Proc. of the 5th World Multiconference on Systemics, Cybernetics and Informatics (SCI 2001).
Orlando, Florida, Jul. 2001, pg. 254-259

01-28

Bayesian Interpretation of Least Squares Support Vector Machines for Financial Time Series Prediction

Tony Van Gestel[1], Johan A.K. Suykens[1]*, Gert Lanckriet[1], Annemie Lambrechts[1], Dirk-Emma Baestaens[2], Bart De Moor[1] and Joos Vandewalle[1]

[1] Dept. Electrical Engineering ESAT-SISTA, Katholieke Universiteit Leuven
Kasteelpark Arenberg 10, B-3001 Leuven, Belgium
[2] Financial Markets Research, Fortis Bank Brussels
Warandeberg 3, B-1000 Brussels, Belgium

ABSTRACT

For financial time series, the generation of error bars on the point prediction is important in order to estimate the corresponding risk. In Least Squares Support Vector Machines (LS-SVMs) for nonlinear function estimation, the training problem is presented and formulated so as to obtain a set of linear equations in the dual space, while the training problem of Support Vector Machines involves a (convex) Quadratic Programming (QP) problem in the dual space. In this paper, a Bayesian interpretation is related to the LS-SVM regression formulation within the evidence framework, in a similar way as it has been done for multilayer perceptrons. The LS-SVM formulation allows to derive analytic expressions in the feature space and practical expressions are obtained in the dual space by replacing inner product with the related positive definite kernel function using Mercer's theorem. We illustrate the method on the one step ahead prediction of the Standard & Poor's 500 Financials stock index (S&PTFIN). Significant out-of-sample sign predictions are obtain with respect to the Pesaran-Timmerman test for directional accuracy, while trading on the Sharpe Ratio allows to improve the risk adjusted return.

Keywords: Financial Time Series Prediction, Least Squares Support Vector Machines, Bayesian Inference

1. INTRODUCTION

Motivated by the universal approximation property of multilayer perceptrons (MLPs), neural networks have been applied to model and predict financial time series [1, 6, 10, 15]. The focus of many nonlinear forecasting methods [2, 9, 19] is on predicting of next points of a time series, while the error bars on the prediction are typically less important. In financial time series the noise is often larger than the underlying deterministic signal and one also wants to know the error bars on the prediction. Combining the prediction and corresponding uncertainty, the return/risk ratio of Sharpe Ratio (SR) allows to compare different investments per unit of risk.

In [12, 13], the Bayesian evidence framework was successfully applied to MLPs so as to model nonlinear relations and to infer output probabilities on the corresponding predictions. However, there are drawbacks to the practical design of MLPs like the non-convex optimization problem and the choice of the number of hidden units. In Support Vector Machines (SVMs), the regression problem is formulated and represented as a convex Quadratic Programming (QP) problem [5, 18, 24, 26]. Basically, the SVM regressor maps the inputs into a higher dimensional feature space in which a linear regressor is constructed by minimizing an appropriate cost function. Using Mercer's theorem, the regressor is obtained by solving a finite dimensional QP problem in the dual space avoiding explicit knowledge of the high dimensional mapping and using only the related positive-definite kernel function.

In Least Squares Support Vector Machines (LS-SVMs) [20, 21] for nonlinear classification and regression, the use of the least squares cost function with equality constraints allows to obtain a linear Karush-Kuhn-Tucker system in the dual space. This formulation can also be related to regularization networks [7, 10]. When no bias term is used in the LS-SVM formulation, as proposed in kernel ridge regression [16], the expressions in the dual space correspond to Gaussian Processes [27]. However, the additional insight of using the feature space has been used in kernel PCA [17], while the use of equality constraints and the primal-dual interpretations of LS-SVMs have allowed to make extensions towards recurrent neural networks [22] and nonlinear optimal control [23]. In this paper, a Bayesian interpretation is related to LS-SVM regression [25] in order to infer the point prediction and the corresponding error bars. While error bars are obtained for MLPs using a local quadratic approximation to the non-convex cost function, no approximation has to be made for LS-SVMs since a quadratic cost function is used. The resulting return/risk or Sharpe Ratio is then used to trade on the S&P 500 Financials stock index.

This paper is organized as follows. The LS-SVM regression formulation is reviewed in Section 2. In Section 3 a probabilistic framework is related to the LS-SVM regression formulation by applying Bayes' rule in the feature space. Expressions in the dual space for the probabilistic interpretation of the prediction are derived in Section 4. The daily one step ahead prediction of the S&P 500 Financials stock index is discussed in Section 5.

