Support Vector Method for Robust ARMA System Identification

José Luis Rojo-Álvarez, *Member, IEEE*, Manel Martínez-Ramón, *Member, IEEE*,
Mario de Prado-Cumplido, *Student Member, IEEE*, Antonio Artés-Rodríguez, *Senior Member, IEEE*, and
Aníbal R. Figueiras-Vidal, *Senior Member, IEEE*

***Abstract*—This paper presents a new approach to auto-regressive and moving average (ARMA) modeling based on the support vector method (SVM) for identification applications. A statistical analysis of the characteristics of the proposed method is carried out. An analytical relationship between residuals and SVM-ARMA coefficients allows the linking of the fundamentals of SVM with several classical system identification methods. Additionally, the effect of outliers can be cancelled. Application examples show the performance of SVM-ARMA algorithm when it is compared with other system identification methods.**

***Index Terms*—ARMA modeling, cross-correlation, support vector method, system identification, time series.**

I. Introduction

A FREQUENT approach to digital signal processing is to propose a model composed by two discrete-time processes (DTP), which are the input and the output to a linear, time-invariant system (LTIS). An LTIS is usually approximated by means of a rational transfer function that can be an all-zero or moving average (MA), an all-pole or auto-regressive (AR), or a general zero-pole (ARMA) system. ARMA modeling is useful in problems such as system identification, time-series analysis, spectral analysis, and deconvolution. A number of applications for ARMA modeling can be enumerated. For example, in digital communications, channel identification provides a way of reducing intersymbol interference; in addition, channels of multiuser detection schemes can be viewed as multiple-input multiple-output systems, and the MA structure is used to compensate the effect of multipath signal propagation [1]. In radar signal processing, adaptiveness is better achieved by means of ARMA implementations, getting advantage with respect to the Fourier transform in the robust determination of Doppler spectrum [2], [3]. Biomedical signals can be analyzed to build models for diagnosis purposes; spectral analysis of the heart-rate variability and its relationship to blood pressure variability are useful in arrhythmia risk stratification of patients with previous myocardial infarction [4]; cardiovascular models of the aortic vein state can be estimated from blood pressure and flow measurements [5]; and spectral analyses of electroencephalographic potentials enable a deeper knowledge of the brain [6].

Classical ARMA estimation methods present some limitations.

- Analysis of DTP with atypical samples (outliers) is neither easy nor immediate, and it is usually achieved by heuristic or even visual inspection methods [12]. So, the use of robust cost functions to avoid the effect of outliers often arises numerical optimization disadvantages.
- In general terms, ARMA methods require a previous determination of model complexity or number of parameters in the model, and they are quite sensitive to wrong order choice [9].
- Finally, error surfaces are not convex in many cases. Even worse, sample-based approximations to convex theoretical surfaces can exhibit local minima.

In this paper, we introduce a new approach to ARMA modeling that is based on the support vector method (SVM). The SVM was first proposed to obtain maximum margin separating hyperplanes in classification problems [13], but this technique has become a general learning theory [14], [15]. A comprehensive description of this method for classification and regression problems can be found in [16] and [17], respectively. Although some previous work has been done with SVM regression and time series analysis [7], [8], the algorithm that we present is formulated for system identification: a framework where robustness is specially needed. Robustness is just a main feature that SVM controls, as it will be shown in this paper. Besides, we develop a statistical analysis of the characteristics of our method. In addition, we pay special attention to the cost function used in this SVM approach, which is a scarcely discussed aspect in SVM regression literature, although it has a key relevance in signal processing environments.

The approach to ARMA system identification drawn from SVM minimizes a regularized cost functional of the residuals. Potential advantages of SVM are the following.

- It provides a unique solution.
- It is a strongly regularized method, which is appropriate for ill-posed problems.
- It extracts the maximum information from the available samples, although the statistical distribution is unknown.

Consequently, SVM can diminish some of the limitations of classical ARMA system identification methods, as we will see below in more detail.

The structure of the paper is as follows. In Section II, we introduce SVM-ARMA system identification equations, and we propose a robust cost function that leads to the corresponding

Manuscript received December 10, 2002; revised April 10, 2003. This work was supported in part by CICYT under Grants TIC 1999-0216 and TIC 2000-0380-C03-03. The associate editor coordinating the review of this paper and approving it for publication was Prof. Ioan Tabus.

The authors are with the Department of Signal Theory and Communications, University Carlos III de Madrid, 28911 Leganés, Madrid, Spain.

Digital Object Identifier 10.1109/TSP.2003.820084

SVM-ARMA algorithm. Then, the relationship between residuals and SVM-ARMA coefficients is analyzed in Section III, yielding a useful statistical interpretation of SVM-ARMA terms. This allows the establishment of a comparison with classical system identification methods. Simulation and application examples are included in Section IV. Finally, in Section V, conclusions are drawn.

II. SVM-ARMA Formulation

Let us consider two DTPs $\{x_n\}$ and $\{y_n\}$, which are the input and the output, respectively, of a rational LTIS. The corresponding difference equation is

$$y_n = \sum_{i=1}^{P} a_i y_{n-i} + \sum_{j=1}^{Q} b_j x_{n-j+1} + e_n \tag{1}$$

where $\{a_n\}$ and $\{b_n\}$ are the P AR and Q MA coefficients of the system, respectively, and $\{e_n\}$ is a DTP standing for the effect of measurement errors.

Let us also consider a set of N consecutive samples of $\{x_n\}$ and a set of N consecutive samples of $\{y_n\}$ observed at the same time instants. The difference equation can be used to set a relationship among observations, estimated parameters, and errors. In this case, error terms $\{e_n\}$, or residuals, comprehend both measurement and model approximation errors. In order to consider initial conditions, (1) is required only for time-lags $n = k_o, \ldots, N$, where $k_o = \max(P+1, Q)$. Assuming that P and Q (also known as model order) can be properly chosen beforehand, ARMA (P, Q) coefficients are usually estimated from minimizing a cost functional, depending on the residuals.

In general, SVM algorithms for linear classification and regression problems minimize a cost function for the residuals (CFR) that is called Vapnik's (or ε-insensitive) loss function, which is given by

$$L_\varepsilon(e_n) = \begin{cases} |e_n| - \varepsilon, & \text{if } |e_n| \geq \varepsilon \\ 0, & \text{if } |e_n| < \varepsilon \end{cases} \tag{2}$$

and this CFR is regularized with the L_2 norm of the model parameters [13], [18]. Whereas $\varepsilon \neq 0$, SVM allows to obtain sparse solutions where the estimated classification or regression function depends only on a reduced number of the set of observed samples; these samples are called support vectors. The extension of SVM to nonlinear classifiers and regressors can be easily achieved by using Mercer's kernels.

An additional free parameter C must be previously fixed to control the tradeoff between the cost of the residuals and the regularization term. This tradeoff is chosen according to some *a priori* knowledge of the problem or by using cross-validation techniques. The resulting functional is usually optimized via quadratic programming (QP) [19].

