Common Seasonality in Multivariate Time Series

Fabio H. Nieto

Universidad Nacional de Colombia *

Daniel Peña

Universidad Carlos III de Madrid

dpena@est-econ.uc3m.es

Dagoberto Saboyá

Universidad Nacional de Colombia

dsaboyac@unal.edu.co

October 21, 2015

Abstract

Common factors for seasonal multivariate time series are usually obtained by first filtering the series to eliminate the seasonal component and then extracting the nonseasonal common factors. This approach has two drawbacks. First, we cannot detect common factors with seasonal structure; second, it is well known that a deseasonalized time series may exhibit spurious cycles that the original data do not contain, which can make more difficult the detection of

*Corresponding author: fhnietos@unal.edu.co Phone (+57) 1-3165000

the nonseasonal factors. In this paper we propose a procedure using the original data to estimate the dynamic common factors when some, or all, of the time series are seasonal. We assume that the factor may be stationary or nonstationary and seasonal or not. The procedure is based on the asymptotic behavior of the sequence of the so-called sample generalized autocovariance matrices and of the sequence of canonical correlation matrices, and it includes a statistical test for detecting the total number of common factors. The model is estimated by the Kalman Filter. The procedure is illustrated with an environmental example where two interesting seasonal common factors are found.

Keywords: Common seasonality, Dynamic common factors, Multivariate time series.

1 Introduction

Common factors for time series have received much attention in the last years. Restricted Dynamic Factor Models (RDEM) assume a contemporaneous relationship between the series and a small number of factors. Usually these models assume stationarity (Peña and Box (1987); Stock and Watson (1988, 2002); Ahn (1997); Bai and Ng (2002); and Lam and Yao (2012), among others) and use the rank of the lag covariance matrices of the process to identify the number of factors. The estimation of the factors is closely related to the principal components (PC) of the time series (see Tipping and Bishop (1999) and Doz, Giannone, and Reichlin (2012)). Some generalizations to the nonstationary case are Bai (2004), Bai and Ng (2004), Peña and Poncela (2006), and Barigozzi, Lippi, and Luciani (2014) for integrated processes, Pan and Yao (2008) for general nonstationary processes, Eichler, Motta, and Von Sachs (2011) and Motta, Hafner, and Von Sachs (2011) for locally stationary and non-stationarity in the variance, and Luciani and Veredas (2015) for fractional integrated processes.

Generalized Dynamic Factor Models (GDFM) assume a lag relationship between series and factors. Forni, Hallin, Lippi, and Reichlin (2000) proposed a GDFM model allowing for an infinite number of factor lags and low correlation between any two idiosyncratic components. They show that one can consistently estimate the common component of the time series increasing the number of series to infinity. The relationship between RDFM and GDFM has been studied in Forni, Giannone, Lippi, and Reichlin (2009) who proposed a model that can be seen either as restricted or generalized, and developed estimation methods for the factor structure. Common factors models are used in all branches of science including Medicine (Mamede and Schmid (2004)), Chemistry and Envinonmetrics (Yidanaa, Ophoria, and Banoeng-Yakubob (2008)), Engineering (Carpio, Juan, and López (2014)), and Economics and Business (Stock and Watson (2002)). None of these approaches considers seasonal factors.

It is well known that a deseasonalized time series may have spurious behaviors and therefore this adjustment should be avoided, if possible, when this is not the goal of the analysis. Thus, an important issue is to include directly the seasonal characteristic in the common-factors modeling procedure, avoiding deseasonalization a priori of the time series. Melo, Nieto, Posada, Betancourt, and Barón (2001) analyzed a model with only a (nonstationary) common factor and nine seasonal variables with the seasonal characteristic of each variable specified as a deterministic dummy variable. Busetti (2006) developed a procedure for handling seasonal common factors under the multivariate structural model of Harvey (1989), but without including stationary or (nonseasonal) nonstationary factors. Alonso, Rodríguez, García-Martos, and Sánchez (2011) and García-Martos, Rodríguez, and Sánchez (2011) proposed a RDFM where the factors follow a seasonal multiplicative VARIMA model. The model is very general, but it does not assume orthogonality among the factors and does not separate the different types of factors. Therefore, it is not easy to identify from the data which and how many factors determine the common trends and how

many determine the common seasonality.

The paper is organized as follows: in Section 2 we present our common factors model in which we assume three sets of factors: (i) nonstationary nonseasonal factors affecting the trend; (ii) nonstationary seasonal factors affecting the seasonal pattern; and (iii) stationary common factors. In Section 3 we define the sample generalized autocovariance matrices for seasonal data and find their asymptotic behavior in terms of weak convergence. We include two theorems that describe the limit behaviour of the eigenvalues of the sample generalized autocovariance matrices and canonical correlation matrices, and present a test for the total number of commom factors. Some simulations to illustrate in finite samples the performance of the proposed statistical test and the properties of the sequences of eigenvalues are reported in Section 4. In Section 5 we indicate how the factorial model can be estimated in State Space form. Section 6 presents an application to environmental data. Finally, Section 7 concludes.

2 Factor model specification

Let $\{y_t = (y_{1t}, ..., y_{mt})^{\mathrm{T}}\}$ be an observable multivariate time series generated by an r-dimensional latent process $\{f_t\}$, where $r \leq m$, with

$$y_t \;=\; P f_t \;+\; e_t \;,\; t \in \mathbb{Z} \;, \tag{1}$$

where $\mathbb{Z}$ is the set of integer numbers, P is an $m \times r$ factor loading matrix, and the process $\{e_t\}$ is a multivariate Gaussian white noise process with mean 0 and full-rank diagonal variance matrix Σ_e. The symbol "T"means matrix transposition. We assume that $f_t = (f_{1t}^{\mathrm{T}}, f_{2t}^{\mathrm{T}}, f_{3t}^{\mathrm{T}})^{\mathrm{T}}$, where the process $\{f_{1t}\}$ is nonstationary and nonseasonal, with dimension r_1 and follows the model

$$\phi_1(B)\nabla^d f_{1t} = \theta_1(B) a_{1t}, \tag{2}$$

where $\nabla = (1 - B)$ and $d \geq 1$. The process $\{f_{2t}\}$ is seasonal (nonstationary) with period S and dimension r_2, such that

$$\phi_2(B^S)\nabla_S^D f_{2t} = \theta_2(B^S)a_{2t}, \tag{3}$$

where $\nabla_S = (1 - B^S)$ and $D \geq 1$. Finally, $\{f_{3t}\}$ is stationary with dimension r_3 and follows the model

$$\phi_3(B)f_{3t} = \theta_3(B)a_{3t}. \tag{4}$$

For each $i = 1, 2, 3$, $\{a_{it}\}$ is a Gaussian white noise process with mean 0 and full-rank variance matrix Σ_i and the determinants of the matrix polynomials $\phi_i(\cdot)$ have their roots outside the unit circle. Here, $r_1 + r_2 + r_3 = r$ and we write $P = [P_1, P_2, P_3]$, where the submatrix P_i is of dimension $m \times r_i$, $i = 1, 2, 3$. For future reference, we set $f_{it} = (f_{i1,t}, ..., f_{ir_i,t})^{\mathrm{T}}$ for all $i = 1, 2, 3$ and for all $t \in \mathbb{Z}$.