*Corresponding author. E-mail: {tony.vangestel, johan.suykens}@esat.kuleuven.ac.be

2. LEAST SQUARES SUPPORT VECTOR MACHINES

In Support Vector Machines [5, 16, 18, 21, 24, 26] for nonlinear regression, the nonlinear function $y_i = f(x_i) + e_i$, disturbed by additive noise e_i, is assumed to be of the following form

$$y_i = w^T \varphi(x_i) + b + e_i. \tag{1}$$

For financial time series, the output $y_i \in \mathbb{R}$ is typically a return of an asset or exchange rate at the time index i. The input vector $x_i \in \mathbb{R}^n$ may consists of lagged returns, volatility measures and macro-economic explanatory variables. The mapping $\varphi(\cdot) : \mathbb{R}^n \to \mathbb{R}^{n_f}$ is a nonlinear function that maps the input vector x into a higher (possibly infinite) dimensional feature space $\mathbb{R}^{n_f}$. However, the weight vector $w \in \mathbb{R}^{n_f}$ and the function $\varphi(\cdot)$ are never calculated explicitly. Instead, Mercer's condition $\varphi(x_i)^T \varphi(x) = K(x_i, x)$ is applied to relate the function $\varphi(\cdot)$ with the positive definite kernel function K. For $K(x_i, x)$ one typically has the following choices:

- $K(x_i, x) = x_i^T x$ (linear SVM);
- $K(x_i, x) = (x_i^T x + 1)^d$ (polynomial SVM of degree d);
- $K(x_i, x) = \exp(-||x - x_i||_2^2/\sigma^2)$ (SVM with RBF-kernel), where σ is a tuning parameter.

In the sequel of this paper, we will focus on the use of an RBF-kernel.

In Least Squares Support Vector Machines (LS-SVMs) [16, 20, 21], the model parameters w and b of the nonlinear function $f(x)$ are inferred from given data $D = \{(x_i, y_i)\}_{i=1}^N$ by solving the following least squares minimization problem with ridge regression:

$$\min_{w,e} \mathcal{J}_1(w, e) = \frac{\mu}{2} w^T w + \frac{1}{2} \sum_{i=1}^N \zeta_i e_i^2, \tag{2}$$

subject to the equality constraints

$$y_i = w^T \varphi(x_i) + b + e_i, \qquad i = 1, \ldots, N. \tag{3}$$

The meaning of the hyperparameters μ and ζ_i ($i = 1, \ldots, N$) will become clear in the next Sections. When $\zeta_i = \zeta$ for $i = 1, \ldots, N$, the minimization problem (2)-(3) corresponds to the standard LS-SVM formulation [20, 21]. For non-constant ζ_i, the error term corresponds to a weighted least squares cost function. In the probabilistic interpretation of Section 3, non-constant ζ_i are related to the time varying volatility $1/\sqrt{\zeta_i}$ of the time series [3, 4]. Substitution of (3) into (2) yields the following weighted least squares cost function with ridge regression:

$$\min_{w,b} \mathcal{J}_1(w, b) = \mu E_W + \textstyle\sum_{i=1}^N \zeta_i E_{D,i} \tag{4}$$

with

$$E_W = \tfrac{1}{2} w^T w, \tag{5}$$

$$E_{D,i} = \tfrac{1}{2} e_i^2 = \tfrac{1}{2}(y_i - w^T \varphi(x_i) - b)^2. \tag{6}$$

This formulation will be used in the next Section. Notice that the solution w, b of (2)-(3) only depends on the ratio's $\gamma_i = \frac{\zeta_i}{\mu}$ ($i = 1, \ldots, N$), as introduced in [20, 21].

To solve the minimization problem (2)-(3), one constructs the Lagrangian $\mathcal{L}_1(w, b, e; \alpha) = \mathcal{J}_1(w, e) - \sum_{i=1}^N \alpha_i (w^T \varphi(x_i) + b + e_i - y_i)$, where $\alpha_i \in \mathbb{R}$ are the Lagrange multipliers (also called support values). The conditions for optimality are given by:

$$\begin{cases} \frac{\partial \mathcal{L}_1}{\partial w} = 0 \to w = \sum_{i=1}^N \alpha_i \varphi(x_i) \\ \frac{\partial \mathcal{L}_1}{\partial b} = 0 \to \sum_{i=1}^N \alpha_i = 0 \\ \frac{\partial \mathcal{L}_1}{\partial e_i} = 0 \to \alpha_i = \gamma_i e_i, \quad i = 1, \ldots, N \\ \frac{\partial \mathcal{L}_1}{\partial \alpha_i} = 0 \to b = y_i - w^T \varphi(x_i) - e_i, \quad i = 1, \ldots, N. \end{cases} \tag{7}$$

As in standard SVMs, w and $\varphi(x_i)$ are never calculated and by elimination of w and e the following linear system is obtained [20, 21, 25]:

$$\left[\begin{array}{c|c} 0 & 1_v^T \\ \hline 1_v & \Omega + D_\gamma^{-1} \end{array}\right] \left[\begin{array}{c} b \\ \hline \alpha \end{array}\right] = \left[\begin{array}{c} 0 \\ \hline y \end{array}\right], \tag{8}$$

with[1] $y = [y_1; \ldots; y_N]$, $1_v = [1; \ldots; 1]$, $e = [e_1; \ldots; e_N]$, $\alpha = [\alpha_1; \ldots; \alpha_N]$, $D_\gamma = \mathrm{diag}([\gamma_1, \ldots, \gamma_N])$ and where Mercer's condition [5, 18, 26] is applied within the Ω matrix

$$\Omega_{ij} = \varphi(x_i)^T \varphi(x_j) = K(x_i, x_j). \tag{9}$$

In the optimum we have $w = \sum_{i=1}^N \alpha_i \varphi(x_i)$ and the LS-SVM regressor is obtained by applying the Mercer condition:

$$f(x) = w^T \varphi(x) + b = \textstyle\sum_{i=1}^N \alpha_i K(x, x_i) + b. \tag{10}$$

Efficient algorithms exist in numerical linear algebra for solving large scale systems. By reformulating the linear system (8) into two linear systems with positive definite data matrices as in [21], iterative methods can be applied such as, e.g., the Hestenes-Stiefel conjugate gradient algorithm. The sparseness property of standard SVMs [7, 24, 26] is lost by the introduction of the 2-norm. However, sparseness can be obtained by sequentially pruning the support value spectrum [21].

3. BAYESIAN INTERPRETATION IN THE FEATURE SPACE

Given the data points $D = \{(x_i, y_i)\}_{i=1}^N$ and the hyperparameters μ and $\zeta_{1:N} = [\zeta_1, \ldots, \zeta_N]$ of the model $\mathcal{H}$ (LS-SVM with kernel function K), we obtain the model parameters by maximizing the posterior $P(w, b|D, \log\mu, \log\zeta_{i,1:N}, \mathcal{H})$. Application of Bayes' rule at the first level of inference [2, 13] gives:

$$P(w, b|D, \log\mu, \log\zeta_{1:N}, \mathcal{H}) = \frac{P(D|w, b, \log\mu, \log\zeta_{1:N}, \mathcal{H}) P(w, b|\log\mu, \log\zeta_{1:N}, \mathcal{H})}{P(D|\log\mu, \log\zeta_{1:N}, \mathcal{H})}, \tag{11}$$

where the evidence $P(D|\log\mu, \log\zeta_{i,1:N}, \mathcal{H})$ is a normalizing constant such that integration over all possible w and b parameters will give probability 1.

[1] The matlab notation $[X_1; X_2]$ is used, where $[X_1; X_2] = [X_1^T\ X_2^T]^T$. The diagonal matrix $D_a = \mathrm{diag}(a) \in \mathbb{R}^{N \times N}$ has diagonal elements $D_a(i, i) = a(i)$, $i = 1, \ldots, N$, with $a \in \mathbb{R}^N$.

