Structured Kernel Based Modeling: An Exploration in Short-Term Load Forecasting

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

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

Abstract

This paper considers an exploratory modeling strategy applied to a large scale real-life problem of power load forecasting. Different model structures are considered, including Autoregressive models with eXogenous inputs (ARX), Nonlinear Autoregressive models with eXogenous inputs (NARX), both of which are also extended to incorporate residuals that follow an Autoregressive (AR) process (AR-(N)ARX). These models are parameterized either as a full black-box model or as partially linear structures. Starting from the Least Squares Support Vector Machines (LS-SVMs) formulation for regression, relevant properties from the problem at hand can be imposed as additional constraints which can be incorporated in a straightforward way within the primal-dual optimization formulation and be solved through convex optimization. For the case study of short-term load forecasting, it is shown that a structured model can reach results comparable to the full black-box model, with the additional benefit of having interpretable coefficients. It obtains satisfactory results when the model structure and the hyperparameter selection procedure are tailored towards the application, taking into account prior knowledge or relevant properties of the problem.

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

1 Introduction

This paper addresses the issue of selecting a suitable model structure by using prior knowledge and physical insight taken from the problem under analysis.

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

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 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 predictive accuracy because it has been"tailored" to the application at hand.

The work in this paper has been developed using the Least Squares Support Vector Machines (LS-SVMs [2]) nonlinear regression formulation as a basis technique for nonlinear system identification. LS-SVMs belong to the class of kernel methods, which is a subject with several directions studied in different fields. LS-SVMs and Support Vector Machines (SVMs) [3–5] follow the approach of a primal-dual optimization formulation, where both techniques make use of a so-called feature space where the inputs have been transformed by means of a (possibly infinite dimensional) nonlinear mapping $\boldsymbol{\varphi}$. This is converted to the dual space by means of Mercer's theorem and the use of a positive definite kernel, without computing explicitly the mapping $\boldsymbol{\varphi}$. The SVM model solves a quadratic programming problem in dual space, obtaining a sparse solution [6]. The LS-SVM formulation, on the other hand, solves a linear system under a least squares cost function with equality constraints, where the sparseness property can be obtained e.g. by sequentially pruning the support value spectrum. The LS-SVM training procedure involves a selection of the kernel parameter and the regularization parameter of the cost function, that usually can be done by cross-validation or by using Bayesian techniques [7]. Other directions in kernel methods follow different approaches. In Reproducing Kernel Hilbert Spaces (RKHS) [8] the problem of function estimation is treated as a variational problem ; Gaussian Processes (GP) [9,10] follow a probabilistic-Bayesian setting. Although these different approaches have links with each other, e.g. for the simple case of static regression without a bias term it is well-known that GP, regularization networks and LS-SVM lead to the same set of linear equations to be solved at the dual level, in general the methodologies are different. Particularly, the primal-dual formulation of LS-SVM makes it easy to add additional constraints, which in the context of the present work makes it straightforward to incorporate more structure into the models, as it will be illustrated throughout the paper. In addition, the LS-SVM models can be estimated in primal space directly with a sparse representation by using Nyström methods (which originated in the GP literature) for the case of large samples, which is also exploited in this paper.

In this work we consider the LS-SVM formulation for regression. However, the LS-SVM framework is more general [2]. Given that the LS-SVM formulation works with equality constraints and a L_2 loss function, this optimization-based kernel methodology can be extended to a wide range of problems, including

kernel versions of PCA (Principal Components Analysis) [11], FDA (Fisher's Discriminant Analysis) [12,13], CCA (Canonical Correlation Analysis) [14], PLS (Partial Least Squares), spectral clustering, subspace methods [15], recurrent networks, control, and others. In general, LS-SVMs can be seen as core formulations from which more sophisticated kernel machines can be built by using additional constraints or using different loss functions for robustness or sparsity. Within the LS-SVM framework, different model structures and parameterizations can often be estimated as a solution of a set of linear equations obtained from a convex problem.

As a real-life application, we consider the problem of Short-Term Load Forecasting (STLF), which is an important area of applied quantitative research. The goal of STLF is to build a model that can be used for the prediction of the power load over the next few 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 [16]. This situation translates into autocorrelation patterns across the corresponding hour on previous days [17]. In addition, unlike other time series problems, in STLF accurate forecasts are required particularly for the hour of the daily peak. 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 [17] in terms of temperature sensitivity, growth trends, useful 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 [2]). Previous work on short-term load forecasting using black-box models estimated with LS-SVMs only considers the case of a general NARX black-box formulation [18,19]. The work presented in this paper extends the framework towards the use of structured models which are tailored to the application under study.

This paper is organized as follows. The description of the different model structures used in this paper is presented in Section 2. The estimation techniques based on LS-SVM are described in Section 3, where the model structures are parameterized in several ways. In Section 4, the implementation using Fixed-Size LS-SVMs [2] 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 in Section 5.

2 Model Formulation

In this paper we consider two 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 $\boldsymbol{u}_t \in \mathbb{R}^{N_u}$, with $n = p + N_u q$. A (N)ARX model structure [1,20,21] 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 [1,22,23] 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 used in applied modeling work [24], as in the seminal work of Engle et al. on electricity prices [25].

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. This model can be estimated as a linear regression, using e.g. Ordinary Least Squares (OLS).

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 [2]:

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

where b is a constant (bias) term, and $\boldsymbol{\varphi}(\cdot) : \mathbb{R}^n \rightarrow \mathbb{R}^{N_h}$ 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 [2,4], 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) [26] 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 $\boldsymbol{z}_t = [\boldsymbol{z}_{A,t}; \boldsymbol{z}_{B,t}]$, where $\boldsymbol{\varphi}(\cdot) : \mathbb{R}^n \rightarrow \mathbb{R}^{N_h}$ is the feature map from input space to feature space.

