ROBUST COMPLEXITY CRITERIA FOR NONLINEAR REGRESSION IN NARX MODELS

Jos De Brabanter * **Kristiaan Pelckmans** * **Johan A.K. Suykens** *
Bart De Moor * **Joos Vandewalle** *

* *K.U.Leuven, ESAT-SCD-SISTA*
Kasteelpark Arenberg 10, B-3001 Leuven - Heverlee (Belgium)
{*Jos.Debrabanter,Johan.Suykens*}*@esat.kuleuven.ac.be*

Abstract: Many different methods have been proposed to construct a smooth regression function, including local polynomial estimators, kernel estimators, smoothing splines and LS-SVM estimators. Each of these estimators use hyperparameters. In this paper a robust version for general cost functions based on the Akaike information criterion is proposed.

Keywords: Akaike information criterion (AIC), smoother matrix, weighted LS-SVM, nonparametric variance estimator, influence function, breakdown point

1. INTRODUCTION

Most efficient learning algorithms in neural networks, support vector machines and kernel based methods (Vapnik, 1998) require the tuning of some extra learning parameters, or hyperparameters, denoted here by θ. In statistics, various hyperparameter selection criteria have been developed. One class of selection criteria is based on resampling. Cross-validation (Stone, 1974) and bootstrap (Efron, 1979) are methods in problems with i.i.d. data. Applications to time series and other dependent data (Markov chains and stochastic processes that are not indexed by time) are not straightforward. Another class, denoted by complexity criteria, (e.g. Akaike Information Criterion (Akaike, 1973), Baysian Information Criterion (Schwartz, 1979) and C_p statistic (Mallows, 1973) take the general form of a prediction error which consists of the sum of the training set error and a complexity term. The complexity term represents a penalty which grows as the number of free parameters in the model grows.

In the context of dependent data, it is often preferable to have a complexity criterion to select θ. These methods can handle both dependent and independent data. Complexity criteria have been developed often in the context of linear models, for assessing the generalization performance of trained models without the use of validation data. (Moody, 1992) has generalized such criteria to deal with nonlinear models and to allow for the presence of a regularization term. One advantage of cross-validation and bootstrap over complexity criteria is that they do not require estimates of the error variance. This means that complexity criteria require roughly correct working model to obtain the estimate of the error variance. In order to overcome such a problem, a nonparametric variance estimator (independent of the model) is used and the complexity criteria will work well even if the models being assessed are far from correct. (Akaike, 1973) found a relationship between the Kullback-Leibler distance and Fisher's maximized log-likelihood function. This relationship leads to an effective and very general methodology for selecting a parsimonious model for the analysis of empirical data. But when there are outliers in the observations (or if the distribution of the random errors has a heavy tail so that $E[|e_k|] = \infty$), it becomes very difficult to obtain good asymptotic results for the complexity criteria. In order to overcome such problems, robust complexity criteria are proposed in this paper.

In Section 2 we discuss the classical model selection criteria. Section 3 explains the robustness problem (extreme sensitivity) of a complexity criterion. In Sec-

tion 4 we propose a new robust criterion for prediction. Section 5 illustrates the applications.

2. CLASSICAL MODEL SELECTION CRITERIA

Consider the nonparametric regression model

$$y_k = f(x_k) + e_k \tag{1}$$

where $x_1, ..., x_N$ are points in the Euclidean space $\mathbb{R}^d$, $f : \mathbb{R}^d \to \mathbb{R}$ is an unknown real-valued smooth function that we wish to estimate and the e_k are random errors with $E[e_k] = 0$ and $Var[e_k] < \infty$. Let $\mathcal{P}$ be a finite set of parameters. For $p \in \mathcal{P}$, let $\mathcal{F}_p$ be a set of functions f , let $Q_N(p) \in \mathbb{R}^+$ be a complexity term for $\mathcal{F}_p$ and let $\hat{f}$ be an estimator of f in $\mathcal{F}_p$. A crucial step in estimating $f(x)$ is choosing the hyperparameters. The hyperparameters are chosen to be the minimizer of a cost function defined as

$$J(\theta) = \frac{1}{N}\sum_{k=1}^{N} L\left(y_k, \hat{f}\left(x_k;\theta\right)\right) + \lambda\left(Q_N(p)\right)\hat{\sigma}_e^2 \tag{2}$$

where $\sum_{k=1}^{N} L(y_k, \hat{f}(x_k;\theta))$ is the residual sum of squares (RSS). The general cost function depends ony on $\hat{f}$ and the data. If $\hat{f}$ is defined by minimizing the empirical L_2 risk over some linear vector space $\mathcal{F}_p$ of functions with dimension K_p, then $J(\theta)$ will be of the form:

