Journal of Forecasting
J. Forecast. **25**, 77–100 (2006)
Published online in Wiley InterScience
(www.interscience.wiley.com) **DOI**: 10.1002/for.975

A Bayesian Nonlinear Support Vector Machine Error Correction Model

TONY VAN GESTEL,[1,2*] MARCELO ESPINOZA,[2]
BART BAESENS,[3] JOHAN A. K. SUYKENS,[2]
CARINE BRASSEUR[4] AND BART DE MOOR[2]
[1] *Dexia Group, Belgium*
[2] *Katholieke Universiteit Leuven, Belgium*
[3] *School of Management, University of Southampton, UK*
[4] *Fortis Bank Brussels, Belgium*

ABSTRACT

The use of linear error correction models based on stationarity and cointegration analysis, typically estimated with least squares regression, is a common technique for financial time series prediction. In this paper, the same formulation is extended to a nonlinear error correction model using the idea of a kernel-based implicit nonlinear mapping to a high-dimensional feature space in which linear model formulations are specified. Practical expressions for the nonlinear regression are obtained in terms of the positive definite kernel function by solving a linear system. The nonlinear least squares support vector machine model is designed within the Bayesian evidence framework that allows us to find appropriate trade-offs between model complexity and in-sample model accuracy. From straightforward primal–dual reasoning, the Bayesian framework allows us to derive error bars on the prediction in a similar way as for linear models and to perform hyperparameter and input selection. Starting from the results of the linear modelling analysis, the Bayesian kernel-based prediction is successfully applied to out-of-sample prediction of an aggregated equity price index for the European chemical sector.

KEY WORDS financial time series prediction; least squares support vector machines; Bayesian inference; error correction mechanism; kernel-based learning

INTRODUCTION

Financial time series forecasting is a dynamic field with important contributions coming from many disciplines. The issue of forecasting financial time series has traditionally been seen as a difficult task, since the data generating process is dominated by stochastic rather than deterministic components. Although the efficient markets hypothesis (Bachelier, 1900; Fama, 1965) states that stock returns are unpredictable, recent modelling techniques seem to suggest that stock returns are pre-

* Correspondence to: Tony Van Gestel, Credit Risk Modelling, Risk Management, Dexia Group, Square Meeus 1, B-1000 Brussels, Belgium. E-mail: tony.vangestel@dexia.com

dictable to some degree (Brock *et al.*, 1992; Campbell *et al.*, 1997; Lo *et al.*, 2000; Sullivan *et al.*, 1999).

Within the linear framework, usually forecasting models are based on stationarity considerations of the series at hand (Campbell *et al.*, 1997; Granger and Newbold, 1986; Hamilton, 1994). In this sense, evidence of possible cointegration can be exploited into an error correction mechanism formulation. This model is typically estimated by ordinary least squares, applying input selection to control the model complexity and avoid overfitting on the training set.

The universal approximation property of multilayer perceptrons (MLPs) motivated their use for nonlinear financial time series prediction (Granger and Terasvirta, 1993; Hutchinson *et al.*, 1994; Refenes and Zapranis, 1999). While powerful design techniques like the Bayesian evidence framework (MacKay, 1995) have been developed, the practical use of neural networks suffers from drawbacks like the nonconvex optimization problem with multiple local minima and the choice of the number of hidden neurons. In support vector machines (SVMs), least squares support vector machines (LS-SVMs) and related kernel-based prediction techniques (Schölkopf and Smola, 2002; Suykens *et al.*, 2002; Vapnik, 1998), the solution follows from a convex optimization problem. Basically these methods map the inputs in a nonlinear way, first into a high kernel-induced feature space, in which ridge regression is applied in the case of LS-SVMs. The solution follows from a linear Karush–Kuhn–Tucker system in the dual space in terms of the positive definite kernel function by applying Mercer's theorem (Schölkopf and Smola, 2002; Suykens *et al.*, 2002; Vapnik, 1998).

In this paper, a nonlinear error correction model (ECM) formulation is estimated using LS-SVMs to predict an aggregated equity price index for the European chemical sector. First a stationarity and cointegration analysis is performed to define a good linear model formulation. This model is used as a starting point to design the nonlinear kernel-based model within the Bayesian evidence framework. The parameters and inputs of the LS-SVM are estimated[1] in the Bayesian evidence framework (Van Gestel *et al.*, 2001, 2002). The Bayesian framework embodies Occam's razor to find an optimal trade-off between training set accuracy and model complexity in a similar way as the Akaike and Bayesian information criteria (Akaike, 1974; Schwarz, 1978). The linear and nonlinear predictions are compared in terms of directional accuracy and market timing ability.

This paper is organized as follows. The initial problem definition, stationarity analysis and linear model specification are reviewed and applied in the next section. The design and application of nonlinear kernel-based regression within the evidence framework is presented in the third section. The final results are discussed in the fourth section.

LINEAR MODELLING

In financial engineering applications, the importance of having good forecasting modelling tools is straightforward. Risk and portfolio management are mainly based on such tools, therefore any improvement over traditional techniques can lead to competitive advantages. Traditional forecasting based on linear models is built upon the concepts of stationarity and cointegration.

Stationarity

A linear model formulation to predict an output $y \in \mathbb{R}$ based on n explanatory input variables $\boldsymbol{x} = [x_1; \ldots; x_n] = [x_1, \ldots, x_n]^T \in \mathbb{R}^n$ can be written as

[1] A Matlab toolbox for the LS-SVM formulation and Bayesian inference is available from http://www.esat.kuleuven.ac.be/sista/lssvmlab.

DOI: 10.1002/for

$$y = \boldsymbol{w}^T \boldsymbol{x} + b + e \tag{1}$$

with $\boldsymbol{w} \in \mathbb{R}^n$ is a coefficient vector and $b \in \mathbb{R}$ a bias term. Having a set of $n_{\mathcal{D}}$ observations $\mathcal{D} = \{(\boldsymbol{x}_t, y_t)\}_{t=1}^{n_{\mathcal{D}}}$, the most usual technique to estimate a linear model is by using the ordinary least squares (OLS) estimator

$$\min_{w,b} \sum_{t=1}^{n_{\mathcal{D}}} (y_t - (\boldsymbol{w}^T \boldsymbol{x}_t + b))^2 \tag{2}$$

with error $e_t = y_t - (\boldsymbol{w}^T \boldsymbol{x}_t + b)$. Defining $\boldsymbol{y} = [y_1; y_2; \ldots; y_{n_{\mathcal{D}}}] = [y_1, y_2, \ldots, y_{n_{\mathcal{D}}}]^T \in \mathbb{R}^{n_{\mathcal{D}}}$, $\mathbf{1} = [1, 1, \ldots, 1]^T \in \mathbb{R}^{n_{\mathcal{D}}}$ and $\boldsymbol{X} = [\boldsymbol{x}_1, \boldsymbol{x}_2, \ldots, \boldsymbol{x}_{n_{\mathcal{D}}}]^T \in \mathbb{R}^{n_{\mathcal{D}} \times n}$ the solution to (2) is obtained from the linear set of equations

$$[\hat{\boldsymbol{w}}; \hat{b}] = \left([\boldsymbol{X}, \boldsymbol{1}]^{\mathbf{T}} [\boldsymbol{X}, \boldsymbol{1}]\right)^{-1} [\boldsymbol{X}, \boldsymbol{1}]^{\mathbf{T}} \boldsymbol{y} \tag{3}$$

Although the general assumptions underlying the application of the OLS are described in the Gauss–Markov conditions, in the particular scope of time series forecasting it is required that the series involved should be stationary. It has been widely recognized that performing linear regression with nonstationary series has the potential to lead to serious inference errors (Granger and Newbold, 1974). Some of the known problems when performing OLS estimations with nonstationary series are, for instance, the identification of spurious relationships between unrelated variables, or the non-convergence of the $\hat{\boldsymbol{w}}$ estimates (Maddala and Kim, 1998). Formally, for a process y_t to be (weakly) stationary, it must satisfy the following set of properties: $E[y_t] = \mu_y$, $E[(y_t - \mu_y)^2] = \text{var}(y_t) = \sigma_y^2 = \gamma(0)$, $E[(y_t - \mu_y)(y_{t-\tau} - \mu_y)] = \text{cov}(y_t, y_{t-\tau}) = \gamma(\tau)$, where the mean and variance of y_t are constant, and the covariances depend only on the time *interval* τ and not on the particular moment of time t.

One of the most common tests for stationarity of a time series y_t is based on the so-called augmented Dickey–Fuller (ADF) regression (Dickey and Fuller, 1979): $\Delta y_t = \alpha + \rho y_{t-1} + \Sigma_{j=1}^{q} \beta_j \Delta y_{t-j} + e_t$. Under the null hypothesis of nonstationarity ($H_0: \rho = 1$), the t-statistic of the estimated coefficient $\hat{\rho}$ will follow a nonstandard distribution, usually known as the Dickey–Fuller (DF) distribution. If the corresponding t-statistic for the coefficient y_{t-1} is above the critical value of the ADF test, then the null hypothesis of nonstationarity cannot be rejected (Rao, 1994).

Cointegration

If a series $y_t^{(l)}$ of levels is found to be nonstationary on its original levels, one usual transformation is to take first differences and work with the transformed variable $y_t^{(d)} = \Delta y_t^{(l)} = y_t^{(l)} - y_{t-1}^{(l)}$. However, before attempting to transform all nonstationary variables into first differences, it is useful to explore for possible cointegration between the dependent variable and any subset of the explanatory variables. For the case of two nonstationary variables $y_t^{(l)}$, $x_t^{(l)}$ testing for cointegration involves testing for stationarity of the residuals in the regression

$$y_t^{(l)} = \beta_0 + \beta_1 x_t^{(l)} + e_t \tag{4}$$

Thus, finding stationary residuals from the regression above is equivalent to finding a cointegrating relationship between the variables, where the stationary cointegrating linear combination can be estimated as $z_t = y_t^{(l)} - \hat{\beta}_0 - \hat{\beta}_1 x_t^{(l)}$.

DOI: 10.1002/for

If cointegration exists, then it is possible to take advantage of this long-term relationship and use it to model the short-term behaviour of the system. It was proved that any cointegrating system can have an equivalent ECM representation (Engle and Granger, 1987). Following from the example above, if the series $x_t^{(I)}$ and $y_t^{(I)}$ do cointegrate, then their corresponding linear ECM equals

$$y_t^{(d)} = \alpha_0 + \alpha_1 z_{t-1} + \sum_{j=1}^{p} b_j y_{t-j}^{(d)} + \sum_{j=1}^{p} c_j x_{t-j}^{(d)} + e_t \tag{5}$$

which can be written more generally as

$$y_t^{(d)} = f\left(z_{t-1}; y_{t-1}^{(d)}, \ldots, y_{t-p}^{(d)}; x_{t-1}^{(d)}, \ldots, x_{t-p}^{(d)}\right) + e_t \tag{6}$$

It is possible to also include some additional external variables that can help to improve the model, but the central concept of the ECM is as shown above. The extension[2] using more than two variables is straightforward.

The number of lags p for an autoregressive AR(p) model can be heuristically defined based on the partial autocorrelation function pacf(ρ) (Hamilton, 1994). For an autoregressive formulation of the stationary series $y(t)^{(d)}$ of order p (like the ECM above), it can be shown that the pacf(ρ) function will drop to zero after $\rho > p$ (Box and Jenkins, 1970).

NONLINEAR KERNEL-BASED MODELLING AND PREDICTION

Least squares support vector machines

A straightforward way to extend the linear models (1), (5) and (19) to a nonlinear model is to preprocess the inputs $\boldsymbol{x}$ in a nonlinear way by a mapping

$$\boldsymbol{\varphi}: \mathbb{R}^n \to \mathbb{R}^{n_\varphi} : \mathbf{x} \mapsto \varphi(\mathbf{x}) \tag{7}$$

where the feature vector $\boldsymbol{\varphi}(\boldsymbol{x})$ is typically high (or even infinite)-dimensional. Given the nonlinear mapping, the ECM model (6) is assumed to be of the following form:

$$y = \boldsymbol{w}^T \boldsymbol{\varphi}(\boldsymbol{x}) + b + e \tag{8}$$

where $f(\boldsymbol{x}) \simeq \boldsymbol{w}^T \boldsymbol{\varphi}(\boldsymbol{x}) + b$. Given this nonlinear mapping, the coefficient vector $\boldsymbol{w}$ is estimated by solving the (regularized) least squares problem in the primal or feature space:

$$\min_{w,b,e} J_1(\boldsymbol{w}, b) = \frac{\mu}{2} \boldsymbol{w}^T \boldsymbol{w} + \frac{\zeta}{2} \sum_{t=1}^{n_D} e_t^2 \tag{9}$$

$$\text{s.t.} \quad e_t = y_t - (\boldsymbol{w}^T \boldsymbol{\varphi}(\boldsymbol{x}_t) + b), \quad t = 1, \ldots, n_D \tag{10}$$

[2] The model depicted so far is known as the Engle–Granger two-step approach for cointegration. When using vector formulations, the so-called Johansen procedure (Johansen, 1988) is applied.

