Confidence Bands for Least Squares Support Vector Machine Classifiers: A Regression Approach

K. De Brabanter[a,b,*], P. Karsmakers[a,c], J. De Brabanter[a,b,d], J.A.K. Suykens[a,b], B. De Moor[a,b]

[a] *Department of Electrical Engineering ESAT-SCD, Katholieke Universiteit Leuven, Kasteelpark Arenberg 10, B-3001 Leuven, Belgium*
[b] *IBBT-K.U.Leuven Future Health Department, Kasteelpark Arenberg 10, B-3001 Leuven, Belgium*
[c] *K.H.Kempen (Associatie K.U.Leuven), Dep. IBW, Kleinhoefstraat 4, B-2440 Geel*
[d] *Hogeschool KaHo Sint-Lieven (Associatie K.U.Leuven), Departement I.I. - E&A, G. Desmetstraat 1, B-9000 Gent*

Abstract

This paper presents bias-corrected $100(1-\alpha)\%$ simultaneous confidence bands for least squares support vector machine classifiers based on a regression framework. The bias, which is inherently present in every nonparametric method, is estimated using double smoothing. In order to obtain simultaneous confidence bands we make use of the volume-of-tube formula. We also provide extensions of this formula in higher dimensions and show that the width of the bands are expanding with increasing dimensionality. Simulations and data analysis support its usefulness in practical real life classification problems.

Keywords: kernel based classification, bias, variance, linear smoother, higher-order kernel, simultaneous confidence intervals

1. Introduction

Nonparametric techniques e.g. regression estimators, classifiers,... are becoming more and more standard tools for data analysis [7, 28]. The popularity is mostly due to their ability to generalize well on new data and relative ease of implementation. Nonparametric classification, in particular Support Vector Machines (SVM) and Least Squares Support Vector Machines (LS-SVM), are widely known and used in many different application areas see e.g. [17, 21, 14]. On the one hand the methods gain in popularity, on the other hand the construction of intervals estimates such as confidence bands for regression and classification has been studied less frequently. Although, the practical implementation of these methods is straightforward, their statistical properties (bias and variance) are more difficult to obtain and to analyze than classical linear methods. For example, the construction of interval estimates is often troubled with the inevitable presence of bias accompanying these nonparametric methods [30].

The goal of this paper is to statistically investigate the bias-variance properties of LS-SVM for classification and its application in the construction of confidence bands. Therefore, the classification problem is written as a regression problem [27]. By writing the classification problem as a regression problem the linear smoother properties of the LS-SVM can be used to derive suitable bias and variance expressions [6] with applications to confidence and prediction intervals for regression, further explained in Section 3. This paper provides new insights of [6] in the sense that it extends the latter to the classification case.

Finally, using the estimated bias and variance $\hat{\mathbf{V}}$, we are searching for the width of the bands c given a confidence level $\alpha \in (0,1)$ such that

$$\inf_{m\in\mathcal{M}} \mathbf{P}\left\{\hat{m}(x) - c\sqrt{\hat{\mathbf{V}}(x)} \leq m(x) \leq \hat{m}(x) + c\sqrt{\hat{\mathbf{V}}(x)}, \forall x \in \mathcal{X}\right\} = 1-\alpha,$$

for some suitable class of smooth functions $\mathcal{M}$ with $\hat{m}$ an estimate of the true function m and $\mathcal{X} \subseteq \mathbb{R}^d$. The width of the bands can be determined by several methods e.g. Bonferroni corrections, Šidák corrections [23], length heuristic [8], Monte Carlo based techniques [22] and volume-of-tube formula [25]. The first three are easy to calculate but produce conservative confidence bands. Monte Carlo based techniques are known to obtain very accurate results but are often computational intractable in practice. The volume-of-tube formula tries to combine the best of both worlds i.e. relatively easy to calculate and producing results similar to Monte Carlo based techniques. In the remaining of the paper the volume-of-tube formula will be used. We will also provide an extension of this formula which is valid in the d dimensional case.

This paper is organized as follows: The relation between classification and regression in the LS-SVM framework is clarified in Section 2. Bias and variance estimates as well as the volume-of-tube formula are discussed in Section 3. The construction of bias-corrected $100(1-\alpha)\%$ simultaneous confidence bands are formulated in Section 4. Simulations and how to interpret the confidence bands in case of classification are given in Section 5. Finally, Section 6 states the conclusions.

*Corresponding author
Email addresses: kris.debrabanter@esat.kuleuven.be (K. De Brabanter), peter.karsmakers@khk.be (P. Karsmakers), jos.debrabanter@kahosl.be (J. De Brabanter), johan.suykens@esat.kuleuven.be (J.A.K. Suykens), bart.demoor@esat.kuleuven.be (B. De Moor)

2. Classification versus Regression

Given a training set defined as $\mathcal{D}_n = \{(X_k, Y_k) : X_k \in \mathbb{R}^d, Y_k \in \{-1, +1\}; k = 1, \ldots, n\}$, where X_k is the k-th input pattern and Y_k is the k-th output pattern. In the primal weight space, LS-SVM for classification is formulated as [26, 27]

$$\begin{aligned} \min_{w,b',e} \mathcal{J}_c(w, e) &= \tfrac{1}{2} w^T w + \tfrac{\gamma}{2} \sum_{i=1}^{n} e_i^2 \\ \text{s.t.} \quad & Y_i[w^T \phi(X_i) + b'] = 1 - e_i, \quad i = 1, \ldots, n, \end{aligned} \tag{1}$$

where $\phi : \mathbb{R}^d \to \mathbb{R}^{n_h}$ is the feature map to the high dimensional feature space (can be infinite dimensional) as in the standard Support Vector Machine (SVM) case [29], $w \in \mathbb{R}^{n_h}, b' \in \mathbb{R}$ and $\gamma \in \mathbb{R}_0^+$ is the regularization parameter. On the target value an error variable e_i is allowed such that misclassifications can be tolerated in case of overlapping distributions. By using Lagrange multipliers, the solution of (1) can be obtained by taking the Karush-Kuhn-Tucker (KKT) conditions for optimality. The result is given by the following linear system in the dual variables μ