- Let $\lambda = 2$ and $Q_N(p) = N^{-1}K_p$

$$AIC(\theta) = \frac{1}{N}RSS + 2\left(\frac{K_p}{N}\right)\hat{\sigma}_e^2$$

- Let $\lambda = \log N$ and $Q_N(p) = N^{-1}K_p$

$$BIC(\theta) = \frac{1}{N}RSS + (\log N)\left(\frac{K_p}{N}\right)\hat{\sigma}_e^2$$

- Let $\lambda = \log\log N$ and $Q_N(p) = N^{-1}K_p$

$$J_1(\theta) = \frac{1}{N}RSS + (\log\log N)\left(\frac{K_p}{N}\right)\hat{\sigma}_e^2,$$

  which has been proposed by (Hannon and Quinn, 1979) in the context of autoregressive model order determination.

The AIC was originally designed for parametric models as an approximately unbiased estimate of the expected Kullback-Leibler information. For linear regression and time series models, Hurvich and Tsai (Hurvich and Tsai, 1989) demonstrated that in small samples the bias of the AIC can be quite large, especially as the dimension of the candidate model approaches the sample size (leading to overfitting of the model), and they proposed a corrected version, $AICC$, which was found to be less biased than the AIC. The $AICC$ for hyperparameter selection is given by

$$AICC(\theta) = \frac{1}{N}RSS + \left(1 + \frac{2(K_p+1)}{N-K_p-2}\right)\hat{\sigma}_e^2 \tag{3}$$

where $\lambda = 1$ and $Q_N(p) = 1 + \frac{2(K_p+1)}{N-K_p-2}$.

An important property is that all the estimators are linear, in the sense that $\hat{f}(x;\hat{\theta}) = S(x;\hat{\theta})y$, where the matrix $S(x;\hat{\theta})$ is called the smoother matrix and depends on x but not on y (e.g. regression spline estimators, wavelet and LS-SVM estimators are linear estimators). Let $\mathcal{Q}_N$ be a finite set of effective number of parameters. For $q \in \mathcal{Q}_N$, let $\mathcal{F}_{N,q}$ be a set of functions f, let $Q_N(q) \in \mathbb{R}^+$ be a complexity term for $\mathcal{F}_{N,q}$ and let $\hat{f}$ be an estimator of f in $\mathcal{F}_{N,q}$. Based on (Moody, 1992) and by analogy, the hyperparameters are chosen to be the minimizer of a more generalized cost function defined as

$$J_C(\theta) = \frac{1}{N}RSS + \left(1 + \frac{2\text{tr}(S(x;\hat{\theta})) + 2}{N - \text{tr}(S(x;\hat{\theta})) - 2}\right)\hat{\sigma}_e^2. \tag{4}$$

Each of these selectors depends on $S(x;\hat{\theta})$ through its trace ($\text{tr}(S(x;\hat{\theta})) < N-2$), which can be interpreted as the effective number of parameters used in the fit. Because $J_C(\theta)$ is based on least squares estimation (via $L\left(y_k, \hat{f}\left(x_k;\theta\right)\right), \hat{\sigma}_e^2$), it is very sensitive to outliers and other deviations from the normality assumption on the error distribution.

3. ANALYSIS OF ROBUSTNESS

3.1 *Contamination scheme*

In order to understand why certain estimators behave the way they do, it is necessary to look at the various measures of robustness. The effect of one outlier, on the expected Kullback-Leibler information estimator, in the sample can be described by the influence function (Hampel, 1974) which (roughly speaking) formalizes the bias caused by one outlier. Another measure of robustness is which amount of contaminated data an expected Kullback-Leibler information estimator can stand before it becomes useless. This aspect is covered by the breakdown point of the estimator. Given a random sample $\{(x_1, y_1), ..., (x_N, y_N)\}$ with common distribution F. To allow for the occurrence of outliers and other departures from the classical model we will assume that the actual distribution F of the data belongs to the contamination neighborhood

$$F_\epsilon = \{F : F = (1-\epsilon)F_0 + \epsilon H, \;\; H \text{ arbitrary}\} \tag{5}$$

where $0 \le \epsilon < 0.5$. The interpretation is that we have a model F_0 which is not thought to hold exactly and to allow for this one considers an amount of contamination ϵ which is represented by the distribution H. If samples are generated from a F then on average a proportion $(1-\varepsilon)$ of the observations will come from the model F_0 and a proportion ϵ of observations, the so called contaminants, will come from the distribution H. Next we give definitions of influence function, breakdown point and show that the $J_C(\theta)$ as described above is not robust.