DOI: 10.1002/for

As the nonlinear mapping $\boldsymbol{\varphi}$ is high-dimensional, the regularization term $\frac{\mu}{2}\boldsymbol{w}^T\boldsymbol{w}$ is introduced to avoid overfitting the training data. The parameters μ and ζ determine the trade-off between regularization $J_w = (1/2)\boldsymbol{w}^T\boldsymbol{w}$ and error minimization $J_e = (1/2)\Sigma_{t=1}^{n_\mathcal{D}} e_t^2$.

A key element of support vector machines and kernel-based learning methods is that the nonlinear mapping $\boldsymbol{\varphi}(\boldsymbol{x})$ is not explicitly known. Instead it is implicitly defined from Mercer's theorem in terms of the positive-definite kernel function

$$K(\boldsymbol{x}_i, \boldsymbol{x}_j) = \boldsymbol{\varphi}(\boldsymbol{x}_i)^T \boldsymbol{\varphi}(\boldsymbol{x}_j) \tag{11}$$

Some commonly used kernel functions are

$$\begin{aligned}
&1.\ K(\boldsymbol{x}_i, \boldsymbol{x}_j) = \boldsymbol{x}_i^T \boldsymbol{x}_j && \text{(linear kernel)}\\
&2.\ K(\boldsymbol{x}_i, \boldsymbol{x}_j) = (1 + \boldsymbol{x}_i^T \boldsymbol{x}_j / c)^d && \text{(polynomial kernel of degree } d\text{)}\\
&3.\ K(\boldsymbol{x}_i, \boldsymbol{x}_j) = \exp\left(-\frac{\|\boldsymbol{x}_i - \boldsymbol{x}_j\|_2^2}{\sigma^2}\right) && \text{(radial basis function kernel)}
\end{aligned} \tag{12}$$

where $d \in \mathbb{N}$ and $c,\ \sigma \in \mathbb{R}^+$ are tunable parameters.

In order to solve the constrained optimization problem (9) and (10), one constructs the Lagrangian

$$\mathcal{L}(\mathbf{w}, \mathbf{e}, \boldsymbol{\alpha}, b) = \frac{\mu}{2}\mathbf{w}^T\mathbf{w} + \frac{\zeta}{2}\sum_{t=1}^{n_\mathcal{D}} e_t^2 + \sum_{t=1}^{n_\mathcal{D}} \alpha_t (y_t - (\mathbf{w}^T\boldsymbol{\varphi}(\mathbf{x}_t) + b) - e_t) \tag{13}$$

where the scalars $\alpha_t \in \mathbb{R}$ are the Lagrange multipliers associated with the equality constraints (10) and are called support values. The conditions for optimality are

$$\begin{cases}
\dfrac{\partial \mathcal{L}}{\partial \boldsymbol{w}} = 0 \leftrightarrow \mu\boldsymbol{w} = \sum_{t=1}^{n_\mathcal{D}} \alpha_t \boldsymbol{\varphi}(\boldsymbol{x}_t) & \leftrightarrow \quad \mu\boldsymbol{w} - \boldsymbol{\Phi}^T\boldsymbol{\alpha} = 0\\
\dfrac{\partial \mathcal{L}}{\partial b} = 0 \leftrightarrow \sum_{t=1}^{n_\mathcal{D}} \alpha_t = 0 & \leftrightarrow \quad \boldsymbol{\alpha}^T 1 = 0\\
\dfrac{\partial \mathcal{L}}{\partial e_t} = 0 \leftrightarrow \alpha_t = \zeta e_t, \quad (t = 1, \ldots, n_\mathcal{D}) & \leftrightarrow \quad \boldsymbol{\alpha} - \zeta\boldsymbol{e} = 0\\
\dfrac{\partial \mathcal{L}}{\partial \alpha_t} = 0 \leftrightarrow e_t = y_t - \boldsymbol{w}^T\boldsymbol{\varphi}(\boldsymbol{x}_t) - b, \quad (t = 1, \ldots, n_\mathcal{D}) & \leftrightarrow \quad \boldsymbol{\Phi}\boldsymbol{w} + b1 + \boldsymbol{e} = \boldsymbol{y}
\end{cases} \tag{14}$$

in which we have defined $\boldsymbol{\Phi} = [\boldsymbol{\varphi}(\boldsymbol{x}_1), \ldots, \boldsymbol{\varphi}(\boldsymbol{x}_{n_\mathcal{D}})]^T \in \mathbb{R}^{n_\mathcal{D} \times n_\varphi}$, $\boldsymbol{\alpha} = [\alpha_1, \ldots, \alpha_{n_\mathcal{D}}]^T \in \mathbb{R}^{n_\mathcal{D}}$, $\boldsymbol{e} = [e_1, \ldots, e_{n_\mathcal{D}}]^T \in \mathbb{R}^{n_\mathcal{D}}$, $\boldsymbol{y} = [y_1, \ldots, y_{n_\mathcal{D}}]^T \in \mathbb{R}^{n_\mathcal{D}}$ and $\mathbf{1} = [1, \ldots, 1]^T \in \mathbb{R}^{n_\mathcal{D}}$. For the linear case (e.g. a linear kernel) one typically has $n_{\boldsymbol{\varphi}} = n << n_\mathcal{D}$ and after elimination of $\boldsymbol{e}$ and $\boldsymbol{\alpha}$, one solves the $(n_\varphi + 1) \times (n_\varphi + 1)$ linear system in the primal space

$$\begin{bmatrix} \boldsymbol{\Phi}^T\boldsymbol{\Phi} + \dfrac{\mu}{\zeta}\boldsymbol{I}_{n_\varphi} & \boldsymbol{\Phi}^T 1 \\ 1^T\boldsymbol{\Phi} & n_\mathcal{D} \end{bmatrix} \begin{bmatrix} \boldsymbol{w} \\ b \end{bmatrix} = \begin{bmatrix} \boldsymbol{\Phi}^T\boldsymbol{y} \\ 1^T\boldsymbol{y} \end{bmatrix} \tag{15}$$

DOI: 10.1002/for

In nonlinear kernel-based regression, one usually has $n_\varphi >> n_\mathcal{D}$ and, moreover, the feature vector $\boldsymbol{\varphi}(\boldsymbol{x})$ is only implicitly defined in terms of the kernel function K from (11). Eliminating $\boldsymbol{w}$ and $\boldsymbol{e}$ from (14), one obtains the linear Karush–Kuhn–Tucker (KKT) system of dimension $(n_\mathcal{D} + 1) \times (n_\mathcal{D} + 1)$ in the dual space (Suykens *et al.*, 2002)

$$\begin{bmatrix} \frac{1}{\mu}\boldsymbol{\Omega} + \frac{1}{\zeta}\boldsymbol{I}_{n_\mathcal{D}} & \mathbf{1} \\ \mathbf{1}^T & 0 \end{bmatrix} \begin{bmatrix} \boldsymbol{\alpha} \\ b \end{bmatrix} = \begin{bmatrix} \boldsymbol{y} \\ 0 \end{bmatrix} \tag{16}$$

where the Mercer condition (11) is applied in the matrix $\boldsymbol{\Omega} = \boldsymbol{\Phi}\boldsymbol{\Phi}^T \in \mathbb{R}^{n_\mathcal{D} \times n_\mathcal{D}}$ with elements $\boldsymbol{\Omega}_{ij} = K(\boldsymbol{x}_i, \boldsymbol{x}_j)$, $i, j = 1, \ldots, n_\mathcal{D}$ and guarantees that $\boldsymbol{\Omega} \geq 0$. The primal–dual formulations also allow us to make extensions to nonlinear generalized least squares regression in a straightforward way, as typically used in financial forecasting (Campbell *et al.*, 1997).

Given the support values $\boldsymbol{\alpha}$ and bias term b, one obtains the predicted value $\hat{y}$ corresponding to a new input $\boldsymbol{x}$ as a weighted sum of the kernel functions evaluated in the new data point and the training data points:

$$\hat{y} = \hat{\boldsymbol{w}}^T \boldsymbol{\varphi}(\boldsymbol{x}) + \hat{b} = \frac{1}{\mu} \sum_{t=1}^{n_\mathcal{D}} \alpha_t K(\boldsymbol{x}, \boldsymbol{x}_t) + \hat{b} \tag{17}$$

Bayesian inference for model design

Given the primal–dual formulations, it is clear how to estimate the model parameters $\hat{\boldsymbol{w}}$, $\hat{b}$ and point prediction $\hat{y}$. However, the regularization and kernel function parameters still have to be tuned from the given training data. In this subsection, the design is done within the Bayesian evidence framework (Van Gestel *et al.*, 2001, 2002), depicted in Figure 1.

The model parameters $\boldsymbol{w}$, b, the hyperparameters μ, ζ and model structure $\mathcal{H}$ (corresponding, e.g., to the input set and/or tunable kernel parameters) are inferred by applying Bayes' formula on three different levels:

1. On the first level, it is assumed that the hyperparameters μ, ζ and model $\mathcal{H}$ are given. Applying Bayes' formula, the posterior probability of the model parameters $\boldsymbol{w}$ and b is obtained:

$$p(\boldsymbol{w}, b|\mathcal{D}, \log\mu, \log\zeta, \mathcal{H}) = \frac{p(\mathcal{D}|\boldsymbol{w}, b, \log\mu, \log\zeta, \mathcal{H})\, p(\boldsymbol{w}, b|\log\mu, \log\zeta, \mathcal{H})}{p(\mathcal{D}|\log\mu, \log\zeta, \mathcal{H})}$$

   The evidence shows that $p(\mathcal{D}|\log\mu, \log\zeta, \mathcal{H})$ does not depend upon the model parameters $\boldsymbol{w}$ and b and is a normalizing constant such that the left-hand side is a probability density function $\int\int \ldots \int p(\boldsymbol{w}, b|\mathcal{D}, \log\mu, \log\zeta, \mathcal{H}) dw_1 \ldots dw_{n_\varphi} db = 1$. The ridge regression cost function (9) and (10) is obtained by taking the negative logarithm of the posterior $p(\boldsymbol{w}, b|\mathcal{D}, \log\mu, \log\zeta, \mathcal{H})$ using proper choices for the prior $p(\boldsymbol{w}, b|\log\mu, \log\zeta, \mathcal{H})$ and the likelihood $p(\mathcal{D}|\boldsymbol{w}, b, \log\mu, \log\zeta, \mathcal{H})$. In the dual space, the parameters $\boldsymbol{\alpha}$ and b are obtained from the linear KKT-system (16).
2. The hyperparameters μ and ζ are inferred on the second level:

$$p(\log\mu, \log\zeta|\, \mathcal{D}, \mathcal{H}) = \frac{p(\mathcal{D}|\log\mu, \log\zeta, \mathcal{H})\, p(\log\mu, \log\zeta, \mathcal{H})}{p(\mathcal{D}|\mathcal{H})}$$

DOI: 10.1002/for

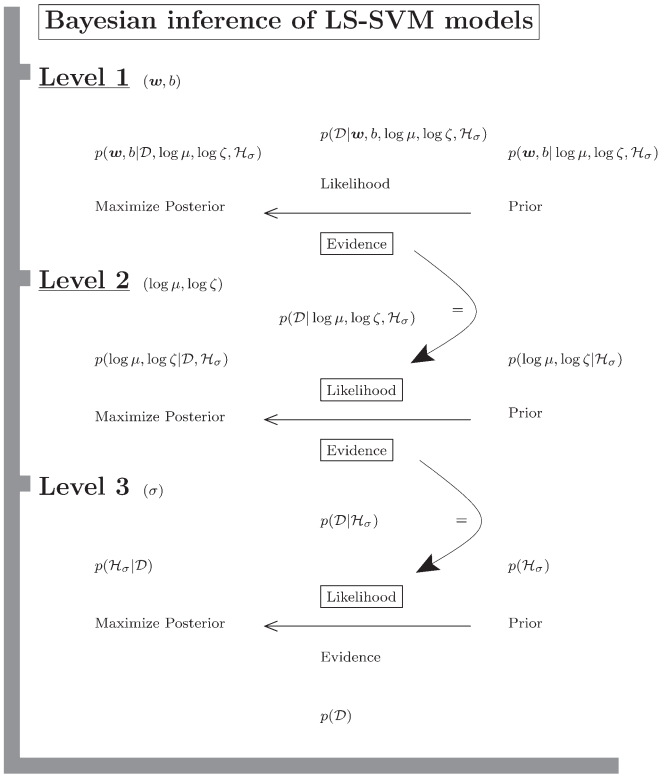

Figure 1. Different levels of Bayesian inference. The posterior probability of the model parameters $\boldsymbol{w}$ and b is inferred from the data $\mathcal{D}$ by applying Bayes' formula on the first level for given hyperparameters μ (prior) and ζ (likelihood) and the model structure $\mathcal{H}$. The model parameters are obtained by maximizing the posterior. The evidence on the first level becomes the likelihood on the second level when applying Bayes' formula to infer μ and ζ (with $\gamma = \zeta/\mu$) from the given data $\mathcal{D}$. The optimal hyperparameters μ_{MP} and ζ_{MP} are obtained by maximizing the corresponding posterior on level 2. Model comparison is performed on the third level in order to compare different model structures, e.g., with different candidate input sets and/or different kernel parameters

Taking the negative logarithm of the posterior $p(\log\mu, \log\zeta|\mathcal{D},\mathcal{H})$, the cost function (34) is obtained in order to optimize μ and ζ, which can be further simplified into optimizing $\gamma = \zeta/\mu$ from (38).

