Robust Statistics for Kernel based NARX Modeling

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
Tel: 32/16/32 17 09 - Fax: 32/16/32 19 70
Email: johan.suykens@esat.kuleuven.ac.be
† Sint-Lieven Hogeschool (Associatie KULeuven)

*Corresponding author

Abstract

In this paper we study nonlinear ARX models in relation to a class of kernel based models which make use of kernel induced feature spaces, a methodology which is common in the area of support vector machines (SVMs). Methods are proposed for extending the use of least squares support vector machine (LS-SVM) models towards a robust setting. In order to understand the robustness of these estimators against outliers and non-Gaussian noise, we use the influence functions and maxbias curves. Together with robust versions of LS-SVMs, robust counterparts for the final prediction error criteria are proposed. It is discussed how these robust techniques can be applied to the primal as well as dual representations of LS-SVMs. Examples are given on artificial and real-life data.

Keywords. Kernel induced feature spaces, Least Squares Support Vector Machines, Robust Fixed-Size LS-SVM, Weighted LS-SVM, Robust Final Prediction Error Criterion, Influence Functions, Empirical Influence Curves, Maxbias Curves, Breakdown point, NARX models.

1 Introduction

Most efficient learning algorithms in neural networks, support vector machines and kernel based methods (Bishop 1995, Cherkassky and Mulier 1998, Vapnik 1998, Suykens *et al.* 2002*b*) require the tuning of some extra learning parameters, or *tuning parameters*, denoted here by θ. The tuning parameter selection methods can be divided into three broad classes:

1. Cross validation and bootstrap.

2. Plug-in methods. The bias of an estimate of an unknown real-valued smooth function is usually approximated through Taylor series expansions. A pilot estimate of the unknown function is then "plugged in" to derive an estimate of the bias and hence an estimate of the mean integrated squared error. The optimal tuning parameters minimizes this estimated measure of fit.

3. Complexity criteria. Mallows' C_p (Mallows 1973) Akaike's information criterion (Akaike 1973), Final Prediction Error criterion (Akaike 1970), Bayes Information Criterion (Schwartz 1979) and Vapnik-Chervonenkis dimension (Vapnik 1998).

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. 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 (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 (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.

The complexity criteria estimate the generalization error via an estimate of the complexity term and then add it to training error. In contrast, the cross-validation and bootstrap methods, are direct estimates of the generalization error. The tuning parameter selection methods (e.g., Generalized Cross-validation, and other complexity criteria) used in this paper are described in the following sections.

A common view on robustness is to provide alternatives to least squares methods and Gaussian theory. In fact, a statistical procedure based on L_2 works well in situations where many assumptions (such as normality, no outliers) are valid. These assumptions are commonly made, but are usually at best approximations to reality. For example, non-Gaussian noise and outliers are common in data-sets and are dangerous for many statistical procedures (Hampel et al., 1986). Modern theory of robust statistics emerged more recently, led by Huber's (1964) classic minimax approach and Hampel's (1974) introduction of influence functions.

Huber (Huber, 1964) gave the first theory of robustness. He considered also the class of M-estimators of location (also called estimating equation, generalized maximum likelihood estimators) described by some suitable function. The Huber estimator is a minimax solution: it minimizes the maximum asymptotic variance over all F in the gross-error model. The gross-error model can be interpreted as yielding exactly normal data with probability $(1 - \epsilon)$, and gross errors (or some other, "contaminating" distribution) with small probability ϵ (usually between 0% and 5%).

Huber developed a second theory (Huber, 1965, 1968) and (Huber and Strassen, 1973) for censored likelihood ratio tests and exact finite-sample confidence intervals, using more general neighborhoods of the normal model. This approach may be mathematically deepest, but seems very hard to generalize and therefore plays hardly any role in applications. Hampel (Hampel, 1968, 1971, 1974; Hampel et al., 1986) introduced three main concepts:

qualitative robustness, which is essentially continuity of the estimator viewed as functional in the weak topology; the influence curve (IC) or influence function (IF) and the breakdown point (BP).

In this paper we investigate these insights from robust statistics and concepts in relation to a class of kernel models related to support vector machines (SVMs) (Vapnik 1998), more specifically least squares support vector machines (LS-SVMs) (Suykens *et al.* 2002*b*). LS-SVMs have been proposed in (Suykens and Vandewalle 1999, Suykens *et al.* 2002*b*) as a class of kernel machines with primal-dual formulations related to kernel Fisher discriminant analysis (kFDA), kernel ridge regression (kRR), kernel principal component analysis (kPCA), kernel canonical correlation analysis (kCCA), kernel partial least squares (kPLS), recurrent networks and control. As for SVMs an optimization approach is taken to this class of kernel based models which are also closely related to methods of Gaussian processes, kriging and regularization networks. One advantage of the LS-SVM formulations is that the problem can be viewed both in the primal space (as a parametric estimation problem, which is more convenient for handling large data sets) and the dual space (semi-parametric problem). Robust kernel based modeling for nonlinear ARX models is investigated in this paper in relation to the primal as well as dual problem of the LS-SVM formulations.

In Section 2 we discuss the considered nonlinear model structure. Section 3 reviews a number of aspects of kernel induced feature spaces and support vector machines with primal-dual representations and least squares support vector machines. In Section 4 a number of model selection criteria are reviewed. In Section 5 robust estimation is discussed together with robust model selection criteria in relation to least squares support vector machines. In Section 6 the approach is successfully demonstrated on artificial and real-life data sets.

2 Model Structure

Given a measurement vector $y(t) \in \mathbb{R}$ and a deterministic input vector $u(t) \in \mathbb{R}$ from a SISO system. We assume that the measurement vector $y(t)$ is an autoregressive stochastic process $\{y(t), t \in \mathbb{Z}\}$ where:

- $E[y^2(t)] < \infty$, $\forall t \in \mathbb{Z}$
- $E[y(t)] = \mu$, $\forall t \in \mathbb{Z}$
- $\text{cov}[y(k), y(l)] = \text{cov}[y(k+t), y(l+t)]$, $\forall k, l, t \in \mathbb{Z}$

Consider a nonlinear autoregressive model with eXogeneous inputs, defined as

$$y(t) = f(y(t-1), \ldots, y(t-q), u(t-1), \ldots, u(t-p)) + e(t) \tag{1}$$

where $f : \mathbb{R}^{q+p} \to \mathbb{R}$ is an unknown smooth function. The measurement noise $e(t) \in \mathbb{R}$ is a stochastic process $\{e(t), t \in \mathbb{Z}\} \sim \text{W}\mathcal{N}(0, \sigma^2)$ and $\text{cov}[e(i), e(j)] = 0$, $\forall i \neq j$. The predicted output is equal to

$$\hat{y}(t) = f\left(y(t-1), \ldots, y(t-q), u(t-1), \ldots, u(t-p)\right). \tag{2}$$

We will now consider a class of functions $\mathcal{F}$ for the model structure, in relation to kernel methods with primal-dual representations in view of support vector machines.

3 Kernel Induced Feature Spaces and Support Vector Machines

3.1 Primal dual representation

Let $X \in \mathcal{X} \subseteq \mathbb{R}^d$ denote a real valued random input vector, and $Y \in \mathcal{Y} \subseteq \mathbb{R}$ a real valued random output variable and let $\Psi \subseteq \mathbb{R}^{n_f}$ denote a high-dimensional feature space. A central idea in support vector machines is the following: it maps the random input vector into the high-dimensional feature space Ψ through some nonlinear mapping $\varphi : \mathcal{X} \to \Psi$. In this space, one considers the class of functions

$$\mathcal{F}_\Psi = \left\{ f : f(x) = w^T \varphi(x) + b,\ \varphi : \mathcal{X} \to \Psi,\ w \in \mathbb{R}^{n_f},\ b \in \mathbb{R} \right\}. \tag{3}$$

However, even if the linear function in the feature space (3) generalizes well and can theoretically be defined, the problem of how to treat the high-dimensional feature space remains. Note that for constructing the linear function (3) in the feature space Ψ, one

does not need to consider the feature space in explicit form. One only replaces the inner product in the feature space $\varphi(x_k)^T\varphi(x_l)$ with the corresponding kernel $K(x_k, x_l)$ satisfying Mercer's condition (Mercer 1909) .

Theorem 1 (Mercer) *Let $K \in L^2(C)$, $g \in L^2(C)$ where C is a compact subset of $\mathbb{R}^d$ and $K(t, z)$ describes an inner product in some feature space. To guarantee that a continuous symmetric function K has an expansion*

$$K(t, z) = \sum_{k=1}^{\infty} a_k \phi_k(t)\phi_k(z) \tag{4}$$

with positive coefficients $a_k > 0$ (i.e. $K(t, z)$ describes an inner product in some feature space), it is necessary and sufficient that the condition

$$\int_C \int_C K(t, z)g(t)g(z)dtdz \geq 0 \tag{5}$$

be valid for all $g \in L^2(C)$.

Assume $w = \sum_{k=1}^{n} \alpha_k \varphi(x_k)$ and based on Mercer theorem, the class of linear functions in feature space (3) has the following equivalent representation in input space $\mathcal{X}$:

$$\mathcal{F}_{\mathcal{X}} = \left\{ f : f(x) = \sum_{k=1}^{n} \alpha_k K(x, x_k) + b, b \in \mathbb{R}, \alpha_k \in \mathbb{R} \right\} \tag{6}$$

where x_k are vectors and $K(x, x_k)$ is a given function satisfying Mercer's condition.

In the primal space the parameter w is not a Euclidean vector but ranges over an "infinite-dimensional" parameter set, while in the dual problem the dimension of the parameter vector α grows with the given number of training points.

3.2 LS-SVM regression: the unweighted case

Consider now the case where there is noise in the description of the functions. Given a training set defined as $\mathcal{D}_n = \{(x_k, y_k) : x_k \in \mathcal{X}, y_k \in \mathcal{Y}\}_{k=1}^{n}$ of size n drawn i.i.d. from an unknown distribution F_{XY} according to

$$y_k = f(x_k) + e_k, k = 1, \ldots, n \tag{7}$$

where $e_k \in \mathbb{R}$ are assumed to be i.i.d. random errors with $E[e_k|X = x_k] = 0$, $\mathrm{Var}[e_k] = \sigma^2 < \infty$, $f(x) \in \mathcal{F}_\Psi$ is an unknown real-valued smooth function and $E[y_k|x = x_k] = f(x_k)$. Our goal is to find the parameters w and b (primal space) that minimize the empirical risk functional

$$\mathcal{R}_n(w, b) = \frac{1}{n} \sum_{k=1}^{n} \left((w^T \varphi(x_k) + b) - y_k \right)^2 \tag{8}$$

under constraint $\|w\|_2 \leq a$, $a \in \mathbb{R}_+$. One can approach the optimization problem of finding the vector w and $b \in \mathbb{R}$ to solve the following optimization problem

$$\min_{w,b,e} \mathcal{J}(w, e) = \frac{1}{2} w^T w + \frac{\gamma}{2} \sum_{k=1}^{n} e_k^2 \tag{9}$$

such that

$$y_k = w^T \varphi(x_k) + b + e_k, k = 1, \ldots, n. \tag{10}$$

Note that the cost function $\mathcal{J}$ consists of a Residual Sum of Squares (RSS) and a regularization term, which is also a standard procedure for the training of multi-layer perceptrons and is related to ridge regression (Golub and Van Loan 1989). 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. These formulations of this form have been investigated independently in (Saunders *et al.* 1998) (without bias term) and (Suykens and Vandewalle 1999).

To solve the optimization problem (in the dual space) one defines the Lagrangian functional $\mathcal{L}(w, b, e; \alpha) = \mathcal{J}(w, e) - \sum_{k=1}^{n} \alpha_k (w^T \varphi(x_k) + b + e_k - y_k)$ with Lagrangian multipliers $\alpha_k \in \mathbb{R}$ (called support values). The conditions for optimality are given by $\frac{\partial \mathcal{L}}{\partial w} = 0$, $\frac{\partial \mathcal{L}}{\partial b} = 0$, $\frac{\partial \mathcal{L}}{\partial e_k} = 0$, $\frac{\partial \mathcal{L}}{\partial \alpha_k} = 0$. After elimination of w, e one obtains the solution

$$\left[\begin{array}{c|c} 0 & 1_n^T \\ \hline 1_n & \Omega + \frac{1}{\gamma} I_n \end{array} \right] \left[\begin{array}{c} b \\ \hline \alpha \end{array} \right] = \left[\begin{array}{c} 0 \\ \hline y \end{array} \right] \tag{11}$$

with $y = (y_1, ..., y_n)^T$, $1_n = (1, ..., 1)^T$, $\alpha = (\alpha_1; ...; \alpha_n)^T$ and $\Omega_{kl} = \varphi(x_k)^T \varphi(x_l)$ for k, $l = 1, ..., n$. According to Mercer's theorem, the resulting LS-SVM model for function estimation becomes

$$\hat{f}(x) = \sum_{k=1}^{n} \hat{\alpha}_k K(x, x_k) + \hat{b}, \tag{12}$$

where $\hat{\alpha}$, $\hat{b}$ are the solution to (11).

