Approximate Confidence and Prediction Intervals for Least Squares Support Vector Regression

Kris De Brabanter, Jos De Brabanter, Johan A. K. Suykens, *Senior Member, IEEE,* and Bart De Moor, *Fellow, IEEE*

***Abstract*— Bias-corrected approximate $100(1-\alpha)\%$ pointwise and simultaneous confidence and prediction intervals for least squares support vector machines are proposed. A simple way of determining the bias without estimating higher order derivatives is formulated. A variance estimator is developed that works well in the homoscedastic and heteroscedastic case. In order to produce simultaneous confidence intervals, a simple Šidák correction and a more involved correction (based on upcrossing theory) are used. The obtained confidence intervals are compared to a state-of-the-art bootstrap-based method. Simulations show that the proposed method obtains similar intervals compared to the bootstrap at a lower computational cost.**

***Index Terms*— Bias, confidence interval, heteroscedasticity, homoscedasticity, kernel-based regression, variance.**

I. Introduction

NONPARAMETRIC function estimators are very popular data analytic tools [1]–[3]. Many of their properties have been rigorously investigated and are well understood. An important activity immediately accompanying these estimators is the construction of interval estimates, e.g., confidence intervals. In the area of kernel-based regression, a popular tool to construct interval estimates is the bootstrap (see [4], references therein). This technique produces very accurate intervals at the cost of heavy computational burden.

In the field of neural networks, [5], [6] have proposed confidence and prediction intervals. In case of nonlinear regression, [7] proposed confidence intervals based on least squares estimation and using the linear Taylor expansion of the nonlinear model. For Gaussian processes (GPs) several type of methods have been developed to construct interval estimates. Early works of [8] and [9] address the construction of interval estimates via a Bayesian approach and a Markov chain Monte Carlo method to approximate the posterior noise variance. An extension of the latter was proposed by [10]. In general, Bayesian intervals (which are often called Bayesian credible intervals) do not exactly coincide with frequentist confidence intervals (as proposed in this paper) for two reasons. First, credible intervals incorporate problem-specific contextual information from the prior distribution, whereas confidence intervals are based only on the data. Second, credible intervals and confidence intervals treat nuisance parameters in radically different ways (see [11], references therein).

Recently, [12] and [13] propose to use the leave-one-out cross-validation estimate of the conditional mean in fitting the model of the conditional variance in order to overcome the inherent bias in maximum likelihood estimates of the variance.

In this paper, we will address some of the difficulties in constructing these interval estimates and develop a methodology for interval estimation in case of least squares support vector machines (LS-SVM) for regression which is not based on bootstrap.

Consider the bivariate data $(X_1, Y_1), \ldots, (X_n, Y_n)$, which form an independent and identically distributed (i.i.d.) sample from a population (X, Y). Our interest is to estimate the regression function $m(X) = \mathbf{E}[Y|X]$, with $\mathbf{E}[Y|X] = \int_{\mathbb{R}} y f_{Y|X}(y|x)\, dy$, where $f_{Y|X}$ is the conditional distribution of the random variables (X, Y). We regard the data as being generated from the model

$$Y = m(X) + \sigma(X)\varepsilon \tag{1}$$

where $\mathbf{E}[\varepsilon|X] = 0$, $\mathbf{Var}[\varepsilon|X] = \mathbf{E}[\varepsilon^2|X] - \mathbf{E}^2[\varepsilon|X] = 1$, and X and ε are independent. In setting (1), it is immediately clear that $\mathbf{E}[Y|X] = m(X)$ and $\mathbf{Var}[Y|X] = \sigma^2(X) > 0$. Two possible situations can occur: 1) $\sigma^2(X) = \sigma^2 =$ constant, and 2) the variance is a function of the random variable X. The first is called homoscedasticity and the latter heteroscedasticity.

The problem that we will address is the construction of confidence intervals for m. Specifically, given $\alpha \in (0, 1)$ and an estimator $\hat{m}$ for m, we want to find a bound g_α such that

$$\mathbf{P}\left(\sup_x |\hat{m}(x) - m(x)| \leq g_\alpha\right) \geq 1 - \alpha \tag{2}$$

Manuscript received July 6, 2010; revised August 30, 2010, October 6, 2010, and October 10, 2010; accepted October 10, 2010. Date of publication November 1, 2010; date of current version January 4, 2011. This work was supported in part by the following agencies: Research Council KUL: GOA AMBioRICS, GOA MaNet, CoE EF/05/006 Optimization in Engineering (OPTEC), IOF-SCORES4CHEM, several PhD/post-doc & fellow Grants; Flemish Government: FWO: PhD/postdoc grants, projects G.0452.04 (new quantum algorithms), G.0499.04 (Statistics), G.0211.05 (Nonlinear), G.0226.06 (cooperative systems and optimization), G.0321.06 (Tensors), G.0302.07 (SVM/Kernel), G.0320.08 (convex MPC), G.0558.08 (Robust MHE), G.0557.08 (Glycemia2), G.0588.09 (Brain-machine) research communities (ICCoS, ANMMM, MLDM), G.0377.09 (Mechatronics MPC); IWT: PhD Grants, McKnow-E, Eureka-Flite+, SBO LeCoPro, SBO Climaqs, POM; Belgian Federal Science Policy Office: IUAP P6/04 (DYSCO, Dynamical systems, control and optimization, 2007-2011); EU: ERNSI; FP7-HD-MPC (INFSO-ICT-223854), COST intelliCIS, EMBOCOM; Contract Research: AMINAL; Other: Helmholtz, viCERP, ACCM, Bauknecht, Hoerbiger.

K. De Brabanter, J. A. K. Suykens, and B. De Moor are with the Department of Electrical Engineering, Research Division SCD, Katholieke Universiteit Leuven, Leuven 3001, Belgium (e-mail: kris.debrabanter@esat.kuleuven.be; johan.suykens@esat.kuleuven.be; bart.demoor@esat.kuleuven.be).

J. De Brabanter is with the Department Industrieel Ingenieur, KaHo Sint Lieven (Associatie K.U. Leuven), Gent 9000, Belgium, (e-mail: jos.debrabanter@kahosl.be).

Color versions of one or more of the figures in this paper are available online at http://ieeexplore.ieee.org.

Digital Object Identifier 10.1109/TNN.2010.2087769

1045–9227/$26.00 © 2010 IEEE

at least for large sample sizes. Also, bear in mind that $\hat{m}$ depends on n.

A major difficulty in finding a solution to (2) is the fact that nonparametric estimators for m are biased (kernel estimators in particular). As a consequence, confidence interval procedures must deal with estimator bias to ensure that the interval is correctly centered and proper coverage is attained [14].

In order to avoid the bias estimation problem, several authors have studied the limiting distribution of $\sup_x |\hat{m}(x) - m(x)|$ for various estimators $\hat{m}$ of m. A pioneering article in this field is due to [15] for kernel density estimation. Extensions of [15] to kernel regression are given in [16]–[18].

A second way to avoid calculating the bias explicitly is to undersmooth. If we smooth less than the optimal amount, then the bias will decrease asymptotically relative to the variance. It was theoretically shown in [4] that undersmoothing in combination with a pivotal statistic based on bootstrap results in the lowest reduction in coverage error of confidence intervals. Unfortunately, there does not seem to be a simple and practical rule for choosing just the right amount of undersmoothing.

A third and more practical way is to be satisfied with indicating the level of variability involved in a nonparametric regression estimator, without attempting to adjust for the inevitable presence of bias. Bands of this type are easier to construct but require careful interpretation. Formally, the bands indicate pointwise variability intervals for $\mathbf{E}[\hat{m}(X)|X]$. Based on this idea, often misconceptions exist between confidence intervals and error bars. In this paper, we propose the construction of confidence intervals based on the central limit theorem for linear smoothers combined with bias correction and variance estimation.

Finally, if it is possible to obtain a reasonable bias estimate, we can use it to construct confidence intervals for m. The application of this approach can be found in local polynomial regression [19], [20], where a bias estimate can be easily calculated. In this paper, we will develop a similar approach to estimate the bias.

Confidence intervals are widely used, and applications can be found, e.g., in the chemical industry, fault detection/diagnosis, and system identification/control. These intervals give the user the ability to see how well a certain model explains the true underlying process while taking statistical properties of the estimator into account. In control applications, these intervals are used for robust design, while the applicability of these intervals in fault detection is based upon reducing the number of false alarms. For further reading on this topic, we refer the reader to the work of [21] and [22].

The rest of this paper is organized as follows. The bias and variance estimation of the LS-SVM are discussed in Section II. The construction of approximate $100(1-\alpha)\%$ pointwise and simultaneous confidence (prediction) intervals are discussed in Section III. Section IV briefly summarizes the ideas of bootstrap-based methods. Simulations and comparison with the current state-of-the-art bootstrap-based method [4] is given in Section V. Some closing comments are made in Section VI.

