LS-SVM REGRESSION WITH AUTOCORRELATED ERRORS

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

K.U. Leuven, ESAT-SCD-SISTA
Kasteelpark Arenberg 10, B-3001 Leuven, Belgium
{*marcelo.espinoza,johan.suykens*}*@esat.kuleuven.be*

Abstract: The problem of nonlinear AR(X) system identification with correlated residuals is addressed. Using LS-SVM regression as a nonlinear black-box technique, it is illustrated that neglecting such correlation can have negative effects on the identification stage. We show that when the correlation structure of the residuals is explicitly incorporated into the model, this information is embedded into the kernel level in the dual space solution. The solution can be obtained from a convex problem, in which the correlation coefficients are considered to be tuning parameters. The dynamical structure of the model is explored in terms of an equivalent NAR(X)-AR representation, for which the optimal one-step-ahead predictor is expressed in terms of the approximated nonlinear function and the correlation structure.

1. INTRODUCTION

This paper addresses the topic of black-box NAR(X) system identification using Least Squares Support Vector Machines (LS-SVM) (Suykens *et al.*, 2002) with correlated errors. Typically, the estimation of a model from a set of observations involves the assumption that the error terms are independently and identically distributed (Ljung, 1987; Sjöberg *et al.*, 1995; Juditsky *et al.*, 1995). Although this assumption is mostly satisfied when working with controlled experimental conditions, or with a clear knowledge of the dynamical behavior of the true underlying system, in practice it can happen otherwise (e.g. when working with observed data (Engle *et al.*, 1986)). When neglected, the presence of correlation in the error sequence can lead to severe problems not only on the identification of the function under study, but also on the future predictions of the system. Within the linear AR(X) system identification framework, the presence of correlated errors leads to the so-called ARAR(X) model structure (Ljung, 1987), which can be solved by exploiting the linearity of the model (Guidorzi, 2003). For the nonlinear/nonparametric case, it has been noted (Altman, 1990) that the presence of correlation in the error terms can mislead the identification of the nonlinear function when using a black-box identification technique. In plain terms, the black-box technique "learns" the structure in the nonlinear function together with the correlation structure in the errors. This problem can be solved by incorporating the knowledge of the correlation structure into the modelling stage. In this paper, starting from prior knowledge about the correlation structure, we show the derivations of the expressions for the case of LS-SVM regression with autocorrelated errors. We show that the solution embeds the correlation information into the kernel level for the approximation of the nonlinear function, and that the model structure leads to an optimal predictor which also incorporates the correlation structure.

This paper is organized as follows. Section 2 presents the general model formulation and its properties, with a solution for the case of AR(1) errors. Section 3 gives the expressions for the optimal predictors of the NAR(X)-AR model structure and discusses about related model representations. Section 4 shows illustrative examples where the inclusion of the prior knowledge about correlation improves substantially over the case where the correlation is neglected.

2. LS-SVM REGRESSION WITH CORRELATED ERRORS

In this section, the model formulation is derived using LS-SVM as a nonlinear system identification technique.

2.1 Model Structure

We focus on the identification of the nonlinear function f in the model with autocorrelated residuals,

$$\begin{cases} y(k) = f(\boldsymbol{x}(k)) + e(k) \\ (1 - a(z^{-1}))e(k) = r(k) \end{cases} \tag{1}$$

for $k = 2, \ldots, N$. The input vector $\boldsymbol{x}(k) \in \mathbb{R}^p$ can contain past values of the output $y(k) \in \mathbb{R}$, leading to a NAR(X) model structure. The residuals $e(k)$ of the first equation are uncorrelated with the input vector $\boldsymbol{x}(k)$, and the sequence $e(k)$ follows an invertible covariance-stationary AR(q) process described by $(1 - a(z^{-1}))e(k) = r(k)$, where $r(k)$ is a white noise sequence with zero mean and constant variance σ_u^2, and where $a(z^{-1})$ is a polynomial in the lag operator z^{-1} with unknown parameters $\rho_i, i = 1, \ldots, q$,

$$a(z^{-1})e(k) = \rho_1 e(k-1) + \rho_2 e(k-2) + \ldots + \rho_q e(k-q). \tag{2}$$

Throughout this paper we assume that we have prior knowledge about the existence and AR(q) structure of the correlation. Therefore, we do not address the problem of detecting correlation. At the same time, we consider the AR(q) parameters $\rho_i, i = 1 \ldots, q$, as tuning parameters rather than to be optimized at the training sample. As a result, the problem remains convex.

