Seasonal Dynamic Factor Analysis and Bootstrap Inference: Application to Electricity Market Forecasting

Andrés M. ALONSO

Department of Statistics
Instituto Flores de Lemus
Universidad Carlos III de Madrid
Getafe, Spain

Carolina GARCÍA-MARTOS

Escuela Técnica Superior Ingenieros Industriales
Universidad Politécnica de Madrid
Madrid, Spain
(*garcia.martos@upm.es*)

Julio RODRÍGUEZ

Facultad de Ciencias Económicas y Empresariales
Universidad Autónoma de Madrid
Madrid, Spain

María JESÚS SÁNCHEZ

Escuela Técnica Superior Ingenieros Industriales
Universidad Politécnica de Madrid
Madrid, Spain

In this work, we propose the Seasonal Dynamic Factor Analysis (SeaDFA), an extension of Nonstationary Dynamic Factor Analysis, through which one can deal with dimensionality reduction in vectors of time series in such a way that both common and specific components are extracted. Furthermore, common factors are able to capture not only regular dynamics (stationary or not) but also seasonal ones, by means of the common factors following a multiplicative seasonal VARIMA$(p, d, q) \times (P, D, Q)_s$ model. Additionally, a bootstrap procedure that does not need a backward representation of the model is proposed to be able to make inference for all the parameters in the model. A bootstrap scheme developed for forecasting includes uncertainty due to parameter estimation, allowing enhanced coverage of forecasting intervals. A challenging application is provided. The new proposed model and a bootstrap scheme are applied to an innovative subject in electricity markets: the computation of long-term point forecasts and prediction intervals of electricity prices. Several appendices with technical details, an illustrative example, and an additional table are available online as Supplementary Materials.

KEY WORDS: Dimensionality reduction; Energy prices; Nonstationary; Seasonality; Unobserved components; VARIMA models.

1. INTRODUCTION

When trying to model and forecast high-dimensional vectors of time series, the number of parameters to be estimated grows and the "curse of dimensionality" arises. In addition, if seasonality is present in the data, not only regular dynamics but also seasonal ones must be estimated, making the problem even greater.

This is why dimensionality reduction techniques for vectors of time series have been extensively studied. Sargent and Sims (1977) and Geweke (1977) were the first to propose a dynamic factor model. On the one hand, Stock and Watson (2002) explored dimensionality reduction in panel data used to explain one variable. On the other hand, Peña and Box (1987) proposed a simplifying structure for a vector of time series valid only for the stationary case. Lee and Carter (1992) extended Principal Component Analysis to the case in which the variables are time series, and computed long-run forecasts of mortality and fertility rates by means of extracting a single common factor. Most recently, Peña and Poncela (2004, 2006) extended the Peña–Box model to the nonstationary case.

However, there is no specific dimension reduction technique that can be applied to vectors of time series with a seasonal pattern. There are many examples of this kind of data, such as vectors of macroeconomic variables, meteorological data, and time series coming from electricity markets (load and prices). Until now, when reducing dimension in vectors of time series with seasonal behavior, the only possible alternative was to deseasonalize and then apply some of the methodology cited above to reduce the number of parameters to be estimated.

In this work, two contributions are introduced. First of all, Seasonal Dynamic Factor Analysis (hereafter referred to as SeaDFA) is presented. It allows the extraction of the common factors of a vector of time series, and the estimation of a seasonal multiplicative Vector Auto Regressive Integrated Moving Average (VARIMA) model, so that both regular and seasonal dynamics can be modeled. Second, with respect to inference procedures, we propose an alternative bootstrap scheme to those derived by Stoffer and Wall (1991) and Wall and Stoffer (2002), valid for all models that can be expressed under the state-space formulation. Bootstrap methods are considered for this purpose instead of other alternatives such as the Fisher Information Matrix (Shumway and Cavanaugh 1996), since asymptotic results are not applicable if time series are not fairly long or if the parameters fall near the boundary of the valid parameter space.

Moreover, an interesting real data example is provided: analysis of time series of electricity prices, which are relevant from both the economic and engineering points of view. In fact, a great number of recent references have focused on several complex problems affecting different agents involved in energy markets, such as load forecasting (Cottet and Smith 2003), wind energy forecasting (Sánchez 2006), and electricity price modeling (Koopman, Ooms, and Carnero 2007). In this work, SeaDFA is applied to compute year-ahead forecasts of electricity prices.

The rest of the article is organized as follows. In Section 2, the Seasonal Dynamic Factor Analysis (SeaDFA) and its estimation algorithm are introduced and explained. In Section 3, the bootstrap scheme developed for the SeaDFA is presented. In Section 4, a Monte Carlo simulation study is provided to check the behavior of the bootstrap procedure. In Section 5, the model and its bootstrap scheme are applied to forecasting electricity prices in the Spanish market. Finally, some conclusions are provided in Section 6.

2. SEASONAL DYNAMIC FACTOR ANALYSIS (SeaDFA)

In this section we present the Seasonal Dynamic Factor Analysis, allowing us to deal with common factors following a VARIMA$(p, d, q) \times (P, D, Q)_s$ model with constant.

2.1 The Model

Let $\mathbf{y}_t$ be an m-dimensional vector of observed time series generated by an r-dimensional vector of unobserved common factors ($r < m$). We assume that vector $\mathbf{y}_t$ can be written as a linear combination of the unobserved common factors, $\mathbf{f}_t$, plus $\boldsymbol{\varepsilon}_t$, to which we will refer from now on as specific components or specific factors:

$$\mathbf{y}_t = \boldsymbol{\Omega}\mathbf{f}_t + \boldsymbol{\varepsilon}_t, \tag{1}$$

where $\boldsymbol{\Omega}$ is the $m \times r$ loading matrix that relates the r-dimensional set of common factors to the vector of observed time series $\mathbf{y}_t$, and $\boldsymbol{\varepsilon}_t$ is the m-dimensional vector of specific components. The common dynamic structure of the m observed time series is included in the r common factors. We suppose that the specific components, $\boldsymbol{\varepsilon}_t$, are white noise. So the vector $\boldsymbol{\varepsilon}_t$ is Gaussian and has zero mean, $E[\boldsymbol{\varepsilon}_t\boldsymbol{\varepsilon}'_\tau] = \mathbf{0}$ if $t \neq \tau$, and diagonal covariance matrix $\mathbf{S} = E[\boldsymbol{\varepsilon}_t\boldsymbol{\varepsilon}'_t]$. It is straightforward to extend this model to the case in which the specific factors evolve over time according to univariate AR models following the ideas in the work of Bai (2003), Peña and Poncela (2004), and Ortega and Poncela (2005). We apply such a model in Section 5 to our analysis of electricity prices.

The unobserved common factors $\mathbf{f}_t$ can be nonstationary, including not only seasonal or regular unit roots but also a seasonal or regular autoregressive and/or moving average pattern. We assume that $\mathbf{f}_t$ follows a seasonal multiplicative VARIMA model $(p, d, q) \times (P, D, Q)_s$ with constant:

$$(\mathbf{I} - B)^d(\mathbf{I} - B^s)^D\boldsymbol{\phi}(B)\boldsymbol{\Phi}(B^s)\mathbf{f}_t = \mathbf{c} + \boldsymbol{\theta}(B)\boldsymbol{\Theta}(B^s)\mathbf{w}_t, \tag{2}$$

where $\boldsymbol{\phi}(B) = (\mathbf{I} - \boldsymbol{\phi}_1 B - \boldsymbol{\phi}_2 B^2 - \cdots - \boldsymbol{\phi}_p B^p)$, $\boldsymbol{\Phi}(B^s) = (\mathbf{I} - \boldsymbol{\Phi}_1 B^s - \boldsymbol{\Phi}_2 B^{2s} - \cdots - \boldsymbol{\Phi}_P B^{Ps})$, $\boldsymbol{\theta}(B) = (\mathbf{I} - \boldsymbol{\theta}_1 B - \boldsymbol{\theta}_2 B^2 - \cdots - \boldsymbol{\theta}_q B^q)$, and $\boldsymbol{\Theta}(B^s) = (\mathbf{I} - \boldsymbol{\Theta}_1 B - \boldsymbol{\Theta}_2 B^{2s} - \cdots - \boldsymbol{\Theta}_Q B^{Qs})$ are $r \times r$ polynomial matrices, B is the backshift operator such that $B\mathbf{y}_t = \mathbf{y}_{t-1}$, the roots of $|\boldsymbol{\phi}(B)| = 0$ and $|\boldsymbol{\Phi}(B^s)| = 0$ are outside the unit circle as well as the roots of $|\boldsymbol{\theta}(B)| = 0$ and $|\boldsymbol{\Theta}(B^s)| = 0$ are outside the unit circle, and $\mathbf{w}_t \sim \mathbf{N}_r(\mathbf{0}, \mathbf{Q})$ is serially uncorrelated, $E[\mathbf{w}_t\mathbf{w}'_{t-h}] = \mathbf{0}, h \neq 0$. We also assume that the noise term of the common factors and the observed series are uncorrelated for all lags, $E[\mathbf{w}_t\boldsymbol{\varepsilon}'_{t-h}] = \mathbf{0}, \forall h$. $\mathbf{c} = \mathbf{C}_1 \cdot \mathbf{1} = (c_1, \ldots, c_r)$ is the vector of constants of the model of the common factors, where matrix $\mathbf{C}_1 = \text{diag}(c_1, \ldots, c_r)$. Its inclusion in (2) could be relevant when trying to compute long-term forecasts in the nonstationary case, which is our purpose for the vector of nonstationary series of electricity prices.

It should be noted that the model is not identifiable under rotations, since for any $r \times r$ nonsingular matrix $\mathbf{H}$, the observed vector of time series can be expressed as a linear combination of a new set of factors, $\mathbf{y}_t = \boldsymbol{\Omega}^r\mathbf{f}^r_t + \boldsymbol{\varepsilon}_t$, where $\mathbf{f}^r_t = \mathbf{H}\mathbf{f}_t$ and $\boldsymbol{\Omega}^r = \boldsymbol{\Omega}\mathbf{H}^{-1}$. To solve this identification problem, we can always choose either $\mathbf{Q} = \mathbf{I}$ or $\boldsymbol{\Omega}'\boldsymbol{\Omega} = \mathbf{I}$. These kinds of restrictions are sufficient for the static case or even when there is a single common dynamic factor; otherwise the model is not yet identified, and we need to introduce an additional constraint to be able to estimate the model. Harvey (1989) imposed the condition $\omega_{ij} = 0$, for $j > i$, where $\boldsymbol{\Omega} = [\omega_{ij}]$. This condition is not restrictive since the factor model can be rotated for better interpretation when needed.

2.2 State-Space Formulation and Its Relationship With SeaDFA

The model presented in Section 2.1 will be estimated under the state-space formulation. For this reason it is necessary to provide the relationship between SeaDFA and state-space (SS hereafter) models. In general, a linear unobserved component model with exogenous variables and time-invariant system matrices can be written as a state-space model as follows:

$$\mathbf{x}_t = \mathbf{A}\boldsymbol{\alpha}_t + \mathbf{B}\boldsymbol{\beta}_t + \mathbf{C}\boldsymbol{\gamma}_t, \tag{3}$$

$$\boldsymbol{\alpha}_t = \mathbf{T}\boldsymbol{\alpha}_{t-1} + \mathbf{F}\boldsymbol{\beta}_t + \mathbf{G}\boldsymbol{\delta}_t. \tag{4}$$

Equation (3) is known as the measurement or observation equation and relates the observed m-dimensional series $\mathbf{x}_t$ with the k-dimensional latent or unobserved state $\boldsymbol{\alpha}_t$ components. $\mathbf{A}$ is the loading matrix, $\boldsymbol{\beta}_t$ is the vector of exogenous variables, and matrix $\mathbf{B}$ relates the vector of observed series with the vector of exogenous variables. The matrix $\mathbf{S}$ is assumed to be an $m \times m$ covariance matrix of the observation noise $\boldsymbol{\gamma}_t$, which is related to $\mathbf{x}_t$ by means of $\mathbf{C}$. The second equation, (4), is the transition equation. It relates the state-vector $\boldsymbol{\alpha}_t$ with the state vector at time $t - 1$ by means of the transition matrix $\mathbf{T}$. The matrix $\mathbf{Q}$ is assumed to be an $r \times r$ covariance matrix of the additive noise, $\boldsymbol{\delta}_t$, of the transition equation and it is related to $\boldsymbol{\alpha}_t$ by means of $\mathbf{F}$, and uncorrelated with $\boldsymbol{\gamma}_t$ at all leads and lags.

