Submitted to Statistical Science

Robustness by Reweighting for Kernel Estimators: An Overview

Kris De Brabanter and Jos De Brabanter

Department of Statistics (ISU), Department of Industrial and Manufacturing Systems Engineering (ISU) & Department of Electrical Engineering (KU Leuven)

Abstract. Using least squares techniques, there is an awareness of the dangers posed by the occurrence of outliers present in the data. In general, outliers may totally spoil an ordinary least squares 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. In this overview paper we illustrate that robust solutions can be acquired by solving a reweighted least squares problem even though the initial solution is not robust. This overview paper relates classical results from robustness to the most recent advances of robustness in least squares kernel based regression, with an emphasis on theoretical results as well as practical examples. Software for iterative reweighting is also made freely available to the user.

Key words and phrases: kernel based regression, robustness, iterative reweighting, influence function, robust model selection.

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 least squares (LS) analysis, see e.g. Rousseeuw & Leroy (2003). 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. Huber (1981) gave the following definition: a robust method is resistant to errors in the results, produced by deviations from assumptions. This means that if the assumptions are only approximately met, the robust estimator will still have a reasonable efficiency, and reasonably small bias, as well as being asymptotically unbiased, meaning having a bias tending towards 0 as the sample size tends towards infinity. One of the most important cases is distributional robustness (Huber, 1981). Classical statistical

Department of Statistics & Department of Industrial and Manufacturing Systems Engineering Iowa State University 2419 Snedecor Hall, Ames, IA, 50011, U.S.A. (e-mail: kbrabant@iastate.edu) Department of Electrical Engineering (ESAT-SCD) Katholieke Universiteit Leuven Kasteelpark Arenberg 10, B-3001 Leuven, Belgium (e-mail: jos.debrabanter@esat.kuleuven.be)

procedures are typically sensitive to "longtailedness" (e.g., when the distribution of the data has longer tails than the assumed normal distribution). This implies that they will be strongly affected by the presence of outliers in the data, and the estimates they produce may be heavily distorted if there are extreme outliers in the data, compared to what they would be if the outliers were not included in the data.

A *first attempt* was done by Edgeworth (1887). 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 (1960), Huber (1964) (minimax approach) and Hampel (1974) (influence functions). Huber (1964) gave the first theory of robustness by considering the general gross-error model or ϵ-contamination model

$$\mathcal{G}_\epsilon = \{F : F(x) = (1-\epsilon)F_0(x) + \epsilon G(x), 0 \leq \epsilon \leq 1\}, \tag{1.1}$$

where F_0 is some given distribution (the ideal nominal model), G an arbitrary continuous distribution and ϵ 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. Below are two examples.

EXAMPLE 1.1. *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 1.2. *ϵ-contamination model for the mixture of normal and Laplace distribution*

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

Note that G is an arbitrary continuous distribution and hence symmetry of G is not required. Huber also considered 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 (1.1). Huber (1964) developed a theory that finds the best strategy for choosing the loss function using only general information about the noise model (see Appendix B)

Huber (1965), Huber (1968), Huber & Strassen (1973) and Huber & Strassen (1974) developed a second theory 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 (1974), is closely related to robustness theory which is more generally applicable than Huber's first and second theory. Four main concepts are introduced:

1. Qualitative robustness, which is essentially continuity of the estimator viewed as functional in the weak topology.
2. Influence Curve (IC) or Influence Function (IF), which describes the first derivative of the estimator, as far as existing.

DEFINITION 1.1 (Influence Function). *Let F be a fixed distribution and $T(F)$ a statistical functional defined on a set $\mathcal{G}_\epsilon$ of distributions satisfying that T is Gâteaux differentiable at the distribution F in domain T. The influence function (IF) of T at F is given by*

$$\text{IF}(z;T,F) = \lim_{\epsilon \downarrow 0} \frac{T[(1-\epsilon)F + \epsilon\Delta_z] - T(F)}{\epsilon} = \lim_{\epsilon \downarrow 0} \frac{T(F_{\epsilon,z}) - T(F)}{\epsilon} \tag{1.2}$$

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

For a robust estimator, we want a bounded influence function, that is, one which does not go to infinity as z becomes arbitrarily large.
Robustness measures defined from the IF are: gross error sensitivity, local shift sensitivity and rejection point (Hampel *et al.*, 1986). Further, there are close relations to Tukey's Jackknife, and to Hoeffding's U-statistics (Hampel, 1974).

