BAYESIAN INPUT SELECTION FOR NONLINEAR REGRESSION WITH LS-SVMS

T. Van Gestel [*,1] **M. Espinoza** [*,1] **J.A.K. Suykens** [*] **C. Brasseur** [**]
B. De Moor [*]

[*] *K.U. Leuven, ESAT-SCD-SISTA, Leuven, Belgium.*
[**] *Fortis Bank, Financial Markets, Brussels, Belgium.*

Abstract: Input selection for linear and nonlinear modelling is an important problem, related to the trade-off between model complexity and in sample model accuracy. For linear modelling, well-known complexity criteria like the Akaike and Bayesian Information Criteria have been developed. In this paper, we explain the Bayesian evidence framework for Least Squares Support Vector Machines (LS-SVMs) and explain its use for input selection.

Keywords: Bayesian Learning, Input Selection, Kernel Based Learning, Least Squares Support Vector Machines, Nonlinear Regression

1. INTRODUCTION

Input selection for linear and nonlinear modelling is an important problem. Given a set of candidate explanatory input variables, a the trade-off between model complexity and in sample model accuracy is made. For linear modelling well-known complexity criteria like the Akaike and Bayesian Information Criteria (AIC and BIC) have been developed consisting of an in sample error term and a complexity term that penalizes too complex models (Akaike, 1978; Schwarz, 1978). These complexity terms are typically estimated using asymptotical approximations that yield the optimal model with respect to some distance measure.

While powerful design techniques for multilayer perceptrons like the Bayesian evidence framework (MacKay, 1992) have been developed, the practical use of neural networks for nonlinear modelling 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 SVMs (LS-SVMs) and related kernel based techniques (Schölkopf and Smola, 2002; Suykens *et al.*, 2002; Vapnik, 1998; Wahba, 1990), 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 [2] . The solution follows from a linear system in the dual space in terms of the positive definite kernel function by applying Mercer's Theorem.

In this paper, the Bayesian evidence framework (Van Gestel *et al.*, 2001*a*; Van Gestel *et al.*, 2002*a*) to perform input selection for nonlinear regression is explained. Bayes' formula is applied on different levels of inference to tune hyperparameters and to select the relevant inputs. The Bayesian framework embodies Occam's razor to find an optimal trade-off between in sample accuracy and model complexity in a similar way as the AIC and BIC.

This paper is organized as follows. The primal-dual LS-SVM formulations are reviewed in Section 2. The Bayesian framework is explained in Section 3 and applied in Section 4.

[1] Corresponding authors, email: {tony.vangestel,marcelo.espinoza}@esat.kuleuven.ac.be. T. Van Gestel is also a senior quantitative analyst with Dexia Group (Credit Risk Management): tony.vangestel@dexia.com.

[2] The LS-SVM formulation is related to regularization networks (Poggio and Girosi, 1990) and Gaussian Processes (Neal, 1996), but the explicit primal-dual formulations allow to make extensions to nonlinear control, recurrent predictions and kernel principal component analysis (Suykens *et al.*, 2002; Suykens *et al.*, 2003).

2. FROM LINEAR TO NONLINEAR KERNEL BASED MODELING AND PREDICTION

2.1 *Kernel Functions*

An important concept of nonlinear kernel based regression and prediction is the interpretation of applying linear regression techniques in the kernel induced feature space. The linear model formulation to predict the output y based on the n explanatory variables $x = [x_1; \ldots; x_n] = [x_1, \ldots, x_n]^T$ can be written as

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

with output $y \in \mathbb{R}$, input vector $x \in \mathbb{R}^n$, coefficient vector $w \in \mathbb{R}^n$ and bias term $b \in \mathbb{R}$. A straightforward way to extend the linear model (1) to a nonlinear model is to preprocess the inputs x in an nonlinear way by the mapping

$$\varphi : \mathbb{R}^n \to \mathbb{R}^{n_f} : x \mapsto \varphi(x) \tag{2}$$

where the feature vector $\varphi(x)$ is typically high (or even infinite) dimensional. In kernel based learning, the mapping is defined implicitly in terms of the positive definite kernel function

$$K(x_1, x_2) = \varphi(x_1)^T \varphi(x_2) \tag{3}$$

from Mercer's theorem (Mercer, 1909) and the nonlinear mapping is tuned implicitly in terms of the parameters of the kernel function. Some commonly used kernel functions are the linear kernel $K(x_1, x_2) = x_1^T x_2$, the polynomial kernel $K(x_1, x_2) = (1 + x_1^T x_2/c)^d$ of degree d and the Radial Basis Function (RBF) kernel $K(x_1, x_2) = \exp(-\frac{||x_1 - x_2||_2^2}{\sigma^2})$, where d and c, σ are constants. Notice that the Mercer condition holds for all $\sigma, c \in \mathbb{R}^+$ values in the RBF (Radial Basis Function) and the polynomial case.

Hence, one can first extend the results of linear modelling in a straightforward way to nonlinear modelling by first applying the (linear) reasoning in the kernel induced feature space. Practical expressions are obtained in the dual space by application of the Mercer condition (Mercer, 1909).

2.2 *Primal-Dual Formulations*

Given the training data $\mathscr{D} = \{(x_i, y_i)\}_{i=1}^{n_\mathscr{D}}$, one obtains the coefficient vector w and bias term b from minimization of the following ridge regression problem

$$\min_{w,b,e} \mathscr{J}(w, b) = \frac{\mu}{2} w^T w + \frac{\zeta}{2} \sum_{i=1}^{n_\mathscr{D}} e_i^2 \tag{4}$$

$$\text{s.t.} \quad e_i = y_i - (w^T \varphi(x_i) + b), \quad i = 1, \ldots, n_\mathscr{D}. \tag{5}$$

The hyperparameters $\mu, \zeta \in \mathbb{R}^+$ allow to find an appropriate trade-off between regularization $\frac{1}{2} w^T w$ and training set error minimization $\frac{1}{2} \sum_{i=1}^{n_\mathscr{D}} e_i^2$. This balance between (in sample) training set accuracy and model complexity is made on the second level of inference. The use of the regularization term in ridge regression (Hoerl and Kennard, 1970) has been used to increase the quality of the model parameter estimates in identification, e.g., in in (Golub *et al.*, 1979; Sjöberg *et al.*, 1995; Van Gestel *et al.*, 2001*b*). Observe that no regularization is applied on the bias term (Brown, 1977).

In order to solve the constrained optimization problem (4)-(5), one constructs the Lagrangian $\mathscr{L} = \frac{\mu}{2} w^T w + \frac{\zeta}{2} \sum_{i=1}^{n_\mathscr{D}} e_i^2 + \sum_{i=1}^{n_\mathscr{D}} \alpha_i (y_i - (w^T \varphi(x_i) + b) - e_i)$, where the support values $\alpha_i \in \mathbb{R}$ are the Lagrange multipliers associated with the equality constraints (5). Defining the matrix notation $\Phi = [\varphi(x_1), \ldots, \varphi(x_{n_\mathscr{D}})]^T \in \mathbb{R}^{n_\mathscr{D} \times n_f}$, $\alpha = [\alpha_1, \ldots, \alpha_{n_\mathscr{D}}]^T \in \mathbb{R}^{n_\mathscr{D}}$, $e = [e_1, \ldots, e_{n_\mathscr{D}}]^T \in \mathbb{R}^{n_\mathscr{D}}$, $y = [y_1, \ldots, y_{n_\mathscr{D}}]^T \in \mathbb{R}^{n_\mathscr{D}}$ and $1 = [1, \ldots, 1]^T \in \mathbb{R}^{n_\mathscr{D}}$, the conditions for optimality correspond to the symmetric linear system

$$\begin{bmatrix} -\mu I_{n_f} & \Phi^T & \mathbf{0} & \mathbf{0} \\ \Phi & \mathbf{0} & 1 & I_{n_\mathscr{D}} \\ \mathbf{0} & 1^T & \mathbf{0} & \mathbf{0} \\ \mathbf{0} & I_{n_\mathscr{D}} & \mathbf{0} & -\zeta I_{n_\mathscr{D}} \end{bmatrix} \cdot \begin{bmatrix} w \\ \alpha \\ b \\ e \end{bmatrix} = \begin{bmatrix} 0 \\ y \\ 0 \\ 0 \end{bmatrix}. \tag{6}$$

For the linear case (e.g. a linear kernel) one typically has $n_f = n \ll n_\mathscr{D}$ and after elimination of e and α, one solves the $(n_f + 1) \times (n_f + 1)$ linear system in the primal space

$$\begin{bmatrix} \Phi^T \Phi + \frac{\mu}{\zeta} I_{n_f} & \Phi^T 1 \\ 1^T \Phi & n_\mathscr{D} \end{bmatrix} \begin{bmatrix} w \\ b \end{bmatrix} = \begin{bmatrix} \Phi^T y \\ 1^T y \end{bmatrix}. \tag{7}$$

In nonlinear kernel based regression, one usually has $n_f \gg n_\mathscr{D}$ and, moreover, the feature vector $\varphi(x)$ is only implicitly defined in terms of the kernel function K from (3). Eliminating w and e from (6), one obtains the linear Karush-Kuhn-Tucker system of dimension $(n_\mathscr{D} + 1) \times (n_\mathscr{D} + 1)$ in the dual space (Suykens and Vandewalle, 1999; Suykens *et al.*, 2002)

$$\begin{bmatrix} \frac{1}{\mu} \Omega + \frac{1}{\zeta} I_{n_\mathscr{D}} & 1 \\ 1^T & 0 \end{bmatrix} \begin{bmatrix} \alpha \\ b \end{bmatrix} = \begin{bmatrix} y \\ 0 \end{bmatrix}, \tag{8}$$

where the Mercer condition (3) is applied in the matrix $\Omega = \Phi \Phi^T \in \mathbb{R}^{n_\mathscr{D} \times n_\mathscr{D}}$ with elements $\Omega_{ij} = K(x_i, x_j)$, $i, j = 1, \ldots, n_\mathscr{D}$ and guarantees that $\Omega \geq 0$. Defining the idempotent centering matrix $\mathbf{N}_\mathrm{c} = I_{n_\mathscr{D}} - \frac{1}{n_\mathscr{D}} 1 1^T \in \mathbb{R}^{nd \times n_\mathscr{D}}$, one obtains explicit formulas for α and b: $(\frac{1}{\mu} \mathbf{N}_\mathrm{c} \Omega \mathbf{N}_\mathrm{c} + \frac{1}{\zeta} I_{n_\mathscr{D}}) \alpha = \mathbf{N}_\mathrm{c} y$ and $n_\mathscr{D} b = 1^T (\frac{1}{\mu} \Omega + \frac{1}{\zeta} I_{n_\mathscr{D}})^{-1} y$, resp. Up to the centering, these expressions correspond to the (only) dual expressions in Gaussian Processes (Neal, 1996). Applying no regularization on the bias term results into a zero mean training set error and is the preferred form of rigde regression as has been motivated in (Brown, 1977).

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

$$\hat{y} = w^T \varphi(x) + b = \frac{1}{\mu} \sum_{t=1}^{n_\mathscr{D}} \alpha_t K(x, x_t) + b. \tag{9}$$

3. BAYESIAN INFERENCE

Given the primal-dual formulations, it is clear how to estimate the model parameters w, b and point prediction y. However, the regularization and kernel function parameters still have to be tuned from the given training data. In this Section, the model parameters w, b, the hyperparameters μ, ζ and model structure $\mathscr{M}$ (corresponding, e.g., to the input set and/or tunable kernel parameters) are inferred from the data $\mathscr{D}$ by applying Bayes formula on three different levels (Van Gestel *et al.*, 2002*a*; Van Gestel *et al.*, 2001*a*).

3.1 *Inference of the Model Parameters (Level 1)*

3.1.1. *Probabilistic Interpretation of the LS-SVM Formulation* Given the data points $\mathscr{D} = \{(x_t, y_t)\}_{t=1}^{n_{\mathscr{D}}}$ and the hyperparameters μ and ζ of the model $\mathscr{M}$ (LS-SVM with given input set and kernel function K), we obtain the model parameters w and b by maximizing the posterior $p(w,b|\mathscr{D},\log\mu,\log\zeta,\mathscr{M})$. Application of Bayes' rule at the first level of inference (MacKay, 1992) gives

$$p(w,b|\mathscr{D},\log\mu,\log\zeta,\mathscr{M}) \tag{10}$$
$$= \frac{p(\mathscr{D}|w,b,\log\mu,\log\zeta,\mathscr{M})\,p(w,b|\log\mu,\log\zeta,\mathscr{M})}{p(\mathscr{D}|\log\mu,\log\zeta,\mathscr{M})},$$

where the evidence $p(\mathscr{D}|\log\mu,\log\zeta,\mathscr{M})$ is a normalizing constant that will be related to model complexity.

The regularization parameters μ and ζ are related to the prior $p(w,b|\log\mu,\log\zeta,\mathscr{M})$ and likelihood $p(\mathscr{D}|w,b,\log\mu,\log\zeta,\mathscr{M})$, respectively. We take the prior $p(w,b|\log\mu,\log\zeta,\mathscr{M})$ independent of the hyperparameter ζ, i.e., $p(w,b|\log\mu,\log\zeta,\mathscr{M}) = p(w,b|\log\mu,\mathscr{M})$. Both w and b are assumed to be independent. The coefficients w are assumed to be multivariate Gaussian distributed with mean zero and covariance matrix $\frac{1}{\mu}I_{n_f}$:

$$p(w|\log\mu,\mathscr{M}) = \left(\frac{\mu}{2\pi}\right)^{\frac{n_f}{2}} \exp(-\tfrac{\mu}{2}w^Tw). \tag{11}$$

This means that 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 $w_k = 0$ ($k = 1,\ldots,n_f$). A uniform distribution for the prior on b is taken, which can also be approximated as a Gaussian distribution $p(b|\log\sigma_b,\mathscr{M}) = (2\pi\sigma_b)^{-1/2}\exp(-b^2/(2\sigma_b^2))$, with $\sigma_b \to \infty$. This is called a uniform prior and explains why there is no regularization term on b in (4)-(5). The negative logarithm of the prior corresponds to the regularization term $\frac{\mu}{2}w^Tw$ in (4)-(5).

