Revista Colombiana de Estadística

No.33 y No.34

CÁLCULO DEL NÚMERO MÍNIMO DE DATOS NECESARIOS PARA ESTIMAR EL VECTOR DE OBSERVACIONES FALTANTES EN UNA SERIE TEMPORAL GENERADA POR UN MODELO AR(p)

HENRY GALLARDO PEREZ

UNIVERSIDAD PEDAGÓGICA Y TECNOLÓGICA DE COLOMBIA

FABIO NIETO

UNIVERSIDAD NACIONAL DE COLOMBIA

RESUMEN. La determinación del número mínimo de datos que se deben utilizar para estimar el valor de observaciones faltantes en una serie temporal univariada, es importante porque permite optimizar el tiempo de computación en el sentido en que sí se utilizara un número mayor de datos, el proceso de estimación resultaría redundante. En este trabajo se determina cuál es número mínimo y cuáles son los datos que se deben utilizar para estimar el vector de observaciones faltantes, cuando el proceso estocástico obedece un modelo autoregresivo de orden p, AR(p). Se utiliza para ello el método de estimación de Peña-Maravall (1991) y el proceso recurrente de Nieto-Martínez (1994). Se presentan adicionalmente algunos ejemplos teóricos en los cuales se aplican los resultados obtenidos.

PALABRAS CLAVES: Datos faltantes, Función de autocorrelación dual, Modelo AR(p).

1.INTRODUCCION

El problema de estimación de datos faltantes en una serie temporal univariada que obedece un modelo ARIMA lineal e invertible, puede ser resuelto utilizando diferentes enfoques. Uno, el de Kohn y Ansley (1986) y Gómez y Maravall (1994), quienes usan el algoritmo de suavizador de punto fijo (basado en el filtro de Kalman). Otro, el

de Peña y Maravall (1991) y Maravall y Peña (1992), quienes usan básicamente los coeficientes de autocorrelación dual para obtener los estimadores de los datos faltantes y sus errores cuadráticos medios. Otra posibilidad está dada por el método recurrente de Nieto y Martínez (1994), el cual parte de una estimación inicial del vector de datos faltantes y luego actualiza la estimación usando cada vez un nuevo dato hasta agotar los datos observados en la serie.

Los enfoques citados anteriormente, salvo posiblemente el de la Función de Autocorrelación Dual, utilizan toda la información contenida en la serie para estimar el vector de datos faltantes. Sin embargo, si el modelo de la serie temporal univariada es AR(p), se puede encontrar un número mínimo de datos observados para realizar la estimación de los datos faltantes. Con este resultado se pueden optimizar los algorítmos de computación para la estimación de los datos faltantes, en especial para los procedimientos recurrentes de suavizador de punto fijo y para el método de Nieto y Martínez.

El número mínimo antes mencionado se encuentra en este trabajo, usando el método de la FACD de Peña y Maravall (1991) y el de Nieto y Martínez (1994).

El trabajo se organiza de la siguiente manera: en la sección 2 se incluyen fórmulas básicas de cálculo de los métodos de Peña-Maravall y Nieto-Martínez; en la sección 3 se deduce el número mínimo de datos, utilizando el método de Peña-Maravall; en la sección 4 se hace una deducción equivalente utilizando el método de Nieto-Martínez; en la sección 5 se presentan algunos ejemplos teóricos utilizando ambas metodologías, finalmente, en la sección 6, se dan algunas conclusiones.

2.Formulas de calculo de los métodos de Peña -Maravall y de Nieto-Martínez

Supóngase que el proceso $\{Z_t\}$ obedece el modelo AR(p)

$$Z_t = \phi_1 Z_{t-1} + \phi_2 Z_{t-2} + ... + \phi_p Z_{t-p} + a_t,$$

o bien

$$\phi(B)\, Z_t = a_t\,, \tag{1}$$

donde $\phi(B) = 1 - \phi_1 B - \phi_2 B^2 - ... - \phi_p B^p$ es un polinomio de grado p en el operador de retardo B y $\{a_t\}$ es un proceso de ruido blanco Gaussiano con media cero y varianza σ_a^2 . El operador autoregresivo $\phi(B)$ contiene tanto las raíces estacionarias como también las no estacionarias (incluyendo posiblemente estacionales).

Debido a que el modelo (1) es autoregresivo puro se puede escribir en la forma

$$\pi(B)\, Z_t = a_t\,, \tag{2}$$

con $\pi(B) \equiv \phi(B)$, esto es, $\pi_i = \phi_i$ para $i = 1, 2, ..., p$.

El modelo dual de (1) está dado por:

$$Z_t^d = \phi(B)\, a_t$$

o, equivalentemente

$$Z_t^d = \pi(B)\, a_t \tag{3}$$

Su varianza y su función generadora de autocorrelación (dual) están dadas por:

$$V_d = \sigma_a^2 \left(1 + \pi_1^2 + \pi_2^2 + ... + \pi_p^2\right) \tag{4}$$

$$\rho_{(B)}^d = \frac{\sigma_a^2 \pi(B)\pi(F)}{V_d} \tag{5}$$

respectivamente, donde $F = B^{-1}$.