$$\left[\begin{array}{c|c} 0 & Y^T \\ \hline Y & \Omega^{(c)} + \frac{1}{\gamma} I_n \end{array}\right] \left[\begin{array}{c} b' \\ \hline \mu \end{array}\right] = \left[\begin{array}{c} 0 \\ \hline 1_n \end{array}\right],$$

with $Y = (Y_1, \ldots, Y_n)^T$, $1_n = (1, \ldots, 1)^T$, $\mu = (\mu_1, \ldots, \mu_n)^T$ and $\Omega_{il}^{(c)} = Y_i Y_l \phi(X_i)^T \phi(X_l) = Y_i Y_l K(X_i, X_l)$ for $i, l = 1, \ldots, n$ with $K(\cdot, \cdot)$ a positive definite kernel. Such a positive definite kernel K guarantees the existence of the feature map ϕ but ϕ is often not explicitly known. Based on Mercer's theorem, the resulting LS-SVM model for classification in the dual space becomes

$$\hat{y}(x) = \operatorname{sign}\left[\sum_{i=1}^{n} \hat{\mu}_i Y_i K(x, X_i) + \hat{b}'\right],$$

where $K : \mathbb{R}^d \times \mathbb{R}^d \to \mathbb{R}$. For example, the Gaussian kernel $K(X_i, X_j) = (1/\sqrt{2\pi}) \exp(-\|X_i - X_j\|_2^2 / 2h^2)$ with bandwidth $h > 0$.

However, by noticing that the class labels satisfy $\{-1, +1\} \subset \mathbb{R}$ one can also interpret (1) as a nonparametric regression problem. In this case the training set is $\mathcal{D}_n$ of size n drawn i.i.d. from an unknown distribution F_{XY} according to

$$Y = m(X) + \sigma(X)\varepsilon,$$

where $\varepsilon \in \mathbb{R}$ are assumed to be i.i.d. random errors with $\mathbf{E}[\varepsilon] = 0$, $\mathbf{Var}[\varepsilon] = 1$, $\mathbf{Var}[Y|X] = \sigma^2(X) < \infty$, $m \in C^z(\mathbb{R})$ with $z \geq 2$, is an unknown real-valued smooth function and $\mathbf{E}[Y|X] = m(X)$. Two possible situations can occur: (i) $\sigma^2(X) = \sigma^2 =$ constant and (ii) the variance is a function of the random variable X. The first is called homoscedasticity and the latter heteroscedasticity [9].

The optimization problem of finding the vector $w \in \mathbb{R}^{n_h}$ and $b' \in \mathbb{R}$ for regression can be formulated as follows [27]

$$\begin{aligned} \min_{w,b',\varepsilon} \mathcal{J}(w, \varepsilon) &= \tfrac{1}{2} w^T w + \tfrac{\gamma}{2} \sum_{i=1}^{n} \varepsilon_i^2 \\ \text{s.t.} \quad & Y_i = w^T \phi(X_i) + b' + \varepsilon_i, \quad i = 1, \ldots, n. \end{aligned} \tag{2}$$

By using Lagrange multipliers, the solution of (2) can be obtained by taking the Karush-Kuhn-Tucker (KKT) conditions for optimality. The result is given by the following linear system in the dual variables α

$$\left[\begin{array}{c|c} 0 & 1_n^T \\ \hline 1_n & \Omega + \frac{1}{\gamma} I_n \end{array}\right] \left[\begin{array}{c} b' \\ \hline \alpha \end{array}\right] = \left[\begin{array}{c} 0 \\ \hline Y \end{array}\right],$$

with $Y = (Y_1, \ldots, Y_n)^T$, $1_n = (1, \ldots, 1)^T$, $\alpha = (\alpha_1, \ldots, \alpha_n)^T$ and $\Omega_{il} = \phi(X_i)^T \phi(X_l) = K(X_i, X_l)$ 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

$$\hat{m}(x) = \sum_{i=1}^{n} \hat{\alpha}_i K(x, X_i) + \hat{b}'. \tag{3}$$

Hence, a model for classification based on regression can be obtained by taking the sign function of (3). As for the classification case the constraints in (2) can be rewritten as $Y_i(w^T \phi(X_i) + b') = 1 - e_i$. This corresponds to the substitution $e_i = Y_i \varepsilon_i$, which does not change the objective function ($Y_i^2 = 1$), so both formulations are equivalent.

3. Statistical Properties of LS-SVM: Bias and Variance

The reason for writing a classification problem as a regression problem is as follows. Since LS-SVM for regression is a linear smoother many statistical properties of the estimator can be derived e.g. bias and variance [6]. Therefore, all properties obtained for the regression case can be directly applied to the classification case.

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 (smoother) vector $L(x) = (l_1(x), \ldots, l_n(x))^T \in \mathbb{R}^n$ such that*

$$\hat{m}(x) = \sum_{i=1}^{n} l_i(x) Y_i, \tag{4}$$

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

On training data, (4) 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 $\hat{m}(X_i)$. It can be shown for LS-SVM that [6]

$$L(x) = \left[\Omega_{x^\star}^T \left(Z^{-1} - Z^{-1} \frac{J_n}{\kappa} Z^{-1}\right) + \frac{J_1^T}{\kappa} Z^{-1}\right]^T,$$

with $\Omega_{x^\star} = (K(x^\star, X_1), \ldots, K(x^\star, X_n))^T$ the kernel vector evaluated at point $x^\star$, $\kappa = 1_n^T \left(\Omega + \frac{I_n}{\gamma}\right)^{-1} 1_n$, $Z = \Omega + \frac{I_n}{\gamma}$, J_n a square matrix with all elements equal to 1 and $J_1 = (1, \ldots, 1)^T$. Based on this linear smoother property a bias estimate of the LS-SVM can be derived.

Theorem 1 (De Brabanter *et al.*, 2010 [6])**.** *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*

$$\widehat{\text{bias}}[\hat{m}(x)|X = x] = L(x)^T \hat{m} - \hat{m}(x). \tag{5}$$

Techniques as (5) are known as plug-in bias estimates and can be directly calculated from the LS-SVM (see also [13, 6]). However, it is possible to construct better bias estimates, at the expense of extra calculations, by using a technique called double smoothing [11] which can be seen as a generalization of the plug-in based technique. Before explaining the double smoothing, we need to introduce the following definition:

Definition 2 (Jones & Foster (1993) [15])**.** *A kernel $\mathcal{K}$ is called a kth-order kernel if*

$$\begin{cases} \int \mathcal{K}(u)\,du = 1 \\ \int u^j \mathcal{K}(u)\,du = 0, \quad j = 1, \ldots, k-1 \\ \int u^k \mathcal{K}(u)\,du \neq 0, \end{cases}$$

where in general $\mathcal{K}$ is an isotropic kernel i.e. if it only depends on the Euclidean distance between points.

For example, the Gaussian kernel satisfies this condition but the linear and polynomial kernels do not. There are several rules for constructing higher-order kernels, see e.g. [15]. Let $\mathcal{K}_{[k]}$ be a kth-order symmetric kernel (k even) which is assumed to be differentiable. Then

$$\mathcal{K}_{[k+2]}(u) = \frac{k+1}{k}\mathcal{K}_{[k]}(u) + \frac{1}{k}u\mathcal{K}'_{[k]}(u), \tag{6}$$

is a $(k+2)$th-order kernel. Hence, this formula can be used to generate higher-order kernels in an easy way. Consider for example the standard normal density function φ (Gaussian kernel) which is a second-order kernel. Then a fourth-order kernel can be obtained via (6):

$$\begin{aligned} \mathcal{K}_{[4]}(u) &= \frac{3}{2}\varphi(u) + \frac{1}{2}u\varphi'(u) \\ &= \frac{1}{2}(3-u^2)\varphi(u) \\ &= \frac{1}{2}\frac{1}{\sqrt{2\pi}}(3-u^2)\exp\left(-\frac{u^2}{2}\right), \end{aligned} \tag{7}$$

where $u = \|X_i - X_j\|_2/g$. In the remaining of the paper the Gaussian kernel (second and fourth order) will be used. Figure 2 shows the standard normal density function $\varphi(\cdot)$ together with the fourth-order kernel $\mathcal{K}_{[4]}(\cdot)$ derived from it. It can be easily verified using Bochner's lemma [4] that $\mathcal{K}_{[4]}$ is an admissible positive definite kernel.

The idea of double smoothing bias estimation is then given as follows:

Theorem 2 (Double Smoothing [20, 11])**.** *Let $L(x)$ be the smoother vector evaluated in a point x, let $\hat{m}_g = (\hat{m}_g(X_1), \ldots, \hat{m}_g(X_n))^T$ be another LS-SVM smoother based on the same data set and a kth-order kernel with $k > 2$ and different bandwidth g. Then, the double smoothing bias estimate of the conditional bias for LS-SVM is defined by*

$$\hat{b}(x) = L(x)^T \hat{m}_g - \hat{m}_g(x). \tag{8}$$

Note that the bias estimate (8) can be thought of as an iterated smoothing algorithm. The pilot smooth $\hat{m}_g$ (with fourth-order kernel $\mathcal{K}_{[4]}$ and bandwidth g e.g. the kernel constructed in (7)) is resmoothed with kernel K (Gaussian kernel), incorporated in the smoother matrix L, and bandwidth h. Because we smooth twice, this is called double smoothing. A second part needed for the construction of confidence bands is the variance of LS-SVM: see Theorem 3.

Theorem 3 (De Brabanter *et al.*, 2010 [6])**.** *Let $L(x)$ be the smoother vector evaluated in a point x and let $S \in \mathbb{R}^{n\times n}$ be the smoother matrix corresponding to the smoothing of squared residuals. Denote by $S(x)$ the smoother vector in an arbitrary point x and if the smoother preserves constant vectors e.g. $S1_n = 1_n$, then the conditional variance of the LS-SVM is given by*

$$\hat{\mathbf{V}}(x) = \mathbf{Var}[\hat{m}(x)|X = x] = L(x)^T \hat{\Sigma}^2 L(x), \tag{9}$$

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

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

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

4. Construction of Confidence Bands

4.1. Theoretical Aspects

There exist many different types of confidence bands e.g. pointwise [9], uniform or simultaneous [16, 6], Bayesian credible bands [2], tolerance bands [1],... It is beyond the scope of this paper to discuss them all, but it is noteworthy to stress the difference between pointwise and simultaneous confidence bands. We will clarify this by means of a simple example. Suppose our aim is to estimate some function $m(x)$. For example, $m(x)$ might be the proportion of people of a particular age (x) who support a given candidate in an election. If x is measured at the precision of a single year, we can construct a "pointwise" 95% confidence interval (band) for each age. Each of these confidence intervals covers the corresponding true value $m(x)$ with a coverage probability of 0.95. The "simultaneous" coverage probability of a collection of confidence intervals is the probability that all of them cover their corresponding true values simultaneously. From this, it is clear that simultaneous confidence bands will be wider than pointwise confidence bands. A more formal definition for simultaneous coverage probability of the confidence bands (in case of an unbiased estimator) is given by

$$\inf_{m\in\mathcal{M}} \mathbf{P}\left\{\hat{m}(x) - c\sqrt{\hat{\mathbf{V}}(x)} \leq m(x) \leq \hat{m}(x) + c\sqrt{\hat{\mathbf{V}}(x)}, \forall x \in \mathcal{X}\right\} = 1 - \alpha,$$

where $\mathcal{M}$ is a suitable large class of smooth functions, $\mathcal{X} \subseteq \mathbb{R}^d$ and denote by $(x_1, \ldots, x_d)$ a point $x \in \mathcal{X}$, the constant c represents the width of the bands (see Proposition 1 and 2) and α denotes the significance level (typically $\alpha = 0.05$). For the remaining of the paper we will discuss the uniform or simultaneous confidence bands.

