Exploratory Data Modeling with LS-SVMs: An Application in Short Term Load Forecasting

Marcelo Espinoza *, Johan A.K. Suykens, Bart De Moor

ESAT-SCD-SISTA, Katholieke Universiteit Leuven, Kasteelpark Arenberg 10, B-3001 Leuven (Heverlee), Belgium

Abstract

Based on the primal-dual formulation of Least Squares Support Vector Machines (LS-SVMs), an exploratory modeling methodology is applied to a large scale real-life case study. ARX and NARX model structures are considered, which are parameterized either as a full black-box model or as partially linear structures, with or without the addition of autocorrelated residuals. For the case study of short-term load forecasting, it is shown that a structured model can reach better results than a full black-box model, obtaining satisfactory results when the model structure and the hyperparameter selection procedure are tailored towards the application, taking into account prior knowledge or relevant features of the problem.

Key words: Least-Squares Support Vector Machines, Fixed-Size LS-SVM, Primal Space Regression, Load Forecasting, Time Series.

1 Introduction

This paper addresses the problem of selecting a suitable model structure by using prior knowledge and physical insight taken from the problem under analysis. In real-life applications, the designer can make use of different model specifications, from nonlinear black-box models to linear parametric models [1]. Following the well known golden rule of "do not estimate what you already know", it is important to include as much information as possible about the

* Corresponding Author.
Email addresses: marcelo.espinoza@esat.kuleuven.be (Marcelo Espinoza), johan.suykens@esat.kuleuven.be (Johan A.K. Suykens).

problem into the modeling stage. This prior information may relate to several levels: the available data, the model structure, model selection criteria, in such a way that the final model obtains a good performance in terms of future forecasts because it has been"tailored" to the application at hand.

As a real-life application, we consider the problem of Short-Term Load Forecasting (STLF), which is an important area of quantitative research where the goal is to build a model that can be used for prediction of the next hours. Usually a load series shows important seasonal patterns (yearly, weekly, intra-daily patterns) that need to be taken into account in the modeling strategy [2]. This situation translates into autocorrelation patterns across the corresponding hour on previous days [3]. On the other hand, good forecasts are required particularly for the daily peak load. The accuracy of the predicted daily peak load has to be as good as possible, in order to minimize the risk of a grid congestion. Typically, a linear parametric model may be preferred over a full nonlinear black-box, due to the insights that can be derived from the identified coefficients [3] in terms of temperature sensitivity, growth trends, for management and planning purposes. By considering all these aspects as prior knowledge, in this paper we illustrate an exploratory modeling methodology for model structure selection within the STLF framework, using estimation techniques based on Least Squares Support Vector Machines (LS-SVM [4]). We demostrate that, starting from the same input variables, a tailored partially linear [5] model structure with correlated residuals [6] can obtain better results than a full generic black-box model structure. Within the LS-SVM framework, different model structures and parameterizations can be estimated as a solution of a linear system obtained from a convex problem.

This paper is organized as follows. The description of the different model structures to be used in the paper is presented on Section 2. The estimation techniques based on LS-SVM are described on Section 3, where the model structures are parameterized in several ways. On Section 4, the implementation using Fixed-Size LS-SVMs [4] is described, with a description of the estimation methodology for each one of the model structures and parameterizations. The application to the STLF case study and the estimation results are presented on Section 5.

2 Model Formulation

In this paper we consider 2 model structures: a (N)ARX model, which is a (Nonlinear) AutoRegression with eXogenous inputs, and a AR-(N)ARX model, which is a (Nonlinear) AutoRegression with eXogenous inputs and AutoRegressive residuals. Each one of these model structures can be formulated in terms of a linear parameterization, a full black-box, or a partially

linear parameterization.

2.1 Model Structures

Consider the regression vector $\boldsymbol{z}_t \in \mathbb{R}^n$ of a (N)ARX model, $\boldsymbol{z}_t = [y_{t-1}; \cdots; y_{t-p}; \boldsymbol{u}_t; \boldsymbol{u}_{t-1}; \cdots; \boldsymbol{u}_{t-q}]$ containing p past values of the output $y_t \in \mathbb{R}$ and q past values of input vectors $\boldsymbol{u}_t \in \mathbb{R}^{N_u}$. A (N)ARX model structure [1,7,8] can be formulated as

$$y_t = g(\boldsymbol{z}_t) + e_t \tag{1}$$

where the error term e_t is assumed to be i.i.d with zero mean and constant variance σ_e^2. The AR-(N)ARX [1,9,6] model structure incorporates a covariance-stationary process on the error terms e_t,

$$\begin{cases} y_t = g(\boldsymbol{z}_t) + e_t \\ e_t = \rho e_{t-\tau} + r_t \end{cases} \tag{2}$$

where $\tau \geq 1$, r_t is a white noise sequence with zero mean and constant variance σ_r^2 and $|\rho| < 1$. When $\tau = 1$, the noise process is AR(1) $e_t = \rho e_{t-1} + r_t$, which is often use in applied modeling work [10].

2.2 Model Parameterizations

For the parameterization of the function $g(\cdot)$ from (1) or (2) the following alternatives are considered.

2.2.1 Linear Parameterization

The function $g(\cdot)$ can be parameterized as

$$g(\boldsymbol{z}_t) = \boldsymbol{\theta}^T \boldsymbol{z}_t \tag{3}$$

with $\boldsymbol{\theta} \in \mathbb{R}^{N_d}$. In this case the function g is linear, and the corresponding models structures from (1) and (2) are an ARX and AR-ARX, respectively.

2.2.2 Black-Box Parameterization

The nonlinear function $g(\cdot)$ for a NARX (1) or AR-NARX (2) structure is parameterized under a black-box formulation in primal space using LS-SVMs

[4]:

$$g(\boldsymbol{z}_t) = \boldsymbol{w}^T \boldsymbol{\varphi}(\boldsymbol{z}_t) + b \tag{4}$$

where b is a constant (bias) term, and $\boldsymbol{\varphi}(\cdot) : \mathbb{R}^n \to \mathbb{R}^{N_h}$ is the feature map from the input space to the so-called feature space (of dimension N_h which can be possibly infinite). In the area of support vector machines and kernel-based learning this feature map is used in relation to a Mercer kernel [4,11], in such a way that the feature map is not computed explicitly.

2.2.3 Partially Linear Parameterization

In this case, some of the regressors are included as linear terms, and others are included under a nonlinear black-box term. Consider a partition of the regression vector $\boldsymbol{z}_t$ as follows. Let $\mathcal{Z} = \{x : x \text{ is a component of the vector } \boldsymbol{z}_t\}$. Define an arbitrary partition $\mathcal{Z} = \mathcal{Z}_A \cup \mathcal{Z}_B$ with $\mathcal{Z}_A \cap \mathcal{Z}_B = \emptyset$. Define a vector $\boldsymbol{z}_{A,t} \in \mathbb{R}^{N_a}$ with regressors $\boldsymbol{x} \in \mathcal{Z}_A$, and a vector $\boldsymbol{z}_{B,t} \in \mathbb{R}^{N_b}$ with regressors $\boldsymbol{x} \in \mathcal{Z}_B$. The original regression vector is thus partitioned as $\boldsymbol{z}_t = [\boldsymbol{z}_{A,t}; \boldsymbol{z}_{B,t}]$ The subscript A (resp. B) represents the subset of regressors that enters linearly (resp. nonlinearly) into the model. The nonlinear component of this Partially Linear (PL) [5] parameterization is, in turn, expressed as a black-box formulation using LS-SVMs. Thus, the nonlinear function $g(\cdot)$ for a PL-NARX (1) or a PL-AR-NARX (2) is parameterized as

$$g(\boldsymbol{z}_t) = \boldsymbol{\beta}^T \boldsymbol{z}_{A,t} + \boldsymbol{w}^T \boldsymbol{\varphi}(\boldsymbol{z}_{B,t}) + b \tag{5}$$

for a given partition $\boldsymbol{z}_t = [\boldsymbol{z}_{A,t}; \boldsymbol{z}_{B,t}]$.

3 Model Estimation

The different nonlinear model structures can be estimated using the LS-SVM regression framework. Starting from a given dataset $\{\boldsymbol{z}_i, y_i\}_{i=1}^N$, the different estimation problems are presented for each of the model structures outlined in the previous section.

3.1 NARX Model: Black-box approach

For the NARX model (1), with $g(\cdot)$ parameterized as in (4), the following optimization problem with a regularized cost function is formulated:

$$\min_{\boldsymbol{w},b,e_i} \quad \frac{1}{2}\boldsymbol{w}^T\boldsymbol{w} + \gamma\frac{1}{2}\sum_{i=1}^{N} e_i^2 \tag{6}$$

$$\text{s.t.} \quad y_i = \boldsymbol{w}^T \boldsymbol{\varphi}(\boldsymbol{z}_i) + b + e_i, \quad i = 1, \ldots, N.$$

where γ is a regularization constant. The solution is formalized in the following lemma.

Lemma 1 [4] *Given a positive definite kernel function $K : \mathbb{R}^n \times \mathbb{R}^n \to \mathbb{R}$, with $K(\boldsymbol{z}_i, \boldsymbol{z}_j) = \boldsymbol{\varphi}(\boldsymbol{z}_i)^T \boldsymbol{\varphi}(\boldsymbol{z}_j)$, the solution to (6) is given by the dual problem*

$$\left[\begin{array}{c|c} 0 & \mathbf{1}^T \\ \hline \mathbf{1} & \boldsymbol{\Omega} + \gamma^{-1}\mathbf{I} \end{array}\right] \left[\begin{array}{c} b \\ \hline \boldsymbol{\alpha} \end{array}\right] = \left[\begin{array}{c} 0 \\ \hline \boldsymbol{y} \end{array}\right], \tag{7}$$

with $\boldsymbol{y} = [y_1, \ldots, y_N]^T, \boldsymbol{\alpha} = [\alpha_1, \ldots, \alpha_N]^T$, and $\boldsymbol{\Omega}$ is the kernel matrix with $\boldsymbol{\Omega}_{ij} = K(\boldsymbol{z}_i, \boldsymbol{z}_j)$, $\forall i, j = 1 \ldots, N$.