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 (primal problem):

$$\begin{aligned} \min_{\boldsymbol{w},b,e_i} \quad & \frac{1}{2}\boldsymbol{w}^T\boldsymbol{w} + \gamma\frac{1}{2}\sum_{i=1}^{N} e_i^2 \\ \text{s.t.} \quad & y_i = \boldsymbol{w}^T\boldsymbol{\varphi}(\boldsymbol{z}_i) + b + e_i, \quad i = 1,\ldots,N. \end{aligned} \tag{6}$$

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

Lemma 1 [2,27] *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:

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

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:

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

The solution is formalized in the following lemma.

Lemma 2 [23] *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:

$$\begin{aligned} \min_{\boldsymbol{w},b,e_i,\boldsymbol{\beta}} \quad & \frac{1}{2}\boldsymbol{w}^T\boldsymbol{w} + \gamma\frac{1}{2}\sum_{i=1}^{N} e_i^2 \\ \text{s.t.} \quad & y_i = \boldsymbol{\beta}^T\boldsymbol{z}_{A,i} + \boldsymbol{w}^T\boldsymbol{\varphi}(\boldsymbol{z}_{B,i}) + b + e_i, \quad i = 1,\ldots,N, \end{aligned} \tag{14}$$

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

Lemma 3 [26] *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:

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

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 [4] $\boldsymbol{\varphi}(\boldsymbol{z}_{B,i})^T\boldsymbol{\varphi}(\boldsymbol{z}_{B,j}) = K(\boldsymbol{z}_{B,i}, \boldsymbol{z}_{B,j})$ with a positive definite kernel K, we can eliminate $\boldsymbol{w}$ and r_i, obtaining

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

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

The estimated model in dual space becomes

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

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

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

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

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 as studied in the context of Gaussian Processes in [28].

4.1 Nonlinear Approximation in Primal Space

For clarity of presentation, consider the generic kernel matrix $\boldsymbol{\Omega}^{\text{eq}}$ with entries $\boldsymbol{\Omega}_{ij}^{\text{eq}} = K_{\text{model}}^{\text{eq}}(\boldsymbol{x}_i, \boldsymbol{x}_j)$, where the kernel function $K_{\text{model}}^{\text{eq}}$ 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

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

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

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

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 [28])

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

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 in 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

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

leading to the $M-$dimensional approximation

$$\hat{\boldsymbol{\varphi}}(\boldsymbol{x}) = [\hat{\boldsymbol{\varphi}}_1(\boldsymbol{x}), \hat{\boldsymbol{\varphi}}_2(\boldsymbol{x}), \ldots, \hat{\boldsymbol{\varphi}}_M(\boldsymbol{x})]^T. \tag{29}$$

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 [28] 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 [2] 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 [29]

$$H_R = -\log \int p(\boldsymbol{x})^2 d\boldsymbol{x}, \tag{30}$$

which can be approximated by using $\int \hat{p}(\boldsymbol{x})^2 d\boldsymbol{x} = \frac{1}{M^2}\mathbf{1}^T \boldsymbol{\Omega}^{\mathrm{eq}} \mathbf{1}$. 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 $K^{\mathrm{eq}}_{\mathrm{ar-narx}}$ and $K^{\mathrm{eq}}_{\mathrm{pl-ar-narx}}$ 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 $\boldsymbol{\Omega}^{\mathrm{eq}}$ evaluated over a data subsample of fixed-size M, the approximation $\hat{\boldsymbol{\varphi}}$ is obtained [2] using (28) for each of the components. Let us denote by $\hat{\boldsymbol{\varphi}}^{\mathrm{model}}(\boldsymbol{z}_t)$ the finite dimensional 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

$$\min_{\boldsymbol{w},b,e_i} \quad \frac{1}{2}\boldsymbol{w}^T\boldsymbol{w} + \gamma\frac{1}{2}\sum_{i=1}^{N} e_i^2 \tag{31}$$

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

  With the explicit expression for $\hat{\boldsymbol{\varphi}}^{\mathrm{narx}}(\boldsymbol{z}_i)$, the model is solved in primal space by eliminating e_i from (31),

$$\min_{\boldsymbol{w},b} \quad \frac{1}{2}\boldsymbol{w}^T\boldsymbol{w} + \gamma\frac{1}{2}\sum_{i=1}^{N}(y_i - \boldsymbol{w}^T\hat{\boldsymbol{\varphi}}^{\mathrm{narx}}(\boldsymbol{z}_i) - b)^2. \tag{32}$$

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

$$\min_{\boldsymbol{w},b,r_i} \quad \frac{1}{2}\boldsymbol{w}^T\boldsymbol{w} + \gamma\frac{1}{2}\sum_{i=\tau+1}^{N} r_i^2 \tag{33}$$

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

where the autocorrelation structure for the nonlinear function is embedded into the evaluation of $\hat{\boldsymbol{\varphi}}^{\text{ar-narx}}(\boldsymbol{z}_i)$ computed from the kernel matrix $\boldsymbol{\Omega}^{\text{eq}}$. With the explicit expression for $\hat{\boldsymbol{\varphi}}^{\text{ar-narx}}(\boldsymbol{z}_i)$, the model is solved in primal space by eliminating r_i from (33),

$$\min_{\boldsymbol{w},b} \quad \frac{1}{2}\boldsymbol{w}^T\boldsymbol{w} + \gamma\frac{1}{2}\sum_{i=1}^{N}(y_i - \rho y_{i-\tau} - \boldsymbol{w}^T\hat{\boldsymbol{\varphi}}^{\text{ar-narx}}(\boldsymbol{z}_i) - b)^2. \tag{34}$$

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

$$\min_{\boldsymbol{w},b,e_i,\boldsymbol{\beta}} \quad \frac{1}{2}\boldsymbol{w}^T\boldsymbol{w} + \gamma\frac{1}{2}\sum_{i=1}^{N} e_i^2 \tag{35}$$

