Time Series Prediction using LS-SVMs

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

{marcelo.espinoza,tillmann.falck,johan.suykens}@esat.kuleuven.be
K.U.Leuven, ESAT - SCD
Kasteelpark Arenberg 10, B-3001 Leuven (Heverlee), Belgium

Abstract. This paper describes the use of LS-SVMs as an estimation technique in the context of the time series prediction competition of ESTSP 2008 (Finland). Given three different time series, a model is estimated for each series, and subsequent simulations of several points after the last available sample are produced. For the first series, a NARX model is formulated after a careful selection of the relevant lags of inputs and outputs. The second and third series show cyclical or seasonal patterns. Series 2 is modelled by adding deterministic "calendar" variables into the nonlinear regression. Series 3 is first cleaned from the seasonal patterns, and a NAR model is estimated using LS-SVM on the deseasonalized series. In all cases, hyperparameters selection and input selection are made on a cross-validation basis.

1 Introduction

Time series prediction can be treated as a special case of system identification [1]. In nonlinear system identification [2, 3] it is common to use past lags of the output variable $y \in \mathbb{R}$ and, if available, of exogenous input variables $\boldsymbol{u} \in \mathbb{R}^d$ to build a NARX model of the form

$$y_t = f(y_{t-1}, y_{t-2}, \ldots, y_{t-p}, \boldsymbol{u}_{t-1}, \boldsymbol{u}_{t-2}, \ldots, \boldsymbol{u}_{t-q}) + e_t, \tag{1}$$

where e_t is assumed to be a white noise process. The estimated model can then be used for prediction or simulation. Nonlinear effects can be identified when the function f is parameterized as a nonlinear function. In this article, we use Least-Squares Support Vector Machines (LS-SVMs) [4] as a tool for estimating f in NARX models in order to produce the required simulations for three time series available in the context of the time series competition of the Second European Symposium of Time Series Prediction (ESTSP 2008, Porvoo, Finland, http://www.estsp.org/).

LS-SVMs belong to the class of kernel methods [4, 5, 6], which use positive-definite kernel functions to build a nonlinear representation of the original inputs in a high-dimensional feature space. An optimization problem consisting of a

*The work presented was funded by Research Council KUL: GOA AMBioRICS, CoE EF/05/006, IOF-SCORES4CHEM, several PhD/postdoc & fellow grants; FWO: PhD/postdoc grants, projects G.0452.04, G.0499.04, G.0211.05, G.0226.06, G.0321.06, G.0302.07, G.0320.08, G.0558.08, G.0557.08, research communities (ICCoS, ANMMM, MLDM); IWT: PhD Grants, McKnow-E, Eureka-Flite+, Helmholtz: viCERP, IUAP P6/04; EU: ERNSI;Contract Research: AMINAL. Johan Suykens is a professor and Bart De Moor is a full professor at the Katholieke Universiteit Leuven, Belgium.

regularized least-squares cost function and equality constraints is formulated in primal space and solved in the dual variables. This formulation is flexible in the sense that it allows the incorporation of different elements of knowledge about the problem at hand. In the primal domain the model is parametric and it is easy to incorporate some types of prior knowledge that gets automatically embedded in the dual representation. Examples for this are the inclusion of partially linear models [7], autocorrelated residuals [8, 9] and monotonicity [10].

The remainder of the paper is structured as follows. A general overview of LS-SVMs and a discussion about practical elements are given in Section 2. The study cases for time series of the competition are discussed in Sections 3 and 4.

2 General Methodology

2.1 Least Squares Support Vector Machine Regression

LS-SVMs belong to the class of kernel methods, which use positive-definite kernel functions to build a nonlinear representation of the original inputs in a high-dimensional feature space. We start by parameterizing NARX model (1) as

$$y_t = \boldsymbol{w}^T \boldsymbol{\varphi}(\boldsymbol{x}_t) + b + e_t \tag{2}$$

