

# ESTIMACIÓN MÁXIMO-VEROSÍMIL DE MODELOS DE FUNCIÓN DE TRANSFERENCIA CON INTERRUPCIONES

Copyright	1995, <b>Bayes</b> Inference, S.A.
Título	ESTIMACIÓN MÁXIMO-VEROSÍMIL DE MODELOS DE FUNCIÓN DE TRANSFERENCIA CON INTERRUPCIONES
Asunto	Estimación máximo-verosímil e interrupciones en el mecanismo generador de los datos
Clave	TIPO DE DOCUMENTO NOTA12
Archivo	n:\document\bayes\concept\arma\estconint2.doc
Edición	26/06/99 18:24
Impresión	28/01/00 17:07
Distribución	Interna

## 1. Objeto

El objeto de la presente nota es describir la estimación máximo-verosímil de modelos de la clase Función de Transferencia, con especial atención a la existencia de interrupciones en la observación.

Se supone que los valores iniciales de los inputs observados necesarios para iniciar la correspondiente recursión descrita en la forma ecuación en diferencias del modelo son conocidos. Ello significa bien que los inputs son deterministas bien que existe algún procedimiento externo al cálculo de la función de verosimilitud que proporciona tantos valores iniciales como sean necesarios.

A lo largo del texto se consideran dos tipos de función de respuesta del output a los distintos inputs. En el primer caso la función de respuesta consiste en un polinomio finito en el operador de retardo B. Posteriormente se considera el caso de funciones de respuesta de la clase fracción racional de polinomios en B.

El caso más general donde desconocemos las condiciones iniciales de ciertos inputs estocásticos pero estos son representables por un modelo lineal se tratarán en un próximo documento sobre estimación máximo-verosímil de modelos ARMA vectoriales.

Una vez establecidos los resultados en el caso en que el sistema carece de interrupciones, aquellos se generalizan a esta última situación bajo dos supuestos diferentes:

- el sistema guarda memoria de su comportamiento inercial y, alternativamente,
- el sistema carece de inercia inicial. En este último caso puede ocurrir que el sistema generador de los datos atraviese por una época transitoria de aprendizaje o reaprendizaje que debe modelizarse explícitamente.

La estrategia de presentación se inicia con modelos esencialmente equivalentes a la conocida clase ARMAX escalar. Posteriormente los resultados se generalizan a modelos con función de respuesta racional. Por último se presentan los casos con interrupciones en sus dos variedades con memoria inicial y sin ella.

## 2. El modelo ARMAX escalar y el modelo de Función de Transferencia

Un modelo ARMAX (Hannan y Deistler, 1988) escalar tiene la forma

$$[2.1] \quad \phi(B)z_t = \sum_i \omega_i(B)x_{i,t} + \theta(B)a_t$$

mientras que un modelo de Función de Transferencia (Box y Jenkins, 1976) es

$$[2.2] \quad z_t = \sum_i \frac{\omega_i(B)}{\delta_i(B)} x_{i,t} + \frac{\theta(B)}{\phi(B)} a_t$$

Es claro que si en [2.1] dividimos ambos lados de la expresión por  $\phi(B)$  el modelo ARMAX queda como un modelo de Función de Transferencia con la restricción de que todas las funciones de respuesta en [2.2] tienen idéntico denominador.

### 2.1 Forma del modelo

El modelo inicial que tratamos es

$$[2.1.1] \quad \phi(B)z_t = \phi(B) \sum_i \omega_i(B)x_{i,t} + \theta(B)a_t$$

donde

$\phi(B), \theta(B)$  son polinomios en B con sus raíces fuera del círculo unidad, mientras que  $\omega(B)$  es un polinomio en B con raíces arbitrarias,  $z_t, x_{i,t}, a_t$  son, respectivamente, la serie output, el input i-ésimo y un input no observado que se distribuye normal e independientemente con media cero y desviación típica  $\sigma$  (ruido blanco gaussiano).

Suponemos que los inputs son bien deterministas o bien conocemos  $d_i$  datos iniciales, donde  $d_i$  es el grado del polinomio  $\omega_i(B)$ . El modelo [2.1.1] puede escribirse como

$$\begin{aligned} [2.1.2] \quad z_t &= \sum_i \omega_i(B) x_{i,t} + r_t \\ \phi(B) r_t &= \theta(B) a_t \end{aligned}$$

Razones de comodidad expositiva y de limitación de supuestos acerca de las condiciones iniciales aconsejan que tratemos el modelo de Función de Transferencia posteriormente.

Las ecuaciones [2.1.2] pueden escribirse en forma matricial como

$$\begin{aligned} [2.1.3] \quad Z &= \Omega X + R \\ \Phi R &= \Theta A + H U \end{aligned}$$

donde

$$\begin{aligned} Z^T &= [z_1, \dots, z_n], \\ X^T &= [x_{1,1-d_1}, \dots, x_{1,1}, \dots, x_{1,n}, x_{2,1-d_2}, \dots, x_{2,1}, \dots, x_{k,n}], \\ R^T &= [r_1, \dots, r_n], \\ A^T &= [a_1, \dots, a_n], \\ U^T &= [r_{1-p}, \dots, r_0, a_{1-q}, \dots, a_0] \end{aligned}$$

donde  $n$  es el número de datos y  $k$  el número de inputs observados,

$$\Omega = [\Omega_1, \Omega_2, \dots, \Omega_k]$$

donde cada matriz  $\Omega_i$  posee  $n$  filas y  $n+d_i$  columnas, siendo  $d_i$  el grado del polinomio  $\omega_i(B)$ , y

$$\Omega_i = \begin{bmatrix} \omega_{i,d_i} & & \omega_0 & 0 & 0 \\ 0 & \omega_{i,d_i} & & \omega_0 & \\ & & & & \\ 0 & & & \omega_{i,d_i} & \omega_0 \end{bmatrix},$$

