UNIVERSIDAD CARLOS III DE MADRID working papers

Working Paper 08-14
Statistic and Econometric Series 06
March 2008

Departamento de Estadística
Universidad Carlos III de Madrid
Calle Madrid, 126
28903 Getafe (Spain)
Fax (34-91) 6249849

SEASONAL DYNAMIC FACTOR ANALYSIS AND BOOTSTRAP INFERENCE: APPLICATION TO ELECTRICITY MARKET FORECASTING *

Andrés M. Alonso,[1] Carolina García-Martos,[2] Julio Rodríguez,[3] and María Jesús Sánchez[2]

Abstract

Year-ahead forecasting of electricity prices is an important issue in the current context of electricity markets. Nevertheless, only one-day-ahead forecasting is commonly tackled up in previous published works. Moreover, methodology developed for the short-term does not work properly for long-term forecasting.

In this paper we provide a seasonal extension of the Non-Stationary Dynamic Factor Analysis, to deal with the interesting problem (both from the economic and engineering point of view) of long term forecasting of electricity prices. Seasonal Dynamic Factor Analysis (SeaDFA) allows to deal with dimensionality reduction in vectors of time series, in such a way that extracts common and specific components. Furthermore, common factors are able to capture not only regular dynamics (stationary or not) but also seasonal one, by means of common factors following a multiplicative seasonal VARIMA(p,d,q)×(P,D,Q)$_s$ model.

Besides, a bootstrap procedure is proposed to be able to make inference on all the parameters involved in the model. A bootstrap scheme developed for forecasting includes uncertainty due to parameter estimation, allowing to enhance the coverage of forecast confidence intervals. Concerning the innovative and challenging application provided, bootstrap procedure developed allows to calculate not only point forecasts but also forecasting intervals for electricity prices.

Keywords: Dynamic Factor Analysis, Bootstrap, Forecasting, Confidence intervals.
JEL Classification: C32 and C53

* Financial support from Projects MTM2005-08897, SEJ2005-06454 and SEJ2007-64500 (Ministerio de Educación y Ciencia, Spain) is gratefully acknowledged.

[1] Departamento de Estadística. Universidad Carlos III de Madrid
[2] Escuela Técnica Superior de Ingenieros Industriales. Universidad Politécnica de Madrid.
[3] Facultad de Ciencias Económicas y Empresariales. Universidad Autónoma de Madrid.

1 Introduction

Forecasting electricity prices has been developed recently, since just a few years ago only demand was predicted in centralized markets. Nowadays, however, electricity is traded under competitive rules, like other *commodities,* and this has opened a new field of research. The novelty of the problem and the special features that electricity presents (not being able to be stored, and demand to be satisfied instantaneously, which are responsible for the largely unpredictable behavior of its price) has created the need of developing specific models that deal with this problem, since it has a great importance for a strategic sector in the economy of any country.

Nowadays there are several ways to trade with electricity:

1. Forward markets and options, which are only well developed in some electricity markets, like the European Energy Exchange (EEX) in Germany.

2. The *pool*, in which both the producers and consumers submit their respective generation and consumption bids (for each hour) to the market operator. In the market of mainland Spain the marginal price is defined as the bid submitted by the last generation unit needed to satisfy the whole demand. This mechanism is shown in Fig. 1, and means that for each day a 24-dimensional vector of prices is generated. Also this structure of the data appears when analyzing other well known electricity markets like Nord Pool or the PJM Interconnection (a regional transmission organization that plays a vital role in the U.S. electric power system), where there is a load data and price data for each hour. On the other hand, other electricity markets like New South Wales in Australia operate in such a way that the resulting clearing price and accepted production and consumption bids are determined every half an hour, so for each day 48 data are available.

3. Bilateral contracts: customers and generators can agree to trade a certain amount of power at a certain price. But every contract implies a risk since the seller must purchase every day in the Pool the amount of energy agreed. Having accurate long-term forecasts (covering at least the length of the bilateral contract) reduces this risk. In this work we compute forecasts for electricity prices, with forecasting horizon ranging from one day to one year.

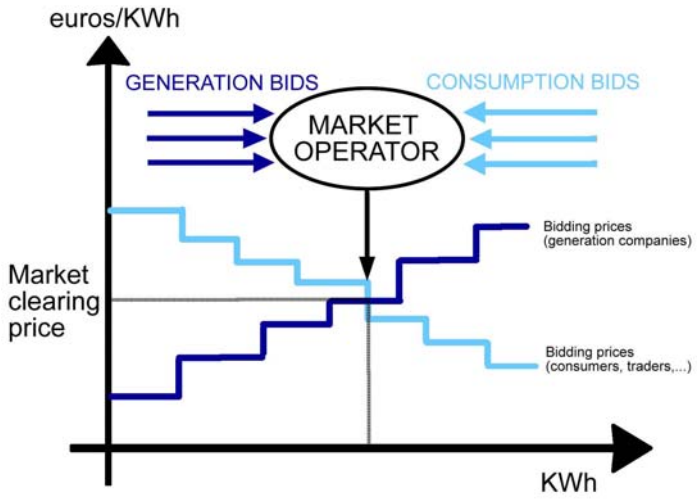

Figure 1: Market clearing price. Spanish Market.

Most previous works are focused on computing short-term (one-day-ahead) forecasts of electricity prices or load forecasts. Nogales et al. (2002) applied transfer functions and dynamic regression to forecast electricity prices. Contreras et al. (2003) forecasted electricity prices of the Californian and Spanish markets by applying ARIMA models. Troncoso et al. (2002) compared the kWNN (k Weighted Nearest Neighbours) technique with dynamic regression. Crespo-Cuaresma et al. (2002) have suggested a group of univariate models to predict electricity prices in the Leipzig market, the most important spot market in Germany. Conejo et al. (2005) compared several methods including wavelet approximation, ARIMA models and neural networks, the extensive analysis is conducted using data from the PJM Interconnection.. Nogales et al. (2006) forecasted the prices in the PJM Interconnection, through transfer functions, showing that the inclusion of explanatory variables (like demand) does not significantly reduce the prediction errors. For all of them the average error is around 13-15%. García-Martos et al. (2007) computed short-term forecasts for every hour in the period 1998-2003, obtaining a prediction error around 12.61%. On the other hand, Cottet and Smith (2003) and Koopman et al. (2007) provided load forecasts and periodic extensions of dynamic long-memory regression for the analysis of daily spot prices, respectively. Nevertheless, long-term forecasting of electricity prices has been scarcely studied, and there is no published works dealing with this issue. Moreover, methodology applied to short-term forecasting does

not work properly in the long-run, as it will be shown in this work, so an specific methodology must be developed for long-term forecasting of electricity prices.

Since a 24-dimensional vector of time series is considered instead of the complete time series, following the well known *parallel approach*, a high-dimensional vector of series must be modelled, so the number of parameters to be estimated growths and the "curse of dimensionality" arises. Besides, seasonality is present (due to strong dependence of prices and demand on weather and economic and social activities), so not only regular dynamics but also seasonal one must be estimated, the problem is more important.

Dimensionality reduction techniques for vectors of time series have been deeply studied and many references can be quoted in this direction. Sargent and Sims (1977) and Geweke (1977) were the first to propose a dynamic factor model. On the one hand, Stock and Watson (2002) explores 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 Components Analysis to the case in which the variables are time series, and compute 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 Non-Stationary case.

However, there is not a specific dimension reduction technique that can be applied to vectors of time series that present a seasonal pattern. There are many examples of this kind of data, such as vectors of macroeconomic variables, metheorological data and time series coming from electricity markets (load and prices). Bearing this in mind, till 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 references cited above in order 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 as SeaDFA) is presented. It allows extracting the common factors of a vector of time series, and estimating the seasonal multiplicative VARIMA model that they follow, so regular and seasonal dynamics can be modeled. Secondly, concerning 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 Fisher information matrix (Shumway and Cavanaugh (1996)), since asymptotic results are not of application if time series are not fairly long or the parameters fall near the boundary of the parameter space.

Development of these two contributions are motivated by the analysis of time series of electricity prices, which are relevant, both from the engineering and economic point of view. Actually, a great number of recent references (Cottet and Smith (2003), Koopman, Ooms and Carnero (2007)) focus that complex problem affecting all the agents involved in electricity markets. SeaDFA is applied to compute year-ahead forecasts of electricity prices.

The rest of the paper is organized as follows. In Section 2 the Seasonal Dynamic Factor Analysis (SeaDFA) and its estimation algorithm is 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 proposed bootstrap procedure. In Section 5 the model and its bootstrap scheme is applied to forecasting electricity prices in the Spanish market. Finally, in Section 6 some conclusions are provided.

2 Seasonal Dynamic Factor Analysis (SeaDFA)

In this Section we present the Seasonal Dynamic Factor Model, that allows to deal with common factors following a VARIMA$(p, d, q) \times (P, D, Q)_s$ model with constant. In Subsection 2.1 we present the model specification, in Subsection 2.2 we present the relationship between SeaDFA and state-space formulation. In Subsection 2.3 the estimation procedure is described.

2.1 The model

Let $\mathbf{y}_t$ be a m-dimensional vector of observed time series generated by a 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.

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

where Ω is the $m \times r$ load matrix that relates the r-dimensional set of common factors to the vector of observed time series $\mathbf{y}_t$, and ε_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, while the specific dynamic is captured by the specific factors. We suppose that the specific components, $\boldsymbol{\varepsilon}_t$, are white noise. Vector $\boldsymbol{\varepsilon}_t$ has zero mean, $E[\boldsymbol{\varepsilon}_t \boldsymbol{\varepsilon}'_\tau] = 0$ if $t \neq \tau$, and diagonal covariance matrix $\mathbf{S} = E[\boldsymbol{\varepsilon}_t \boldsymbol{\varepsilon}'_t]$.

Unobserved common factors $\mathbf{f}_t$ can be non-stationary, 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 as follows,

$$(1 - B)^d (1 - B^s)^D \boldsymbol{\phi}(B) \Phi(B^s) \mathbf{f}_t = c + \boldsymbol{\theta}(B) \Theta(B^s) \mathbf{w}_t, \tag{2}$$

where $\boldsymbol{\phi}(B) = (\mathbf{I} - \boldsymbol{\phi}_1 B - \boldsymbol{\phi}_2 B^2 - ... - \boldsymbol{\phi}_p B^p)$, $\Phi(B^s) = (\mathbf{I} - \Phi_1 B^s - \Phi_2 B^{2s} - ... - \Phi_p B^{Ps})$, $\boldsymbol{\theta}(B) = (\mathbf{I} - \boldsymbol{\theta}_1 B - \boldsymbol{\theta}_2 B^2 - ... - \boldsymbol{\theta}_q B^q)$ and $\Theta(B^s) = (\mathbf{I} - \Theta_1 B - \Theta_2 B^{2s} - ... - \Theta_q B^{Qs})$ are polynomial matrices $r \times r$, B is the backshift operator such that $B\mathbf{y}_t = \mathbf{y}_{t-1}$, the roots of $|\boldsymbol{\phi}(B)| = 0$ and $|\Phi(B^s)| = 0$ are on or outside the unit circle, the roots of $|\boldsymbol{\theta}(B)| = 0$ and $|\Theta(B^s)| = 0$ are outside the unit circle and $\mathbf{w}_t \sim \mathbf{N}_r(\mathbf{0}, \mathbf{Q})$ is serially uncorrelated, $E[\mathbf{w}_t \mathbf{w}'_{t-h}] = 0, h \neq 0$. We also assume that the noise term of the common factors and the observed series are also uncorrelated for all lags, $E[\mathbf{w}_t, \boldsymbol{\varepsilon}_{t-h}] = 0, \forall h$. c is the constant of the model of the common factors.

It should be noted that the model is not identifiable, since for any $r \times r$ non-singular matrix $\mathbf{H}$, the observed vector of time series can be expressed as a linear combination of a new set of factors. To solve this identification problem, we can always choose either $\mathbf{Q} = \mathbf{I}$ or $\Omega'\Omega = \mathbf{I}$. Therefore, the model is not yet identified under rotations, and we need to introduce an additional constraint to be able to estimate the model. Harvey (1989) imposes that $\omega_{ij} = 0$, for $j > i$, where $\Omega = [\omega_{ij}]$. This condition is not restrictive since the factor model can be rotated for better interpretation when needed.

The inclusion of the constant can be relevant if trying to compute long-term forecasts in

the non-stationary case.

2.2 State-space formulation and its relationship with SeaDFA

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:

$$\begin{aligned} x_t &= A\alpha_t + B\beta_t + C\gamma_t \\ \alpha_t &= D\alpha_{t-1} + F\beta_t + G\delta_t \end{aligned}$$

The first equation is known as the measurement or observation equation and relates the observed $m-$dimensional series x_t with the k-dimensional latent or unobserved state α_t components. A is the loading matrix, β_t is the vector of exogenous variables and matrix B relates the vector of observed series with the vector of exogenous variables. The additive observation noise γ_t is assumed to be Gaussian with $m \times m$ covariance matrix S, and it is related with x_t by means of C. The second one is the transition equation. It relates the state-vector α_t with the state vector at time $t-1$ by means of the transition matrix D. The additive noise of the transition equation is δ_t, assumed to be Gaussian with $r \times r$ covariance matrix Q, and related with α_t by means of F, and uncorrelated with γ_t at all leads and lags.