The system matrices $\mathbf{A}$, $\mathbf{B}$, $\mathbf{C}$, $\mathbf{T}$, $\mathbf{F}$, $\mathbf{G}$, $\mathbf{Q}$, and $\mathbf{S}$ are assumed to be predetermined in the sense that they are known at time $t - 1$, and since they are fixed the model is said to be time-invariant.

Bearing this in mind, (1) and (2) can be directly considered, respectively, as an observation equation without exogenous variables ($\mathbf{A} = \boldsymbol{\Omega}$ and $\mathbf{C} = \mathbf{I}$), a transition equation in which $\mathbf{T} = \boldsymbol{\Psi}$, $\mathbf{F} = \mathbf{C}_1$, and, for the SeaDFA, $\mathbf{G} = \mathbf{I}$. The VARIMA

model in (2) can be reformulated as a transition equation just writing the VARIMA model adequately, using the multivariate extension of the state-space formulation for ARIMA models proposed by Ansley and Kohn (1986):

$$\mathbf{y}_t = \mathbf{\Omega}\mathbf{f}_t + \boldsymbol{\varepsilon}_t, \tag{5}$$

$$\mathbf{f}_t = \mathbf{C}_1 \cdot \mathbf{1} + \mathbf{\Psi}\mathbf{f}_{t-1} + \mathbf{w}_t. \tag{6}$$

As a particular case, an exogenous variable of ones, $\boldsymbol{\beta}_t = \mathbf{1} = (1, 1, \ldots, 1)'$, will be introduced in the transition equation in order to estimate the constant in the model of the common factors. It is not necessary to include exogenous variables in the measurement equation that relates the vector of observed time series with the set of common factors and specific factors. Thus, in the SeaDFA, $\mathbf{B}$ is the null matrix and $\mathbf{F} = \mathbf{C}_1$.

Once the general formulation of state-space models given by (3) and (4) has been linked with the equations of the SeaDFA, (1)–(2), and only for illustration purposes, in the Supplementary Materials available online we provide an example of how to build the transition matrix $\mathbf{\Psi}$ in the particular case in which the observed series $\mathbf{y}_t$ present seasonality.

It should be noticed that some nonlinear constraints appear between the elements of the transition matrix, and this will be important in the estimation procedure described in the following subsection.

2.3 SeaDFA Estimation

The previous state-space formulation depends on a set of parameters $\mathbf{\Lambda} = \{\mathbf{c}, \mathbf{\Psi}, \mathbf{\Omega}, \mathbf{S}, \mathbf{Q}\}$ that must be estimated from the observed vector of time series. Also the common factors, $\mathbf{f}_t$, and the specific ones, $\boldsymbol{\varepsilon}_t$, must be estimated. Notice that for the case of SeaDFA we impose $\mathbf{Q} = \mathbf{I}_r$ because otherwise the model remains unidentified under rotations, so $\mathbf{Q}$ does not need to be estimated.

For estimation, maximum likelihood is used under the assumption that the initial state is normal, $\mathbf{f}_0 \sim \mathbf{N}_r(\boldsymbol{\mu}_0, \mathbf{P}_0^0)$, where $\boldsymbol{\mu}_0$ and $\mathbf{P}_0^0$ are the initial mean and covariance, and are both assumed to be known conditional on $\mathbf{\Lambda}$ [for extensions on initialization of the Kalman filter, see the book by Durbin and Koopman (2001)]. The errors $\boldsymbol{\varepsilon}_t$ and $\mathbf{w}_t$ are jointly normal and serially uncorrelated. Furthermore, for simplicity we also assume that $\boldsymbol{\varepsilon}_t$ and $\mathbf{w}_t$ have no cross-correlation.

In addition to nonstationarity included in the articles by Peña and Poncela (2004, 2006), in this work we include the possibility of common factors following a multiplicative seasonal VARIMA model with constant. Seasonality introduces some additional nonlinear constraints between parameters in the matrix $\mathbf{\Psi}$, as described in the example included in the online Supplementary Materials, and this implies that these nonlinear restrictions must be imposed when the estimation is carried out.

For simplicity we will explain the estimation algorithm for models without a moving average component.

In general, the log-likelihood for any model that can be expressed under SS formulation, ignoring a constant (see Shumway and Stoffer 2006, for details on SS models), is given by the expression

$$\log L(\mathbf{\Lambda}) = -\frac{1}{2}\sum_{t=1}^{n}\log|\mathbf{\Sigma}_t(\mathbf{\Lambda})| - \frac{1}{2}\sum_{t=1}^{n}\boldsymbol{\epsilon}_t(\mathbf{\Lambda})'(\mathbf{\Sigma}_t(\mathbf{\Lambda}))^{-1}\boldsymbol{\epsilon}_t(\mathbf{\Lambda}), \tag{7}$$

where $\boldsymbol{\epsilon}_t$ are the innovations and $\mathbf{\Sigma}_t$ is their variance–covariance matrix, which are obtained by running the Kalman filter (these recursions as well as those related to the Kalman smoother are given in Appendix A in the online Supplementary Materials). It is important to highlight the dependence of the innovations $\boldsymbol{\epsilon}_t$ in the parameters included in $\mathbf{\Lambda}$. The log-likelihood given by (7) is a complicated function that is highly nonlinear in the unknown parameters. Estimation proceeds by numerically maximizing $\log L(\mathbf{\Lambda})$, using a Newton–Raphson algorithm. The common factors, $\mathbf{f}_t$, will be obtained from the Kalman filter in the last iteration, thus $\widehat{\mathbf{f}}_t = \mathbf{f}_{t|t} = \mathbf{f}_t^t = E[\mathbf{f}_t|\mathbf{Y}_t]$, where $\mathbf{Y}_t = \{\mathbf{y}_1, \ldots, \mathbf{y}_t\}$. (See Appendix A in the online Supplementary Materials for the Kalman filter and smoother recursions.)

Direct maximization of (7) can be difficult in practice if there is a large number of parameters to be estimated. This is the case for our SeaDFA model. For an m-dimensional observed vector of data $\mathbf{y}_t$, with r common factors following a VARIMA$(p, d, 0) \times (P, D, 0)_s$, the parameters to be estimated and their dimensions are:

- the constant of the model of the common factors, $\mathbf{c}$, which is r by 1.
- the transition matrix $\mathbf{\Psi}$ including the dynamics of the common factors, where the number of parameters to be estimated is $r^2(p + P)$.
- the loading matrix $\mathbf{\Omega}$ that relates the observed vector of time series $\mathbf{y}_t$ to the unobserved set of common factors $\mathbf{f}_t$, whose dimensions are m by r. However, for $j > i$ $\omega_{ij} = 0$, so the number of elements to be estimated in $\mathbf{\Omega}$ is $mr - \frac{r(r-1)}{2}$.
- the variance–covariance matrix of the noise of the measurement equation (variance–covariance matrix of the specific components), $\mathbf{S}$, which is a diagonal matrix, so the number of parameters to be estimated is m.

Therefore, the number of parameters to be estimated in $\mathbf{\Lambda}$ is $r + r^2(p + P) + mr - \frac{r(r-1)}{2} + m$. Since m is usually large (this is the reason for reducing dimensionality to r), even for low orders of the regular and seasonal AR, p and P, and for a relatively small number of factors, r, say 2 or 3, the number of parameters to be estimated is high. An alternative to direct maximization of (7) is the Expectation–Maximization (EM) algorithm presented by Shumway and Stoffer (1982).

For both estimation procedures (direct optimization or EM algorithm), an important issue is the selection of the number of common factors, r, as well as the orders (d, D, p, and P) of the seasonal VARI model for the common factors. For a preliminary selection of the number of common factors, the tests proposed by Peña and Poncela (2006) or Forni et al. (2000) are used. A crucial stage is the selection of unit roots (regular and seasonal). In general, when dealing with data that present

a strong seasonal pattern, it can be useful to start considering a seasonal unit root, that is, $D = 1$. Thus, SeaDFA can be estimated considering r common factors (r is an output obtained by performing the tests mentioned above) that follow an $I(1)_s$ model. Once this has been done, the stationarity of the estimated specific components, $\widehat{\boldsymbol{\varepsilon}}_t = \mathbf{y}_t - \widehat{\boldsymbol{\Omega}}\widehat{\mathbf{f}}_t$, must be checked. If they are nonstationary the number of unit roots must be increased (trying, e.g., $d = 1$, $d = 2$, or even $D = 2$ if necessary).

When the number of regular and seasonal unit roots have been selected, the specific factors must be checked for cross-correlation. If present, this would indicate further common dynamic structure that should be incorporated in the common component. It could even point to a need to increase the number of common factors initially considered. Some iteration may be necessary to achieve specific factors with negligible cross-correlations.

2.4 EM Algorithm for SeaDFA

The main idea behind the EM algorithm is that if in addition to the observations $\mathbf{Y}_n = \{\mathbf{y}_1, \ldots, \mathbf{y}_n\}$ we could observe the state variables, that is, the common factors $\mathbf{F}_n = \{\mathbf{f}_0, \mathbf{f}_1, \ldots, \mathbf{f}_n\}$ in the particular case of the SeaDFA, then we would consider $\{\mathbf{Y}_n, \mathbf{F}_n\}$ as the complete data, and its joint density function would be given by the expression (see Shumway and Stoffer 2006, for details):

$$f_{\boldsymbol{\Lambda}}(\mathbf{Y}_n, \mathbf{F}_n) = f_{\boldsymbol{\mu}_0, \mathbf{P}_0^0}(\mathbf{f}_0) \prod_{t=1}^{n} f_{\boldsymbol{\Psi}, \mathbf{Q}}(\mathbf{f}_t | \mathbf{f}_{t-1}) \prod_{t=1}^{n} f_{\mathbf{S}}(\mathbf{y}_t | \mathbf{f}_t). \quad (8)$$

Assuming Gaussianity, the log-likelihood of the complete data $\{\mathbf{Y}_n, \mathbf{F}_n\}$ is given by the following expression (see Appendix B in the online Supplementary Materials for details):

$$\begin{aligned} \log L_{\mathbf{Y},\mathbf{F}}(\boldsymbol{\Lambda}) = -\frac{1}{2}\Bigg\{ & \ln|\mathbf{P}_0^0| + \operatorname{tr}((\mathbf{P}_0^0)^{-1}(\mathbf{f}_0 - \boldsymbol{\mu}_0)(\mathbf{f}_0 - \boldsymbol{\mu}_0)') \\ & + n \ln|\mathbf{S}| + \operatorname{tr} \sum_{t=1}^{n} \mathbf{c}\mathbf{c}' \\ & + \sum_{t=1}^{n} \operatorname{tr}((\mathbf{f}_t - \boldsymbol{\Psi}\mathbf{f}_{t-1})(\mathbf{f}_t - \boldsymbol{\Psi}\mathbf{f}_{t-1})') \\ & - 2 \cdot \operatorname{tr} \sum_{t=1}^{n} \mathbf{c}'(\mathbf{f}_t - \boldsymbol{\Psi}\mathbf{f}_{t-1}) \\ & + \sum_{t=1}^{n} \operatorname{tr}(\mathbf{S}^{-1}(\mathbf{y}_t - \boldsymbol{\Omega}\mathbf{f}_t)(\mathbf{y}_t - \boldsymbol{\Omega}\mathbf{f}_t)') \Bigg\}. \end{aligned} \quad (9)$$

The complete data $\{\mathbf{Y}_n, \mathbf{F}_n\}$ is not available since the state variables, the unobserved common factors $\mathbf{F}_t$ in the case of SeaDFA, must also be estimated. The EM algorithm provides an iterative method for finding the MLEs of $\boldsymbol{\Lambda}$, based on the incomplete data $\mathbf{Y}_n$ by successively maximizing the conditional expectation of the complete data likelihood. This procedure consists of maximizing the expression of the conditional expectation $E\{\log L_{\mathbf{Y},\mathbf{F}}(\boldsymbol{\Lambda})|\mathbf{Y}_n, \boldsymbol{\Lambda}^{(j-1)}\}$ with respect to the parameters. The vector $\widehat{\boldsymbol{\Lambda}}^{(j)} = \{\widehat{\mathbf{c}}^{(j)}, \widehat{\boldsymbol{\Psi}}^{(j)}, \widehat{\boldsymbol{\Omega}}^{(j)}, \widehat{\mathbf{S}}^{(j)}\}$ includes all parameters estimated at the jth iteration.