In general, simultaneous confidence bands for a function m are constructed by studying the asymptotic distribution of $\sup_{a \le x \le b} |\hat{m}(x) - m(x)|$, where $\hat{m}$ denotes the estimated function. The approach of [3] relates this to a study of the distribution of $\sup_{a \le x \le b} |Z(x)|$, with $Z(x)$ a (standardized) Gaussian process satisfying certain conditions, which they show to have an asymptotic extreme value distribution. A closely related approach and the one we will use here, is to construct confidence bands based on the volume-of-tube formula. In [24] the tail probabilities of suprema of Gaussian random processes were studied. These turn out to very useful in constructing simultaneous confidence bands. Proposition 1 and 2 summarize the results of [24] and [25] in the two and d dimensional case respectively when the error variance σ^2 is not known a priori and has to be estimated. It is important to note that the justification for the tube formula assumes the errors have a normal distribution (this can be relaxed to spherically symmetric distributions [18]), but does not require letting $n \to \infty$. As a consequence, the tube formula does not require embedding finite sample problems in a possibly artificial sequence of problems, and the formula can be expected to work well at small sample sizes.

Proposition 1 (two-dimensional [25])**.** *Suppose $\mathcal{X}$ is a rectangle in $\mathbb{R}^2$. Let κ_0 be the area of the continuous manifold $\mathfrak{M} = \{T(x) = L(x)/\|L(x)\|_2 : x \in \mathcal{X}\}$, let ζ_0 be the volume of the boundary of $\mathfrak{M}$ denoted by $\partial\mathfrak{M}$. Let $T_j(x) = \partial T(x)/\partial x_j$, $j = 1 \ldots, d$ and let c be the width of the bands. Then,*

$$\alpha = \frac{\kappa_0}{\sqrt{\pi^3}} \frac{\Gamma\left(\frac{(\nu+1)}{2}\right)}{\Gamma\left(\frac{\nu}{2}\right)} \frac{c}{\sqrt{\nu}} \left(1 + \frac{c^2}{\nu}\right)^{-\frac{\nu+1}{2}} + \frac{\zeta_0}{2\pi} \left(1 + \frac{c^2}{\nu}\right)^{-\frac{\nu}{2}} + \mathbf{P}\left(|t_\nu| > c\right) + o\left(\exp\left(-\frac{c^2}{2}\right)\right), \tag{10}$$

with $\kappa_0 = \int_{\mathcal{X}} \det^{1/2}(A^T A)\, dx$, $\zeta_0 = \int_{\partial\mathcal{X}} \det^{1/2}(A_\star^T A_\star)\, dx$ where $A = (T_1(x), \ldots, T_d(x))$ and $A_\star = (T_1(x), \ldots, T_{d-1}(x))$. t_ν is a t-distributed random variable with $\nu = n - \mathrm{trace}(L)$, with $L \in \mathbb{R}^{n \times n}$, degrees of freedom.

Proposition 2 (d-dimensional [25])**.** *Suppose $\mathcal{X}$ is a rectangle in $\mathbb{R}^d$ and let c be the width of the bands. Then,*

$$\alpha = \kappa_0 \frac{\Gamma\left(\frac{(\nu+1)}{2}\right)}{\pi^{\frac{d+1}{2}}} \mathbf{P}\left(F_{d+1,\nu} > \frac{c^2}{d+1}\right) + \frac{\zeta_0}{2} \frac{\Gamma\left(\frac{d}{2}\right)}{\pi^{\frac{d}{2}}} \mathbf{P}\left(F_{d,\nu} > \frac{c^2}{d}\right) + \frac{\kappa_2 + \zeta_1 + z_0}{2\pi} \frac{\Gamma\left(\frac{(\nu-1)}{2}\right)}{\pi^{\frac{d-1}{2}}} \mathbf{P}\left(F_{d-1,\nu} > \frac{c^2}{d-1}\right) + O\left(c^{d-4} \exp\left(-\frac{c^2}{2}\right)\right), \tag{11}$$

where κ_2, ζ_1 and z_0 are certain geometric constants. $F_{q,\nu}$ is a F-distributed random variable with q and $\nu = \eta_1^2/\eta_2$ degrees of freedom [5] where $\eta_1 = \mathrm{trace}\left[(I_n - L^T)(I_n - L)\right]$ and $\eta_2 = \mathrm{trace}\left[\left((I_n - L^T)(I_n - L)\right)^2\right]$.

The equations (10) and (11) contain quantities which are often rather difficult to compute in practice. Therefore, the following approximations can be made: (i) according to [19], $\kappa_0 = \frac{\pi}{2}(\mathrm{trace}(L) - 1)$ and (ii) it is shown in the simulations of [25] that the third term is negligible in (11). More details on the computation of these constants can be found in [25] and [22]. To compute the value c, the width of the bands, any method for solving nonlinear equations can be used.

To illustrate the effect of increasing dimensionality on the c value in (11) we conduct the following Monte Carlo study. For increasing dimensionality we calculate the c value for a Gaussian density function in d dimensions. 1000 data points were generated uniformly on $[-3, 3]^d$. Figure 1 shows the result of the calculated c value averaged over 20 times for each dimension. It can be concluded that the width of the bands is increasing for increasing dimensionality. Theoretical derivations confirming this simulation can be found in [12, 29]. Simply put, estimating a regression function (in a high-dimensional space) is especially difficult because it is not possible to densely pack the space with finitely many sample points [10]. The uncertainty of the estimate is becoming larger for increasing dimensionality, hence the confidence bands are wider.

