado, la ecuación de estado indica que el vector de variables de estado en el instante t depende de la situación de las variables de estado en el instante $t-1$ más un error de medida.

La representación de un sistema mediante las ecuaciones (10.18) y (10.19) no es única. Siempre es posible aumentar la dimensión del vector de estado poniendo ceros en las matrices que le multiplican y decimos que el vector de estado tiene dimensión mínima cuando no es posible representar el sistema con menos de p variables de estado. Una vez fijada la dimensión el vector de estado tampoco es único. Dado un vector de estado $\boldsymbol{\alpha}_t$ el sistema puede igualmente representarse con el vector de estado $\boldsymbol{\alpha}_t^* = \mathbf{A}\boldsymbol{\alpha}_t$, donde $\mathbf{A}$ es cualquier matriz cuadrada no singular. En efecto escribiendo la ecuación de observación como

$$\mathbf{z}_t = \mathbf{H}_t \mathbf{A}^{-1} \mathbf{A} \boldsymbol{\alpha}_t + \boldsymbol{\epsilon}_t = \mathbf{H}_t^* \boldsymbol{\alpha}_t^* + \boldsymbol{\epsilon}_t$$

y la de evolución del estado como

$$\boldsymbol{\alpha}_t^* = \boldsymbol{\Omega}_t^* \boldsymbol{\alpha}_{t-1}^* + \mathbf{u}_t$$

donde ahora $\boldsymbol{\Omega}_t^* = \mathbf{A}\boldsymbol{\Omega}_t\mathbf{A}^{-1}$. En adelante supondremos que el sistema tiene dimensión mínima.

Cualquier modelo ARMA(p, q) puede escribirse en esta formulación como sigue. Definamos $m = \max(p, q+1)$ y llamemos $\boldsymbol{\alpha}_t = (\alpha_{1,t}, \alpha_{2,t}, ..., \alpha_{m,t})'$ al vector de variables de estado, que seguirá la ecuación de estado:

$$\begin{bmatrix} \alpha_{1,t} \\ \alpha_{2,t} \\ ... \\ \alpha_{m,t} \end{bmatrix} = \begin{bmatrix} \phi_1 & 1 & ... & 0 \\ \phi_2 & 0 & ... & 0 \\ ... & 0 & ... & 1 \\ \phi_m & 0 & ... & 0 \end{bmatrix} \begin{bmatrix} \alpha_{1,t-1} \\ \alpha_{2,t-1} \\ ... \\ \alpha_{m,t-1} \end{bmatrix} + \begin{bmatrix} 1 \\ -\theta_1 \\ ... \\ -\theta_m \end{bmatrix} a_t \tag{10.20}$$

Observemos que en esta ecuación la matriz de estado, $\boldsymbol{\Omega}_t$, tiene la forma

$$\boldsymbol{\Omega}_t = \begin{bmatrix} \boldsymbol{\phi}_{m-1} & \mathbf{I} \\ \phi_m & \mathbf{0}' \end{bmatrix}$$

donde $\boldsymbol{\phi}_{m-1}$ es un vector columna de dimensión m-1, $\mathbf{I}$ es la matriz identidad y $\mathbf{0}'$ es un vector de ceros. Por otro lado el vector de ruidos en esta ecuación es

$$\mathbf{u}_t = \boldsymbol{\theta} a_t$$

donde $\boldsymbol{\theta}' = (1, -\theta_1, ..., -\theta_m)$. La matriz de varianzas y covarianzas de $\mathbf{u}$ es

$$\mathbf{R}_t = \boldsymbol{\theta}\boldsymbol{\theta}'\sigma^2.$$

Vamos a comprobar que sustituyendo sucesivamente en las variables de estado se obtiene la representación del proceso ARMA. La primera ecuación es

$$\alpha_{1,t} = \phi_1 \alpha_{1,t-1} + \alpha_{2,t-1} + a_t \tag{10.21}$$

y la segunda

$$\alpha_{2,t} = \phi_2 \alpha_{1,t-1} + \alpha_{3,t-1} - \theta_1 a_t \tag{10.22}$$

Sustituyendo $\alpha_{2,t-1}$ en (10.21) de acuerdo con la expresión (10.22), tenemos que

$$\alpha_{1,t} = \phi_1 \alpha_{1,t-1} + \phi_2 \alpha_{1,t-2} + \alpha_{3,t-2} + a_t - \theta_1 a_{t-1} \tag{10.23}$$

La tercera ecuación es

$$\alpha_{3,t} = \phi_3 \alpha_{1,t-1} + \alpha_{4,t-1} - \theta_2 a_t$$

y sustituyendo ahora en (10.23) $\alpha_{3,t-2}$ por su expresión anterior, vamos recuperando el proceso ARMA en la variable $\alpha_{1,t}$. La ecuación de observación sirve simplemente para hacer la variable observada, z_t, que es escalar, igual a la primera componente del vector de estado:

$$z_t = (1, 0, ...0)\boldsymbol{\alpha}_t \tag{10.24}$$

Las ecuaciones (10.20) y (10.24) constituyen una forma de representar el modelo ARMA en el espacio de los estados. Observemos que son un caso particular de la (10.18) y (10.19). En la ecuación de observación (10.18) el vector de datos es ahora un escalar, el valor de la serie observada en cada instante, el vector de estado es un vector de dimensiones $m = \max(p, q+1)$, la matriz $\mathbf{H}_t$, es siempre el vector $(1, 0, ..., 0)$ y no existe error de medida o ruido en la matriz de observación. En la ecuación de estado la matriz $\boldsymbol{\Omega}_t$ es invariante en el tiempo, y la matriz de covarianzas de $\mathbf{u}_t$ es singular de rango uno.

La representación de un modelo ARMA en el espacio de los estados no es única, véase Box, Jenkins y Reinsel (1994), Brockwell y Davies (1996) y Peña, Tiao y Tsay (2002) para representaciones alternativas.

Una forma alternativa de escribir modelos de series temporales que proporciona una representación inmediata en el espacio de los estados son los modelos estructurales, estudiados por Harrison y Stevens (1976), West and Harrison (1989) y Harvey (1989) entre otros. En estos modelos la serie se expresa como suma de componentes asociados a la tendencia y la estacionalidad más un ruido blanco, como

$$z_t = \mu_t + S_t + a_t$$

Un formulación habitual de la tendencia es:

$$\mu_t = \mu_{t-1} + \beta_{t-1} + v_t \tag{10.25}$$
$$\beta_t = \beta_{t-1} + u_t \tag{10.26}$$

y la estacionalidad se modela con coeficientes estacionales constantes que pueden ser arbotrarios o seguir una estructura trigonométricas, como vimos en el capítulo 2. Suponiendo coeficientes constantes, los coeficientes S_t deben verificar

$$\sum_{j=1}^{s} S_{t-j} = \epsilon_t \tag{10.27}$$

donde las variables v_t, u_t, y ϵ_t son ruidos blancos independientes. Este modelo tiene un vector de estado de dimensión 13, formado por las variables de estado μ_t, β_t y los 11 coeficientes estacionales. Definiendo:

$$\boldsymbol{\alpha}_t = (\mu_t, \beta_t, S_t, S_{t-1}, ..., S_{t-10})'$$

el modelo se escribe en el espacio de los estados con una ecuación de observación

$$z_t = (1, 0, 1, 0, ..., 0)\boldsymbol{\alpha}_t + a_t$$

y una ecuación de estado:

$$\begin{bmatrix} \mu_t \\ \beta_t \\ S_t \\ ... \\ S_{t-9} \\ S_{t-10} \end{bmatrix} = \begin{bmatrix} 1 & 1 & 0 & ... & 0 & 0 \\ 0 & 1 & 0 & ... & 0 & 0 \\ 0 & 0 & -1 & ... & -1 & -1 \\ ... & ... & ... & ... & ... & .. \\ 0 & 0 & 0 & ... & 0 & 0 \\ 0 & 0 & 0 & ... & 1 & 0 \end{bmatrix} \begin{bmatrix} \mu_{t-1} \\ \beta_{t-1} \\ S_{t-1} \\ ... \\ S_{t-10} \\ S_{t-11} \end{bmatrix} + \begin{bmatrix} v_t \\ u_t \\ \epsilon_t \\ ... \\ 0 \\ 0 \end{bmatrix}$$

La tres primeras ecuaciones del vector de estado son las (10.25), (10.26) y (10.27). Las restantes son identidades, del tipo $S_{t-j} = S_{t-j}$.