3. The posterior probability of the model $\mathcal{H}$ is obtained on the third level as

$$p(\mathcal{H}|\mathcal{D}) \propto p(\mathcal{D}|\mathcal{H})p(\mathcal{H})$$

As there are infinitely many models, the evidence is omitted here. The candidate models $\mathcal{H}_i$ (with different kernel parameters σ_i or different sets of explanatory inputs I_i) are compared using the expression (44).

Observe that the evidence on level 1 and 2 is equal to likelihoods on level 2 and 3, respectively, which implies that one also needs expressions of the lower levels in order to perform inference on the higher levels. The mathematical details of the Bayesian inference are given in Appendix B.

DOI: 10.1002/for

The LS-SVM is designed in the Bayesian framework using the following steps:

1. Preprocess the data by completing missing values and handling outliers. Standardize the inputs to zero mean and unit variance.
2. Define models $\mathcal{H}_i$ by choosing a candidate input set I_i, a kernel function K_i and a kernel parameter, e.g., σ_i in the RBF kernel case. For all models $\mathcal{H}_i$, with $i = 1, \ldots, n_{\mathcal{H}}$ (with $n_{\mathcal{H}}$ the number of models to be compared), compute the level 3 posterior:
   (a) Find the optimal hyperparameters μ_{MP} and ζ_{MP} by solving the scalar optimization problem (38) in $\gamma = \zeta/\mu$ related to maximizing the level 2 posterior.[3] With the resulting γ_{MP}, compute the effective number of parameters from (35), the hyperparameters μ_{MP} and ζ_{MP}.
   (b) Evaluate the level 3 posterior (44) for model comparison.
3. Select the model $\mathcal{H}_i$ with maximal evidence. If desired, refine the model tuning parameters K_i, σ_i, I_i to further optimize the classifier and go back to step 2; else, go to step 4.
4. Given the optimal $\mathcal{H}_i^*$, calculate $\boldsymbol{\alpha}$ and b from (16), with kernel K_i, parameter σ_i and input set I_i.

Given the optimized model and its parameters, the prediction for a new observation is obtained by first standardizing the inputs in exactly the same way as the training set and then evaluating (17) for the prediction $\hat{y}$ and (48) for the variance σ_y^2 indicating the uncertainty on the prediction due to the noise and model uncertainty.

CASE STUDY: STOCK MARKET PREDICTION

In this application, the goal is to predict the performance of an aggregated equity price index for the European chemical sector. The data set consists of 13 variables in weekly values ranging from April 1986 to February 2001, provided by Datastream (787 observations). The dependent variable is labelled CHMCLEM, and the prediction will be made in a one-week-ahead schedule. The set of (candidate) explanatory variables selected by the financial analyst includes macroeconomic indices (industrial production index, gross domestic product, consumer price index), some specific market series (oil price, raw materials price) and some financial variables (bonds, exchange rate dollar/euro, fibor 3-month interest rate). The first 600 observations are selected for initial model estimation (from April 1986 to mid-September 1997). Details on the behaviour of the data (in logarithms) within the estimation sample are described in Table I.

Performance measures

The model is evaluated in a forward way on the time period $t = 601, \ldots, 787$. The out-of-sample forecasts for $\text{CHMCLEM}_t^{(d)}$ are computed as follows. The first forecast ($t = 601$) is computed from the initial model. Then, the first observation is dismissed ($t = 1$) and the new observation is incorporated in the model for re-estimation ($t = 601$). In this way, each forecast is computed from a model estimated with the last 600 observations available. Only the variables found to be relevant are used in this moving window approach, it is assumed that the relevance found in the initial 600 data points will also hold out-of-sample.

[3] Observe that this implies in each iteration step maximizing the level 1 posterior in $\boldsymbol{w}$ and b.

DOI: 10.1002/for

Table I. Name and description of the dependent or output variable and of the candidate explanatory or input variables as selected by the financial analyst. The mean, maximum, minimum and standard deviation of the data in the training sample are also reported

Variable	Description	Min	Max	Mean	St.Dev.
CHMCLEM	Index/Chemical Sector	5.6503	6.8340	6.1555	0.2668
CPI	Consumer Price Index	4.2885	4.6201	4.4607	0.1080
FIBOR3M	Interest Rate	1.1309	2.2935	1.7294	0.3750
EUOOCIPDG	Industrial Production Index	4.4224	4.6482	4.5423	0.0586
OILBREN	Oil Price	2.1955	3.7098	2.8871	0.1895
ETYEUSP	Ethylene Price	5.6922	6.7557	6.0901	0.2562
PHREUSN	Specific Polymer Price	−1.0564	−0.0030	−0.3875	0.2582
GDP	European GDP	6.9418	7.2249	7.1061	0.0841
USEURWD	Exchange Rate US/Euro	−0.0739	0.4240	0.2123	0.0857
CHLORED	Chlorine Price	1.7373	2.2330	2.0123	0.1231
BMBD02Y	Bond 2 years (index)	4.5641	4.6774	4.6331	0.0305
BMBD05Y	Bond 5 years (index)	4.4890	4.6659	4.5862	0.0533
BMBD010Y	Bond 10 years (index)	4.4323	4.6955	4.5689	0.0654

The quality of the forecasts is assessed as follows. One usual way to quantify the quality of the forecasts is by using magnitude accuracy measures, such as the mean squared error (MSE) or the mean absolute error (MAE). Additionally, in financial applications, the percentage of correct sign predictions (PCSP) is often used, or the success in forecasting only the *direction* of the change rather than its magnitude (in plain terms, if the stock price rises or falls). The PCSP significance is assessed by using the Pesaran–Timmerman test statistic (PTstat) and the corresponding p-value (Pesaran and Timmerman, 1992). This test discriminates if the PCSP is obtained randomly or not. A PTstat above 2 allows us to reject the null hypothesis of no dependency between the predictions and the observations.

Nevertheless, a simple trading exercise is performed using a transaction cost of 0.1% (10 bps as in Refenes and Zapranis, 1999) to assess the market timing ability. In investment strategy 1 (IS1), a naive allocation of 100% equities or cash is implemented, based on the sign of the prediction. The corresponding Sharpe ratio (SR, defined as the ratio between the return and the risk of a particular asset), equivalent yearly return (Re) and risk (Ri) on the test set are computed. A more advanced trading rule involves the use of the uncertainty or moderated output on the prediction for doing the actual trade: trading is still based on the sign of the prediction, but only when the ratio $\hat{y}/\sigma_y$ exceeds a threshold value (investment strategy 2). For comparison, the same indicators for a simple buy&hold strategy (buy the asset today and sell it at the end period) are calculated.

Stationarity and cointegration analysis

All series were found to be nonstationary by using the ADF test. Nevertheless, evidence of linear cointegration between the variables $\text{CHMCLEM}^{(I)}$, $\text{FIBOR3M}^{(I)}$, $\text{CPI}^{(I)}$ and $\text{IP}^{(I)}$ was found. The residuals e_t in the linear regression

$$\text{CHMCLEM}_t^{(I)} = a_0 + a_1\text{CPI}_t^{(I)} + a_2\text{FIBOR3M}_t^{(I)} + a_3\text{IP}_t^{(I)} + e_t \tag{18}$$

are found to be stationary, as reported in Table II. It can be seen from the critical values of the ADF test that the null hypothesis of nonstationarity of the residuals can be rejected.

DOI: 10.1002/for

Table II. Estimates of the cointegrating regression, where the variables are used in nonstationary levels. The high R^2 value and the low Durbin–Watson (DW) statistics are clearly a sign of a misleading regression in terms of inference and forecasts. Usually the DW statistic, measuring the serial correlation between the residuals of a regression, is between 0 and 4 with a value of 2.00 showing no serial correlation. The ADF statistic is equal to –4.85, which is low given the critical values –4.64 (1%), –4.10 (5%) and –3.81 (10%)

Cointegrating regression: $R^2 = 0.95$, DW = 0.08, ADF = –4.85			
Variable	Coeff.	St.Dev.	*t*-stat
Constant	–8.4264	0.2378	–35.4397
CPI	0.7897	0.0617	12.8098
FIBOR3M	–0.3422	0.0083	–41.4730
EUOOCIPDG	2.5658	0.0977	26.2700

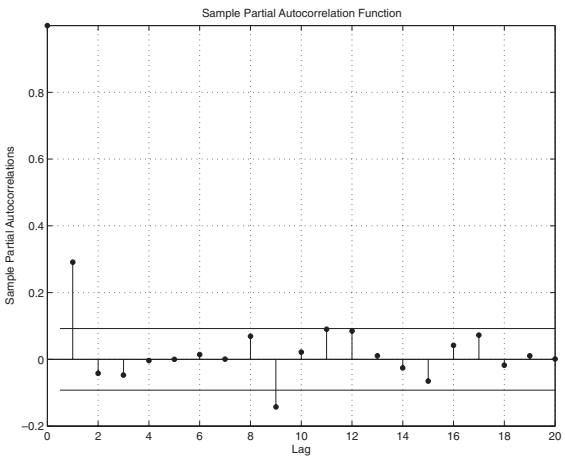

Figure 2. Partial autocorrelation function (pacf) for the output or dependent variable $\text{CHCLEM}^{(d)}$ in the ECM specification

The evidence of cointegration between the dependent original variable and a subset of the explanatory variables allow us to implement an ECM specification. The series $\text{CHMCLEM}^{(l)}$, $\text{FIBOR3M}^{(l)}$, $\text{CPI}^{(l)}$ and $\text{IP}^{(l)}$ are used as first differences. Also, from Figure 2 we see that the partial autocorrelation function of the dependent variable in the ECM ($\text{CHMCLEM}_t^{(d)}$) drops to zero after one lag. Therefore, we will include all variables lagged at $t-1$ and $t-2$ (one additional lag is used for conservativeness). The remaining explanatory variables (those not included in the cointegration) will also be included in the model up to two lags, in first differences, as exogeneous variables.

According to the ECM specification, we have the following model:

$$\begin{aligned}\text{CHMCLEM}_t^{(d)} = f(&z_{t-1}, \text{CHMCLEM}_{t-i}^{(d)}, \text{FIBOR3M}_{t-i}^{(d)}, \text{CPI}_{t-i}^{(d)}, \text{CPI}_{t-i}^{(d)},\\ &\text{IP}_{t-i}^{(d)}, \text{OILBREN}_{t-i}^{(d)}, \text{ETYEUSP}_{t-i}^{(d)}, \text{PHREUSN}_{t-i}^{(d)}, \text{GDP}_{t-i}^{(d)}, \text{USEURWD}_{t-i}^{(d)},\\ &\text{CHLORED}_{t-i}^{(d)}, \text{BMBD02Y}_{t-i}^{(d)}, \text{BMBD05Y}_{t-i}^{(d)}, \text{BMBD010Y}_{t-i}^{(d)}) + e_t\end{aligned} \tag{19}$$

DOI: 10.1002/for

Table III. Final estimates of the linear regression based on the ECM specification

Linear ECM regression: $R^2 = 0.10$, DW = 2.01			
Variable	Coeff.	St.Dev.	*t*-stat
z_{t-1}	–0.0397	0.0100	–3.9718
$\text{CHMCLEM}^{(d)}_{t-1}$	0.2213	0.0391	5.6589
$\text{FIBOR3M}^{(d)}_{t-1}$	0.1555	0.0513	3.0318
$\text{FIBOR3M}^{(d)}_{t-2}$	–0.2216	0.0514	–4.3144
$\text{ETYEUSP}^{(d)}_{t-2}$	–0.0659	0.0327	–2.0141

where $i = 1,2$ and $z_{t-1} = \text{CHMCLEM}^{(l)}_{t-1} - (\hat{a}_0 + \hat{a}_1\text{CPI}^{(l)}_{t-1} + \hat{a}_2\text{FIBOR3M}^{(l)}_{t-1} + \hat{a}_3\text{IP}^{(l)}_{t-1})$, where the coefficients are estimated from (18).

Linear ECM model

With this initial definition of input variables, the function f will be estimated by the linear OLS regression. The data set contains 787 observations. The model is first estimated using the initial 600 data points, in order to obtain the relevant variables. In the linear regression, the selection of relevant inputs is based on the asymptotic *t*-tests of individual significance. With this methodology, the relevant variables are found to be the following: z_{t-1}, $\text{CHMCLEM}^{(d)}_{t-1}$, $\text{FIBOR3M}^{(d)}_{t-1}$, $\text{FIBOR3M}^{(d)}_{t-2}$ and $\text{ETYEUSP}^{(d)}_{t-2}$. The detailed results for the linear regression are reported in Table III.

The linear forecasts yield the following performance indicators: $\text{MSE} = 6.63 \times 10^{-4}$, MAE = 0.021, PCSP = 53.2%, PTstat = 0.73, *p*-value = 0.462. In this case, the low PTstat for the 53% of predictional accuracy shows that it is not significantly different from a random case. The results for investment strategy 1 are Sharpe ratio = 0.596, return = 8.94 and risk = 15.00, while investment strategy 2 yields Sharpe ratio = 0.487, return = 7.19 and risk = 14.76, respectively. Compared to the buy&hold strategy (Sharpe ratio = 0.208, return = 3.98, risk = 19.14), we can see that the linear model defined in this section allows us to increase the possible profits compared to a simple buy&hold strategy.

Nonlinear ECM model

Nonlinear modelling was performed using the ECM specification with the same candidate input set as in (19). The model is designed on the same training data set and is evaluated on the remaining test set using the same moving window approach.

Backward input selection is applied, removing in each step one input until the model probability $p(\mathcal{H}|\mathcal{D})$ stops increasing. The evolution of the level 3 and level 2 cost function as a function of the number of input pruning steps is depicted in Figure 3, together with the evolution of d_{eff}, γ and of the directional accuracy measures PTstat and PCSP. From the initial 27 input variables, 21 inputs have been removed. The optimal input set for the nonlinear model is reported in Table IV, from which it is observed that the variables found to relevant for the nonlinear models are almost the same as for the linear models. The cointegrating vector z_t is kept as an important input, its removal would lead to a significantly worse predictive performance of the nonlinear model. The optimal regularization and kernel function parameters that are inferred from the training data $\mathcal{D}$ are: $\mu_{MP} = 258.29$, $\zeta_{MP} = 2.93 \times 10^3$, $\gamma_{MP} = 11.34$ and $\sigma_{MP} = 0.25$.

With the nonlinear forecasts, we obtain $\text{MSE} = 6.87 \times 10^{-4}$, MAE = 0.021, while the performance measures for directional accuracy are PCSP = 60.2%, PTstat = 2.66, *p*-value = 0.8%. The high PTstat

DOI: 10.1002/for

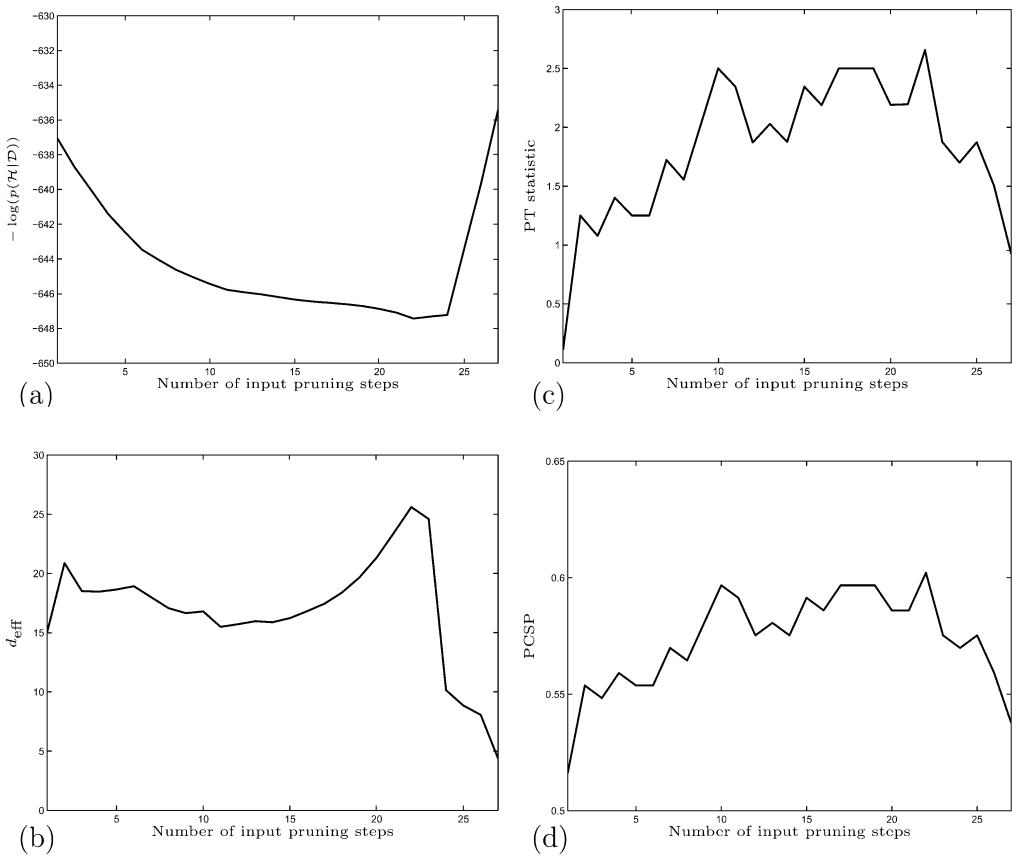

Figure 3. Backward input selection for nonlinear kernel-based regression. The evolution of the level 3 cost function (a) as a function of the number of input pruning steps yields an optimum at step 22. The corresponding values for the effective number of parameters d_{eff}, the out-of-sample Pesaran–Timmerman test statistic PTstat and the percentage of correct sign predictions (PCSP) are reported in panels (b)–(d). Notice that the PT statistic and PCSP become maximal at the minimum of the level 3 cost function

for the 60.2% of predictional accuracy shows that it is significantly different from random sign predictions. This is also observed in better performances with both trading strategies. With investment strategy 1 we obtain Sharpe ratio = 0.826, return = 12.75 and risk = 15.44, while investment strategy 2 yields Sharpe ratio = 0.841, return = 12.78 and risk = 15.21. The results for the buy&hold strategy, the linear and nonlinear ECM models are summarized in Table V, while the cumulative profits are depicted in Figure 4. Using almost the same input variables, the nonlinear model achieves clearly better out-of-sample performance.

CONCLUSIONS

In financial time series modelling and prediction, it is important to have reliable forecasting and modelling techniques. Based on stationarity and cointegration analysis, a linear ECM model is spec-

DOI: 10.1002/for

Table IV. Optimal input sets for the linear and nonlinear models

Linear		LS-SVM	
Variables	Lags	Variables	Lags
z	$t-1$	z	$t-1$
CHMCLEM	$t-1$	CHMCLEM	$t-1$
FIBOR3M	$t-1$, $t-2$	FIBOR3M	$t-1$, $t-2$
ETYEUSP	$t-2$	ETYEUSP	$t-2$
		BMBD010Y	$t-2$

Table V. Test set performances of the LS-SVM model obtained on the one-week-ahead prediction of the aggregated chemical index. The LS-SVM time series model with RBF-kernel is compared with linear ECM and a buy&hold strategy. The RBF-LS-SVM clearly achieves a better directional accuracy, better return (Re), risk (Ri) and resulting Sharpe ratio (SR) in combination with investment strategies IS1 and IS2

	Residuals		Directional accuracy			IS1			IS2		
	MSE	MAE	PCSP	PT	p-value	SR_1	Re_1	Ri_1	SR_2	Re_2	Ri_2
LS-SVM	6.87×10^{-4}	0.021	60.2	2.66	0.008	0.826	12.75	15.44	0.841	12.78	15.21
Linear	6.63×10^{-4}	0.021	53.2	0.73	0.462	0.596	8.94	15.00	0.487	7.19	14.76
B&H	—	—	—	—	—	0.208	3.98	19.14	0.208	3.98	19.14

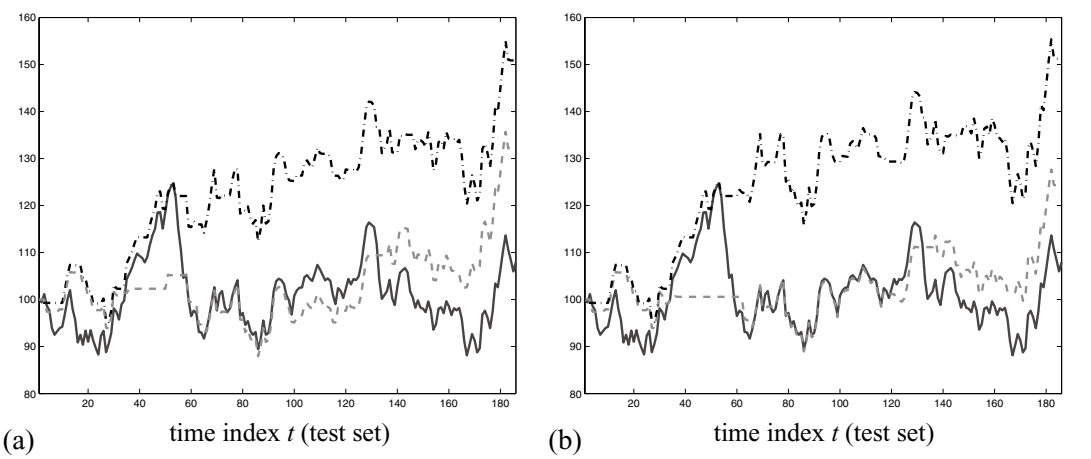

Figure 4. Cumulative returns using the sign predictions (transaction cost 0.1%) on the out-of-sample test set obtained with: (1) LS-SVM regressor with nonlinear RBF-kernel (dash-dotted line); (2) linear model (dashed line); and (3) buy&hold strategy (full line). The LS-SVM regressor yields the highest annualized return and corresponding Sharpe ratio as reported in Table V. Panels (a) and (b) depict the results of investment strategies 1 and 2, respectively

DOI: 10.1002/for

ified and estimated in order to produce out-of-sample linear stock market forecasts. The specified input variables of the linear ECM formulation are used in this paper as an initial candidate input set for nonlinear kernel-based regression. The nonlinear model was designed within the Bayesian evidence framework, getting an appropriate trade-off between model complexity and in-sample model accuracy. The regularization, kernel function parameters and relevant inputs are obtained by applying Bayes' formula on different levels of inference. For the prediction of an aggregated index for the European chemical sector, it was found that the optimal linear and nonlinear forecasts are based on almost the same set of relevant variables including the cointegrating vector. Comparing both techniques, the nonlinear model achieves significantly nonrandom out-of-sample sign predictions and also yields better a Sharpe ratio when implemented in a simple trading strategy.

APPENDIX A: BAYESIAN INFERENCE FOR LS-SVM REGRESSION

Inference of model parameters (level 1)

Bayes' formula

Applying Bayes' formula on level 1, one obtains the posterior probability of the model parameters $\boldsymbol{w}$ and b:

$$\begin{aligned} p(\boldsymbol{w}, b|\mathcal{D}, \log\mu, \log\zeta, \mathcal{H}) &= \frac{p(\mathcal{D}|\boldsymbol{w}, b, \log\mu, \log\zeta, \mathcal{H})\, p(\boldsymbol{w}, b|\log\mu, \log\zeta, \mathcal{H})}{p(\mathcal{D}|\log\mu, \log\zeta, \mathcal{H})} \\ &\propto p(\mathcal{D}|\boldsymbol{w}, b, \log\mu, \log\zeta, \mathcal{H})\, p(\boldsymbol{w}, b|\log\mu, \log\zeta, \mathcal{H}) \end{aligned} \tag{20}$$

where the last step is obtained since the evidence $p(\mathcal{D}|\log\mu, \log\zeta, \mathcal{H})$ is a normalizing constant that does not depend upon $\boldsymbol{w}$ and b.

For the *prior*, no correlation between $\boldsymbol{w}$ and b is assumed: $p(\boldsymbol{w}, b|\log\mu, \mathcal{H}) = p(\boldsymbol{w}|\log\mu, \mathcal{H})p(b|\mathcal{H}) \propto p(\boldsymbol{w}|\log\mu, \mathcal{H})$, with a multivariate Gaussian prior on $\boldsymbol{w}$ with zero mean and covariance matrix $\mu^{-1}\boldsymbol{I}_{n_\varphi}$ and an uninformative, flat prior on b:

$$\begin{aligned} p(\boldsymbol{w}|\log\mu, \mathcal{H}) &= \left(\frac{\mu}{2\pi}\right)^{\frac{n_f}{2}} \exp\left(-\frac{\mu}{2}\boldsymbol{w}^T\boldsymbol{w}\right) \\ p(b|\mathcal{H}) &= \text{constant} \end{aligned} \tag{21}$$

The uniform prior distribution on b can be approximated by a Gaussian distribution with standard deviation $\sigma_b \to \infty$. The negative logarithm of (21) corresponds to the regularization term $\mu J_w = \mu/2\boldsymbol{w}^T\boldsymbol{w}$. The prior states a belief that without any learning from data, the coefficients are zero with an uncertainty denoted by the variance $1/\mu$; *a priori* we do not expect a functional relation between the feature vector φ and the observation y. Before the data are available, the most likely model has zero weights $w_k = 0$ ($k = 1, \ldots, n_\varphi$), corresponding to the efficient market hypothesis (Bachelier, 1900; Campbell *et al*., 1997; Fama, 1965).

It is assumed that the errors $e_t = y_t - (\boldsymbol{w}^T\varphi(\boldsymbol{x}_t) + b)$ are independently identically normally distributed with zero mean and variance $1/\zeta$ for expressing the *likelihood*

DOI: 10.1002/for

$$p(\mathcal{D}|\boldsymbol{w}, b, \log\zeta, \mathcal{H}) \propto \prod_{t=1}^{n_\mathcal{D}} p(e_t|\boldsymbol{x}_t, \boldsymbol{w}, b, \log\zeta, \mathcal{H}) \tag{22}$$

with

$$p(e_t|\boldsymbol{w}, b, \log\zeta, \mathcal{H}) = \sqrt{\frac{\zeta}{2\pi}} \exp\left(-\frac{\zeta}{2}(y_t - \boldsymbol{w}^T\varphi(\boldsymbol{x}_t) - b)^2\right) \tag{23}$$

The negative logarithm of the likelihood (22) corresponds to the sum squared error term $\zeta J_e = \frac{\zeta}{2}\sum_{t=1}^{n_\mathcal{D}} e_t^2$.

Substituting (21) and (22) into (20), neglecting all constants and taking the negative logarithm, Bayes' rule at the first level of inference corresponds to the constrainted minimization problem (9) and (10) that can be solved for $\boldsymbol{w}$ and b in the primal space from (15) in the linear case when $n \leq n_\mathcal{D}$ or $\boldsymbol{\alpha}$ and b in the dual space from (16) in the nonlinear kernel-based regression case and in the linear case when $n \geq n_\mathcal{D}$. In the remainder of this paper, the maximum *a posteriori* parameter estimates are denoted by the subscript 'MP', e.g., $\boldsymbol{w}_{MP}$ and b_{MP}.

Given that the prior (21) and likelihood (22) are multivariate distributions, the *posterior* (20) is a multivariate normal distribution[4] in $[\boldsymbol{w}; b]$ with mean $[\boldsymbol{w}_{MP}; b_{MP}] \in \mathbb{R}^{n_\varphi+1}$ and covariance matrix $\boldsymbol{Q} \in \mathbb{R}^{(n_\varphi+1)\times(n_\varphi+1)}$. An alternative expression for the posterior is obtained by substituting (21) and (22) into (20). These approaches yield

$$p(\boldsymbol{w}, b|\mathcal{D}, \log\mu, \log\zeta, \mathcal{H}) = \sqrt{\frac{\det(\boldsymbol{Q}^{-1})}{(2\pi)^{n_\varphi+1}}} \exp\left(-\frac{1}{2}[\boldsymbol{w} - \boldsymbol{w}_{MP}; b - b_{MP}]\boldsymbol{Q}^{-1}[\boldsymbol{w} - \boldsymbol{w}_{MP}; b - b_{MP}]\right) \tag{24}$$

$$\propto \left(\frac{\mu}{2\pi}\right)^{\frac{n_f}{2}} \exp\left(-\frac{\mu}{2}\boldsymbol{w}^T\boldsymbol{w}\right)\left(\frac{\zeta}{2\pi}\right)^{\frac{n_\mathcal{D}}{2}} \exp\left(-\frac{\zeta}{2}\sum_{i=1}^{n_\mathcal{D}} e_i^2\right) \tag{25}$$

respectively.

The *evidence* is a normalizing constant in (20) independent of $\boldsymbol{w}$ and b such that $\int\int\ldots\int p(\boldsymbol{w}, b|\mathcal{D}, \log\mu, \log\zeta, \mathcal{H})dw_1 \ldots dw_{n_\varphi}db = 1$. Substituting the expressions for the prior (21), likelihood (22) and posterior (25) into (20), one obtains

$$p(\mathcal{D}|\log\mu, \log\zeta, \mathcal{H}) = \frac{p(\boldsymbol{w}_{MP}|\log\mu, \mathcal{H})p(\mathcal{D}|\boldsymbol{w}_{MP}, b_{MP}, \log\zeta, \mathcal{H})}{p(\boldsymbol{w}_{MP}, b_{MP}|\mathcal{D}, \log\mu, \log\zeta, \mathcal{H})} \tag{26}$$

Computation and interpretation

The model parameters with maximum posterior probability are obtained by minimizing the negative logarithm of (24) and (25):

[4] The notation $[x; y] = [x, y]^T$ is used here.

DOI: 10.1002/for

$$(\boldsymbol{w}_{MP}, b_{MP}) = \arg\min_{w,b} J_1(\boldsymbol{w}, b)$$

$$= J_1(\boldsymbol{w}_{MP}, b_{MP}) + \frac{1}{2}\left([\boldsymbol{w} - \boldsymbol{w}_{MP}; b - b_{MP}]^T \boldsymbol{Q}^{-1}[\boldsymbol{w} - \boldsymbol{w}_{MP}; b - b_{MP}]\right) \quad (27)$$

$$= \frac{\mu}{2}\boldsymbol{w}^T\boldsymbol{w} + \frac{\zeta}{2}\sum_{i=1}^{n_{\mathcal{D}}} e_i^2 \quad (28)$$

where constants are neglected in the optimization problem. Both expressions yield the same optimization problem and the covariance matrix $\boldsymbol{Q}$ is equal to the inverse of the Hessian $\boldsymbol{H}$ of J_1. The Hessian is expressed in terms of the matrix $\Phi = [\varphi(\boldsymbol{x}_1), \ldots \varphi(\boldsymbol{x}_{n_{\mathcal{D}}})]^T$ with regressors, as derived in Appendix B.

The optimal $\boldsymbol{w}_{MP}$ and b_{MP} are computed in the dual space from the linear KKT-system (16), while the prediction $\hat{y} = \boldsymbol{w}_{MP}^T\varphi(\boldsymbol{x}) + b_{MP}$ is expressed in terms of the dual parameters $\boldsymbol{\alpha}$ and bias term b_{MP} via (17).

Substituting (21), (22) and (25) into (26), one obtains

$$p(\mathcal{D}|\log\mu, \log\zeta, \mathcal{H}) \propto \left(\frac{\mu^{n_\varphi}\zeta^{n_{\mathcal{D}}}}{\det \boldsymbol{H}}\right)^{\frac{1}{2}} \exp(-J_1(\boldsymbol{w}_{MP}, b_{MP})) \quad (29)$$

As $J_1(\boldsymbol{w}, b) = \mu J_w(\boldsymbol{w}) + \zeta J_e(\boldsymbol{w}, b)$, the evidence can be rewritten as

$$\underbrace{p(\mathcal{D}|\log\mu, \log\zeta, \mathcal{H})}_{\text{evidence}} \propto \underbrace{p(\mathcal{D}|\boldsymbol{w}_{MP}, b_{MP}, \log\zeta, \mathcal{H})}_{\text{likelihood}|\boldsymbol{w}_{MP}, b_{MP}} \underbrace{p(\boldsymbol{w}_{MP}|\log\mu, \mathcal{H})(\det \boldsymbol{H})^{-1/2}}_{\text{Occam factor}}$$

The model evidence consists of the likelihood of the data and an Occam factor that penalizes for too complex models. The Occam factor consists of the regularization term $1/2\boldsymbol{w}_{MP}^T\boldsymbol{w}_{MP}$ and the ratio $(\mu^{n_\varphi}/\det \boldsymbol{H})^{1/2}$, which is a measure for the volume of the posterior probability divided by the volume of the prior probability. Strong contractions of the posterior versus prior space indicates too many free parameters and, hence, overfitting on the training data. The evidence will be maximized on level 2, where also dual space expressions are derived.

Inference of hyperparameters (level 2)

Bayes' formula

The optimal regularization parameters μ and ζ are inferred from the given data $\mathcal{D}$ by applying Bayes' formula on the second level (Van Gestel *et al*., 2001, 2002)

$$p(\log\mu, \log\zeta|\mathcal{D}, \mathcal{H}) = \frac{p(\mathcal{D}|\log\mu, \log\zeta, \mathcal{H})p(\log\mu, \log\zeta)}{p(\mathcal{D}|\mathcal{H})} \quad (30)$$

The *prior* $p(\log\mu, \log\zeta) = p(\log\mu|\mathcal{H})p(\log\zeta|\mathcal{H}) =$ constant is taken to be a flat uninformative prior ($\sigma_{\log\mu}, \sigma_{\log\zeta} \to \infty$). The level 2 *likelihood* $p(\mathcal{D}|\log\mu, \log\zeta, \mathcal{H})$ is equal to the level 1 evidence (29). In this way, Bayesian inference implicitly embodies Occam's razor: on level 2 the evidence of level 1 is optimized so as to find a trade-off between the model fit and a complexity term to avoid overfitting (MacKay, 1995). The level 2 *evidence* is obtained in a similar way as on level 1, as the like-

DOI: 10.1002/for

lihood for the maximum *a posteriori* times the ratio of the volume of the posterior probability and the volume of the prior probability:

$$p(\mathcal{D}|\mathcal{H}) \simeq p(\mathcal{D}|\log\mu_{MP}, \log\zeta_{MP}, \mathcal{H}) \frac{\sigma_{\log\mu|\mathcal{D}}\sigma_{\log\zeta|\mathcal{D}}}{\sigma_{\log\mu}\sigma_{\log\zeta}} \tag{31}$$

where one typically approximates the posterior probability by a multivariate normal probability function with diagonal covariance matrix $\mathrm{diag}([\sigma^2_{\log\mu|\mathcal{D}}, \sigma^2_{\log\mu|\mathcal{D}}]) \in \mathbb{R}^{2\times 2}$.

Neglecting all constants, Bayes' formula (30) becomes

$$p(\log\mu, \log\zeta|\mathcal{D}, \mathcal{H}) \propto p(\mathcal{D}|\log\mu, \log\zeta, \mathcal{H}) \tag{32}$$

where the expressions for the level 1 evidence are given by (26) and (29).

Computation and interpretation

In the primal space, the hyperparameters are obtained by minimizing the negative logarithm of (29) and (32)

$$\begin{aligned}(\mu_{MP}, \zeta_{MP}) = \arg\min_{\mu,\zeta} J_2(\mu, \zeta) = \mu J_w(\boldsymbol{w}_{MP}) + \zeta J_e(\boldsymbol{w}_{MP}, b_{MP}) \\ + \frac{1}{2}\log\det \boldsymbol{H} - \frac{n_\varphi}{2}\log\mu - \frac{n_{\mathcal{D}}}{2}\log\zeta\end{aligned} \tag{33}$$

Observe that in order to evaluate (33) one needs also to calculate $\boldsymbol{w}_{MP}$ and b_{MP} for the given μ and ζ and evaluate the level 1 cost function. The determinant of $\boldsymbol{H}$ is equal to (see Appendix B for details)

$$\det(\boldsymbol{H}) = (\zeta n_{\mathcal{D}})\det(\mu \boldsymbol{I}_{n_\varphi} + \zeta \boldsymbol{\Phi}^T \mathbf{N}_c \boldsymbol{\Phi})$$

with the idempotent centring matrix $\mathbf{N}_c = \boldsymbol{I}_{n_{\mathcal{D}}} - 1/n_{\mathcal{D}}\mathbf{1}\mathbf{1}^T = \mathbf{N}_c^2 \in \mathbb{R}^{n_{\mathcal{D}}\times n_{\mathcal{D}}}$. The determinant is also equal to the product of the eigenvalues. The n_e nonzero eigenvalues $\lambda_1, \ldots, \lambda_{n_e}$ of $\boldsymbol{\Phi}^T\mathbf{N}_c\boldsymbol{\Phi}$ are equal to the n_e nonzero eigenvalues of $\mathbf{N}_c\boldsymbol{\Phi}\boldsymbol{\Phi}^T\mathbf{N}_c = \mathbf{N}_c\boldsymbol{\Omega}\mathbf{N}_c \in \mathbb{R}^{n^{\mathcal{D}}\times n^{\mathcal{D}}}$, which can be calculated in the dual space. Substituting the determinant $\det(\boldsymbol{H}) = \zeta n_{\mathcal{D}}\mu^{n_\varphi - n_e}\prod_{i=1}^{n_e}(\mu+\zeta\lambda_i)$ into (33), one obtains the optimization problem in the dual space

$$J_2(\mu, \zeta) = \mu J_w(\boldsymbol{w}_{MP}) + \zeta J_e(\boldsymbol{w}_{MP}, b_{MP}) + \sum_{i=1}^{n_e}\frac{\log(\mu+\zeta\lambda_i)}{2} - \frac{n_e}{2}\log\mu - \frac{n_e - 1}{2}\log\zeta \tag{34}$$

where it can be shown by matrix algebra (see Appendix B) that $\mu J_w(\boldsymbol{w}_{MP}) + \zeta J_e(\boldsymbol{w}_{MP}, b_{MP}) = \frac{1}{2}\boldsymbol{y}^T\mathbf{N}_c(\frac{1}{\mu}\mathbf{N}_c\boldsymbol{\Omega}\mathbf{N}_c + \frac{1}{\zeta}\boldsymbol{I}_{n_{\mathcal{D}}})^{-1}\mathbf{N}_c\boldsymbol{y}$.

An important concept in neural networks and Bayesian learning in general is the *effective number of parameters*. Although there are $n_\varphi + 1$ free parameters $w_1, \ldots, w_{n_\varphi}, b$ in the primal space, the use of these parameters (28) is restricted by the use of the regularization term $1/2\boldsymbol{w}^T\boldsymbol{w}$. The effective number of parameters d_{eff} is equal to $d_{\mathrm{eff}} = \Sigma_i \lambda_{i,u}/\lambda_{i,r}$, where $\lambda_{i,u}$, $\lambda_{i,r}$ denote the eigenvalues of the Hessian of the unregularized cost function $J_{1,u} = \zeta E_D$ and the regularized cost function $J_{1,r} = \mu E_W$

DOI: 10.1002/for

$+ \zeta E_D$ (Bishop, 1995; MacKay, 1995). For LS-SVMs, the effective number of parameters is equal to

$$d_{\text{eff}} = 1 + \sum_{i=1}^{n_e} \frac{\zeta\lambda_i}{\mu + \zeta\lambda_i} = 1 + \sum_{i=1}^{n_e} \frac{\gamma\lambda_i}{1 + \gamma\lambda_i} \tag{35}$$

with $\gamma = \zeta/\mu \in \mathbb{R}^+$. The term +1 appears because no regularization is applied on the bias term b. As shown, one has that $n_e \leq n_{\mathcal{D}} - 1$ and, hence, also that $d_{\text{eff}} \leq n_{\mathcal{D}}$, even in the case of high-dimensional feature spaces.

The conditions for optimality for (34) are obtained by putting $\partial J_2/\partial_\mu = \partial J_2/\partial\zeta = 0$. One obtains[5]

$$\partial J_2/\partial\mu = 0 \rightarrow 2\mu_{MP} J_w(\boldsymbol{w}_{MP}; \mu_{MP}, \zeta_{MP}) = d_{\text{eff}}(\mu_{MP}, \zeta_{MP}) - 1 \tag{36}$$

$$\partial J_2/\partial\zeta = 0 \rightarrow 2\zeta_{MP} J_e(\boldsymbol{w}_{MP}; b_{MP}; \mu_{MP}, \zeta_{MP}) = n_{\mathcal{D}} - d_{\text{eff}} \tag{37}$$

where the latter equation corresponds to the unbiased estimate of the noise variance $1/\zeta_{MP} = \frac{1}{2}\sum_{i=1}^{n_{\mathcal{D}}} e_i^2/(n_{\mathcal{D}} - d_{\text{eff}})$.

Instead of solving the optimization problem in μ and ζ, one may also reformulate (34) using (36) and (37) in terms of $\gamma = \zeta/\mu$ and solve the following scalar optimization problem (Van Gestel *et al.*, 2002):

$$\min_{\gamma} \sum_{i=1}^{n_{\mathcal{D}}-1} \log\left(\lambda_i + \frac{1}{\gamma}\right) + (n_{\mathcal{D}} - 1)\log(J_{\mathbf{w}}(\mathbf{w}_{MP}) + \gamma J_{\mathbf{e}}(\mathbf{w}_{MP}, b_{MP})) \tag{38}$$

with

$$J_e(\boldsymbol{w}_{MP}, b_{MP}) = \frac{1}{2\gamma^2} \boldsymbol{y}^T \boldsymbol{N}_c \boldsymbol{V} (\Lambda + \boldsymbol{I}_N/\gamma)^{-2} \boldsymbol{V}^T \boldsymbol{N}_c \boldsymbol{y} \tag{39}$$

$$J_w(\boldsymbol{w}_{MP}) = \frac{1}{2} \boldsymbol{y}^T \boldsymbol{N}_c \boldsymbol{V} \Lambda (\Lambda + \boldsymbol{I}/\gamma)^{-2} \boldsymbol{V}^T \boldsymbol{N}_c \boldsymbol{y} \tag{40}$$

$$J_w(\boldsymbol{w}_{MP}) + \gamma J_e(\boldsymbol{w}_{MP}, b_{MP}) = \frac{1}{2} \boldsymbol{y}^T \boldsymbol{N}_c \boldsymbol{V} (\Lambda + \boldsymbol{I}_N/\gamma)^{-1} \boldsymbol{V}^T \boldsymbol{N}_c \boldsymbol{y} \tag{41}$$

and with the eigenvalue decomposition $\mathbf{N}_c\boldsymbol{\Omega}\mathbf{N}_c = \boldsymbol{V}^T\boldsymbol{\Lambda}\boldsymbol{V}$. Given the optimal γ_{MP} from (38) one finds the effective number of parameters d_{eff} from $d_{\text{eff}} = 1 + \Sigma_{i=1}^{n_e} \gamma\lambda_i/(1 + \gamma\lambda_i)$. The optimal μ_{MP} and ζ_{MP} are obtained from $\mu_{MP} = (d_{\text{eff}} - 1)/(2J_w(\boldsymbol{w}_{MP}))$ and $\zeta_{MP} = (n_{\mathcal{D}} - d_{\text{eff}})/(2J_e(\boldsymbol{w}_{MP}, b_{MP}))$.

[5] In this derivation, one uses that (MacKay, 1995; Suykens *et al.*, 2002; Van Gestel *et al.*, 2002) $\partial(J_1(\boldsymbol{w}_{MP}, b_{MP}))/\partial_\mu = \delta(J_1(\boldsymbol{w}_{MP}, b_{MP}))/\delta_\mu + \delta(J_1(\boldsymbol{w}_{MP}, b_{MP}))/\delta[\boldsymbol{w}; b]|_{[\boldsymbol{w}_{MP}; b_{MP}]} \times \delta([\boldsymbol{w}_{MP}; b_{MP}])/\delta_\mu = J_w(\boldsymbol{w}_{MP})$, since $\delta J_1(\boldsymbol{w}_{MP}, b_{MP}))/\delta[\boldsymbol{w}; b]|_{[\boldsymbol{w}_{MP}; b_{MP}]} = 0$.

DOI: 10.1002/for

Model comparison (level 3)

Bayes' formula

The model structure $\mathcal{H}$ determines the remaining parameters of the kernel-based model: the selected kernel function (linear, RBF, . . .), the kernel parameter (RBF kernel parameter σ) and selected explanatory inputs. The model structure is inferred on level 3.

Consider, for example, the inference of the RBF-kernel parameter σ, where the model structure is denoted by $\mathcal{H}_\sigma$. Bayes' formula for the inference of $\mathcal{H}_\sigma$ is equal to

$$p(\mathcal{H}_\sigma|\mathcal{D}) \propto p(\mathcal{D}|\mathcal{H}_\sigma)p(\mathcal{H}_\sigma) \tag{42}$$

where no evidence $p(\mathcal{D})$ is used in the expression on level 3 as it is in practice impossible to integrate over all model structures. The prior probability $p(\mathcal{H}_\sigma)$ is assumed to be constant. The likelihood is equal to the level 2 evidence (31).

Computation and interpretation

Substituting the evidence (31) into (42) and taking in the constant prior, the Bayes' rule (31) becomes

$$p(\mathcal{H}|\mathcal{D}) \simeq p(\mathcal{D}|\log\mu_{MP}, \log\zeta_{MP}, \mathcal{H}) \frac{\sigma_{\log\mu|\mathcal{D}}\sigma_{\log\zeta|\mathcal{D}}}{\sigma_{\log\mu}\sigma_{\log\zeta}} \tag{43}$$

As uninformative priors are used on level 2, the standard deviations $\sigma_{\log\mu}$ and $\sigma_{\log\zeta}$ of the prior distribution both tend to infinity and are omitted in the comparisons of different models in (43). The posterior error bars can be approximated analytically as $\sigma^2_{\log\mu|\mathcal{D}} \simeq 2/(d_{\text{eff}} - 1)$ and $\sigma^2_{\log\zeta|\mathcal{D}} \simeq 2/(n_{\mathcal{D}} - d_{\text{eff}})$, respectively (MacKay, 1995). The level 3 posterior becomes

$$\begin{aligned} p(\mathcal{H}_\sigma|\mathcal{D}) &\simeq p(\mathcal{D}|\log\mu_{MP}, \log\zeta_{MP}, \mathcal{H}_\sigma) \frac{\sigma_{\log\mu|\mathcal{D}}\sigma_{\log\zeta|\mathcal{D}}}{\sigma_{\log\mu}\sigma_{\log\zeta}} \\ &\propto \sqrt{\frac{\mu_{MP}^{n_e}\zeta_{MP}^{n_{\mathcal{D}}-1}}{(d_{\text{eff}}-1)(N-d_{\text{eff}})\prod_{i=1}^{n_e}(\mu_{MP}+\zeta_{MP}\lambda_i)}} \end{aligned} \tag{44}$$

where all expressions can be calculated in the dual space. A practical way to infer the kernel parameter σ is to calculate (44) for a grid of possible kernel parameters $\sigma_1, \ldots, \sigma_m$ and to compare the corresponding posterior model parameters $p(\mathcal{H}_{\sigma_1}|\mathcal{D}), \ldots, p(\mathcal{H}_{\sigma_m}|\mathcal{D})$.