We need the following assumptions in order to establish our main results.

Assumption A1. For all $i, j = 1, 2, 3$, with $i \neq j$, and all $t \in \mathbb{Z}$, the random vectors a_{it} and a_{jt} are orthogonal.

Assumption A2. The processes $\{a_t = (a_{1t}^{\mathrm{T}}, a_{2t}^{\mathrm{T}}, a_{3t}^{\mathrm{T}})^{\mathrm{T}}\}$ and $\{e_t\}$ are orthogonal, so that a_{it} and e_s are orthogonal for all $i = 1, 2, 3$ and all $t, s \in \mathbb{Z}$.

It is easy to see that A2 implies that f_t and e_s are orthogonal for each $t, s \in Z$.

Assumption A3. For model identifiability $\Sigma_a = \mathrm{Var}(a_t) = I_r$, where I_r is the identity matrix of order r.

This assumption on Σ_a and the linear representation of a VARIMA process (Nieto (2007)) imply that f_{it} and f_{jt} are orthogonal for all $i, j = 1, 2, 3$ with $i \neq j$ and all $t \in \mathbb{Z}$. Furthermore, the components of vector f_{it} are pairwise orthogonal for all $t \in \mathbb{Z}$ and all $i = 1, 2, 3$.

Assumption A4. The matrix polynomial operators $\phi_i(B)$ and $\theta_i(B)$ are diagonal with entries $\phi_{ij}(B)$ and $\theta_{ij}(B)$, respectively, for $j = 1, ..., r_i$ and $i = 1, 3$. Analogously,

the matrix polynomial operators $\phi_2(B^S)$ and $\theta_2(B^S)$ are diagonal with entries $\phi_{2j}(B^S)$ and $\theta_{2j}(B^S)$, respectively, for $j = 1, ..., r_2$.

Assumption A5. With $\psi_{ij}(B) = \phi_{ij}^{-1}(B)\theta_{ij}(B) = \sum_{k=0}^{\infty} \psi_{ij,k}B^k$, $i = 1, 2, 3$,and $j = 1, ..., r_1$, $\sum_{k=0}^{\infty} k|\psi_{ij,k}| < \infty$ and $\psi_{ij}(1) \neq 0$.

Under assumption *A5*, the matrices $\Psi_i(1) = \text{diag}\{\psi_{i1}(1), ..., \psi_{ir_i}(1)\}$ are of rank $r_i, i = 1, 2, 3$.

3 Some properties of the seasonal factor model

3.1 Theoretical characteristics

We assume for simplicity that $d = D$. Let N be the sample size. We define the sample generalized autocovariance (SGCV) matrices $C(k, N)$ as

$$C(k, N) = \frac{S^{2d}}{N^{2d}} \sum_{t=k+1}^{N} (y_{t-k} - \bar{y})(y_t - \bar{y})^{\mathrm{T}} \ , \tag{5}$$

where $k = 0, 1, 2, ..., N - 1$ and $\bar{y} = \frac{1}{N}\sum_{t=1}^{N} y_t$. The weight of the cross-products sum in $C(k, N)$ is equal to $1/(\frac{N}{S})^{2d}$, where $\frac{N}{S}$ is the number of seasons in the sample, whenever N is an integer multiple of S. Our definition of $C(k, N)$ is different from that of Peña and Poncela (2006) since it takes into account the presence of seasonality. The canonical correlation matrices $M(k, N)$, $k = 1, 2, ..., N - 1$, are defined as

$$M(k, N) = [\sum_{t=k+1}^{N} y_t y_t^{\mathrm{T}}]^{-1} \sum_{t=k+1}^{N} y_t y_{t-k}^{\mathrm{T}} [\sum_{t=k+1}^{N} y_{t-k} y_{t-k}^{\mathrm{T}}]^{-1} \sum_{t=k+1}^{N} y_{t-k} y_t^{\mathrm{T}} \ , \tag{6}$$

and it is well known that their eigenvalues are the squared canonical correlations between y_{t-k} and y_t.

Theorem 1. If *A1-A5* hold and K is a positive integer such that $K/N \to 0$ as $N \to \infty$, then, for each $k = 0, 1, ..., K$, we have that

(i) as $N \longrightarrow \infty$, the sequence $\{C(k,N)\}_N$ converges weakly to the random matrix

$$\Gamma_{Y,S}(k) = \left[P_1\Psi_1(1)\Sigma_1^{\frac{1}{2}}, P_2\Psi_2(1)\Sigma_2^{\frac{1}{2}}\right] \begin{bmatrix} A_Y & B_{Y,S} \\ B_{Y,S}^{\mathrm{T}} & C_S(k) \end{bmatrix} \begin{bmatrix} \left(\Sigma_1^{\frac{1}{2}}\right)^{\mathrm{T}} \Psi_1(1)^{\mathrm{T}} P_1^{\mathrm{T}} \\ \left(\Sigma_2^{\frac{1}{2}}\right)^{\mathrm{T}} \Psi_2(1)^{\mathrm{T}} P_2^{\mathrm{T}} \end{bmatrix}, \quad (7)$$

where $A_Y = S^{2d} \int_0^1 V_{0,d-1}(t)V_{0,d-1}(t)^{\mathrm{T}}dt$, $B_{Y,S} = \sum_{s=1}^{S} \int_0^1 V_{0,d-1}(t)V_{s,d-1}(t)^{\mathrm{T}}dt$, and $C_S(k) = \sum_{s=1}^{S} \int_0^1 V_{s,d-1}(t)V_{l_s,d-1}(t)^{\mathrm{T}}dt$, with $V_{i,d}(t) = F_{i,d}(t) - \int_0^1 F_{i,d}(t)dt$ for all $i = 0, 1, ..., S$ and l_s is a natural number that depends on k and s. Here, the process $\{F_{i,d}(t)\}$ is defined recursively by $F_{i,d}(t) = \int_0^t F_{i,d-1}(\tau)d\tau$ for all $d \in \mathbb{N}$, with $F_{i,0}(t) = W_i(t)$ an r_1-dimensional Brownian motion if $i = 0$ and r_2-dimensional if $i \geqslant 1$, and such that $\{W_i(t)\}$ is independent of $\{W_j(t)\}$ for all $i, j = 0, 1, ..., S$ with $i \neq j$.

(ii) The random eigenvalues of $\Gamma_{Y,S}(k)$ are such that for $k = jS$, and for all $j \in \mathbb{N}$, we have, almost surely, $r_1 + r_2$ nonzero eigenvalues which are positive and $m - (r_1 + r_2)$ eigenvalues which are equal to zero, and for $k \neq jS$, for all $j \in \mathbb{N}$, we have r_1 nonzero positive eigenvalues and $m - r_1$ eigenvalues equal to zero.

Proof. See the Appendix. There l_s depends on k and s and, consequently, $C_S(k)$ depends on k.