3. Maxbias Curve: gives the maximal bias that an estimator can suffer from when a fraction of the data comes from a contaminated distribution. By letting the fraction vary between zero and the breakdown value a curve is obtained.

DEFINITION 1.2. *Let $T(F)$ denote a statistical functional and let the contamination neighborhood of F be defined by $\mathcal{G}_\epsilon$ for a fraction of contamination ϵ. The maxbias curve is defined as*

$$B(\epsilon, T, F) = \sup_{F \in \mathcal{G}_\epsilon} |T(F) - T(F_0)|. \tag{1.3}$$

A robust estimator is expected to have a relatively small and stable (asymptotic) bias as F ranges over $\mathcal{G}_\epsilon$. The overall bias performance of $T(F)$ on the neighborhood $\mathcal{G}_\epsilon$ can then be measured by the maximum asymptotic bias 1.3. The maxbias is a function that depends on the fraction of contamination ϵ. A plot of $B(\epsilon, T, F)$ versus ϵ, called maxbias curve, conveys a complete robustness information for the given estimate.

4. Breakdown Point (BP): a global robustness measure describing how many percent gross errors are still tolerated before the estimator totally breaks down. The higher the breakdown point of an estimator, the more robust it is. Intuitively, the breakdown point cannot exceed 50% because if more than half of the observations are contaminated, it is not possible to distinguish between the underlying distribution and the contaminating distribution (Rousseeuw & Leroy, 2003). Hence, the maximum breakdown point is 0.5 and there are estimators which achieve such a breakdown point. For example, for the sample mean $\overline{X}_n$ we have

$$\overline{X}_n = \frac{1}{n}\sum_{i=1}^{n} X_i = \frac{1}{n}\left[\sum_{i=1}^{n-1} X_i + X_n\right] = \frac{n-1}{n}\overline{X}_{n-1} + \frac{1}{n}X_n$$

and if X_n is large enough, then $\overline{X_n}$ can be made arbitrarily large regardless of the other $n-1$ values. Hence the (finite) breakdown point of the sample mean is $1/n$ and asymptotic breakdown point is zero. In contrast, the sample median has an asymptotic breakdown point of 0.5.

DEFINITION 1.3 (Breakdown Point). *Let the estimator $T(\hat{F}_n)$ of $T(F)$ be the functional of the sample distribution $\hat{F}_n$. The breakdown point $\epsilon^\star$ of an estimator $T(\hat{F}_n)$ for the functional $T(F)$ at F is defined by*

$$\epsilon^\star(T,F) = \inf\{\epsilon > 0 | B(\epsilon, T, F) = \infty\}. \tag{1.4}$$

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, see Hubert (2001). 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 by Hampel *et al.* (1986), Rousseeuw & Leroy (2003) and Maronna *et al.* (2006). See also the book by Jurečková & Picek (2006) for robust statistical methods with R providing a systematic treatment of robust procedures with an emphasis on practical applications. Hettmansperger & McKean (2010) cover univariate tests and estimators with extensions to linear models, multivariate models, times series models, experimental designs and mixed models. Wilcox (2012) provides a thorough explanation of the foundation of robust methods, incorporating the latest updates in R and S-Plus, robust ANOVA (Analysis of Variance) and regression. Riazoshams (2018) discuss applications using R and cover a variety of theories and applications of nonlinear robust regression.

This overview paper is organized as follows. Section 2 summarizes the differences between parametric and nonparametric regression with a focus on robustness aspects. Section 3 illustrates and discusses the problems with outliers in nonparametric regression. Section 4 summarizes some recent results in the area of kernel based regression and robustness analysis. Section 5 relates influence functions to the leave-one-out cross-validation criterion. Section 6 discusses several properties of well-known weight functions and finally, Section 7 illustrates some results on real and toy data sets.