where $y_t \in \mathbb{R}$, $\boldsymbol{x}_t \in \mathbb{R}^n$ is the regression vector $[y_{t-1}, y_{t-2}, \ldots, y_{t-p}, \boldsymbol{u}_{t-1}, \boldsymbol{u}_{t-2}, \ldots, \boldsymbol{u}_{t-q}]$, $b \in \mathbb{R}$ is a bias term, $\boldsymbol{w} \in \mathbb{R}^{n_h}$ is an unknown coefficient vector, and $\boldsymbol{\varphi} : \mathbb{R}^n \rightarrow \mathbb{R}^{n_h}$ is a nonlinear feature map, which transforms the original input $\boldsymbol{x}_t \in \mathbb{R}^n$ to a high-dimensional vector $\boldsymbol{\varphi}(\boldsymbol{x}_t) \in \mathbb{R}^{n_h}$, which can be infinite dimensional [6]. Consider the following constrained optimization problem with a regularized cost function:

$$\min_{\boldsymbol{w},b,e_t} \quad \frac{1}{2}\boldsymbol{w}^T\boldsymbol{w} + \gamma\frac{1}{2}\sum_{t=1}^{N} e_t^2 \tag{3}$$

$$\text{s.t.} \quad y_t = \boldsymbol{w}^T\boldsymbol{\varphi}(\boldsymbol{x}_t) + b + e_t, \quad t = 1, \ldots, N,$$

where γ is a regularization constant and K is a p.d. kernel function. With the application of the Mercer's theorem for the kernel matrix $\boldsymbol{\Omega}$ as $\boldsymbol{\Omega}_{ij} = K(\boldsymbol{x}_i, \boldsymbol{x}_j) = \boldsymbol{\varphi}(\boldsymbol{x}_i)^T\boldsymbol{\varphi}(\boldsymbol{x}_j), i, j = 1, \ldots, N$ it is not required to compute explicitly the nonlinear mapping $\boldsymbol{\varphi}(\cdot)$ as this is done implicitly through the use of positive definite kernel functions K. For $K(\boldsymbol{x}_i, \boldsymbol{x}_j)$ there are usually the following choices: $K(\boldsymbol{x}_i, \boldsymbol{x}_j) = \boldsymbol{x}_i^T\boldsymbol{x}_j$ (linear kernel); $K(\boldsymbol{x}_i, \boldsymbol{x}_j) = (\boldsymbol{x}_i^T\boldsymbol{x}_j + c)^d$ (polynomial of degree d, with $c \geq 0$ a tuning parameter); $K(\boldsymbol{x}_i, \boldsymbol{x}_j) = \exp(-||\boldsymbol{x}_i - \boldsymbol{x}_j||_2^2/\sigma^2)$ (radial basis function, RBF), where σ is a tuning parameter.

The problem (3) is solved using Lagrange multipliers and the solution is expressed in dual form [4]. The final expression for the estimated f is given by

$$\hat{f}(\boldsymbol{x}) = \sum_{t=1}^{N} \alpha_t K(\boldsymbol{x}_t, \boldsymbol{x}) + b. \tag{4}$$

The one-step-ahead prediction is simply $\hat{y}_{N+1} = \hat{f}(\boldsymbol{x}_{N+1})$ using the estimated $\hat{f}$; simulation $n-$steps ahead can be obtained by iteratively applying the prediction equation replacing future outputs by its predictions [1, 2].

2.2 Practical Implementation

The training process of LS-SVM involves the selection of kernel parameters and the regularization constant γ. A good choice of these parameters is crucial for the performance of the estimator. In this paper, we use 5-fold cross-validation for selecting these parameters. The second important choice is the selection of regressors, i.e., which lags of inputs and outputs are going to be included in the regression vector $\boldsymbol{x}_t$. This selection is done by using a large number of initial components and then performing a greedy search to prune non-informative lags on a cross-validation basis. Therefore an initial model containing all regressors is estimated and optimal choices for the parameters are made. On each stage of the greedy backwards elimination process, a regressor is removed if the cross-validation Mean Squared Error (CV-MSE) improves. The final set of regressors is then used for the final predictions. For the purpose of model estimation, all series are normalized to zero mean and unit variance.