The system matrices A, B, C, D, F, G, Q and 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 as an observation equation without exogenous variables ($A = \Omega$ and $C = \mathbf{I}$) and 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).

$$\mathbf{y}_t = \Omega\mathbf{f}_t + \boldsymbol{\varepsilon}_t \tag{3}$$

$$\mathbf{f}_t = \mathbf{c}\cdot\mathbf{1} + \Psi\mathbf{f}_{t-1} + \mathbf{w}_t \tag{4}$$

As a particular case, an exogenous variable equal to one 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 the specific ones, thus $\beta_t = 1, \forall t$, B is the null matrix and $F = c$.

For example, if there are r common factors, $\mathbf{f}_t$, that follow a VARIMA$(2,0,0) \times (1,0,0)_4$ model, $\boldsymbol{\phi}(B)\Phi(B^4)\mathbf{f}_t = c + \mathbf{w}_t$, where $\boldsymbol{\phi}(B) = (\mathbf{I} - \boldsymbol{\phi}_1 B - \boldsymbol{\phi}_2 B^2)$ and $\Phi(B^4) = (\mathbf{I} - \Phi_1 B^4)$ are $r \times r$ polynomial matrices. The corresponding transition equation would be $\mathbf{f}_t = \Psi \mathbf{f}_{t-1} + c + \mathbf{w}_t$, where $\mathbf{f}_t$ is a $(r \cdot (2 + 1 \cdot 4))$-dimensional vector containing the common factors at time t and their five lags. The r-dimensional vectors c and $\mathbf{w}_t$ are respectively $(c_1, c_2, ..., c_r)'$ and $(w_{1,t}, ..., w_{m,t}, 0_{r\cdot(p-1)})'$. The $(6r) \times (6r)$ transition matrix is:

$$\Psi = \begin{pmatrix} \Psi_1 & \Psi_2 & \Psi_3 & \Psi_4 & \Psi_5 & \Psi_6 \\ I_r & 0_r & \cdots & \cdots & 0_r & 0_r \\ 0_r & \ddots & \ddots & & & \vdots \\ \vdots & & & & & \\ \vdots & & & \ddots & \ddots & \vdots \\ 0_r & \cdots & & \cdots & I_r & 0_r \end{pmatrix} = \begin{pmatrix} \boldsymbol{\phi}_1 & \boldsymbol{\phi}_2 & 0_r & \Phi_1 & -\boldsymbol{\phi}_1\Phi_1 & -\boldsymbol{\phi}_2\Phi_1 \\ I_r & 0_r & \cdots & \cdots & 0_r & 0_r \\ 0_r & \ddots & \ddots & & & \vdots \\ \vdots & & & & & \\ \vdots & & & \ddots & \ddots & \vdots \\ 0_r & \cdots & & \cdots & I_r & 0_r \end{pmatrix}.$$

2.3 SeaDFA Estimation using the EM Algorithm

The previous state-space formulation depends on a set of parameters $(c, \Psi, \Omega, \mathbf{S}, ...)$ that must be estimated from the observed vector of time series. In this Subsection we present an estimation procedure.

We use maximum likelihood under the assumption that the initial state is normal, $\mathbf{f}_0 \sim N(\boldsymbol{\mu}_0, \mathbf{P}_0^0)$, where $\boldsymbol{\mu}_0$ and $\mathbf{P}_0^0$ are the initial mean and covariance, and the errors $\boldsymbol{\varepsilon}_t$ and $\mathbf{w}_t$ are jointly normal and uncorrelated vector variables. For simplicity we also assume that $\boldsymbol{\varepsilon}_t$ and $\mathbf{w}_t$ are uncorrelated. Although it is not necessary to assume Gaussianity in $\boldsymbol{\varepsilon}_t$ and $\mathbf{w}_t$, certain additional conditions would have to apply and adjustments to the asymptotic covariance matrix would have to be made (Shumway and Stoffer (2006)).

In addition to the non-stationarity included in Peña and Poncela (2004 and 2006), in this work we include the possibility of common factors following a multiplicative seasonal model

with constant. Seasonality introduces some additional non-linear constraints between parameters in matrix Ψ, and modifying the estimation procedure as we will show. Just for simplicity we will explain the estimation algorithm for models with common factors without moving average component.

The estimation is carried out by means of the EM algorithm (Shumway and Stoffer (1982)) using the state space formulation and the Kalman filter. The main problem EM algorithm solves is that the common factors must also be estimated since they are also unknown. The log-likelihood cannot be maximized directly, so it must be maximized iteratively, till convergence is reached. The EM algorithm will be used for this purpose.

We use $\Lambda = \{\mathbf{c}, \Psi, \Omega, \mathbf{S}, \boldsymbol{\mu}_0, \mathbf{P}_0^0\}$ to represent the vector of parameters containing:

- The constant of the model of the common factors, c.
- The transition matrix Ψ including the dynamics of the common factors.
- The loading matrix Ω that relates the observed vector of time series $\mathbf{y}_t$ to the unobserved set of common factors $\mathbf{f}_t$.
- 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.
- The initial condition for the mean and variance of the state variables, $\boldsymbol{\mu}_0$ and $\mathbf{P}_0^0$ respectively.

One should bear in mind that the variance-covariance matrix of the residuals of the model for the common factors, $\mathbf{Q}$ will be fixed as equal to the identity matrix, as explained in Subsection 2.1.

The common factors, $\mathbf{f}_t$, that 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 $\mathbf{Y}_t = \{\mathbf{y}_1, ..., \mathbf{y}_t\}$.

If unobserved common factors $\mathbf{F}_T = \{\mathbf{f}_0, \mathbf{f}_1, ..., \mathbf{f}_T\}$ were known, in addition to the observations $\mathbf{Y}_T = \{\mathbf{y}_0, \mathbf{y}_1, ..., \mathbf{y}_T\}$, then, we would consider the vector of time series $V = \{\mathbf{Y}_T, \mathbf{F}_T\}$,

as the complete data, and its joint density function would be given by the expression:

$$f_{\Lambda}(\mathbf{Y}_T, \mathbf{F}_T) = f_{\boldsymbol{\mu}_0, \mathbf{P}_0^0}(\mathbf{f}_0) \prod_{t=1}^{T} f_{\Psi,Q}(\mathbf{f}_t|\mathbf{f}_{t-1}) \prod_{t=1}^{T} f_S(\mathbf{y}_t|\mathbf{f}_t)$$

Assuming Gaussianity of $\boldsymbol{\varepsilon}_t$ and $\mathbf{w}_t$, the log-likelihood of complete data, i. e., the likelihood of vector $V = \{\mathbf{Y}_T, \mathbf{F}_T\}$, has the following expression:

$$\begin{aligned} \ln L_{Y,F}(\Lambda) &= \frac{-1}{2}\{\ln\left|\mathbf{P}_0^0\right| + (\mathbf{f}_0 - \boldsymbol{\mu}_0)'(\mathbf{P}_0^0)^{-1}\mathbf{f}_0(\mathbf{f}_0 - \boldsymbol{\mu}_0) + \\ &+T\ln|\mathbf{Q}| + \sum_{t=1}^{T}(\mathbf{f}_t - \Psi\mathbf{f}_{t-1} - F\beta_t)'\mathbf{Q}^{-1}(\mathbf{f}_t - \Psi\mathbf{f}_{t-1} - F\beta_t) + \\ &+T\ln|\mathbf{S}| + \sum_{t=1}^{T}(\mathbf{y}_t - \Omega\mathbf{f}_t - B\beta_t)'\mathbf{S}^{-1}(\mathbf{y}_t - \Omega\mathbf{f}_t - B\beta_t)\}. \end{aligned} \tag{5}$$

Using properties of trace, traspose and inverse, and for the particular case in which the exogenous varible is included to estimate the constant, equation (5) can be expressed as:

$$\begin{aligned} -2\ln L_{Y,F}(\Lambda) &= \ln\left|\mathbf{P}_0^0\right| + tr((\mathbf{P}_0^0)^{-1}(\mathbf{f}_0 - \boldsymbol{\mu}_0)(\mathbf{f}_0 - \boldsymbol{\mu}_0)') + \\ &+T\ln|\mathbf{Q}| + \sum_{t=1}^{T} tr(\mathbf{Q}^{-1}(\mathbf{f}_t - \Psi\mathbf{f}_{t-1})(\mathbf{f}_t - \Psi\mathbf{f}_{t-1})') \\ &-2{\cdot}tr\sum_{t=1}^{T} c'\mathbf{Q}^{-1}(\mathbf{f}_t - \Psi\mathbf{f}_{t-1}) + tr\sum_{t=1}^{T}\mathbf{Q}^{-1}cc' \\ &+T\ln|\mathbf{S}| + \sum_{t=1}^{T} tr(\mathbf{S}^{-1}(\mathbf{y}_t - \Omega\mathbf{f}_t)(\mathbf{y}_t - \Omega\mathbf{f}_t)'). \end{aligned} \tag{6}$$

The vector $\widehat{\Lambda}^{(j)} = \left\{\widehat{c}^{(j)}, \widehat{\Psi}^{(j)}, \widehat{\Omega}^{(j)}, \widehat{S}^{(j)}, \widehat{\boldsymbol{\mu}}_0^{(j)}, \widehat{\mathbf{P}}_0^{0(j)}\right\}$ includes all parameters estimated at the j^{th} iteration. Since (6) cannot be maximized directly as the common factors are unknown the EM algorithm provides an iterative method for finding the MLEs of Λ, by successively maximizing the conditional expectation of the complete data likelihood, using only a vector of multivariate time series $\mathbf{y}_t$ and subsequently maximizing the conditional expectation of likelihood. It consists of two steps:

- **E-step** (Expectation step): We calculate the conditional expectation of $-2\ln L_{Y,F}$ defined in (6) given $\mathbf{y}_t$ and $\Lambda^{(j-1)}$. Here we use the properties derived from the Kalman

smoother (Shumway and Stoffer (2006)). The desired conditional expectations are obtained as smoothers.

$$
\begin{aligned}
E\{-2\ln L_{Y,F}(\Lambda)|\mathbf{y}_t,\Lambda^{(j-1)}\} &= \ln\left|\mathbf{P}_0^0\right| + tr((\mathbf{P}_0^0)^{-1}[\mathbf{P}_0^T + (\mathbf{f}_0^T-\boldsymbol{\mu}_0)(\mathbf{f}_0^T-\boldsymbol{\mu}_0)']) + \\
&+n\ln|\mathbf{Q}| + \sum_{t=1}^{T} tr\{\mathbf{Q}^{-1}(\mathbf{S}_{11}-\mathbf{S}_{10}\Psi'-\Psi\mathbf{S}_{10}+\Psi\mathbf{S}_{00}\Psi)\} \\
&-2\cdot tr\{\mathbf{Q}^{-1}\sum_{t=1}^{T}(\mathbf{f}_t^T-\Psi\mathbf{f}_{t-1}^T)c'\} + T\cdot tr(\mathbf{Q}^{-1}cc') \\
&+n\ln|\mathbf{S}| + tr\{S^{-1}\sum_{t=1}^{T}(\mathbf{y}_t-\Omega\mathbf{f}_t^T)(\mathbf{y}_t-\Omega\mathbf{f}_t^T)' + \Omega\mathbf{P}_t^T\Omega`\}. \quad (7)
\end{aligned}
$$

where:

$$
\mathbf{f}_t^t = E(\mathbf{f}_t|\mathbf{Y}_t), \quad \mathbf{P}_t^T = var(\mathbf{f}_t|\mathbf{Y}_t), \quad \mathbf{P}_{t,t-1}^t = cov(\mathbf{f}_t,\mathbf{f}_{t-1}|\mathbf{Y}_t), \quad \mathbf{Y}_t = \{\mathbf{y}_0,\mathbf{y}_1,...,\mathbf{y}_t\}
$$

$$
\mathbf{S}_{00} = \sum_{t=1}^{T}(\mathbf{P}_{t-1}^T + \mathbf{f}_{t-1}^T\mathbf{f}_{t-1}^{T\prime})), \quad \mathbf{S}_{10} = \sum_{t=1}^{T}(\mathbf{P}_{t,t-1}^T + \mathbf{f}_{t-1}^T\mathbf{f}_{t-1}^{T\prime})), \quad \mathbf{S}_{11} = \sum_{t=1}^{T}(\mathbf{P}_t^T + \mathbf{f}_t^T\mathbf{f}_t^{T\prime}))
$$

are obtained after running the Kalman filter and smoother, Shumway and Stoffer (2006).