We take the prior $P(w,b|\log\mu,\log\zeta_{1:N},\mathcal{H})$ independent of the hyperparameters ζ_i, i.e., $P(w,b|\log\mu,\log\zeta_{1:N},\mathcal{H}) = P(w,b|\log\mu,\mathcal{H})$. Both w and b are assumed to be independent. The weight parameters w are assumed to be Gaussian distributed with zero mean: $P(w|\log\mu,\mathcal{H}) = \left(\frac{\mu}{2\pi}\right)^{\frac{n_f}{2}} \exp(-\frac{\mu}{2}w^Tw)$. This means that a priori we do not expect a functional relation between the feature vector $\varphi(x)$ and the observation y. Before the data are available, the most likely model has zero weights $w_k = 0$ $(k = 1,\ldots,n_f)$, corresponding to the Efficient Market Hypothesis. A uniform distribution for the prior on b is taken, which can also be approximated as a Gaussian distribution $P(b|\log\sigma_b,\mathcal{H}) = \frac{1}{\sqrt{2\pi\sigma_b^2}}\exp(-\frac{b^2}{2\sigma_b^2})$, with $\sigma_b \to \infty$. The prior $P(w,b|\log\mu,\log\zeta_{1:N},\mathcal{H})$ in (11) then becomes:

$$P(w,b|\log\mu,\mathcal{H}) \quad \propto \quad \left(\frac{\mu}{2\pi}\right)^{\frac{n_f}{2}} \exp(-\frac{\mu}{2}w^Tw). \quad (12)$$

The negative logarithm of the prior (12) corresponds to the regularization term μE_W in (5).

We take the likelihood of the observed data $D = \{(x_i,y_i)\}_{i=1}^N$ independent of the hyperparameter μ and assume that all data points (x_i,y_i) are independent:

$$\begin{aligned} P(D|w,b,\log\zeta_{1:N},\mathcal{H}) &= \textstyle\prod_{i=1}^N P(x_i,y_i|w,b,\log\zeta_i,\mathcal{H}) \\ &= \textstyle\prod_{i=1}^N P(y_i|x_i,w,b,\log\zeta_i,\mathcal{H})P(x_i|w,b,\log\zeta_i,\mathcal{H}) \\ &\propto \textstyle\prod_{i=1}^N P(y_i|x_i,w,b,\log\zeta_i,\mathcal{H}), \end{aligned} \quad (13)$$

where the probability $P(x_i|w,b,\log\zeta_i,\mathcal{H})$ is independent from the model $\mathcal{H}$, i.e., $P(x_i|w,b,\log\zeta_i,\mathcal{H}) = P(x_i)$. The probability $P(x_i)$ is assumed to be constant. Assuming that the additive noise e_i is drawn from a Gaussian distribution with variance $1/\zeta_i$, we have:

$$\begin{aligned} P(y_i|x_i,w,b,\log\zeta_i,\mathcal{H}) &= \sqrt{\tfrac{\zeta_i}{2\pi}}\exp(-\tfrac{\zeta_i}{2}e_i^2) \\ &= \sqrt{\tfrac{\zeta_i}{2\pi}}\exp(-\tfrac{\zeta_i}{2}(y_i - w^T\varphi(x_i) - b)^2). \end{aligned} \quad (14)$$

Other distributions with heavier tails like, e.g., the student-t distribution, are sometimes assumed in the literature; a Gaussian distribution with time-varying variance ζ_i^{-1} is used here [3].

Substituting (12) and (14) into (11) and neglecting all constants, Bayes' rule at the first level of inference gives:

$$\begin{aligned} &P(w,b|D,\log\mu,\log\zeta_{1:N},\mathcal{H}) \\ &\propto \exp(-\tfrac{\mu}{2}w^Tw - \textstyle\sum_{i=1}^N \tfrac{\zeta_i}{2}e_i^2) = \exp(-\mathcal{J}_1(w,b)), \end{aligned} \quad (15)$$

with the LS-SVM cost function $\mathcal{J}_1(w,b)$ defined in (4). Hence, the maximum a posteriori estimates w_{MP} and b_{MP} are obtained by minimizing the negative logarithm of (15), corresponding to the minimization problem (2)-(3). As already explained in the previous Section, the least squares problem (2)-(3) is not explicitly solved in w and b. Instead, the linear system (8) in α and b is solved in the dual space.