We are now ready to formulate the proposed confidence bands. In the unbiased case, simultaneous confidence intervals would be of the following form

$$\hat{m}(x) \pm c\sqrt{\hat{\mathbf{V}}(x)}, \tag{12}$$

where the width of the bands c can be determined by solving (11) in the d dimensional case. However, since all nonparametric regression techniques have an inherent bias, modification of the confidence intervals (12) is needed to attain the required coverage probability. Therefore, we propose the following: Let $\mathcal{M}_\delta$ be the class of smooth functions and $m \in \mathcal{M}_\delta$ where

$$\mathcal{M}_\delta = \left\{ m : \sup_{x \in \mathcal{X}} \left| \frac{b(x)}{\sqrt{\mathbf{V}(x)}} \right| \le \delta \right\},$$

then bands of the form

$$\left(\hat{m}(x) - (\delta + c)\sqrt{\hat{\mathbf{V}}(x)}, \hat{m}(x) + (\delta + c)\sqrt{\hat{\mathbf{V}}(x)}\right) \tag{13}$$

are a confidence band for $m(x)$ where the bias $b(x)$ and variance $V(x)$ can be estimated using (8) and (9) respectively. Notice that the confidence interval (13) expands the bands in the presence of bias rather than recentering the bands to allow for bias. It is shown in [25] that bands of the form (13) lead to a lower bound for the true coverage probability of the form

$$\inf_{m \in \mathcal{M}_\delta} \mathbf{P}\left\{ |\hat{m}(x) - m(x)| < c\sqrt{\hat{\mathbf{V}}(x)}, \forall x \in \mathcal{X} \right\} = 1 - \alpha - O(\delta)$$

as $\delta \to 0$. The error term can be improved to $O(\delta^2)$ if one considers classes of functions with bounded derivatives.

We conclude this Section by summarizing the construction of simultaneous confidence intervals given in Algorithm 1.

Algorithm 1 Construction of Confidence Bands

1: Given the training data $\{(X_1, Y_1), \ldots, (X_n, Y_n)\}$, calculate $\hat{m}$ on training data using (3)
2: Calculate residuals $\hat{\varepsilon}_k = Y_k - \hat{m}(X_k), k = 1, \ldots, n$
3: Calculate the variance of the LS-SVM by using (9)
4: Calculate the bias using double smoothing (8)
5: Set significance level e.g. $\alpha = 0.05$
6: Calculate c from (11)
7: Use (13) to obtain simultaneous confidence bands.

4.2. Illustration and interpretation of the method in two dimensions

In this paragraph we graphically illustrate the proposed method on the Ripley data set. First, the regression model from the Ripley data set is estimated according to (3). From the obtained model, confidence bands can be calculated using (13). Figure 3 shows the obtained results in three dimensions and its two dimensional projection respectively. In the latter the two outer bands represent 95% confidence intervals for the classifier. An interpretation of this result can be given as follows. For every point within (out) the two outer bands, the classifier casts doubt with significance level α (is confident with significance level α) on its label. However, in higher dimensions, the previous figures cannot be made anymore. Therefore, we can visualize the classifier via its latent variables, i.e. the output of (3) (before taking the sign function), and show the corresponding confidence intervals, see Figure 4. The middle full line represents the sorted latent variables of the classifier. The dots above and below the full line are the 95% confidence intervals of the classifier. These dots correspond to the confidence bands in Figure 3. The dashed line at zero is the decision boundary. The rectangle visualizes the critical area for the latent variables. Hence, for all points with latent variables between the two big dots the classifier, i.e. |latent variable| ≤ 0.51, casts doubt on the corresponding label with significance level of 5%. Such a visualization can always be made and can assist the user in assessing the quality of the classifier.

5. Simulations

All the data sets in the simulation are freely available from `http://archive.ics.uci.edu/ml` and/or `http://www.csie.ntu.edu.tw/~cjlin/libsvmtools/datasets/`.

First, take the Pima Indians data set ($d = 8$). Figure 5 shows the $100(1-\alpha)$% confidence bands for the classifier based on the latent variables where α is varied from 0.05 to 0.1 respectively. Figure 5(a) illustrates the 95% confidence bands for the classifier based on the latent variables and Figure 5(b) the 90% confidence bands. It is clear that the width of the confidence band will decrease when α increases. Hence, the 95% and 90% confidence band for the latent variables is given roughly by $(-0.70, 0.70)$ and $(-0.53, 0.53)$.

Second, consider the Fourclass data set ($d = 2$). This is an example of a non-linear separable classification problem (see Figure 6(a)). We can clearly see in Figure 6(a) that the 95% confidence bands are not so wide, indicating no overlap between classes. Figure 6(b) shows the 95% confidence bands for the classifier based on the latent variables. The two black dots indicate the critical region. Therefore, if for any point |latent variable| ≤ 0.2 we have less than 95% confidence on the decision of the class label.

As a third example, consider the Haberman's Survival data set ($d = 3$). The data set contains cases from a study that was conducted between 1958 and 1970 at the University of Chicago's Billings Hospital on the survival of patients who had undergone surgery for breast cancer. The task is to predict whether the patient will survive five years or longer. Figure 7 shows the 95% confidence bands for the latent variables. Every point lying left from the big dot i.e. latent variable < 0.59, the classifier casts doubt on its label on a significance level of 5%. This is a quite difficult classification task as can be seen from the elongated form of the sorted latent variables and is also due to the unbalancedness in the data.

A fourth example is the Wisconsin Breast Cancer data set ($d = 10$). Figure 8 shows the 95% confidence bands for the latent variables. Thus every point lying between the big dots, the classifier casts doubt on its label on a significance level of 5%.

As a last example, we used the Heart data set ($d = 13$) which was scaled to $[-1, 1]$. Figure 9 shows the 95% confidence bands for the latent variables. The interpretation of the results is the same as before.

6. Conclusions

In this paper we proposed bias-corrected $100(1-\alpha)$% data-driven confidence bands for kernel based classification, more specifically for LS-SVM in the classification context. We have illustrated how suitable bias and variance estimates for LS-SVM can be calculated in a relatively easy way. In order to obtain simultaneous confidence intervals we used the volume-of-tube formula and provided extensions in the d dimensional case. A simple simulation study pointed out that these bands are becoming wider for increasing dimensionality. Simulations of the proposed method on classification data sets provide insight in the classifier's confidence and can assist the user in the interpretation of the classifier's result.

