Revista Colombiana de Estadística
Diciembre 2008, volumen 31, no. 2, pp. 183 a 192

Modelo factorial dinámico *threshold*

Threshold Dynamic Factor Model

María Elsa Correal[1,a], Daniel Peña[2,b]

[1]Departamento de Ingeniería Industrial, Universidad de los Andes, Bogotá, Colombia

[2]Departamento de Estadística y Economía, Universidad Carlos III de Madrid, Madrid, España

Resumen

En este artículo se introduce el modelo factorial dinámico *threshold*, el cual permite analizar sistemas de series temporales que presenten comportamientos no lineales del tipo umbral. Se propone un método de estimación que combina el algoritmo EM con un procedimiento de búsqueda directa utilizando los algoritmos del filtro y de suavización de Kalman. El procedimiento estima factores comunes con comportamientos que cambian de régimen de acuerdo con una variable umbral.

***Palabras clave*:** series de tiempo no lineales, análisis factorial, modelo threshold, algoritmo EM, filtro de Kalman.

Abstract

This paper introduces a threshold dynamic factor model for the analysis of vector time series which shows non-linear behavior of threshold type. We propose an estimation procedure combining an EM algorithm with a grid search procedure by the ways of the Kalman filter and smoothing recursions. We estimate common latent threshold factors that may explain the dynamic relationships within the group of variables.

***Key words*:** Nonlinear time series, Factor analysis, Threshold model, EM algorithm, Kalman filter.

1. Introducción

En este artículo se presenta un procedimiento para estimar factores comunes en series temporales que presenten comportamientos no lineales del tipo *threshold*.

[a]Profesora asociada. E-mail: mcorreal@uniandes.edu.co
[b]Profesor catedrático. E-mail: dpena@est-econ.uc3m.es

Tanto los procesos multivariados como la no linealidad comprenden desarrollos metodológicos de especial interés dentro del estudio de series de tiempo. Uno de los modelos no lineales para series de tiempo más difundido es el modelo autorregresivo umbral, TAR (*Threshold AutoRegressive*), propuesto inicialmente por Tong & Lim (1980). Este modelo está representado mediante diferentes procesos autorregresivos que se activan cuando determinada variable sobrepasa un valor umbral. El análisis de los modelos TAR en el caso multivariado es más reciente. Tsay (1998) es tal vez el primero en proponer un procedimiento de estimación y una prueba de no linealidad para el caso vectorial. La inferencia estadística en los modelos *threshold* ha sido estudiada entre otros por Hansen (1997, 2000), Gonzalo & Pitarakis (2002) y, para el caso multivariado, por Tsay (1998). Al igual que en los modelos vectoriales $VARMA$ (*Vector AutoRegressive-MovingAverage*), en los modelos TAR multivariados existen múltiples estructuras con características similares y no existe una solución simple para la identificación de los parámetros. La proliferación de parámetros puede ser tan alta como para hacer la estimación intratable en la práctica. Un modelo factorial no solamente reduce la dimensión del sistema, sino que permite dejar al descubierto componentes comunes al conjunto de variables que explican las interrelaciones existentes entre ellas. El modelo factorial dinámico de Peña & Box (1987) representa las variables observadas mediante una suma de dos componentes latentes ortogonales: una común a todas las variables, descrita por un proceso $ARMA$ (*AutoRegressive-MovingAverage*) de dimensión reducida; y otra específica a cada variable particular, que no está correlacionada con la componente común. Este modelo inicialmente formulado para series estacionarias ha sido generalizado a series no estacionarias en Peña & Poncela (2004, 2006); recientemente se han presentado técnicas para su identificación en Hu & Chou (2004). Este modelo debe distinguirse del modelo factorial dinámico utilizado por Stock & Watson (2002) y Forni et al. (2005), en el cual se asume que el número de variables o la dimensión del sistema tiende a infinito. La presencia o ausencia de este supuesto es determinante en los procesos de identificación y estimación. En el modelo factorial que se define en este trabajo no se hace este supuesto. El objetivo de este trabajo es extender el modelo factorial dinámico de Peña y Box para permitir tener en cuenta efectos no lineales del tipo umbral.

El modelo factorial dinámico *threshold* se define en la segunda sección del documento y sus propiedades se analizan en la tercera. En la cuarta se presenta el método de estimación. La estrategia consiste en realizar la estimación secuencialmente por concentración de la función de verosimilitud, combinando el algoritmo *EM (Expectation-Maximization)* con un método de búsqueda directa. En la quinta sección, la metodología se aplica a un sistema de caudales de ríos colombianos en el cual hay dos regímenes que se activan mediante la variable del Índice de Oscilación del Sur.

2. Formulación del modelo

Definición 1. Sea $\mathbf{Z}_t$ una serie temporal k-dimensional, $\mathbf{Z}_t = (z_{1t}, z_{2t}, \ldots, z_{kt})'$ con media cero. Diremos que $\mathbf{Z}_t$ se representa mediante un *modelo factorial dinámico threshold* con c regímenes de órdenes $p_1, p_2, \ldots, p_c$ y variable umbral w_t,

si

$$\begin{cases} \mathbf{Z}_t = \Lambda \mathbf{f}_t + \mathbf{u}_t; \\ \mathbf{f}_t = \sum_{i=1}^{p_j} \phi_i^{(j)} \mathbf{f}_{t-i} + \Upsilon^{(j)} \mathbf{a}_t, \quad \text{si } w_{t-d} \in (\gamma_{j-1}, \gamma_j], \quad j = 1, \ldots, c. \end{cases} \tag{1}$$

donde w_t es una variable aleatoria unidimensional observable y estacionaria, $\mathbf{f}_t$ es un vector aleatorio r-dimensional no observable con media cero, $\mathbf{u}_t$ es ruido blanco k-dimensional con matriz de varianza-covarianza Σ_u diagonal y definida positiva, $\mathbf{a}_t$ es un ruido blanco r-dimensional con matriz de varianza-covarianza la identidad I_r y tales que $\mathbf{u}_t$ sea independiente de $\mathbf{f}_{t-h}$ para $h \geq 0$, $\mathbf{a}_t$ independiente de $\mathbf{f}_{t-h}$ para $h \geq 1$, $\{w_t\}$, $\{\mathbf{u}_t\}$ y $\{\mathbf{a}_t\}$ independientes entre sí. Los parámetros del modelo son los denominados parámetros umbral: $-\infty = \gamma_0 < \gamma_1 < \cdots < \gamma_{c-1} < \gamma_c = -\infty$, el entero positivo d, rezago de la variable umbral, la matriz de carga Λ, de dimensión $(k \times r)$ que debe ser tal que $rango(\Lambda) = r$ y $\Lambda'\Lambda = I_r$, I_r matriz identidad de orden r, y los parámetros que determinan la dinámica del factor, $\phi_i^{(j)}$, $\Upsilon^{(j)}$ $j = 1, \ldots, c$, matrices de dimensión $(r \times r)$, con $\Upsilon^{(j)}$ diagonal y definida positiva.

El modelo propuesto detecta componentes comunes no lineales que puedan ser representadas por modelos umbral y que involucren la dinámica propia del sistema. La idea general es representar el vector temporal mediante la suma de dos componentes latentes ortogonales: una común a las componentes del vector, descrita por un proceso vectorial autorregresivo *threshold* TAR, de dimensión menor, y otra específica a cada componente particular. La variable umbral w puede ser una de las componentes estacionarias del vector observado, z_{jt}, $j \leq k$, o una variable exógena estacionaria que afecte el estado del sistema, o una combinación de las componentes de $\mathbf{Z}_t$. Esta combinación debe ser estacionaria.

Mediante este modelo pueden salir a relucir características significativas en un régimen y no en el otro. El proceso formulado para los factores permite tener en consideración autorregresiones con órdenes diferentes en los regímenes. La serie de los factores es generada por procesos diferentes en diferentes instantes de tiempo; su cambio es consecuencia de un estado del sistema que se mantiene hasta que determinada variable sobrepasa un valor umbral.

3. Propiedades del modelo

3.1. Identificación

El modelo propuesto hereda los problemas de identificación presentes en el modelo factorial estático debido a la no observabilidad de f. La imposición de una estructura TAR con c regímenes para los factores no evita la no identificación de los parámetros. Efectivamente, si $\mathbf{f}_t$ es un TAR r-dimensional con c regímenes de órdenes $p_1, \ldots, p_c$ y variable umbral w_t, entonces para cualquier matriz $(r \times r)$ no singular C, el vector $\mathbf{f}_t^* = C\mathbf{f}_t$ será también un TAR r-dimensional para la misma variable umbral, el mismo número de regímenes c y los mismos órdenes

autorregresivos dentro de cada régimen. Específicamente, $\mathbf{f}_t^*$ se expresa como

$$\mathbf{f}_t^* = \sum_{i=1}^{p_j} \phi_i^{*(j)} \mathbf{f}_{t-i}^* + \Upsilon^{*(j)} \mathbf{a}_t, \qquad \text{si } w_{t-d} \in (\gamma_{j-1}, \gamma_j],\ j = 1, \ldots, c$$

donde $\phi_i^{*(j)} = C\phi_i^{(j)} C^{-1}$ y $\Upsilon^{*(j)} = C\Upsilon^{(j)}$. Dicho de otra forma, para cualquier matriz C no singular, los conjuntos

$$\left\{\Lambda, \phi_1^{(1)}, \ldots, \phi_{p_c}^{(c)}, \Upsilon^{(1)}, \ldots, \Upsilon^{(c)}, \Sigma_u, d, \gamma_1, \ldots, \gamma_c\right\}$$

y

$$\left\{\Lambda C^{-1}, C\phi_1^{(1)} C^{-1}, \ldots, C\phi_{p_c}^{(c)} C^{-1}, C\Upsilon^{(1)}, \ldots, C\Upsilon^{(c)}, \Sigma_u, d, \gamma_1, \ldots, \gamma_c\right\}$$

no pueden distinguirse a partir de las observaciones.

Proposición 1. *Las restricciones $\Lambda'\Lambda = I_r$ y $\Upsilon^{(j)}$ matriz diagonal y positiva definida para $j = 1, \ldots, c$, eliminan esta fuente de indeterminación.*

Demostración. En efecto, si Λ y $\Lambda^* = \Lambda C^{-1}$ satisfacen la primera restricción, $\Lambda'\Lambda = I_r$ y $(\Lambda^*)'\Lambda^* = (C^{-1})'C^{-1} = I_r$, entonces C será matriz ortogonal. Además, si $\Upsilon^{(j)}$ y $\Upsilon^{*(j)} = C\Upsilon^{(j)}$ satisfacen la segunda restricción, entonces $C = \Upsilon^{*(j)}(\Upsilon^{(j)})^{-1}$ y, por tanto, C es diagonal. Puesto que la única matriz ortogonal y diagonal es la matriz identidad, se concluye que $C = I_r$. □

Vale la pena mencionar que la representación TAR de los factores en el modelo factorial dinámico *threshold* (1) puede escribirse como $\mathbf{f}_t = \sum_{i=1}^{p_j} \phi_i^{(j)} \mathbf{f}_{t-i} + \mathbf{a}_t^{(j)}$ con $\mathbf{a}_t^{(j)} = \Upsilon^{(j)} \mathbf{a}_t$. Puede verse entonces que la matriz de varianza covarianza de $\mathbf{a}_t$ puede restringirse a la identidad sin pérdida de generalidad.

3.2. Estructura para las matrices de covarianza rezagadas

Las matrices de covarianza para diferentes rezagos contienen información acerca de la dinámica de las interrelaciones entre las diferentes componentes del proceso. Sean $\Gamma_Z(h) = E(\mathbf{Z}_{t-h}\mathbf{Z}_t')$, $\Gamma_f(h) = E(\mathbf{f}_{t-h}\mathbf{f}_t')$ y $\Sigma_u(h) = E(\mathbf{u}_{t-h}\mathbf{u}_t')$ para $h = 0, 1, 2, \ldots$ las matrices de covarianza rezagadas de $\mathbf{Z}, \mathbf{f}$ y $\mathbf{u}$ respectivamente.

Proposición 2. *Si Z_t se representa mediante un modelo autorregresivo* threshold*,*

$$rango(\Gamma_Z(h)) = r, \qquad \text{para } h = 1, 2, \ldots$$

Demostración. En efecto, si $\mathbf{f}_t$ es estacionario de segundo orden, $\mathbf{Z}_t$ también lo es, y puesto que $\mathbf{Z}_t = \Lambda\mathbf{f}_t + \mathbf{u}_t$, y $\mathbf{u}_t$ es ruido blanco,

$$\Gamma_Z(h) = \Lambda\Gamma_f(h)\Lambda', \qquad \text{para } h = 1, 2, \ldots$$ □

Esta propiedad es muy útil en la etapa de identificación del número de factores comunes.

3.3. Modelo de rezagos distribuidos por regímenes

Proposición 3. $\mathbf{Z}_t$ *puede expresarse como un modelo de rezagos distribuidos por regímenes*

$$\mathbf{Z}_t = \sum_{i=1}^{p_j} \Lambda_i^{(j)} \mathbf{f}_{t-i} + \varepsilon_t^{(j)}, \quad \text{si } w_{t-d} \in (\gamma_{j-1}, \gamma_j], \quad j = 1, \ldots, c$$

donde los espacios nulos de $\Lambda_i^{(j)}$ *comparten un subespacio común de dimensión* $(k - r)$.

Demostración. En efecto, remplazando la segunda ecuación de (1) en la primera se obtiene $\Lambda_i^{(j)} = \Lambda\phi_i^{(j)}$; por tanto, si M es una matriz $k \times (k - r)$ cuyas columnas generan el espacio nulo de Λ', $M'\Lambda_i^{(j)} = 0$ para $i = 1, \ldots, p_j$, $j = 1, \ldots, c$. Sin embargo, las matrices $\Lambda_i^{(j)}$ no necesariamente tienen rango completo y $rango\Big(\Lambda_i^{(j)}\Big) = rango\Big(\phi_i^{(j)}\Big)$.

El ruido asociado al modelo expresado en rezagos distribuidos es $\varepsilon_t^{(j)} = \Lambda\Upsilon^{(j)}\mathbf{a}_t + \mathbf{u}_t$, y su matriz de covarianza, no necesariamente diagonal, viene dada por

$$\Psi_{\varepsilon}^{(j)} = \Lambda D^{(j)}\Lambda' + \Psi_u, \qquad j = 1, \ldots, c$$

con $D^{(j)} = \Upsilon^{(j)}\Upsilon^{(j)}$ matriz diagonal. □

4. Estimación

Se propone estimar los parámetros del modelo por máxima verosimilitud mediante un algoritmo que combina el principio del algoritmo EM con un método de búsqueda directa. El procedimiento maximiza el logaritmo de la función de verosimilitud L_Z^w de forma secuencial, primero sobre $\psi_1 = \left\{\Lambda, \mathbf{\Phi}^{(1)}, \mathbf{\Phi}^{(2)}, \Upsilon^{(1)}, \Upsilon^{(2)}, \Sigma_u\right\}$ y luego sobre $\psi_2 = \{d, \gamma\}$. Para d y γ fijos, el máximo sobre ψ_1 se obtiene mediante un algoritmo EM. La utilización del algoritmo EM en el contexto de factores dinámicos fue propuesta inicialmente por Shumway & Stoffer (1982) y Watson & Engle (1983) y ha sido utilizada posteriormente por Wu et al. (1996) y Peña & Poncela (2006), entre otros. En la segunda etapa, por búsqueda directa se encuentran los valores $\widehat{d}$ y $\widehat{\gamma}$ que maximicen L_Z^w. La búsqueda se realiza para $d \in \{1, \ldots, \bar{d}\}$, $\bar{d}$ una cota para el retardo y $\gamma \in \{\gamma_1, \ldots, \gamma_L\}$, conjunto formado por los cuantiles muestrales de la variable umbral escogidos de forma tal que en cada régimen se tengan suficientes observaciones para estimar adecuadamente los parámetros asociados. Siendo así las cosas, se tendrán que realizar $\bar{d}\,L$ veces el procedimiento EM.

Para d y γ fijos, el logaritmo de la función de verosimilitud de los datos completos, $L^w_{Z,f}(\psi_1;\psi_2)$, puede separarse por una función indicadora:

$$\begin{aligned}L^w_{Z,f}(\psi_1;\psi_2) = cte &- \frac{T}{2}\log|\Sigma_u| - \frac{1}{2}\sum_{t=1}^{T}(\mathbf{Z}_t - \Lambda\mathbf{f}_t)'\Sigma_u^{-1}(\mathbf{Z}_t - \Lambda\mathbf{f}_t)\\ &- \frac{1}{2}\sum_{t\in I_1}\log\left|\Upsilon^{(1)}\Upsilon^{(1)}\right| - \frac{1}{2}\sum_{t\in I_2}\log\left|\Upsilon^{(2)}\Upsilon^{(2)}\right|\\ &- \frac{1}{2}\sum_{t\in I_1(d,\gamma)}\left(\mathbf{f}_{t+1} - \varphi^{(1)}\mathbf{X}_t\right)'\left(\Upsilon^{(1)}\Upsilon^{(1)}\right)^{-1}\left(\mathbf{f}_{t+1} - \varphi^{(1)}\mathbf{X}_t\right)\\ &- \frac{1}{2}\sum_{t\in I_2(d,\gamma)}\left(\mathbf{f}_{t+1} - \varphi^{(2)}\mathbf{X}_t\right)'\left(\Upsilon^{(2)}\Upsilon^{(2)}\right)^{-1}\left(\mathbf{f}_{t+1} - \varphi^{(2)}\mathbf{X}_t\right)\end{aligned}$$

donde $I_1(d,\gamma) = \{t \in \{1,\ldots,T\}/w_{t-d} < \gamma\}$, $I_2(d,\gamma) = \{t \in \{1,\ldots,T\}/w_{t-d} \geq \gamma\}$, $\mathbf{X}_t = (\mathbf{f}_t', \mathbf{f}_{t-1}', \ldots, \mathbf{f}_{t-p+1}')'$ vector $rp \times 1$, y $\varphi^{(i)} = \left[\phi_1^{(i)}\middle|\phi_2^{(i)}\middle|\cdots\middle|\phi_p^{(i)}\right]$ matriz $r \times rp$, para $i = 1, 2$.

Utilizando la solución de la k-ésima iteración, $\widetilde{\psi}_1^{(k)}$, la evaluación del paso E del algoritmo

$$Q\left(\psi_1;\widetilde{\psi}_1^{(k)}\right) = E_{\widetilde{\psi}_1^{(k)}}\left(L^W_{Z,f}(\psi_1;\psi_2)\middle|\mathbf{Z}_1,\ldots,\mathbf{Z}_T,w_1,\ldots,w_T\right) \tag{2}$$

se obtiene involucrando los algoritmos del filtro de Kalman y de suavización de intervalo fijo aplicados a una representación espacio-estado del modelo. Como resultado de este paso, se obtienen las sucesiones $\widehat{\mathbf{X}}^{(k)}_{t|T}$, $P^{(k)}_{t|T}$, $M^{(k)}_{t|T}$, $t = 1,\ldots,T$, donde

$$\begin{aligned}\widehat{\mathbf{X}}^{(k)}_{t|T} &= E_{\widetilde{\psi}_1^{(k)}}\left(\mathbf{X}_t\middle|\mathbf{Z}_1,\ldots,\mathbf{Z}_T,w_1,\ldots,w_T\right)\\ P^{(k)}_{t|T} &= E_{\widetilde{\psi}_1^{(k)}}\left[\left(\mathbf{X}_t - \widehat{\mathbf{X}}_{t|T}\right)\left(\mathbf{X}_t - \widehat{\mathbf{X}}_{t|T}\right)'\middle|\mathbf{Z}_1,\ldots,\mathbf{Z}_T,w_1,\ldots,w_T\right]\\ M^{(k)}_{t|T} &= E_{\widetilde{\psi}_1^{(k)}}\left[\left(\mathbf{X}_{t+1} - \widehat{\mathbf{X}}_{t+1|T}\right)\left(\mathbf{X}_t - \widehat{\mathbf{X}}_{t|T}\right)'\middle|\mathbf{Z}_1,\ldots,\mathbf{Z}_T,w_1,\ldots,w_T\right]\end{aligned}$$

Teniendo en cuenta que $\mathbf{f}_t = S_t'\mathbf{X}_t$ para $S' = \left[I_r \mid 0_r \mid \cdots \mid 0_r\right]$, también se obtienen las sucesiones

$$\begin{aligned}\widehat{\mathbf{f}}^{(k)}_{t|T} = E_{\widetilde{\psi}_1^{(k)}}\left(\mathbf{f}_t\middle|\mathbf{Z}_1,\ldots,\mathbf{Z}_T,w_1,\ldots,w_T\right) &= S'\widehat{\mathbf{X}}^{(k)}_{t|T}\\ E_{\widetilde{\psi}_1^{(k)}}\left(\mathbf{f}_t\mathbf{f}_t'\middle|\mathbf{Z}_1,\ldots,\mathbf{Z}_T,w_1,\ldots,w_T\right) &= S'\left(P^{(k)}_{t|T} + \widehat{\mathbf{X}}^{(k)}_{t|T}\widehat{\mathbf{X}}^{(k)\prime}_{t|T}\right)S\\ E_{\widetilde{\psi}_1^{(k)}}\left(\mathbf{f}_{t+1}\mathbf{X}_t'\middle|\mathbf{Z}_1,\ldots,\mathbf{Z}_T,w_1,\ldots,w_T\right) &= S'\left(M^{(k)}_{t|T} + \widehat{\mathbf{X}}^{(k)}_{t|T}\widehat{\mathbf{X}}^{(k)\prime}_{t|T}\right)\end{aligned}$$

En la k-ésima iteración del paso M del algoritmo se maximiza $Q\Big(\psi_1;\widetilde{\psi}_1^{(k)}\Big)$ con respecto a ψ_1. Se mostrará a continuación que la solución viene dada por

$$\widetilde{\Lambda}_{(k+1)} = \left(\sum_{t=1}^{T}\mathbf{Z}_t\widehat{\mathbf{f}}_{t|T}^{(k)\prime}\right)\left(\sum_{t=1}^{T}\widehat{\mathbf{f}}_{t|T}^{(k)}\widehat{\mathbf{f}}_{t|T}^{(k)\prime}+P^{(k)}\right)^{-1}$$

$$\widetilde{\Sigma}_v^{(k+1)} = \frac{1}{T}\left(\sum_{t=1}^{T}\Big(\mathbf{Z}_t-\widetilde{\Lambda}_{(k+1)}\widehat{\mathbf{f}}_{t|T}^{(k)}\Big)\Big(\mathbf{Z}_t-\widetilde{\Lambda}_{(k+1)}\widehat{\mathbf{f}}_{t|T}^{(k)}\Big)'+\widetilde{\Lambda}_{(k+1)}P^{(k)}\widetilde{\Lambda}'_{(k+1)}\right)$$

$$\widetilde{\varphi}_{(k+1)}^{(j)} = \left(\sum_{t\in I_j}\widehat{\mathbf{f}}_{t+1|T}^{(k)}\widehat{\mathbf{X}}_t^{(k)\prime}+V_j^{(k)}\right)\left(\sum_{t\in I_j}\widehat{\mathbf{f}}_{t+1|T}^{(k)}\widehat{\mathbf{f}}_{t+1|T}^{(k)\prime}+W_j^{(k)}\right)^{-1}, \qquad j=1,2$$

$$\widetilde{\Upsilon}_{(k+1)}^{(j)} = \frac{1}{T_j}\left(\sum_{t\in I_j}\Big(\widehat{\mathbf{f}}_{t+1|T}^{(k)}-\widetilde{\varphi}_{(k+1)}^{(j)}\widehat{\mathbf{X}}_t^{(k)}\Big)'\Big(\widehat{\mathbf{f}}_{t+1|T}^{(k)}-\widetilde{\varphi}_{(k+1)}^{(j)}\widehat{\mathbf{X}}_t^{(k)}\Big)+\widetilde{\varphi}_{(k+1)}^{(j)}W_j^{(k)}\widetilde{\varphi}_{(k+1)}^{(j)}\right)$$

donde T_j número de casos en el régimen j y las matrices $P^{(k)}$, $V_j^{(k)}$ y $W_j^{(k)}$ están dadas por

$$P^{(k)}=\sum_{t=1}^{T}S'P_{t|T}^{(k)}S, \quad V_j^{(k)}=\sum_{t\in I_j}S'M_{t|T}^{(k)}, \quad W_j^{(k)}=\sum_{t\in I_j}S'P_{t|T}^{(k)}S$$

para $S'=\big[I_r\mid 0_r\mid\cdots\mid 0_r\big]$.

En efecto, el término que involucra a Λ en (2) es

$$E\left(\sum_{t=1}^{T}\big(\mathbf{Z}_t-\Lambda\mathbf{f}_t\big)'\Sigma_u^{-1}\big(\mathbf{Z}_t-\Lambda\mathbf{f}_t\big)\Big|\mathbf{Z}_1,\ldots,\mathbf{Z}_T,w_1,\ldots,w_T\right)$$

que es igual a

$$traza\left(\Sigma_u^{-1}\sum_{t=1}^{T}\left(\mathbf{Z}_t\mathbf{Z}_t'-\Lambda\widehat{\mathbf{f}}_{t|T}^{(k)}\mathbf{Z}_t'-\mathbf{Z}_t\widehat{\mathbf{f}}_{t|T}^{(k)\prime}\Lambda'+\Lambda\Big(\widehat{\mathbf{f}}_{t|T}^{(k)}\widehat{\mathbf{f}}_{t|T}^{(k)\prime}+S'P_{t|T}^{(k)}S\Big)\Lambda'\right)\right)$$

y, por tanto,

$$\frac{\partial Q}{\partial\Lambda}=-\Sigma_u^{-1}\sum_{t=1}^{T}\mathbf{Z}_t\widehat{\mathbf{f}}_{t|T}^{(k)\prime}+\Sigma_u^{-1}\Lambda\sum_{t=1}^{T}\Big(\widehat{\mathbf{f}}_{t|T}^{(k)}\widehat{\mathbf{f}}_{t|T}^{(k)\prime}+S'P_{t|T}^{(k)}S\Big)$$

Igualando a cero este sistema de derivadas parciales, se obtiene el resultado para $\widetilde{\Lambda}_{(k+1)}$.

Ahora, fijando $\widetilde{\Lambda}_{(k+1)}$, la parte de $E\big(L_{Z,f}^{W}(\psi_1;\psi_2)\big|\mathbf{Z}_1,\ldots,\mathbf{Z}_T,w_1,\ldots,w_T\big)$ que involucra a Σ_u puede escribirse como

$$\frac{T}{2}\log|\Sigma_u|-\frac{1}{2}\times$$
$$traza\left[\Sigma_u^{-1}\sum_{t=1}^{T}\left(\Big(\mathbf{Z}_t-\widetilde{\Lambda}_{(k+1)}\widehat{\mathbf{f}}_{t|T}^{(k)}\Big)\Big(\mathbf{Z}_t-\widetilde{\Lambda}_{(k+1)}\widehat{\mathbf{f}}_{t|T}^{(k)}\Big)'+\widetilde{\Lambda}_{(k+1)}S'P_{t|T}^{(k)}S\widetilde{\Lambda}_{(k+1)}\Big)\right]$$

y entonces vale el resultado para $\widetilde{\Sigma}_u^{(k+1)}$.

El resultado para $\widetilde{\varphi}_{(k+1)}^{(j)}$ y $\widetilde{\Upsilon}_{(k+1)}^{(j)}$ se prueba de forma similar. En este caso, aparece $E(\mathbf{f}_{t+1}\mathbf{X}_t)$, que involucra a $V^{(k)} = S' M_{t|T}^{(k)}$. Más detalles pueden consultarse en Correal (2007).

El algoritmo proporciona también el estimador óptimo de los factores:

$$\widehat{\mathbf{f}}_{t|T} = E\big(\mathbf{f}_t \mid \mathbf{Z}_1, \ldots, \mathbf{Z}_T, w_1, \ldots, w_T\big)$$

5. Aplicación

El modelo y el método de estimación se aplican a un vector de dimensión 5 conformado por los caudales de los ríos colombianos Calima, Cauca, Grande, Ovejas y Prado, que pertenecen a la cuenca del Magdalena. Los datos históricos disponibles abarcan un periodo de 36 años y corresponden al periodo comprendido entre enero de 1955 y diciembre de 1990, para un total de 432 observaciones mensuales.

El procedimiento se realizó en tres etapas resumidas a continuación. Los resultados detallados pueden consultarse en Correal (2007). En la primera se probó la hipótesis de que los caudales presentan un comportamiento no lineal del tipo *threshold* con variable umbral Índice de Oscilación del Sur, IOS, variable atmosférica relacionada con el evento climático del fenómeno de El Niño. La hipótesis se contrastó mediante el test propuesto por Tsay (1989). Este se basa en autorregresiones reordenadas de acuerdo con la variable umbral w_{t-d}. El test se aplicó para los retardos $d = 1, 2, \ldots, 12$ y para los datos $z_{it} = (1 - \theta B^{12})^{-1}(1 - B^{12}) \log c_{it}$, $i = 1, \ldots, 5$; c_{it} caudal del $i-$ésimo río en el instante t. De los nueve ríos considerados originalmente, los cinco utilizados en esta aplicación dieron significativos.

En la segunda etapa, se procedió a identificar el número de factores comunes. Para esto se realizaron dos pruebas, ambas basadas en los vectores propios de las matrices de autocovarianza rezagadas observadas $\Gamma_Z(k) = E(Z_{t-k} Z_t')$, y cuyos detalles pueden consultarse en Peña & Poncela (2006) y Hu & Chou (2004). Los resultados de estas pruebas llevaron a plantear un modelo con dos factores comunes.

En la tercera etapa se implementó el algoritmo para estimar los parámetros del modelo. El algoritmo de búsqueda se realizó sobre el par de conjuntos $\{1, 2, \ldots, 12\}$ para d y $\{-2.6, -2.5, \ldots, 2.3\}$ para γ, con lo que el algoritmo EM se corrió 60 veces.

El estimador del valor umbral fue $\widehat{\gamma} = -2.3$. Puesto que los episodios del fenómeno de El Niño se presentan acompañados de valores negativos del Índice de Oscilación al Sur, el régimen 1 puede asociarse a una de las fases del fenómeno. El resultado para el rezago fue $\widehat{d} = 1$ y la estimación para la matriz de carga del modelo factorial dinámico *threshold* es

$$\widehat{\Lambda} = \begin{bmatrix} 0.29 & 0.54 & 0.34 & 0.47 & 0.52 \\ 0.94 & -0.05 & -0.06 & -0.23 & -0.22 \end{bmatrix}'$$

La estimación de la matriz de varianzas de los términos específicos es $\widehat{\Sigma}_u = diag(0.016, 0.006, 0.032, 0.040, 0.201)$, y el modelo estimado para el factor es

$$\begin{bmatrix} f_{1t} \\ f_{2t} \end{bmatrix} = \begin{bmatrix} 0.70 & 0.00 \\ 0.00 & 0.55 \end{bmatrix} \begin{bmatrix} f_{1,t-1} \\ f_{2,t-1} \end{bmatrix} + \begin{bmatrix} a_{1t} \\ a_{2t} \end{bmatrix} \qquad \text{si } IOS_{t-1} < -2.3$$

$$\begin{bmatrix} f_{1t} \\ f_{2t} \end{bmatrix} = \begin{bmatrix} 0.78 & 0.00 \\ 0.00 & 0.67 \end{bmatrix} \begin{bmatrix} f_{1,t-1} \\ f_{2,t-1} \end{bmatrix} + \begin{bmatrix} a_{1t} \\ a_{2t} \end{bmatrix} \qquad \text{si } IOS_{t-1} \geq -2.3$$

donde $cov\left(\mathbf{a}_t^{(1)}\right) = diag(0.30, 0.08)$, $cov\left(\mathbf{a}_t^{(2)}\right) = diag(0.27, 0.04)$ para $\mathbf{a}_t^{(1)} = \left(a_{1t}^{(1)}, a_{2t}^{(1)}\right)'$ y $\mathbf{a}_t^{(2)} = \left(a_{1t}^{(2)}, a_{2t}^{(2)}\right)'$.

6. Conclusiones

El modelo presentado en este trabajo permite analizar sistemas de series temporales que presenten efectos no lineales del tipo *threshold*. El modelo puede interpretarse o bien como una reparametrización del modelo TAR vectorial, que reduce significativamente el número de parámetros, o bien como una extensión del modelo de Peña y Box que permite tener en cuenta efectos no lineales. El vector de los factores comunes se representa mediante diferentes procesos autorregresivos que se activan cuando determinada variable sobrepasa un valor umbral. Los diferentes regímenes pueden relacionarse con los estados de una economía o con estados propios de la naturaleza, como el caso que se estudia en la aplicación, donde los estados están asociados a la presencia o ausencia del fenómeno de El Niño.

Agradecimientos

Este artículo es producto del trabajo de tesis del primer autor (Correal 2007) para obtener el título de doctor en Estadística de la Universidad Nacional de Colombia.

[Recibido: marzo de 2008 — Aceptado: septiembre de 2008]

Referencias

Correal, M. E. (2007), Modelo factorial dinámico con efectos umbral, Tesis doctoral, Departamento de Estadística, Facultad de Ciencias, Universidad Nacional de Colombia.

Forni, M., Hallin, M., Lippi, M. & Reichlin, L. (2005), 'The Generalized Dynamic Factor Model: One-Sided Estimation and Forecasting', *Journal of the American Statistical Association* **100**, 830–840.

Gonzalo, J. & Pitarakis, J. Y. (2002), 'Estimation and Model Selection Based Inference in Single and Multiple Threshold Models', *Journal of Econometrics* **110**, 319–352.

Hansen, B. E. (1997), 'Inference in TAR Models', *Studies in Nonlinear Dynamics and Econometrics* **2**, 1–14.

Hansen, B. E. (2000), 'Sample Splitting and Threshold Estimation', *Econometrica* **68**, 575–603.

Hu, Y. P. & Chou, R. J. (2004), 'On the Peña-Box Model', *Journal of Time Series Analysis* **25**, 811–830.

Peña, D. & Box, G. E. P. (1987), 'Identifying a Simplifying Structure in Time Series', *Journal of the American Statistical Association* **82**, 836–843.

Peña, D. & Poncela, P. (2004), 'Forecasting with Nonstationary Dynamic Factor Models', *Journal of Econometrics* **119**, 291–321.

Peña, D. & Poncela, P. (2006), 'Nonstationary Dynamic Factor Models', *Journal of Statistical Planning and Inference* **136**, 1237–1257.

Shumway, R. H. & Stoffer, D. S. (1982), 'An Approach to Time Series Smoothing and Forecasting Using the EM Algorithm', *Journal of Time Series Analysis* **3**, 253–264.

Stock, J. H. & Watson, M. W. (2002), 'Forecasting Using Principal Components From a Large Number of Predictors', *Journal of the American Statistical Association* **97**, 1167–1179.

Tong, H. & Lim, K. S. (1980), 'Threshold Autoregression, Limit Cycles and Cyclical Data', *Journal of The Royal Statistics Society* **Ser. B**(4), 245–292.

Tsay, R. S. (1989), 'Outliers, Level Shifts and Variance Changes in Time Series', *Journal of Forecasting* **7**, 1–20.

Tsay, R. S. (1998), 'Testing and Modeling Multivariate Threshold Models', *Journal of the American Statistical Association* **93**, 1188–1202.

Watson, M. W. & Engle, R. F. (1983), 'Alternative Algorithms for the Estimation of Dynamic Factor, Mimic and Varying Coefficient Regression Models', *Journal of Econometrics* **23**, 385–400.

Wu, L. S., Pai, J. S. & Hosking, J. R. M. (1996), 'An Algorithm for Estimating Parameters of State-Space Models', *Statistics & Probability Letters* **28**, 99–106.