3 Analysis for Time Series 1

The available data for time series consists of a sequence $\{y_t, u_t, v_t\}_{t=1}^{N}$ for the output y and the two exogenous inputs u and v with $N = 354$ datapoints. The series are depicted in Figure 1. The goal is to produce a sequence of simulated future values $\hat{y}_{N+1}$ until $\hat{y}_{N+18}$. Based on autocorrelation analysis, it can be noticed that all three variables share some periodic behavior with a period of 12 samples.

3.1 Modeling Strategy

We test two model specifications, with and without the exogenous inputs (NAR and NARX models, respectively), to be estimated using LS-SVM, as follows:

- (NAR) $\hat{y}_t = \hat{f}(y_{t-1}, \ldots, y_{t-p})$ with $p \leq 48$
- (NARX) $\hat{y}_t = \hat{f}(y_{t-1}, \ldots, y_{k-p}, u_{t-1}, \ldots, u_{t-q_1}, v_{t-1}, \ldots, v_{t-q_2})$ with p, q_1, $q_2 \leq 48$

The performance measure used for 5-fold cross-validation is the the CV-MSE over two different prediction scenarios:

- One step ahead prediction for each test fold,
- Simulation of the time series for 18 time steps for every sample in the test fold, averaging the values at each time step over predictions (1 step, 2 steps, ..., 18 steps)

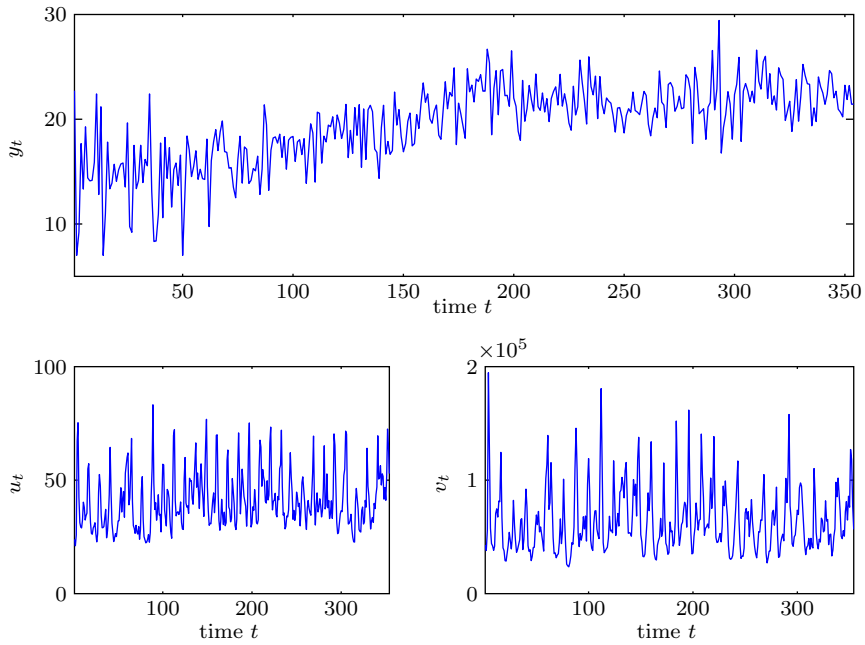


Fig. 1: Time plots of target variable y (top) and the exogenous variables u (bottom left) and v (bottom right) of time series 1

3.2 Results

Consider the NAR(48) model formulation. The selected lags to be included in the regression vector are pruned using a greedy backwards search. The CV-MSE as a function of the number of regressors included in the model is shown as the blue line in Figure 2. By actively selecting the regressors the MSE can be improved by 33%. The best NAR model uses 12 regressors and achieves a CV-MSE of 0.2834. However, further correlation analysis between the model residuals and the exogenous inputs reveal that some information has not been captured by the NAR model alone. This means we should test a NARX formulation.

