# Financial Time Series Prediction using Least Squares Support Vector Machines within the Evidence Framework

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

[1] Katholieke Universiteit Leuven
Dept. of Electrical Engineering ESAT-SISTA
Kasteelpark Arenberg 10, B-3001 Leuven, Belgium
Tel. +32/16/32.17.09, Fax. +32/16/32.19.70
E-mail: {tony.vangestel, johan.suykens}@esat.kuleuven.ac.be

[2] Fortis Bank Brussels
Warandeberg 3, B-1000 Brussels, Belgium
Tel. +32/02/565.85.54, Fax. +32/02/565.62.36
E-mail: dirk.baestaens@fortisbank.com

*Corresponding author.

**Abstract**

For financial time series, the generation of error bars on the point prediction is important in order to estimate the corresponding risk. The Bayesian evidence framework, already successfully applied to design of multilayer perceptrons, is applied in this paper to Least Squares Support Vector Machine (LS-SVM) regression in order to infer nonlinear models for predicting a time series and the related volatility. On the first level of inference, a statistical framework is related to the LS-SVM formulation which allows to include the time-varying volatility of the market by an appropriate choice of several hyperparameters. By the use of equality constraints and a 2-norm, the model parameters of the LS-SVM are obtained from a linear Karush-Kuhn-Tucker system in the dual space. Error bars on the model predictions are obtained by marginalizing over the model parameters. The hyperparameters of the model are inferred on the second level of inference. The inferred hyperparameters, related to the volatility, are used to construct a volatility model within the evidence framework. Model comparison is performed on the third level of inference in order to automatically tune the parameters of the kernel function and to select the relevant inputs. The LS-SVM formulation allows to derive analytic expressions in the feature space and practical expressions are obtained in the dual space replacing the inner product by the related kernel function using Mercer's theorem. The one step ahead prediction performances obtained on the prediction of the weekly 90-day T-bill rate and the daily DAX30 closing prices show that significant out of sample sign predictions can be made with respect to the Pesaran-Timmerman test statistic.

**Keywords:** Financial Time Series Prediction, Least Squares Support Vector Machines, Bayesian Inference, Volatility Modelling, Hyperparameter Selection, Model Comparison

# 1 Introduction

Motivated by the universal approximation property of multilayer perceptrons (MLPs), neural networks have been applied to learn nonlinear relations in financial time series [3, 12, 19]. The aim of many nonlinear forecasting methods [5, 14, 25] is to predict next points of a time series. 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. These density (volatility) predictions give information on the corresponding risk of the investment and they will, e.g., influence the trading behavior. A second reason why density forecasts have become important is that the risk has become a tradable quantity itself in options and other derivatives. In [15, 16], the Bayesian evidence framework was successfully applied to MLPs so as to infer output probabilities and the amount of regularization.

The practical design of MLPs suffers from drawbacks 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 [7, 24, 31, 32]. 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 kernel function. In this paper, we apply the evidence framework to Least Squares Support Vector Machines (LS-SVMs) [26, 27], where one uses equality constraints instead of inequality constraints and a least squares error term in order to obtain a linear set of equations in the dual space. This formulation can also be related to regularization networks [10, 12]. When no bias term is used in the LS-SVM formulation, as proposed in kernel ridge regression [20], the expressions in the dual space correspond to Gaussian Processes [33]. However, the additional insight of using the feature space has been used in kernel PCA [21], while the use of equality constraints and the primal-dual interpretations of LS-SVMs have allowed to make extensions towards recurrent neural networks [28] and nonlinear optimal control [29].

In this paper, the Bayesian evidence framework [15, 16] is applied to LS-SVM regression [26, 27] in order to estimate nonlinear models for financial time series and the related volatility. On the first level of inference, a probabilistic framework is related to the LS-SVM regressor inferring the time series model parameters from the data. Gaussian probability densities of the predictions are obtained within this probabilistic framework.

The hyperparameters of the time series model, related to the amount of regularization and the variance of the additive noise, are inferred from the data on the second level on inference. Different hyperparameters for the variance of the additive noise are estimated, corresponding to the time varying volatility of financial time series [23]. While volatility was typically modelled using (Generalized) Autoregressive Conditionally Heteroskedastic ((G)ARCH) models [1, 6, 30], more recently alternative models [9, 13, 17] have been proposed that basically model the observed absolute return. In this paper, the latter approach is related to the Bayesian estimate of the volatility on the second level of inference of the time series model. These volatility estimates are used to infer the volatility model.

On the third level of inference, the time series model evidence is estimated in order to select the tuning parameters of the kernel function and to select the most important set of inputs. In a similar way as the inference of the time series model, the volatility model is constructed using the inferred hyperparameters of the time series model. A schematic overview of the inference of the time series and volatility model is depicted in Figure 1. The LS-SVM formulation allows to derive analytical expressions in the feature space for all levels of inference, while practical expressions are obtained in a second step by using matrix algebra and the related kernel function.

This paper is organized as follows. The three levels for inferring the parameters of the LS-SVM time series model are described in Sections 2, 3 and 4, respectively. The inference of the volatility model is discussed in Section 5. An overview of the design of the LS-SVM time series and volatility model within the evidence framework is given in Section 6. Application examples of the Bayesian LS-SVM framework are discussed in Section 7.

# 2 Inference of the Model Parameters (Level 1)

A probabilistic framework [15, 16] is related to the LS-SVM regression formulation [26, 27] by applying Bayes' rule on the first level of inference. Expressions in the dual space for the probabilistic interpretation of the prediction are derived.

## 2.1 Probabilistic Interpretation of the LS-SVM Formulation

In Support Vector Machines [7, 24, 27, 32] for nonlinear regression, the data are generated by the nonlinear function $y_i = f(x_i) + e_i$ which is assumed to be of the following form

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

with model parameters $w \in \mathbb{R}^{n_f}$ and $b \in \mathbb{R}$ and where $e_i$ is additive noise. For financial time series, the output $y_i \in \mathbb{R}$ is typically a return of an asset or exchange rate, or some measure of the volatility 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 $\varphi(\cdot) : \mathbb{R}^n \rightarrow \mathbb{R}^{n_f}$ 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 theorem $K(x_i, x) = \varphi(x_i)^T \varphi(x)$ is applied to relate the function $\varphi(\cdot)$ with the symmetric and 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 $\sigma$ is a tuning parameter. In the sequel of this paper, we will focus on the use of an RBF-kernel.

Given the data points $D = \{(x_i, y_i)\}_{i=1}^N$ and the hyperparameters $\mu$ and $\zeta_{1:N} = [\zeta_1, \zeta_2, ..., \zeta_N]$ of the model $\mathcal{H}$ (LS-SVM with kernel function $K$), we obtain the model parameters by maximizing the posterior $P(w, b|D, \log\mu, \log\zeta_{1:N}, \mathcal{H})$. Application of Bayes' rule at the first level of inference [5, 15] gives:

$$P(w, b|D, \log\mu, \log\zeta_{1:N}, \mathcal{H}) = \frac{P(D|w, b, \log\mu, \log\zeta_{1:N}, \mathcal{H}) P(w, b|\log\mu, \log\zeta_{1:N}, \mathcal{H})}{P(D|\log\mu, \log\zeta_{1:N}, \mathcal{H})}, \tag{2}$$

where the evidence $P(D|\log\mu, \log\zeta_{1:N}, \mathcal{H})$ follows from normalization and is independent of $w$ and $b$.

We take the prior $P(w, b|\log\mu, \log\zeta_{1:N}, \mathcal{H})$ independent of the hyperparameters $\zeta_i$, 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 have a Gaussian distribution $P(w|\log\mu, \mathcal{H}) = \left(\frac{\mu}{2\pi}\right)^{\frac{n_f}{2}} \exp(-\frac{\mu}{2} w^T w)$ with zero mean, 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 $P(b|\log\sigma_b,\mathcal{H}) = \left(\frac{\sigma_b^{-2}}{2\pi}\right)^{-\frac{1}{2}}\exp(-\frac{b^2}{2\sigma_b^2})$, with $\sigma_b \to \infty$. We then obtain the following prior:

$$P(w,b|\log\mu,\mathcal{H}) = \left(\tfrac{\mu}{2\pi}\right)^{\frac{n_f}{2}}\exp(-\tfrac{\mu}{2}w^Tw)\tfrac{1}{\sqrt{2\pi\sigma_b^2}}\exp(-\tfrac{b^2}{2\sigma_b^2}) \propto \left(\tfrac{\mu}{2\pi}\right)^{\frac{n_f}{2}}\exp(-\tfrac{\mu}{2}w^Tw). \quad (3)$$

Assuming Gaussian distributed additive noise $e_i$ ($i = 1,...,N$) with zero mean and variance $\zeta_i^{-1}$, the likelihood $P(D|w,b,\log\zeta_{1:N},\mathcal{H})$ can be written as [16]

$$P(D|w,b,\log\zeta_{1:N},\mathcal{H}) \;=\; \prod_{i=1}^N \left(\tfrac{\zeta_i}{2\pi}\right)^{\frac{1}{2}}\exp(-\tfrac{\zeta_i}{2}e_i^2)\,P(x_i). \quad (4)$$

Taking the negative logarithm and neglecting all constants, we obtain that the likelihood (4) corresponds to the error term $\sum_{i=1}^N \zeta_i E_{D,i}$. 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 $\zeta_i^{-1}$ is used here [1, 6, 30] and is recently motivated by [2]. The corresponding optimization problem corresponds to taking the 2-norm of the error and results into a linear Karush-Kuhn-Tucker system in the dual space [20, 26, 27] while SVM formulations use different norms with inequality constraints which typically result into a Quadratic Programming Problem [7, 24, 31, 32].

Substituting (3) and (4) into (2) and neglecting all constants, application of Bayes' rule yields $P(w,b|D,\log\mu,\log\zeta_{1:N},\mathcal{H}) \propto \exp(-\frac{\mu}{2}w^Tw)\exp(-\sum_{i=1}^N \frac{\zeta_i}{2}e_i^2)$. Taking the negative logarithm, the maximum a posteriori model parameters $w_{MP}$ and $b_{MP}$ are obtained as the solution to the following optimization problem:

$$\min_{w,b} \mathcal{J}_1(w,b) = \mu E_W + \sum_{i=1}^N \zeta_i E_{D,i} \quad (5)$$

with

$$E_W \;=\; \tfrac{1}{2}w^Tw, \quad (6)$$

$$E_{D,i} \;=\; \tfrac{1}{2}e_i^2 = \tfrac{1}{2}(y_i - w^T\varphi(x_i) - b)^2. \quad (7)$$

The least squares problem (10)-(11) is not explicitly solved in $w$ and $b$. Instead, the linear system (13) in $\alpha$ and $b$ is solved in the dual space as explained in the next Subsection.