$\Phi$  es una matriz triangular de  $n$  filas y columnas con cada término de la diagonal principal igual a 1, la subdiagonal  $i$ -ésima compuesta por términos iguales a  $-\phi_i$ . Es decir

$$\Phi = \begin{bmatrix} 1 & & & \\ -\phi_1 & 1 & & \\ -\phi_2 & -\phi_1 & 1 & \\ & & & -\phi_2 & -\phi_1 & 1 \end{bmatrix}$$

Del mismo modo, la matriz  $\Theta$  es triangular con n filas y columnas, con cada término de la diagonal principal igual a 1 y cada término de la subdiagonal i-ésima igual a  $-\theta_i$ . Finalmente,

$$H = \begin{bmatrix} H_1 & H_2 \\ 0_{n-p,p} & 0_{n-q,q} \end{bmatrix}$$

donde

$$H_1 = \begin{bmatrix} \phi_p & \phi_{p-1} & \phi_1 \\ & \phi_p & \phi_2 \\ & & \phi_p \end{bmatrix} \text{ y}$$

$$H_2 = \begin{bmatrix} -\theta_q & -\theta_{q-1} & -\theta_1 \\ & -\theta_q & -\theta_2 \\ & & -\theta_q \end{bmatrix}$$

Premultiplicando la segunda ecuación de [2.1.3] por  $\Phi^{-1}$  y sustituyendo en la primera tenemos que

$$[2.1.4] \quad Z = \Omega X + \Psi A + VU$$

Premultiplicando por  $\Psi^{-1}$  en [2.1.4] tenemos

$$[2.1.5] \quad A = \Pi Z - \Pi \Omega X - \Theta^{-1} H U = \Pi R - \Theta^{-1} H U = \Pi R - L U$$

## 2.2 El máximo de la función de verosimilitud

La función de densidad de una muestra de n observaciones generadas por un modelo ARMAX, con parámetros  $\alpha = (\phi_1, \dots, \phi_p, \theta_1, \dots, \theta_q, \omega_{1,0}, \dots, \omega_{1,d_1}, \dots, \omega_{k,d_k})^T$  y  $\sigma$  es

$$[2.2.1] \quad f(Z|\alpha, \sigma^2) = (2\pi\sigma^2)^{-n/2} |\Sigma|^{-1/2} \exp\left(-\frac{1}{2\sigma^2} (Z - \Omega X)^T \Sigma^{-1} (Z - \Omega X)\right)$$

donde  $\sigma^2 \Sigma$  es la matriz de covarianzas del ruido. Tomando logaritmos en la función de densidad y utilizando que  $Z - \Omega X = R$ , tenemos que el logaritmo de la función de densidad es

$$\log(f(Z|\alpha, \sigma^2)) = \frac{-n}{2} (\log 2\pi + \log \sigma^2) - \frac{1}{2} \log |\Sigma| - \frac{1}{2\sigma^2} R^T \Sigma^{-1} R$$

y maximizando respecto de  $\sigma^2$ , tenemos que

$$\frac{-n}{2\sigma^2} + \frac{R^T \Sigma^{-1} R}{2\sigma^4} = 0 \Rightarrow \sigma^2 = \frac{R^T \Sigma^{-1} R}{n}$$

que sustituyendo en [2.2.1] muestra que la función de verosimilitud se maximiza minimizando

$$[2.2.2] \quad \max(l(\alpha)) = |\Sigma|^{-\frac{1}{n}} (Z - \Omega X)^T \Sigma^{-1} (Z - \Omega X) = |\Sigma|^{-\frac{1}{n}} R^T \Sigma^{-1} R$$

## 2.3 Estrategia para maximizar la función de verosimilitud

La estrategia para maximizar la función de verosimilitud se basa en explotar el hecho que

$$[2.3.1] \quad p(Z, U|\alpha, \sigma) = p(Z|U, \alpha, \sigma) p(U|\alpha, \sigma) = p(U|Z, \alpha, \sigma) p(Z|\alpha, \sigma)$$

Así, primero obtenemos una expresión para  $p(U|\alpha, \sigma)$  y  $p(Z|U, \alpha, \sigma)$  para a continuación determinar  $p(U|Z, \alpha, \sigma)$  y de ahí la probabilidad deseada.

## 2.4 Función de densidad conjunta de las observaciones y las condiciones iniciales

Para obtener la  $p(Z|U, \alpha, \sigma)$  consideremos primero la función de probabilidad de una muestra de tamaño  $n$  de ruido blanco

$$p(A|\alpha, \sigma) = p(A) = (2\pi\sigma^2)^{-\frac{n}{2}} \exp\left(-\frac{1}{2\sigma^2} A^T A\right)$$

Usando [2.1.5] y observando que el jacobiano de la transformación es unitario podemos escribir

$$[2.4.1] \quad p(Z|U, \alpha, \sigma) = (2\pi\sigma^2)^{-\frac{n}{2}} \exp\left(-\frac{1}{2\sigma^2} (\Pi(Z - \Omega X) - LU)^T (\Pi(Z - \Omega X) - LU)\right)$$

Por su parte, las condiciones iniciales se distribuyen normalmente de forma conjunta, de modo que su función de densidad puede expresarse como

$$[2.4.2] \quad p(U|\alpha, \sigma) = (2\pi\sigma^2)^{-\frac{p+q}{2}} |\Sigma_*|^{-\frac{1}{2}} \exp\left(-\frac{1}{2\sigma^2} U^T \Sigma_*^{-1} U\right)$$

donde, naturalmente  $p$  y  $q$  son los órdenes de los polinomios AR y MA del ruido y  $\sigma^2 \Sigma_*$  es la matriz de covarianzas de  $U$ .