We propose to use Vapnik's CFR plus an L_2 regularization, including both the AR and the MA model coefficients, to create what we call SVM-ARMA modeling. This corresponds to the unconstrained minimization of

$$L_P(a_i, b_j, e_n) = \frac{1}{2}\left(\sum_{i=1}^{P} a_i^2 + \sum_{j=1}^{Q} b_j^2\right) + C\sum_{n=k_o}^{N} L_\varepsilon(e_n). \tag{3}$$

An equivalent formulation of the above process [13] is to minimize

$$L_P(a_i, b_j, \xi_n, \xi_n^*) = \frac{1}{2}\left(\sum_{i=1}^{P} a_i^2 + \sum_{j=1}^{Q} b_j^2\right) + C\sum_{n=k_o}^{N} (\xi_n + \xi_n^*) \tag{4}$$

where ξ_n, and ξ_n^* are slack variables or losses constrained to

$$y_n - \sum_{i=1}^{P} a_i y_{n-i} - \sum_{j=1}^{Q} b_j x_{n-j+1} \leq \varepsilon + \xi_n \tag{5}$$

$$-y_n + \sum_{i=1}^{P} a_i y_{n-i} + \sum_{j=1}^{Q} b_j x_{n-j+1} \leq \varepsilon + \xi_n^* \tag{6}$$

$$\xi_n^{(*)} \geq 0 \tag{7}$$

for $n = k_o, \ldots, N$, where $\xi_n^{(*)}$ denotes both ξ_n and ξ_n^*. The primal-dual or Lagrange functional for this problem is obtained by introducing a nonnegative coefficient (Lagrange multiplier) for each constraint [19], yielding

$$\begin{aligned} &L_{PD}(a_i, b_j, \xi_n, \xi_n^* \alpha_n, \alpha_n^*, \beta_n, \beta_n^*) \\ &\quad = C\sum_{n=k_o}^{N} (\xi_n + \xi_n^*) \\ &\qquad + \frac{1}{2}\left(\sum_{i=1}^{P} a_i^2 + \sum_{j=1}^{Q} b_j^2\right) - \sum_{n=k_o}^{N} (\beta_n \xi_n + \beta_n^* \xi_n^*) \\ &\qquad + \sum_{n=k_o}^{N} \alpha_n \left(y_n - \sum_{i=1}^{P} a_i y_{n-i} \right. \\ &\qquad\qquad \left. - \sum_{j=1}^{Q} b_j x_{n-j+1} - \varepsilon - \xi_n\right) \\ &\qquad + \sum_{n=k_o}^{N} \alpha_n^* \left(-y_n + \sum_{i=1}^{P} a_i y_{n-i} \right. \\ &\qquad\qquad \left. + \sum_{j=1}^{Q} b_j x_{n-j+1} - \varepsilon - \xi_n^*\right) \end{aligned} \tag{8}$$

where the multipliers are constrained to $\alpha_n^{(*)} \geq 0$ and $\beta_n^{(*)} \geq 0$, and (7) also stands. Equation (8) has to be minimized with respect to primal variables a_i, b_j, and $\xi_n^{(*)}$ and maximized with respect to Lagrange multipliers (also called dual variables) $\alpha_n^{(*)}$ and $\beta_n^{(*)}$.

From

$$\frac{\partial L_{PD}}{\partial a_i} = 0; \quad \frac{\partial L_{PD}}{\partial b_j} = 0; \quad \frac{\partial L_{PD}}{\partial \xi_n^{(*)}} = 0 \tag{9}$$

several consequences are drawn. First, dual variables are shown to be constrained between an upper and a lower bound

$$0 \leq \alpha_n^{(*)} \leq C \tag{10}$$

for $n = k_0, \ldots, N$. Second, we can observe that

$$a_i = \sum_{n=k_o}^{N} (\alpha_n - \alpha_n^*)\, y_{n-i} \tag{11}$$

$$b_j = \sum_{n=k_o}^{N} (\alpha_n - \alpha_n^*)\, x_{n-j+1} \tag{12}$$

show the analytical relationship among the model coefficients, the dual coefficients, and the observations. Third, these conditions can be introduced into Lagrange functional (8) to remove the primal variables. After that, another term grouping can be done in L_{PD} by writing down

$$R_y^P(m,k) = \sum_{i=1}^{P} y_{m-i} y_{k-i} \tag{13}$$

$$R_x^Q(m,k) = \sum_{j=1}^{Q} x_{m-j+1} x_{k-j+1}. \tag{14}$$

These equations denote the time-local Qth- and Pth-order sample estimators of the values of the autocorrelation functions of the input and the output DTP $R_x(m,k)$ and $R_y(m,k)$, respectively.

By introducing (13) and (14) into (8), the problem can be expressed in terms of the residual constraints and the data, and it is stated in vector-matrix form as the maximization of

$$\begin{aligned} L_D = -\frac{1}{2}(\boldsymbol{\alpha}-\boldsymbol{\alpha}^*)^T \left[\mathbf{R_x}^Q + \mathbf{R_y}^P\right](\boldsymbol{\alpha}-\boldsymbol{\alpha}^*) \\ + (\boldsymbol{\alpha}-\boldsymbol{\alpha}^*)^T \boldsymbol{y} - \varepsilon \mathbf{1}^T(\boldsymbol{\alpha}+\boldsymbol{\alpha}^*) \end{aligned} \tag{15}$$

under constrains (10), and where $\boldsymbol{\alpha}^{(*)} = \left[\alpha_{k_o}^{(*)}, \ldots, \alpha_N^{(*)}\right]^T$ and $\boldsymbol{y} = [y_{k_o}, \ldots, y_N]^T$.

This is a QP problem, and if we write

$$\mathbf{z} = [\boldsymbol{\alpha}^T, \;\; \boldsymbol{\alpha}^{*T}]^T \tag{16}$$

$$\mathbf{H} = -\frac{1}{2}\begin{bmatrix} \mathbf{R_x}^Q + \mathbf{R_y}^P, & -\mathbf{R_x}^Q - \mathbf{R_y}^P \\ -\mathbf{R_x}^Q - \mathbf{R_y}^P, & \mathbf{R_x}^Q + \mathbf{R_y}^P \end{bmatrix} \tag{17}$$

$$\mathbf{b} = [\boldsymbol{y}^T - \varepsilon, \;\; -\boldsymbol{y}^T - \varepsilon]^T \tag{18}$$

then the aim is to maximize $\mathbf{z^T H z} + \mathbf{b^T z}$ with respect to $\mathbf{z}$ under some linear constraints. It is clear that matrix $\mathbf{H}$ is not invertible and the solution is unfeasible. Like in SVM regression [17], this problem can be *ad hoc* regularized by adding a small value diagonal matrix, i.e., by replacing the square matrix by $\mathbf{H}' = \mathbf{H} + \gamma\mathbf{I}$. Matrix $\mathbf{H}'$ can be easily shown to be definite positive, and the resulting constrained QP problem has a single minimum, thus avoiding the local minima problems in the LS solution of the normal ARMA system identification equations [9].

Note that the above-mentioned numerical regularization is apparently different from the L_2-norm regularization of the primal functional in (4); in fact, this numerical regularization is not always reported in SVM regression literature, and it is frequently treated as a simple numerical trick. However, the numerical regularization in the dual problem can be seen as an additional term in the dual functional, given by the L_2 norm of the Lagrange coefficients. This means that the QP problem we are really solving is the maximization of

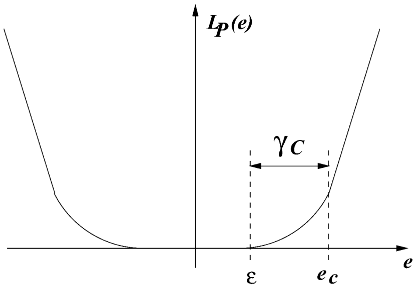


Fig. 1. Robust CFR for the SVM-ARMA algorithm. There are three different (possible) regions, allowing us to deal with different kinds of noise.

$$\begin{aligned} L_D^{SVM} = &-\frac{1}{2}(\boldsymbol{\alpha}-\boldsymbol{\alpha}^*)^T \left[\mathbf{R_x}^Q + \mathbf{R_y}^P\right](\boldsymbol{\alpha}-\boldsymbol{\alpha}^*) \\ &+ (\boldsymbol{\alpha}-\boldsymbol{\alpha}^*)^T \boldsymbol{y} - \varepsilon \mathbf{1}^T(\boldsymbol{\alpha}+\boldsymbol{\alpha}^*) \\ &- \frac{\gamma}{2}\left(\boldsymbol{\alpha}^T \mathbf{I} \boldsymbol{\alpha} + \boldsymbol{\alpha}^{*T} \mathbf{I} \boldsymbol{\alpha}^*\right) \end{aligned} \tag{19}$$

constrained to (10).

Several robust cost functions have been used in SVM regression, such as Vapnik's loss function [13], Huber's robust cost [7], or the ridge regression approach [20]. Here, we propose a more general cost function that has the above-mentioned ones as particular cases. Proposed robust CFR is depicted in Fig. 1, and it can be expressed as the following piecewise-defined function:

$$L^P(e_n) = \begin{cases} 0, & |e_n| \le \varepsilon \\ \frac{1}{2\gamma}(|e_n| - \varepsilon)^2, & \varepsilon \le |e_n| \le e_C \\ C(|e_n| - \varepsilon) - \frac{1}{2}\gamma C^2, & |e_n| \ge e_C \end{cases} \tag{20}$$

where $e_c = \varepsilon + \gamma C$. The three different intervals of L^{SVM} serve to deal with different kinds of noise. Insensitive zone $|e_n| < \varepsilon$ is adequate for low-frequency variations such as wander or baseline deviations. The quadratic cost zone takes into account the observation noise, the L_2 norm in this zone being appropriate for Gaussian processes. The linear cost zone limits the effect of either outliers or jitter noise on the model parameter estimation.

Thus, the proposed robust SVM-ARMA algorithm is stated as the minimization of this robust CFR plus the regularization term:

$$\begin{aligned} L_P^{SVM}(a_i, b_j, \xi_n, \xi_n^*) = \frac{1}{2}\left(\sum_{i=1}^{P} a_i^2 + \sum_{j=1}^{Q} b_j^2\right) \\ + \frac{1}{2\gamma}\sum_{n \in I_1}\left(\xi_n^2 + \xi_n^{*2}\right) + C\sum_{n \in I_2}(\xi_n + \xi_n^*) \end{aligned} \tag{21}$$

constrained to (5)–(7), where I_1 is the set of samples for which $\varepsilon \le \xi_n^{(*)} \le e_C$, and I_2 is the set of samples for which $\xi_n^{(*)} > e_C$. Appendix A shows that the dual problem corresponding to (21) is equivalent to (19) in the sense that both reach their optimum for the same values of Lagrange coefficients. This fact permits a twofold interpretation.

- When considering Vapnik's loss function as the cost on the residuals, the regularized cost function for SVM-ARMA we are really working with is not (4), but rather (21), due to the effect of the numerical regularization.

- Far from being a disadvantage, considering the quadratic cost zone will allow to work with a more general cost function than Vapnik's cost and for some kinds of DTP, it can be useful to set an appropriate value for free parameter γ.

Note that three free parameters (ε, e_C and C) are to be tuned. These parameters can be *a priori* fixed according to the statistical model of DTPs. Selecting $\varepsilon = 0$ leads to Huber's robust cost function [21]. In addition, $\gamma = 0$ represents Vapnik's function, showing that it becomes nondifferentiable at $e_n = \varepsilon$ in the absence of numerical regularization. In a number of applications, some knowledge of the statistical properties of DTPs can be available. So, it will be desirable relating (20) to other cost criteria by establishing a general class of robust cost functions. When noise is Gaussian, the quadratic cost will be the most appropriate; however, in this case, insensitivity should be removed ($\varepsilon = 0$), and sparsity will not be achieved. For sub-Gaussian noise, a better performance could be obtained with a cubic or higher degree cost zone. If super-Gaussian (for instance, impulsive or heavy-tails) noise is present, it will be convenient to set a low value for product γC but with $\gamma \neq 0$ so that we will not have an ill-posed problem.

III. Statistical Interpretation

The consideration of the cost function in (20) allows us to trace an analytical relationship between residuals and Lagrange multipliers in the SVM-ARMA algorithm (21). Let $\boldsymbol{\alpha}^o, \boldsymbol{\alpha}^{*o}$ be the Lagrange coefficients in the solution. Then, residuals e_n and coefficients are related in the form

$$\alpha_n^o - \alpha_n^{o*} = f(e_n) = \begin{cases} -C, & e_n \leq -e_C \\ \frac{1}{\gamma}(e_n + \varepsilon), & -e_C \leq e_n \leq -\varepsilon \\ 0, & -\varepsilon \leq e_n \leq \varepsilon \\ \frac{1}{\gamma}(e_n - \varepsilon), & \varepsilon \leq e_n \leq e_C \\ C, & e_C \leq e_n. \end{cases} \quad (22)$$

This analytical relationship is proved in Appendix B from Karush-Khun-Tucker (KKT) conditions when the solution of the QP problem is reached.

Analytical relationship (22) between Lagrange multipliers and residuals in SVM-ARMA algorithm (21) allows us to compare the latter with classical algorithms. Although a really vast number of methods for system identification have been suggested, Ljung points out in [9] that there are two main general procedures. On the one hand, prediction-error methods (PEMs) are based on the minimization of a function of the residual power for a given model. PEMs contain well-known procedures such as the least-squares (LS) method, and they are closely related to Bayesian maximum a posteriori estimation [10]. On the other hand, correlation methods (CMs) seek for the minimization of the cross-correlation between a function of the residuals and a transformation of the data, possibly depending on the parameter vector; this approach includes the instrumental-variable 4 method (from now, Iv4), as well as several procedures for rational transfer function modeling [11]. Here, we present a brief comparison of SVM-ARMA with both kinds of methods.

Comparison with PEM: If $\varepsilon = 0$, $e_C = +\infty$, then Lagrange multipliers are just proportional to the residuals. To highlight the bias-variance dilemma in SVM-ARMA versus LS system identification, a geometrical interpretation based on the orthogonality principle will be carried out. For the sake of simplicity, let us consider the projection of a vector $\mathbf{v} \in \mathbf{R}^3$ onto the subspace generated by data $\{\mathbf{x}, \mathbf{y}\}$, as depicted in Fig. 2. According to LS criterion, the projection is denoted as $\mathbf{v_{LS}} = v_1\mathbf{x} + v_2\mathbf{y}$, whereas SVM projection is $\mathbf{v_{SVM}} = w_1\mathbf{x} + w_2\mathbf{y}$. SVM projection is obtained as the result of minimizing

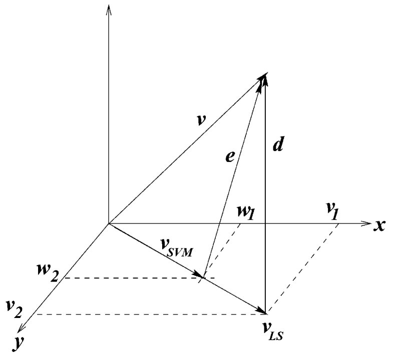


Fig. 2. Geometrical interpretation of error bias. LS solution $\mathbf{v_{LS}}$ produces an unbiased averaged error $\mathbf{d}$ that is orthogonal to the space spanned by the data, whereas the SVM solution $\mathbf{v}_{\mathrm{SVM}}$ can be seen as a regularized but biased solution with increased approximation averaged error $\mathbf{e}$.

$$L = \frac{1}{2}(w_1^2 + w_2^2) + \frac{1}{2\gamma}\|\mathbf{e}\|^2 \quad (23)$$

constrained to $\mathbf{e} = \mathbf{v} - \mathbf{v_{SVM}}$. From

$$\frac{\partial L}{\partial w_1} = 0; \quad \frac{\partial L}{\partial w_2} = 0 \quad (24)$$

the following relationships can be established for the coefficients of both solutions:

$$w_1 = \frac{1}{1+\gamma}v_1; \quad w_2 = \frac{1}{1+\gamma}v_2 \quad (25)$$

from which

$$\langle \mathbf{e}, \mathbf{x} \rangle = v_1 - w_1 = (1+\gamma)w_1 - w_1 = \gamma w_1 \quad (26)$$

$$\langle \mathbf{e}, \mathbf{y} \rangle = v_2 - w_2 = (1+\gamma)w_2 - w_2 = \gamma w_2. \quad (27)$$

Accordingly, SVM-ARMA coefficients given by (11) and (12) can be seen as the minimum norm projection of the solution onto the subspace that is generated by the input signal, the output signal, and their delayed versions. The model coefficients are calculated as the dot product between the nonlinearized error and the subspace base vectors.

If the observation noise is Gaussian, and the observed data have actually been generated by an ARMA model, i.e., $\mathbf{v} = u_1\mathbf{x} + u_2\mathbf{y}$, then $\mathbf{v_{LS}}$ is an unbiased estimator of the model coefficients [9], and for the ith coefficient $(i = 1, 2)$, we have

$$E\{v_i\} = u_i \quad (28)$$

where $E\{\cdot\}$ denotes statistical expectation of a random variable. However, $\mathbf{v_{SVM}}$ is a biased estimator since

$$E\{w_i\} = E\left\{\frac{1}{1+\gamma}v_i\right\} = \frac{1}{1+\gamma}u_i \quad (29)$$

and the bias is due to the regularization term. Additionally, the variance of the LS estimator is

$$\sigma_{v_i}^2 = E\{(v_i - u_i)^2\} = E\{v_i^2\} - u_i^2 \tag{30}$$

whereas the variance of the SVM estimator is

$$\sigma_{w_i}^2 = E\{(w_i - u_i)^2\} = \frac{1}{1+\gamma}\sigma_{v_i}^2 + \frac{\gamma}{1+\gamma}u_i^2. \tag{31}$$

In this case, biased SVM-ARMA will provide better estimators only if its variance is below LS variance, i.e.,

$$\sigma_{w_i}^2 \leq \sigma_{v_i}^2 \Leftrightarrow \sigma_{v_i}^2 \geq u_i^2. \tag{32}$$

As a consequence, in the presence of Gaussian noise and for an underlying ARMA model, the LS parameter estimator is the optimum solution in the sense of unbiased error. Here, SVM-ARMA algorithm will be able to provide a biased (yet reduced variance) estimator whenever LS estimator variance is greater than the true ARMA coefficient; this can be the case in situations having a low number of observation samples, true coefficients of low amplitude, or low signal-to-noise ratio (SNR).

Comparison with CM: As previously mentioned, CM are based on the assumption that a good model produces residuals, regardless of past data. They also seek the minimization of the cross-correlation between a function of the residuals and a transformation of the data. The open issue in this setting is how to find both the residual function and the data transformation.

In order to relate SVM-ARMA algorithm with CM, we start from the case $\varepsilon = 0$, $\gamma = 1$, and $e_C = +\infty$ (quadratic loss only), which again leads to Lagrange multipliers that are equal to the residuals. In this case, prediction coefficients $\{a_n\}$ and $\{b_n\}$ are sample estimates (except for a scaling factor) of the cross-correlation between the residuals and the data. Using (22) in (11) and (12), and since

$$R_{ye}(l) = E\{y_{n+l}e_n\} \simeq \frac{1}{N_s}\sum_{n\in N_s} y_{n+l}e_n \tag{33}$$

$$R_{xe}(l) = E\{x_{n+l}e_n\} \simeq \frac{1}{N_s}\sum_{n\in N_s} x_{n+l}e_n \tag{34}$$

with $N_s = N - k_0 + 1$, the coefficients can be seen as

$$a_i = N_s\hat{R}_{ye}(-i) \tag{35}$$

$$b_j = N_s\hat{R}_{xe}(-j+1). \tag{36}$$

This emphasizes the fact that the SVM-ARMA solution leads to low correlation (in the given lags) between the residuals and the data.

If $\varepsilon \neq 0$ or $e_C \neq +\infty$ (or both), then $\boldsymbol{\alpha}^o - \boldsymbol{\alpha}^{*o}$ is a nonlinear transformation of the residuals given by (22). In this case, the model coefficients are estimated from the cross-correlation of the transformed residuals and the data

$$a_i = N_s\hat{R}_{yf(e)}(-i) \tag{37}$$

$$b_j = N_s\hat{R}_{xf(e)}(-j+1). \tag{38}$$

As long as we impose the minimum norm of the coefficients, the sample correlation between this nonlinear function of the residuals and the data is reduced. Taking into account (11) and (12), one can see the relationship to the general CM equation [9]. Note that there is no data transformation, but instead, the residual nonlinear function is based on the statistical knowledge of the errors, which is incorporated into the free parameters of CFR.

Finally, the comparison with CM suggests the possibility of a new algorithm by forcing the residuals to be uncorrelated with the data for more lags than P and Q. This could be easily achieved by forcing coefficients $\{a_i, b_j\}$ to be small or even zero for $i > P, j > Q$.

IV. Application Examples

In the applications that we will present here for the system identification SVM-ARMA algorithm, attention will be paid to some of its main features. First, a simulation example allows us to appreciate the performance when outliers are present. Second, a classical example (Feedback's Process Trainer [9]) is used to compare SVM-ARMA with several system identification methods. Finally, a real-world example is reported (the relationship between simultaneous heart rate and diastolic blood pressure), where erroneous measurement values often appear. Two performance measurements are used: the error of the approximation to the true impulse response of the LTIS when it is known (example A) and the prediction error on a validation set when real data are analyzed (examples B and C).

A. Insensitivity to Outliers

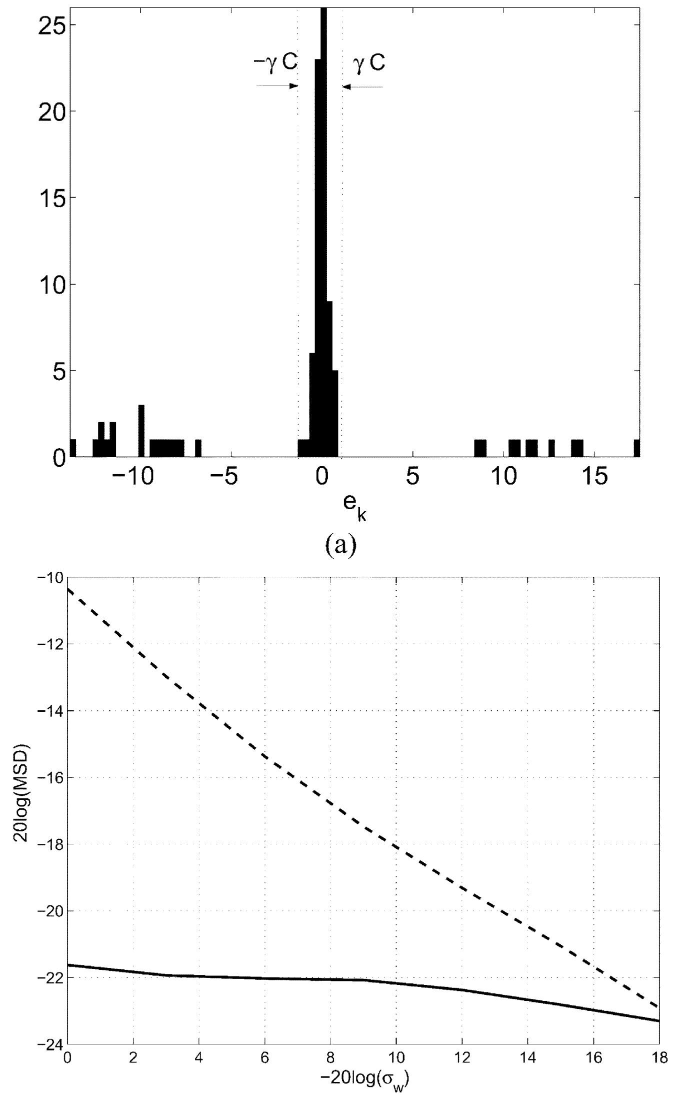
The test system to be identified is

$$y_n = 0.03y_{n-1} - 0.01y_{n-2} + 3x_n - 0.5x_{n-1} + 0.2x_{n-2}. \tag{39}$$

This system is chosen because its impulse response has samples with amplitudes of different orders of magnitude. Input DTP is a white, Gaussian noise sequence of unit variance, which will be denoted $\{x_n\} \sim N(0,1)$. The corresponding output DTP is corrupted by an additive, small variance random process $\{e_n\} \sim N(0, 0.1)$, modeling the measurement errors. This leads to an observed process $\{o_n\} = \{y_n\} + \{e_n\}$. The number of observed samples is chosen to be low, $N = 100$ because SVM algorithms are expected to work well in low-sized data sets, mainly due to their strong regularization.