II. Estimation of Bias and Variance

A. LS-SVM Regression and Smoother Matrix

In the primal weight space, LS-SVM is formulated as follows [23]:

$$\begin{aligned} \min_{w,b,e} \mathcal{J}(w,e) &= \tfrac{1}{2} w^T w + \tfrac{\gamma}{2} \sum_{i=1}^{n} e_i^2 \\ \text{s.t.} \quad Y_i &= w^T \varphi(X_i) + b + e_i \qquad i = 1, \ldots, n \end{aligned} \tag{3}$$

where $e_i \in \mathbb{R}$ are assumed to be i.i.d. random errors with $\mathbf{E}[e|X] = 0$, $\mathbf{Var}[e|X] < \infty$, $m \in C^z(\mathbb{R})$[1] with $z \geq 2$, is an unknown real-valued smooth function, and $\mathbf{E}[Y|X] = m(X)$, $\varphi : \mathbb{R}^d \to \mathbb{R}^{n_h}$ is the feature map to the high-dimensional feature space (can be infinite dimensional) as in the standard SVM case [24], $w \in \mathbb{R}^{n_h}, b \in \mathbb{R}$ and $\gamma \in \mathbb{R}_0^+$ is the regularization parameter. By using Lagrange multipliers, the solution of (3) can be obtained by taking the Karush-Kuhn-Tucker conditions for optimality. The result is given by the following linear system in the dual variables α:

$$\left[\begin{array}{c|c} 0 & 1_n^T \\ \hline 1_n & \Omega + \frac{1}{\gamma} I_n \end{array}\right] \left[\begin{array}{c} b \\ \hline \alpha \end{array}\right] = \left[\begin{array}{c} 0 \\ \hline Y \end{array}\right] \tag{4}$$

with $Y = (Y_1, \ldots, Y_n)^T$, $1_n = (1, \ldots, 1)^T$, $\alpha = (\alpha_1, \ldots, \alpha_n)^T$, and $\Omega_{il} = \varphi(X_i)^T \varphi(X_l) = K(X_i, X_l)$ for $i, l = 1, \ldots, n$ with $K(\cdot, \cdot)$ a positive definite kernel. Based on Mercer's theorem, the resulting LS-SVM model for function estimation becomes

$$\hat{m}(x) = \sum_{i=1}^{n} \hat{\alpha}_i K(x, X_i) + \hat{b} \tag{5}$$

where $K : \mathbb{R}^d \times \mathbb{R}^d \to \mathbb{R}$, d is the number of dimensions. For example, the radial basis function (RBF) kernel $K(X_i, X_j) = \exp(-\|X_i - X_j\|_2^2 / h^2)$ with bandwidth $h > 0$. Other possibilities are the polynomial kernel $K(X_i, X_j) = (X_i^T X_j + \tau)^p$ of degree p with $\tau \geq 0$ and linear kernel $K(X_i, X_j) = X_i^T X_j$.

By noticing that LS-SVM is a linear smoother, suitable bias and variance formulations can be found.

Definition 1 (Linear Smoother): An estimator $\hat{m}$ of m is a linear smoother if, for each $x \in \mathbb{R}^d$, there exists a vector $L(x) = (l_1(x), \ldots, l_n(x))^T \in \mathbb{R}^n$ such that

$$\hat{m}(x) = \sum_{i=1}^{n} l_i(x) Y_i \tag{6}$$

where $\hat{m}(\cdot) : \mathbb{R}^d \to \mathbb{R}$.

On training data, (6) can be written in matrix form as $\hat{m} = LY$, where $\hat{m} = (\hat{m}(X_1), \ldots, \hat{m}(X_n))^T \in \mathbb{R}^n$, and $L \in \mathbb{R}^{n \times n}$ is a smoother matrix whose ith row is $L(X_i)^T$, thus $L_{ij} = l_j(X_i)$. The entries of the ith row show the weights given to each Y_i in forming the estimate $\hat{m}(X_i)$. Thus we have

Theorem 1: The LS-SVM estimate (5) is

$$\hat{m}(x) = \sum_{i=1}^{n} l_i(x) Y_i$$

[1] $m \in C^z(\mathbb{R})$ stands for m is z-times continuously differentiable in $\mathbb{R}$

and $L(x) = (l_1(x), \ldots, l_n(x))^T$ is the smoother vector

$$L(x) = \left[\Omega_x^{\star T} \left(Z^{-1} - Z^{-1} \frac{J_n}{c} Z^{-1} \right) + \frac{J_1^T}{c} Z^{-1} \right]^T \quad (7)$$

with $\Omega_x^\star = (K(x, X_1), \ldots, K(x, X_n))^T$ the kernel vector evaluated at point x, $c = 1_n^T (\Omega + (I_n/\gamma))^{-1} 1_n$, $Z = \Omega + (I_n/\gamma)$, J_n a square matrix with all elements equal to 1 and $J_1 = (1, \ldots, 1)^T$.

Then the estimator, under model (1), has conditional mean

$$\mathbf{E}[\hat{m}(x)|X = x] = \sum_{i=1}^{n} l_i(x) m(x_i)$$

and conditional variance

$$\mathbf{Var}[\hat{m}(x)|X = x] = \sum_{i=1}^{n} l_i(x)^2 \sigma^2(x_i). \quad (8)$$

Proof: see Appendix I. ■

Note that one should not confuse linear smoothers, i.e., smoothers of the form (6), with linear regression, in which one assumes that the regression function m is linear.

B. Bias Estimation

Using Theorem 1, the conditional bias can be written as

$$\mathbf{E}[\hat{m}(x)|X = x] - m(x) = \sum_{i=1}^{n} l_i(x) m(x_i) - m(x).$$

It can be shown, in the 1-D case, by using a Taylor series expansion around the fitting point $x \in \mathbb{R}$, that for $|x_i - x| \leq h$ the conditional bias is equal to

$$\mathbf{E}[\hat{m}(x)|X = x] - m(x) = m'(x) \sum_{i=1}^{n} (x_i - x) l_i(x) + \frac{m''(x)}{2} \sum_{i=1}^{n} (x_i - x)^2 l_i(x) + O(\rho(h, \gamma))$$

where ρ is an unknown function describing the relation between the two tuning parameters.

Although the above expression gives insight on how the conditional bias behaves asymptotically, it is quite hard to use this in practice since it involves the estimation of first- and second-order derivatives of the unknown m. In fact, this procedure can be rather complicated in the multivariate case, especially when estimating derivatives.

Therefore, we opt for a procedure that does not rely completely on the asymptotic expression, but stays "closer" to the exact expression for the conditional bias. As a result, this will carry more information about the finite sample bias.

Theorem 2: Let $L(x)$ be the smoother vector evaluated in a point x and denote $\hat{m} = (\hat{m}(X_1), \ldots, \hat{m}(X_n))^T$. Then, the estimated conditional bias for LS-SVM is given by

$$\widehat{\text{bias}}[\hat{m}(x)|X = x] = L(x)^T \hat{m} - \hat{m}(x). \quad (9)$$

Proof: see Appendix II. ■

The above rationale is also closely related to the technique of double smoothing, see [25] and iterative bias reduction.

C. Variance Estimation

Our goal is to derive a fully automated procedure to estimate the variance function $\sigma^2(\cdot)$. Due to the simple decomposition $\sigma^2(x) = \mathbf{Var}[Y|X = x] = \mathbf{E}[Y^2|X = x] - \{\mathbf{E}[Y|X = x]\}^2$, we tend to use the following obvious and direct estimator [26]:

$$\hat{\sigma}_d^2(x) = \mathbf{E}[Y^2|X = x] - \{\hat{m}(x)\}^2.$$

In fact, there are some drawbacks in using such an estimator. For example, $\hat{\sigma}_d^2(x)$ is not always nonnegative due to estimation error, especially if different smoothing parameters are used in estimating the regression function and $\mathbf{E}[Y^2|X]$. Furthermore, such a method can result in a very large bias [27]. Before stating the estimator, we first need a condition on the weights of the LS-SVM (Lemma 1). The resulting variance estimator is given in Theorem 3.

Lemma 1: The weights $\{l_i(x)\}$ of the LS-SVM smoother are normal

$$\sum_{i=1}^{n} l_i(x) = 1.$$

Proof: see Appendix III. ■

Theorem 3 (Variance Estimator): Assume model (1), and let L denote the smoother matrix corresponding to the initial smooth. Let $S \in \mathbb{R}^{n \times n}$ denote the smoother matrix corresponding to a natural means of estimating $\sigma^2(\cdot)$ based on smoothing the squared residuals. Denote by $S(x)$ the smoother vector in an arbitrary point x, where $S(\cdot) : \mathbb{R}^d \to \mathbb{R}^n$. Assume that S preserves constant vectors, i.e., $S1_n = 1_n$, then an estimator for the variance function evaluated at a point x is given by

$$\hat{\sigma}^2(x) = \frac{S(x)^T \operatorname{diag}(\hat{\varepsilon}\hat{\varepsilon}^T)}{1 + S(x)^T \operatorname{diag}(LL^T - L - L^T)} \quad (10)$$

where $\hat{\varepsilon}$ denote the residuals and $\operatorname{diag}(A)$ is the column vector containing the diagonal entries of the square matrix A.