The algebraic derivations needed for the E-step and M-step are given in detail in Appendix B in the online Supplementary Materials. They are respectively:

- the final expression of the conditional expectation $E\{\log L_{\mathbf{Y},\mathbf{F}}(\boldsymbol{\Lambda})|\mathbf{Y}_n, \boldsymbol{\Lambda}^{(j-1)}\}$, and
- maximization of $E\{\log L_{\mathbf{Y},\mathbf{F}}(\boldsymbol{\Lambda})|\mathbf{Y}_n, \boldsymbol{\Lambda}^{(j-1)}\}$ with respect to the parameters that must be estimated.

The advantage of using the EM instead of direct maximization of (7) is that we obtain closed expressions for the estimates of $\boldsymbol{\Omega}$ and $\mathbf{S}$ ($\widehat{\boldsymbol{\Omega}}$ and $\widehat{\mathbf{S}}$, respectively), so the numerical optimization with the nonlinear constraints due to seasonality in common factors illustrated in the example presented in the online Supplementary Materials, only involves $r + r^2(p + P)$ parameters. However, the log-likelihood function given by (7) has $r + r^2(p + P) + mr - \frac{r(r-1)}{2} + m$ parameters, and taking into account that $m \gg r$, this implies a great reduction in the number of variables involved in the optimization procedure.

The E and M steps are repeated alternately until convergence in the log-likelihood is reached:

0. Initialize the procedure, giving initial values to $\boldsymbol{\Lambda}^{(0)} = \{\mathbf{c}^{(0)}, \boldsymbol{\Psi}^{(0)}, \boldsymbol{\Omega}^{(0)}, \mathbf{S}^{(0)}\}$. $\mathbf{Q}$ is fixed to be the identity matrix.

   For the subsequent iterations, $j = 1, 2, \ldots$.
1. Run the Kalman filter and smoother using the recursions given in (A.1)–(A.9) in Appendix A in the online Supplementary Materials, and get the value of the incomplete data log-likelihood, $\log L_{\mathbf{Y}}(\boldsymbol{\Lambda}^{(j-1)})$.
2. Perform the E-step, calculating $E\{\log L_{\mathbf{Y},\mathbf{F}}(\boldsymbol{\Lambda})|\mathbf{Y}_n, \boldsymbol{\Lambda}^{(j-1)}\} = E^{(j)}$, using (B.12) in Appendix B in the online Supplementary Materials.
3. Perform the M-step to update the estimate of the hyperparameter $\boldsymbol{\Lambda} = \{\mathbf{c}, \boldsymbol{\Psi}, \boldsymbol{\Omega}, \mathbf{S}\}$, getting $\boldsymbol{\Lambda}^{(j)} = \{\mathbf{c}^{(j)}, \boldsymbol{\Psi}^{(j)}, \boldsymbol{\Omega}^{(j)}, \mathbf{S}^{(j)}\}$, from (B.13)–(B.15) in Appendix B in the online Supplementary Materials, obtaining $\boldsymbol{\Omega}^{(j)}$ and $\mathbf{S}^{(j)}$, and maximizing (B.16) to get $\mathbf{c}^{(j)}$ and $\boldsymbol{\Psi}^{(j)}$.
4. If $E^{(j)} - E^{(j-1)} < \varepsilon$, with ε small enough, then stop. If convergence has not been reached, then steps 1, 2, and 3 are iteratively repeated.

3. BOOTSTRAP SCHEME FOR SEASONAL DYNAMIC FACTOR ANALYSIS

In this section we provide a bootstrap scheme for assessing uncertainty in the maximum likelihood estimates of parameters of our Seasonal Dynamic Factor model, as well as computing forecast intervals.

Furthermore, since SeaDFA is a particular case of a model that can be written using the state-space formulation (as shown in Section 2.1), this bootstrap scheme is able to assess the precision of estimates of any state-space model. This is an advantage, since a wide range of statistical and econometric models can be represented under this formulation. In fact, many authors have focused on estimation of time series models by state-space methods (see Harvey 1989 and Durbin and Koopman 2001).

Application of classical inference methods relying on asymptotic theory is subject to the availability of large datasets, as investigated by Ansley and Newbold (1980), among others. For

this reason bootstrap techniques are a powerful alternative to inference procedures based on the Fisher Information Matrix. Moreover, bootstrap methods have a large advantage because they allow the uncertainty due to parameter estimation to be taken into account, which enhances the coverage of the forecasting intervals.

The existence, under certain conditions, of asymptotic theory involving the consistency of parameter estimates obtained by maximum likelihood and state estimators obtained from the Kalman filter (see Ljung and Caines 1979 or Spall and Wall 1984) has allowed other authors (Stoffer and Wall 1991; Wall and Stoffer 2002; Rodríguez and Ruiz 2009) to be able to develop procedures for bootstrapping state-space models, resampling from the innovations and generating bootstrap replicas of the model under study using the innovation form representation (see Anderson and Moore 1979).

Our bootstrap procedure is an alternative to those of Stoffer and Wall (1991), Wall and Stoffer (2002), and Rodríguez and Ruiz (2009), which is not based on the innovations form representation and allows for the computation of percentile-based forecasting intervals, as will be explained in Section 3.2.

3.1 Inference on the Parameters of the SeaDFA

By means of the new bootstrap procedure we will obtain percentile-based confidence intervals for each element in the loading matrix $\mathbf{\Omega}$, as well as for the parameters of the VARIMA model for the common factors, $\mathbf{\Psi}$, the constant $\mathbf{c}$, and the variance–covariance matrix of the specific factors, $\mathbf{S}$. We will be able to test the significance of the elements in these matrices.

The bootstrap scheme consists of the following steps:

1. The model defined by (5)–(6) is estimated following the EM algorithm described in Section 2. Once this has been completed, the parameters involved, $\widehat{\mathbf{c}}, \widehat{\mathbf{\Psi}}, \widehat{\mathbf{\Omega}}, \widehat{\mathbf{S}}, \widehat{\boldsymbol{\mu}}_0, \widehat{\mathbf{P}}_0^0$, are available. In addition, we have consistent estimates, $\widehat{\mathbf{f}}_t$, of the state variables $\mathbf{f}_t$, derived from the Kalman filter at the last iteration [conditions for consistency in the estimation of the latent state-variables are given in the article by Bai (2003)].
2. Obtain the estimated specific factors, $\widehat{\boldsymbol{\varepsilon}}_t = \mathbf{y}_t - \widehat{\mathbf{\Omega}}\widehat{\mathbf{f}}_t$, and the estimated residuals of the transition equation, $\widehat{\mathbf{w}}_t = \widehat{\mathbf{f}}_t - \widehat{\mathbf{\Psi}}\widehat{\mathbf{f}}_{t-1}$.
3. Standardize and rescale the estimated specific factors, $\widehat{\boldsymbol{\varepsilon}}_t$, as well as the estimated residuals of the transition equation, $\widehat{\mathbf{w}}_t$.

One should bear in mind the relationship between $\widehat{\boldsymbol{\varepsilon}}_t$ and $\boldsymbol{\varepsilon}_t$, and their variance–covariance matrices, $\mathbf{S} = E[\boldsymbol{\varepsilon}_t\boldsymbol{\varepsilon}_t']$ and $\mathbf{S}_{\widehat{\boldsymbol{\varepsilon}}} = E[\widehat{\boldsymbol{\varepsilon}}_t\widehat{\boldsymbol{\varepsilon}}_t']$. It can be shown that $\mathbf{S} = \mathbf{S}_{\widehat{\boldsymbol{\varepsilon}}} + \operatorname{var}(\boldsymbol{\varepsilon}_t - \widehat{\boldsymbol{\varepsilon}}_t) = \mathbf{S}_{\widehat{\boldsymbol{\varepsilon}}} + \mathbf{S}_{\text{correction}}$. The matrix $\mathbf{S}_{\text{correction}} = \operatorname{var}(\boldsymbol{\varepsilon}_t - \widehat{\boldsymbol{\varepsilon}}_t)$ can be considered as a *correction factor*, as in the model proposed by Harvey, Ruiz, and Sentana (1992). An expression for $\operatorname{var}(\boldsymbol{\varepsilon}_t - \widehat{\boldsymbol{\varepsilon}}_t)$ is derived in Appendix C in the online Supplementary Materials. The same correction applies to the relationship between $\widehat{\mathbf{w}}_t$ and $\mathbf{w}_t$:

$$\begin{aligned} \widetilde{\boldsymbol{\varepsilon}}_t &= (\widehat{\boldsymbol{\varepsilon}}_t - \overline{\boldsymbol{\varepsilon}})(\mathbf{S}_{\widehat{\boldsymbol{\varepsilon}}})^{-1/2}\widehat{\mathbf{S}}^{1/2}, \\ \widetilde{\mathbf{w}}_t &= (\widehat{\mathbf{w}}_t - \overline{\mathbf{w}})(\mathbf{Q}_{\widehat{\mathbf{w}}})^{-1/2}, \end{aligned} \tag{10}$$

where $\mathbf{Q}_{\widehat{\mathbf{w}}} = E[\widehat{\mathbf{w}}_t\widehat{\mathbf{w}}_t']$. Note that for SeaDFA, $\widehat{\mathbf{Q}} = \mathbf{I}$. In a general SS model, (10) would be $\widetilde{\mathbf{w}}_t = (\widehat{\mathbf{w}}_t - \overline{\mathbf{w}})(\mathbf{Q}_{\widehat{\mathbf{w}}})^{-1/2}\widehat{\mathbf{Q}}^{1/2}$.

4. Draw an *iid* resample, $\varepsilon_{e,t}^*$, from $F_{\widetilde{\varepsilon}_e}$, for $e = 1, \ldots, m$, where $F_{\widetilde{\varepsilon}_e}$ is the empirical distribution function of each corrected specific factor, $F_{\widetilde{\varepsilon}_e}(x) = \frac{1}{n}\sum_{t=1}^{n} I(\widetilde{\varepsilon}_{e,t} \le x)$.
5. Draw an *iid* resample, $w_{i,t}^*$, from $F_{\widetilde{w}_i}$, for $i = 1, \ldots, r$, where $F_{\widetilde{w}_i}$ is the empirical distribution function of each corrected series of residuals of the model for the common factors, $F_{\widetilde{w}_i}(x) = \frac{1}{n}\sum_{t=1}^{n} I(\widetilde{w}_{i,t} \le x)$.
6. Build a bootstrap replica of the common factors using the transition equation: $\mathbf{f}_t^* = \widehat{\mathbf{\Psi}}\mathbf{f}_{t-1}^* + \mathbf{w}_t^*$, where $\mathbf{w}_t^* = (w_{1,t}^*, \ldots, w_{r,t}^*)'$.
7. Build a bootstrap replica of the data, using bootstrap replicas of the common and specific factors obtained in steps 4 and 6, $\mathbf{f}_t^*$ and $\boldsymbol{\varepsilon}_t^*$, respectively: $\mathbf{y}_t^* = \widehat{\mathbf{\Omega}}\mathbf{f}_t^* + \boldsymbol{\varepsilon}_t^*$, where $\boldsymbol{\varepsilon}_t^* = (\varepsilon_{1,t}^*, \ldots, \varepsilon_{m,t}^*)'$.
8. Repeat steps 4 to 7, N times, where N is the number of bootstrap replicates.

Notice that although both $\widehat{\boldsymbol{\varepsilon}}_t$ and $\widehat{\mathbf{w}}_t$ could be dependent, the bootstrap replicas $\varepsilon_{e,t}^*$ and $w_{i,t}^*$ are not, since they are obtained by sampling independently from the respective empirical distributions, $F_{\widetilde{\varepsilon}_e}(x)$ and $F_{\widetilde{w}_i}(x)$. This approach also holds for simpler models, as the estimated residuals can present correlation; see the article by Ljung (1986). Sampling independently (*iid*) from the empirical distribution function is the usual approach to guarantee independence in the bootstrap replicas [see, for instance, the work of Thombs and Shucany (1990) and Pascual, Romo, and Ruiz (2004)]. Moreover, when sampling from the empirical distribution functions, $F_{\widetilde{\varepsilon}_e}(x)$ and $F_{\widetilde{w}_i}(x)$, we are doing it for each $e = 1, \ldots, m$ and each $i = 1, \ldots, r$, respectively. In this way we are not only ensuring that $\boldsymbol{\varepsilon}_t^*$ and $\boldsymbol{\varepsilon}_s^*$ are independent, but also $\varepsilon_{e,t}^*$ and $\varepsilon_{i,t}^*$, for a given t. The same holds for the $\mathbf{w}_t^*$.