The posterior $P(w,b|D,\log\mu,\log\zeta_{1:N},\mathcal{H})$ can also be written as the Gaussian distribution:

$$P(w,b|D,\log\mu,\log\zeta_{1:N},\mathcal{H}) \;=\; \frac{1}{\sqrt{(2\pi)^{n_f+1}\det Q}}\exp(-\tfrac{1}{2}g^TQ^{-1}g), \quad (8)$$

with $g = [w - w_{MP};\ b - b_{MP}]$ and $Q = \mathrm{covar}(w, b) = \mathcal{E}(g^T g)$, 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 $\mathcal{J}_1(w, b)$:

$$Q = H^{-1} = \begin{bmatrix} H_{11} & H_{12} \\ H_{12}^T & H_{22} \end{bmatrix}^{-1} = \begin{bmatrix} \frac{\partial^2 \mathcal{J}_1}{\partial w^2} & \frac{\partial^2 \mathcal{J}_1}{\partial w \partial b} \\ \frac{\partial^2 \mathcal{J}_1}{\partial b \partial w} & \frac{\partial^2 \mathcal{J}_1}{\partial b^2} \end{bmatrix}^{-1}. \tag{9}$$

## 2.2 Moderated Output of the LS-SVM Regressor

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 $b$ [16] one obtains that the prediction $\hat{y}_{N+1}$ is Gaussian distributed with mean $\hat{y}_{MP,N+1} = z_{MP} = w_{MP}^T \varphi(x) + b_{MP}$ and variance $\sigma^2_{\hat{y}_{N+1}} = \zeta_{N+1}^{-1} + \sigma_z^2$. The first term $\zeta_{N+1}^{-1}$ corresponds to the volatility at the next time step and has to be predicted by volatility model. In Section 5 we discuss the inference of an LS-SVM volatility model to predict $\hat{\zeta}_{MP,N+1}^{-1/2} = \zeta_{N+1}^{-1/2}$. The second term $\sigma_z^2$ is due to the Gaussian uncertainty on the estimated model parameters $w$ and $b$ in the linear transform $z = w^T \varphi(x) + b$.

### 2.2.1 Expression for $z_{MP}$

Taking the expectation with respect to the Gaussian distribution over the model parameters $w$ and $b$ the mean $z_{MP}$ is obtained:

$$z_{MP} = \mathcal{E}\{z\} = w_{MP}^T \varphi(x) + b_{MP}.$$

In order to obtain a practical expression in the dual space, one solves the following optimization problem corresponding to (5):

$$\min_{w,e} \mathcal{J}_1(w, e) = \tfrac{\mu}{2}\, w^T w + \tfrac{1}{2} \textstyle\sum_{i=1}^N \zeta_i e_i^2 \tag{10}$$

$$\text{s.t.} \qquad y_i = w^T \varphi(x_i) + b + e_i, \qquad i = 1, \ldots, N. \tag{11}$$

To solve the minimization problem (10)-(11), 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:

$$\begin{cases} \frac{\partial \mathcal{L}_1}{\partial w} = 0 & \rightarrow \quad w = \sum_{i=1}^{N} \alpha_i \varphi(x_i) \\ \frac{\partial \mathcal{L}_1}{\partial w} = 0 & \rightarrow \quad \sum_{i=1}^{N} \alpha_i = 0 \\ \frac{\partial \mathcal{L}_1}{\partial w} = 0 & \rightarrow \quad \alpha_i = \gamma_i e_i, \qquad i = 1, \ldots, N \\ \frac{\partial \mathcal{L}_1}{\partial w} = 0 & \rightarrow \quad b = y_i - w^T \varphi(x_i) - e_i, \qquad i = 1, \ldots, N, \end{cases} \tag{12}$$

with $\gamma_i = \zeta_i / \mu$ $(i = 1, ..., N)$. Eliminating $w$ and $e$, one obtains the following linear Karush-Kuhn-Tucker system in $\alpha$ and $b$ [26, 27]

$$\left[\begin{array}{c|c} 0 & 1_v^T \\ \hline 1_v & \Omega + D_\gamma^{-1} \end{array}\right] \left[\begin{array}{c} b \\ \hline \alpha \end{array}\right] = \left[\begin{array}{c} 0 \\ \hline y \end{array}\right], \tag{13}$$

with[1] $y = [y_1; ...; y_N]$, $1_v = [1; ...; 1]$, $e = [e_1; ...; e_N]$, $\alpha = [\alpha_1; ...; \alpha_N]$ and $D_\gamma = \mathrm{diag}([\gamma_1; ...; \gamma_N])$. Mercer's theorem [7, 24, 32] is applied within the $\Omega$ matrix

$$\Omega_{ij} = \varphi(x_i)^T \varphi(x_j) = K(x_i, x_j). \tag{14}$$

The LS-SVM regressor is then obtained as:

$$\boxed{z_{MP} = \sum_{i=1}^{N} \alpha_i K(x, x_i) + b_{MP}} \tag{15}$$

Efficient algorithms for solving large scale systems such as e.g., the Hestenes-Stiefel conjugate gradient algorithm from numerical linear algebra can be applied to solve (13) by reformulating it into two linear systems with positive definite data matrices [27]. Also observe that Interior Point methods for solving the QP problem related to SVM regression solve a linear system of the same form as (13) in each iteration step. Although the effective number of parameters $\gamma_{eff}$ are controlled by the regularization term, the sparseness property of standard SVMs [11] is lost by taking the 2-norm. However, sparseness can be obtained by sequentially pruning the support value spectrum [27].

[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 $a \in \mathbb{R}^N$.

### 2.2.2 Expression for $\sigma_z^2$

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

$$\sigma_z^2 = \mathcal{E}\{(z - z_{MP})^2\} = \mathcal{E}\{[(w^T\varphi(x) + b) - (w_{MP}^T\varphi(x) + b_{MP})]^2\} = \psi(x)^T H^{-1}\psi(x), \quad (16)$$

with $\psi(x) = [\varphi(x);\, 1]$. The computation for $\sigma_z^2$ 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 $\sigma_z^2$ in the dual space is derived in Appendix A:

$$\boxed{\begin{aligned}\sigma_z^2 = {} & \theta(x)^T U_G Q_D U_G^T \theta(x) U^T + \tfrac{1}{\mu}K(x,x) - \tfrac{2}{s_\zeta}\theta(x)^T U_G Q_D U_G^T \Omega D_\zeta 1_v \\ & + \tfrac{2}{s_\zeta}\mu^{-1}\theta(x)^T D_\zeta 1_v + \tfrac{1}{s_\zeta} + \tfrac{1}{s_\zeta^2} 1_v^T D_\zeta \Omega U_G Q_D U_G^T \Omega D_\zeta 1_v + \tfrac{1}{\mu s_\zeta^2} 1_v^T D_\zeta \Omega D_\zeta 1_v\end{aligned}} \quad (17)$$

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 $D_G = \mathrm{diag}([\lambda_{G,1}, \ldots, \lambda_{G,N_{eff}}])$, where $\nu_{G,i}$ and $\lambda_{G,i}$ are the solution to the eigenvalue problem (45):

$$(D_\zeta - \tfrac{1}{s_\zeta} D_\zeta 1_v 1_v^T D_\zeta)\Omega\nu_{G,i} \;=\; \lambda_{G,i}\nu_{G,i}\ , \quad i = 1, \ldots, N_{eff} \leq N-1. \quad (18)$$

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

# 3 Inference of the Hyperparameters (Level 2)

In this Section, Bayes' rule is applied on the second level of inference [15, 16] in order to infer the hyperparameters $\mu$ and $\zeta_i$. Whereas it is well-known that the Bayesian estimate of the variance is biased, this problem is mainly due to the marginalization (see also (31) in this Section). The cost function related to the Bayesian inference of the hyperparameters is derived first. We then discuss the inference of $\mu$ and $\zeta = \zeta_i$ $(i = 1, ..., N)$ and the inference of non-constant $\zeta_i$.

### 3.1 Cost Function for Inferring the Hyperparameters

The hyperparameters $\mu$ and $\zeta_i$ ($i = 1, ..., N$) are inferred from the data $D$ by applying Bayes' rule on the second level:

$$\begin{aligned} P(\log\mu, \log\zeta_{1:N}|D,\mathcal{H}) &= \frac{P(D|\log\mu,\log\zeta_{1:N},\mathcal{H})P(\log\mu,\log\zeta_{1:N}|\mathcal{H})}{P(D|\mathcal{H})} \\ &\propto P(D|\log\mu,\log\zeta_{1:N},\mathcal{H}), \end{aligned} \tag{19}$$

where a flat, non-informative prior is assumed on the hyperparameters $\mu$ and $\zeta_i$. The probability $P(D|\log\mu,\log\zeta_{1:N},\mathcal{H})$ is equal to the evidence in (2) of the previous level. By substitution of (3), (4) and (8) into (19), one obtains:

$$P(\log\mu,\log\zeta_{1:N}|D,\mathcal{H}) \propto \frac{\sqrt{\mu^{n_f}}\prod_{i=1}^{N}\sqrt{\zeta_i}}{\sqrt{\det H}}\frac{\exp(-\mathcal{J}_1(w,b))}{\exp(-\frac{1}{2}g^THg)} \propto \sqrt{\frac{\mu^{n_f}\prod_{i=1}^{N}\zeta_i}{\det H}}\exp(-\mathcal{J}_1(w_{MP},b_{MP})). \tag{20}$$

Using the expression for $\det H$ from Appendix A and taking the negative logarithm, we find the maximum a posteriori estimates $\mu_{MP}$ and $\zeta_{MP,i}$ by minimizing the level 2 cost function:

$$\boxed{\begin{aligned} \mathcal{J}_2(\mu,\zeta_{1:N}) = {} & \mu E_W(w_{MP}) + \textstyle\sum_{i=1}^{N}\zeta_i E_{D,i}(w_{MP},b_{MP}) \\ & +\tfrac{1}{2}\textstyle\sum_{i=1}^{N_{eff}}\log(\mu+\lambda_{G,i}) - \frac{N_{eff}}{2}\log\mu - \frac{1}{2}\sum_{i=1}^{N}\log\zeta_i + \frac{1}{2}\log(\sum_{i=1}^{N}\zeta_i) \end{aligned}} \tag{21}$$

This is an optimization problem in $N+1$ unknowns and may require long computations. Therefore, we will first discuss the inference in the case of constant $\zeta_i = \zeta$. This value for the hyperparameters will then be used to infer the non-constant $\zeta_i$.