Proof: see Appendix III. ■

The class of variance function estimators (10) can be viewed as a generalization of those commonly used in parametric modeling (see [28]).

We next approximate the conditional variance of LS-SVM in (8). Given the estimator of the error variance function (10), an estimate of the conditional variance of LS-SVM with heteroscedastic errors (8) is given by

$$\mathbf{Var}[\hat{m}(x)|X = x] = L(x)^T \hat{\Sigma}^2 L(x) \quad (11)$$

with $\hat{\Sigma}^2 = \operatorname{diag}(\hat{\sigma}^2(X_1), \ldots, \hat{\sigma}^2(X_n))$.

III. Confidence and Prediction Intervals

A. Pointwise Confidence Intervals

The estimated bias (9) and variance (11) can be used to construct pointwise confidence intervals. Under certain regularity conditions [29], the central limit theorem for linear smoothers is valid and one can show asymptotically

$$\frac{\hat{m}(x) - \mathbf{E}[\hat{m}(x)|X = x]}{\sqrt{\mathbf{Var}[\hat{m}(x)|X = x]}} \xrightarrow{D} \mathcal{N}(0, 1)$$

where $\xrightarrow{D}$ denotes convergence in distribution. If the estimate is conditionally unbiased, i.e., $\mathbf{E}[\hat{m}(x)|X = x] = m(x)$,

approximate $100(1-\alpha)\%$ pointwise (in point x) confidence intervals may take the form

$$\hat{m}(x) \pm z_{1-\alpha/2}\sqrt{\mathbf{Var}[\hat{m}(x)|X=x]} \tag{12}$$

where $z_{1-\alpha/2}$ denotes the $(1-\alpha/2)$th quantile of the standard Gaussian distribution. Formally, the interval (12) is a confidence interval for $\mathbf{E}[\hat{m}(x)|X=x]$. It is a confidence interval for $m(x)$ under the assumption $\mathbf{E}[\hat{m}(x)|X=x]=m(x)$.

However, since LS-SVM is a biased smoother, the interval (12) has to be adjusted to allow for bias. The exact bias is given by $\text{bias}[\hat{m}(x)|X=x]=\mathbf{E}[\hat{m}(x)|X=x]-m(x)$. Since the exact bias is unknown, this quantity can be estimated by (9). Therefore, a bias-corrected approximate $100(1-\alpha)\%$ pointwise confidence interval is given by

$$\hat{m}_c(x) \pm z_{1-\frac{\alpha}{2}}\sqrt{\mathbf{Var}[\hat{m}(x)|X=x]} \tag{13}$$

with $\hat{m}_c(x)=\hat{m}(x)-\widehat{\text{bias}}[\hat{m}(x)|X=x]$.

The reader can easily verify that bias-corrected approximate $100(1-\alpha)\%$ pointwise confidence intervals can also be obtained for the classical ridge regression case by replacing the LS-SVM smoother matrix by the one of the ridge regression estimators [30]. In a similar way as before, suitable bias and variance estimates can be computed.

B. Simultaneous Confidence Intervals

The confidence intervals presented so far in this paper are pointwise. For example, by looking at two pointwise confidence intervals in Fig. 1 (Fossil dataset [31]), we can make the following two statements *separately*:

1) $(0.70743, 0.70745)$ is an approximate 95% pointwise confidence interval for $m(105)$;
2) $(0.70741, 0.70744)$ is an approximate 95% pointwise confidence interval for $m(120)$.

However, as is well known in multiple comparison theory, it is wrong to state that $m(105)$ is contained in (0.70743, 0.70745) and simultaneously $m(120)$ is contained in (0.70741, 0.70744) with 95% confidence. Therefore, it is not correct to connect the pointwise confidence intervals to produce a band around the estimated function. In order to make these statements, we have to modify the interval to obtain simultaneous confidence intervals. Three major groups exist to modify the interval. Monte Carlo simulations, Bonferroni, Šidák or other type of corrections, and upcrossing theory [32], [33]. The latter is also well known in GPs (see [34]).

Although Monte Carlo-based modifications are accurate (even when the number of data points n is relatively small), they are computationally expensive. Therefore, we will not discuss this type of methods in this paper. The interested readers can refer to [31] and reference therein.

Šidák and Bonferroni corrections are one of the most popular methods since they are easy to calculate and produce quite acceptable results. In what follows, the rationale behind the Šidák correction (generalization of Bonferroni) is elucidated. This correction is derived by assuming that individual tests are independent. Let the significance threshold for each test be β (significance level of pointwise confidence interval), then the probability that at least one of the tests is significant under this threshold is (1 – the probability that none of them is significant). Since we are assuming that they are independent, the probability that all of them are not significant is the product of the probabilities that each of them is not significant, or $1-(1-\beta)^n$. Now we want this probability to be equal to α, which is the significance level for the entire series of tests (or simultaneous confidence interval). By solving for β, we get $\beta=1-(1-\alpha)^{1/n}$. To obtain an approximate $100(1-\alpha)\%$ simultaneous confidence intervals, based on a Šidák correction, (13) has to be modified as

$$\hat{m}_c(x) \pm z_{1-\beta/2}\sqrt{\mathbf{Var}[\hat{m}(x)|X=x]}$$

with $\hat{m}_c(x)=\hat{m}(x)-\widehat{\text{bias}}[\hat{m}(x)|X=x]$.

The last method analytically approximates the modification of the interval. As a result, an approximate $100(1-\alpha)\%$ simultaneous confidence interval will be of the form

$$\hat{m}_c(x) \pm \nu_{1-\alpha}\sqrt{\mathbf{Var}[\hat{m}(x)|X=x]}. \tag{14}$$

The value for $\nu_{1-\alpha}$ is given by [35]

$$\nu_{1-\alpha}=\sqrt{2\log\left(\frac{\kappa_0}{\alpha\pi}\right)} \tag{15}$$

where

$$\kappa_0=\int_{\mathcal{X}}\frac{\sqrt{\|L(x)\|^2\|L'(x)\|^2-(L(x)^TL'(x))^2}}{\|L(x)\|^2}\,dx$$

where $\mathcal{X}$ is the set of x values of interest and $L'(x)=(d/dx)L(x)$, with the differentiation applied elementwise. Computation of κ_0 requires numerical integration and is slow in multiple dimensions or at small bandwidths h. It is shown in [32] that κ_0 is strongly related to the degrees of freedom of the fit. A good approximation of κ_0 results in

$$\kappa_0 \approx \frac{\pi}{2}(\text{trace}(L)-1).$$

For the Fossil dataset ($n=106$), set $\alpha=0.05$, then $z_{1-\alpha/2}=1.96$ and $z_{1-\beta/2}=3.49$. The simultaneous intervals obtained by using a Šidák correction are about 1.78 (= 3.49/1.96) times wider than the pointwise intervals. In fact, using Monte Carlo simulations [or upcrossing theory (15)] resulted in a factor 3.2 (3.13) instead of 3.49 obtained

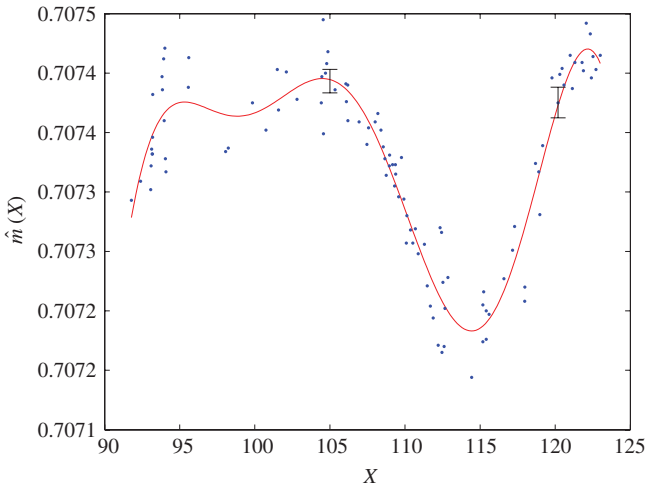


Fig. 1. Fossil data with two pointwise 95% confidence intervals.

Algorithm 1 Confidence Intervals

1: Given the training data $\{(X_1, Y_1), \ldots, (X_n, Y_n)\}$, calculate $\hat{m}$ on training data using (5)
2: Calculate the bias by using (9) on training data
3: Calculate residuals $\hat{\varepsilon}_k = Y_k - \hat{m}(X_k), k = 1, \ldots, n$
4: Calculate the variance of the LS-SVM by using (10) and (11)
5: Set significance level e.g. $\alpha = 0.05$
6: For pointwise confidence intervals: use (13). For simultaneous confidence intervals: use (15) and (14).