10.5.2 El filtro de Kalman

El filtro de Kalman es un algoritmo recursivo para obtener predicciones de futuras observaciones y proporciona rápidamente los errores de predicción a un paso y sus varianzas. Vamos a presentar el algoritmo en su formulación general y después indicaremos su particularización para calcular la función de verosimilitud de un proceso ARMA.

El algoritmo funciona en tres pasos. En el primero, predecimos el estado futuro a partir de la información del estado actual. En el segundo, predecimos nuevas observaciones. En el tercero, que se realiza cuando llega una nueva observación al sistema, se revisa la estimación del estado en ese instante a la vista de la nueva información. Vamos a revisar estas tres etapas.

La primera etapa es la predicción del estado futuro a partir de una estimación del estado actual. Supongamos que disponemos de la información $Z_{t-1} = \{z_1, ..., z_{t-1}\}$ y que diponemos de un estimador del vector de estado, $\widehat{\boldsymbol{\alpha}}_{t-1}$, y deseamos predecir $\widehat{\boldsymbol{\alpha}}_{t|t-1}$,la estimación futura del estado utilizando la información disponible, Z_{t-1}. Esta estimación se obtiene tomando esperanzas en 10.19 condicionadas a Z_{t-1} y obtendremos:

$$\widehat{\boldsymbol{\alpha}}_{t|t-1} = \boldsymbol{\Omega}_t \widehat{\boldsymbol{\alpha}}_{t-1} \tag{10.28}$$

donde hemos utilizado la notación $\widehat{\boldsymbol{\alpha}}_{t-1|t-1} = \widehat{\boldsymbol{\alpha}}_{t-1}$. Llamaremos $\mathbf{S}_{t|t-1}$ a la matriz de covarianzas de esta estimación:

$$\mathbf{S}_{t|t-1} = E\left[(\boldsymbol{\alpha}_t - \widehat{\boldsymbol{\alpha}}_{t|t-1})(\boldsymbol{\alpha}_t - \widehat{\boldsymbol{\alpha}}_{t|t-1})'|Z_{t-1}\right]$$

y para calcularla utilizaremos que restando de (10.19) la ecuación (10.28):

$$\boldsymbol{\alpha}_t - \widehat{\boldsymbol{\alpha}}_{t|t-1} = \boldsymbol{\Omega}_t \left(\boldsymbol{\alpha}_{t-1} - \widehat{\boldsymbol{\alpha}}_{t-1}\right) + \mathbf{u}_t,$$

y sustituyendo esta expresión en la definición de $\mathbf{S}_{t|t-1}$ y llamando $\mathbf{S}_{t-1} = \mathbf{S}_{t-1/t-1}$, se obtiene que

$$\mathbf{S}_{t|t-1} = \boldsymbol{\Omega}_t \mathbf{S}_{t-1} \boldsymbol{\Omega}_t' + \mathbf{R}_t. \tag{10.29}$$

Esta ecuación tiene un clara interpretación intuitiva: la incertidumbre al predecir el nuevo estado con información hasta $t-1$, es la suma de la incertidumbre que teníamos respecto al estado anterior con esa información, mediada por $\mathbf{S}_{t-1}$, y la incertidumbre del ruido en la ecuación de estado, $\mathbf{R}_t$. La matriz $\boldsymbol{\Omega}_t$ aparece para relacionar los componentes del estado en el instante $t-1$ y en el instante t. Si esta matriz fuese la identidad, el estado evoluciona como un paseo aleatorio, $\boldsymbol{\Omega}_t = \mathbf{I}$, la incertidumbre de estimación del estado aumentaría continuamente a través de la suma de la matriz $\mathbf{R}_t$. Como en general esta matriz no es

la identidad, el aumento de incertidumbre depende de su estructura. Por ejemplo, supongamos un AR(1). Entonces el vector de estado es escalar, y $\Omega_t = \phi < 1$. La varianza de la estimación sigue el proceso

$$s_{t|t-1} = \phi^2 s_{t-1} + \sigma^2$$

y solamente una parte de la incertidumbre de $t-1$ se traslada al instante t.

El segundo paso del filtro es la predicción de la nueva observación $\mathbf{z}_t$ dada la información hasta $t-1$. Esta predicción se calcula de nuevo con la esperanza condicional dada Z_{t-1} y obtenemos:

$$\widehat{\mathbf{z}}_{t|t-1} = E(\mathbf{z}_t|Z_{t-1}) = \mathbf{H}_t\widehat{\boldsymbol{\alpha}}_{t|t-1} \tag{10.30}$$

Esta predicción tendrá una incertidumbre que mediremos por la matriz de varianzas y covarianzas de los errores de predicción:

$$\mathbf{e}_t = \mathbf{z}_t - \widehat{\mathbf{z}}_{t|t-1}$$

definida por:

$$\mathbf{P}_{t|t-1} = E\left[\mathbf{e}_t\mathbf{e}_t'\right].$$

Para calcular esta matriz restando la predicción (10.30) de la ecuación de observación (10.18), tenemos que:

$$\mathbf{e}_t = \mathbf{z}_t - \widehat{\mathbf{z}}_{t|t-1} = \mathbf{H}_t(\boldsymbol{\alpha}_t - \widehat{\boldsymbol{\alpha}}_{t|t-1}) + \boldsymbol{\epsilon}_t \tag{10.31}$$

y sustituyendo esta expresión en la definición de $\mathbf{P}_{t|t-1}$ se obtiene que

$$\mathbf{P}_{t|t-1} = \mathbf{H}_t\mathbf{S}_{\mathbf{t|t-1}}\mathbf{H}_t' + \mathbf{V}_t. \tag{10.32}$$

Esta ecuación, indica que la incertidumbre de la predicción acumula la incertidumbre en el estado y la del error de medida de la ecuación de observación. El error de predicción que viene de la estimación del estado se modula dependiendo de la matriz $\mathbf{H}_t$. Si esta matriz es la identidad, lo que supone que las observaciones $\mathbf{z}_t$ son mediciones de las variables de estado más un error aleatorio, al error de las variables de estado se añade el error de medición de las observaciones.

El tercer y último paso del filtro es revisar la estimación del estado a la vista de la nueva información. Supongamos que se ha observado $\mathbf{z}_t$ con lo que la información disponible pasa a ser $Z_t = (Z_{t-1}, \mathbf{z}_t)$. La nueva estimación del estado, $\widehat{\boldsymbol{\alpha}}_t = \widehat{\boldsymbol{\alpha}}_{t|t} = E(\boldsymbol{\alpha}_t|Z_t)$, se calcula por regresión como

$$E(\boldsymbol{\alpha}_t|Z_{t-1}, \mathbf{z}_t) = E(\boldsymbol{\alpha}_t|Z_{t-1}) + cov(\boldsymbol{\alpha}_t, \mathbf{z}_t|Z_{t-1})var(\mathbf{z}_t|Z_{t-1})^{-1}(\mathbf{z}_t - E(\mathbf{z}_t|Z_{t-1})). \tag{10.33}$$