Impulsive noise is generated as a sparse sequence, for which 30% of the samples, randomly placed, are of high-amplitude, having the form $\pm 10 + U(0,1)$ [where $U()$ denotes the uniform distribution in the given interval]. The remaining are zero samples. This noise sequence is denoted by $\{j_n\}$. The observations consist of DTP input $\{x_n\}$ and the observed output plus impulsive noise, i.e., $\{o_n\} + \sigma_w\{j_n\}$. Values of σ_w go from -18 to 0 dB.

As insensitivity to outliers is expected to be reduced by the linear zone of the empirical cost, $\varepsilon = 0$ is used, and true orders for both numerator and denominator are introduced into the model. An extremely low value of γ leads to a higher emphasis on minimizing the losses so that overfitting can occur. Then, we select a moderately high $\gamma = 0.1$. The appropriate choice of C can be addressed by considering that according to (35) and (36), the solution is a function of Lagrange multipliers



(a)

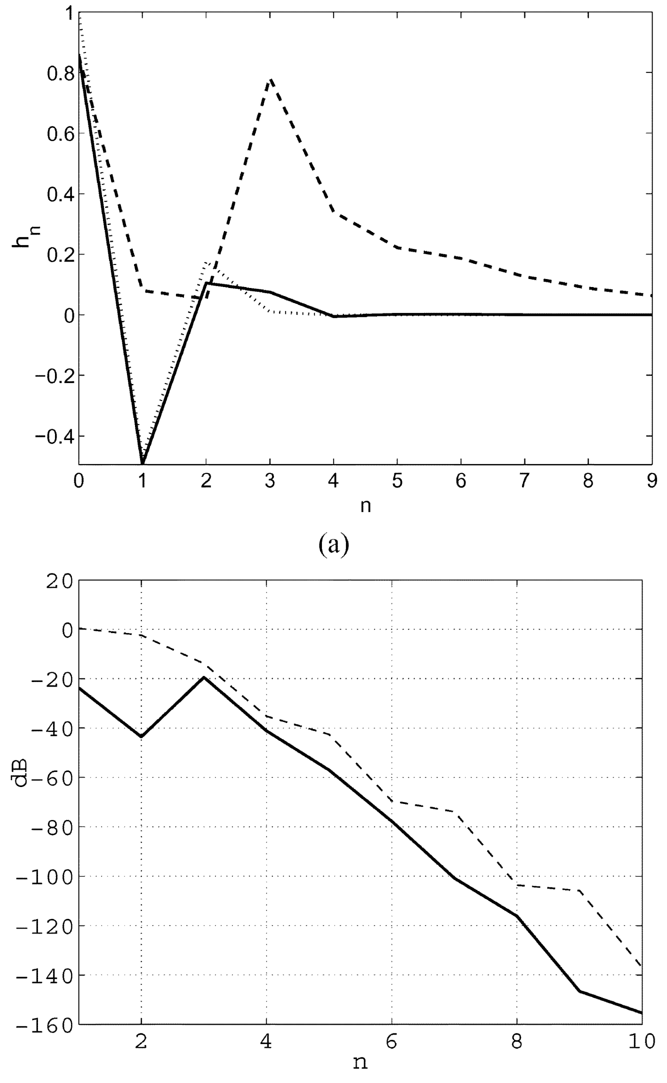
(b)

Fig. 3. (a) Insensitivity of SVM-ARMA to outliers. Histogram of the residuals and control of the outlier impact on the solution by means of C. (b) MSD between true and estimated impulse response for SVM (continuous line) and ARX (dashed line) for different powers of impulsive noise (averaged for 100 realizations).

and data. Besides, (22) reveals that a high amplitude residual, corresponding to an outlier, will produce a high amplitude multiplier, which will distort the solution, but if the maximum value that the multiplier can take is properly limited by C, the impact of the outlier on the solution is weakened. Fig. 3(a) shows that γC should be low enough to exclude residual amplitudes that are greater than a base level. In our case, we fix single value $C = 1$ for all the simulations to be able to deal with high-amplitude and low-amplitude outliers.

The merit figure chosen for comparison purposes is the mean squared deviation (MSD) between the ten first samples of the true and the estimated impulse response

$$\mathrm{MSD} = \frac{1}{10}\sum_{n=0}^{9} |h_n - \hat{h}_n|^2 \qquad (40)$$



(a)

(b)

Fig. 4. (a) Comparison between SVM (solid), ARX (dashed), and true (dotted) impulse responses for $\sigma_w = 0$ dB. (b) Averaged squared deviation between the true impulse response and SVM (solid) ARX (dashed) estimates for the example in (a).

where h_n and $\hat{h}_n$ are the true and the estimated impulse responses, respectively. Results are compared with the conventional LS algorithm (ARX model).

Fig. 3(b) shows MSD for the true and the estimated impulse response for both SVM and ARX algorithms, averaged over 100 realizations. For ARX approach, MSD is very high for high values of impulsive noise power, whereas MSD of SVM remains around -22 dB. Fig. 4(a) shows a realization for 0 dB, and Fig. 4(b) shows the squared deviation $|h_n - \hat{h}_n|^2$ between true and estimated impulse responses along the ten first samples for both ARX and SVM, averaged over 1000 realizations.

B. Feedback's Process Trainer

Feedback's Process Trainer, a classical example, is a variable heating-controller device governed by a binary voltage. The objective proposed in [9] is the identification of an LTIS relating the input (binary voltage) and the output (outlet air temperature).

A delay of three samples is introduced in input DTP. As far as a previous study in [9]shows, this represents the best delay for modeling LTIS. First, $N = 80$ samples of both input and output DTPs are used as training series in order to work with a low number of data, and the next 800 samples are used as the validation set.

SVM-ARMA free parameters were fixed by the following procedure. Initially, 40 samples are used to build a model with a given set of free parameters, and 40 samples are used as a test set to evaluate the prediction error. Explored ranks are $\gamma \in (10^{-10}, 1)$ (logarithmic scale), $e_C \in \left(10^{-5}, \max_n |y_n|\right)$, and $\varepsilon \in (0, 0.10 \max_n |y_n|)$, using 20 points for each parameter exploration. Once appropriate free parameters have been selected for a model, the whole 80-sample set is used to obtain the model coefficients. Comparison with several methods is reported: ARX, output error (OE), and Iv4. A variety of model orders was explored, which are $P \in (1, 10)$ with fixed $Q = 6$, and $Q \in (1, 10)$ with fixed $P = 5$.

Fig. 5 shows the prediction error in the validation set of samples. Both OE and Iv4 provide poor models in terms of prediction error, whereas ARX and SVM-ARMA exhibit better performances. Besides, there is no significant difference between both methods, and in fact, the free parameters of SVM-ARMA that are found ($\varepsilon = 0$, $\gamma \simeq 10^{-8}$ and $e_C \rightarrow \infty$) lead to a cost function that is very close to the ARX one. However, note that, for $P = 7$, SVM-ARMA does not provide the best model. This is due to a failure in the free parameter selection strategy.

Fig. 6 shows the prediction error in the validation set in terms of the free parameters when 40 or 80 length sequences are used to train the model. We can see that the heuristic choice (usually $\gamma = 10^{-8}$) is not the best one, due to the fact that very few samples are available, and in this case, it would lead to a loss of 5 dB in the prediction capabilities. As there are no outliers, $e_C \rightarrow \infty$ is appropriate. In addition, $\varepsilon = 0$ is the best value for both training data sets, and a different value could lead to severe losses.

Therefore, we may conclude that SVM-ARMA provides as good results in this example as the best method (ARX). However, special care must be taken with the free parameter selection strategy, and more sophisticated approaches than those used here should be developed.

Finally, we should note that QP optimization presents some numerical problems in this case because the low value of γ is comparable with the order of magnitude of algorithmic tolerances. Its careful control increases the required time for convergence. This effect could be avoided by using weighted LS to train SVMs, as proposed in [22].

C. Heart Rate and Systolic Blood Pressure

The control of the cardiovascular system by the autonomous nervous system can be studied from its effect on the cardiac rhythm [which causes the so-called heart rate variability (HRV)] and from its effect on the diastolic blood pressure (DBP). In this setting, system identification techniques have been used to model the relationship between HRV and DBP [4]. However, a very common problem of the HRV signal is the presence of anomalous data, which can be due to missing sinusal (normal) beats, ectopic (nonsinusal originated) cardiac beats, or noncardiac artifacts. The manual detection of these beats is a time-consuming task, and a robust analysis method is very desirable. SVM-ARMA system identification properties make this algorithm a useful approach for this application.

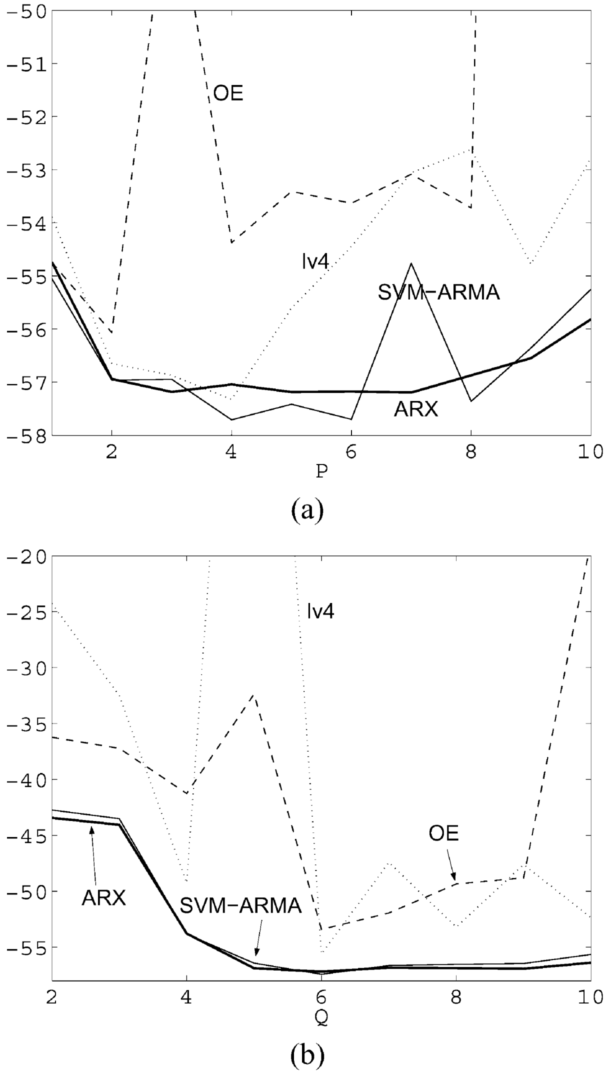


Fig. 5. Results for Feedback's Process Trainer example. (a) Validation prediction error for $Q = 6$ and $P \in (1, 10)$. (b) Validation prediction error for $P = 5$ and $Q \in (1, 10)$.

Fig. 7 shows two simultaneous HRV and DBP signals. The aim consists of obtaining an ARMA model where DBP and HRV are considered its input and output, respectively. The training, test, and validation samples are indicated in this figure. In brief, the training samples of HRV contain a significant number (over 15%) of anomalous measurements, with either too-high or too-low amplitudes, due to electrocardiographic acquisition failure during some time intervals.

A comparison like that in the preceding example is shown in Fig. 8, where different model orders and system identification procedures are shown as a function of the prediction error in the validation set. SVM-ARMA clearly outperforms ARX and Iv4, and it provides almost the same performance for all the presented model orders.

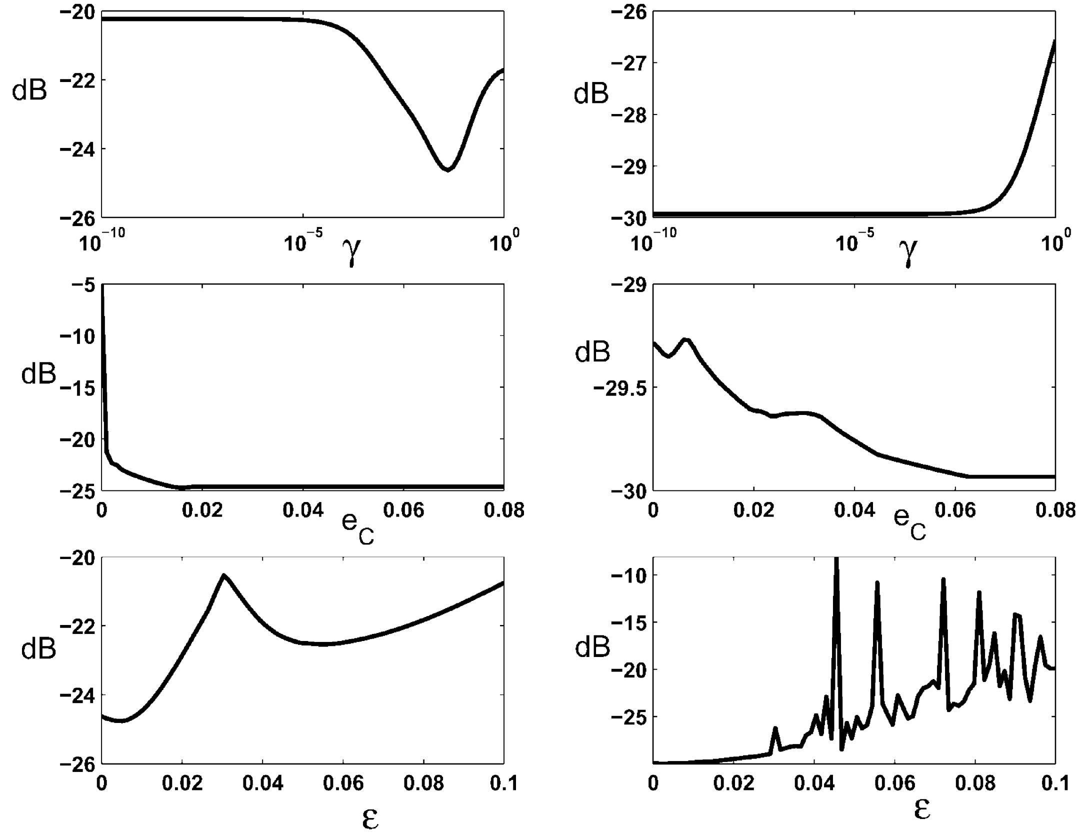


Fig. 6. Free parameter selection for Feedback's Process Trainer. Prediction error in the validation set by using 40 (left) and 80 (right) training samples.

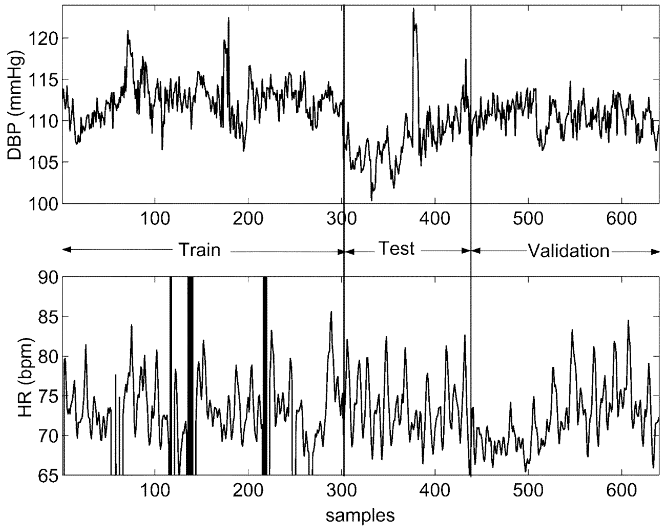


Fig. 7. Signals of blood pressure (up) and heart rate (down) used as training, test, and validation sets for the example in C.

However, we must point out that the OE method reaches a better model than all the other methods, though only occasionally. The reason is that this method gives an output estimate that is better than the extremely noisy observations. This suggests the possibility of an algorithm that can exhibit the advantages of both SVM-ARMA and OE models by using auxiliary variables representing LTIS output and including additional constrains relating the output model and the observations.

V. Conclusion

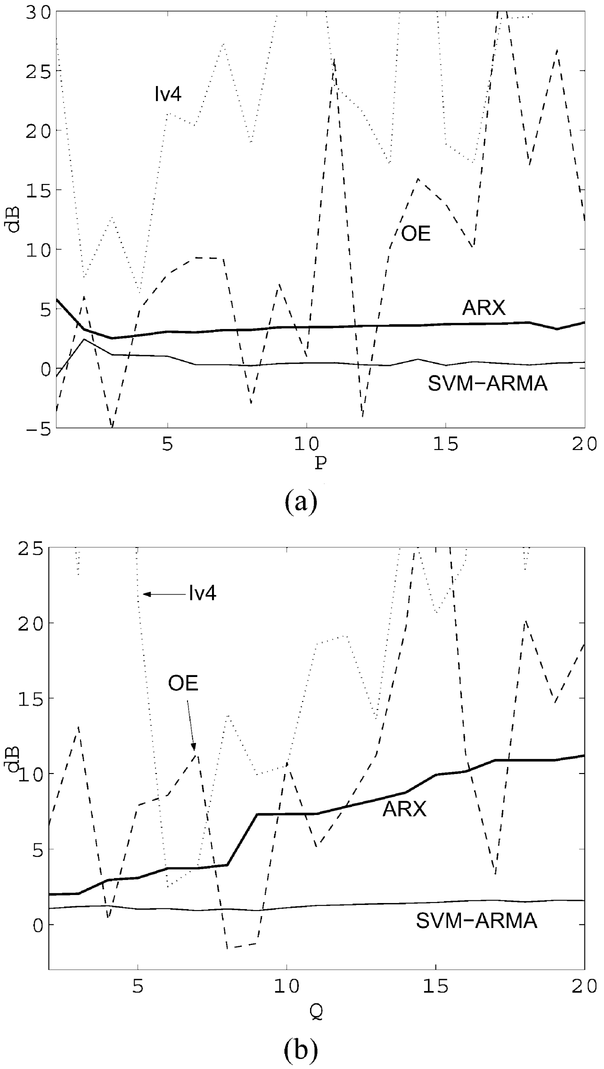
A new approach to estimate ARMA models based on SVM has been introduced. The corresponding algorithm has been formulated, and a statistical interpretation of this SVM-ARMA has been presented. Theoretically, the SVM-ARMA algorithm allows the reduction of residual and data cross-correlation. Simulation and application examples have been used to test the method and to compare it with classical algorithms for ARMA identification, showing that insensitivity to outliers can be achieved by an appropriate choice of the free parameters of SVM-ARMA algorithm.

The application of more efficient and adaptive SVM training schemes [22], [23], as well as the extension of this approach to other signal processing problems, are obvious research lines. Using nonlinear kernels, which allows us to obtain a transformation of the input and the output state, is also an interesting possibility that will lead to robust, nonlinear system identification procedures.

Appendix A
Regularized Cost Function

Our objective is to find the dual problem of the minimization of (21) under constrains (5)–(7). The primal-dual functional is

$$L_{PD}^{C}\left(a_i, b_j, \xi_n, \xi_n^*, \alpha_n, \alpha_n^*, \beta_n, \beta_n^*, \delta_n, \delta_n^*\right)$$
$$= \frac{1}{2}\left(\sum_{i=1}^{P} a_i^2 + \sum_{j=1}^{Q} b_j^2\right)$$



(a)

(b)

Fig. 8. Results for the heart rate and blood pressure example. (a) Validation prediction error for $Q = 5$ and $P \in (1, 20)$. (b) Validation prediction error for $P = 5$ and $Q \in (1, 10)$.

$$
\begin{aligned}
&+ \frac{1}{2\gamma} \sum_{n \in I_1} \left(\xi_n^2 + \xi_n^{*2} \right) + C \sum_{n \in I_2}^{N} \left(\xi_n + \xi_n^* \right) \\
&- \sum_{n \in I_1} \left(\delta_n \xi_n + \delta_n^* \xi_n^* \right) - \sum_{n \in I_2} \left(\beta_n \xi_n + \beta_n^* \xi_n^* \right) \\
&+ \sum_{n=k_o}^{N} \alpha_k \left(y_n - \sum_{i=1}^{P} a_i y_{n-i} \right. \\
&\qquad\qquad \left. - \sum_{j=1}^{Q} b_j x_{n-j+1} - \varepsilon - \xi_n \right) \\
&+ \sum_{n=k_o}^{N} \alpha_n^* \left(-y_n + \sum_{i=1}^{P} a_i y_{n-i} \right. \\
&\qquad\qquad \left. + \sum_{j=1}^{Q} b_j x_{n-j+1} - \varepsilon - \xi_n^* \right)
\end{aligned}
\tag{41}
$$

constrained to

$$\alpha_n, \alpha_n^*, \beta_n, \beta_n^*, \delta_n, \delta_n^* \geq 0. \tag{42}$$

By zeroing the derivative of L_{PD}^C with respect to the primal variables, we obtain the expression for the prediction coefficients as a function of Lagrange multipliers. However, a difference appears when calculating the derivatives with respect to $\xi_n^{(*)}$ with $n \in I_1$, for which

$$\frac{\partial L_{PD}}{\partial \xi_n^{(*)}} = \frac{1}{\gamma} \xi_n^{(*)} - \alpha_n^{(*)} - \delta_n^{(*)} = 0 \quad (n \in I_1). \tag{43}$$

It is clear that $\xi_n^{(*)} \neq 0$ and, necessarily, $\delta_n^{(*)} = 0$, so that

$$\alpha_n^{(*)} = \frac{1}{\gamma} \xi_n^{(*)} \quad (n \in I_1) \tag{44}$$

and denoting by Z the constant value

$$Z = -\frac{\gamma}{2} \sum_{n \in I_2} \left(\alpha_n^2 + \alpha_n^{*2} \right) = -\frac{\gamma}{2} N_2 C^2 \tag{45}$$

it is easy to see that

$$L_D(\cdot) = L_D^C(\cdot) + Z \tag{46}$$

where $L_D^C(\cdot)$ is given by (19). This reveals that both functions reach the minimum at the same values of Lagrange multipliers.

Appendix B
Residuals and Lagrange Coefficients

The analytical relationship presented in (22) must be separately shown for three different cases, according to the different KKT conditions for each one.

Case 1) For $\varepsilon \leq e_n \leq e_C$, we find that $\alpha_n^o > 0$ and $\alpha_n^{*o} = 0$. In the solution, the values reached by the primal and the dual functional are equal. Therefore, taking into account (11) and (12), the following equation holds:

$$
\begin{aligned}
L_D(\boldsymbol{\alpha}^o) = &\frac{1}{2} \sum_{i=1}^{P} a_{\ i}^{o2} + \frac{1}{2} \sum_{j=1}^{Q} b_{\ j}^{o2} - \sum_{n=k_o}^{N} \left(\alpha_n^o - \alpha_n^{o*} \right) y_n \\
&+ \varepsilon \sum_{n=k_o}^{N} \left(\alpha_n^o + \alpha_n^{o*} \right) + \frac{\gamma}{2} \sum_{n=k_o}^{N} \left(\alpha_n^{o2} + \alpha_n^{o*2} \right).
\end{aligned}
\tag{47}
$$

Derivatives of the constraints corresponding to α_n^o vanish at the solution. Thus, using the chain rule, we obtain, by zeroing the derivative of (47)

$$\sum_{i=1}^{P} a_i^o y_{n-i} + \sum_{j=1}^{Q} b_j^o x_{n-j+1} - y_n - \gamma \alpha_n^o - \varepsilon = 0 \tag{48}$$

and hence

$$\alpha_n^o = \frac{1}{\gamma} (e_n - \varepsilon). \tag{49}$$

The derivation is similar for $-\varepsilon \geq e_n \geq -e_C$, just considering that, in this case, $\alpha_n^o = 0$, and $\alpha_n^{*o} > 0$.