Algorithm 2 Prediction Intervals

1: Given the training data $\{(X_1, Y_1), \ldots, (X_n, Y_n)\}$ and test data $\{(X_{n+1}, Y_{n+1}), \ldots, (X_m, Y_m)\}$, calculate $\hat{m}(X_k)$ for $k = n+1, \ldots, m$ using (5)
2: Calculate the bias by using (9) on test data
3: Calculate residuals $\hat{\varepsilon}_k = Y_k - \hat{m}(X_k), k = 1, \ldots, n$
4: Calculate the variance of the LS-SVM by using (10) and (11) on test data
5: Set significance level e.g. $\alpha = 0.05$
6: For pointwise prediction intervals: use (16). For simultaneous prediction intervals: use (15) and (17).

by a Šidák correction. This is the reason why Šidák (and also Bonferroni) corrections result in conservative confidence intervals. In the rest of this paper, (15) will be used to determine the modification of the interval.

We conclude this paragraph with a final remark regarding the difference between confidence intervals and error bars. Although the latter are very popular, they differ from confidence intervals. Statistically, pointwise confidence intervals are defined according to (2) and a bound g_α has to be found. Also, the constructed confidence intervals must deal with estimator bias to ensure proper coverage rate and centering of the interval. On the other hand, error bars do not take this bias into account. They only give an idea of the variability of the estimator $\hat{m}$ and do not provide a solution to (2).

C. Prediction Intervals

In some cases, one may also be interested in the uncertainty on the prediction for a new observation. This type of requirement is fulfilled by the construction of a prediction interval. Assume that the new observation $Y^\star$ at a point $x^\star$ is independent of the estimation data. Then

$$\begin{aligned}\mathbf{Var}[Y^\star - \hat{m}(x^\star)|X{=}x^\star] &= \mathbf{Var}[Y^\star|X{=}x^\star] + \mathbf{Var}[\hat{m}(x^\star)|X{=}x^\star] \\ &= \sigma^2(x^\star) + \sum_{i=1}^{n} l_i(x^\star)^2 \sigma^2(x_i).\end{aligned}$$

In the last step, we have used the model assumptions (1) and (8). Thus, an approximate pointwise $100(1-\alpha)\%$ prediction interval in a new point $x^\star$ is constructed by

$$\hat{m}_c(x^\star) \pm z_{1-\frac{\alpha}{2}} \sqrt{\hat{\sigma}^2(x^\star) + \mathbf{Var}[\hat{m}(x^\star)|X = x^\star]} \qquad (16)$$

with $\hat{m}_c(x^\star) = \hat{m}(x^\star) - \widehat{\text{bias}}[\hat{m}(x^\star)|X = x^\star]$.

Algorithm 3 Bootstrap-based on residuals (homoscedasticity)

1: Given the data $\{(X_1, Y_1), \ldots, (X_n, Y_n)\}$, calculate $\hat{m}(X_k)$ using (5)
2: Calculate residuals $\hat{\varepsilon}_k = Y_k - \hat{m}(X_k)$
3: Re-center residuals $\tilde{\varepsilon}_k = \hat{\varepsilon}_k - n^{-1} \sum_{j=1}^{n} \hat{\varepsilon}_j$
4: Generate bootstrap samples $\{\tilde{\varepsilon}_k^\star\}_{k=1}^n$ by uniform sampling with replacement from $\{\tilde{\varepsilon}_k\}_{k=1}^n$
5: Generate $\{Y_k^\star\}_{k=1}^n$ from $Y_k^\star = \hat{m}(X_k) + \tilde{\varepsilon}_k^\star$
6: Calculate $\hat{m}^\star$ from $\{(X_1, Y_1^\star), \ldots, (X_n, Y_n^\star)\}$
7: Repeat steps 4–6 B times

Algorithm 4 Bootstrap-based on residuals (heteroscedasticity)

1: Given the data $\{(X_1, Y_1), \ldots, (X_n, Y_n)\}$, calculate $\hat{m}(X_k)$ using (5)
2: Calculate residuals $\hat{\varepsilon}_k = Y_k - \hat{m}(X_k)$
3: Re-center residuals $\tilde{\varepsilon}_k = \hat{\varepsilon}_k - n^{-1} \sum_{j=1}^{n} \hat{\varepsilon}_j$
4: Generate bootstrap data $\tilde{\varepsilon}_k^\star = \tilde{\varepsilon}_k \eta_k$ where η_k are Rademacher variables, defined as

$$\eta_k = \begin{cases} 1, & \text{with probability } 1/2; \\ -1, & \text{with probability } 1/2. \end{cases}$$

5: Generate $\{Y_k^\star\}_{k=1}^n$ from $Y_k^\star = \hat{m}(X_k) + \tilde{\varepsilon}_k^\star$
6: Calculate $\hat{m}^\star$ from $\{(X_1, Y_1^\star), \ldots, (X_n, Y_n^\star)\}$
7: Repeat steps 4–6 B times

As before, an approximate simultaneous $100(1-\alpha)\%$ prediction interval in a new point $x^\star$ is given by

$$\hat{m}_c(x^\star) \pm v_{1-\alpha} \sqrt{\hat{\sigma}^2(x^\star) + \mathbf{Var}[\hat{m}(x^\star)|X = x^\star]}. \qquad (17)$$

We conclude this section by summarizing the construction of confidence intervals and prediction intervals given in Algorithms 1 and 2, respectively.

IV. Bootstrap-Based Confidence and Prediction Intervals

In this section we will briefly review the current state-of-the-art regarding bootstrap-based confidence and prediction intervals, which are used for comparison in the experimental section.

A. Bootstrap-Based on Residuals

It is shown in [36] that the standard bootstrap [37] based on residuals does not work for nonparametric heteroscedastic regression models. A technique used to overcome this difficulty is the wild or external bootstrap, developed in [38] following suggestions in [39] and [40]. Further theoretical refinements are found in [41]. Algorithms 3 and 4 have to be applied when the errors are homoscedastic and heteroscedastic, respectively.

Other possibilities for the two-point distribution of the η_k also exist (see [38]). The Rademacher distribution was chosen because it was empirically shown in [42] that this distribution came out as best among six alternatives.

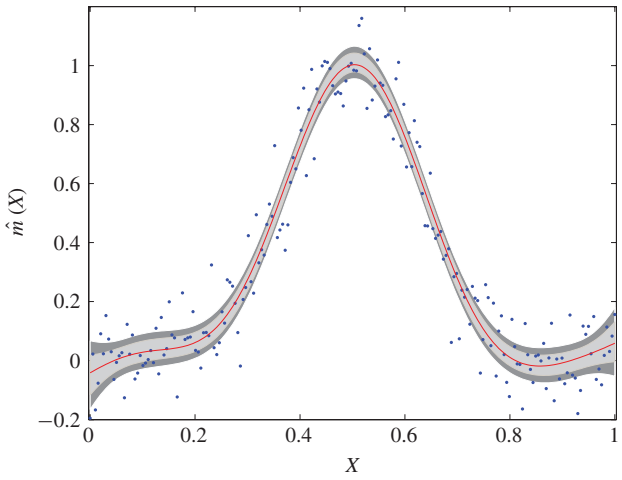


Fig. 2. Pointwise and simultaneous 95% confidence intervals. The outer (inner) region corresponds to simultaneous (pointwise) confidence intervals. The full line (in the middle) is the estimated LS-SVM model. For illustration purposes, the 95% pointwise confidence intervals are connected.

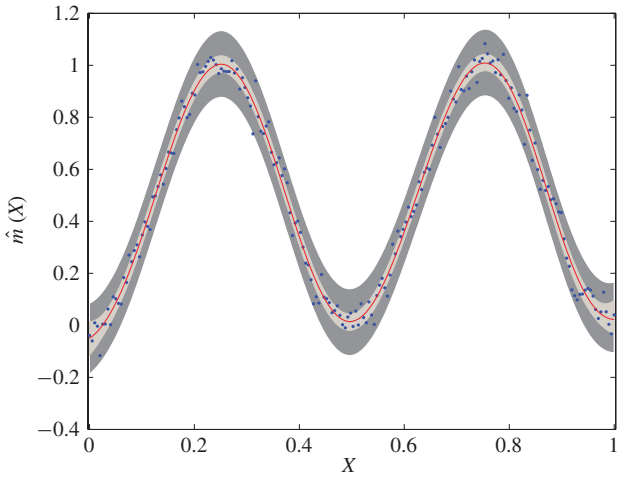


Fig. 3. Simultaneous 95% confidence and prediction intervals. The outer (inner) region corresponds to simultaneous prediction (confidence) intervals. The full line (in the middle) is the estimated LS-SVM model.