2.2 LS-SVM with correlated errors

The derivations are presented here for the case of $q = 1$, for which $a(z^{-1})e(k) = \rho e(k-1)$. It can be extended to the general AR(q) straightforwardly. The case of $q = 1$ is often use in applied work (e.g. load analysis (Engle *et al.*, 1986)). The inclusion of correlated errors to the LS-SVM regression can be formulated as follows. Given the sample of N points $\{\boldsymbol{x}(k), y(k)\}_{k=1}^N$ and the model structure (1), the following optimization problem with a regularized cost function is formulated:

$$\begin{aligned} \min_{\boldsymbol{w},b,r(k)} \quad & \frac{1}{2}\boldsymbol{w}^T\boldsymbol{w} + \gamma\frac{1}{2}\sum_{k=q+1}^{N} r(k)^2 \\ \text{s.t.} \quad & \begin{cases} y(k) = \boldsymbol{w}^T\boldsymbol{\varphi}(\boldsymbol{x}(k)) + b + e(k) \\ (1 - a(z^{-1}))e(k) = r(k) \end{cases} \end{aligned} \tag{3}$$

for $k = 2, \ldots, N$, γ is a regularization constant and the AR(1) coefficient ρ is a tuning parameter satisfying $|\rho| < 1$ (invertibility condition of the AR(1) process). The nonlinear function f from (1) has been parameterized as $f(\boldsymbol{x}(k)) = \boldsymbol{w}^T\boldsymbol{\varphi}(\boldsymbol{x}(k)) + b$, where the feature map $\boldsymbol{\varphi}(\cdot) : \mathbb{R}^p \to \mathbb{R}^{n_h}$ is the mapping to a high dimensional (and possibly infinite dimensional) feature space. By eliminating $e(k)$, the following equivalent problem is obtained:

$$\begin{aligned} \min_{\boldsymbol{w},b,r(k)} \quad & \frac{1}{2}\boldsymbol{w}^T\boldsymbol{w} + \gamma\frac{1}{2}\sum_{k=q+1}^{N} r(k)^2 \\ \text{s.t.} \quad & (1 - a(z^{-1}))y(k) = \\ & (1 - a(z^{-1}))[\boldsymbol{w}^T\boldsymbol{\varphi}(\boldsymbol{x}(k)) + b] + r(k), \\ & k = 2, \ldots, N \end{aligned} \tag{4}$$

which, when $q = 1$, corresponds to the case of standard LS-SVM regression for nonlinear identification of the NAR(X)-AR model structure

$$\begin{aligned} y(k) = \rho y(k-1) + \boldsymbol{w}^T\boldsymbol{\varphi}(\boldsymbol{x}(k)) - \rho\boldsymbol{w}^T\boldsymbol{\varphi}(\boldsymbol{x}(k-1)) \\ + b(1-\rho) + r(k). \end{aligned} \tag{5}$$

The residuals $r(k)$ of this new model (5) are uncorrelated by construction and therefore standard LS-SVM regression can be applied to identify (5). The solution is formalized in the following lemma.

Lemma 1. Given a positive definite kernel function $K : \mathbb{R}^n \times \mathbb{R}^n \to \mathbb{R}$, with $K(\boldsymbol{x}_i, \boldsymbol{x}_j) = \boldsymbol{\varphi}(\boldsymbol{x}_i)^T\boldsymbol{\varphi}(\boldsymbol{x}_j)$, the solution to (4) for $q = 1$ and $a(z^{-1})e(k) = \rho e(k-1)$ is given by the dual problem

$$\left[\begin{array}{c|c} 0 & \mathbf{1}^T \\ \hline \mathbf{1} & \boldsymbol{\Omega}^{(\rho)} + \gamma^{-1}\mathbf{I} \end{array}\right]\left[\begin{array}{c} b \\ \hline \boldsymbol{\alpha} \end{array}\right] = \left[\begin{array}{c} 0 \\ \hline \tilde{\boldsymbol{y}} \end{array}\right], \tag{6}$$