In this paper, we focus on the choice of an RBF kernel $K(x_k, x_l; h) = \exp\{-\|x_k - x_l\|_2^2/h^2\}$. Let $\theta = (h, \gamma)^T$ and for all training data $\{x_k, y_k\}_{k=1}^n$, one obtains $\hat{f}_n = \Omega\alpha + 1_n b = S(\theta)y$. The LS-SVM for regression corresponds to the case with $\hat{f}_n$ defined by (13) and

$$S(\theta) = \Omega\left(Z^{-1} - Z^{-1}\frac{J_N}{c}Z^{-1}\right) + \frac{J_n}{c}Z^{-1}. \tag{13}$$

where $c = 1_n^T\left(\Omega + \frac{1}{\gamma}I_n\right)^{-1}1_n$, $Z = \left(\Omega + \frac{1}{\gamma}I_n\right)$, J_n is a square matrix with all elements equal to 1, $y = (y_1, \ldots, y_n)^T$ and $\hat{f}_n = (\hat{f}_n(x_1), \ldots, \hat{f}_n(x_n))^T$. As the estimated function in (13) is a linear transformation of the vector y, the LS-SVM for regression is an example of a linear smoother. By analogy one defines the effective degrees of freedom of the LS-SVM for regression (effective number of parameters) to be

$$d_{\text{eff}}(\theta) = \text{tr}[S(\theta)]. \tag{14}$$

An important property of the smoother matrix based on a RBF kernel is that the $\text{tr}[S(\theta)] < n$, except in the case ($h \to 0$, $\gamma \to \infty$) where $\text{tr}[S(\theta)] \to n$.

3.3 Fixed-size LS-SVM

3.3.1 Estimation in primal weight space

For large data sets it is advantageous if one solves the problem in the primal space (Suykens *et al.* 2002*a*). However, one then needs explicitly an expression for some nonlinear mapping $\varphi : \mathcal{X} \to \Psi$. Let $x_k \in \mathbb{R}^d$, $\forall k = 1, \ldots, n$ be a random sample from an unknown distribution $F_X(x)$. Let C be a compact subset of $\mathbb{R}^d$, let $V = L^2(C)$ and let $M(V, V)$ be a class of linear operators from V into V. Consider the eigenfunction expansion of a kernel function

$$K(x, z) = \sum_{i=1}^{s} \lambda_i \phi_i(x)\phi_i(z), \tag{15}$$

where $s \leq \infty$, $K(x, u) \in V$, $\lambda_i \in \mathbb{C}$ and $\phi_i \in V$ are respectively the eigenvalues and the eigenfunctions, defined by the Fredholm integral equation of the first kind

$$\begin{aligned}(T\phi_i)(z) &= \int_C K(x, z)\phi_i(x)dF_X(x) \\ &= \lambda_i\phi_i(z)\end{aligned} \tag{16}$$

where $T \in M(V,V)$.

One can discretize (16) on a finite set of evaluation points $\{x_1, \ldots, x_n\} \in C$ with associated weights $v_k \in \mathbb{R}$, $k = 1, \ldots, n$. Define a quadrature method Q^n for an $n \in \mathbb{N}$

$$Q^n = \sum_{k=1}^{n} v_k \psi(x_k). \tag{17}$$

Let $v_k = \frac{1}{n}$, $k = 1, \ldots, n$, then the Nyström method (Williams and Seeger 2001) approximates the integral by means of Q_n and determines an approximation λ_i^C and ϕ_i^C by

$$\lambda_i^C \phi_i^C(z) \approx \frac{1}{n} \sum_{k=1}^{n} K(x_k, z) \phi_i^C(x_k), \forall t \in C. \tag{18}$$

Let $z = x_j$ and denote the corresponding approximation by λ_i^n and ϕ_i^n, in matrix notation one obtains then

$$\Omega U_{n \times n} = U_{n \times n} \Lambda_{n \times n} \tag{19}$$

where $\Omega_{kj} = K(x_k, x_j)$ are the elements of the kernel matrix, $U_{n \times n} = (u_1^n, \ldots, u_n^n)$ is a $n \times n$ matrix of eigenvectors of Ω and $\Lambda_{n \times n}$ is a $n \times n$ diagonal matrix of nonnegative eigenvalues in a decreasing order $\Lambda_{n \times n} = \mathrm{diag}(\lambda_1^n, \ldots, \lambda_n^n)$ and $\lambda_1^n \geq \cdots \geq \lambda_n^n$. Expression (18) delivers direct approximations of the eigenvalues and eigenfunctions for the $x_k \in \mathbb{R}^d$, $k = 1, \ldots, n$ points

$$\phi_i(x_j) \approx \sqrt{n} u_{ij}^n \tag{20}$$

and

$$\lambda_i \approx \frac{1}{n} \lambda_i^n, \tag{21}$$

Substituting (20) and (21) in (18) gives an approximation $\hat{\phi}_i$ of an eigenfunction evaluation in point $x \in C$

$$\hat{\phi}_i(x) \approx \frac{\sqrt{n}}{\lambda_i^n} \sum_{k=1}^{n} K(x_k, x) u_{ik}^n. \tag{22}$$

One obtains, based on the Nyström approximation, an explicit expression for the entries of the approximated nonlinear mapping $\varphi_i : \mathcal{X} \rightarrow \Psi$:

$$\begin{aligned} \hat{\varphi}_i^n(x) &= \sqrt{\lambda_i} \hat{\phi}_i(x) \\ &= \frac{1}{\sqrt{\lambda_i^n}} \sum_{k=1}^{n} K(x_k, x) u_{ik}^n. \end{aligned} \tag{23}$$

In order to introduce parsimony, one chooses as fixed size n_{FS} with $n_{\mathrm{FS}} \ll n$ for a working subsample. A likewise n_{FS}-approximation $\hat{\varphi}^{n_{\mathrm{FS}}}$ can be made and the model takes the form

$$\begin{aligned} f^{n_{\mathrm{FS}}}(x) &= w^T \hat{\varphi}^{n_{\mathrm{FS}}}(x) + b \\ &= \sum_{i=1}^{n_{\mathrm{FS}}} w_i \frac{1}{\sqrt{\lambda_i^{n_{\mathrm{FS}}}}} \sum_{k=1}^{n_{FS}} K(x_k, x) u_{ik}^{n_{FS}} + b. \end{aligned} \tag{24}$$

One can solve now the following ridge regression problem in the primal weight space with unknowns $w \in \mathbb{R}^{n_{FS}}$, $b \in \mathbb{R}$

$$\min_{w,b} \frac{1}{2} \sum_{i=1}^{n_{\mathrm{FS}}} w_i^2 + \gamma \frac{1}{2} \sum_{k=1}^{n} \left(y_k - \sum_{i=1}^{n_{\mathrm{FS}}} (w_i \hat{\varphi}_i^{n_{\mathrm{FS}}}(x_k) + b) \right)^2. \tag{25}$$

This approach gives explicit links between primal and the dual space representation. Such fixed-size LS-SVM algorithms have provided high quality nonlinear models e.g. on the silver box identification problem (Espinoza *et al.* 2004).

3.3.2 Active selection of a subsample

In order to make a more suitable selection of the subsample instead of a random selection, one can relate the Nyström method to an entropy criteria as proposed in (Suykens *et al.* 2002*b*) based on (Girolami 2002). Let $x_k \in \mathbb{R}^d$, $k = 1, \ldots, n$, be a set of input samples from a random variable $X \in \mathbb{R}^d$. The success of a selection method depends on how much information about the original input sample $x_k \in \mathbb{R}^d$, $k = 1, \ldots, n$ is contained in a subsample $x_j \in \mathbb{R}^d$, $j = 1, \ldots, n_{\mathrm{FS}}$ such that $n_{FS} \ll n$. The information content of a sample of size n is measured by its entropy denoted as $H_n(X)$. Thus, the purpose of a subsample selection is to extract n_{FS} samples from $\{x_1, \ldots, x_n\}$ such that $H_{n_{\mathrm{FS}}}(X)$ comes close to $H_n(X)$.

One estimates the density function $f(x)$ by the Parzen kernel density estimation

$$\hat{f}(x) = \frac{1}{n_{\mathrm{FS}} h^d} \sum_{k=1}^{n_{\mathrm{FS}}} K\left(\frac{x - x_k}{h} \right) \tag{26}$$

where h is the bandwidth and the kernel $K : \mathbb{R}^d \to \mathbb{R}$ satisfying $\int_{\mathbb{R}^d} K(u)du = 1$. When

the differential entropy (Shannon 1948) defined by

$$\begin{aligned} H_S(X) &= E\left[-\log f(x^{(1)},\ldots,x^{(d)})\right] \\ &= -\int\ldots\int f(x^{(1)},\ldots,x^{(d)})\log f(x^{(1)},\ldots,x^{(d)})\; dx^{(1)}\ldots dx^{(d)} \end{aligned} \tag{27}$$

is used along with the kernel density estimate $\hat{f}(x)$, the estimation of the entropy $H_{S,n_{FS}}(X)$ becomes very complex. But Renyi's entropy of order $q = 2$ (also called quadratic entropy) leads to a simpler estimate of entropy $H_{S,n_{FS},R_2}(X)$. Renyi's entropy of order q is defined as:

$$H_{R_q}(X) = -\frac{1}{1-q}\log\int f(x^{(1)},\ldots,x^{(d)})^q dx^{(1)}\ldots dx^{(d)},\;\; q>0, q\neq 1. \tag{28}$$

The differential entropy can be viewed as one member of the Renyi's entropy family, because $\lim_{q\to 1} H_{R_q}(X) = H_S(X)$. Although Shannon's entropy is the only one which possesses all properties (e.g., continuity, symmetry, extremal property, recursiveness and additivity) for an information measure, the Renyi's entropy family is equivalent with regards to entropy maximization. In real problems, which information measure to use depends upon other requirements such as ease of implementation. Combining (26) and (28), Renyi's quadratic entropy estimator becomes

$$\begin{aligned} H_{n_{\mathrm{FS}},R_2}(X) &= -\log\frac{1}{n_{\mathrm{FS}}^2 h^{2d}}\sum_{k=1}^{n_{\mathrm{FS}}}\sum_{l=1}^{n_{\mathrm{FS}}} K(\frac{x_k - x_l}{h}) \\ &= -\log\frac{1}{n_{\mathrm{FS}}^2 h^{2d}} 1_{n_{\mathrm{FS}}}^T \Omega 1_{n_{\mathrm{FS}}}. \end{aligned} \tag{29}$$

One chooses a fixed size n_{FS} ($n_{\mathrm{FS}} \ll n$) for a working set of data points and actively selects points from the pool of training input samples as a candidate for the working set. In the working set a point is randomly selected and replaced by a randomly selected point from the training input sample if the new point improves Renyi's quadratic entropy criterion. This leads to the following fixed size LS-SVM algorithm:

Algorithm 1 (Fixed size LS-SVM)

1. *Given a training set* $\mathcal{D}_n = \{(x_1, y_1),\ldots,(x_n, y_n)\}$, *construct a standardized input training set* $\mathcal{S}_n = \left\{\tilde{x} : \tilde{x} = \frac{x - E[x]}{\mathrm{Var}[x]}, \tilde{x}_k \in \mathcal{X}, k = 1,\ldots,n\right\}$, *choose a working set* $\mathcal{W}_{n_{\mathrm{FS}}} = \{\tilde{x} : \tilde{x}_j \in \mathcal{X}; j = 1,\ldots,n_{\mathrm{FS}} \ll n\} \subset \mathcal{S}_n$.

2. *Randomly select a sample point* $\tilde{x}^* \in \mathcal{W}_l$ *and* $\tilde{x}^{**} \in \mathcal{S}_n$*, swap* $(\tilde{x}^*, \tilde{x}^{**})$.

3. *If* $H_{n_{\text{FS}},R_2}(\tilde{x}_1, \ldots, \tilde{x}_{n_{\text{FS}}-1}; \tilde{x}^{**}) > H_{n_{\text{FS}},R_2}(\tilde{x}_1, \ldots, \tilde{x}_i^*, \ldots, \tilde{x}_{n_{\text{FS}}})$ *then* $\tilde{x}^{**} \in \mathcal{W}_{n_{\text{FS}}}$*,* $\tilde{x}^* \notin \mathcal{W}_{n_{\text{FS}}}$*,* $\tilde{x}^* \in \mathcal{S}_n$.