En esta ecuación las esperanzas $E(\boldsymbol{\alpha}_t|Z_{t-1}) = \widehat{\boldsymbol{\alpha}}_{t|t-1}$ y $E(\mathbf{z}_t|Z_{t-1}) = \widehat{\mathbf{z}}_{t|t-1}$ son conocidas, así como la matriz $var(\mathbf{z}_t|Z_{t-1}) = \mathbf{P}_{t|t-1}$. Lo único que queda por calcular es la covarianza entre el estado y la nueva observación que viene dado por

$$cov(\boldsymbol{\alpha}_t, \mathbf{z}_t|Z_{t-1}) = E\left[(\boldsymbol{\alpha}_t - \widehat{\boldsymbol{\alpha}}_{t|t-1})(\mathbf{z}_t - \widehat{\mathbf{z}}_{t|t-1})'\right] = E\left[(\boldsymbol{\alpha}_t - \widehat{\boldsymbol{\alpha}}_{t|t-1})\mathbf{e}_t'\right]$$

y sutituyendo (10.31)

$$cov(\boldsymbol{\alpha}_t, \mathbf{z}_t|Z_{t-1}) = E\left[(\boldsymbol{\alpha}_t - \widehat{\boldsymbol{\alpha}}_{t|t-1})((\boldsymbol{\alpha}_t - \widehat{\boldsymbol{\alpha}}_{t|t-1})'\mathbf{H}_t' + \boldsymbol{\epsilon}_t')\right] = \mathbf{S}_{\mathbf{t|t-1}}\mathbf{H}_t'. \tag{10.34}$$

ya que el error de observación $\boldsymbol{\epsilon}_t'$ es ruido blanco e independiente de $\boldsymbol{\alpha}_t - \widehat{\boldsymbol{\alpha}}_{t|t-1}$. Sustituyendo esta covarianza en (10.33), podemos escribir:

$$\widehat{\boldsymbol{\alpha}}_t = \widehat{\boldsymbol{\alpha}}_{t|t-1} + \mathbf{K}_t(\mathbf{z}_t - \widehat{\mathbf{z}}_{t|t-1}) \tag{10.35}$$

donde $\mathbf{K}_t$ es la matriz de coeficientes de regresión que se llama la ganancia del filtro, y viene dada por:

$$\mathbf{K}_t = \mathbf{S}_{\mathbf{t|t-1}}\mathbf{H}_t'\mathbf{P}_{t|t-1}^{-1}.$$

La ecuación (10.35) indica que la revisión que hacemos de la estimación previa del estado depende del error de predicción, $\mathbf{e}_t = \mathbf{z}_t - \widehat{\mathbf{z}}_{t|t-1}$. Si este error es cero no modificamos la estimación, en otro caso hacemos una modificación de la estimación del estado que depende del cociente entre el error en la estimación del

estado, $\mathbf{S}_{t|t-1}$, y el error de predicción $\mathbf{P}^{-1}_{t|t-1}$. La matriz $\mathbf{H}'_t$ permite comparar estas matrices. Una forma equivalente de escribir la ecuación (10.35) es

$$\widehat{\boldsymbol{\alpha}}_t = (\mathbf{I} - \mathbf{K}_t\mathbf{H}_t)\widehat{\boldsymbol{\alpha}}_{t|t-1} + \mathbf{K}_t\mathbf{z}_t$$

que indica que la estimación del estado es una combinación lineal de las dos fuentes de información de que disponemos. Por un lado, la estimación previa, $\widehat{\boldsymbol{\alpha}}_{t|t-1}$, y por otra, la observación $\mathbf{z}_t$ que también aporta información sobre el estado. Puede demostrarse (véase Peña, 1995) que las ponderaciones de las dos fuentes de información son iguales a su precisión relativa. La matriz de covarianzas de esta estimación será

$$\mathbf{S}_t = E\left[(\boldsymbol{\alpha}_t - \widehat{\boldsymbol{\alpha}}_t)(\boldsymbol{\alpha}_t - \widehat{\boldsymbol{\alpha}}_t)'|Z_t\right]$$

y sustituyendo $\widehat{\boldsymbol{\alpha}}_t$ por su expresión en la ecuación (10.35), tenemos que

$$\mathbf{S}_t = E\left[(\boldsymbol{\alpha}_t - \widehat{\boldsymbol{\alpha}}_{t|t-1} - \mathbf{K}_t\mathbf{e}_t)(\boldsymbol{\alpha}_t - \widehat{\boldsymbol{\alpha}}_{t|t-1} - \mathbf{K}_t\mathbf{e}_t)'|Z_t\right]$$

y utilizando (10.34) y operando se obtiene finalmente que

$$\mathbf{S}_t = \mathbf{S_{t|t-1}} - \mathbf{S_{t|t-1}}\mathbf{H}'_t\mathbf{P^{-1}_{t|t-1}}\mathbf{H}_t\mathbf{S_{t|t-1}}. \tag{10.36}$$

Las ecuaciones (10.28), (10.29), (10.30), (10.32), (10.35) y (10.36) constituyen el filtro de Kalman.

La aplicación del filtro para obtener la función de verosimilitud del modelo ARMA requiere escribir el modelo en el espacio de los estados, como hemos visto, y calcular los errores de prediccion $e_t = z_t - \widehat{z}_{t|t-1}$, que en este caso son escalares y sus varianzas, $\mathbf{P}_{t/t-1}$. En este caso la matriz $\mathbf{V}_t$ es cero. Para comenzar el filtro hace falta indicar un valor inicial para las variables de estado, $\boldsymbol{\alpha}_0$, y para su matriz de covarianzas $\mathbf{S_0}$. Estos valores iniciales no son cruciales, porque el filtro depende poco de las condiciones iniciales. El lector interesado en la implantación del algoritmo puede acudir a Harvey (1989), y Gómez y Maravall (1994).

10.6 Propiedades de los estimadores

Puede demostrarse que las propiedades asintóticas del método de máxima verosimilitud son válidas, en condiciones de regularidad generales, para los estimadores MV de modelos ARMA. Estas condiciones exigen que el proceso sea estacionario y que el modelo ARMA que estimamos no contiene factores comunes en su parte AR y MA. Ya vimos en el capítulo anterior que si un proceso AR(1) tiene el parámetro igual a la unidad, el valor verdadero está en la frontera del intervalo parámetrico [0,1], y no se cumplen las condiciones habituales de regularidad necesarias para obtener las propiedades asintóticas del estimado MV.

Para procesos estacionarios en muestras grandes los estimadores MV tendrán distribución aproximadamente normal y serán centrados y eficientes. En particular, la matriz de segundas derivadas del soporte en su máximo proporciona directamente las varianzas y covarianzas de los estimadores:

$$\mathbf{Var}\left(\widehat{\boldsymbol{\beta}}_{MV}\right) = -\left[\frac{\partial^2 L\left(\widehat{\boldsymbol{\beta}}_{MV}\right)}{\partial\boldsymbol{\beta}\partial\boldsymbol{\beta}'}\right]^{-1}$$

Este resultado se basa en que la función de verosimilitud es aproximadamente cuadrática en el máximo. Para comprobarlo, conviene siempre que sea posible estudiar la forma de la función soporte alrededor del estimador MV, calculando numéricamente y dibujando sus curvas de nivel.

La condición de que no existan factores comunes en la parte AR y MA es importante. Por ejemplo, si ω_t es ruido blanco y estimamos el modelo

$$(1 - \phi B)\,\omega_t = (1 - \theta B)\,a_t$$

todos los valores de los parámetros con la condición $\phi = \theta$ son compatibles con los datos y puede demostrarse que la varianza de los estimadores es infinita. En general, si el modelo está sobreparametrizado y contiene simultáneamente factores redundantes AR y MA tendremos una situación de fuerte multicolinealidad que

puede dar lugar, primero, a múltiples máximos en la función de verosimilitud y, segundo, a que la aproximación cuadrática en el máximo sea inadecuada.

Ejemplo 10.1

Vamos a estimar los modelos identificados para la serie de matriculación de vehículos en España que se analizó en el capítulo anterior. Utilizaremos tres programas. El programa TSW, que utiliza el filtro de Kalman y maxima verosimilitud exacta, el SCA que también utiliza MV exacta y Minitab que lo hace condicional. Se observa que los dos primeros programas dan valores muy similares, mientras que con el tercero la estimación de los términos estacionales no es muy precisa. El TSW sólo permite AR estacionales de orden uno, por lo que no hemos podido estimar el AR(2) estacional con dicho programa. Se observa que el mejor modelo desde el punto de vista de la varianza residual es el que tiene un ARMA(1,1) para la parte estacional.