B. Construction of Confidence and Prediction Intervals

The construction of confidence and prediction intervals for nonparametric function estimation falls into two parts, i.e., the construction of a confidence or a prediction interval based on a pivotal method for the expected value of the estimator, and bias correction through undersmoothing. Then, a confidence interval is constructed by using the asymptotic distribution of a pivotal statistic. The latter can be obtained by bootstrap. Before illustrating the construction of intervals based on bootstrap, we give a formal definition of a key quantity used in the bootstrap approach.

Definition 2 (Pivotal Quantity): Let $X = (X_1, \ldots, X_n)$ be random variables with unknown joint distribution F and denote by $T(F)$ a real-valued parameter of interest (e.g., the regression function). A random variable $\mathcal{T}(X, T(F))$ is a pivotal quantity (or pivot) if the distribution of $\mathcal{T}(X, T(F))$ is independent of all parameters.

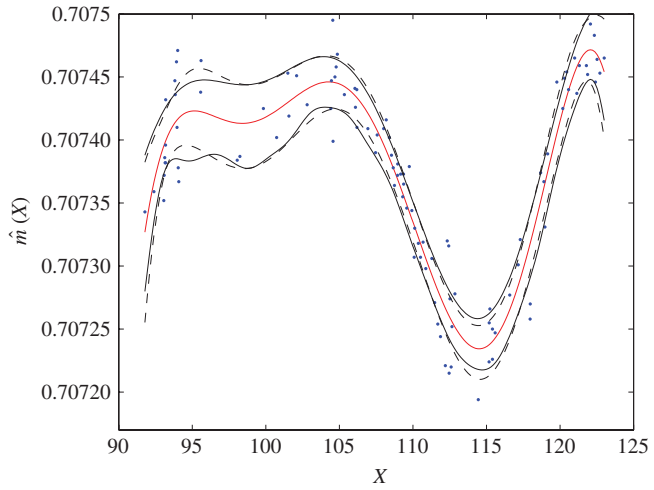


Fig. 4. Simultaneous 95% confidence intervals for the Fossil dataset. The dashed lines correspond to the proposed simultaneous confidence intervals and the full lines are the bootstrap confidence intervals. The full line (in the middle) is the estimated LS-SVM model.

Hall [4] suggested the following approach: estimate the distribution of the pivot

$$\mathcal{T}(m(x), \hat{m}(x)) = \frac{\hat{m}(x) - m(x)}{\sqrt{\mathbf{Var}[\hat{m}(x)|X = x]}}$$

by the bootstrap. Depending on homoscedastic or heteroscedastic errors, Algorithms 3 or 4 should be used to estimate this distribution. Now, the distribution of the pivotal statistic $\mathcal{T}(m(x), \hat{m}(x))$ is approximated by the corresponding distribution of the bootstrapped statistic

$$\mathcal{V}(\hat{m}^\star(x), \hat{m}_g(x)) = \frac{\hat{m}^\star(x) - \hat{m}_g(x)}{\sqrt{\mathbf{Var}^\star[\hat{m}(x)|X = x]}}$$

where $\hat{m}_g$ denotes the undersmoother with bandwidth g and $\star$ denote bootstrap counterparts. Practically, we choose bandwidth $g = h/2$. Hence, a $100(1-\alpha)\%$ pointwise confidence interval is given by

$$\left[\Psi_{\frac{\alpha}{2}}, \Psi_{1-\frac{\alpha}{2}}\right]$$

with

$$\Psi_{\frac{\alpha}{2}} = \hat{m}(x) + Q_{\frac{\alpha}{2}}\sqrt{\mathbf{Var}[\hat{m}(x)|X = x]}$$

and

$$\Psi_{1-\frac{\alpha}{2}} = \hat{m}(x) + Q_{1-\frac{\alpha}{2}}\sqrt{\mathbf{Var}[\hat{m}(x)|X = x]}.$$

Q_α denotes the αth quantile of the bootstrap distribution of the pivotal statistic. A $100(1-\alpha)\%$ simultaneous confidence interval can be constructed by applying a Šidák correction, see Section III. Similar to the workflow in Section III, $100(1-\alpha)\%$ pointwise and simultaneous prediction intervals are obtained.

A question that remains unanswered is how to determine B, the number of bootstrap replications in Algorithms 3 and 4. The construction of confidence (and prediction) intervals demands accurate information of the low and high quantiles of the limit distribution. Therefore, enough resamplings are needed in order for bootstrap to accurately reproduce this distribution. Typically, B is chosen in the range of 1.000–2.000 for pointwise intervals and more than 10.000 for simultaneous intervals.

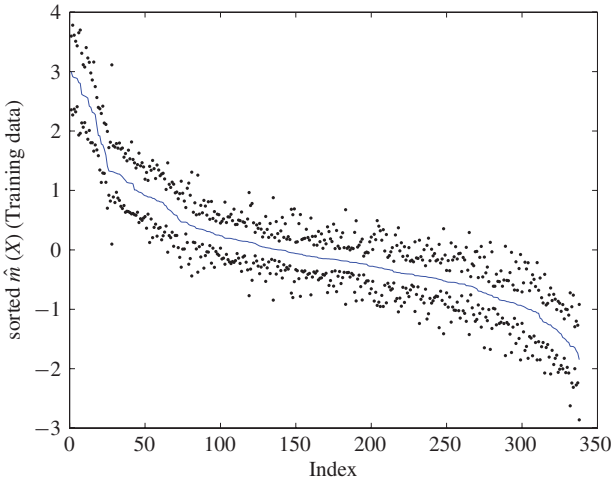


Fig. 5. Simultaneous 95% confidence intervals for the Boston housing dataset (dots). Sorted outputs are plotted against their index.

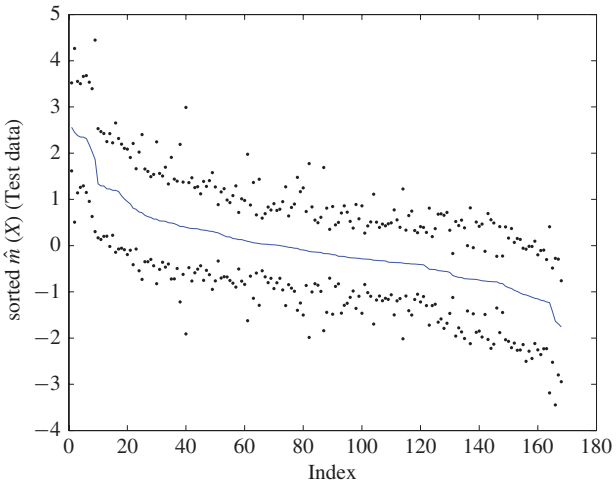


Fig. 6. Simultaneous 95% prediction intervals for the Boston housing dataset (dots). Sorted outputs are plotted against their index.

V. Examples

In all simulations, the RBF kernel was used and $\alpha = 0.05$. The tuning parameters (regularization parameter γ and kernel bandwidth h) of the LS-SVM were obtained via leave-one-out cross validation.

A. Homoscedastic Examples

In the first example, data were generated from (1) using normal errors and following the regression curve:

$$m(x) = e^{-32(x-0.5)^2}.$$

The sample size is taken to be $n = 200$, and $\sigma(x) = \sigma = 0.1$. Pointwise and simultaneous 95% confidence intervals are shown in Fig. 2. The line in the middle represents the LS-SVM model. For illustration purposes, the 95% pointwise confidence intervals are connected.

In a second example, we generate data from (1) following the regression curve (normal errors with $\sigma = 0.05$)

$$m(x) = \sin^2(2\pi x).$$

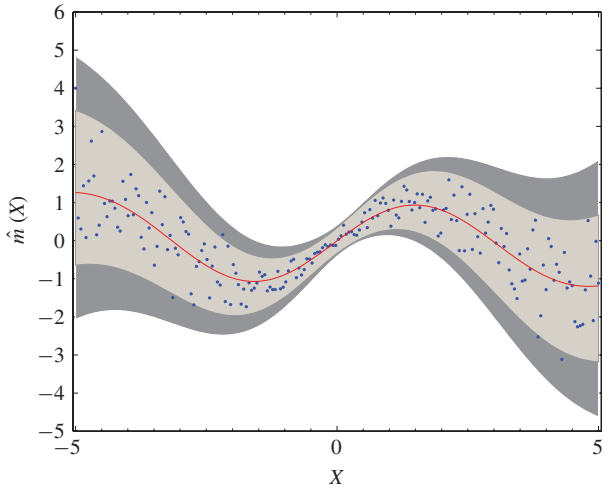


Fig. 7. Pointwise and simultaneous 95% prediction intervals for heteroscedastic errors. The outer (inner) region corresponds to simultaneous (pointwise) prediction intervals. The full line (in the middle) is the estimated LS-SVM model. For illustration purposes, the 95% pointwise prediction intervals are connected.