Two different NARX models are evaluated. The first one includes all lags up to 48 for y, u and v, NARX(48,48,48). The second model assumes a delay between inputs and output of 18 time steps. The NARX(48,30,30) model contains the lags 19 up to 48 for the inputs u and v. Figure 2 shows the feature selection process for the models, both of which clearly outperform the NAR model. The model using all regressors achieves a CV-MSE of 0.2437 with 34 regressors, whereas the delayed model uses 25 regressors and has a CV-MSE of 0.2489. The improvement of the full over the delayed model is rather small with 2%. For the purpose of generating the results for the time series competition, the full model would require to simulate the exogenous inputs as well, which might be an additional source of errors for the simulated outputs. On the other hand,

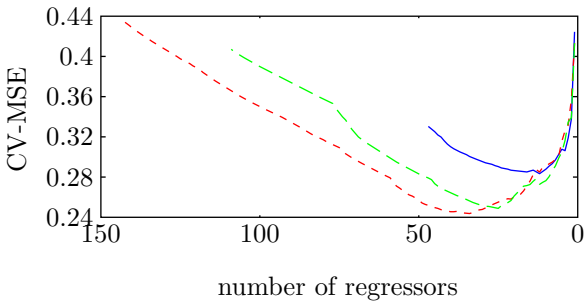


Fig. 2: Feature selection using greedy backward search. Solid blue: NAR model (lags y: 1-48), green long dashes: NARX (lags y: 1-48, u: 1-48, v: 1-48), red short dashes: NARX (lags y: 19-48, u: 19-48, v: 19-48)

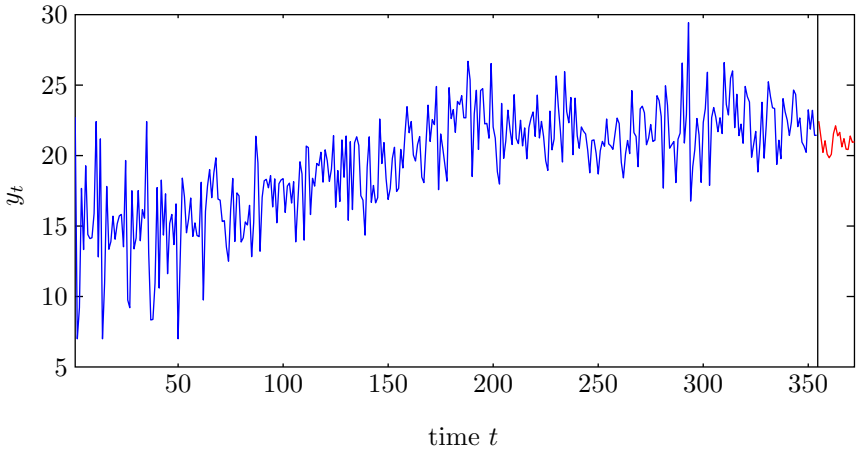


Fig. 3: Simulations for the next 18 points for Series 1, shown after the vertical line.

the delayed model does not have this problem, and we can always use actual observations to generate the simulated outputs. Considering that the difference in performance is small, we select the delayed NARX model to simulate 18 consecutive new samples. The simulated values together with the training data are shown in Figure 3.

4 Analysis for Time Series 2 and 3

Series 2 and Series 3 are modelled following a similar methodology. Series 2 (Figure 4, top) consists of 1,300 observations, and the goal is to simulate the next 100 points. Series 3 contains 31,614 samples (Figure 5, top), and the goal is to predict the next 200 points. The two series display a similar cyclical behavior. Series 2 display a strong correlation every 7 samples, as can be seen

in the autocorrelation plot (Figure 4, center). Such pattern would correspond to a "weekly" seasonal cycle for a series consisting of consecutive daily values. In the same manner, Series 3 displays cyclical patterns similar to those of "daily", "monthly" and "yearly" cycles for a series consisting of hourly values.

4.1 Modeling Strategy

The seasonal patterns detected in the series suggest to use a specific seasonal modeling strategy, following the golden-rule of "do not estimate what you already know" [2]. One of the most common approaches is the use of deterministic calendar information to keep track of the sequence of patterns involved. For the case of Series 2, we use the binary-valued vector $\boldsymbol{W}_t \in \{0,1\}^7$, which is a vector of zeros with a 1 in the position of the day of the week at time t to keep track of the day-to-day cycle. For example, Monday corresponds to $\boldsymbol{W}_t = [1, 0, \cdots, 0]$. In the same way, the variable $\boldsymbol{M}_t \in \{0,1\}^{12}$ is defined as a vector of zeros with a 1 in the position of the month at time t. A binary-valued vector $\boldsymbol{H}_t \in \{0,1\}^{24}$ is similarly defined to keep track of the hour of the day at time t. These variables are considered as exogenous inputs to the corresponding models for each series.