The model representation in dual form is given by

$$\hat{y}_t = \sum_{i=1}^{N} \alpha_i K(\boldsymbol{z}_i, \boldsymbol{z}_t) + b. \tag{8}$$

Remark 1 [Kernel functions.] *For a positive definite kernel function K some common choices are: $K(\boldsymbol{x}_i, \boldsymbol{x}_j) = \boldsymbol{x}_j^T \boldsymbol{x}_j$ (linear kernel); $K(\boldsymbol{x}_i, \boldsymbol{x}_j) = (\boldsymbol{x}_i^T \boldsymbol{x}_j + c)^d$ (polynomial of degree d, with $c > 0$ a tuning parameter); $K(\boldsymbol{x}_i, \boldsymbol{x}_j) = \exp(-||\boldsymbol{x}_j - \boldsymbol{x}_j||_2^2/\sigma^2)$ (RBF kernel), where σ is a tuning parameter.*

3.2 AR-NARX Model: Incorporating a noise model

Consider the AR-NARX model (2), with $g(\cdot)$ parameterized as in (4). With the inclusion of a noise model, the following regularized optimization problem is formulated:

$$\begin{aligned} \min_{\boldsymbol{w}, b, r_i} \quad & \frac{1}{2}\boldsymbol{w}^T\boldsymbol{w} + \gamma\frac{1}{2}\sum_{i=\tau+1}^{N} r_i^2 \\ & \text{s.t.} \begin{cases} y_i = \boldsymbol{w}^T \boldsymbol{\varphi}(\boldsymbol{z}_i) + b + e_i, \quad i = \tau+1, \ldots, N. \\ e_i = \rho e_{i-\tau} + r_i, \quad i = \tau+1, \ldots, N. \end{cases} \end{aligned} \tag{9}$$

where γ is a regularization constant and the noise model coefficient ρ is a tuning parameter satisfying $|\rho| < 1$ (invertibility condition of the process). At this stage, the parameter ρ is considered to be given. By eliminating e_i, the

following problem is obtained:

$$\begin{aligned}\min_{\boldsymbol{w},b,r_i} \quad & \frac{1}{2}\boldsymbol{w}^T\boldsymbol{w} + \gamma\frac{1}{2}\sum_{i=\tau+1}^{N} r_i^2 \\ \text{s.t.} \quad & y_i = \rho y_{i-\tau} + \boldsymbol{w}^T\boldsymbol{\varphi}(\boldsymbol{z}_i) - \rho\boldsymbol{w}^T\boldsymbol{\varphi}(\boldsymbol{z}_{i-\tau}) \\ & \qquad + b(1-\rho) + r_i, \quad i = \tau+1,\ldots,N. \end{aligned} \tag{10}$$

The solution is formalized in the following lemma.

Lemma 2 [6] *Given a positive definite kernel function $K : \mathbb{R}^n \times \mathbb{R}^n \to \mathbb{R}$, with $K(\boldsymbol{z}_i, \boldsymbol{z}_j) = \boldsymbol{\varphi}(\boldsymbol{z}_i)^T\boldsymbol{\varphi}(\boldsymbol{z}_j)$, the solution to (10) is given by the dual problem*

$$\left[\begin{array}{c|c} 0 & \mathbf{1}^T \\ \hline \mathbf{1} & \boldsymbol{\Omega}^{(\rho)} + \gamma^{-1}\mathbf{I} \end{array}\right]\left[\begin{array}{c} b \\ \hline \boldsymbol{\alpha} \end{array}\right] = \left[\begin{array}{c} 0 \\ \hline \tilde{\boldsymbol{y}} \end{array}\right], \tag{11}$$

with $\tilde{\boldsymbol{y}} = [y_{1+\tau} - \rho y_1, \ldots, y_{N+\tau} - \rho y_{N-1}]^T$, $\boldsymbol{\alpha} = [\alpha_1, \ldots, \alpha_{N-\tau}]^T$, and $\boldsymbol{\Omega}^{(\rho)}$ is the kernel matrix with entries $\boldsymbol{\Omega}_{ij}^{(\rho)} = K(\boldsymbol{z}_{i+\tau}, \boldsymbol{z}_{j+\tau}) - \rho K(\boldsymbol{z}_i, \boldsymbol{z}_{j+\tau}) - \rho K(\boldsymbol{z}_{i+\tau}, \boldsymbol{z}_j) + \rho^2 K(\boldsymbol{z}_i, \boldsymbol{z}_j)$, $\forall i, j = 1\ldots, N-\tau$.

The model can be represented in dual form as

$$\hat{y}_t = \rho y_{t-\tau} + h(\boldsymbol{z}_t) - \rho h(\boldsymbol{z}_{t-\tau}), \tag{12}$$

where $h(\boldsymbol{z}_t)$ is given by

$$h(\boldsymbol{z}_t) = \sum_{i=\tau+1}^{N} \alpha_{i-\tau}\left[K(\boldsymbol{z}_i, \boldsymbol{z}_t) - \rho K(\boldsymbol{z}_{i-\tau}, \boldsymbol{z}_t)\right] + b. \tag{13}$$

Remark 2 [Considering ρ as a tuning parameter.] *For the estimation of the AR-NARX and PL-AR-NARX models, we have considered the parameter ρ as a tuning parameter in order to work with a feasible convex optimization problem in which the Mercer's Theorem can be applied and a unique solution can be obtained. The parameter ρ, therefore, is determined on another level (e.g. via cross-validation, together with the kernel parameters and the regularization term) to yield a good generalization performance of the model. In this way, the selected ρ will be the value that gives the best cross-validation performance.*

3.3 PL-NARX Model: Considering a Partially Linear Structure

For the PL-NARX model (2), with $g(\cdot)$ parameterized as in (5), the following optimization problem with a regularized cost function is formulated:

$$\min_{\boldsymbol{w},b,e_i,\boldsymbol{\beta}} \quad \frac{1}{2}\boldsymbol{w}^T\boldsymbol{w} + \gamma\frac{1}{2}\sum_{i=1}^{N} e_i^2 \tag{14}$$

$$\text{s.t.} \quad y_i = \boldsymbol{\beta}^T \boldsymbol{z}_{A,i} + \boldsymbol{w}^T \boldsymbol{\varphi}(\boldsymbol{z}_{B,i}) + b + e_i, \quad i = 1, \ldots, N,$$

where γ is a regularization constant. The solution is given by the following lemma.

Lemma 3 [5] *Given the problem (14) and a positive definite kernel function $K : \mathbb{R}^{N_b} \times \mathbb{R}^{N_b} \to \mathbb{R}$, the solution to (14) is given by the dual problem*

$$\left[\begin{array}{c|c|c} \mathbf{0}_{N_a \times N_a} & \mathbf{0}_{N_a \times 1} & \boldsymbol{Z}^T \\ \hline \mathbf{0}_{1 \times N_a} & 0 & \mathbf{1}^T \\ \hline \boldsymbol{Z} & \mathbf{1} & \boldsymbol{\Omega} + \gamma^{-1}\mathbf{I} \end{array}\right] \left[\begin{array}{c} \boldsymbol{\beta} \\ \hline b \\ \hline \boldsymbol{\alpha} \end{array}\right] = \left[\begin{array}{c} \mathbf{0}_{N_a \times 1} \\ \hline 0 \\ \hline \boldsymbol{y} \end{array}\right], \tag{15}$$

where $\boldsymbol{Z} = [\boldsymbol{z}_{A,1}^T; \boldsymbol{z}_{A,2}^T; \cdots; \boldsymbol{z}_{A,N}^T] \in \mathbb{R}^{N \times N_a}$ is the matrix of linear regressors; $\boldsymbol{y} = [y_1; y_2; \ldots; y_N]$, and $\boldsymbol{\Omega}$ is the kernel matrix with $\boldsymbol{\Omega}_{ij} = K(\boldsymbol{z}_{B,i}, \boldsymbol{z}_{B,j}) = \boldsymbol{\varphi}(\boldsymbol{z}_{B,i})^T \boldsymbol{\varphi}(\boldsymbol{z}_{B,j})$, $\forall i, j = 1, \ldots, N$.

The estimated model in dual space becomes

$$\hat{y}_t = \boldsymbol{\beta}^T \boldsymbol{z}_{A,t} + \sum_{i=1}^{N} \alpha_i K(\boldsymbol{z}_{B,i}, \boldsymbol{z}_{B,t}). \tag{16}$$

3.4 PL-AR-NARX Model: Combining it all

Consider now the PL-AR-NARX model, with $g(\cdot)$ parameterized as in (5). With the inclusion of a noise model, the following regularized optimization problem is formulated:

$$\begin{aligned} \min_{\boldsymbol{w},b,r_i} \quad & \frac{1}{2}\boldsymbol{w}^T\boldsymbol{w} + \gamma\frac{1}{2}\sum_{i=\tau+1}^{N} r_i^2 \\ \text{s.t.} \quad & \begin{cases} y_i = \boldsymbol{\beta}^T \boldsymbol{z}_{A,t} + \boldsymbol{w}^T \boldsymbol{\varphi}(\boldsymbol{z}_i) + b + e_i, & i = \tau+1, \ldots, N. \\ e_i = \rho e_{i-\tau} + r_i, & i = \tau+1, \ldots, N. \end{cases} \end{aligned} \tag{17}$$

where γ is a regularization constant and the noise model coefficient ρ is a given tuning parameter satisfying $|\rho| < 1$. By eliminating e_i, the following problem is formulated:

$$\begin{aligned} \min_{\boldsymbol{w},b,r_i} \quad & \frac{1}{2}\boldsymbol{w}^T\boldsymbol{w} + \gamma\frac{1}{2}\sum_{i=\tau+1}^{N} r_i^2 \\ & \text{s.t.} \quad y_i = \rho y_{i-\tau} + \boldsymbol{w}^T \boldsymbol{\varphi}(\boldsymbol{z}_{B,i}) - \rho \boldsymbol{w}^T \boldsymbol{\varphi}(\boldsymbol{z}_{B,i-\tau}) \\ & \quad + b(1-\rho) + \boldsymbol{\beta}^T(\boldsymbol{z}_{A,i} - \rho \boldsymbol{z}_{A,i-\tau}) + r_i, \quad i = \tau+1, \ldots, N. \end{aligned} \tag{18}$$

The solution is formalized in the following lemma.

Lemma 4 *Given the problem (18) and a positive definite kernel function* $K : \mathbb{R}^{N_b} \times \mathbb{R}^{N_b} \to \mathbb{R}$*, the solution to (18) is given by the dual problem*

$$\left[\begin{array}{c|c|c} \mathbf{0}_{N_a\times N_a} & \mathbf{0}_{N_a\times 1} & \tilde{\boldsymbol{Z}}^T \\ \hline \mathbf{0}_{1\times N_a} & 0 & \mathbf{1}^T \\ \hline \tilde{\boldsymbol{Z}} & \mathbf{1} & \boldsymbol{\Omega}^{(\rho)} + \gamma^{-1}\mathbf{I} \end{array}\right] \left[\begin{array}{c} \boldsymbol{\beta} \\ \hline b \\ \hline \boldsymbol{\alpha} \end{array}\right] = \left[\begin{array}{c} \mathbf{0}_{N_a\times 1} \\ \hline 0 \\ \hline \tilde{\boldsymbol{y}} \end{array}\right], \tag{19}$$

where $\tilde{\boldsymbol{Z}} = [\boldsymbol{z}_{A,1+\tau}^T - \rho\boldsymbol{z}_{A,1}^T; \cdots; \boldsymbol{z}_{A,N}^T - \rho\boldsymbol{z}_{A,N-\tau}] \in \mathbb{R}^{(N-\tau)\times N_a}$ *is the matrix of linear regressors;* $\tilde{\boldsymbol{y}} = [y_{1+\tau} - \rho y_1, \ldots, y_N - \rho y_{N-\tau}]^T, \boldsymbol{\alpha} = [\alpha_1, \ldots, \alpha_{N-\tau}]^T$*, and* $\boldsymbol{\Omega}^{(\rho)}$ *is the kernel matrix with entries* $\boldsymbol{\Omega}_{ij}^{(\rho)} = K(\boldsymbol{z}_{B,i+\tau}, \boldsymbol{z}_{B,j+\tau}) - \rho K(\boldsymbol{z}_{B,i}, \boldsymbol{z}_{B,j+\tau}) - \rho K(\boldsymbol{z}_{B,i+\tau}, \boldsymbol{z}_{B,j}) + \rho^2 K(\boldsymbol{z}_{B,i}, \boldsymbol{z}_{B,j})$*,* $\forall i, j = 1 \ldots, N - \tau$*.*

Proof. Consider the Lagrangian of problem (18)

$$\begin{aligned}\mathcal{L}(\boldsymbol{w}, b, e, \boldsymbol{\beta}; \alpha) = \frac{1}{2}\boldsymbol{w}^T\boldsymbol{w} + \gamma\frac{1}{2}\sum_{i=\tau+1}^{N} r_i^2 - \sum_{k=\tau+1}^{N} \alpha_{k-\tau}[\boldsymbol{\beta}^T(\boldsymbol{z}_{A,i} - \rho\boldsymbol{z}_{A,i-\tau}) \\ + \boldsymbol{w}^T\boldsymbol{\varphi}(\boldsymbol{z}_{B,i}) - \rho\boldsymbol{w}^T\boldsymbol{\varphi}(\boldsymbol{z}_{B,i-\tau}) + \rho y_{i-\tau} - y_i - r_i],\end{aligned}$$

where $\alpha_j \in \mathbb{R}, j = 1, \ldots, N - \tau$ are the Lagrange multipliers. Taking the optimality conditions $\frac{\partial\mathcal{L}}{\partial\boldsymbol{w}} = 0$, $\frac{\partial\mathcal{L}}{\partial b} = 0$, $\frac{\partial\mathcal{L}}{\partial r_i} = 0$, $\frac{\partial\mathcal{L}}{\partial\alpha_j} = 0$, $\frac{\partial\mathcal{L}}{\partial\boldsymbol{\beta}} = 0$ yields

$$\begin{aligned}\boldsymbol{w} &= \sum_{k=\tau+1}^{N} \alpha_{(k-\tau)}[\boldsymbol{\varphi}(\boldsymbol{z}_{B,i}) - \rho\boldsymbol{\varphi}(\boldsymbol{z}_{B,i-\tau})], \\ r_i &= \alpha_{i-\tau}/\gamma, \quad i = \tau+1, \ldots, N, \\ 0 &= \sum_{k=1}^{N-\tau} \alpha_k, \\ 0 &= \sum_{i=\tau+1}^{N} \alpha_{i-\tau}(\boldsymbol{z}_{A,i} - \rho\boldsymbol{z}_{A,i-\tau}), \\ y_i &= \rho y_{i-\tau} + \boldsymbol{w}^T\boldsymbol{\varphi}(\boldsymbol{z}_{B,i}) - \rho\boldsymbol{w}^T\boldsymbol{\varphi}(\boldsymbol{z}_{B,i-\tau}) \\ &\quad + b(1-\rho) + \boldsymbol{\beta}^T(\boldsymbol{z}_{A,i} - \rho\boldsymbol{z}_{A,i-\tau}) + r_i, \quad i = \tau+1, \ldots, N.\end{aligned} \tag{20}$$

With application of Mercer's theorem [11] $\boldsymbol{\varphi}(\boldsymbol{z}_{B,i})^T\boldsymbol{\varphi}(\boldsymbol{z}_{B,j}) = K(\boldsymbol{z}_{B,i}, \boldsymbol{z}_{B,j})$ with a positive definite kernel K, we can eliminate $\boldsymbol{w}$ and r_i, obtaining

$$\begin{aligned}y_i - \rho y_{i-\tau} = \sum_{k=\tau+1}^{N} \alpha_{k-\tau}[K(\boldsymbol{z}_{B,i}, \boldsymbol{z}_{B,k}) - \rho K(\boldsymbol{x}_{B,i-\tau}, \boldsymbol{x}_{B,k}) - \rho K(\boldsymbol{z}_{B,i}, \boldsymbol{z}_{B,k-\tau}) \\ + \rho^2 K(\boldsymbol{z}_{B,i-\tau}, \boldsymbol{z}_{B,k-\tau})] + b + \boldsymbol{\beta}^T(\boldsymbol{z}_{A,i} - \rho\boldsymbol{z}_{A,i-\tau}) + \frac{\alpha_{k-\tau}}{\gamma}.\end{aligned} \tag{21}$$

Building the kernel matrix $\boldsymbol{\Omega}_{ij}^{(\rho)}$ and writing the equations in matrix notation gives the final system (19) □

Table 1
Summary of Nonlinear Model Structures and Representations by LS-SVMs.

NARX Model	
Primal	$\hat{y}_t = \boldsymbol{w}^T \boldsymbol{\varphi}(\boldsymbol{z}_t) + b$
Dual	$\hat{y}_t = \sum_{i=1}^{N} \alpha_i K(\boldsymbol{z}_i, \boldsymbol{z}_t) + b$
AR-NARX Model	
Primal	$\hat{y}_t = \rho y_{t-\tau} + \boldsymbol{w}^T \boldsymbol{\varphi}(\boldsymbol{z}_t) - \rho \boldsymbol{w}^T \boldsymbol{\varphi}(\boldsymbol{z}_{t-\tau}) + (1-\rho) b$
Dual	$\hat{y}_t = \rho y_{t-\tau} + h(\boldsymbol{z}_t) - \rho h(\boldsymbol{z}_{t-\tau})$ with $h(\boldsymbol{z}_t) = \sum_{i=\tau+1}^{N} \alpha_{i-\tau} \left[K(\boldsymbol{z}_i, \boldsymbol{z}_t) - \rho K(\boldsymbol{z}_{i-\tau}, \boldsymbol{z}_t) \right] + b$
PL-NARX Model	
Primal	$\hat{y}_t = \boldsymbol{\beta}^T \boldsymbol{z}_{A,t} + \boldsymbol{w}^T \boldsymbol{\varphi}(\boldsymbol{z}_{B,t}) + b$
Dual	$\hat{y}_t = \boldsymbol{\beta}^T \boldsymbol{z}_{A,t} + \sum_{i=1}^{N} \alpha_i K(\boldsymbol{z}_{B,i}, \boldsymbol{z}_{B,t})$
PL-AR-NARX Model	
Primal	$\hat{y}_t = \rho y_{t-\tau} + \boldsymbol{w}^T \boldsymbol{\varphi}(\boldsymbol{z}_{B,t}) - \rho \boldsymbol{w}^T \boldsymbol{\varphi}(\boldsymbol{z}_{B,t-\tau}) + b(1-\rho) + \boldsymbol{\beta}^T (\boldsymbol{z}_{A,t} - \rho \boldsymbol{z}_{A,t-\tau})$
Dual	$\hat{y}_t = \rho y_{t-\tau} + h(\boldsymbol{z}_{B,t}) - \rho h(\boldsymbol{z}_{B,t-\tau}) + \boldsymbol{\beta}^T (\boldsymbol{z}_{A,i} - \rho \boldsymbol{z}_{A,i-\tau})$ with $h(\boldsymbol{z}_{B,t}) = \sum_{i=\tau+1}^{N} \alpha_{i-\tau} \left[K(\boldsymbol{z}_{B,i}, \boldsymbol{z}_{B,t}) - \rho K(\boldsymbol{z}_{B,i-\tau}, \boldsymbol{z}_{B,t}) \right] + b$

The estimated model in dual space becomes

$$\hat{y}_t = \rho y_{t-\tau} + h(\boldsymbol{z}_{B,t}) - \rho h(\boldsymbol{z}_{B,t-\tau}) + \boldsymbol{\beta}^T (\boldsymbol{z}_{A,i} - \rho \boldsymbol{z}_{A,i-\tau}), \tag{22}$$

where $h(\boldsymbol{z}_{B,t})$ is given by

$$h(\boldsymbol{z}_t) = \sum_{i=\tau+1}^{N} \alpha_{i-\tau} \left[K(\boldsymbol{z}_{B,i}, \boldsymbol{z}_{B,t}) - \rho K(\boldsymbol{z}_{B,i-\tau}, \boldsymbol{z}_{B,t}) \right] + b. \tag{23}$$

A summary of the different nonlinear model structures and representations is given in Table 1.

Table 2
Kernel function $K^{\text{eq}}_{\text{model}}$ for the different model structures.

$K^{\text{eq}}_{\text{narx}}(\boldsymbol{z}_i, \boldsymbol{z}_j) =$	$K(\boldsymbol{z}_i, \boldsymbol{z}_j)$
$K^{\text{eq}}_{\text{ar}-\text{narx}}(\boldsymbol{z}_i, \boldsymbol{z}_j) =$	$K(\boldsymbol{z}_{i+\tau}, \boldsymbol{z}_{j+\tau}) - \rho K(\boldsymbol{z}_i, \boldsymbol{z}_{j+\tau}) - \rho K(\boldsymbol{z}_{i+\tau}, \boldsymbol{z}_j) + \rho^2 K(\boldsymbol{z}_i, \boldsymbol{z}_j)$
$K^{\text{eq}}_{\text{pl}-\text{narx}}(\boldsymbol{z}_i, \boldsymbol{z}_j) =$	$K(\boldsymbol{z}_{B,i}, \boldsymbol{z}_{B,j})$
$K^{\text{eq}}_{\text{pl}-\text{ar}-\text{narx}}(\boldsymbol{z}_i, \boldsymbol{z}_j) =$	$K(\boldsymbol{z}_{B,i+\tau}, \boldsymbol{z}_{B,j+\tau}) - \rho K(B, \boldsymbol{z}_{B,i}, \boldsymbol{z}_{B,j+\tau}) - \rho K(\boldsymbol{z}_{B,i+\tau}, \boldsymbol{z}_{B,j}) + \rho^2 K(\boldsymbol{z}_{B,i}, \boldsymbol{z}_{B,j})$

4 Implementation using Fixed-Size LS-SVM

All the models of the previous section are given in terms of the dual solution, which requires solving a linear system of size $(N - \tau + N_a) \times (N - \tau + N_a)$ (for a NARX model, $\tau = 0, N_a = 0$). This system is obtained with the application of Mercer's theorem, without having to compute the nonlinear mapping $\boldsymbol{\varphi}$ explicitly. However, for large sample sizes this may become unpractical. For such a case, it is possible to obtain a finite dimensional approximation $\hat{\boldsymbol{\varphi}}(\cdot)$ of the nonlinear mapping from any kernel matrix $\boldsymbol{\Omega}$ using the Nyström technique.

4.1 Nonlinear Approximation in Primal Space

For clarity of presentation, consider the generic kernel matrix $\boldsymbol{\Omega}^{\text{eq}}$ with entries $\boldsymbol{\Omega}^{\text{eq}}_{ij} = K^{\text{eq}}_{\text{model}}(\boldsymbol{x}_i, \boldsymbol{x}_j)$, where the kernel function $K^{\text{eq}}_{\text{model}}$ can be defined for each of the model structures, as listed on Table 2, and where the vector $\boldsymbol{x}$ represents $\boldsymbol{z}$ or $\boldsymbol{z}_B$ depending on the model. Explicit expressions for $\hat{\boldsymbol{\varphi}}$ can be obtained by means of an eigenvalue decomposition of the kernel matrix $\boldsymbol{\Omega}^{\text{eq}}$. Given the integral equation

$$\int K^{\text{eq}}_{\text{model}}(\boldsymbol{x}, \boldsymbol{x}_j)\phi_i(\boldsymbol{x})p(\boldsymbol{x})d\boldsymbol{x} = \lambda_i \phi_i(\boldsymbol{x}), \tag{24}$$

with solutions λ_i and ϕ_i for a variable $\boldsymbol{x}$ with probability density $p(\boldsymbol{x})$, we can write

$$\boldsymbol{\varphi} = [\sqrt{\lambda_1}\phi_1, \sqrt{\lambda_2}\phi_2, \ldots, \sqrt{\lambda}_{N_h}\phi_{N_h}]^T. \tag{25}$$

Given the dataset $\{\boldsymbol{x}_i, y_i\}_{i=1}^N$, where the vector $\boldsymbol{x}_i$ represents any of the regression vectors $\boldsymbol{z}_i$ or $\boldsymbol{z}_{B,i}$ depending on the model being estimated, it is possible to approximate the integral by a sample average. This will lead to the eigenvalue

problem (Nyström approximation [12])

$$\frac{1}{N}\sum_{k=1}^{N} K^{\text{eq}}_{\text{model}}(\boldsymbol{x}_k, \boldsymbol{x}_j)u_i(\boldsymbol{x}_k) = \lambda_i^{(s)} u_i(\boldsymbol{x}_j), \tag{26}$$

where the he sample eigenvalues $\lambda_i^{(s)}$ and eigenvectors u_i of the kernel matrix $\boldsymbol{\Omega}^{\text{eq}}$ can be used for an approximation of the eigenvalues λ_i and eigenfunctions ϕ_i from the continuous problem (24), as

$$\hat{\lambda}_i = \frac{1}{N}\lambda_i^{(s)}, \hat{\phi}_i = \sqrt{N}u_i. \tag{27}$$

Based on this approximation, it is possible to compute the eigendecomposition of the kernel matrix $\boldsymbol{\Omega}^{\text{eq}}$ and use its eigenvalues and eigenvectors to compute the ith required component of $\hat{\boldsymbol{\varphi}}(\boldsymbol{x}^{(v)})$ for any point $\boldsymbol{x}^{(v)}$ by means of

$$\hat{\boldsymbol{\varphi}}_i(\boldsymbol{x}^{(v)}) = \frac{1}{\sqrt{\lambda_i^{(s)}}}\sum_{k=1}^{N} u_{ki} K^{\text{eq}}_{\text{model}}(\boldsymbol{x}_k, \boldsymbol{x}^{(v)}). \tag{28}$$

This finite dimensional approximation $\hat{\boldsymbol{\varphi}}(\boldsymbol{x})$ can be used in the primal problem (6) to estimate $\boldsymbol{w}$ and b.

4.2 Sparse Representation and Large Scale Problems

It is important to emphasize that the use of the entire training sample of size N to compute the approximation of $\boldsymbol{\varphi}$ will yield at most N components, each one of which can be computed by (27) for all $\boldsymbol{x} \in \{\boldsymbol{x}_i\}_{i=1}^{N}$. However, if we have a large scale problem, it has been motivated [12] to use of a subsample of $M \ll N$ datapoints to compute the $\hat{\boldsymbol{\varphi}}$. In this case, up to M components will be computed. It has been further proposed [4] to use an external criteria such as entropy maximization for an optimal selection of the subsample: given a fixed-size M, the aim is to select the support vectors that maximize the quadratic Renyi entropy [13]

$$H_R = -\log \int p(\boldsymbol{x})^2 d\boldsymbol{x}, \tag{29}$$

which can be approximated by using $\int \hat{p}(\boldsymbol{x})^2 d\boldsymbol{x} = \frac{1}{M^2}\mathbf{1}^T\boldsymbol{\Omega}^{\text{eq}}\mathbf{1}$. The use of this active selection procedure can be important for large scale problems, as it is related to the underlying density distribution of the sample. In this sense, the optimality of this selection is related to the final accuracy that can be obtained in the modeling exercise.

4.3 Selection of Support Vectors in the presence of correlated noise

For the uncorrelated case, building the small kernel matrix only requires evaluation of the kernel function at the selected points. However, the kernel matrices for the AR-NARX and PL-AR-NARX structures require the evaluations of $K^{\text{eq}}_{\text{ar}-\text{narx}}$ and $K^{\text{eq}}_{\text{pl}-\text{ar}-\text{narx}}$ on *past* values of the selected points. This means that the selection of support vectors for this case also requires the knowledge of the points that are located τ timesteps in the past.

4.4 Final Estimation in Primal Space

From a kernel matrix $\boldsymbol{\Omega}^{\text{eq}}$ evaluated over a data subsample of fixed-size M, the approximation $\hat{\boldsymbol{\varphi}}$ is obtained [4] using (28) for each of the components. Let denote by $\hat{\boldsymbol{\varphi}}^{\text{model}}(\boldsymbol{z}_t)$ the approximation of the feature map for a datapoint $\boldsymbol{z}_t$ for a given model structure. The model can now be estimated in primal space directly by using ridge regression techniques with regularization parameter γ. This is, the problem can be solved by minimizing a regularized least-squared cost function as follows.

- For the NARX model in Primal Space, the solutions $\boldsymbol{w}, b$ are obtained from

$$\min_{\boldsymbol{w},b,e_i} \quad \frac{1}{2}\boldsymbol{w}^T\boldsymbol{w} + \gamma\frac{1}{2}\sum_{i=1}^{N} e_i^2 \tag{30}$$

$$\text{s.t.} \quad y_i = \boldsymbol{w}^T\hat{\boldsymbol{\varphi}}^{\text{narx}}(\boldsymbol{z}_i) + b + e_i, \quad i = 1,\ldots,N.$$

  With the explicit expression for $\hat{\boldsymbol{\varphi}}^{\text{narx}}(\boldsymbol{z}_i)$, the model is solved in primal space by eliminating e_i from (30),

$$\min_{\boldsymbol{w},b} \quad \frac{1}{2}\boldsymbol{w}^T\boldsymbol{w} + \gamma\frac{1}{2}\sum_{i=1}^{N}(y_i - \boldsymbol{w}^T\hat{\boldsymbol{\varphi}}^{\text{narx}}(\boldsymbol{z}_i) - b)^2. \tag{31}$$

- For the AR-NARX model, the solutions $\boldsymbol{w}, b$ are obtained from

$$\min_{\boldsymbol{w},b,r_i} \quad \frac{1}{2}\boldsymbol{w}^T\boldsymbol{w} + \gamma\frac{1}{2}\sum_{i=\tau+1}^{N} r_i^2 \tag{32}$$

$$\text{s.t.} \quad y_i = \rho y_{i-\tau} - \boldsymbol{w}^T\hat{\boldsymbol{\varphi}}^{\text{ar}-\text{narx}}(\boldsymbol{z}_i) + b + r_i, \quad i = \tau+1,\ldots,N.$$

  where the autocorrelation structure for the nonlinear function is embedded into the evaluation of $\hat{\boldsymbol{\varphi}}^{\text{ar}-\text{narx}}(\boldsymbol{z}_i)$ computed from the kernel matrix $\boldsymbol{\Omega}^{\text{eq}}$. With the explicit expression for $\hat{\boldsymbol{\varphi}}^{\text{ar}-\text{narx}}(\boldsymbol{z}_i)$, the model is solved in primal

space by eliminating r_i from (32),

$$\min_{\boldsymbol{w},b} \quad \frac{1}{2}\boldsymbol{w}^T\boldsymbol{w} + \gamma\frac{1}{2}\sum_{i=1}^{N}(y_i - \rho y_{i-\tau} - \boldsymbol{w}^T\hat{\boldsymbol{\varphi}}^{\mathrm{ar-narx}}(\boldsymbol{z}_i) - b)^2. \tag{33}$$

- For the PL-NARX model, the solutions $\boldsymbol{\beta}, \boldsymbol{w}, b$ are obtained from

$$\min_{\boldsymbol{w},b,e_i,\boldsymbol{\beta}} \quad \frac{1}{2}\boldsymbol{w}^T\boldsymbol{w} + \gamma\frac{1}{2}\sum_{i=1}^{N} e_i^2 \tag{34}$$

$$\text{s.t.} \quad y_i = \boldsymbol{\beta}^T\boldsymbol{z}_{A,i} + \boldsymbol{w}^T\hat{\boldsymbol{\varphi}}^{\mathrm{pl-narx}}(\boldsymbol{z}_{B,i}) + b + e_i. \;\; i = 1, \ldots, N,$$

With the explicit expression for $\hat{\boldsymbol{\varphi}}^{\mathrm{pl-narx}}(\boldsymbol{z}_{B,i})$, the model is solved in primal space by eliminating e_i from (34),

$$\min_{\boldsymbol{w},b,\boldsymbol{\beta}} \quad \frac{1}{2}\boldsymbol{w}^T\boldsymbol{w} + \gamma\frac{1}{2}\sum_{i=1}^{N}(y_i - \boldsymbol{\beta}^T\boldsymbol{z}_{A,i} - \boldsymbol{w}^T\hat{\boldsymbol{\varphi}}^{\mathrm{pl-narx}}(\boldsymbol{z}_{B,i}) - b)^2. \tag{35}$$

- Finally, for the PL-AR-NARX model structure, the solutions $\boldsymbol{\beta}, \boldsymbol{w}, b$ are obtained from

$$\min_{\boldsymbol{w},b,r_i,\boldsymbol{\beta}} \quad \frac{1}{2}\boldsymbol{w}^T\boldsymbol{w} + \gamma\frac{1}{2}\sum_{i=\tau+1}^{N} r_i^2 \tag{36}$$

$$\text{s.t.} \quad y_i = \rho y_{i-\tau} + \boldsymbol{\beta}^T(\boldsymbol{z}_{A,i} - \rho\boldsymbol{z}_{A,i-\tau}) + \boldsymbol{w}^T\hat{\boldsymbol{\varphi}}^{\mathrm{pl-ar-narx}}(\boldsymbol{z}_{B,i}) + b + r_i,$$

for $i = \tau + 1, \ldots, N$. With the explicit expression for $\hat{\boldsymbol{\varphi}}^{\mathrm{pl-ar-narx}}(\boldsymbol{z}_{B,i})$, the model is solved in primal space by eliminating r_i from (36),

$$\min_{\boldsymbol{w},b,\boldsymbol{\beta}} \quad \frac{1}{2}\boldsymbol{w}^T\boldsymbol{w} + \gamma\frac{1}{2}\sum_{i=\tau+1}^{N}(y_i - \rho y_{i-\tau} - \boldsymbol{\beta}^T(\boldsymbol{z}_{A,i} - \rho\boldsymbol{z}_{A,i-\tau}) - \boldsymbol{w}^T\hat{\boldsymbol{\varphi}}^{\mathrm{pl-ar-narx}}(\boldsymbol{z}_{B,i}) - b)^2. \tag{37}$$

5 Case Study: Short-Term Load Forecasting

In this section the practical application is described, in terms of the problem context, methodological issues and results.

5.1 Description and Objective

The main goal in load forecasting to estimate a model that can capture all the dynamics between possible explanatory variables for the load, in order to obtain forecasts from one hour ahead, up to one day ahead [14]. In the literature there is a broad consensus concerning important explanatory variables: past

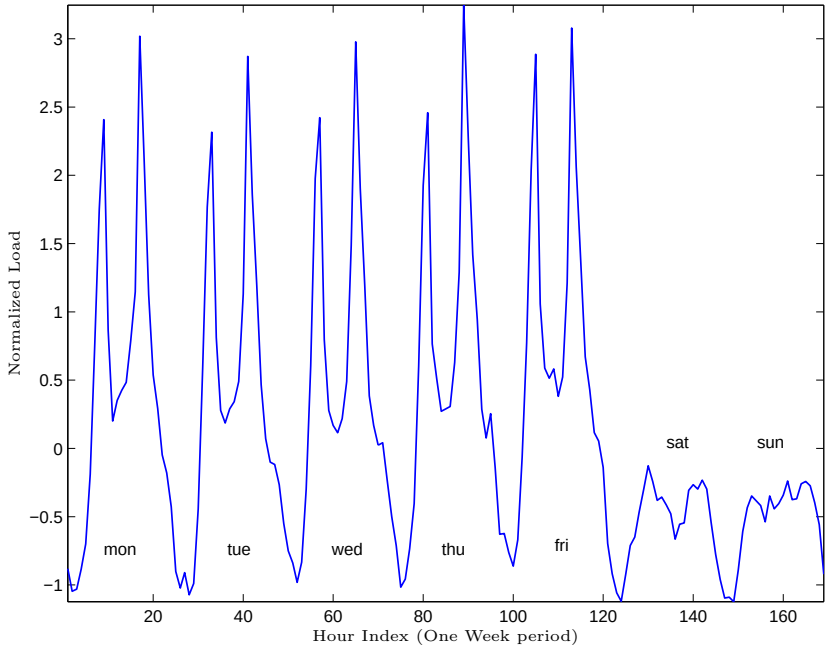

Fig. 1. *Example of the Load series within a week. Daily cycles are visible, as well as the weekend effects. Also visible are the intra-day phenomena, such as the peaks (morning, noon, evening) and the night hours.*

values of the load, weather information, calendar information, etc. For further details, the interested reader is referred to [3,15,16]. Our objective is to show an exploratory data modeling methodology, where prior knowledge about the problem can be incorporated at different levels, obtaining a model structure tailored to the STLF application.

5.2 Data Description

The data consists of a 3.5-years series of hourly load values from a particular substation in Belgium, thus containing more than 26,000 datapoints. We define our training set to consist of 18,000 datapoints, and we validate the final models using a test set of 30 days out of sample. Figure 1 shows an example of the load series in a week, at hourly values starting at 00:00 hrs on Monday, until 24:00 hrs on Sunday. The daily shapes, the daily peaks and the weekly cycle are clearly visible.

5.3 Model Structures and Estimation

The model structures described previously are applied to the STLF problem with the available data. The general regression vector $\boldsymbol{z}$ contains an autore-

gressive part of 48 lagged load values $y_{t-1}, \ldots, y_{t-48}$ (the past 2 days); and exogenous inputs $\boldsymbol{u}_t$ containing the temperature-related variables (measuring the effect of temperature on cooling and heating requirements [17]) and calendar information in the form of dummy variables for month of the year, day of the week and hour of the day [3]. This leads to a set of 48+3+12+7+24= 94 explanatory variables. Most of the prior information about the seasonal properties of the load are incorporated into the modeling stage in the form of the mentioned dummy variables for calendar information. In addition, it is known that the hourly load series shows a strong correlation every 24 hours, therefore we use $\tau = 24$ for the noise correlation description for the AR-NARX and PL-AR-NARX models.

The procedure for model estimation using the large-scale fixed-size LS-SVM methodology can be described in the following steps:

(1) Normalize inputs and outputs to have zero mean and unit variance
(2) Select an initial subsample of size M
(3) Choose a Model Structure (NARX, AR-NARX, PL-NARX, PL-AR-NARX)
(4) Build the $(M \times M)$ kernel matrix for the model structure and compute its eigendecomposition
(5) Build the nonlinear feature map approximation for the rest of the data
(6) Estimate the model in primal space according to the primal representation from Table 1
(7) Estimate the nonlinear feature map approximation for a test data
(8) Use the model estimates with the test data nonlinear mapping to obtain the out of sample forecasts

5.4 Performance Measurements

The performance of the predictions obtained from any given model can be assessed by using different indicators. Denoting by y_i and $\hat{y}_i$ the observed values and the predictions, respectively, in an evaluation sample $mathcalS$, the Mean Squared Error (MSE) is defined as

$$\text{MSE} = \sum_{i \in \mathcal{S}} \frac{(y_i - \hat{y}_i)^2}{N_s}, \tag{38}$$

and the Mean Absolute Percentage Error (MAPE) is defined as

$$\text{MAPE} = \sum_{i \in \mathcal{S}} \frac{|y_i - \hat{y}_i|}{y_i}. \tag{39}$$

The MSE is typically used within the context of applied modeling, but for the STLF problem it is also important to measure the absolute errors in terms of

the load. For this reason, in this case study the performance of the models is assessed using both indicators.

5.5 Hyperparameters Selection

The tuning of the hyperparameters σ, γ and ρ is performed by 10 fold cross-validation (CV) in the training sample. By using cross-validation, it is possible to incorporate another element of the prior knowledge into the modeling. Typically, for the training of a model using CV, the Mean-Squared-Error (MSE) is used as a performance measure, and therefore the hyperparameters to be used in the out of sample predictions are those that minimize the cross-validation MSE (MSE-CV). However, the predictions for the daily peak load are critical. An error in the prediction at the peak hour has a larger impact into the planning and unit commitment scheduling [14]. For this reason, we also implement the tuning in such a way that the selected parameters minimize the Mean Absolute Prediction Error under cross-validation (MAPE-CV), which is computed only for the daily peak loads. Under this scheme, we look for the hyperparameters that are able to produce accurate forecasts for the daily peaks.

For any model structure, the hyperparameters can be tuned using MSE-CV and MAPE-CV. Once the hyperparameters are selected, then the model is used on the test sample, where its performance is measured by both the MSE and MAPE.

The size M of the subsample from which the feature map approximation is computed can affect the performance of the final model. For this paper we use $M = 1000$, which accounts for 4% of the available dataset. The size of $M = 1000$ can produce very good results for this problem without requiring too much computational time. For an illustration of the the effect of increasing sizes of M, the interested reader is referred to [18].

6 Results

The results of the model selection methodology, the identified model structures and the final prediction results are described on this section.

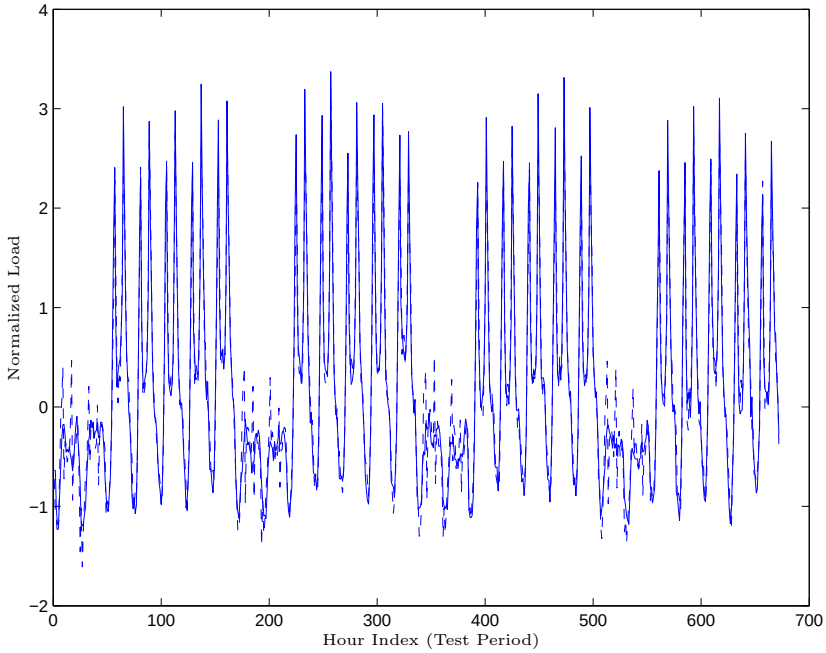

Fig. 2. *Performance of the linear model. The observed load (full line) is compared to the prediction from the linear ARX (dashed) model.*

6.1 (N)-ARX Models

First, a linear ARX model (3) is estimated with Ordinary Least Squares using the $\boldsymbol{z}_t$ with the input variables defined above. The performance of the linear model on the test set is 0.04 (MSE) and 0.143 (MAPE). A second model is estimated, this time a full black-box NARX model ('NARX' on Table 1). The hyperparameters for this second model can be tuned by minimizing the MSE-CV or the MAPE-CV. For the case of MSE-CV minimization, the test set results are 0.019 (MSE) and 0.037 (MAPE). For the case of MAPE-CV minimization, the results are 0.020 (MSE) and 0.036 (MAPE). It is already clear that the NARX model shows a better performance in terms of prediction accuracy than the linear ARX model, which means that there are some effects that are captured by the linear structure, and in turn they are captured in the full black-box NARX model. However, it is interesting to check whether these effects are limited to the effect of a subset of the input variables, or to the correlation structure. So far, it is clear that the NARX model is not substantially improved by moving from the minimization of the MSE-CV to the minimization of the MAPE-CV. The quality of the ARX and NARX predictions for the test set are shown on Figures (2) and (3), respectively. It can be observed from the figures that the linear model generates bad predictions for the weekend days. The residuals for the both models are shown on Figure 4. The difference between the ARX and NARX models is very significant.

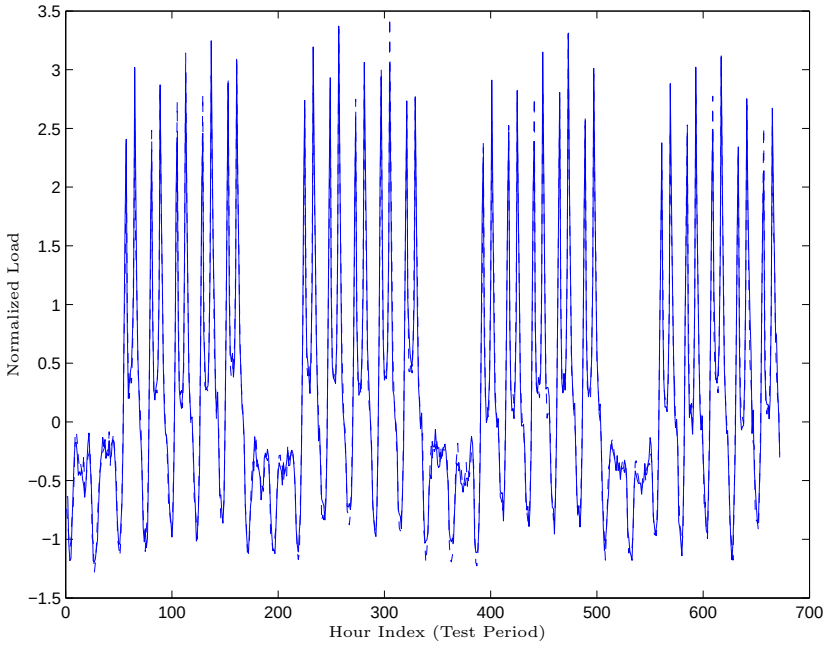

Fig. 3. *Performance of the NARX model. The observed load (full line) is compared to the prediction from the NARX (dashed) model.*

6.2 AR-NARX

The next step is to check if the NARX black-box model can be improved by imposing an autocorrelation with $\tau = 24$. The autocorrelation parameter ρ is tuned by minimization of the MAPE-CV, obtaining an optimal ρ^* close to -0.4, as shown on Figure (5). According to this result, there exist a negative correlation on the residuals of the NARX model. The estimated AR-NARX model, however, does not show a significant improvement over the NARX model. The AR-NARX performance on the test set is 0.020 (MSE) and 0.036 (MAPE). A linear AR-ARX model obtains a small positive ρ parameter of 0.2, with a small improvement to 0.032 (MSE) and 0.120 (MAPE).

6.3 PL-NARX

Now, we turn into the partially linear parameterizations. Several linear models have been proposed in the load forecasting literature [19,20] with good results, it is interesting to check if a partially linear parameterization can improve over the linear ARX model into the direction of the full black-box NARX model. Since the seminal work of partially linear models applied to electricity prices [17], it has been common practice to separate the inputs which are the past values of the load, from those inputs which are calendar and temperature effects. In our notation, starting from our original regression vector $\boldsymbol{z}_t$, the

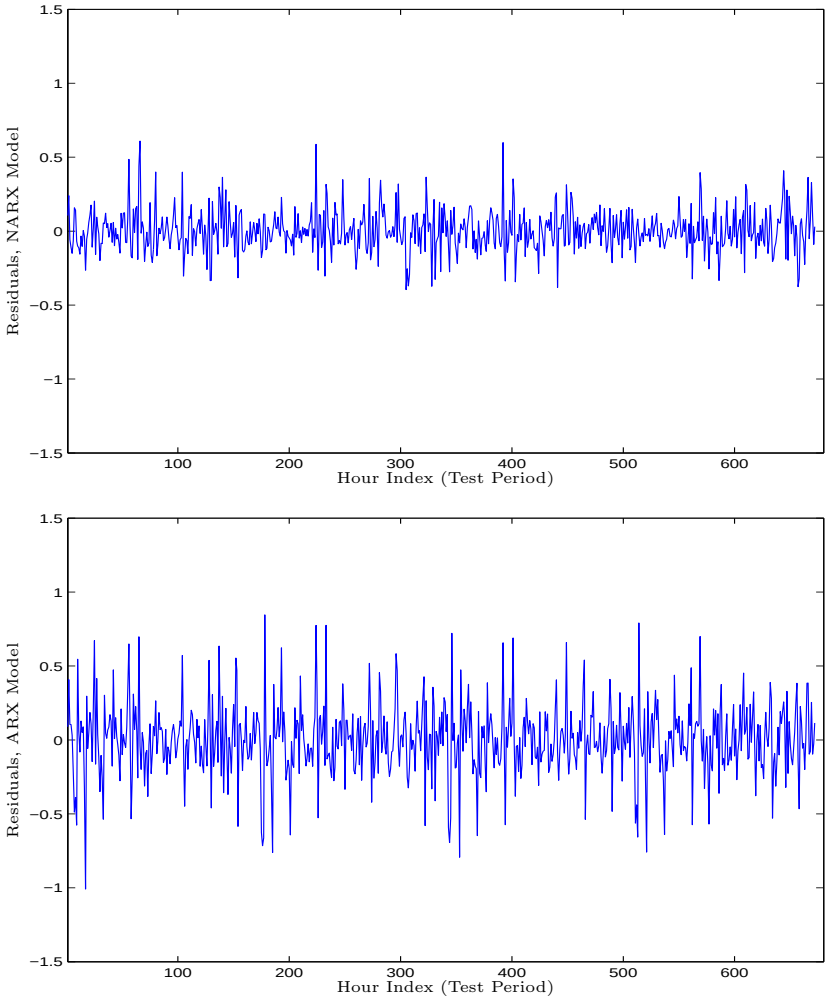

Fig. 4. *Residuals on the Test Set, for the NARX model (Top panel) and the ARX model (Bottom panel).*

following partitions are defined to be used in the PL-NARX models:

- Use the past values of the load $y_{t-i}, i = 1, \ldots, 48$ as nonlinear regressors, and the exogenous inputs $\boldsymbol{u}_t$ (calendar and temperature effects) as linear regressors. This is, $\boldsymbol{z}_{A,t} = \boldsymbol{u}_t, \boldsymbol{z}_{B,t} = [y_{t-1}; y_{t-2}, \ldots, y_{t-48}]$. Using this partition, the PL-NARX model is estimated. Let us denote this model by 'PL-NARX-1'.
- Use the past values of the load $y_{t-i}, i = 1, \ldots, 48$ as linear regressors, and the exogenous inputs $\boldsymbol{u}_t$ (calendar and temperature effects) as nonlinear regressors. In this case, $\boldsymbol{z}_{A,t} = [y_{t-1}; y_{t-2}, \cdots, y_{t-48}], \boldsymbol{z}_{B,t} = \boldsymbol{u}_t$. Let us denote by 'PL-NARX-2' the model estimated using this partition.

The performance of the models PL-NARX-1 and PL-NARX-2 differ from each other. When the tuning parameters are found by minimization of the MSE-CV, the performance of PL-NARX-1 is 0.021 (MSE) and 0.039 (MAPE) on the test set. The performance of PL-NARX-2 is 0.029 (MSE) and 0.073 (MAPE). Under this tuning scheme, the model with linear terms on the calendar and

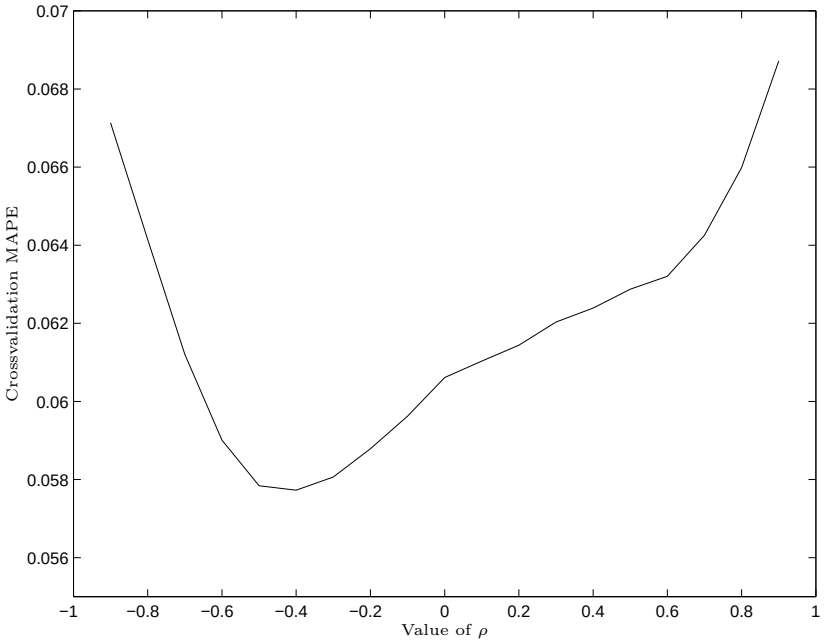

Fig. 5.
MAPE-CV performance of the AR-NARX model for different values of ρ

temperature information (PL-NARX-2) performs worse than the model PL-NARX, which, in turn, shows a result comparable to the NARX model above.

However, when the tuning parameters are found by minimization of the MAPE-CV, model PL-NARX-2 improves considerably. Under this tuning scheme, the performance of PL-NARX-1 is 0.021 (MSE) and 0.037 (MAPE), whereas PL-NARX-2 improves to 0.019 (MSE) and 0.036 (MAPE). This last result is already comparable to the performance of the AR-NARX model. This example shows that moving from the MSE-CV to the MAPE-CV for hyperparameter selection can improve substantially the performance of a model.

The improvement of the model PL-NARX-2 is shown on Figure 6, where the prediction obtained with the PL-NARX-2 model tuned with MSE-CV (dotted) is compared to the prediction obtained with the PL-NARX-2 model tuned with MAPE-CV (dashed). For an easier visualization, only part of the test set data is shown on the plot. By looking to the actual values (thin line) it can be observed that the model improves around the peak of each day, which is the purpose of using the MAPE performance criterion for hyperparameters tuning.

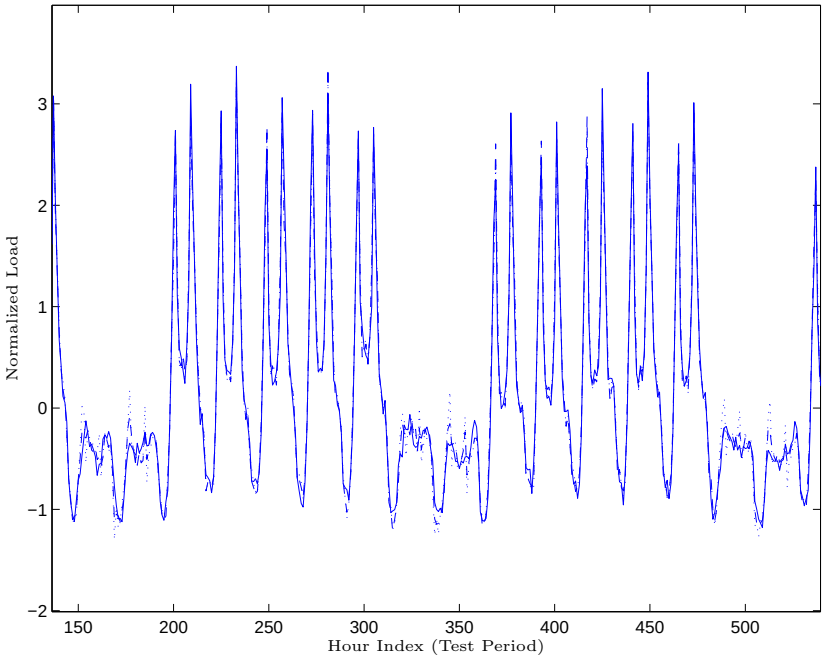

Fig. 6. *Improvement of the PL-NARX-2 model. The observed load (full line) is compared to the prediction from the PL-NARX-2 model tuned with MSE-CV (dotted) and the prediction from the PL-NARX-2 model tuned with MAPE-CV (dashed).*

6.4 PL-AR-NARX

Finally, it is still possible to check the effect of imposing an autocorrelation structure on both PL-NARX models. Using the partitions over the original regression vector defined above, two models are estimated, PL-AR-NARX-1 and PL-AR-NARX-2.

For the case of the PL-AR-NARX-1 model (which uses $\boldsymbol{z}_{A,t} = \boldsymbol{u}_t, \boldsymbol{z}_{B,t} = [y_{t-1}; y_{t-2}, \ldots, y_{t-48}]$), there is a small improvement when testing different values of ρ. For this model, the optimal ρ^* is found to be -0.2, with a final performance on the test set of 0.021 (MSE) and 0.034 (MAPE).

For the case of the PL-AR-NARX-2 model (which uses $\boldsymbol{z}_{A,t} = [y_{t-1}; y_{t-2}, \ldots, y_{t-48}]$, $\boldsymbol{z}_{B,t} = vecU_t$), the optimal ρ^* is found at -0.5, as shown on Figure 7. For this model, the performance on the test set is now 0.019 (MSE) and 0.031 (MAPE). The PL-AR-NARX-2 model is therefore better (in terms of MAPE) than the full black-box NARX model estimated above.

The improvement of the PL-AR-NARX-2 model over the initial full black-box NARX model can be visualized by comparing the residuals on the test set, as shown on Figure 8. The large peaks found on the residuals from the NARX models are not present on the residuals from the PL-AR-NARX-2 model.

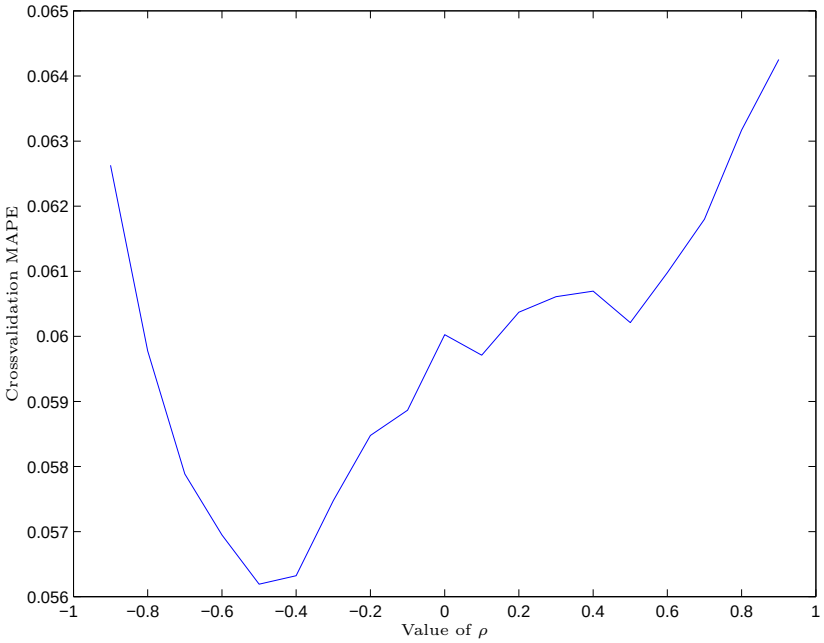

Fig. 7. *MAPE CV performance of the PL-NARX-2 model for different values of* ρ

Table 3
Results on Test Set: Summary

Model	Tuning with MSE-CV		Tuning with MAPE-CV			
	No Noise Model		No Noise Model		With AR Noise Model	
	MSE	MAPE	MSE	MAPE	MSE	MAPE
ARX	0.040	0.143	0.040	0.143	0.032	0.120
NARX	0.019	0.037	0.020	0.036	0.020	0.036
PL-NARX-1	0.021	0.039	0.021	0.037	0.021	0.034
PL-NARX-2	0.029	0.073	0.019	0.036	0.019	0.031

6.5 Summary of Results

The results of the exploratory data modeling are summarized on Table 3. It is found that the black-box NARX model does not improve substantially when adding a noise structure. However, the partially linear structure PL-NARX-2 which is linear on the past values of the load and nonlinear on the calendar and temperature information, can improve substantially when an AR noise structure is included. The improvement is such that it can match, and even surpass, the performance of the full black-box NARX model.

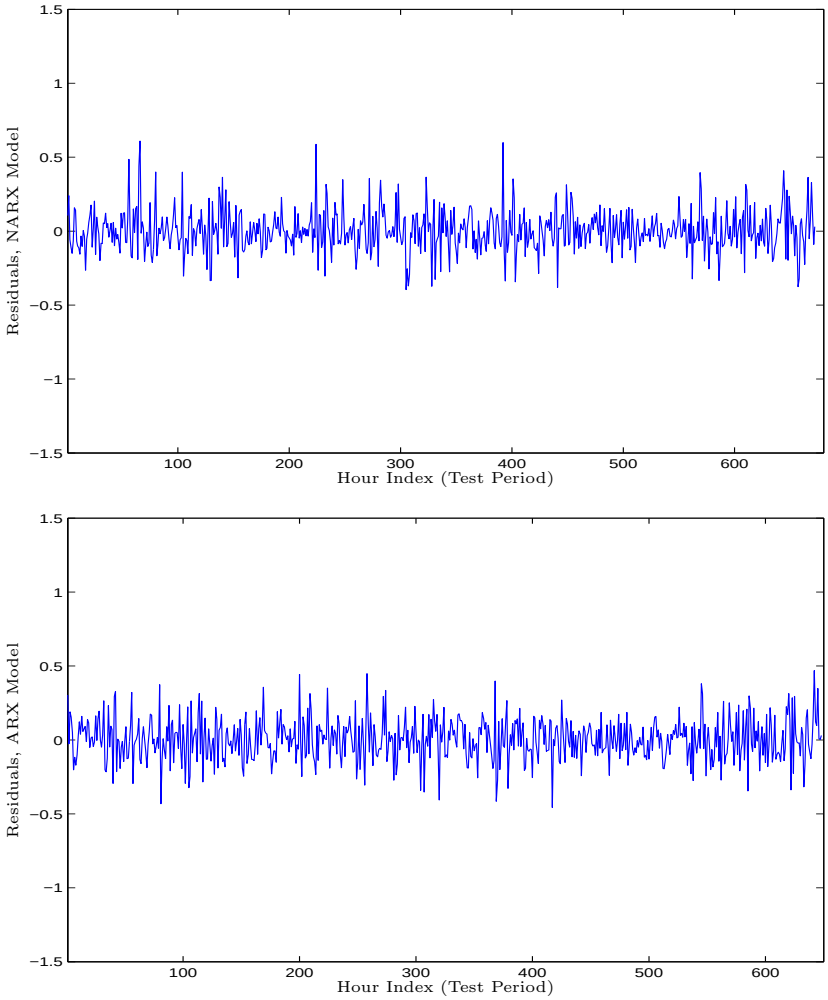

Fig. 8. *Residuals on the Test Set, for the NARX model (Top panel) and the PL-AR-NARX2 model (Bottom panel).*

7 Conclusion

In this paper we have used an exploratory modeling methodology for a real-life case study. Based on the LS-SVMs formulation, several model structures and parameterizations have been tested, from a linear ARX model, to a full black box NARX model, to partially linear structures with autocorrelated residuals. In all these cases, the solution of each model comes out of a linear system which is obtained from a convex problem. It is illustrated that a modeling strategy can benefit from prior knowledge of the problem under research. In this way, the obtained model is clearly more tailored towards the application, with better performance on the test sample. For the case study of short-term load forecasting, it has been shown how the modeling strategy can benefit from a tailored definition of the cross-validation criteria. By moving from the minimization of the MSE-CV to the minimization of the MAPE-CV over the

daily peaks, and by including an autocorrelation with delay $\tau = 24$, the results improve substantially for the partially linear structure. This structured model, which is linear on the past values of the load and nonlinear on the calendar and temperature information, shows a final performance on the test set which is better than the performance of the full black-box NARX model.

Acknowledgments. This work is supported by grants and projects for the Research Council K.U.Leuven (GOA- Mefisto 666, GOA- Ambiorics, several PhD/ Postdocs & fellow grants), the Flemish Government (FWO: PhD/ Postdocs grants, projects G.0211.05, G.0240.99, G.0407.02,G.0197.02,G.0141.03,G.0491.03,G.0120.03, G.0452.04, G.0499.04, ICCoS, ANMMM; AWI; IWT: PhD grants, GBOU (McKnow) Soft4s), the Belgian Federal Government (Belgian Federal Science Policy Office: IUAP V-22; PODO-II (CP/ 01/40), the EU (FP5- Quprodis; ERNSI, Eureka 2063- Impact; Eureka 2419- FLiTE) and Contracts Research / Agreements (ISMC /IPCOS, Data4s, TML, Elia, LMS, IPCOS, Mastercard). J. Suykens and B. De Moor are an associate professor and a full professor at the K.U.Leuven, Belgium, respectively. The scientific responsibility is assumed by its authors.

References

[1] L. Ljung, *System Identification: Theory for the User*, Prentice Hall, New Jersey, 1987.

[2] S. Hylleberg, *Modelling Seasonality*, Oxford University Press, 1992.

[3] M. Espinoza, C. Joye, R. Belmans, B. De Moor, Short term load forecasting, profile identification and customer segmentation: A methodology based on periodic time series, *IEEE Transactions on Power Systems* 20 (3) (2005) 1622–1630.

[4] J.A.K. Suykens, T. Van Gestel, J. De Brabanter, B. De Moor, J. Vandewalle, *Least Squares Support Vector Machines*, World Scientific, Singapore, 2002.

[5] M. Espinoza, J.A.K. Suykens, B. De Moor, Kernel based partially linear models and nonlinear identification, *IEEE Transactions on Automatic Control, Special Issue: Linear vs. Nonlinear* To appear.

[6] M. Espinoza, J.A.K. Suykens, B. De Moor, LS-SVM regression with autocorrelated errors, Tech. Rep. 05-177, ESAT-SISTA, K.U.Leuven (2005).

[7] J. Sjöberg, Q. Zhang, L. Ljung, A. Benveniste, B. Deylon, P. Glorennec, H. Hjalmarsson, A. Juditsky, Nonlinear Black-box Modelling in System Identification: a Unified Overview, *Automatica* 31 (1995) 1691–1724.

[8] A. Juditsky, H. Hjalmarsson, A. Benveniste, B. Deylon, L. Ljung, J. Sjöberg, Q. Zhang, Nonlinear Black-box Modelling in System Identification: mathematical foundations, *Automatica* 31 (1995) 1725–1750.

[9] R. Guidorzi, *Multivariable System Identification*: From Observations to Models, Bononia University Press, 2003.

[10] J. Hamilton, *Time Series Analysis*, Princeton University Press, 1994.

[11] V. Vapnik, *Statistical Learning Theory*, Wiley, New-York, 1998.

[12] C. Williams, M. Seeger, Using the nyström method to speed up kernel machines, in: V. T.Leen, T.Dietterich (Ed.), *Proc. NIPS 2000*, Vol. 13, MIT press, Vienna, Austria, 2000.

[13] M. Girolami, Orthogonal series density estimation and the kernel eigenvalue problem, *Neural Computation* 10 (6) (1998) 1455–1480.

[14] E. Mariani, S. Murthy, *Advanced Load Dispatch for Power Systems*, Advances in Industrial Control, Springer-Verlag, 1997.

[15] H. Steinherz, C. Pedreira, R. Castro, Neural networks for short-term load forecasting: A review and evaluation, *IEEE Transactions on Power Systems* 16 (1) (2001) 44–55.

[16] D. Fay, J. Ringwood, M. Condon, M. Kelly, 24-h electrical load data-a sequential or partitioned time series?, *Neurocomputing* (55) (2003) 469–498.

[17] R. Engle, C. Granger, J. Rice, A. Weiss, Semiparametric estimates of the relation between weather and electricity sales, *Journal of the American Statistical Association* 81 (394) (1986) 310–320.

[18] M. Espinoza, J. Suykens, B. De Moor, Load forecasting using fixed-size least squares support vector machines, in: J. Cabestany, A. Prieto, F. Sandoval (Eds.), *Proceedings of the 8th International Work-Conference on Artificial Neural Networks*, Vol. 3512 of Lecture Notes in Computer Science, Springer-Verlag, 2005, pp. 1018–1026.

[19] N. Amjady, Short-term hourly load forecasting using time-series modeling with peak load estimation capability, *IEEE Transactions on Power Systems 16 (4) (2001)* 798–805.

[20] R. Ramanathan, R. Engle, C. Granger, C. Vahid-Aragui, F.and Brace, Short-run forecasts of electricity load and peaks, *International Journal of Forecasting* (13) (1997) 161–174.