4. *Calculate* $H_{n_{FS},R_2}(\tilde{x})$ *for the present* $\mathcal{W}_{n_{\text{FS}}}$

5. *Stop if the change in entropy value (29) is small.*

6. *Estimate* w, b *in the primal space after estimating the eigenfunctions from the Nyström approximation according to (25).*

4 Model Selection

4.1 Final Prediction Error (FPE) criterion, Mallows' C_p, AIC and BIC

Let $\mathcal{P}$ be a finite set of parameters. For $\beta \in \mathcal{P}$, let $\mathcal{F}_\beta$ be a set of functions

$$\mathcal{F}_\beta = \left\{ f : f(x, \beta) = \beta_0 + \sum_{l=1}^{d} \beta_l x^{(l)}, x \in \mathbb{R}^d \text{ and } \beta \in \mathcal{P} \right\} \tag{30}$$

where $x^{(l)}$ denotes the l-th component of $x \in \mathbb{R}^d$. let $Q_n(\beta) \in \mathbb{R}^+$ be a complexity term for $\mathcal{F}_{n,\beta}$ and let $\hat{f}$ be an estimator of f in $\mathcal{F}_{n,\beta}$. The learning parameters are chosen to be the minimizer of a cost function defined as

$$J_\beta(\lambda) = \frac{1}{n} \sum_{k=1}^{n} L(y_k, \hat{f}(x_k; \beta)) + \lambda(Q_n(\beta))\hat{\sigma}_e^2 \tag{31}$$

where $\sum_{k=1}^{n} L(y_k, \hat{f}(x_k; \beta))$ is the residual sum of squares (RSS), $Q_n(\beta)$ is a complexity term, $\lambda > 0$ is a cost complexity parameter and the term $\hat{\sigma}_e^2$ is an estimate of the error variance. The Final Prediction Error criterion depends only on $\hat{f}$ and the data. If $\hat{f}$ is defined by minimizing the empirical L_2 risk over some linear vector space $\mathcal{F}_{n,\beta}$ of functions with dimension d_β, then $J_\beta(\lambda)$ will be of the form:

- (AIC): Let $\lambda = 2$ and $Q_n(\beta) = n^{-1}d_\beta$

$$C_p(\lambda) = \frac{1}{n}RSS + 2\left(\frac{d_\beta}{n}\right)\hat{\sigma}_e^2. \tag{32}$$

The Akaike information criterion (AIC) is a similar but more generally applicable estimate when a log-likelihood loss function is used. For the Gaussian model (with variance $\sigma_e^2 = \hat{\sigma}_e^2$ assumed known), the AIC statistic is equivalent to C_p.

- (BIC): Let $\lambda = \log n$ and $Q_n(\beta) = n^{-1}d_\beta$

$$\mathrm{BIC}(\lambda) = \frac{1}{n}RSS + (\log n)\left(\frac{d_\beta}{n}\right)\hat{\sigma}_e^2$$

- (J_1): Let $\lambda = \log\log n$ and $Q_n(\beta) = n^{-1}d_\beta$

$$J_1(\lambda) = \frac{1}{n}RSS + (\log\log n)\left(\frac{d_\beta}{n}\right)\hat{\sigma}_e^2,$$

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

- (AICC): 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 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

$$\mathrm{AICC}_\beta(\lambda) = \frac{1}{n}RSS + \left(1 + \frac{2(d_\beta + 1)}{n - d_\beta - 2}\right)\hat{\sigma}_e^2 \tag{33}$$

where $\lambda = 1$ and $Q_n(\beta) = 1 + \frac{2(d_\beta+1)}{n-d_\beta-2}$.

For $\theta \in \mathcal{Q}$, let $\mathcal{F}_{n,\theta}$ be a set of functions

$$\mathcal{F}_{n,\theta} = \left\{ f : f(x,\theta), x \in \mathbb{R}^d, y \in \mathbb{R}^n, \theta \in \mathcal{Q}, \hat{f}_n(\hat{\theta}) = S(\hat{\theta})y \right\}, \tag{34}$$

let $Q_n(\theta) \in \mathbb{R}^+$ be a complexity term for $\mathcal{F}_{n,\theta}$ and let $\hat{f}$ be an estimator of f in $\mathcal{F}_{n,\theta}$. For example, regression spline estimators, wavelet and LS-SVM estimators are linear estimators, in the sense that $\hat{f}_n(\hat{\theta}) = S(\hat{\theta})y$, where the matrix $S(\hat{\theta})$ is called the smoother matrix

(13). Based on (Moody 1992) and by analogy, the learning parameters are chosen to be the minimizer of a more generalized cost function defined as

$$JC_\theta(\lambda) = \frac{1}{n} RSS + \left(1 + \frac{2\mathrm{tr}(S(\hat{\theta})) + 2}{n - \mathrm{tr}(S(\hat{\theta})) - 2}\right) \hat{\sigma}_e^2. \tag{35}$$

Each of these selectors depends on $S(\hat{\theta})$ through its trace ($\mathrm{tr}(S(\hat{\theta})) < n-2$), which can be interpreted as the effective number of parameters used in the fit (14).

4.2 Generalized cross-validation

The GCV criterion was first proposed by (Craven and Wahba 1979) for the use in the context of a nonparametric regression with a roughness penalty. However, (Golub *et al.* 1979) showed that GCV can be used to solve a wide variety of problems involving estimation of minimizers for γ in (9):

$$\mathrm{GCV}(\theta) = \frac{1}{n} \frac{\sum_{k=1}^{n} (y_k - \hat{f}(x_k;\theta))^2}{(1 - n^{-1}\mathrm{tr}[S(\theta)])^2}. \tag{36}$$

5 Robust Methods

The theory of robust inference studies the effect of outliers or non-Gaussian noise models on given estimator. Let x be a continuous random variable. The general gross-error model or ϵ-contamination model is used to formalize the noise model

$$\mathcal{U}(F_0, \epsilon) = \{F : F(x) = (1-\epsilon)F_0(x) + \epsilon G(x), 0 \leq \epsilon \leq 1\} \tag{37}$$

where $F_0(x)$ is some given distribution (the ideal nominal model), $G(x)$ is an arbitrary continuous distribution and ϵ is the first parameter of contamination. The contamination scheme describes the case where, with large probability $(1-\epsilon)$ the data occurs with distribution $F_0(x)$ and with small probability ϵ outliers occur according to the distribution $G(x)$.

In order to understand why certain estimators behave the way they do in the case of non-Gaussian noise models and outliers, it is necessary to look at the various measures of robustness.

5.1 Measures of Robustness

There exist a large variety of approaches towards the robustness problem. The approach based on influence functions (Hampel 1968, Hampel 1974) will be used here. The effect of one outlier on the estimator can be described by the influence function (IF) which (roughly speaking) formalizes the bias caused by one outlier. The breakdown value characterizes how much contaminated data an estimator can tolerate before it becomes useless. A more quantitative measure of robustness of an estimator is the maxbias curve, which gives the maximal bias that an estimator can suffer from when a fraction of the data come from a contaminated distribution. By letting the fraction vary between zero and the breakdown value a curve is obtained.

5.1.1 Influence functions and breakdown points

Let F be a fixed distribution and $T(F)$ a statistical functional defined on a set $\mathcal{U}$ of distributions satisfying some regularity conditions (Hampel *et al.* 1986). Let the estimator $T_n = T(\hat{F}_n)$ of $T(F)$ be the functional of the sample distribution F_n.

Definition 1 (Influence function) *The influence function $IF(x; T, F)$ is defined as*

$$IF(x; T, F) = \lim_{\epsilon \downarrow 0} \frac{T\left[(1-\epsilon)F + \epsilon\Delta_x\right] - T(F)}{\epsilon}. \tag{38}$$

Here Δ_x denotes 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. Note that this kind of differentiation of statistical functionals is a differentiation in the sense of von Mises (Fernholz 1983). From the influence function, several robustness measures can be defined: the gross error sensitivity, the local shift sensitivity and the rejection point (Hampel 1968, Hampel 1974). Mathematically speaking, the influence function is the set of all partial derivatives of the functional T in the direction of the point masses. For functionals, there exist several concepts of differentiation; Gâteaux, Hadamard of compact, and Fréchet derivative have been used in statistics, the Fréchet derivative being the strongest concept and previously considered to be very rarely applicable; but the main reason for this belief seems to be the non-robustness of most

classical estimators, while at least some (if not most) smooth M-estimators are indeed Fréchet-differentiable (Bedmarski 1993). The IF describes the derivative of a functional in whatever sense it exists.

Definition 2 (Maxbias curve) *Let $T(F)$ denote a statistical functional and let the contamination neighborhood of F be defined by (37) for a fraction of contamination ϵ. The maxbias curve is then defined by*

$$B(\epsilon; T, F) = \sup_{F \in \mathcal{U}(F_0, \epsilon)} \left| \frac{T(F) - T(F_0)}{\sigma_0} \right| \tag{39}$$

where σ_0 denotes the standard deviation of the uncontaminated distribution F_0.

Definition 3 (Breakdown point) *The breakdown point ϵ^* of the estimator $T_n = T(\hat{F}_n)$ for the functional $T(F)$ at F is defined by*

$$\epsilon^*(T, F) = \inf \{ \epsilon > 0 \,|B(\epsilon, T(F), F) = \infty \} . \tag{40}$$

This notion defines the largest fraction of gross errors that still keeps the bias bounded.

5.1.2 Empirical influence curves

The most important empirical versions of (38) are the sensitivity curve (Tukey 1970) and the Jackknife (Quenouille 1949) and (Tukey 1958). There are two versions, one with addition and one with replacement.

In the case of an additional observation, one starts with a sample $(x_1, \ldots, x_{n-1})$ of size $n-1$. Let $T(F)$ be an estimator and let $T(\hat{F}_{n-1}) = T(x_1, \ldots, x_{n-1})$ denote the estimate. The change in the estimate when the nth observation $x_n = x$ is included is $T(x_1, \ldots, x_{n-1}, x) - T(x_1, \ldots, x_{n-1})$. One multiply the change by n and the result is the sensitivity curve.

Definition 4 (sensitivity curve) *One obtains the sensitivity curve if one replaces F by $\hat{F}_{n-1}$ and ϵ by $\frac{1}{n}$ in (38):*

$$\begin{aligned} SC_{n-1}(x; T, \hat{F}_{n-1}) &= \frac{T\left[\frac{n-1}{n}\hat{F}_{n-1} + \frac{1}{n}\Delta_x\right] - T(\hat{F}_{n-1})}{\frac{1}{n}} \\ &= (n-1)T(\hat{F}_{n-1}) + T(\Delta_x) - nT(\hat{F}_{n-1}) \\ &= n\left[T_n(x_1, \ldots, x_{n-1}, x) - T_{n-1}(x_1, \ldots, x_{n-1})\right] . \end{aligned} \tag{41}$$

Example 1 (median) *The sample median is defined by*

$$\text{med} = \begin{cases} x_{n(k+1)} & \text{if } n = 2k+1 \\ \frac{1}{2}(x_{n(k)} + x_{n(k+1)}) & \text{if } n = 2k \end{cases} \tag{42}$$

where $x_{n(1)} \leq \cdots \leq x_{n(n)}$ *are the order statistics. The sensitivity curve of the mean is then*

$$SC_{n-1}(x; \text{med}, \hat{F}_{n-1}) = n(\text{med}(x_1, \ldots, x_{n-1}, x) - \text{med}(x_1, \ldots, x_{n-1})). \tag{43}$$

Depending on the rank of x*, the sensitivity curve of the median is given by*

$$SC_{n-1}(x; \text{med}, \hat{F}_{n-1}) = \begin{cases} n(x_{(k)} - \text{med}(x_1, \ldots, x_{n-1})) & \text{for } x < x_{(k)} \\ x & \text{for } x_{(k)} \leq x \leq x_{(k+1)} \\ n(x_{(k+1)} - \text{med}(x_1, \ldots, x_{n-1})) & \text{for } x > x_{(k+1)} \end{cases} \tag{44}$$

Given an univariate data set $x = (x_1, \ldots, x_{100})^T$ where $x \sim \mathcal{N}(0, 1^2)$. Location estimators applied to this sample are the sample mean and the sample median. We show the sensitivity curve for the location estimators in Figure 2. The most important aspect is that the sensitivity curve of the mean becomes unbounded for both $x \to \infty$ and $x \to -\infty$, whereas the median remains constant.

Example 2 (spread) *Let* $T(F) = \sigma^2$ *denote the variance in a population and let* $x_1, \ldots, x_n$ *denote a sample from that population. Then* $\hat{\sigma}^2 = \frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x}_n)^2$ *is the maximum likelihood estimate of* σ^2*. Shift the horizontal axis such that* $\sum_{k=1}^{n-1} x_k = 0$*. The sensitivity curve of the variance is then*

$$\begin{aligned} SC_{n-1}(x; \hat{\sigma}^2, \hat{F}_{n-1}) &= n(\hat{\sigma}_n^2 - \hat{\sigma}_{n-1}^2) \\ &= n\left(\frac{1}{n}\sum_{i=1}^{n-1}(x_i - \bar{x}_n)^2 + (x - \bar{x}_n)^2 - \hat{\sigma}_{n-1}^2\right) \\ &= \left(\sum_{i=1}^{n-1}(x_i - \bar{x}_n)^2 + \left(\frac{n-1}{n}\right)x^2 - n\hat{\sigma}_{n-1}^2\right) \\ &= (n-1)\hat{\sigma}_{n-1}^2 + \left(\left(\frac{n-1}{n}\right) + \frac{1}{n^2}\right)x^2 - n\hat{\sigma}_{n-1}^2 \\ &= \left(\left(\frac{n-1}{n}\right) + \frac{1}{n^2}\right)x^2 - \hat{\sigma}_{n-1}^2 \simeq x^2 - \hat{\sigma}_{n-1}^2. \end{aligned}$$

Scale estimators applied to the sample $x = (x_1, \ldots, x_{100})^T$ where $x \sim \mathcal{N}(0, 1^2)$, are the sample variance, the Mean Deviation and the Median Absolute Deviation. The Mean Deviation is defined as

$$T(\hat{F}_n) = \frac{1}{n} \sum_{k=1}^{n} |x_k - \bar{x}|. \tag{45}$$

This estimator is non-robust to outliers and has a breakdown point $\epsilon^* = 0$. The Median Absolute Deviation (MAD), one of the more robust scale estimators, is defined as

$$T(\hat{F}_n) = \text{med}_{k=1,\ldots,n} |x_k - \text{med}_{l=1,\ldots,n}(x_l)|. \tag{46}$$

This estimator has a high breakdown point $\epsilon^* = 0.5$. We graph the sensitivity curve for the scale estimators in Figure 4.