We discuss now an alternative representation of the cost function (4) and the likelihood (15), which will be used in the next Subsection. By observing that the cost function (4) is a quadratic cost function, the posterior $P(w,b|D,\log\mu,\log\zeta_{i,1:N},\mathcal{H})$ can also be written as the Gaussian distribution:

$$\begin{aligned} &P(w,b|D,\log\mu,\log\zeta_{1:N},\mathcal{H}) \\ &\quad = \frac{1}{\sqrt{(2\pi)^{n_f+1}\det Q}}\exp(-\frac{1}{2}g^TQ^{-1}g), \end{aligned} \quad (16)$$

with $g = [w - w_{MP}; b - b_{MP}]$ and $Q = \mathrm{covar}(w,b) = \mathcal{E}(g^Tg)$, where the expectation is taken with respect to w and b. The covariance matrix Q is related to the Hessian H of the LS-SVM cost function (4):

$$Q = H^{-1} = \begin{bmatrix} \frac{\partial^2\mathcal{J}_1}{\partial w^2} & \frac{\partial^2\mathcal{J}_1}{\partial w\partial b} \\ \frac{\partial^2\mathcal{J}_1}{\partial b\partial w} & \frac{\partial^2\mathcal{J}_1}{\partial b^2} \end{bmatrix}^{-1}. \quad (17)$$

3. MODERATED OUTPUT OF THE LS-SVM IN THE DUAL SPACE

The uncertainty on the estimated model parameters results into an additional uncertainty for the one step ahead prediction $\hat{y}_{MP,N+1} = w_{MP}^T\varphi(x) + b_{MP}$, where the input vector $x \in \mathbb{R}^n$ may be composed of lagged returns $y_N, y_{N-1},\ldots$ and of other explanatory variables available at the time index N. By marginalizing over the nuisance parameters w and b [12, 13, 25] one obtains that the prediction $\hat{y}_{N+1}$ is Gaussian distributed with mean

$$\hat{y}_{MP,N+1} = z_{MP} = w_{MP}^T\varphi(x) + b_{MP} \quad (18)$$

and variance

$$\sigma_{\hat{y}_{N+1}}^2 = \zeta_{N+1}^{-1} + \sigma_z^2. \quad (19)$$

The variance is thus composed of two terms: the first term ζ_{N+1}^{-1} corresponds to the volatility of the noise e_{N+1} in the next time step $N+1$. Different volatility models can be constructed. In this paper, we use a moving average approach based on the last 20 business days. The second term σ_z^2 is due to the Gaussian uncertainty on the estimated model parameters w and b in the linear transform $z = w^T\varphi(x)+b$. We will now derive expressions for z_{MP} and σ_z^2 in the dual space.

Expression for z_{MP}

Now, we further discuss the expressions for the mean $z_{MP} = \mathcal{E}(z)$ and the variance $\sigma_z^2 = \mathcal{E}[(z - z_{MP})^2]$, taking the expectation with respect to the Gaussian distribution over the model parameters w and b. The mean z_{MP} is obtained as: $z_{MP} = \mathcal{E}\{z\} = w_{MP}^T\varphi(x) + b_{MP}$, with $w = \sum_{i=1}^N \alpha_i\varphi(x_i)$ [20, 21, 25]. Applying the Mercer condition, we obtain

$$\boxed{z_{MP} = \sum_{i=1}^N \alpha_i K(x,x_i) + b_{MP}} \quad (20)$$

Expression for σ_z^2

Since z is a linear transformation of the Gaussian distributed model parameters w and b, the variance σ_z^2 in the feature space is given by

$$\begin{aligned} \sigma_z^2 &= \mathcal{E}\{(z - z_{MP})^2\} \\ &= \mathcal{E}\{[(w^T\varphi(x)+b) - (w_{MP}^T\varphi(x) + b_{MP})]^2\} \\ &= \psi(x)^TH^{-1}\psi(x), \end{aligned} \quad (21)$$

with $\psi(x) = [\varphi(x); 1]$. The computation for σ_z^2 can be obtained without explicit knowledge of the mapping $\varphi(\cdot)$. Using matrix algebra and replacing inner products by the related kernel function, the expression for σ_z^2 in the dual space is derived in [25]:

$$\begin{aligned}\sigma_z^2 = &\theta(x)^T U_G Q_D U_G^T \theta(x) U^T + \tfrac{1}{\mu s_\zeta^2} 1_v^T D_\zeta \Omega D_\zeta 1_v \\ &- \tfrac{2}{s_\zeta} \theta(x)^T U_G Q_D U_G^T \Omega D_\zeta 1_v + \tfrac{\zeta}{s_\zeta} \mu^{-1} \theta(x)^T D_\zeta 1_v \\ &+ \tfrac{1}{s_\zeta} + \tfrac{1}{s_\zeta^2} 1_v^T D_\zeta \Omega U_G Q_D U_G^T \Omega D_\zeta 1_v + \tfrac{1}{\mu} K(x,x)\end{aligned} \quad (22)$$