2. PARAMETRIC VS. NONPARAMETRIC (ROBUST) REGRESSION

Parametric models are models for which the parameter vector is finite dimensional. The dimensionality may be less or greater than the dimensionality of the explanatory variables or covariates. In parametric regression it is of interest to estimate that parameter vector. In this scenario, the user assumes the function that describes the relationship between the response and explanatory variables to be known. For example, the user can assume that this function is linear and errors follow a normal distribution. In general, the function can be assumed linear or nonlinear in the parameters. However, after fitting the model, the user should verify whether the assumptions of this model hold or not via residual analysis. For an in depth discussion of linear models, we refer the interested reader to Kutner *et al.* (2005) and Fox (2016).

One of the major differences between parametric and nonparametric regression is that for the latter no pre-specified form of the regression function is required, but instead the model is determined from the data. A common misconception

is to assume that nonparametric models do not have any parameters, rather their number of parameters are flexible and not fixed beforehand (the parameter space is of infinite dimension such as a Reproducing Kernel Hilbert Space). In the parametric case where we are interested in estimating the unknown (finite dimensional) parameter vector; in the nonparametric case, the set of parameters is a subset of an infinite dimensional vector space. Further, in the nonparametric case, the relationship between the response and explanatory variables is unknown. The function to be estimated can take on any shape maybe linear or nonlinear (but unknown to the user). We refer the interested reader to Wasserman (2005) and references therein for an overview of nonparametric modelling.

There exist a vast majority of excellent books and monographs (Hampel *et al.*, 1986; Rousseeuw & Leroy, 2003; Maronna *et al.*, 2006) and articles (Croux *et al.*, 1994; Wilcox, 1996; Yu & Yao, 2017) regarding parametric regression with outliers. Although the theoretical techniques to analyze nonparametric estimators strongly resemble those of their parametric counterparts (discussed in the aforementioned references), there are some major differences between both approaches. First, it is well-known that one single outlier can have a large effect on an ordinary least squares estimate. Consequently, this will lead to a global breakdown of a parametric regression estimator. In general, a global breakdown is unlikely to happen in case of nonparametric regression (depending on the weight or kernel function) for groups of outliers (see Figure 1(a)). Instead, there will be some local breakdown of the estimator close to the outlier or group of outliers. However, there are scenarios, with errors following the ϵ-contamination model (1.1), that can cause a complete breakdown of the estimator (see Figure 1(b)). Second, a robust loss function is not enough to obtain a robust solution in case of nonparametric regression. Most nonparametric regression estimators depend on so-called tuning parameters, often called bandwidth and/or smoothing parameter, which are often determined via cross-validation. The cross-validation also has to be robust (i.e. be based on a robust loss function) in order for the nonparametric estimator to be robust. In addition, if the weight/kernel function K satisfies $K(u) \to 0$ as $u \to \pm\infty$ has some implications for leverage points i.e. outliers in the X-direction (Christmann & Steinwart, 2007) (see also next sections). In case of parametric regression, a robust loss function does not imply robustness against leverage points.

3. PROBLEMS WITH OUTLIERS IN NONPARAMETRIC REGRESSION

Consider 200 uniformly distributed 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 1(a) shows the mean function with normally distributed errors with variance $\sigma^2 = 0.3^2$ and two distinct groups of outliers. Figure 1(b) shows the same mean function, but the errors are generated from the gross error or ϵ-contamination model (1.1). In this example $F_0 \sim N(0, 0.3^2)$, $G \sim N(0, 10^2)$ and $\epsilon = 0.3$. This simple example clearly shows that the nonparametric estimates based on the L_2 norm with classical L_2 cross-validation (CV) are influenced in a certain region or even break down (in case of the gross error model) in contrast to estimates based on robust kernel based regression (KBR) with robust CV. A fully robust KBR method will be discussed in Section 4.

