Revista Colombiana de Estadística
Diciembre 2009, volumen 32, no. 2, pp. 189 a 212

Comparación entre dos métodos de reducción de dimensionalidad en series de tiempo

Comparison between Two Dimensionality Reduction Methods in Time Series

Hanwen Zhang[1,a]

[1]Centro de Investigaciones y Estudios Estadísticos (CIEES), Facultad de Estadística, Universidad Santo Tomás, Bogotá, Colombia

Resumen

En este trabajo se analizan dos métodos de reducción de dimensionalidad en series de tiempo multivariadas estacionarias: el método de Peña y Box, basado en el dominio del tiempo, y el método de Brillinger, basado en el dominio de las frecuencias. Se encontraron dos fallas en el método de Peña y Box, y se propusieron correcciones a estas. También se compararon los dos métodos con respecto a la capacidad para identificar el número de factores latentes mediante simulaciones y se realizó una aplicación empírica.

***Palabras clave*:** series de tiempo multivariadas, reducción de dimensionalidad, dominio del tiempo, dominio de las frecuencias.

Abstract

Two methods of dimensionality reduction of multivariate stationary time series are analyzed: Peña-Box's methodology in the time domain and Brillinger's methodology in the frequency domain. Two failures of Peña-Box's methodology were found, and their corrections are given. Also the two methods are compared regarding to their capacities to identify the number of latent factors by simulations and an empirical application.

***Key words*:** Multivariate time series, Reduction of dimensionality, Time domain, Frequency domain.

1. Introducción

Dos de los aspectos más importantes en la estadística son las variables y los individuos, y existen métodos estadísticos de reducción tanto para el número de

[a]Docente investigadora. E-mail: hanwenzhang@usantotomas.edu.co

individuos como para el número de variables. Entre de los métodos de reducción de variables se encuentran el método de componentes principales (ACP) y el modelo factorial. Estos métodos tienen como objetivo no solo reducir el número de variables sino hacerlo de forma óptima y así poder representar los datos en una dimensión inferior logrando una interpretación más simple y compacta; siendo de gran importancia y uso frecuente en diferentes áreas de la estadística, incluyendo su extensión al análisis multivariado de series de tiempo.

El aspecto más importante en el análisis de series de tiempo es que los datos son correlacionados, entonces cuando se trabaja con una serie de tiempo multivariada, el número de parámetros en los modelos es muy grande debido a la presencia de correlación tanto dentro de cada serie como entre las series univariadas que la componen. Por consiguiente, si se logra reducir el número de series, se puede simplificar considerablemente la estructura del modelo; además, descubrir los factores latentes que generan las series observadas es muy importante en series económicas, demográficas, meteorológicas, etc.

En el análisis de series de tiempo existen dos enfoques: análisis en el dominio del tiempo y en el dominio de las frecuencias. Las herramientas principales del enfoque en el dominio del tiempo son las funciones de autocorrelación y autocorrelación parcial; mientras que el enfoque en el dominio de las frecuencias asume que la serie es la suma de ondas de diferentes frecuencias y se estudian las frecuencias más importantes. Estos dos enfoques son matemáticamente equivalentes pues la función espectral en el dominio de las frecuencias es simplemente la transformada de Fourier de la función de autocovarianzas en el dominio del tiempo.

En el campo de reducción de dimensionalidad en series de tiempo, también se han usado estos dos enfoques. Dentro del enfoque del dominio de tiempo se encuentran el modelo de index propuesto por Reinsel (1983), el modelo de rango reducido de Ahn & Reinsel (1988) y el modelo de componente escalar de Tiao & Tsay (1989). Uno de los dos métodos considerados en esta investigación, el método de Peña & Box (1987), está restringido a series estacionarias, y tiene como herramienta principal las matrices de autocovarianzas muestrales. Debido a la popularidad de este método, ha sido extendido para series no estacionarias por Peña & Poncela (2006) y para series no lineales por Correal & Peña (2008). Por otro lado, en el enfoque del dominio de las frecuencias, se encuentra el método de Stoffer (1999) para detectar señales comunes en una determinada frecuencia y el método de Brillinger (1981). Este último es similar al método clásico de componentes principales, pues utiliza la varianza recogida por cada componente principal para determinar la reducción del número de variables. El método de Peña y Box y el método de Brillinger tienen básicamente el mismo objetivo, y es natural, al momento de tener la necesidad de reducir la dimensión de una serie multivariada, preguntar cuál de estos dos métodos es mejor o, equivalentemente, cuál método utilizar. Sin embargo, en la literatura estadística, no se ha presentado ningún estudio comparativo de estos métodos[1]. Por consiguiente, el objetivo de esta investigación es llenar este vacío

[1]Tampoco existe, en la literatura, una relación matemática directa entre estos dos métodos y el método clásico de componentes principales.

comparando los dos métodos en términos de su capacidad de identificación del número de factores latentes usando tanto simulaciones como datos reales.

2. Método de Peña y Box

2.1. El modelo básico

El método de Peña & Box (1987) supone que el proceso observable $\{z_t\}$ centrado y de dimensión k es generado por factores $\{y_t\}$ de dimensión r con $r \leq k$ más un término de error ϵ_t, específicamente:

$$z_t = Py_t + \epsilon_t \tag{1}$$

donde P es una matriz de tamaño $k \times r$ cuyo elemento p_{ij} representa el peso del j-ésimo factor sobre la i-ésima serie observada, y ϵ_t es un proceso ruido blanco con matriz de covarianzas Σ_ϵ de rango completo k. Igual que en el modelo factorial clásico, la ecuación (1) no está determinada de manera única, pues para cualquier matriz invertible de tamaño $r \times r$, H, si se definen $P^* = PH$ y $y_t^* = H^{-1}y_t$ se tiene que $z_t = P^*y_t^* + \epsilon_t$, y se logra la misma representación con nuevas matrices P^* y y_t^*. Para evitar este problema de identificación se toma $P'P = I$.

Otro supuesto es que el proceso $\{y_t\}$ obedece a un proceso $VARMA(p_y, q_y)$, donde los polinomios autorregresivos y promedio móvil se denotan por $\Phi_y(B)$ y $\Theta_y(B)$ con $\Phi_y(B) = I - \Phi_{1y}B - \cdots - \Phi_{p_y y}B^{p_y}$ y $\Theta_y(B) = I - \Theta_{1y}B - \cdots - \Theta_{q_y y}B^{q_y}$, entonces el modelo de $\{y_t\}$ es $\Phi_y(B)y_t = \Theta_y(B)a_t$. También se supone que las raíces de los determinantes $|\Phi_y(B)|$ y $|\Theta_y(B)|$ están fuera del círculo unitario; el ruido del proceso $VARMA$ $\{a_t\}$ es un ruido blanco gaussiano con matriz de covarianzas Σ_a, además $E(a_t\epsilon_s') = 0$ para todo t y s.

2.1.1. Factores independientes

En primera instancia se supone que las componentes de y_t son mutuamente independientes y las matrices Φ_{iy} y Θ_{jy} con $i = 1, \ldots, p_y$ y $j = 1, \ldots, q_y$ son diagonales. Sea $\Gamma_z(k)$ la función matricial de covarianzas del proceso observado $\{z_t\}$ y $\Gamma_y(k)$ la función matricial de covarianzas del proceso estocástico $\{y_t\}$. Algunas propiedades de estas matrices son:

- $\Gamma_z(0) = P\Gamma_y(0)P' + \Sigma_\epsilon$
- $\Gamma_z(k) = P\Gamma_y(k)P'$, para $k \geq 1$
- $r(\Gamma_z(k)) = r$, para $k \geq 1$,

donde $r(A)$ es el rango de la matriz A. Bajo el supuesto de que los factores son independientes y la matriz Σ_a es diagonal se tiene que $\Gamma_y(k)$ es diagonal, entonces $\Gamma_z(k)$ es simétrica para $k \geq 1$ y las columnas de la matriz P serán los vectores columnas de $\Gamma_z(k)$ asociado a los r valores propios distintos de cero, para todo $k \geq 1$. Nótese también que si $\Gamma_y(k) = 0$ para todo $k \neq 0$, entonces $\Gamma_z(k) = 0$ para

todo $k \neq 0$, es decir, el proceso $\{z_t\}$ es un proceso ruido blanco, en cuyo caso se tiene el modelo factorial clásico representado por la ecuación (1).