with $Q_D = (\mu I_{N_{eff}} + D_G)^{-1} - \mu^{-1} I_{N_{eff}}$ and the scalar $s_\zeta = \sum_{i=1}^N \zeta_i$. The vector $\theta(x) \in \mathbb{R}^N$ and the matrices $U_G \in \mathbb{R}^{N \times N_{eff}}$ and $D_G \in \mathbb{R}^{N_{eff} \times N_{eff}}$ are defined as follows: $\theta_i(x) = K(x, x_i), i = 1, \ldots, N$; $U_G(:, i) = (\nu_{G,i} \Omega \nu_{G,i})^{\frac{1}{2}} \nu_{G,i}, i = 1, \ldots, N_{eff} \leq N - 1$ and $D_G = \text{diag}([\lambda_{G,1}, \ldots, \lambda_{G,N_{eff}}])$, where $\nu_{G,i}$ and $\lambda_{G,i}$ are the solution to the eigenvalue problem [25]:

$$(D_\zeta - \tfrac{1}{s_\zeta} D_\zeta 1_v 1_v^T D_\zeta) \Omega \nu_{G,i} = \lambda_{G,i} \nu_{G,i}, \quad (23)$$

$i = 1, ..., N_{eff} \leq N - 1$. The number of non-zero eigenvalues is denoted by $N_{eff} < N$. The matrix $D_\zeta = \text{diag}([\zeta_1, ..., \zeta_N]) \in \mathbb{R}^{N \times N}$ is a diagonal matrix with diagonal elements $D_\zeta(i,i) = \zeta_i$. Further details can be found in [25].

5. PREDICTING THE S&P 500 FINANCIALS INDEX

We apply the Bayesian interpretation of the LS-SVM to the prediction of the Standard and Poor's 500 Financials (S&PTFIN) index from till . The first 600 points are used for training, while the 200 subsequent data points are used for validation. These 200 points were used to select the inputs or explanatory variables of the inputs using the p-value of the Pesaran-Timmerman test for directional accuracy [14]. We selected $\gamma = 800$ and $\sigma = 110$ corresponding to a Pesaran-Timmerman statistic (PTstat) of 2.23 and corresponding p-value of 0.025. The following inputs were selected:

- S&P 500 Financials Index: lags -1, -2, -3, ..., -9 and moving average over 20 and 40 days
- Dow Jones Industrial Index: lags -1, -2, ..., -5 and moving average over 10 days
- US$-Yen Exchange Rate: lags -2, -3, ..., -9.

All data were retrieved from Datastream. All inputs were normalized to zero mean and unit variance [2], while the output was normalized to unit variance for convenience.

The inputs and hyperparameters γ and σ were then kept fixed and the LS-SVM regressor is validated on the out-of-sample dataset consisting of 1493 data points covering the period from 22/03/1993 till 09/12/1998 (using the dd/mm/yy convention), which includes the Asian crisis in 1998. To model possible time varying relations of the financial markets, the predictions were made using the rolling approach, re-estimating the model parameters w and b after 200 data points. The corresponding Percentage of Correct Sign Predictions (PCSP) and the Pesaran-Timmerman test statistic with corresponding p-values are reported in Table 1. The performance of the LS-SVM with RBF-kernel (RBF-LS-SVM) is here compared with linear regressor autoregressive model with exogenous inputs (ARX), which is estimated using Ordinary Least Squares using exactly the same inputs. While both models yield a significant result with respect to the Directional Accuracy test, the nonlinear model gives a better performance than the linear ARX model. Relating the sign predictions to a classifier performance, the McNemar test [8] can be applied to compare the RBF-LS-SVM model with the ARX model. We obtained a test statistic of 2.34, which corresponds to a p-value of 0.0627.

We also illustrate the use of taking different noise levels and corresponding weightings into account (RBF-LS-SVM$_w$) and ARX-model (ARX$_w$), as has also been applied in [24]. Different algorithms exists for modeling the volatility of financial time series, related to the variance of the additive noise [3, 4]. Here we use a simple approach taking the 10 days moving average estimate of volatility [4]. Since the point prediction of the LS-SVM only depends on the ratio $\gamma = \zeta/\mu$ (with $\gamma = 800$), we first estimated ζ from the validation set in order to obtain $\mu = \zeta/\gamma$. We

Directional Accuracy	PCSP	PTstat	p-value
RBF-LS-SVM	55.94%	3.69	0.00022
RBF-LS-SVM$_w$	56.43%	3.92	0.00009
ARX	54.97%	3.02	0.00252
ARX$_w$	55.18%	3.08	0.00203

Table 1: Out-of-sample test set performances obtained on the one step ahead prediction of S&PTFIN with different models: RBF-LS-SVM, RBF-LS-SVM$_w$, ARX and ARX$_w$. The Directional Accuracy is assessed by means of the Percentage of Correct Sign Predictions (PCSP), the Pesaran-Timmerman test statistic (PTstat) and the corresponding p-value.