Case 2) When $-\varepsilon \geq e_n \geq \varepsilon$, it is straightforward to see that corresponding α_n vanish.

Case 3) When $e_n \geq e_C$, the Lagrange multiplier direction cannot be greater than C, and hence, $\alpha_n^o = C$. Similar considerations can be made for $e_n \leq -e_C$.

References

[1] N. D. Sidiropoulos and G. Z. Dimic, "Blind multiuser detection in W-CDMA systems with large delay spread," in *IEEE Signal Processing Lett.*, vol. 8, Mar. 2001, pp. 87–89.

[2] S. Haykin and A. Steinhart, *Adaptive Radar Detection and Estimation*. New York: Wiley, 1992.

[3] S. de Waele and P. M. T. Broersen, "Modeling radar data with time series models," in *Proc. EUSIPCO*, Tampere, Finland, Sept. 2000, Ref. 116.

[4] Task force of the European Society of Cardiology and the North American Society of Pacing and Electrophysiology, "Heart rate variability. Standards of measurement, physiological interpretation and clinical use," *Eur. Heart J.*, vol. 17, pp. 354–381, Mar. 1996.

[5] C. Vlachopoulos and M. O'Rourke, "Genesis of normal and abnormal arterial pulse," *Current Prob. Cardiol.*, vol. 25, pp. 297–368, 2000.

[6] J. D. Bronzino, *The Biomedical Engineering Hadbook*. Boca Raton, FL: CRC/IEEE, 1995.

[7] K. R. Müller *et al.*, "Predicting time series with support vector machines," in *Advances in Kernel Methods—Support Vector Learning*, B. Scholkopf *et al.*, Eds. Cambridge, MA: MIT Press, 1999, pp. 243–254.

[8] S. Mukherjee, E. Osuna, and F. Girosi, "Nonlinear prediction of chaotic time series using support vector machines," in *Proc. IEEE NNSP*, Amelia Island, FL, Sept. 1997, pp. 511–519.

[9] L. Ljung, *System Identification. Theory for the User*, 2nd ed. Englewood Cliffs, NJ: Prentice-Hall, 1999.

[10] T. C. Hsia, *Identification: Least Squares Methods*. Lexington, MA: Lexington Books, 1977.

[11] T. Söderström and P. Stoica, *System Identification*. Englewood Cliffs, NJ: Prentice-Hall, 1987.

[12] J. Li, K. Miyashita, T. Kato, and S. Miyazaki, "GPS time series modeling by autoregressive moving average method: Application to the crustal deformation in central Japan," *Earth Planets Space*, vol. 52, pp. 155–162, Feb. 2000.

[13] V. Vapnik, *The Nature of Statistical Learning Theory*. New York: Springer-Verlag, 1995.

[14] B. Schölkopf and K. Sung, "Comparing support vector machines with Gaussian kernels to radial basis function classifiers," *IEEE Trans. Signal Processing*, vol. 45, pp. 2758–2765, Nov. 1997.

[15] M. Pontil and A. Verri, "Support vector machines for 3D object recognition," *IEEE Trans. Pattern Anal. Machine Intell.*, vol. 20, pp. 637–646, June 1998.

[16] C. J. C. Burges, "A tutorial on support vector machines for pattern recognition," *Data Mining Knowl. Discovery*, vol. 2, pp. 1–32, Jan. 1998.

[17] A. J. Smola and B. Schölkopf, "A tutorial on support vector regression," ESPRIT, Neural and Computational Learning Theory NeuroCOLT2 NC2-TR-1998-030, 1998.

[18] A. N. Tikhonov and V. Y. Arsenin, *Solution to Ill-Posed Problems*. Washington, DC: V.H. Winston, 1977.

[19] D. G. Luenberger, *Linear and Nonlinear Programming*. Reading, MA: Addison-Wesley, 1984.

[20] N. Cristianini and J. S. Taylor, *An Introduction to Support Vector Machines and Other Kernel-Based Learning Methods*. Cambridge, U.K.: Cambridge Univ. Press, 2000.

[21] P. J. Huber, "Robust statistics: A review," *Ann. Stat.*, vol. 43, pp. 1041–67, 1972.

[22] A. Navia-Vázquez, F. Pérez-Cruz, A. Artés-Rodríguez, and A. R. Figueiras-Vidal, "Weighted least squares training of support vector classifiers leading to compact and adaptive schemes," *IEEE Trans. Neural Networks*, vol. 12, pp. 1047–1059, Sept. 2001.

[23] F. Pérez-Cruz, A. Navia-Vázquez, P. L. Alarcón-Diana, and A. Artés-Rodríguez, "An IRWLS procedure for SVR," in *Proc. EUSIPCO*, Tampere, Finland, Sept. 2000, Ref. 156.

José Luis Rojo-Álvarez (M'01) received the telecom engineer degree from University of Vigo, Vigo, Spain, in 1996 and the Ph.D. degree in telecommunication from the Polytechnical University of Madrid, Madrid, Spain, in 2000.

He is an Assistant Professor with the Department of Signal Theory and Communications, University Carlos III, Madrid. His main research interests include statistical learning theory, digital signal processing, and complex system modeling, with applications both to digital communications and to cardiac signal processing. He has published work on cardiac arrhythmia and arrhythmia-genesis mechanisms, robust analysis, and echocardiographic image and hemodynamic function evaluation.

Manel Martínez-Ramón (M'00) received the telecom engineer degree from Politechnical University of Catalunya, Barcelona, Spain, in 1994 and the Ph.D. degree in telecommunication from the University Carlos III, Madrid, Spain, in 1999.

He is a Visitant Professor with the Department of Signal Theory and Communications, University Carlos III. His main research interests include statistical learning theory, digital signal processing, and communications, with applications to digital communications, areas of research on he has published several works.

Mario de Prado-Cumplido (S'02) was born in Madrid, Spain, on December 1977. He received the M.Sc. degree in telecommunication engineering from the Universidad Politécnica de Madrid in 2000. He is currently pursuing the Ph.D. degree at Universidad Carlos III, Madrid, where he is doing research on machine learning algorithms, feature selection, and probability density estimation techniques and their application to biomedical problems, with particular attention to cardiovascular pathologies.

Antonio Artés-Rodríguez (M'89–SM'01) was born in Alhama de Almería, Spain, in 1963. He received the Ingeniero de Telecomunicación and Doctor Ingeniero de Telecomunicación degrees, both from the Universidad Politécnica de Madrid, Spain, in 1988 and 1992, respectively.

He is now a Professor with the Departamento de Teoría de la Señal y Comunicaciones, Universidad Carlos III, Madrid. His research interests include detection, estimation, and statistical learning methods and their application to signal processing and communication.

Aníbal R. Figueiras-Vidal (S'74–M'76–SM'84) received the Telecom. Engineer degree from Universidad Politécnica de Madrid, Madrid, Spain, in 1973 (ranked number 1; National Award to graduation) and the Doctor degree (Honors) from Universidad Politécnica de Barcelona, Barcelona, Spain, in 1976.

He is a Professor of signal theory and communications with Universidad Carlos III, Madrid. His research interests are digital signal processing, digital communications, neural networks, and learning theory. He has (co)authored more than 300 journal and conference papers in these areas.

Dr. Figueiras received an "Honoris Causa" Doctorate degree in 1999 from Universidad de Vigo, Vigo, Spain. He is currently General Secretary of the Academy of Engineering of Spain.