Así la densidad conjunta de las condiciones iniciales y las observaciones queda

$$[2.4.3] \quad p(Z, U|\alpha, \sigma) = (2\pi\sigma^2)^{-\frac{n+p+q}{2}} |\Sigma_*|^{-\frac{1}{2}} \exp\left(-\frac{1}{2\sigma^2} ((\Pi R - LU)^T (\Pi R - LU) + U^T \Sigma_*^{-1} U)\right)$$

## 2.5 La densidad marginal de las observaciones

La forma cuadrática  $S$  de la función exponencial de [2.4.3] puede escribirse como

$$[2.5.1] \quad S = U^T (\Sigma_*^{-1} + L^T L) U - 2U^T L^T \Pi R + R^T \Pi^T \Pi R$$

El modo en el que podemos obtener la densidad marginal del vector de observaciones es descomponiendo dicha forma cuadrática en dos formas, una de las cuales es independiente de  $U$ , que es precisamente la función de densidad marginal buscada. El objetivo se alcanza minimizando  $S$  respecto a  $U$  y reescribiendo  $S$ .

Derivando e igualando a 0 vemos que la forma cuadrática  $S$  toma un mínimo cuando

$$[2.5.2] \quad (\Sigma_*^{-1} + L^T L) U = L^T \Pi R$$

Por tanto, podemos escribir  $S = S_1 + S_2$ , donde (Ver Apéndice 1)

$$[2.5.3] \quad S_1 = (U - \bar{U})^T (\Sigma_*^{-1} + L^T L) (U - \bar{U})$$

y

$$[2.5.4] \quad S_2 = (\Pi R - L\bar{U})^T (\Pi R - L\bar{U}) + \bar{U}^T \Sigma_*^{-1} \bar{U}$$

con

$$[2.5.5] \quad \bar{U} = (\Sigma_*^{-1} + L^T L)^{-1} L^T \Pi R$$

satisfaciendo la ecuación [2.5.2]. Es claro que  $S_2$  no depende de  $U$  y por contra [2.5.3] depende de  $Z$  (a través de [2.5.2]). Esto implica que [2.4...3] puede factorizarse

$$p(U|Z, \alpha, \sigma) = (2\pi\sigma^2)^{-\frac{p+q}{2}} |\Sigma_*^{-1} + L^T L|^{\frac{1}{2}} \exp\left(-\frac{1}{2\sigma^2} S_1\right)$$

y

$$[2.5.6] \quad p(Z|\alpha, \sigma) = (2\pi\sigma^2)^{-\frac{n}{2}} |\Sigma_*|^{-\frac{1}{2}} |\Sigma_*^{-1} + L^T L|^{-\frac{1}{2}} \exp\left(-\frac{1}{2\sigma^2} S_2\right)$$

Naturalmente,  $S_2 = R^T \Sigma^{-1} R = (Z - \Omega X)^T \Sigma^{-1} (Z - \Omega X)$  y

$$[2.5.7] \quad |\Sigma| = |\Sigma_*| |\Sigma_*^{-1} + L^T L|$$

Teniendo en cuenta [2.1.5] podemos definir

$$[2.5.8] \quad \bar{A} = \Pi R - L \bar{U}$$

De modo

$$[2.5.9] \quad S_2 = \bar{A}^T \bar{A} + \bar{U}^T \Sigma_*^{-1} \bar{U}$$

Las expresiones [2.5.8] y [2.5.6] proporcionan un método de cálculo de [2.2.2]. Es sabido que la raíz enésima del determinante [2.5.6] tiende a la unidad con el aumento de  $n$ . Este hecho y la poca variabilidad del determinante [2.5.6] a distintos valores de los parámetros hace que, al menos en muestras grandes aproximemos los parámetros máximo-verosímiles por aquellos que minimizan  $S_2$ . Puede observarse que  $S_2$  es una suma de cuadrados, y por tanto pueden usarse algoritmos estándar de minimización cuadrática no lineal. De acuerdo con los resultados que se establecen en el documento **Matrices, Polinomios en B y F y Matrices de Covarianzas de Procesos Lineales** las únicas inversiones necesarias en la determinación de  $S_2$ , son las relativas al cálculo de  $\bar{U}$ . En efecto, el cálculo de cada término  $\bar{a}_i$  puede realizarse de forma recursiva mediante las ecuaciones [2.1.1] o, alternativamente, [2.1.2].

### 3. Minimización cuadrática

#### 3.1 No linealidad de los residuos respecto a los parámetros

Expresemos ahora [2.5.8] por  $S_2(\alpha)$  para subrayar su dependencia respecto al vector de parámetros del modelo y escribamos

$$[3.1] \quad S_2(\alpha) = A_\alpha^T A_\alpha + (M_\alpha^{-1} U_\alpha)^T (M_\alpha^{-1} U_\alpha) = \tilde{A}_\alpha^T \tilde{A}_\alpha$$

donde  $\tilde{A}_\alpha = \begin{bmatrix} A_\alpha \\ M_\alpha^{-1} U_\alpha \end{bmatrix}$  y  $\Sigma_{*\alpha} = M_\alpha M_\alpha^T$  es la descomposición de Cholesky de la matriz covarianzas de las condiciones iniciales.

Es fácil observar que la función  $f(\alpha) = \tilde{A}_\alpha$  no es lineal. En efecto, sea  $\tilde{a}_t^\alpha$  un término del vector  $A_\alpha$ . Tomando en consideración que

$$a_t = \theta^{-1}(B) \phi(B) (z_t - \sum_i \omega_i(B) x_{i,t})$$

vemos que

$$[3.2] \quad \begin{aligned} \frac{\partial a_t}{\partial \phi_k} &= -B^k \theta^{-1}(B) r_t + \frac{\partial r_t}{\partial \phi_k} \\ \frac{\partial a_t}{\partial \theta_k} &= B^k \theta^{-2}(B) \phi(B) r_t + \frac{\partial r_t}{\partial \theta_k} \\ \frac{\partial a_t}{\partial \omega_{i,k}} &= -B^k \theta^{-1}(B) \phi(B) (x_{i,t} + \frac{\partial r_t}{\partial \omega_{i,k}}) \end{aligned}$$

De donde, bajo el supuesto de que los segundos sumandos del lado derecho de las tres expresiones precedentes son nulos, es decir, bajo el supuesto de que las condiciones iniciales son fijas

$$[3.3] \quad \begin{aligned} \phi(B) \frac{\partial a_t}{\partial \phi_k} &= -a_{t-k} \\ \theta(B) \frac{\partial a_t}{\partial \theta_k} &= a_{t-k} \\ \theta(B) \frac{\partial a_t}{\partial \omega_{i,k}} &= -\phi(B) x_{i,t-k} \end{aligned}$$

La no-linealidad de la función  $f(\alpha)$  motiva que la determinación del vector  $\hat{\alpha}$  que minimiza la función deba realizarse por métodos de aproximación iterativa.