4.1.1 Model for Series 2

The estimated model is the following NARX formulation:

$$y_t = f(y_{t-1}, \ldots, y_{t-p}, \boldsymbol{W}_t, \boldsymbol{M}_t) + e_t \tag{5}$$

estimated using LS-SVM. The order p, the hyperparameters and the relevant regressors are determined on a 5-fold cross-validation basis using the greedy search optimization procedure described previously.

4.1.2 Model for Series 3

This series is modelled differently because of the strong presence of the seasonal patterns. Series 3 shows a very strong combination of seasonal patterns that can be considered the "backbone" of the observed series. Following a standard approach in seasonal modeling [11], we decompose the original series as the sum of a regular and an irregular component, $y_t = r_t + z_t$, where $r_t = \beta_1^T \boldsymbol{H}_t + \beta_2^T \boldsymbol{M}_t + \beta_3^T \boldsymbol{W}_t$ is the contribution of the deterministic seasonal variables. The identification of r_t and z_t is obtained by estimating the following linear regression:

$$y_t = \beta_1^T \boldsymbol{H}_t + \beta_2^T \boldsymbol{M}_t + \beta_3^T \boldsymbol{W}_t + z_t, \tag{6}$$

with $\beta_1 \in \mathbb{R}^{24}, \beta_2 \in \mathbb{R}^{12}, \beta_3 \in \mathbb{R}^7$ estimated with ordinary least-squares. The irregular component z_t corresponds to the residual of this regression. Figure 5 shows the original series (top) and the decomposition in the regular component r_t (center, left) and the irregular component z_t (center, right).

Predicting the regular component into the future is straightforward. For the prediction of the irregular component z_t, we estimate a NAR model using

LS-SVM,

$$z_t = f(z_{t-1}, \ldots, z_{t-p}) + e_t. \tag{7}$$

The irregular component z_t still displays significant autocorrelation with a 24 hours period (Figure 5 bottom left), which should be captured by the NAR model. Given that z_t is much more stationary than the original series, the LS-SVM model is estimated only using the last 1,000 observations. The order of this model, hyperparameters and relevant lags are determined as in the other models.

4.2 Results

For Series 2, the best order of the NARX model is found to be $p = 14$, which gives a total number of regressors of 33 (14 past values, 7 days of the week, 12 months in a year). The model with 33 regressors shows a CV-MSE= 0.23. From the 33 regressors, only 11 are found to be relevant, lowering the CV-MSE to 0.20. This final model with 11 regressors is used to produce the final simulations. The final simulations are shown on the bottom panel of Figure 4.

For Series 3, the best order of the NAR model on y_s is $p = 48$. The selection of relevant regressors yields no significant improvement. The autocorrelation present in z_t has been captured by the model, and the residuals e_t show no such correlation (Figure 5 bottom right). Overall, the CV-MSE obtained following this strategy is 0.004. The NAR model is used to produce the simulations for y_s, which are added to the simulations for the regular component. The final simulation requested for the competition is shown on Figure (6).

5 Conclusions

This paper described the use of LS-SVMs as an estimation technique in the context of the time series prediction competition of the Second European Symposium ESTSP 2008 (Porvoo, Finland). Given three different time series, a model is estimated for each series, and subsequent simulations of several points after the last available sample are produced. For the first series, a NARX model is formulated after a careful selection of the relevant lags of inputs and outputs. Supported by correlation analysis, it is determined that the exogenous inputs provided for this series have a significant influence on modeling of the output series.

The second and third series are modelled independently, yet following a similar methodology. Both series show seasonal variations, and we used deterministic seasonal variables as exogenous inputs. Series 2 is modelled by using a NARX formulation including the deterministic variables; Series 3 is first "deseasonalized" by using a linear regression of the series on the deterministic variables, and later a NAR model is estimated with LS-SVM on the residuals of the first regression.