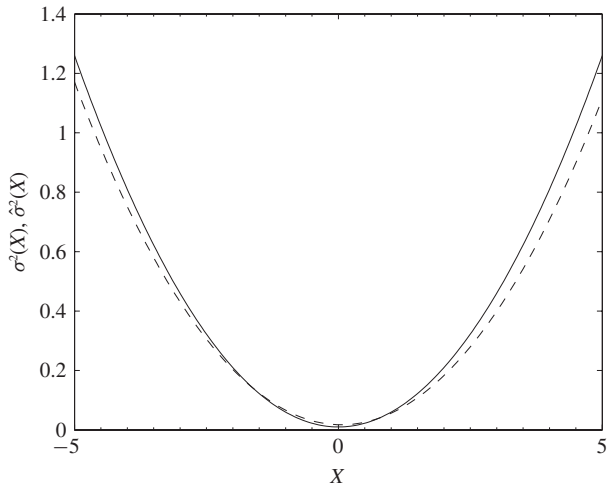


Fig. 8. Variance function estimation. The full line represents the real variance function and the dashed line is the estimated variance function obtained by Theorem 3.

Fig. 3 illustrates the 95% simultaneous confidence and prediction intervals. The outer (inner) region corresponds to the prediction (confidence) interval.

In a third example, we compare the proposed simultaneous confidence intervals with the bootstrap method, see Section IV, on the Fossil dataset. The number of bootstrap replications were $B = 15.000$. From Fig. 4, it is clear that both methods produce similar confidence intervals. Although the proposed method is based on an asymptotic result (central limit theorem for smoothers) and the number of data points is small ($n = 106$), it produces good confidence intervals close the ones obtained by bootstrap. It can be expected that, when n is very small, the bootstrap-based confidence intervals will be more accurate than the proposed confidence intervals since the former reconstructs the limit distribution from the given data.

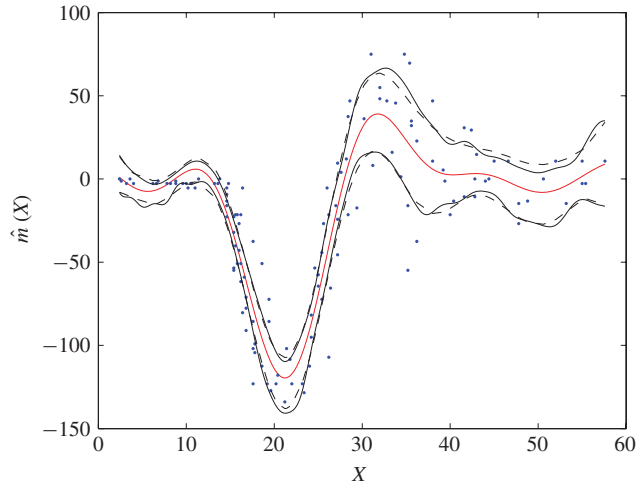


Fig. 9. Simultaneous 95% confidence intervals for the Motorcycle dataset. The dashed lines correspond to the proposed simultaneous confidence intervals and the full lines are the bootstrap confidence intervals. The full line (in the middle) is the estimated LS-SVM model.

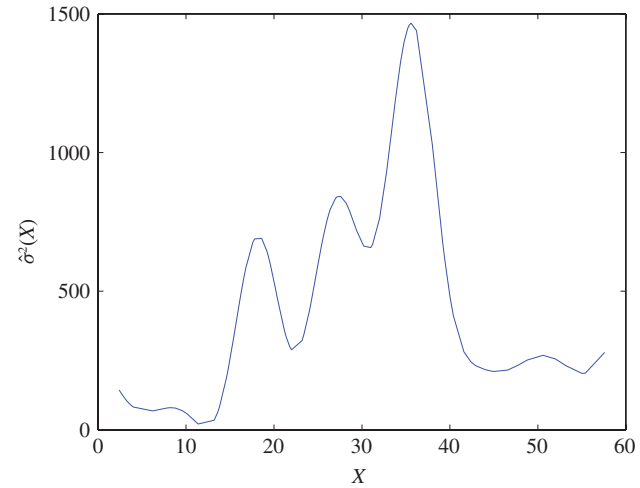


Fig. 10. Variance function estimation of the Motorcycle dataset obtained by Theorem 3.

As a last example, consider the Boston housing dataset (multivariate example). We selected randomly 338 training data points and 168 test data points. The corresponding simultaneous confidence and prediction intervals are shown in Figs. 5 and 6, respectively. The outputs on training and test data are sorted and plotted against their corresponding index. Also, the respective intervals are sorted accordingly.

B. Heteroscedastic Examples

The data is generated according to the following model (with normal errors):

$$Y_k = \sin(x_k) + \sqrt{0.05x_k^2 + 0.01}\ \varepsilon_k, \quad k = 1, \ldots, 200$$

where the x_k are equally spaced over the interval $[-5, 5]$. Figs. 7 and 8 show the 95% pointwise and simultaneous prediction intervals for this model and the estimated (and true) variance function, respectively. The variance function was obtained by Theorem 3. The latter clearly demonstrates the capability of the proposed methodology for variance estimation.

As a last example, consider the Motorcycle dataset. We compare the proposed simultaneous confidence intervals with the wild bootstrap method, see Section IV. The number of bootstrap replications was $B = 15.000$. The result is given in Fig. 9. As before, both intervals are very close to each other. Fig. 10 shows the estimated variance function of this data set.

VI. Conclusion

In this paper, we have studied the properties of data-driven confidence bands for kernel-based regression, more specifically for LS-SVM in the regression context. We have illustrated how to calculate a bias estimate for LS-SVM without computing higher order derivatives. Also, we proposed a simple way to obtain the variance function if the errors are heteroscedastic. These two estimates can be combined to obtain approximate $100(1-\alpha)\%$ pointwise and simultaneous confidence and prediction intervals. In order to correspond with multiple comparison theory, a Šidák correction and a more involved result from upcrossing theory were used to construct approximate $100(1-\alpha)\%$ simultaneous confidence and prediction intervals. Furthermore, we compared our method with a state-of-the-art bootstrap-based method. From this simulation, it is clear that the two produce similar intervals. However, it can be expected, when the number of data points is small, that the intervals based on bootstrap will be more accurate than the proposed confidence (prediction) intervals since the former reconstructs the limit distribution from the given data and does not rely on asymptotic results.

Appendix I
Proof of Theorem 1

In matrix form, the resulting LS-SVM model (5) on training data is given by

$$\hat{m} = \Omega\hat{\alpha} + 1_n\hat{b}$$

with

$$\hat{\alpha} = \left(\Omega + \frac{I_n}{\gamma}\right)^{-1}(Y - 1_n\hat{b})$$

and

$$\hat{b} = \frac{1_n^T\left(\Omega + \frac{I_n}{\gamma}\right)^{-1}Y}{1_n^T\left(\Omega + \frac{I_n}{\gamma}\right)^{-1}1_n}.$$

Plugging this in to the above expression results in

$$\begin{aligned}\hat{m} &= \left[\Omega\left(Z^{-1} - Z^{-1}\frac{J_n}{c}Z^{-1}\right) + \frac{J_n}{c}Z^{-1}\right]Y \\ &= LY\end{aligned}$$

with $c = 1_n^T\left(\Omega + (I_n/\gamma)\right)^{-1}1_n$, $Z = \Omega + (I_n/\gamma)$, and J_n is a square matrix with all elements equal to 1.

The above derivation is valid when all points x are considered as training data. However, evaluating the LS-SVM at an arbitrary point x can be written as

$$\begin{aligned}\hat{m}(x) &= \Omega_x^{\star T}\hat{\alpha} + 1_n\hat{b} \\ &= \left[\Omega_x^{\star T}\left(Z^{-1} - Z^{-1}\frac{J_n}{c}Z^{-1}\right) + \frac{J_1^T}{c}Z^{-1}\right]Y \\ &= L(x)^T Y\end{aligned}$$

with $\Omega_x^\star = (K(x, X_1), \ldots, K(x, X_n))^T$ the kernel vector evaluated in point $x^\star$ and $J_1 = (1, \ldots, 1)^T$.