3.2 *Measure of robustness*

Let F be a fixed distribution and $T(F)$ a statistical functional defined on a set $\mathcal{F}$ of distributions satisfying some regularity conditions (Hampel, 1974). Statistics which are representable as functionals $T(F_N)$ of the sample distribution F_N are called statistical functionals. Let $\mathcal{X}$ be a random variable, the relevant functional for the variance parameter σ^2 is $T(F) = \int (x - \int x dF(x))^2 dF(x)$. Let the estimator $T_N = T(F_N)$ of $T(F)$ be the functional of the sample distribution F_N. Then the influence function $IF(x; T, F)$ is defined as

$$IF(x, T; F) = \lim_{\epsilon \downarrow 0} \frac{T\left[(1-\epsilon) F + \epsilon \Delta_x\right] - T(F)}{\epsilon}. \tag{6}$$

Here Δ_x is the pointmass distribution at the point x. The IF reflects the bias caused by adding a few outliers at the point x, standardized by the amount ϵ of contamination.

3.3 *Influence function of the expected Kullback-Leibler information estimator*

Let $\mathcal{U}$ be a random variable with realizations $\xi_k = (y_k - \hat{f}(x_k; \theta))$. The corresponding statistical functional for the expected Kullback-Leibler information estimator is given by

$$T(F) = T_1(F) + T_2(F) = \int L(\xi)\, dF(\xi) + A \int \left(\xi - \int \xi dF(\xi)\right)^2 dF(\xi), \tag{7}$$

where $A = 1 + \frac{2\left[\mathrm{tr}\left(S(x;\hat{\theta})\right)+1\right]}{N - \mathrm{tr}\left(S(x;\hat{\theta})\right)-2}$ and by model assumption $\int \xi dF(\xi) = 0$. From (6), (7) and $IF(x; T, F) = IF(x; T_1, F) + IF(x; T_2, F)$, it follows that $IF(\xi; T, F) = L(\xi) - T_1(F) + A\xi^2 - T_2(F) = A\xi^2 + \xi^2 - T(F_N)$. We see that the influence function is unbounded in $\mathbb{R}$. This means that an added observation at a large distance from $T(F_N)$ gives a large value in absolute sense for the influence function.

3.4 *L-estimators*

Consider the class of L-estimators in the location model, $x_k = \vartheta + e_k,\ k = 1, ..., N$. Let $x_{N(1)} \leq ... \leq x_{N(N)}$ be the order statistics corresponding to $x_1, ..., x_N$ with distribution F. The L-estimator of ϑ in the location model is of the form

$$T_N = \sum_{k=1}^{N} C_k x_{N(k)}, \tag{8}$$

where $\sum_{k=1}^{N} C_k = 1$ and the weights are usually chosen to be nonnegative and symmetric (i.e., $C_k \geq 0$ and $C_k = C_{N-k+1},\ k = 1, ..., N$). This class covers a big scale of estimators from the sample mean to the sample median. To get a robust L-estimator of ϑ, we should put $C_k = 0$ for $k \leq g$ and for $k \geq N - g + 1$ with a proper $g > 0$. A typical example of such an estimator is the β-trimmed mean $(0 < \beta < 0.5)$

$$T_N(\beta) = \frac{1}{N - 2g} \sum_{k=g+1}^{N-g} x_{N(k)}, \tag{9}$$

where the trimming proportion β is selected so that $g = \lfloor N\beta \rfloor$. When F is no longer symmetric, it may sometimes be preferable to trim asymmetrically if the tail is expected to be heavier in one direction than the other. If the trimming proportions are β_1 on the left and β_2 on the right, the (β_1, β_2)-trimmed mean is defined as

$$T_N(\beta_1, \beta_2) = \frac{1}{N - (g_1 + g_2)} \sum_{k=g_1+1}^{N-g_2} x_{N(k)}, \tag{10}$$

where β_1 and β_2 are selected so that $g_1 = \lfloor N\beta_1 \rfloor$ and $g_2 = \lfloor N\beta_2 \rfloor$.

The trimmed mean has a simple structure and it is easy to use for practitioners. The corresponding statistical functional for the $(0, \beta_2)$-trimmed mean is given in terms of a quantile function and is defined as

$$T(0, \beta_2; F) = \frac{1}{1 - \beta_2} \int_0^{F^-(1-\beta_2)} x dF(x). \tag{11}$$

The quantile function of a cumulative distribution function F is the generalized inverse $F^- : (0, 1) \to \mathbb{R}$ given by $F^-(q) = \inf\{x : F(x) \geq q\}$. In the absence of information concerning the underlying distribution function F of the sample, the empirical distribution function F_N and the empirical quantile function F_N^- are reasonable estimates for F and F^-, respectively. The empirical quantile function is related to the order statistics $x_{N(1)} \leq, \ldots, \leq x_{N(N)}$ of the sample and $F^-(q) = x_{N(k)},\ \text{ for } q \in \left(\frac{k-1}{N}, \frac{k}{N}\right)$.

4. ROBUST MODEL SELECTION CRITERIA

4.1 *A robust cost function*

A natural approach to robustify the expected Kullback-Leibler information estimator $J_C(\theta)$ is to replace: the $RSS = \sum_{k=1}^{N} (y_k - \hat{f}(x_k; \theta))^2$ by the corresponding robust counterpart RSS_{robust}:

$$RSS_{robust} = \sum_{k=1}^{N} v_k^2 \left(y_k - \hat{f}_{robust}(x_k; v_k, \theta)\right)^2 \tag{12}$$

where $\hat{f}_{robust}(x_k; v_k, \theta)$ is a weighted representation of the function estimate based on (e.g. M-estimator (Huber, 1964) or weighted LS-SVM (Suykens *et al.*, 2002*a*)). The weighting of $\hat{f}_{robust}(x_k; v_k, \theta)$ corresponding with $\{x_k, y_k\}$ is denoted by v_k. This means that data point k will be retained in the procedure if its residual is small to moderate, but disregarded if it is an outlier. An alternative is to replace the variance estimator by a robust estimator of variance.

4.2 *Estimation of variance*

The class of estimators of σ^2 that covers all approaches is that of quadratic forms in the output variable. The first subclass contains variance estimators based on the residuals from an estimator of the regression itself (Neumann, 1994). Difference-based variance estimators lead to the second subclass (Gasser *et al.*, 1986). These estimators are independent of a model fit, but the order of difference (the number of related outputs involved to calculate a local residual) has some influence.

Consider the NARX model (Ljung, 1987)

$$\hat{y}(t) = f(y(t-1), ..., y(t-l), u(t-1), ..., u(t-l)). \tag{13}$$

In practice, it is usually the case that only the ordered observed data $y(t)$ according to the discrete time index t, are known. The variance estimator suggested by (Gasser *et al.*, 1986) is used

$$\hat{\sigma}_e^2(y(t)) = \frac{1}{N-2}\sum_{k=2}^{N-1}\frac{(y(t-1)a + y(t+1)b - y(t))^2}{a^2+b^2+1} \tag{14}$$

where $a = \frac{(t+1)-t}{(t+1)-(t-1)}$ and $b = \frac{t-(t-1)}{(t+1)-(t-1)}$.

Let $\mathcal{Z} = \mathcal{U}^2$ be a function of a random variable with distribution $F(\mathcal{Z})$. Let $F_N(\mathcal{Z})$ be the empirical estimate of $F(\mathcal{Z})$. In (14), a realization of the random variable $\mathcal{U}$ is given by

$$\xi_k = \frac{(y(t-1)a + y(t+1)b - y(t))}{\sqrt{a^2+b^2+1}}. \tag{15}$$

The variance estimator (14) can now be written as an average of the random sample $\xi_1^2, ..., \xi_N^2$:

$$\hat{\sigma}_e^2 = \frac{1}{n-2}\sum_{k=2}^{n-1}\xi_k^2.$$

This can be viewed as an indication of the central value of the density function $g(\mathcal{Z})$. The mean is therefore sometimes referred to as a location parameter. The median, a robust counterpart of the mean, is also a location parameter. But in the asymmetric case the median is not necessarily equal to the mean. Other robust location estimators are: M-estimators (obtained as solutions of equations), L-estimators (linear functions of order statistics) and R-estimators (rank scores). Based on the knowledge of the distribution $F_N(\mathcal{Z})$, in practice $F_N(\mathcal{Z})$ is always concentrated on the positive axis with an asymmetric distribution, we can select the candidate location estimators. First, the Huber-type M-estimators have been found to be quite robust.For asymmetric distributions on the other hand, the computation of the Huber type M-estimators requires rather complex iterative algorithms and its convergence cannot be guaranteed in some important cases. Secondly, the trimmed mean (L-estimator) can be used when the distribution is symmetric and even when the distribution is asymmetric. In this paper we use the trimmed mean.

4.3 *Robust Kullback-Leibler information estimator*

The final robust expected Kullback-Leibler information estimator is given by

$$J_C(\theta)_{robust} = \frac{1}{\sum_{k=1}^N v_k^2} RSS_{robust} + \left(1 + \frac{2[\text{tr}(S^*(x; v_k, \hat{\theta})) + 1]}{n - \text{tr}(S^*(x; v_k, \hat{\theta})) - 2}\right)(\hat{\sigma}_e^2)_{robust} \tag{16}$$

where the smoother matrix $S^*(x; v_k, \hat{\theta})$ is now based on the weighting elements v_k and the robust variance estimator is given by

$$\left(\hat{\sigma}_e^2\right)_{robust} = \frac{1}{M - \lfloor M\beta_2 \rfloor}\sum_{m=\lfloor M\beta_2 \rfloor + 1}^{M-\lfloor M\beta_2 \rfloor}\xi_{M(m)}^2 \tag{17}$$

where $m = 1, ..., M,\ \ M = N-1,\ \ m = k-2$.

The corresponding statistical functional for the robust expected expected Kullback-Leibler information estimator is given by

$$T(F) = \int v^2(\xi)L(\xi)dF(\xi) + A^*\frac{1}{(1-\beta_2)}\int_0^{F^-(1-\beta_2)}\xi^2 dF(\xi) \tag{18}$$

where $A^* = [1 + \frac{2tr(S^*(x,\theta))+2}{N-tr(S^*(x,\theta))-2}]$. ¿From (6) and (18) it follows that $IF(\xi, T_1; F) = v^2(\xi)\xi^2 - T_1(F_N)$ and

$$IF(\xi, T_2; F) = \begin{cases} \frac{\xi^2 - \beta_2 A^*(F^-(1-\beta_2))^2}{(1-\beta_2)} - T_2(F_N) \\ \text{if } |\xi| \le F^{-1}(1-\beta_2) \\ \\ A^*(F^-(1-\beta_2))^2 - T_2(F_N) \\ \text{if } |\xi| > F^-(1-\beta_2). \end{cases}$$

We see that the influence functions are bounded in $\mathbb{R}$. This means that an added observation at a large distance from $T(F_N)$ gives a bounded value in absolute sense for the influence functions.

5. APPLICATION

5.1 *LS-SVM for nonlinear function estimation*

Given a training data set of N points $\{(x_k, y_k)\}_{k=1}^N$ with output data $y_k \in \mathbb{R}$, at input data $x_k \in \mathbb{R}^n$ according to equation (1). where the e_k are i.i.d. random errors with $E[e_k] = 0$, $Var[e_k] = \sigma^2 I_N$ and $f(x)$ is an unknown real-valued regression function that we wish to estimate. One considers the following optimization problem in primal weight space:

$$\min_{w,b,e}\mathcal{J}(w,e) = \frac{1}{2}w^T w + \frac{1}{2}\gamma\sum_{k=1}^N e_k^2 \tag{19}$$

such that $y_k = w^T\varphi(x_k) + b + e_k, \quad k = 1, \ldots, N$ with $\varphi(.) : \mathbb{R}^n \to \mathbb{R}^{n_h}$ a function which maps the input space into a so-called higher dimensional (possibly infinite dimensional) feature space, weight

vector $w \in \mathbb{R}^{n_h}$ in primal weight space, random errors $e_k \in \mathbb{R}$ and bias term b. Note that the cost function $\mathcal{J}$ consists of a RSS fitting error and a regularization term, which is also a standard procedure for the training of MLPs and is related to ridge regression.The relative importance of these terms is determined by the positive real constant γ. In the case of noisy data one avoids overfitting by taking a smaller γ value. Formulations of this form have been investigated in (Suykens *et al.*, 2002*b*). Without the bias term the mathematical solution is equivalent to regularization networks (Poggio and Girosi, 1990).

In primal weight space one has the model $y(x) = w^T\varphi(x) + b$ The weight vector w can be infinite dimensional, which makes a calculation in w from (19) impossible in general. Therefore, one computes the model in the dual space instead of the primal space. One defines the Lagrangian $\mathcal{L}(w,b,e;\alpha) = \mathcal{J}(w,e) - \sum_{k=1}^{N}\alpha_k\left(w^T\varphi(x_k) + b + e_k - y_k\right)$ with Lagrangian multipliers $\alpha_k \in \mathbb{R}$ (called support values). The conditions for optimality are given by $\frac{\partial\mathcal{L}}{\partial w} = \frac{\partial\mathcal{L}}{\partial b} = \frac{\partial\mathcal{L}}{\partial e_k} = \frac{\partial\mathcal{L}}{\partial\alpha_k} = 0$. After elimination of w and e one obtains the solution

$$\left[\begin{array}{c|c} 0 & 1_N^T \\ \hline 1_N & \Omega + \frac{1}{\gamma}I_N \end{array}\right]\left[\begin{array}{c} b \\ \hline \alpha \end{array}\right] = \left[\begin{array}{c} 0 \\ \hline y \end{array}\right] \qquad (20)$$

with $y = [y_1;...;y_N]$, $1_N = [1;...;1]$, $\alpha = [\alpha_1;...;\alpha_N]$ and $\Omega_{kl} = \varphi(x_k)^T\varphi(x_k)$ for $k, l = 1,...,N$. According to Mercer's theorem, there exists a mapping φ and an expansion $K(t,z) = \sum_k \varphi_k(t)\varphi_k(z)$, $t, z \in \mathbb{R}^n$ As a result, one can choose a kernel $K(.,.)$ such that $K(x_k,x_l) = \varphi(x_k)^T\varphi(x_l)$, $k, l = 1,...,N$. The resulting LS-SVM model for function estimation becomes

$$\hat{f}(x;\theta) = \sum_{k=1}^{N}\hat{\alpha}_k K(x,x_k) + \hat{b} \qquad (21)$$

where $\hat{\alpha}$, $\hat{b}$ are the solution to (20). We focus on the choice of an RBF kernel $K(x_k,x_l;h) = \exp\left\{-\|x_k - x_l\|_2^2/h^2\right\}$. One obtains $\hat{f}(x;\theta) = \Omega\alpha + 1_N b = S(x;\theta)y$ where $\theta = (h,\gamma)^T$. The LS-SVM for regression corresponds to

$$S(x;\theta) = \Omega(Z^{-1} - Z^{-1}\frac{J_N}{c}Z^{-1}) + \frac{J_N}{c}Z^{-1} \qquad (22)$$

where $c = 1_N^T(\Omega + \frac{1}{\gamma}I_N)^{-1}1_N$, J_N is a square matrix of 1's and $Z = (\Omega + \frac{1}{\gamma}I_N)$. Therefore, the LS-SVM for regression is an example of a linear smoother. The linear operator $S(x;\lambda,\gamma)$ is known as the smoother matrix. An important property of the smoother matrix based on RBF kernels is that the $\mathrm{tr}[S(x;\theta)] < N$, except in the case $(h \to 0, \gamma \to \infty)$ where $\mathrm{tr}[S(x;\theta)] \to N$.

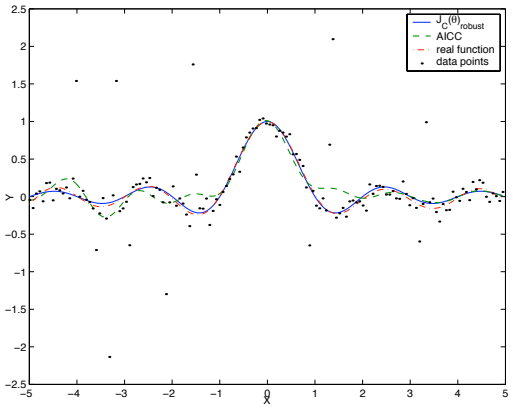

Fig. 1. *Experimental results comparing the LS-SVM tuned by AICC and the weighted LS-SVM tuned by $J_C(\theta)_{robust}$ on a noisy sinc dataset with heavy contamination.*

5.2 *Weighted LS-SVM for nonlinear function estimation*

In order to obtain a robust estimate based upon the previous LS-SVM solution, in a subsequent step, one can weight the error variables $e_k = \alpha_k/\gamma$ by weighting factors v_k (Suykens *et al.*, 2002*a*). This leads to the optimization problem:

$$\min_{w^*,b^*,e^*}\mathcal{J}(w^*,e^*) = \frac{1}{2}w^{*T}w^* + \frac{1}{2}\gamma\sum_{k=1}^{N}v_k e_k^{*2} \qquad (23)$$

such that $y_k = w^{*T}\varphi(x_k) + b^* + e_k^*, \quad k = 1,\ldots,N$ The Lagrangian is constructed in a similar way as before. The unknown variables for this weighted LS-SVM problem are denoted by the $*$ symbol. From the conditions for optimality and elimination of w^*, e^* one obtains the Karush-Kuhn-Tucker system:

$$\left[\begin{array}{c|c} 0 & 1_N^T \\ \hline 1_N & \Omega + V_\gamma \end{array}\right]\left[\begin{array}{c} b^* \\ \hline \alpha^* \end{array}\right] = \left[\begin{array}{c} 0 \\ \hline y \end{array}\right] \qquad (24)$$

where the diagonal matrix V_γ is given by $V_\gamma = \mathrm{diag}\left\{\frac{1}{\gamma v_1},...,\frac{1}{\gamma v_N}\right\}$ The choice of the weights v_k is determined based upon the error variables $e_k = \alpha_k/\gamma$ from the (unweighted) LS-SVM case (20). Robust estimates are obtained then (see (Rousseeuw and Leroy, 1986)) e.g. by taking

$$v_k = \begin{cases} 1 & \text{if } |e_k/\hat{s}| \le c_1 \\ \frac{c_2 - |e_k/\hat{s}|}{c_2 - c_1} & \text{if } c_1 \le |e_k/\hat{s}| \le c_2 \\ 10^{-4} & \text{otherwise} \end{cases} \qquad (25)$$

where $\hat{s} = 1.483\,\mathrm{MAD}(e_k)$ is a robust estimate of the standard deviation of the LS-SVM error variables e_k and MAD stands for the median absolute deviation.

5.3 *Illustrative examples*

Two examples illustrate the advantage of the proposed criteria. The first example is a static regression problem for a noisy sinc with heavy contamination (see Fig. 1). The second example considers a stochastic version of the logistic map $x_{t+1} = cx_t(x_t -$

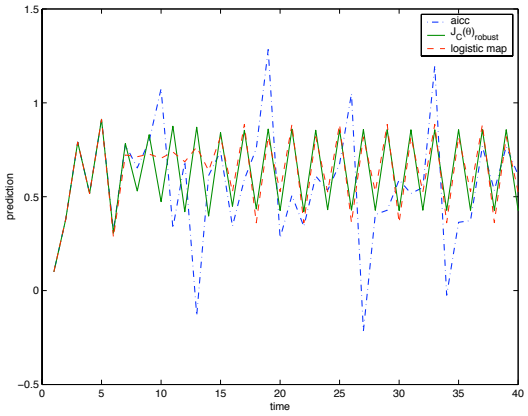

Fig. 2. *Recurrent predictions based on the logistic map, the LS-SVM tuned by AICC and the weighted LS-SVM tuned by* $J_C(\theta)_{robust}$.

Method	L_2	L_1	L_∞
Static Regression			
LS-SVM & AICC:	0.0133	0.0816	0.3241
WLS-SVM & $J_C(\theta)_{robust}$	0.0009	0.0247	0.1012
Stochastic Logistic Map, 40 points predicted recurrently			
LS-SVM & AICC:	0.1088	0.2251	1.0338
WLS-SVM & $J_C(\theta)_{robust}$	0.0091	0.0708	0.2659

Table 1. *Numerical comparison in different norms*

$1) + e_t$ with $c = 3.55$ and contaminating process noise e_t. The recurrent prediction illustrates the difference of the NAR model based on the LS-SVM tuned by AICC and the weighted LS-SVM tuned by $J_C(\theta)_{robust}$ (see Fig 2). The numerical test set performances of both are shown in Table 1 with improved results (in L_2, L_1, L_∞ norm) by applying the robust model selection criteria.

6. CONCLUSION

In this paper, robust complexity criteria based on AIC have been introduced. This can be used to tune the hyperparameters of weighted LS-SVMs in the case of non-Gaussian noise and outliers on the data for function approximation and NARX models.

Acknowledgements. This research work was carried out at the ESAT laboratory of the Katholieke Universiteit Leuven. It is supported by grants from several funding agencies and sources: Research Council KU Leuven: Concerted Research Action GOA-Mefisto 666 (Mathematical Engineering), IDO (IOTA Oncology, Genetic networks), several PhD/postdoc & fellow grants; Flemish Government: Fund for Scientific Research Flanders (several PhD/postdoc grants, projects G.0407.02 (support vector machines), G.0256.97 (subspace), G.0115.01 (bio-i and microarrays), G.0240.99 (multilinear algebra), G.0197.02 (power islands), research communities ICCoS, ANMMM), AWI (Bil. Int. Collaboration Hungary/ Poland), IWT (Soft4s (softsensors), STWW-Genprom (gene promotor prediction), GBOU-McKnow (Knowledge management algorithms), Eureka-Impact (MPC-control), Eureka-FLiTE (flutter modeling), several PhD grants); Belgian Federal Government: DWTC (IUAP IV-02 (1996-2001) and IUAP V-10-29 (2002-2006) (2002-2006): Dynamical Systems and Control: Computation, Identification & Modelling), Program Sustainable Development PODO-II (CP/40: Sustainibility effects of Traffic Management Systems); Direct contract research: Verhaert, Electrabel, Elia, Data4s, IPCOS. JS is a postdoctoral researcher with the National Fund for Scientific Research FWO - Flanders and a professor with KU Leuven. BDM and JVDW are full professors at KU Leuven, Belgium.

7. REFERENCES

Akaike, H. (1973). Statistical predictor identification. *Ann. Inst. Statist. Math.* **22**, 203–217.

Efron, B. (1979). Bootstrap methods: another look at the jackknife. *Ann. of Statist.* **7**(1), 1–26.

Gasser, T., L. Sroka and C. Jennen-Steinmetz (1986). Residual variance and residual pattern in nonlinear regression. *Biometrika* **73**, 625–633.

Hampel, F.R. (1974). The influence curve and its role in robust estimation. *J. Am. Statist. Ass.* **69**, 383–393.

Hannon, E.J and B.G. Quinn (1979). The determination of the order of autoregression. *J. Roy. Statist. Soc. Ser.* **B**(41), 190–195.

Huber, P.J. (1964). Robust estimation of a location parameter. *Ann. Math. Statist.* **35**, 73–101.

Hurvich, C.M. and C.L. Tsai (1989). Regression and time series model selection in small samples. *Biometrika* **76**, 297–307.

Ljung, L. (1987). *System Identification, Theory for the User*. Prentice Hall.

Mallows, C.L. (1973). Some comments on C_p. *Technometrics* **15**, 661–675.

Moody, J.E. (1992). The effective number of parameters: An analysis of generalization and regularization in nonlinear learning systems. In: *Neural Information Processing Systems*. Vol. 4. Morgan Kaufmann. San Mateo CA. pp. 847–854.

Neumann, M.H. (1994). Fully data-driven nonparametric variance estimators. *Statistics* **25**, 189–212.

Poggio, T. and F. Girosi (1990). Networks for approximation and learning. *Proceedings of the IEEE* **78**(9), 1481–1497.

Rousseeuw, P.J. and A.M. Leroy (1986). *Robust Regression and Outlier Detection*. Wiley & sons.

Schwartz, G. (1979). Estimating the dimension of a model. *Ann. of Statist.* **6**, 461–464.

Stone, M. (1974). Cross-validatory choice and assessment of statistical predictions. *J. Roy. Statist. Soc. Ser.* **B**(36), 111–147.

Suykens, J.A.K., J. De Brabanter, l. Lukas and J. Vandewalle (2002*a*). Weighted least squares support vector machines: robustness and sparse approximation. *Neurocomputing* **48**(1-4), 85–105.

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

Vapnik, V.N. (1998). *Statistical Learning Theory*. Wiley and Sons.