The selected models over the three series are those who provided the best

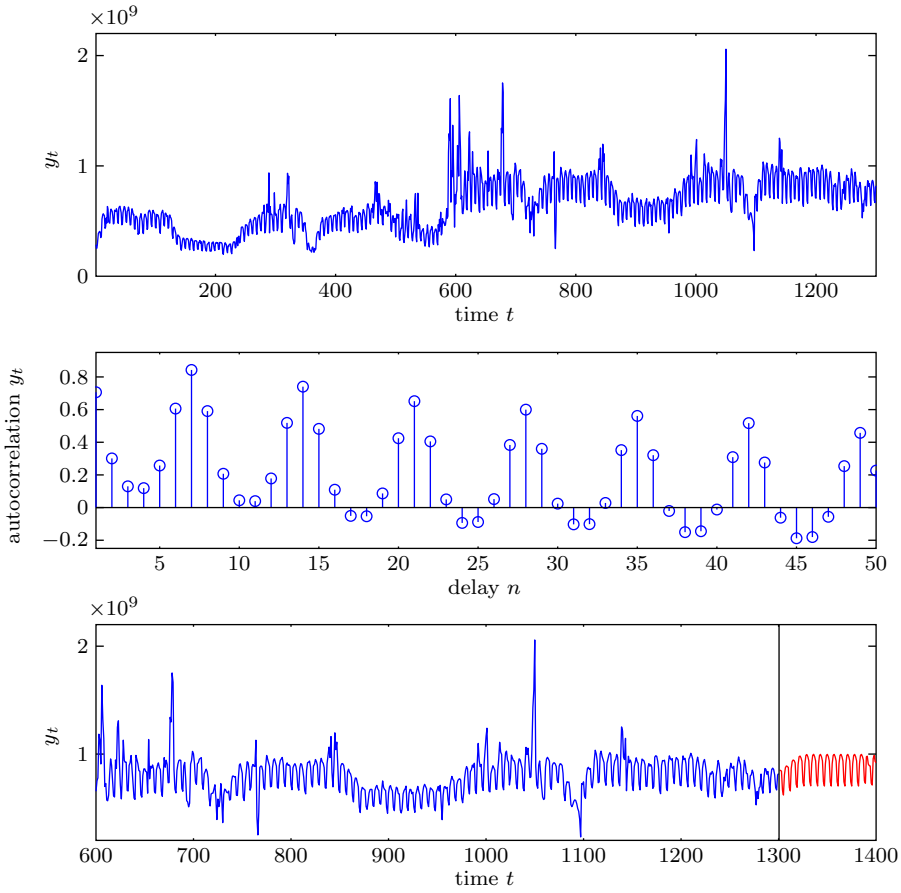


Fig. 4: The original data for Series 2 (top) shows a seasonal pattern visible in the autocorrelation plot (center) of the series. Using a NARX formulation, the requested simulations are computed for the next 100 points (bottom), shown after the vertical line.