Supóngase que la serie finita $\{Z_t\}$ tiene k datos faltantes en los tiempos $T_1, T_2, ..., T_k$ donde $T_i < T_j$ para $i < j$ y que la longitud del período completo de observación (incluyendo los tiempos de los datos faltantes) es N. En las siguientes dos subsecciones se presentan los estimadores para los datos faltantes de cada uno de los métodos a considerar en el presente trabajo.

2.1.El Método de Peña-Maravall

Sean $\mathbf{Z}_m = (\mathbf{Z}_{T_1}, ..., \mathbf{Z}_{T_k})'$,el vector de datos faltantes, $Z_m = (Z_{T_1}, ..., Z_{T_k})'$ un vector de datos inventados (outliers aditivos) y R_d la matriz de autocorrelaciones duales de Z_m, definida de la siguiente manera

$$R_d = \begin{bmatrix} 1 & \rho_{T_2-T_1}^d & \rho_{T_3-T_1}^d & \cdots & \rho_{T_k-T_1}^d \\ & 1 & \rho_{T_3-T_2}^d & \cdots & \rho_{T_k-T_2}^d \\ & & 1 & \cdots & \rho_{T_k-T_3}^d \\ & & \cdots & \cdots & \cdots \\ & & & 1 & \rho_{T_k-T_{k-1}}^d \\ & & & & 1 \end{bmatrix},$$

Peña y Maravall (1991) demuestran que

$$\hat{\mathbf{Z}}_m = Z_m - R_d^{-1}\rho_{(B)}^d Z_m \tag{6}$$

es el estimador de error cuadrático medio mínimo (ECMM) de $\mathbf{Z}_m$ con matriz de error cuadrático medio dado por

$$ECM\left(\hat{\mathbf{Z}}_m\right) = \frac{\sigma_a^2}{V_d} R_d^{-1} \tag{7}$$

Es importante notar que el estimador de Z_m no depende de los datos inventados (Peña-Maravall, 1991).

2.2.El Método de Nieto-Martínez

Sean $\mathbf{Z}_m$ como en la sección 2.1, $Y = (Z_{n_1}, Z_{n_2}, ..., Z_{m_k+1}, ..., Z_N)'$ el vector de datos observados después del período n, donde n representa el número de períodos antes del primer dato faltante, y $\mathbf{Z} = (Z_1, Z_2, ..., Z_n)'$ el vector de datos observados antes de primer dato faltante.

Supóngase que dado n_{j-1}, $j > 1$, se conoce el estimador de ECMM de $\mathbf{Z}_m$, digamos $\tilde{\mathbf{Z}}_{m,j-1}$, y su matriz de ECM, digamos $P_{m,j-1}$. Usando la observación Z_{n_j},se obtiene que el nuevo estimador del vector de datos faltantes está dado por

$$\tilde{\mathbf{Z}}_{m,j} = \tilde{\mathbf{Z}}_{m,j-1} + A_{m,j}\left(Z_{n_j} - \hat{Z}_{n_j,j-1}\right) \tag{8}$$

y la nueva matriz de ECM por

$$P_{m,j} = P_{m,j-1} - A_{m,j} V_{21}^{(i)} \quad , \tag{9}$$

donde $\hat{Z}_{n_j,j-1}$ es la proyección ortogonal de Z_{n_j} sobre la información dada hasta el paso $j-1$,

$A_{m,j} = \frac{1}{v_{22}^{(j)}} V_{12}^{(j)}$,

$v_{22}^{(j)}$ es el EMC de $\hat{Z}_{n_j,j-1}$,

$V_{12}^{(j)}$ es la covarianza entre $Z_{n_j} - \hat{Z}_{n_j,j-1}$ y $\mathbf{Z}_m - \mathbf{Z}_{m,j-1}$,

$V_{21}^{(j)} = V_{12}^{(j)\prime}$.

Mas aún, Nieto y Martínez (1994) demuestran que

$$v_{22}^{(j)} = \sigma_a^2 + \sum_{\iota=1}^{k_j} \pi_{\delta(\iota)}^2 \left(P_{m,j-1}\right)_{k_j+1-\iota,k_j+1-\iota} + 2 \sum_{\substack{i,\iota=1 \\ i<j}}^{k_j} \pi_{\delta(i)} \pi_{\delta(\iota)} \left(P_{m,k-1}\right)_{k_j+1-i,k_j+1-\iota} \tag{10}$$

y la entrada s-ésima de $V_{12}^{(j)}$ está dada por

$$V_{12(s)}^{(j)} = \begin{cases} -\sum_{\iota=1}^{k_j} \pi_{\delta(\iota)} \left(P_{m,j-1}\right)_{s,k_j+1-\iota} & , \ s = 1,2,...,k_j \\ \psi_{T_s - T_j} \sigma_a^2 - \sum_{\iota=1}^{k_j} \pi_{\delta(\iota)} \left(P_{m,j-1}\right)_{s,k_j+1-\iota} & , \ s = k_j+1,...,k \end{cases} \tag{11}$$

donde k_j denota el número de datos faltantes antes de n_j , $\delta(\iota) = n_j - T_{k_j+1-\iota}$ para

$\iota = 1, ..., k_j$, y los coeficientes ψ_y se encuentran a partir de la relación

$$(1 - \phi_1 B - ... - \phi_p B^p)\left(1 + \psi_1 B + \psi_2 B^2 + ...\right) = 1,$$

igualando los coeficientes de las potencias de B (ver Bell, 1984).

3.Determinación del mínimo de datos requeridos para estimar el vector de datos faltantes en un proceso generado por un modelo AR(p) utilizando el método de Peña-Maravall.

Teorema 3.1

Sea $\{Z_t\}$ una serie temporal cuyo modelo AR(p) está descrito como en (1), con k datos faltantes en los períodos $T_1, T_2, ..., T_K$. Entonces utilizando el método de Peña-Maravall se requieren

$$2kp - \sum_{j=1}^{k-1} \xi_j$$

datos para estimar el vector de observaciones faltantes, donde ξ_j está dado por:

$$\xi_j = \begin{cases} \#\text{ de períodos de tiempo en el intervalo } \left[T_{j+1} - p,\ T_j + p\right] & , si\ T_{j+1} - p \leq T_j + p \\ 0 & , si\ T_{j+1} - p > T_j + p \end{cases}$$

para $j = 1, 2, ..., k-1$.

Demostración: Se usará inducción sobre k , el número de datos faltantes, para probar el teorema.

En primer lugar sea $k = 1$. La observación faltante se reemplaza con un número

arbitrario $\mathbf{Z}_{T_1}$ y se construye la serie observada

$$\mathbf{Z}_t = \begin{cases} Z_t + W_1 & , \; si \; t = T_1 \\ Z_t & , \; si \; t \neq T_1 \end{cases} \tag{12}$$

donde W_1 es un parámetro desconocido.

La serie "observada" se puede escribir así:

$$\mathbf{Z}_t = Z_t + W_1 d_t^1 \tag{13}$$

o también de la siguiente manera:

$$Z_t = \mathbf{Z}_t - W_1 d_t^1 \tag{14}$$

donde $d_t^1 = \begin{cases} 1 & si \; t = T_1 \\ 0 & si \; t \neq T_1 \end{cases}$

Este es un modelo de intervención y se puede escribir como

$$\pi(B)\,\mathbf{Z}_t = W_1 \pi(B)\, d_t^1 + a_t, \tag{15}$$

o también, como

$$Y_t = W_1 X_{1t} + a_t, \tag{16}$$

con $Y_t = \pi(B)\,\mathbf{Z}_t$ y $X_{1t} = \pi(B)\, d_t^1$.

Peña y Maravall (1991), demostraron que

$$\hat{W}_1 = \mathbf{Z}_{T_1} + \sum_{i=1}^{p} \rho_i^d \left(\mathbf{Z}_{T_1-i} + \mathbf{Z}_{T_1+i}\right) , \tag{17}$$

donde los ρ_i^d, $i = 1, ..., p$, son los coeficientes de la FACD.

De (17) se concluye que se requieren $2p$ datos para estimar el dato Z_{T_1}, ya que en este caso $\xi_j = 0$. Esto es, para $k = 1$

$$2kp - \sum_{j=1}^{k-1} \xi_j = 2\,(1)\,p - 0 = 2p.$$

Con el fin de ilustrar los detalles para la demostración general, se considera ahora el caso en el cual hay dos datos faltantes en los períodos T_1 y T_2 con $T_1 < T_2$. de nuevo usando el resultado de Peña y Maravall se encuentra que

$$\widehat{\vec{W}} = \left(\vec{X}'\vec{X}\right)^{-1} \vec{X}'\vec{Y} = \frac{\sigma_a^2}{V_d} R_d^{-1} \frac{V_d}{\sigma_a^2} \rho_{(B)}^d \mathbf{Z}^2 = R_d^{-1} \rho_{(B)}^d \mathbf{Z}^2.$$

Puesto que la función de autorrelación dual (FACD) del proceso es conocida (ya que el modelo es conocido) entonces R_d^{-1} está plenamente determinado. Por tanto nos ocuparemos solamente de analizar el vector $\rho_{(B)}^d \mathbf{Z}^2$.

$$\begin{aligned} \rho_{(B)}^d \mathbf{Z}^2 = \rho_{(B)}^d \begin{pmatrix} \mathbf{Z}_{T_1} \\ \mathbf{Z}_{T_2} \end{pmatrix} &= \left(1 + \textstyle\sum_{1=1}^{p} \rho_i^d \left(B^i + F^i\right)\right) \begin{pmatrix} \mathbf{Z}_{T_1} \\ \mathbf{Z}_{T_2} \end{pmatrix} \\ &= \begin{pmatrix} \mathbf{Z}_{T_1} \\ \mathbf{Z}_{T_2} \end{pmatrix} + \begin{pmatrix} \sum_{i=1}^{p} \rho_i^d \left(\mathbf{Z}_{T_1-i} + \mathbf{Z}_{T_1+i}\right) \\ \sum_{i=1}^{p} \rho_i^d \left(\mathbf{Z}_{T_2-i} + \mathbf{Z}_{T_2+i}\right) \end{pmatrix} \end{aligned} \tag{18}$$

$$\text{Sea } \xi_1 = \begin{cases} \text{Número de períodos de tiempo en el intervalo } [T_2 - p,\ T_1 + p] & , \text{si } T_2 - p \leq T_1 + p \\ 0 & , \text{si } T_2 - p > T_1 + p \end{cases}$$

Para estimar la primera componente del vector (18) se requiere información desde $T_1 - p$ hasta $T_1 + p$ y para estimar la segunda componente se requiere información desde $T_2 - p$ hasta $T_2 + p$.

Ahora bien, si $\xi_1 = 0$ se necesitarán $4p$ datos para estimar (18), y por consiguiente estimar a $\vec{W}$ (véase la tabla 1 para una ilustración tanto de este caso como de los dos subsiguientes).

Si $0 < \xi_1 \leq p$ hay ξ_1 datos (traslapados) en el intervalo $[T_2 - p, T_1 + p]$, los cuales se utilizan simultáneamente para estimar las dos componentes de (18), por lo tanto se concluye que se requiere $4p - \xi_1$ datos para estimar las componentes de (18).

Finalmente , si $\xi_1 > p$ entonces en el intervalo $[T_2 - p, T_1 + p]$, hay $2p - (T_2 - T_1) - 1$ datos observados y puesto que en los períodos de tiempo T_1 y T_2 no hay datos observados entonces se tiene que $\xi_1 = 2p - (T_2 - T_1) + 1$ y en este caso $4p - \xi_1 = 4p - 2p + (T_2 - T_1) - 1 = 2p + [(T_2 - T_1) - 1]$, que no es otra cosa que p datos observados anteriores a T_1, más p datos observados a continuación de T_2, más los datos observados entre T_1 y T_2.

En los tres casos se concluye que para estimar (18) se require utilizar

$4p - \xi_1 = 2\,(2)\,p - \xi_1$ datos.

A manera de ilustración consideramos el caso particular en el cual el modelo para $\{Z_t\}$ es un AR(3). En este caso (18) se puede escribir así:

$$\rho^4_{(B)}\mathbf{Z}^2 = \begin{pmatrix} \mathbf{Z}_{T_1} \\ \mathbf{Z}_{T_2} \end{pmatrix} + \begin{pmatrix} \rho_1^4\left(\mathbf{Z}_{T_1-1} + \mathbf{Z}_{T_1+1}\right) + \rho_2^4\left(\mathbf{Z}_{T_1-2} + \mathbf{Z}_{T_1+2}\right) + \rho_3^4\left(\mathbf{Z}_{T_1-3} + \mathbf{Z}_{T_1+3}\right) \\ \rho_1^4\left(\mathbf{Z}_{T_2-1} + \mathbf{Z}_{T_2+1}\right) + \rho_2^4\left(\mathbf{Z}_{T_2-2} + \mathbf{Z}_{T_2+2}\right) + \rho_3^4\left(\mathbf{Z}_{T_2-3} + \mathbf{Z}_{T_2+3}\right) \end{pmatrix}$$

$$\xi_1 = \begin{cases} \text{Número de períodos de tiempo en el intervalo } [T_2 - 3,\ T_1 + 3] & , \text{ si } T_2 - 3 \leq T_1 + 3 \\ 0 & , \text{ si } T_2 - 3 > T_1 + 3 \end{cases}$$

En la tabla 1 se representa en la recta T las diferentes ubicaciones que podrían tenerse para T_1 y T_2 y con un $(*)$ se marcan los períodos en los cuales se deben tener datos observados para estimar a $\bar{W}$

Tabla 1: Diferentes ubicaciones para T_1 y T_2

																	ξ_1	$4p - \xi$
.	*	*	*	T_1	T_2	*	*	*	.	.	.	.	.	.	.	.	6	6
.	*	*	*	T_1	*	T_2	*	*	*	.	.	.	.	.	.	.	5	7
.	*	*	*	T_1	*	*	T_2	*	*	*	.	.	.	.	.	.	4	8
.	*	*	*	T_1	*	*	*	T_2	*	*	*	.	.	.	.	.	3	9
.	*	*	*	T_1	*	*	*	*	T_2	*	*	*	.	.	.	.	2	10
.	*	*	*	T_1	*	*	*	*	*	T_2	*	*	*	.	.	.	1	11
.	*	*	*	T_1	*	*	*	*	*	*	T_2	*	*	.	.	.	0	12
.	*	*	*	T_1	*	*	*	.	*	*	*	T_2	*	*	*	.	0	12
.	*	*	*	T_1	*	*	*	.	.	*	*	*	T_2	*	*	*	0	12
																	$\vdots$	$\vdots$

Supongamos ahora que hay $(k-1)$ datos faltantes en los períodos $T_1, T_2, ..., T_{k-1}$ y que se requiere utilizar $2(k-1)p - \sum_{j=1}^{k-2} \xi_j$ datos para estimar el vector de datos faltantes $\mathbf{Z}_{m_1} = (\mathbf{Z}_{T_1}, ..., \mathbf{Z}_{T_{k-1}})'$, donde

$$\xi_j = \begin{cases} \text{Número de períodos de tiempo en el intervalo } \left[T_{j+1} - p,\ T_j + p\right] & , \text{ si } T_{j+1} - p \leq T_j + p \\ 0 & , \text{ si } T_{j+1} - p > T_j + p \end{cases}$$

$j = 1, 2, ..., k-2.$

Como en el período T_k (con $T_k > T_{k-1}$) se presenta un nuevo dato faltante, se define

$$\xi_{k-1} = \begin{cases} \text{Número de períodos de tiempo en el intervalo } \left[T_k - p,\ T_{k-1} + p\right] & , si\ T_k - p \leq T_{k-1} + p \\ 0 & , si\ T_k - p > T_{k-1} + p \end{cases}$$

Entonces, según se probó en el caso $k = 1$, se requieren $2p$ datos observados para estimar a $\mathbf{Z}_{T_k}$; sin embargo, existen ξ_{k-1} de estos que pueden estar traslapados con los datos necesarios para estimar $\mathbf{Z}_{T_{k-1}}$.

Por lo tanto para estimar el vector completo de datos faltantes

$\mathbf{Z}_m = (\mathbf{Z}_{T_1}, \mathbf{Z}_{T_2}, ..., \mathbf{Z}_{T_{k-1}}, \mathbf{Z}_{T_k})$ es necesario utilizar, como mínimo

$$2(k-1)p - \sum_{j=1}^{k-2} \xi_j + 2p - \xi_{k-1} = 2kp - \sum_{j=1}^{k-1} \xi_j$$

datos.

4.Determinación del mínimo de datos requeridos para estimar el vector de datos faltantes en un proceso generado por un modelo AR(p) utilizando el método de Nieto-Martínez.

Teorema 4.1

Sea $\{Z_t\}$ una serie temporal cuyo modelo AR(p) está descrito como en (1) con k datos faltantes en los períodos $T_1, ..T_k$. Entonces, utilizando el método recurrente de Nieto-Martínez se requiere utilizar mínimo $2p + (T_k - T_1 - k + 1)$ datos para obtener la estimación del vector $\mathbf{Z}_m$.

Demostración: Estudiaremos dos casos particulares antes de abordar el resultado general.

1. Un dato faltante en el período de tiempo T_1 $(T_1 = n + 1)$:

El "vector " de datos faltantes será $\mathbf{Z}_m = (\mathbf{Z}_{T_1})'$ y en este caso se tiene que

$$\tilde{Z}_{m,1} = \tilde{Z}_{T_1} + A_{m,1}\left(Z_{n_1} - \tilde{Z}_{n_1}\right) = \tilde{Z}_{T_1} + A_{m,1}\left(Z_{n+2} - \tilde{Z}_{n+2}\right) \tag{19}$$

con

$$A_{m,1} = \frac{\phi_1}{1+\phi_1^2} \tag{20}$$

$$\tilde{Z}_{T_1} = E\left(Z_{T_1} \mid \mathbf{Z}\right) = \phi_1 Z_n + \phi_2 Z_{n-1} + \ldots + \phi_p Z_{n-p+1} \tag{21}$$

y con ECMM dado por:

$$P_{m,1} = \sigma_a^2(1-\phi_1) \tag{22}$$

Incluimos ahora la observación $Z_{n_2} = Z_{n+3}$ y con ella se realiza una nueva estimación del vector de datos faltantes. Para ello necesitamos definir

$$\delta(1) = n_2 - T_1 = (n+3) - (n+1) = 2 \tag{23}$$

La matriz $P_{0,2}$ estará dada por

$$P_{0,2} = \begin{bmatrix} P_{m,1} & V_{12}^{(2)} \\ V_{21}^{(2)} & V_{22}^{(2)} \end{bmatrix}$$

donde $P_{m,1}$ está dado por (22) y

$$V_{12}^{(2)} = \left(-\sum_{i=1}^{1} \pi_{\delta(i)} \left(P_{m,1}\right)_{1,2-i}\right) = \left(-\pi_2(1-\phi_1)\sigma_a^2\right) = -\phi_2(1-\phi_1)\sigma_a^2 \tag{24}$$

$$v_{22}^{(2)} = \sigma_a^2 + \pi_{\delta(1)}^2 (P_{m,1})_{1,1} = \sigma_a^2 \left[1 + \phi_2^2 (1 - \phi_1)\right] \tag{25}$$

por lo tanto

$$A_{m,2} = \frac{1 - \phi_1}{1 + \phi_2^2(1 - \phi_1)} \tag{26}$$

y el nuevo estimador de $\mathbf{Z}_m$ estará dado por la expresión

$$\tilde{Z}_{m,2} = \tilde{Z}_{m,1} + A_{m,2} \left(Z_{n_2} - \tilde{Z}_{n_2,1}\right) = \tilde{Z}_{m,1} + A_{m,2} \left(Z_{n+3} - \tilde{Z}_{n+3,1}\right). \tag{27}$$

De esta manera se continúa hasta llegar al dato en el período $n_j = n + p + 1$, para el cual $\delta(1) = p$, y $V_{12}^{(i)} = -\phi_p (P_{m,j-1})_{1,1}$, obteniéndose el siguiente estimador de $\mathbf{Z}_m$:

$$\tilde{Z}_{m,p} = \tilde{Z}_{m,p-1} + A_{m,p} \left(Z_{n_p} - \tilde{Z}_{n_p,1}\right) = \tilde{Z}_{m,p-1} + A_{m,p} \left(Z_{n+p+1} - \tilde{Z}_{n+p+1,1}\right) \tag{28}$$

Nótese que si $n_i > n+p+1$ entonces $\delta(1) > p$ y en este caso $\pi_{\delta(1)} = 0$, lo cual implica que $V_{12}^{(j)} = 0$ y por tanto $A_{m,j} = 0$.

De esta ultima observación se concluye que

$$\tilde{Z}_{m,p+1} = \tilde{Z}_{m,p+2} = \tilde{Z}_{m,p+3} = \ldots = \tilde{Z}_{m,p}$$

De las ecuaciones (21) y (28), y de los procedimientos intermedios se concluye que para estimar a $\mathbf{Z}_m$ se requiere utilizar $2p$ datos (los p datos anteriores a T_1 y los p datos siguientes a T_1).

(2)Supongamos ahora que la serie temporal $\{Z_t\}$ tiene dos faltantes en los períodos de tiempo T_1 y T_2 (con $T_1 < T_2$).

Según se observó en el caso anterior, todo depende de $V_{12}^{(j)}$ ya que si este vector es

cero la estimación de $\mathbf{Z}_m$ será la misma del paso anterior y obviamente con el mismo ECMM.

Para $j = 1, ..., T_2 - T_1 - 1$ se tiene que $k_j = 1$ y por tanto $\iota = 1$, luego

$$\delta(1) = n_j - T_{k_j+1-1} = n_j - T_1 = n + 1 + j - (n+1) = j, \text{ y}$$

$$\psi_{T_{k_j+1}-n_j} = \psi_{T_2-n_j}.$$

Bajo estas circunstancias el vector $V_{12}^{(j)}$ será de la forma:

$$V_{12}^{(j)'} = \begin{pmatrix} -\pi_j \left(P_{m,j-1}\right)_{1,1} \\ \psi_{T_2-n_j}\sigma_a^2 - \pi_j \left(P_{m,j-1}\right)_{2,1} \end{pmatrix} \tag{29}$$

Nótese que si $T_2 - n_j - 1 > p$, los ultimos $T_2 - n_j - 1 - p$ vectores $V_{12}^{(j)}$ que se obtienen en (29) tendrán la primera componente igual a cero, pero la segunda siempre será diferente de cero.

Para $j \geq T_2 - T_1$ se tiene que $k_j = 2$ y por lo tanto ahora ι toma dos valores: $\iota = 1, 2$; luego:

$$\delta(1) = n_j - T_{k_j+1-1} = n_j - T_2 \quad \text{y} \quad \delta(2) = n_j - T_{k_j+1-2} = n_j - T_1$$

y el vector $V_{12}^{(j)}$ será de la forma

$$V_{12}^{(j)} = \begin{pmatrix} -\pi_{n_j-T_2} \left(P_{m,j-1}\right)_{1,2} \\ -\pi_{n_j-T_1} \left(P_{m,j-1}\right)_{1,1} \end{pmatrix} \tag{30}$$

Nótese que cuando $n_j - T_2 > p$ el vector $V_{12}^{(j)}$ dado en (30) será igual al vector cero

(En la tabla 2 se ilustra este resultado).

Así se concluye de (29) y (30) que el vector $V_{12}^{(j)}$ es distinto de cero siempre que, en el proceso de estimación, se incluya una observación comprendida entre T_1 y T_2 y también lo es los primeros p-períodos después de T_2.

Puesto que en el primer paso se utilizan los p datos precedentes a T_1 y en los demás pasos se utilizan todos las datos comprendidos entre T_1 y T_2 y además los p datos siguientes a T_2, entonces en total se utilizan $2p + (T_2 - T_1 - 1)$ datos para estimar el vector $\mathbf{Z}_m$.

A manera de ilustración consideremos una serie $\{Z_t\}$ con modelo AR(3) y con datos faltantes en $T_1 = n+1$ y $T_2 = n+10$; lo cual corresponde al último caso citado en la tabla 1.

En la tabla 2 se registran los valores de j, k_j, ι, $\delta(\iota)$ y el correspondiente vector $V_{12}^{(j)}$:

Tabla 2:Diferentes posibilidades para $V_{12}^{(j)}$

j	k_j	ι	$\delta(\iota)$	$\mathbf{V_{12}^{(j)}}$
1	1	1	1	$\left(-\pi_1 (P_{m,0})_{1,1}\ , \psi_8\sigma_a^2 - \pi_1 (P_{m,0})_{2,1}\right)$
2	1	1	2	$\left(-\pi_2 (P_{m,1})_{1,1}\ , \psi_7\sigma_a^2 - \pi_2 (P_{m,1})_{2,1}\right)$
3	1	1	3	$\left(-\pi_3 (P_{m,2})_{1,1}\ , \psi_6\sigma_a^2 - \pi_3 (P_{m,1})_{2,1}\right)$
4	1	1	4	$\left(0\ , \psi_5\sigma_a^2\right)$
5	1	1	5	$\left(0\ , \psi_4\sigma_a^2\right)$
6	1	1	6	$\left(0\ , \psi_3\sigma_a^2\right)$
7	1	1	7	$\left(0\ , \psi_2\sigma_a^2\right)$
8	1	1	8	$\left(0\ , \psi_1\sigma_a^2\right)$
9	2	1;2	1;9	$\left(-\pi_1 (P_{m,8})_{1,2}\ , -\pi_1 (P_{m,8})_{2,2}\right)$
10	2	1;2	2;10	$\left(-\pi_2 (P_{m,9})_{1,2}\ , -\pi_2 (P_{m,9})_{2,2}\right)$
11	2	1;2	3;11	$\left(-\pi_3 (P_{m,10})_{1,2}\ , -\pi_3 (P_{m,10})_{2,2}\right)$
12	2	1;2	4;12	$(0,0)$
13	2	1;2	5;13	$(0,0)$
⋮	⋮	⋮	⋮	⋮

Si se comparan los dos ejemplos se notará que el método de Nieto-Martínez utiliza los ocho datos comprendidos entre T_1 y T_2 en tanto que el método de Peña-Maravall sólo utiliza seis de ellos para estimar el vector $\mathbf{Z}_m$.

En este punto podemos generalizar nuestro resultado para una serie $\{Z_t\}$ con el modelo AR(p) dado por (1), con k datos faltantes en los períodos $T_1, ..., T_k$.

El problema se resuelve mostrando que el vector $V_{12}^{(j)}$ es diferente de cero cuando se incluyan los datos observados desde $T_1 + 1$ hasta $T_k + p$ y que vale cero si el dato que se incluye está ubicado en un período posterior a $T_k + p$.

Si n_j es tal que $T_1 + 1 \leq n_j < T_k$ entonces $T_k - n_j \geq 1$ y en (11) la última componente del vector $V_{12}^{(j)}$ sería diferente de cero, lo cual garantiza que el vector $V_{12}^{(j)}$ no es necesariamente el vector nulo.

Si n_j es tal que $T_k < n_j \leq T_k + p$ entonces para cada $n_j = T_k + 1, ..., T_k + p$, se tiene que $\delta(1) = n_j - T_k$ tomará los valores $1, ..., p$, respectivamente. Luego para cada observación Z_{n_j} que se incluya en el proceso recurrente de estimación de $\mathbf{Z}_m$, el vector $V_{12}^{(j)}$ tendrá sus k−componentes no todas nulas según se puede deducir de (11).

Si j es tal que $n_j > T_k + p$ entonces $\delta(\iota) = n_j - T_{k+1-\iota} > p$ y en (11) todos los $\pi_{\delta(\iota)}$ serán iguales a cero, y por lo tanto $V_{12}^{(j)} = 0$.

En conclusión, con base en (21) y en la discusión anterior, para estimar el vector de datos faltantes $\mathbf{Z}_m$ se requiere utilizar los p datos anteriores a T_1, los p datos posteriores a T_k y los $T_k - T_1 - k + 1$ datos observados comprendidos entre T_1 y T_2. Esto es, utilizar como mínimo $2p + (T_k - T_1 - k + 1)$ datos.

Comparando los resultados se encuentra que el método de Nieto-Martínez requiere utilizar todos los datos observados entre T_1 y T_k, mientras que el método de Peña-

Maravall puede realizar la estimación con menos información, cuando se presentan traslapes entre T_1 y T_k.

5. Algunos ejemplos teóricos

Ejemplo 1

Supongamos que $\{Z_t\}$ sigue el modelo AR(1)

$$Z_t = \phi_1 Z_{t-1} + a_t \tag{31}$$

y que para el proceso tiene dos datos faltantes en $T_1 = n+1$ y $T_2 = n+3$.

Usando el método de Peña-Maravall se encuentra que se requiere utilizar tres datos para estimar a $\mathbf{Z_m}$; estos datos se encuentran ubicados en los períodos n, $n+2$, $n+4$.

Usando el método de Nieto-Martínez el número de datos necesarios para realizar la estimación es también de tres y son los mismos datos utilizados en el caso anterior.

Ejemplo 2

Si $\{Z_t\}$ sigue el modelo AR(1) dado en (31) con los datos faltantes en los períodos $T_1 = n+1$ y $T_2 = n+5$, al usar el método de Peña-Maravall para estimar el vector de datos faltantes, el número de datos necesarios para realizar dicha estimación es de cuatro, los cuales se encuentran en los períodos n, $n+2$, $n+4$, $n+6$.

Usando el método de Nieto-Martínez para estimar el vector de datos faltantes se requieren cinco datos ubicados en los períodos n, $n+2$, $n+3$, $n+4$, $n+6$.

Ejemplo 3

Sea $\{ Z_t \}$ generada por un modelo AR(7) con cinco datos faltantes en $T_1 = n+1, T_2 = n+4, T_3 = n+8, T_4 = n+9, T_5 = n+11$. utilizando el *teorema 3.1*

se tiene que $\xi_1 = 12$, $\xi_2 = 11$, $\xi_3 = 14$ y $\xi_4 = 13$, y por lo tanto se requieren

$$2kp - \sum_{j=1}^{4} \xi_j = 2(5)(7) - (12 + 11 + 14 + 13) = 20$$

observaciones para estimar el vector de datos faltantes, ubicadas en los períodos $T_1 - 7 = n - 6$ hasta $T_5 = n + 18$, excepto en los cinco períodos en los cuales se presentaron los datos faltantes. Las mismas observaciones se requieren para realizar la estimación por el método de Nieto-Martínez.

6.Conclusiones

Cuando una serie temporal univariada $\{\mathbf{Z}_t\}$ obedece a un modelo AR(p), no se necesita recurrir a toda la información observada para estimar los datos faltantes, basta con utilizar los p datos anteriores al primer dato faltante, más los p datos posteriores al último dato faltante, más todos o parte de los datos observados en los períodos de tiempo comprendidos entre el primero y último dato faltante.

Este resultado es trascendente desde el punto de vista computacional, en el sentido que permite detener el programa de computador para estimar el vector de datos faltantes utilizando el método de Nieto-Martínez, cuando el modelo de la serie es AR(p). Es importante recalcar que al utilizar los datos observados ubicados en el período $T_k + p + 1$ en adelante, la estimación es la misma que la obtenida hasta utilizar el dato ubicado en el período $T_k + p$; por lo tanto al detener el programa no se está truncando la estimación.

El resultado aquí obtenido indica también que computacionalmente es preferible el método de Nieto-Martínez a los métodos que usan Suavizador de Punto Fijo, ya

que estos realizan el proceso de estimación hasta agotar los datos observados en la serie.

El método de Peña-Maravall para estimar el vector de datos faltantes, en el caso en que el modelo de la serie temporal sea AR(p), puede utilizar menos información (a lo sumo la misma) que la usada por el método de Nieto-Martínez.

REFERENCIAS

Bell, W.R. (1984). *"Signal Extraction for Nonstationary Time Series "*, The Annals of Statistics, 12, 646-664.

Box,G.E.P. and G.M Jenkins (1976). *Times Series Analysis Forecasting and Control.* Holden-day, Oakland,California.

Gómez, V. and A. Maravall (1994). *"Estimation, Prediction and Interpolation for Nonstationary Series with the Kalman Filter"* , Journal of the American Statistical Association, Vol 89, 611-624.

Kohn, R. and C.F.Ansley (1986). *"Estimation, Prediction and Interpolation for ARIMA Models with Missing Data "*, Journal of the American Statistical Association, Vol 81, 751-761.

Nieto, F. and J. Martínez (1994). *"Recursive Approach for Estimating Missing Observations in a Time Series when the ARIMA Model for de Process Known"*, Reporte interno No. 35, Unidad de investigación, Dpto. de Matemáticas y Estadítica, Universidad Nacional de Colombia, Bogotá.

Peña, D. and Maravall (1991). *"Interpolation, Outliers and Inverse Autocorrelations "*, Commun. Statist., Theory Meth., 20(10), 3175-3186.