The conditional mean and conditional variance of the LS-SVM can then be derived as follows:

$$\begin{aligned}\mathbf{E}[\hat{m}(x)|X = x] &= \sum_{i=1}^n l_i(x)\,\mathbf{E}[Y_i|X = x_i] \\ &= \sum_{i=1}^n l_i(x)m(x_i)\end{aligned}$$

and

$$\begin{aligned}\mathbf{Var}[\hat{m}(x)|X = x] &= \sum_{i=1}^n l_i(x)^2\,\mathbf{Var}[Y_i|X = x_i] \\ &= \sum_{i=1}^n l_i(x)^2\sigma^2(x_i).\end{aligned}$$

Appendix II
Proof of Theorem 2

The exact conditional bias for LS-SVM is given by (in matrix form)

$$\mathbf{E}[\hat{m}|X] - m = (L - I_n)m$$

where $m = (m(X_1), \ldots, m(X_n))^T$ and $\hat{m} = LY$. Observe that the residuals are given by

$$\hat{\varepsilon} = Y - \hat{m} = (I_n - L)Y.$$

Taking expectations yields

$$\begin{aligned}\mathbf{E}[\hat{\varepsilon}|X] &= m - Lm \\ &= -\,\text{bias}[\hat{m}|X].\end{aligned}$$

This suggests estimating the conditional bias by smoothing the negative residuals

$$\begin{aligned}\widehat{\text{bias}}[\hat{m}|X] &= -L\hat{\varepsilon} \\ &= -L(I_n - L)Y \\ &= (L - I_n)\hat{m}.\end{aligned}$$

Therefore, evaluating the estimated conditional bias at a point x can be written as

$$\begin{aligned}\widehat{\text{bias}}[\hat{m}(x)|X = x] &= \sum_{i=1}^n l_i(x)m(x_i) - \hat{m}(x) \\ &= L(x)^T m - \hat{m}(x).\end{aligned}$$

Appendix III
Proof of Lemma 1 and Theorem 3

Before we prove Theorem 3, we prove Lemma 1. First we show that for all training data $L1_n = 1_n$

$$\begin{aligned}L1_n &= \left[\Omega\left(Z^{-1} - Z^{-1}\frac{J_n}{c}Z^{-1}\right) + \frac{J_n}{c}Z^{-1}\right]1_n \\ &= \Omega\left(Z^{-1} - Z^{-1}\frac{J_n}{c}Z^{-1}\right)1_n + \frac{J_n}{c}Z^{-1}1_n \\ &= \Omega Z^{-1}\left(1_n - \frac{J_n}{c}Z^{-1}1_n\right) + \frac{J_n}{c}Z^{-1}1_n.[4pt]\end{aligned}$$

It suffices to show that $(J_n/c)Z^{-1}1_n = 1_n$ to complete the proof

$$\frac{J_n}{c}Z^{-1}1_n = \frac{1_n 1_n^T\left(\Omega + \frac{I_n}{\gamma}\right)^{-1}1_n}{1_n^T\left(\Omega + \frac{I_n}{\gamma}\right)^{-1}1_n} = 1_n.$$

We can now formulate the result for any point x. Let $L(x)$ be the smoother vector in a point x, then

$$\begin{aligned}\sum_{i=1}^n l_i(x) &= L(x)^T 1_n \\ &= \left[\Omega_x^{\star T}\left(Z^{-1} - Z^{-1}\frac{J_n}{c}Z^{-1}\right) + \frac{J_1^T}{c}Z^{-1}\right]1_n.\end{aligned}$$

Similar to the derivation given above, we have to show that $(J_1^T/c)Z^{-1}1_n = 1$ to conclude the proof

$$\frac{J_1^T}{c}Z^{-1}1_n = \frac{1_n^T\left(\Omega + \frac{I_n}{\gamma}\right)^{-1}1_n}{1_n^T\left(\Omega + \frac{I_n}{\gamma}\right)^{-1}1_n} = 1.$$

Theorem 3 is proved as follows. Let $L \in \mathbb{R}^{n\times n}$ be the smoother matrix corresponding to an initial smooth of the data, and put

$$\hat{\varepsilon} = (I_n - L)Y$$

the vector of residuals. Then a natural means of estimating the variance function $\sigma^2(\cdot)$ is to smooth the squared residuals to obtain $S\,\text{diag}(\hat{\varepsilon}\hat{\varepsilon}^T)$. It is also reasonable that the estimator should be unbiased when the errors are homoscedastic. Thus, under homoscedasticity and $\Sigma = \mathbf{E}[(Y - m)^2|X]$ and $B_1 = \mathbf{E}[LY|X] - m$, we obtain

$$\begin{aligned}\mathbf{E}[S\text{diag}(\hat{\varepsilon}\hat{\varepsilon}^T)|X] &= S\,\mathbf{E}\left[\text{diag}\left\{(I_n - L)YY^T(I_n - L)^T\right\}|X\right] \\ &= S\left[\text{diag}\left\{(I_n - L)\,\mathbf{E}(YY^T|X)(I_n - L)^T\right\}\right] \\ &= S\left[\text{diag}\left\{(I_n - L)(mm^T + \Sigma)(I_n - L)^T\right\}\right] \\ &= S\left[\text{diag}\left(B_1B_1^T\right) + \sigma^2\text{diag}\left((I_n - L)(I_n - L)^T\right)\right] \\ &= S\left[\text{diag}\left(B_1B_1^T\right) + \sigma^2(1_n + \Delta)\right]\end{aligned}$$

where $\Delta = \text{diag}\left(LL^T - L - L^T\right)$. Since $\mathbf{E}[S\,\text{diag}(\hat{\varepsilon}\hat{\varepsilon}^T)|X] = \sigma^2(1_n + S\Delta)$ when LY is conditionally unbiased, i.e., $B_1 = 0$,

and using Lemma 1, this motivates the variance estimator at a point x

$$\hat{\sigma}^2(x) = \frac{S(x)^T \operatorname{diag}(\hat{\varepsilon}\hat{\varepsilon}^T)}{1 + S(x)^T \operatorname{diag}\left(LL^T - L - L^T\right)}$$

where $S(x)$ is the smoother matrix in an arbitrary point x [see also (7)].

REFERENCES

[1] C.-J. Ong, S. Shao, and J. Yang, "An improved algorithm for the solution of the regularization path of support vector machine," *IEEE Trans. Neural Netw.*, vol. 21, no. 3, pp. 451–462, Mar. 2010.

[2] Z. Sun, Z. Zhang, H. Wang, and M. Jiang, "Cutting plane method for continuously constrained kernel-based regression," *IEEE Trans. Neural Netw.*, vol. 21, no. 2, pp. 238–247, Feb. 2010.

[3] A. B. Tsybakov, *Introduction to Nonparametric Estimation*. New York: Springer-Verlag, 2009.

[4] P. Hall, "On bootstrap confidence intervals in nonparametric regression," *Ann. Statist.*, vol. 20, no. 2, pp. 695–711, Jun. 1992.

[5] G. Chryssoloiuris, M. Lee, and A. Ramsey, "Confidence interval prediction for neural network models," *IEEE Trans. Neural Netw.*, vol. 7, no. 1, pp. 229–232, Jan. 1996.

[6] G. Papadopoulos, P. J. Edwards, and A. F. Murray, "Confidence estimation methods for neural networks: A practical comparison," *IEEE Trans. Neural Netw.*, vol. 12, no. 6, pp. 1278–1287, Nov. 2001.

[7] I. Rivals and L. Personnaz, "Construction of confidence intervals for neural networks based on least squares estimation," *Neural Netw.*, vol. 13, nos. 4–5, pp. 463–484, Jun. 2000.

[8] C. M. Bishop and C. S. Qazaz, "Regression with input-dependent noise: A Bayesian treatment," in *Advances in Neural Information Processing Systems 9*. Cambridge, MA: MIT Press, 1997, pp. 347–353.

[9] P. W. Goldberg, C. K. Williams, and C. M. Bishop, "Regression with input-dependent noise: A Gaussian process treatment," in *Advances in Neural Information Processing Systems 10*. Cambridge, MA: MIT Press, 1998.

[10] K. Kersting, C. Plagemann, P. Pfaff, and W. Burgard, "Most likely heteroscedastic Gaussian process regression," in *Proc. 24th ICML*, Corvalis, OR, 2007, pp. 393–400.

[11] J. M. Bernardo and A. F. M. Smith, *Bayesian Theory*. New York: Wiley, 2000.

[12] G. C. Cawley, N. L. C. Talbot, R. J. Foxall, S. R. Dorling, and D. P. Mandic, "Heteroscedastic kernel ridge regression," *Neurocomputing*, vol. 57, pp. 105–124, Mar. 2004.

[13] G. C. Cawley, N. L. C. Talbot, and O. Chapelle, "Estimating predictive variances with kernel ridge regression," in *Machine Learning Challenges Workshop*. Berlin, Germany: Springer-Verlag, 2006, pp. 56–77.

[14] R. L. Eubank and P. L. Speckman, "Confidence bands in nonparametric regression," *J. Amer. Statist. Assoc.*, vol. 88, no. 424, pp. 1287–1301, Dec. 1993.

[15] P. J. Bickel and M. Rosenblatt, "On some global measures of the deviations of density function estimates," *Ann. Statist.*, vol. 1, no. 6, pp. 1071–1095, 1973.

[16] P. Hall and D. M. Titterington, "On confidence bands in nonparametric density estimation and regression," *J. Multivariate Anal.*, vol. 27, no. 1, pp. 228–254, Oct. 1998.

[17] W. Härdle, "Asymptotic maximal deviation of M-smoothers," *J. Multivariate Anal.*, vol. 29, no. 2, pp. 163–179, May 1989.