with $\tilde{\boldsymbol{y}} = [y_2 - \rho y_1, \ldots, y_N - \rho y_{N-1}]^T, \boldsymbol{\alpha} = [\alpha_1, \ldots, \alpha_{N-1}]^T$, and $\boldsymbol{\Omega}^{(\rho)}$ is the kernel matrix with entries $\boldsymbol{\Omega}_{i,j}^{(\rho)} = K(\boldsymbol{x}_{(i+1)}, \boldsymbol{x}_{(j+1)}) - \rho K(\boldsymbol{x}_i, \boldsymbol{x}_{(j+1)}) - \rho K(\boldsymbol{x}_{(i+1)}, \boldsymbol{x}_j) + \rho^2 K(\boldsymbol{x}_i, \boldsymbol{x}_j)$ $\forall i, j = 1 \ldots, (N-1)$.

Proof. Consider the Lagrangian of problem (4)

$$\begin{aligned} \mathcal{L}(\boldsymbol{w}, b, e; \alpha) = & \frac{1}{2}\boldsymbol{w}^T\boldsymbol{w} + \gamma\frac{1}{2}\sum_{i=2}^{N} r(k)^2 \\ & - \sum_{k=2}^{N}\alpha_{k-1}[\boldsymbol{w}^T\boldsymbol{\varphi}(\boldsymbol{x}(k)) - \rho\boldsymbol{w}^T\boldsymbol{\varphi}(\boldsymbol{x}(k-1)) \\ & + \rho y(k-1) - y(k) - r(k)], \end{aligned}$$

where $\alpha_i \in \mathbb{R}.i = 1,\ldots,(N-1)$ are the Lagrange multipliers. Taking the optimality conditions $\frac{\partial \mathcal{L}}{\partial \boldsymbol{w}} = 0$, $\frac{\partial \mathcal{L}}{\partial b} = 0$, $\frac{\partial \mathcal{L}}{\partial r(k)} = 0$, $\frac{\partial \mathcal{L}}{\partial \alpha_{k-1}} = 0$ yields

$$
\begin{aligned}
\boldsymbol{w} &= \sum_{k=2}^{N} \alpha_{(k-1)}[\boldsymbol{\varphi}(\boldsymbol{x}(k)) - \rho\boldsymbol{\varphi}(\boldsymbol{x}(k-1))],\\
r(k) &= \alpha_{k-1}/\gamma, \quad k = 2,\ldots,N,\\
0 &= \sum_{k=1}^{N-1} \alpha_k,\\
y(k) &= \rho y(k-1) + \boldsymbol{w}^T\boldsymbol{\varphi}(\boldsymbol{x}(k)) - \rho\boldsymbol{w}^T\boldsymbol{\varphi}(\boldsymbol{x}(k-1))\\
&\quad +b(1-\rho) + r(k), \quad k = 2,\ldots,N.
\end{aligned}
$$

With the application of Mercer's theorem (Vapnik, 1998) $\boldsymbol{\varphi}(\boldsymbol{x}_i)^T\boldsymbol{\varphi}(\boldsymbol{x}_j) = K(\boldsymbol{x}_i,\boldsymbol{x}_j)$ with a positive definite kernel K, we can eliminate $\boldsymbol{w}$ and $r(k)$, obtaining $y(k)-\rho y(k-1) = \sum_{k=2}^{N}\alpha_{k-1}(K(\boldsymbol{x}_i,\boldsymbol{x}_j)-\rho K(\boldsymbol{x}_{i-1},\boldsymbol{x}_j) - \rho K(\boldsymbol{x}_i,\boldsymbol{x}_{j-1}) + \rho^2 K(\boldsymbol{x}_{i-1},\boldsymbol{x}_{j-1}))$ $+b + \frac{\alpha_k}{\gamma}$. Building the kernel matrix $\boldsymbol{\Omega}_{i,j}^{(\rho)}$ and writing the equations in matrix notation gives the final system (6)

□

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

Remark 2. (*Equivalent Kernel*). The final approximation for f in the original model (1) with $q = 1$ can be expressed in dual space as

$$
\hat{f}(\boldsymbol{x}(k)) = \sum_{j=2}^{N} \alpha_{j-1} K_{eq}(\boldsymbol{x}(j),\boldsymbol{x}(k)) + b \qquad (7)
$$

where $K_{eq}(\boldsymbol{x}(j),\boldsymbol{x}(k)) = K(\boldsymbol{x}(j),\boldsymbol{x}(k))-\rho K(\boldsymbol{x}(j-1),\boldsymbol{x}(k))$ is the equivalent kernel which embodies the information about the error correlation.