Investment Strategy 1	SR_1	Re_1	Ri_1
1. RBF-LS-SVM	1.1296	17.36	15.35
2. RBF-LS-SVM$_w$	1.3757	20.79	15.36
3. ARX	0.9437	14.25	15.11
4. ARX$_w$	1.0755	15.74	14.63
5. Buy&Hold	0.9498	18.24	19.21
Investment Strategy 2	SR_2	Re_2	Ri_2
6. RBF-LS-SVM$_w$	1.6354	14.56	8.91
7. ARX$_w$	1.3673	11.15	8.15

Table 2: Annualized Returns (Re), Risk (Ri) and corresponding Sharpe Ratio (SR) of Investment Strategy 1 and Investment Strategy 2 (transaction cost of 0.1%) on the test set comparing the RBF-LS-SVM, RBF-LS-SVM$_w$, ARX and ARX$_w$ with a Buy and Hold strategy.

involves the solution of a QP-problem, the LS-SVM regressor is obtained from a linear Karush-Kuhn-Tucker system. The least squares cost function also allows to obtain analytical expressions for the variance, while both in SVMs and MLPs use a local quadratic approximation to the cost function in order to estimate the Hessian. We applied the nonlinear LS-SVM to the daily one step ahead prediction of the S&P 500 Financials stock index. Significant out-of-sample test set predictions were obtained for the Directional Accuracy of the Pesaran-Timmerman test. The corresponding uncertainty of the predictions is used in trading on the corresponding risk-adjusted return.

ACKNOWLEDGMENTS

[11] T. Kailath, Linear Systems, Prentice Hall, Englewood Cliffs, 1980.

[12] D.J.C. MacKay, "Bayesian Interpolation", Neural Computation, vol. 4, pp. 415-447, 1992.

[13] D.J.C. MacKay, "Probable Networks and Plausible Predictions - A Review of Practical Bayesian Methods for Supervised Neural Networks", Network: Computation in Neural Systems, vol. 6, pp. 469-505, 1995.

[14] M.H. Pesaran and A. Timmerman, "A simple nonparametric test of predictive performance", Journal of Business and Economic Statistics, vol. 10, pp. 461-465, 1992.

[15] A.N. Refenes, A.N. Burgess and Y. Bentz, "Neural Networks in Financial Engineering: A Study in Methodol-

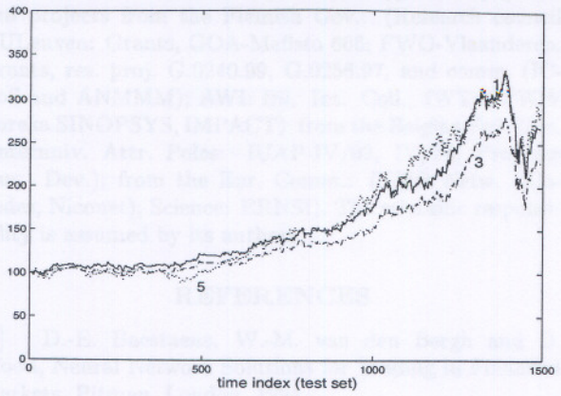

Figure 1: Cumulative returns on the test set using Investment Strategy 1 (IS1) assuming homo-scedastic noise: (1) RBF-LS-SVM (IS1, full line, no marker), (3) ARX (IS1, dash-dotted line, no marker), (5) Buy&Hold (dotted line, no marker).