Programa	Modelo	$\widehat{\sigma}_a$
TSW	$\nabla\nabla_{12}\ln M_t = \underset{(0,04)}{(1-0.61B)}\underset{(0,03)}{\left(1-0.78B^{12}\right)}a_t$	0.123
TSW	$\underset{(0,06)}{\left(1-0.21B^{12}\right)}\nabla\nabla_{12}\ln M_t = \underset{(0,06)}{(1-0.61B)}\underset{(0,05)}{\left(1-0.89B^{12}\right)}a_t$	0.121
SCA	$\nabla\nabla_{12}\ln M_t = \underset{(0,04)}{(1-0.61B)}\underset{(0,03)}{\left(1-0.78B^{12}\right)}a_t$	0.122
SCA	$\underset{(0,05)}{\left(1-0.14B^{12}\right)}\nabla\nabla_{12}\ln M_t = \underset{(0,04)}{(1-0.61B)}\underset{(0,05)}{\left(1-0.85B^{12}\right)}a_t$	0.117
SCA	$\left(1+\underset{(.04)}{0.53}B^{12}+\underset{(.04)}{.32}B^{24}\right)\nabla\nabla_{12}\ln M_t = \underset{(0,04)}{(1-0.62B)}a_t$	0.122
Minitab	$\nabla\nabla_{12}\ln M_t = \underset{(0,03)}{(1-0.62B)}\underset{(0,03)}{\left(1-0.84B^{12}\right)}a_t$	0.119
Minitab	$\underset{(0,06)}{(1-0.25B)}\nabla\nabla_{12}\ln M_t = \underset{(0,06)}{(1-0.62B)}\underset{(0,03)}{\left(1-0.95B^{12}\right)}a_t$	0.116

10.7 Criterios de Selección de modelos

Supongamos que hemos estimado un conjunto de modelos $M_1, ..., M_m$ y deseamos seleccionar el que mejor explica la serie observada. El criterio de ajuste dentro de la muestra no resulta adecuado, ya que claramente si comparamos un AR(1) con un AR(2) el modelo con más parámetros siempre conducirá a una mayor verosimilitud y a una menor suma de cuadrados de los errores dentro de la muestra. Por tanto, para seleccionar entre modelos debemos acudir a otros principios.

El problema puede verse como un problema de discriminación: tenemos distintos modelos M_i y una serie, $\mathbf{z}_T = (z_1, ..., z_T)$, y queremos seleccionar el modelo más compatible con la serie observada. El problema puede abordarse desde el punto de vista clásico o Bayesiano. Comenzando con el enfoque clásico, vemos que no es útil comparar la verosimilitud de distintos modelos, porque siempre el modelo con más parámetros tendrá mayor verosimilitud. Podemos sin embargo calcular el valor esperado de la verosimilitud para cada uno de los modelos, es decir, el valor que esperamos obtener para la verosimilitud sobre muchas realizaciones del proceos si en cada una estimamos los parámetros por máxima verosimilitud (véase Galeano y Peña, 2004) y seleccionar aquel modelo que produzca un valor esperado más alto de esta verosimilitud esperada. Este es el enfoque que conduce al criterio de Akaike, que exponemos a continuación.

Si disponemos de probabilidades a priori para cada modelo, $P(M_i)$, podríamos utilizar el enfoque Bayesiano y seleccionar el modelo con probabilidad máxima dados los datos. Es decir calculamos

$$P(M_i|\mathbf{z}_T) = \frac{P(\mathbf{z}_T|M_i)P(M_i)}{\sum_{i=1}^{m} P(\mathbf{z}_T|M_j)P(M_j)} \tag{10.37}$$

y seleccionamos el modelo más probable a la vista de los datos. Observemos que este planteamiento no requiere que la serie sea estacionaria, por lo que puede aplicarse para comparar modelos con distinto número de diferencias y utilizarse para decidir como alternativa a los contrastes de raices unitarias. Si suponemos que las probabilidades a priori de todos los modelos son las mismas, este enfoque conduce al criterio BIC, que presentamos a continuación.

Un enfoque alternativo para seleccionar el mejor modelo es validación cruzada. En este enfoque estimamos el modelo en una parte de la muestra y utilizamos la otra parte para calcular el error de predicción fuera de la muestra. Un problema importante es como dividir la muestra en estas dos partes. Puede demostrarse que en muestras grandes existe una equivalencia entre los métodos de validación cruzada y los criterios de selección de modelos.

10.7.1 El criterio AIC de Akaike

La función de verosimilitud de un modelo ARIMA viene dada por (10.4). Multiplicando por -2 y tomando esperanzas en esta expresión tenemos que

$$E(-2L(\boldsymbol{\beta})) = T\log\sigma^2 + \sum_{t=1}^{T}\log v_{t|t-1} + E\left[\sum_{t=1}^{T}\frac{e_t^2}{\sigma^2 v_{t|t-1}}\right]$$

Se demuestra en el apéndice 10.2 que si: (1) suponemos que los parámetros no son conocidos, sino que se estima con los datos;y (2) calculamos la esperanza de la expresión de la verosimilitud anterior, llamando k al número de parámetros estimados para calcular las predicciones a un paso, se obtiene que

$$AIC = E(-2L(\boldsymbol{\beta})) = T\log\widehat{\sigma}^2_{MV} + 2k. \tag{10.38}$$

donde T es el tamaño muestral utilizado para estimar el modelo, $\widehat{\sigma}^2_{MV}$ el estimador MV de la varianza de las innovaciones y k el número de parámetros. Por tanto, seleccionar el modelo qut tenga la verosimilitud esperada máxima equivale a escoger el que minimiza la verosimilitud con signo negativo, que es el que minimiza el criterio AIC dado por (10.38). Este criterio se conoce como criterio AIC, y es debido a Akaike.

El problema con el AIC es que tiende a sobre estimar el número de parámetros en el modelo y este afecto puede ser muy grande en pequeñas muestras. Una alternativa que corrige esta sobrestimación es el criterio AIC corregido, AICC, dado por

$$AICC = T\log\widehat{\sigma}^2_{MV} + T\frac{(1+k/T)}{1-(k+2)/T}.$$

Si utilizamos el citerio AIC para comparar modelos ajustados a una serie es importante que T, el número efectivo de observaciones utilizado para estimar el modelo sea el mismo para todos ellos. El número de datos estacionarios es igual a los datos originales menos $d - sD$, siendo d el número de diferencias regulares, s el periodo estacional y D el número de diferencias estacionales. Si estimamos el modelo por máxima verosimilitud exacta podemos calcular residuos en todos los puntos y por tanto el número de datos efectivo es $T - d - SD$. Si consideramos modelos con distinto número de diferencias y llamamos $d_{\max}$ y $D_{\max}$ los grados de diferenciación más altos de los modelos que comparamos, entonces el número de datos efectivos es

$$T = T_0 - d_{\max} - sD_{\max} \tag{10.39}$$

Sin embargo, si utilizasemos un programa que realiza estimación condicionada para calcular los residuos no tenemos estos datos sino un número menor, ya que, por ejemplo, para un AR(1) como condicionamos al primer dato el residuo para esa observación no puede calcularse. En general, si hacemos estimación condicional y llamamos $r = \max(p,q)$, y $R = \max(P,Q)$ el número efectivo de observaciones con las que podemos calcular los residuos en un modelo dado es $T - d - sD - r - R$, y puede variar mucho de unos modelos a otros. Esta es una razón adicional de utilizar máxima verosimilitud exacta, de manera que siempre consideramos los residuos en las mismas T observaciones en las comparaciones entre los modelos. Si no hacemos esto, es posible que parte de las diferencias observadas entre los modelos se deban a los distintos residuos calculados.