## 3.2 El algoritmo de Marquardt

Un algoritmo por el que dicha aproximación iterativa puede realizarse es el proporcionado por Marquardt (1963). El algoritmo se basa en combinar el algoritmo de Gauss-Newton y el método de máximo descenso que resultan apropiados en diferentes situaciones de aproximación al mínimo.

El algoritmo de Gauss-Newton consiste en sustituir la función  $A(\alpha) = \tilde{A}_\alpha$  por su aproximación de primer orden en un punto que se supone próximo al mínimo global. La función así aproximada se minimiza y el punto mínimo obtenido se utiliza para construir una nueva aproximación de primer orden que se vuelve a minimizar y así sucesivamente hasta que el descenso de la función es menor que cierta cota preestablecida.

Así definiendo  $J$  como la matriz cuyo término  $(t,j)$  es  $\frac{\partial \tilde{a}_t(\alpha)}{\partial \alpha_j}$  podemos escribir

$\tilde{A}(\alpha + \xi) \approx \tilde{A}(\alpha) + J\xi$ . A partir de ahí podemos minimizar  $(\tilde{A}(\alpha) + J\xi)^T (\tilde{A}(\alpha) + J\xi)$ , respecto a  $\xi$ , que como es sabido equivale a resolver el sistema lineal

$$[3.2.1] \quad -J^T \tilde{A}(\alpha) = J^T J \xi$$

Los métodos basados en el gradiente eligen ajustes de  $\alpha$  en la dirección  $-J^T \tilde{A}(\alpha)$ . Una vez determinada la dirección el método del máximo descenso consiste en encontrar el mínimo de la función  $\tilde{A}(\alpha - tJ^T \tilde{A}(\alpha))$ .

La secuencia de valores del vector  $\alpha$  obtenida por Gauss-Newton podría no converger hacia el mínimo si los valores iniciales están mal elegidos. De otra parte la dirección localmente óptima de los métodos basados en el gradiente puede producir una secuencia con convergencia muy lenta. El algoritmo de Marquardt combina ambos métodos proponiendo la resolución del sistema

$$[3.2.2] \quad -J^T \tilde{A}(\alpha) = (J^T J + \lambda I) \xi$$

Es claro que [3.2.2] cuando  $\lambda$  tiende a cero es idéntico al método de Gauss-Newton y cuando  $\lambda$  crece el sistema se aproxima a los métodos basados en el gradiente.

## 4. La matriz de covarianzas de los parámetros estimados

Es sabido que, en el contexto de la inferencia clásica, bajo ciertas condiciones de regularidad que en nuestro caso se satisfacen (Klimov (1986)), los parámetros poseen una distribución asintótica normal con vector de medias los estimadores máximo-verosímiles y matriz de covarianzas el inverso de la matriz de información.

La matriz de información se define en nuestro caso por

$$[3.1] \quad I(\alpha) = E_{\alpha} \left( \frac{\partial(\log(p(Z|\alpha)))}{\partial \alpha_i} \frac{\partial(\log(p(Z|\alpha)))}{\partial \alpha_j} \right)$$

Aproximando  $\log(p(Z|\alpha, \sigma)) \approx -\frac{1}{2\sigma^2} S_2 = -\frac{1}{2\sigma^2} \tilde{A}_{\alpha}^T \tilde{A}_{\alpha}$ , usando el hecho de que el primer sumando de  $S_2$  en [2.5.8] domina el comportamiento global cuando  $n$  se hace grande, tenemos que el término (i,j) de la matriz  $I(\alpha)$ , es

$$[3.2] \quad I_{i,j}(\alpha) \approx \frac{1}{\sigma^4} E_{\alpha} \sum \tilde{a}_t^2 \frac{\partial \tilde{a}_t}{\partial \alpha_i} \frac{\partial \tilde{a}_t}{\partial \alpha_j} = \frac{1}{\sigma^2} \sum E_{\alpha} \left( \frac{\partial \tilde{a}_t}{\partial \alpha_i} \frac{\partial \tilde{a}_t}{\partial \alpha_j} \right) = \frac{1}{\sigma^2} J^T J$$

## 5. La forma función de transferencia

Consideremos la siguiente generalización de la forma [2.1.2]

$$[5.1.1] \quad z_t = \sum_i \frac{\omega_i(B)}{\delta_i(B)} x_{i,t} + r_t$$

$$\phi(B)r_t = \theta(B)a_t$$

Suponemos que conocemos no sólo un conjunto de valores iniciales de cada variable  $x_{i,t}$  sino que además conocemos los valores iniciales correspondientes de la variable

$$y_t = \frac{\omega_i(B)}{\delta_i(B)} x_{i,t}$$

Es claro que el mecanismo de estimación, bajo las condiciones enunciadas no cambia ya que la evaluación del ruido puede realizarse recursivamente recurriendo a la forma en diferencias del modelo.

La alimentación de datos inicial debe realizarse explícitamente por parte del analista o en caso contrario se supone que los valores iniciales son iguales a cero.

## 6. Estimación con interrupciones

Pueden presentarse dos tipos de interrupciones en los datos: las debidas a una interrupción de la observación y las debidas a una interrupción de la actividad del

mecanismo generador. Una interrupción de la observación debida a la ausencia de actividad del mecanismo generador puede conllevar una pérdida de la memoria existente en el proceso, sobre todo si dicha interrupción es prolongada.