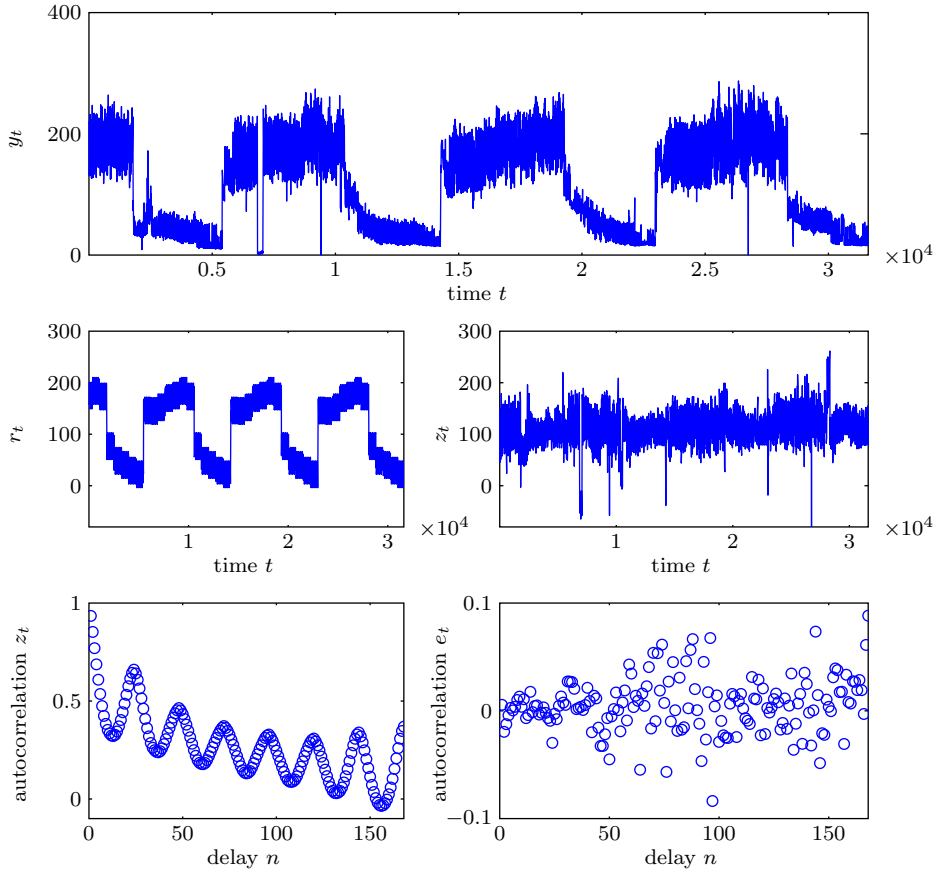


Fig. 5: The original data for Series 3 (top) can be decomposed as the sum of a regular component r_t (center, left) and an irregular component z_t (center, right). Using a NAR formulation to model the irregular component z_t, the final model captures the autocorrelation that was still present in z_t. This is visible by comparing the autocorrelation plot for z_t (Bottom, left) with that of the NAR model residuals e_t (bottom, right).

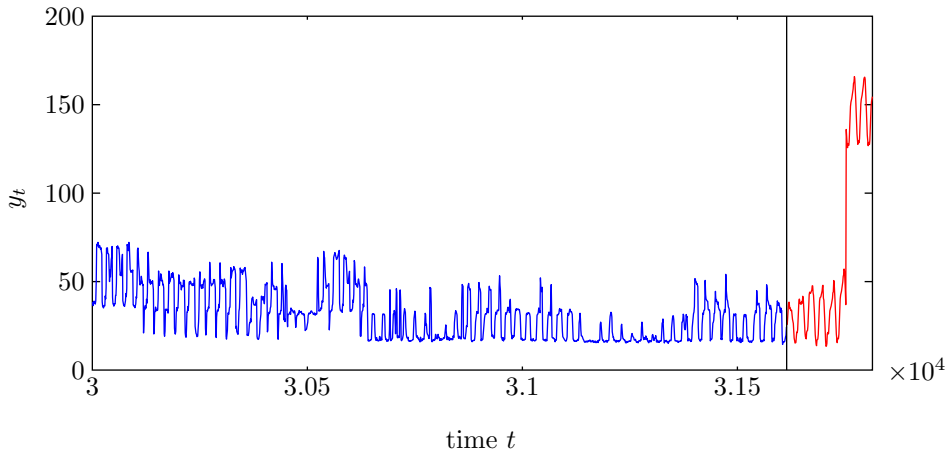


Fig. 6: Simulations for the next 200 points for Series 3, shown after the vertical line.

performance on a cross-validation basis, in terms of order selection, kernel parameters, regularization constant and relevant regressors.

References

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

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

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

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

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

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

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

[8] M. Espinoza, J.A.K. Suykens, and B. De Moor. LS-SVM regression with autocorrelated errors. In *Proc. of the 14th IFAC Symposium on System Identification (SYSID 2006)*, pages 582–587.

[9] M. Espinoza, J.A.K. Suykens, R. Belmans, and B. De Moor. Electric load forecasting: Using kernel-based modeling for nonlinear system identification. *IEEE Control Systems Magazine*, 27(5):43–57, 2007.

[10] K. Pelckmans, M. Espinoza, J. De Brabanter, J.A.K. Suykens, and B. De Moor. Primal-dual monotone kernel regression. *Neural Processing Letters*, 22(2):171–182, 2005.

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