Estimating the SeaDFA for each of the N replicas obtained in step 8, we have $\widehat{\mathbf{c}}^*$, $\widehat{\mathbf{\Psi}}^*$, $\widehat{\mathbf{\Omega}}^*$, and $\widehat{\mathbf{S}}^*$ and their respective bootstrap distribution functions, $\widehat{F}_{\widehat{\mathbf{c}}^*}^*$, $\widehat{F}_{\widehat{\mathbf{\Psi}}^*}^*$, $\widehat{F}_{\widehat{\mathbf{\Omega}}^*}^*$, and $\widehat{F}_{\widehat{\mathbf{S}}^*}^*$. They are used to compute percentile-based confidence intervals for all these parameters using the following expression: $[q^*(\alpha/2), q^*(1-\alpha/2)]$, where, for example, when calculating intervals for the elements c_i in the constant of the model, $\mathbf{c} = (c_1, \ldots, c_r)'$, $q^*(\cdot) = \widehat{F}_{\widehat{c}_i^*}^{*-1}$. Finally, bootstrap confidence intervals for the loads, ω_{ij}, and VARIMA parameters, Ψ_{ij}, are obtained from the corresponding bootstrap distribution functions, $F_{\omega_{i,j}^*}^*$ and $F_{\Psi_{i,j}^*}^*$, of the elements (i,j) of matrices $\mathbf{\Omega}^*$ and $\mathbf{\Psi}^*$, respectively.

The percentile-based confidence intervals for loads and VARIMA parameters will allow, for example, for testing the equality of loads, or whether the parameters of the VARIMA model of the common factors are significant. The results obtained can be used to impose constraints among loads or VARIMA parameters that can be applied in a subsequent estimation of the SeaDFA.

3.2 Bootstrap Procedure for Forecasting

As far as forecasting is concerned, the main objective is to obtain not only point forecasts but also an uncertainty measure. Bootstrap techniques have been applied for this purpose (Thombs and Schucany 1990; García-Jurado et al. 1995; Alonso, Peña, and Romo 2002). The previous scheme introduced in Section 3.1 can be modified if we want to obtain

bootstrap confidence intervals for the forecasts of vector $\mathbf{y}_t$. The conditional distribution of future observations given the observed vector of time series should be replicated. We proceed as in the article by Cao et al. (1997), fixing the last f observations of the common factors. In the particular case of the common factors following a VARI$(p,d)\times(P,D)_s$, $f=(P+D)s+(p+d)$. In this way, we generate bootstrap trajectories of the future observations conditioning on the last f observed values. The first steps of the bootstrap procedure for forecasting coincide with steps 1 to 7 proposed in the previous subsection. The forecasting steps are the following:

8. Each bootstrap replica of future values for common factors is calculated using the relationship: $\mathbf{f}^*_{t+h}=\widehat{\boldsymbol{\Psi}}^*\times\mathbf{f}^*_{t+h-1}+\mathbf{w}^*_{t+h}$, for $h=1,\ldots,H$, where H is the forecasting horizon, $\mathbf{f}^*_n=\widehat{\mathbf{f}}_n$, and $\mathbf{w}^*_{t+h}=(w^*_{1,t+h},w^*_{2,t+h},\ldots,w^*_{m,t+h})'$ and each $w^*_{i,t+h}$ is generated by resampling from $F_{\widetilde{w}_i}$. The bootstrap resamples of future specific components $\boldsymbol{\varepsilon}^*_{t+h}=(\varepsilon^*_{1,t+h},\varepsilon^*_{2,t+h},\ldots,\varepsilon^*_{m,t+h})'$ are generated by resampling from $(F_{\widetilde{\varepsilon}_1},\ldots,F_{\widetilde{\varepsilon}_m})$, respectively.
9. Future bootstrap observations are calculated for vector $\mathbf{y}_t$ using the relationship $\mathbf{y}^*_{t+h}=\widehat{\boldsymbol{\Omega}}^*\mathbf{f}^*_{t+h}+\boldsymbol{\varepsilon}^*_{t+h}$, for $h=1,\ldots,H$.
10. Repeat steps 8 and 9, N times, where N is the number of future bootstrap observations for forecasting horizons varying from 1 to H, that are calculated for the vector $\mathbf{y}_t$.

Finally the bootstrap distribution function of $\mathbf{y}^*_{t+h}$ is used as estimator of the conditional distribution of $\mathbf{y}_{t+h}$ given the observed sample. Bootstrap confidence intervals for $\mathbf{y}_{t+h}=(y_{1,t+h},\ldots,y_{m,t+h})$ are obtained using the quantiles of the corresponding bootstrap distribution functions. The $(1-\alpha)\cdot100\%$ forecast interval for $y_{k,t+h}$ is $[q^*(\alpha/2),q^*(1-\alpha/2)]$, where $q^*(\cdot)=\widehat{F}^{*-1}_{y^*_{k,t+h}}$ are the quantiles of the estimated bootstrap distribution. Additionally, the prediction steps of our method can be modified in order to compute conditional forecast errors. This can be done by subtracting the conditional forecast and the future bootstrap observation generated using steps 8 to 10. The computation of the conditional forecast can be done by imposing $\boldsymbol{\varepsilon}^*_{t+h}=0$ and $\mathbf{w}^*_{t+h}=0$.

4. SIMULATION STUDY

In this section the performance of the bootstrap procedure introduced in the previous section is illustrated by means of a Monte Carlo simulation study. We also check the performance of the SeaDFA. We report the results for the following two models, which have been selected to check the behavior of the bootstrap scheme under different conditions:

- *Model 1*: In the first experiment we consider a model with a common nonstationary factor [common trend, $I(1)$ with constant, $c=3$], for $m=20$ observed series. The loading matrix is $\boldsymbol{\Omega}=(1,1,1,1,2,1,1,1,1,-0.5,1,1,1,1,1,1,1,1,1,1)'$, $E(\boldsymbol{\varepsilon}_t)=\mathbf{0}$, $\mathrm{var}(\boldsymbol{\varepsilon}_t)=\mathbf{S}=0.1\cdot\mathbf{I}_{20}$ and $E(\mathbf{w}_t)=\mathbf{0}$, $\mathrm{var}(\mathbf{w}_t)=\mathbf{Q}=\sigma^2_w=1$. This model has been selected because it is similar to the model that appears in the article by Peña and Poncela (2004), and we have added the constant to validate its estimation, since we have included this possibility in our model.
- *Model 2*: In the second model under study we check the performance of our procedure when there is a seasonal pattern affecting two common factors. There are two common nonstationary factors following a seasonal multiplicative VARIMA model:

$$[\mathbf{I}-B^7]\left[\mathbf{I}-\begin{pmatrix}0.1&0\\0&-0.15\end{pmatrix}B^7\right]\times\left[\mathbf{I}-\begin{pmatrix}0.4&0\\0&0.6\end{pmatrix}B\right]\mathbf{f}_t=\mathbf{w}_t$$

for $m=25$ observed series, the loading matrix $\boldsymbol{\Omega}=[\boldsymbol{\Omega}_1\ \boldsymbol{\Omega}_2]$, where $\boldsymbol{\Omega}_1=(1,1,1,1,2,1,1,1,1,-0.5,1,1,1,1,1,1,1,1,1,1,1,1)'$ and $\boldsymbol{\Omega}_2=0.3\cdot(0,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,3)'$, $E(\boldsymbol{\varepsilon}_t)=\mathbf{0}$, $\mathrm{var}(\boldsymbol{\varepsilon}_t)=0.1\cdot\mathbf{I}_{25}$, $E(\mathbf{w}_t)=\mathbf{0}_{2\times1}$, $\mathrm{var}(\mathbf{w}_t)=\mathbf{I}_2$.

$R=100$ realizations of each model have been generated and estimated, and $P=1000$ future values $\mathbf{y}_{T+h}$ have been generated for different forecasting horizons, $h=1$ and 10, while two sample sizes, T, have been considered: 50 and 200. For each vector of series simulated and estimated $N=500$ bootstrap resamples have been generated as described in the previous section, and the corresponding model was estimated. For $k=1,\ldots,m$, the $(1-\alpha)\cdot100\%$ prediction intervals $[Q^*_M(\alpha/2),Q^*_M(1-\alpha/2)]$ were computed. The coverage is estimated as $C_M=\#\{Q^*_M(\alpha/2)\le y^P_{k,T+h}\le Q^*_M(1-\alpha/2)\}$, where $y^P_{k,T+h}$ is the vector of future values generated in the first step. Meanwhile, using $L_T=y^{P(1-\alpha/2)}_{k,T+h}-y^{P(\alpha/2)}_{k,T+h}$ and $L_B=Q^*_M(1-\alpha/2)-Q^*_M(\alpha/2)$, we obtain the "theoretical" and bootstrap interval lengths. L_T is the estimated "true" mean interval length, and C_T is the nominal coverage.

The results for Models 1 and 2, and nominal coverage of 95%, are shown in Table 1. Since the number of series, m, is 20 in Model 1 and 25 in Model 2, we provide results only for some of the observed series, chosen after consideration of the structure of the loading matrix $\boldsymbol{\Omega}$. It can be observed that even for small or moderate sample sizes the values obtained for coverages are close to the nominal coverage, even in Model 2 in which the transition matrix must be estimated and nonlinear constraints appear between its elements. As expected, when the sample size increases, coverages are nearer to their nominal values. Although we focused on coverages for the forecasts, the results obtained show that the coverages for the parameters of the model are also correct.

Our scheme incorporates the uncertainty due to parameter estimation by reestimating the SeaDFA for each of the N bootstrap replicates generated in step 8 of Section 3.1. Although some authors found that incorporating this uncertainty can improve the results (Alonso, Peña, and Romo 2002; Pascual, Romo, and Ruiz 2004), for the two models presented in this section there are no significant differences (in terms of coverages) between reestimating and not the SeaDFA for each bootstrap replicate. This means that under some circumstances we can benefit from some computational savings, not attempting to account for parameter uncertainty. This can be done by generating

$$\mathbf{f}^*_{t+h}=\widehat{\boldsymbol{\Phi}}\mathbf{f}^*_{t+h-1}+\mathbf{w}^*_{t+h},$$
$$\mathbf{y}^*_{t+h}=\widehat{\boldsymbol{\Omega}}\mathbf{f}^*_{t+h}+\boldsymbol{\varepsilon}^*_{t+h},$$