Acknowledgments

Research supported by Onderzoeksfonds K.U.Leuven/Research Council KUL: GOA/11/05 Ambiorics, GOA/10/09 MaNet , CoE EF/05/006 Optimization in Engineering (OPTEC) en PFV/10/002 (OPTEC), IOF-SCORES4CHEM, several PhD/postdoc & fellow grants; Flemish Government: FWO: PhD/postdoc grants, projects: G0226.06 (cooperative systems and optimization), G0321.06 (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 (WOG: ICCoS, ANMMM, MLDM); G.0377.09 (Mechatronics MPC) IWT: PhD Grants, Eureka-Flite+, SBO LeCo-Pro, SBO Climaqs, SBO POM, O&O-Dsquare Belgian Federal Science Policy Office: IUAP P6/04 (DYSCO, Dynamical systems, control and optimization, 2007-2011); IBBT EU: ERNSI; FP7-HD-MPC (INFSO-ICT-223854), COST intelliCIS, FP7-EMBOCON (ICT-248940), FP7-SADCO (MC ITN-264735), ERC HIGHWIND (259 166) Contract Research: AMINAL Other: Helmholtz:

viCERP ACCM. BDM is a full professor at the Katholieke Universiteit Leuven, Belgium.

References

[1] R.B. Abernethy, The New Weibull Handbook (5th ed.), 2004.

[2] J.M. Bernardo and A.F.M. Smith, Bayesian Theory, John Wiley & Sons, 2000.

[3] P.J. Bickel and M. Rosenblatt, On Some Global Measures of the Deviations of Density Function Estimates, Annals of Statistics, 1(6) (1973) 1071–1095.

[4] S. Bochner, Lectures on Fourier integrals, Princeton University Press, 1959.

[5] W.S. Cleveland and S.J. Devlin, Locally Weighted Regression: An Approach to Regression Analysis by Local Fitting, Journal of the American Statistical Association, 83(403) (1988) 596–610.

[6] K. De Brabanter, J. De Brabanter, J.A.K. Suykens and B. De Moor, Approximate Confidence and Prediction Intervals for Least Squares Support Vector Regression, IEEE Transactions on Neural Networks, 22(1) (2010) 110–120.

[7] L. Devroye, L. Györfi and G. Lugosi, A Probabilistic Theory of Pattern Recognition, Springer, 1996.

[8] B. Efron, The length heuristic for simultaneous hypothesis tests, Biometrika, 84(1) (1997) 143–157.

[9] J. Fan and I. Gijbels, Local Polynomial Modelling and its Applications, Chapman & Hall, 1996.

[10] L. Györfi, M. Kohler, A. Krzyżak and H. Walk, A Distribution-Free Theory of Nonparametric Regression, Springer, 2002.

[11] W. Härdle, P. Hall and J.S. Marron, Regression Smoothing Parameters That Are Not Far From Their Optimum, Journal of the American Statistical Association, 87(417) (1992) 227–233.

[12] T. Hastie, R. Tibshirani and J. Friedman, The Elements of Statistical Learning: Data Mining, Inference, and Prediction (2nd ed.), Springer, 2009.

[13] N.W. Hengartner, P-A. Cornillon and E. Matzner–Lober, Recursive Bias Estimation and L_2 Boosting, Technical Report LA-UR-09-01177, Los Alamos National Laboratory, 2009. `http://www.osti.gov/energycitations/product.biblio.jsp?osti_id=956561`

[14] G. Huang, H. Chen, Z. Zhou, F. Yin and K. Guo, Two-Class Support Vector Data Description, Pattern Recognition, 44(2)(2011) 320–329.

[15] M.C. Jones and P.J. Foster, Generalized Jackknifing and Higher Order Kernels, Journal of Nonparametric Statistics, 3(11) (1993) 81–94.

[16] T. Krivobokova, T. Kneib and G. Claeskens, Simultaneous Confidence Bands for Penalized Splines, Journal of the American Statistical Association, 105(490) (2010) 852–863.

[17] S. Li, J.T. Kwok, H. Zhu and Y. Wang, Texture Classification Using the Support Vector Machines, Pattern Recognition, 36(12) (2003) 2883–2893.

[18] C. Loader and J. Sun, Robustness of Tube Formula Based Confidence Bands, Journal of Computational and Graphical Statistics, 6(2) (1997) 242–250.

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

[20] H. Müller, Empirical Bandwidth Choice for Nonparametric Kernel Regression by Means of Pilot Estimators, Statistical Decisions, 2 (1985), 193–206.

[21] F. Orabona, C. Castellini, B. Caputo, L. Jie and G. Sandini, On-Line Independent Support Vector Machines, Pattern Recgnition, 43(4) (2010) 1402–1412.

[22] D. Ruppert, M.P. Wand and R.J. Carroll, Semiparametric Regression, Cambridge University Press, 2003.

[23] Z. Šidák, Rectangular Confidence Region for the Means of Multivariate Normal Distributions, 62(318) (1967) 626–633.

[24] J. Sun, Tail Probabilities of the Maxima of Gaussian Random Fields, The Annals of Probability, 21(1) (1993) 852–855.

[25] J. Sun and C.R. Loader, Simultaneous Confidence Bands for Linear Regression and Smoothing, Annals of Statistics, 22(3) (1994) 1328-1345.

[26] J.A.K. Suykens and J. Vandewalle, Least Squares Support Vector Machine Classifiers, Neural Processing Letters, 9(3) (1999) 293–300.

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

[28] A.B. Tsybakov. Introduction to Nonparametric Estimation, Springer, 2009.

[29] V.N. Vapnik, Statistical Learning Theory, John Wiley & Sons Inc, 1999.

[30] L. Wasserman, All of Nonparametric Statistics, Springer, 2006.

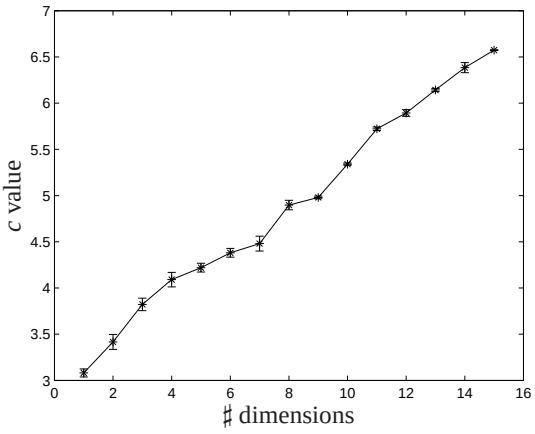


Figure 1: Result of the calculated c value averaged over 20 times for each dimension (for a Gaussian density function) with corresponding standard error.

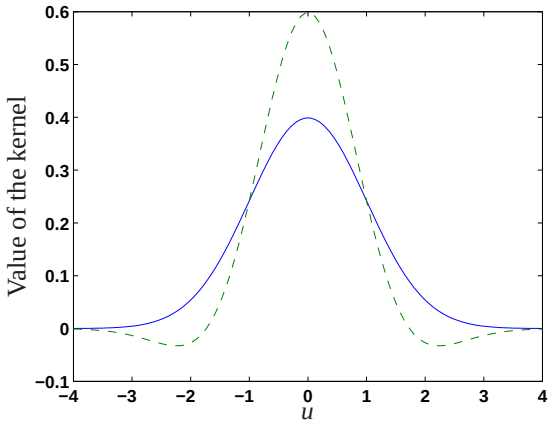


Figure 2: Plot of $\mathcal{K}_{[2]}(u) = \varphi(u)$ (solid curve) and $\mathcal{K}_{[4]}(u)$ (dashed curve).

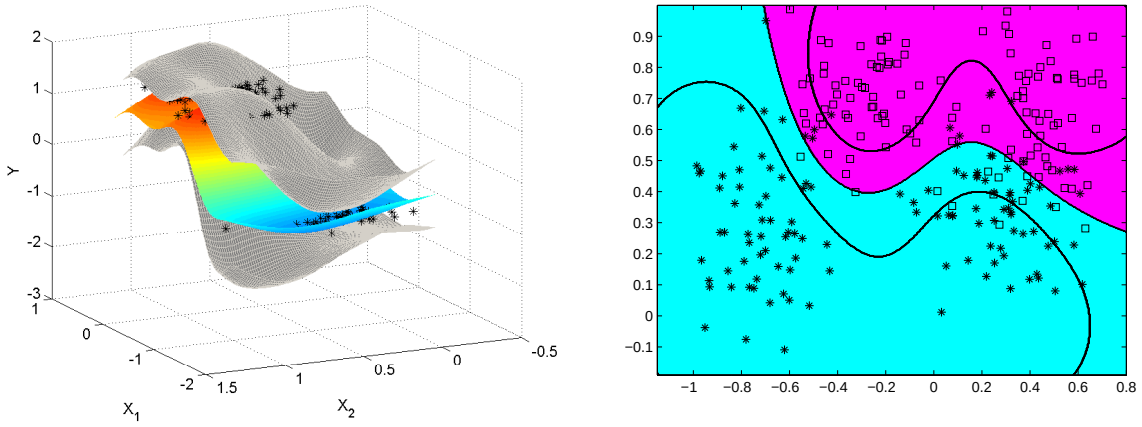


Figure 3: Ripley data set (a) Regression on the Ripley data set with corresponding 95% confidence intervals obtained with (13) where X_1, X_2 are the corresponding abscissa and Y the function value; (b) Two dimensional projection of (a) obtained by cross-secting the regression surfaces with the decision plane $Y = 0$. The two outer lines represent 95% confidence intervals on the classifier. The line in the middle is the resulting classifier.

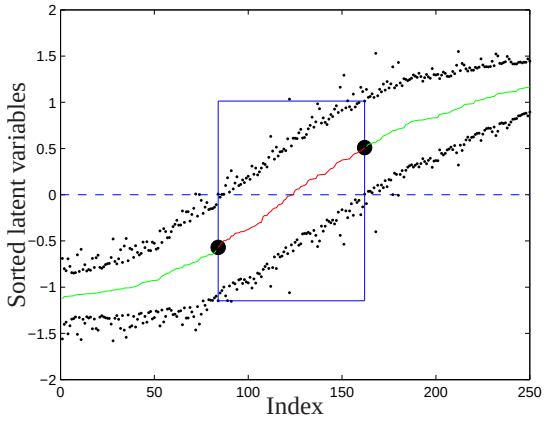
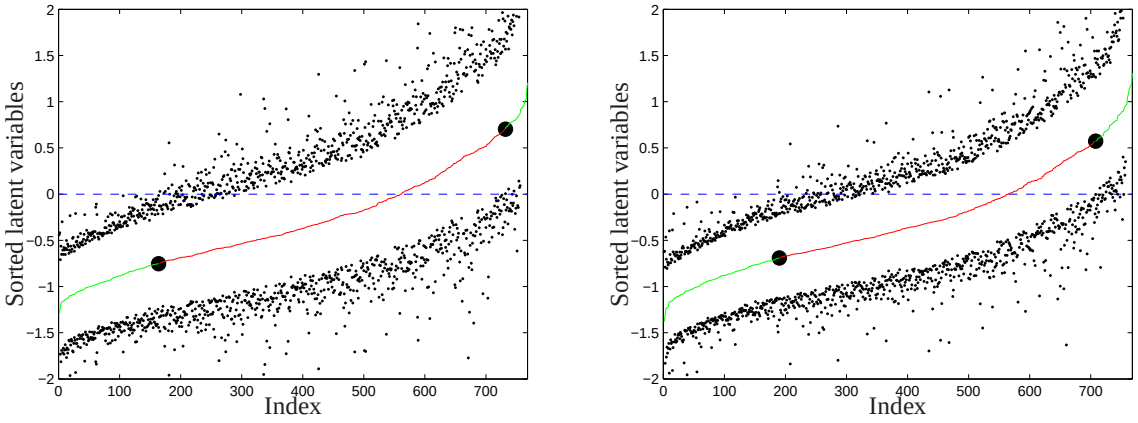
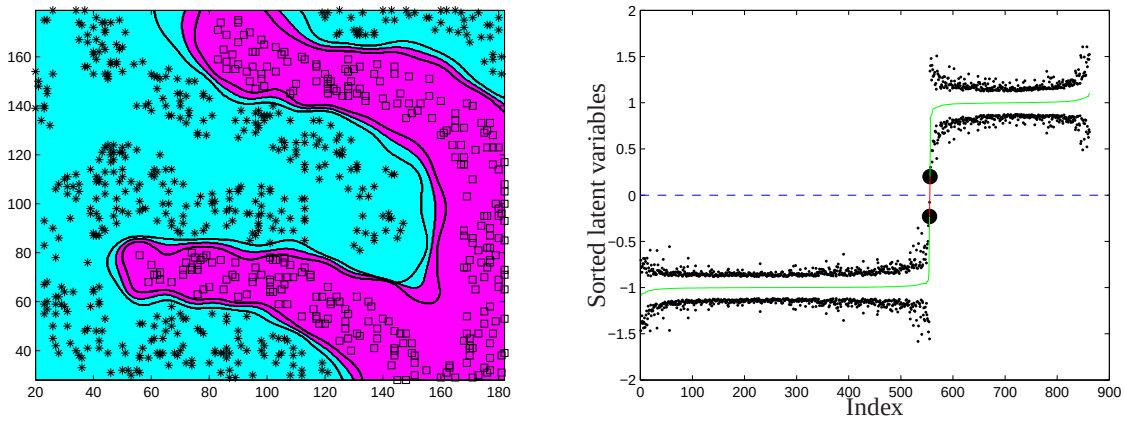


Figure 4: Visualization of the 95% confidence bands (small dots above and below the middle line) for the Ripley data set based on the latent labels (middle full line). The rectangle visualizes the critical area for the latent variables, therefore, every point lying between the two big dots the classifier casts doubt on its label with significance level of 5%. The dashed line is the decision boundary.



(a) (b)

Figure 5: Pima Indians data set. Effect of larger significance level on the width of the confidence bands. The bands will become wider when the significance level decreases. (a) 95% confidence band on the latent variables; (b) 90% confidence band on the latent variables



(a) (b)

Figure 6: Fourclass data set (a) Two dimensional projection of the classifier (inner line) and its corresponding 95% confidence bands (two outer lines); (b) Visualization of the 95% confidence bands (small dots above and below the middle line) for classification based on the latent labels (middle full line). For every latent variable lying between the two big dots the classifier casts doubt on its label. The dashed line is the decision boundary.

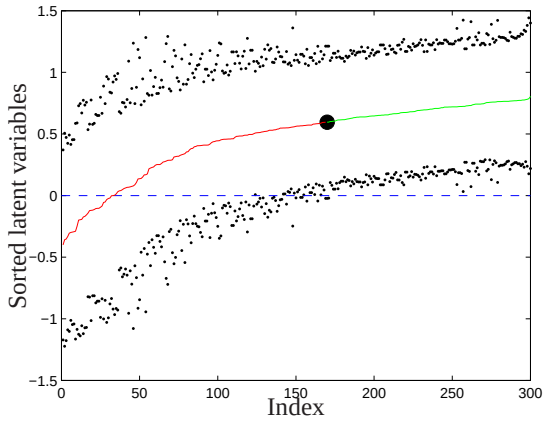


Figure 7: Visualization of the 95% confidence bands (small dots above and below the middle line) for Haberman's survival data set based on the latent labels (middle full line). For every latent variable lying left from the big dot the classifier casts doubt on its label on a 5% significance level. The dashed line is the decision boundary.

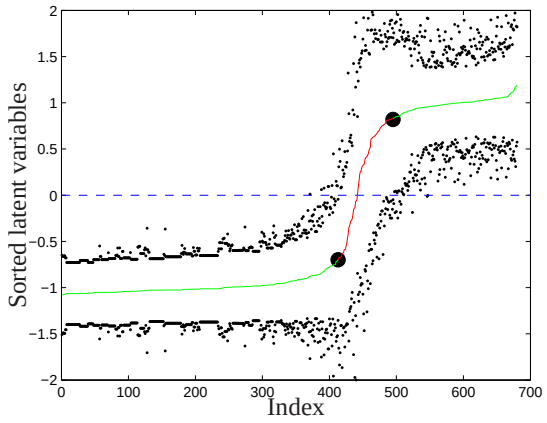


Figure 8: Visualization of the 95% confidence bands (small dots above and below the middle line) for Wisconsin Breast Cancer data set based on the latent labels (middle full line). For every latent variable lying between the big dots the classifier casts doubt on its label on a 5% significance level. The dashed line is the decision boundary.

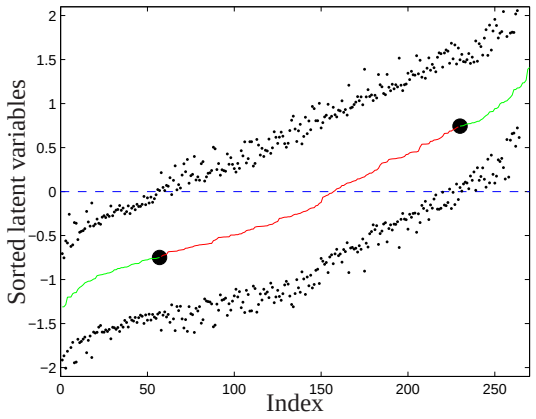


Figure 9: Visualization of the 95% confidence bands (small dots above and below the middle line) for the Heart data set (scaled to $[-1, 1]$) based on the latent labels (middle full line). For every latent variable lying left from the big dot the classifier casts doubt on its label on a 5% significance level. The dashed line is the decision boundary.