Remark 3. (*Partially Linear Structure*). The existence of correlated errors in (1) induces new dynamics into the system, leading to the model structure (5) which is a partially linear model (Speckman, 1988; Espinoza *et al.*, 2005) with a very specific restriction on the coefficients: the past output $y(k-1)$ is included as a linear term with coefficient ρ, and the past input vector $\boldsymbol{x}(k-1)$ is included under the nonlinear function which, in turn, is weighted by the value $-\rho$.

Remark 4. (*Considering ρ as an unknown*). If we consider ρ as an unknown instead of a tuning parameter in (4), an additional optimality condition from the Lagrangian $\frac{\partial \mathcal{L}}{\partial \rho} = 0$ gives

$$
\sum_{k=2}^{N} \alpha_{(k-1)}[y(k-1) - \boldsymbol{w}^T\boldsymbol{\varphi}(\boldsymbol{x}(k-1)) - b] = 0.
$$

Noting that $e(k-1) = y(k-1)-\boldsymbol{w}^T\boldsymbol{\varphi}(\boldsymbol{x}(k-1))-b$ and $\alpha_{(k-1)} = r(k)\gamma = [e(k) - \rho e(k-1)]\gamma$, this means that the estimate $\hat{\rho}$ is obtained as a solution of

$$
\sum_{k=2}^{N}[e(k) - \rho e(k-1)]e(k-1) = 0, \qquad (8)
$$

or,

$$
\hat{\rho} = \frac{\sum_{k=2}^{N} e(k)e(k-1)}{\sum_{k=2}^{N} e(k-1)^2}, \qquad (9)
$$

which corresponds to the ordinary least squares (OLS) estimator of the slope parameter from a linear regression of $e(k)$ on $e(k-1)$. This is a very intuitive result, but unfortunately the sequence $e(k)$ is unobserved. Moreover, considering ρ as an unknown parameter in (4) gives rise to a non-convex problem, as the remaining optimality conditions include products of ρ with the other unknowns. Thus, considering ρ as an unknown in (4), makes the optimization problem very difficult to solve.

Remark 5. (*Considering ρ as a tuning parameter*). 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) to yield a good generalization performance of the model, although this does not necessarily mean that the optimality condition (9), obtained for the case where ρ is an unknown in (4), is enforced. In this way, the selected ρ will be the value that gives the best cross-validation performance. This approach may increase the computational load, as each time a grid of possible values has to be defined for ρ, which may become computationally intensive for a general AR(q) case with $q > 1$. However, the definition of possible values can be guided from theoretical ranges for allowed values of ρ, which can be derived from the invertibility condition of the AR(q) process: for $q = 1$, we have $|\rho| < 1$; for $q = 2$, a sufficient condition is $|\rho_1 + \rho_2| < 1$. In general it is required for all the roots of the equation $1 - a(z^{-1}) = 0$ to be outside the unit circle (Hamilton, 1994).

3. OPTIMAL PREDICTOR AND RELATED REPRESENTATIONS

In this section, further discussion about the model properties are addressed.

3.1 Optimal Predictor

If there would be no correlation, the optimal one-step-ahead predictor $y(k|k-1)$ for time k given information known at $(k-1)$ is simply

$$y(k|k-1) = \hat{f}(\boldsymbol{x}(k)), \tag{10}$$

which corresponds to the outcome of the nonlinear identification problem (7). For the case of correlation, however, the optimal one-step-ahead predictor $y(k|k-1)$ for the model structure (1) is given by (Guidorzi, 2003)

$$y(k|k-1) = a(z^{-1})y(k) + (1 - a(z^{-1}))\hat{f}(\boldsymbol{x}(k)). \tag{11}$$

It is clear that the correlation information is incorporated into (12) in different levels:

- The first level is the optimal predictor expression itself. The prediction $y(k|k-1)$ depends not only on $\hat{f}(\boldsymbol{x}(k))$ but also on past values of $y(k)$ and $\hat{f}(\boldsymbol{x}(k))$ which are generated by the correlation structure contained in $a(z^{-1})$.
- The second level is the expression for $\hat{f}$, which contains the temporal correlation structure embedded at the kernel level.

This becomes clear when expressing $\hat{f}$ in terms of the kernel expression (7), in which case the optimal predictor is given (for $q = 1$) by,

$$\begin{aligned} y(k|k-1) = \rho y(k-1) + \sum_{j=2}^{N} \alpha_{j-1}[K(\boldsymbol{x}(j), \boldsymbol{x}(k)) - \\ \rho K(\boldsymbol{x}(j-1), \boldsymbol{x}(k)) - \rho K(\boldsymbol{x}(j), \boldsymbol{x}(k-1)) + \\ \rho^2 K(\boldsymbol{x}(j-1), \boldsymbol{x}(k-1))] + b(1-\rho). \end{aligned} \tag{12}$$

3.2 Links with other model representations

The above expression (11) is valid for $\boldsymbol{x}(k)$ containing past values of the output $y(k)$. However, interesting links with existing and well known model representations can be established for the case where $\boldsymbol{x}(k)$ does not contain past values of the output, i.e., the nonlinear function $f(\boldsymbol{x}(k))$ is a static nonlinearity. Considering $\boldsymbol{x}(k)$ as an exogenous input, the model structure

$$(1 - a(z^{-1}))y(k) = (1 - a(z^{-1}))f(\boldsymbol{x}(k)) + r(k) \tag{13}$$

is equivalent to a Hammerstein system (Crama *et al.*, 2004)

$$y(k) = \sum_{i=1}^{r} c_i y(k-i) + \sum_{i=0}^{s} d_i f(\boldsymbol{x}(k-i)) + r(k) \tag{14}$$

with $r = s = q$ (the order is given by the order of the AR(q) residual process), and the following conditions on the coefficients: $c_i = \rho_i$, $d_i = -\rho_i$, $i = 1, \ldots, q$ and $d_0 = 1$.

Alternatively, additional insights into the model structure can be obtained when considering the model formulation as a state-space description. For clarity of presentation, consider the case for $q = 1$

$$\begin{cases} e(k+1) = \rho e(k) + r(k+1) \\ y(k) = e(k) + f(\boldsymbol{x}(k)) \end{cases} \tag{15}$$

and now the AR(1) process corresponds to the state equation. In this interpretation, $e(k)$ corresponds to the unobserved state of the system, $r(k+1)$ is the process noise, and ρ is the parameter for the state equation of this system. The output equation consists of the state $e(k)$ with coefficient equal to 1, and an input which is described as a nonlinear function of the vector $\boldsymbol{x}$. The above description gives explicit expressions for optimal prediction, where not only the nonlinear function f has to be approximated, but also the corresponding state should be predicted as well. Under this interpretation, the optimal predictor for $k+1$ given the information up to time k can be easily obtained in terms of both the predictors of the future state $e(k+1|k)$ and the output $y(k+1|k)$ via e.g. Kalman filter applied to (15), and is equivalent to the optimal predictor (11) for the case of a static nonlinearity.

4. EXAMPLES

In this section, two examples are considered to illustrate the effect of autocorrelated residuals with $q = 1$. The first case is a static regression model, the second case is a NARX formulation. On each case, an RBF kernel is used, and the hyperparameters are tuned by 10-fold cross-validation. By assumption, $|\rho| < 1$, therefore the considered values for the tuning parameter ρ range from -0.9 to 0.9 every 0.1 steps. Each example involves the estimation of the correlation-corrected LS-SVM (C-LS-SVM) and standard LS-SVM for comparison.

4.1 Static Nonlinearity

Consider the following example where the true underlying system (1) is defined to contain a static formulation $f(x) = 1 - 6x + 36x^2 - 53x^3 + 22x^5$. The input values $x(k)$ are sampled i.i.d. from a uniform distribution between 0 and 1, with $N = 100$ datapoints. The error sequence $e(k)$ is built using $\rho = 0.7$ and $\sigma_u^2 = 0.5$. In this case, the original system is static, and the correlation induces a dynamical behavior in the observed values. Figure 1 (bottom) shows the plot of y on x, in order to visualize the true polynomial function as a function of x. The true f function is shown as a thin line, and the estimated function from (7) is

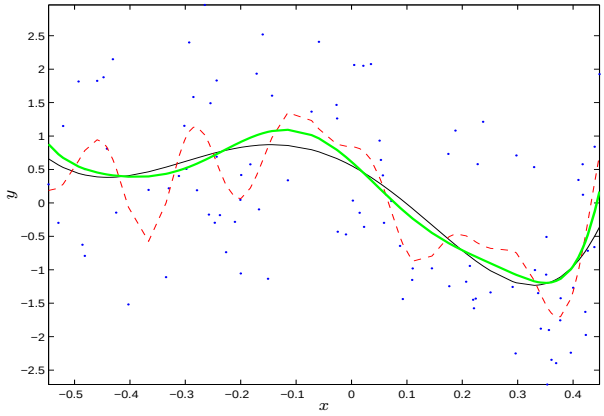


Fig. 1. True (thin) function and the identified functions estimated with C-LSSVM (thick) and standard LS-SVM (dashed) for Example 1.

shown with a thick line. For comparison, the estimated function with standard LS-SVM (neglecting correlation) is shown in dashed-line. It is clear that the estimation with the corrected LS-SVM can better identify the true function, whereas the standard LS-SVM mixes the true function with the correlation structure. The parameter ρ that minimizes the cross-validation mean squared error (MSE) coincides with the true AR(1) parameter 0.7. This example of a static nonlinearity already shows the effect of the error correlation, where the apparently independent sequence of inputs and outputs obtains a temporal correlation via the residuals of the equation.

4.2 NAR-AR model

This example considers the identification of a NAR-AR model

$$\begin{cases} y(k) = 2 \cdot \mathrm{sinc}(y(k-1)) + e(k) \\ e(k) - \rho e(k-1) = r(k) \end{cases} \tag{16}$$

generated with $\rho = 0.6$, $\sigma_u = 0.1$ for 150 datapoints. The first 100 points are used for identification, and the remaining 50 points are used for out-of-sample assessment of the prediction performance.

- *Identification of the AR(1) parameter.* Following the standard methodology, 10-fold cross-validation is performed to select the hyperparameters γ (regularization term), σ (RBF kernel parameter) and the ρ (the AR(1) parameter). Figure 2 (top) shows the cross-validation MSE for different combinations of hyperparameters, plotted for the values of ρ. In other words, for a given ρ, different MSE results are obtained depending on the combinations of σ and γ. The best performance is obtained for ρ=0.6 which corresponds to the true value of the AR(1) process.

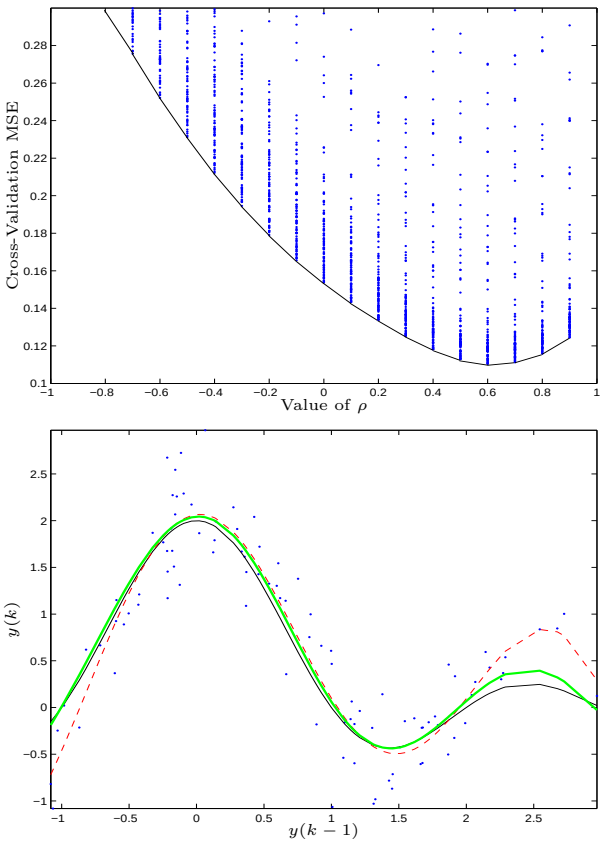


Fig. 2. (Top) Evolution of the cross-validation MSE for different combination of hyperparameters. The optimal performance is found at ρ=0.6.(Bottom) True (thin) function and the identified functions estimated with C-LSSVM (thick) and standard LS-SVM (dashed) for Example 2.

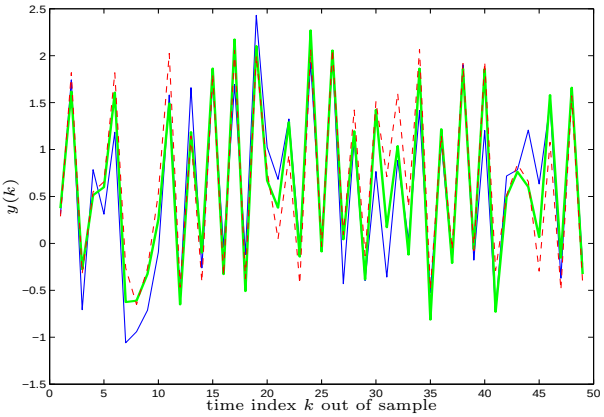


Fig. 3. Out-of-sample predictions obtained with C-LSSVM (thick) and standard LS-SVM (dashed) compared to the actual values (thin line) for Example 2.

- *Identification of the nonlinear function.* Once the hyperparameters are selected, the approximation of f is obtained from (7). Figure 2 (bottom) shows the training points (dots), the identified function $\hat{f}$ (thick line), the true function (thin line) and the approximation obtained with standard LS-SVM (dashed line) for comparison. As in the previous example, the corrected LS-SVM is able to sepa-

Performance	LS-SVM	C-LS-SVM
MSE in-sample	0.13	0.09
MSE cross-validation	0.17	0.10
MSE out-of-sample	0.18	0.09

Table 1. In-sample, cross-validation and out-of-sample performance of the models for Example 2.

rate the correlation effects from the nonlinear function.

- *Prediction Performance.* Using the expression (12), out-of-sample predictions are computed for the system (16) for the next 50 datapoints. Table 1 shows the MSE calculated over the test set, compared with the results obtained from prediction using standard LS-SVM. The better performance of the correlation-corrected LS-SVM reflects the fact that the optimal predictor includes all information about the model structure, whereas the standard LS-SVM considers that all dynamical effects are due to the nonlinear function only. Figure (3) shows the actual values (thin line) and the predictions generated by C-LSSVM (thick line) and standard LS-SVM (dashed line).

5. CONCLUSIONS

In this paper the problem of LS-SVM regression with correlated residuals has been addressed. Starting from the prior knowledge of the correlation structure, the modelling is treated as a convex problem with the coefficients of the AR residual process as tuning parameters. The dual solution of the model incorporates the correlation information into the kernel level. Additionally, the optimal one-step-ahead predictor includes the correlation structure explicitly. The correlation structure induces a very specific dynamical behavior into the final model, which can be linked to a restricted Hammerstein system and a state-space representation for the case of a static nonlinearity. Practical examples show how the inclusion of the correlation structure into the model gives a much better identification of the nonlinear function, and better out-of-sample performance in terms of prediction and simulation.

ACKNOWLEDGEMENTS

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 with K.U.Leuven, Belgium, respectively. The scientific responsibility is assumed by its authors.

REFERENCES

Altman, N.S. (1990). Kernel smoothing of data with correlated errors. *Journal of the American Statistical Association* **85**, 749–759.

Crama, P., J. Schoukens and R. Pintelon (2004). Generation of enhanced initial estimates for hammerstein systems. *Automatica* **40**, 1269–1273.

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

Espinoza, M., J.A.K. Suykens and B. De Moor (2005). Kernel based partially linear models and nonlinear identification. *IEEE Transactions on Automatic Control, Special Issue: Linear vs. Nonlinear*. To appear.

Guidorzi, R. (2003). *Multivariable System Identification: From Observations to Models*. Bononia University Press.

Hamilton, J. (1994). *Time Series Analysis*. Princeton University Press.

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

Ljung, L. (1987). *System Identification: Theory for the User*. Prentice Hall. New Jersey.

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

Speckman, P. (1988). Kernel smoothing in partial linear models. *J. R. Statist. Soc. B*.

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

Vapnik, V. (1998). *Statistical Learning Theory*. Wiley. New-York.