Table 1. Models 1 and 2. Nominal coverages 95 percent

	Horizon h	Sample size T	Series j	C_M Theoret. 95%	(se) C_M	Cov (below) 2.5%	Cov (above) 2.5%	L_T	L_M	(se) L_M
Model 1	1	50	1	93.742	0.203	3.136	3.122	4.120	4.064	0.027
			5	93.828	0.226	3.214	2.958	7.920	7.881	0.061
			10	93.392	0.209	3.202	3.406	2.334	2.257	0.014
			15	94.166	0.185	2.988	2.846	4.107	4.102	0.026
			20	93.750	0.199	3.164	3.086	4.120	4.055	0.026
		200	1	94.352	0.174	2.888	2.760	4.089	4.059	0.019
			5	94.292	0.184	2.892	2.816	7.921	7.826	0.039
			10	94.306	0.167	2.832	2.862	2.328	2.296	0.011
			15	94.320	0.159	2.972	2.708	4.107	4.067	0.021
			20	94.158	0.186	3.024	2.818	4.092	4.038	0.020
	10	50	1	91.756	0.406	4.318	3.926	12.487	12.235	0.054
			5	91.704	0.435	4.336	3.960	24.811	24.379	0.107
			10	91.608	0.449	3.944	4.448	6.339	6.220	0.026
			15	91.844	0.419	4.344	3.812	12.459	12.220	0.054
			20	91.700	0.433	4.400	3.900	12.450	12.239	0.051
		200	1	94.238	0.165	3.090	2.672	12.480	12.370	0.054
			5	94.154	0.171	3.156	2.690	24.813	24.622	0.099
			10	94.256	0.174	2.760	2.984	6.300	6.275	0.027
			15	94.184	0.167	3.158	2.658	12.443	12.358	0.048
			20	94.668	0.139	2.670	2.662	6.910	6.894	0.027
Model 2	1	50	1	92.258	0.280	3.748	3.994	4.138	3.975	0.034
			5	92.290	0.287	3.804	3.906	8.081	7.782	0.070
			10	91.614	0.329	4.426	3.960	2.621	2.511	0.014
			15	93.328	0.200	3.654	3.018	4.278	4.177	0.030
			25	93.214	0.207	4.048	2.738	5.431	5.300	0.026
		200	1	94.360	0.179	2.842	2.798	4.136	4.105	0.020
			5	94.516	0.185	2.784	2.700	8.031	8.019	0.046
			10	93.384	0.293	3.206	3.410	2.592	2.563	0.010
			15	94.448	0.158	2.858	2.694	4.283	4.255	0.020
			25	94.106	0.182	2.944	2.950	5.409	5.411	0.023
	10	50	1	93.178	0.208	3.500	3.322	4.436	4.308	0.031
			5	93.152	0.213	3.606	3.242	8.700	8.439	0.059
			10	92.374	0.261	3.780	3.846	2.875	2.719	0.015
			15	93.456	0.188	3.562	2.982	4.650	4.504	0.025
			25	93.004	0.206	3.666	3.330	6.182	5.890	0.033
		200	1	94.400	0.175	2.926	2.674	4.448	4.415	0.025
			5	94.254	0.168	2.868	2.878	8.780	8.673	0.047
			10	94.148	0.170	2.938	2.914	2.855	2.839	0.014
			15	94.164	0.158	2.954	2.882	4.695	4.610	0.026
			25	94.264	0.158	2.850	2.886	6.169	6.115	0.029

that is, replacing $\widehat{\mathbf{\Phi}}^*$ by $\widehat{\mathbf{\Phi}}$ and $\widehat{\mathbf{\Omega}}^*$ by $\widehat{\mathbf{\Omega}}$ in the scheme we initially provided in Section 3.2.

Bearing in mind that for the models here considered there are not significant differences between the results obtained with the reduced scheme that allows reducing computational costs and the complete one explained in Section 3.2, an additional step consisting of running the filter with the estimates of the current bootstrap replica and the actual data values [as an extension in the direction of the proposal by Rodríguez and Ruiz (2009)] might not be necessary. Comparing our results in terms of coverages with those provided by Rodríguez and Ruiz (2009), we can state that ours for SeaDFA are similar to theirs for the Univariate Local Level model. Although it is beyond the scope of this article, it would be of interest for further research to compare for different univariate, multivariate, and factor models, the performance of the three aforementioned schemes.

5. FORECASTING ELECTRICITY PRICES IN THE SPANISH MARKET

In this section the SeaDFA and its bootstrap scheme are applied to compute point forecasts and forecast intervals for electricity prices in the Spanish market.

Currently, electricity is traded under competitive rules, as is the case for other *commodities*, and this has opened a new field of research. The special features that electricity presents (nonstorability and the need to satisfy demand instantaneously) are

responsible for the largely unpredictable behavior of its price. This, together with the great importance of this strategic sector in the national economy, supports the need to develop specific models for predicting electricity prices. Moreover, from the engineering point of view, the availability of accurate forecasts of electricity prices allows the appropriate scheduling of generation units.

Before the liberalization of electricity markets, demand was the only variable of interest, and a lot of works focused on this issue (Cottet and Smith 2003), although in the current context of liberalized markets, forecasting electricity prices, both in the short and long run, is of great interest.

Short-term or one-day-ahead forecasting is useful for planning the production of the generation units, minimizing costs, and improving bidding strategies so as to maximize profits. Some well-known references in this field include the articles by Nogales et al. (2002), Contreras et al. (2003), and Conejo et al. (2005), all of which used time series models to produce one-step-ahead forecasts for electricity prices in some weeks, in both the Spanish and the Californian markets. Koopman, Ooms, and Carnero (2007) analyzed spot prices coming from several European markets. Recently, García-Martos, Rodríguez, and Sánchez (2007) provided a computational experiment to obtain the combination of univariate time series models with the best global performance in the period under study, 1998–2003.

Medium- and long-term forecasting (where the forecasting horizon is between one month and one year) are useful to reduce the risk that every bilateral contract implies. By means of bilateral contracts, customers and generators can agree to trade a certain amount of power at a certain price. However, every contract implies a risk since the seller must purchase the amount of energy agreed for every day in the Pool. Having accurate long-term forecasts (covering at least the length of the bilateral contract) is crucial to maximize profits and/or reduce risks. Nevertheless, there are very few published works concerning long-term forecasting of electricity prices. Vehviläinen and Pyykkonen (2005) provided medium-term forecasts for monthly electricity prices in the NordPool. They included exogenous variables that affect the prices, such as temperature. Conejo et al. (2010) carried out a discretization which consisted of considering for each month four peak loads and four base loads, that is, they predicted 48 values per year and incorporated additional information on financial derivatives. By doing so, they obtained year-ahead forecasts for the prices in the German Market (the European Energy Exchange, EEX) and generated realistic scenarios that characterize all the possible realizations of the generating process of the prices. The novelty of this work is the long forecasting horizon. They introduced the forward prices as an explanatory variable, but these products are only well developed in a few electricity markets, such as the EEX in Germany. Finally, with respect to factor analysis and electricity markets, Frestad (2008) demonstrated that electricity swap returns could be explained by a set of uncorrelated common and unique risk factors in the Scandinavian market.

Here, we have selected the Spanish market, which according to several authors (Nogales et al. 2002; Contreras et al. 2003) is less predictable than, for instance, the Pennsylvania–Jersey–Maryland (PJM) interconnection or the NordPool, due to its higher proportion of outliers and a lesser degree of competition as well as the fact that during peak hours the Spanish market shows an even higher dispersion. This fact causes more uncertainty in periods of high demand, producing less accurate forecasts.

For these reasons computing long-term forecasts for electricity prices in the Spanish market presents a good challenge for testing the performance of the SeaDFA, and also a novelty since there are no related published works on long-term forecasting for this market. Our objective is to compute forecasts for every hour in the year 2004 using data from 1st January 1998 through 31st December 2003, so that the forecasting horizon ranges from one day up to one year.

Moreover, in some previous works (Angelus 2001) the importance of calculating not only point forecasts but also their uncertainty is emphasized. In our work the bootstrap procedure is applied to compute forecasting intervals for the prices in the period under study, the whole year 2004. In practice, this is crucial for long-term risk management of the utilities.

5.1 Estimation of SeaDFA for Electricity Prices in the Spanish Market

A 24-dimensional vector of time series can be built when considering the series of prices in the 24 hours of each day. This is known as the parallel approach (Grady et al. 1991; Cottet and Smith 2003; Smith and Cottet 2006).

In this subsection we provide the results obtained when modeling this 24-dimensional vector of prices. The SeaDFA was estimated for centered transformed prices in the period 1998–2003 in the Spanish zone of the Iberian market, and the transformed prices $\mathbf{y}_t$ are calculated from the original prices $\mathbf{p}_t$ as $\mathbf{y}_t = \log(\mathbf{p}_t + K)$. The constant K is added to avoid taking logs of a zero price. Although it very rarely occurs, the marginal prices in the liberalized Iberian power market could be zero. This is due to the generation technologies that are in the base of the load curve (wind and nuclear power plants in this case). In Spain it is regulated by law that all the energy produced in the wind farms must be dispatched. Nuclear power plants could also offer the energy they produce at zero price since they cannot be stopped and restarted quickly due to technical and safety reasons. Of course the forecasting errors computed in the following subsections were calculated in the original untransformed space.

For the transformed prices, $\mathbf{y}_t$, we first select the number of common factors as well as their multivariate model. Then the estimation is carried out and parameters of the SeaDFA are obtained. Inference is done by means of the new bootstrap scheme developed.

Seasonality must be dealt with when using this vector of series from electricity market data, which in this case is weekly ($s = 7$). Additionally, the yearly seasonality could be considered, as was done by Sáfadi and Peña (2008), who extracted this kind of seasonality by means of a linear combination of sines and cosines. All the empirical applications presented in this section were also carried out following this procedure, but there was no significant improvement in terms of prediction error. The results obtained were similar or even slightly worse than those presented here (without removing the yearly seasonality). Six years of data are used and the estimation of the coefficients of the sines and cosines may lack sufficient precision.

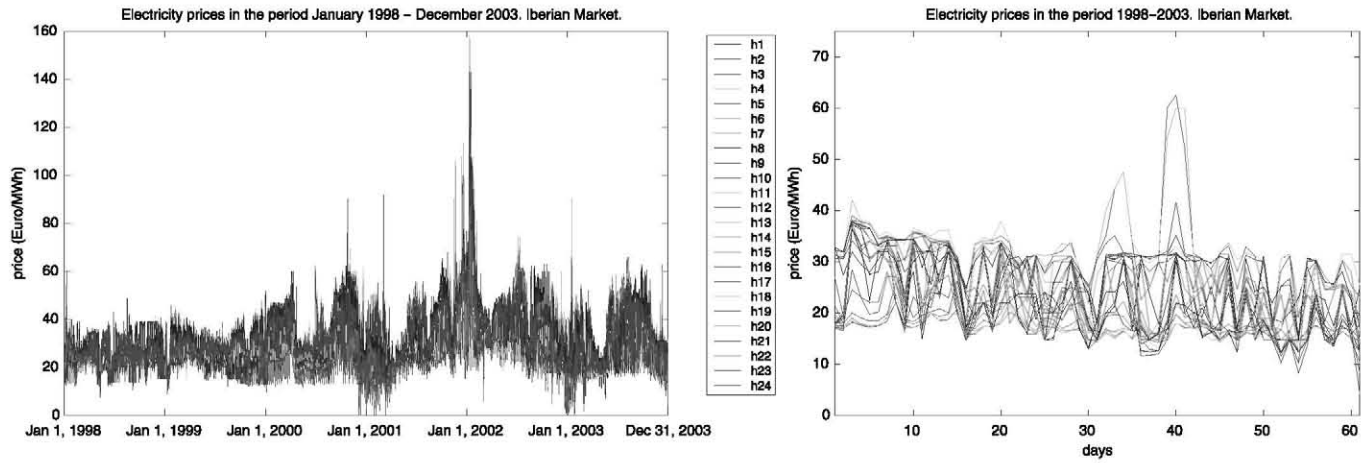

Figure 1. 24 hourly time series of prices in the period 1998–2003 and detail of the months November–December 2003.

Figure 1 shows the 24 hourly time series considered in the period 1998–2003, as well as a detail of the last two months in 2003. A common dynamics in the conditional mean of hourly prices can be observed.

First, it is important to decide on the number of common factors, r. For this purpose the test proposed by Peña and Poncela (2006) is used. We performed their test sequentially, increasing r at each step, and for lags $k = 1, \ldots, 5$, for our vector of 24 time series. For each lag k we rejected a maximum of zero common factors, and therefore there is at least one common factor that is very persistent (probably nonstationary); a second factor also appears for all the lags, indicating the possibility of a second common nonstationary factor (its autocorrelation does not die faster than for the first factor). Therefore, the number of common factors is two.

Apart from selecting the number of common factors, with the decision based on the results of the Peña and Poncela test, it is also necessary to choose the model for these common factors. We have fitted a VARIMA$(1,0,0) \times (1,1,0)_7$. Given the fact that the common factors are nonstationary, we fix a seasonal unit root, based on the importance of seasonality in electricity market data. Moreover, a single AR lag either in the regular or seasonal part was not able to capture all the common dynamics, so both are needed. The equation of the VARIMA model $(1,0,0) \times (1,1,0)_7$ is

$$\begin{aligned}[\mathbf{I} - B^7]\left[\mathbf{I} - \begin{pmatrix} \Phi_{1,11} & \Phi_{1,12} \\ \Phi_{1,21} & \Phi_{1,22} \end{pmatrix} B^7\right] \\ \times \left[\mathbf{I} - \begin{pmatrix} \phi_{1,11} & \phi_{1,12} \\ \phi_{1,21} & \phi_{1,22} \end{pmatrix} B\right] \begin{pmatrix} f_{1,t} \\ f_{2,t} \end{pmatrix} \\ = \begin{pmatrix} c_1 \\ c_2 \end{pmatrix} + \begin{pmatrix} w_{1,t} \\ w_{2,t} \end{pmatrix}. \end{aligned} \tag{11}$$

We have estimated this model, and used the bootstrap procedure previously described to make inference on the parameters involved. We have detected that the constants c_1 and c_2 are not significant, since their 95% confidence intervals are respectively $[-0.1254, 0.0683]$ and $[-0.0409, 0.0352]$.