- **M-step** (Maximization step): The conditional expectation of the log-likelihood, $\ln L_{Y,F}(\Lambda)$, is maximized with respect to the parameters we want to estimate. An explicit expression for $\widehat{\Psi}$ cannot be obtained, since some non-linear constraints appear between the elements of the transition matrix. As an example for illustrating these nonlinear constraints we consider, for instance, for quarterly data, a multiplicative seasonal operator having the form:

$$
(\mathbf{I}_r - \Phi_1 B^4)(\mathbf{I}_r - \boldsymbol{\phi}_1 B - \boldsymbol{\phi}_2 B^2)
$$

The corresponding VAR operator is $\Psi(B) = \mathbf{I}_r - \Psi_1 B - ... - \Psi_6 B^6 = \mathbf{I}_r - \boldsymbol{\phi}_1 B - \boldsymbol{\phi}_2 B^2 - \Phi_1 B^4 + \Phi_1\boldsymbol{\phi}_1 B^5 + \Phi_1\boldsymbol{\phi}_2 B^6$, so that $\Psi_1 = \boldsymbol{\phi}_1, \Psi_2 = \boldsymbol{\phi}_2, \Psi_3 = \mathbf{0}_k, \Psi_4 = \Phi_1, \Psi_5 = -\Phi_1\boldsymbol{\phi}_1, \Psi_6 = -\Phi_1\boldsymbol{\phi}_2$.

Hence, the coefficients $\Psi_1, ..., \Psi_6$, are determined by $\boldsymbol{\phi}_1, \boldsymbol{\phi}_2$ and Φ_1. Due to these non-linear constraints, the terms in equation (7) involving Ψ, where the vector of constants also appears.

$$\sum_{t=1}^{T} tr\{\mathbf{Q}^{-1}(\mathbf{S}_{11} - \mathbf{S}_{10}\Psi' - \Psi\mathbf{S}_{10} + \Psi\mathbf{S}_{00}\Psi)\} - 2{\cdot}tr\{\mathbf{Q}^{-1}\sum_{t=1}^{T}(\mathbf{f}_t^T - \Psi\mathbf{f}_{t-1}^T)c'\} + T{\cdot}tr(\mathbf{Q}^{-1}cc') \tag{8}$$

should be minimized respect to Ψ by means of an optimization procedure with non-linear restrictions. We have used a subspace trust region method and it is based on the interior-reflective Newton method described in Coleman and Li (1994, 1996). Each iteration involves the approximate solution of a large linear system using the method of Preconditioned Conjugate Gradients (PCG). The TOMLAB toolbox for Matlab 6.5 has been used for its computational implementation.

The other parameters we want to estimate (not affected by non-linear constraints), can be obtained maximizing (7) with respect to them, which yields:

$$vec\widehat{\Omega} = \mathbf{E}_2\mathbf{S}_{11}^{-1} + (\mathbf{S}_{11}^{-1} \otimes S)\mathbf{F}_2'(\mathbf{F}_2(\mathbf{S}_{11}^{-1} \otimes S)\mathbf{F}_2')^{-1}(\mathbf{g}_2 - \mathbf{F}_2 vec(\mathbf{E}_2\mathbf{S}_{11}^{-1})), \tag{9}$$

where $\mathbf{E}_2 = \sum_{t=1}^{T}(\mathbf{y}_t\mathbf{f}_t^{T\prime} - \mathbf{M}\mathbf{f}_t^{T\prime})$, $\mathbf{M} = \sum_{t=1}^{T}(\mathbf{y}_t - \Omega\mathbf{f}_t^T)$, and $\mathbf{F}_2\ vec\Omega = \mathbf{g}_2$ allows to include linear restrictions between the parameters in Ω (see Wu, Pai and Hosking (1996)). $\mathbf{F}_2$ is a matrix whose number of rows is equal to the number of linear restrictions to be imposed on Ω and its number of column coincides with the length of $vec\Omega$. And $\mathbf{g}_2$ is a vector whose length is equal to the number of restrictions.

Finally, the value of S that maximizes the conditional expectation of the likelihood is:

$$\widehat{S} = \frac{1}{T}\sum_{t=1}^{T}((\mathbf{y}_t - \widehat{\Omega}\mathbf{f}_t^T)(\mathbf{y}_t - \widehat{\Omega}\mathbf{f}_t^T)' + \widehat{\Omega}\mathbf{P}_t^T\widehat{\Omega}`), \quad \widehat{\boldsymbol{\mu}}_0 = \mathbf{f}_0^T, \quad \widehat{\mathbf{P}}_0^0 = \mathbf{P}_0^T. \tag{10}$$

The E and M step are repeated alternatively till convergence is reached, i.e., till the difference between conditional likelihood in two consecutive iterations is small enough.

Summarizing, the steps to estimate the SeaDFA using EM would be:

1. Initialize the procedure, giving initial values to $\mathbf{c}, \Psi, \Omega, S, \boldsymbol{\mu}_0$ and $\mathbf{P}_0^0$. $\mathbf{Q}$ is fixed to be the identity matrix.

2. Specify F_2 and g_2, bearing in mind the linear constraints affecting Ω.

3. Obtain $\mathbf{f}_t^t, \mathbf{f}_t^T$, $\mathbf{P}_t^T$, $\mathbf{P}_{t,t-1}^T$ $\forall t$ from the Kalman filter and smoother.

4. Calculate $\Psi^{(i+1)}$ maximizing expression (8) subject to the corresponding nonlinear constraints, and $\Omega^{(i+1)}$, $\mathbf{S}^{(i+1)}$, $\widehat{\boldsymbol{\mu}}_0$ and $\widehat{\mathbf{P}}_0^0$ from equations (9) and (10), as well as the conditional likelihood (7).

5. If $(E(-2\ln L_{Y,F}(\Lambda)) - E(-2\ln L_{Y,X}(\Lambda^{i-1})) < \varepsilon$, with ε small enough and prefixed, then stop.

If convergence has not been reached, then steps 3, 4 and 5 are iteratively repeated.

3 Bootstrap scheme for Seasonal Dynamic Factor Analysis

In this Section we provide a bootstrap scheme for assessing uncertainty to 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 Subsection 2.1). This bootstrap scheme is able to asses precision of estimates of any linear 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 model by state-space methods (see Harvey (1992), Durbin and Koopman (2001)).

Application of classical inference methods relying on asymptotic theory is subject to the disposal of large data sets, as investigated by Ansley and Newbold (1980), among others. For this reason bootstrap techniques are a powerful alternative to inference procedures based on Fisher Information Matrix. Moreover, bootstrap methods have a main advantage, since they allow to take into account the uncertainty due to parameter estimation, which enhance the coverage of the confidence intervals for the forecasts.

The existence, under certain conditions, of this 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)) 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 procedure is an alternative that generates replicas of the model resampling not only from the observation equation residuals but also from the transition equation, once consistent estimates of parameters and state variables have been obtained.

In Subsection 3.1 we present the bootstrap scheme for making inference on the parameters of the model. Bootstrap scheme for computing forecast intervals is provided in Subsection 3.2. A simulation study is carried out in Section 4.

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 loading matrix Ω, as well as for the parameters of the VARMA model for the common factors, Ψ, and the constant $\mathbf{c}$, and 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 seven subsequent steps:

1. The model defined by (3), (4) is estimated following the EM algorithm described in Section 2. Once this has been completed, as explained in section 2.3, the parameters involved, $\widehat{\mathbf{c}}, \widehat{\Psi}, \widehat{\Omega}, \widehat{\mathbf{S}}, \widehat{\boldsymbol{\mu}}_0, \widehat{\mathbf{P}}_0^0$, are available. Besides, we have consistent estimates, $\widehat{\mathbf{f}}_t$, of the state variables, $\mathbf{f}_t$ derived from the Kalman filter at the last iteration.

2. The specific factors are calculated: $\widehat{\boldsymbol{\varepsilon}}_t = \mathbf{y}_t - \widehat{\Omega}\widehat{\mathbf{f}}_t$.

   One should bear in mind the relationship between $\widehat{\boldsymbol{\varepsilon}}_t$ and $\boldsymbol{\varepsilon}_t$, and their variance-covariance matrices. $\mathbf{S} = \mathbf{E}[\boldsymbol{\varepsilon}_\mathbf{t}\boldsymbol{\varepsilon}_\mathbf{t}']$ and $\mathbf{S}_{\widehat{\boldsymbol{\varepsilon}}} = \mathbf{E}[\widehat{\boldsymbol{\varepsilon}}_\mathbf{t}\widehat{\boldsymbol{\varepsilon}}_\mathbf{t}']$. Since $\boldsymbol{\varepsilon}_t = \widehat{\boldsymbol{\varepsilon}}_t + (\boldsymbol{\varepsilon}_t - \widehat{\boldsymbol{\varepsilon}}_t)$ it is verified that $\mathbf{S} = \mathbf{S}_{\widehat{\boldsymbol{\varepsilon}}} + var(\boldsymbol{\varepsilon}_t - \widehat{\boldsymbol{\varepsilon}}_t) = \mathbf{S}_{\widehat{\boldsymbol{\varepsilon}}} + \mathbf{S}_{correction}$, because the cross products are zero. The term $\mathbf{S}_{correction} = var(\boldsymbol{\varepsilon}_t - \widehat{\boldsymbol{\varepsilon}}_t)$ can be considered as a *correction factor*, as in the model pro-

posed in Harvey, Ruiz and Sentana (1992). An expression for $var(\boldsymbol{\varepsilon}_t - \widehat{\boldsymbol{\varepsilon}}_t)$ in terms of the estimated loads and the variance of the state variables is derived in Appendix A.

3. Draw B resamples, $\boldsymbol{\varepsilon}^*_{e,t}$, from $F_{\widetilde{\boldsymbol{\varepsilon}}}$, the empirical distribution function of the centered and corrected specific factors,

$$F_{\widetilde{\boldsymbol{\varepsilon}}}(x) = \frac{1}{T}\sum_{t=1}^{T} I(\widetilde{\boldsymbol{\varepsilon}}_{e,t} \leq x), \widetilde{\boldsymbol{\varepsilon}}_{e,t} = \mathbf{S}^{-1/2}_{\widehat{\boldsymbol{\varepsilon}}}(\widehat{\boldsymbol{\varepsilon}}_t - \overline{\boldsymbol{\varepsilon}})\widehat{\mathbf{S}}^{1/2}.$$

4. Calculate $\widehat{\mathbf{w}}_t = \widehat{\mathbf{f}}_t - \widehat{\Psi}\widehat{\mathbf{f}}_{t-1}$.

5. Draw B resamples, $\mathbf{w}^*_{j,t}$, from $F_{\widetilde{\mathbf{w}}}$, the empirical distribution function of the standardized residuals of the VARIMA model for the common factors.

$$F_{\widetilde{\mathbf{w}}}(x) = \frac{1}{T}\sum_{t=1}^{T} I(\widetilde{\mathbf{w}}_{j,t} \leq x) \text{ and } \widetilde{w}_{t,j} = \frac{\widehat{w}_{t,j} - \overline{w}_{t,j}}{\sigma_{\widehat{w}_{j,t}}}, \text{ where } j = 1, ..., r.$$

   We use the standardized residuals since one of the constraints imposed for identifiability of the SeaDFA is $\mathbf{Q} = \mathbf{I}$. In the general case of a linear state-space model we would reescale $\widehat{\mathbf{w}}_t$ using $\widehat{\mathbf{Q}}$ and $\widehat{\mathbf{Q}}_{\widehat{\mathbf{w}}} = E[\widehat{\mathbf{w}}_t\widehat{\mathbf{w}}'_t]$, as it was done for the estimated specific factors, $\widehat{\boldsymbol{\varepsilon}}_t$ to obtain $\widetilde{\boldsymbol{\varepsilon}}_t$.

6. Generate B bootstrap replicas of the common factors using the transition equation:

$$\mathbf{f}^*_t = \widehat{\Psi}\mathbf{f}^*_{t-1} + \mathbf{w}^*_t.$$

7. Generate B bootstrap replicates of the SeaDFM, using bootstrap replicas of the common and specific factors obtained in steps 3 and 6, respectively:

$$\mathbf{y}^*_t = \widehat{\Omega}\mathbf{f}^*_t + \boldsymbol{\varepsilon}^*_t.$$

From estimating the SeaDFA for each replica obtained in step 7, we have $\widehat{\mathbf{c}}^*, \widehat{\Psi}^*, \widehat{\Omega}^*$ and $\widehat{\mathbf{S}}^*$ and their respective bootstrap distribution functions, $\widehat{F}^*_{\widehat{\mathbf{c}}^*}\widehat{F}^*_{\widehat{\Psi}^*}\widehat{F}^*_{\widehat{\Omega}^*}\widehat{F}^*_{\widehat{\mathbf{S}}^*}$. 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 constant of the model, $q^*(\cdot) = \widehat{F}^{*-1}_{\widehat{\mathbf{c}}^*}$.

Finally, bootstrap confidence intervals for the loads, ω_{ij}, and VARMA 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 Ω^* and Ψ^* respectively.

The percentile based confidence intervals for loads and VARMA parameters will allow, for example, testing the equality of loads, or if the parameters of the VARMA model of the common factors are significant or not. The results obtained can be used to impose constraints between loads or VARMA parameters that can be applied in a subsequent estimation of the SeaDFA.