Así pues consideraremos dos casos:

- **Continuidad inercial** correspondiente a la situación en que el comportamiento de la serie se reanuda respondiendo a las características inerciales derivadas del modelo y de los datos observados previa a la interrupción. Naturalmente, cuando el mecanismo generador no se interrumpe el sistema conserva el comportamiento inercial del pasado, es decir, guarda memoria de su pasado en la forma prescrita por el modelo. No obstante, ciertas interrupciones breves de la actividad pueden conservar su memoria inercial y por tanto tratarse en este primer contexto.
- **Reinicialización del proceso con presencia eventual de un mecanismo de recuperación de memoria** que se corresponde a los casos en los que la actividad misma se ha interrumpido, sobre todo si esto ha ocurrido de forma prolongada. En el caso en que el mecanismo generador se interrumpe, el sistema pierde la memoria de su pasado reinicializando el proceso que vuelve a tener las mismas características probabilísticas, pero partiendo de nuevas condiciones iniciales. Adicionalmente, un proceso interrumpido, sufre al reinicializarse de un proceso de aprendizaje especial que se representa mediante una variable que vale 1 a partir de la nueva apertura y cero en los puntos temporales previos a la apertura y con una función de respuesta del tipo función de transferencia, descrita en el punto anterior. Si ocurren rupturas de la observación reiteradas, cada proceso de reequilibrio debería ser representado por distintas variables pero con idéntica función de respuesta.

## 6.1 Interrupciones con continuidad inercial

Suponemos que existe una única interrupción en los datos. Ello, como se verá, no provoca pérdida de la generalidad del argumento.

Así  $Z^T = [Z_1^T, Z_2^T]$  donde entre el primer vector y el segundo vector existe una discontinuidad de orden  $c$ . Tomando en cuenta que

$$p(Z|\alpha) = p(Z_2|Z_1, \alpha)p(Z_1|\alpha)$$

es claro que  $p(Z_1|\alpha)$  puede calcularse por el método reseñado en los puntos 2 y 3 de esta nota.

Sea

$$E = [e_{m-1}, \dots, e_{m-p}, a_{m-1}, \dots, a_{m-q}]^T$$

donde  $m$  es el subíndice temporal mínimo de la segunda submuestra y

$$e_{m-k} = \begin{cases} 0 & \text{si } c < k \\ \sum_{i=0}^{m-k-N_1-1} \psi_i a_{m-k-i} & \text{si } c \geq k \end{cases}$$

Debe notarse que  $e_{m-k}$  es, simplemente, el error de previsión de  $z_{m-k}$  desde el origen  $N_1$  que es el subíndice temporal máximo de la primera submuestra. Si la interrupción es menor que  $q$ , los términos  $a_{m-k}$  donde  $k$  es mayor que la longitud de la interrupción son conocidos e iguales a los errores correspondientes estimados en la primera parte de la muestra. Por tanto, el número de términos independientes del vector  $E$  es igual al tamaño de la interrupción si ésta es de duración menor o igual a  $p+q$ , o a  $p+q$ , en caso contrario.

De forma paralela al razonamiento del punto [2.3] tenemos que

$$p(Z_2, E | Z_1, \alpha) = p(Z_2 | Z_1, \alpha) p(E | Z_1, Z_2, \alpha) = p(Z_2 | Z_1, \alpha) p(E | Z_2, \alpha) = p(Z_2 | E, Z_1, \alpha) p(E | \alpha)$$

donde en la segunda igualdad utilizamos la independencia de  $Z_1$  y  $E$ .

Sea  $A_2$  el vector de innovaciones correspondiente a la segunda parte de la muestra. Puede notarse que

$$A_2 = \Pi_2(Z_2 - \Omega_2 X_2) - L_2 U_2$$

con el subíndice 2 haciendo referencia a la segunda parte de la muestra y a la definición homóloga de las matrices correspondientes a la realizada en el punto [2.1] y con

$$U_2 = E + \hat{Z}$$

donde  $\hat{Z} = [\hat{z}_{m-1}, \dots, \hat{z}_{m-p}, \underbrace{0, \dots, 0}_{q \text{ terminos}}]^T$

En el caso en que el tamaño de la interrupción es  $c \geq p+q$  la matriz de covarianzas de  $E$  es

$$\text{Cov}(E, E^T) = \sigma^2 \Sigma_*^{(2)} = \sigma^2 \begin{bmatrix} \Xi_p & \Psi_{q,p}^T \\ \Psi_{q,p} & I_q \end{bmatrix}$$

donde

$$\Xi_p = \begin{bmatrix} \sum_{i=0}^{m-N_1-2} \psi_i^2 & \sum_{i=0}^{m-N_1-3} \psi_i \psi_{i+1} & \cdots & \cdots & \sum_{i=0}^{m-N_1-p-1} \psi_i \psi_{i+p-1} \\ \sum_{i=0}^{m-N_1-3} \psi_i \psi_{i+1} & \sum_{i=0}^{m-N_1-3} \psi_i^2 & & & \vdots \\ \vdots & & \ddots & & \vdots \\ \vdots & & & \ddots & \vdots \\ \sum_{i=0}^{m-N_1-p-1} \psi_i \psi_{i+p-1} & \cdots & \cdots & \cdots & \sum_{i=0}^{m-N_1-p-1} \psi_i^2 \end{bmatrix}$$

$I_q$  es la matriz unidad de orden  $q$  y  $\Psi_{q,p}$  es una matriz de  $q$  filas y  $p$  columnas con la siguiente estructura:

$$\Psi_{q,p} = \begin{bmatrix} 1 & 0 & \cdots & \cdots & 0 & \cdots & 0 \\ \psi_1 & 1 & \ddots & & & & \vdots \\ \vdots & \ddots & \ddots & \ddots & & & \vdots \\ \psi_{\max(p,q)-1} & \cdots & \psi_1 & 1 & 0 & \cdots & 0 \end{bmatrix}$$

donde a la derecha de la diagonal de unos todos los términos son cero.

En otro caso, es decir cuando  $c < p + q$  podemos escribir

$$[6.1.1] \quad \begin{bmatrix} e_{m-1} \\ \vdots \\ e_{m-\min(p,c)} \\ a_{m-1} \\ \vdots \\ a_{m-\min(q,c)} \end{bmatrix} = \begin{bmatrix} \Psi_{*c,\min(q,c)} \\ I_{\min(q,c)}, 0_{\max(0,c-q)} \end{bmatrix} \begin{bmatrix} a_{m-1} \\ \vdots \\ a_{m-c} \end{bmatrix}$$

donde

$$\Psi_{*c,\min(q,c)} = \begin{bmatrix} 1 & \psi_1 & \cdots & \psi_{\min(q,c)} \\ 0 & 1 & \cdots & \psi_{\min(q,c)-1} \end{bmatrix}$$

Adaptando el argumento ofrecido en el punto [2] de esta nota se obtiene que el máximo de la función de verosimilitud de una serie con una interrupción donde no existe pérdida de la memoria inercial se obtiene minimizando la cantidad

$$[6.1.2] \quad S_2^{(1,2)} = \hat{A}_1^T \hat{A}_1 + \hat{U}_1^T \Sigma_*^{-1} \hat{U}_1 + \hat{A}_2^T \hat{A}_2 + \hat{E}^T (\Sigma_*^{(2)})^{-1} \hat{E}$$

Puesto que el vector  $E$  puede expresarse en función de un vector de dimensión  $\min(p+q, c)$ , donde  $c$  es la longitud de la interrupción, la última forma cuadrática de la expresión [6.1.1] puede expresarse también en términos de dicha base.

## 6.2 Análisis de Intervención y datos no observados

La última expresión del punto [6.1] proporciona un método general para tratar el caso de interrupciones en los datos, a condición de que el sistema conserve su comportamiento inercial, independientemente de la longitud de la interrupción. Es obvio, también, que el caso con múltiples interrupciones responde a la fórmula

$$[6.2.1] \quad S_2^{(1,2,\dots,v)} = \sum_{i=0}^v \hat{A}_i^T \hat{A}_i + \hat{U}_1^T \Sigma_*^{-1} \hat{U}_1 + \sum_{i=2}^v \hat{E}_i^T (\Sigma_*^{(i)})^{-1} \hat{E}_i$$

Volvamos ahora al caso en el que existe una sola interrupción y supongamos que la longitud de la interrupción es menor o igual a  $p+q$ . Puesto que

$$\Sigma_*^{(2)} = \begin{bmatrix} \Psi_{*c, \min(q,c)} \\ I_{\min(q,c)}, 0_{\max(0, c-q)} \end{bmatrix} \begin{bmatrix} \Psi_{*c, \min(q,c)} \\ I_{\min(q,c)}, 0_{\max(0, c-q)} \end{bmatrix}^T$$

la expresión  $\hat{E}^T (\Sigma_*^{(2)})^{-1} \hat{E} = \hat{A}_{[N_1+1, m-1]}^T \hat{A}_{[N_1+1, m-1]}$  donde el subíndice del último vector de innovaciones indica el rango de índices que contiene dicho vector. Este último resultado significa, simplemente, que en el caso en que la interrupción es de orden menor que  $p+q$  la minimización cuadrática puede realizarse sobre el total de la secuencia de innovaciones mas la suma de cuadrados de las condiciones iniciales de la primera muestra. La interrupción puede por tanto llenarse con valores arbitrarios e introducir un conjunto de impulsos en cada uno de los periodos de interrupción. Puesto que bajo este último planteamiento se minimiza el total de la suma de cuadrados mas la forma cuadrática inicial, los resultados obtenidos bajo uno y otro planteamiento resultan ser idénticos.

## 6.3 Interrupciones con pérdida de memoria

Cuando se dan interrupciones de la observación con pérdida de memoria, basta con considerar condiciones iniciales independientes de las submuestras precedentes. Así la función de verosimilitud es el producto de las funciones de verosimilitud para cada submuestra y la forma cuadrática objeto de minimización queda

$$S_2^{(1,2,\dots,v)} = \sum_{i=0}^v (\hat{A}_i^T \hat{A}_i + \hat{U}_i^T \hat{U}_i)$$

## 6.4 Procesos de reaprendizaje

Una serie temporal que experimenta interrupciones con pérdida de memoria suele experimentar asimismo un proceso de readaptación dinámica en el inicio de cada submuestra. Este fenómeno puede representarse con la introducción de un input que toma el valor 1 a lo largo de la submuestra correspondiente y cero en el inicio, con una función de respuesta del tipo

$$v(B) = \frac{\omega_0}{1 - \delta B}$$

Cada submuestra experimenta un proceso de aprendizaje con idéntica función de respuesta hasta situarse en su senda dinámica normal.

## 7. Reducción al caso MA puro

Dado un modelo ARIMA cualquiera

$$\phi(B)r_t = \theta(B)a_t$$

siempre existe un modelo MA equivalente

$$r'_t = \theta(B)a_t$$

con la simple sustitución

$$r'_t = \phi(B)r_t$$

En este modelo, hay  $p$  residuos menos y no son necesarios valores iniciales para la serie output pues se utilizan los valores conocidos. Se trata pues, de estimar el modelo ARIMA con todos sus parámetros pero calculando los valores iniciales solamente para los residuos, aplicando previamente las diferencias y la parte autorregresiva a la serie de ruido.

### 7.1 Cálculo de los valores iniciales

En este caso la matriz de covarianzas es la identidad y la ecuación [2.5.5] se reduce a

$$[7.1.1] \quad \bar{U} = (I_q + L^T L)^{-1} L^T \Theta^{-1} R' = (I_q + L^T L)^{-1} L^T \Theta^{-1} \Phi R = (I_q + L^T L)^{-1} L^T \Pi R$$

donde la matriz  $L$  se simplifica del siguiente modo

$$[7.1.2] \quad L = \Theta^{-1} \begin{bmatrix} H_2 \\ 0_{n-q,q} \end{bmatrix} = (\Theta^{-1})_q H_2 \in \Re^{(N-p) \times q}$$

Si  $q$  es pequeño la matriz  $I_q + L^T L$  se invierte sin apenas coste. Cuando no es así suele ser porque el polinomio MA es multiplicativo de la forma

$$[7.1.3] \quad \theta(B) = \theta_0(B^{s_0}) \cdot \theta_1(B^{s_1}) \cdot \dots \cdot \theta_r(B^{s_r}) \wedge s_r \geq \dots > s_1 \geq s_0 \geq 1$$

con lo que el polinomio  $\theta(B)$  tiene relativamente pocos coeficientes no nulos y la matriz  $I_q + L^T L$  es rala ( en inglés *sparse* ) con lo que existen métodos especiales de inversión que aprovechan la forma de la matriz para acelerar su inversión. Uno de ellos es el de Sherman-Morrison (ver apéndice A3).

## 7.2 Cálculo de los residuos y del jacobiano

De la ecuación [2.1.5] se obtiene

$$[7.1.4] \quad A = \Pi R - LU = \Pi R - L(I_q + L^T L)^{-1} L^T \Pi R = MR$$

definiendo

$$[7.1.5] \quad M = (I_N - L(I_q + L^T L)^{-1} L^T) \Pi \in \Re^{(N-p) \times N}$$

Si  $x$  es una variable del modelo de la que los polinomios ARMA son independientes, entonces las matrices  $L$  y  $\Pi$ , y por ende  $M$ , son independientes de  $x$  y la ecuación [7.1.4] se cumple también para los jacobianos

$$[7.1.6] \quad \frac{\partial A}{\partial x} = M \frac{\partial R}{\partial x}$$

De esta forma, en el caso de la función de transferencia se puede calcular analíticamente las derivadas parciales de los residuos con respecto a los parámetros de transferencia pues se conocen las del ruido que se deducen directamente de [5.1.1]

$$[7.1.7] \quad \frac{\partial r_t}{\partial \omega_{i,j}} = - \sum_i \frac{B^j}{\delta_i(B)} x_{i,t}$$

$$[7.1.8] \quad \frac{\partial r_t}{\partial \delta_{i,j}} = + \sum_i \frac{\omega_i(B) B^j}{(\delta_i(B))^2} x_{i,t}$$



Esto es mucho más ventajoso que utilizar un método numérico para el cálculo del jacobiano, necesario para aplicar el método de Marquardt; sobre todo cuando el número de parámetros de transferencia es muy grande, ya que cada derivada parcial implicaría una evaluación completa del modelo.

### 7.3 Estimación con interrupciones

En el caso de estimación con interrupciones, si denotamos los datos censurados con un asterisco superíndice

$$[7.2.1] \quad r_t = r_t^* + c_t; \quad R = R^* + C$$

donde las interrupciones sólo toman valores en un subconjunto de índices  $I$

$$[7.2.2] \quad c_t = 0 \quad \forall t \notin I = \{t_1, \dots, t_k\} \subset \{1 \dots N\}$$

De la ecuación [2.6.6] se deduce

$$[7.2.3] \quad A = MR = MR^* + MC = A^* + MC$$

Los residuos se anulan en  $I$

$$[7.2.4] \quad a_t = 0 \quad \forall t \in I \subset \{1 \dots N\}$$

así que indicando con el subíndice  $I$  a las submatrices de los elementos de filas y columnas con índices en  $I$  se tiene el siguiente sistema lineal

$$[7.2.5] \quad A_I = A_I^* + M_{II} C_I = 0$$

Para que el sistema tenga solución necesitamos que haya un residuo que anular para cada interrupción, o sea, una ecuación para cada variable, lo cual se consigue si no hay interrupciones anteriores al grado del polinomio ARI, es decir

$$[7.2.6] \quad I \subset \{p+1 \dots N\}$$

En el caso de haber interrupciones anteriores la única solución es incluir series input de tipo pulso para intervenirlas.

Así pues  $M_{II}$  es cuadrada y regular por construcción, luego es inversible y además, al igual que  $I_q + L^T L$ , cuando es grande es rara y se puede utilizar el método de Sherman-Morrison para calcular los valores de las interrupciones despejando directamente de [7.2.5]

$$[7.2.7] \quad C_I = -(\mathbf{M}_{II})^{-1} A_I^*$$

Como  $C_I$  determina completamente a  $C$ , basta aplicar [7.2.3] para hallar los residuos no censurados a partir de los censurados.

De forma análoga, se calcula el jacobiano de los residuos no censurados a partir del jacobiano de los censurados puesto que para toda variable  $x$  de la que los polinomios ARMA sean independientes se cumplirá

$$[7.2.8] \quad \frac{\partial A}{\partial x} = \frac{\partial A^*}{\partial x} + \mathbf{M} \frac{\partial C}{\partial x}$$

$$[7.20] \quad \frac{\partial C_I}{\partial x} = -(\mathbf{M}_{II})^{-1} \frac{\partial A_I^*}{\partial x}$$

## 8. Solución exacta

$$[8.1.1] \quad (\Sigma_*^{-1} + L^T L)U = L^T \Theta^{-1} R$$

$$\Sigma_* = L_* L_*^T$$

$$\Sigma_*^{-1} = L_*^{-1T} L_*^{-1}$$

$$L_a = \begin{pmatrix} L_*^{-1} \\ L \end{pmatrix} = \begin{pmatrix} L_*^{-1} \\ \Theta^{-1} H \end{pmatrix} \in \Re^{(N+n) \times n}$$

$$L_a = \begin{pmatrix} I_{n \times n} \\ \Theta^{-1} H L_* \end{pmatrix} L_*^{-1}$$

$$L_a^T L_a = \Sigma_*^{-1} + L^T L$$

$$v = \begin{pmatrix} 0 \\ \Theta^{-1} R \end{pmatrix}$$

$$L_a^T L_a u = L \Theta^{-1} R = L_a^T v$$

$$\min_u \|L_a u - v\|_2$$

$$u = L_a^+ v$$

$$u = (\Sigma_*^{-1} + L^T L)^{-1} L \Theta^{-1} R$$

$$u = \Sigma_* (I + \Sigma_*^{-1} L^T L)^{-1} L \Theta^{-1} R$$

## 9. Apéndices

### 9.1 A1: Sumas cuadráticas y mínimos cuadrados generalizados

En el texto principal usamos el siguiente resultado general

$$[A1.1] \quad \sum_i (a_i - A_i x)^T (a_i - A_i x) = \sum_i (a_i - A_i \bar{x})^T (a_i - A_i \bar{x}) + (x - \bar{x})^T \left( \sum_j A_j^T A_j \right) (x - \bar{x})$$

donde

$$[A1.2] \quad \bar{x} = \left( \sum_j A_j^T A_j \right)^{-1} \left( \sum_i A_i^T a_i \right)$$

es el vector que minimiza la suma de cuadrados del lado izquierdo de [A1.1] como puede comprobarse fácilmente.

[A1.1] se verifica desarrollando el lado izquierdo de la expresión y sustituyendo  $x$  por  $(x - \bar{x}) + \bar{x}$ . Esto es

$$\begin{aligned} & \sum_i a_i^T a_i - 2((x - \bar{x}) + \bar{x})^T A_i^T a_i + ((x - \bar{x}) + \bar{x})^T A_i^T A_i ((x - \bar{x}) + \bar{x}) = \\ & = \sum_i (a_i - A_i \bar{x})^T (a_i - A_i \bar{x}) + (x - \bar{x})^T \left( \sum_j A_j^T A_j \right) (x - \bar{x}) + \\ & + \left( \sum_k -2(x - \bar{x})^T A_k^T a_k \right) + 2(x - \bar{x})^T \left( \sum_l A_l^T A_l \right) \bar{x} \end{aligned}$$

Y sustituyendo  $\hat{x}$  por su valor en [A1.2] se sigue el resultado buscado.

### 9.2 A2: La matriz de covarianzas de las condiciones iniciales

La matriz de covarianzas de las condiciones iniciales es

$$E(UU^T) = \begin{bmatrix} \Sigma_p & \Gamma_{p,q} \\ \Gamma_{p,q}^T & \sigma^2 I_q \end{bmatrix}$$

donde  $E(Z_p Z_p^T) = \Sigma_p$  es la matriz de autocovarianzas del proceso de orden p,  $E(A_q A_q^T) = \sigma^2 I_q$  y  $E(Z_p A_q^T) = \Gamma_{p,q}$  es la matriz de covarianzas entre las p valores iniciales de la variable output y los q valores iniciales de las innovaciones (ver el documento **Matrices de Autocovarianza en Procesos ARMA. ArmaPolCov.doc**).

### 9.3 A3: El método de Sherman-Morrison

Sea una matriz real cuadrada y regular  $A \in \mathbb{R}^{n \times n}$ , y un par de vectores cualesquiera  $u, v \in \mathbb{R}^n$ . Se trata de invertir la matriz  $A' = A + u \cdot v^T$  conociendo la inversa de  $A$

Obviamente, podemos escribir

$$(A + u \cdot v^T)^{-1} = (I + A^{-1} \cdot u \cdot v^T)^{-1} \cdot A^{-1}$$

Por otro lado, para toda matriz cuadrada se cumple que

$$(I + B)^{-1} = \sum_{k=0}^{\infty} (-B)^k$$

luego

$$(A + u \cdot v^T)^{-1} = \sum_{k=0}^{\infty} (-A^{-1} \cdot u \cdot v^T)^k \cdot A^{-1} =$$

$$A^{-1} - A^{-1} \cdot u \cdot v^T \cdot A^{-1} + A^{-1} \cdot u \cdot v^T \cdot A^{-1} \cdot u \cdot v^T \cdot A^{-1} - A^{-1} \cdot u \cdot v^T \cdot A^{-1} \cdot u \cdot v^T \cdot A^{-1} \cdot u \cdot v^T \cdot A^{-1} + \dots$$

Si definimos

$$\lambda = v^T \cdot A^{-1} \cdot u \in \mathbb{R}$$

entonces

$$(A + u \cdot v^T)^{-1} = A^{-1} - A^{-1} \cdot u \cdot v^T \cdot A^{-1} (1 - \lambda + \lambda^2 - \dots) = \frac{A^{-1} - A^{-1} \cdot u \cdot v^T \cdot A^{-1}}{1 + \lambda}$$

En el caso en que

$$u = (0 \dots \overset{i}{\alpha} \dots 0)^T$$

$$v = (0 \dots \overset{j}{1} \dots 0)^T$$

$$a'_{i,j} = a_{i,j} + \alpha$$

$$a'_{k,l} = a_{k,l} \forall (k,l) \neq (i,j)$$

Así pues, si tenemos una matriz cuya inversa conocemos podemos invertir una matriz que se diferencie de ella en relativamente pocos términos aplicando sucesivamente el método explicado.

Para las matrices  $I_q + L^T L$  y  $M_{II}$  se cumple siempre que no tienen ningún elemento nulo en la diagonal con lo que se puede partir de la matriz que consta únicamente de la diagonal cuya inversa es trivial.

## 9.4 A4: Métodos especiales de resolución de sistemas lineales

Sea una matriz real cuadrada y regular  $A \in \Re^{n \times n}$  de la cual se conoce su inversa

$$(A + B) \cdot u = b$$

$$A^{-1} \cdot (b - B \cdot u) = u$$

$$u_0 = 0$$

$$u_k = A^{-1} \cdot (b - B \cdot u_{k-1})$$