Model comparison is also used to infer the set of most relevant inputs (Van Gestel *et al*., 2001) out of the given set of candidate explanatory variables by making pairwise comparisons of models with different input sets. In a backward input selection procedure, one starts from the full candidate input set and removes in each input pruning step that input that yields the best model improvement (or smallest decrease) in terms of the model probability (44). The procedure is stopped when no significant decrease of the model probability is observed. In the case of equal prior model probabilities $p(\mathcal{H}_i) = p(\mathcal{H}_j)(\forall i, j)$ the models $\mathcal{H}_i$ and $\mathcal{H}_j$ are compared according to their Bayes factor

$$\mathcal{B}_{ij} = \frac{p(\mathcal{D}|\mathcal{H}_i)}{p(\mathcal{D}|\mathcal{H}_j)} = \frac{p(\mathcal{D}|\log\mu_i, \log\zeta_i, \mathcal{H}_i)}{p(\mathcal{D}|\log\mu_j, \log\zeta_j, \mathcal{H}_j)} \frac{\sigma_{\log\mu_i|\mathcal{D}}\sigma_{\log\zeta_i|\mathcal{D}}}{\sigma_{\log\mu_j|\mathcal{D}}\sigma_{\log\zeta_j|\mathcal{D}}} \tag{45}$$

DOI: 10.1002/for

According to Jeffreys (1961), a value $2\ln \mathcal{B}_{ij}$ corresponding to 0–2, 2–5, 5–10 and >10 indicates a very weak, positive, strong and decisive evidence against the null hypothesis of no difference in model performance between the models $\mathcal{H}_i$ and $\mathcal{H}_j$.

Moderated output

The uncertainty on the estimated model parameters results in an additional uncertainty for the one-step-ahead prediction $y_{MP} = \boldsymbol{w}_{MP}^T\varphi(\boldsymbol{x}) + b_{MP} = \sum_{t=1}^{n_{\mathcal{D}}}\frac{\alpha_t}{\mu}K(\boldsymbol{x}, \boldsymbol{x}_t) + b_{MP}$. Given the normal distribution (27) of the model parameters $[\boldsymbol{w}; b]$ with mean $[\boldsymbol{w}_{MP}; b_{MP}]$ and covariance matrix $\boldsymbol{Q}$, and the additive noise with mean zero and noise variance ζ, it is well known that the predicted output is normally distributed with mean

$$\hat{y}_{MP} = \boldsymbol{w}_{MP}^T\varphi(\boldsymbol{x}) + b_{MP} \tag{46}$$

and variance

$$\sigma_y^2 = \frac{1}{\zeta} + [\varphi(\boldsymbol{x}); 1]^T \boldsymbol{Q}[\varphi(\boldsymbol{x}); 1] \tag{47}$$

where the first term is due to the additive noise e_t and the second term is due to the posterior uncertainty (27) on the model parameters $\boldsymbol{w}$ and b.

The dual space expression for y_{MP} is given in (17). The expression for the variance σ_y^2 involves the inversion of the Hessian $\boldsymbol{H} = \boldsymbol{Q}^{-1}$ in the feature space. Given the expressions (51) and (53), the following practical expression is obtained in the dual space by applying linear matrix algebra:

$$\begin{aligned}\sigma_y^2 = \frac{1}{\zeta} + \frac{1}{n_{\mathcal{D}}\zeta} + \frac{1}{\mu}\left(K(\mathbf{x},\mathbf{x}) + \frac{1}{n_{\mathcal{D}}^2}1^T\boldsymbol{\Omega}1 - \frac{2}{n_{\mathcal{D}}}\mathbf{k}(\mathbf{x})^T1\right) \\ - \frac{\zeta}{\mu}\left(\mathbf{k}(\mathbf{x}) - \frac{1}{n_{\mathcal{D}}}\boldsymbol{\Omega}1\right)^T \mathbf{N}_c(\mu\mathbf{I}_{n_{\mathcal{D}}} + \zeta\boldsymbol{\Omega}_c)^{-1}\mathbf{N}_c\left(\mathbf{k}(\mathbf{x}) - \frac{1}{n_{\mathcal{D}}}\boldsymbol{\Omega}1\right)\end{aligned} \tag{48}$$

with the vector $\boldsymbol{k}(\boldsymbol{x}) = \Phi\varphi(\boldsymbol{x}) = [K(\boldsymbol{x}, \boldsymbol{x}_1), \ldots, K(\boldsymbol{x}, \boldsymbol{x}_{n_{\mathcal{D}}})]^T \in \mathbb{R}^{n_{\mathcal{D}}}$. This dual space expression allows us to compute the variance on the point prediction y_{MP} when the nonlinear mapping φ is implicitly defined by the (nonlinear) kernel function K or when $n \geq n_{\mathcal{D}}$ in the linear case.

APPENDIX B: MATHEMATICS

Expression for the Hessian and covariance matrix

The level 1 posterior probability $p([\boldsymbol{w}; b]|\mathcal{D}, \mu, \zeta, \mathcal{H})$ is a multivariate normal distribution in $\mathbb{R}^{n_\varphi}$ with mean $[\boldsymbol{w}_{MP}; b_{MP}]$ and covariance matrix $\boldsymbol{Q} = \boldsymbol{H}^{-1}$, where $\boldsymbol{H}$ is the Hessian of the least squares cost function (9). Defining the matrix of regressors $\boldsymbol{\Phi}^T = [\varphi(\boldsymbol{x}_1), \ldots, \varphi(\boldsymbol{x}_{n_\varphi})]$, the identity matrix $\boldsymbol{I}$ and the vector with all ones $\mathbf{1}$ of appropriate dimension; the Hessian is equal to

DOI: 10.1002/for

$$\boldsymbol{H} = \begin{bmatrix} \boldsymbol{H}_{11} & \boldsymbol{h}_{12} \\ \boldsymbol{h}_{21} & h_{22} \end{bmatrix} = \begin{bmatrix} \mu \boldsymbol{I}_{n_\varphi} + \zeta \boldsymbol{\Phi}^T \boldsymbol{\Phi} & \zeta \boldsymbol{\Phi}^T \mathbf{1} \\ \zeta \mathbf{1}^T \boldsymbol{\Phi} & \zeta n_{\mathcal{D}} \end{bmatrix} \tag{49}$$

with corresponding block matrices $\boldsymbol{H}_{11} = \mu \boldsymbol{I}_{n_\varphi} + \zeta \boldsymbol{\Phi}^T \boldsymbol{\Phi}$, $\boldsymbol{h}_{12} = \boldsymbol{h}_{21}^T = \boldsymbol{\Phi}^T \mathbf{1}$ and $h_{22} = n_{\mathcal{D}}$. The inverse Hessian $\boldsymbol{H}^{-1}$ is then obtained via a Schur complement type argument:

$$\begin{aligned} \boldsymbol{H}^{-1} &= \left(\begin{bmatrix} \boldsymbol{I}_{n_\varphi} & \boldsymbol{X} \\ 0^T & 1 \end{bmatrix} \begin{bmatrix} \boldsymbol{I}_{n_\varphi} & -\boldsymbol{X} \\ 0^T & 1 \end{bmatrix} \begin{bmatrix} \boldsymbol{H}_{11} & \boldsymbol{h}_{12} \\ \boldsymbol{h}_{12}^T & h_{22} \end{bmatrix} \begin{bmatrix} \boldsymbol{I}_{n_\varphi} & 0 \\ -\boldsymbol{X}^T & 1 \end{bmatrix} \begin{bmatrix} \boldsymbol{I}_{n_\varphi} & 0 \\ -\boldsymbol{X}^T & 1 \end{bmatrix} \right)^{-1} \\ &= \left(\begin{bmatrix} \boldsymbol{I}_{n_\varphi} & \boldsymbol{X} \\ 0^T & 1 \end{bmatrix} \begin{bmatrix} \boldsymbol{H}_{11} - \boldsymbol{h}_{12} h_{22}^{-1} \boldsymbol{h}_{12}^T & 0 \\ 0^T & h_{22} \end{bmatrix} \begin{bmatrix} \boldsymbol{I}_{n_\varphi} & 0 \\ \boldsymbol{X}^T & 1 \end{bmatrix} \right)^{-1} \qquad (50) \\ &= \begin{bmatrix} (\boldsymbol{H}_{11} - \boldsymbol{h}_{12} h_{22}^{-1} \boldsymbol{h}_{12}^T)^{-1} & -\boldsymbol{F}_{11}^{-1} \boldsymbol{h}_{12} h_{22}^{-1} \\ -h_{22}^{-1} \boldsymbol{h}_{12}^T \boldsymbol{F}_{11}^{-1} & h_{22}^{-1} + h_{22}^{-1} \boldsymbol{h}_{12}^T \boldsymbol{F}_{11}^{-1} \boldsymbol{h}_{12} h_{22}^{-1} \end{bmatrix} \qquad (51) \end{aligned}$$

with $\boldsymbol{X} = \boldsymbol{h}_{12} h_{22}^{-1}$ and $\boldsymbol{F}_{11} = \boldsymbol{H}_{11} - \boldsymbol{h}_{12} h_{22}^{-1} \boldsymbol{h}_{12}^T$. In matrix expressions, it is useful to express $\boldsymbol{\Phi}^T \boldsymbol{\Phi} - \frac{1}{n_{\mathcal{D}}} \boldsymbol{\Phi}^T \mathbf{1} \mathbf{1}^T \boldsymbol{\Phi}$ as $\boldsymbol{\Phi}^T \mathbf{N}_c \boldsymbol{\Phi}$ with the idempotent centring matrix $\mathbf{N}_c = \boldsymbol{I}_{n_{\mathcal{D}}} - \frac{1}{n_{\mathcal{D}}} \mathbf{1}\mathbf{1}^T \in \mathbb{R}^{N \times N}$ having $\mathbf{N}_c = \mathbf{N}_c^2$. Given that $\boldsymbol{F}_{11}^{-1} = (\mu \boldsymbol{I}_{n_\varphi} + \zeta \boldsymbol{\Phi}^T \mathbf{N}_c \boldsymbol{\Phi})^{-1}$, the inverse Hessian $\boldsymbol{H}^{-1} = \boldsymbol{Q}$ is equal to

$$\boldsymbol{Q} = \begin{bmatrix} (\mu \boldsymbol{I}_{n_\varphi} + \zeta \boldsymbol{\Phi}^T \mathbf{N}_c \boldsymbol{\Phi})^{-1} & -\frac{1}{n_\Delta} (\mu \boldsymbol{I}_{n_\varphi} + \zeta \boldsymbol{\Phi}^T \mathbf{N}_c \boldsymbol{\Phi})^{-1} \boldsymbol{\Phi}^T \mathbf{1} \\ -\frac{1}{n_{\mathcal{D}}} \mathbf{1}^T \boldsymbol{\Phi} (\mu \boldsymbol{I}_{n_\varphi} + \zeta \boldsymbol{\Phi}^T \mathbf{N}_c \boldsymbol{\Phi})^{-1} & \frac{1}{\zeta n_{\mathcal{D}}} + \frac{1}{n_{\mathcal{D}^2}} \mathbf{1}^T \boldsymbol{\Phi} (\mu \boldsymbol{I}_{n_\varphi} + \zeta \boldsymbol{\Phi}^T \mathbf{N}_c \boldsymbol{\Phi})^{-1} \boldsymbol{\Phi}^T \mathbf{1} \end{bmatrix}$$

Expression for the determinant

The determinant of $\boldsymbol{H}$ is obtained from (50) using the fact that the determinant of a product is equal to the product of the determinants and is thus equal to

$$\begin{aligned} \det(\boldsymbol{H}) &= \det(\boldsymbol{H}_{11} - \boldsymbol{h}_{12}^T h_{22}^{-1} \boldsymbol{h}_{12}) \times \det(h_{22}) \\ &= \det(\mu \boldsymbol{I}_{n_\varphi} + \zeta \boldsymbol{\Phi}^T \mathbf{N}_c \boldsymbol{\Phi}) \times (\zeta n \mathcal{D}) \end{aligned} \tag{52}$$

which is obtained as the product of $\zeta n \mathcal{D}$ and the eigenvalues λ_i ($i = 1, \ldots, n_\varphi$) of $\mu \boldsymbol{I}_{n_\varphi} + \zeta \boldsymbol{\Phi}^T \mathbf{N}_c \boldsymbol{\Phi}$, denoted as $\lambda_i(\mu \boldsymbol{I}_{n_\varphi} + \zeta \boldsymbol{\Phi}^T \mathbf{N}_c \boldsymbol{\Phi})$. Because the matrix $\boldsymbol{\Phi}^T \mathbf{N}_c \boldsymbol{\Phi} \in \mathbb{R}^{n_\varphi \times n_\varphi}$ is rank deficient with rank $n_e \leq n_{\mathcal{D}} - 1$, $n_\varphi - n_e$ eigenvalues are equal to μ.

The dual space expressions can be obtained in terms of the singular value decomposition

$$\boldsymbol{\Phi}^T \mathbf{N}_c = \boldsymbol{U} \boldsymbol{S} \boldsymbol{V}^T = [\boldsymbol{U}_1 \quad \boldsymbol{U}_2] \begin{bmatrix} \boldsymbol{S}_1 & 0 \\ 0 & 0 \end{bmatrix} [\boldsymbol{V}_1 \quad \boldsymbol{V}_2] \tag{53}$$

with $\boldsymbol{U} \in \mathbb{R}^{n_\varphi \times n_\varphi}$, $\boldsymbol{S} \in \mathbb{R}^{n_\varphi \times n_{\mathcal{D}}}$, $\boldsymbol{V} \in \mathbb{R}^{n_{\mathcal{D}} \times n_{\mathcal{D}}}$ and with the block matrices $\boldsymbol{U}_1 \in \mathbb{R}^{n_\varphi \times n_e}$, $\boldsymbol{U}_2 \in \mathbb{R}^{n_\varphi \times (n_\varphi - n_e)}$, $\boldsymbol{S}_1 = \text{diag}([s_1, s_2, \ldots, s_{n_e}]) \in \mathbb{R}^{n_e \times n_e}$, $\boldsymbol{V}_1 \in \mathbb{R}^{n_{\mathcal{D}} \times n_e}$ and $\boldsymbol{V}_2 \in \mathbb{R}^{n_{\mathcal{D}} \times (n_{\mathcal{D}} - n_e)}$, with $0 \leq n_e \leq n_{\mathcal{D}} - 1$. Due to the

DOI: 10.1002/for

orthonormality property we have $\boldsymbol{U}\boldsymbol{U}^T = \boldsymbol{U}_1\boldsymbol{U}_1^T + \boldsymbol{U}_2\boldsymbol{U}_2^T = \boldsymbol{I}_{n_\varphi}$ and $\boldsymbol{V}\boldsymbol{V}^T = \boldsymbol{V}_1\boldsymbol{V}_1^T + \boldsymbol{V}_2\boldsymbol{V}_2^T = \boldsymbol{I}_{n_\mathcal{D}}$. Hence, one obtains the primal and dual eigenvalue decompositions

$$\boldsymbol{\Phi}^T\mathbf{N}_c\boldsymbol{\Phi} = \boldsymbol{U}_1\boldsymbol{S}_1^2\boldsymbol{U}_1^T \tag{54}$$

$$\mathbf{N}_c\boldsymbol{\Phi}\boldsymbol{\Phi}^T\mathbf{N}_c = \mathbf{N}_c\boldsymbol{\Omega}\mathbf{N}_c = \boldsymbol{V}_1\boldsymbol{S}_1^2\boldsymbol{V}_1^T \tag{55}$$

The n_φ eigenvalues of $\mu\boldsymbol{I}_{n_\varphi} + \zeta\boldsymbol{\Phi}^T\mathbf{N}_c\boldsymbol{\Phi}$ are equal to $\lambda_1 = \mu + \zeta s_1^2, \ldots, \lambda_{n_e} = \mu + \zeta s_{n_e}^2$, $\lambda_{n_e+1} = \mu, \ldots, \lambda_{n_\varphi} = \mu$, where the nonzero eigenvalues s_i^2 ($i = 1, \ldots, n_e$) are obtained from the eigenvalue decomposition of $\mathbf{N}_c\boldsymbol{\Phi}\boldsymbol{\Phi}^T\mathbf{N}_c$ from (55). The expression for the determinant is equal to $N\zeta\mu^{n_\mathcal{D}-n_e}\prod_{i=1}^{n_e}(\mu+\zeta\lambda_i(\mathbf{N}_c\boldsymbol{\Omega}\mathbf{N}_c)$, with $\mathbf{N}_c\boldsymbol{\Omega}\mathbf{N}_c = \boldsymbol{V}_1\mathrm{diag}([\lambda_1, \ldots, \lambda_{n_e}])\boldsymbol{V}_1^T$ and $\lambda_i = s_i^2$, $i = 1, \ldots, n_e$.

Expression for the level 1 cost function

The dual space expression for $J_1(\boldsymbol{w}_{MP}, b_{MP})$ is obtained by substituting $[\boldsymbol{w}_{MP}; b_{MP}] = \boldsymbol{H}^{-1}[\boldsymbol{\Phi}^T\boldsymbol{y}; \mathbf{1}^T\boldsymbol{y}]$ in (9). Applying a similar reasoning and algebra as for the calculation of the determinant, one obtains the dual space expression:

$$J_1(\boldsymbol{w}, b) = \mu J_w(\boldsymbol{w}_{mp}) + \zeta J_e(\boldsymbol{w}_{MP}, b_{MP}) = \frac{1}{2}\boldsymbol{y}^T\boldsymbol{N}_c(\mu^{-1}\boldsymbol{N}_c\boldsymbol{\Omega}\boldsymbol{N}_c + \zeta^{-1}\boldsymbol{I}_{n_\mathcal{D}})^{-1}\boldsymbol{N}_c\boldsymbol{y} \tag{56}$$

Given that $\mathbf{N}_c\boldsymbol{\Omega}\mathbf{N}_c = \boldsymbol{V}\boldsymbol{\Lambda}\boldsymbol{V}^T$, with $\boldsymbol{\Lambda} = \mathrm{diag}([s_1^2, \ldots, s_{n_e}^2, 0, \ldots, 0])$, one obtains (41). In a similar way, one obtains (39) and (40).

ACKNOWLEDGEMENTS

The authors would like to thank Peter Van Dijcke (Dexia Bank), Joao Garcia, Luc Leonard, Eric Hermann (Dexia Group) and Dirk Baestaens (Fortis Bank) for many helpful comments. This work was partially supported by grants and projects from the K.U.Leuven (GOA-Mefisto 666, GOA-Ambiorics), the Flemish Government (FWO Projects G.0407.02, G.0499.04, G.0211.05., ICCoS, ANMMM, IWT, GBOU), the Belgian Federal Government (IUAP V-22, PODO-II) and the EU (Ernsi, Eureka 2063, 2419). Scientific responsibility is assumed by its authors.

REFERENCES

Akaike H. 1974. A new look at statistical model identification. *IEEE Transactions on Automatic Control* **19**: 716–723.

Bachelier L. 1900. *Théorie de la spéculation*. Gauthier-Villars: Paris.

Bishop CM. 1995. *Neural Networks for Pattern Recognition*. Oxford University Press: Oxford.

Box GEP, Jenkins GM. 1970. *Time Series Analysis, Forecasting and Control*. Holden-Day: San Francisco.

Brock W, Lakonishok J, Le Baron B. 1992. Simple technical trading rules and the stochastic properties of econometrics. *Journal of Finance* **47**: 1731–1764.

Campbell JY, Lo AW, MacKinlay AC. 1997. *The Econometrics of Financial Markets*. Princeton University Press: Princeton, NJ.

Dickey DA, Fuller WA. 1979. Distribution of the estimators for autoregressive time series with a unit root. *Journal of the American Statistical Association* **74**: 427–431.

DOI: 10.1002/for

Engle RF, Granger CWJ. 1987. Cointegration and error correction: representations, estimation and testing. *Econometrica* **55**: 252–276.
Fama EF. 1965. The behaviour of stock market prices. *Journal of Business* **38**: 34–105.
Granger CWJ, Newbold P. 1986. *Forecasting Economic Time Series*. Academic Press: New York.
Granger CWJ, Newbold P. 1974. Spurious regression in econometrics. *Journal of Econometrics* **2**: 111–120.
Granger CWJ, Terasvirta T. 1993. *Modelling Nonlinear Economic Relationships*. Oxford University Press: Oxford.
Hamilton J. 1994. *Time Series Analysis*. Princeton University Press: Princeton, NJ.
Hutchinson JM, Lo AW, Poggio T. 1994. A nonparametric approach to pricing and hedging derivative securities via learning networks. *Journal of Finance* **49**: 851–889.
Jeffreys H. 1961. *Theory of Probability*. Oxford University Press: Oxford.
Johansen S. 1988. Statistical analysis of cointegration vectors. *Journal of Economics Dynamics and Control* **12**: 231–254.
Lo A, Mamaysky H, Wang J. 2000. Foundations of technical analysis: computational algorithms, statistical inference, and empirical implementation. *Journal of Finance* **55**: 1705–1765.
MacKay DJC. 1995. Probable networks and plausible predictions—a review of practical Bayesian methods for supervised neural networks. *Network: Computation in Neural Systems* **6**: 469–505.
Maddala GS, Kim IM. 1998. *Cointegration, Unit Roots and Structural Change*. Cambridge University Press: Cambridge.
Pesaran MH, Timmerman A. 1992. A simple nonparametric test of predictive performance. *Journal of Business and Economic Statistics* **10**: 461–465.
Rao B. 1994. *Cointegration for the Applied Economist*. MacMillan: London.
Refenes AP, Zapranis AD. 1999. Neural model identification, variable selection and model adequacy. *Journal of Forecasting* **18**: 299–332.
Schölkopf B, Smola A. 2002. *Learning with Kernels*. MIT Press: Cambridge, MA.
Schwarz G. 1978. Estimating the dimension of a model. *Annals of Statistics* **6**: 461–464.
Sullivan R, Timmerman A, White H. 1999. Data-snooping, technical trading rule performance, and the bootstrap. *Journal of Finance* **54**: 1647–1691.
Suykens JAK, Van Gestel T, De Brabanter J, De Moor B, Vandewalle J. 2002. *Least Squares Support Vector Machines*. World Scientific: Singapore.
Van Gestel T, Suykens JAK, Baestaens DE, Lambrechts A, Lanckriet G, Vandaele B, De Moor B, Vandewalle J. 2001. Predicting financial time series using least squares support vector machines within the evidence framework. *IEEE Transactions on Neural Networks (Special Issue on Financial Engineering)* **12**: 809–821.
Van Gestel T, Suykens JAK, Lanckriet G, Lambrechts A, De Moor B, Vandewalle J. 2002. A Bayesian framework for least squares support vector machine classifiers, Gaussian processes and kernel Fisher discriminant analysis. *Neural Computation* **14**: 1115–1147.
Vapnik V. 1998. *Statistical Learning Theory*. John Wiley & Sons: New York.

Authors' biographies:

Tony Van Gestel obtained an electromechanical engineering degree and a PhD in applied sciences (subject: mathematical modelling for financial engineering) in 1997 and 2002 at the Katholieke Universiteit Leuven. He currently works as a senior quantitative analyst at Dexia Group and is a free postdoctoral researcher at the Katholieke Universiteit Leuven.

Marcelo Espinoza obtained a degree in civil engineering and an MSc in applied economics in 1998 at the University of Chile, and the degree of Master in Artificial Intelligence in 2002 at the Katholieke Universiteit Leuven. After 4 years' experience in international commodity trading, he is currently pursuing a PhD at the Katholieke Universiteit Leuven.

Bart Baesens obtained the degree of Master in Management Informatics and a PhD in applied economic sciences at the K.U.Leuven (Belgium) in 1998 and 2003, respectively. He currently works as a lecturer (assistant professor) at the School of Management, University of Southampton (UK), and the Department of Applied Economics, K.U.Leuven (Belgium).

Johan Suykens obtained a degree in electro-mechanical engineering and a PhD (subject: artificial neural networks) in applied sciences from the Katholieke Universiteit Leuven in 1989 and 1995, respectively. In 1996

DOI: 10.1002/for

he was a visiting postdoctoral researcher at the University of California, Berkeley. He is currently an associated professor at the Department of Electrical Engineering, Katholieke Universiteit Leuven.

Bart De Moor received his doctoral degree in applied sciences in 1988 at the Katholieke Universiteit Leuven, Belgium. He was a visiting research associate (1988–1989) at the Department of Computer Science and Electrical Engineering of Stanford University, California. Bart De Moor is a full professor at the Katholieke Universiteit Leuven.

Carine Brasseur received her doctoral degree in economic sciences in 2000 at the Université catholique de Louvain, Belgium. She currently works as a senior strategist in the Research Department at Global Markets, Fortis Bank Belgium.

Authors' addresses:
Tony Van Gestel, Credit Risk Modelling, Risk Management, Dexia Group, Square Meeus 1, B-1000 Brussels, Belgium.

Tony Van Gestel, **Marcelo Espinoza**, **Johan A. K. Suykens** and **Bart De Moor**, Katholieke Universiteit Leuven, Department of Electrical Engineering ESAT-SISTA, Kasteelpark Arenberg 10, B-3001 Leuven, Belgium.

Bart Baesens, School of Management, University of Southampton, Southampton SO17 1BJ, UK.

Carine Brasseur, Financial Markets, Fortis Bank Brussels, Warandeberg 3, B-1000 Brussels, Belgium.

DOI: 10.1002/for