3.2 Bootstrap procedure for forecasting

Related to forecasting, the main objective is to obtain not only point forecasts but also an uncertainty measure for them. Bootstrap techniques have been applied for this purpose (Alonso et al. (2002), Thombs and Schucany (1990)). The previous scheme 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. The final state is fixed for the common factors $\mathbf{f}_t$.

The first steps of the bootstrap procedure for forecasting coincides with the seven steps proposed in the previous Subsection. The following are the forecasting steps.

8. The future bootstrap observations for common factors are calculated using the relationship: $\mathbf{f}^*_{t+h} = \widehat{\Psi}^* \mathbf{f}^*_{t+h-1} + \mathbf{w}^*_{t+h}$,
   where $\mathbf{f}^*_t = \widehat{\mathbf{f}}_t$ if $t \leq T$, T is the length of the vector of time series, $\mathbf{f}^*_{t+h} = (f^*_{t+h,1}, f^*_{t+h,2}, ...f^*_{t+h,r})'$ and the future bootstrap observations $\boldsymbol{\varepsilon}^*_{e,t+h} = (\varepsilon^*_{e,t+h}, \varepsilon^*_{e,t+h}, ..., \varepsilon^*_{e,t+h})'$ are generated resampling from $F_{\widetilde{\boldsymbol{\varepsilon}}}$.

9. The future bootstrap observations are calculated for vector $\mathbf{y}_t$ using the relation $\mathbf{y}^*_{t+h} = \widehat{\Omega}^* \mathbf{f}^*_{t+h} + \boldsymbol{\varepsilon}^*_{t+h}$.

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 are obtained

using the quantiles of the bootstrap distribution function $F^*_{y^*_{t+h}}$. The $(1-\alpha)\%$ forecast interval for $\mathbf{y}_{t+h}$ is the following:

$$[q^*(\alpha/2), q^*(1-\alpha/2)],$$

where $q^*(\cdot) = \widehat{F}^{*-1}_{\mathbf{y}^*_{t+h}}$ are the quantiles of the estimated bootstrap distribution.

4 Simulation study

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

Model 1: In the first experiment a common nonstationary factor (common trend, I(1) with constant ($c = 3$)), for $m = 4$ observed series. $\Omega = (1, 1, 1, 1)'$, $E(\boldsymbol{\varepsilon}_t) = 0$, $\text{var}(\boldsymbol{\varepsilon}_t) = \mathbf{S} = \mathbf{I}_4$ and $E(\mathbf{w}_t) = 0$, $\text{var}(\mathbf{w}_t) = \mathbf{Q} = \boldsymbol{\sigma}^2_w = 1$. This model has been selected because it appears in 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: The second model considers a common nonstationary factor $(1-B)(1-0.5B)\mathbf{f}_t = \mathbf{w}_t$, for $m = 3$ observed series, $\Omega = (1, 1, 1)'$, $E(\boldsymbol{\varepsilon}_t) = 0$, $\text{var}(\boldsymbol{\varepsilon}_t) = \mathbf{S} = \mathbf{I}_4$, $E(\mathbf{w}_t) = 0$, $\text{var}(\mathbf{w}_t) = \mathbf{Q} = \boldsymbol{\sigma}^2_w = 1$.

Model 3: In the third model under study we check the performance of our procedure when there is a seasonal pattern. There is a common nonstationary factor following a seasonal multiplicative ARIMA model $(1 - B^7)(1 - 0.4B)(1 - 0.15B^7)\mathbf{f}_t = \mathbf{w}_t$, for $m = 4$ observed series, $\Omega = (1, 1, 1, 1)'$, $E(\boldsymbol{\varepsilon}_t) = 0$, $\text{var}(\boldsymbol{\varepsilon}_t) = \mathbf{I}_3$, $E(\mathbf{w}_t) = 0$, $\text{var}(\mathbf{w}_t) = \boldsymbol{\sigma}^2_w = 1$.

$R = 100$ 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, $h = 1, 3$, while three sample sizes, T, have been considered: $50, 100$ and 200. For each vector of series simulated and estimated $B = 500$ bootstrap resamples have been generated as described in the previous section, and the corresponding model was estimated. The $(1 - \alpha)\%$ prediction intervals

$[Q^*_M(\alpha/2), Q^*_M(1-\alpha/2)]$ were computed.

The coverage is estimated as $C_M = \sharp\{Q^*_M(\alpha/2) \leq \mathbf{y}^P_{T+h} \leq Q^*_M(1-\alpha/2)\}$, where $\mathbf{y}^P_{T+h}$ is the vector of future values generated in first step. Meanwhile, using $L_T = \mathbf{y}^{P(1-\alpha/2)}_{T+h} - \mathbf{y}^{P(\alpha/2)}_{T+h}$, 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 $\overline{C}_T$ is the nominal coverage.

The results for Model 1, and nominal coverage 80% and 95%, are shown in Table 1 and Table 2 respectively. It can be observed that even for small or moderate lengths the values obtained for coverages are correct. Besides, and although we focused on coverages for the forecasts, when checking the coverages for the parameters of the model, the results obtained are also correct.

The results for Model 2 and 3, and nominal coverages 80% and 95% are shown in Tables 3, 4, 5 and 6 respectively. The bootstrap scheme developed for SeaDFA also performs reasonably well in Model 2 and Model 3 since the coverage and length tend to the nominal values as the sample size grows.

The results for Model 2, and nominal coverages 80% and 95%, are shown in Tables 3 and 4 respectively.

5 Application: 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. Most previous works compute short term forecasts for electricity prices (Nogales et al. (2002), Contreras et al. (2003), Conejo et al. (2005), García-Martos, Rodríguez y Sánchez (2007)), load forecasts (Cottet and Smith (2003)), or analysis of several markets spot prices (Koopman, Ooms and Carnero (2007)), but long term forecasting of electricity prices is a difficult issue not commonly tackled up to now. Besides, we have selected the Spanish market, which is less predictable than others sometimes considered for study, such as PJM interconnection or Nordpool. Less predictability

Table 1: Model 1. Coverage 80 percent.

Lag	Sample size	Series	C_M (se)	Coverage (below/above)	L_T	L_B (se)
h	T	m	Theoretical 80%	10.0%/10.0%		
1	50	1	76.697 (0.524)	11.578/11.725	3.601	3.561 (0.027)
		2	77.252 (0.446)	10.563/12.185	3.603	3.589 (0.025)
		3	76.522 (0.407)	10.599/12.879	3.610	3.575 (0.025)
		4	76.539 (0.512)	10.666/12.795	3.626	3.572 (0.026)
	100	1	76.757 (0.429)	11.871/11.372	3.619	3.561 (0.019)
		2	77.504 (0.483)	11.710/10.786	3.621	3.614 (0.020)
		3	77.418 (0.418)	12.137/10.445	3.610	3.614 (0.021)
		4	77.036 (0.410)	12.270/10.694	3.624	3.591 (0.023)
	200	1	77.443 (0.493)	10.470/12.087	3.609	3.610 (0.020)
		2	76.698 (0.461)	11.120/12.182	3.627	3.555 (0.017)
		3	77.285 (0.409)	10.794/11.921	3.625	3.592 (0.018)
		4	76.749 (0.455)	10.979/12.272	3.629	3.571 (0.019)
h	T	m	Theoretical 80%	10%/10%	L_T	L_B (se)
3	50	1	78.325 (0.322)	10.458/11.217	5.138	5.115 (0.028)
		2	78.402 (0.368)	10.101/11.497	5.114	5.109 (0.031)
		3	78.227 (0.310)	9.964/11.809	5.101	5.091 (0.023)
		4	78.319 (0.340)	10.033/11.648	5.121	5.113 (0.030)
	100	1	78.300 (0.290)	10.889/10.811	5.135	5.101 (0.022)
		2	78.259 (0.276)	11.321/10.420	5.142	5.098 (0.023)
		3	78.621 (0.272)	11.261/10.118	5.126	5.135 (0.024)
		4	78.492 (0.290)	11.384/10.124	5.119	5.101 (0.023)
	200	1	78.437 (0.327)	10.235/11.328	5.098	5.094 (0.023)
		2	78.653 (0.317)	10.198/11.149	5.105	5.120 (0.023)
		3	78.714 (0.296)	10.077/11.209	5.105	5.121 (0.023)
		4	78.307 (0.318)	10.319/11.374	5.129	5.080 (0.022)

Table 2: Model 1. Coverage 95 percent.

Lag	Sample size	Series	C_M (se)	Coverage (below/above)	L_T	L_B (se)
h	T	m	Theoretical 95%	2.5%2.5%		
1	50	1	93.161 (0.327)	2.989/3.850	5.559	5.513 (0.041)
		2	93.376 (0.262)	2.663/3.961	5.515	5.536 (0.038)
		3	92.897 (0.297)	2.684/4.419	5.549	5.488 (0.037)
		4	92.849 (0.323)	2.786/4.365	5.541	5.497 (0.040)
	100	1	93.323 (0.243)	3.283/3.394	5.549	5.506 (0.030)
		2	93.496 (0.297)	3.297/3.207	5.539	5.556 (0.035)
		3	93.484 (0.242)	3.340/3.176	5.527	5.545 (0.035)
		4	93.477 (0.275)	3.429/3.094	5.528	5.523 (0.034)
	200	1	93.411 (0.322)	2.909/3.680	5.517	5.504 (0.028)
		2	93.296 (0.274)	2.987/3.717	5.531	5.474 (0.030)
		3	93.358 (0.270)	3.027/3.615	5.537	5.510 (0.030)
		4	93.069 (0.316)	3.117/3.814	5.552	5.488 (0.032)
h	T	m	Theoretical 95%	2.5%/2.5%	L_T	L_B (se)
3	50	1	94.123 (0.193)	2.473/3.404	7.845	7.819 (0.040)
		2	93.961 (0.195)	2.544/3.495	7.834	7.839 (0.048)
		3	93.790 (0.169)	2.346/3.864	7.798	7.786 (0.033)
		4	94.005 (0.186)	2.452/3.543	7.853	7.844 (0.045)
	100	1	93.927 (0.185)	2.788/3.285	7.838	7.784 (0.035)
		2	94.150 (0.171)	2.943/2.907	7.850	7.848 (0.039)
		3	94.083 (0.161)	3.015/2.902	7.845	7.852 (0.037)
		4	93.981 (0.177)	3.030/2.989	7.840	7.806 (0.038)
	200	1	94.009 (0.176)	2.694/3.297	7.835	7.794 (0.033)
		2	94.006 (0.170)	2.758/3.236	7.837	7.785 (0.033)
		3	94.063 (0.174)	2.814/3.123	7.815	7.798 (0.038)
		4	93.962 (0.192)	2.781/3.257	7.806	7.766 (0.032)

Table 3: Model 2. Coverage 80 percent.

Lag	Sample size	Series	C_M (se)	Coverage (below/above)	L_T	L_B (se)
h	T	m	Theoretical 80%	10%/10%		
1	50	1	74.822 (0.979)	11.855/13.098	3.634	3.749 (0.040)
		2	74.330 (1.041)	12.268/13.402	3.639	3.747 (0.039)
		3	74.993 (0.901)	11.979/12.834	3.623	3.754 (0.039)
	100	1	73.969 (0.815)	11.438/14.974	3.634	3.643 (0.029)
		2	75.105 (0.835)	11.638/13.255	3.639	3.718 (0.027)
		3	74.870 (0.878)	11.492/13.961	3.623	3.657 (0.026)
	200	1	73.833 (0.693)	13.407/12.477	3.632	3.577 (0.021)
		2	74.793 (0.658)	13.131/12.376	3.621	3.563 (0.018)
		3	73.904 (0.679)	13.479/12.487	3.636	3.561 (0.019)
h	T	m	Theoretical 80%	10%/10%	L_T	L_B (se)
3	50	1	75.372 (0.621)	11.931/12.255	6.668	6.525 (0.078)
		2	75.444 (0.676)	12.236/12.211	6.672	6.532 (0.082)
		3	75.462 (0.597)	12.011/12.303	6.668	6.514 (0.082)
	100	1	76.873(0.431)	10.503/12.823	6.668	6.617 (0.043)
		2	77.703 (0.397)	10.791/12.350	6.672	6.658 (0.046)
		3	77.219 (0.455)	10.912/12.605	6.668	6.598 (0.047)
	200	1	77.787 (0.424)	11.654/10.988	6.659	6.698 (0.044)
		2	78.014 (0.431)	11.242/11.229	6.662	6.675 (0.046)
		3	78.021 (0.391)	11.690/11.222	6.651	6.669 (0.048)

Table 4: Model 2. Coverage 95 percent.

Lag	Sample size	Series	C_M (se)	Coverage (below/above)	L_T	L_B (se)
h	T	m	Theoretical 95%	2.5%/2.5%		
1	50	1	91.842 (0.681)	3.721/4.182	5.559	5.866 (0.084)
		2	91.702 (0.638)	3.729/4.432	5.571	5.880 (0.083)
		3	92.169 (0.546)	3.577/4.084	5.543	5.852 (0.080)
	100	1	92.116 (0.467)	3.481/4.916	5.559	5.674 (0.052)
		2	92.501 (0.583)	3.685/4.232	5.571	5.763 (0.048)
		3	92.194 (0.557)	3.185/4.560	5.543	5.715 (0.047)
	200	1	91.716 (0.483)	4.431/4.143	5.550	5.515 (0.032)
		2	91.943 (0.422)	4.309/3.907	5.542	5.519 (0.033)
		3	91.972 (0.441)	4.404/3.924	5.552	5.537 (0.036)
h	T	m	Theoretical 95%	2.5%/2.5%	L_T	L_B (se)
3	50	1	92.122 (0.423)	3.764/4.034	10.212	10.031 (0.123)
		2	92.150 (0.433)	3.967/3.955	10.212	10.047 (0.131)
		3	92.406 (0.387)	3.642/3.753	10.169	10.094 (0.127)
	100	1	93.801 (0.231)	2.981/3.800	10.212	10.242 (0.076)
		2	93.743 (0.232)	2.948/3.421	10.212	10.370 (0.079)
		3	93.418 (0.267)	3.077/3.528	10.169	10.256 (0.085)
	200	1	93.982 (0.252)	3.125/3.004	10.079	10.390 (0.071)
		2	94.071 (0.229)	2.974/3.128	10.161	10.369 (0.069)
		3	94.053 (0.255)	3.005/3.223	10.120	10.320 (0.070)

Table 5: Model 3. Coverage 80 percent.

Lag	Sample size	Series	C_M (se)	Coverage (below/above)	L_T	L_B (se)
h	T	m	Theoretical 80%	10.0%/10%		
1	50	1	72.143 (0.794)	13.435/14.422	3.642	3.409 (0.033)
		2	71.917 (0.753)	13.514/14.569	3.628	3.350 (0.030)
		3	71.020 (0.737)	14.066/14.914	3.624	3.305 (0.030)
		4	71.552 (0.733)	13.834/14.614	3.618	3.340 (0.030)
	100	1	70.871 (0.725)	15.964/13.165	3.623	3.277 (0.026)
		2	71.339 (0.599)	15.337/13.324	3.601	3.275 (0.020)
		3	71.427 (0.641)	15.283/13.290	3.631	3.276 (0.023)
		4	71.208 (0.538)	15.154/13.638	3.627	3.263 (0.024)
	200	1	71.427 (0.461)	12.902/15.671	3.623	3.259 (0.018)
		2	70.759 (0.569)	13.784/15.457	3.629	3.223 (0.022)
		3	70.851 (0.577)	13.837/15.312	3.637	3.225 (0.021)
		4	71.484 (0.553)	13.053/15.463	3.609	3.251 (0.020)
h	T	m	Theoretical 80%	10.0%/10%		
3	50	1	72.564 (0.670)	14.088/13.348	3.785	3.499 (0.032)
		2	72.856 (0.728)	14.046/13.098	3.799	3.490 (0.031)
		3	71.837 (0.686)	14.690/13.473	3.795	3.432 (0.030)
		4	72.019 (0.677)	14.589/13.392	3.794	3.459 (0.030)
	100	1	72.827 (0.509)	14.060/13.113	3.801	3.447 (0.027)
		2	72.504 (0.503)	14.373/13.123	3.799	3.422 (0.023)
		3	72.185 (0.454)	14.319/13.496	3.789	3.405 (0.028)
		4	72.254 (0.450)	14.182/13.564	3.805	3.425 (0.025)
	200	1	71.7800. (480)	13.606/14.614	3.789	3.391 (0.019)
		2	72.095 (0.471)	13.512/14.393	3.793	3.421 (0.022)
		3	71.823 (0.469)	13.752/14.425	3.796	3.395 (0.020)
		4	71.735 (0.473)	13.678/14.587	3.784	3.387 (0.021)

Table 6: Model 3. Coverage 95 percent.

Lag	Sample size	Series	C_M (se)	Coverage (below/above)	L_T	L_B (se)
h	T	m	Theoretical 95%	2.5%/2.5%		
1	50	1	91.841 (0.503)	3.676/4.483	5.552	5.501 (0.049)
		2	91.597 (0.457)	3.855/4.548	5.555	5.400 (0.043)
		3	91.386 (0.406)	4.02874.586	5.543	5.356 (0.046)
		4	91.841 (0.447)	3.874/4.285	5.536	5.413 (0.048)
	100	1	90.985 (0.474)	5.001/4.014	5.513	5.310 (0.042)
		2	91.176 (0.390)	4.827/3.997	5.521	5.312 (0.034)
		3	91.319 (0.363)	4.671/4.010	5.555	5.312 (0.041)
		4	90.996 (0.368)	4.943/4.061	5.577	5.269 (0.040)
	200	1	90.621 (0.349)	4.020/5.359	5.547	5.180 (0.037)
		2	90.320 (0.378)	4.346/5.334	5.562	5.097 (0.040)
		3	90.368 (0.369)	4.348/5.284	5.546	5.131 (0.034)
		4	90.636 (0.372)	4.080/5.284	5.484	5.125 (0.038)
h	T	m	Theoretical 95%	2.5%/2.5%		
3	50	1	91.693 (0.460)	4.075/4.232	5.758	5.619 (0.050)
		2	91.295 (0.456)	4.404/4.301	5.808	5.559 (0.052)
		3	90.866 (0.438)	4.627/4.507	5.817	5.510 (0.042)
		4	91.043 (0.428)	4.469/4.488	5.828	5.529 (0.047)
	100	1	92.071 (0.340)	4.025/3.904	5.780	5.549 (0.044)
		2	92.029 (0.315)	4.296/3.675	5.814	5.563 (0.040)
		3	91.579 (0.294)	4.296/4.125	5.816	5.478 (0.045)
		4	91.936 (0.299)	4.091/3.973	5.800	5.526 (0.044)
	200	1	91.490 (0.376)	4.189/4.321	5.791	5.503 (0.050)
		2	91.559 (0.329)	4.063/4.378	5.793	5.509 (0.049)
		3	91.359 (0.344)	4.286/4.355	5.798	5.438 (0.043)
		4	91.441 (0.311)	3.952/4.607	5.792	5.481 (0.041)

can be justified from two points of view:

1. There is a higher proportion of outliers and a lesser degree of competition.
2. During peak hours the Spanish market shows even higher dispersion. This fact causes more uncertainty in periods of high demand, producing less accurate forecasts.

Because of these reasons computing long-term forecasts for electricity markets in the Spanish market is a good challenge for testing the performance of the SeaDFA, and also a novelty since there is no related previous published works.

Our objective is calculating forecasts for the whole year 2004 using data from 1^{st} January 1998 up to 31^{st} December 2003, so the forecasting horizon ranges from 1 day up to 1 year. In Subsection 5.1 the numerical results concerning the estimation and inference of SeaDFA for electricity prices are shown. In Subsection 5.2 forecasting results are provided.

5.1 Estimation of SeaDFA for electricity prices in the Spanish Market

The main objective is forecasting the hourly prices for the whole year 2004 using the SeaDFA estimated for 1998-2003 data. A 24-dimensional vector of time series can be built if considering the series of prices in the 24 hours of each day. This is known as the parallel approach, some references including this modelling are Grady et al. (1991) and Cottet and Smith (2003). The seasonality we must deal with when using electricity market data is weekly. The frequency has been reduced when splitting the complete time series into 24 hourly time series and daily seasonal pattern does not appear. Figure 2 shows the 24 hourly time series considered. And Figure 3 a detail of the last two months in 2003. A common dynamic in the vector of time series of hourly prices can be observed.

The test proposed by Peña and Poncela (2006) for our vector of 24 hourly time series, gives a preliminary idea about the number of common factors, in our case two factors. Furthermore, we also check the diagnostics by means of the *auxiliary residuals*, using the procedure proposed by Harvey and Koopman (1992), and the previous result about the number of factors is

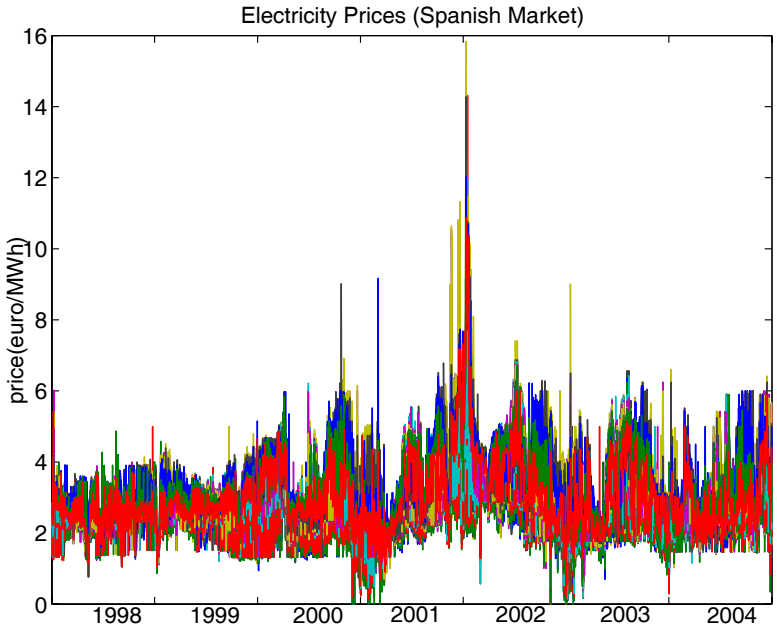

Figure 2: 24 hourly time series of electricity prices (1998-2003).

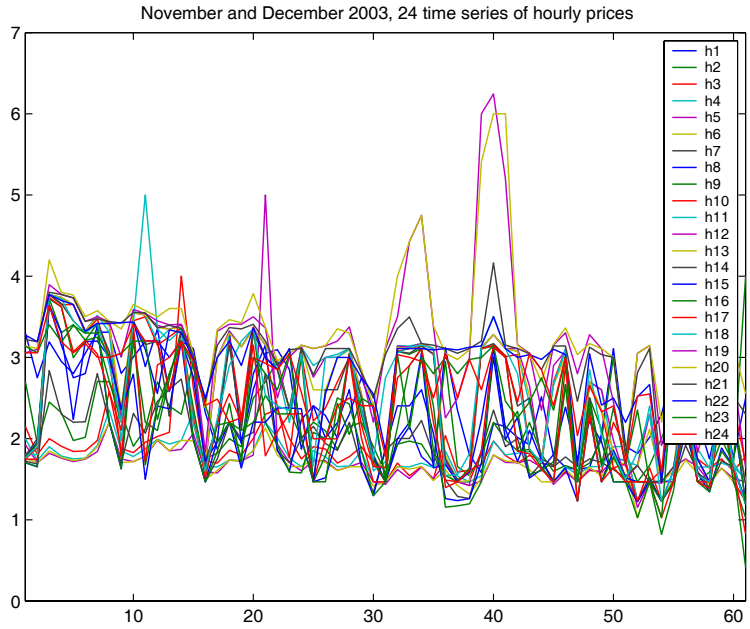

Figure 3: 24 hourly time series of electricity prices (September-December 2003).

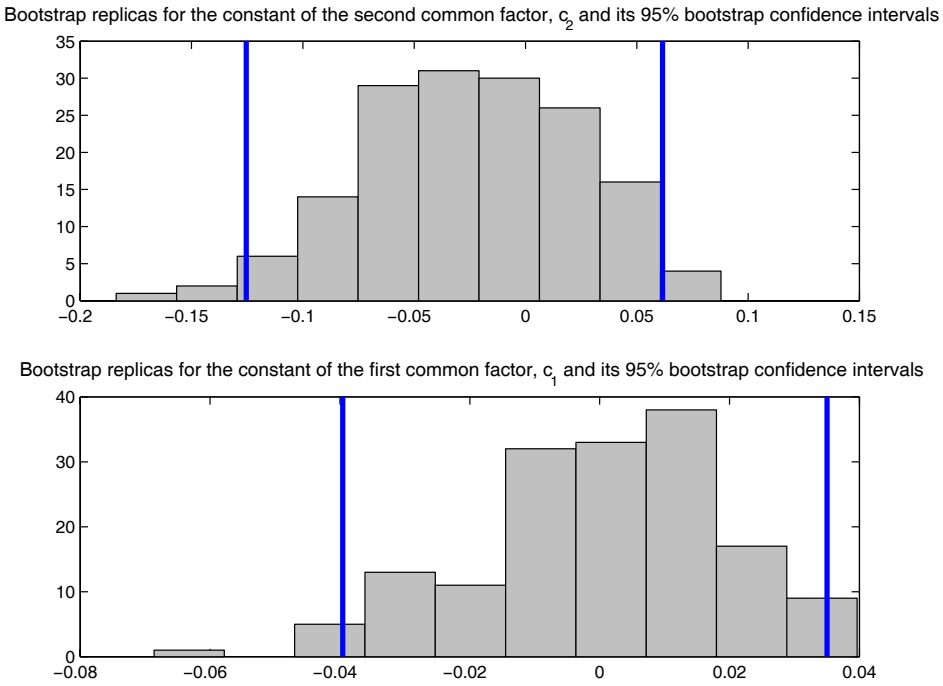

Figure 4: Histogram for bootstrap replicates of the constant of the models, c_1 and c_2.

confirmed. It is not the case of our application but if the diagnostics was not correct we would increase the number of common factors and reestimate the model.

For the two unobserved common factors, we have fitted a VARIMA $(1,0,0)\times(1,1,0)_7$. The equation of this model is given in (11).

$$[I-B^7]\left[I-\begin{pmatrix}\Phi_{1,11} & \Phi_{1,12}\\ \Phi_{1,21} & \Phi_{1,22}\end{pmatrix}B^7\right]\left[I-\begin{pmatrix}\phi_{1,11} & \phi_{1,12}\\ \phi_{1,21} & \phi_{1,22}\end{pmatrix}B\right]\begin{pmatrix}f_{1,t}\\ f_{1,t}\end{pmatrix}=\begin{pmatrix}c_1\\ c_2\end{pmatrix}+\begin{pmatrix}u_{1,t}\\ u_{1,t}\end{pmatrix} \quad (11)$$

We have estimated this model, and used the bootstrap procedure described to make inference on the parameters involved. We have detected that the constant is not significant (this can be observed in Figure 4, note that zero is included in the 95% percentile based confidence interval of the constants).

The model is reestimated including these constraints, i.e.:

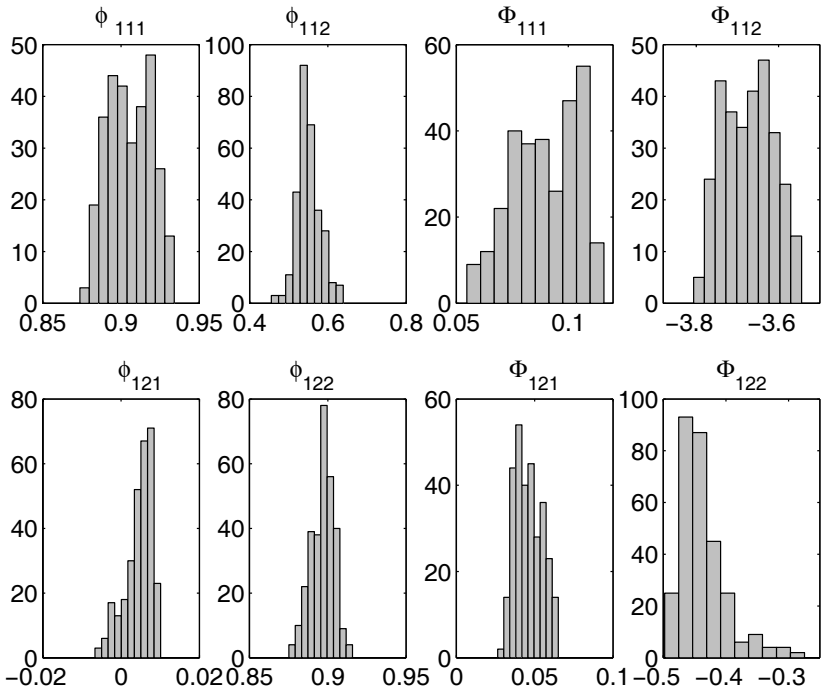

Figure 5: Histogram for bootstrap replicates of the parameters of the models.

$$[I - B^7]\left[I - \begin{pmatrix} \Phi_{111} & \Phi_{112} \\ \Phi_{121} & \Phi_{122} \end{pmatrix} B^7\right]\left[I - \begin{pmatrix} \phi_{111} & \phi_{112} \\ \phi_{121} & \phi_{122} \end{pmatrix} B\right]\begin{pmatrix} f_{1,t} \\ f_{2,t} \end{pmatrix} = \begin{pmatrix} w_{1,t} \\ w_{2,t} \end{pmatrix}$$

The coefficient ϕ_{121}, that relates the first common factor $f_{1,t}$ with the first lag of the second factor is not significant as it is shown in Figure 5. Once again the test for significance is percentile based, using the bootstrap distribution function.

When the SeaDFA is again reestimated including the constraint $\phi_{121} = 0$, all the other coefficients remain significant, finally the model estimated for the set of two unobserved common factors is given in 12.

$$[I - B^7]\left[I - \begin{pmatrix} \Phi_{111} & \Phi_{112} \\ \Phi_{121} & \Phi_{122} \end{pmatrix} B^7\right]\left[I - \begin{pmatrix} \phi_{111} & \phi_{112} \\ 0 & \phi_{122} \end{pmatrix} B\right]\begin{pmatrix} f_{1,t} \\ f_{2,t} \end{pmatrix} = \begin{pmatrix} w_{1,t} \\ w_{2,t} \end{pmatrix} \quad (12)$$

Once all the parameters of the seasonal VARIMA model are significant, we present the

loading matrix obtained in Figure 6. There is a clear relationship between loads and boxplot of hourly prices as shown in Figure 7.

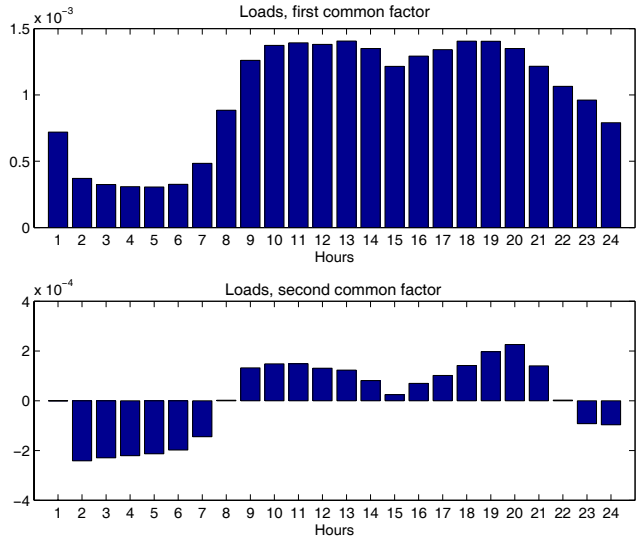

Figure 6: Loads, unobserved common factors.

The common part of each hourly time series of prices are obtained by multiplying loads by unobserved common factors obtained. The first common factor gives positive loads to every hourly time series, larger in those hours in which both the level and variance of the prices are higher. The second one separates between night and day hours.

According to Figure 7 we have selected hours 4, 9, 12 and 21 because of their representativeness and we provide in Figure 8 the centered log-prices in these hours in the period September-December 2003 (the final period considered to estimate SeaDFA), as well as the part explained by the common factors. The difference between each series and the common part explained by the unobserved common factors is the specific component.

5.2 Point forecasts and prediction intervals

We will now provide the results obtained when calculating forecasts for electricity prices in 2004, using the data from 1^{st} January 1998 through 31^{st} December 2003. Thus, the forecasting horizon is varying from 1 day up to 1 year, since the last data we used corresponds to the last

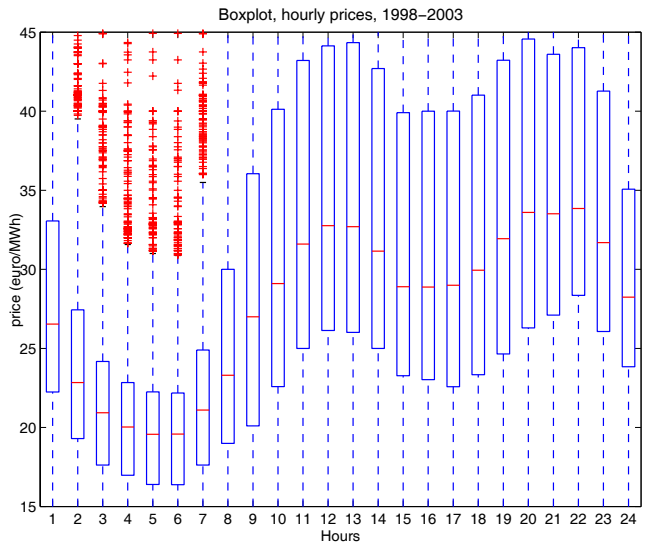

Figure 7: Boxplot of hourly prices (1998-2003).

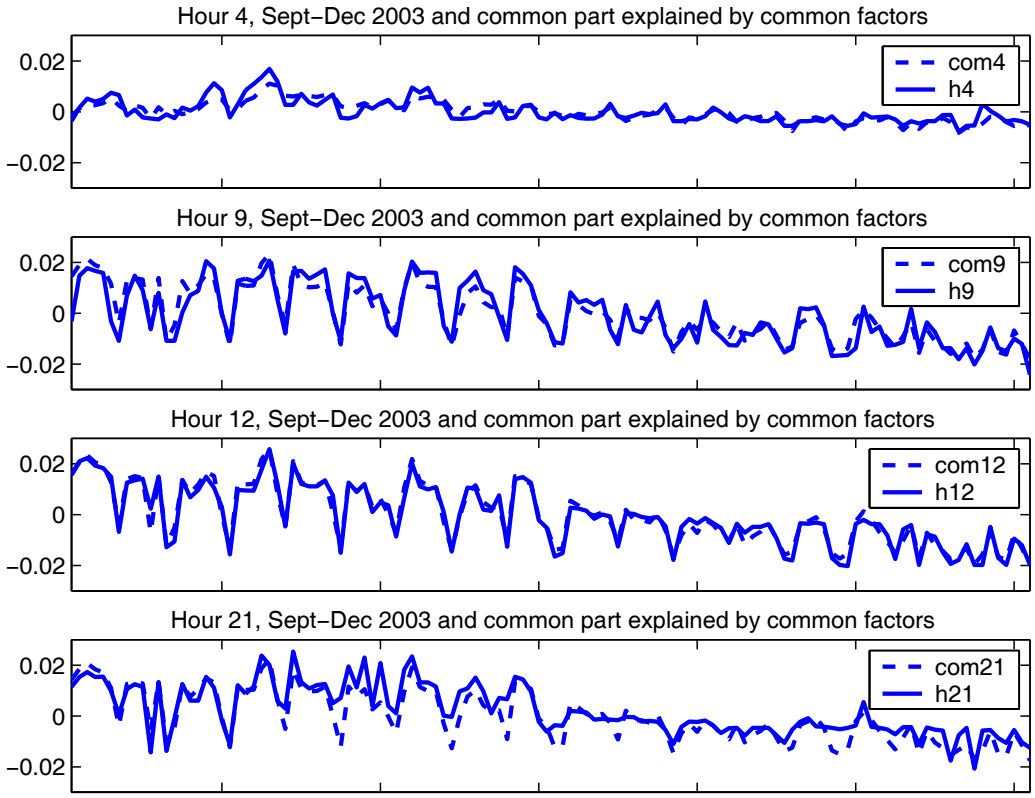

Figure 8: Some hourly time series and the common part explained by common factors.

day in 2003, no matter for which day of 2004 we are computing the forecast.

Given the complexity of the problem we have decided to follow the idea proposed by Ortega and Poncela (2002), where autoregressive processes were fitted for the specific factors. However, since the main objective of this work is long-term forecasting, the effect of specific factors on long-term forecasts it is not relevant.

The accuracy metrics we use (Mean Average Percentage Error, and MAPE2, see definitions below), have been selected because they have been used in the literature for evaluating the forecasting performance of the models developed for forecasting in electricity markets, (Conejo et al. (2005)). Let $p_{t,d}$ be the price in day d in the H^{th} hour, and $\widehat{p}_{t,d}$ its forecast computed, then the error measurement $e_{t,d}$ (for each hour of each day) is defined as $e_{t,d} = |p_{t,d} - \widehat{p}_{t,d}| / p_{t,d}$. Using $e_{t,d}$, the subsequent accuracy metrics can be defined for each day:

$$emedian_d = median(e_{1,d}, e_{2,d}, ...e_{24,d}) \tag{13}$$

$$emean_d = \frac{1}{24} \sum_{t=1}^{24} e_{t,d} \tag{14}$$

And using expression (13) and (14) MAPE and MAPE2 are obtained for a D-day period.

$$\begin{aligned} \text{MAPE} &= \frac{1}{D} \sum_{d=1}^{D} emean_d \\ \text{MAPE2} &= \frac{1}{D} \sum_{d=1}^{D} emedian_d \end{aligned}$$

Since there is no published work on long-run forecasting of electricity prices we will relate our results with short-term ones. We will also compare our results to those obtained with some methods that has been specifically developed for short-term forecasting in the Spanish market. We will use as a benchmark model the best mixed model proposed in García-Martos et al. (2007), that uses a combination of several seasonal ARIMA models for different lengths of time series. We will illustrate the fact that the available accurate models for short-term forecasting do not work properly in the long run.

Besides, we will also check short-term forecasting performance of SeaDFA. Although SeaDFA was not developed for this purpose, we get accurate forecasts even comparing them with others obtained by methods specifically designed for the short-term.

Table 7: Monthly prediction errors. SeaDFA, DFA (Peña and Poncela (2004), Mixed Model (García-Martos et al. (2007))

	SeaDFA		DFA (2004)		Mixed model (2007)	
	MAPE (%)	MAPE2 (%)	MAPE (%)	MAPE2 (%)	MAPE (%)	MAPE2 (%)
January	25.02	24.42	23.86	22.49	31.9	29.87
February	18.61	16.2	21.93	19.84	35.72	33.34
March	23.55	22.11	24.57	22.06	46.2	46
April	20.3	19.66	34.41	30.5	39.42	38.8
May	18.96	16.71	33.68	29.1	41.82	42.22
June	19.36	17.57	29.77	33.6	45.31	46.37
July	20.6	19.08	33.6	26.28	46.52	48.21
August	14.55	13.94	31.7	23.17	45.72	48.61
September	25.39	26.19	25.48	24.82	53.36	61.27
October	18.44	17.72	25.84	23.92	51.94	61.48
November	23.26	21.98	26.21	24.52	52.92	56.67
December	30.67	29.11	27.09	26.04	56.59	60.29
Year 2004	21.56	20.39	28.17	25.52	45.62	47.76

The numerical results obtained for year 2004 are shown in Table 7.

The MAPE for the whole year is 21.56% and the MAPE2 is 20.39%. These results should be related with those obtained for the short-term, which are around 13-15%, and the forecasting horizon is 24 hours, (Contreras et al. (2003), Conejo et al. (2005), Nogales et al. (2006)). With the mixed model provided in García-Martos, Rodríguez and Sánchez (2007), the short-term prediction error obtained when computing forecasts for every hour in the period 1998-2003 is 12.61%. We will use this as benchmark model, because it is the best one in the short-term for the Spanish Market. However, when using this mixed model for long-term forecasting the MAPE for the whole year is 45.62% and MAPE2 is 47.76%. This illustrates the fact that the models developed for the short-term do not work properly when the forecasting horizon increases.

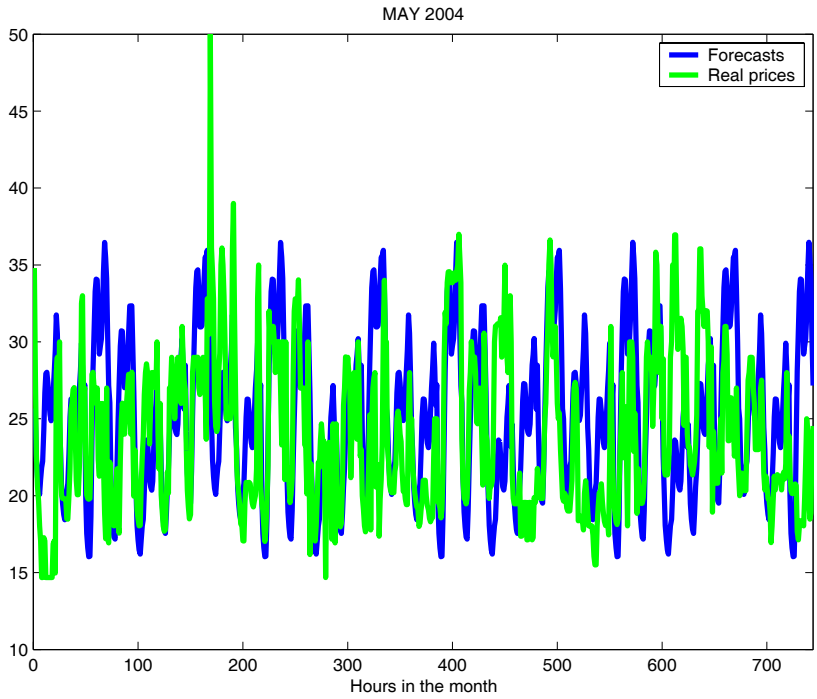

Figure 9: Forecasts and real prices, May 2004.

Non-stationary Dynamic Factor Analysis derived in Peña and Poncela (2004, 2006) could be an alternative in which only regular part of the dynamics can be modeled. Fitting a VARIMA(9,1,0), MAPE and MAPE2 are respectively 28.17% and 25.52, which both illustrates the great reduction in the error compared to the mixed model, as well as the importance of including seasonality in DFA, since SeaDFA reduces MAPE to 21.56% and MAPE2 to 20.39%.

In Figure 9 we provide results for May 2004, so forecasting horizon varies from four months and one day (1^{st} May 2004) up to five months (31^{st} May 2004). MAPE is 18.96% and MAPE2 is 16.71%. MAPE obtained when using the model described in García-Martos, Rodríguez and Sánchez (2007) is 41.82% and MAPE2 is 42.22%, and MAPE and MAPE2 obtained with DFA (Peña and Poncela (2004, 2006) are respectively 33.68% and 26.28%. Furthermore, the level of the prices has been adequately captured. This point is relevant for long run forecasting.

In Figure 10 the results for August 2004 are shown. The forecasting horizon varies from seven months and one day (1^{st} August 2004) up to eight months (31^{st} August 2004). MAPE is 14.55% and MAPE2 is 13.94%. The MAPE obtained when using the model described in

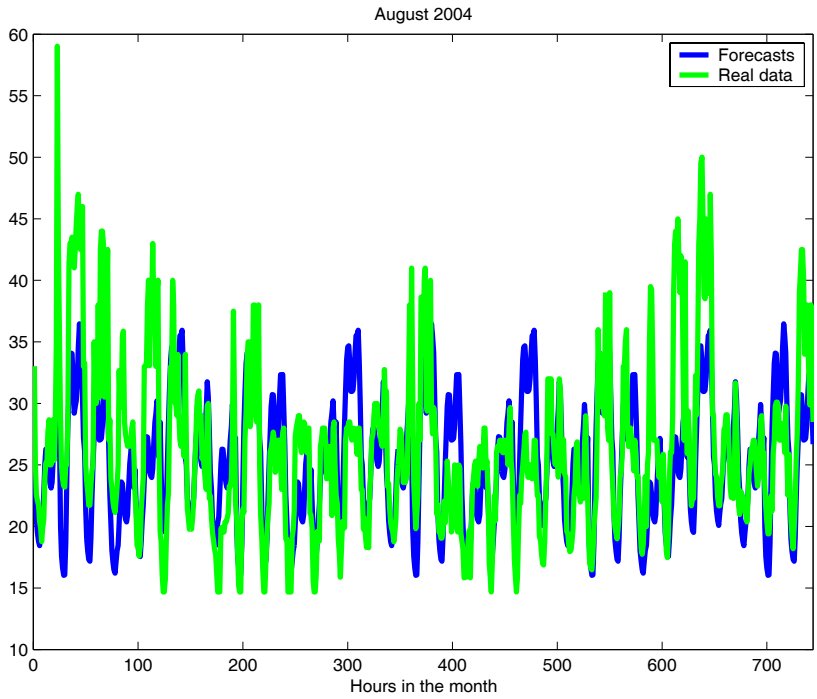

Figure 10: Forecasts and real prices 2004.

García-Martos, Rodríguez and Sánchez (2007) is 45.72% and the MAPE2 is 48.61%. MAPE and MAPE2 with DFA are 31.7% and 23.17% respectively.

Furthermore, when taking into account only the third week in each month (often the last or the one before last is used to check accuracy of forecasting models, Conejo et al. (2005) and Contreras et al. (2003)) and calculating the MAPE with SeaDFA for these twelve weeks 19.75% is obtained, which again reflects the accuracy of SeaDFA and its great performance. Considering only the third week in each month the result is slightly better than considering all weeks (MAPE 21.56%).

Figure 11 shows the results for the third week in February (16^{th}-22^{nd} February 2004). Forecasting errors are provided in Table 8. These results have been obtained using once more the SeaDFA estimated for the prices in 1998-2003, so the forecasting horizon varies from 7 weeks up to 8 weeks. The MAPE for this week is 16.38%. Using the mixed model in García-Martos the MAPE is 34.78%. Besides, and to illustrate that the SeaDFA is not only valid for medium and long term forecasting but also for the short-run, we provide in Figure 12 and Table 8

Table 8: Prediction errors, long term forecasting, 16-22 February 2004.

	Day 1	Day 2	Day 3	Day 4	Day 5	Day 6	Day 7	MWE
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%

respectively the forecasts and errors obtained for the same week in February 2004, but having estimated the model using data up to 15^{th} February, and reestimating the model six more times updating the data, which means that for each day in this week the forecasting horizon is 1 day. The daily prediction errors are shown in Table 8. See Conejo et al. (2005) or García-Martos, Rodríguez and Sánchez (2007) to check that the forecasting errors obtained by SeaDFA are of the same magnitude or even lower in comparison with those provided there.

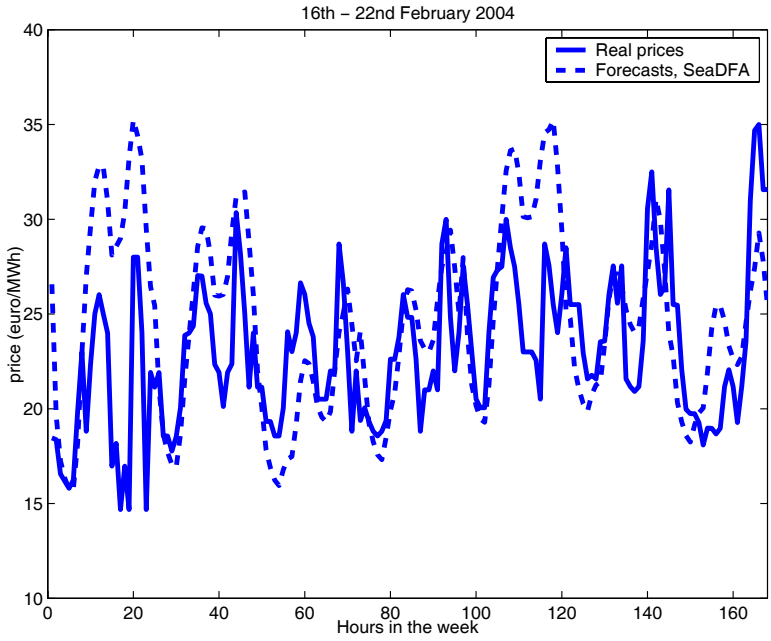

Figure 11: Forecasts and real prices, 16^{th}-22^{nd} February 2004. Last data used to estimate the model: 31^{st} December 2003.

Once numerical results for long and short run point forecasts have been provided, it is of interest to report the results obtained when forecasting intervals were computed by means of the bootstrap scheme proposed. In Figure 13 the percentile based confidence intervals for the point prediction, which include uncertainty due to parameter estimation, are provided for

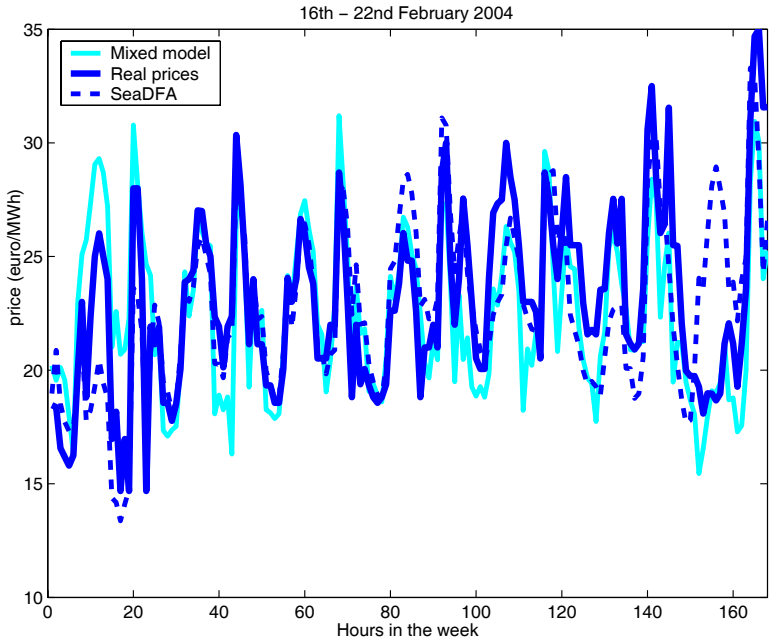

Figure 12: Forecasts and real prices, 16^{th}-22^{nd} February 2004. One-step-ahead forecasts.

Table 9: Prediction errors, one-day-ahead forecasting, 16-22 February 2004.

	Day 1	Day 2	Day 3	Day 4	Day 5	Day 6	Day 7	MWE
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%

the last complete week in May 2004, which is usually selected for performance checking in electricity price forecasting (Contreras et al. (2003) and Conejo et al. (2005)). The intervals have been computed with the SeaDFA estimated for the data in 1998-2003, so the forecasting horizon varies from 21 up to 22 weeks, which means medium or long term forecasting.

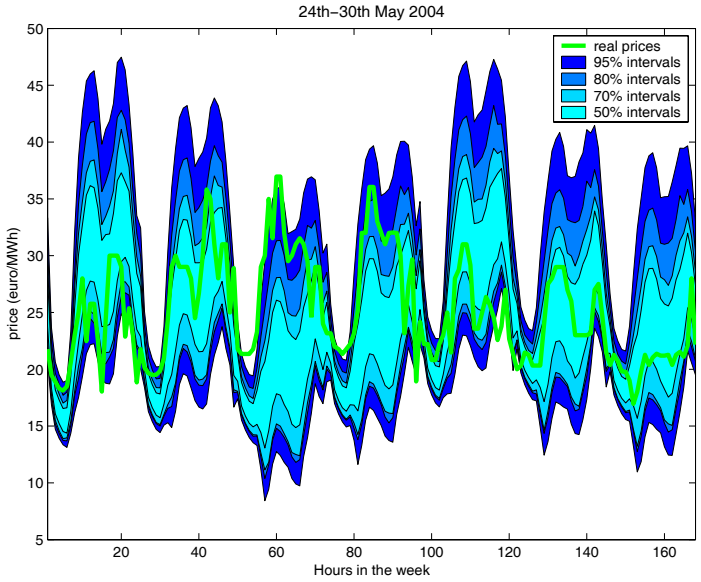

Figure 13: Percentile based confidence intervals including uncertainty due to parameter estimation. (24^{th}-30^{th} May 2004).

In Figure 14 the percentile based confidence intervals for the point prediction, which include uncertainty due to parameter estimation, are provided for the last complete of the year 2004 (20^{th}-26^{th} December 2004), which contains 24^{th} and 25^{th}, which usually corresopond to unusual days in demand and prices. The forecasting horizon is almost one year, which means long-term forecasting. This results illustrates the great coverage of the confidence intervals obtained using bootstrap techniques.

6 Conclusions

In this work we have provided the seasonal extension to Non-Stationary Dynamic Factor Analysis. Till now the only possible way to deal with dimensionality reduction in vectors of

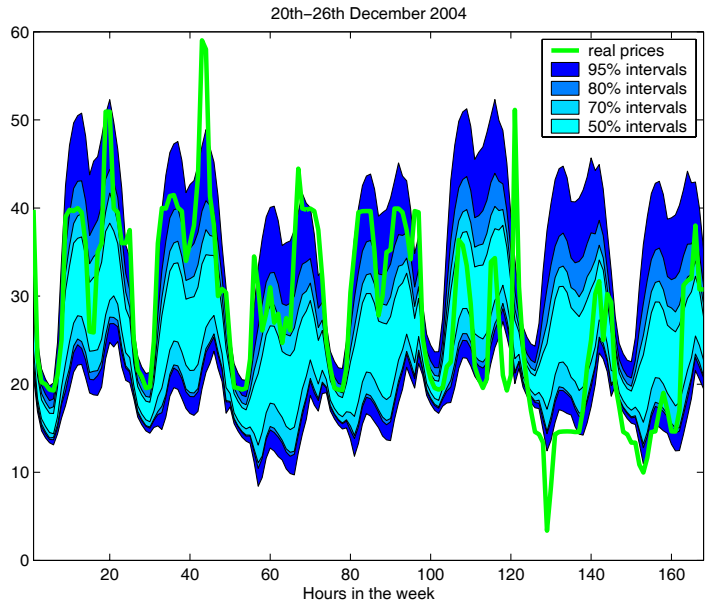

Figure 14: Percentile based confidence intervals including uncertainty due to parameter estimation. (20^{th}-26^{th} December 2004).

time series with a seasonal pattern was to deseasonalize and then apply one of the existing techniques, Peña and Box (1987), Lee and Carter (1992) or Peña and Poncela (2004, 2006), among others. However, deseasonalizing can only remove the seasonality in the mean and variance, but the seasonality in serial dependence structure remains. In order to deal with this problem we propose the Seasonal Dynamic Factor Analysis (SeaDFA), which is able to estimate common factors that follow a multiplicative seasonal VARIMA model with constant.

A modification in the estimation procedure due to seasonality and to including the constant has been presented. The constant has been included by means of an exogenous variable, which let us improve long-run forecasts in the case of non-stationary processes.

Besides, 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.

In the application to electricity price modeling we have obtained percentile based confidence intervals for each element in the loading matrix Ω, studying the long term significance of each

of the 24 hourly time series, as well as for the parameters of the VARMA model for the common factors. The numerical results provided for an interesting, difficult to forecast data set (electricity prices in the Spanish market) are accurate, with a prediction error around 20% for a horizon between one day and and one year and can be compared with short-term predictions (forecasting horizon between 24 and 72 hours).

A Appendix

The specific factors, $\boldsymbol{\varepsilon}_t$, can be expressed as follows:

$\boldsymbol{\varepsilon}_t = (\boldsymbol{\varepsilon}_t - \widehat{\boldsymbol{\varepsilon}}_t) + \widehat{\boldsymbol{\varepsilon}}_t$ and $\boldsymbol{\varepsilon}_t\boldsymbol{\varepsilon}_t' = (\boldsymbol{\varepsilon}_t - \widehat{\boldsymbol{\varepsilon}}_t)(\boldsymbol{\varepsilon}_t - \widehat{\boldsymbol{\varepsilon}}_t)' + \widehat{\boldsymbol{\varepsilon}}_t\widehat{\boldsymbol{\varepsilon}}_t'$ since cross-product is zero because the estimate from the Kalman filter is fixed and known at time t.

And substituting $\widehat{\boldsymbol{\varepsilon}}_t = \mathbf{y}_t - \widehat{\Omega}\widehat{\mathbf{f}}_t$,

gives

$$(\boldsymbol{\varepsilon}_t - \widehat{\boldsymbol{\varepsilon}}_t) = \widehat{\Omega}\widehat{\mathbf{f}}_t - \Omega\mathbf{f}_t, so \mathbf{E}[\boldsymbol{\varepsilon}_\mathbf{t}\boldsymbol{\varepsilon}_\mathbf{t}'] = \mathbf{E}[\widehat{\boldsymbol{\varepsilon}}_\mathbf{t}\widehat{\boldsymbol{\varepsilon}}_\mathbf{t}'] + \mathbf{E}[(\widehat{\Omega}\widehat{\mathbf{f}}_\mathbf{t} - \Omega\mathbf{f}_\mathbf{t})(\widehat{\Omega}\widehat{\mathbf{f}}_\mathbf{t} - \Omega\mathbf{f}_\mathbf{t})'].$$

Besides, $\widehat{\Omega}\widehat{\mathbf{f}}_t - \Omega\mathbf{f}_t = \widehat{\Omega}\widehat{\mathbf{f}}_t + \Omega\widehat{\mathbf{f}}_t - \Omega\widehat{\mathbf{f}}_t - \Omega\mathbf{f}_t = (\widehat{\Omega} - \Omega)\widehat{\mathbf{f}}_t + \Omega(\widehat{\mathbf{f}}_t - \mathbf{f}_t)$.And taking into account *projection theorem* gives:

$$(\widehat{\Omega}\widehat{\mathbf{f}}_t - \Omega\mathbf{f}_t)(\widehat{\Omega}\widehat{\mathbf{f}}_t - \Omega\mathbf{f}_t)' = (\widehat{\Omega} - \Omega)\widehat{\mathbf{f}}_t\widehat{\mathbf{f}}_t'(\widehat{\Omega} - \Omega)' + \Omega(\widehat{\mathbf{f}}_\mathbf{t} - \mathbf{f}_\mathbf{t})(\widehat{\mathbf{f}}_\mathbf{t} - \mathbf{f}_\mathbf{t})\Omega'.$$

Finally, the expression for $var(\boldsymbol{\varepsilon}_t - \widehat{\boldsymbol{\varepsilon}}_t)$:

$$E[(\widehat{\Omega}\widehat{\mathbf{f}}_t - \Omega\mathbf{f}_t)(\widehat{\Omega}\widehat{\mathbf{f}}_t - \Omega\mathbf{f}_t)'] = E[(\widehat{\Omega} - \Omega)\widehat{\mathbf{f}}_t\widehat{\mathbf{f}}_t'(\widehat{\Omega} - \Omega)'] + E[\Omega(\widehat{\mathbf{f}}_\mathbf{t} - \mathbf{f}_\mathbf{t})(\widehat{\mathbf{f}}_\mathbf{t} - \mathbf{f}_\mathbf{t})\Omega'] = \mathbf{S}_{correction}.$$

When $T \to \infty$, $\mathbf{S}_{correction} = \Omega E[(\widehat{\mathbf{f}}_\mathbf{t} - \mathbf{f}_\mathbf{t})(\widehat{\mathbf{f}}_\mathbf{t} - \mathbf{f}_\mathbf{t})]\Omega'$, given the consistency of $\widehat{\Omega}$. For details on $(\widehat{\mathbf{f}}_\mathbf{t} - \mathbf{f}_\mathbf{t})$ expectancy and covariance, see Spall and Wall (1984).

B Acknowledgements

The authors would like to thank José Mira for his help and comments.

This work was supported by Project MTM2005-08897, Ministerio de Educación y Ciencia, Spain. Besides, Carolina García-Martos would like to thank International Institute of Forecasters for the International Symposium on Forecasting 2007 Travel Award, and Andrés M. Alonso thanks financial support by Projects SEJ2007-64500 and SEJ2005-06454, Ministerio de Educación y Ciencia, Spain.

References

[1] Alonso, A.M., Peña, D., Romo, J. (2002), "Forecasting time series with sieve bootstrap," *Journal of Statistical Planning and Inference,* 100, 1, 1-11.

[2] 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.

[3] Ansley, C.F. and Newbold, P. (1980), "Finite sample properties of estimators for Autorregressive Moving average properties," *The Journal of Econometrics,* 13,159-183.

[4] Anderson, B.D.O. and Moore, J.B., (1979), *Optimal Filtering,* Englewood Cliffs, Prentice-Hall, New Jersey.

[5] Coleman, T.F. and Y. Li, (1994), "On the Convergence of Reflective Newton Methods for Large-Scale Nonlinear Minimization Subject to Bounds," *Mathematical Programming*, 67, 2, 189-224.

[6] Coleman, T.F. and Y. Li, (1996), "An Interior, Trust Region Approach for Nonlinear Minimization Subject to Bounds," *SIAM Journal on Optimization*, 6, 418-445.

[7] Conejo A. J., Contreras J., Espínola R., Plazas M. A. (2005), "Forecasting electricity prices for a day-ahead pool-based electric energy market,". *International Journal of Forecasting,* 21, 3, 435-462.

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

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

[10] Crespo-Cuaresma J., Hlouskova, Kossmeier S., Obersteiner M., (2004), "Forecasting electricity spot-prices using linear univariate time-series models". *Applied Energy*, 77, 1 , 87-106.

[11] Durbin, J. and Koopman, S.J. (2001), *Time Series Analysis by State Space Methods*, Oxford University Press.

[12] 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.

[13] 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, 1404-10.

[14] Geweke, J. (1977), "The dynamic factor analysis of economic time series," *Latent variables in socio-economic models*, Amsterdam, North Holland.

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

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

[17] 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.

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

[19] Ljung, L. and Caines, P.E. (1979), "Asymptotic normality of prediction error estimators for approximate system models," *Stochastics*, 3, 29-46.

[20] Nogales F. J., Conejo A. J, (2006), "Electricity Price Forecasting through Transfer Function Models", *Journal of the Operational Research Society*, 57, 350-356.

[21] 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.

[22] 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.

[23] Peña, D. and Poncela, P. (2004), "Forecasting with Nonstationary Dynamic Factor Models," *The Journal of Econometrics*, 119, 291-321.

[24] Peña, D. and Poncela, P. (2006), "Nonstationary Dynamic Factor Analysis," *Journal of Statistical Planning and Inference*, 136, 1237-1257.

[25] Sánchez, I. (2006), "Recursive estimation of Dynamic Models using Cook's distance, with application to wind energy forecast," *Technometrics, ,* 48, 61-73.

[26] Sargent, T. J. and C. A. Sims (1977), *Business cycle modeling without pretending to have too much a priori economic theory*. In C. A. S. et al. (Ed.), New methods in business cycle research. Minneapolis: Federal Reserve Bank of Minneapolis.

[27] 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, 1 347-355.

[28] 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.

[29] Shumway, R.H. and Stoffer, D.S. (2006), *Time Series Analysis and Its Applications With R Examples*, Springer Texts in Statistics.

[30] 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.

[31] Stock, J. H. and M. Watson (2002), "Forecasting using principal components from a large number of predictors," *Journal of the American Statistical Association,* 97, 1167–79.

[32] 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.

[33] Thombs, L.A. and Schucany, W.R., (1990), "Bootstrap prediction intervals for autoregression," *Journal of the American Statistical Association,* 85, 486-492.

[34] Troncoso A., Riquelme J., Riquelme J., Gómez A., Martínez J.L. (2002), "A Comparison of Two Techniques for Next-Day Electricity Price Forecasting". *Lecture Notes In Computer Science*, 2453, Proceedings of the 13th International Conference on Database and Expert Systems Applications.

[35] 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.

[36] Wu, L.S., Pai, J., Hosking, J.R.M. (1996), "An algorithm for estimating parameters of state-space models". *Statistics and Probability Letters*, 28, 2, 99-106.