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

With the explicit expression for $\hat{\boldsymbol{\varphi}}^{\text{pl-narx}}(\boldsymbol{z}_{B,i})$, the model is solved in primal space by eliminating e_i from (35),

$$\min_{\boldsymbol{w},b,\boldsymbol{\beta}} \quad \frac{1}{2}\boldsymbol{w}^T\boldsymbol{w} + \gamma\frac{1}{2}\sum_{i=1}^{N}(y_i - \boldsymbol{\beta}^T\boldsymbol{z}_{A,i} - \boldsymbol{w}^T\hat{\boldsymbol{\varphi}}^{\text{pl-narx}}(\boldsymbol{z}_{B,i}) - b)^2. \tag{36}$$

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

$$\min_{\boldsymbol{w},b,r_i,\boldsymbol{\beta}} \quad \frac{1}{2}\boldsymbol{w}^T\boldsymbol{w} + \gamma\frac{1}{2}\sum_{i=\tau+1}^{N} r_i^2 \tag{37}$$

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

for $i = \tau+1,\ldots,N$. With the explicit expression for $\hat{\boldsymbol{\varphi}}^{\text{pl-ar-narx}}(\boldsymbol{z}_{B,i})$, the model is solved in primal space by eliminating r_i from (37),

$$\min_{\boldsymbol{w},b,\boldsymbol{\beta}} \quad \frac{1}{2}\boldsymbol{w}^T\boldsymbol{w} + \gamma\frac{1}{2}\sum_{i=\tau+1}^{N}(y_i - \rho y_{i-\tau} - \boldsymbol{\beta}^T(\boldsymbol{z}_{A,i} - \rho\boldsymbol{z}_{A,i-\tau}) - \boldsymbol{w}^T\hat{\boldsymbol{\varphi}}^{\text{pl-ar-narx}}(\boldsymbol{z}_{B,i}) - b)^2. \tag{38}$$

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

Short-Term Load Forecasting (STLF) is an important area of quantitative research. In order to deal with the everyday process of planning, scheduling and unit-commitment [30], the need for accurate short-term forecasts has led to the development of a wide range of models based on different techniques. In recent literature, some interesting examples are related to traditional time series analysis [31–33], and neural networks applications [34–38]. 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 several days ahead [30]. In the literature there is a broad consensus concerning important explanatory variables: past values of the load, weather information, calendar information, etc. It has also been shown in the literature that it is possible to incorporate a priori information concerning the seasonalities at several levels (daily, weekly, yearly, etc.) of linear models by appropriately choosing the model structure and the estimation method. Examples are stochastic seasonality (Box-Jenkins seasonal ARIMA models [39]), seasonal unit roots [16], nonparametric models with seasonal components [40]; or periodic autoregressions [41,17]. Our objective is to show an exploratory data modeling methodology, where prior knowledge about the problem can be incorporated at different levels, obtaining a nonlinear model structure based on LS-SVMs which is 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 HV-LV substation in Belgium, thus containing more than 26,000 datapoints. A first linear regression containing only a linear trend is estimated for this time series, to remove any growth trend present in the sample. Finally, the series is normalized using the mean and standard deviation of the observed values, in order to scale it to zero-mean and unit variance. We define our training set to consist of 18,000 datapoints, and we validate the final models using different out of sample test sets of 30 days each. 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-

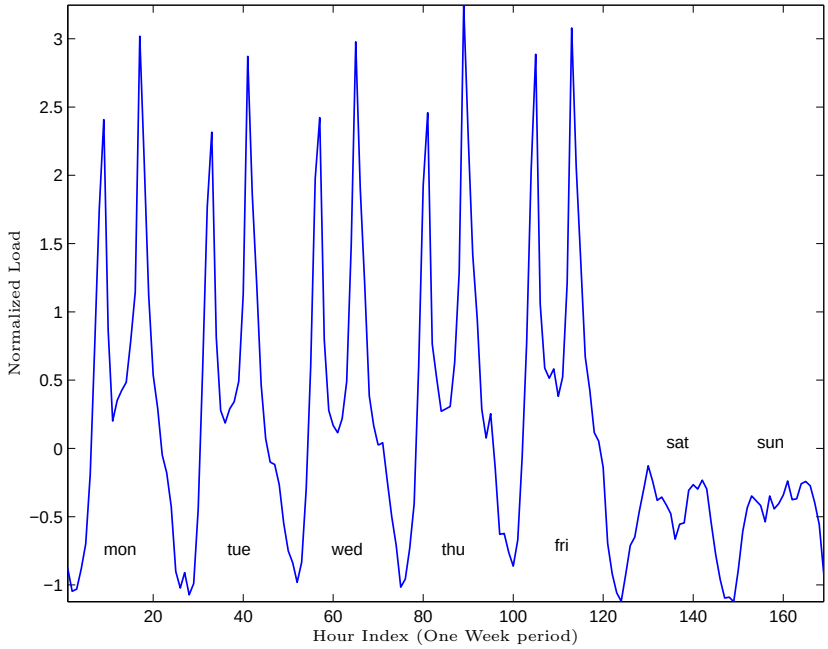

Fig. 1. *Example of the Load series within a week. Daily cycles are visible, as well as the weekend effects. Also visible are the intra-day phenomena, such as the peaks (morning, noon, evening) and the night hours.*

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 [25]) and calendar information in the form of dummy variables for month of the year, day of the week and hour of the day [17]. 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 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 Assessment

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 $\mathcal{S}$, the Mean Squared Error (MSE) is defined as

$$\text{MSE} = \frac{1}{N_s} \sum_{i \in \mathcal{S}} (y_i - \hat{y}_i)^2, \tag{39}$$

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

$$\text{MAPE} = \frac{1}{N_s} \sum_{i \in \mathcal{S}} \frac{|y_i - \hat{y}_i|}{y_i}. \tag{40}$$

The MSE is typically used within the general context of applied modeling. In this paper the MAPE is also used because it is common practice in the particular context of the STLF. 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. For a given kernel function (built with σ and ρ) and a given regularization parameter γ, the training sample is divided in 10 parts, 9 of which are used for model estimation. The performance of the estimated model is assessed in the remaining data part (which is used as a test set). By repeating the process over the 10 parts, the cross-validation performance of the model is the average of the 10 individual performances. The whole process is repeated for different hyperparameters.