### 3.2 Inferring $\mu$ and $\zeta_i = \zeta$

We will now further discuss the inference of the hyperparameters for the special case of constant $\zeta_i = \zeta$. In this case, one can observe that the eigenvalues $\lambda_{G,i}$ in (18) are equal to $\lambda_{G,i} = \zeta\lambda'_{G,i}$, where the eigenvalues $\lambda'_{G,i}$ are obtained from the eigenvalue problem

$$(I_N - \tfrac{1}{N}1_v 1_v^T)\Omega\nu_{G,i} = \lambda'_{G,i}\nu_{G,i}\,, \quad i = 1,\ldots,N_{eff} \leq N-1, \tag{22}$$

with the identity matrix $I_N \in \mathbb{R}^{N\times N}$ and with the corresponding diagonal matrix $D'_G = \mathrm{diag}([\lambda'_{G,1},\ldots,\lambda'_{G,N_{eff}}]) = \zeta^{-1}D_G$. The eigenvalue problem (22) is now independent[2] from

[2]Observe that in this case, the eigenvalue problem (22) is related to the eigenvalue problem used in kernel PCA [21]. The corresponding eigenvalues are also used to derive improved bounds in VC-theory [22]. In the evidence framework, capacity is controlled by the prior.

the hyperparameter $\zeta$. By defining $E_D = \sum_{i=1}^{N} E_{D,i}$ and by using $s_\zeta = N\zeta$, the level 2 optimization problem (21) becomes:

$$\min_{\mu,\zeta} \mathcal{J}_3(\mu,\zeta) = \mathcal{J}_1(w_{MP}, b_{MP}) + \tfrac{1}{2}\sum_{i=1}^{N_{eff}} \log(\mu + \zeta\lambda'_{G,i}) - \tfrac{N_{eff}}{2}\log\mu - \tfrac{N-1}{2}\log\zeta. \quad (23)$$

The gradients of $\mathcal{J}_3(\mu,\zeta)$ towards $\mu$ and $\zeta$ are [15]

$$\frac{\partial \mathcal{J}_3}{\partial \mu} = E_W(w_{MP}) + \tfrac{1}{2}\sum_{i=1}^{N_{eff}} \frac{1}{\mu+\zeta\lambda'_{G,i}} - \frac{N_{eff}}{2\mu} \quad (24)$$

$$\frac{\partial \mathcal{J}_3}{\partial \zeta} = E_D(w_{MP}, b_{MP}) + \tfrac{1}{2}\sum_{i=1}^{N_{eff}} \frac{\lambda'_{G,i}}{\mu+\zeta\lambda'_{G,i}} - \frac{N-1}{2\zeta}. \quad (25)$$

Since the LS-SVM cost function consists of an error term $E_D$ with regularization term $E_W$ (ridge regression), the effective number of parameters [5, 16] is decreased by applying regularization. For the LS-SVM, the effective number of parameters $\gamma_{eff}$ is equal to

$$\boxed{\gamma_{eff} = 1 + \sum_{i=1}^{N_{eff}} \frac{\zeta\lambda'_{G,i}}{\mu+\zeta\lambda'_{G,i}}} \quad (26)$$

where the first term is due to the fact that no regularization is applied on the bias term $b$ of the LS-SVM model. Since $N_{eff} \leq N-1$, we cannot estimate more effective parameters than the number of data points $N$, even if the parameterization of the model $[w; b]$ has $n_f + 1$ degrees of freedom before one starts training, with typically $n_f \gg N$.

In the optimum of the level 2 cost function $\mathcal{J}_3(\mu,\zeta)$, both the partial derivatives (24) and (25) are zero. Putting (24) equal to zero, one obtains $2\mu_{MP}E_W(w_{MP}) = \gamma_{eff} - 1$, while one obtains $2\zeta_{MP}E_D(w_{MP}, b_{MP}) = N - \gamma_{eff}$ from (25). This equation corresponds to the unbiased estimate of the variance $\zeta_{MP}^{-1} = 2E_D/(N-\gamma_{eff})$ within the evidence framework.

These optimal hyperparameters $\mu_{MP}$ and $\zeta_{MP}$ are obtained by solving the optimization problem (23) with gradients (24) and (25). Therefore, one needs the expressions for $E_D = \sum_{i=1}^{N} E_{D,i}$ and $E_W = \frac{1}{2}w_{MP}^T w_{MP}$. These terms can be expressed in the dual variables using the conditions (12) in the optimum of level 1. The first term $E_{D,i} = \frac{1}{2}e_i^2$ is the easiest to calculate. Using the relation $\alpha_i = \gamma_i e_i$ of (12), we obtain:

$$E_{D,i} = \frac{1}{2}\frac{\alpha_i^2}{\gamma_i^2} = \frac{1}{2}\frac{(\mu^2\alpha_i)^2}{\zeta_i^2}. \quad (27)$$

The regularization term $E_W$ is calculated by combining the first and last condition in (12):

$$E_W = \tfrac{1}{2}\sum_{i=1}^{N} \alpha_i[w_{MP}^T\varphi(x_i)] = \tfrac{1}{2}\sum_{i=1}^{N} \alpha_i[y_i - \tfrac{\mu\alpha_i}{\zeta_i} - b_{MP}]. \quad (28)$$

In the case of constant $\zeta_i$, the parameters $\gamma_i$ are also constant $\gamma_i = \gamma = \zeta/\mu$.

### 3.3 Inferring $\mu$ and $\zeta_i$

In the previous Subsection, the conditions for optimality of the level 2 cost function $\mathcal{J}_3(\mu, \zeta)$ with respect to $\mu_{MP}$ and $\zeta_{MP}$ were related to the number of effective parameters $\gamma_{eff}$. In this Subsection, we will derive the conditions in the optimum of $\mathcal{J}_2(\mu, \zeta_i)$ with respect to $\zeta_i$ to infer the Bayesian estimate of the volatility.

The gradient $\frac{\partial \mathcal{J}_2}{\partial \mu}$ is derived in a similar way as the gradient $\frac{\partial \mathcal{J}_3}{\partial \mu}$ and is obtained by formally replacing $\zeta \lambda'_{G,i}$ by $\lambda_{G,i}$ in (24). By defining the effective number of parameters as

$$\gamma_{eff} = 1 + \textstyle\sum_{i=1}^{N} \frac{\lambda_{G,i}}{\mu + \lambda_{G,i}}, \tag{29}$$

a similar relation between $\mu_{MP}$ and $\gamma_{eff}$ holds in the optimum of $\mathcal{J}_3$ as in Subsection 3.2.

For the gradient $\frac{\partial \mathcal{J}_2}{\partial \zeta_i}$, one obtains (starting from the negative logarithm of (20)):

$$\frac{\partial \mathcal{J}_2}{\partial \zeta_i} = E_{D,i} + \frac{1}{2}\mathrm{Tr}[H^{-1}\frac{\partial H}{\partial \zeta_i}] - \frac{1}{2\zeta_i} = E_{D,i} + \frac{1}{2}\sigma_{z_i}^2 - \frac{1}{2\zeta_i}, \qquad i = 1, \ldots, N, \tag{30}$$

where $\mathrm{Tr}[H^{-1}\frac{\partial H}{\partial \zeta_i}] = \mathrm{Tr}[H^{-1}\psi(x_i)\psi(x_i)^T)] = \sigma_{z_i}^2$ using (16) and the expression for $H$. In the optimum, the gradient is zero, which yields:

$$2\zeta_{MP,i} E_{D,i}(w_{MP}, b_{MP}) = 1 - \sigma_{z_i}^2 \zeta_{MP,i}, \qquad i = 1, \ldots, N. \tag{31}$$

The last equation has to be interpreted as the unbiased estimate of the variance in the Bayesian framework, as mentioned in the introduction of this Section. The maximum a posteriori estimate of the variance $1/\zeta_{MP,i}$ is equal to the squared error, corrected by the relative model output uncertainty $e_i^2/(1 - \zeta_{MP,i}\sigma_{z_i}^2)$. Since the estimates $\zeta_{MP,i}$ are essentially only based on one observation of the time series, these estimates will be rather noisy.

Therefore, we will infer the hyperparameters $\zeta_{MP,i}$ by assuming that we are close to the optimum:

$$\zeta_{MP,i}^{-1} \;\; \simeq \;\; e_i^2 + \sigma_{z_i}^2, \qquad i = 1, \ldots, N, \tag{32}$$

where both $e_i$ and $\sigma_{z_i}^2$ are obtained from the LS-SVM model with constant $\zeta_{MP}$. The above assumption corresponds to an iterative method for training MLPs with constant $\zeta$ [15, 16] but does not guarantee convergence. We did not observe convergence problems in our experiments. The 'noisy' estimates will not be used to infer the LS-SVM time series model with non-constant hyperparameters $\zeta_{MP,i}$. Instead, the estimates $\zeta_{MP,i}$ are used to infer the LS-SVM volatility model in Section 5. The modelled $\hat{\zeta}_{MP,i}$ of the LS-SVM volatility model are far less noisy estimates of the corresponding volatility and will be used to infer the LS-SVM model time series model using a weighted least squares error term.

# 4 Model Comparison (Level 3)

In this Section Bayes' rule is applied to rank the evidence of different models $\mathcal{H}_j$ [15, 16]. For SVMs, different models $\mathcal{H}_j$ correspond to different choices for the kernel function; e.g., for an RBF kernel with tuning parameter $\sigma_j$, the probability of the corresponding models $\mathcal{H}_j$ is calculated in order to select the tuning parameter $\sigma_j$ with the greatest model evidence. Model comparison can also be used to select the relevant set of inputs by ranking the evidence of models inferred with different sets of inputs. The model selection of the time series model is performed before inferring the $\hat{\zeta}_{MP,i}$, obtained as the outputs of the volatility model; and therefore we will assume a constant $\zeta_i = \zeta$, $i = 1, \ldots, N$ in this Section.

By applying Bayes' rule on the third level, we obtain the posterior for the model $\mathcal{H}_j$:

$$P(\mathcal{H}_j|D) \propto P(D|\mathcal{H}_j)P(\mathcal{H}_j). \tag{33}$$

At this level, no evidence or normalizing constant is used since it is impossible to compare all possible models $\mathcal{H}_j$. The prior $P(\mathcal{H}_j)$ over all possible models is assumed to be uniform here. Hence, (33) becomes $P(\mathcal{H}_j|D) \propto P(D|\mathcal{H}_j)$. The likelihood $P(D|\mathcal{H}_j)$ corresponds to the evidence (19) of the previous level. For the prior $P(\log\mu_{MP}, \log\zeta_{MP}|\mathcal{H}_j)$ on the positive scale parameters $\mu$ and $\zeta$, a separable Gaussian with error bars $\sigma_{\log\mu}$ and $\sigma_{\log\zeta}$ is taken. We assume that these a priori error bars are the same for all models $\mathcal{H}_j$. To calculate the posterior approximation analytically, it is assumed [15] that the evidence $P(\log\mu, \log\zeta|D, \mathcal{H}_j)$ can be very well approximated by using a separable Gaussian with error bars $\sigma_{\log\mu|D}$ and $\sigma_{\log\zeta|D}$. As in Section 3, the posterior $P(D|\mathcal{H}_j)$ then becomes [16]

$$P(D|\mathcal{H}_j) \propto P(D|\log\mu_{MP}, \log\zeta_{MP}, \mathcal{H}_j)\frac{\sigma_{\log\mu|D}\sigma_{\log\zeta|D}}{\sigma_{\log\mu}\sigma_{\log\zeta}}. \tag{34}$$

Ranking of models according to model quality $P(D|\mathcal{H}_j)$ is thus based on the goodness of the fit (20) and the Occam factor [15], which punishes for overparameterized models.

We refer to [16] for a discussion on relations between the evidence framework and other theories of generalization like, e.g., Minimum Description Length and VC-theory.

Following a similar reasoning as in [15, 16] approximate expressions for the errors bars $\sigma_{\log\mu|D}$ and $\sigma_{\log\zeta|D}$ are obtained by differentiating (21) twice with respect to $\mu$ and $\zeta$: $\sigma^2_{\log\mu|D} \simeq 2/(\gamma_{eff}-1)$ and $\sigma^2_{\log\zeta|D} \simeq 2/(N-\gamma_{eff})$. One then obtains

$$\boxed{P(D|\mathcal{H}_j) \propto \sqrt{\frac{\mu_{MP}^{N_{eff}}\,\zeta_{MP}^{N-1}}{(\gamma_{eff}-1)(N-\gamma_{eff})\prod_{i=1}^{N_{eff}} \mu_{MP}+\zeta_{MP}\lambda'_{G,i}}}} \tag{35}$$

# 5 Volatility Modelling

Since the volatility is not an observed variable of the time series $\{y_i\}_{i=1}^N$, we will use the inferred hyperparameters $\zeta_{MP,i}$, $i = 1, \ldots, N$, from (32) to train the LS-SVM volatility model. The inverse values $1/\zeta_{MP,i}$ correspond to the estimated variances of the noise $e_i$ on the observations $y_i$. Instead of modelling and predicting the inferred $\zeta_{MP,i}$ or $\zeta_{MP,i}^{-1}$ directly, we will model $\zeta_{MP,i}^{-1/2}$, which corresponds to the prediction of absolute returns [13, 17, 30]. Indeed, one can observe that when the model output uncertainty $\sigma^2_{z,i}$ is small ($\zeta_{MP,i}\sigma^2_{z_i} \ll 1$), then (30) becomes $\zeta_{MP,i} \simeq 1/e_i^2$. In this case, the prediction of $\zeta_{MP,i}^{-1/2}$ corresponds to predicting the absolute values $|e_i|$, which corresponds to the prediction of the absolute returns when the no time series model is used (see, e.g., [13, 17, 30]). We briefly discuss the three levels of inference and point out differences with the inference of the LS-SVM time series model.

The outputs $\tilde{y}_i \in \mathbb{R}$ of the LS-SVM volatility model $\tilde{f}(\tilde{x}) = \tilde{w}^T\tilde{\varphi}(\tilde{x}) + \tilde{b}$ are the inferred $\zeta_{MP,i}^{-1/2}$ values of the second level of the time series model, i.e., $\tilde{y}_i = \zeta_{MP,i}^{-1/2}$. The inputs $\tilde{x}_i \in \mathbb{R}^{\tilde{n}}$ are determined by the user and may consist of lagged absolute returns [13, 30] and other explanatory variables. Input pruning can be performed on the third level as explained in the previous Section. In a similar way as in Section 2, the model parameters $\tilde{w}$ and $\tilde{b}$ are inferred from the data $\tilde{D} = \{(\tilde{x}_i, \tilde{y}_i)\}_{i=1}^N$ by minimizing $\tilde{\mathcal{J}}_1(\tilde{w}, \tilde{b}) = \tilde{\mu}\,\tilde{w}^T\tilde{w} + \tilde{\zeta}\,\tilde{E}_D$, with $\tilde{E}_D = \frac{1}{2}\sum_{i=1}^N \tilde{e}_i^2$ and $\tilde{e}_i = \tilde{y}_i - (\tilde{w}^T\tilde{\varphi}(\tilde{x}_i) + \tilde{b})$, $i = 1, ..., N$. By introducing the Lagrange multipliers $\tilde{\alpha}_i \in \mathbb{R}$, the following linear set of equations is obtained in the dual space:

$$\left[\begin{array}{c|c} 0 & 1_v^T \\ \hline 1_v & \tilde{\Omega} + D_{\tilde{\gamma}}^{-1} \end{array}\right] \left[\begin{array}{c} \tilde{b} \\ \hline \tilde{\alpha} \end{array}\right] = \left[\begin{array}{c} 0 \\ \hline \tilde{y} \end{array}\right], \tag{36}$$

with $\tilde{y} = [\tilde{y}_1; \ldots; \tilde{y}_N]$, $\tilde{\alpha} = [\tilde{\alpha}_1; \ldots; \tilde{\alpha}_N]$ and $D_{\tilde{\gamma}} = \tilde{\gamma} I_N$, where $\tilde{\gamma} = \tilde{\zeta}/\tilde{\mu}$. The matrix $\tilde{\Omega} \in \mathbb{R}^{N\times N}$ has elements $\tilde{\Omega}_{ij} = \tilde{K}(\tilde{x}_i, \tilde{x}_j)$. The expected value $\hat{\zeta}^{-1/2}_{MP,N+1} = z_{MP}$ of the LS-SVM volatility model is obtained as:

$$\tilde{z}_{MP} = \tilde{f}(\tilde{x}) = \tilde{w}_{MP}^T \tilde{\varphi}(\tilde{x}) + \tilde{b}_{MP} = \textstyle\sum_{i=1}^N \tilde{\alpha}_i \tilde{K}(\tilde{x}, \tilde{x}_i) + \tilde{b}_{MP}. \tag{37}$$

In a similar way as in Subsection 2.2, one may derive error bars on the predicted volatility measure $\tilde{y}_{MP}$. This uncertainty on the volatility forecasts is not in the scope of this paper.

The hyperparameters $\tilde{\mu}$ and $\tilde{\zeta}_i = \tilde{\zeta}$ $(i = 1, ..., N)$ of the volatility model correspond to the regularization term $\tilde{E}_W$ and error term $\tilde{E}_D$, respectively. Observe that we assume a constant variance of the noise in the volatility model. The hyperparameters $\tilde{\mu}$ and $\tilde{\zeta}$ are obtained by minimizing

$$\min_{\tilde{\mu},\tilde{\zeta}} \tilde{\mathcal{J}}_3(\tilde{\mu}, \tilde{\zeta}) \quad = \quad \tilde{\mathcal{J}}_1(\tilde{w}, \tilde{b}) + \tfrac{1}{2}\textstyle\sum_{i=1}^{N_{eff}} \log(\tilde{\mu} + \zeta \lambda'_{\tilde{G},i}) - \tfrac{N_{eff}}{2}\log\tilde{\mu} - \tfrac{N-1}{2}\log\tilde{\zeta}, \tag{38}$$

where $\tilde{E}_D(\tilde{w}_{MP}, \tilde{b}_{MP})$, $\tilde{E}_W(\tilde{w}_{MP})$ and $\lambda'_{\tilde{G},i}$ are obtained in a similar way as in Section 3 from (27), (28) and (22), respectively. Similar relations as in Subsection 3.2 exists, relating the regularization term and the error term to the effective number of parameters $\tilde{\gamma}_{eff} = 1 + \sum_{i=1}^N \frac{\tilde{\zeta}_{MP}\lambda'_{\tilde{G},i}}{\tilde{\mu}_{MP} + \tilde{\zeta}_{MP}\lambda_{\tilde{G},i}}$ of the volatility model $\tilde{\mathcal{H}}$.

In a similar way as in Section 4, the probability of different volatility models $\tilde{\mathcal{H}}_j$ can be ranked. This then yields a similar expression as (39):

$$P(\tilde{D}|\tilde{\mathcal{H}}_j) \propto \sqrt{\frac{\tilde{\mu}_{MP}^{\tilde{N}_{eff}}\tilde{\zeta}_{MP}^{N-1}}{(\tilde{\gamma}_{eff}-1)(N-\tilde{\gamma}_{eff})\prod_{i=1}^{\tilde{N}_{eff}} \tilde{\mu}_{MP} + \tilde{\zeta}_{MP}\lambda'_{\tilde{G},i}}}. \tag{39}$$

# 6 Design of the Bayesian LS-SVM

We will apply the theory from the previous Sections to the design of the LS-SVM time series and volatility model within the evidence framework.

## 6.1 Design of the LS-SVM Time Series Model

The design of the LS-SVM time series model consists of the following steps (see also Figure 1):

1. The selected inputs are normalized to zero mean and unit variance [5]. The normalized training data are denoted by $D = \{(x_i, y_i)\}_{i=1}^N$, with $x_i \in \mathbb{R}^n$ the normalized inputs and $y_i \in \mathbb{R}$ the corresponding outputs, transformed to become stationary.

2. Select the model $\mathcal{H}_j$ by choosing a kernel type $K_j$, e.g, an RBF-kernel with parameter $\sigma_j$. For this model, the hyperparameters $\mu_{MP}$ and $\zeta_{MP,i} = \zeta_{MP}$ are inferred from the data on the second level. This is done as follows: (a) Solve the eigenvalue problem (22) to find the $N_{eff}$ important eigenvalues $\lambda'_{G,i}$ and the corresponding eigenvectors $\nu_{G,i}$. (b) Minimize $\mathcal{J}_3(\mu, \zeta)$ from (23) with respect to $\mu$ and $\zeta$. The cost function (23) and gradients (24)-(25) are evaluated by using the optimal time series model parameters $w_{MP}$ and $b_{MP}$. These are obtained from the first level of inference in the dual space by solving the linear system (13). (c) Calculate the number of effective parameters $\gamma_{eff}$ defined in (26). (d) Calculate the volatility estimates $\zeta_{MP,i}^{-1/2}$ with $\zeta_{MP,i}$ from (32) (these values will be used to infer the volatility model $\tilde{\mathcal{H}}_j$).

3. Calculate the model evidence $P(D|\mathcal{H}_j)$ from (35). For an RBF-kernel, one may refine $\sigma_j$ such that a higher model evidence is obtained. This is done by maximizing $P(D|\mathcal{H}_j)$ with respect to $\sigma_j$ by evaluating the model evidence for the refined kernel parameter starting from step 2(a).

4. Select the model $\mathcal{H}_j$ with maximal model evidence $P(D|\mathcal{H}_j)$. If the predictive performance is insufficient, select a different kernel function $K_j$ (step 2) or select a different set of inputs (step 1).

5. Use the outputs $\hat{\zeta}_{MP,i}^{-1/2}$ of the volatility model to refine the time series model. This is done in the following steps: (a) Solve the eigenvalue problem (18) to find the $N_{eff}$ important eigenvalues $\lambda_{G,i}$ and the corresponding eigenvectors $\nu_{G,i}$. (b) Refine the amount of regularization $\mu$. This is done by optimizing $\mathcal{J}_2(\mu, \zeta_i)$ in (21) with respect to $\mu$, while keeping $\zeta_i = \hat{\zeta}_i$. The gradient $\frac{\partial \mathcal{J}_2}{\partial \mu}$ is obtained by formally replacing $\zeta \lambda'_{G,i}$ by $\lambda_{G,i}$ in (24). The cost function and the gradient are evaluated as in step 2(b) by inferring $\alpha$ and $b_{MP}$ in the dual space on the first level and calculating $E_D$ and $E_W$ from (27) and (28), respectively. (c) Calculate the effective number of parameters $\gamma_{eff}$ from (29).

Notice that for a kernel function without tuning parameter like, e.g., the polynomial kernel with fixed degree $d$, steps 2 and 3 are trivial. No tuning parameter of the kernel function has to be chosen in step 2 and no refining is needed in step 3. The model evidence $P(D|\mathcal{H}_j)$ can be used in step 4 to rank different kernel types. The model evidence can also be used to rank models with different input sets, in order to select the most appropriate inputs.

## 6.2 Design of the LS-SVM Volatility Model

The design of the LS-SVM volatility model is similar to the design of the time series model. In step 1, the inputs are normalized to zero mean and unit variance [5]. The normalized training data are denoted by $\tilde{D} = \{(\tilde{x}_i, \tilde{y}_i)\}_{i=1}^{N}$, where $\tilde{x}_i \in \mathbb{R}^n$, $i = 1, \ldots, N$, are the normalized inputs and where $\tilde{y}_i = \zeta_{MP,i}^{-1/2} \in \mathbb{R}$ are the corresponding outputs, with $\zeta_{MP,i}$ from (32) of the time series model $\mathcal{H}_j$. In step 2, one selects the model $\tilde{\mathcal{H}}_j$ by choosing a kernel type $\tilde{K}_j$, e.g., an RBF-kernel with parameter $\tilde{\sigma}_j$. For this model, the hyperparameters $\tilde{\mu}_{MP}$ and $\tilde{\zeta}_{MP}$ are inferred from the data on the second level as in steps 2(a), 2(b) and 2(c) of the time series model. The model evidence $P(\tilde{D}|\tilde{\mathcal{H}}_j)$ is calculated from (39) in step 3. In step 4, one selects the model $\tilde{\mathcal{H}}_j$ with maximal model evidence $P(D|\mathcal{H}_j)$. Go to step 2 or step 1 if the performance is insufficient. For an RBF-kernel, one may refine $\tilde{\sigma}_j$ such that a higher model evidence is obtained. Finally, one calculates the $\hat{\zeta}_{MP,i}$ from (37) in step 5.

## 6.3 Generating Point and Density Predictions

Given the designed LS-SVM time series and volatility model $\mathcal{H}_j$ and $\tilde{\mathcal{H}}_j$, the point prediction $\hat{y}_{MP,N+1}$ and corresponding error bar $\sigma_{\hat{y}_{N+1}}$ are obtained as follows:

1. Let the input $x$ of the time series model be normalized in the same way as the training data $D$. The point prediction $\hat{y}_{MP,N+1}$ is then obtained as $\hat{y}_{MP,N+1} = z_{MP}$ from (15).

2. Normalize the input of the volatility model in the same way as the training data $\tilde{D}$. The normalized input is denoted by $\tilde{x}$. Predict the volatility measure $\hat{\zeta}_{MP,N+1}^{-1/2} = \tilde{f}(\tilde{x})$ from (37). Calculate error bar $\sigma_z$ due to the model uncertainty from (17). The total uncertainty on the prediction is then $\sigma_{\hat{y}_{N+1}}^2 = \hat{\zeta}_{MP,N+1}^{-1} + \sigma_z^2$.

# 7 Examples

The design of the LS-SVM regressor in the evidence framework is applied to two cases. First, the performance of the LS-SVM time series model is compared with results from the literature [4, 8] for the case of one step ahead prediction of the US short term interest rate. Second, we illustrate the use of the LS-SVM time series and volatility model for the one step ahead prediction of the DAX30 index. All simulations were carried out in Matlab.

## 7.1 Prediction of the US Short Term Interest Rate

The LS-SVM times series model is designed within the evidence framework for the one step ahead prediction weekly Friday observations of the 90-day US T-bill rate on secondary markets from 4 Jan. 1957 to 17 Dec. 1993, which is the period studied in [4, 8]. The first differences of the original series are studied, which is stationary at the 5% level according to the augmented Dickey-Fuller test. Using the same inputs as in [8], the input vector is constructed using past observations with lags from 1 to 6. The time series model was constructed assuming a constant volatility.

The first 1670 observations (1957-1988) were used to infer the optimal hyperparameters $\mu_{MP} = 0.0057$ and $\zeta_{MP,i} = \zeta_{MP} = 1.23$ and the optimal tuning parameter $\sigma_{MP} = 12$ resulting into an effective number of parameters $\gamma_{eff} = 108.88$. These hyper- and kernel parameters were kept fixed for the out of sample one step ahead prediction on the 254 observations of the period 1989-1993. In the first experiment, the model parameters $w_{MP}$ and $b_{MP}$ were kept fixed (NRo, No Rolling approach, [8]); in the second experiment the Rolling approach (Ro) was applied, i.e., re-estimating the model parameters $w$ and $b$ or $\alpha$ and $b$ each time a new observation becomes available. In Table 1, the out of sample prediction performances of the LS-SVM an AutoRegressive model (AR14) with lags at 1, 4, 7 and 14 (this is the optimal model structure selected in [8] using Akaike's information criterion (AIC)). The performances of a kernel-based non-parametric conditional mean predictor (NonPar), with mean squared error cost function (MSE) [8], are quoted in the last row of Table 1.

The MSE and corresponding sample standard deviations of the different models are reported in the first column. The MSE for a random walk model is 0.186 with sample standard deviation (0.339), which indicates that only a small part of the signal is explained by the models. The reduction obtained with the LS-SVM is of the same magnitude as the reduction obtained by applying a nearest neighbor technique on quarterly data [4]. The next columns show that the LS-SVM regressor clearly achieves a higher Percentage of Correct Sign Predictions (PCSP). The high values of the Pesaran-Timmerman (PT) statistic for directional accuracy [18] allow to reject the H0 hypothesis of no dependency between predictions and observations at significance levels below 1%.

## 7.2 Prediction of the DAX 30

We design the LS-SVM time series model in the evidence framework to predict the daily closing price return of the German DAX30 index (Deutscher Aktien Index). Then we used the inferred hyperparameters of the time series model to construct the LS-SVM volatility model. The modeled volatility level is then used to refine the LS-SVM model using the weighted least squares cost function and to calculate the return per unit risk $\hat{y}_{MP,N+1}/\sigma_{\hat{y}_{MP,N+1}}$ (Sharpe Ratio [14, 19, 30] neglecting riskfree return) of the prediction. The following inputs were used: lagged returns of closing prices of DAX30, Germany 3-Month Middle Rate, US 30-year bond, S&P500, FTSE, CAC40. All inputs were normalized to zero mean and unit variance [5], while the output was normalized to unit variance for convenience. We started with a total number of 38 inputs for the LS-SVM time series model. The performance of the LS-SVM model was compared with the performance of an ARX model (ARX10) with 10 inputs and an AR model (AR20) of order 20 with lags (1, 3, 4, 9, 17, 20), estimated with Ordinary Least Squares (OLS). The inputs of the AR and ARX model were sequentially pruned using AIC, starting from 20 lags and the 38 inputs of the LS-SVM model, respectively. The performances are also compared with a simple Buy-and-Hold strategy (B&H). The training set consists of 600 training data points from 17.04.92 till 17.03.94. The next 200 data points were used as a validation set. An out of sample test set of 1234 points was used, covering the period 23.12.94-10.12.98, which includes the Asian crises in 1998.

The LS-SVM model was inferred as explained in Section 6. From level 3 inference, we obtained the kernel parameter $\sigma = 20$. The effective parameters of the LS-SVM model with weighted error term is $\gamma_{eff} = 146.4$. Predictions were made using the rolling approach updating the model parameters after 200 predictions. The performances of the models are compared with respect to the Success Ratio (SR) and the Pesaran-Timmerman test statistic [18] for directional accuracy (PT) with corresponding $p$-value. The market timing ability of the models was estimated by using the prediction in 2 investment strategies assuming a transaction cost of 0.1% (10 bps as in [19]). Investment Strategy 1 (IS1) implements a naive allocation of 100% equities or cash, based on the sign of the prediction. This strategy will result in many transactions (588 for the LS-SVM) and profit will be eroded by the commissions[3] In Investment Strategy 2 (IS2) one changes the position (100% cash/0% equities - 0%/100% cash/equities) according to the sign of the prediction only when the absolute value of the Sharpe Ratio $\hat{y}_{MP,N+1}/\sigma_{\hat{y}_{MP,N+1}}$ exceeds a threshold, determined on the training set. This strategy reduces the number of transactions (424 for the LS-SVM) changing positions only when a clear trading signal is given. The volatility measure $\hat{\zeta}_{N+1}^{-1}$ in $\sigma_{\hat{y}_{MP,N+1}}$ is predicted by the LS-SVM volatility model as explained below. The cumulative returns obtained with the different models using strategy IS2 are visualized in Figure 2. The annualized return and risk characteristics of the investment strategy are summarized in Table 2. The LS-SVM with RBF-kernel has a better out of sample performance than the ARX and AR model with respect to the Directional Accuracy, where the predictive performance of the ARX is mainly due to lagged interest rate values. Also in combination with both investment strategies IS1 and IS2, the LS-SVM yields the best annualized risk/return ratio (Sharpe Ratio, SR), while strategy S2 illustrates the use of the uncertainty[4] on the predictions.