Por otro lado, la identificación del número de factores r se puede realizar observando la matriz de autocorrelaciones parciales $\wp(l)$ (Tiao & Box 1981). Se puede demostrar que el rango de $\wp(l)$ es a lo más r. Existe una tercera forma de identificar el número de factores, que es consecuencia del siguiente teorema (Peña & Box 1987).

Teorema 1. *Si $z_t = Py_t + \epsilon_t$, donde $y_t \sim VARMA(p_y, q_y)$, P es de rango r con $r \leq k$, $\{\epsilon_t\}$ es un proceso ruido blanco, entonces $\{z_t\}$ sigue un proceso $VARMA(p_z, q_z)$ con $p_z = p_y$ y $q_z = max(p_y, q_y)$.*

De este teorema se obtiene otra forma de identificar el número de factores: las matrices Ψ_i en la representación $z_t = \sum_{i=1}^{\infty} \Psi_i \epsilon_{t-i}$ también tienen rango igual a r. En conclusión, existen diferentes formas para encontrar el número de factores r, pero el cálculo de las matrices $\wp(k)$ y Ψ es más complicado que el de las matrices $\Gamma(k)$, y por eso la metodología de esta investigación se basa en el rango de las matrices $\Gamma(k)$.

2.1.2. Transformación canónica

Se puede encontrar una transformación del proceso $\{z_t\}$ de la cual se obtiene una reducción de dimensión del proceso. Esta transformación está dada por la matriz M definida por:

$$M = \begin{bmatrix} P^- \\ B \end{bmatrix} \tag{2}$$

donde P^- es la inversa de Moore-Penrose de la matriz P y B es una matriz con $BP = 0$. Es posible definir B como la matriz formada por los $k - r$ vectores propios ligados a los valores propios nulos de la matriz PP'. Así, podemos obtener las siguientes igualdades:

$$x_t = Mz_t = \begin{bmatrix} P^- z_t \\ Bz_t \end{bmatrix} = \begin{bmatrix} y_t + P^- \epsilon_t \\ BPy_t + B\epsilon_t \end{bmatrix} = \begin{bmatrix} y_t + P^- \epsilon_t \\ B\epsilon_t \end{bmatrix} = \begin{bmatrix} x_{1t} \\ x_{2t} \end{bmatrix} \tag{3}$$

Es decir, los primeros r componentes de x_t son iguales a los factores y_t más un ruido y los restantes $k - r$ componentes son iguales a un ruido blanco. De esta forma, la estructura del proceso z_t queda resumida dentro del proceso x_{1t} que tiene dimensión inferior, y así se logra una reducción de la dimensionalidad del proceso z_t.

2.1.3. Factores dependientes

Cuando los factores generadores $\{y_t\}$ son dependientes, se trabaja con las matrices de coeficientes del modelo $VARMA$ en vez de las matrices de covarianzas, y se tienen las siguientes propiedades:

(1) $\Phi_y(s) = W_s F_s W_s^{-1}$ donde F_s de tamaño $r \times r$ es diagonal y contiene los valores propios de $\Phi_y(s)$ y W_s contiene los vectores propios de $\Phi_y(s)$,

(2) $\Phi_z(s)PW_s = PW_s F_s$. Por ser F_s diagonal, esta contiene los valores propios de $\Phi_z(s)$ y la matriz PW_s de tamaño $k \times r$ contiene los vectores propios de $\Phi_z(s)$.

En conclusión, se puede establecer que los valores propios de $\Phi_y(s)$ son los mismos de $\Phi_z(s)$. Por el otro lado, se considera la primera matriz de coeficientes autorregresivos $\Phi_z(1)$ y suponga que H es la matriz de tamaño $k \times k$ que contiene a todos sus vectores propios. Entonces las r columnas de H conforman la matriz PW_1 y estas corresponden a los vectores propios asociados a los valores propios no nulos de $\Phi_z(1)$.

$$M = \begin{bmatrix} W_1^{-1}P' \\ V' \end{bmatrix} \tag{4}$$

donde V es la matriz de vectores propios de la matriz $PP' = (PW_1)(W_1^{-1}P')$ asociados a los valores propios nulos. Definida M de esta forma podemos obtener una transformación similar al caso de factores independientes:

$$x_t = Mz_t = \begin{bmatrix} W_1^{-1}y_t + W_1^{-1}\epsilon_t \\ V'\epsilon_t \end{bmatrix} = \begin{bmatrix} x_{1t} \\ x_{2t} \end{bmatrix} \tag{5}$$

Análogo al caso de los factores independientes, se obtiene un proceso x_{1t} con dimensión inferior que la del proceso original z_t.

2.2. Implementación práctica

Esta sección introduce dos inconvenientes generados por el método de Peña y Box al momento de su implementación en la práctica. Así mismo, se proponen dos alternativas teóricas para lidiar con estos inconvenientes sin afectar los resultados finales del método.

2.2.1. Identificación del número de factores

El método de Peña y Box sugiere determinar el número de factores, observando los valores propios de las matrices de covarianzas muestrales de las series observadas, $\widehat{\Gamma}_z(h)$. Sin embargo, estas matrices $\widehat{\Gamma}_z(h) = \sum_{t=1}^{n-h}(z_t - \overline{z})(z_{t+h} - \overline{z})'/n$ claramente no son simétricas, y por consiguiente pueden tener valores propios negativos y, más aún, complejos, lo cual dificulta la determinación del número de factores puesto que no es correcto usar las magnitudes de los valores propios para determinar el número de factores como lo ilustra el siguiente ejemplo de simulación. Se simuló una serie z_t de dimensión 4 de acuerdo con la ecuación (1) con

$$P = \begin{bmatrix} 0.51 & 0.30 \\ 0.54 & 0.60 \\ 0.60 & -0.54 \\ 0.30 & -0.51 \end{bmatrix}$$

$\Sigma_\epsilon = I$, $y_t \sim VARMA(1,0)$ con $\Phi = diag(0.3, 0.5)$ y

$$\Sigma_a = \begin{bmatrix} 1.0 & 0.6 \\ 0.6 & 1.0 \end{bmatrix}$$

La matriz de autocovarianzas muestrales de orden 1 está dada por:

$$\widehat{\Gamma}_z(1) = \begin{bmatrix} -0.11 & -0.13 & -0.23 & -0.01 \\ 0.02 & -0.1 & 0.26 & -0.01 \\ 0.16 & 0.11 & -0.09 & -0.27 \\ 0.02 & 0.16 & -0.26 & -0.18 \end{bmatrix} \tag{6}$$

Al calcular los valores propios de la matriz $\widehat{\Gamma}_z(1)$, se encontró que estos son -0.46, $0.043 \pm 0.15i$ y -0.11. De tal forma que las magnitudes son: 0.46, 0.16, 0.16 y 0.11, respectivamente. Los anteriores valores llevan a la conclusión que el número de factores es 1 (considerando a 0.16 como pequeño) o 3 (considerando a 0.16 significativo), lo cual es erróneo de cualquier forma puesto que el número correcto de factores es 2. Afortunadamente la anterior inconveniencia se puede resolver modificando la metodología de Peña y Box, de la forma sugerida en esta investigación, teniendo en cuenta que el número de factores es igual al rango de las matrices $\widehat{\Gamma}_z(h)$, y este es igual al número de valores singulares no nulos. En primer lugar, es necesario recordar la definición de los valores singulares de una matriz cuadrada (Jiménez 2004).

Definición 1. Los valores singulares de una matriz real A de tamaño $n \times n$ son las raíces cuadradas de los valores propios asociados a la matriz simétrica $A'A$ (listados con sus multiplicidades algebraicas). Estos valores se denotan por $\sigma_1, \sigma_2, \ldots, \sigma_n$, y se colocan en orden decreciente:

$$\sigma_1 \geq \sigma_2 \geq \cdots \sigma_n \geq 0$$

Nótese que por definición, a diferencia de los valores propios, los valores singulares de una matriz son siempre reales positivos. El siguiente teorema establece la relación entre el rango de una matriz cuadrada y sus valores singulares.

Teorema 2. *(Descomposición en Valores Singulares) Sea A una matriz real de tamaño $n \times n$ con rango r, $r < n$. Entonces existen matrices ortogonales U y V de tamaño $n \times n$, tales que*

$$A = USV^t \tag{7}$$

donde S es la matriz particionada de tamaño $m \times n$, dada por

$$S = \begin{bmatrix} D_r & 0 \\ 0 & 0 \end{bmatrix}$$

con

$$D_r = \begin{bmatrix} \sigma_1 & \cdots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \cdots & \sigma_r \end{bmatrix}$$

donde los σ_i con $i = 1, \cdots, r$ son los valores singulares no nulos de A.

En el teorema anterior, como V^t es una matriz invertible, entonces A y US tienen el mismo rango; usando el mismo argumento y el hecho que U también es invertible, por consiguiente A y S tienen el mismo rango, y por construcción el rango de S es r, el número de valores singulares no nulos de A. Por lo tanto, se concluye que el rango de una matriz cuadrada es igual al número de valores singulares no nulo, y por consiguiente es más conveniente usar los valores singulares de $\widehat{\Gamma}_z(h)$, para determinar el rango, que usar los valores propios. Retomando el ejemplo formulado al principio de esta sección, donde se vio que el uso de valores propios no es siempre adecuado, se tiene que con la propuesta de usar valores singulares de la matriz $\widehat{\Gamma}(1)$ dada por (6), se llega a que estos son 0.51, 0.34, 0.16, 0.05. De esta manera, si consideramos los dos primeros valores singulares los más importantes, se puede llegar a la conclusión correcta de 2 factores; mientras que con el uso de los valores propios, el número de factores identificado será 1 o 3, mas nunca el valor correcto, 2. Ahora, observar directamente los valores de los valores singulares y decidir a simple vista que los dos primeros son importantes no es un método riguroso para identificar r, pues se necesitaría cierto tipo de prueba estadística para decidir sobre si estos son significativos o no; sin embargo, en la literatura estadística no existe tal prueba basada en valores singulares, y será un tema abierto para futuras investigaciones.

2.2.2. Estimación del modelo

Según lo mencionado en la sección 2.1, solo se puede encontrar la matriz P cuando se asume que las componentes de y_t son independientes para cada t. En este caso, existen dos formas de hacerlo:

1. Usando los vectores propios de las matrices $\widehat{\Gamma}_z(k)$.

2. Usando los vectores propios de las matrices de coeficientes $\Phi_z(k)$ del modelo $VARMA$ ajustado a la serie z_t.

En ambos casos los vectores propios deben ser ortonormales para garantizar que $P'P = I$, de tal manera que el modelo quede determinado de forma única. Sin embargo, existen dificultades al momento de ejecutar cualquiera de estas dos alternativas en la práctica.

En el caso de usar $\widehat{\Gamma}_z(k)$, como se indica en la sección 2.1.1, bajo el supuesto (1), las matrices de autocovarianzas $\Gamma_z(k)$ son simétricas; pero en la práctica la estimación de esta, $\widehat{\Gamma}_z(k) = \frac{1}{n}\sum_{t=1}^{n-k}(z_{t+k} - \overline{z})(z_t - \overline{z})'$ puede no ser simétrica, y no siempre es posible encontrar vectores propios ortonormales de una matriz no simétrica y, por lo tanto, no siempre se puede hallar la matriz P. El mismo problema surge cuando se usan los vectores propios de $\Phi_z(k)$ para construir P. Sin embargo, este problema se puede resolver usando la representación del modelo de estados y el filtro de Kalman como lo sugieren por Peña & Poncela (2006), método que se describe a continuación.

Siguiendo a Brockwell & Davis (1996), un modelo de estados está dado por las siguientes ecuaciones:

$$z_t = G_t x_t + w_t \tag{8a}$$
$$x_{t+1} = F_t x_t + R_t v_t \tag{8b}$$

donde $\{z_t\}$ es de dimensión $w \times 1$, que se expresa como una función lineal del vector de estado x_t de dimensión $v \times 1$ mediante la matriz G_t de tamaño $w \times v$, $\{w_t\} \sim RB(0, R_t)$, $\{v_t\} \sim RB(0, Q_t)$, $E(w_t v_s') = 0$ para todo t y s. Además, el vector de estado inicial x_1 es incorrelacionado con los errores v_t y w_t para todo $t \geq 1$. La primera ecuación se llama ecuación de observación y la segunda, ecuación de estado o de transición.

La presentación del modelo de Peña y Box en forma de un modelo de estados es como sigue:

$$z_t = \widetilde{P} x_t + \epsilon_t \tag{9a}$$
$$x_t = F x_{t-1} + R v_t \tag{9b}$$

donde $x_t = [y_t', y_{t-1}', \ldots, y_{t-l+1}']'$ es de tamaño $rl \times 1$ con $l = max(p_y, q_y + 1)$, donde p_y y q_y denotan los órdenes del modelo $VARMA$ de $\{y_t\}$ y r es el número de factores. $\widetilde{P}$ es una matriz de tamaño $r \times rl$ que contiene la matriz P y matrices nulas,

$$F = \begin{bmatrix} \Phi_{1y} & \Phi_{2y} & \cdots & \Phi_{l-1,y} & \Phi_{ly} \\ I_r & 0 & \cdots & 0 & 0 \\ 0 & I_r & \cdots & 0 & 0 \\ \vdots & \vdots & \ddots & \vdots & \vdots \\ 0 & 0 & \cdots & I_r & 0 \end{bmatrix}, \text{ y } \; R = \begin{bmatrix} I_r & \Theta_{1y} & \cdots & \Theta_{l-1,y} \\ 0 & 0 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & 0 \end{bmatrix} \tag{10}$$

Cabe resaltar que Peña & Poncela (2006) proponen el modelo de estados para la estimación puesto que su trabajo está enmarcado dentro del contexto de factores no estacionarios. Según lo señalado en este trabajo, en el caso de factores estacionarios, también es adecuado utilizar la misma metodología. Por otro lado, para estimar el modelo (9), es necesario conocer los órdenes p_y y q_y. Para esto se hace uso del teorema 1, donde se establece que, bajo los supuestos en (1), se tiene que $p_z = p_y$ y $q_z = \text{máx}(p_y, q_y)$. Entonces el procedimiento será (i) obtener los valores adecuados para p_z y q_z usando las series observadas $\{z_t\}$; (ii) usar las relaciones antes mencionadas para encontrar valores candidatos de p_y y q_y; por ejemplo, si $p_z = q_z = 2$, entonces se tiene que $p_y = p_z$ y los valores candidatos para q_y serán 2, 1 y 0; (iii) usar criterios adecuados (por ejemplo, los criterios de información) para escoger los valores finales de p_y y q_y.

3. Método de Brillinger

3.1. Método teórico

El método de Brillinger (1981) hace supuestos inversos a los supuestos del método de Peña y Box, en el sentido de que este último expresa el proceso observado en términos de los factores generadores, pero el método de Brillinger expresa los factores generadores como combinación del proceso observado. Suponga que el proceso observado X_t es de tamaño $k \times 1$, y los factores generadores, que Brillinger llama los componentes principales, ζ_t son de tamaño $r \times 1$ con $r \leq k$. Brillinger plantea la siguiente relación entre estos dos procesos:

$$\zeta_t = \sum_u b_{t-u} X_u, \tag{11}$$

donde las matrices b_t son de tamaño $r \times k$. Dada esta ecuación, la serie observada X_t se puede estimar mediante los ζ_t de la siguientes forma:

$$\widehat{X}_t = \sum_u c_{t-u} \zeta_u \tag{12}$$

Entonces el problema de encontrar los componentes principales ζ_t se convierte en buscar las matrices b_u y c_u tales que $\widehat{X}_t$ sea cercano a X_t. El siguiente teorema provee la solución para este problema:

Teorema 3. *(Brillinger, 1981) Sea $\{X_t\}$ de dimensión $k \times 1$, un proceso estacionario en sentido débil con media cero, función matricial de autocovarianzas absolutamente sumable y función matricial de densidad espectral $f(\omega)$. Entonces las matrices b_t y c_t que minimizan $E\Big\{\big[\overline{X_t - \sum_u c_{t-u}\zeta_u}\big]^t \big[X_t - \sum_u c_{t-u}\zeta_u\big]\Big\}$ están dadas por*

$$b_t = \frac{1}{2\pi} \int_0^{2\pi} B(\alpha) e^{it\alpha} d\alpha \tag{13}$$

y

$$c_t = \frac{1}{2\pi} \int_0^{2\pi} C(\alpha) e^{it\alpha} d\alpha \tag{14}$$

donde

$$B(\lambda) = \begin{bmatrix} \overline{V_1(\lambda)^t} \\ \vdots \\ \overline{V_r(\lambda)^t} \end{bmatrix} \tag{15}$$

y

$$C(\lambda) = \overline{B(\lambda)^t} = \begin{bmatrix} V_1(\lambda) & \cdots & V_r(\lambda) \end{bmatrix} \tag{16}$$

Aquí, $V_j(\lambda)$ denota el j-ésimo vector propio de la matriz de densidad espectral $f(\lambda)$ ligado al valor propio $\mu_j(\lambda)$, para $j = 1, \ldots, r$ y cada $\lambda \in [0, 2\pi]$.

Para encontrar el j-ésimo componente principal $\zeta_j(t)$ se debe escribir X_t en términos de la representación de Cramér dado por

$$X_t = \int_0^{2\pi} e^{i\lambda t} dZ_X(\lambda) \tag{17}$$

donde

$$2\pi Z_X(\lambda) = \int_0^{2\lambda} d_X(\alpha) d\alpha \tag{18}$$

y

$$d_X(\lambda) = \sum_{t=-\infty}^{\infty} X(t) e^{-i\lambda t} \tag{19}$$

Se obtiene que $\zeta(t)$ está dado por

$$\zeta(t) = \int_0^{2\pi} \overline{B(\lambda)^t} e^{i\lambda t} dZ_X(\lambda) \tag{20}$$

Los componentes principales se pueden definir de otra manera como indica el siguiente teorema:

Teorema 4. *Brillinger (1981) suponga que se satisfacen las condiciones del teorema 3, entonces la componente j-ésima de ζ_t, $\zeta_j(t)$, está dada por*

$$\int_0^{2\pi} B_j(\lambda) e^{i\lambda t} dZ_X(\lambda) \tag{21}$$

$j = 1, \ldots, r$ y $t \in \mathbb{Z}$, donde $B_j(\lambda)$ es de tamaño $1 \times r$ que satisface $B_j(\lambda)\overline{B_j(\lambda)^t} = 1$, $\zeta_j(t)$ tiene varianza máxima y coherencia 0 con $\zeta_k(t)$ con $k < j$. La varianza máxima alcanzada por $\zeta_j(t)$ es

$$\int_0^{2\pi} \mu_j(\alpha) d\alpha \tag{22}$$

En la práctica, se obtiene la estimación de $f(\lambda)$, y sus valores y vectores propios y finalmente los componentes principales. La forma de calcular los componentes principales se presenta en la siguiente sección.

3.2. Implementación práctica

Análogo al método clásico de componentes principales, se puede escoger el número de componentes usando la cantidad de varianza recogida en cada componente definida en (22). En la práctica, esta integral se puede evaluar usando la definición de las integrales de Riemann-Stieltjes, esto es,

$$\int_0^{2\pi} \mu_j(\alpha) d\alpha \approx \sum_{i=1}^{n-1} \mu_j \left(\frac{t_{i+1} + t_i}{2} \right) (t_{i+1} - t_i) \tag{23}$$

donde $(t_i, t_{i+1}]$, con $i = 1, \ldots, n-1$, es una partición regular del intervalo $[0, 2\pi]$. De manera análoga se evalúa las expresiones en las ecuaciones (18), (19) y (20).

Nótese que para calcular estas varianzas recogidas dadas por (23), es necesario calcular los valores propios de las matrices de densidad espectral; la estimación de estas matrices son matrices hermitianas, por lo tanto tienen valores propios reales, y finalmente las varianzas recogidas por cada componente también serán números reales. Esta es una ventaja frente al método de Peña y Box, pero como se verá en las simulaciones de la sección 4, si se modifica la metodología de Peña y Box siguiendo la sugerencia de la sección anterior, los dos métodos tendrán la misma capacidad para identificar el mismo número de factores.

4. Comparación de los métodos

Para realizar la comparación de los dos métodos, recurrimos a simulaciones de procesos estacionarios debido a la complejidad teórica que se presenta al intentar relacionar valores y vectores propios de las matrices de autocovarianzas con los de las matrices de densidad espectral. La idea de las simulaciones es crear unas series observadas z_t que seas generadas por un número conocido, r, de series no observables y_t, y se aplican los dos métodos a fin de comparar sus capacidades para identificar el número r.

Sin embargo, el método de Peña y Box y el método de Brillinger tienen supuestos diferentes con respecto a la relación existente entre las series observadas y las no observables. Y debido a esa diferencia, la forma de crear las series observadas no puede ser ninguno de estos dos métodos, pues de lo contrario se estaría dando ventaja a uno de los dos. Por lo tanto, se decide crear las series observadas a partir de cierta estructura de matriz de autocovarianzas de rango inferior a la dimensión, lo cual es una consecuencia del método de Peña y Box, y a la vez no está restringido a un modelo específico, pues, dada una función de autocovarianzas, pueden existir varios modelos que tienen a esa función como su función de autocovarianzas.

En esta investigación se derivó el siguiente teorema que permite crear una serie multivariada a partir de una estructura de matriz de autocovarianzas, y extiende los resultados del teorema 1.5.1 de Brockwell & Davis (1991, Pág.27).

Teorema 5. *Una sucesión de matrices, $\Gamma(k)$, $k = 0, 1, 2, \ldots$, de tamaño $p \times p$, es una función matricial de autocovarianzas de un proceso estocástico multivariado $\{X_t\}$, con X_t de dimensión p, si para todo n entero positivo, la matriz*

$$\Sigma = \begin{bmatrix} \Gamma(0) & \Gamma(1) & \cdots & \Gamma(n-1) \\ \Gamma(1)' & \Gamma(0) & \cdots & \Gamma(n-2) \\ \vdots & \vdots & \ddots & \vdots \\ \Gamma(n-1)' & \Gamma(n-2)' & \cdots & \Gamma(0) \end{bmatrix} \tag{24}$$

de tamaño $np \times np$ es definida no negativa.

Demostración. Dado n cualquier entero positivo, sea $t = (t_1, t_2, \ldots, t_{np})' \in \mathbb{Z}^{np}$ con $t_1 < t_2 \cdots < t_{np}$, y sea F_t la función de distribución con función característica

dada por

$$\phi_t(u) = \exp\{-u'\Sigma u/2\} \tag{25}$$

con $u = (u_1, u_2, \ldots, u_{np})' \in \mathbb{R}^{np}$. Como la matriz Σ es definida no negativa, $\phi_t(u)$ es la función característica de una distribución $N_{np}(0, \Sigma)$. Entonces F_t es consistente y, por el teorema de Kolmogorov, existe un proceso estocástico $\{X_t\}$ con F_t como su función de distribución finito-dimensional y $\phi_t(u)$ como su función característica.

Por otro lado, si se toma cualquier vector del proceso $\{X_t\}$, lo podemos escribir como

$$(X_{11}, X_{21}, \ldots, X_{p1}, X_{12}, X_{22}, \ldots, X_{p2}, \ldots, X_{1n}, X_{2n}, \ldots, X_{pn})'$$

y se toma el operador *vec* inversa para convertirlo en una matriz de tamaño $p \times n$, vamos a obtener la siguiente matriz

$$\mathbb{X} = \begin{bmatrix} X_{11} & X_{12} & \cdots & X_{1n} \\ X_{21} & X_{22} & \cdots & X_{2n} \\ \vdots & \vdots & \ddots & \vdots \\ X_{p1} & X_{p2} & \cdots & X_{pn} \end{bmatrix} = [X_1, X_2, \ldots, X_n] \tag{26}$$

donde X_i con $i = 1, \ldots, n$ son vectores de dimensión p. Y, por construcción, tenemos que $Cov(X_{t+h}, X_t) = \Gamma(h)$, $t = 1, \ldots, n-h$ y $h = 0, 1, \ldots, n-1$. A $\{X_1, \ldots, X_n\}$ se puede considerar como un proceso estacionario de longitud n. □

A continuación se consideran casos cuando k series observadas son generadas por r series no observadas con $r \leq k$. Para el método de Peña y Box, se calcula los porcentajes acumulados de los valores singulares de $\widehat{\Gamma}_z(h)$. Como los valores singulares siempre son reales positivos, y el número de valores singulares no nulos es el rango de $\widehat{\Gamma}_z(h)$, entonces se pueden usar estos porcentajes acumulados para determinar el rango de $\widehat{\Gamma}_z(h)$. Por ejemplo, si estos porcentajes acumulados de $\widehat{\Gamma}_z(h)$ son $(0.65, 0.92, 0.97, 0.99, 1)$, se puede decir que el rango de $\Gamma(h)$ es 2. Por otro lado, para el método de Brillinger, se calcula el porcentaje acumulado de varianza. En esta investigación se consideraron los siguientes casos:

- Tres series observadas generadas por uno, dos y tres factores.
- Cinco series observadas generadas por uno, dos, tres, cuatro y cinco factores.

Para cada uno de los anteriores casos, las simulaciones se realizaron con 1000 réplicas, y las series simuladas son de tamaño 100. Para el método de Peña y Box, se calculó en cada iteración los porcentajes acumulados de los valores singulares de $\widehat{\Gamma}(h)$ con $h = 0, 1, 2, 3$; para el método de Brillinger, se calculó en cada iteración el porcentaje acumulado de varianzas. Si la serie de dimensión k es generada por r factores, un indicio de que el método de Peña y Box es adecuado es el hecho de que el r-ésimo porcentaje acumulado de los valores singulares toma un valor cercano a 100 %; el criterio para evaluar el método de Brillinger es análogo, pero aplicado al

porcentaje acumulado de varianzas. Por esta razón, los resultados de las simulaciones corresponden al número de veces de las 1000 réplicas donde los porcentajes acumulados se ubican en rangos determinados. Por limitación de espacio, aquí se presentan únicamente los resultados de los siguientes tres casos:

Modelo 1: 3 series generadas por 2 factores, con estructura de covarianzas de los factores dada por $\Gamma(0) = diag(1.09, 1.01)$, $\Gamma(1) = diag(0.3, -0.1)$ y $\Gamma(h) = 0$ para $h > 1$; los resultados se encuentran en la tabla 1.

Modelo 2: 3 series generadas por 3 factores, con estructura de covarianzas de los factores dada por
$\Gamma(0) = diag(1.39, 2.79, 2.11)$,
$\Gamma(1) = diag(-0.28, 0.155, 1.46)$, $\Gamma(2) = diag(-0.61, 2.22, 0.68)$ y $\Gamma(h) = \Gamma(h-1)diag(-0.3, 0.1, 0.9) + \Gamma(h-2)diag(-0.5, -0.8, -0.3)$ para $h > 2$; los resultados se encuentran en la tabla 2.

Modelo 3: 5 series generadas por 3 factores, con estructura de covarianzas de los factores dada por
$\Gamma(0) = diag(1.39, 2.79, 2.11)$, $\Gamma(1) = diag(-0.28, 0.155, 1.46)$,
$\Gamma(2) = diag(-0.61, 2.22, 0.68)$ y $\Gamma(h) = \Gamma(h-1)diag(-0.3, 0.1, 0.9) + \Gamma(h-2)diag(-0.5, -0.8, -0.3)$ para $h > 2$; los resultados se encuentran en la tabla 3.

Tabla 1: El número de iteraciones donde el porcentaje acumulado de valores singulares (para el método Peña y Box) y de varianzas (para el método de Brillinger) se ubica en determinados rangos para el modelo 1.

		$> 90\,\%$	$80 \sim 90$	$70 \sim 80$	$60 \sim 70$	$< 60\,\%$
$\widehat{\Gamma}(0)$	1	0	443	536	21	0
	2	1000	0	0	0	0
$\widehat{\Gamma}(1)$	1	245	312	204	158	81
	2	1000	0	0	0	0
$\widehat{\Gamma}(2)$	1	384	315	152	143	6
	2	1000	0	0	0	0
$\widehat{\Gamma}(3)$	1	352	210	241	151	46
	2	1000	0	0	0	0
Brillinger	1	0	519	472	9	0
	2	1000	0	0	0	0

De la tabla 1, se observa en primer lugar que para las matrices $\widehat{\Gamma}(h)$ con $h = 0, 1, 2, 3$, el número de veces que el segundo porcentaje acumulado de los valores singulares es mayor a $90\,\%$ es igual al número de réplicas 1000; por otro lado, para el método de Brillinger, el porcentaje de la varianza acumulada del segundo componente también es mayor a $90\,\%$ para cada una de las 1000 réplicas. Lo anterior indica que los dos métodos tienen la misma capacidad para identificar el número de factores. Nótese que, dentro del método de Peña y Box, con el uso de $\widehat{\Gamma}(0)$, en ninguna de las 1000 iteraciones, el primer valor singular pesa más

del 90 %, mientras que con $\widehat{\Gamma}(h)$ con $h > 0$, hay más posibilidad de subestimar el número de factores. Con respecto a la tabla 2, nótese que en el modelo 2 el número

Tabla 2: El número de iteraciones donde el porcentaje acumulado de valores singulares (para el método Peña y Box) y de varianzas (para el método de Brillinger) se ubica en determinados rangos para el modelo 2.

		$> 90\,\%$	$80 \sim 90$	$70 \sim 80$	$60 \sim 70$	$< 60\,\%$
$\Gamma(0)$	1	0	0	23	155	822
	2	0	105	809	81	5
	3	1000	0	0	0	0
$\Gamma(1)$	1	0	21	272	409	298
	2	481	507	12	0	0
	3	1000	0	0	0	0
$\Gamma(2)$	1	0	163	345	291	201
	2	845	133	21	1	0
	3	1000	0	0	0	0
$\Gamma(3)$	1	10	156	238	271	325
	2	535	428	37	0	0
	3	1000	0	0	0	0
Brillinger	1	0	0	169	613	218
	2	8	842	143	7	0
	3	1000	0	0	0	0

de factores es igual al número de series observadas, esto es, 3; entonces un buen método debe tener poca posibilidad de subestimar el número de factores. Con el uso de $\widehat{\Gamma}(0)$ del método de Peña y Box, en ninguna de las 1000 réplicas el segundo porcentaje acumula más del 90 %; también el número de veces que acumula entre 80 y 90 % es bajo (105 de las 1000 réplicas). Mientras que con el uso de $\widehat{\Gamma}(h)$ con $h > 0$, hay más iteraciones donde se identifica erróneamente 2 factores. Por otro lado, con el método de Brillinger, el número de veces que el segundo porcentaje acumula más del 90 % también es muy bajo (8 de las 1000 réplicas). En conclusión, el desempeño de Peña y Box usando $\widehat{\Gamma}(0)$ es levemente mejor que el desempeño del método de Brillinger, y el desempeño de este es mejor que el de $\widehat{\Gamma}(h)$ con $h > 0$. Con respecto a la tabla 3, para los resultados de los métodos aplicados a datos simulados a partir del modelo 3, se observa, en primer lugar, que el desempeño del método de Brillinger es, de nuevo, similar al de $\widehat{\Gamma}(0)$ del método de Peña y Box. Por otro lado, el desempeño de $\widehat{\Gamma}(1)$ es mejor que el de $\widehat{\Gamma}(0)$ puesto que el número de veces que el primer porcentaje de valores singulares de $\widehat{\Gamma}(1)$ sea mayor de 90 % (149 de las 1000 réplicas) es menor que el de $\widehat{\Gamma}(0)$ (484 de las 1000 iteraciones). Es decir, con el uso de $\widehat{\Gamma}(1)$ hay menos posibilidad de identificar erróneamente un factor que con el uso de $\widehat{\Gamma}(0)$. De la misma manera, se puede ver que con el uso de $\widehat{\Gamma}(1)$ hay menos posibilidad de identificar erróneamente 2 factores que con el uso de $\widehat{\Gamma}(0)$.

Tabla 3: El número de iteraciones donde el porcentaje acumulado de valores singulares (para el método Peña y Box) y de varianzas (para el método de Brillinger) se ubica en determinados rangos para el modelo 3.

		$> 90\,\%$	$80 \sim 90$	$70 \sim 80$	$60 \sim 70$	$< 60\,\%$
$\widehat{\Gamma}(0)$	1	484	364	144	8	0
	2	955	37	8	0	0
	3	1000	0	0	0	0
	4	1000	0	0	0	0
$\widehat{\Gamma}(1)$	1	149	332	341	95	83
	2	815	164	21	0	0
	3	1000	0	0	0	0
	4	1000	0	0	0	0
$\widehat{\Gamma}(2)$	1	933	54	13	0	0
	2	1000	0	0	0	0
	3	1000	0	0	0	0
	4	1000	0	0	0	0
$\widehat{\Gamma}(3)$	1	394	332	214	60	0
	2	958	42	0	0	0
	3	1000	0	0	0	0
	4	1000	0	0	0	0
Brillinger	1	495	356	59	9	0
	2	969	31	0	0	0
	3	1000	0	0	0	0
	4	1000	0	0	0	0

Para las simulaciones cuyos resultados por limitación de espacio, no se presentan aquí, se muestran comportamientos análogos a los presentados anteriormente. En síntesis, se observan los siguientes comportamientos:

- Para el caso cuando el número de factores, r, es estrictamente menor que el número de series observadas, k:
  1. El uso de $\widehat{\Gamma}_z(0)$ conduce al mismo resultado que el método de Brillinger.
  2. El uso de $\widehat{\Gamma}_z(1)$ es mejor que el de $\widehat{\Gamma}_z(h)$ para $h > 1$.
  3. El uso de $\widehat{\Gamma}_z(0)$ y $\widehat{\Gamma}_z(1)$ conduce a los mismos resultados.
- Para el caso cuando el número de factores, r, es igual al número de series observadas, k: el uso de $\widehat{\Gamma}_z(0)$ es levemente mejor que el de $\widehat{\Gamma}_z(h)$ para $h > 1$ y que el método de Brillinger, pues con $\widehat{\Gamma}_z(h)$ o el método de Brillinger hay más posibilidad de subestimar el número de factores.

5. Una aplicación empírica

En esta sección se aplican los dos métodos a 9 variables económicos de Colombia analizadas en Melo et al. (2001): situación económica actual de la industria, volumen actual de pedidos por atender de la industria, índice de producción real de la

industria manufacturera sin trilla de café, índice de empleo de obreros de la industria, producción de cemento, demanda de energía más consumo de gas residencial e industrial, importaciones reales exceptuando las de bienes de capital y duraderos, cartera neta real en moneda legal y saldo de efectivo en términos reales. En Melo et al. (2001) se analizaron estas series desde enero de 1980 hasta agosto de 2001; en esta aplicación se amplió el periodo de observación hasta diciembre de 2005. Los datos se desestacionalizaron usando métodos de suavizamiento y están libres de datos atípicos e intervenciones (Martínez 2007). La gráfica de estas nueve series económicas se encuentra en la figura 1.

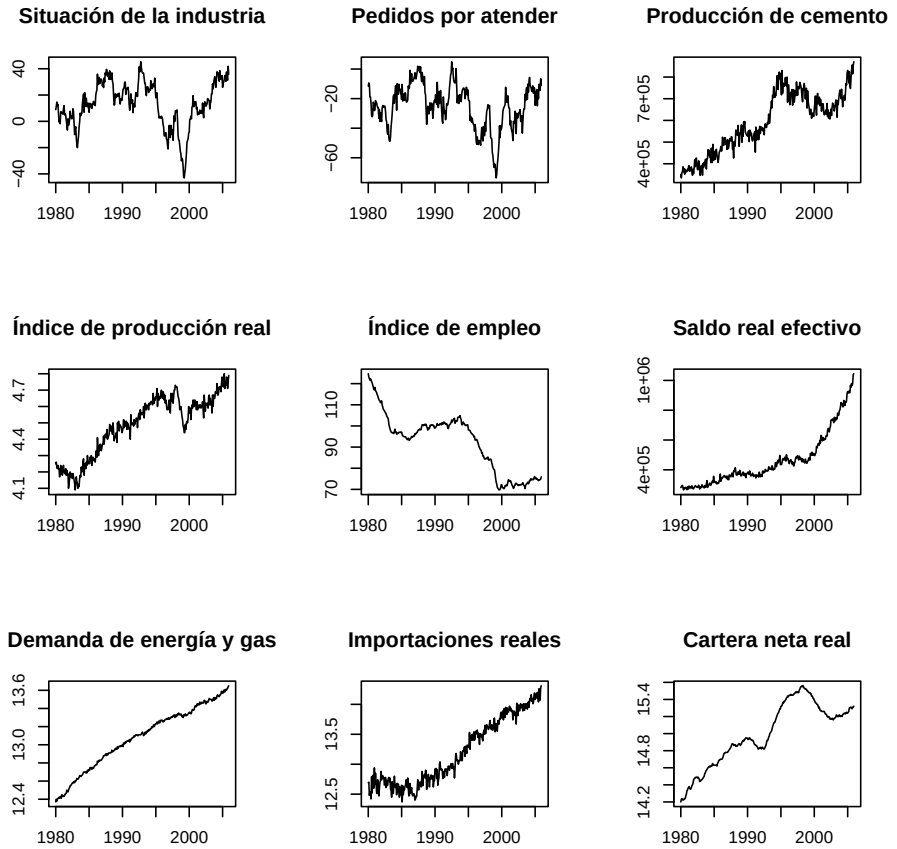

Figura 1: Series económicas de Colombia, concernientes a la aplicación de la sección 5, sin diferenciar.

Los correlogramas y los p-valores de la prueba de raíz unitaria de Dickey y Fuller sugieren que cada una de las nueve series tiene una raíz unitaria, y con la aplicación de la diferenciación de orden 1 se obtiene la estacionariedad, de donde se encuentra que cada una de las nueve series son integradas de orden 1 ($I(1)$). Un

análisis de cointegración muestra que las series son cointegradas de orden $CI(1,1)$. Aplicando la prueba de Johansen (1991) con la estadística de prueba $\lambda_{\text{máx}}$ para hallar el rango de integración s, se obtuvo que, con un nivel de significación del 5 %,

- para la hipótesis nula de $H_0 : s \leq 6$, $\lambda_{\text{máx}} = 11.27$, el valor crítico es 21.07, de donde no se rechaza H_0;
- para la hipótesis nula de $H_0 : s \leq 5$, $\lambda_{\text{máx}} = 28.77$, el valor crítico es 27.14, de donde se rechaza H_0.

Por lo tanto, se concluye que el rango de cointegración es 6, y por consiguiente las series comparten en total 3 tendencias estocásticas comunes. Nótese que un interesante estudio es aplicar el método de Peña & Poncela (2006) para series no estacionarias y comparar con el resultado del anterior análisis de cointegración.

A continuación se aplican los métodos de Peña y Box y el de Brillinger. Para tal fin las series deben ser estacionarias; por lo tanto, en adelante se trabajará con las series diferenciadas de orden 1 las cuales se presentan en la figura 2.

5.1. Identificación del número de factores

Para identificar el número de factores comunes, se calculan los valores propios de las matrices de autocovarianzas muestrales siguiendo lo sugerido por Peña y Box. En la tabla 4 se muestran dichos valores; se observa que los dos primeros valores propios son los más importantes, lo cual indica que las nueve series se pueden reducir a dos; sin embargo, los primeros dos valores propios de las matrices de covarianzas muestrales de rezago 2 y 4 son complejos, por eso se examinan los valores singulares de la tabla 5. Estos conducen a la misma conclusión de dos factores. Se puede observar cómo el uso de los valores singulares corrige las inconveniencias que se presentan al usar los valores propios. Por el lado del método de Brillinger, la varianza acumulada de la primera componente pesa un 81 %, y los dos primeros componentes acumulan el 100 % de la varianza, de donde se concluye que también se identifican 2 factores.

Tabla 4: Valores propios de las matrices de autocovarianzas muestrales de las series económicas de la aplicación empírica para los rezagos $h = 0, 1, 2, 3, 4$.

$\Gamma(0)$	$\Gamma(1)$	$\Gamma(2)$	$\Gamma(3)$	$\Gamma(4)$
1.28E9	-5.8E8	-2E7+1.4E7i	8.05E7	-4.5E7+1.6E7i
2.99E8	-8.7E7	-2E7-1.4E7i	6.78E7	-4.5E7-1.6E7i
436	-8.2	-0.82+0.86i	3.22	2.23
8.52	-2.1	-0.82-0.86i	1.02	-0.98
0.48	0.095	0.08	0.08	-0.09
0.01	0.00	0.01	0.00	-0.01
0.00	0.00	0.00	0.01	0.00
0.00	0.00	0.00	0.00	0.00
0.00	0.00	0.00	0.00	0.00

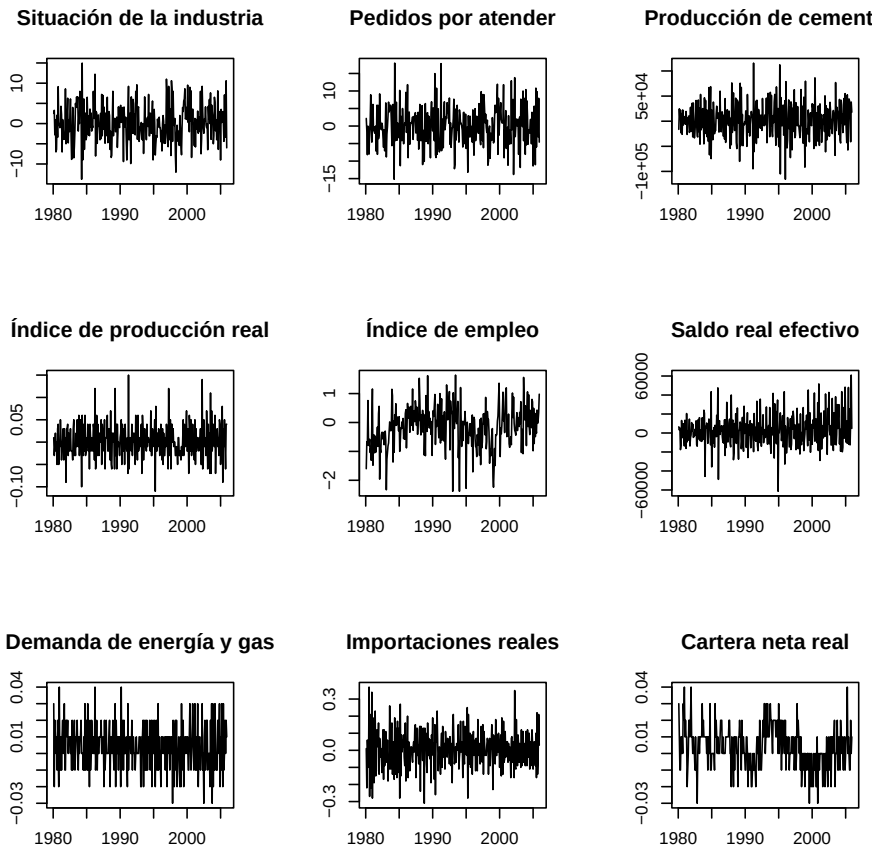

FIGURA 2: Series económicas de Colombia, concernientes a la aplicación de la sección 5, después de diferenciar una vez.

TABLA 5: Valores singulares de las matrices de autocovarianzas muestrales de las series económicas de la aplicación empírica para los rezagos $h = 0, 1, 2, 3, 4$.

$\Gamma(0)$	$\Gamma(1)$	$\Gamma(2)$	$\Gamma(3)$	$\Gamma(4)$
1.28E9	6.02E8	5.52E7	9.37E7	1.08E8
2.99E8	8.56E7	1.14E7	5.8E7	2.09E7
436	8.75	3.71	5.27	3.27
8.52	2.01	0.67	0.636	0.95
0.48	0.10	0.05	0.08	0.09
0.01	0.00	0.00	0.00	0.01
0.00	0.00	0.00	0.00	0.00
0.00	0.00	0.00	0.00	0.00
0.00	0.00	0.00	0.00	0.00

5.2. Extracción de los factores

Para estimar el modelo de Peña y Box con el modelo de estados, se necesitan los valores de p_y y q_y. Siguiendo lo indicado en la sección 2.2.2, primero se identifican los valores de p_z y q_z. Para esto se examinaron las matrices de correlación muestral $\widehat{\rho}(k)$ cuyo i,j-ésimo elemento está dado por

$$\frac{\sum_{t=1}^{n-k}(z_{i,t}-\overline{z}_i)(z_{j,t+k}-\overline{z}_j)}{\left[\sum_{t=1}^{n}(z_{i,t}-\overline{z}_i)^2\sum_{t=1}^{n}(z_{j,t}-\overline{z}_j)^2\right]^{1/2}} \tag{27}$$

y las estimaciones de matrices de autocorrelación parcial

$$\wp(k)=\begin{cases}\Gamma'(1)[\Gamma(0)]^{-1} & s=1\\ \Gamma'(k)-c'(k)[A(k)]^{-1}b(k)\Gamma(0)-b'(k)[A(k)]^{-1}b(k)^{-1} & s>1\end{cases} \tag{28}$$

donde

$$A(k)=\begin{bmatrix}\Gamma(0) & \Gamma'(1) & \cdots & \Gamma'(k-2)\\ \Gamma(1) & \Gamma(0) & \cdots & \Gamma'(k-3)\\ \vdots & \vdots & & \vdots\\ \Gamma(k-2) & \Gamma(k-3) & \cdots & \Gamma(0)\end{bmatrix}$$

$$b(k)=\begin{bmatrix}\Gamma'(k-1)\\ \Gamma'(k-2)\\ \vdots\\ \Gamma'(1)\end{bmatrix}$$

y

$$c(k)=\begin{bmatrix}\Gamma(1)\\ \Gamma(2)\\ \vdots\\ \Gamma(k-1)\end{bmatrix}$$

Ninguna de las matrices $\widehat{\rho}(k)$ y $\widehat{\wp}(k)$ muestra tendencia de extinguirse a medida que crece k, lo cual indica que el modelo apropiado para z_t no es VAR o VMA puro, sino un modelo $VARMA$ mixto; pues para un proceso $VAR(p)$ las matrices $\wp(k)$ son iguales a 0 para $k>p$, mientras que para un proceso $VMA(q)$ las matrices de correlación son nulas para rezagos mayores que q, (Wei 2006, pg. 402). Debido a la dificultad que tiene identificar y estimar un modelo $VARMA$ para un número grande de series, se procede a estimar el modelo de estados (9) y (10) directamente para diferentes valores de p_y y q_y. En la tabla 6 se encuentran los valores de algunos criterios de selección para diferentes valores de p_y y q_y, donde se escoge el modelo $VARMA(2,1)$ para los factores y_t.

Para facilitar la interpretación de los factores, los modelos (9) y (10) fueron estimados con la restricción que los componentes de factores y_t son independientes. Esta restricción es equivalente a que las matrices Φ y Θ del modelo $VARMA$ para

Tabla 6: Valores de criterios de selección de diferentes modelos $VARMA$ para el proceso de los factores latentes.

	órdenes del $VARMA$							
	(1,0)	(0,1)	(1,1)	(2,0)	(0,2)	(2,1)	(1,2)	(2,2)
AIC	-82.56	-41.34	-41.64	-82.56	-42.94	-84.32	-52.78	-56.41
BIC	-82.53	-41.35	-41.62	-82.31	-42.94	-84.30	-52.76	-56.12

y_t son diagonales y el término de error, a_t, sea ruido blanco gaussiano con $\Sigma_a = I$. Finalmente el modelo estimado es:

$$z_t = \begin{bmatrix} 0.11 & 0.31 \\ -0.02 & 0.08 \\ 0.94 & -0.11 \\ -0.27 & -0.13 \\ -0.03 & -0.56 \\ -0.04 & -0.74 \\ 0.08 & 0.04 \\ 0.01 & 0 \\ 0.10 & -0.09 \end{bmatrix} y_t + \epsilon_t \tag{29}$$

y

$$y_t = \begin{bmatrix} 0.28 & 0 \\ 0 & 0.87 \end{bmatrix} y_{t-1} + \begin{bmatrix} -0.13 & 0 \\ 0 & 0.52 \end{bmatrix} y_{t-2} + \begin{bmatrix} 0.1 & 0 \\ 0 & -0.17 \end{bmatrix} a_{t-1} + a_t \tag{30}$$

con $\Sigma_a = I$ y $\Sigma_\epsilon = diag(0, 0.05, 0.04, 0.22, 0.14, 0.29, 0.05, 0.28, 0.02)$. De las estimaciones obtenidas[2], vemos que el primer factor está asociado principalmente con la variable $PRCEM$ diferenciada, pues esta tiene el mayor peso, 0.94, sobre el primer factor; mientras que el segundo factor está asociado de manera negativa con las variables $PRCEM$, IPR, $IEMOB$ y $EFECRC$.

5.3. Análisis de residuales

Para verificar que los residuales ϵ_t son aproximadamente un proceso ruido blanco, se utiliza la estadística Q, siguiendo a Li (2004), definida como sigue:

$$Q(m) = n \sum_{k=1}^{m} tr\left(C_k' C_0^{-1} C_k C_0^{-1}\right) \tag{31}$$

donde

$$C_k = \frac{1}{n} \sum_{t=k+1}^{n} (\widehat{\epsilon}_t - \overline{\epsilon})(\widehat{\epsilon}_t - \overline{\epsilon})' \tag{32}$$

[2]Con el anterior modelo estimado, se puede extraer los dos factores identificados usando la presentación en modelos de estados; por el otro lado, aunque con el método de Brillinger también se puede calcular numéricamente los componentes principales, estos son complejos, lo que dificulta la interpretación, por lo cual se recomienda usar el método de Peña y Box en la práctica.

se tiene que la estadística se distribuye asintóticamente chi-cuadrado con l^2m donde l es la dimensión de la serie ϵ_t. Se aplicó la anterior prueba a los residuales $\widehat{\epsilon}_t$ del modelo (29); el valor de la estadística $Q(m)$ fue de 150.5, un valor grande a primera vista, pero al tener en cuenta que en esta aplicación la dimensión de $\widehat{\epsilon}_t$ es $l = 9$, entonces para cualquier valor de $m > 1$ se tiene que el percentil 95 % de la distribución chi-cuadrado es mayor que 150.5, y así se acepta a $\widehat{\epsilon}_t$ como la realización de un proceso ruido blanco.

6. Conclusiones y sugerencias

En este trabajo se comparó el método de Peña y Box y el método de Brillinger para la reducción de dimensionalidad bajo los supuestos respectivos de los dos métodos. Se encontró que tanto en las simulaciones como en la aplicación empírica los dos métodos tienen la misma capacidad para identificar el número de factores comunes. Cabe resaltar que en las simulaciones y en la aplicación empírica se utilizó el método de Peña y Box modificado, según los resultados de este trabajo. Adicionalmente se encontró que:

- Con respecto a la identificación del número de factores comunes, el método de Brillinger conduce a los mismos resultados que el método de Peña y Box usando $\widehat{\Gamma}_z(0)$.
- Cuando el número de factores es cercano al número de series observadas, el uso del método de Brillinger o el uso de $\widehat{\Gamma}_z(0)$ es mejor que el de $\widehat{\Gamma}_z(1)$, pues $\widehat{\Gamma}_z(1)$ tiende a subestimar el número de factores.
- Con respecto a la extracción de factores comunes, se encontró que el método de Brillinger conduce a factores complejos que no tienen interpretación en la práctica, mientras que el método de Peña y Box, mediante el uso de modelos de estados, permite una interpretación más diáfana.

Como resumen, en la tabla 7 se encuentran los principales inconvenientes de cada uno de los dos métodos. Para algunos se proveen algunas soluciones respectivos y en la tabla 8 se encuentran los principales resultados de comparación de los dos métodos.

Como motivo para futuras investigaciones, nótese que en el método de Peña y Box modificado propuesto en este trabajo, la identificación del número de factores se lleva a cabo observando la magnitud de los valores singulares de las matrices $\widehat{\Gamma}_z(k)$ de manera empírica o heurística, pues en la literatura aún no se conoce la distribución probabilística de estos valores singulares, y por lo tanto tampoco existe una prueba estadística basada en valores singulares para determinar el número de factores. Otro tema abierto a investigación es probar matemáticamente que los dos métodos tienen las mismas capacidades de identificación del número de factores, y adicionalmente su capacidad predictiva. El tema de investigación de este artículo continúa abierto pues los resultados encontrados acá están supeditados a los supuestos originales de los autores de cada método. Sin embargo es

Tabla 7: Inconvenientes y mejoras de los dos métodos.

Método de Peña-Box	
Inconveniente	Solución
Valores propios de $\widehat{\Gamma(k)}$ pueden ser negativos o complejos, causando problema para determinar el número de factores r.	Usar valores singulares en vez de valores propios, pues son siempre reales y positivos.
No se puede encontrar los vectores propios ortonormales de $\widehat{\Gamma}(k)$, pues estas no son simétricas.	Usar el modelo de estado para estimar el modelo y extraer los factores
Método de Brillinger	
Inconveniente	Solución
Conduce a factores complejos que dificultan la interpretación	No hay solución

Tabla 8: Comparaciones entre los dos métodos.

Método de Peña-Box	Método de Brillinger
Tienen la misma capacidad para identificar el número de factores latentes.	
La metodología original usa valores propios de $\widehat{\Gamma}(k)$ que pueden ser complejos o negativos	Usa varianza recogida por cada factor que siempre es real y positivo.
Procedimientos relativamente sencillos	Procedimientos numéricos largos y costos
Conduce a factores reales	Conduce a factores complejos
Factores con fácil interpretación	Factores carecen de interpretación
Se puede extender para series no estacionarias	En la literatura no se conoce todavía extensión para series no estacionarias

posible plantearse qué tan robustos son los métodos mediante el uso de residuales caracterizados porque su función de densidad tuviese colas pesadas, o mostrasen evidencia de heterocedasticidad u otras características. Por otro lado, los resultados encontrados en este artículo se basan en el método de simulación de series multivariadas dado por el teorema 5. Cabe resaltar que no es la única alternativa para la simulación de las series. Por ejemplo, otra opción está dada por la simulación de una serie multivariada que sea una combinación convexa entre un proceso generado de acuerdo con el método de Peña y Box y otro respecto al método de Brillinger.

Agradecimientos

El autor da gracias a Dios por su bondad, al profesor Fabio Nieto por su paciencia, a Andrés Gutiérrez por su ayuda en la redacción y a los árbitros por los valiosos comentarios.

[Recibido: julio de 2008 — Aceptado: septiembre de 2009]

Referencias

Ahn, S. K. & Reinsel, G. C. (1988), 'Nested Reduced-Rank Autoregressive Models for Multiple Time Series', *Journal of the American Statistical Association* **83**, 849–856.

Brillinger, D. R. (1981), *Time Series: Data Analysis and Theory*, Holden-Day, San Francisco, United States.

Brockwell, P. J. & Davis, R. A. (1991), *Time Series: Theory and Methods*, 2 edn, Springer, New York, United States.

Brockwell, P. J. & Davis, R. A. (1996), *Introduction to Time Series and Forecasting*, 1 edn, Springer, New York, United States.

Correal, M. E. & Peña, D. (2008), 'Modelo factorial dinámico threshold', *Revista Colombiana de Estadística* **31**(2), 183–192.

Jiménez, J. A. (2004), *Álgebra Lineal II, con aplicaciones en estadística*, 1 edn, Unibiblos, Bogotá, Colombia.

Johansen, S. (1991), 'Estimation and Hypothesis Testing of Cointegration Vectors in Gaussian Vector Autoregressive Models', *Econometrica* **59**(6), 1551–1580.

Li, W. K. (2004), *Diagnostic Checks in Time Series*, 1 edn, Chapman & Hall/CRC.

Martínez, W. (2007), Uso de tendencias comunes en la construcción de índices coincidentes, Tesis de maestría, Departamento de Estadística, Facultad de Ciencias, Universidad Nacional de Colombia, Bogotá, Colombia.

Melo, L. F., Nieto, F., Posada, C. E., Betancourt, Y. R. & Barón, J. D. (2001), 'Un índice coincidente para la actividad económica de Colombia', *Ensayos Sobre Política Económica* **40**, 46–88.

Peña, D. & Box, G. B. P. (1987), 'Identifying a Simplifying Structure in Time Series', *Journal of the American Statistical Association* **82**(399), 836–843.

Peña, D. & Poncela, P. (2006), 'Nonstationary Dynamic Factor Analysis', *Journal of Statistical Planning and Inference* **136**(4), 1237–1257.

Reinsel, G. C. (1983), 'Some Results on Multivariate Autoregressive Index Models', *Biometrika* **70**, 145–156.

Stoffer, D. S. (1999), 'Detecting Common Signals in Multiple Time Series using the Spectral Envelope', *Journal of the American Statistical Association* **94**(448), 1341–1356.

Tiao, G. C. & Box, G. E. P. (1981), 'Modelling Multiple Time Series with Applications', *Journal of the American Statistical Association* **76**(376), 802–816.

Tiao, G. C. & Tsay, R. S. (1989), 'Model Specification in Multivariate Time Series', *Journal of the Royal Statistical Society, B* **51**(2), 157–213.

Wei, W. S. (2006), *Time Series Analysis: Univariate and Multivariate Methods*, 1 edn, Pearson, Boston, United States.