The model is re-estimated including the previous result on significance of the constant, that is, imposing $(c_1, c_2)' = (0,0)'$, and the coefficient $\phi_{1,21}$ that relates the first common factor $f_{1,t}$ to the first lag of the second factor is not significant, as is shown in Figure 2. Once again the test for significance is percentile-based, using the bootstrap distribution function.

When the SeaDFA is again re-estimated including the constraint $(c_1, c_2)' = (0,0)'$ and $\phi_{1,21} = 0$, all the other coefficients remain significant. We present the loading matrix, $\widehat{\boldsymbol{\Omega}}$, obtained in Figure 3(a). There is a clear relationship between hourly loads and the boxplot of hourly prices as shown in Figure 3(b) as well.

In Figure 4(a), the common factors are provided for the period October–December 2003, and the grid has been placed to indicate when a week starts (Monday), to be able to interpret in terms of the day of the week. The first common factor is seasonal (of course, one seasonal root is present), but the differences between weekdays and weekends are much more important in the second one.

The part of the ith hourly time series explained by the first common factor is obtained by multiplying ω_{i1} by $f_{1,t}$, and the same holds for the part explained by the second common factor, which is obtained as $\omega_{i2} f_{2,t}$. According to Figure 3(b) we have

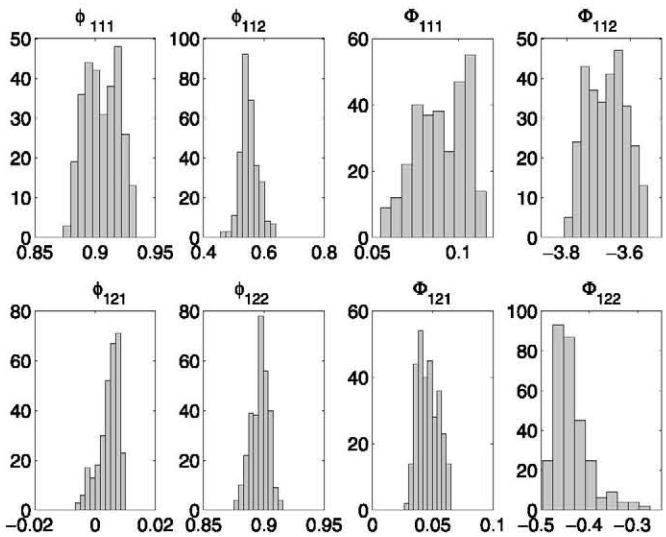

Figure 2. Histogram for bootstrap replicates of the parameters of the models.

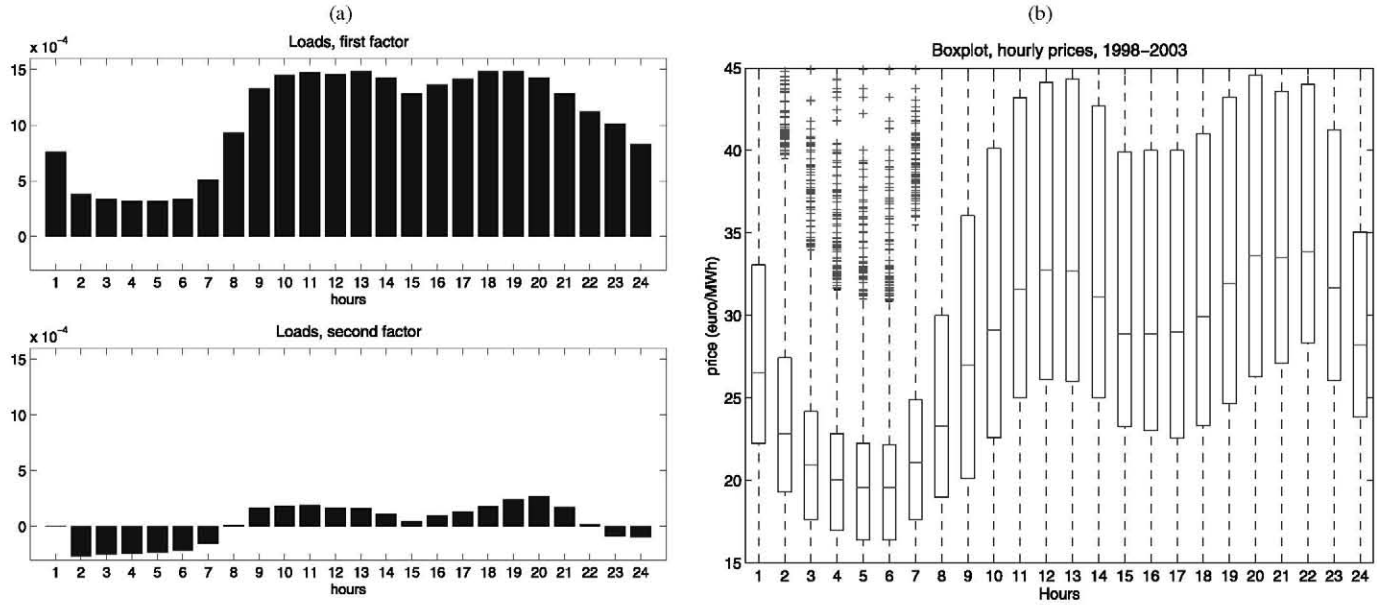

Figure 3. (a) Loads of the unobserved common factors, and (b) boxplot of the hourly prices.

selected hours 4, 11, and 20 because of their representativeness and we provide in Figure 4(b) the prices for these hours during the period October–December 2003 (the final period considered in estimating SeaDFA), as well as the part explained by the first common factor and the part of these hourly series explained by the two common factors. The difference between each series and the common part explained by the unobserved common factors is the specific component.

The loads corresponding to the first factor are all positive [Figure 3(a)], larger in those hours in which both the level and variance of the prices are higher. The absolute value of loads corresponding to the first factor are much larger than those for the second one, so the first one explains much more of the variability of each time series of hourly prices. Thus, for each series, the commonality of the second factor only implies a small "correction" (addition or subtraction) over the commonality of this hour considering only f_{1t}. This addition or subtraction depends on the day of the week (weekday or weekend) and nighttime or daytime. The loads corresponding to this second factor are positive during the night and negative for daytime hours [in Figure 3(a) ω_{i1} and $-\omega_{i2}$, $i = 1, \ldots, 24$, are plotted].

According to Figures 3(a) and 4(a):

- For hour 4 (nighttime), in weekdays (first part of the week, the dates that appear in the graph correspond to Mondays) we observe that a negative term is added, and a positive one for weekends (end of weeks).
- For hour 11 or 20 (daytime, morning and early evening), in weekdays (first part of the week, the dates that appear in the graph correspond to Mondays) we observe that a positive amount is added, and a negative one for weekends (end of weeks).

These patterns can be observed in Figure 4(b).

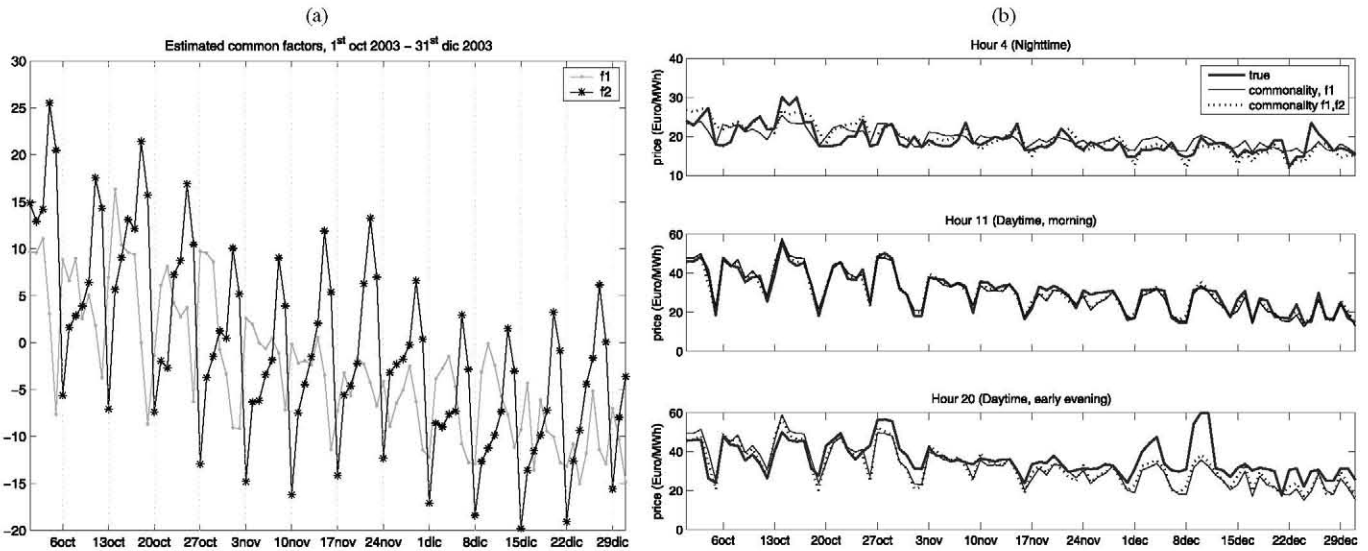

Figure 4. October–December 2003. (a) Common factors. (b) Hourly series and commonalities.

Finally, it is worth explicitly mentioning that, given the complexity of the data analyzed, as a second stage we have fitted univariate autoregressive processes for the estimated specific factors, $\widehat{\boldsymbol{\varepsilon}}_t = (\widehat{\varepsilon}_{1t}, \ldots, \widehat{\varepsilon}_{mt})'$, just to capture any remaining structure that might affect them. This idea was first applied by Peña and Poncela (2004) and Ortega and Poncela (2005). From a theoretical point of view this is feasible; Bai (2003) stated that even if some correlation is present in the specific factors, consistent estimates can be obtained for the parameters involved in the model. However, since the main objective of this work is long-term forecasting, the effect of specific factors on long-term forecasts is not relevant. Even in the short run, and given that the variance of the specific factors is very small, the influence of this modeling on our short-term prediction errors was not significant.

5.2 Point Forecasts

We provide the results obtained when calculating forecasts for electricity prices in 2004, using the data from 1st January 1998 through 31st December 2003. Thus, the forecasting horizon varies from one day up to one year, since the last datum we used was the last day in 2003, no matter for which day of 2004 we are computing the forecast. The data were selected to be able to compare results with models developed for short-run forecasting.

The accuracy metrics we consider have been selected because they have been previously used in the literature to evaluate the performance of the models developed for forecasting in electricity markets (Conejo et al. 2005). Let $p_{H,d}$ be the price in day d in the Hth hour, and $\widehat{p}_{H,d}$ its computed forecast; then the error measurement $e_{H,d}$ (for each hour of each day) is defined as $e_{H,d} = |p_{H,d} - \widehat{p}_{H,d}|/p_{H,d}$. Using $e_{H,d}$, the subsequent accuracy metrics can be defined for each day: $emean_d = (1/24)\sum_{H=1}^{24} e_{H,d}$ and $emedian_d = median(e_{1,d}, e_{2,d}, \ldots, e_{24,d})$. Finally, using the expressions of $emean_d$ and $emedian_d$, the MAPE and MAPE2 are obtained for a period of D' days as

$$\text{MAPE} = \frac{1}{D'}\sum_{d=1}^{D'} emean_d \qquad \text{and}$$

$$\text{MAPE2} = \frac{1}{D'}\sum_{d=1}^{D'} emedian_d.$$

5.2.1 Long-Run Forecasting. The period selected (computing forecasts for the year 2004 using the data for years 1998 to 2003) was chosen to be able to compare the results with previous ones. Since there is no published work on long-run forecasting of electricity prices, we will compare our results with those obtained by methods specifically developed for short-term forecasting in the Spanish market.

With respect to benchmarking models, in a different application context, Taylor (2008) carried out a comparison of six different forecasting methods for seasonal data. He drew two main conclusions: first, that there is a strong potential for the use of seasonal ARIMA modeling and the extension of Holt–Winters for predicting up to about two or three days, and second, for longer lead times, a simplistic historical average called "Seasonal Mean" is difficult to beat.

In this article we use as benchmarking models the following:

1. The Mixed Model approach proposed by García-Martos, Rodríguez, and Sánchez (2007), which uses a combination of several univariate seasonal ARIMA models for different lengths of time series. We have selected this model because, with respect to prediction errors in the short term, it obtains the best results among models published for the Spanish market, and it computes forecasts for every hour in a very large span of hours (all in the period 1998–2003).
2. The nonstationary Dynamic Factor Analysis (DFA) derived by Peña and Poncela (2004, 2006) could be an alternative when only the regular part of the dynamics can be modeled. In this case a VARIMA$(9, 1, 0)$, has been fitted for the $r = 2$ unobserved common factors extracted.
3. The "Seasonal Mean" model by Taylor (2008). For each lead time, this would be the mean of the prices for the same hour of the week as the period to be predicted. He set the moving window to be equal to 24 weeks in length.

The numerical results obtained for the year 2004, MAPE for each month, are shown in Figure 5 and Table 2.

In Figure 5 and Table 2 the monthly MAPEs are provided, as well as the MAPE calculated for the whole year 2004, so after computing year-ahead forecasts for the whole year 2004, we have obtained hourly forecasting errors, $e_{H,d}$, $H = 1, \ldots, 24$, $d = 1, \ldots, 366$. MAPE for the whole year is 21.56% and the MAPE2 is 20.39%. Although the Mixed Model provided by García-Martos, Rodríguez, and Sánchez (2007) is very good at short-term forecasting, the errors obtained for the long run are very large, since MAPE for the whole year is 45.62% and MAPE2 is 47.76%. Although this Mixed Model is a sophisticated ARIMA specially designed for the Spanish market and selected for the best global performance over a period of six years, it does not perform well over the long run. Taylor (2008) pointed out this handicap of ARIMA models.

On the other hand, when using the DFA provided by Peña and Poncela, extracting $r = 2$ common factors and fitting for them a VARIMA$(9, 1, 0)$, MAPE and MAPE2 are respectively 28.17% and 25.52%, which illustrates both: (1) the great reduction in the error when using dimensionality reduction techniques compared to the Mixed Model and (2) the importance of including seasonality in DFA, allowing for extracting seasonal common factors, since SeaDFA reduces the MAPE to 21.56% and the MAPE2 to 20.39%.

Taylor's (2008) "Seasonal Mean," which has the benefit of simplicity, performs much better than the Mixed Model, but slightly worse than the DFA, with a MAPE of 29.60% and MAPE2 of 27.40%.

To summarize the results in Figure 5 and Table 2: the estimation of SeaDFA is worth the effort since a MAPE of 21.56% is obtained, in comparison with a MAPE of 28.17% for DFA, 29.60% for the "Seasonal Mean," and 45.62% for the Mixed Model by García-Martos, Rodríguez, and Sánchez (2007). Moreover, as observed in the histograms depicted in Figure 5, the errors obtained with the SeaDFA are not only the smallest in mean but also those that present the lowest variability.

Moreover, to emphasize the nice performance of SeaDFA, it is interesting to compare the year-ahead forecasting error of

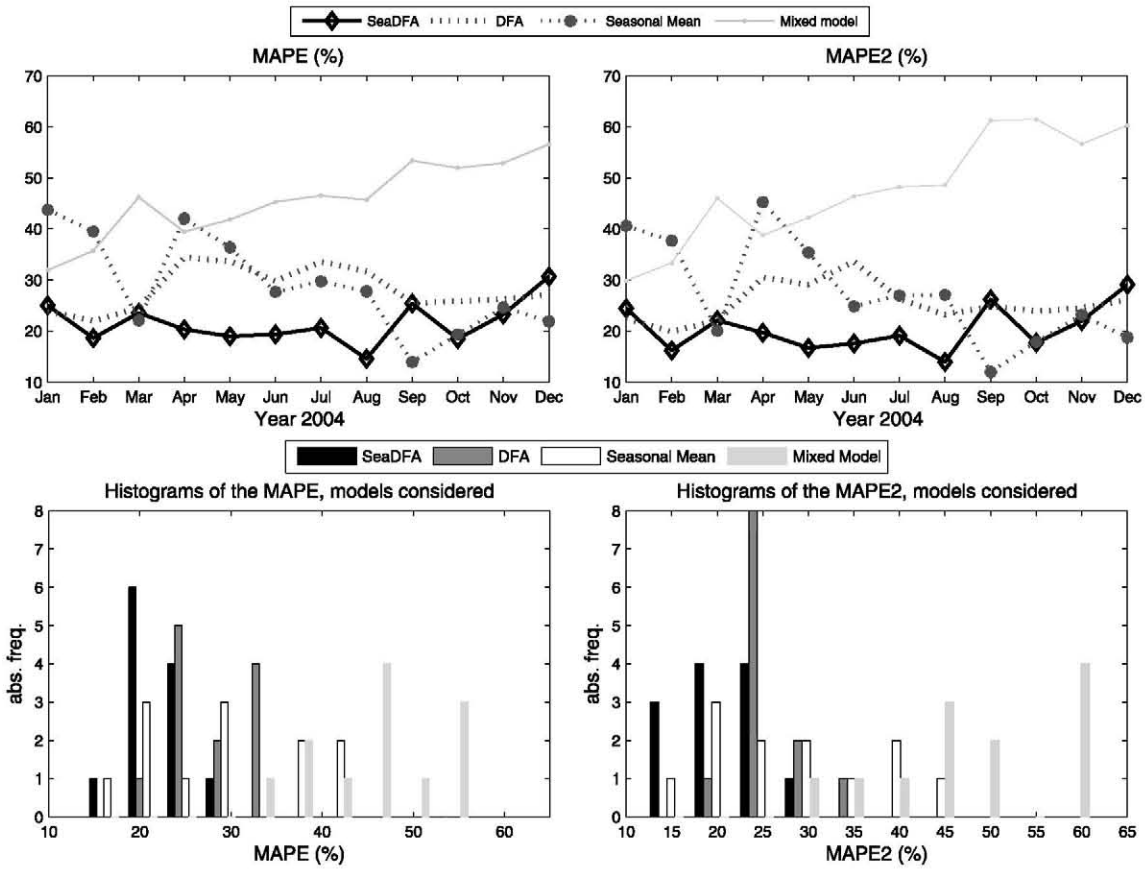

Figure 5. Monthly prediction errors (MAPE and MAPE2). SeaDFA, DFA (Peña and Poncela 2004), Seasonal Mean (Taylor 2008), and Mixed Model (García-Martos, Rodríguez, and Sánchez 2007). The complete table corresponding to these data is included in the online Supplementary Materials. The online version of this figure is in color.

21.56% obtained with SeaDFA to the one-day-ahead forecasting error of 12.61% obtained with the Mixed Model.

After presenting global results of SeaDFA for the entire year of 2004, we will focus on some particular weeks, specifically the third week in each month (often this week is used to check accuracy of forecasting models: Conejo et al. 2005; Contreras et al. 2003). Calculating the MAPE with SeaDFA, for these twelve weeks we get a MAPE equal to 19.75%, which again reflects the good performance of SeaDFA. Notice that for the first of these twelve weeks selected (the third week in January 2004) the forecasting horizon ranges from two to three weeks, while for the third week in February the forecasting horizon ranges from seven to eight weeks and so on. In the last of the twelve weeks considered the forecasting horizon ranges from 51 to 52 weeks.

In Table 3 the results for the third week in February (16th–22nd February 2004) are provided. These results have again been obtained using the SeaDFA estimated for the prices in 1998–2003, so the forecasting horizon varies from seven to eight weeks. The MAPE for this week is 16.38%. Using the Mixed Model of García-Martos, Rodríguez, and Sánchez (2007), the MAPE is 34.78%.

Finally, and despite the fact that they incorporated the information from the futures market, Conejo et al. (2010) computed year-ahead forecasts in the EEX, obtaining MAPEs that vary from 13% up to 44%.

5.2.2 Short-Term Forecasting. In addition, although it is not the main goal of this article, we will also check the short-term forecasting performance of SeaDFA. Although SeaDFA was not developed for this purpose, we get good forecasts in terms of prediction errors, even when comparing them with other forecasts obtained by methods specifically designed for the short-term time horizons.

To illustrate that the SeaDFA is valid not only for medium- and long-term forecasting but also for the short run, we provide

Table 2. Average MAPE and MAPE2 for year-ahead forecasts for 2004. Monthly forecasting errors are plotted in Figure 5, and detailed in the online Supplementary Materials

	SeaDFA		DFA		Seasonal Mean		Mixed Model	
(%)	MAPE	MAPE2	MAPE	MAPE2	MAPE	MAPE2	MAPE	MAPE2
Year 2004	21.56	20.39	28.17	25.52	29.06	27.46	45.62	47.76

Table 3. Prediction errors, long-term forecasting, and one-step-ahead, for the week 16–22 February 2004

		Day 1	Day 2	Day 3	Day 4	Day 5	Day 6	Day 7	MWE
Long-term	SeaDFA	37.17%	13.00%	12.00%	9.83%	18.99%	9.94%	14.70%	16.38%
	Mixed Model	27.04%	29.30%	29.62%	29.21%	34.50%	44.54%	49.22%	34.78%
One-step-ahead	SeaDFA	12.44%	3.53%	4.35%	8.60%	6.54%	11.02%	20.3%	9.54%
	Mixed Model	19.78%	6.34%	5.17%	5.39%	10.70%	8.07%	11.0%	9.49%

in Table 3 the forecasts and errors computed for the same week in February 2004, but obtained by estimating the model using data up to the 15th of February and then re-estimating the model six more times while subsequently updating the data, which means that for each day in this week we are computing one-day-ahead forecasts. The results are compared with those obtained with the Mixed Model by García-Martos, Rodríguez, and Sánchez (2007), which was specifically designed for one-day-ahead forecasting. Since this Mixed Model is based on seasonal ARIMA modeling, it falls into the class of models provided by Taylor (2008) as performing best in the short term. The daily prediction errors obtained by SeaDFA are of the same magnitude or even lower in comparison with these best-performing short-term models.

5.3 Forecasting Intervals

Once numerical results for long- and short-run point forecasts have been provided, it is of interest to report those obtained when forecasting intervals are computed by means of the bootstrap scheme proposed. Computation of forecasting intervals is relevant for decision making of agents involved in power markets. In Figure 6 the percentile-based forecasting intervals for the point prediction, which include uncertainty due to parameter estimation, are provided for the week of the 24th–30th of May 2004. This week in May has been used in previous works to check the performance of forecasting methods because its behavior in terms of load is special (Contreras et al. 2003).

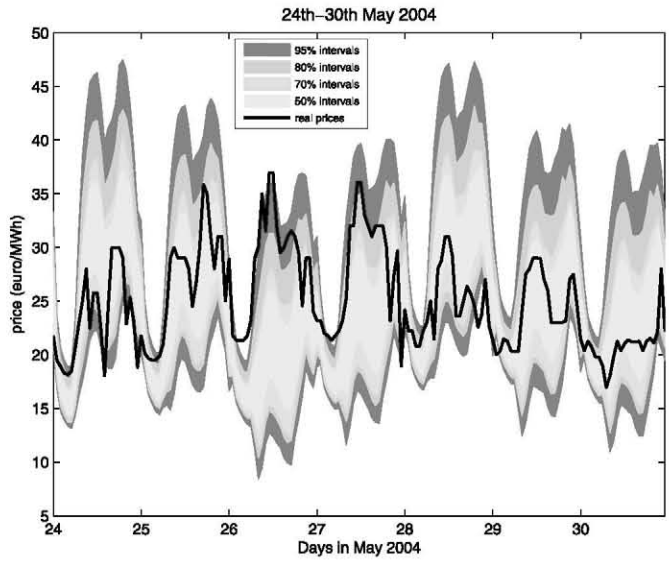

Figure 6. Percentile-based confidence intervals including uncertainty due to parameter estimation (24–30 May 2004).

The forecasting horizon is almost five months, since the last data used to estimate the model are from the 31st of December 2003 (as in all the forecasts computed) which means medium-term forecasting. Figure 6 illustrates the coverage of the prediction intervals obtained using bootstrap techniques. For the week shown in this figure the percentage of real prices which are inside the 95% intervals is 89.88%, since 17 real prices lie outside them, and there are 168 hourly prices in a week. The coverages shown in Section 4 for simulated data are closer to the nominal value (95%), but it should be considered that we are now dealing with real data, and outliers can be encountered, which make the task more difficult. Anyway, the value obtained is almost 90%, and can be considered acceptable under the described circumstances.

6. CONCLUSIONS

In this work we have provided a seasonal extension to Nonstationary Dynamic Factor Analysis, which is a powerful tool for dealing with the important problem of long-term forecasting of electricity prices. This problem has remained unsolved until now. In addition, the dimensionality reduction technique proposed here can be applied for modeling and forecasting any dataset consisting of a high-dimensional vector of time series with a seasonal pattern. Seasonal Dynamic Factor Analysis (SeaDFA) avoids previous seasonal differencing (which can remove the seasonality only in the mean and variance, but the seasonality in serial dependence structure remains), and allows not only for regular unit roots and dynamics but also seasonal ones. Seasonal Dynamic Factor Analysis (SeaDFA) is able to estimate common factors that follow a multiplicative seasonal VARIMA model with a constant.

SeaDFA requires a modification in the estimation procedure of the existing Dynamic Factor Analysis, due to seasonality and to the inclusion of the constant presented. The inclusion of the constant allows us to improve long-run forecasts in the case of nonstationary processes.

Also, we have proposed a bootstrap procedure for making inference on all the parameters involved in the model. Furthermore, the bootstrap scheme introduced in this work can be applied to all models that can be expressed under state-space formulation and needs neither a backward representation of the model nor the innovations form representation.

We apply our ideas to an interesting dataset that is difficult to forecast, electricity prices in the Spanish market. Our forecasts are accurate, with a typical error of around 20% for year-ahead, which compares well with one-day-ahead prediction errors of about 13% found using some recent and accurate pub-

lished models (García-Martos, Rodríguez, and Sánchez 2007). Apart from very accurate forecasts for the medium and long run, the SeaDFA is also competitive in the short term. The calculation of prediction intervals is done by the bootstrap scheme here proposed.

The results obtained are precise enough to be able to use them for long-term risk management of utilities.

SUPPLEMENTARY MATERIALS

Example, table, and appendices: *Illustration Example 1*: Example on how to build the transition matrix in the particular case of observed data that present seasonality. *Table of monthly forecasting errors*: Includes monthly forecasting errors obtained with SeaDFA and other benchmarking models (this information is plotted in Figure 5 of the article). *Appendix A*: Kalman filter and smoother recursions. *Appendix B*: EM derivations for the SeaDFA. *Appendix C*: Derivation of the correction term for the bootstrap procedure. (A_GM_R_S_supplemTEX.pdf)

ACKNOWLEDGMENTS

This work was supported by Projects SEJ2007-64500 ECO2008-05080, ECO2009-10287, and MTM2009-12419, Spanish Ministry of Science and Innovation. We are grateful to Pilar Poncela for her suggestions and comments, as well as to José Mira and Javier Nogales for their help. We also thank Prof. David Steinberg, an associate editor, as well as two anonymous referees, whose insightful comments let us improve the article.

[Received March 2009. Revised February 2011.]

REFERENCES

Alonso, A. M., Peña, D., and Romo, J. (2002), "Forecasting Time Series With Sieve Bootstrap," *Journal of Statistical Planning and Inference*, 100 (1), 1–11. [141,142]

Anderson, B. D. O., and Moore, J. B. (1979), *Optimal Filtering*, Englewood Cliffs, NJ: Prentice-Hall. [141]

Angelus, A. (2001), "Electricity Price Forecasting in Deregulated Markets," *Electricity Journal*, 14 (3), 32–41. [144]

Ansley, C. F., and Kohn, R. (1986), "Estimation Prediction and Interpolation for ARIMA Models With Missing Data," *Journal of the American Satistical Association*, 81 (395), 751–761. [139]

Ansley, C. F., and Newbold, P. (1980), "Finite Sample Properties of Estimators for Autorregressive Moving Average Properties," *The Journal of Econometrics*, 13, 159–183. [140]

Bai, J. (2003), "Inferenctial Theory for Factor Models of Large Dimensions," *Econometrica*, 71 (1), 135–171. [138,141,147]

Cao, R., Febrero-Bande, M., González-Manteiga, W., Prada-Sánchez, J. M., and García-Jurado, I. (1997), "Saving Computer Time in Constructing Consistent Bootstrap Prediction Intervals for Autoregressive Processes," *Communications in Statistics Simulation and Computation*, 71 (1), 135–171. [142]

Conejo, A. J., Carrión, M., Nogales, F. J., and Morales, J. M. (2010), "Electricity Pool Prices: Long-Term Uncertainty Characterization for Futures-Market Trading and Risk Management," *Journal of the Operational Research Society*, 61, 235–245. [144,148]

Conejo, A. J., Contreras, J., Espínola, R., and Plazas, M. A. (2005), "Forecasting Electricity Prices for a Day-Ahead Pool-Based Electric Energy Market," *International Journal of Forecasting*, 21 (3), 435–462. [144,147,148]

Contreras, J., Espínola, R., Nogales, F. J., and Conejo, A. J. (2003), "ARIMA Models to Predict Next-Day Electricity Prices," *IEEE Transactions on Power Systems*, 18 (3), 1014–1020. [144,148,149]

Cottet, R., and Smith, M. (2003), "Bayesian Modeling and Forecasting of Intraday Electricity Load," *Journal of the American Statistical Association*, 98 (464), 839–849. [138,144]

Durbin, J., and Koopman, S. J. (2001), *Time Series Analysis by State Space Methods*, Oxford, U.K.: Oxford University Press. [139,140]

Forni, M., Hallin, M., Lippi, M., and Reichlin, L. (2000), "The Generalized Dynamic-Factor Model: Identification and Estimation," *The Review of Economics and Statistics*, 82 (4), 540–554. [139]

Frestad, D. (2008), "Common and Unique Factors Influencing Daily Swap Returns in the Nordic Electricity Market," *Energy Economics*, 30, 1081–1097. [144]

García-Jurado, I., González-Manteiga, W., Prada-Sánchez, J. M., Febrero-Bande, M., and Cao, R. (1995), "Predicting Using Box–Jenkins, Nonparametric and Bootstrap Techniques," *Technometrics*, 37 (3), 303–310. [141]

García-Martos, C., Rodríguez, J., and Sánchez, M. J. (2007), "Mixed Models for Short-Run Forecasting of Electricity Prices: Application for the Spanish Market," *IEEE Transactions on Power Systems*, 2 (2), 544–552. [144,147-150]

Geweke, J. (1977), "The Dynamic Factor Analysis of Economic Time Series," in *Latent Variables in Socio-Economic Models*, Amsterdam: North Holland. [137]

Grady, W. M., Groce, L. A., Huebner, T. M., Lu, Q. C., and Crawford, M. M. (1991), "Enhancement, Implementation, and Performance of an Adaptive Short-Term Load Forecasting Algorithm," *IEEE Transactions on Power Systems*, 6 (4), 1404–1410. [144]

Harvey, A. C. (1989), *Forecasting Structural Time Series Models and the Kalman Filter*, Cambridge: Cambridge University Press. [138,140]

Harvey, A. C., Ruiz, E., and Sentana, E. (1992), "Unobservable Component Time Series Models With ARCH Disturbances," *The Journal of Econometrics*, 52, 129–158. [141]

Koopman, S. J., Ooms, M., and Carnero, M. A. (2007), "Periodic Seasonal Reg–ARFIMA–GARCH Models for Daily Electricity Spot Prices," *Journal of the American Satistical Association*, 102 (477), 16–27. [138,144]

Lee, R. D., and Carter, L. R. (1992), "Modeling and Forecasting U.S. Mortality," *Journal of the American Statistical Association*, 87 (419), 659–671. [137]

Ljung, G. (1986), "Diagnostic Testing of Univariate Time Series Models," *Biometrika*, 73 (3), 725–730. [141]

Ljung, L., and Caines, P. E. (1979), "Asymptotic Normality of Prediction Error Estimators for Approximate System Models," *Stochastics*, 3, 29–46. [141]

Nogales, F. J., Contreras, J., Conejo, A. J., and Espínola, R. (2002), "Forecasting Next-Day Electricity Prices by Time Series Models," *IEEE Transactions on Power Systems*, 17 (2), 342–348. [144]

Ortega, J. A., and Poncela, P. (2005), "Joint Forecasts of Southern European Fertility Rates With Non-Stationary Dynamic Factor Models," *International Journal of Forecasting*, 21, 539–550. [138,147]

Pascual, L., Romo, J., and Ruiz, E. (2004), "Bootstrap Predictive Inference for ARIMA Models," *Journal of Time Series Analysis*, 25 (4), 449–465. [141, 142]

Peña, D., and Box, G. E. P. (1987), "Identifying a Simplifying Structure in Time Series," *Journal of the American Statistical Association*, 82 (399), 836–843. [137]

Peña, D., and Poncela, P. (2004), "Forecasting With Nonstationary Dynamic Factor Models," *The Journal of Econometrics*, 119, 291–321. [137-139,142, 147,148]

——— (2006), "Nonstationary Dynamic Factor Analysis," *Journal of Statistical Planning and Inference*, 136, 1237–1257. [137,139,145,147]

Rodríguez, A., and Ruiz, E. (2009), "Bootstrap Predicition Intervals in State-Space Models," *Journal of Time Series Analysis*, 30, 167–178. [141,143]

Sáfadi, T., and Peña, D. (2008), "Bayesian Analysis of Dynamic Factor Models: An Application to Air Pollution and Mortality in São Paulo, Brazil," *Environmetrics*, 19, 582–601. [144]

Sánchez, I. (2006), "Recursive Estimation of Dynamic Models Using Cook's Distance, With Application to Wind Energy Forecast," *Technometrics*, 48, 61–73. [138]

Sargent, T. J., and Sims, C. (1977), "Business Cycle Modeling Without Pretending to Have too Much a priori Theory," in *New Methods of Business Cycle Research*, ed. C. Sims, Minneapolis: Federal Reserve Bank of Minneapolis. [137]

Shumway, R. H., and Cavanaugh, J. E. (1996), "On Computing the Expected Fisher Information Matrix for State-Space Model Parameters," *Statistics and Probability Letters*, 26 (4), 347–355. [137]

Shumway, R. H., and Stoffer, D. S. (1982), "An Approach to Time Series Smoothing and Forecasting Using the EM Algorithm," *Journal of Time Series Analysis*, 3, 253–264. [139]

——— (2006), *Time Series Analysis and Its Applications With R Examples. Springer Texts in Statistics*, New York: Springer. [139,140]

Smith, M., and Cottet, R. (2006), "Estimation of a Longitudinal Multivariate Stochastic Volatility Model for the Analysis of Intra-Day Electricity Prices," Working Paper ECMT2006-1, Sydney University, School of Economics and Political Science. [144]

Spall, J. C., and Wall, K. D. (1984), "Asymptotic Distribution Theory for the Kalman Filter State Estimator," *Communications in Statistics: Theory and Methods*, 13, 1981–2003. [141]

▶Stock, J. H., and Watson, M. (2002), "Forecasting Using Principal Components From a Large Number of Predictors," *Journal of the American Statistical Association*, 97, 1167–1179. [137]

▶Stoffer, D. S., and Wall, K. (1991), "Bootstrapping State Space Models: Gaussian Maximum Likelihood Estimation and the Kalman Filter," *Journal of the American Statistical Association*, 86, 1024–1033. [137,141]

▶Taylor, J. W. (2008), "A Comparison of Univariate Time Series Methods for Forecasting Intraday Arrivals at a Call Center," *Management Science*, 54, 253–265. [147-149]

▶Thombs, L. A., and Schucany, W. R. (1990), "Bootstrap Prediction Intervals for Autoregression," *Journal of the American Statistical Association*, 85, 486–492. [141]

▶Vehvilainen, I., and Pyykkonen, T. (2005), "Stochastic Factor Model for Electricity Spot Price: The Case of the Nordic Market," *Energy Economics*, 27, 351–367. [144]

▶Wall, K., and Stoffer, D. S. (2002), "A State Space Approach to Bootstrapping Conditional Forecasts in ARMA Models," *Journal of Time Series Analysis*, 23, 733–751. [137,141]