The most important aspect is that the sensitivity curves of the variance and the Mean Deviation become unbounded for both $x \to \infty$ and $x \to -\infty$, whereas the sensitivity curve of the Median Absolute Deviation is bounded.

Another approach to approximating the IF, but only at the sample values $x_1, \ldots, x_n$ themselves, is the Jackknife.

Definition 5 (The Jackknife approximation) *If one substitutes $\hat{F}_n$ for F and $-\frac{1}{(n-1)}$ for ϵ in (38), one obtains*

$$\begin{aligned} J_{IF}(x_i; T, F_n) &= \frac{T\left[\left(\frac{n}{n-1}\right) F_n - \frac{1}{n-1}\Delta_{x_i}\right] - T(F_n)}{\frac{-1}{n-1}} \\ &= -(n-1)\left[(T(F_n) - T(\Delta_{x_i})) - T(F_n)\right] \\ &= (n-1)\left[T_n(x_1, \ldots, x_n) - T_{n-1}(x_1, \ldots, x_{i-1}, x_{i+1}, \ldots, x_n)\right]. \end{aligned} \tag{47}$$

In some cases, namely when the influence function does not depend smoothly on F, the Jackknife is in trouble.

5.2 Location estimators

We will now discuss some location estimators.

Definition 6 (Statistical location model) *Letting $\xi = L(e) = (\xi_1, \ldots, \xi_n)^T$ and $\delta = (\delta_1, \ldots, \delta_n)^T$, we then write the statistical location model as*

$$\xi_k = \eta + \delta_k, k = 1, \ldots, n, \tag{48}$$

where η is an unknown one-dimensional parameter and δ_k is normally distributed with mean zero and standard deviation σ.

Definition 7 (Location estimator) *Given a location model and a norm $\|\cdot\|$, an estimator $\hat{\eta} = T(\hat{F}_n)$ of η induced by the norm is*

$$\hat{\eta} = \arg\min_{\eta} \|\xi - 1_n\eta\|, \tag{49}$$

where 1_n denotes the vector whose components are 1. The estimator $\hat{\eta} = T(\hat{F}_n)$ is called a univariate location estimator.

An important location estimator is the least squares (LS) estimator given as

$$T(\hat{F}_n) = \arg\min_{\eta} \sum_{k=1}^{n} (\xi_k - \eta)^2, \tag{50}$$

which leads to the arithmetic mean. This estimator has a poor performance in the presence of contamination. Therefore Huber (Huber, 1964) has lowered the sensitivity of the LS loss function by replacing the square by a suitable function ρ. This leads to the location M-estimator.

5.2.1 M-estimators

An important subclass of M-estimates is introduced by Huber (Huber 1964). Let $\xi_1, \ldots, \xi_n$ be a random sample from a distribution F with density $p(\xi - \eta)$, where η is the location parameter. Assume that F is a symmetric unimodal distribution, then η is the center of symmetry to be estimated. The M-estimator $\hat{\eta} = T(\hat{F}_n)$ of the location parameter is defined then as some solution of the following minimization problem

$$T(\hat{F}_n) = \arg\min_{\eta} \sum_{k=1}^{n} \rho(\xi_k - \eta), \tag{51}$$

where $\rho(\cdot)$ is an even non-negative function called the contrast function (Phanzagl 1969), $\rho(\xi_k - \eta)$ is the measure of discrepancy between the observation ξ_k and the estimated center. For a given density f the choice $\rho(t) = -\log f(t)$ yields the Maximum Likelihood Estimator (MLE). It is convenient to formulate the M-estimators in terms of the derivative of the

contrast function $D(t) = d\rho(t)/dt$ called the score function. In this case, the M-estimator is defined as a solution to the following implicit equation

$$\sum_{k=1}^{n} D(\xi_k - \hat{\eta}) = 0. \tag{52}$$

A well-known example of a location parameter estimator is the Huber M-estimator. Huber considers minimization of $\sum_{k=1}^{n} \rho(\xi_k - \eta)$, where

$$\rho(t) = \begin{cases} \frac{1}{2}t^2 & |t| \le c \\ c\,|t| - \frac{1}{2}c^2 & |t| > c. \end{cases} \tag{53}$$

The score function is

$$D(t) = \begin{cases} -c, & t < -c \\ t & |t| \le c \\ c & t > c. \end{cases} \tag{54}$$

The corresponding M-estimate is a type of Winsorized mean (explained in further detail in next subsection). It turns out to be the sample mean of the modified ξ_k's, where ξ_k becomes replaced by $\hat{\eta} \pm c$, whichever is nearer, if $|\xi_k - \hat{\eta}| > c$. The contrast function and score function are sketched in Figure 5.

5.2.2 *L*-estimators

L-estimators were originally proposed by Daniel (Daniel 1920) and since then have been forgotten for many years, with a revival now in robustness studies. The description of L-estimators can be formalized as follows.

Let $\xi_1, \ldots, \xi_n$ be a random sample on a distribution F, the ordered sample values $\xi_{n(1)} \le \cdots \le \xi_{n(n)}$ is called the order statistic. A linear combination of (transformed) order statistics, or L-statistic, is a statistic of the form

$$T(\hat{F}_n) = \sum_{j=1}^{n} C_{n(j)} a(x_{n(j)}), \tag{55}$$

for some choice of constants $C_{n(1)}, \ldots, C_{n(n)}$ where $\sum_{j=1}^{n} C_{n(j)} = 1$ and $a(\cdot)$ is some fixed function. The simplest example of an L-statistic is the sample mean. More interesting,

a compromise between mean and median (trade-off between robustness and asymptotic efficiency), is the β_2-trimmed mean (Figure 1) defined as

$$\hat{\mu}_{(\beta_2)} = \frac{1}{n-2g} \sum_{j=g+1}^{n-g} \xi_{n(j)}, \tag{56}$$

where the trimming proportion β is selected so that $g = \lfloor n\beta_2 \rfloor$ and $a(\xi_{n(j)}) = \xi_{n(j)}$ is the identity function. The β-trimmed mean is a linear combination of the order statistics given zero weight to a number g of extreme observations at each end. It gives equal weight $1/(n-2g)$ to the number of $(n-2g)$ central observations. 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

$$\hat{\mu}_{(\beta_1,\beta_2)} = \frac{1}{n-(g_1+g_2)} \sum_{j=g_1+1}^{n-g_2} \xi_{n(j)}, \tag{57}$$

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 (β_1, β_2)-trimmed mean is a linear combination of the order statistics giving zero weight to g_1 and g_2 extreme observations at each end and equal weight $1/(n-g_1-g_2)$ to the $(n-g_1-g_2)$ central observations.

5.3 Robust Regression

5.3.1 Weighted LS-SVM (dual space)

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:

$$\min_{w^*,b^*,e^*} \mathcal{J}(w^*, e^*) = \frac{1}{2} w^{*T} w^* + \frac{\gamma}{2} \sum_{k=1}^{n} v_k e_k^{*2} \tag{58}$$

such that $y_k = w^*\varphi(x_k) + b^* + e_k^*, \;\; k = 1, \ldots, n$. 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^* and e^* one obtains the Karush-Kuhn-Tucker system:

$$\left[\begin{array}{c|c} 0 & 1_n^T \\ \hline 1_n & \Omega + D_\gamma \end{array}\right] \left[\begin{array}{c} b^* \\ \hline \alpha^* \end{array}\right] = \left[\begin{array}{c} 0 \\ \hline y \end{array}\right] \tag{59}$$

where the diagonal matrix D_γ is given by $D_\gamma = \mathrm{diag}\left\{\frac{1}{\gamma v_1}, \ldots, \frac{1}{\gamma v_n}\right\}$. 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. Robust estimates are obtained then (Rousseeuw and Leroy 1986) e.g. by taking

$$v_k = \begin{cases} 1 & \text{if } |e_k/\hat{s}| \leq c_1 \\ \frac{c_2 - |e_k/\hat{s}|}{c_2 - c_1} & \text{if } c_1 \leq |e_k/\hat{s}| \leq c_2 \\ 10^{-4} & \text{otherwise} \end{cases} \tag{60}$$

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. One assumes that e_k has a symmetric distribution which is usually the case when (h, γ) are well-determined by an appropriate model selection method. The constants c_1, c_2 are typically chosen as $c_1 = 2.5$ and $c_2 = 3$ (Rousseeuw and Leroy 1986). Using these weightings one can correct for outliers (y-direction). This leads us to the following algorithm:

Algorithm 2 (Weighted LS-SVM)

1. *Given training data $\mathcal{D}_n = \{(x_1, y_1), \ldots, (x_n, y_n)\}$, find an optimal (h, γ) combination (e.g. by cross-validation, FPE criterion) by solving linear systems (11). For the optimal (h, γ) combination one computes $e_k = \alpha_k/\gamma$ from (11).*

2. *Compute $\hat{s} = 1.483\,\mathrm{MAD}(e_k)$ from the e_k distribution.*

3. *Determine the weights v_k dased upon e_k, $\hat{s}$ and the constants c_1, c_2.*

4. *Solve the weighted LS-SVM (59), giving the model $\hat{f}^*(x) = \sum_{k=1}^n \alpha_k^* K(x, x_k) + b^*$.*

First, we show the sensitivity curve (one with replacement) for $(x, \hat{f}^*(x)) \in \mathcal{A}$ and $(x_i, \hat{f}^*(x_i)) \notin \mathcal{A}$ in Figure 2.b. The most important aspect is that the sensitivity curve of the $\hat{f}^*(x)$ be-

comes bounded ($x \in \mathcal{A}$) for both $y \to \infty$ and $y \to -\infty$, whereas the $\hat{f}^*(x_i)$ remains constant ($x_i \notin \mathcal{A}$).

5.3.2 Robust fixed size LS-SVM regression (primal space)

The influence function of the fixed size LS-SVM regression and its robust counterpart is given. The training data $\mathcal{D}_n = \{(x_1, y_1), \ldots, (x_n, y_n)\}$ are assumed to be zero mean. We known from Subsection 3.3 that one can construct a new training set defined as $\mathcal{D}_n^{(\text{feature})} = \{(\hat{\varphi}(x_k), y_k) | \hat{\varphi}(x_k) \in \mathbb{R}^{n_{\text{FS}}}, y_k \in \mathcal{Y}; \forall k = 1, \ldots, n, n_{FS} \leq n\}$ based on Nyström approximation. The following fixed size LS-SVM model is assumed

$$y = Aw + e, \tag{61}$$

where $y = (y_1, \ldots, y_n)^T$, $e = (e_1, \ldots, e_n)^T$ and the $n \times n_{FS}$ matrix A is defined as

$$A = \begin{pmatrix} a_1^T \\ . \\ . \\ . \\ a_n^T \end{pmatrix} = \begin{pmatrix} \hat{\varphi}_1(x_1), & \ldots, & \hat{\varphi}_{n_{FS}}(x_1) \\ . & & . \\ . & & . \\ . & & . \\ \hat{\varphi}_1(x_n), & \ldots, & \hat{\varphi}_{n_{FS}}(x_n) \end{pmatrix}. \tag{62}$$

Assume that the rows of A are i.i.d. observations from a n_{FS}-dimensional distribution F_a. Let F be the joint distribution for the $(n_{FS}+1)$- dimensional distribution, a^T is a random row of A and y is the associated dependent variable. Let $\hat{F}_n$ be the empirical distribution which put mass $\frac{1}{n}$ on each of the n rows (a_k^T, y_k), $k = 1, \ldots, n$, of the matrix $[A|y]$. Define the real-valued functional $B(F) = E[ya]$. It maps the class of all distributions on $\mathbb{R}^{n_{FS}+1}$ into $\mathbb{R}^{n_{\text{FS}}}$ (for which the expectation $E[ya]$ exists), $y \in \mathbb{R}$ and $a \in \mathbb{R}^{n_{\text{FS}}}$. Evaluated at $\hat{F}_n$

$$B(\hat{F}_n) = \frac{1}{n} \sum_{k=1}^{n} y_k a_k = \frac{1}{n} A^T y. \tag{63}$$

Define $Q(F) = E[aa^T]$, evaluated at $\hat{F}_n$

$$Q(\hat{F}_n) = \frac{1}{n} \sum_{k=1}^{n} a_k a_k^T = \frac{1}{n} A^T A. \tag{64}$$

The ridge regression functional $T(F) = w_{\text{ridge}} = (A(F) + \frac{1}{\gamma}I)^{-1}B(F)$ yields

$$\hat{w}_{\text{ridge}} = T(\hat{F}_n) = (A^T A + \frac{1}{\gamma} I_n)^{-1} A^T y. \tag{65}$$

This can also be written as a linear smoothing estimator $\hat{y} = Hy$ where $H = A(A^T A + \frac{1}{\gamma} I_{n_{\text{FS}}})^{-1} y$.

Lemma 1 *The influence function of $T(F)$ at (a^T, y^*) with $y^* \in \mathbb{R}$ is*

$$IF((a^T, y^*); T, F) = (Q(F) + \frac{1}{\gamma} I)^{-1}(B(H) - Q(F)T(F)). \tag{66}$$

Proof: By definition, the influence function gives for each (a_k^T, y_k) (where $a_k \in \mathbb{R}^{n_{\text{FS}}}, y_k \in \mathbb{R}$ and $\forall k = 1, \ldots, n$) the directional derivative (Gâteau derivative) of T at F in the direction of $H = \Delta - F$

$$\begin{aligned}
IF((a^T, y^*); T, F) &= \lim_{\epsilon \downarrow 0} \frac{T[(1-\epsilon)F + \epsilon \Delta_{[a^T y]}] - T(F)}{\epsilon} \\
&= \frac{d}{d\epsilon}[T(F + \epsilon H)]_{\epsilon=0} \\
&= \left[\frac{d}{d\epsilon} \left(Q(F) + \epsilon Q(H) + \frac{1}{\gamma} I \right)^{-1} (B(F) + \epsilon B(H)) \right]_{\epsilon=0} \\
&+ \left[\left(Q(F) + \epsilon Q(H) + \frac{1}{\gamma} I \right)^{-1} B(H) \right]_{\epsilon=0} \\
&= \left(Q(F) + \frac{1}{\gamma} I \right)^{-1} B(H) \\
&- \left(\left(Q(F) + \frac{1}{\gamma} I \right)^{-1} Q(H) \left(Q(F) + \frac{1}{\gamma} I \right)^{-1} \right) B(F) \\
&= \left(Q(F) + \frac{1}{\gamma} I \right)^{-1} ((B(H) - Q(F)T(F))) \\
&= \left(Q(F) + \frac{1}{\gamma} I \right)^{-1} (a(y^* - a^T T(F))).
\end{aligned} \tag{67}$$

□

This influence factors into an influence of position $\left(Q(F) + \frac{1}{\gamma} I \right)^{-1} a$ and the influence of the residual $(y^* - a^T T(F))$. The influence function is unbounded in $\mathbb{R}$, the observations with large residuals have a much larger contribution to the influence function.

Consider the fixed-size LS-SVM regression where $\gamma \to \infty$. Let $T(\hat{F}_n)$ be denote a robust fixed-size LS-SVM estimator based on an M-estimator.

Lemma 2 *The influence function of the robust $T(F)$ at $(a^T, y^*) \in \mathbb{R}^{n_{\text{FS}}+1}$ is*

$$IF((a^T, y^*); T, F) = \frac{\rho\left(a, y^* - a^T T(F)\right)}{\int D\left(a, y^* - a^T T(F)\right) a a^T dF(a, y^*)}. \tag{68}$$

The functional $T(F)$ corresponding to a robust fixed size LS-SVM is the solution of

$$\int \rho\left(a, y^* - a^T T(F)\right) a dF(a, y^*) = 0. \tag{69}$$

Then the influence function of T at a distribution F is given by

$$\begin{aligned} IF((a^T, y^*); T, F) &= \lim_{\epsilon \downarrow 0} \frac{T\left[(1-\epsilon)F + \epsilon \Delta_{[a^T y^*]}\right] - T(F)}{\epsilon} \\ &= \frac{d}{d\epsilon}\left[T(F + \epsilon H)\right]_{\epsilon=0} \\ &= \frac{\rho\left(a, y^* - a^T T(F)\right)}{\int D\left(a, y^* - a^T T(F)\right) a a^T dF(a, y^*)}. \end{aligned} \tag{70}$$

where $D(t) = d\rho(t)/dt$. This influence function is bounded $\mathbb{R}$. For a complete proof, see (Hampel *et al.* 1986).

5.4 Robust tuning parameter selection

Based on location estimators (e.g., mean, median, M-estimators, L-estimators, R-estimators), one can find robust counterparts of model selection criteria (e.g. Cross-Validation, Final Prediction Error criterion).

Given a random sample $e = (e_1, \ldots, e_n)^T$ from a distribution $F(e)$. Let Γ_θ be a model selection criterion and let $\xi = L(e)$ be a function of the random variable e. One can transform the cost function of Γ_θ, based on $\xi = L(e)$, into a simple location problem. The robust counterpart of the model selection criterion is now based on a robust regression estimation and a robust location estimator (e.g. median, M-estimators, L-estimators, R-estimators). The choice of a robust location estimator depends on the distribution $F(\xi)$ and its robustness and efficiency properties.

5.4.1 Robust Final Prediction Error (FPE) criterion

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 measure of complexity on $\mathcal{F}_{n,q}$ and let $\hat{f}$ be an estimator of f in $\mathcal{F}_{n,q}$. The model parameters $\theta \in \Theta$, are chosen to be the minimizer a generalized FPE criterion defined as

$$J_C(\theta) = \frac{1}{n} RSS + \left(1 + \frac{2\mathrm{tr}(S(\hat{\theta})) + 2}{n - \mathrm{tr}(S(\hat{\theta})) - 2}\right) \hat{\sigma}_e^2. \tag{71}$$

Each of these selectors depends on $S(\hat{\theta})$ through its trace ($\mathrm{tr}(S(\hat{\theta})) < n-2$), 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 $L(y_k, \hat{f}(x_k;\theta))$, $\hat{\sigma}_e^2$), it is very sensitive to outliers and other deviations from the normality assumption on the error distribution. A natural approach to robustify the FPE criterion $J_C(\theta)$ is as follows:

1. A robust estimator $\hat{f}_{\mathrm{robust}}(x,\theta)$ based on (e.g. M-estimator (Huber 1964) or weighted LS-SVM (Suykens *et al.* 2002*b*)) replaces $\hat{f}(x,\theta)$. The corresponding smoother matrix $S^*(v_1,\ldots,v_n;\hat{\theta})$ is now based on the weighting elements v_k for $k = 1,\ldots,n$ of $\hat{f}_{\mathrm{robust}}(x,\theta)$.

2. The RSS is rewritten as a location estimator which can be robustified straightforwardly. Let $\xi = L(e)$ be a function of a random variable e. A realization of the random variable e is given by $e_k = (y_k - \hat{f}(x_k;\theta))$, $k = 1,\ldots,n$, and the $\frac{1}{n}\mathrm{RSS} = J_1(\theta)$ can be written as a location problem

$$J_1(\theta) = \frac{1}{n}\sum_{k=1}^{n} L(e_k) = \frac{1}{n}\sum_{k=1}^{n} \xi_k, \tag{72}$$

where $\xi_k = e_k^2$, $k = 1,\ldots,n$. Using a robust analog of the sum ($(0,\beta_2)$-trimmed mean), the robust $J_1(\theta)$ is defined by

$$J_1^{\mathrm{robust}}(\theta) = \frac{1}{n - \lfloor n\beta_2 \rfloor} \sum_{k=1}^{n-\lfloor n\beta_2 \rfloor} \xi_{n(k)}, \tag{73}$$

where $\xi_{n(1)},\ldots,\xi_{n(n)}$ and $e_k = y_k - \hat{f}_{\mathrm{robust}}(x_k;\theta)$.

3. The variance estimator $\hat{\sigma}_e^2$ is substituted by a robust counterpart $\hat{\sigma}_{e,robust}^2$. Consider the NARX model (Ljung 1987)

$$\hat{y}(t) = f(y(t-1), \ldots, y(t-q), u(t-1), \ldots, u(t-p)). \tag{74}$$

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

$$\hat{\sigma}_e^2(y(t)) = \frac{1}{n-2} \sum_{t=2}^{n-1} \frac{(y(t-1)a + y(t+1)b - y(t))^2}{a^2 + b^2 + 1} \tag{75}$$

where $a = \frac{y(t+1)-y(t)}{y(t+1)-y(t-1)}$ and $b = \frac{y(t)-y(t-1)}{y(t+1)-y(t-1)}$. Let $\zeta = L(\vartheta)$ be a function of a random variable, a realization of the random variable ϑ is given by

$$\vartheta_k = \frac{(y(t-1)a + y(t+1)b - y(t))}{\sqrt{a^2 + b^2 + 1}}. \tag{76}$$

The variance estimator (75) can now be written as an average of the random sample $\vartheta_1^2, \ldots, \vartheta_n^2$ (a location problem):

$$\hat{\sigma}_e^2 = \frac{1}{n-2} \sum_{k=2}^{n-1} \zeta_k, \tag{77}$$

where $\zeta_k = \vartheta_k^2$, $k = 2, \ldots, n-1$. Using a robust analog of the sum ($(0, \beta_2)$-trimmed mean), the robust $\hat{s}_{e,\text{robust}}^2$ is defined by

$$\hat{\sigma}_{e,\text{robust}}^2 = \frac{1}{m - \lfloor m\beta_2 \rfloor} \sum_{l=1}^{m-\lfloor m\beta_2 \rfloor} \zeta_{n(l)}, \tag{78}$$

where $m = n - 2$.

The final robust FPE criterion is given by

$$J_C^{\text{robust}}(\theta) = J_1^{\text{robust}}(\theta) + \left(1 + \frac{2[\text{tr}(S^*(v_k, \hat{\theta})) + 1]}{n - \text{tr}(S^*(v_k, \hat{\theta})) - 2}\right) \hat{\sigma}_{e,\text{robust}}^2. \tag{79}$$

5.4.2 Influence function of the Robust FPE criterion

Because $J_C(\theta)$ is based on least squares estimation (via $L(y_k, \hat{f}(x_k;\theta))$, $\hat{s}_e^2$), it is very sensitive to outliers and other deviations from the normality assumption on the error distribution. The influence function of the Final Prediction Error (FPE) criterion is unbounded in $\mathbb{R}$. The corresponding statistical functional for the robust FPE criterion is given by

$$T(F) = \frac{1}{1-\beta_2}\int_0^{F^-(1-\beta_2)} \xi dF(\xi) + M^{\text{robust}}\left(\frac{1}{1-\beta_2}\int_0^{F^-(1-\beta_2)} \zeta dF(\zeta)\right), \qquad (80)$$

where $M^{\text{robust}} = \left(1 + \frac{2[\text{tr}(S^*(v_k,\hat{\theta}))+1]}{n-\text{tr}(S^*(v_k,\hat{\theta}))-2}\right)$. From the definition of the influence function and (80), it follows that

$$IF(\xi,\zeta;T,F) = IF(\xi;T_1,F) + M^{\text{robust}}(IF(\zeta;T_2,F)), \qquad (81)$$

where (See Appendix A)

$$IF(\xi;T_1,F) = \begin{cases} \frac{\xi-\beta_2 F(1-\beta_2)}{(1-\beta_2)} - T_1(F) & 0 \leq \xi \leq F^-(1-\beta_2) \\ F^-(1-\beta_2) - T_1(F) & F^-(1-\beta_2) < \xi \end{cases} \qquad (82)$$

and the influence function $IF(\zeta;T_2,F)$ can be calculated. We can see that the influence functions are bounded in $\mathbb{R}$. This means that an added observation at a large distance from $T(\hat{F}_n)$ gives a bounded value in absolute sense for the influence functions.

5.4.3 Robust Generalized Cross-validation

A natural approach to robustify the GCV is by replacing the linear procedure of averaging by the corresponding robust counterparts. Let $\xi = L(\vartheta)$ be a function of a random variable ϑ. In the GCV case, a realization of the random variable $\vartheta = g(e)$ is given by

$$\vartheta_k = \left(\frac{y_k - \hat{f}^*(x_k;\theta)}{1-(1/\sum_k v_k)\text{tr}(S^*)}\right), \quad k = 1,\ldots,n \qquad (83)$$

where $\hat{f}^*(x_k;\theta)$ is the weighted LS-SVM as described in Section 5.3.1, the weighting of $f^*(x_k;\theta)$ corresponding with $\{x_k, y_k\}$ is denoted by v_k and the smoother matrix based on these weightings is defined as in Eq. (13) where Z is replaced by $Z^* = (\Omega + V_\gamma)$ with

$V_\gamma = \text{diag}\left\{\frac{1}{\gamma v_1}, \ldots, \frac{1}{\gamma v_n}\right\}$. The GCV can now be written as

$$\text{GCV}(\theta) = \frac{1}{n}\sum_{k=1}^{n} L(\vartheta_k) = \frac{1}{n}\sum_{k=1}^{n} \vartheta_k^2. \tag{84}$$

Using a robust analog of the sum ($(0, \beta_2)$-trimmed mean), the robust GCV is defined by

$$\text{GCV}_{\text{robust}}(\theta) = \frac{1}{n - \lfloor n\beta_2 \rfloor} \sum_{k=1}^{n-\lfloor n\beta_2 \rfloor} I_{[\vartheta_{n(1)}, \vartheta_{n(n-\lfloor n\beta_2 \rfloor)}]}(\vartheta^2) \tag{85}$$

where $I_{[\cdot,\cdot]}(\cdot)$ is an indicator function which outputs the value 1 for the indicated interval. For the robust properties of the trimmed mean, see Appendix A.

6 Applications

6.1 Example 1: Artificial example

An example illustrates the advantage of proposed robust criteria. It considers a stochastic version of the logistic map $y(t) = cy(t-1)(y(t-1)-1) + e(t)$ with $c = 3.55$ and contaminating process noise $e(t) \sim (1-\epsilon)\mathcal{N}(0, 0.05) + \epsilon\mathcal{N}(0, 2)$. The estimates of the logistic map illustrates the difference of the NAR model based on the LS-SVM tuned by GCV and AICC, and the weighted LS-SVM tuned by $J_C^{\text{robust}}(\theta)$ and robust GCV (see Figure 6).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 as well in the one-step ahead prediction as the iterative prediction.

6.2 Example 2: Process data

The process is a liquid-satured steam heat exchanger, where water is heated by pressurized saturated steam through a copper tube. The output variable is the outlet liquid temperature. The input variables are the liquid flow rate, the steam temperature, and the inlet liquid temperature. In this experiment the steam temperature and the inlet liquid temperature are kept constant to their nominal values (See dataset 97-002 of DaISy, (De Moor 1998)).

A number of different models are tried on the dataset (for $t = 1, \ldots, 3000$). A final comparison of the estimated models is done and measured on an independent testset (for $t = 3001, \ldots, 4000$)

1. At first, classical linear tools were used to explore properties of the data. An appropriate ARX model was selected using AICC and GCV. Although its one-step ahead predictor is excellent, iterative predictions are bad because of the overestimation of the orders of the data.

2. A way to overcome this, is to use robust methods for the parameter estimation in ARX modeling (based on Huber's cost function) and the robust counterparts of AICC and GCV. Although the performance in the one-step-ahead prediction on the test set is slightly worse compared to the previous non-robust models, the iterative prediction is the iterative prediction based on robust procedures outperforms the non-robust methods. Note that small orders are selected by the robust procedures.

3. Thirdly, the fixed-size LS-SVM (RBF kernel) was used for model identification of a NARX model (Table 3). The performance in the one-step-ahead prediction on test data is slightly worse compared to the linear models, while in the iterative prediction the fixed-size LS-SVM outperforms (in the L_1, L_2 and L_∞ norm) the linear models. The best results (L_1, L_2 and L_∞ norm) are obtained by robust selection criteria combined with a robust fixed-size LS-SVM.

7 Conclusions

In this paper we have proposed robust estimation and robust model selection techniques for the use of least squares support vector machines with nonlinear ARX models. LS-SVMs is a class of kernel models with primal-dual optimization formulations. Robust techniques have been proposed for fixed-size LS-SVMs in the primal space as well as for the dual problem. Several examples illustrate that these methods can further improve standard non-robust techniques in the case of outliers and non-Gaussian noise distributions.

Acknowledgments

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 (soft-sensors), 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. Johan Suykens is an associate professor at KU Leuven. Bart De Moor and Joos Vandewalle are full professors at KU Leuven, Belgium.

References

Akaike, H. (1970). Statistical predictor identification. *Ann. Inst. Statist. Math.* **22**, 203–217.

Akaike, H. (1973). Information theory and an extension of the maximum likelihood principle. In: *Second International Symposium on Information Theory*. pp. 267–281.

Bedmarski, T. (1993). *New Directions in Statistical Data Analysis and Robustness*. Chap. Fréchet differentiability of statistical functionals and implications to robust statistics, pp. 25–34. Birkhäuser Verlag. Basel.

Bishop, C. (1995). *Neural Networks for Pattern Recognition*. Oxford University Press.

Cherkassky, V. and F. Mulier (1998). *Learning from Data*. Wiley. New York.

Craven, P. and G. Wahba (1979). Smoothing noisy data with spline functions. *Numer. Math.* **31**, 377–390.

Daniel, C. (1920). Observations weighted according to order. *Amer.J. Math* **42**, 222–236.

De Moor, B. (Ed.) (1998). *DaISy: Database for the Identification of Systems*. Department of Electrical Engineering, ESAT/SISTA, K.U.Leuven. Belgium. http://www.esat.kuleuven.ac.be/sista/daisy/.

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

Espinoza, M., K. Pelckmans, L. Hoegaerts, J. Suykens, B. De Moor (2004). A comparative study of LS-SVMs applied to the Silver Box identification problem. Internal Report 04-17, ESAT-SISTA, K.U.Leuven (Leuven, Belgium), *submitted for publication*.

Fernholz, L.T. (1983). *von Mises calculus for statistical functionals, lecture notes in statistics*. Springer-Verlag.

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

Girolami, M. (2002). Orthogonal series density estimation and the kernel eigenvalue problem. *Neural Computation* **14**(3), 669–688.

Golub, G.H. and C.F. Van Loan (1989). *Matrix Computations*. John Hopkins University Press.

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

Hampel, F.R. (1968). Contributions to the Theory of Robust Estimation. PhD thesis. University of California. Berkeley.

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

Hampel, F.R., E.M. Ronchetti, P.J. Rousseeuw and W.A. Stahel (1986). *Robust statistics, the approach based on influence functions*. Wiley & Sons, New York.

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.

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

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.

Phanzagl, T. (1969). On measurability and consistency of minimum contrast estimates. *Metrika* **14**, 248–278.

Quenouille, M. (1949). Approximate tests of correlation in times series. *J. Roy. Statist. Soc. Ser. B* **11**, 18–84.

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

Saunders, C., A. Gammerman and V. Vovk (1998). Ridge regression learning algorithm in dual variables. In: *Proc. of the 15th Int. Conf. on Machine learning (ICML'98)*. Morgan Kaufmann. pp. 515–521.

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

Shannon, C.E. (1948). A mathematical theory of communication. *Bell System Tech. I* **27**, 379–423, 623–656.

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

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

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, J. De Brabanter, B. De Moor and J. Vandewalle (2002*b*). *Least Squares Support Vector Machines*. World Scientific, Singapore.

Tukey, J.W. (1958). Bias and confidence in not quite large samples. *Abstract. Ann. Math. Statist.* **29**, 614.

Tukey, J.W. (1970). *Exploraty data analysis*. Addison-Wesley.

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

Wahba, G. (1990). *Splines for Observational Data*. SIAM.

Williams, C.K.I., M. Seeger (2001). Using the Nyström method to speed up kernel machines. In T.K. Leen, T.G. Dietterich, and V. Tresp (Eds.), *Advances in neural information processing systems*, **13**, 682–688, MIT Press.

A The $(0, \beta_2)$-trimmed mean

The corresponding statistical functional for the $(0, \beta_2)$-trimmed mean is given in terms of a quantile function and is defined as

$$\mu_{(0,\beta_2)} = T_{(0,\beta_2)}(F) = \frac{1}{1-\beta_2} \int_0^{F^-(1-\beta_2)} x dF(x) = \frac{1}{1-\beta_2} \int_0^{(1-\beta_2)} F^-(q) d(q). \tag{86}$$

The *quantile function* of a cumulative distribution function F is the generalized inverse $F^- : (0,1) \to \mathbb{R}$ given by

$$F^-(q) = \inf\{x : F(x) \geq q\}. \tag{87}$$

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^{-1} are reasonable estimates for F and F^-, respectively. The empirical quantile function is related to the order statistics $x_{n(1)} \leq \cdots \leq x_{n(n)}$ of the sample through

$$F^-(q) = x_{n(i)}, \quad \text{for } q \in \left(\frac{i-1}{n}, \frac{i}{n} \right). \tag{88}$$

To derive the influence function $IF(x; F^-(q), F)$ for the qth quantile functional $F^-(q)$, assume that F has a density f which is continuous and positive at $x_q = F^-(q)$. Let $F_\epsilon = F + \epsilon(\Delta_x - F)$ and apply (88)

$$\begin{aligned} T\left[F + \epsilon(\Delta_{x_0} - F)\right] &= \inf\left\{x : F(x) + \epsilon(\Delta_{x_0}(x) - F(x)) \geq q\right\} \\ &= \inf\left\{x : F(x) + \epsilon\left[I(x \geq x_0) - F(x)\right] \geq q\right\} \\ &= \inf\left\{x : F(x) \geq \frac{q - \epsilon\left[I(x \geq x_0)\right]}{(1-\epsilon)}\right\}. \end{aligned} \tag{89}$$

One finds $IF(x; F^-(q), F) = (\partial/\partial\epsilon)\left[F_\epsilon^{-1}(q)\right]_{\epsilon=0}$ indirectly by first calculating $(d/d\epsilon)\left[F_\epsilon^{-1}(q)\right]$ for $\epsilon > 0$ and then taking $\lim_{\epsilon \downarrow 0}(d/d\epsilon)\left[F_\epsilon^{-1}(q)\right]$. From (89) we have $F_\epsilon^{-1}(q) = F^-\left(\frac{q-\epsilon[I(x \geq x_0)]}{(1-\epsilon)}\right)$. Thus

$$\begin{aligned} \frac{d}{d\epsilon}\left[F^-\left(\frac{q - \epsilon\left[I(x \geq x_0)\right]}{(1-\epsilon)}\right)\right] &= \frac{\frac{d}{d\epsilon}\left(\frac{q-\epsilon[I(x \geq x_0)]}{(1-\epsilon)}\right)}{f\left(F^-\left(\frac{q-\epsilon[I(x \geq x_0)]}{(1-\epsilon)}\right)\right)} \\ &= \frac{q - I(x_o \leq F^-(q))}{f\left(F^-\left(\frac{q-\epsilon[I(x \geq x_0)]}{(1-\epsilon)}\right)\right)}, \end{aligned}$$

such that

$$\lim_{\epsilon\downarrow 0}\frac{d}{d\epsilon}\left[F^{-}\left(\frac{q-\epsilon\left[I(x\geq x_0)\right]}{(1-\epsilon)}\right)\right]=\frac{q-I\left(x_0\leqslant F^{-}(q)\right)}{f(F^{-}(q))}. \tag{90}$$

The $IF(x;F^{-}(q),F)$ is

$$IF(x;F^{-}(q),F)=\begin{cases}\frac{(q-1)}{f(F^{-}(q))} & x_0<F^{-}(q)\\ 0 & x_0=F^{-}(q)\\ \frac{q}{f(F^{-}(q))} & x_0>F^{-}(q).\end{cases} \tag{91}$$

Now we can calculate the influence function of the $(0,\beta_2)$-trimmed means. Define

$$\begin{aligned}T_{(0,\beta_2)}(F_\epsilon)&=\frac{1}{1-\beta_2}\int_0^{F_\epsilon^{-1}(1-\beta_2)}ydF_\epsilon(y)\\&=\frac{1}{1-\beta_2}\left[\int_0^{F_\epsilon^{-1}(1-\beta_2)}ydF(y)+\epsilon\int_0^{F_\epsilon^{-1}(1-\beta_2)}yd(\Delta_x-F)(y)\right].\end{aligned} \tag{92}$$

We will find $IF(x;\mu_{(0,\beta_2)},F)=(d/d\epsilon)\left[T_{(0,\beta_2)}(F_\epsilon)\right]_{\epsilon=0}$ indirectly by first calculating $(d/d\epsilon)\left[T_{(0,\beta_2)}(F_\epsilon)\right]$ for $\epsilon>0$ and then taking $\lim_{\epsilon\downarrow 0}(d/d\epsilon)\left[T_{(0,\beta_2)}(F_\epsilon)\right]$. From (92)

$$\begin{aligned}\frac{d}{d\epsilon}\left[T_{(0,\beta_2)}(F_\epsilon)\right]&=\frac{F_\epsilon^{-1}(1-\beta_2)}{(1-\beta_2)}f\left(F_\epsilon^{-1}(1-\beta_2)\right)\frac{d}{d\epsilon}\left[F_\epsilon^{-1}(1-\beta_2)\right]\\&\quad+(\frac{1}{1-\beta_2})\left[\int_0^{F_\epsilon^{-1}(1-\beta_2)}yd(\Delta_x-F)(y)\right]\\&\quad+\epsilon(\frac{1}{1-\beta_2})\frac{d}{d\epsilon}\left[\int_0^{F_\epsilon^{-1}(1-\beta_2)}yd(\Delta_x-F)(y)\right],\end{aligned} \tag{93}$$

so that

$$\begin{aligned}IF(x;\mu_{(0,\beta_2)},F)&=\lim_{\epsilon\downarrow 0}\frac{d}{d\epsilon}\left[T_{(0,\beta_2)}(F_\epsilon)\right]\\&=\frac{F^{-}(1-\beta_2)}{(1-\beta_2)}f(F^{-}(1-\beta_2))IF(x;F^{-}(q),F)\\&\quad+\frac{1}{1-\beta_2}\left[\int_0^{F^{-}(1-\beta_2)}yd\Delta_x(y)-\int_0^{F^{-}(1-\beta_2)}ydF(y)\right]\\&=\frac{F^{-}(1-\beta_2)}{(1-\beta_2)}f(F^{-}(1-\beta_2))IF(x;F^{-}(q),F)-\mu_{(0,\beta_2)}\\&\quad+\frac{x}{(1-\beta_2)}I\left(x\leq F^{-}(1-\beta_2)\right).\end{aligned} \tag{94}$$

Substituting the influence function $IF(x; F^{-}(q), F)$ with $q = (1 - \beta_2)$ as given in (91) into (94) yields:

$$IF(x; \mu_{(0,\beta_2)}, F) = \begin{cases} \frac{x - \beta_2 F^{-}(1-\beta_2)}{1-\beta_2} - \mu_{(0,\beta_2)} & 0 \leq x \leq F^{-}(1-\beta_2) \\ F^{-}(1-\beta_2) - \mu_{(0,\beta_2)} & F^{-}(1-\beta_2) < x. \end{cases} \tag{95}$$

The IF of the $(0, \beta_2)$-trimmed mean is sketched in Figure 1. Note that it is both continuous and bounded in $\mathbb{R}$. The finite sample breakdown point of the $(0, \beta_2)$-trimmed mean has $\epsilon_n^* = (\lfloor n\beta_2 \rfloor + 1)/n$ and the limiting value $\lim_{n\to\infty} \epsilon_n^* = \beta_2$.

Captions of Tables

Table 1: Performance of kernel based NARX model and its robust counterpart on time-series generated by logistic map with contaminated equation noise.

Table 2: Numerical performance measure on test data for the results of the process dataset. The results compare the performance of linear ARX models tuned by different model selection criteria (AICC, GCV) and its robust counterparts based on the Huber loss function and robust complexity criteria (J_C^{robust} and robust GCV). The robust procedures outperform the classical methods in the iterative prediction.

Table 3: Numerical performance measure on test data for the results of the process dataset. The results compare the performance of linear ARX models with NARX using fixed size LS-SVM (RBF kernel) tuned by different model selection criteria (AICC, GCV) and its robust counterparts based on robust fixed size LS-SVM and robust complexity criteria (J_C^{robust} and robust GCV). Again, the robust procedures outperform the classical methods in the iterative prediction.

Captions of Figures

Figure 1: (a) Sensitivity curve of the sample mean and the sample median; (b) influence function of the $(0, \beta_2)$ - trimmed mean.

Figure 2: Empirical influence curve of $\hat{f}(x)$ as a function of $(x - x_k)$. The dataset used for this experiment is $y_k = \mathrm{sinc}(x_k) + e_k$ where the errors $e_k \sim \mathcal{N}(0, 0.2^2)$ for all $k = 1, \ldots, n$ and $n = 150$.
(a) In case where $\hat{f}$ is the classical LS-SVM regression function estimator, the influence curve is unbounded in $\mathbb{R}$.
(b) In case where $\hat{f}$ is the weighted LS-SVM regression function estimator, the influence curve is bounded in $\mathbb{R}$.

Figure 3: (a) Given 150 training data (Wahba 1990) corrupted with noise $\sim \mathcal{N}(0, 1^2)$. Consider the region $\mathcal{A}$ between $x = 1$ and $x = 2$. In each step the data in the region $\mathcal{A}$ is contaminated by replacing a *good* point (given as "$\circ$") by a *bad* point (denoted as "$*$") until the estimation becomes useless. (b) Maxbias curves of the corresponding LS-SVM regression estimator $\hat{f}(x)$ and the weighted LS-SVM regression estimator $\hat{f}^*(x)$ for increasing contamination.

Figure 4: Sensitivity curves of scale parameters (shown are the mean deviation, median absolute deviation and the sample variance).

Figure 5: The contrast function and the score function of the Huber type of M-estimators.

Figure 6: The logistic map with contaminated equation noise. (a) Plot of the training data $y(t)$ with respect to the previous value $y(t-1)$. The estimated map as given by the LS-SVM tuned for γ and σ by the AICC and the GCV criterion and the weighted LS-SVM tuned for γ and σ by $\mathrm{GCV}_{\mathrm{robust}}$ and $J_c^{\mathrm{robust}}(\theta)$. (b) Time plot of the validation data and its iterative predictions using mentioned LS-SVM estimators.

Figure 7: The process is a liquid-satured steam heat exchanger, where water is heated by pressurized saturated steam through a copper tube.

Figure 8: Maxbias curve of the estimated fixed size LS-SVM and its robust counterpart.

Method	L_2	L_1	L_∞
One-step-ahead			
AICC + LS-SVM	0.0128	0.0841	0.2195
GCV + LS-SVM	0.0573	0.1713	0.4913
$J_C(\theta)_{robust}$ + wLS-SVM	0.0024	0.0460	0.0702
robust GCV + wLS-SVM	0.0025	0.0430	0.0817
Iterative			
AICC + LS-SVM	0.0086	0.0799	0.2005
GCV + LS-SVM	0.0894	0.2397	0.4586
$J_C(\theta)_{robust}$ + wLS-SVM	0.0072	0.0590	0.1726
robust GCV + wLS-SVM	0.0067	0.0612	0.2189

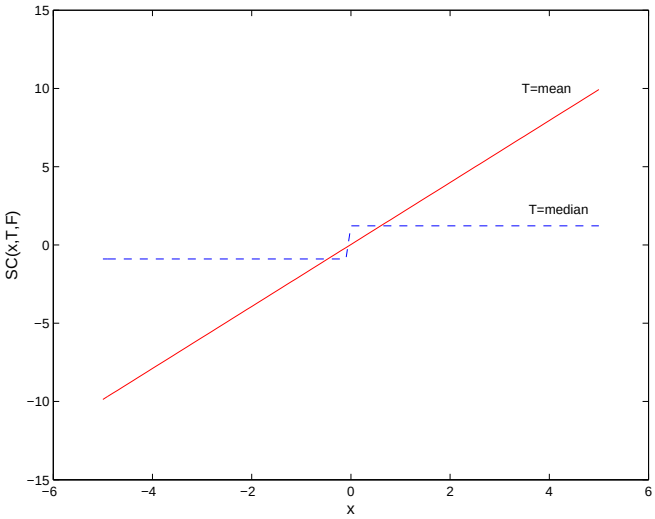
Table 1:

		One-step-ahead			**Iterative**		
Model	**Model Sel.**	L_2	L_1	L_∞	L_2	L_1	L_∞
Linear Models							
$J_C = 0.052$, $GCV = 0.059$							
ARX(40,40,8)	AICC	0.0823	0.226	1.10	3.04	1.42	4.48
ARX(40,18,8)	GCV	0.0819	0.226	1.10	0.722	0.708	2.28
Robust Linear Models							
$J_C^{robust} = 0.033$, $GCV^{robust} = 0.033$							
rARX(1,18,8)	robust AICC	0.0964	0.248	1.25	0.467	0.555	2.38
rARX(1,3,8)	robust GCV	0.0982	0.251	1.22	0.507	0.583	2.16

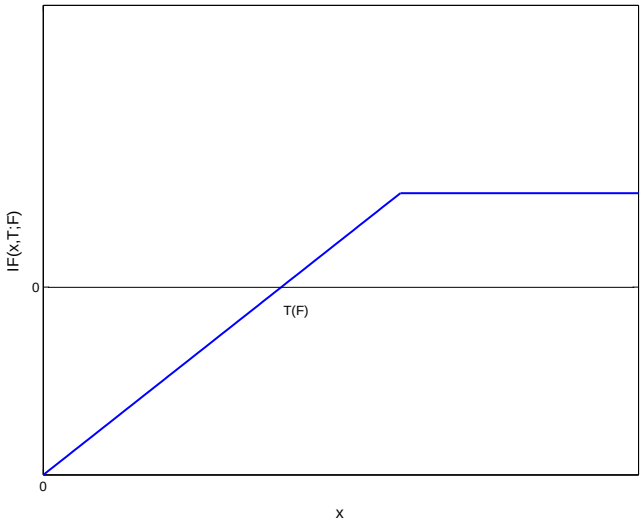
Table 2:

		One-step-ahead			**Iterative**		
Model	**Model Sel.**	L_2	L_1	L_∞	L_2	L_1	L_∞
Linear Models							
$J_C = 0.052$, $GCV = 0.059$							
ARX(40,40,8)	AICC	0.0823	0.226	1.10	3.04	1.42	4.48
ARX(40,18,8)	GCV	0.0819	0.226	1.10	0.722	0.708	2.28
Fixed-size LS-SVM (RBF) based models							
$J_C = 0.048$, $GCV = 0.054$							
NARX(3,4,8)	AICC	0.0967	0.250	1.12	0.560	0.609	2.16
NARX(3,4,8)	GCV	0.0967	0.250	1.12	0.560	0.609	2.16
Robust fixed-size LS-SVM (RBF) based models							
$J_C^{robust} = 0.024$, $GCV^{robust} = 0.030$							
rNARX(3,3,8)	robust AICC	0.0958	0.246	1.11	0.501	0.589	2.09
rNARX(2,1,8)	robust GCV	0.0914	0.245	1.08	0.496	0.566	1.99

Table 3:

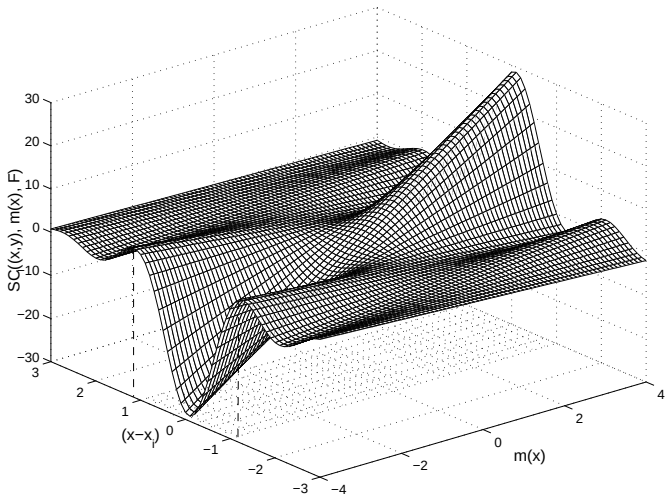


(a)

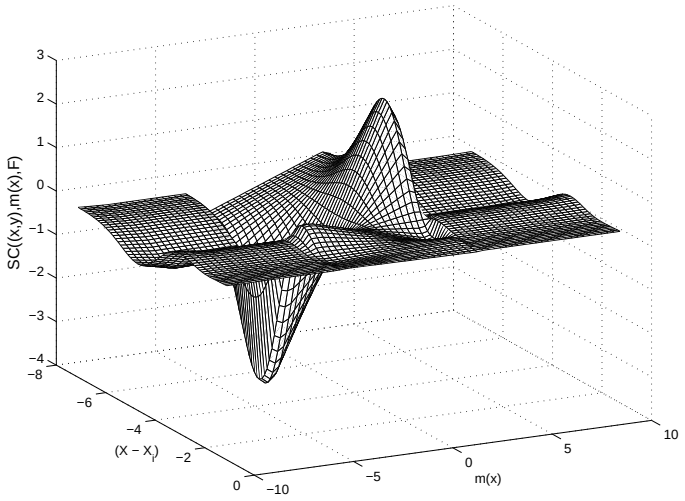


(b)

Figure 1:

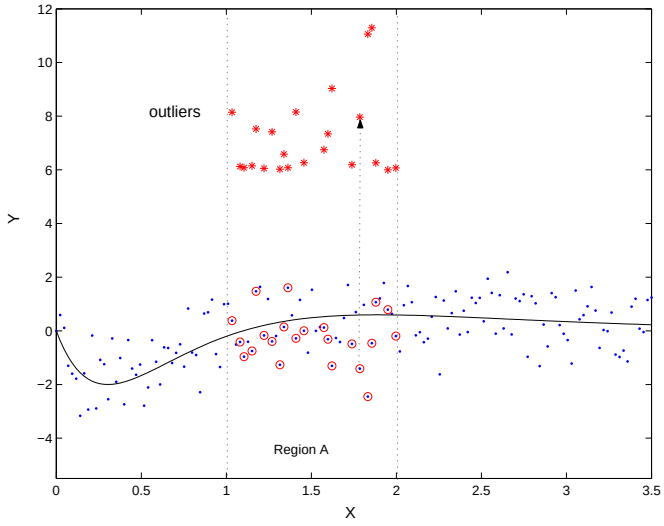

SC((x,y), m(x), F)
30
20
10
0
−10
−20
−30
3
2
1
0
−1
−2
−3
(x−x_i)
4
2
0
−2
−4
m(x)

(a)

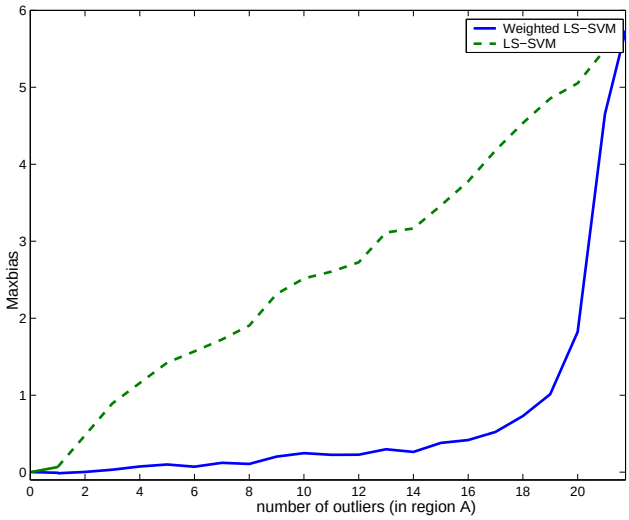

SC((x,y),m(x),F)
3
2
1
0
−1
−2
−3
−4
−8
−6
−4
−2
0
(X − X_i)
10
5
0
−5
−10
m(x)

(b)

Figure 2:


outliers
Region A
Y
X

(a)


Weighted LS−SVM
LS−SVM
Maxbias
number of outliers (in region A)

(b)

Figure 3:

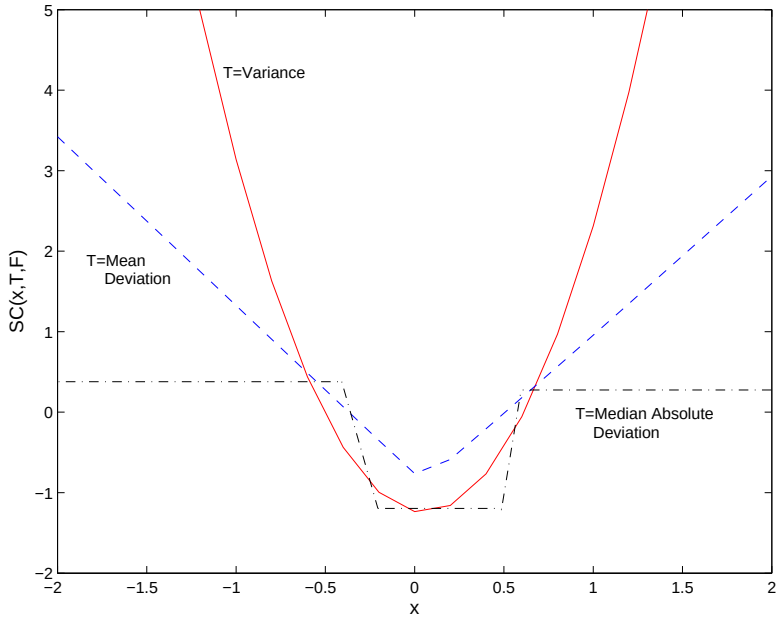


Figure 4:

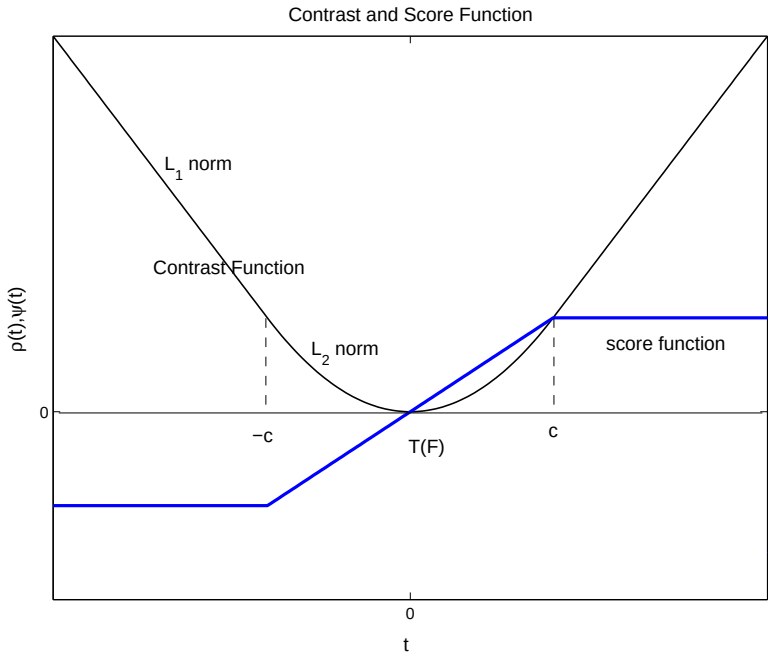
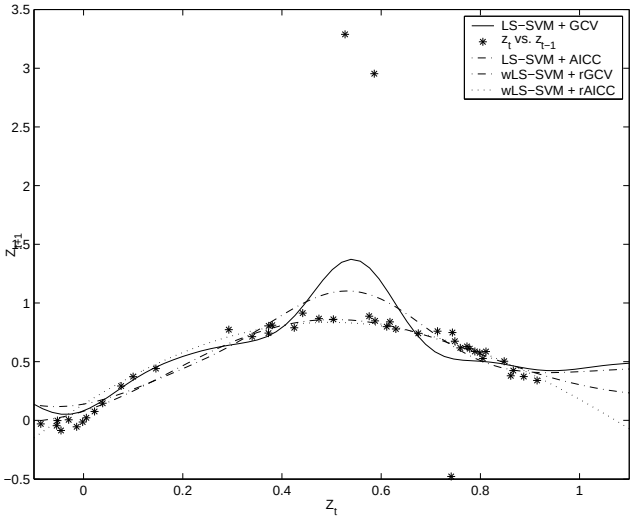
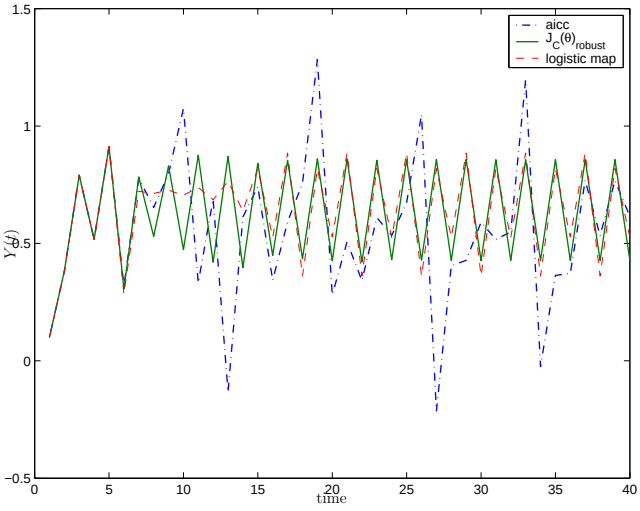


Figure 5:


LS−SVM + GCV
z_t vs. z_{t-1}
LS−SVM + AICC
wLS−SVM + rGCV
wLS−SVM + rAICC
z_{t+1}
z_t

(a)


aicc
$J_C(\theta)_{robust}$
logistic map
$Y(t)$
time

(b)

Figure 6:

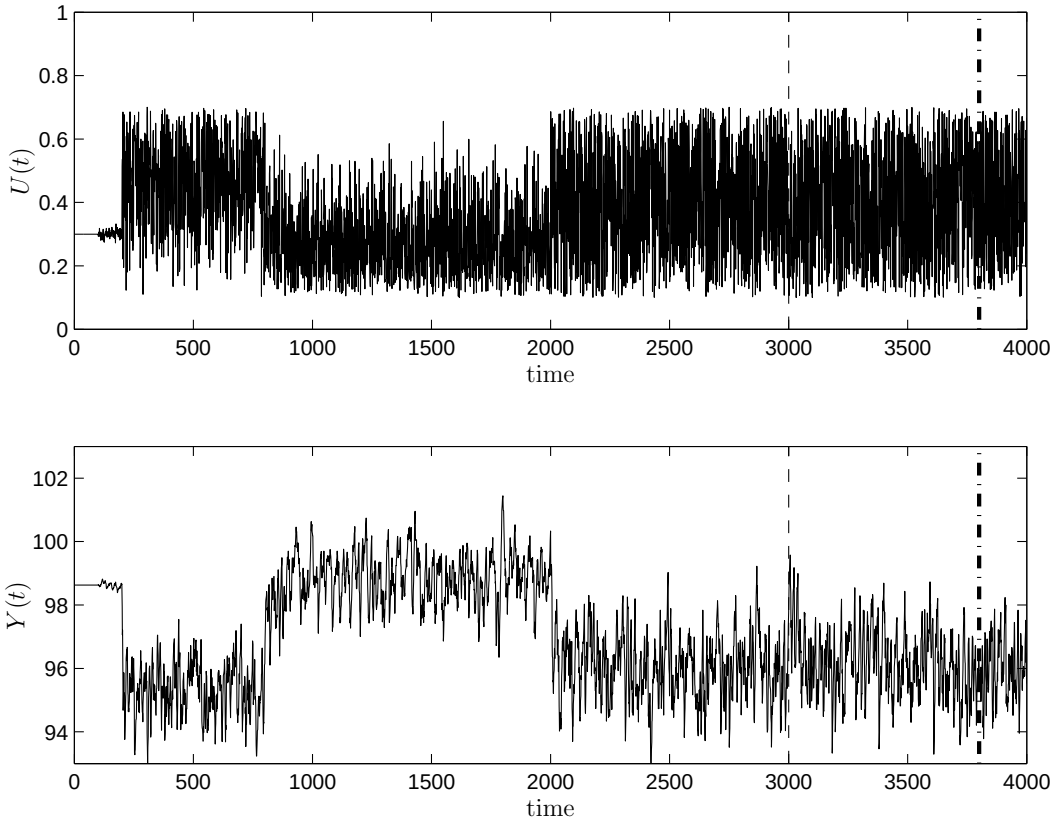

1
0.8
0.6
0.4
0.2
0
$U(t)$
0
500
1000
1500
2000
2500
3000
3500
4000
time
102
100
98
96
94
$Y(t)$
0
500
1000
1500
2000
2500
3000
3500
4000
time

Figure 7:

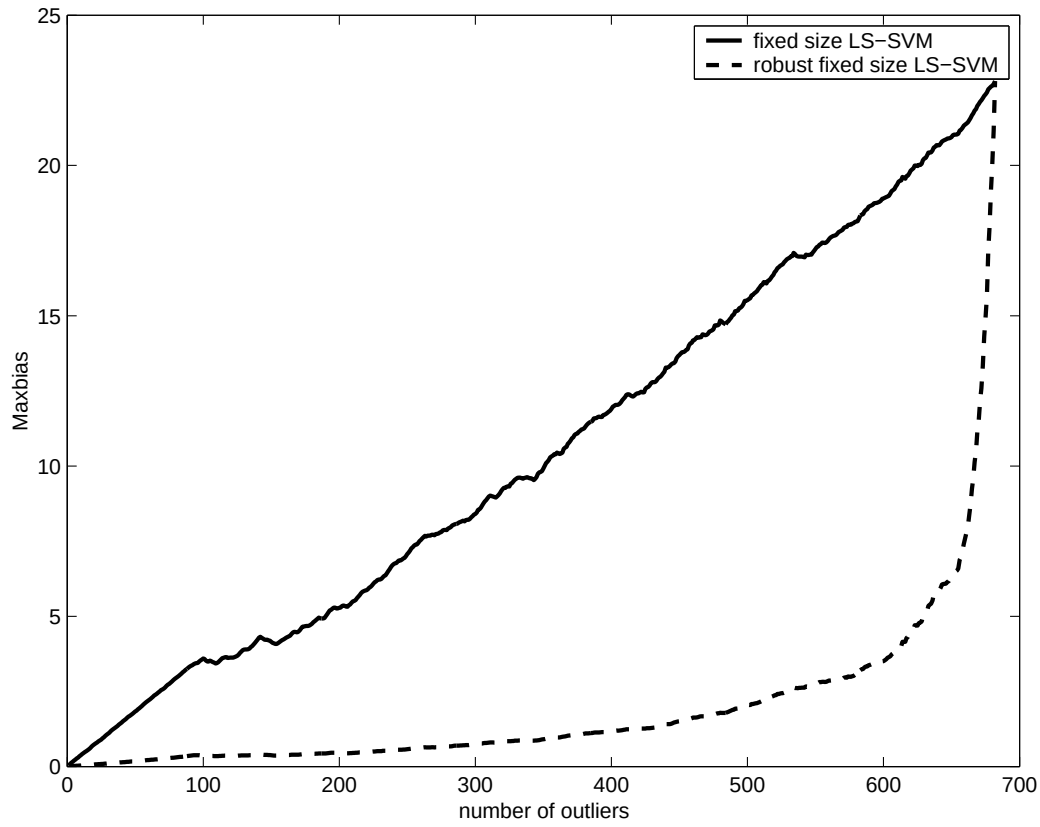

fixed size LS−SVM
robust fixed size LS−SVM
Maxbias
number of outliers
0
5
10
15
20
25
0
100
200
300
400
500
600
700

Figure 8: