CMA IADCOM - UBA CENTRO DE INVESTIGACIÓN EN METODOS CUANTITATIVOS APLICADOS A LA ECONOMÍA Y LA GESTIÓN

Universidad de Buenos Aires Facultad de Ciencias Económicas

LATINDEX
http://www.latindex.org/latindex/ficha?folio=2751

http://www.economicas.uba.ar/institutos_y_centros/rimf/

CRITERIOS DE INFORMACIÓN Y COMPLEJIDAD ESTOCÁSTICA

Mirta L. GONZALEZ y Alberto H. LANDRO

Universidad de Buenos Aires. Facultad de Ciencias Económicas. Instituto de Investigaciones en Administración, Contabilidad y Métodos Cuantitativos para la Gestión (IADCOM). Centro de Investigaciones en Econometría (CIE). Ciudad Autónoma de Buenos Aires, Argentina.

cie@fce.uba.ar

Resumen

Recibido: 10/2017

Aceptado: 02/2018

Palabras clave

Criterio de Akaike.

Criterio de Scharwz.

Orden de autorregresividad.

Complejidad estocástica.

Church-Turing.

Los criterios objetivos de selección del orden de un modelo autorregresivo pueden ser clasificados en no-Bayesianos -basados en la minimización del error de predicción y las medidas de información- y Bayesianos. La diferencia entre ambos radica en que los primeros asumen como punto de partida la validez de la hipótesis de que todo proceso está afectado por su infinito pasado y proporcionan estimadores asintóticamente eficientes en tanto que los Bayesianos se basan en la negación de la tesis de Church-Turing y proporcionan estimadores consistentes. A fin de evitar la disyuntiva que genera esta clasificación, en este trabajo se propone caracterizar al modelo utilizando la definición de complejidad estocástica. La aplicación de este concepto y los postulados de los teoremas de convergencia de las medidas de complejidad permiten demostrar, además, la condición de óptimo del término de penalización del criterio de selección de Schwarz.

ISSN: 2250-687X - ISSN (En línea): 2250-6861

INFORMATION CRITERIA AND STOCHASTIC COMPLEXITY

Mirta L. GONZALEZ y Alberto H. LANDRO

Universidad de Buenos Aires. Facultad de Ciencias Económicas. Instituto de Investigaciones en Administración, Contabilidad y Métodos Cuantitativos para la Gestión (IADCOM). Centro de Investigaciones en Econometría (CIE). Ciudad Autónoma de Buenos Aires, Argentina.

cie@fce.uba.ar

Abstract

KEYWORDS

Akaike criterion.

Schwartz criterion.

Autoregressive order.

Stochastic complexity.

Church-Turing.

The objective criteria for the selection of the order of an autoregressive model can be classified into non-Bayesians, which are those based on the minimization of the prediction error and on the information measures, and Bayesians. The former group assume the validity of the hypothesis that every process is affected by its infinite past and provides asymptotically efficient estimators, while the Bayesians rely on the denial of the Church-Turing thesis and provide consistent estimators. To avoid the disjunctive generated by this classification, it is proposed to characterize the model through the definition of stochastic complexity. The application of this concept and the postulates of the convergence theorems of the complexity measures allow in addition to demonstrate the optimal condition of the penalty term of the Schwarz selection criterion.

ISSN: 2250-687X - ISSN (En línea): 2250-6861

1.- INTRODUCCIÓN

La especificación de un modelo consiste en la selección, a partir de un conjunto de familias, de la familia de distribuciones de probabilidades condicionadas que mejor represente un conjunto de observaciones.

Sea, en particular, un proceso estocástico unidimensional estacionario (o integrado de orden cero) discreto en el dominio del tiempo, continuo en el dominio de las variables $\{Y_t\}: I(0)\ \ (t = 1,2,...)$ cuya estructura autorregresiva admite una aproximación mediante una representación $AR(p)$ de la forma $\Phi_p(B)Y_t = \varepsilon_t$ (donde las variables Y_t son centradas y $\{\varepsilon_t\}: WN - N(0, \sigma_\varepsilon^2)$). El problema de la especificación del modelo correspondiente consistirá, en este caso, en la adopción de un criterio de selección del orden de autorregresividad condicionante (p) óptimo, entre un conjunto de modelos candidatos, a partir de la información que proporciona la serie cronológica $\{Y_t\}\ \ (t = 1,2,...,n)$ de sus realizaciones pasadas.

Los criterios de selección del orden *p* pueden ser clasificados, en principio, en subjetivos y objetivos. Los subjetivos pueden ser divididos, a su vez, en: i) los que se basan en tests de hipótesis y ii) los que se basan en la función de autocorrelaciones. Los objetivos incluyen los métodos basados en: i) el error de predicción para el período inmediatamente posterior a la última observación; ii) las medidas de información y iii) la inferencia Bayesiana[1].

2.- CRITERIOS DE SELECCIÓN BASADOS EN PRUEBAS DE HIPÓTESIS

Sea $\phi = \left[\phi_1 \ \ \phi_2 \ \ ... \ \ \phi_p\right]^T$ el vector de coeficientes de una representación $AR(p)$ que puede ser particionado de la siguiente forma: $\phi = \left[\phi_j^{(1)T} \ \ \phi_{p-j}^{(2)T}\right]^T$ (donde $\phi_j^{(1)} = \left[\phi_1 \ \ \phi_2 \ \ ... \ \ \phi_j\right]^T$ y $\phi_{p-j}^{(2)} = \left[\phi_{j+1} \ \ \phi_{j+2} \ \ ... \ \ \phi_p\right]^T$). Anderson, T.W. (1963) propuso un test que consiste en contrastar en forma secuencial las hipótesis nulas:

$$H_0^{(1)}: \phi_p = 0$$
$$H_0^{(2)}: \phi_p = \phi_{p-1} = 0$$
$$H_0^{(3)}: \phi_p = \phi_{p-1} = \phi_{p-2} = 0$$
$$.....$$
$$H_0^{(p)}: \phi_p = \phi_{p-1} = ... = \phi_1 = 0$$

con respecto a la alternativa H_1: *que algún de los coeficientes* ϕ_j *sea significativo* . Obsérvese que si en esta secuencia se considera que existen razones suficientes para rechazar la j-ésima hipótesis nula, $H_0^{(j)}: \phi_p = \phi_{p-1} = ... = \phi_{p-(j-2)} = \phi_{p-(j-1)} = 0$, con un nivel de confiabilidad dado, entonces se rechazarán

[1] Para una visión más detallada del estado del arte en el ámbito de la selección de modelos para procesos unidimensionales de parámetro discreto ver Rudra, A. (1954), van der Boom, A.J.W. (1974), Unbenhauen, H.; Göhring, B. (1974), de Gooijer, J.G.; Abraham, B.; Gould, A.; Robinson, L. (1985), Rao, C.R.; Wu, Y. (2001), Landro, A.H.; González, M.L. (2009).

todas las hipótesis nulas siguientes y se seleccionará como más adecuado el modelo $AR(p-j+1)$. Whittle, P. (1952)(1954) demostró que, para contrastar la hipótesis: H_0:*el modelo supuesto es un* $AR(j)$ respecto de la alternativa: H_1:*la verdadera representacion es* $AR(p)$, el cociente de verosimilitudes asume aproximadamente la forma $LR = c(n)\left\{\ln\left[\hat{\sigma}_\varepsilon^2(AR(j))\right] - \ln\left[\hat{\sigma}_\varepsilon^2(AR(p))\right]\right\}$ (donde $\hat{\sigma}_\varepsilon^2(AR(j))$ y $\hat{\sigma}_\varepsilon^2(AR(p))$ denotan los estimadores máximo-verosímiles de σ_ε^2 bajo las hipótesis H_0 y H_1, respectivamente y $c(n)$ denota una función del número de observaciones, que habitualmente se aproxima por $c(n) \approx -(n-p)$) y que, bajo el supuesto de que la hipótesis nula es verdadera, converge en distribución a una función $\chi^2_{(p-j)}$.

Por su parte, Anderson, T.W. (1971) propuso un test alternativo consistente en contrastar la hipótesis nula $H_0: \phi_{j+1} = \phi_{j+2} = ... = \phi_p = 0$ $(j = 0,1,2,..., p-1)$ en una representación $AR(p)$, utilizando el estadístico de Wald de la forma $W = \frac{n}{\hat{\sigma}_\varepsilon^2(AR(p))}\hat{\phi}^{(2)T}\left[\hat{R}_{22} - \hat{R}_{21}\hat{R}_1^{-1}\hat{R}_{12}\right]\phi^{(2)} \xrightarrow{d} \chi^2_{(p-j)}$ (donde: i) las matrices de coeficientes de autocorrelaciones $\hat{R}_{11}$, de orden $(j \times j)$, $\hat{R}_{12}$, de orden $\left[j \times (p-j)\right]$, $\hat{R}_{21}$, de orden $\left[(p-j) \times j\right]$ y $\hat{R}_{22}$, de orden $\left[(p-j) \times (p-j)\right]$ se obtienen de particionar la matriz $\hat{R}_p = \begin{bmatrix} 1 & \hat{\rho}_{-1}(Y) & \hat{\rho}_{-2}(Y) & ... & \hat{\rho}_{-(p-1)}(Y) \\ \hat{\rho}_1(Y) & 1 & \hat{\rho}_{-1}(Y) & ... & \hat{\rho}_{-(p-2)}(Y) \\ ... & ... & ... & ... & ... \\ \hat{\rho}_{p-1}(Y) & \hat{\rho}_{p-2}(Y) & \hat{\rho}_{p-3}(Y) & ... & 1 \end{bmatrix} = \begin{bmatrix} \hat{R}_{11} & \hat{R}_{12} \\ \hat{R}_{21} & \hat{R}_{22} \end{bmatrix}$; ii) el vector $\hat{\phi}^{(2)}$, de orden $(p-j)$, se obtiene de particionar el vector de los estimadores $\hat{\phi} = \begin{bmatrix} \hat{\phi}^{(1)} \\ \hat{\phi}^{(2)} \end{bmatrix}$ y

iii) $\hat{\rho}_j(Y) = \frac{\sum_{t=1}^{n-j} Y_t Y_{t+j}}{\sum_{t=1}^{n} Y_t^2}$).

3.- CRITERIOS DE SELECCIÓN BASADOS EN LA FUNCIÓN DE AUTOCORRELACIONES

A fin de evitar los inconvenientes que trae aparejada la construcción de los modelos para cada valor de p, requerida por los criterios de selección analizados en la sección precedente, Cleveland, W.S. (1972), Chatfield, C. (1979) propusieron la utilización de autocorrelaciones inversas. Para el caso de un proceso $AR(p)$ se comprueba que la función generatriz de autocorrelaciones inversas $\rho_Y^{(I)}(w) = \frac{\gamma_Y^{(I)}(w)}{\gamma_0^{(I)}(w)}$ $(j = 1,2,...)$, (donde $\gamma_Y^{(I)}(w)$ denota la función generatriz de autocovarianzas inversas) asume valores significativos hasta el orden p y después se anula. Bhansali, R.J. (1980) demostró que, en el caso que $\{Y_t\}$:*WN* , el estimador

$$\hat{\rho}_Y^{(I)}(w) = \frac{\hat{\gamma}_Y^{(I)}(w)}{\hat{\gamma}_0^{(I)}} \xrightarrow{d} N\left(0, \frac{1}{n}\right).$$

Otro método que utiliza las autocorrelaciones para identificar el orden p de un modelo autorregresivo es el propuesto por Gray, H.L.; Kelly, G.D.; McIntire, D.D. (1978) y Woodward, W.A.; Gray, H.L. (1978), basado en el comportamiento de dos arreglos de números definidos como $R_j\left[\rho_i(Y)\right] = \frac{H_j\left[\rho_i(Y)\right]}{H_j\left[1, \rho_i(Y)\right]}$ y $S_j\left[\rho_i(Y)\right] = \frac{H_{j+1}\left[1, \rho_i(Y)\right]}{H_j\left[\rho_i(Y)\right]}$ donde:

$$H_j\left[\rho_i(Y)\right] = \begin{vmatrix} \rho_i(Y) & \rho_{i+1}(Y) & \dots & \rho_{i+j-1}(Y) \\ \rho_{i+1}(Y) & \rho_{i+2}(Y) & \dots & \rho_{i+j}(Y) \\ \dots & \dots & \dots & \dots \\ \rho_{i+j-1}(Y) & \rho_{i+j}(Y) & \dots & \rho_{i+2j-2}(Y) \end{vmatrix}$$

$$H_{j+1}\left[\rho_i(Y)\right] = \begin{vmatrix} 1 & 1 & \dots & 1 \\ \rho_i(Y) & \rho_{i+1}(Y) & \dots & \rho_{i+j}(Y) \\ \dots & \dots & \dots & \dots \\ \rho_{i+j-1}(Y) & \rho_{i+j}(Y) & \dots & \rho_{i+2j-1}(Y) \end{vmatrix}$$

(para $j = 1,2,\dots$) y $H_0\left[\rho_i(Y)\right] = 1$. Para una representación $AR(p)$ se verifica que: i) para $j > p$ e $i \neq -p$, $R_j\left[\rho_i(Y)\right] = 0$ y ii) para $j > p$ e $i \neq -p$, $S_j\left[\rho_i(Y)\right]$ es indefinido.

Un método alternativo basado también en un arreglo de autocorrelaciones, es el denominado "*corner method*", debido a Beguin, J.M.; Gourieroux, C.; Montfort, A. (1980). El punto de partida de este criterio consiste en la definición de un determinante de orden $(j \times j)$ de la forma:

$$\Delta(i,j) = \begin{vmatrix} \rho_i(Y) & \rho_{i-1}(Y) & \dots & \rho_{i-j+1}(Y) \\ \rho_{i+1}(Y) & \rho_i(Y) & \dots & \rho_{i-j+2}(Y) \\ \dots & \dots & \dots & \dots \\ \rho_{i+j-1}(Y) & \rho_{i+j-2}(Y) & \dots & \rho_i(Y) \end{vmatrix}$$

respecto del cual se verifica que, dado un proceso $\{Y_t\}: I(0)$, la condición necesaria y suficiente para poder asegurar que la representación AR óptima es de orden p, es que: i) para $i \geq 1$ y $j \geq p+1$, se verifique que $\Delta(i,j) = 0$; ii) para $i \geq 0$, se verifique que $\Delta(i,p) \neq 0$ y iii) para $j \geq p$, se verifique que $\Delta(0,i) \neq 0$. La selección se realiza simplemente a partir de la observación de los estimadores $\hat{\Delta}(i,j)$ (ver también Gooijer, J.G.; Heuts, R.M.J. (1981), Petrucelli, J.D.; Davies, N. (1984)).

Por su parte, Woodward, W.A.; Gray, H.L. (1981) propusieron un "coeficiente de autocorrelación parcial generalizado" de la forma:

$$\phi_{jj}^{(i)}(Y) = \begin{cases} \dfrac{\rho_{i+1}(Y)}{\rho_i(Y)} & si \quad j = 1 \\ \dfrac{\Delta^*(i,j)}{\Delta(i,j)} & si \quad j > 1 \end{cases}$$

(donde $\Delta^*(i,j)$ denota el determinante formado por las primeras $i-1$ columnas de $\Delta(i,j)$, con la i-ésima columna formada por el vector $\left[\rho_{i+1}(Y) \;\; \rho_{i+2}(Y) \;\; \dots \;\; \rho_{i+j}(Y)\right]^T$ y los coeficientes $\phi_{jj}^{(0)}(Y)$ $(j = 1,2,\dots)$ coinciden con los coeficientes de autocorrelación parcial), tal que, para $i \geq 1$ y $j > p$, $\phi_{jj}^{(i)}(Y)$ es indefinido.

Takemura, A. (1984), a partir de la propuesta de Bartlett, M.S.; Dinanda, P.H. (1950), definió una "función de autocorrelaciones generalizada" de la forma:

$$\rho_{i+1,j+1}(Y) = \left[\gamma_{i+j+1}(Y) - \gamma_{(j+1,i)}^{T}(Y)\Gamma^{-1}(j,i)\gamma_{(i+j,-i)}(Y)\right]\bullet$$
$$\bullet\left[\gamma_0(Y) - 2\gamma_{(j,i)}^{T}(Y)\left(\Gamma^{T}(j,i)\right)^{-1}\gamma_{(j+1,i)}(Y) + \gamma_{(j+1,i)}^{T}(Y)\Gamma^{-1}(j,i)\Gamma(0,i)\left(\Gamma^{T}(j,i)\right)^{-1}\gamma_{(j+1,i)}(Y)\right]^{-1}$$

(para $\left|\Gamma(j,i)\right| \neq 0$) y $\rho_{i+1,j+1}(Y) = 0$ (para $\left|\Gamma(j,i)\right| = 0$), donde $\gamma_{(j,i)}(Y) =$

$= \left[\gamma_j(Y) \quad \gamma_{j+1}(Y) \quad \ldots \quad \gamma_{j+i-1}(Y)\right]^{T}$ y $\gamma_{(j,-i)}(Y) = \left[\gamma_j(Y) \quad \gamma_{j-1}(Y) \quad \ldots \quad \gamma_{j-i+1}(Y)\right]^{T}$, para $i > 0$ y:

$$\Gamma(j,i) = \begin{bmatrix} \gamma_j(Y) & \gamma_{j+1}(Y) & \ldots & \gamma_{j+i-1}(Y) \\ \gamma_{j-1}(Y) & \gamma_j(Y) & \ldots & \gamma_{j+i-2}(Y) \\ \ldots & \ldots & \ldots & \ldots \\ \gamma_{j-i+1}(Y) & \gamma_{j-i+2}(Y) & \ldots & \gamma_j(Y) \end{bmatrix}$$

y demostró, además, que la función $\rho_{(i,j)}(Y)$ $(i > 0, j > 0)$ posee las siguientes propiedades: i) para $j \geq 1$, se verifica que $\rho_{(1,j)}(Y) = \rho_{(j)}(Y)$, ii) para $i \geq 1$, se verifica que $\left|\rho_{(i,j)}(y)\right| \leq 1$ y iii) para una representación *AR(p); i>p* y $j \geq 1$, y , se verifica que $\rho_{(i,j)}(Y) = 0$.

Tiao, G.C.; Tsay, R.S. (1983a)(1983b) y Tsay, R.S.; Tiao, G.C. (1984) utilizaron estimadores consistentes de los parámetros ϕ_j para definir el estimador de la llamada "función de autocorrelaciones extendida", $\hat{\rho}_{j(p)} = \hat{\rho}_j\left(Y_t - \sum_{h=1}^{p}\hat{\phi}_{j(h)}^{(k)}Y_{t-h}\right)$. Sea un proceso $\{Y_t\}: AR(p)$ y sea $\hat{\phi}_{p(j)}^{(i)}$ el estimador del coeficiente que acompaña al regresor Y_{t-j} , obtenido de la *i*-ésima iteración de la regresión $AR(p)$. Para calcular el estimador de la función de autocorrelaciones extendida es necesario calcular, en primer término, los estimadores mínimo cuadráticos $\hat{\phi}_{p(j)}^{CM(0)}$ $\left(p = 1,2,\ldots,p^{*}; j = 1,2,\ldots,p\right)$ sucesivamente, desde un *AR(1)* hasta $AR\left(p^{*}\right)$ (siendo p^{*} un valor predeterminado). Tsay, R.C.; Tiao, G.C. (1984) demostraron que:

$$\hat{\rho}_{k(j)} = \begin{cases} c(j-p,k) & para \quad 0 \leq k \leq j-p \\ 0 & para \quad k > j-p \geq 0 \end{cases}$$

(donde $c(j-p,k)$ denota una constante no-nula o una variable aleatoria continua en el dominio $[-1,1]$). Bajo el supuesto que $\Phi_k^{(i)}(B)Y_t = \varepsilon_t: WN$, se verifica que

$\sigma^2\left(\hat{\rho}_{k(j)}\right) \approx \frac{1}{n-j-k}$ (ver Bartlett, M.S. (1946)).

4.- CRITERIOS DE SELECCIÓN BASADOS EN LA MINIMIZACIÓN DEL ERROR DE PREDICCIÓN

Dada la imposibilidad de especificar objetivamente el orden *p* utilizando los criterios analizados en las secciones precedentes, se generaron métodos de selección que atienden al propósito particular para el cual se construye el modelo, por ejemplo, la predicción de los valores futuros de un proceso.

Estos criterios se caracterizan por considerar que el comportamiento autorregresivo de cualquier fenómeno $\{Y_t\}$ está condicionado por su infinito pasado (es decir, que su explicación asume un aspecto semántico) y que, si satisface la condición de estacionariedad de segundo orden, su representación autorregresiva puede ser aproximada a partir de la información que proporciona un conjunto finito de sus realizaciones pasadas.

Desde un punto de vista formal, estos criterios se basan en la minimización de la distancia entre cada uno de los modelos candidatos y el modelo autorregresivo de orden finito que mejor se aproxime al modelo $AR(\infty)$. Esta distancia está definida por una función (o métrica) $D(u,v)$ (donde u y v pueden representar escalares o vectores) que debe satisfacer las siguientes condiciones: i) $D(u,v) > 0$, para $u \neq v$ y, $D(u,v) = 0$ para $u = v$; ii) de simetría $D(u,v) = D(v,u)$, y iii) de desigualdad triangular, $D(u,v) \leq D(u,w) + D(w,v)$. Del conjunto de estas funciones, los métodos basados en el error de predicción tomaron en consideración la distancia $L_2 = \frac{1}{n-p}\sum_{t=p+1}^{n}\left[\hat{Y}_t - Y_t^*\right]^2$ (donde: i) $\hat{Y}_t$ denota el valor de estimado a partir de la aplicación del modelo candidato y ii) $Y_t^* = E\left(Y_t / Y_{t-1}, Y_{t-2}, \ldots\right)$ denota la estructura autorregresiva de $\{Y_t\}$, $Y_t = Y_t^* + \varepsilon_t$, siendo $\{\varepsilon_t\}: WN - N\left(0, \sigma_\varepsilon^2\right)$).

Akaike, H. (1969)(1970) fue el primero en proponer un criterio de selección de este tipo. Sea un proceso $\{Y_t\}: I(0)$ y tal que $m(Y) = 0$, respecto del cual se cuenta con una serie cronológica $\{Y_1, Y_2, \ldots, Y_n\}$ que permite construir un modelo de la forma $\hat{Y}_t = \hat{\phi}_1 Y_{t-1} + \hat{\phi}_2 Y_{t-2} + \ldots + \hat{\phi}_p Y_{t-p}$ $(t = p+1, p+2, \ldots, n)$ (donde $\hat{\phi}_j$ $(j = 1,2,\ldots,p)$ denota el estimador mínimo-cuadrático del coeficiente ϕ_j). El predictor obtenido a partir de este modelo para el período siguiente a la última observación, $\hat{Y}_{n+1} = \hat{\phi}_1 Y_n + \hat{\phi}_2 Y_{n-1} + \ldots + \hat{\phi}_p Y_{n-p+1}$, es tal que, para *n* suficientemente grande, el error medio cuadrático de la predicción puede ser aproximado como $E\left[\left(Y_{n+1} - \hat{Y}_{n+1}\right)^2\right] = \left(1 + \frac{p}{n-p}\right)\sigma_\varepsilon^2$ (ver Box, G.E.P.; Jenkins, G.M. (1976), Yamamoto, Y. (1976)). Suponiendo que el verdadero modelo sea un $AR\left(p^*\right)$ y que $p < p^*$, resulta que el valor esperado del estimador máximo-verosímil del error medio cuadrático de la predicción es de la forma $E\left(\hat{\sigma}_\varepsilon^{(MV)2}\right) = \left(1 - \frac{p}{n-p}\right)\sigma_\varepsilon^2$. De modo que $\frac{\hat{\sigma}_\varepsilon^{(MV)2}}{1 - \frac{p}{n-p}}$ es un estimador insesgado de σ_ε^2. Sustituyendo en la expresión original σ_ε^2 por este estimador, se obtiene el siguiente estimador del error final de la predicción: $FPE(p) = \hat{\sigma}_{p\varepsilon}^2 \frac{n}{n-2p}$ ("final

prediction error"). El valor $\hat{p}$ que minimiza esta expresión es un estimador asintóticamente insesgado y asintóticamente eficiente de p .

Dada una representación $AR(p)\ \left(p = 1,2,...,p^*\right)$ cuyo modelo es estimado n veces, eliminando cada vez una observación y calculando el estimador mínimo-cuadrático de dicha observación, queda definido el llamado "criterio de validación cruzada", basado en la suma de los cuadrados de los errores de predicción ponderados, $\sum_{t=p+1}^{n} \psi_{t(i)}\left(Y_t - \hat{Y}_{t(i)}\right)^2$ (donde $\psi_{t(i)}$ denota la sucesión de ponderaciones y $\hat{Y}_{t(i)}$ denota el estimador de la t-ésima observación calculado a partir de un modelo $AR(p)$ aplicado a la serie cronológica de la cual se ha excluido la i -ésima observación). El estimador del orden p óptimo del modelo se obtiene de minimizar esta suma para $p = 1,2,...,p^*$.

Obsérvese que, a pesar de la hipótesis ya comentada sobre la que se fundan estos criterios, la definición de Akaike (y, en general, de todos los criterios basados en aproximaciones y, como se verá en la próxima sección, en la distancia de Kullback-Leibler) considera al modelo de orden finito que se aproxima mejor al ideal de orden infinito como el verdadero modelo.

Mallows, C.L. (1973)(1995) propuso una aproximación diferente a los criterios basados en la distancia L_2 . Sea la función:

$$J_p = \frac{1}{\sigma_\varepsilon^{*2}} \sum_{t=p+1}^{n} \left[\left(\hat{\phi}_1 Y_{t-1} + ... + \hat{\phi}_p Y_{t-p}\right) - \left(\phi_1 Y_{t-1} + ... + \phi_p Y_{t-p}\right)\right]^2 = \frac{1}{\sigma_\varepsilon^{*2}} \sum_{t=p+1}^{n} \left(\hat{Y}_t - Y_t\right)^2 = \frac{n}{\sigma_\varepsilon^{*2}} L_2$$

(donde σ_ε^{*2} denota la varianza de la verdadera representación), se demuestra que $E\left(J_p\right) = V_p + \frac{B_p}{\sigma_\varepsilon^{*2}}$

(donde $V_p = p$ denota la varianza y B_p denota el sesgo) y que $\frac{1}{\sigma_\varepsilon^{*2}} \sum_{t=p+1}^{n} \hat{\varepsilon}_t^2 - (n - 3p)$ es un estimador insesgado de J_p . Sustituyendo en esta expresión σ_ε^{*2} por su estimador insesgado (calculado a partir del modelo candidato de mayor orden de autorregresividad), se obtiene el criterio de selección $C_p = \frac{1}{\hat{\sigma}_{p^*\varepsilon}^2} \sum_{t=p+1}^{n} \hat{\varepsilon}_t^2 - (n - 3p)$.

Otro criterio de selección conceptualmente similar al C_p , basado en una aproximación a partir de la teoría espectral, es el propuesto por Parzen, E. (1974). Dada una serie $\{Y_t\}\ (t = 1,2,...,n)$, este criterio parte del supuesto que su mecanismo generador autorregresivo es un proceso $AR(\infty)$ de la forma $\Phi_\infty(B)Y_t = \varepsilon_t$ y que este proceso puede ser aproximado adecuadamente por una representación $AR(p)$ de orden finito y propone la siguiente definición de distancia entre estas dos representaciones: $J_p = \frac{1}{2\pi} \int_{-\pi}^{\pi} \left| \frac{\Phi_p\left(e^{iw}\right) - \Phi_\infty\left(e^{iw}\right)}{\Phi_\infty\left(e^{iw}\right)} \right|^2 dw$. Si se sustituyen en esta expresión los coeficientes de la representación $AR(p)$ por sus estimadores de Yule-Walker, de acuerdo con Kromer, R.E. (1969), se verifica que $E\left\{ \frac{1}{2\pi} \int_{-\pi}^{\pi} \left| \frac{\hat{\Phi}_p\left(e^{iw}\right) - \Phi_\infty\left(e^{iw}\right)}{\Phi_\infty\left(e^{iw}\right)} \right|^2 dw \right\} = 1 - \frac{\sigma_{\infty\varepsilon}^2}{\sigma_{p\varepsilon}^2} + \frac{p}{n}$ (donde $1 - \frac{\sigma_{\infty\varepsilon}^2}{\sigma_{p\varepsilon}^2}$

representa el sesgo debido a la aproximación de $\Phi_\infty(B)$ por $\Phi_p(B)$ y $\frac{p}{n}$ y representa la varianza de los $\hat{\phi}_j$ en el operador $\Phi_p(B)$). El estimador óptimo (asintóticamente eficiente) de p se obtiene de minimizar esta expresión.

Posteriormente, Parzen, E. (1975) modificó este método de selección y propuso un criterio, al que denominó CAT ("*criterion for autoregressive transfer functions*"), definido por:

$$CAT(p) = \begin{cases} \frac{1}{n^2}\sum_{j=1}^{p}\frac{n-j}{\hat{\sigma}_{j\varepsilon}^2} - \frac{n-p}{n\hat{\sigma}_{p\varepsilon}^2} & \left(p = 1,2,...,p^*\right) \\ -\left(1+\frac{1}{n}\right) & \left(p = 0\right) \end{cases}$$

(para una discusión más detallada sobre este criterio y sus modificaciones, ver Parzen, E. (1977)(1978) (1979)(1980), Parzen, E.; Pagano, M. (1979), Beamish, N.; Priestley, M.B. (1981)).

5.- CRITERIOS DE SELECCIÓN BASADOS EN MEDIDAS DE INFORMACIÓN

Otra función de distancia utilizada en los métodos de predicción es la conocida como distancia dirigida de Kullback-Leibler. Sea Y una variable aleatoria continua con función de densidad conocida, $f(Y/\beta)$ (donde $\beta = \left[\beta_1 \quad \beta_2 \quad ... \quad \beta_k\right]^T$ denota un vector k -dimensional de coeficientes desconocidos), sea un vector β^* formado por los supuestos verdaderos valores de β y sea $f(Y/\beta^*)$ la correspondiente supuesta verdadera función de densidad. Si el problema consiste en seleccionar los coeficientes β que se aproximen en la mayor medida de lo posible al vector β^*, la distancia entre las distribuciones $f(Y/\beta^*)$ y $f(Y/\beta)$ puede ser caracterizada por una medida de entropía de la forma (ver Akaike, H. (1978b)):

$$D(\beta^*,\beta) = \int_{-\infty}^{\infty} f(y/\beta^*)\ln\left[f(y/\beta)\right]dy - \int_{-\infty}^{\infty} f(y/\beta^*)\ln\left[f(y/\beta^*)\right]dy$$

(donde el primer sumando del segundo miembro representa la capacidad para adecuarse de $f(Y/\beta)$ respecto de $f(Y/\beta^*)$ y el segundo sumando, para una función $f(Y/\beta^*)$ dada, es una constante). La minimización de la medida de entropía implica la minimización del criterio de información[2]:

$$KL(\beta^*,\beta) = -D(\beta^*,\beta) = \int_{-\infty}^{\infty}\left\{\ln\left[f(y/\beta^*)\right] - \ln\left[f(y/\beta)\right]\right\} f(y/\beta^*)dy$$

Suponiendo que $\beta = \beta^* + \Delta\beta$ (donde $\Delta\beta = \left[\Delta\beta_1 \quad \Delta\beta_2 \quad ... \quad \Delta\beta_k\right]^T$ es un vector de norma arbitrariamente pequeña), entonces el criterio $KL(\beta^*,\beta)$ admite un desarrollo en serie de Taylor de la forma:

$$KL(\beta^*,\beta^*+\Delta\beta) \approx \int_{-\infty}^{\infty}\left\{-\sum_i (\Delta\beta)_i \frac{\partial\log\left[f(y/\beta^*)\right]}{\partial\beta_i^*} - \frac{1}{2}\sum_i\sum_j (\Delta\beta)_i(\Delta\beta)_j \frac{\partial^2\log\left[f(y/\beta^*)\right]}{\partial\beta_i^*\partial\beta_j^*}\right\} \bullet$$
$$\bullet\, f(y/\beta^*)dy$$

[2] Ver Kullback, S. (1959).

Si $f\left(y/\beta^*\right)$ es una función regular, el primer término del segundo miembro de esta expresión se anula y, en consecuencia, resulta que $KL\left(\beta^*,\beta^*+\Delta\beta\right)\approx\frac{1}{2}\|\Delta\beta\|_I^2$ (donde $\|\Delta\beta\|_I^2=\Delta\beta^T I\left(\beta^*\right)\Delta\beta$, siendo $\|\bullet\|_I^2$ la norma Euclidiana e $I(\bullet)$ la matriz de información de Fisher). Supóngase que β esté incluido en un espacio s-dimensional Θ_s $(s=1,2,...,k-1)$, en tanto que el vector de los verdaderos valores de los coeficientes, β^*, está incluido en un espacio k-dimensional ($k>s$). Denotando por β_s^* a la proyección de β^* sobre Θ_s en el sentido de la norma Euclidiana, se demuestra que $2KL\left(\beta^*,\beta_s\right)\approx\left\|\beta_s^*-\beta^*\right\|_I^2+\left\|\beta_s-\beta_s^*\right\|_I^2$ (donde $\beta_s\in\Theta_s$ y se verifica que $\beta_s\approx\beta_s^*$). Reemplazando β_s por el vector de variables aleatorias $\hat{\beta}_s$ formado por los estimadores máximo-verosímiles restringidos de β^* en Θ_s y, teniendo en cuenta que, para valores de n suficientemente grandes, $n\left\|\beta_s^*-\hat{\beta}_s\right\|_I^2\xrightarrow{d}\chi_{(s)}^2$, se verifica que $2E\left[KL\left(\beta^*,\hat{\beta}_s\right)\right]\approx\left\|\beta_s^*-\beta^*\right\|_I^2+\frac{s}{n}$. Esta expresión constituye una medida de los desvíos de $\hat{\beta}_s$ respecto del vector β^* y permite concluir que el valor esperado de este desvío incluye una componente que representa el error relacionado con la selección de un espacio de coeficientes aproximado por β_s^* y otra que representa el error debido a la estimación del vector de los coeficientes. Akaike demostró que, bajo ciertas condiciones de regularidad, el cociente de verosimilitudes $LR(Y)=-2\sum_{j=1}^{n}\log\left[\frac{f\left(y_i/\hat{\beta}_s\right)}{f\left(y_i/\hat{\beta}^{(MV)}\right)}\right]\xrightarrow{d}\chi_{NC(k-s)}^2\left(\left\|\beta_s^*-\beta^*\right\|_I^2\right)$ y, por lo tanto, que $\frac{1}{n}\left[LR(Y)+2s-k\right]$ es un estimador insesgado de la medida $E\left[KL\left(\beta^*,\hat{\beta}_s\right)\right]$. El criterio de información de Akaike (AIC)[3] consiste en minimizar el logaritmo de la función de verosimilitud $-2L_n\left(Y,\hat{\beta}_s\right)+2s$ $(s=1,2,...,k-1)$ en la cual el primer término representa la medida del error debido a la falta de capacidad para adecuarse a la aproximación y el segundo término define el factor de penalización. Bajo el supuesto de Normalidad del supuesto verdadero modelo, su función de densidad asume la forma $f\left(\hat{Y}^*\right)=\frac{1}{\left(\sigma_\varepsilon^*\sqrt{2\pi}\right)^{n-p}}\exp\left\{-\frac{1}{2\sigma_\varepsilon^{*2}}\sum_{t=p+1}^{n}\left[Y_t-Y_t^*\right]^2\right\}$ y la función de verosimilitud del modelo candidato ($\hat{Y}_t^p$) será de la forma $f\left(\hat{Y}^p\right)=\frac{1}{\left(\sigma_{p\varepsilon}\sqrt{2\pi}\right)^{n-p}}\exp\left[-\frac{1}{2\sigma_{p\varepsilon}^2}\sum_{t=p+1}^{n}\left(Y_t-\phi_1Y_{t-1}-...-\phi_pY_{t-p}\right)^2\right]$. Por lo tanto, la distancia de Kullback-Leibler asumirá la forma:

$$KL=\frac{2}{n-p}E\left[\ln\left(\frac{f\left(\hat{Y}^*\right)}{f\left(Y^p\right)}\right)/f\left(\hat{Y}^*\right)\right]=$$

$$=\ln\left(\frac{\sigma_{p\varepsilon}^2}{\sigma_\varepsilon^{*2}}\right)+\frac{\sigma_\varepsilon^{*2}}{\sigma_{p\varepsilon}^2}\frac{1}{n-p}\sum_{t=p+1}^{n}\left[m\left(\hat{Y}_t^*\right)-\phi_1Y_{t-1}-...-\phi_pY_{t-p}\right]^2-1$$

[3] *"Akaike´s information criterion"*.

De modo que, sustituyendo en esta expresión los coeficientes ϕ_j, σ_ε^{*2} y $\sigma_{p\varepsilon}^2$ por sus estimadores máximo-verosímiles, se obtiene que $KL = \ln\left(\frac{\hat{\sigma}_{p\varepsilon}^{(MV)2}}{\hat{\sigma}_\varepsilon^{*(MV)2}}\right) + \frac{\hat{\sigma}_\varepsilon^{*(MV)2}}{\hat{\sigma}_{p\varepsilon}^{(MV)2}} + \frac{L_2}{\hat{\sigma}_{p\varepsilon}^{(MV)2}} - 1$. A partir de esta definición resulta el siguiente criterio de selección: $AIC(p) = \ln\left(\hat{\sigma}_{p\varepsilon}^{(MV)2}\right) + \frac{2(p+1)}{n-p}$, que permite obtener un estimador $\hat{p} = \min_p AIC(p)$ asintóticamente eficiente.

6.- CRITERIOS DE SELECCIÓN BAYESIANOS

Estos criterios se caracterizan por considerar que la verdadera representación del comportamiento de $\{Y_t\}$ es de orden finito, es decir, que la verdadera representación está incluida en el conjunto de las representaciones candidatas, de modo que el método de selección permite identificar asintóticamente el orden de autorregresividad de dicha representación.

Sea una variable aleatoria continua Y y sea $y = [y_1 \quad y_2 \quad \ldots \quad y_n]^T$ una muestra sobre dicha variable. Supóngase que existan dos modelos alternativos capaces de proporcionar una representación adecuada del conjunto de observaciones de la muestra, caracterizados por las hipótesis H_1 y H_2 representadas por las funciones de densidad $f(Y/\beta_1)$ y $f(Y/\beta_2)$, donde β_1 y β_2 denotan los vectores de orden k_1 y k_2, respectivamente, de los coeficientes desconocidos. Desde un punto de vista Bayesiano, la selección entre H_1 y H_2 puede ser hecha a partir del cociente entre probabilidades "a posteriori" $K_{12} = \frac{p(H_1/Y)}{p(H_2/Y)} = \frac{p(H_1)}{p(H_2)}\frac{p(Y/H_1)}{p(Y/H_2)}$, (donde $p(H_i)\,(i=1,2)$ denota la probabilidad "a priori" de que la hipótesis H_i sea verdadera). Este cociente de probabilidades "a posteriori" es una medida de la potencia relativa del grado de creencia sobre las hipótesis alternativas, condicionada por la evidencia que proporciona la muestra. Seleccionar el modelo al cual le corresponde la mayor probabilidad "a posteriori", $p(H_i/Y)\,(i=1,2)$, es equivalente a minimizar la pérdida esperada asociada a la aceptación de una de las dos hipótesis. Según que K_{12} sea mayor o menor que uno, el modelo elegido será H_1 o H_2, si $K_{12}=1$, la selección es indiferente en términos de pérdida esperada (obviamente esta regla de selección resulta óptima sólo en el caso de simetría de las funciones de pérdida de los modelos). El problema que plantea este método de selección radica en la dificultad para especificar completamente la probabilidad "a priori", $p(H_i)\,(i=1,2)$. De acuerdo con Jeffrey, H. (1961), este problema se resuelve asignando la misma probabilidad a ambos modelos. A partir de esta sugerencia Chow, G.C. (1981) demostró que, con una información "a priori" mínima, se obtiene que:

$$\ln\left[p(Y/H_i)\right] = \ell_n(Y,\beta_i) - \frac{k}{2}\ln(n) - \frac{1}{2}\ln\left[\left|I(\hat{\beta}_i)\right|\right] + \frac{k_i}{2}\ln(2\pi) + \ln\left[f_{H_i}(\hat{\beta})\right] + O(n^{-1/2})$$

(donde: i) ℓ_n denota el logaritmo de la función de verosimilitud; ii) $f_{H_i}(\hat{\beta})\ (i=1,2)$ denota la función de densidad "a priori" de los estimadores $\hat{\beta}$ cuando el modelo H_i es el verdadero; iii) $\hat{\beta}$ es el

estimador máximo-verosímil del vector $\beta_i\ (i=1,2)$; iv) $\left|I\left(\hat{\beta}_i\right)\right|$ denota el valor del determinante de la matriz de información para $\beta_i=\hat{\beta}_i$ y v) $O_p\left(n^{-1/2}\right)$ denota el orden de la probabilidad). El problema, en este caso, radica en la dificultad para definir las funciones de densidad "a priori" para cada modelo, $f_{H_i}(\beta)$. La solución propuesta por Schwarz, G. (1978) consiste, en considerar solamente los dos primeros términos del segundo miembro de la expresión que figura más arriba y seleccionar, para n suficientemente grande, el modelo que maximice la expresión $SIC=\ell_n\left(Y,\hat{\beta}_i\right)-\frac{k_i}{2}\ln(n)$. A diferencia del AIC, este criterio es fuertemente consistente para modelos autorregresivos y resulta asintóticamente equivalente a la minimización de la función $S(p)=(n-p)\ln\left(\sigma_\varepsilon^2\right)+p\ln(n-p)\ \ \left(p=1,2,...,p^*\right)$ (ver Geweke, J.F.; Meese, R.A. (1981)). A fin de evitar el problema relacionado con la especificación de $f_{H_i}(\beta)$, Akaike, H. (1977)(1978a)(1979) propuso como criterio de selección Bayesiano el mínimo del valor esperado "a posteriori" de la función de pérdida, $E_{H_i}\left\{\ln\left[f\left(Y/\beta\right)\right]\right\}$: $SIC(p)=(n-k_i)\ln\left[\frac{\hat{S}_i(Y)}{n-k_i}\right]+\frac{1}{2}\ln\left[\frac{n\sigma_Y^2-S_i(Y)}{k_i}\right]$ (donde $\hat{\sigma}_y^2$ denota la varianza de las observaciones). En el caso particular de una representación $AR(p)$, esta expresión asume la forma:

$$SIC(p)=(n-p)\ln\left(\hat{\sigma}_\varepsilon^{(MV)2}\right)-(n-2p)\ln\left(1-\frac{p}{n-p}\right)+p\ln(n-p)+p\ln\left[\frac{1}{p}\left(\frac{\hat{\sigma}_Y^{(MV)2}}{\hat{\sigma}_\varepsilon^{(MV)2}}-1\right)\right]$$

Como se verá en la próxima sección, Rissanen, J. (1978)(1980), basándose en el principio de minimización del número de dígitos binarios necesario para reproducir una sucesión de observaciones, propuso un criterio de selección del orden p, minimizando la función $BIC^*(p)=\ln\left(\hat{\sigma}_\varepsilon^{(MV)2}\right)+\frac{p\ln(n-p)}{n-p}$ y demostró que el estimador del orden p que se obtiene es consistente.

Por su parte, Hannan, E.J.; Quinn, B.G. (1979), a partir de los postulados de los teoremas de Heyde, C.C. (1974) y Heyde, C.C.; Scott, D.J. (1973) propusieron un criterio de selección de la forma $HQ(p)=\ln\left(\hat{\sigma}_{p\varepsilon}^{(MV)2}\right)+\frac{cp\ln\left[\ln(n-p)\right]}{n-p}\ \ \left(p=1,2,...,p^*\right)$, donde c es una constante a ser determinada y demostraron que, si se selecciona $c>2$, el estimador $\hat{p}$ que se obtiene es consistente.

7.- CRITERIOS DE SELECCIÓN Y TEORÍA DE LA COMPLEJIDAD ESTOCÁSTICA

De las consideraciones realizadas en las secciones anteriores, se concluye que los criterios de selección del orden p en una representación $AR(p)$ también pueden ser clasificados en asintóticamente eficientes y consistentes. Como se vio, los primeros coinciden con aquellos que consideran que la verdadera representación autorregresiva del proceso no pertenece al conjunto de los modelos candidatos, por el contrario, los criterios consistentes surgen de considerar que la verdadera representación autorregresiva es de orden finito.

En otros términos, la consistencia del estimador del orden sólo se obtiene si se supone que el proceso es anticipatorio (de acuerdo con la nomenclatura de Rosen, R. (1985)), es decir, si se supone que la tesis de Turing-Church es falsa. Esta tesis supone que la complejidad de la estructura autorregresiva de un proceso unidimensional involucra más vínculos que cualquier expresión sintáctica (modelo) de su comportamiento. Por lo tanto, su negación implica la aceptación de la interpretación reduccionista y el reconocimiento de que la explicación del comportamiento de todo proceso consiste únicamente en la búsqueda de la verdadera sintaxis.

A fin de superar la polémica acerca de la existencia o no de un mecanismo generador de las realizaciones de un proceso estimable a partir de un conjunto finito de observaciones y de los verdaderos valores de los coeficientes de su representación, se propone considerar una clase de modelos que no posee un número determinado de coeficientes y cuya única pretensión es la de proporcionar una sintaxis capaz de permitir simplemente la descripción de las regularidades locales observadas en el comportamiento del proceso, a partir de una serie cronológica finita.

Sea un proceso $\{Y_t\}$ formado por variables aleatorias binomiales, el descubrimiento de posibles regularidades en sus realizaciones eliminando la información redundante, se puede lograr a partir de la determinación de la medida del grado de complejidad algorítmica del conjunto de observaciones, entendida como la medida del menor programa (modelo) necesario para generar la sucesión $\{Y_t\}$ $\left(t=1,2,...,n\,;\,\Omega(Y_t)=\{0,1\}\right)$. Este criterio, debido a Risanen, J. (1978)(1983) y conocido como "de la medida mínima para la descripción de $\{Y_t\}$ " (MDL, "minimum description length"), se basa en los trabajos de Kolmogorov, A.N. (1956), Solomonoff, R.J. (1964), Chaitin, G. (1975)(1987)(1990) y Thom, R. (1975) y permite la estimación de los coeficientes junto con la estimación del número de coeficientes. Igual que en el caso de la estimación por máxima verosimilitud, se selecciona una clase paramétrica de funciones de probabilidades $p_{(k,\theta_1,\theta_2,...,\theta_k)}(Y_1,Y_2,...,Y_n)$ $(k=0,1,2,...)$ que satisfacen la condición de compatibilidad de Kolmogorov, según la cual $\lim_{\substack{Y_{j+1}\to\infty \\ \cdots\cdots \\ Y_n\to\infty}} p_{(k,\theta_1,...,\theta_k)}(Y_1,Y_2,...,Y_n)=p_{(k,\theta_1,...,\theta_k)}(Y_1,Y_2,...,Y_j)$. Cada una de estas funciones le asigna una probabilidad a una sucesión $\{Y_1,Y_2,...,Y_n\}$, en la que cada realización puede ser expresada como una serie binaria de α dígitos (es decir, cada realización puede ser expresada con una precisión $\alpha<\infty$), de modo que la sucesión de n realizaciones $y=\{y_1,y_2,...,y_n\}$, puede ser descripta con algo más de $n\alpha$ "bits". Obviamente, si se tienen en cuenta las posibles correlaciones entre las variables Y_t, el número de "bits" necesarios para la descripción de la sucesión y disminuirá y, de acuerdo con el criterio MDL, se hará mínimo cuando se consideren todas las regularidades existentes en dicha sucesión, es decir, cuando la codificación de la sucesión se realice utilizando la verdadera función de probabilidades, $p_{(k,\theta_1,...,\theta_k)}(Y_1,Y_2,...,Y_n)$, generadora de las realizaciones. Debe tenerse en cuenta que la sucesión binaria de símbolos que la función de codificación ($C(y)$) le asigna a cada sucesión $\{y_1,y_2,...,y_n\}$, es de una medida ($L(y)$) que satisface la desigualdad de Kraft $\sum_y 2^{-L(y)}\le 1$, (ver Abramson, N. (1968)). La desigualdad se verifica si $C(y)$ puede ser considerada como una descripción auto contenida de la sucesión $\{y_1,y_2,...,y_n\}$ que incluye la información sobre su propia extensión (es decir si $C(y)$ posee la llamada "propiedad del prefijo", que implica que ninguna función $C(y)$ constituye un prefijo de otra función $C'(y)$). Si $C(y)$ es tal que posee una longitud media mínima, entonces se verifica la igualdad y $2^{-L(y)}$ define una distribución de probabilidades.

Si se selecciona un modelo $p_{(\theta_1,\theta_2,...,\theta_k)}(y_1,y_2,...,y_n)$ y se asigna a la sucesión $\{y_1,y_2,...,y_n\}$ un código binario formado por $-\ln\left[p_{(\theta_1,\theta_2,...,\theta_k)}(y_1,y_2,...,y_n)\right]$ símbolos, entonces la longitud promedio del código sobre todas las sucesiones de medida n será tal que $-\sum p_{\theta^*}(y)\ln\left[p_\theta(y)\right] \le -\sum p_{\theta^*}(y)\ln\left[p_{\theta^*}(y)\right]$ (donde θ^* denota el vector formado por los verdaderos valores de los coeficientes). Entonces, dada una distribución de probabilidades $p_\theta(y)$ con θ fijo, es posible hallar, para cada sucesión $\{y_1,y_2,...,y_n\}$, un función de codificación $C(y)$ definida como una sucesión binaria de longitud ideal $-\ln\left[p_\theta(y)\right]$. Por otra parte, dada una familia paramétrica de distribuciones, para definir el "mejor código" es necesario seleccionar el vector $\hat{\theta}$ que minimice $-\ln\left[p_\theta(y)\right]$, es decir, seleccionar el estimador máximo verosímil.

Ahora bien, el problema que permanece se refiere a la codificación de los coeficientes y la medida de dicha codificación. Independientemente de la definición de la función de codificación, la longitud del código puede ser determinada sólo si se satisface la condición $\sum 2^{-L(\theta)} \le 1$ (donde θ varía en el dominio de los estimadores $\hat{\theta}$).

Dada una representación con un vector de coeficientes $\theta=[\theta_1 \quad \theta_2 \quad ... \quad \theta_k]^T$, la medida total ideal ($L(y)$) del código necesario para explicar la sucesión $y=\{y_1,y_2,...,y_n\}$ está dada por el criterio MDL, $L(y)=-\ln\left[p_\theta(y)\right]+\frac{k}{2}\ln(n)$, bajo el supuesto de que no se cuenta con ningún conocimiento "a priori" acerca de los coeficientes θ.

Se denominan estimadores MDL, $\hat{\theta}=\left[\hat{\theta}_1 \quad \hat{\theta}_2 \quad ... \quad \hat{\theta}_k\right]^T$, a aquellos que minimizan la medida $L(y)$ (obsérvese que, cuando la información "a priori" permite determinar el valor de k, los estimadores MDL se asimilan a los estimadores máximo-verosímiles).

Se dice que un código es regular si su medida, $L(y_1,y_2,...,y_n)$, además de satisfacer la desigualdad de Kraft para todo n, es tal que $L(y_1,y_2,...,y_{n+1}) \ge L(y_1,y_2,...,y_n)$. En particular, si se verifica que $\sum 2^{-L(y_1,y_2,...,y_n)}=1$, entonces se puede asegurar que existe una correspondencia biunívoca entre códigos regulares y procesos estocásticos. De acuerdo con Risanen, J. (1984), la medida $L(y)$ calculada a partir de estimadores MDL es óptima en el dominio de los códigos regulares y representa la medida de la información contenida en las realizaciones $y=[y_1 \quad y_2 \quad ... \quad y_n]^T$ con respecto a un modelo dado. Esta medida es una combinación de la información probabilística de Shanon y la información combinatoria o algorítmica de Kolmogorov y se origina en dos interpretaciones de la complejidad estocástica, la primera como la menor medida de la sucesión binara necesaria para su codificación y la segunda como el menor error de predicción.

Esta segunda interpretación considera una codificación predictiva de las observaciones (criterio MDL-predictivo, estrechamente vinculado con el principio presecuencial de Dawid, A.P. (1984)) que permite: i) definir un límite inferior para la medida del modelo necesario para codificar una sucesión larga de observaciones, dada una clase determinada de modelos; ii) definir un límite inferior universal para el valor esperado de los errores de predicción cualquiera sea la función de

predicción adoptada y iii) realizar un análisis de la calidad de los estimadores, incluyendo el estimador del número de coeficientes[4].

Sea un modelo estocástico capaz de codificar el conjunto de observaciones $\{y_1, y_2, ..., y_t\}$ con una precisión implícita α, definido por una función de probabilidades de la forma $p_{(k,\hat{\theta}_1,...,\hat{\theta}_k)}(y_1, y_2, ..., y_t)$ $(k = 0,1,2,...)$ y sea $f_{(k,\hat{\theta}_1,...,\hat{\theta}_k)}(Y_{t+1} / y_1, y_2, ..., y_t)$ la función de densidad de la variable Y_{t+1}, condicionada por la sucesión de observaciones $\{y_1, y_2, ..., y_t\}$. Esta función de densidad permite codificar la observación y_{t+1} con un código ideal de medida $-\ln\left[p_{(k,\hat{\theta}_1,...,\hat{\theta}_k)}(Y_{t+1} / y_1, y_2, ..., y_t)\right]$ y con una precisión α, la cual está representada entonces por $-\ln\left[f_{(k,\hat{\theta}_1,...,\hat{\theta}_k)}(Y_{t+1} / y_1, y_2, ..., y_t)\right]$ (que la medida del código sea ideal implica que, si la variable Y_{t+1} se distribuye de acuerdo con la función de probabilidades del modelo, no existe ningún prefijo con medida promedio menor que dicho código). Sumando todas las medidas de los códigos ideales, se obtiene la medida de la codificación predictiva de la sucesión $\{y_1, y_2, ..., y_t\}$, $L^{(\Pr)}(y / k) = -\sum_{t=0}^{n-1} \ln\left[f_{(k,\hat{\theta}_1,...,\hat{\theta}_k)}(Y_{t+1} / y_1, y_2, ..., y_t)\right]$. Esta puede ser minimizada con respecto a k, a fin de obtener un estimador $\hat{k}(y_1, y_2, ..., y_t)$ que permita calcular el estimador final $\hat{\theta}(y_1, y_2, ..., y_t) = \hat{\theta}(k(t)) = \left[\hat{\theta}_1(t) \;\; \hat{\theta}_2(t) \;\; ... \;\; \hat{\theta}_{k(n)}(t)\right]^T$. Risanen, J. (1984b) demostró que, en condiciones bastante amplias con respecto a los regresores, los estimadores $\hat{k}(t) = \hat{k}(y_1, y_2, ..., y_t)$ son consistentes.

A fin de evitar los problemas conceptuales y técnicos vinculados a la especificación de las funciones de densidad "a priori" de los coeficientes θ para la minimización de $L(y / k)$, es posible adoptar una metodología consistente en: i) especificar una función de densidad $f_{Y_1}(y_1)$ para la primera variable aleatoria (lo cual es, indudablemente, menos complicado que especificar una función de densidad para todos los coeficientes θ); ii) codificar la primera observación, y_1, con un código de medida $-\ln\left[f_{Y_1}(y_1)\right]$; iii) codificar las restantes observaciones utilizando la misma función de densidad, hasta que un coeficiente pueda ser estimado y iv) en forma similar, incrementar sucesivamente el número de coeficientes estimados hasta completar los k coeficientes necesarios para la minimización de $L(y / k)$.

De acuerdo con Risanen, J. (1983), la codificación del número natural k requiere $L^*(k) = \ln^*(k) + \ln(c)$ de dígitos binarios (donde $\ln^*(k) = \ln(k) + \ln[\ln(k)] + ...$, incluyendo la suma todas las iteraciones positivas y c es una constante aproximadamente igual a 2,865, que hace que $\sum_{n=1}^{\infty} 2^{-L^*(k)} = 1$). Por lo tanto, se puede definir una medida de la "complejidad estocástica semi-predictiva" de la sucesión $y = \{y_1, y_2, ..., y_t\}$, para una determinada clase de modelos, como:

$$L^{(S\,\Pr)}(Y / k) = (y) = \min_k \left[L(y / k) + \ln^*(k) + c\right]$$

(la calificación de esta complejidad como semi-predictiva se debe a que la optimización del número de coeficientes no se obtuvo en forma predictiva).

Asimismo, de acuerdo con Risanen, J. (1978)(1983) existe una medida de la complejidad no-predictiva de la sucesión $\{y_1, y_2, ..., y_t\}$ definida de la siguiente forma:

$$L^{(N\,\Pr)}(Y / k) = \min_{k,\theta}\left\{-\ln\left[f_{(k,\theta)}(y)\right] + \frac{k}{2}\ln(t)\right\}$$

donde el término $\frac{k}{2}\ln(t)$ representa el número de dígitos requerido para codificar k coeficientes con una precisión óptima (debe tenerse en cuenta que esta expresión es sólo formalmente equivalente al criterio de Schwarz, G. (1978)).

[4] Si el predictor es el que minimiza el error medio cuadrático de la predicción, entonces el principio MDL se asimila al criterio de predicción de los cuadrados mínimos (ver, Risanen, J. (1986)).

Dada una función real $L(y)$ que satisface la desigualdad $L(y) \geq -\ln\left[f_Y(y)\right]$ (donde $f_Y(y)$ denota la función de densidad del proceso estocástico $\{Y_t\}$) y, por lo tanto, la relación de Kraft, se puede concluir fácilmente que la desigualdad se verifica: i) para la complejidad semi-predictiva, si $f_Y(y) = \sum_{k=1}^{n} 2^{-L^*(k)} 2^{-L(y/k)}$ (donde $L(y/k) = \sum_{t=0}^{n-1} \ln\left[f_{k,\hat{\theta}(t)}(Y_{t+1} / y_1, y_2, ..., y_n)\right]$); ii) para la complejidad predictiva, si $f_Y(y) = 2^{-L^{(\Pr)}(Y)}$ y iii) para la complejidad no-predictiva, para los términos de orden menor o igual que t. Además, Risanen, J. (1986) (teorema 1) demuestra que:

$$E_{k,\theta}\left[L(y_1, y_2, ..., y_t)\right] \geq H_{k,\theta}(t) + \left(\frac{1}{2} - \varepsilon\right) k \ln(t)$$

(donde $H_{k,\theta}(t)$ denota la medida de la entropía de la sucesión $y = \{y_1, y_2, ..., y_t\}$). Se puede concluir, entonces, que la desigualdad de Shanon es un caso particular de esta expresión, para $k = 0$ y que los resultados anteriores permiten considerar a la noción de complejidad estocástica como un concepto capaz de proporcionar una base racional a un criterio de selección de modelos, independientemente del número de sus coeficientes.

Los teoremas 2 y 3 de Risanen, J. (1986) demuestran que las tres complejidades convergen casi-con-certeza al límite inferior $-\ln\left[f_{k,\theta}(y)\right] + \frac{k}{2}\ln(n)$, lo cual permite concluir que el término $\frac{k}{2}\ln(n)$ puede ser considerado óptimo para representar la complejidad del modelo que, en el caso particular de los procesos unidimensionales, coincide con su grado de autorregresividad.

8.- CONCLUSIONES

Los criterios objetivos de selección del orden de autorregresividad en los modelos $AR(p)$ pueden clasificarse en aquellos que consideran la minimización del error de predicción, aquellos que minimizan la función de distancia de Kullback-Leibler y aquellos basados en métodos de inferencia Bayesiana.

Los dos primeros se fundan en la hipótesis que la estructura autorregresiva del proceso unidimensional $\{Y_t\}$ está condicionado por su infinito pasado, es decir, que involucra más vínculos que los que puede contener cualquier modelo de su comportamiento, en otros términos, que su verdadera representación autorregresiva es de orden infinito y que, por lo tanto, no está incluida en el conjunto de las representaciones candidatas de orden indiscutiblemente finito. Los estimadores que proporcionan estos criterios son asintóticamente eficientes.

Por el contrario, los criterios Bayesianos cuyo fundamento conceptual se origina en la negación de la tesis de Church-Turing y, por lo tanto, en el supuesto de la existencia de un mecanismo generador que admite una representación asintóticamente identificable por un modelo $AR(p)$ $\left(p = 1,2,..., p^*\right)$ de orden finito, proporcionan estimadores consistentes.

Como una forma de evitar esta disyuntiva entre estimadores asintóticamente eficientes y consistentes, este trabajo propone utilizar el concepto de complejidad estocástica (debido a Kolmogorov, Solomonoff, Thom y Chaitin) como argumento que permite considerar a un modelo como una estructura capaz de proporcionar una sintaxis que describa las regularidades locales observadas en el comportamiento del proceso a partir de un conjunto finito de su realizaciones y que, en consecuencia, no posee un número determinado de coeficientes.

La aplicación de la noción de complejidad estocástica, que considera la estimación simultánea de los coeficientes y del número de coeficientes del modelo, y los postulados de los teoremas de Risanen sobre la convergencia de las medidas de complejidad predictiva y no-predictiva, permitieron demostrar, además, que el criterio de Schwarz puede ser considerado como aquél que contiene el término de penalización óptimo en la identificación del orden p de un modelo AR .

REFERENCIAS BIBLIOGRAFICAS

Abraham, B.; Ledolter, J. (1984): A note on inverse autocorrelations. *Biometrika*, vol. 71, pp. 609-612.

Abramson, N. (1968): *Information theory and coding*. McGraw-Hill.

Akaike, H. (1969): Fitting autoregressive models for prediction. *Annals of the Institute of Statistical Mathematics*, vol. 21, pp. 225-242.

Akaike, H. (1970): Statistical predictor identification. *Annals of the Institute of Statistical Mathematics,* vol. 22, pp. 203-217.

Akaike, H. (1973): *Information theory and an extension of the maximum likelihood principle*. En Petrov, B.N.; Csáki, F. (eds.).

Akaike, H. (1977): *On entropy maximization principle*. En Krishnaiah, P.R. (ed.).

Akaike, H. (1978): A Bayesian analysis of the minimum AIC procedure. *Annals of the Institue of Statistical Mathematics*, vol. 30, pp. 9-14.

Akaike, H. (1978): On the likelihood of a time series model. *The Statistician*, vol. 27, pp.237-242.

Akaike, H. (1979): A Bayesian extension of the minimum AIC procedure. *Biometrika,* vol. 66, 1979.

Allen, D.M. (1971): Mean square error of prediction as a criterion for selecting variables. *Technometrics*, vol. 13, pp. 237-242.

Anderson, O.D. (1980): *Time series*. North-Holland.

Anderson, T.W. (1963): Determination of the order of dependence in normally distributed time series. En Rosemblatt, M. (ed.).

Anderson, T.W. (1971): *The statistical analysis of time series*. Wiley.

Bartlett, M.S. (1946): On the theoretical specification and sampling properties of autocorrelated time series. *JRSS, Serie B, vol. 8*, pp. 27-41.

Bartlett, M.S.; Dinanda, P.H. (1950): Extensions of Quenouille's test for autoregressive schemes. *JRSS, Serie B, vol. 12,* pp. 108-115.

Beguin, J.M.; Gourieroux, C.; Montfort, A. (1980): *Identification of a mixed autoregressive-moving average process: The corner method.* En Anderson, O.D. (ed.) .

Bhansali, R.J. (1980): The autoregressive and the window estimate of the inverse correlation function. *Biometrika, vol. 67*, pp. 551-566.

Bhansali, R.J. (1983): The inverse partial autocorrelaction function of a time series and its application. *Journal of Multivariate Analysis, vol. 13,* pp. 310-327.

Box, G.E.P.; Jenkins, G.M. (1976): *Time series analysis: Forecasting and control*. Holden-Day.

Chaitin, G. (1975): A theory of program size formally identical to information theory. *Journal ACM, Vol. 22,* pp. 329-340.

Chaitin, G. (1987): *Algorithmic information theory*. Cambridge University Press.

Chaitin, G. (1990): *Information, randomness, and incompletness*. World Scientific, 2da. Edición..

Chatfield, C. (1979): Inverse autocorrelations. *JRSS, Seire A, vol. 142*, pp. 363-377.

Chow, G.C. (1981): A comparison of the information and posterior probability criteria for model selection. *Journal of Econometrics, vol. 16,* pp. 21-33.

Cleveland, W.S. (1972): The inverse autocorrelations of a time series and their applications. *Technometrics, vol. 14,* pp. 277-293.

Davies, N.; Petrucelli, J.D. (1984): On the use of the general partial autocorrelation function for order determination in ARMA(p,q) processes. *JASA, vol. 79*, pp. 374-377.

Davisson, L.D. (1965): The prediction error of stationary Gaussian time series of unknown covariance. *IEEE Transactions, Information Theory* IT-11, pp. 527-532.

Dawid, A.P. (1984): Present position and potential developments: Some personal views, statistical theory, the prequential approach. *JRSS, Serie A, vol. 147*, pp. 278-292.

Gani, J.; Piestley, M.B. (1986): Essays in time series and allied processes. *Applied Probability Trust*.

Geweke, J.F.; Meese, R.A. (1981): Estimating regression models of finite but unknown order. *International Economic Review, vol. 22*, pp. 55-70.

Gray, H.L.; Kelly, G.D.; McIntire, D.D. (1978): "A new approach to ARMA modelling". Communications in Statistics, Serie B, vol. 7, pp. 1-77.

Gooijer, J.G.; Heuts, RM.J. (1981): The corner method: An investigation of an order determination procedure for general ARMA processes. *Journal of Operations Reaearch Society, vol. 32,* pp. 1039-1042.

Hannan, E.J.; Quinn, B.G. (1979): The determination of the order of an autoregression. *JRSS, Serie B, vol. 41*, pp. 190-195.

Heyde, C.C. (1974): An iterated logarithm result for autocorrelations of stationary linear process. *Annals of Probability*, *vol. 2,* pp. 328-332.

Heyde, C.C.; Scott, D.J. (1973): Invariance principles for the laws of the iterated logarithm for martingales and processes with stationary increments. *Annals of Probability, vol. 1,* pp. 428-436.

Hipel, K.W.; McLeod, A.I.; Lennox, W.C. (1977): Advances in Box-Jenkins modelling, I: Model construction. *Water Resources Research, vol. 13*, pp. 567-575.

Hosking, J.R.M. (1980): Lagrange-multiplier tests on time series models. *JRSS, Serie B, vol. 42*, pp. 170-181.

Jacobs, O.; Davis, M.; Dempster, M.; Harris, C.; Parks, P. (eds.) (1980): *Analysis and optimization of stochastic systems*. Academic Press.

Kolmogorov, A.N. (1956): *Foundation of the theory of probability*. Chelsea.

Krishnaiah, P.R. (ed.) (1977): *Multivariate analysis*. IV. North-Holland.

Krishnaiah, P.R. (ed.) (1977): *Applications in statistics*. North-Holland.

Kromer, R.E. (1969): *Asymptotic properties of the autoregressive spectral estimator*. Ph D. Thesis, Stanford University.

Kullback, S. (1959): *Information theory and statistics*. Wiley.

Landro, A.H.; González, M.L. (2009): *Elementos de econometría de los fenómenos dinámicos*. Ediciones Cooperativas.

Mallows, C.L. (1973): Some comments on Cp. *Technometrics, vol. 15*, pp. 661-675.

Mallows, C.L. (1995): More comments on Cp" *Technometrics, vol. 37,* pp. 362-372.

McLeod, A.I.; Hipel, K.W.; Lennox, W.C. (1977): Advances in Box-Jenkins modelling, 2: Applications. *Water Resources Research, vol. 13,* pp. 577-586.

Newbold, P.; Bos, T. (1983): On conditioned partial correlations. *Journal of Time Series Analysis, vol. 4,* pp. 53-55.

Owen, D.B. (ed.) (1975): *The search for oil.* Dekker.

Parzen, E. (1974): Some recent advances in time series modelling. *IEEE Transactions Automatic Control, vol. 19,* pp. 723-730.

Parzen E. (1975): Some solutions to the time series modelling and prediction problem. En Owen, D.B. (ed.).

Parzen, E. (1977): *Multiple time series: Determining the order of approximating autoregressive schemes*. En Krishnaiah, P.R. (ed.).

Petrov, B.N.; Csáki, F. (eds.) (1973): *Second international symposium on information theory*. Akadémai Kiadó.

Petrucelli, J.D.; Davies, N. (1984): Some restrictions on the use of corner method hypothesis tests. *Communications in Statistics, vol. 13,* pp. 543-551.

Rao, C.R.; Wu, Y. (1989): A strongly consistent procedure for model selection in a regression problem. *Biometrika, vol. 76*, pp 369-374.

Rao, C.R.; Wu, Y. (2001): On model selection. *IMS Lecture Notes. Monograph Series, vol. 38, pp.* 1-57.

Risanen, J. (1978): Modeling by shortest data description. *Automatica, vol. 14*, pp. 465-471.

Risanen, J. (1980): Consistent order-estimates of atoregressive processes by shortest description of data. En Jacobs, O, Davis, M.; Dempster, M.; Harris, C.; Parks, P. (eds.).

Risanen, J. (1983): A universal prior for integers and estimation by minimum description lenght. *Annals of Statistics, vol. 11*, pp. 416-431.

Risanen, J. (1984): Universal coding information, prediction and estimation. *IEEE Transactions of Information Theory, IT-30*, pp. 629-636.

Risanen, J. (1986): A predictive least squares principle. *IMA Journal of Mathematics, Control and Information, vol. 3,* pp. 211-222.

Risanen, J. (1986): Stochastic complexity and modeling. *Annals of Statistics, vol. 14,* pp. 1080-1100.

Rosen R. (1985*): Anticipatory Systems*. Pergamon Press.

Rosenblatt, M. (ed.) (1971): *Proceedings of the Symposium of time series analysis*. Wiley.

Schwarz, G. (1978): Estimating the dimension of a model. *Annals of Statistics, vol. 6,* pp. 461-474.

Shaman, P. (1975): An approximate inverse for the covariance matrix of moving average and autoregressive processes. *Annals of Statistics, vol. 3,* pp. 532-538.

Shaman, P. (1976): Approximations for stationary covariances matrices and their inverses with applications to ARMA models. *Annals of Statistics, vol. 4,* pp. 292-301.

Solomonoff, R.J. (1964): A formal theory of inductive inference Part I. *Information and Control, vol. 7,* pp. 1-22.

Solomonoff, R.J. (1964): A formal theory of inductive inference- Part II. *Information and Control, vol. 7,* pp. 224-254.

Takemura, A. (1984): A generalization of autocorrelation and partial autocorrelation functions useful for identification of ARMA(p,q) processes. *Technical Report nº 84-16*, Department of Statistics, Purdue University, 1984.

Thom, R. (1975): *Structural stability and morphogenesis*. Benjamin, 1975.

Tiao, G.C.; Tsay, R.S. (1983): Consistency properties of least squares estimates of autoregressive parameters in ARMA models. *Annals of Statistics, vol. 11*, pp. 856-871.

Tiao, G.C.; Tsay, R.S. (1983): Multiple time series modelling and extended sample cross-correlation. *Journal of Business Economy and Statistics, vol. 1,* pp. 43-56.

Tsay, R.S.; Tiao, G.C. (1984): Consistent estimates of autoregressive parameters and extended sample autocorrelation function for stationary and non-stationary ARMA models. *JASA, vol. 79*, pp. 84-96.

Tukey, J.W.; Akaike, H.; Robinson, E.A.; Granger, C.W.J. (1978): Comments on Gray, Kelley and McIntire. *Communications in Statistics B7*, vol. 7, pp. 1-77.

Whittle, P. (1952): Tests of fit in time series. *Biometrika, vol. 29*, pp. 309-318.

Whittle, P. (1954): Some recent contributions to the theory of stationary time series. En Wold, H. (ed.).

Wold, H. (ed.) (1954): *A study in the analysis of stationary time series.* Almqvist & Wicksell.

Woodward, W.A.; Gray, H.L. (1978): New ARMA models for Wölfer suspot data. *Communications in Statistics, Serie B, vol. 7,* pp. 97-116.

Woodward, W.A.; Gray, H.L. (1981): On the relationship between the S array and the Box-Jenkins method of ARMA model identification. *JASA, vol. 76,* pp. 579-587.

Yamamoto, Y. (1976): Asymptotic mean square prediction error for an auteregressive model with estimated coefficients. *Applied Statistics, vol. 25,* pp. 123-127.