10.7.2 El criterio BIC

Un criterio alternativo ha sido propuesto por Schwarz (1978) desde el enfoque Bayesiano, aproximando asintóticamente las probabilidades a posteriori de cada modelo supuesto que las probabilidades a priori son

Modelo	$\widehat{\sigma}^2$	T	k	BIC	AIC
ARIMA(0,1,1)×$(0,1,1)_{12}$	0.122^2	466	2	8.08	-0.20
ARIMA(0,1,1)×$(1,1,1)_{12}$	0.117^2	466	3	14.14	1.70
ARIMA(0,1,1)×$(2,1,0)_{12}$	0.122^2	466	3	14.22	1.79

Tabla 10.1: Valores de los criterios de seleccion con la serie de matriculación

las mismas para todos los modelos. Como según (10.37) $P(M_i|\mathbf{z})$ es proporcional a $P(\mathbf{z}|M_i)P(M_i)$, si las probabilidades a priori son las mismas la probabilidad del modelo es proporcional a $P(\mathbf{z}|M_i)$. Seleccionar el modelo que maximice esta probabilidad es equivalente a seleccionar el modelo que minimiza -2log$P(\mathbf{z}|M_i)$. Puede demostrarse, véase el apéndice 10.3, que sustituyendo los parámetros por sus estimaciones máximo verosímiles el modelo que minimiza esta cantidad es el que minimiza el criterio :

$$BIC = T \log \widehat{\sigma}^2_{MV} + k \log T \tag{10.40}$$

donde, como en el caso anterior, T es el tamaño muestral, $\widehat{\sigma}^2_{MV}$ el estimador MV de la varianza y k el número de parámetros. Si comparemos esta expresion con la (10.38) vemos que el BIC tiene una penalización mayor que el AIC por introducir nuevos parámetros, con lo que tiende a elegir modelos más parsimoniosos. La diferencia entre ellos puede ser grande si T es grande.

Para calcular el BIC de varios modelos ajustados a una serie hay que tener en cuenta que el número de observaciones de todos los modelos debe ser el mismo para tener comparaciones homogéneas y evitar que las diferencias entre modelos se deban a diferencias entre residuos includos en unos modelos y no en otros. El tamaño T debe calcularse con (10.39).

10.7.3 Comparacion entre criterios

Los criterios estudiados pueden expresarse, dividiendo por T las expresiones anteriores, como

$$\text{mimimizar}(\log \widehat{\sigma}^2_{MV} + kc(k,T))$$

donde $c(k,T)$ es un término de corrección que tiende a cero al aumentar el tamaño muestral. El criterio AIC toma $c(k,T) = 2/T$, y el BIC $c(k,T) = (\log T)/T$. Puede demostrarse que el criterio BIC es consistente, en el sentido de que cuando los datos han sido generados por un modelo ARIMA el BIC selecciona el orden adecuado del modelo con probabilidad uno. Por el contrario, el criterio AIC es eficiente, en el sentido de que si los datos han sido generados por un modelo que puede ser de orden infinito, y consideramos una secuencia de estimadores con orden que aumenta con el tamaño muestral, el predictor seleccionado es el de menor error de predicción esperado.

La diferencia entre estas dos propiedades es que el BIC supone un modelo correcto entre los estimados, mientras que el AIC admite que el modelo correcto puede no ser de orden finito. Además el BIC es una aproximación a un criterio que puede aplicarse exactamente, calcular las probabilidades a posteriori y seleccionar la más alta, mientras que AIC supone una situación más ficticia asintótica donde vamos incrementado el orden del modelo con el tamaño muestral, que puede ser menos relevante ante una muestra concreta. Existe mucha evidencia de que el criterio AIC tiende a seleccionar un orden demasiado alto, y recomendamos el criterio BIC que funciona mejor en la práctica. Por otro lado conviene notar que no siempre el modelo correcto es el que proporciona las mejores predicciones. Por ejemplo, Sánchez y Peña (2001) han demostrado que cuando los datos se generan con un AR(p+1), con una raiz próxima a la unidad, las predicciones obtenidas diferenciando y estimando un AR(p) tienen menor error cuadrático medio que las del AR(p+1). La razón es que cuando el coeficiente está próximo a uno, al fijarlo en uno por la diferencia el error es menor que al estimarlo en la muestra.

Ejemplo 10.2

Vamos a aplicar los criterios de seleccion de modelos para seleccionar el mejor de los tres estimados en el ejercicio 10.1 La tabla 10.1 indica el modelo, la varianza residual, el número de parámetros y el valor del criterio de selección correspondiente. La serie tiene 479 datos menos 13 debido a las diferencias, 466, ya que hemos utilizado máxima verosimilitud exacta y ese el valor de T que utilizaremos en las comparaciones.

10.8 Lecturas Complementarias

El libro de Harvey (1989) desarrolla con detalle la representación en el espacio de los estados y el filtro de Kalman. Una referencia clásica sobre este tema es Anderson y Moore (1979). La estimación Bayesiana de modelos en el espacio de los estados puede consultarse en West y Harrison (1997). Fuller (1996) describe con detalle las propiedades de los estimadores . Buenas presentaciones de la estimación pueden verse en Brockwell and Davis (1991) y Shumway and Stoffer (2000).

Para los métodos de selección de modelos véase Galeano y Peña (2004). La relación entre validación cruzada y selección de modelos fué encontrada por Stone(1979). Para métodos actuales de realizar la validación cruzada con modelos ARIMA y su relación con criterios de selección de modelos véase Peña y Sánchez (2004).

Ejercicios 10

10.1 Demostrar que la varianza del estimador de μ para un AR(p) obtenida por estimación condicionada es mayor que la varianza de la media muestral de todo el proceso.

10.2 Demostrar que en un proceso AR(2) la distribución marginal de la primera observación es normal con media $E[\omega_1] = \mu$ y $Var[\omega_1] = \dfrac{\sigma^2}{1-\phi_1^2-\phi_2^2}$ y que para la segunda observación $E[\omega_2|\omega_1] = \mu + \phi_1(\omega_1 - \mu)$ y $Var[\omega_2|\omega_1] = \sigma^2 \dfrac{(1-\phi_1^2)}{1-\phi_1^2-\phi_2^2}$.

10.3 Demostrar que el estimador MV de $\widehat{\sigma}_a^2$ para un proceso AR(p) es $\widehat{\sigma}_a^2 = \sum_{t=p+1}^{T} \left(\omega_t - \widehat{\mu} - \Sigma_{i=1}^{p} \widehat{\phi}_i(\omega_{t-i} - \widehat{\mu}\right)^2 /T$

10.3 Comprobar que la representación en el espacio de los estados de un modelo AR(1) tiene $\Omega = \phi$, $H = 1$ y $R = \sigma^2$.

10.4 Escribir las ecuaciones del filtro de Kalman para prever con un AR(1) y comprobar que se reducen a $\widehat{z}_{t|t-1} = \phi z_t$ con varianza $p_{t|t-1} = \sigma^2$.

10.5 Escribir las ecuaciones en el espacio de los estados para un MA(1) y comprobar que se verifica $\Omega = \begin{bmatrix} 0 & 1 \\ 0 & 0 \end{bmatrix}$, H=(0,1) y $R = \sigma^2 \begin{bmatrix} 1 & -\theta \\ -\theta & \theta^2 \end{bmatrix}$.

10.6 Utilizando que para matrices cuadradas A, C y rectangulares con las dimensiones apropiadas B, D se verifica que $(A+BCD)^{-1} = A^{-1} - A^{-1}B(DA^{-1}B + C^{-1})^{-1}DA^{-1}$, demostrar que la ecuación de revisión de las covarianzas de estimación del estado puede escribirse como $\mathbf{S}_t^{-1} = \mathbf{S}_{\mathbf{t|t-1}}^{\mathbf{-1}} + \mathbf{H}_t^{'}\mathbf{V}_{\mathbf{t}}^{\mathbf{-1}}\mathbf{H}_t$.

10.7 Llamando precisión a la inversa de la varianza. justificar que la expresión anterior se interpreta como que la precisión final es la suma de la precisión inicial y la precisión aportada por la última observación.

10.8 Escribir las ecuaciones del filtro de Kalman para un AR(2) y relacionar el método de cálculo de las predicciones con el estudiado en el capítulo 8.

Apéndice 10.1 Algoritmos de Optimización no lineal

Supongamos una función $f(\boldsymbol{\theta})$ con primeras y segundas derivadas continuas cuyo mínimo se quiere calcular. (Si se tratara de su máximo convertiríamos el problema en hallar el mínimo de $-f(\boldsymbol{\theta})$.) La condición necesaria y suficiente para que $\boldsymbol{\theta}^*$ sea un mínimo local es que el vector de primeras derivadas, que llamaremos vector gradiente, sea nulo en $\boldsymbol{\theta}^*$y que la matriz de segundas derivadas (hessiana) sea definida positiva en $\boldsymbol{\theta}^*$. Llamando:

$$\mathbf{g}(\boldsymbol{\theta}) = \frac{\partial f(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}_i}$$

al vector gradiente y

$$\mathbf{H}(\boldsymbol{\theta}) = \frac{\partial^2 f(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}_i \partial \boldsymbol{\theta}_i}$$

a la matriz hessiana, las condiciones de mínimo local son:

$$\mathbf{g}(\boldsymbol{\theta}^*) = \mathbf{0}$$

$$\mathbf{H}(\boldsymbol{\theta}^*) = \text{ definida positiva}$$

Los algoritmos de optimización no lineal son procedimientos iterativos para pasar de un valor $\boldsymbol{\theta}_i$ a otro $\boldsymbol{\theta}_{i+1}$ más próximo al mínimo, según la relación:

$$\boldsymbol{\theta}_{i+1} = \boldsymbol{\theta}_i + \lambda_i \mathbf{d}_i \tag{10.41}$$

siendo $\mathbf{d}_i$ un vector de dirección y λ_i la "amplitud" de paso. El valor final $\boldsymbol{\theta}_{i+1}$ se convierte en el inicial de la iteracción siguiente, y el proceso continúa hasta obtener convergencia, definida por $|\boldsymbol{\theta}_{i+1} - \boldsymbol{\theta}_i| < \varepsilon_1$, $|f(\boldsymbol{\theta}_{i+1}) - f(\boldsymbol{\theta}_i)| < \varepsilon_2$ y $|\mathbf{g}(\boldsymbol{\theta}_{i+1})| < \varepsilon_3$, siendo ε_1, ε_2, ε_3, valores pequeños fijados previamente, que dependen de la precisión deseada.

Los algoritmos difieren principalmente por su elección de $\mathbf{d}_i$, ya que λ_i puede determinarse siempre por una búsqueda en dicha dirección. Una posible elección de $\mathbf{d}_i$ es:

$$\mathbf{d}_i = -\mathbf{g}(\boldsymbol{\theta})$$

ya que una función disminuye siempre con un pequeño movimiento en la dirección del gradiente negativo. Este es el *método del gradiente* que consiste en calcular este vector y desplazarse en la dirección asociada a sus valores negativos, determinando λ_i de manera que la función disminuya lo máximo posible en cada iteración. Este método permite avanzar rápidamente cuando nos encontramos lejos del mínimo pero es muy lento al acercarnos a éste.

Un mejor procedimiento cuando partimos de un valor próximo al mínimo, es el de *Newton–Raphson*, que se basa en que cerca del óptimo la función debe ser aproximadamente cuadrática. Aproximando la función en el punto de llegada de la iteración $\boldsymbol{\theta}_{i+1}$, por una función cuadrática:

$$f(\boldsymbol{\theta}_{i+1}) \simeq f(\boldsymbol{\theta}_i) + (\boldsymbol{\theta}_{i+1} - \boldsymbol{\theta}_i)' \mathbf{g}(\boldsymbol{\theta}_i) + 1/2 (\boldsymbol{\theta}_{i+1} - \boldsymbol{\theta}_i)' \mathbf{H}(\boldsymbol{\theta}_i) (\boldsymbol{\theta}_{i+1} - \boldsymbol{\theta}_i) \tag{10.42}$$

derivando esta expresión respecto a $\boldsymbol{\theta}_{i+1}$ y considerando constantes los términos que dependen de $\boldsymbol{\theta}_i$, se obtiene que:

$$\mathbf{g}(\boldsymbol{\theta}_{i+1}) \simeq \mathbf{g}(\boldsymbol{\theta}_i) + \mathbf{H}(\boldsymbol{\theta}_i)(\boldsymbol{\theta}_{i+1} - \boldsymbol{\theta}_i) \tag{10.43}$$

es decir, aproximar la función por una cuadrática equivale a aproximar el gradiente linealmente en un entorno de $\boldsymbol{\theta}_{i+1}$. Imponiendo la condición de que $\boldsymbol{\theta}_{i+1}$ sea un mínimo local, $\mathbf{g}(\boldsymbol{\theta}_{i+1}) = 0$, y despejando en (10.43) resulta el algoritmo:

$$\boldsymbol{\theta}_{i+1} = \boldsymbol{\theta}_i - \mathbf{H}^{-1}(\boldsymbol{\theta}_i) \mathbf{g}(\boldsymbol{\theta}_i) \tag{10.44}$$

que es el algoritmo de Newton-Raphson.

Este algoritmo puede presentar problemas cuando se comienza con un valor inicial lejos del mínimo y la aproximación cuadrática no es buena. Por ejemplo, si la función es aproximadamente lineal la matriz hessiana será nula, y el método fallará. Además, aunque la matriz hessiana, $\mathbf{H}(\boldsymbol{\theta}_i)$, no sea singular, es posible que lejos del mínimo no sea definida positiva, con lo que el valor en $\boldsymbol{\theta}_{i+1}$ puede ser mayor que en $\boldsymbol{\theta}_i$, y el método puede no converger. En contrapartida, cerca del mínimo es un buen algoritmo, aunque costoso computacionalmente ya que requiere el cálculo de primeras y segundas derivadas.

Se han desarrollado métodos para aproximar (10.44) utilizando una matriz $\mathbf{H}$ que no tenga que calcularse en cada etapa. Si f es una función soporte, una solución propuesta por R. A. Fisher es tomar en todas las iteraciones el valor esperado de $\mathbf{H}$, que es la matriz de información. Este algoritmo, conocido por el *algoritmo de la tasa de discriminación*, (scoring method) es:

$$\boldsymbol{\theta}_{i+1} = \boldsymbol{\theta}_i + \mathbf{IE}(\boldsymbol{\theta}_i)^{-1} \mathbf{g}(\boldsymbol{\theta}_i) \tag{10.45}$$

donde $\mathbf{IE}(\boldsymbol{\theta})$es el valor esperado de la matriz $(-\mathbf{H}(\boldsymbol{\theta}))$.

En la estimación condicional de modelos ARMA la función a minimizar puede escribirse

$$L(\boldsymbol{\theta}) = \sum a_t^2(\boldsymbol{\theta})$$

y equivale a minimizar una suma de cuadrados. Para aplicar el procedimiento de Newton-Raphson a esta función es necesario calcular su gradiente, que en este caso será:

$$\mathbf{g}(\boldsymbol{\theta}) = \sum 2a_t(\boldsymbol{\theta}) \frac{\partial a_t(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}} = \sum 2a_t(\boldsymbol{\theta}) \mathbf{g}(a_t) \tag{10.46}$$

llamando $\mathbf{g}(a_t)$ al vector de primera derivadas de a_t respecto a cada componente del vector $\boldsymbol{\theta}$. El hessiano será, derivando de nuevo en (10.46):

$$\mathbf{H}(\boldsymbol{\theta}) = 2\sum \mathbf{g}(a_t)\mathbf{g}'(a_t) + 2\sum a_t(\boldsymbol{\theta})\mathbf{H}(a_t)$$

siendo $\mathbf{H}(a_t)$ la matriz hessiana de segundas derivadas de a_t respecto a los términos de $\boldsymbol{\theta}$. Supongamos ahora que $\mathbf{H}(a_t)$ es despreciable en cada iteración frente al primer término de primeras derivadas. Esto equivale a una aproximación lineal para los residuos del tipo:

$$a_{t,i+1} = a_t(\boldsymbol{\theta}_{i+1}) \overset{\circ}{=} a_t(\boldsymbol{\theta}_i) + (\boldsymbol{\theta}_{i+1} - \boldsymbol{\theta}_i)' \mathbf{g}_i(a_t) \tag{10.47}$$

ya que suponemos $\mathbf{H}(a_t)$ despreciable frente a $\mathbf{g}(a_t)$. Entonces, la expresión general (10.44) se reduce a:

$$\boldsymbol{\theta}_{i+1} = \boldsymbol{\theta}_i - \left[\sum \mathbf{x}_{ti}\mathbf{x}'_{ti}\right]^{-1} \sum \mathbf{x}_{ti}a_{ti} \tag{10.48}$$

donde $\mathbf{x}_{ti} = \mathbf{g}(a_{ti})$ es el vector de derivadas de los residuos respecto a los parámetros $\boldsymbol{\theta}_i$ y a_{ti} los residuos calculados con ellos. Esta ecuación indica que pasaremos de un punto al siguiente añadiendo al valor inicial el vector de parámetros resultante de realizar una regresión con variable dependiente a_{ti} $(t = 1, ..., n)$ y vector de regresores $\mathbf{x}_{ti}$. Este mismo resultado se deduce de (10.47). Escribiendo:

$$a_{ti} = a_t(\boldsymbol{\theta}_i) = \mathbf{x}'_{ti}\boldsymbol{\beta} + a_{t,t+1} \tag{10.49}$$

siendo $\boldsymbol{\beta} = -(\boldsymbol{\theta}_{i+1} - \boldsymbol{\theta}_i)$, es claro que $\boldsymbol{\beta}$ se obtiene por regresión y que los nuevos residuos resultantes, $a_{t,t+1}$ tendrá siempre menor varianza que la variable dependiente –a_{ti}–, con lo que avanzaremos hacia el mínimo de la suma de cuadrados. Este método se conoce como algoritmo de Gauss-Newton.

El algoritmo (10.48) puede mejorarse en dos aspectos: cuando estamos lejos del óptimo el avance a través de aproximaciones cuadráticas es muy lento mientras que, como sabemos el método del gradiente proporciona una dirección de avance más rápida. Prescindiendo de subíndices y llamando $\mathbf{X}'\mathbf{X} = \sum \mathbf{x}_i\mathbf{x}'_i$; $\mathbf{X}'\mathbf{Y} = \sum \mathbf{x}_t a_{t.}$, esto sugiere un esquema del tipo:

$$\boldsymbol{\theta}^* = \boldsymbol{\theta} - [\mathbf{X}'\mathbf{X} + \delta\mathbf{I}]^{-1}\mathbf{X}'\mathbf{Y} \tag{10.50}$$

donde δ se determina en función del avance de la función y es muy grande al principio –para que $\mathbf{X}'\mathbf{X}$ sea despreciable frente a δ veces la matriz unidad $\mathbf{I}$, y la dirección sea la del gradiente $\mathbf{g}(\boldsymbol{\theta})$ (10.46)- y muy pequeña cerca del mínimo, con lo que (10.50) equivale a Gauss-Newton. Las iteraciones pueden interpretarse ahora como realizar regresiones-cresta (ridge regression). Este algoritmo es debido a Marquard (1963).

La implantación de estos algoritmos para estimar procesos ARMA es directa. Por ejemplo, la estimación condicional se realiza como sigue:

1. Comenzar con un valor inicial $\widehat{\boldsymbol{\beta}}_0$ de los parámetros calculado con el algortimo de Hannan-Rissanen.

2. Calcular los residuos del modelo con el valor supuesto de los parámetros mediante (10.15).

3. Calcular las derivadas primeras de los residuos respecto al vector de parámetros. Este cálculo puede realizarse numéricamente definiendo:

$$\frac{\partial a_t\left(\widehat{\boldsymbol{\beta}}_0\right)}{\partial \beta_i} \simeq \frac{a_t\left(\widehat{\boldsymbol{\beta}}_0 + \mathbf{1}_i\varepsilon\right) - a_t\left(\widehat{\boldsymbol{\beta}}_0\right)}{\varepsilon} \quad i = 1, ..., p+1$$

   donde $\mathbf{1}_i$ es un vector con un uno en la componente i y cero en otro caso y ε es un número pequeño (0,001 o similar). El cálculo de las derivadas requiere repetir $(p+1)$ veces el paso (2), modificando cada vez ligeramente un componente del vector $\widehat{\boldsymbol{\beta}}_0$.

4. Estimar los coeficientes de una regresión entre los residuos calculados en (2) como la variable dependiente y un vector de $(p+1)$ variables explicativas $\mathbf{X}_t$, cuyos componentes son las derivadas parciales de los residuos calculadas en (3) y cambiadas de signo. Es decir:

$$x_{it} = -\frac{\partial a_t\left(\widehat{\boldsymbol{\beta}}_0\right)}{\partial \beta_i}$$

5. Sea $\mathbf{b}$ el vector de parámetros estimado con la regresión anterior que mide el efecto lineal sobre los residuos de cambios en el vector de parámetros. Definir un nuevo estimador $\widehat{\boldsymbol{\beta}}_1$ con:

$$\widehat{\boldsymbol{\beta}}_1 = \widehat{\boldsymbol{\beta}}_0 + \mathbf{b} \tag{10.51}$$

y volver al paso (2) tomando ahora $\widehat{\boldsymbol{\beta}}_1$ como estimador inicial. Repetir el proceso hasta obtener convergencia, definida por $\left|\boldsymbol{\beta}_{i+1} - \boldsymbol{\beta}_i\right| < \alpha$, y $\left|\partial L\left(\boldsymbol{\beta}_i\right)/\partial \boldsymbol{\beta}\right| < \gamma$, para todos sus componentes, siendo α y γ constantes pequeñas.

Apéndice 10.2 El criterio FPE y el criterio AIC

Vamos a deducir el criterio AIC cuando el modelo es AR(p). Como el criterio es asintótico utilizaremos para simplificar la verosimilitud condicional, de manera que T es el número de datos utilizado en la estimación y $v_{t|t-1} = 1$. La esperanza de la verosimilitud será entonces

$$E(-2L(\mu, \boldsymbol{\phi},\sigma^2)) = T\log\sigma^2 + E\left[\sum_{t=1}^{T^*}\frac{e_t^2}{\sigma^2}\right]$$

El error de predicción cuando los parámetros se estiman con la muestra puede escribirse, llamando $\boldsymbol{\omega}_{t-p} = (\omega_{t-1}, ..., \omega_{t-p})'$ al vector de regresores y $\boldsymbol{\phi} = (\phi_1, ..., \phi_p)'$ el vector de parámetros, como:

$$e_t = \omega_t - \widehat{\boldsymbol{\phi}}'\boldsymbol{\omega}_{t-p} = \omega_t - \boldsymbol{\phi}'\boldsymbol{\omega}_{t-p} + (\boldsymbol{\phi} - \widehat{\boldsymbol{\phi}})'\boldsymbol{\omega}_{t-p} = a_t + (\boldsymbol{\phi} - \widehat{\boldsymbol{\phi}})'\boldsymbol{\omega}_{t-p}, \tag{10.52}$$

donde se suman los errores de predicción conociendo los parámtros, las innovaciones, y el error de estimar los parámetros. Elevando al cuadrado, sumando para todos los datos y tomando esperazas tenemos que

$$\frac{1}{\sigma^2}E\left[\sum_{t=1}^{T^*}e_t^2\right] = \frac{1}{\sigma^2}\left(T\sigma^2 + \sigma^2 E\left[(\boldsymbol{\phi} - \widehat{\boldsymbol{\phi}})'(\sum \boldsymbol{\omega}_{t-p}\sigma^{-2}\boldsymbol{\omega}'_{t-p})(\boldsymbol{\phi} - \widehat{\boldsymbol{\phi}})\right]\right)$$

El segundo miembro es una ji cuadrado con p grados de libertad y su esperanza será p. Por tanto

$$E(-2L(\mu, \boldsymbol{\phi},\sigma^2)) = T\log\sigma^2 + T + p$$

Para calcular esta expresión necesitamos un estimador de σ^2. Utilizando como estimador la varianza residual del modelo, que es insesgada para σ^2, dada por

$$\widehat{\sigma}^2 = \frac{T}{T-p}\widehat{\sigma}^2_{MV}$$

donde $\widehat{\sigma}^2_{MV}$ es el estimador MV podemos escribir

$$E(-2L(\mu, \boldsymbol{\phi},\sigma^2)) = -T\log(\frac{T-p}{T}) + T\log\widehat{\sigma}^2 + T + p \approx T\log\widehat{\sigma}^2 + T + 2p$$

donde hemos utilizado que, para T grande, $\log(1+p/T) \approx p/T$. Esta expresión indica que debemos seleccionar el modelo que minimize$T\log\widehat{\sigma}^2 + 2p$.

Una forma alternativa de obtener el criterio AIC es imponer la condición de minimizar los errores de predicción fuera de la muestra. Supongamos que queremos seleccionar un modelo AR(p) de manera que se minimize el error cuadrático de predicción. Este error viene dado por

$$PSE(\omega_t) = E\left[\omega_t - \widehat{\boldsymbol{\phi}}'\boldsymbol{\omega}_{t-p}\right]^2 \tag{10.53}$$

En esta expresión la esperanza se toma respecto a la distribución conjunta de las variables $(\omega_t, \widehat{\boldsymbol{\phi}}'\boldsymbol{\omega}_{t-p})$. Utilizando 10.40 tenemos que

$$PSE(\omega_t) = \sigma^2 + E\left[(\boldsymbol{\phi} - \widehat{\boldsymbol{\phi}})'\boldsymbol{\omega}_{t-p}\boldsymbol{\omega}'_{t-p}(\boldsymbol{\phi} - \widehat{\boldsymbol{\phi}})\right]$$

Esta expresión descompone el error de predicción como suma de la variabilidad debido a las innovaciones y la debida a los parámetros. Para calcular la esperanza vamos a suponer que la variable $\boldsymbol{\omega}_{t-p}$ es independiente de $\widehat{\boldsymbol{\phi}}$, lo que supone que los parámetros se han calculado con una muestra que no incluye los valores utilizado en la predicción (hipótesis que es razonable para tamaño muestral alto). Entonces la esperanza $E\left[(\boldsymbol{\phi} - \widehat{\boldsymbol{\phi}})'\boldsymbol{\omega}_{t-p}\boldsymbol{\omega}'_{t-p}(\boldsymbol{\phi} - \widehat{\boldsymbol{\phi}})\right]$ puede calcularse tomando primero la esperanza respecto a $\boldsymbol{\omega}_{t-p}$, lo que conduce a $E(\boldsymbol{\omega}_{t-p}\boldsymbol{\omega}'_{t-p}) = \Gamma_p$, y podemos escribir

$$PSE(z_{T+1}) = \sigma^2 + E\left[(\boldsymbol{\phi} - \widehat{\boldsymbol{\phi}})'\Gamma_p(\boldsymbol{\phi} - \widehat{\boldsymbol{\phi}})\right].$$

Para calcular la esperanza respecto a $\widehat{\boldsymbol{\phi}}$, como $\sqrt{T}(\boldsymbol{\phi} - \widehat{\boldsymbol{\phi}})$ tiene distribución asintoticamente normal con media cero y covarianzas $\sigma^2\Gamma_p^{-1}$, la forma cuadrática $(\boldsymbol{\phi} - \widehat{\boldsymbol{\phi}})'\Gamma_p(\boldsymbol{\phi} - \widehat{\boldsymbol{\phi}})T/\sigma^2$ es asintoticamente una χ^2_p y

$$PSE(z_{T+1}) = \sigma^2(1 + p/T).$$

Tomando como estimador de σ^2 el centrado $T\widehat{\sigma}^2_{MV}/(T-p)$ e insertando este valor en la ecuación obtenemos

$$FPE = \widehat{\sigma}^2(T+p)/(T-p).$$

Este criterio indica que debemos seleccionar el orden del AR de manera que se minimice este criterio. Una forma alternativa de escribir este criterio es

$$\log FPE = \log\widehat{\sigma}^2 + \log T(1 + p/T) - \log T(1 - p/T),$$

y utilizando que $\log(1+x) \approx x$ para x pequeña esta expresión puede aproximarse por $\log\widehat{\sigma}^2 + 2p/T$. Multiplicando esta ecuación por T obtenemos el criterio AIC

$$AIC = T\log\widehat{\sigma}^2 + 2p.$$

Apéndice 10.3 El criterio BIC

Las probabilidades a posteriori se obtienen con el teorema de Bayes

$$P(M_i|\boldsymbol{\omega}) = \frac{P(\boldsymbol{\omega}|M_i)P(M_i)}{P(\boldsymbol{\omega})}$$

y suponiendo que las probabilidades a priori son las mismas y como

$$P(\boldsymbol{\omega}) = \sum_{i=1}^{H} P(\boldsymbol{\omega}|M_i)P(M_i)$$

es una constante para todos los modelos, tenemos que

$$-2\log P(M_i|\boldsymbol{\omega}) = -2\log P(\boldsymbol{\omega}|M_i) + c$$

donde c es una constante. Para calcular el segundo miembro utilizamos que

$$P(\boldsymbol{\omega}, \boldsymbol{\theta}|M_i) = P(\boldsymbol{\omega}|\boldsymbol{\theta}, M_i)P(\boldsymbol{\theta}|M_i) = P(\boldsymbol{\theta}|\boldsymbol{\omega}, M_i)P(\boldsymbol{\omega}|M_i)$$

de donde despejamos $P(\boldsymbol{\omega}|M_i)$, tomando logaritmos

$$-2\log P(\boldsymbol{\omega}|M_i) = -2\log P(\boldsymbol{\omega}|\boldsymbol{\theta}, M_i) - 2\log P(\boldsymbol{\theta}|M_i) + 2\log P(\boldsymbol{\theta}|\boldsymbol{\omega}, M_i). \tag{10.54}$$

El primer término del segundo miembro es la verosimilitud, el segundo la probabilidad a priori de los parametros que suponemos es constante sobre la región de interés y el tercero la probabilidad a posteriori de los parametros. Supongamos ahora para simplificar el análisis que tenemos procesos AR(p) de distintos órdenes. Pensando para simplificar en la verosimilitud condicionada dada por (10.9). Si sustituimos los parámetros por sus estimaciones MV, tenemos que, para T grande de manera que $T-p$ y T son similares,

$$\log P(\boldsymbol{\omega}|\widehat{\boldsymbol{\theta}},M_i) = -\frac{T}{2}\log\widehat{\sigma}^2_{MV} + c_1$$

Por otro lado la distribución a posteriori de los parámetros puede demostrarse que es nomal con media $\widehat{\boldsymbol{\theta}}$, el estimador MV de los parámetros y con matriz de varianzas y covarianzas que depende del tamaño muestral. Puede demostrarse que, sustituyendo los parámetros por sus etimadores MV, para tamaño muestral grande

$$\log P(\widehat{\boldsymbol{\theta}}|\boldsymbol{\omega},M_i) = \frac{k}{2}\log T + c_2$$

siendo k la dimension del vector de parámetros. Sustituyendo estas expresiones en (10.54), tenemos que, para tamaño muestral grande

$$BIC = -2\log P(\boldsymbol{\omega}|M_i) = T\log\widehat{\sigma}^2_{MV} + k\log T \tag{10.55}$$

que es el criterio BIC (Véase Peña y Galeano, 2004 para otras expresiones del criterio mejores en muestras pequeñas). Por otro lado de la derivación anterior concluimos que si conocemos los valores del criterio BIC para varios modelos podemos deducir sus probabilidades a posteriori. De (10.55) concluimos que

$$P(\boldsymbol{\omega}|M_i) = ce^{-\frac{1}{2}BIC_i}$$

por lo que las probabilidades a posteriori son

$$P(M_i|\boldsymbol{\omega}) = \frac{e^{-\frac{1}{2}BIC_i}}{\sum_{j=1}^{H} e^{-\frac{1}{2}BIC_j}}$$