By using cross-validation, it is also 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. 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 hours of the daily peak loads. The set $\mathcal{S}$ in (40) thus only contains the index of the hours where the peak of each day occurs. 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 then be tuned using both MSE-CV and MAPE-CV.

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 3% of the available training 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 [18,19].

5.6 Final Assessment on Test Data

Once the hyperparameters are selected, then the final models are used to compute predictions for the test sample (which is never part of the training set), where its performance is measured by both the MSE and MAPE. The test sample starts after the last datapoint of the training sample. Usually models for short-term load forecasting may be used for prediction during a certain number of days, after which the models are re-estimated again with the new available information. Typically a model might be re-estimated after no more than one week, in some cases in a matter of a few days. Rarely a model will remain for more than 7 days without re-estimation. In this context, for the purposes of evaluation of the models in this study, 50 different non-overlapping weeks are used to assess the performance of the model. Denote by d the last day contained in the training dataset. The first test period is the week occurring immediately after the training data, covering from $d+1$ until $d+7$. The second test week begins on $d+8$, finishing on $d+14$, the third period begins on $d+15$ until $d+21$, and so on, until the last testing period which begins at $d+344$ and ends on $d+350$. The models performance are assessed by using MSE on each of the test periods, after which the mean and standard deviation are reported for each case.

6 Results

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

6.1 (N)-ARX Models

First, a linear ARX model (3) is estimated with Ordinary Least Squares using $\boldsymbol{z}_t$ with the input variables defined above. The mean performance of the linear model on the test sets is 0.047 (MSE) and 0.050 (MAPE). This model is simple to estimate, and it has a high degree of interpretability in the sense that each one of the explanatory variables obtains an estimated coefficient which reflects its importance for predicting the load. However, as it is typically the case, the linear functional form neglects nonlinear effects which may be relevant. A second model is estimated, this time a full black-box NARX model ('NARX' on Table 1). The hyperparameters for this second model are tuned by minimizing the MSE-CV or the MAPE-CV. For the case of MSE-CV minimization, the mean test sets results are 0.0201 (MSE) and 0.0301 (MAPE). For the case of MAPE-CV minimization, the results are 0.0202 (MSE) and 0.0310 (MAPE). It is already clear, by comparing the mean test set performances, that the NARX models shows a better performance in terms of prediction accuracy than the linear ARX model. The average MSE is reduced to less than half. This means that there are revelant nonlinear effects that are neglected by the linear structure, but they are captured in the full black-box NARX model. However, it is interesting to check whether these effects are produced only by a subset of the input variables, or if they are due to the correlation structure. So far, it is clear that the performance of the NARX model does not substantially change when its hyperparameters are selected either by minimization of the MSE-CV or the minimization of the MAPE-CV. The quality of the ARX and NARX predictions for the first test set (immediately after the training 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.

6.2 AR-(N)ARX

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, a negative correlation on the residuals of the NARX model can be captured using the AR-NARX parameterization. The estimated AR-NARX model, however, does not show a significant improvement over the NARX model. The AR-NARX mean performance on the test sets is 0.020 (MSE) and 0.031 (MAPE).

For comparison, the linear ARX is extended to contain the same residual

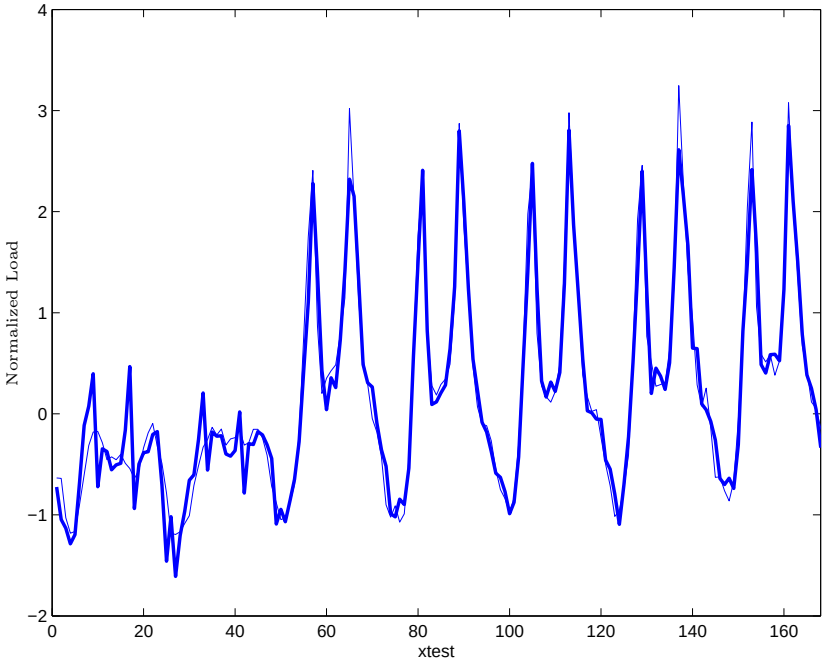

Fig. 2. *Performance of the linear model over the first test set. The observed load (full line) is compared to the prediction from the linear ARX (thick line) model.*

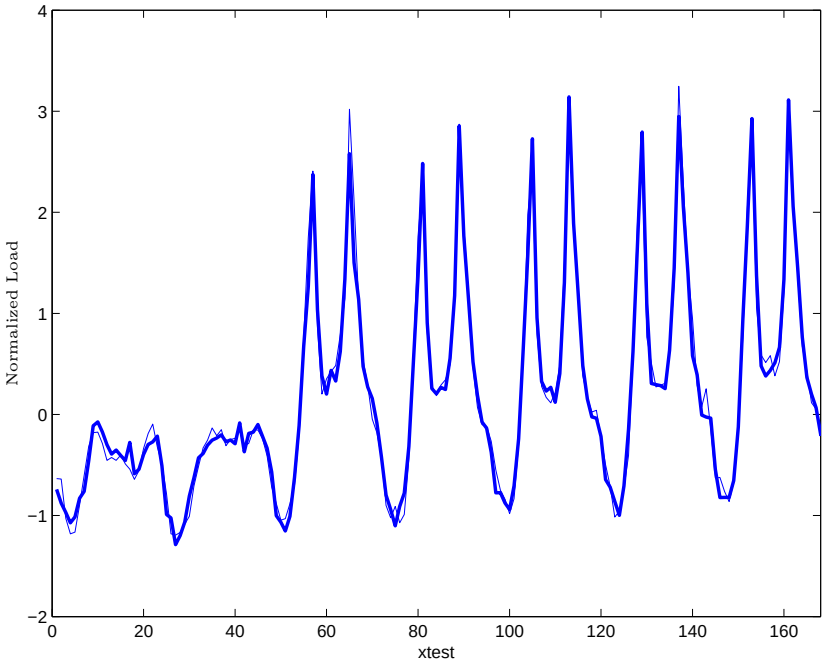

Fig. 3. *Performance of the NARX model over the first test set. The observed load (full line) is compared to the prediction from the NARX (thick line) model.*

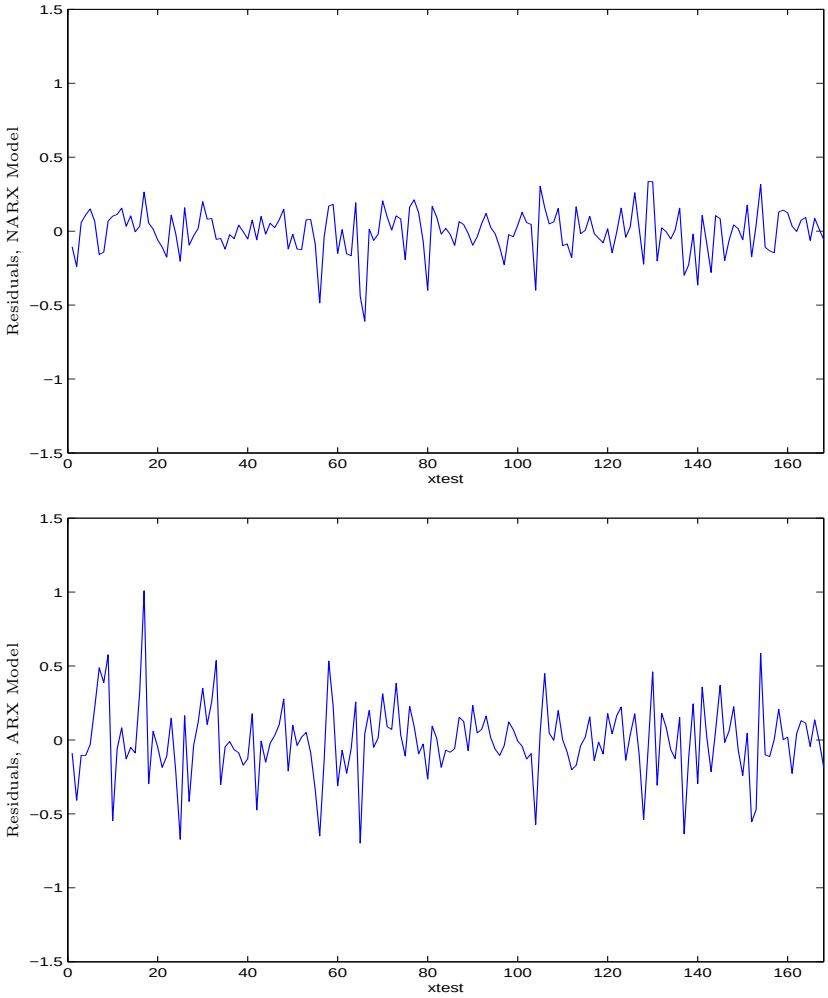

Fig. 4. *Residuals on the Test Set, for the NARX model (Top panel) and the ARX model (Bottom panel) on the first test set.*

autocorrelation process model with $\tau = 24$. The AR-ARX model is estimated using the Cochrane-Orcutt procedure for correlated residuals [42,43]. This is a classical econometric technique which iterates in order to obtain the optimal autocorrelation coefficient ρ, by exploiting the linear property of the model structure. This method gives an estimate of the correlation to be positive ($\hat{\rho} = 0.15$), and the mean performance on the test sets is given by 0.0365 (MSE) and 0.0451 (MAPE). This is an improvement from the previous ARX linear model without correlated residuals, but still it is above the levels reached by the nonlinear counterparts.

6.3 PL-NARX

Now, we turn into the partially linear parameterizations. Several linear models have been proposed in the load forecasting literature [31,32] with good results,