Finally, we illustrate input pruning for the case of the time series model. This is done by sequentially pruning the inputs of the model comparing the full model evidence with the input pruned model evidences. We start from the time series model with 38 inputs, which yields a PCSP of 57.7% on the validation set. In the first pruning step, we compare 38

[3]For zero transactions cost, the LS-SVM, ARX10, AR20 and B&H achieves annualized returns (Re) 32.7%, 21.8%, 8.7% and 16.4% with corresponding risk (Ri) 14.6%, 15.2%, 15.3% and 20.3% resulting in Sharpe Ratios (SR) 2.23, 1.44, 0.57 and 0.81, respectively.

[4]In order to illustrate the use of the model uncertainty for the LS-SVM model, trading on the signal $\hat{y}_{MP,N+1}/\hat{\zeta}_{N+1}^{-1/2}$ with IS2 yields a SR, Re and Ri of 1.28, 18.8 and 14.8, respectively.

models and remove the input corresponding to the lowest model evidence. After the first pruning step, the PCSP remained 57.7%. The pruning of the input corresponding to the highest model evidence would have resulted in a significantly lower PCSP of 55.2%. We restart now from the first model with 37 inputs and compare again the model evidence with 37 prunded model evidences. The pruning process is stopped when the model evidences of the pruned model are lower than the full model of the previous pruning step. Here, we performed 5 pruning steps, resulting in no loss with respect to the PCSR on the validation set. One may notice that the pruning is rather time consuming. An alternative way is to start from one input and sequentially add inputs within the evidence framework.

The volatility model is inferred as explained in Section 5. The input vector $\tilde{x}$ consists of ten lagged absolute returns, while the outputs of the training set are obtained from the LS-SVM Time Series Model. The hyperparameters $\tilde{\mu}_{MP} = 2.87$ and $\tilde{\zeta}_{MP} = 4.31$ and the kernel parameter $\tilde{\sigma} = 18$ were inferred on the second and third level, respectively, yielding $\tilde{\gamma}_{eff} = 8.61$. The performance of the volatility model was compared on the same targets with a GARCH(1,1) model [1, 6, 23, 30] and with three autoregressive models of order ten ($|\text{AR10}|$, $|\text{AR10}|^{1.1}$ and $|\text{AR10}|^2$) for the absolute returns [9, 13, 17, 30] using power transformations 1, 1.1 and 2, respectively. Since these models do not guarantee positive outputs, also an AR model (logAR10) is estimated on the logarithms of the data predicting where the predicted volatility corresponds to the exponential of the output of the logAR10 model. The AR models are estimated using OLS and pruning the inputs according to AIC, while the power transformation 1.1 was selected from a power transformation matrix [9, 17] according to AIC. The MSE and Mean Average Error (MAE) test set performances of the 5 models are reported together with the corresponding sample standard deviations in Table 2. In the last two columns, the models are compared with respect to the Negative Log Likelihood (NLL) $-\log \prod_{i=1}^{N_{\text{test}}} P(e_i)$ of the observation $e_i$ given the modeled volatility. Although guaranteeing a positive output, the logAR10 yields clearly lower performances. The nonlinear LS-SVM model with RBF-kernel yields a better performance than the AR models. Also, all AR models yield better performances than the GARCH(1,1) model on the MSE and MAE criteria, while vice versa the GARCH(1,1) yields a better NLL. This corresponds to the different training criteria of the different models. The LS-SVM model yields comparable results than with the GARCH(1,1) model.

# 8 Conclusions

In financial time series, the deterministic signal is masked by heteroskedastic noise and density predictions are important because one wants to know the associated risk, e.g., to make optimal investment decisions. In this paper, the Bayesian evidence framework is combined with Least Squares Support Vector Machines (LS-SVMs) for nonlinear regression in order to infer nonlinear models of a time series and the corresponding volatility. The time series model was inferred from the past observations of the time series. On the first level of inference, a probabilistic framework is related to the LS-SVM regressor in which the model parameters are inferred for given hyperparameters and given kernel functions. Error bars on the prediction are obtained in the defined probabilistic framework. The hyperparameters of the time series model are inferred from the data on the second level of inference. Since the volatility is not a directly observed variable of the time series, the volatility model is inferred within the evidence framework from past absolute returns and the hyperparameters of the time series model related to the volatility inferred in the second level. The volatility forecasts of the volatility model are used in combination with the model output uncertainty in order to generate the error bars in the density prediction. Model comparison is performed on the third level to infer the tuning parameter of the RBF-kernel by ranking the evidences of the different models. The design of the LS-SVM regressor within the evidence framework is validated on the prediction of the weekly US short term T-bill rate and the daily closing prices of the DAX30 stock index.

## Acknowledgments

The authors would like to thank David MacKay for helpful comments related to the second and third level of inference. T. Van Gestel and J. Suykens are a Research Assistant and a Postdoctoral Researcher with the Fund for Scientific Research-Flanders (FWO-Vlaanderen), respectively. This work was partially supported by grants and projects from the Flemish Gov.: (Research council KULeuven: Grants, GOA-Mefisto 666; FWO-Vlaanderen: proj. G.0240.99, G.0256.97, and comm. (ICCoS, ANMMM); IWT: STWW Eureka SINOPSYS, IMPACT); from the Belgian Fed. Gov. (IUAP-IV/02, IV/24; Program Dur. Dev.); from the Eur. Comm.: (TMR Netw. (Alapedes, Niconet); Science: ERNSI). The scientific responsibility is assumed by its authors.

# References


[1] T.G. Anderson, T. Bollerslev, "Answering the Skeptics: Yes, Standard Volatility Models Do Provide Accurate Forecasts," *International Economic Review,* vol. 39, pp. 885-905, 1998.

[2] T.G. Anderson, T. Bollerslev, F.X. Diebold and P. Labys, "Exchange Rate Returns Standardized by Realized Volatility are (Nearly) Gaussian," Working Paper 262, Dept. of Finance, Kellogg Graduate School of Management, Northwestern University, Evanston, U.S.A., 2000.

[3] D.-E. Baestaens, W.-M. van den Bergh and D. Wood, *Neural Network Solutions for Trading in Financial Markets*, Pitman, London, 1994.

[4] J.T. Barkoulas, C.F. Baum and J. Onochie, "A Nonparametric Investigation of the 90-day T-bill rate," *Review of Financial Economics,* vol. 6, pp. 187-198, 1997.

[5] C.M. Bishop, *Neural Networks for Pattern Recognition*, Oxford University Press.

[6] T. Bollerslev, R.F. Engle and D.B. Nelson, "ARCH Models," in *The Handbook of Econometrics,* vol. 4, R.F. Engle and D.L. McFadden, eds., Elsevier Science, Amsterdam, The Netherlands, 1994.

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

[8] J.G. De Gooijer and D. Zerom, "Kernel-based Multistep-ahead Predictions of the US Short-term Interest Rate," *Journal of Forecasting,* vol. 19, pp. 335-353, 2000.

[9] Z. Ding, C.W.J. Granger and R.F. Engle, "A Long Memory Property of Stock Market Returns and a New Model," *Journal of Empirical Finance,* vol. 1, pp. 83-106, 1993.

[10] T. Evgeniou, M. Pontil and T. Poggio, "Regularization Networks and Support Vector Machines," *Advances in Computational Mathematics,* vol. 13, pp. 1-50, 2001.

[11] F. Girosi, "An equivalence between sparse approximation and support vector machines," *Neural Computation*, vol. 10, pp. 1455-1480, 1998.

[12] J.M. Hutchinson, A.W. Lo and T. Poggio, "A Nonparametric Approach to Pricing and Hedging Derivative Securities Via Learning Networks," *Journal of Finance*, vol. 49, pp. 851-889, 1994.

[13] C.W.J. Granger and C.-Y. Sin, "Modelling the Absolute Returns of Different Stock Indices: Exploring the Forecastability of an Alternative Measure of Risk," *Journal of Forecasting,* vol. 19, pp. 277-298, 2000.

[14] C.W.J. Granger and T. Terasvirta, *Modelling Nonlinear Economic Relationships,* Oxford University Press, 1993.

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

[16] 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.

[17] M.D. McKenzie, "Power Transformation and forecasting the magnitude of exchange rates," *International Journal of Forecasting,* vol. 15, pp. 49-55, 1999.

[18] 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.

[19] A.N. Refenes, A.N. Burgess and Y. Bentz, "Neural Networks in Financial Engineering: A Study in Methodology," *IEEE Trans. on Neural Networks*, vol. 8, pp. 1222-1267, 1997.

[20] C. Saunders, A. Gammerman and V. Vovk, "Ridge regression learning algorithm in dual variables," in *Proc. 15th Int. Conf. on Machine Learning ICML-98*, pp. 515-521, Madison-Wisconsin, 1998.

[21] B. Schölkopf, A. Smola and K.-M. Müller, "Nonlinear component analysis as a kernel eigenvalue problem," *Neural Computation*, vol. 10, pp. 1299-1319, 1998.

[22] B. Schölkopf, J. Shawe-Taylor, A. Smola and R.C. Williamson, "Kernel-dependent Support Vector Error Bounds," in *Proc. of the 9th Int. Conf. on Artificial Neural Networks (ICANN-99)*, pp. 304-309, Edinburgh, UK, 1999.

[23] G.W. Schwert, "Why Does Stock Market Volatility Change Over Time," *Journal of Finance,* vol. 44, pp. 1115-1153, 1989.

[24] A. Smola and B. Schölkopf, "On a kernel-based method for pattern recognition, regression, approximation and operator inversion," *Algorithmica*, vol. 22, pp. 211-231, 1998.

[25] J.A.K. Suykens and J. Vandewalle, eds., *Nonlinear Modeling: Advanced Black-Box Techniques,* Kluwer Academic Publishers, Boston, 1998.

[26] J.A.K. Suykens and J. Vandewalle, "Least squares support vector machine classifiers," *Neural Processing Letters,* vol. 9, pp. 293-300, 1999.

[27] J.A.K. Suykens, "Least Squares Support Vector Machines for Classification and Nonlinear Modelling," *Neural Network World*, vol. 10, pp. 29-48, 2000.

[28] J.A.K. Suykens and J. Vandewalle, "Recurrent least squares support vector machines," *IEEE Transactions on Circuits and Systems-I*, vol. 47, pp. 1109-1114, 2000.

[29] J.A.K. Suykens, J. Vandewalle and B. De Moor, "Optimal control by least squares support vector machines," *Neural Networks*, vol. 14, pp. 23-35, 2001.

[30] S. Taylor, *Modelling Financial Time Series,* John Wiley, New-York, 1986.

[31] T. Van Gestel, J.A.K. Suykens, B. De Moor and D.-E. Baestaens, "Volatility Tube Support Vector Machines," *Neural Network World*, vol. 10, pp. 287-297, 2000.

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

[33] C.K.I. Williams, "Prediction with Gaussian Processes: from Linear Regression to Linear Prediction and Beyond," in *Learning and Inference in Graphical Models*, M.I. Jordan, ed., Kluwer Academic Press, 1998.

# A Expressions for the variance $\sigma_z^2$ and $\det H$

The expression (16) for the variance $\sigma_z^2$ cannot be evaluated in its present form, since $\varphi(\cdot)$ is not explicitly known and hence also $\psi(x)$ and $H^{-1}$ are unknown. By defining $\Upsilon = [\varphi(x_1), \ldots, \varphi(x_N)]$, with $\Omega = \Upsilon^T\Upsilon$, the expressions for the block matrices in the Hessian (9) can be written as follows: $H_{11} = \mu I_{n_f} + \Upsilon D_\zeta \Upsilon^T$, $H_{12} = \Upsilon D_\zeta 1_v$ and $H_{22} = \sum_{i=1}^N \zeta_i = s_\zeta$. The diagonal matrix $D_\zeta \in \mathbb{R}^{N\times N}$ is defined as follows $D_\zeta = \mathrm{diag}([\zeta_1, \ldots, \zeta_N])$. The inverse Hessian $H$ is then obtained via a Schur complement type argument:

$$H^{-1} = \left( \begin{bmatrix} I_{n_f} & X \\ 0 & 1 \end{bmatrix} \begin{bmatrix} H_{11} - H_{12}H_{22}^{-1}H_{12}^T & 0 \\ 0 & H_{22} \end{bmatrix} \begin{bmatrix} I_{n_f} & 0 \\ X^T & 1 \end{bmatrix} \right)^{-1}, \tag{40}$$

with $X = H_{12}H_{22}^{-1}$. By defining $G = \Upsilon(D_\zeta - \frac{1}{s_\zeta}D_\zeta 1_v 1_v^T D_\zeta)\Upsilon^T$, we obtain that

$$H_{11} - H_{12}H_{22}^{-1}H_{12}^T = \mu I_{n_f} + G. \tag{41}$$

Notice that the maximum rank of $D_\zeta - \frac{1}{s_\zeta}D_\zeta 1_v 1_v^T D_\zeta$, with dimension $N \times N$, is equal to $N-1$, since $1_v$ is the eigenvector corresponding to the zero eigenvalue. Finally (40) becomes

$$H^{-1} = \begin{bmatrix} (\mu I_{n_f} + G)^{-1} & -(\mu I_{n_f} + G)^{-1}H_{12}H_{22}^{-1} \\ -H_{22}^{-1}H_{12}^T(\mu I_{n_f} + G)^{-1} & H_{22}^{-1} + H_{22}^{-1}H_{12}^T(\mu I_{n_f} + G)^{-1}H_{12}H_{22}^{-1} \end{bmatrix}. \tag{42}$$

The expression (16) for the variance $\sigma_z^2$ now becomes $\sigma_z^2 = \varphi(x)^T(\mu I_{n_f} + G)^{-1}\varphi(x) - \frac{2}{s_\zeta}\varphi(x)^T(\mu I_{n_f} + G)^{-1}\Upsilon D_\zeta 1_v + \frac{1}{s_\zeta} + \frac{1}{s_\zeta^2}1_v^T D_\zeta \Upsilon^T(\mu I_{n_f} + G)^{-1}\Upsilon D_\zeta 1_v$. The next step is to express the inverse $(\mu I_{n_f} + G)^{-1}$ in terms of the mapping $\varphi(x_i)$, $i = 1, \ldots, N$ using properties of linear algebra. The inverse will be calculated using the eigenvalue decomposition of the symmetric matrix $G = G^T = P_1^T D_G P_1 + cP_2^T P_2$, with $P = [P_1\, P_2]$ a unitary matrix and where $c = 0$. The matrix $P_1$ corresponds to the eigenspace corresponding to the non-zero eigenvalues and the null space is denoted by $P_2$. Indeed, since $D_\zeta - \frac{1}{s_\zeta}D_\zeta 1_v 1_v^T D_\zeta$ is a positive semi-definite matrix with rank $N-1$, there are maximally $N-1$ eigenvalues $\lambda_{G,i} > 0$ and their corresponding eigenvectors $v_{G,i}$ are a linear combination of $\Upsilon$: $v_{G,i} = c_{G,i}\Upsilon\nu_{G,i}$, with $c_{G,i}$ a normalization constant such that $v_{G,i}^T v_{G,i} = 1$. The eigenvalue problem we need to solve is the following: $\Upsilon(D_\zeta - s_\zeta^{-1}D_\zeta 1_v 1_v^T D_\zeta)\Upsilon^T v_{G,i} = \lambda_{G,i}v_{G,i}$ or

$$\Upsilon(D_\zeta - s_\zeta^{-1}D_\zeta 1_v 1_v^T D_\zeta)\Upsilon^T\Upsilon\nu_{G,i} = \lambda_{G,i}\Upsilon\nu_{G,i}. \tag{43}$$

Multiplication of the last equation to the left with $\Upsilon^T$ and applying the Mercer condition yields:

$$\Omega(D_\zeta - s_\zeta^{-1} D_\zeta 1_v 1_v^T D_\zeta)\Omega\nu_{G,i} \;=\; \lambda_{G,i}\Omega\nu_{G,i}, \tag{44}$$

which is a generalized eigenvalue problem of dimension $N$. If $\Omega$ is invertible, this corresponds to the eigenvalue problem

$$(D_\zeta - s_\zeta^{-1} D_\zeta 1_v 1_v^T D_\zeta)\Omega\nu_{G,i} \;=\; \lambda_{G,i}\nu_{G,i}. \tag{45}$$

When $\Omega$ is not invertible, one can always proceed with the non-zero eigenvalues of the generalized eigenvalue problem. The remaining $n_f - N_{eff}$ dimensional orthonormal null space $P_2$ of $G$ can not be explicitly calculated, but using the fact that $[P_1\, P_2]$ is a unitary matrix will allow us to use $P_2 P_2^T = I_{n_f} - P_1 P_1^T$. This finally yields

$$(\mu I_{n_f} + G)^{-1} = P(\mu I_{n_f} + D_G)^{-1} P^T = P_1(\mu I_{N_{eff}} + D_G)^{-1} P_1^T + \mu^{-1} P_2 P_2^T. \tag{46}$$

By defining $\theta(x) = \Upsilon^T \varphi(x)$ with $\theta_i(x) = K(x, x_i)$, $i = 1, \ldots, N$, the variance $\sigma_z^2$ can now be calculated by using Mercer's theorem and one obtains (17).

Finally, an expression for $\det(H)$ is given using the eigenvalues of $G$. The Hessian is nonsingular, mainly because of the use of a regularization term $\mu E_W$ when $n_f + 1 > N$. Thus the inverse exists and we can write $\det H^{-1} = 1/\det H$. Since $\det(H)$ is not changed the block diagonalizing (40). By combination with (41), we obtain: $\det H = N\zeta \det(\mu I_{n_f} + \zeta G)$. Since the determinant is the product of the eigenvalues, this yields:

$$\det H = s_\zeta \mu^{n_f - N_{eff}} \prod_{i=1}^{N_{eff}} (\mu + \zeta\lambda_{G,i}). \tag{47}$$

## Captions of Figures

Figure 1:

Illustration of the different steps for the modelling of financial time series using LS-SVMs within the evidence framework. The model parameters $w$, $b$, the hyperparameters $\mu$, $\zeta_i$ and the kernel parameter and relevant inputs of the time series model $\mathcal{H}$ are inferred from the data $D$ on different levels of inference. The inferred hyperparameters $\zeta_{MP,i}$ are used to estimate the parameters $\tilde{w}$, $\tilde{b}$, $\tilde{\mu}$, $\tilde{\zeta}$ and $\tilde{\sigma}$ of the volatility model $\tilde{\mathcal{H}}$. The predicted volatility is used to calculate error bars $\sigma^2_{\hat{y}_{MP,N+1}}$ on the point prediction $\hat{y}_{MP,N+1}$.

Figure 2:

Cumulative returns using Investment Strategy 2 (IS2) (transaction cost 0.1%) on the test set obtained with: (1) LS-SVM regressor with RBF-kernel (full line); (2) the ARX model (dashed-dotted); (3) the Buy-and-Hold strategy (dashed) and (4) the AR model (dotted). The LS-SVM regressor yields the highest annualized return and corresponding Sharpe Ratio as denoted in Table 2.

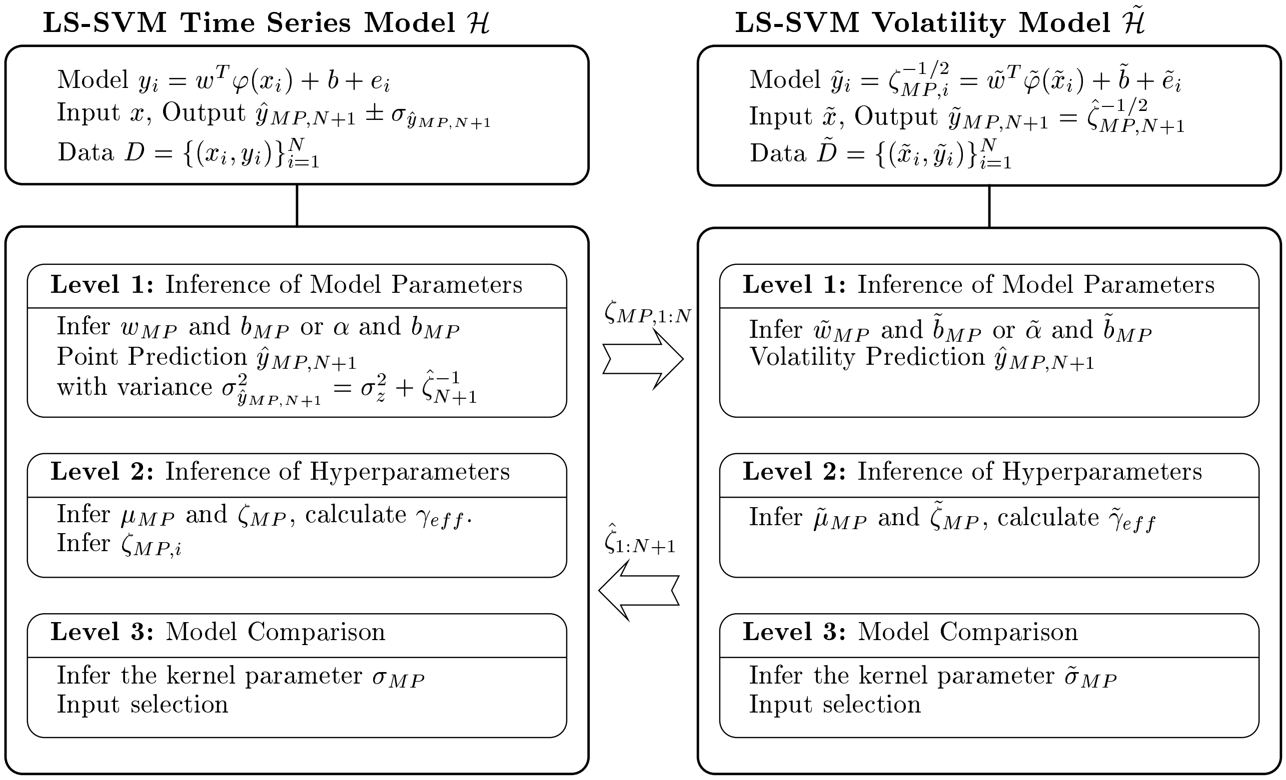
LS-SVM Time Series Model $\mathcal{H}$
Model $y_i = w^T\varphi(x_i) + b + e_i$
Input $x$, Output $\hat{y}_{MP,N+1} \pm \sigma_{\hat{y}_{MP,N+1}}$
Data $D = \{(x_i, y_i)\}_{i=1}^N$
Level 1: Inference of Model Parameters
Infer $w_{MP}$ and $b_{MP}$ or $\alpha$ and $b_{MP}$
Point Prediction $\hat{y}_{MP,N+1}$
with variance $\sigma^2_{\hat{y}_{MP,N+1}} = \sigma_z^2 + \hat{\zeta}_{N+1}^{-1}$
Level 2: Inference of Hyperparameters
Infer $\mu_{MP}$ and $\zeta_{MP}$, calculate $\gamma_{eff}$.
Infer $\zeta_{MP,i}$
Level 3: Model Comparison
Infer the kernel parameter $\sigma_{MP}$
Input selection
$\zeta_{MP,1:N}$
$\hat{\zeta}_{1:N+1}$
LS-SVM Volatility Model $\tilde{\mathcal{H}}$
Model $\tilde{y}_i = \zeta_{MP,i}^{-1/2} = \tilde{w}^T\tilde{\varphi}(\tilde{x}_i) + \tilde{b} + \tilde{e}_i$
Input $\tilde{x}$, Output $\tilde{y}_{MP,N+1} = \hat{\zeta}_{MP,N+1}^{-1/2}$
Data $\tilde{D} = \{(\tilde{x}_i, \tilde{y}_i)\}_{i=1}^N$
Level 1: Inference of Model Parameters
Infer $\tilde{w}_{MP}$ and $\tilde{b}_{MP}$ or $\tilde{\alpha}$ and $\tilde{b}_{MP}$
Volatility Prediction $\hat{y}_{MP,N+1}$
Level 2: Inference of Hyperparameters
Infer $\tilde{\mu}_{MP}$ and $\tilde{\zeta}_{MP}$, calculate $\tilde{\gamma}_{eff}$
Level 3: Model Comparison
Infer the kernel parameter $\tilde{\sigma}_{MP}$
Input selection

Figure 1:

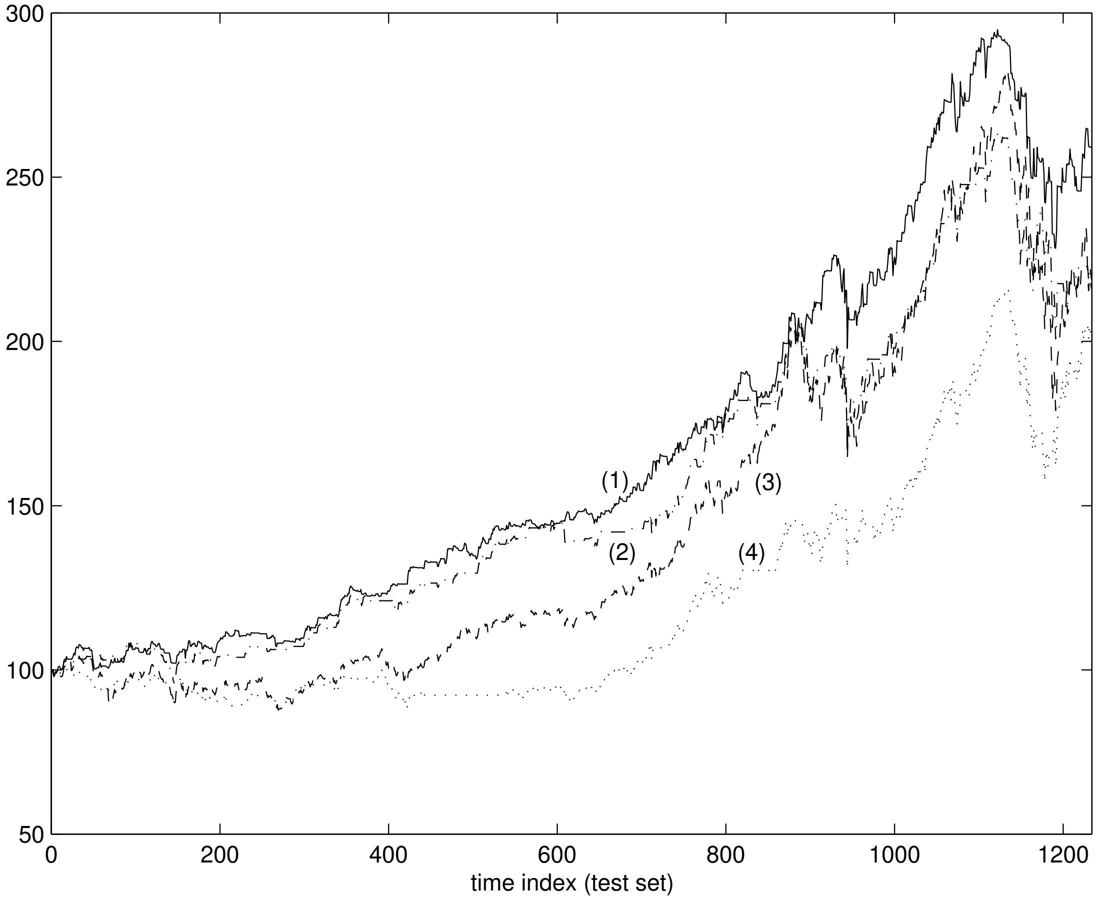
300
250
200
150
100
50
0
200
400
600
800
1000
1200
time index (test set)
(1)
(2)
(3)
(4)

Figure 2:

## Captions of Tables

Table 1:

Out of sample test set performances obtained on one step ahead prediction of the US weekly T-bill rate with different models: LS-SVM with RBF-kernel (RBF-LS-SVM), an AR(14) model and the non-parametric model (NonPar) using both rolling (Ro) and non-rolling (NRo) approaches. First, the sample MSE and corresponding sample standard deviation are reported. Then the Directional Accuracy is assessed by the the Percentage of Correct Sign Predictions (PCSP), the Pesaran-Timmerman statistic (PT) and the corresponding $p$-value. These $p$-values illustrate that the LS-SVM with RBF-kernel (RBF-LS-SVM) clearly performs better than the other models with respect to the directional accuracy criterion.

Table 2:

Test set performances of the LS-SVM times series and volatility model obtained on the one step ahead prediction of the DAX30 index. The LS-SVM time series model with RBF-kernel is compared with an ARX10 and AR20 model and a Buy-and-Hold (B&H) strategy. The RBF-LS-SVM clearly achieves a better directional accuracy. In combination with investment strategies IS1 and IS2 the LS-SVM yields also better annualized returns (Re) and risks (Ri) resulting in a higher Sharpe Ratio (SR). In the second part of the table, the LS-SVM volatility model is compared with three AR10 models using different power transformations, a log transformed AR10 model and the GARCH(1,1) model. The RBF-LS-SVM model achieves better out of sample test set performances than the other models with respect to the MSE, MAE criteria, while a comparable Negative Log Likelihood (NLL) is obtained with respect to the GARCH model.

| | MSE | | PCSP | PT | $p$-value |
|---|---|---|---|---|---|
| RBF-LS-SVM (Ro) | 0.172 | (0.316) | 62% | 3.24 | 0.0011 |
| RBF-LS-SVM (NRo) | 0.173 | (0.318) | 61% | 3.10 | 0.0018 |
| AR(14) (Ro) | 0.183 | (0.346) | 56% | 1.76 | 0.0782 |
| AR(14) (NRo) | 0.184 | (0.347) | 54% | 1.23 | 0.2170 |
| NonPar [8] (Ro) | 0.162 | - | 56% | - | - |

Table 1:

| Time Series Model | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|
| | Directional Accuracy | | | Inv. Strat. 1 (IS1) | | | Inv. Strat. 2 (IS2) | | |
| | PCSP | PT | $p$-value | $SR_1$ | $Re_1$ | $Ri_1$ | $SR_2$ | $Re_2$ | $Ri_2$ |
| LS-SVM | 56.8 | 4.39 | 1.1e-5 | 1.38 | 20.24 | 14.62 | 1.49 | 21.66 | 14.58 |
| ARX10 | 55.0 | 3.20 | 0.001 | 0.65 | 9.99 | 15.19 | 1.03 | 16.63 | 16.03 |
| AR20 | 48.1 | -1.27 | 0.202 | -0.40 | -4.75 | 15.34 | 0.89 | 14.84 | 16.52 |
| B&H | - | - | - | 0.81 | 16.35 | 20.29 | 0.81 | 16.35 | 20.29 |
| Volatility Model | | | | | | | | | |
| | MSE | | | MAE | | | NLL | | |
| LS-SVM | 0.474 | (1.548) | | 0.461 | (0.477) | | 1.374 | (2.563) | |
| $\lvert AR10\rvert$ | 0.490 | (1.655) | | 0.470 | (0.492) | | 1.613 | (2.129) | |
| $\lvert AR10\rvert^{1.1}$ | 0.485 | (1.621) | | 0.471 | (0.486) | | 1.572 | (1.892) | |
| $\lvert AR10\rvert^2$ | 0.489 | (1.479) | | 0.501 | (0.474) | | 1.530 | (1.602) | |
| logAR10 | 0.613 | (2.161) | | 0.497 | (0.606) | | 3.253 | (7.931) | |
| GARCH(1,1) | 0.480 | (1.351) | | 0.516 | (0.452) | | 1.314 | (1.059) | |

Table 2: