

Robustness of Kernel Based Regression: Influence and Weight Functions

Kris De Brabanter^{*†}, Jos De Brabanter^{*†‡}, Johan A.K. Suykens^{*†}, Joos Vandewalle^{*} and Bart De Moor^{*†}

^{*}Department of Electrical Engineering (ESAT-SCD), Research Division SCD

Katholieke Universiteit Leuven, Kasteelpark Arenberg 10, B-3001 Leuven, Belgium

Email: {kris.debrabanter,jos.debrabanter,johan.suykens,joos.vandewalle,bart.demoor}@esat.kuleuven.be

Telephone: (+32) 16 32 86 64

[†]IBBT-KU Leuven Future Health Department

Katholieke Universiteit Leuven, Kasteelpark Arenberg 10, B-3001 Leuven, Belgium

[‡]KaHo Sint Lieven (Associatie KU Leuven), Departement Industrieel Ingenieur

G. Desmetstraat 1, 9000 Gent, Belgium

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 classical statistics, robustness is improved by reweighting the original estimate. We study the influence of reweighting the LS-KBR estimate using three well-known weight functions and one new weight function called Myriad. 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. Finally, we establish an empirical maxbias curve to demonstrate the ability of the proposed methodology.

I. 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 least squares (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 in 1887. He argued that outliers have a very large influence on a LS loss function 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 [1], Huber [2] (minimax approach) and Hampel (influence functions) [3]. Huber [2] gave the first theory of robustness. He considered the general gross-error model or ϵ -contamination model

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

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 .

Example 1: ϵ -contamination model with symmetric contamination

$$F(x) = (1 - \epsilon)\mathcal{N}(0, 1) + \epsilon\mathcal{N}(0, \kappa^2\sigma^2), \quad 0 \leq \epsilon \leq 1, \kappa > 1.$$

Example 2: ϵ -contamination model for the mixture of the Normal and Laplace or double exponential distribution

$$F(x) = (1 - \epsilon)\mathcal{N}(0, 1) + \epsilon\text{Lap}(0, \lambda), \quad 0 \leq \epsilon \leq 1, \lambda > 0.$$

Huber considered also the class of M -estimators of location (also called 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.

Huber developed a second theory [4], [5] for censored likelihood ratio tests and exact finite sample confidence intervals, using more general neighborhoods of the normal model. This approach may be mathematically the most rigorous but seems very hard to generalize and therefore plays hardly any role in applications. A third theory proposed by Hampel [3], [6] is closely related to robustness theory which is more generally applicable than Huber's first and second theory. Three main concepts are introduced: (i) qualitative robustness, which is essentially continuity of the estimator viewed as functional in the weak topology; (ii) the influence curve (IC) or influence function (IF), which describes the first derivative of the estimator, as far as existing; and (iii) the breakdown point (BP), a global robustness measure describing how many percent gross errors are still tolerated before the estimator totally breaks down.

Robustness has provided at least two major insights into statistical theory and practice: (i) Relatively small perturbations from nominal models can have very substantial deleterious

effects on many commonly used statistical procedures and methods (e.g. estimating the mean, F-test for variances). (ii) Robust methods are needed for detecting or accommodating outliers in the data [7], [8].

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, etc. Detailed information about these estimators as well as methods for robustness measuring can be found in the books [3], [9]–[11]. See also the book [12] for robust statistical methods with R providing a systematic treatment of robust procedures with an emphasis on practical applications.

This paper is organized as follows. Section II describes the measures of robustness and introduces some terminology. Section III explains the practical difficulties associated with estimating a regression function when the data is contaminated with outliers. Section IV gives the influence function of least squares kernel based regression (LS-KBR) and derives a new weight function. Section V shows an empirical maxbias curve for the LS-KBR contaminated with the gross error model. Finally, Section VI states the conclusions.

II. MEASURES OF ROBUSTNESS

In order to understand why certain estimators behave the way they do, it is necessary to look at various measures of robustness. There exist numerous approaches towards the robustness problem. The approach based on influence functions will be used here. The effect of one outlier on the estimator can be described by the influence function (IF). The 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 . Another measure of robustness of an estimator is the maxbias curve. The maxbias curve 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. The breakdown value is defined as how much contaminated data an estimator can tolerate before it becomes useless.

Let F be a fixed distribution and $T(F)$ a statistical functional defined on a set \mathcal{G}_ϵ of distributions satisfying that T is Gâteaux differentiable at the distribution F in domain T [3]. Let the estimator $T(\hat{F}_n)$ of $T(F)$ be the functional of the sample distribution F_n .

Definition 1 (Influence Function): The influence function (IF) of T at F is given by

$$\text{IF}(x; T, F) = \lim_{\epsilon \downarrow 0} \frac{T[(1 - \epsilon)F + \epsilon\Delta_x] - T(F)}{\epsilon} \quad (2)$$

in those x where this limit exists. Δ_x denotes the probability measure which puts mass 1 at the point x .

Hence, the IF reflects the bias caused by adding a few outliers at the point x , standardized by the amount ϵ of contamination. Therefore, a bounded IF leads to robust estimators. Note that this kind of differentiation of statistical functionals is a differentiation in the sense of von Mises with a kernel function [13],

[14]. From the influence function, several robustness measures can be defined: the gross error sensitivity, the local shift sensitivity and the rejection point, see [3, Section 2.1c] for an overview. 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 i.e. Gâteaux, Hadamard or compact, Bouligand and Fréchet. An application of the Bouligand IF can be found in [15] in order to investigate the robustness properties of support vector machines (SVM). The Bouligand IF has the advantage of being positive homogeneous which is in general not true for Hampel's influence function (2). Christmann & Van Messen [15] also show that there exists an interesting relationship between the Bouligand IF and the IF: if the Bouligand IF exists, then the IF does also exist and both are equal. Next, we give the definitions of the maxbias curve and the breakdown point. Note that some authors can give a slightly different definition of the maxbias curve, see e.g. [16].