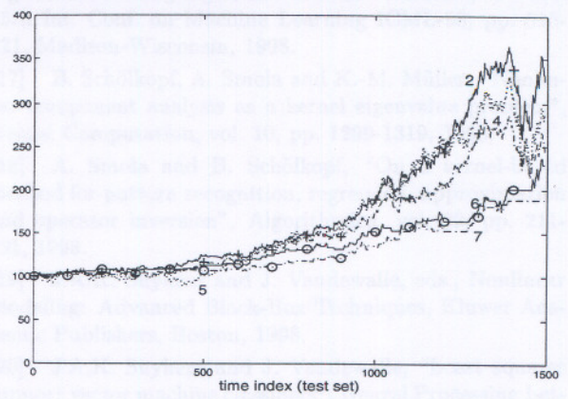

Figure 2: Cumulative returns on the test set using Investment Strategy 1 (IS1) and Investment Strategy 2 (IS2) assuming hetero-scedastic noise: (2) RBF-LS-SVM$_w$ (IS1, full line, marker +), (4) ARX$_w$ (IS1, dash-dotted line, marker +), (6) RBF-LS-SVM$_w$ (IS2, full line, marker ∘)and (7) ARX$_w$ (IS2, dash-dotted line, marker ∘). The cumulative return of the Buy&Hold strategy is denoted by the dotted line (5).

then replaced the constant ζ by time varying ζ_i using the corresponding moving average volatility estimates in order to estimate the non-linear RBF-LS-SVM$_w$ from (8). We compared the results with a linear ARX$_w$ using Generalized Least Squares with diagonal weighting matrix corresponding to the same volatility estimates. The results for Directional Accuracy are summarized in Table 1, which indicate that taking into account the time varying volatility of the markets yields better models. The McNemar test statistic [8] for a comparison between the RBF-LS-SVM$_w$ and ARX$_w$ "classifier" performance was 3.4839, with a p-value of 0.0307.

The model was also validated for trading assuming a transaction cost of 0.1% (10 bps [15]). Investment Strategy 1 implements a naive allocation of 100% equities or 100% cash, depending on the sign of the prediction. The corresponding annualized return/risk ratios (Sharpe Ratio when neglecting the risk free return) for the RBF-LSSVM, RBF-LS-SVM$_w$, ARX, ARX$_w$ and the Buy-and-Hold strategy are summarized in Table 2. In order to improve the annualized Sharpe Ratio SR$_1$, we define a trading signal based on $\hat{y}_{MP,N+1}/\sigma_{\hat{y}_{MP,N+1}}$, with $\sigma_{\hat{y}_{MP,N+1}}$ from (19). When this trading signal goes above or below a threshold, one changes the position (100% equities - 0% cash and vice versa) in Investment Strategy 2. This strategy is implemented for RBF-LS-SVM$_w$ and ARX$_w$, while the thresholds were obtained from the validation set so as to obtain the annualized Sharpe Ratio. This allows to improve the Sharpe Ratio on the test set as we can see from Table 2. The cumulative returns of the investment strategies with the different models depicted in Figures 1 and 2. The Asian crisis can be recognized by the drop in the cumulative return of, e.g., the Buy&Hold strategy.

In order to illustrate the influence on the model uncertainty on the error bars, we also depicted the error bars

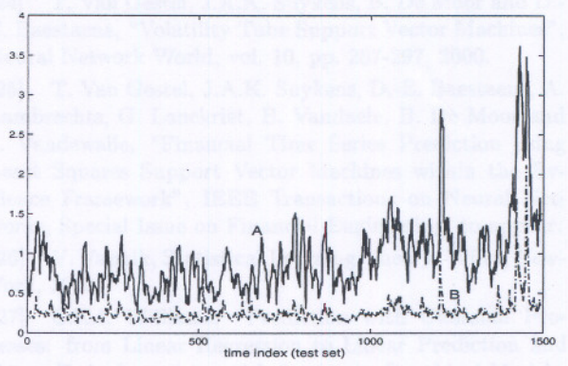

Figure 3: Error bars of the RBF-LS-SVM$_w$ model. (A) Total standard deviation $\sigma_{\hat{y}_{MP,N+1}} = (\zeta_{N+1}^{-1} + \sigma_z^2)^{\frac{1}{2}}$ from (19) (full line) and (B) standard deviation σ_z from (22) due to the uncertainty on the model parameters (dashed-dotted line). Observe the uncertainty on the model parameters during the Asian crisis.

(19) and the standard deviation σ_z from (22) due to the uncertainty on the model parameters are depicted. It can be observed that the model uncertainty becomes very high during the outbreak of the Asian crisis in the last part of the graph.

6. CONCLUSIONS

A probabilistic interpretation has been related to the Least Squares Support Vector Machine (LS-SVM) formulation