[18] G. J. Johnstone, "Probabilities of maximal deviations for nonparametric regression function estimates," *J. Multivariate Anal.*, vol. 12, no. 3, pp. 402–414, Sep. 1982.

[19] J. Fan and I. Gijbels, "Data-driven bandwidth selection in local polynomial fitting: Variable bandwidth and spatial adaptation," *J. R. Statist. Soc. B*, vol. 57, no. 2, pp. 371–394, 1995.

[20] J. Fan and I. Gijbels, *Local Polynomial Modelling and its Applications*. London, U.K.: Chapman & Hall, 1996.

[21] R. Isermann, *Fault-Diagnosis Systems: An Introduction from Fault Detection to Fault Tolerance*. New York: Springer-Verlag, 2005.

[22] M. Witczak, *Modelling and Estimation Strategies for Fault Diagnosis of Non-Linear Systems: From Analytical to Soft Computing Approaches* (Lecture Notes in Control and Information Sciences). New York: Springer-Verlag, 2007.

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

[24] V. N. Vapnik, *Statistical Learning Theory*. New York: Wiley, 1999.

[25] W. Härdle, P. Hall, and J. S. Marron, "Regression smoothing parameters that are not far from their optimum," *J. Amer. Statist. Assoc.*, vol. 87, no. 417, pp. 227–233, Mar. 1992.

[26] W. Härdle and A. Tsybakov, "Local polynomial estimators of the volatility function in nonparametric autoregression," *J. Econometrics*, vol. 81, no. 1, pp. 223–242, Nov. 1997.

[27] J. Fan and Q. Yao, "Efficient estimation of conditional variance functions in stochastic regression," *Biometrika*, vol. 85, no. 3, pp. 645–660, Sep. 1998.

[28] J. A. Rice, *Mathematical Statistics and Data Analysis*, 2nd ed. Pacific Grove, CA: Duxbury Press, 1995.

[29] A. N. Shiryaev, *Probability*, 2nd ed. New York: Springer, 1996.

[30] A. E. Hoerl and R. W. Kennard, "Ridge regression: Biased estimation for nonorthogonal problems," *Technometrics*, vol. 12, no. 1, pp. 55–67, Feb. 1970.

[31] D. Ruppert, M. P. Wand, and R. J. Carroll, *Semiparametric Regression*. Cambridge, U.K.: Cambridge Univ. Press, 2003.

[32] C. Loader, *Local Regression and Likelihood*. New York: Springer-Verlag, 1999.

[33] S. O. Rice, "The distribution of the maxima of a random curve," *Amer. J. Math.*, vol. 61, no. 2, pp. 409–416, Apr. 1939.

[34] J. Sun and C. R. Loader, "Simultaneous confidence bands for linear regression and smoothing," *Ann. Statist.*, vol. 22, no. 3, pp. 1328–1345, Sep. 1994.

[35] G. Knafl, J. Sacks, and D. Ylvisaker, "Confidence bands for regression functions," *J. Amer. Statist. Assoc.*, vol. 80, no. 391, pp. 683–691, Sep. 1985.

[36] W. Härdle, "Resampling for inference from curves," in *Proc. 47th Session Int. Statist. Inst.*, Bonn, Germany, 1989, pp. 59–69.

[37] B. Efron, "Bootstrap methods: Another look at the jackknife," *Ann. Statist.*, vol. 7, no. 1, pp. 1–26, 1979.

[38] R. Y. Liu, "Bootstrap procedure under some non-I.I.D. models," *Ann. Statist.*, vol. 16, no. 4, pp. 1696–1708, Dec. 1988.

[39] C. F. J. Wu, "Jackknife, bootstrap and other resampling methods in regression analysis," *Ann. Statist.*, vol. 14, no. 4, pp. 1261–1295, 1986.

[40] C. F. J. Wu, "Jackknife bootstrap and other resampling methods in regression analysis," *Ann. Statist.*, vol. 14, no. 4, pp. 1295–1298, 1986.

[41] E. Mammen, "Bootstrap and wild bootstrap for high dimensional linear models," *Ann. Statist.*, vol. 21, no. 1, pp. 255–285, Mar. 1993.

[42] J. Davidson, A. Monticini, and D. Peel, "Implementing the wild bootstrap using a two-point distribution," *Econ. Lett.*, vol. 96, no. 3, pp. 309–315, Sep. 2007.



Kris De Brabanter was born in Ninove, Belgium, on February 21, 1981. He received the Masters degree in electronic engineering from the Erasmus Hogeschool Brussel, Brussels, Belgium, in 2005, and the Masters degree in electrical engineering (data mining and automation) from the Katholieke Universteit Leuven (K. U. Leuven), Leuven, Belgium, in 2007. Currently, he is pursuing the Ph.D. degree in the SCD-SISTA Laboratory, Department of Electrical Engineering, K. U. Leuven.

His current research interests include nonparametric statistics and nonlinear systems.



Jos De Brabanter was born in Ninove, Belgium, on January 11, 1957. He received the Masters degree in electronic engineering in 1990, the Safety Engineer degree in 1992, the Master of Environment and Human Ecology degree in 1993, the Master of Artificial Intelligence degree in 1996, the Master of Statistics degree in 1997, and the Ph.D. degree in applied sciences in 2004, all from Katholieke Universiteit Leuven (K. U. Leuven), Leuven, Belgium.

He currently holds an Associated Docent position at K. U. Leuven. His current research interests include statistics and nonlinear systems.



Johan A. K. Suykens (M'02–SM'04) was born in Willebroek, Belgium, on May 18, 1966. He received the Masters degree in electro-mechanical engineering and the Ph.D. degree in applied sciences from the Katholieke Universiteit Leuven (K. U. Leuven), Leuven, Belgium, in 1989 and 1995, respectively.

He was a Visiting Post-Doctoral Researcher at the University of California, Berkeley, in 1996, through the Fund for Scientific Research FWO Flanders. He is currently a Professor (Hoogleraar) at K. U. Leuven. He is the author of *Artificial Neural Networks for Modelling and Control of Non-linear Systems* (Kluwer Academic Publishers) and *Least Squares Support Vector Machines* (World Scientific), a co-author of *Cellular Neural Networks, Multi-Scroll Chaos and Synchronization* (World Scientific), and the editor of *Nonlinear Modeling: Advanced Black Box Techniques* (Kluwer Academic Publishers) and *Advances in Learning Theory: Methods, Models and Applications* (IOS Press).

Dr. Suykens was an Associate Editor of the IEEE TRANSACTIONS ON CIRCUITS AND SYSTEMS from 1997 to 1999 and from 2004 to 2007, and the IEEE TRANSACTIONS ON NEURAL NETWORKS from 1998 to 2009. He received an IEEE Signal Processing Society Best Paper (Senior) Award in 1999, and several Best Paper Awards at International Conferences. He was a recipient of the International Neural Networks Society INNS Young Investigator Award in 2000 for significant contributions in the field of neural networks. He has served as a Director and Organizer of the North Atlantic Treaty Organization Advanced Study Institute on Learning Theory and Practice, Leuven, in 2002, as a Program Co-Chair for the International Joint Conference on Neural Networks in 2004, and the International Symposium on Nonlinear Theory and its Applications in 2005, and as an organizer of the International Symposium on Synchronization in Complex Networks in 2007.



Bart De Moor (M'86–SM'93–F'04) was born in Halle, Belgium, in 1960. He received the M.E. degree in electrical engineering and the Ph.D. degree in engineering from the Katholieke Universiteit Leuven (K. U. Leuven), Leuven, Belgium.

He spent two years as a Visiting Research Associate in the Department of Electrical Engineering (Information Systems Laboratory, under Prof. Kailath) and Computer Science (under Prof. Golub), Stanford University, Stanford, CA, from 1988 to 1990. Currently, he is a Full Professor in the Department of Electrical Engineering, K. U. Leuven, in the research group SCD. He is also the Chairman of the Industrial Research Fund, Hercules (heavy equipment funding in Flanders) and several other scientific and cultural organizations. He is on the board of six spinoff companies (IPCOS, Data4s, TMLeuven, Silicos, Dsquare, Cartagenia) of the Flemish Interuniversity Institute for Biotechnology the Study Center for Nuclear Energy, and of the Institute for Broadband Technology. He is leading a research group of 30 Ph.D. students and 8 post-doctoral fellows. In the recent past, he has guided 55 Ph.D. students.

Prof. De Moor was a recipient of the Leybold-Heraeus Prize in 1986, the Leslie Fox Prize in 1989, the Guillemin-Cauer Best Paper Award of the IEEE TRANSACTION ON CIRCUITS AND SYSTEMS in 1990, the Laureate of the Belgian Royal Academy of Sciences in 1992, the Biannual Siemens Award in 1994, the Best Paper Award of Automatica (the International Federation of Automatic Control) in 1996, and the IEEE Signal Processing Society Best Paper Award in 1999. He was a member of the Academic Council of K. U. Leuven and of its Research Policy Council.