Remarks. (1) Putting $S = 1$, Peña and Poncela's (2006) Theorem 1 is a particular case of our Theorem 1. (2) Comparing our limit random matrix with that of Peña and Poncela (2006), we see that ours deviates from theirs at lags of the form $k = jS$, $j \in \mathbb{N}$, the seasonal lags. (3) Let $(\Omega, \mathfrak{F}, P)$ be the probability space on which all the random elements are defined. Let $\omega \in \Omega$ and $\Lambda_1(k,\omega)$,...,$\Lambda_m(k,\omega)$ be the eigenvalues of the numerical matrix $\Gamma_{Y,S}(k,\omega)$, for some lag k. Then, if $k = jS$, $j \in \mathbb{N}$, the eigenvectors corresponding to the $r_1 + r_2$ positive eigenvalues of $\Gamma_{Y,S}(k,\omega)$ form a basis of the column space of the submatrix $[P_1, P_2]$ and, if $k \neq jS$ for all $j \in \mathbb{N}$, the eigenvectors corresponding to the r_1 positive eigenvalues are a basis of the column

space of the submatrix P_1. This happens almost surely.

Part (ii) of Theorem 1 has empirical implications: the random matrix $\Gamma_{Y,S}(k)$, $\{\Lambda_1(k),...,\Lambda_m(k)\}$ has two disjoint subsets of random eigenvalues: one contains the, almost surely, positive ones, and the other the, almost surely, zero values. Suppose we list the zero eigenvalues as either $\Lambda_{r_1+1}(k) = ... = \Lambda_m(k) = 0$ if k is not a seasonal lag, or $\Lambda_{r_1+r_2+1}(k) = ... = \Lambda_m(k) = 0$ if k is a seasonal lag. Let $\lambda_1(k,N),...,\lambda_m(k,N)$ be the random eigenvalues of the random matrix $C(k,N)$. Then, for each k and for all $i = 1, ..., m$, the sequence $\{\lambda_i(k,N)\}$ converges weakly to a random eigenvalue of the matrix $\Gamma_{Y,S}(k)$ as $N \longrightarrow \infty$. Let $\Lambda_i(k)$ be the limit of such sequence. Then, $\{\lambda_i(k,N)\}$ converges weakly to 0 as $N \longrightarrow \infty$ when $k \neq jS$, $j \in \mathbb{N}$, and $i > r_1$, or when $k = jS$, $j \in \mathbb{N}$, and $i > r_1 + r_2$. Hence, $\{\lambda_i(k,N)\}$ converges in probability to 0 as $N \longrightarrow \infty$ at the nonseasonal lags when $i > r_1$, and at the seasonal lags when $i > r_1 + r_2$. In this way, for N large enough and for any $\epsilon > 0$, $P(|\lambda_i(k,N)| \leq \epsilon)$ is close to 1 for $i > r_1$ when $k \neq jS$, and also for $i > r_1 + r_2$ if $k = jS$.

Now, if $\hat{C}(k,N)$ is the sample SGCV matrix, the corresponding eigenvalues $\hat{\lambda}_i(k,N)$ must be numerically very small for $i = r_1, ..., m$ when $k \neq jS$ and for $i = r_1 + r_2 + 1, ..., m$ when $k = jS$. This suggests plotting the m eigenvalues sequences $\{\hat{\lambda}_i(k,N)\}$ indexed by k in order to check for this numerical characteristic. In Section 4.2 we illustrate with simulated examples the practical utility of these plots to specify the number r_1 of nonstationary and nonseasonal common factors and the number r_2 of seasonally integrated factors. Obviously, finding the rate of convergence to zero of the eigenvalue sequences is an important problem and will be investigated in the future.

Theorem 2. If *A1-A5* hold, $P^{\mathrm{T}}P = I$, and K is a positive integer such that $K/N \to 0$ as $N \to \infty$, then, for each $k = 1, ..., K$, the sequence $\{M(k,N)\}$ converges weakly to a random matrix that has $m - r$ eigenvalues equal to zero.

Proof. See the Appendix. In this result, the limit matrix does not depend on k.

Remark. Peña and Poncela's (2006) Theorem 3 remains valid for this more general

model with seasonal common factors.

Using similar arguments to those used for characterizing the numerical eigenvalue sequences of the SGCV matrices, we find that the last $m - r$ numerical eigenvalues sequences, indexed by k, of the sample matrices $\hat{M}(k, N)$ are expected to have very small values when N is large enough. Thus, plots of the m numerical sequences of eigenvalues of matrices $\hat{M}(k, N)$ might help to specify, in practice, the total number r of common factors.

A test statistic for the null hypothesis that the model has r factors, then can be given by

$$S_{m-r,k}(N) = -(N-k)\sum_{j=1}^{m-r}\ln(1-\lambda_j)\ , \tag{8}$$

where $\lambda_1 \leq \cdots \leq \lambda_m$ are the ordered (random) eigenvalues of matrix $M(k, N)$, in the sense that $P(\{\omega \in \Omega : \lambda_1(\omega) \leq \cdots \leq \lambda_m(\omega)\}) = 1$. Under the assumptions in Theorem 2 and that $m - r > 0$, we get that, for all $k = 1, ..., K$, $\{S_{m-r,k}(N)\}$ converges weakly to a $\chi^2_{(m-r)^2}$-distributed random variable as $N \longrightarrow \infty$. The proof of this claim follows the lines of that in Peña and Poncela's (2006) paper. The test is applied starting with $r = 0$, if the test rejects the null hypothesis of no common factors, we check $r = 1$, and we continue increasing r until the hypothesis of r factors is not rejected.

4 A simulation study

4.1 The performance of the test in finite samples

To check the performance in finite samples of the test at (8), we used six factorial models with seasonal variables and $S = 12$. In all the models the variance of the univariate white noise processes was equal to one and we drew 1000 simulations

(sample paths or time series). The six models are given in Table 1, where $\mathbf{0}_{10\times 2}$ denotes the zero matrix of dimension 10×2.

Model	m	r	P	Factor models
M1	2	1	$P_1^T = \begin{bmatrix} 1/3 & \sqrt{8}/3 \end{bmatrix}$	$\nabla_{12} f_t = (1 - .2B^{12}) a_t$
M2	3	2	$P_2^T = \begin{bmatrix} 1 & 1 & .8 \\ 1 & -1 & .2 \end{bmatrix}$	$(1 - .8B)\nabla f_{1t} = (1 - .2B) a_{1t}$ $(1 - .4B^{12})\nabla_{12} f_{2t} = (1 - .2B^{12}) a_{2t}$
M3	4	2	$P_3^T = \begin{bmatrix} .5 & .2 & .25 & -.81 \\ 0 & .33 & .94 & -.02 \end{bmatrix}$	$\nabla f_{1t} = a_{1t},\ \nabla_{12} f_{2t} = a_{2t}$
M4	10	2	$P_4^T = \begin{bmatrix} P_3^T & P_3^T & .5 I_2 \end{bmatrix}$	$\nabla f_{1t} = a_{1t},\ \nabla_{12} f_{2t} = a_{2t}$
M5	20	2	$P_5^T = \begin{bmatrix} P_4^T & P_4^T \end{bmatrix}$	$\nabla f_{1t} = a_{1t},\ \nabla_{12} f_{2t} = a_{2t}$
M6	50	2	$P_7^T = \begin{bmatrix} P_5^T & P_5^T & \mathbf{0}_{10\times 2}^T \end{bmatrix}$	$\nabla f_{1t} = a_{1t},\ \nabla_{12} f_{2t} = a_{2t}$

Table 1: The models in the simulation study

In the cells of Tables 2 and 3 we present the number of times in which the null hypothesis of r factors was rejected. The test was carried out at the 5% significance level. As all models have a seasonal factor, checking the seasonal lags 12 or 24 is more powerful for detecting the true number of factors than checking just lag one. In Table 1 the sample size, N, is 120, 480, and 1000. The power of the test depends on the lag and the ratio N/m, which measures the effective number of observations for each series and the accuracy of the canonical correlation matrices. For instance, in model M1 the hypothesis of zero factors is rejected with a relative frequency that goes from

0.413 to 1 depending on the lag and the sample size. The test is more powerful at the seasonal lags and when we increase the sample size. A similar situation happens for models M2 to M4, where the hypothesis of one factor is rejected with less power when we decrease the ratio N/m. However, the performance of the test is very good when we include 12 or 24 lags and $N/m \geq 30$ (note the decrease in power in M4 with $m = 10$ and $N = 120$ because then N/m is only 12).

Table 3 presents the results for a moderate number of time series ($m = 20, 50$), and sample sizes $30m$ and $60m$. We have found a clear decrease in the power of the test when $N/m < 20$. In these cases the test may suggest more factors than the true value, although with a small probability unless this ratio is very small (say smaller than 10).

4.2 Numerical behavior of eigenvalues in finite samples

We analyzed the finite-sample behavior of the eigenvalues of the generalized autocovariance matrices, $\hat{C}(k, N)$, and canonical-correlation matrices, $\hat{M}(k, N)$, by using models M1 and M2 of the previous subsection. Since $S = 12$ we covered three seasons for the eigenvalues analysis and compute the matrices for lags $k = 0, 1, ..., 35$, then obtained their m eigenvalues, put them in descendent order according to their absolute values, and conformed m eigenvalues sequences indexed by k. We set $N = 984$ as the sample size (82 complete seasons) for each model. This experiment was repeated in 1000 time series generated by each model. Then, we computed the average value of each eigenvalue at each lag and its standard deviation.

Figure 1 shows the two mean-values sequences of these eigenvalues and the bands of ± 2 standard deviations around the average values for matrices $\hat{C}(k, N)$ computed by data generated from M1. We find that the first eigenvalue has significant values at the seasonal lags, whereas the second eigenvalue is practically zero at all lags. This

Model	m	r	Sample size								
			120			480			1000		
			Lag k			Lag k			Lag k		
			1	12	24	1	12	24	1	12	24
M1	2	0	413	1000	1000	659	1000	1000	706	1000	1000
		1	18	51	43	27	64	41	40	55	61
M2	3	0	1000	1000	1000	1000	1000	1000	1000	1000	1000
		1	448	999	986	642	1000	999	702	1000	1000
		2	23	52	57	31	47	46	36	49	52
M3	4	0	1000	1000	1000	1000	1000	1000	1000	1000	1000
		1	362	986	960	605	1000	1000	694	1000	1000
		2	23	58	43	28	50	55	38	48	49
		3	1	1	3	2	2	4	0	4	2
M4	10	0	1000	1000	1000	1000	1000	1000	1000	1000	1000
		1	314	992	961	442	1000	997	527	1000	1000
		2	31	150	175	17	65	61	25	62	52
		3	3	6	16	0	3	1	2	2	1

Table 2: Frequencies of rejecting the null hypothesis in Models 1 to 4

Model	m	r	Sample size					
			$30m$			$60m$		
			Lag k			Lag k		
			1	12	24	1	12	24
M5	20	0	1000	1000	1000	1000	1000	1000
		1	394	1000	997	452	1000	1000
		2	43	155	122	35	80	67
		3	4	8	8	0	4	2
		4	0	0	1	0	0	0
M6	50	0	1000	1000	1000	1000	1000	1000
		1	630	1000	1000	539	1000	1000
		2	137	333	321	51	140	149
		3	11	43	35	4	2	18
		4	1	1	2	0	1	0
		5	0	0	0	0	0	0
		6	0	1	0	0	0	0

Table 3: Frequencies of rejecting the null hypothesis in Models 5 to 6

fact coincides with the implication of Theorem 1 for this example, for which $r_1 = 0$ and $r_2 = 1$. The same conclusion is obtained from the two eigenvalues sequences of matrices $\hat{M}(k, N)$. They showed the numerical implication of Theorem 2, for which $r = 1$ (we omit this figure because of space restrictions).

We now consider M2. In Figure 2 we plot the eigenvalues sequences, with bands of ± 2 standard deviations, for matrices $\hat{M}(k, N)$ and note that the third eigenvalue is practically zero. This fact coincides with the thesis of Theorem 2, for which $r = 2$. The first eigenvalue sequence here has many significant values that decrease with the lag, as expected in an integrated process. The second eigenvalue has large values at the seasonal lags, showing a cyclical seasonal behavior. The eigenvalues sequences of matrices $\hat{C}(k, N)$ show the same numerical implications (the plot is omitted) implied by Theorem 1.

The numerical implications of Theorems 1 and 2 have been tested in many simulated models, with different numbers of variables and common factors and the empirical results are analogous to those of M1 and M2 (they can be provided by the authors upon request).

5 Fitting the factor model via a state space form

To estimate the model fixed parameters and the common factors we use maximum likelihood and linear prediction theory (Catlin (1989); Brockwell and Davis (1991)), respectively. The prediction optimality criterion is the Minimum Mean Square Error (MMSE). It is well known that if the common-factors predictors are unbiased their MMSEs are equal to their prediction-error variances. Also, if the prediction errors distributions are known we can find prediction intervals for the unobservable factors. This estimation problem can be accomplished using a state space form (SSF). Then, taking into account the Gaussianity assumption, the maximum likelihood estimators

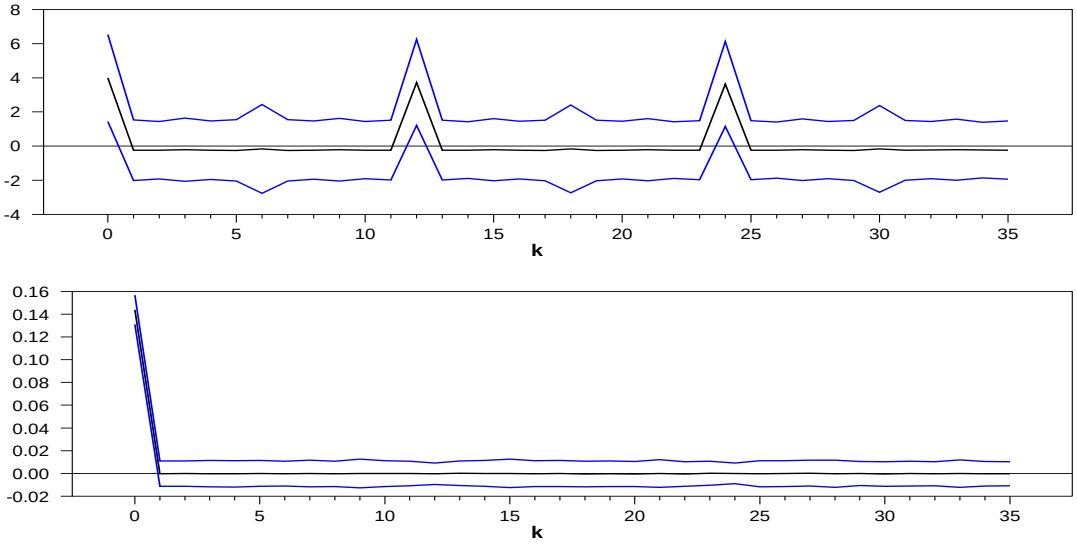

Figure 1: Sequences of the first (top) and the second (bottom) mean eigenvalues for matrices $\hat{C}(k,T)$ in M1, with bands of $\pm$ 2 standard deviations

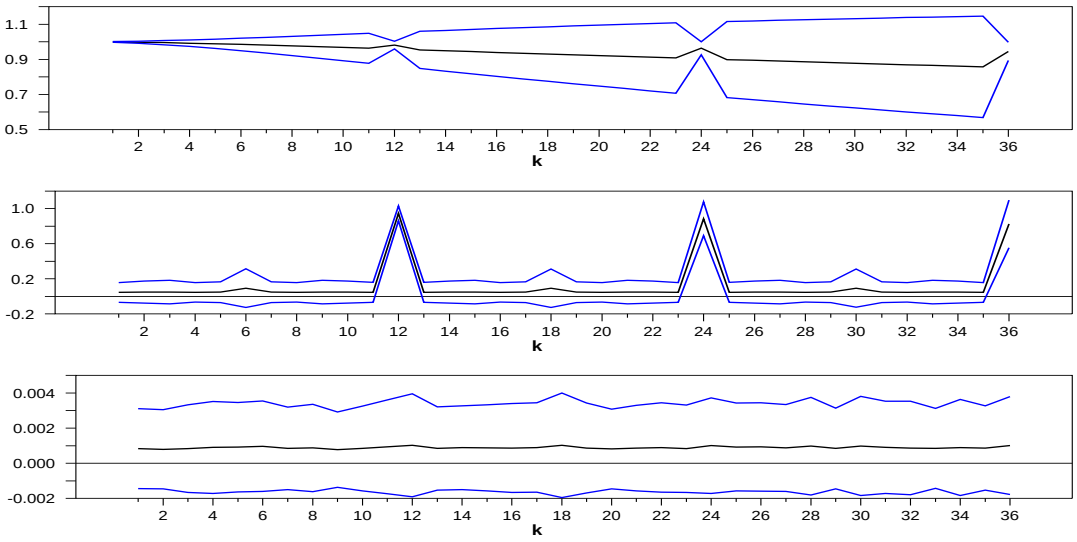

Figure 2: Sequences of the three mean eigenvalues (from top to bottom) for matrices $\hat{M}(k,T)$ in M2, with bands of $\pm$ 2 standard deviations

of the fixed parameters are consistent and asymptotically Normal, and the common factors predictors are unbiased with Gaussian prediction errors (see Harvey (1989)).

In order to obtain the SSF for the factorial model, we need to identify the number r_1 of nonseasonal and nonstationary factors, the number r_2 of seasonally integrated factors, the number r_3 of stationary factors, and the models for the factors. To do this, we propose the following methodology:

Step 1. Finding the number and type of factors. We decide the total number r of common factors by using the statistical test in Section 3. This decision can be confirmed by the eigenvalues sequences of matrices $\hat{M}(k, N)$, as was described in Sections 3 and 4.2. Then, we obtain r_1 and r_2 by using the eigenvalues sequences of the sample SGCV matrices and looking for the number of large eigenvalues at the nonseasonal lags (r_1) as well as at the seasonal lags (r_2), as was noted in Sections 3 and 4.2. Finally, we obtain $r_3 = r - r_1 - r_2$.

Step 2. Finding a model for the factors. These models can be obtained by one of the following procedures. The first computes a preliminary estimation of the submatrix P_1 by using the eigenvectors associated to the first r_1 eigenvalues of $\hat{C}(k, N)$, for some $k \geq 0$, and obtains the transformed time series $z_t = \hat{P}_1^{\mathrm{T}} y_t$ to identify ARIMA models for each of the components of z_t. In the same way, obtain the r_2 transformed time series $w_t = \hat{P}_2^{\mathrm{T}} y_t$ and identify pure seasonal ARIMA models for the components of w_t, as specified in Section 2. The second procedure is to use Harvey's (1989) unobserved components models for extracting the trend-cycle and seasonal components from each variable via, for example, the statistical package STAMP of Koopman, Harvey, Doornik, and Shepard (2011). Then, ARIMA models for the trend-cycle components and seasonally integrated models for the seasonal components can be found. We take the r_1 most frequent models for the trend-cycle component and the r_2 most frequent for the seasonal component, as the candidate models for the nonstationary and nonseasonal common factors and the seasonal common factors,

respectively.

Step 3. Estimating the model. From the above, a state space model for the multivariate time series can be built, as outlined below, which can be estimated by maximum likelihood (for the so-called hyperparameters) and by the fixed-point smoother algorithm (for the common factors predictions).

The state space model.

In order to implement *Step 3* we set $\varphi_{1j}(B) = \phi_{1j}(B)(1-B)^d$, for each $j = 1, ..., r_1$, $\varphi_{2j}(B^S) = \phi_{2j}(B^S)(1-B^S)^D$, for all $j = 1, ..., r_2$, and $\varphi_{3j}(B) = \phi_{3j}(B)$ for $j = 1, ..., r_3$. Let p_{ij} be the degree of polynomial $\varphi_{ij}(\cdot)$ and q_{ij} the degree of polynomial $\theta_{ij}(\cdot)$ so we can write $\varphi_{ij}(B) = 1 + \sum_{l=1}^{p_{ij}} \varphi_{ij,l} B^l$ and $\theta_{ij}(B) = 1 + \sum_{l=1}^{q_{ij}} \theta_{ij,l} B^l$. Let $r_{ij} = \max\{p_{ij}, q_{ij}+1\}$, $j = 1, ..., r_i$, $i = 1, 2, 3$.

Following Gómez and Maravall (1994), we have the state vector $\alpha_t = (\alpha_{1,t}^{\mathrm{T}}, \alpha_{2,t}^{\mathrm{T}}, \alpha_{3,t}^{\mathrm{T}})^{\mathrm{T}}$, where $\alpha_{i,t}^{\mathrm{T}} = (\alpha_{i1,t}^{\mathrm{T}}, ..., \alpha_{ir_i,t}^{\mathrm{T}})^{\mathrm{T}}$, $i = 1, 2, 3$, with $\alpha_{ij,t} = (f_{ij,t}, f_{ij,t+1|t}, ..., f_{ij,t+r_{ij}-1|t})^{\mathrm{T}}$, $j = 1, ..., r_i$, $i = 1, 2, 3$. Here $f_{ij,t+h|t}$, $h \geq 1$, is the orthogonal projection of $f_{ij,t+h}$ onto the closed span of $\{f_{ij,1}, ..., f_{ij,t}\}$. Since the dimension of vector $\alpha_{ij,t}$ is r_{ij}, the dimension of vector $\alpha_{i,t}$ is $\sum_{j=1}^{r_i} r_{ij} = r_i^*$, $i = 1, 2, 3$, and, consequently, the dimension of α_t is $\sum_{i=1}^{3} r_i^* = r^*$.

For each $j = 1, ..., r_i$ and $i = 1, 2, 3$, let

$$A_{ij} = \begin{bmatrix} 0 & & I_{r_{ij}-1} & \\ -\varphi_{ij,r_{ij}} & -\varphi_{ij,r_{ij}-1} & \cdots & -\varphi_{ij,1} \end{bmatrix},$$

where $I_{r_{ij}-1}$ is the identity matrix of order $r_{ij} - 1$ and $\varphi_{ij,l} = 0$ if $l > p_{ij}$. We put $A_i = \mathrm{diag}\{A_{i1}, ..., A_{ir_i}\}$, for each $i = 1, 2, 3$ and then set $A = \mathrm{diag}\{A_1, A_2, A_3\}$ as the system matrix.

The observation matrix we propose is the matrix $C = PH$, where $H = [H(l,k)]$ is of dimension $r \times r^*$ and its entries are given in the following way. For the ith row we have three cases: (i) if $1 \leq i \leq r_1$, we set $H(i, r_{11} + ... + r_{1,i-1} + 1) = 1$ with the

convention $r_{1,0} = 0$. (ii) if $r_1+1 \leq i \leq r_1+r_2$, we put $H(i, r_1^*+r_{21}+...+r_{2,i-r_1-1}+1) = 1$ with $r_{2,0} = 0$. (iii) if $r_1 + r_2 + 1 \leq i \leq r$, we set $H(i, r_1^* + r_2^* + r_{31} + ... + r_{3,i-r_1-r_2-1} + 1) = 1$, defining $r_{3,0} = 0$. The remaining entries are set equal to zero. Now, the variance of the system-equation error process is $W = GG^{\mathrm{T}}$, where G is a matrix of dimension $r^* \times r$ given by $G = \text{diag}\{G_{11}, ..., G_{1r_1}, G_{21}, ..., G_{2r_2}, G_{31}, ..., G_{3r_3}\}$,where $G_{ij} = (1, \psi_{ij,1}, ..., \psi_{ij,r_{ij}-1})^{\mathrm{T}}$, for $j = 1, ..., r_i$, and $i = 1, 2, 3$, with the numbers $\psi_{ij,k}$; $k = 1, ..., r_{ij} - 1$, obtained from the recursive relations (Brockwell and Davis (1991)),

$$\psi_{ij,0} = 1, \ \psi_{ij,k} = \sum_{l=1}^{\min(p_{ij},k)} (-\varphi_{ij,l})\psi_{ij,k-l}, \ k \geq 1 \ .$$

The state space model is given by $y_t = C\alpha_t + e_t$ as the observation equation, and $\alpha_t = A\alpha_{t-1} + w_t$ as the system equation, where $Var(e_t) = \Sigma_e$, $w_t = Ga_t$, and $Var(w_t) = W$. As initial conditions we put $\alpha_0 = 0$ and $Var(\alpha_0) = 10^p I_{r^*}$, for some positive integer number p (relatively large in order to compensate for large uncertainty).

A simulated example. In order to illustrate the estimation of the factorial model we simulated model M3 of Section 4. We simulated 100 multiple time series of the model, each with sample size 480, and estimated the model parameters using the proposed state space form and the Kalman filter. Then we obtained the sample mean of the 100 estimates of each parameter and its standard deviation. These results are in the matrices $\bar{P}$ and $\bar{\Sigma}_e$ below, with standard deviations in parentheses. Figures are rounded to 2 decimal digits.

$$\bar{P}^{\mathrm{T}} = \begin{bmatrix} 0.49(0.04) & 0.20(0.02) & 0.25(0.03) & -0.80(0.06) \\ 0 & 0.30(0.09) & 0.88(0.07) & -0.01(0.01) \end{bmatrix} ,$$

and $\bar{\Sigma}_e = \text{diag}\{0.99(0.07), 1.01(0.07), 1.05(0.12), 1.01(0.09)\}$. Comparing to the true values in Table 1, we conclude that the estimated parameters are close and the intervals of ± 2 standard deviations contain the true values.

6 An empirical application

We present a data application of our proposed methodology. The variables to be considered are monthly measures of rainfall (in mm) from the meteorological stations located at the airports of six cities in Colombia: Bucaramanga (y_1), Cúcuta (y_2), Ibagué (y_3), Medellín (y_4), Manizales (y_5), and Bogotá (y_6). The sample period is January, 1975-June, 2013. In Figure 3 we plot the time series provided by IDEAM, the Colombian official agency for climatic and environmental studies. Colombia is located close to the equator in the Torrid Zone and, in a typical year, two rain epochs occur in the periods April-June and October-December, approximately.

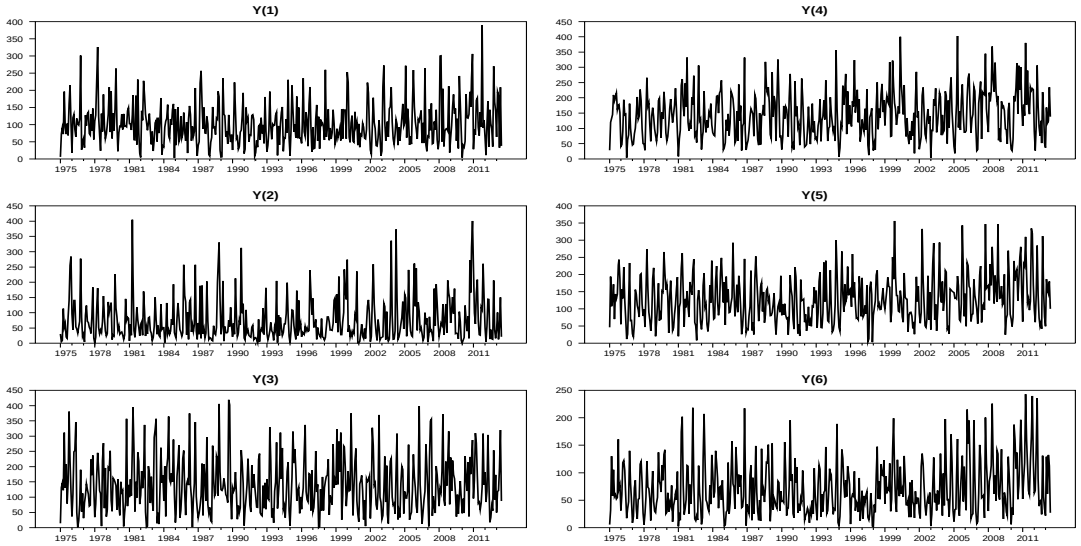

Figure 3: Colombian rainfalls

Step 1. We present in Table 4 the p-values for the test for the number of factors. The test is expected to be more powerful for identifying seasonal factors at lags 12 or 24. This is seen in Table 4 where the hypothesis of two factors is clear at seasonal lags. The plot of the eigenvalues of matrices $\hat{M}(k, N)$, shown in Figure 4, strongly suggests two seasonal factors. In order to confirm the number of nonstationary common factors

and their type, nonseasonal (r_1) and seasonal (r_2), we computed the eigenvalues sequences of matrices $\hat{C}(k, N)$. In Figure 5(a) we plot the first sequence that shows a cyclical pattern and large values (in absolute value) at seasonal lags. In Figure 5(b) we plot the next five eigenvalues and it can be seen that the second eigenvalue has also relatively large values at the seasonal lags. Thus, we conclude that $r_1 = 0$ and $r_2 = 2$.

r	Lag k		
	1	12	24
0	0.0000	0.0000	0.0000
1	0.0000	0.0041	0.0035
2	0.0473	0.9759	0.6805
3	0.4305	0.9461	0.6686
4	0.3571	0.9617	0.6457
5	0.8472	0.8741	0.3112

Table 4: Results for the statistical tests in the data example

Step 2. To identify the stochastic models for the common factors we used the first procedure that was proposed in Section 5, and we specified a SARIMA$(0, 1, 1)_{12}$ model for the first factor $f_{1,t}$ and a SARIMA$(1, 1, 0)_{12}$ model for the second, $f_{2,t}$.

Step 3. Using the DLM instruction of the RATS package (Doan (2011)), we obtained the estimation results $(1-B^{12})f_{1,t} = (1-0.91B^{12})a_{1,t}$ and $(1+0.58B^{12})(1-B^{12})f_{2,t} = a_{2,t}$ as the factors models,
$\hat{\Sigma}_e = \text{diag}\{2712.39, 3070.49, 3794.05, 1709.43, 1463.79, 869.58\}$, and the loading matrix:

$$\hat{P}^T = \begin{bmatrix} 30.74 & 25.58 & 44.62 & 43.94 & 40.83 & 23.38 \\ 0.00 & -0.62 & -0.43 & 0.71 & -0.56 & -0.18 \end{bmatrix}.$$

In the estimation of matrix $P = (p_{ij})$ we set $p_{12} = 0$ as an additional condition

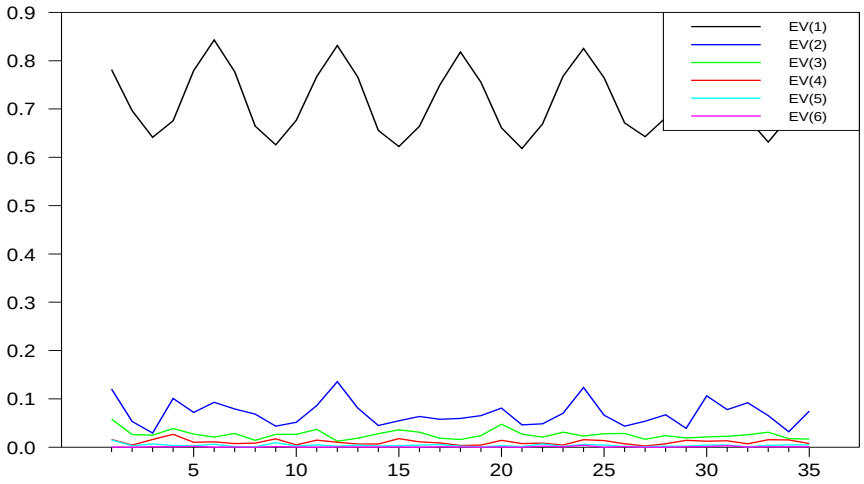

Figure 4: Eigenvalues sequences for the canonical correlation matrices in the rainfall data example

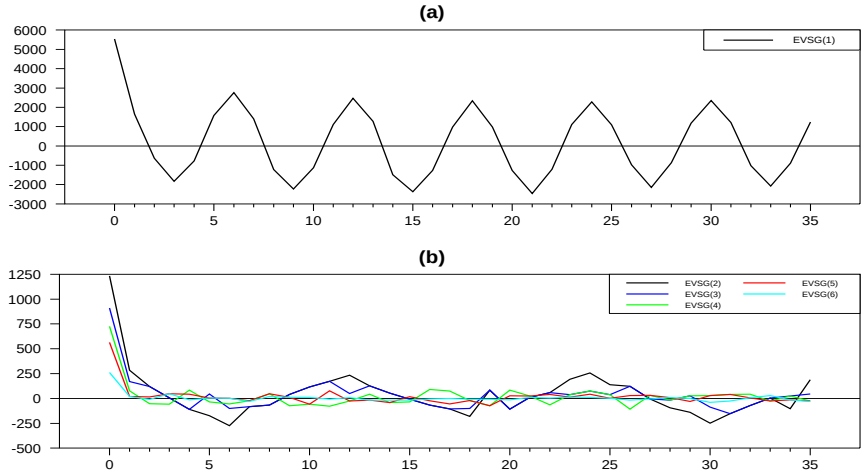

Figure 5: Eigenvalues sequences for the SGCV matrices in the data example: (a) the first eigenvalue; (b) the last five eigenvalues

for model identifiability. All the estimated parameters are significant at the 5% level.

The structure of the factors can be seen in the columns of the P matrix. The first is a weighted average of all the time series and it follows an IMA_{12} model with

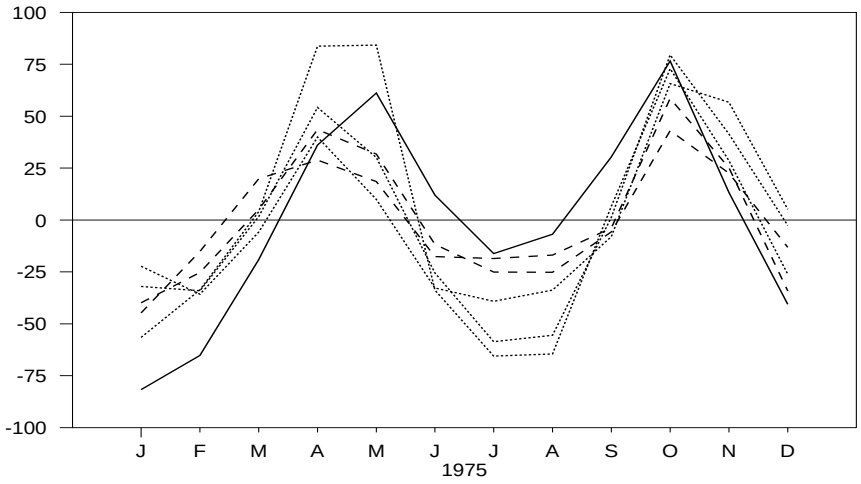

Figure 6: Seasonal effects for the rainfall variables. Medellín (continuous line), Cúcuta, Ibagué, and Manizales (points), Bucaramanga and Bogotá (discontinuous)

a moving average parameter close to one; it represents a stable seasonal pattern in all the series. The second is more complex, separating Medellín rainfall (y_4) from the other cities precipitations, and mostly from the series (y_2, y_3, y_5). The model for this factor indicates that the seasonal pattern is changing over time. After this result we computed a simple estimation of the seasonal coefficients for each time series by the difference between the monthly mean in different years and the global mean of the time series; they are plotted in Figure 6. It can be seen that the seasonal coefficients of Medellín rainfall are different from the other cities and mainly from those of Cúcuta, Ibagué, and Manizales (y_2, y_3, y_5). The precipitation in the period June-September in Medellin, although below the mean of the year, is larger than in the other cities and mainly with respect to (y_2, y_3, y_5). Also, in this period Medellin has a larger precipitation than in the period November-March, whereas Cúcuta, Ibagué and Manizales (y_2, y_3, y_5) have few precipitations in June-September, and of a similar magnitude to the period November-March.

From a meteorological point of view this is a reasonable finding for the rainfalls

studied here because, geographically, Medellín is very close to the Pacific Ocean coast and is influenced by the so-called Low Anchored of Panamá (or of the Pacific Ocean), a phenomenon that causes both high levels and annual large periods of precipitation in the Colombian Pacific-Ocean coast close to Panamá (Zea (2002), Fujita (1962)). In fact Medellin is the city with the largest average precipitation and also with the larger span of rainfall, in agreement with this theory. This explains the need of at least two factors to describe the seasonality on the data. There is a general seasonal behaviour and a specific seasonal pattern due to this geographical effect. The models for the two common factors imply that their seasonal differences have a cycle of period 12 months (besides other cycles); but the autocorrelation function of the first factor is similar to the one usually found in this type of seasonal effect whereas the second factor explains a complex seasonal behavior with dying annual correlation structure that alternates its values.

7 Conclusions

We have presented an extension of the dynamic common factor model with common seasonal stochastic factors. We have shown that the eigenvalues of the random limit matrix (in weak convergence) of the sample generalized autocovariance matrix sequence, are useful for identifying the presence of both nonstationary and nonseasonal common factors and seasonally integrated common factors. Also, we have shown that the sequence of the canonical correlation matrices converges weakly to a random matrix that has $m - r$ eigenvalues equal to zero almost surely, where m is the number of variables and r is the total number of common factors. These results allow a procedure for fitting common factors to seasonal time series that has shown to be useful with data.

Acknowledgements

The authors are very grateful to an anonymous referee and an associate editor, for their very useful comments and suggestions that lead us to substantially improve the paper.

References

Ahn, S.K. (1997). Inference of vector autoregressive models with cointegration and scalar components. *Journal of the American Statistical Association* **437**, 350-356.

Alonso, A.M., Rodríguez, J., García-Martos, C., and Sánchez, M.J. (2011). Seasonal dynamic factor analysis and bootstrap inference: Application to electricity market forecasting. *Technometrics* **53**, 137-151.

Bai, J. (2004). Estimating cross-section common stochastic trends in nonstationary panel data. *Journal of Econometrics* **122**, 137-183.

Bai, J. and Ng, S.(2002). Determining the number of factors in approximate factor models. *Econometrica* **70**, 191-221.

Bai, J. and Ng, S. (2004). A PANIC attack on unit roots and cointegration. *Econometrica* **72**, 1127-1177.

Brockwell, P.J. and Davis, R.A. (1991). *Time Series: Theory and Methods*. Springer-Verlag, New York.

Busetti, F. (2006). Tests of seasonal integration and cointegration in multivariate unobserved component models. *Journal of Applied Econometrics* **21**, 419-438.

Catlin, D. (1989). *Estimation, Control, and The Discrete Kalman Filter*. Springer-Verlag, New York.

Carpio, J., Juan, J., and López, D. (2014). Multivariate exponential smoothing and dynamic factor model applied to hourly electric price analysis. *Technometrics* **56**, 494-503

Doan, T.A. (2011). *WinRATS Pro (v. 8.10)*. Estima, Evanston, IL.

Doz, C., Giannone, D., and Reichlin, L. (2012). A quasi maximum likelihood approach for large approximate dynamic factor models. *Review of Economics and Statistics* **94**, 1014-1024.

Eichler, M., Motta, G., and von Sachs, R. (2011). Fitting dynamic factor models to nonstationary time series. *Journal of Econometrics* **163**, 51-70.

Forni, M. and Lippi, M (2011). The general dynamic factor model: One-sided representation results. *Journal of Econometrics* **163**, 23-28.

Forni, M., Hallin, M., Lippi, M., and Zaffaroni, P. (2015). Dynamic factor models with infinite-dimensional factor spaces: One-sided representations. *Journal of Econometrics* **185**, 359-371.

Forni, M., Hallin, M., Lippi, M., and Reichlin, L. (2000). The generalized dynamic factor model: Identification and estimation. *The Review of Economic and Statistics* **82**, 540-554.

Forni, M., Giannone, D., Lippi, M., and Reichlin, L. (2009). Opening the black box: structural factor models with large cross sections. *Econometric Theory* **25**, 1319-1347.

Fujita, T. (1962). A review of researches on analytical mesometeorology. University of Chicago, Department of Geophysical Sciences. Mesometeorology Project, Paper No. 8.

García-Martos, C., Rodríguez, J., and Sánchez, M.J. (2011). Forecasting electricity prices and their volatilities using unobserved components. *Energy Economics* **33**,

1227-1239.

Gómez, V. and Maravall, A. (1994). Estimation, prediction, and interpolation for nonstationary series with the Kalman Filter. *Journal of the American Statistical Association* **89**, 611-624.

Harvey, A.C. (1989). *Forecasting, Structural Time Series Models, and the Kalman Filter*. Cambridge University Press, Cambridge.

Koopman, S.J., Harvey, A.C., Doornik, J.A., and Shephard, N. (2011). *Stamp 8.0: Structural Time Series Analyser, Modeller and Predictor*. Timberlake Consultants, London.

Lam, C. and Yao, Q. (2012). Factor modeling for high-dimensional time series: inference for the number of factors. *Annals of Statistics* **40**, 694-726.

Luciani, M. and Veredas, D. (2015). Estimating and forecasting large panels of volatilities with approximate dynamic factor models. *Journal of Forecasting* (forthcoming).

Mamede, S. and Schmid, H. (2004). The structure of reflective practice in medicine. *Medical Education* **38**, 1302-1308.

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

Motta, G., Hafner, C.M., and Von Sachs, R. (2011). Locally stationary factor models: Identification and nonparametric estimation. *Econometric Theory* **27**, 1279-1319.

Nieto, F.H. (2007). *Ex Post* and *Ex Ante* prediction of unobserved multivariate time series: a structural-model based approach. *Journal of Forecasting* **26**, 53-76.

Pan, J. and Yao, Q. (2008). Modelling multiple time series via common factors.

Biometrika **95**, 365-379.

Peña, D. and Box, G.E.P. (1987). Identifying a simplifying structure in time series. *Journal of the American Statistical Association* **82**, 836-843.

Peña, D. and Poncela, P. (2006). Nonstationary dynamic factor analysis. *Journal of Statistical Planning and Inference* **136**, 1237-1257.

Stock, J. H. and Watson, M.W. (1988). Testing for common trends. *Journal of the American Statistical Association* **83**, 1097-1107.

Stock, J. H. and Watson, M.W (2002). Forecasting using principal components from a large number of predictors. *Journal of the American Statistical Association* **97**, 1167-1179.

Tanaka, K. (1996). *Time Series Analysis. Nonstationary and Noninvertible Distribution Theory.* John Wiley & Sons: New York.

Tipping, M. and Bishop, C. (1999). Probabilistic principal component analysis. *Journal of the Royal Statistical Society B* **61**, 611-622.

Yidanaa, S. M., Ophoria, D., and Banoeng-Yakubob, B. (2008). A multivariate statistical analysis of surface water chemistry data—The Ankobra Basin, Ghana. *Journal of Environmental Management* **86**, 80-87.

Zea, J.A. (2003). Baja Anclada del Pacífico. *Meteorología Colombiana* **7**, 109-116.