Robustness of Kernel Based Regression: a Comparison of Iterative Weighting Schemes

K. De Brabanter[1], K. Pelckmans[1], J. De Brabanter[1,2],
M. Debruyne[3], J.A.K. Suykens[1], M. Hubert[4], and B. De Moor[1]

[1] KULeuven, ESAT-SCD, Kasteelpark Arenberg 10, 3001 Leuven, Belgium
{Kris.DeBrabanter,Kristiaan.Pelckmans,Johan.Suykens
Bart.DeMoor}@esat.kuleuven.be
[2] KaHo Sint-Lieven (Associatie K.U.Leuven), Departement Ind. Ing., B-9000 Gent
Jos.DeBrabanter@kahosl.be
[3] Universiteit Antwerpen, Department of Mathematics and Computer Science
Middelheimlaan 1G, B-2020 Antwerpen, Belgium
Michiel.Debruyne@ua.ac.be
[4] KULeuven, Department of Statistics, Celestijnenlaan 200B,B-3001 Leuven,Belgium
Mia.Hubert@wis.kuleuven.be

Abstract. It has been shown that Kernel Based Regression (KBR) with a least squares loss has some undesirable properties from robustness point of view. KBR with more robust loss functions, e.g. Huber or logistic losses, often give rise to more complicated computations. In this work the practical consequences of this sensitivity are explained, including the breakdown of Support Vector Machines (SVM) and weighted Least Squares Support Vector Machines (LS-SVM) for regression. In classical statistics, robustness is improved by reweighting the original estimate. We study the influence of reweighting the LS-SVM estimate using four different weight functions. Our results give practical guidelines in order to choose the weights, providing robustness and fast convergence. It turns out that Logistic and Myriad weights are suitable reweighting schemes when outliers are present in the data. In fact, the Myriad shows better performance over the others in the presence of extreme outliers (e.g. Cauchy distributed errors). These findings are then illustrated on toy example as well as on a real life data sets.

Key words: Least Squares Support Vector Machines, Robustness, Kernel methods, Reweighting

1 Introduction

Regression analysis is an important statistical tool routinely applied in most sciences. However, using least squares techniques there is an awareness of the dangers posed by the occurrence of outliers present in the data. Not only the response variable can be outlying, but also the explanatory part, leading to leverage points. Both types of outliers may totally spoil an ordinary LS analysis.

To cope with this problem, statistical techniques have been developed that are not so easily affected by outliers. These methods are called robust or resistant.

A first attempt was done by Edgeworth [1]. He argued that outliers have a very large influence on LS because the residuals are squared. Therefore, he proposed the least absolute values regression estimator (L_1 regression). The second great step forward in this class of methods occurred in the 1960s and early 1970s with fundamental work of Tukey [2], Huber [3] and Hampel [4]. From their work the following methods were developed: M-estimators, Generalized M-estimators, R-estimators, L-estimators, S-estimators, repeated median estimator, least median of squares, ... Detailed information about these estimators as well as methods for robustness measuring can be found in [5],[6], [7] and [8].

All of the above mentioned techniques were originally proposed for parametric regression. In this paper we further investigate these ideas to the nonparametric case, more specifically for Least Squares Support Vector Machines (LS-SVM). Other recent work in this direction is [9], [10] and [11]. LS-SVMs were proposed by Suykens et al. [12] as a reformulation of the Support Vector Machines (SVM) [13], applicable to a wide range of problems in supervised and unsupervised learning. In case of LS-SVMs one works with equality instead of inequality constraints and a sum of squared error cost function is used. Due to this, the regression solution is found by solving a linear system instead of a convex quadratic programming problem. By using an L_2 cost function robustness properties are lost. A successful attempt to improve the robustness was given by Suykens et al. [14]. The technique is based on a two stage approach: first, classical LS-SVM is applied and secondly appropriate weighting values are computed taking the residuals of the first step into account. For LS-SVM this weighting technique can be employed cheaply and efficiently in order to robustify the solution. In this way the weighting procedure serves as an alternative to other robust estimation methods based on L_1 and Huber's loss function without giving rise to complicated computations.

In this paper we show that the weighted LS-SVM breaks down under non Gaussian noise distributions with heavy tails. In order to deal with these distributions a reweighting scheme is proposed. Different weight functions are investigated in order to compare their performance under these heavy tailed distributions. This paper is organized as follows. In Section 2 we briefly review the basic notions of weighted LS-SVM. Section 3 explains the practical difficulties associated with estimating a regression function when the data is contaminated with outliers. Section 4 describes some extensions of existing results in order to deal with outliers in nonparametric regression. The methods are illustrated on a toy example as well as on real life data sets in Section 5.

2 Weighted LS-SVM for Nonlinear Function Estimation

In order to obtain a robust estimate, one can replace the L_2 loss function in the LS-SVM formulation by e.g. L_1 or Huber's loss function. This would lead to a Quadratic Programming (QP) problem and hence increasing the computational load. Instead of using robust cost functions, one can obtain a robust estimate based upon the previous LS-SVM solution. Given a training set defined as $\mathcal{D}_n =$

$\{(X_k, Y_k) : X_k \in \mathbb{R}^d, Y_k \in \mathbb{R}; k = 1, \ldots, n\}$ of size n drawn i.i.d. from an unknown distribution F_{XY} according to $Y_k = g(X_k) + e_k, \quad k = 1, \ldots, n$, where $e_k \in \mathbb{R}$ are assumed to be i.i.d. random errors with $\mathrm{E}[e_k|X = X_k] = 0$, $\mathrm{Var}[e_k] = \sigma^2 < \infty$, $g \in C^z(\mathbb{R})$ with $z \geq 2$, is an unknown real-valued smooth function and $\mathrm{E}[Y_k|X = X_k] = g(X_k)$. The optimization problem of finding the vector w and $b \in \mathbb{R}$ for regression can be formulated as follows [12]

$$\begin{aligned} \min_{w,b,e} \mathcal{J}(w, e) &= \tfrac{1}{2} w^T w + \tfrac{\gamma}{2} \sum_{k=1}^{n} v_k e_k^2 \\ \text{s.t.} \quad Y_k &= w^T \varphi(X_k) + b + e_k, \quad k = 1, \ldots, n, \end{aligned} \tag{1}$$

where the error variables from the unweighted LS-SVM $\hat{e}_k = \hat{\alpha}_k/\gamma$ (case $v_k = 1, \forall k$) are weighted by weighting factors v_k [14] according to (3) and $\varphi : \mathbb{R}^d \to \mathbb{R}^{n_h}$ is the feature map to the high dimensional feature space as in the standard (SVM) [13] case.

By using Lagrange multipliers, the solution of (1) can be obtained by taking the Karush-Kuhn-Tucker (KKT) conditions for optimality. The result is given by the following linear system [12] in the dual variables α

$$\left(\begin{array}{c|c} 0 & 1_n^T \\ \hline 1_n & \Omega + D_\gamma \end{array}\right) \left(\frac{b}{\alpha}\right) = \left(\frac{0}{Y}\right), \tag{2}$$

with $D_\gamma = \mathrm{diag}\left\{\frac{1}{\gamma v_1}, \ldots, \frac{1}{\gamma v_n}\right\}$. The weights v_k are based upon $\hat{e}_k = \hat{\alpha}_k/\gamma$ from the (unweighted) LS-SVM ($D_\gamma = I$). The weights v_k are given by [6]

$$v_k = \begin{cases} 1, & |\hat{e}_k/\hat{s}| \leq c_1; \\ \frac{c_2 - |\hat{e}_k/\hat{s}|}{c_2 - c_1}, & c_1 \leq |\hat{e}_k/\hat{s}| \leq c_2; \\ 10^{-8}, & \text{otherwise}, \end{cases} \tag{3}$$

where $\hat{s} = 1.483\,\mathrm{MAD}(\hat{e}_k)$ is a robust estimate of the standard deviation, where MAD is the Median Absolute Deviation. The constants are set to $c_1 = 2.5$ and $c_2 = 3$. Also $Y = (Y_1, \ldots, Y_n)^T$, $1_n = (1, \ldots, 1)^T$, $\alpha = (\alpha_1, \ldots, \alpha_n)^T$ and $\Omega_{kl} = \varphi(X_k)^T \varphi(X_l) = K(X_k, X_l)$ for $k, l = 1, \ldots, n$, with K a positive definite kernel. The resulting weighted LS-SVM model for function estimation becomes

$$\hat{g}(x) = \sum_{k=1}^{n} \hat{\alpha}_k K(x, X_k) + \hat{b}. \tag{4}$$

3 Problems with Outliers in Nonparametric Regression

A number of problems, some quite fundamental, occur when nonparametric regression is attempted in the presence of outliers. In nonparametric regression, e.g. Nadaraya-Watson kernel estimates, local polynomial kernel estimates, spline estimates and wavelets estimates, the L_2 risk is often used. There are two reasons for considering the L_2 risk: (i) this simplifies the mathematical treatment of the whole problem and (ii) trying to minimize the L_2 risk leads to estimates which

can be computed rapidly. However, the L_2 risk can be very sensitive to regression outliers. A linear kernel (in kernel-based regression) leads to non-robust methods. On the other hand using decreasing kernels, i.e. kernels such that $K(u) \to 0$ as $u \to \infty$, leads to quite robust methods with respect to outliers in the x-space (leverage points). The influence for both $x \to \infty$ and $x \to -\infty$ is bounded in $\mathbb{R}$ when using decreasing kernels. Common choices for decreasing kernels are: $K(u) = \max(1-u^2, 0)$, $K(u) = \exp(-u^2)$ and $K(u) = \exp(-u)$.

This breakdown of kernel based nonparametric regression is illustrated by a simple simulated example in Figure 1. Consider the following 200 observations $\{(X_1, Y_1), \ldots, (X_{200}, Y_{200})\}$ according to the relation $f(X) = 1 - 6X + 36X^2 - 53X^3 + 22X^5$ and $X \sim U[0,1]$. Two different types of outlier sets are added to the underlying function. The errors are normally distributed with variance $\sigma^2 = 0.05$ and $\sigma^2 = 0.1$ in Figure 1d. In Figure 1b and Figure 1c three outliers are added to the data. LS-SVM (unweighted case) cannot cope with the outliers showing a bump between 0.8 and 0.95. Notice that the unweighted LS-SVM only shows a local and not a global breakdown for the regression. SVM [13] on the other hand, deals with these type of outliers since it uses an ϵ-insensitive loss function. Figure 1c shows that the weighted LS-SVM method is able to handle these outliers and has a similar result as SVM. In Figure 1d the distribution of the errors was given by the gross error model or ϵ-contamination model [3] and is defined as follows

$$\mathcal{U}(F_0, \epsilon) = \{F : F(e) = (1-\epsilon)F_0(e) + \epsilon G(e), 0 \leq \epsilon \leq 1\} \tag{5}$$

where F_0 is some given distribution (the ideal nominal model), G is an arbitrary continuous distribution and ϵ is the first parameter of contamination. This contamination model describes the case, where with large probability $(1-\epsilon)$, the data occurs with distribution F_0 and with small probability ϵ outliers occur according to distribution G. In this case the contamination distribution G was taken to be a cubic standard Cauchy distribution and $\epsilon = 0.3$. This distribution is quite special since its moments are not defined. Both robust methods fail to fit the underlying regression model.

4 Iteratively Reweighted Kernel Based Regression

In this Section we describe and compare four types of weight functions. Also convergence properties for each of the weight functions are given.

4.1 Weight Functions

Many weight functions have been proposed in literature, especially for linear regression [6]. Four of these weight functions $V : \mathbb{R} \to [0,1]$, with $V(r) = \frac{\psi(r)}{r}$ satisfying $V(0) = 1$, are shown in Table 1 with corresponding loss function $L(r)$ and score function $\psi(r) = \frac{dL(r)}{dr}$. The first three weight function are quite common and are often used in regression [6, 10]. The fourth function, Myriad with parameter $\delta \in \mathbb{R}_0^+$, has been proposed in the area of statistical nonlinear signal

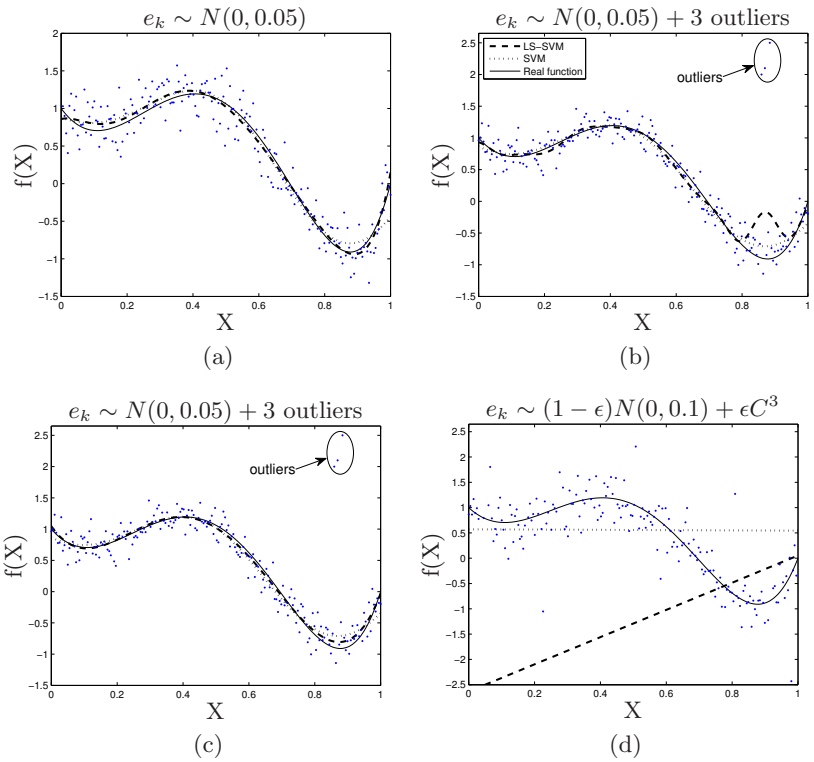
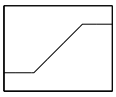
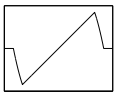
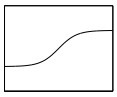
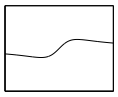


Fig. 1: Simulated data with two types of different outlier sets, fitted with LS-SVM (dashed line) and SVM (dotted line). The full line represents the underlying polynomial function. SVM and weighted LS-SVM (dashed line in (c) and (d)) can both handle the first type of outliers, but fail when the contamination distribution is taken to be a cubic standard Cauchy with $\epsilon = 0.3$. For visual reasons, not all data is displayed in (d).

processing [15]. The Myriad is derived from the Maximum Likelihood (ML) estimation of a Cauchy distribution [16] and is used as a robust location estimator in stable noise environments. When using the Myriad as a location estimator it can be shown that the Myriad offers a rich class of operation modes that can be controlled by varying the parameter δ. When the noise is Gaussian, large values of δ can provide the optimal performance associated with the sample mean, whereas for highly impulsive noise statistics, the resistance of mode-type estimators can be achieved by setting low values of δ. Arce [15] observed experimentally that values on the order of the data range, $\delta \approx X_{(n)} - X_{(1)}$, often make the Myriad an acceptable approximation to the sample average. We denote $X_{(m)}$ as the m-th order statistic for $m = 1, \ldots, n$. Intermediate values of δ assume a sample set with some outliers and some well behaved samples. On the other side, when δ is small i.e. $\delta \approx \min_{i,j} |X_i - X_j|$, the Myriad is to be considered approximately a mode estimator.

Table 1: Definitions for the Huber, Hampel, Logistic and Myriad (with parameter $\delta \in \mathbb{R}_0^+$) weight functions $V(\cdot)$. The corresponding loss $L(\cdot)$ and score function $\psi(\cdot)$ are also given.

	Huber	Hampel	Logistic	Myriad
$V(r)$	$\begin{cases} 1, & \text{if } \lvert r\rvert < \beta; \\ \frac{\beta}{\lvert r\rvert}, & \text{if } \lvert r\rvert \geq \beta. \end{cases}$	$\begin{cases} 1, & \text{if } \lvert r\rvert < b_1; \\ \frac{b_2 - \lvert r\rvert}{b_2 - b_1}, & \text{if } b_1 \leq \lvert r\rvert \leq b_2; \\ 0, & \text{if } \lvert r\rvert > b_2. \end{cases}$	$\frac{\tanh(r)}{r}$	$\frac{\delta^2}{\delta^2 + r^2}$
$\psi(r)$				
$L(r)$	$\begin{cases} r^2, & \text{if } \lvert r\rvert < \beta; \\ \beta\lvert r\rvert - \frac{1}{2}c^2, & \text{if } \lvert r\rvert \geq \beta. \end{cases}$	$\begin{cases} r^2, & \text{if } \lvert r\rvert < b_1; \\ \frac{b_2 r^2 - \lvert r^3\rvert}{b_2 - b_1}, & \text{if } b_1 \leq \lvert r\rvert \leq b_2; \\ 0, & \text{if } \lvert r\rvert > b_2. \end{cases}$	$r\tanh(r)$	$\log(\delta^2 + r^2)$

One can obtain a robust estimate based upon the previous LS-SVM solutions using an iteratively reweighting approach. In the i-th iteration one can weight the error variables $\hat{e}_k^{(i)} = \hat{\alpha}_k^{(i)}/\gamma$ for $k = 1, \ldots, n$ by weighting factors $v^{(i)} = (v_1^{(i)}, \ldots, v_n^{(i)})^T \in \mathbb{R}^n$, determined by one of the four weighting functions in Table 1. One obtains an iterative algorithm, see Algorithm 1, to solve the problem.

Algorithm 1 Iteratively Reweighted LS-SVM

1: Given optimal learning parameters (γ, σ), e.g. by cross-validation, and compute the residuals $\hat{e}_k = \hat{\alpha}_k/\gamma$ from the unweighted LS-SVM ($v_k = 1, \forall k$)
2: **repeat**
3: Compute $\hat{s} = 1.483\,\mathrm{MAD}(e_k^{(i)})$ from the $e_k^{(i)}$ distribution
4: Determine the weights $v_k^{(i)}$ based upon $r^{(i)} = e_k^{(i)}/\hat{s}$ and the chosen weight function V in Table 1
5: Solve the weighted LS-SVM (2) with $D_\gamma = \mathrm{diag}\left\{\frac{1}{\gamma v_1^{(i)}}, \ldots, \frac{1}{\gamma v_n^{(i)}}\right\}$, resulting the model $\hat{m}^{(i)}(x) = \sum_{k=1}^n \hat{\alpha}_k^{(i)} K(x, X_k) + \hat{b}^{(i)}$
6: Set $i = i + 1$
7: **until** consecutive estimates $\alpha_k^{(i-1)}$ and $\alpha_k^{(i)}$ are sufficiently close to each other $\forall k = 1, \ldots, n$. In this paper we take $\max_k(|\alpha_k^{(i-1)} - \alpha_k^{(i)}|) \leq 10^{-4}$.

4.2 Speed of Convergence-Robustness Trade-off

In a functional analysis setting it has been shown in [9] and [10] that the influence function [4] of reweighted Least Squares Kernel Based Regression (LS-KBR) with a bounded kernel converges to bounded influence function, even when the initial LS-KBR is not robust, if

($c1$) $\psi : \mathbb{R} \to \mathbb{R}$ is a measurable, real, odd function,
($c2$) ψ is continuous and differentiable,
($c3$) ψ is bounded,
($c4$) $\mathrm{E}_{P_e}\, \psi'(e) > 0$ where P_e denotes the distribution of the errors. This condition can be relaxed into ψ is increasing.

The influence function (IF) describes the (approximate and standardized) effect of an additional observation in any point x on a statistic T, given a (large) sample with distribution F. Thus an unbounded IF means that an infinitesimal amount of outliers can have an arbitrary large effect.

Define

$$d = \mathrm{E}_{P_e} \frac{\psi(e)}{e} \quad \text{and} \quad c = d - \mathrm{E}_{P_e}\, \psi'(e), \tag{6}$$

then it can be shown [10] that c/d establishes an upper bound on the reduction of the influence function at each step. The upper bound represents a trade-off between the reduction of the influence function (speed of convergence) and the degree of robustness. The higher the ratio c/d the higher the degree of robustness but the slower the reduction of the influence function at each step and vice versa.

In Table 2 this upper bound is calculated at a Normal distribution, a standard Cauchy and a cubic standard Cauchy for the four types of weighting schemes. Note that the convergence of the influence function is quite fast, even at heavy tailed distributions.

For Huber and Myriad weights, the convergence rate decreases rapidly as β respectively δ increases. This behavior is to be expected, since the larger β respectively δ, the less points are downweighted. Also note that the upper bound on the convergence rate approaches 1 as $\beta, \delta \to 0$, indicating a high degree of robustness but slow convergence rate. A good choice between convergence and robustness is therefore Logistic weights. Also notice the small ratio for the Hampel weights indicating a low degree of robustness. The inability of these weights to handle extreme outliers is shown in the next Section. For further elaboration on the topic we refer the reader to [11].

5 Simulations

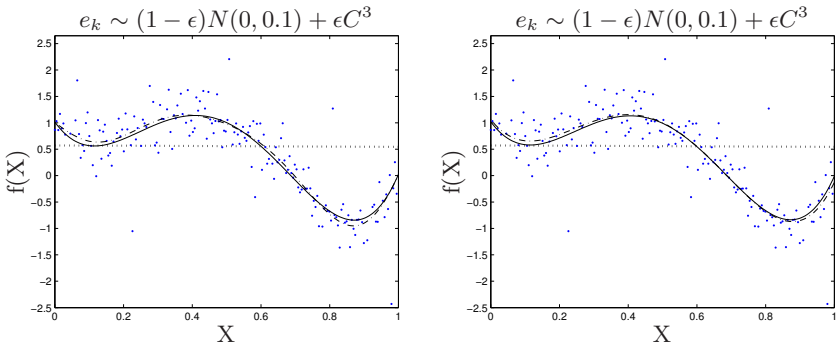
5.1 Toy Example

Recall the low order polynomial function in Section 3 with 200 observations according to $f(X) = 1 - 6X + 36X^2 - 53X^3 + 22X^5$ and $X \sim U[0, 1]$. The distribution of the errors is given by the gross error model, see Section 3, with $\epsilon = 0.3$, $F_0 = N(0, 0.1)$ and $G = C^3(0, 1)$. The results for the four types of weight functions are shown in Figure 2 and performances in the three norms are given in Table 3. For this simulation we set $\beta = 1.345$, $b_1 = 2.5$ and $b_2 = 3$ and $\delta = \frac{1}{2}[\hat{e}^{(i)}_{(\frac{3}{4}n)} - \hat{e}^{(i)}_{(\frac{1}{4}n)}]$ where $\hat{e}^{(i)}_{(m)}$ denotes the m-th order statistic of the residual $\hat{e}$ in the i-th iteration. For all simulations, the learning parameters are tuned via 10-fold robust cross-validation. This simulation shows that the four weight functions are able to handle these extreme outliers. Although Hampel and Myriad weight functions do not satisfy the relaxed condition of (c4), condition (c4) is valid for

Table 2: Values of the constants c, d and c/d for the Huber (with different cutoff values β), Logistic, Hampel and Myriad (for different parameters δ) weight function at a standard Normal distribution, a standard Cauchy and a cubic standard Cauchy. The bold values represent an upper bound for the reduction of the influence function at each step.

Weight function	Parameter settings	$N(0,1)$			$C(0,1)$			$C^3(0,1)$		
		c	d	c/d	c	d	c/d	c	d	c/d
Huber	$\beta = 0.5$	0.32	0.71	**0.46**	0.26	0.55	**0.47**	0.0078	0.034	**0.23**
	$\beta = 1$	0.22	0.91	**0.25**	0.22	0.72	**0.31**	0.0022	0.037	**0.059**
	$\beta = 2$	0.04	0.99	**0.04**	0.14	0.85	**0.17**	0.0002	0.038	**0.0053**
Logistic		0.22	0.82	**0.26**	0.21	0.66	**0.32**	0.004	0.035	**0.12**
Hampel	$b_1 = 2.5$ $b_2 = 3$	0.006	0.99	**0.006**	0.02	0.78	**0.025**	0.00003	0.038	**0.0007**
Myriad	$\delta = 0.1$	0.11	0.12	**0.92**	0.083	0.091	**0.91**	0.007	0.009	**0.83**
	$\delta = 0.6475$	0.31	0.53	**0.60**	0.24	0.40	**0.60**	0.01	0.028	**0.36**
	$\delta = 1$	0.31	0.66	**0.47**	0.25	0.50	**0.50**	0.008	0.032	**0.25**

common error distributions i.e. Normal, Cauchy, Student t, Laplace, ... This simulation shows the best performance for the Myriad weight function. This is to be expected since it was designed for such types of outliers.



(a) (b)

Fig. 2: Low order polynomial function with 200 observations according to $f(X) = 1 - 6X + 36X^2 - 53X^3 + 22X^5$ and $X \sim U[0,1]$. The distribution of the errors is given by the gross error model with $\epsilon = 0.3$, $F_0 = N(0,0.1)$ and $G = C^3(0,1)$. The dotted line is the corresponding SVM fit. The iteratively reweighted LS-SVM with (a) Huber weights (full line) and Hampel weights (dash dotted line); (b) Logistic weights (full line) and Myriad weights (dash dotted line).

5.2 Real Life Data Sets

The octane data [17] consist of NIR absorbance spectra over 226 wavelengths ranging from 1102 to 1552 nm. For each of the 39 production gasoline samples

Table 3: Performances in the three norms (difference between the estimated function and the true underlying function) of the different weight functions used in iteratively reweighted LS-SVM on the low order polynomial. The last column denotes the number of iterations $i_{\max}$ needed to satisfy the stopping criterion in Algorithm 1.

	L_1	L_2	L_∞	$i_{\max}$
Huber	0.06	0.005	0.12	7
Hampel	0.06	0.005	0.13	4
Logistic	0.06	0.005	0.11	11
Myriad	0.03	0.002	0.06	17

the octane number Y was measured. It is well known that the octane data set contains six outliers to which alcohol was added. Table 4 shows the result (medians and mean absolute deviations) of a Monte Carlo simulation (200 times) of the iteratively reweighted LS-SVM (IRLS-SVM), weighted LS-SVM (WLS-SVM) and SVM in different norms on a randomly chosen test set of size 10. As a next example consider the data about the demographical information on the 50 states of the USA in 1980. The data set provides information on 25 variables. The goal is to determine the murder rate per 100,000 population. The result is shown in Table 4 for randomly chosen test sets of size 15. The results of the simulations show that by using reweighting schemes the performance can be improved over weighted LS-SVM and SVM. To illustrate the trade-off between the degree of robustness and speed of convergence, the number of iterations $i_{\max}$ are also given in Table 4. The stopping criterion was taken identically to the one in Algorithm 1. The number of iterations, needed by each weight function, confirms the results in Table 2.

Table 4: Results on the Octane and Demographic data sets. For 200 simulations the medians and mean absolute deviations (between brackets) of three norms are given (on test data). $i_{\max}$ denotes the number of iterations needed to satisfy the stopping criterion in Algorithm 1. The best results are bold faced.

		Octane				Demographic			
	weights	L_1	L_2	L_∞	$i_{\max}$	L_1	L_2	L_∞	$i_{\max}$
IRLS SVM	Huber	**0.19**(0.03)	0.07(0.02)	0.51(0.10)	15	0.31(0.01)	0.14(0.02)	0.83(0.06)	8
	Hampel	0.22(0.03)	0.07(0.03)	0.55(0.14)	2	0.33(0.01)	0.18(0.04)	0.97(0.02)	3
	Logistic	0.20(0.03)	**0.06**(0.02)	0.51(0.10)	18	0.30(0.02)	**0.13**(0.01)	0.80(0.07)	10
	Myriad	0.20(0.03)	**0.06**(0.02)	**0.50**(0.09)	22	**0.30**(0.01)	**0.13**(0.01)	**0.79**(0.06)	12
WLS SVM		0.22(0.03)	0.08(0.02)	0.60(0.15)	1	0.33(0.02)	0.15(0.01)	0.80(0.02)	1
SVM		0.28(0.03)	0.12(0.02)	0.56(0.13)	-	0.37(0.02)	0.21(0.02)	0.90(0.06)	-

6 Conclusion

In this paper we have compared four different type of weight functions and their use in iterative reweighted LS-SVM. We have shown through simulations that

reweighting is useful when outliers are present in the data. By using an upper bound for the reduction of the influence function we have demonstrated the existence of a trade-off between speed of convergence and the degree of robustness. The Myriad weight function is highly robust against (extreme) outliers but has a slow speed of convergence. A good compromise between speed of convergence and robustness can be achieved by using Logistic weights.

Acknowledgements Research supported by: Research Council KUL: GOA AMBioRICS, CoE EF/05/006 Optimization in Engineering(OPTEC), IOF-SCORES4CHEM, several PhD/postdoc & fellow grants; Flemish Government: FWO: PhD/postdoc grants, projects G.0452.04, G.0499.04, G.0211.05, G.0226.06 l, G.0321.06, G.0302.07, G.0320.08, G.0558.08, G.0557.08, research communities (ICCoS, ANMMM, MLDM); IWT: PhD Grants, McKnow-E, Eureka-Flite+; Helmholtz: viCERP. Belgian Federal Science Policy Office: IUAP P6/04 (DYSCO, 2007-2011); EU: ERNSI.

References

1. Edgeworth, F.Y.: On Observations Relating to Several Quantities. Hermathena 6, 279–285 (1887)
2. Tukey, J.W.: A survey of sampling from contaminated distributions. In: Olkin, I. (ed.) Contributions to Probability and Statistics. Stanford University Press, Stanford, CA, pp. 448–485 (1960)
3. Huber, P.J.: Robust Estimation of a Location Parameter. Ann. Math. Stat 35, 73–101 (1964)
4. Hampel, F.R.: A General Definition of Qualitative Robustness. Ann. Math. Stat 42, 1887–1896 (1971)
5. Huber, P.J.: Robust Statistics. Wiley (1981)
6. Rousseeuw, P.J., Leroy, A.M.: Robust Regression and Outlier Detection. Wiley (2003)
7. Maronna, R., Martin, D., Yohai, V.: Robust Statistics. Wiley (2006)
8. Hampel, F.R., Ronchetti, E.M., Rousseeuw, P.J., Stahel, W.A.: Robust Statistics: The Approach Beased on Influence Functions. Wiley (1986)
9. Christmann, A., Steinwart, I.: Consistency and Robustness of Kernel Based Regression in Convex Risk Minimization. Bernoulli 13(3), 799–819 (2007)
10. Debruyne, M., Christmann, A., Hubert, M., Suykens, J.A.K.: Robustness and Stability of Reweighted Kernel Based Regression. Technical Report 06-09, Department of Mathematics, K.U.Leuven (Leuven, Belgium) (2008)
11. Debruyne, M., Hubert, M., Suykens, J.A.K.: Model Selection in Kernel Based Regression using the Influence Function, J. Mach. Learn. Res. 9, 2377–2400 (2008)
12. Suykens, J. A. K., Van Gestel, T., De Brabanter, J., De Moor, B., Vandewalle, J.: Least Squares Support Vector Machines. World Scientific, Singapore (2002)
13. Vapnik, V. N.: Statistical Learning Theory. Wiley (1999)
14. Suykens, J.A.K., De Brabanter J., Lukas L., Vandewalle J.: Weighted Least Squares Support Vector Machines : Robustness and Sparse Approximation. Neurocomputing 48(1–4), 85–105 (2002)
15. Arce, G. R.: Nonlinear Signal Processing: A Statistical Approach. Wiley (2005)
16. Gonzalez, J.G, Arce, G.R.: Weighted Myriad Filters: A Robust Filtering Framework derived from Alpha-Stable Distributions. In Proceedings of the 1996 IEEE Conference one Acoustics, Speech and Signal Processing (1996)
17. Hubert, M., Rousseeuw, P.J., Vanden Branden, K.: ROBPCA: a New Approach to Robust Principal Components Analysis, Technometrics 47, 64-79 (2005)