Definition 2 (Maxbias Curve): Let $T(F)$ denote a statistical functional and let the contamination neighborhood of F be defined by \mathcal{G}_ϵ for a fraction of contamination ϵ . The maxbias curve is defined by

$$B(\epsilon, T, F) = \sup_{F \in \mathcal{G}_\epsilon} |T(F) - T(F_0)|. \quad (3)$$

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

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

From the previous definition it is obvious that the breakdown point defines the largest fraction of gross errors that still keeps the bias bounded. We will give some examples of influence functions and breakdown points for the mean, median and variance.

III. OUTLIERS IN NONPARAMETRIC REGRESSION

Consider 200 observations on the interval $[0, 1]$ and a low-order polynomial mean function $m(X) = 300(X^3 - 3X^4 + 3X^5 - X^6)$. Figure 1a shows the mean function with normally distributed errors with variance $\sigma^2 = 0.3^2$ and two distinct groups of outliers. Figure 1b shows the same mean function, but the errors are generated from the gross error or ϵ -contamination model (1). In this simulation $F_0 \sim N(0, 0.3^2)$, $G \sim N(0, 10^2)$ and $\epsilon = 0.3$. This simple example clearly shows that the estimates based on the L_2 norm with classical cross-validation (CV) (bold line) are influenced in a certain region (similar as before) or even breakdown (in case of the gross error model) in contrast to estimates based on robust kernel based regression (KBR) with robust CV (thin line). A fully robust KBR method will be discussed later in this Section. Another important issue to obtain robustness in nonparametric regression is the kernel function K . Kernels that satisfy $K(u) \rightarrow 0$ as $u \rightarrow \infty$, for $X \rightarrow \infty$ and $X \rightarrow -\infty$, are bounded in \mathbb{R} . These type of kernels are called decreasing kernels. Using decreasing kernels

leads to quite robust methods with respect to outliers in the X -direction (leverage points). Common choices of decreasing kernels are: $K(u) = \max((1 - u^2), 0)$, $K(u) = \exp(-u^2)$, $K(u) = \exp(-|u|), \dots$

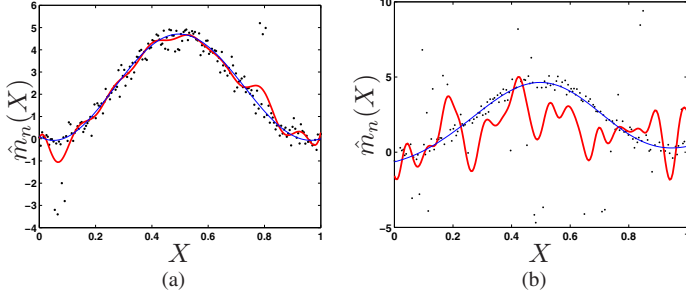


Fig. 1. Kernel based estimate with (a) normal distributed errors and two groups of outliers; (b) the ϵ -contamination model. This clearly shows that the estimates based on the L_2 norm (bold line) are influenced in a certain region or even break down in contrast to estimates based on robust loss functions (thin line).

The last issue to acquire a fully robust estimate is the proper type of cross-validation (CV). When no outliers are present in the data, CV has been shown to produce tuning parameters that are asymptotically consistent [17]. Yang [18] showed that, under some regularity conditions, for an appropriate choice of data splitting ratio, cross-validation is consistent in the sense of selecting the better procedure with probability approaching 1. However, when outliers are present in the data, the use of CV can lead to extremely biased tuning parameters [19] resulting in bad regression estimates. The estimate can also fail when the tuning parameters are determined by standard CV using a robust smoother. The reason is that CV no longer produces a reasonable estimate of the prediction error. Therefore, a fully robust CV method is necessary. Figure 2 demonstrates this behavior on the same toy example (see Figure 1). Indeed, it can be clearly seen that CV results in less optimal tuning parameters resulting in a bad estimate. Hence, to obtain a fully robust estimate, every step has to be robust i.e. robust CV with a robust smoother based on a decreasing kernel.

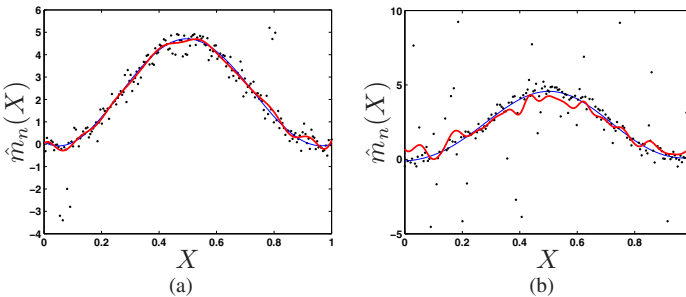


Fig. 2. (a) normal distributed errors and two groups of outliers; (b) the ϵ -contamination model. Bold lines represent the estimate based on classical L_2 CV and a robust smoother. Thin lines represents estimates based on a fully robust procedure.

IV. THEORETICAL BACKGROUND

A. Notation & IF of Kernel Based Regression Methods

KBR methods estimate a functional relationship between a dependent variable X and an independent variable Y , using a sample of n observations $(X_i, Y_i) \in \mathcal{X} \times \mathcal{Y} \subseteq \mathbb{R}^d \times \mathbb{R}$ with joint distribution F_{XY} . First, we give the following definitions taken from [20].

Definition 4 ([20]): Let \mathcal{X} be a non-empty set. Then a function $K : \mathcal{X} \times \mathcal{X} \rightarrow \mathbb{R}$ is called a kernel on \mathcal{X} if there exists a Hilbert space \mathcal{H} with an inner product $\langle \cdot, \cdot \rangle$ and a map $\varphi : \mathcal{X} \rightarrow \mathcal{H}$ such that for all $x, y \in \mathcal{X}$ we have

$$K(x, y) = \langle \varphi(x), \varphi(y) \rangle.$$

φ is called the feature map and \mathcal{H} is a feature space of K .

Definition 5 ([20]): Let \mathcal{X} be a non-empty set and \mathcal{H} be a Hilbert function space over \mathcal{X} , i.e. a Hilbert space that consists of functions mapping from \mathcal{X} into \mathbb{R} .

- A function $K : \mathcal{X} \times \mathcal{X} \rightarrow \mathbb{R}$ is called a reproducing kernel of \mathcal{H} if we have $K(\cdot, x) \in \mathcal{H}$ for all $x \in \mathcal{X}$ and the reproducing property $m(x) = \langle m, K(\cdot, x) \rangle$ holds for all $m \in \mathcal{H}$ and all $x \in \mathcal{X}$.
- The space \mathcal{H} is called a Reproducing Kernel Hilbert Space (RKHS) over \mathcal{X} if for all $x \in \mathcal{X}$ the Dirac functional $\delta_x : \mathcal{H} \rightarrow \mathbb{R}$ defined by

$$\delta_x(m) = m(x), \quad m \in \mathcal{H}$$

is continuous.

Finally, we need the following definition about the joint distribution F_{XY} . For notational ease, we will suppress the subscript XY .

Definition 6 ([20]): Let F be a distribution on $\mathcal{X} \times \mathcal{Y}$, let $a : \mathcal{Y} \rightarrow [0, \infty)$ be a measurable function and let $|F|_a$ be defined as

$$|F|_a = \int_{\mathcal{X} \times \mathcal{Y}} a(y) dF(x, y).$$

If $a(y) = |y|^p$ for $p > 0$ we write $|F|_p$.

Let $L : \mathcal{Y} \times \mathbb{R} \rightarrow [0, \infty)$ be a convex loss function. Then the theoretical regularized risk is defined as

$$m_\gamma = \arg \min_{m \in \mathcal{H}} \mathbf{E}[L(Y, m(X))] + \gamma \|m\|_{\mathcal{H}}^2. \quad (4)$$

Consider the map T which assigns to every distribution F on $\mathcal{X} \times \mathcal{Y}$ with $|F|_a < \infty$, the function $T(F) = m_\gamma \in \mathcal{H}$. An expression for the influence function of (4) was proven in [21].

Proposition 1 ([21]): Let \mathcal{H} be a RKHS of a bounded continuous kernel K on \mathcal{X} with feature map $\varphi : \mathcal{X} \rightarrow \mathcal{H}$, $L : \mathcal{Y} \times \mathbb{R} \rightarrow [0, \infty)$ be a convex loss function satisfying some conditions [21] and denote $L'(y, r) := \partial L(y, r) / \partial r$ w.r.t. the second argument of L . Furthermore, let F be a distribution on $\mathcal{X} \times \mathcal{Y}$ with $|F|_a < \infty$. Then the IF of T exists for all $z = (z_x, z_y) \in \mathcal{X} \times \mathcal{Y}$ and is given by

$$\begin{aligned} \text{IF}(z; T, F) = & S^{-1} \{ \mathbf{E}[L'(Y, m_\gamma(X)) \varphi(X)] \\ & - L'(z_y, m_\gamma(z_x)) S^{-1} \varphi(z_x), \end{aligned}$$

with $m_\gamma = -\frac{1}{2\gamma} \mathbf{E}[h\varphi]$ and $S : \mathcal{H} \rightarrow \mathcal{H}$ defined as $S(m) = 2\gamma m + \mathbf{E}[L''(Y, m_\gamma(X)) \langle \varphi(X), m \rangle \varphi(X)]$.

From this proposition, it follows immediately that the IF only depends on z through the term

$$-L'(z_y, m_\gamma(z_x)) S^{-1} \varphi(z_x).$$

From a robustness point of view, it is important to bound the IF. It is obvious that this can be achieved by using a bounded kernel, e.g. the Gaussian kernel and a loss function with bounded first derivative e.g. L_1 loss or Vapnik's ε -insensitive loss. The L_2 loss on the other hand leads to an unbounded IF and hence is not robust.

B. Robustness by Reweighting

Although loss functions with bounded first derivative are easy to construct, they lead to more complicated optimization procedures such as QP problems. In case of least squares KBR (LS-KBR) this would mean that the L_2 loss should be replaced by e.g. an L_1 loss, what immediately would lead to a QP problem. In what follows we will study an alternative way of achieving robustness by means of reweighting. This has the advantage of easily computable estimates i.e. solving a weighted least squares problem in every iteration. First, we need the following definition concerning the weight function.

Definition 7: For $m \in \mathcal{H}$, let $w : \mathbb{R} \rightarrow [0, 1]$ be a weight function depending on the residual $Y - m(X)$ w.r.t. m . Then the following assumptions will be made on w

- $w(r)$ is a non-negative bounded Borel measurable function;
- w is an even function of r ;
- w is continuous and differentiable with $w'(r) \leq 0$ for $r > 0$.

A sequence of successive minimizers of a weighted least squares regularized risk is defined as follows.

Definition 8: Let $m_\gamma^{(0)} \in \mathcal{H}$ be an initial fit, e.g. obtained by ordinary unweighted LS-KBR. Let w be a weight function satisfying the conditions in Definition 7. Then the $(k+1)^{\text{th}}$ reweighted LS-KBR estimator is defined by

$$m_\gamma^{(k+1)} = \arg \min_{m \in \mathcal{H}} \mathbf{E} \left[w(Y - m_\gamma^{(k)}(X)) (Y - m(X))^2 \right] + \gamma \|m\|_{\mathcal{H}}^2. \quad (5)$$

It was proved by [22] and see also [23] that, under certain conditions, the IF of reweighted LS-KBR estimator (5) is bounded when $k \rightarrow \infty$ and is given as follows.

Proposition 2 ([22]): Denote by T_{k+1} the map $T_{k+1}(F) = m_\gamma^{(k+1)}$. Furthermore, let F be a distribution on $\mathcal{X} \times \mathcal{Y}$ with $|F|_2 < \infty$ and $\int_{\mathcal{X} \times \mathcal{Y}} w(y - m_\gamma^{(\infty)}(x)) dF(x, y) > 0$. Denote by T_∞ the map $T_\infty(F) = m_\gamma^{(\infty)}$. Denote the operators $S_{w,\infty} : \mathcal{H} \rightarrow \mathcal{H}$ and $C_{w,\infty} : \mathcal{H} \rightarrow \mathcal{H}$ given by

$$S_{w,\infty}(m) = \gamma m + \mathbf{E} \left[w \left(Y - m_\gamma^{(\infty)}(X) \right) \langle m, \varphi(X) \rangle \varphi(X) \right]$$

and

$$C_{w,\infty}(m) = -\mathbf{E} \left[w' \left(Y - m_\gamma^{(\infty)}(X) \right) \left(Y - m_\gamma^{(\infty)}(X) \right) \langle m, \varphi(X) \rangle \varphi(X) \right].$$

Further, assume that $\|S_{w,\infty}^{-1} \circ C_{w,\infty}\| < 1$. Then the IF of T_∞ exists for all $z = (z_x, z_y) \in \mathcal{X} \times \mathcal{Y}$ and is given by

$$\begin{aligned} \text{IF}(z; T_\infty, F) = & (S_{w,\infty} - C_{w,\infty})^{-1} \left\{ -\mathbf{E} \left[w \left(Y - m_\gamma^{(\infty)}(X) \right) \left(Y - m_\gamma^{(\infty)}(X) \right) \varphi(X) \right] \right. \\ & \left. + w \left(z_y - m_\gamma^{(\infty)}(z_x) \right) \left(z_y - m_\gamma^{(\infty)}(z_x) \right) \varphi(z_x) \right\}. \end{aligned}$$

The condition $\|S_{w,\infty}^{-1} \circ C_{w,\infty}\| < 1$ is needed to ensure that the IF of the initial estimator eventually disappears. Notice that the operators $S_{w,\infty}$ and $C_{w,\infty}$ are independent of the contamination z . Since $\|\varphi(x)\|_{\mathcal{H}}^2 = \langle \varphi(x), \varphi(x) \rangle = K(x, x)$, the $\text{IF}(z; T_\infty, F)$ is bounded if

$$\|w(r)r\varphi(x)\|_{\mathcal{H}} = w(r)|r|\sqrt{K(x, x)}$$

is bounded for all $(x, r) \in \mathbb{R}^d \times \mathbb{R}$. From Proposition 2, the following result immediately follows

Corollary 1: Assume that the conditions of Proposition 2 and Definition 7 are satisfied, then $\|\text{IF}(z; T_\infty, F)\|_{\mathcal{H}}$ bounded implies $\|\text{IF}(z; T_\infty, F)\|_\infty$ bounded for bounded kernels.

Proof: For any $m \in \mathcal{H} : \|m\|_\infty \leq \|m\|_{\mathcal{H}} \|K\|_\infty$. The result immediately follows for a bounded kernel K . ■

An interesting fact which has practical consequences is the choice of the kernel function. It is readily seen that if one takes a Gaussian kernel, only downweighting the residual is needed as the influence in the X -space is controlled by the kernel. On the other hand, taking an unbounded kernel such as the linear or polynomial kernel requires a weight function that decreases with the residual as well as with x to obtain a bounded IF. See also [24] and [25] for similar results regarding ordinary LS and [26] for iteratively defined statistics.

It does not suffice to derive the IF of the reweighted LS-KBR but also to establish conditions for convergence. The following proposition is due to [22].

Proposition 3 ([22]): Define $w(r) = \frac{\psi(r)}{r}$ with ψ the contrast function. Then, reweighted LS-KBR with a bounded kernel converges to a bounded influence, even if the initial LS-KBR is not robust, if

- (c1) $\psi : \mathbb{R} \rightarrow \mathbb{R}$ is a measurable, real, odd function;
- (c2) ψ is continuous and differentiable;
- (c3) ψ is bounded;
- (c4) $\mathbf{E}_{F_e} \psi'(e) > -\gamma$ where F_e denotes the distribution of the errors.

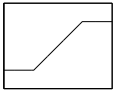
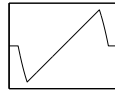
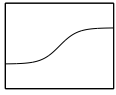
Finally, reweighting is not only useful when outliers are present in the data but it also leads to a more stable method, especially at heavy tailed distributions.

C. Weight Functions

It is without doubt that the choice of weight function w plays a significant role in the robustness aspects of the smoother. We will demonstrate later that the choice of weight function w has an influence on the speed of convergence [22], [25]. Table I illustrates three well-known weight functions (L is an invariant symmetric convex loss function). For further reading we refer the reader to [9]. We will show another kind of weight function, called Myriad, that exhibits some

remarkable properties. The Myriad is derived from the Maximum Likelihood (ML) estimation of a Cauchy distribution with scaling factor δ (see below) and can be used as a robust location estimator in stable noise environments. Given a set of i.i.d. random variables $X_1, \dots, X_n \sim X$ and $X \sim C(\zeta, \delta)$, where the location parameter ζ is to be estimated from data i.e. $\hat{\zeta}$ and $\delta > 0$ is a scaling factor.

TABLE I
DEFINITIONS FOR THE HUBER, HAMPEL AND LOGISTIC WEIGHT FUNCTIONS $w(r) = \psi(r)/r$. THE CORRESPONDING LOSS $L(r)$ AND CONTRAST FUNCTION $\psi(r)$ ARE ALSO GIVEN AND $\beta, b_1, b_2 \in \mathbb{N} \setminus \{0\}$.

	Huber	Hampel	Logistic
$w(r)$	$\begin{cases} 1, & r < \beta \\ \frac{\beta}{ r }, & r \geq \beta \end{cases}$	$\begin{cases} 1, & r < b_1; \\ \frac{b_2 - r }{b_2 - b_1}, & b_1 \leq r \leq b_2 \\ 0, & r > b_2 \end{cases}$	$\frac{\tanh(r)}{r}$
$\psi(r)$			
$L(r)$	$\begin{cases} r^2, & r < \beta \\ \beta r - \frac{\beta^2}{2}, & r \geq \beta \end{cases}$	$\begin{cases} r^2, & r < b_1 \\ \frac{b_2 r^2 - r ^3}{b_2 - b_1}, & b_1 \leq r \leq b_2 \\ 0, & r > b_2 \end{cases}$	$r \tanh(r)$

The ML principle yields the sample Myriad

$$\hat{\zeta}_\delta = \arg \max_{\zeta \in \mathbb{R}} \left(\frac{\delta}{\pi} \right)^n \prod_{i=1}^n \frac{1}{\delta^2 + (X_i - \zeta)^2},$$

which is equivalent to

$$\hat{\zeta}_\delta = \arg \min_{\zeta \in \mathbb{R}} \sum_{i=1}^n \log [\delta^2 + (X_i - \zeta)^2]. \quad (6)$$

Note that, unlike the sample mean or median, the definition of the sample Myriad involves the free parameter δ . We will refer to δ as the linearity parameter of the Myriad. The behavior of the Myriad estimator is markedly dependent on the value of its linearity parameter δ . Tuning the linearity parameter δ adapts the behavior of the myriad from impulse-resistant mode-type estimators (small δ) to the Gaussian-efficient sample mean (large δ). If an observation in the set of input samples has a large magnitude such that $|X_i - \zeta| \gg \delta$, the cost associated with this sample is approximately $\log(X_i - \zeta)^2$ i.e. the log of squared deviation. Thus, much as the sample mean and sample median respectively minimize the sum of square and absolute deviations, the sample myriad (approximately) minimizes the sum of logarithmic squared deviations. Some intuition can be gained by plotting the cost function (6) for various values of δ . Figure 3a depicts the different cost function characteristics obtained for $\delta = 20, 2, 0.75$ for a sample set of size 5. For a set of samples defined as above, an M-estimator of location is defined as the parameter ζ minimizing a sum of the form $\sum_{i=1}^n L(X_i - \zeta)$, where L is the cost or loss function. In general, when $L(x) = -\log f(x)$, with f a density, the M-estimate $\hat{\zeta}$ corresponds to the ML estimator associated with f .

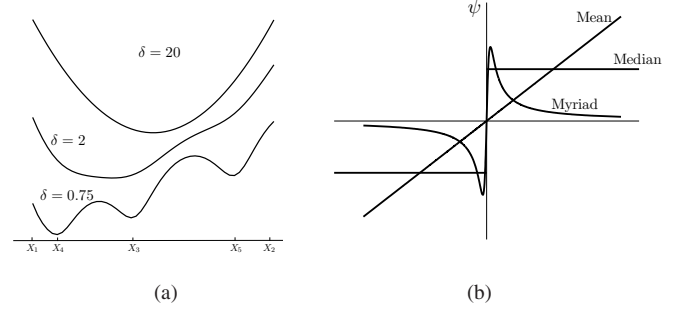


Fig. 3. (a) Myriad cost functions for the observation samples $X_1 = -3, X_2 = 8, X_3 = 1, X_4 = -2, X_5 = 5$ for $\delta = 20, 2, 0.2$; (b) Influence function for the mean, median and Myriad.

According to (6), the cost function associated with the sample Myriad is given by

$$L(x) = \log[\delta^2 + x^2].$$

Some insight in the operation of M-estimates is gained through the definition of the IF. For an M-estimate, the IF is proportional to the score function [3, p. 101]. For the Myriad (see also Figure 3b), the IF is given by

$$L'(x) = \psi(x) = \frac{2x}{\delta^2 + x^2}.$$

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 . Also, the Myriad has a mean property i.e. when $\delta \rightarrow \infty$ then the sample Myriad reduces to the sample mean. The following results were independently shown by [27] and [23].

Theorem 1 (Mean Property): Given a set of samples X_1, \dots, X_n . The sample Myriad $\hat{\zeta}_\delta$ converges to the sample mean as $\delta \rightarrow \infty$, i.e.

$$\begin{aligned} \hat{\zeta}_\infty &= \lim_{\delta \rightarrow \infty} \hat{\zeta}_\delta = \lim_{\delta \rightarrow \infty} \left\{ \arg \min_{\zeta \in \mathbb{R}} \sum_{i=1}^n \log [\delta^2 + (X_i - \zeta)^2] \right\} \\ &= \frac{1}{n} \sum_{i=1}^n X_i. \end{aligned}$$

Proof: First, we establish upper and lower bounds for $\hat{\zeta}_\delta$. Consider the order statistic $X_{(1)} \leq \dots \leq X_{(n)}$ of the sample X_1, \dots, X_n . Then, by taking $\zeta < X_{(1)} = \min\{X_1, \dots, X_n\}$ and for all i

$$\delta^2 + (X_i - X_{(1)})^2 < \delta^2 + (X_i - \zeta)^2,$$

it follows that $\hat{\zeta}_\delta \geq X_{(1)}$. Similarly, one can find that $\hat{\zeta}_\delta \leq$

$X_{(n)}$. Hence, ■

$$\begin{aligned}\hat{\zeta}_\delta &= \arg \min_{X_{(1)} \leq \zeta \leq X_{(n)}} \prod_{i=1}^n [\delta^2 + (X_i - \zeta)^2] \\ &= \arg \min_{X_{(1)} \leq \zeta \leq X_{(n)}} \delta^{2n} + \delta^{2n-2} \sum_{i=1}^n (X_i - \zeta)^2 + O(\delta^{2n-4}) \\ &= \arg \min_{X_{(1)} \leq \zeta \leq X_{(n)}} \sum_{i=1}^n (X_i - \zeta)^2 + \frac{O(\delta^{2n-4})}{\delta^{2n-2}}.\end{aligned}$$

For $\delta \rightarrow \infty$ the last term becomes negligible and

$$\hat{\zeta}_\infty \rightarrow \arg \min_{X_{(1)} \leq \zeta \leq X_{(n)}} \sum_{i=1}^n (X_i - \zeta)^2 = \frac{1}{n} \sum_{i=1}^n X_i.$$

As the Myriad moves away from the linear region (large values of δ) to lower values of δ , the estimator becomes more resistant to outliers. When δ tends to zero, the myriad approaches the mode of the sample.

Theorem 2 (Mode Property): Given a set of samples X_1, \dots, X_n . The sample Myriad $\hat{\zeta}_\delta$ converges to a mode estimator for $\delta \rightarrow 0$. Further,

$$\hat{\zeta}_0 = \lim_{\delta \rightarrow 0} \hat{\zeta}_\delta = \arg \min_{X_j \in \mathcal{K}} \prod_{X_i \neq X_j}^n |X_i - X_j|,$$

where \mathcal{K} is the set of most repeated values.

Proof: Since $\delta > 0$, the sample Myriad (6) can be written as

$$\arg \min_{\zeta \in \mathbb{R}} \prod_{i=1}^n \left[1 + \frac{(X_i - \zeta)^2}{\delta^2} \right].$$

For small values of δ , the first term in the sum, i.e. 1, can be omitted, hence

$$\prod_{i=1}^n \left[1 + \frac{(X_i - \zeta)^2}{\delta^2} \right] = O\left(\frac{1}{\delta^2}\right)^{n-\kappa(\zeta)}, \quad (7)$$

where $\kappa(\zeta)$ is the number of times that ζ is repeated in the sample X_1, \dots, X_n . The right-hand side of (7) is minimized for ζ when the exponent $n - \kappa(\zeta)$ is minimized. Therefore, $\hat{\zeta}_0$ will be a maximum of $\kappa(\zeta)$ and consequently, $\hat{\zeta}_0$ will be the most repeated value in the sample X_1, \dots, X_n or the mode.

Let $\kappa = \max_j \kappa(X_j)$ and $X_j \in \mathcal{K}$. Then,

$$\begin{aligned}\prod_{X_i \neq X_j}^n \left[1 + \frac{(X_i - X_j)^2}{\delta^2} \right] &= \prod_{X_i \neq X_j}^n \left[\frac{(X_i - X_j)^2}{\delta^2} \right] \\ &\quad + O\left(\frac{1}{\delta^2}\right)^{(n-\kappa)-1}.\end{aligned} \quad (8)$$

For small δ , the second term in (8) will be small compared to the first term, since this is of order $O\left(\frac{1}{\delta^2}\right)^{n-\kappa}$. Finally, $\hat{\zeta}_0$ can be computed as follows.

$$\begin{aligned}\hat{\zeta}_0 &= \arg \min_{X_j \in \mathcal{K}} \prod_{X_i \neq X_j}^n \left[\frac{(X_i - X_j)^2}{\delta^2} \right] \\ &= \arg \min_{X_j \in \mathcal{K}} \prod_{X_i \neq X_j}^n |X_i - X_j|.\end{aligned}$$

D. Speed of Convergence-Robustness Tradeoff

Define

$$d = \mathbf{E}_{F_e} \frac{\psi(e)}{e} \quad \text{and} \quad c = d - \mathbf{E}_{F_e} \psi'(e),$$

with F_e denoting the distribution of the errors, then c/d establishes an upper bound on the reduction of the influence function at each step [22]. 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 II this upper bound is calculated for a Normal distribution and a 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 $\beta, \delta \rightarrow 0$, indicating a high degree of robustness but slow convergence rate. Therefore, logistic weights offer a good tradeoff between speed of convergence and degree of robustness. Also notice the small ratio for the Hampel weights indicating a low degree of robustness. The highest degree of robustness is achieved by using Myriad weights.

TABLE II
VALUES OF THE CONSTANTS c , d AND c/d FOR THE HUBER, LOGISTIC, HAMPEL AND MYRIAD WEIGHT FUNCTION AT A STANDARD NORMAL DISTRIBUTION AND A 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/d	c/d
Huber	$\beta = 0.5$	0.46	0.47
	$\beta = 1$	0.25	0.31
Logistic		0.26	0.32
Hampel	$b_1 = 2.5$ $b_2 = 3$	0.006	0.025
Myriad	$\delta = 0.1$	0.92	0.91
	$\delta = 1$	0.47	0.50

E. Robust Selection of Tuning Parameters

It is shown in Figure 2 that also the model selection procedure plays a significant role in obtaining fully robust estimates. It is theoretically shown that a robust CV procedure differs from the Mean Asymptotic Squared Error (MASE) by a constant shift and a constant multiple [19]. Neither of these are dependent on the bandwidth. Further, it is shown that this multiple depends on the score function and therefore, also on the weight function. To obtain a fully robust procedure for LS-KBR one needs also, besides a robust smoother and bounded kernel, a robust model selection criterion. Consider

for example the robust LOO-CV (RLOO-CV) given by

$$\text{RLOO-CV}(\theta) = \frac{1}{n} \sum_{i=1}^n L\left(Y_i, \hat{m}_{n,\text{rob}}^{(-i)}(X_i; \theta)\right), \quad (9)$$

where L is a robust loss function e.g. L_1 , Huber loss, Myriad loss, $\hat{m}_{n,\text{rob}}$ is a robust smoother and $\hat{m}_{n,\text{rob}}^{(-i)}(X_i; \theta)$ denotes the leave-one-out estimator where point i is left out from the training and θ denotes the tuning parameter vector. A similar principle can be used in robust v -fold CV. For robust counterparts of GCV and complexity criteria see e.g. [28], [29] and [30]. Robust CV can also be transformed as a location estimation problem based on L -estimators (trimmed mean and Winsorized mean) to achieve robustness. See also [31] for model selection in kernel based regression using the influence function.

V. SIMULATIONS

A. Empirical Maxbias Curve

We compute the empirical maxbias curve (3) for a LS-KBR method and its robust counterpart iteratively reweighted LS-KBR (IRLS-KBR) on a test point. Given 150 “good” equispaced observations according to the relation [32, Chapter 4, p. 45]

$$Y_k = m(x_k) + e_k, \quad k = 1, \dots, 150,$$

where $e_k \sim \mathcal{N}(0, 0.1^2)$ and

$$m(x_k) = 4.26 (\exp(-x_k) - 4 \exp(-2x_k) + 3 \exp(-3x_k)).$$

Let $\mathcal{A} = \{x : 0.8 \leq x \leq 2.22\}$ denote a particular region (consisting of 60 data points) and let $x = 1.5$ be a test point in that region. In each step, we start to contaminate the region \mathcal{A} by deleting one “good” observation and replacing it by a “bad” point (x_k, Y_k^b) , see Figure 4a. In each step, the value Y_k^b is chosen as the absolute value of a standard Cauchy random variable. We repeat this until the estimation becomes useless. A maxbias plot is shown in Figure 4b where the values of the LS-KBR estimate (non-robust) $\hat{m}_n(x)$ and the robust IRLS-KBR estimate $\hat{m}_{n,\text{rob}}(x)$ are drawn as a function of the number of outliers in region \mathcal{A} . The tuning parameters are tuned with L_2 LOO-CV for KBR and RLOO-CV (9), based on an L_1 loss and Myriad weights, for IRLS-KBR. The maxbias curve of $\hat{m}_{n,\text{rob}}(x)$ increases very slightly with the number of outliers in region \mathcal{A} and stays bounded right up to the breakdown point. This is in strong contrast with the LS-KBR estimate $\hat{m}_n(x)$ which has a breakdown point equal to zero.

B. Real Life Data Sets

The octane data consists of NIR absorbance spectra over 226 wavelengths ranging from 1102 to 1552 nm. For each of the 39 production gasoline samples the octane number was measured. It is well known that the octane data set contains six outliers to which alcohol was added. Table III shows the result (median and median absolute deviation for each method are reported) of a Monte Carlo simulation (200 runs) of the IRLS-KBR and SVM in different norms on a randomly chosen test set of

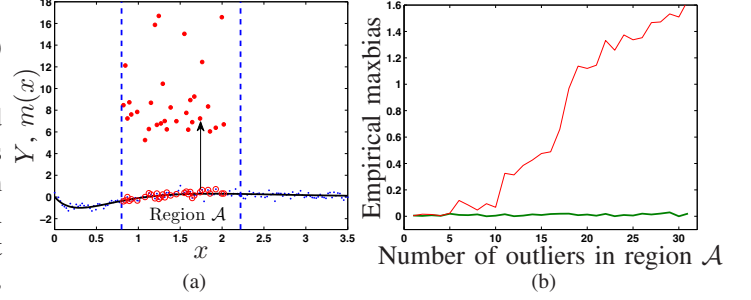


Fig. 4. (a) In each step, one good point (circled dots) of the the region $\mathcal{A} = \{x : 0.8 \leq x \leq 2.22\}$ is contaminated by the absolute value of a standard Cauchy random variable (full dots) until the estimation becomes useless; (b) Empirical maxbias curve of the LS-KBR estimator $\hat{m}_n(x)$ (thine line) and IRLS-KBR estimator $\hat{m}_{n,\text{rob}}(x)$ (bold line) in a test point $x = 1.5$.

size 10. Model selection was performed using robust LOO-CV (9). Minimizing the non-smooth robust LOO-CV surface was done via the procedure described in [33] to escape from local minima by means of a combination of a state-of-the-art global optimization technique with a simplex method.

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 III for randomly chosen test sets of size 15. 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 III. The number of iterations, needed by each weight function, confirms the results in Table II.

TABLE III
FOR 200 SIMULATIONS THE MEDIANS AND MEDIAN ABSOLUTE DEVIATIONS (BETWEEN BRACKETS) OF THE L_1 AND L_∞ NORMS ARE GIVEN (ON TEST DATA). i_{\max} DENOTES THE NUMBER OF ITERATIONS NEEDED TO CONVERGE. THE BEST RESULTS ARE BOLD FACED.

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

C. Importance of Robust Model Selection

An extreme example to show the absolute necessity of a robust model selection procedure (9) is given next. Consider 200 observations on the interval $[0, 1]$ and a low-order polynomial mean function

$$m(X) = 1 - 6X + 36X^2 - 53X^3 + 22X^5$$

and $X \sim \mathcal{U}[0, 1]$. The errors are generated from the gross error (1) model with the Normal distribution $N(0, 1)$ taken as nominal distribution and the contamination distribution is taken to be a cubed standard Cauchy with $\epsilon = 0.3$. We compare SVM, which is known to be robust, based on L_2 -CV and SVM based on robust model selection. The result is shown in Figure 5. This extreme example confirms the fact that, even if the smoother is robust, also the model selection procedure has to be robust in order to obtain fully robust estimates.

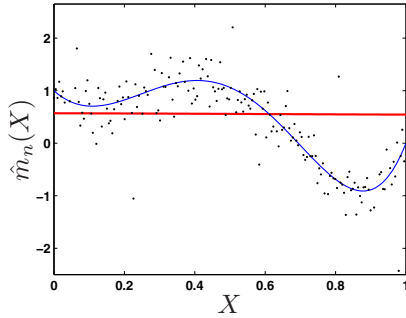


Fig. 5. SVM (bold line) cannot handle these extreme type of outliers and the estimate becomes useless. SVM based on robust model selection (thin line) can handle these outliers and does not break down. For visual purposes, not all data is displayed in the figure.

VI. CONCLUSION

We reviewed some measures of robustness and investigated the robustness of least squares kernel based regression. Although counterintuitive, robustness in the nonparametric regression case can be obtained by using a least squares cost function by means of iterative reweighting. In order to achieve a fully robust procedure, three requirements have to be fulfilled. By means of an upper bound for the reduction of the influence function in each step, we revealed the existence of a tradeoff between speed of convergence and the degree of robustness. Finally, we demonstrated that the Myriad weight function is highly robust against (extreme) outliers but exhibits a slow speed of convergence.

ACKNOWLEDGMENT

Research supported by Onderzoeksfonds K.U.Leuven/Research Council KUL: GOA/11/05 Ambiorics, GOA/10/09 MaNet, CoE EF/05/006 Optimization in Engineering (OPTEC) en PFV/10/002 (OPTEC), IOF-SCORES4CHEM, several PhD/postdoc & fellow grants; Flemish Government: FWO: FWO: PhD/postdoc grants, projects: G0226.06 (cooperative systems and optimization), G0321.06 (Tensors), G.0302.07 (SVM/Kernel), G.0320.08 (convex MPC), G.0558.08 (Robust MHE), G.0557.08 (Glycemia2), G.0588.09 (Brain-machine) research communities (WOG: ICCoS, ANMMM, MLDM); G.0377.09 (Mechatronics MPC), IWT: PhD Grants, Eureka-Flite+, SBO LeCoPro, SBO Climaps, SBO POM, O&O-Dquare, Belgian Federal Science Policy Office: IUAP 16/04 (DYSCO, Dynamical systems, control and optimization, 2007-2011), IBBT, EU: ERNSI, FP7-HD-MPC (INFOS-ICT-223854), COST intellICIS, FP7-EMBOCON (ICT-248940), Contract Research: AMINAL, Other: Helmholtz, viCERP, ACCM, JVDW and BDM are full professors at the Katholieke Universiteit Leuven, Belgium. JS is a professor at the Katholieke Universiteit Leuven, Belgium.

REFERENCES

- [1] J. W. Tukey, in *I. Olkin (Ed.), Contributions to Probability and Statistics*. Stanford University Press, 1960, ch. A survey of sampling from contaminated distributions, pp. 448–485.
- [2] P. J. Huber, “Robust estimation of a location parameter,” *The Annals of Mathematical Statistics*, vol. 35, no. 1, pp. 73–101, 1964.
- [3] F. R. Hampel, E. M. Ronchetti, P. J. Rousseeuw, and W. A. Stahel, *Robust Statistics: The Approach Based On Influence Functions*. New York: Wiley, 1986.
- [4] P. J. Huber, “A robust version of the probability ratio test,” *The Annals of Mathematical Statistics*, vol. 36, no. 6, pp. 1753–1758, 1965.
- [5] P. J. Huber and V. Strassen, “Minimax tests and the Neyman-Pearson lemma for capacities,” *The Annals of Statistics*, vol. 1, no. 2, pp. 251–263, 1973.

- [6] F. R. Hampel, “The influence curve and its role in robust estimation,” *Journal of the American Statistical Association*, vol. 69, no. 346, pp. 383–393, 1974.
- [7] M. Hubert, “Multivariate outlier detection and robust covariance matrix estimation - discussion,” *Technometrics*, vol. 43, no. 3, pp. 303–306, 2001.
- [8] J. A. K. Suykens, J. De Brabanter, L. Lukas, and J. Vandewalle, “Weighted least squares support vector machines: robustness and sparse approximation,” *Neurocomputing*, vol. 48, no. 1–4, pp. 85–105, 2002.
- [9] P. J. Rousseeuw and A. M. Leroy, *Robust Regression and Outlier Detection*. Wiley, 2003.
- [10] R. Maronna, D. Martin, and V. Yohai, *Robust Statistics: Theory and Methods*. Wiley, 2006.
- [11] P. J. Huber and E. M. Ronchetti, *Robust Statistics*, 2nd ed. Wiley, 2009.
- [12] J. Jurečková and J. Picek, *Robust Statistical Methods with R*. Chapman & Hall (Taylor & Francis Group), 2006.
- [13] L. T. Fernholz, *von Mises Calculus for Statistical Functionals*, ser. Lecture Notes in Statistics. Springer, 1983.
- [14] B. R. Clark, “Uniqueness and Fréchet differentiability of functional solutions to maximum likelihood type equations,” *The Annals of Statistics*, vol. 11, no. 4, pp. 1196–1205, 1983.
- [15] A. Christmann and A. Van Messem, “Bouligand derivatives and robustness of support vector machines for regression,” *Journal of Machine Learning Research*, vol. 9, pp. 915–936, 2008.
- [16] C. Croux and G. Haesbroeck, “Maxbias curves of robust scale estimators based on subranges,” *Metrika*, vol. 53, no. 2, pp. 101–122, 2001.
- [17] W. Härdle, P. Hall, and J. S. Marron, “How far are automatically chosen regression smoothing parameters from their optimum?” *Journal of the American Statistical Association*, vol. 83, no. 401, pp. 86–95, 1988.
- [18] Y. Yang, “Consistency of cross validation for comparing regression procedures,” *The Annals of Statistics*, vol. 35, no. 6, pp. 2450–2473, 2007.
- [19] D. H. Y. Leung, “Cross-validation in nonparametric regression with outliers,” *The Annals of Statistics*, vol. 33, no. 5, pp. 2291–2310, 2005.
- [20] I. Steinwart and A. Christmann, *Support Vector Machines*. Springer, 2008.
- [21] A. Christmann and I. Steinwart, “Consistency and robustness of kernel-based regression in convex risk minimization,” *Bernoulli*, vol. 13, no. 3, pp. 799–819, 2007.
- [22] M. Debruyne, A. Christmann, M. Hubert, and J. A. K. Suykens, “Robustness of reweighted least squares kernel based regression,” *Journal of Multivariate Analysis*, vol. 101, no. 2, pp. 447–463, 2010.
- [23] K. De Brabanter, “Least squares support vector regression with applications to large-scale data: a statistical approach,” Ph.D. dissertation, Faculty of Engineering, KU Leuven, April 2011.
- [24] M. B. Dollinger and R. G. Staudte, “Influence functions of iteratively reweighted least squares estimators,” *Journal of the American Statistical Association*, vol. 86, no. 415, pp. 709–716, 1991.
- [25] K. De Brabanter, K. Pelckmans, J. De Brabanter, M. Debruyne, J. A. K. Suykens, M. Hubert, and B. De Moor, “Robustness of kernel based regression: a comparison of iterative weighting schemes,” in *Proc. of the 19th International Conference on Artificial Neural Networks (ICANN)*, pp. 100–110, September 2009.
- [26] M. A. Jorgensen, “Influence function for iteratively defined statistics,” *Biometrika*, vol. 80, no. 2, pp. 253–265, 1993.
- [27] J. G. Gonzalez and G. R. Arce, “Optimality of the myriad filter in practical impulsive-noise environments,” *IEEE Transactions On Signal Processing*, vol. 49, no. 2, pp. 438–441, 2001.
- [28] M. A. Lukas, “Strong robust generalized cross-validation for choosing the regularization parameter,” *Inverse Problems*, vol. 24, no. 3, p. 034006 (16pp), 2008.
- [29] E. Ronchetti, “Robust model selection in regression,” *Statistics & Probability Letters*, vol. 3, no. 1, pp. 21–23, 1985.
- [30] P. Burman and D. Nolan, “A general Akaike-type criterion for model selection in robust regression,” *Biometrika*, vol. 82, no. 4, pp. 877–886, 1995.
- [31] M. Debruyne, M. Hubert, and J. A. K. Suykens, “Model selection in kernel based regression using the influence function,” *Journal of Machine Learning Research*, vol. 9, pp. 2377–2400, 2008.
- [32] G. Wahba, *Spline Models for Observational Data*. Philadelphia, PA: SIAM, 1990.
- [33] K. De Brabanter, J. De Brabanter, J. A. K. Suykens, and B. De Moor, “Optimized fixed-size kernel models for large data sets,” *Computational Statistics & Data Analysis*, vol. 54, no. 6, pp. 1484–1504, 2010.