We take the likelihood of the observed data $\mathscr{D}$ independent of the hyperparameter μ and assume that all data points (x_t, y_t) are independent:

$$\begin{aligned} &p(\mathscr{D}|w,b,\log\zeta,\mathscr{M}) \\ &= \textstyle\prod_{t=1}^{n_{\mathscr{D}}} p(y_t|x_t,w,b,\log\zeta,\mathscr{M})\,p(x_t|w,b,\log\zeta,\mathscr{M}) \\ &\propto \textstyle\prod_{t=1}^{n_{\mathscr{D}}} p(y_t|x_t,w,b,\log\zeta,\mathscr{M}), \end{aligned} \tag{12}$$

with $p(x_t|w,b,\log\zeta,\mathscr{M}) = p(x_t)$. Assuming that the additive noise e_t is drawn from a Gaussian distribution with variance $1/\zeta$, we have $p(y_t|x_t,w,b,\log\zeta,\mathscr{M}) = (\frac{\zeta}{2\pi})^{1/2}\exp(-\frac{\zeta}{2}(y_t - w^T\varphi(x_t) - b)^2)$.. The negative logarithm of the likelihood (12) corresponds to the sum squared error term $\frac{\zeta}{2}\sum_{t=1}^{n_{\mathscr{D}}} e_t^2$.

Substituting (11) and (12) into (10), neglecting all constants and taking the negative logarithm, Bayes' rule at the first level of inference corresponds to the constrainted minimization problem (4)-(5) that can be solved for w and b in the primal space from (7) in the linear case when $n \leq n_{\mathscr{D}}$ or α and b in the dual space from (8) in the nonlinear kernel-based regression case and in the linear case when $n \geq n_{\mathscr{D}}$. In the remainder of this paper, the maximum a posteriori parameter estimates are denoted by the subscript 'mp', e.g., w_{mp} and b_{mp}.

We discuss now an alternative representation of the LS-SVM cost function (4)-(5) and the posterior probability (10), which will be used in the next Subsection. By observing that the LS-SVM cost function is a quadratic cost function, the posterior $p(w,b|\mathscr{D},\log\mu,\log\zeta,\mathscr{M})$ can also be written as the Gaussian distribution with mean $[w_{mp};b_{mp}]$ and covariance matrix Q. $Q = \mathrm{covar}([w;b]) = \mathscr{E}([w-w_{mp};b-b_{mp}][w-w_{mp};b-b_{mp}]^T)$, The covariance matrix Q is related to the Hessian H of the LS-SVM cost function (4)-(5):

$$Q^{-1} = H = \begin{bmatrix} \mu I + \zeta\Phi^T\Phi & \zeta\Phi^T 1 \\ \zeta 1^T\Phi & \zeta n_{\mathscr{D}} \end{bmatrix}, \tag{13}$$

where the inverse is calculated using, e.g., a Schur complement type of argument.

3.1.1.1. *Model Evidence and Complexity Terms* The evidence $p(D|\log\mu,\log\zeta,\mathscr{M})$ is obtained by using the expressions for the prior, likelihood and posterior from Bayes formula (10)

$$p(\mathscr{D}|\log\mu,\log\zeta,\mathscr{M}) \tag{14}$$
$$= \frac{p(\mathscr{D}|w,b,\log\mu,\log\zeta,\mathscr{M})\,p(w,b|\log\mu,\log\zeta,\mathscr{M})}{p(w,b|D,\log\mu,\log\zeta,\mathscr{M})}.$$

Substituting (11) and (12) into (14) yields (Van Gestel *et al.*, 2002*a*; Van Gestel *et al.*, 2001*a*)

$$p(\mathscr{D}|\log\mu,\log\zeta,\mathscr{M}) \propto \tag{15}$$
$$\underbrace{p(\mathscr{D}|w_{mp},b_{mp},\log\zeta,\mathscr{M})}_{\text{model fit}}\ \underbrace{p(w_{mp}|\log\mu,\mathscr{M})(\det H)^{-\frac{1}{2}}}_{\text{Occam factor}}.$$

Within the Bayesian framework, the evidence of the model is equal to the best fit likelihood of the posterior model parameters times the Occam factor that penalizes models that are too complex (MacKay, 1992) and have too many degrees of freedom. As the determinant of the covariance matrix $(\det Q)^{1/2} = (\det H)^{-1/2}$ is a measure for the volume of the posterior distribution, a large contraction in the volume of the prior distribution towards the posterior may tend to cause overfitting because there are too many free parameters,

the Occam factor penalizes complex models with too many degrees of freedom. Taking the negative logarithm of (15), the model evidence consists of the training set error plus a complexity term. This is related to well known complexity criteria like the Akaike and Bayesian Information Criterion (Akaike, 1978; Schwarz, 1978) that are obtained using appropriate (asymptotical) simplifications.

3.2 *Inference of the Hyperparameters (Level 2)*

The dimension of the kernel induced feature space can be larger than the number of training data points and for some nonlinear kenrel functions, like the RBF kernel, it can even become infinite dimensional. Therefore, it is necessary to find an appropriate trade-off between regularization and training set error by tuning the hyperparameters μ and ζ.

3.2.1. *Optimization Problem for μ and ζ* The regularization hyperparameters μ and ζ (MacKay, 1992) are inferred from the data $\mathscr{D}$ by applying Bayes' rule on the second level:

$$\begin{aligned} & p(\log\mu,\log\zeta|\mathscr{D},\mathscr{M}) \\ &= \frac{p(\mathscr{D}|\log\mu,\log\zeta,\mathscr{M})p(\log\mu,\log\zeta|\mathscr{M})}{p(\mathscr{D}|\mathscr{M})} \\ &\propto p(\mathscr{D}|\log\mu,\log\zeta,\mathscr{M}), \end{aligned} \tag{16}$$

where a flat, non-informative prior is assumed on the hyperparameters μ and ζ. The probability $p(\mathscr{D}|\log\mu, \log\zeta,\mathscr{M})$ is equal to the evidence of the previous level. Taking the negative logarithm of (16) and using (10), one obtains the following cost function

$$\begin{aligned} \mathscr{J}(\mu,\zeta) &= \mu\,\mathscr{J}_w(w_{mp}) + \zeta\,\mathscr{J}_e([w_{mp};b_{mp}]) \\ &\quad + \frac{1}{2}\log\det H - \frac{n_f}{2}\log\mu - \frac{n_{\mathscr{D}}}{2}\log\zeta, \end{aligned} \tag{17}$$

where we used the shorthand notation $\mathscr{J}_w(w) = \frac{1}{2}w^Tw$ and $\mathscr{J}_e([w;b]) = \frac{1}{2}e^Te$ with $e = y - [\Phi;1^T][w;b]$. The expression for $\det H$ is equal to $\det H = n_{\mathscr{D}}\zeta\mu^{n_f-n_f^e}\prod_{i=1}^{n_f^e}(\mu+\zeta\lambda_{\Omega_c,i})$, where $\lambda_{\Omega_c,i}$, $i=1,\ldots,n_f^e$ are the n_f^e non-zero eigenvalues of $\mathbf{N}_c\Phi\Phi^T\mathbf{N}_c = \mathbf{N}_c\Omega\mathbf{N}_c = \Omega_c$.

3.2.2. *Scalar Optimization Problem for γ* In the dual space, the regularization parameter $\gamma = \zeta/\mu$ is obtained by minimizing (Van Gestel *et al.*, 2002*a*; Van Gestel *et al.*, 2001*a*)

$$\begin{aligned} \min_{\gamma}\ \mathscr{J}_2(\gamma) &= \textstyle\sum_{i=1}^{n_{\mathscr{D}}-1}\log(\lambda_{\Omega,i}+\gamma^{-1}) \\ &\quad + (n_{\mathscr{D}}-1)\log[\mathscr{J}_w(w)+\gamma\,\mathscr{J}_e(w,b)], \end{aligned} \tag{18}$$

with $\mathscr{J}_w = \frac{1}{2}w^Tw$, $\mathscr{J}_e = \frac{1}{2}e^Te$ and where $\lambda_{\Omega,i}$ ($i = 1,\ldots,n_{\mathscr{D}}-1$) are the $n_{\mathscr{D}}-1$ eigenvalues of the centered kernel matrix $M_c\Omega M_c$, with idempotent centering matrix $M_c = I - 1/n_{\mathscr{D}}11^T$. The level 1 cost function can be expressed as $\mathscr{J}_w(w)+\gamma\mathscr{J}_e(w,b) = \frac{1}{2}y^TM_c(M_c\Omega M_c+\gamma^{-1}I_{n_{\mathscr{D}}})^{-1}M_cy$, which avoids the explicit solution of the linear system (8) for all possible γ values. As these expressions need to be evaluated for different values of γ one can compute first an eigenvalue decomposition of $M_c\Omega M_c$. This allows then to compute $\mathscr{J}_w$, $\mathscr{J}_e$ and $\mathscr{J}_w+\gamma\mathscr{J}_e$ from matrix vector multiplications only, avoiding the (computationally intensive) calculation of the inverse $(M_c\Omega M_c+\gamma^{-1}I_{n_{\mathscr{D}}})^{-1}$ (Van Gestel *et al.*, 2002*a*; Van Gestel *et al.*, 2001*a*) for different values of γ.

Given the optimal γ from (18) one finds the effective number of parameters[3] d_{eff} from $d_{\text{eff}} = (n_{\mathscr{D}} + \gamma\mathscr{J}_e/\mathscr{J}_w)/(1+\gamma\mathscr{J}_e/\mathscr{J}_w)$. The optimal μ and ζ are obtained from $\mu = (d_{\text{eff}}-1)/(2\mathscr{J}_w(w))$ and $\zeta = (n_{\mathscr{D}}-d_{\text{eff}})/(2\mathscr{J}_e(w,b))$, where the latter is related to the unbiased sample estimate of the noise variance.

3.3 *Model Comparison (Level 3)*

The model structure $\mathscr{M}$ with RBF-kernel parameter σ is inferred on Level 3:

$$p(\mathscr{M}|\mathscr{D}) \propto p(\mathscr{D}|\mathscr{M})p(\mathscr{M}), \tag{19}$$

where typically equal prior probabilities are assumed, meaning that $p(\mathscr{M}|\mathscr{D}) \propto p(\mathscr{D}|\mathscr{M})$. The posterior level 3 probability is obtained as the level 2 probability with an additional Occam factor due to the posterior uncertainty on the inferred hyperparameters μ and ζ

$$p(\mathscr{D}|\mathscr{M}) \simeq p(\mathscr{D}|\log\mu,\log\zeta,\mathscr{M})\frac{\sigma_{\log\mu|\mathscr{D}}\sigma_{\log\zeta|\mathscr{D}}}{\sigma_{\log\mu}\sigma_{\log\zeta}},$$

with $\sigma_{\log\mu|\mathscr{D}} \simeq 2/(d_{\text{eff}}-1)$ and $\sigma_{\log\zeta|\mathscr{D}} \simeq 2/(n_{\mathscr{D}}-d_{\text{eff}})$. When flat, uninformative priors are used on level 2, $\sigma_{\log\mu}$ and $\sigma_{\log\zeta}$ are omitted in the comparisons of different models. In the case of equal prior model probabilities $p(\mathscr{M}_i) = p(\mathscr{M}_j)$ $(\forall i,j)$ the models $\mathscr{M}_i$ and $\mathscr{M}_j$ are compared according to their Bayes factor $\mathscr{B}_{ij} = p(\mathscr{D}|\mathscr{M}_i)/p(\mathscr{D}|\mathscr{M}_j)$:

$$\mathscr{B}_{ij} = \frac{p(\mathscr{D}|\log\mu_i,\log\zeta_i,\mathscr{M}_i)}{p(\mathscr{D}|\log\mu_j,\log\zeta_j,\mathscr{M}_j)}\frac{\sigma_{\log\mu_i|\mathscr{D}}\sigma_{\log\zeta_i|\mathscr{D}}}{\sigma_{\log\mu_j|\mathscr{D}}\sigma_{\log\zeta_j|\mathscr{D}}}. \tag{20}$$

3.4 *Bayesian Input Selection*

Model comparison is also used to infer the set of most relevant inputs (Van Gestel *et al.*, 2001*a*) 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 (19). The procedure is stopped when no significance decrease of the model probability is observed.

---

[3] The effective number of parameters can be defined in terms of the trace of the smoother matrix or in terms of the ratios of the eigenvalues of the Hessian of the unregulated and regulated cost function: $d_{\text{eff}} = 1 + \sum_{i=1}^{n_f^e}(\zeta\lambda_{\Omega_c,i})/(\mu+\zeta\lambda_{\Omega_c,i})$.

4. DESIGN & IMPLEMENTATION

A practical scheme for the design of the LS-SVM within the evidence framework consists of the following steps:

(1) The inputs are standardized to zero mean and unit variance. The normalized training data are denoted by $D = \{(x_t, y_t)\}_{t=1}^{n_{\mathscr{D}}}$.
(2) Select the model $\mathscr{M}_j$ by choosing a kernel type K_j (possibly with a kernel parameter, e.g., σ_j for an RBF-kernel). For this model $\mathscr{M}_j$, the optimal hyperparameters μ_{mp} and ζ_{mp} are estimated on the second level of inference. This is done as follows: a) Solve the scalar optimization problem (18) in $\gamma = \zeta/\mu$ using, e.g., a quasi-Newton method. b) Calculate the effective number of parameters d_{eff} and the regularization parameters μ_{mp} and $\zeta_{mp} = \mu_{mp}\gamma_{mp}$.
(3) Calculate the model evidence $p(\mathscr{M}_j|\mathscr{D})$ from (19).
(4) For a kernel K_j with tuning parameters, refine the tuning parameters. E.g., for the RBF-kernel with tuning parameter σ_j, refine σ_j such that a higher model evidence $p(\mathscr{D}|\mathscr{M}_j)$ is obtained. This can be done by maximizing the model evidence with respect to σ_j by evaluating the model evidence for the refined kernel parameter starting from step 2(a).
(5) Select the model $\mathscr{M}_j$ with maximal model evidence $p(\mathscr{D}|\mathscr{M}_j)$. Go to step 2, unless the best model has been selected.

Input selection is done by starting from an initial candidate input set and calculating the model evidence $p(\mathscr{M}_j|\mathscr{D})$ by taking steps 2-4. In a backward input selection procedure, one step consists of comparing the model probabilities of the model with full candidate input sets and all the models where in turn one of the candidate inputs has been removed. The backward input steps are repeated until $p(\mathscr{M}_j|D)$ stops increasing.

5. EXAMPLES

We apply Bayesian input selection on the estimation of a noisy sinc function $y = \text{sinc}(x_3) + e$, (with $\mathscr{E}(e) = 0$ and $\sigma_e = 0.1$), where the correct input x_3 has to be determined from 5 candidate inputs $x = [x_1; x_2; x_3; x_4; x_5]$. All inputs were drawn from a uniform distribution in the interval $[-3, 3]$. In each backward input selection step the optimal hyper and kernel parameters of the LS-SVM with RBF-kernel are determined. These are used as a starting point to calculate the level 3 cost function when removing in turn the candidate inputs (Figure 1.a). The backward input selection is stopped when the level 3 cost function starts increasing (Figure 1.b), in this case after 5 input pruning steps, selecting the 'true' input x_3. The estimated functions using all 5 candidate inputs and only the 'true' input are depiced in Figures 1.c and 1.d, respectively.

The real-life problem we are concerned with is to weekly predict an aggregated index for the European chemical sector based on a set of 27 candidate explanatory input variables selected by a financial analyst and resulting from econometric time series analysis (Van Gestel *et al.*, 2002*b*). Performing backward

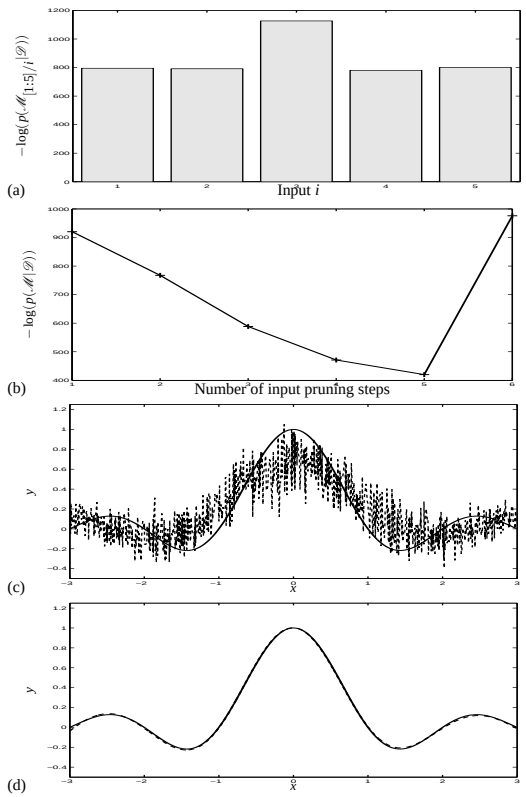


Fig. 1. Bayesian input selection. (a) Level 3 cost function when removing different inputs at step 1. (b) Evolution of level 3 cost function as a function of the number of input pruning steps. (c-d) True and estimated sinc function (full and dashed line, resp.) using 5 and 1 inputs.

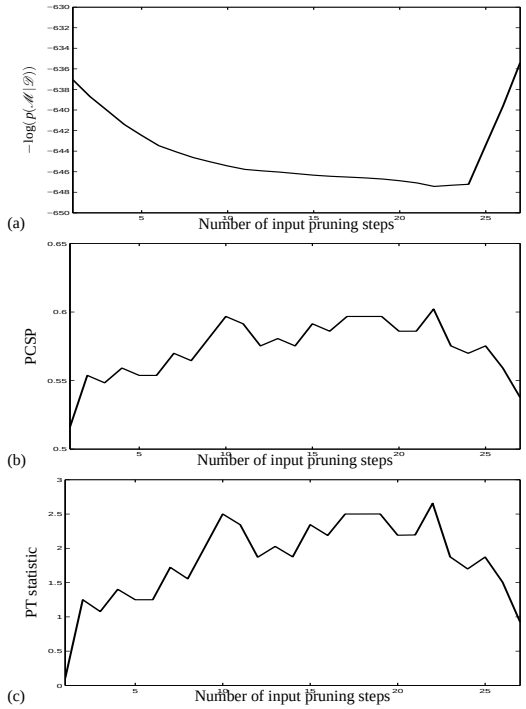


Fig. 2. Evolution of level 3 cost function, percentage of correct sign prediction (PCSP), Pesaran-Timmerman test statistic (PTstat) as a function of the number of input pruning steps.

input selection, the evolution of the level 3 cost function $-\log(p(\mathscr{M}|\mathscr{D}))$, the percentage of out of sample sign predictions and the Pesaran-Timmerman test statistic [4] are depicted in Figure 2 as a function of the number of input pruning steps. Observe that the PT-stat value becomes maximal when the level 3 cost function becomes minimal.

6. CONCLUSIONS

The primal-dual LS-SVM formulations allow to extend the Bayesian framework for linear ridge regression to nonlinear kernel based regression. In this paper, input selection is performed for nonlinear regression within this Bayesian framework. Basically, a trade-off is made between in sample model accuracy and model complexity in a similar way as in the Akaike and Bayesian Information Criteria.

7. ACKNOWLEDGEMENTS

T. Van Gestel is a senior quantitative analyst with Dexia Group (Risk Management Methodology) and holds a honorary position as a Postdoctoral Researchers with the Fund for Scientific Research-Flanders (FWO-Vlaanderen). J. Suykens is a Postdoctoral Researcher with the FWO-Vlaanderen. This work was partially supported by grants and projects from the Research Council K.U.Leuven (GOA-Mefisto 666, IDO, PhD/postdoc & fellow grants), the Flemish Government (FWO-Vlaanderen: PhD/postdoc grants, projects G.0256.97, G.0115.01, G.0240.99, G.0197.02, G.0407.02, ICCoS, ANMMM; AWI, IWT: Soft4s, STWW-Genprom, GBOU-McKnow, Eureka-Impact/FLiTE, PhD grants), the Belgian Federal Government (DWTC: IUAP IV-02, IUAP V-22; PODO-II CP/40). The scientific responsibility is assumed by its authors.

8. REFERENCES

Akaike, H. (1978). A Bayesian analysis of the minimum AIC procedure. *Annals of the Institute of Statistical Mathematics* **30**, 9–14.

Brown, P.J. (1977). Centering and scaling in ridge regression. *Technometrics* **19**, 35–36.

Golub, G.H., M. Heath and G. Wahba (1979). Generalized cross-validation: a method for choosing a good ridge regression parameter. *Technometrics* **21**, 215–223.

Hoerl, A.E. and R.W. Kennard (1970). Ridge regression: Biased estimation for nonorthogonal problems. *Technometrics* **12**, 55–67.

MacKay, D.J.C. (1992). Bayesian interpolation. *Neural Computation* **4**, 415–447.

Mercer, J. (1909). Functions of positive and negative type and their connection with the theory of integral equations. *Philos. Trans. Roy. Soc. London* **209**, 415–446.

Neal, R.M. (1996). *Bayesian Learning for Neural Networks*. Vol. 118 of *Lecture Notes in Statistics*. Springer. New York.

Pesaran, M.H. and A. Timmerman (1992). A simple nonparametric test of predictive performance. *Journal of Business and Economic Statistics* **10**, 461–465.

Poggio, T. and F. Girosi (1990). Networks for approximation and learning. *Proceedings of the IEEE* **78**, 1481–1497.

Schölkopf, B. and A. Smola (2002). *Learning with Kernels*. MIT Press. Cambridge, MA.

Schwarz, G. (1978). Estimating the dimension of a model. *Annals of Statistics* **6**, 461–464.

Sjöberg, J., Q. Zhang, L. Ljung, A. Benveniste, B. Deylon, P. Glorennec, H. Hjalmarsson and A. Juditsky (1995). Nonlinear black-box modelling in system identification: a unified overview. *Automatica* **31**, 1691–1724.

Suykens, J.A.K. and J. Vandewalle (1999). Least squares support vector machine classifiers. *Neural Processing Letters* **9**, 293–300.

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

Suykens, J.A.K., T. Van Gestel, J. Vandewalle and B. De Moor (2003). A support vector machine formulation to pca analysis and its kernel version. *IEEE Transactions on Neural Networks*. Accepted for publication.

Van Gestel, T., J.A.K. Suykens, D.-E. Baestaens, A. Lambrechts, G. Lanckriet, B. Vandaele, B. De Moor and J. Vandewalle (2001*a*). 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., J.A.K. Suykens, G. Lanckriet, A. Lambrechts, B. De Moor and J. Vandewalle (2002*a*). A Bayesian framework for least squares support vector machine classifiers, Gaussian processes and kernel Fisher discriminant analysis. *Neural Computation* **14**, 1115–1147.

Van Gestel, T., J.A.K. Suykens, P. Van Dooren and B. De Moor (2001*b*). Identification of stable models in subspace identification by using regularization. *IEEE Transactions on Automatic Control* **46**, 1416–1420.

Van Gestel, T., M. Espinoza, J. Suykens, B. Carine and B. De Moor (2002*b*). Bayesian nonlinear kernel based stock market prediction. Technical Report 02-151. K.U.Leuven, ESAT/SCD/SISTA. Submitted for publication.

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

Wahba, G. (1990). *Spline Models for Observational Data*. SIAM. Philadelphia.

[4] This test statistic (PTstat) is standard normal distributed under the null hypotesis of random sign predictions (Pesaran and Timmerman, 1992), values above 1.96 and 2.58 are significant at 5% and 1%.