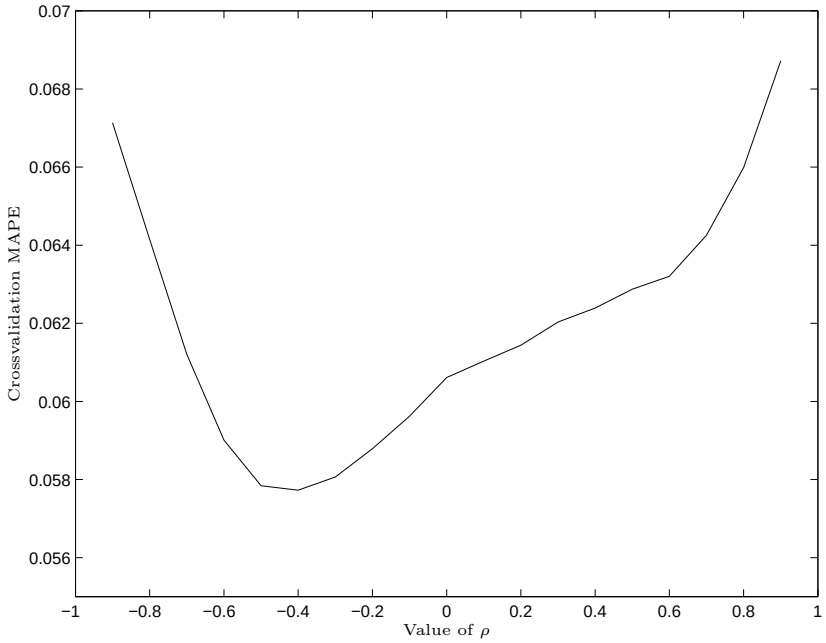

Fig. 5.
MAPE-CV performance of the AR-NARX model for different values of ρ

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 [25], 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 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 mean performance of PL-NARX-1 is 0.0221 (MSE) and 0.0327 (MAPE) on the test sets. The mean performance of PL-NARX-2 is 0.0269 (MSE) and 0.0370 (MAPE). Under this tuning scheme, the model with linear terms on the calendar and temperature information (PL-NARX-2) shows a lower performance than the model PL-NARX-1, which, in turn, shows a result closer to the NARX model.

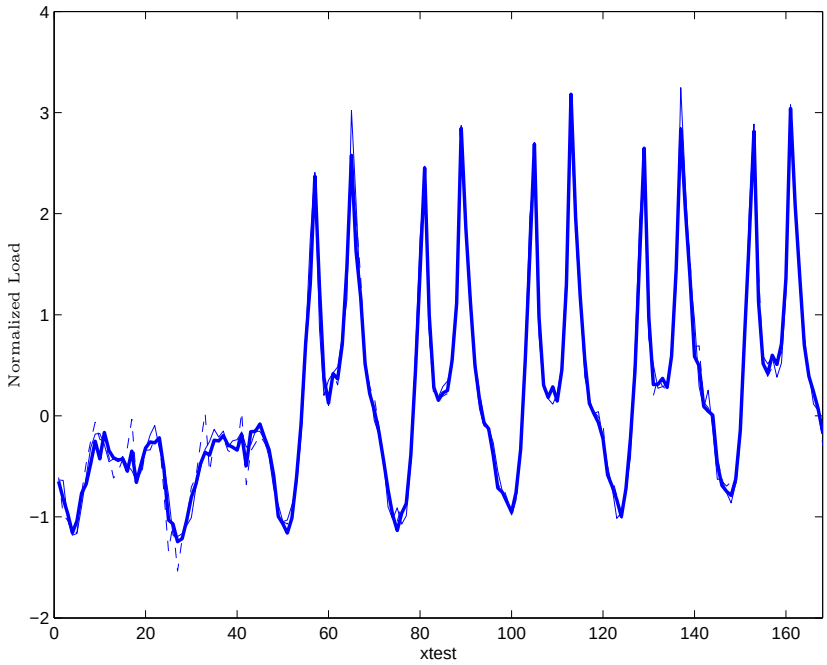

Fig. 6. *Improvement of the PL-NARX-2 model. The observed load (full line) is compared to the prediction from the PL-NARX-2 model tuned with MSE-CV (dashed) and the prediction from the PL-NARX-2 model tuned with MAPE-CV (thick).*

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 0.0221 (MSE) and 0.0326 (MAPE), whereas PL-NARX-2 improves to 0.0217 (MSE) and 0.0322 (MAPE). This last result is much closer to the performance of the AR-NARX model. For this model parameterization, this example shows that moving from the MSE-CV to the MAPE-CV for hyperparameter selection can improve the performance. This could be expected, given the shape of the load series to be forecasted. Errors can occur at the peak hours of the day; hence by selecting the parameters in such a way that the percentage errors for the daily peaks are minimized, then the overall forecasting performance improves.

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), shown for the case of the first test set. 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.

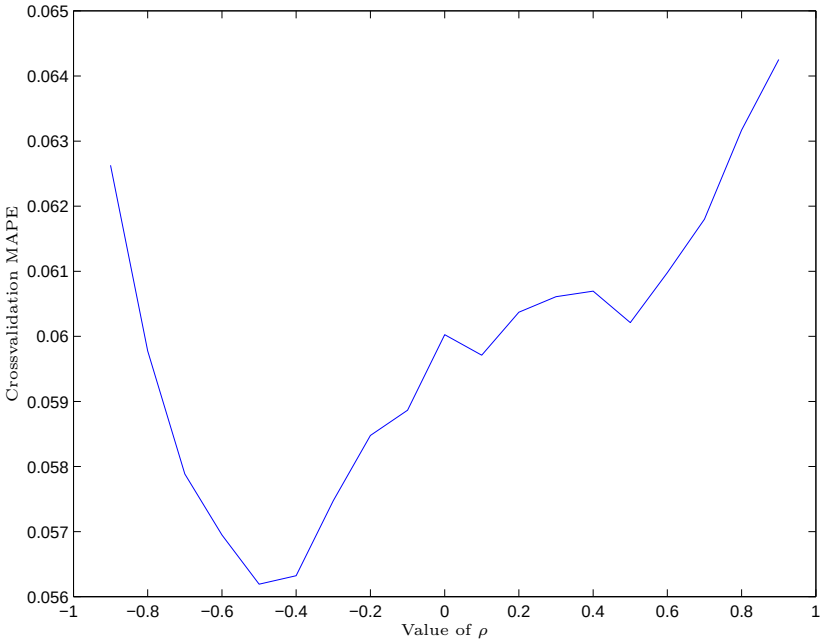

Fig. 7. *MAPE CV performance of the PL-NARX-2 model for different values of* ρ

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 mean performance on the test sets of 0.0221 (MSE) and 0.0326 (MAPE). There is no improvement over the PL-NARX-1 model.

For the case of the PL-AR-NARX-2 model (which uses $\boldsymbol{z}_{A,t} = [y_{t-1}; y_{t-2}, \ldots, y_{t-48}]$, $\boldsymbol{z}_{B,t} = \boldsymbol{u}_t$), the optimal ρ^* is found at -0.5, as shown on Figure 7. For this model, the mean performance on the test set is now 0.0204 (MSE) and 0.0309 (MAPE). The PL-AR-NARX-2 model is therefore better improves over the PL-NARX-2 results. The results obtained are comparable to those from the NARX model.

Table 3
Results on Test Sets. Mean and standard deviation of the performance of each model for the 50 different test sets.

Model	Tuning with MSE-CV		Tuning with MAPE-CV			
	No Noise Model		No Noise Model		AR Noise Model	
	MSE	MAPE	MSE	MAPE	MSE	MAPE
ARX	0.0470 (0.0028)	0.0500 (0.0008)	0.0470 (0.0028)	0.0500 (0.0008)	0.0365 (0.0030)	0.0451 (0.0008)
NARX	0.0201 (0.0012)	0.0307 (0.0014)	0.0202 (0.0011)	0.0310 (0.0012)	0.0207 (0.0012)	0.0314 (0.0013)
PL-NARX-1	0.0221 (0.0012)	0.0327 (0.0015)	0.0221 (0.0013)	0.0326 (0.0016)	0.0221 (0.0016)	0.0326 (0.0021)
PL-NARX-2	0.0269 (0.0013)	0.0370 (0.0005)	0.0217 (0.0017)	0.0322 (0.0013)	0.0204 (0.0016)	0.0309 (0.0013)

6.5 Summary of Results

The results of the exploratory data modeling are summarized on Table 3. The mean over the 50 different test sets is reported, and the standard deviation is given in brackets. 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 the performance of the full black-box NARX model. The PL-NARX-1 models show a performance which is also comparable to the full-black box NARX model. The advantage of using PL-NARX parameterizations is that they provide a set of linear coefficients which can be directly interpretable for the variables of interest. The results are also visualized on the boxplots depicted on Figure 8 for the case of the MSE and the MAPE performances.

7 Conclusion

In this paper we have used an exploratory modeling methodology for a real-life case study in short-term load forecasting. 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

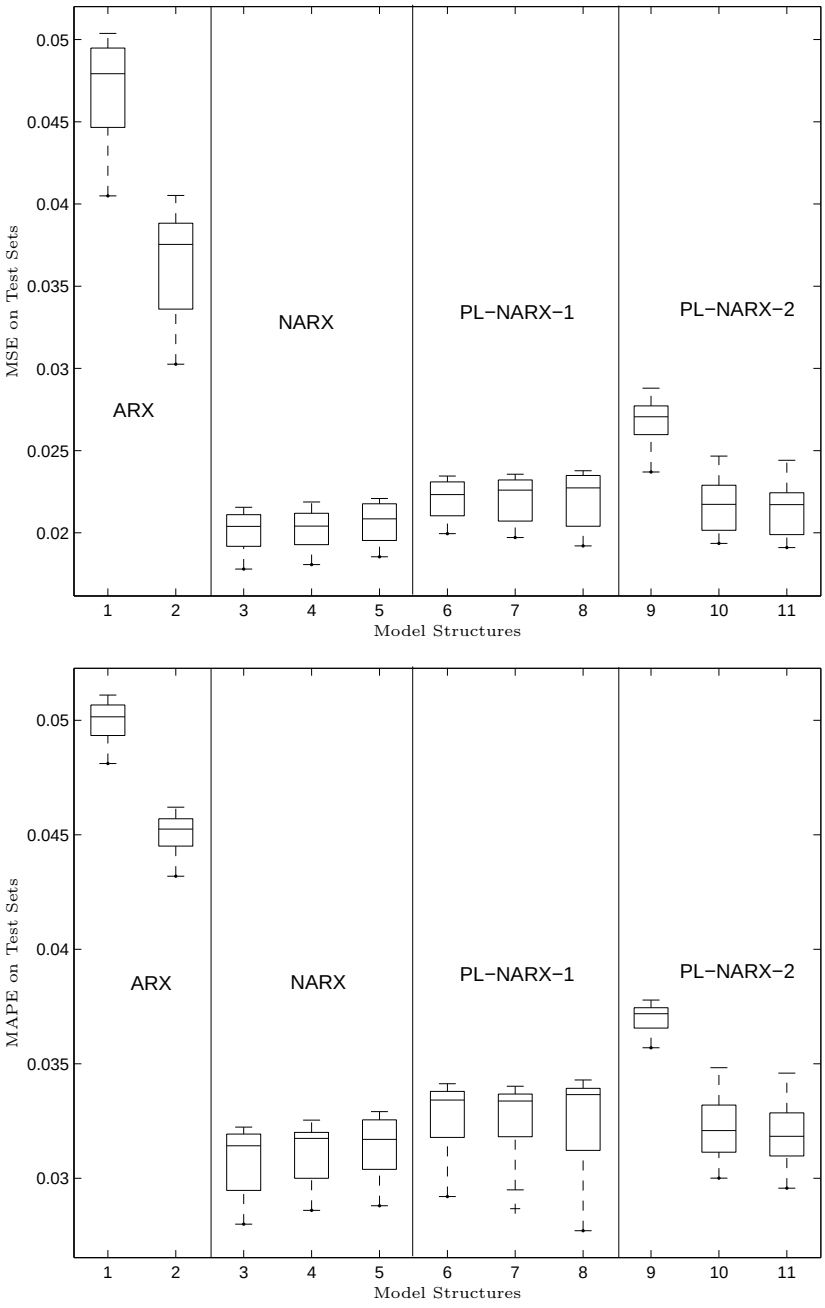

Fig. 8. *Boxplots for the different model performances over the 50 test sets for the MSE (Top) and MAPE (Bottom). The linear models (ARX and AR-ARX) are compared to the NARX, PL-NARX-1, and PL-NARX-2 models. For each of the nonlinear model types, the first box shows the performance of the model where the hyperparameters are tuned with MSE-CV; the second, when they are tuned with MAPE-CV; the third box shows the inclusion of the AR noise model to each structure.*

with autocorrelated residuals. In all these cases, the solution of each model is characterized by a set of linear equations 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 comparable to the full black-box NARX model, and yet it retains a linear part which improves its interpretability with respect to the fully black-box model.

Acknowledgments. Work supported by grants and projects for the Research Council K.U.Leuven (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] J. Suykens, T. Van Gestel, J. De Brabanter, B. De Moor, J. Vandewalle, Least Squares Support Vector Machines, World Scientific, Singapore, 2002.

[3] T. Poggio, F. Girosi, Networks for approximation and learning, Proceedings of the IEEE 78 (1990) 1481–1497.

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

[5] C. Williams, Prediction with gaussian processes: from linear regression to linear prediction and beyond, in: M. Jordan (Ed.), Learning and Inference in Graphical Models, The MIT Press, Cambridge, MA, 1999, pp. 599 – 621.

[6] N. Cristianini, J. Shawe-Taylor, An Introduction to Support Vector Machines, Cambridge University Press, 2000.

[7] D. MacKay, Bayesian interpolation, Neural Computation 4 (1992) 415–447.

[8] G. Wahba, Support vector machines, reproducing kernel hilbert spaces, and randomized gacv, chapter 6, in: B. Schölkopf, C. Burges, A. Smola (Eds.), Advances in Kernel Methods - Support Vector Learning, MIT Press, 1998, pp. 69–87.

[9] C. Williams, Prediction with Gaussian processes: from linear regression to linear prediction and beyond, in: M. Jordan (Ed.), Learning and Inference in Graphical Models, Kluwer Academic Press, 1998, pp. 599–621.

[10] D. MacKay, Introduction to gaussian processes, in: C. Bishop (Ed.), Neural Networks and Machine Learning, Vol. 168 of Springer NATO-ASI Series F, 1998, pp. 133–165.

[11] J. Suykens, T. Van Gestel, J. Vandewalle, B. De Moor, A support vector machine formulation to PCA analysis and its kernel version, IEEE Transactions on Neural Networks 14 (2) (2003) 447–450.

[12] J. Suykens, J. Vandewalle, Least squares support vector machine classifiers, Neural Processing Letters 9 (1999) 293–300.

[13] T. Van Gestel, J. Suykens, G. Lanckriet, A. Lambrechts, B. De Moor, J. Vandewalle, A Bayesian framework for least squares support vector machine classifiers, Gaussian processes and kernel Fisher discriminant analysis, Neural Computation 14 (2002) 1115–1147.

[14] T. Van Gestel, J. Suykens, J. De Brabanter, B. De Moor, J. Vandewalle, Kernel canonical correlation analysis and least squares support vector machines, in: G. Dorffner, H. Bischof, K. Hornik (Eds.), Proceedings International Conference on Artificial Neural Networks (ICANN 2001), Vol. 2130 of Lecture Notes in Computer Science, Springer, Vienna, Austria, 2001, pp. 381–386.

[15] I. Goethals, K. Pelckmans, J. Suykens, B. De Moor, Subspace identification of hammerstein systems using least squares support vector machines, IEEE Transactions on Automatic Control, Special Issue: Linear vs. Nonlinear 50 (10) (2005) 1509–1519.

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

[17] 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.

[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] M. Espinoza, J. Suykens, B. De Moor, Fixed-size least squares support vector machines : A large scale application in electrical load forecasting, Computational Management Science (Special Issue on Support Vector Machines) 3 (2) (2006) 113–129.

[20] 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.

[21] 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.

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

[23] M. Espinoza, J. Suykens, B. De Moor, LS-SVM regression with autocorrelated errors, in: Proc. of the 14th IFAC Symposium on System Identification (SYSID), 2006, pp. 582–587.

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

[25] 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.

[26] M. Espinoza, J. Suykens, B. De Moor, Kernel based partially linear models and nonlinear identification, IEEE Transactions on Automatic Control, Special Issue: Linear vs. Nonlinear 50 (10) (2005) 1602–1606.

[27] C. Saunders, A. Gammerman, V. Vovk, Ridge regression learning algorithm in dual variables, in: Proceedings 15th International Conference on Machine Learning (ICML'98), Morgan Kaufmann, San Francisco, California, 1998, pp. 515–521.

[28] 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, pp. 682–688.

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

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

[31] 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.

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

[33] S.-J. Huang, K.-R. Shih, Short term load forecasting via ARMA model identification including non-gaussian process considerations, IEEE Transactions on Power Systems 18 (2) (2003) 673–679.

[34] 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.

[35] A. Khotanzad, R. Afkhami-Rohani, D. Maratukulam, ANNSTLF-artificial neural network short-term load forecaster-generation three, IEEE Transactions on Power Systems 13 (4) (1998) 1413–1422.

[36] K.-H. Kim, H.-S. Youn, Y.-C. Kang, Short-term load forecasting for special days in anomalous load conditions using neural networks and fuzzy inference method, IEEE Transactions on Power Systems 15 (2) (2000) 559–565.

[37] L. Mohan Saini, M. Kumar Soni, Artificial neural network-based peak load forecasting using conjugate gradient methods, IEEE Transactions on Power Systems 17 (3) (2002) 907–912.

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

[39] G. Box, G. Jenkins, Time Series Analysis, Forecasting and Control, Holden-Day, San Francisco, 1970.

[40] L. Yang, R. Tschernig, Non- and semiparametric identification of seasonal nonlinear autoregression models, Econometric Theory 18 (2002) 1408–1448.

[41] P. Franses, R. Paap, Periodic Time Series Models, Oxford University Press, 2003.

[42] J. Johnston, Econometric Methods, Economics Series, McGraw-Hill, 1991.

[43] D. Cochrane, G. Orcutt, Application of least-squares regressions to relationships containing autocorrelated error terms, Journal of the American Statistical Association 44 (1949) 32–61.