Another important issue to obtain robustness in kernel based regression is

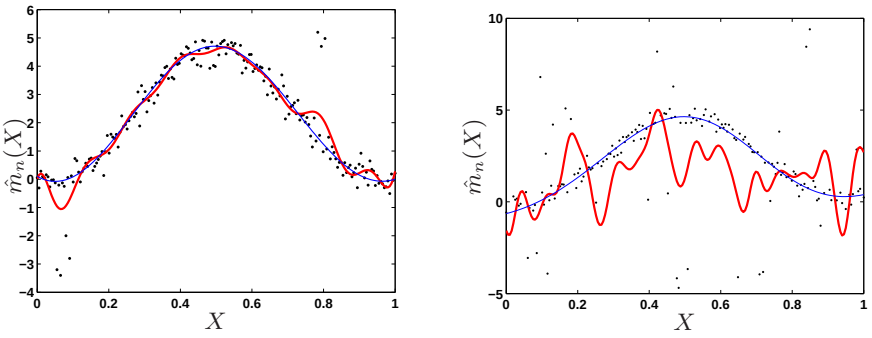


FIGURE 1. Kernel based regression estimates with (a) normal distributed errors and two groups of outliers; (b) the ϵ-contamination model. This clearly shows that the estimates based on the L_2 norm (bold line) are influenced in a certain region or regions in contrast to estimates based on robust loss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

the kernel function K. Continuous kernels that satisfy $K(u) \to 0$ as $u \to \pm\infty$ are bounded in $\mathbb{R}$. These type of kernels are called decreasing kernels. Common choices of decreasing kernels are: Epanechnikov, Gaussian, Laplace etc. Although the decreasing kernel assumption is needed to obtain robustness w.r.t. the Y-direction, it has some implications for leverage points i.e. outliers in the X-direction (Christmann & Steinwart, 2007).

Finally, acquiring a fully robust estimate also requires 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 (Härdle *et al.*, 1988). Under some regularity conditions and for an appropriate choice of data splitting ratio, Yang (2007) showed that 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 (Leung, 2005) resulting in bad regression estimates. The estimate can also fail when the tuning parameters are determined by CV with a squared loss function (L_2-CV) even when using with a robust smoother. The reason is that L_2-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. Indeed, it can be clearly seen that classical 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.

An extreme example to show the absolute necessity of a robust model selection procedure is given next. The errors are generated from the gross error model (1.1) with the same nominal distribution as above and the contamination distribution is taken to be a cubed standard Cauchy with $\epsilon = 0.3$. We compare the support vector machine (Vapnik, 1999), which is known to be robust (Christmann & Van Messem, 2008), with L_2-CV and the fully robust KBR (robust smoother and robust CV). The result is shown in Figure 3. This extreme example confirms the fact that, even if the smoother (based on a decreasing kernel) is robust, also the model selection procedure has to be robust in order to obtain fully robust estimates.

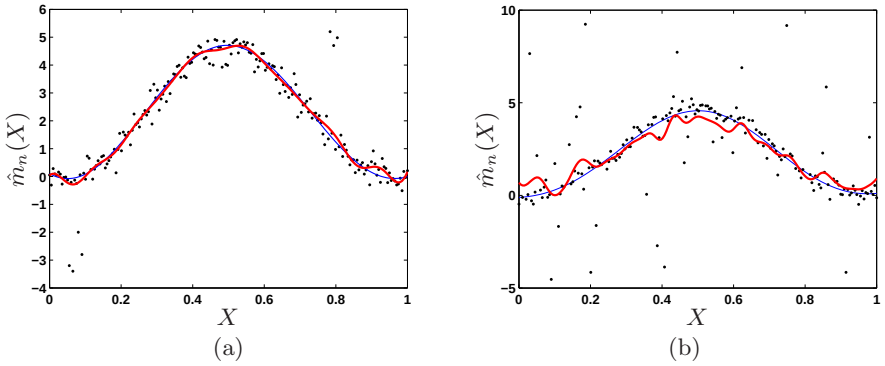


FIGURE 2. *KBR estimates and a similar type of errors as in Figure 1. The bold line represents the estimate based on classical L_2-CV and a robust smoother. The thin line represents estimates based on a fully robust procedure.*

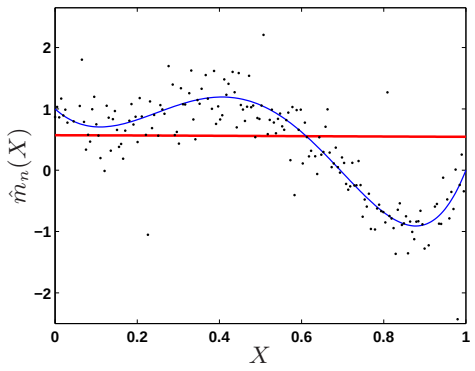


FIGURE 3. *SVM (bold line) cannot handle these extreme type of outliers and the estimate becomes useless. The folly robust KBR (full line) can clearly handle these outliers and does not break down. For visual purposes, not all data is displayed in the figure (the full range of the Y-axis is between -2000 and 2000).*

4. KERNEL BASED REGRESSION AND ITERATIVE REWEIGHTED KERNEL BASED REGRESSION: RESULTS TO DATE

4.1 Kernel based regression

Kernel based regression (KBR) methods estimate a functional relationship between a dependent variable X and an independent variable Y using a sample of n independent and identically distributed (i.i.d.) observations $(X_i, Y_i) \in \mathcal{X} \times \mathcal{Y} \subseteq \mathbb{R}^d \times \mathbb{R}$ with joint distribution F_{XY} from the model

$$Y_i = m(X_i) + e_i, \qquad i = 1, \ldots, n$$

where $\mathbf{E}(e_i \mid X_i) = 0, \mathbf{Var}(e_i \mid X_i) = \sigma^2 < \infty$, $\mathrm{cov}(e_i, e_j \mid X_i, X_j) = 0$ for $i \neq j$ and m is an unknown smooth function. In order to proceed, we need the following definitions from Steinwart & Christmann (2008).

DEFINITION 4.1. *Let $\mathcal{X}$ be a non-empty set. Then a function $K : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ 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 $\varphi : \mathcal{X} \to \mathcal{H}$ such that for all $x, y \in \mathcal{X}$ we have

$$K(x,y) = \langle \varphi(x),\, \varphi(y)\rangle_{\mathcal{H}}\,.$$

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

An example of a frequently used (isotropic[1]) kernel, when $\mathcal{X} = \mathbb{R}^d$, is the RBF kernel $K(u) = \exp(-u^2)$. Since the RBF kernel is an isotropic kernel the notation $K(x,y) = \exp(-\|x-y\|^2/h^2)$ is the same as $K(u) = \exp(-u^2)$ with $u = \|x-y\|/h$. In this case the feature space $\mathcal{H}$ is infinite dimensional (Schölkopf & Smola, 2002, Chap. 2). Also note that the RBF kernel is bounded since

$$\sup_{x,y\in\mathbb{R}^d} K(x,y) = 1.$$

Two other popular kernels when $\mathcal{X} = \mathbb{R}^d$ are the linear and polynomial kernel. For the linear kernel $K(x,x') = x^T x'$ one has that $\mathcal{H}$ equals $\mathbb{R}^d$ with φ the identity map (see to Definition 4.1).

Let $L : \mathcal{Y} \times \mathbb{R} \to [0,\infty)$ be a convex loss function w.r.t its second argument. Then, the theoretical regularized risk (Devito *et al.*, 2004) is defined as

$$m_{\gamma,K,F} = \arg\min_{m\in\mathcal{H}} \mathbf{E}_F\left[L\left(Y, m(X)\right)\right] + \gamma\|m\|^2_{\mathcal{H}}. \tag{4.1}$$

When the sample distribution F_n is used, one has that

$$m_{\gamma,K,F_n} = \arg\min_{m\in\mathcal{H}} \frac{1}{n}\sum_{i=1}^{n} L\left(Y_i, m(X_i)\right) + \gamma\|m\|^2_{\mathcal{H}}. \tag{4.2}$$

As shown by Evgeniou *et al.* (2000), the above minimization problem can also be seen as a particular case of Tikhonov regularization, see Tikhonov & Arsenin (1977) and Mukherjee *et al.* (2006) for a multivariate function approximation problem. The latter is well-known to be ill-posed, see e.g. Evgeniou *et al.* (2000) and Poggio & Smale (2003). Results about the form of the solution of KBR methods are known as representer theorems. They are based on the improvement of $m_{\gamma,K,F_n}(\cdot)$ by its projection to the linear span of $\{K(\cdot,X_i), i \le n\}$. A well-known result in statistical learning theory shows that

$$m_{\gamma,K,F_n}(\cdot) = \frac{1}{n}\sum_{i=1}^{n} \alpha_i K(\cdot,X_i). \tag{4.3}$$

The form of the coefficients α_i strongly depends on the loss function L. Only for the squared loss, the coefficients α_i are characterized as solutions of a system of linear equations (Tikhonov & Arsenin, 1977). For arbitrary convex differentiable loss functions, e.g. the logistic loss, the α_i are the solution of a systems of algebraic equations, see Girosi (1998), Wahba (1999) and Schölkopf *et al.* (2001). For an arbitrary convex loss function L, but possibly nondifferentiable (like the absolute value loss), extensions were obtained by Steinwart (2003) and Devito *et al.* (2004). In practice the variational problem (4.2) and its representation (4.3) are

[1] With such kernels the argument only depends on the distance between two points and not on multiplications between points such as the linear or polynomial kernel

closely related to the methodology of Support Vector Machines i.e., the final result of the variational problem, which fully characterizes the support vectors associated with the solution, coincides with the Karush-Kuhn-Tucker conditions of the dual quadratic programming problem as formulated in Vapnik (1995). A more in depth treatment of this close relationship is given in Devito *et al.* (2004, Section 7). Vapnik (1995) also extended this approach to the regression setting introducing Support Vector Regression (SVR) using the ϵ-insensitive loss function. A dual problem similar to (4.3) is solved, where the coefficients α_i are obtained from a quadratic programming problem. Suykens *et al.* (2002) use the least squares loss function which leads to a linear system of equations.

Before stating the influence function of (4.1) two technical definitions are required. First, we need a description for the growth of the loss function L, see e.g. Christmann & Steinwart (2007).

DEFINITION 4.2. *Let $L : \mathcal{Y} \times \mathbb{R} \to [0, \infty)$ be a loss function, $a : \mathcal{Y} \to [0, \infty)$ be a measurable function and $p \in [0, \infty)$. Then L is a loss function of type (a, p) if there exists a constant $c > 0$ such that*

$$L(y, t) \leq c(a(y) + |t|^p + 1)$$

for all $y \in \mathcal{Y}$ and all $t \in \mathbb{R}$. Furthermore, L is of strong type (a, p) if the first two partial derivatives $L'(y, t) = \frac{\partial}{\partial t} L(y, t)$ and $L''(y, t) = \frac{\partial^2}{\partial^2 t} L(y, t)$ of L with respect to the second argument of L exist and L, L' and L'' are functions of type (a, p).

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

DEFINITION 4.3. *Let F be a distribution on $\mathcal{X} \times \mathcal{Y}$, let $a : \mathcal{Y} \to [0, \infty)$ 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 $|F|_p$.

Regarding the theoretical regularized risk (4.1), Devito *et al.* (2004) proved the following explicit result by differentiating the penalized loss.

THEOREM 4.1. *Let $p = 1$, L be a convex (w.r.t. its second argument) loss function of type (a, p), and F be a distribution on $\mathcal{X} \times \mathcal{Y}$ with $|F|_a < \infty$. Let $\mathcal{H}$ be the reproducing kernel Hilbert space (RKHS) of a bounded, continuous kernel K over $\mathcal{X}$ and $\varphi : \mathcal{X} \to \mathcal{H}$ be the feature map of $\mathcal{H}$. Then with $h(x, y) = L'(y, m_{\gamma,K,F}(x))$ it holds that*

$$m_{\gamma,K,F} = -\frac{1}{2\gamma} \mathbf{E}_F[h\varphi]. \tag{4.4}$$

The assumption on the loss function to be of some type is a technical one; this is needed to ensure the continuity of certain terms appearing in the proof (Devito *et al.*, 2004). Also note that (4.4) does not provide an explicit solution to the problem as h depends on $m_{\gamma,K,F}$. The solution of the theorem above when considering the

empirical distribution F_n is given by (4.3). For didactic purposes, we have moved the definition of reproducing kernel Hilbert space (RKHS) to Appendix A. Consider the map T which assigns to every distribution F on $\mathcal{X} \times \mathcal{Y}$ with $|F|a < \infty$, the function $T(F) = m_{\gamma,K,F} \in \mathcal{H}$. An expression for the influence function (1.2) of T was proven in Christmann & Steinwart (2007). As before, the assumption on the loss function (in the next theorem) to be of some strong type is a technical one; this is needed to ensure the continuity and boundedness of certain terms appearing in the proof.

THEOREM 4.2. *Let $\mathcal{H}$ be a RKHS of a bounded continuous kernel K on $\mathcal{X}$ with feature map $\varphi : \mathcal{X} \to \mathcal{H}$, and $L : \mathcal{Y} \times \mathbb{R} \to [0, \infty)$ be a convex (w.r.t. its second argument) loss function of some strong type (a, p). Furthermore, let F be a distribution on $\mathcal{X} \times \mathcal{Y}$ with $|F|_a < \infty$. Then the IF of $T(F) := m_{\gamma,K,F}$ exists for all $z = (z_x, z_y) \in \mathcal{X} \times \mathcal{Y}$ and is given by*

$$\mathrm{IF}(z; T, F) = S^{-1} \left\{ \mathbf{E}_F \left[L' \left(Y, m_{\gamma,K,F}(X) \right) \varphi(X) \right] \right\} - L' \left(z_y, m_{\gamma,K,F}(z_x) \right) S^{-1} \varphi(z_x),$$

or equivalently using (4.4)

$$\mathrm{IF}(z; T, F) = -S^{-1} \left(2\gamma m_{\gamma,K,F} \right) - L' \left(z_y, m_{\gamma,K,F}(z_x) \right) S^{-1} \varphi(z_x)$$

with $S : \mathcal{H} \to \mathcal{H}$ defined as $S(m) = 2\gamma m + \mathbf{E}_F \left[L''(Y, m_{\gamma,K,F}(X)) \left\langle \varphi(X),\, m \right\rangle \varphi(X) \right]$.

From Theorem 4.2, it follows immediately that the IF only depends on z through the term

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

This theorem basically states that, for some kernel-based regression methods (4.1), not only the influence function exists but it also illustrates how to bound the influence function. 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 RBF kernel and a loss function with bounded first derivative e.g. the logistic loss. The L_2 loss on the other hand leads to an unbounded IF and hence is not robust. Although loss functions with bounded first derivative are easy to construct, they lead to more complicated optimization procedures such as quadratic programming problems e.g. support vector machines.

Note that the above theorem requires a twice continuously differentiable loss function and therefore cannot be used for methods using the L_1, ϵ-insensitive loss or Huber's loss function. Fortunately, an extension of the above theorem that applies to all convex (w.r.t. to its second argument) loss functions of some type (a, p) is given in Christmann & Steinwart (2007) and hence partially solves the problem for non-differentiable loss functions. Stronger results can be obtained if we are willing to consider invariant loss functions i.e., a loss function L is called invariant if there exist a function $l : \mathbb{R} \to [0, \infty)$ with $l(0) = 0$ and $L(y, t) = l(y - t)$ for all $y \in \mathcal{Y}, t \in \mathbb{R}$ (Christmann & Steinwart, 2007).

REMARK 4.1. *The result on the influence functions so far have been mostly expressed as population versions. It is however possible to obtain a finite sample version of Theorem 4.2. These finite sample version could offer some insights on bounding the influence function. Without loss of generality, Consider the L_2 loss*

and empirical distribution F_n. The operator S at F_n in Theorem 4.2 maps $m \in \mathcal{H}$ onto

$$\begin{pmatrix} S_{F_n}(m)(X_1) \\ \vdots \\ S_{F_n}(m)(X_n) \end{pmatrix} = 2\gamma \begin{pmatrix} m(X_1) \\ \vdots \\ m(X_n) \end{pmatrix} + \frac{2}{n} \begin{pmatrix} K(X_1, X_1) & \cdots & K(X_1, X_n) \\ \vdots & \ddots & \vdots \\ K(X_n, X_1) & \cdots & K(X_n, X_n) \end{pmatrix} \begin{pmatrix} m(X_1) \\ \vdots \\ m(X_n) \end{pmatrix}$$
$$= 2S_n \begin{pmatrix} m(X_1) \\ \vdots \\ m(X_n) \end{pmatrix}$$

implying that $2S_n$ is the finite sample version of the operator S at the sample F_n. Then from Theorem 4.2 it follows that the finite sample version of the IF is

$$\begin{pmatrix} IF(z_i; T, F_n)(X_1) \\ \vdots \\ IF(z_i; T, F_n)(X_n) \end{pmatrix} = S_n^{-1} \left\{ (m_{\gamma,K,F_n}(x_i) - y_i) \begin{pmatrix} K(x_i, X_1) \\ \vdots \\ K(x_i, X_n) \end{pmatrix} - \gamma \begin{pmatrix} m_{\gamma,K,F_n}(X_1) \\ \vdots \\ m_{\gamma,K,F_n}(X_n) \end{pmatrix} \right\}.$$

We can now evaluate the influence function at an arbitrary point z_i since m_{γ,K,F_n} and S_n are fully known. It is clear that this (empirical) IF can become very large as we have taken the L_2 loss function i.e. $(m_{\gamma,K,F_n}(x_i) - y_i)$ can be arbitrary large. More information about empirical IF's can be found in Debruyne et al. *(2008).*

In what follows we will look at 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.

4.2 Iterative reweighted least squares kernel based regression

The following definition is needed concerning the weight function w.

DEFINITION 4.4. *Let $m \in \mathcal{H}$. A function $w : R \to [0, 1]$, applied to the residual $Y - m(X)$ w.r.t. m, is called a weight function if the following assumptions hold:*

- *w 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 (Debruyne *et al.*, 2010). Let $m^{(0)}_{\gamma,K,F} \in \mathcal{H}$ be an initial fit, e.g. obtained by ordinary (unweighted) least squares kernel based regression (LS-KBR). Let w be a weight function satisfying the conditions in Definition 4.4. Then, the $(k+1)$th reweighted LS-KBR estimator is defined by

$$m^{(k+1)}_{\gamma,K,F} = \arg\min_{m \in \mathcal{H}} \mathbf{E}_F \left[w(Y - m^{(k)}_{\gamma,K,F}(X))(Y - m(X))^2 \right] + \gamma \|m\|^2_{\mathcal{H}}. \tag{4.5}$$

Let F be a distribution on $\mathcal{X} \times \mathcal{Y}$ with $|F|_2 < \infty$, then by combining results from Devito *et al.* (2004) and Debruyne *et al.* (2010) it follows that the $(k+1)$th reweighted LS-KBR estimator (4.5) can be written as

$$m^{(k+1)}_{\gamma,K,F} = \frac{1}{\gamma} \mathbf{E}_F \left[w(Y - m^{(k)}_{\gamma,K,F})(Y - m^{(k)}_{\gamma,K,F}(X))\, \varphi(X) \right].$$

The above formulation relates the current solution to the previous one without assuming that the solution will converge. In order to obtain such a result we require some assumptions on the weight function w, see Definition (4.4), then Debruyne *et al.* (2010) showed that there exist an $m^{(\infty)}_{\gamma,K,F} \in \mathcal{H}$ such that $m^{(k)}_{\gamma,K,F} \to m^{(\infty)}_{\gamma,K,F}$ as $k \to \infty$ and this limit must satisfy

$$m^{(\infty)}_{\gamma,K,F} = \frac{1}{\gamma}\,\mathbf{E}_F\left[w(Y - m^{(\infty)}_{\gamma,K,F}(X))(Y - m^{(\infty)}_{\gamma,K,F}(X))\,\varphi(X)\right] \tag{4.6}$$

and the solution is the unique minimizer of the theoretical risk (for the weight function w satisfying the assumptions in Definition (4.4)). Assume L is a symmetric convex loss function and suppose L is invariant i.e., there exists a function $l : \mathbb{R} \to [0, \infty)$ such that $L(y, m(x)) = l(y - m(x))$ for all $y \in \mathcal{Y}$, $x \in \mathcal{X}$ and $m \in \mathcal{H}$. Consider the choice $w(r) = l'(r)/(2r)$. If l is such that w satisfies the conditions of Definition 4.4, then from (4.6) it follows that $m^{(\infty)}_{\gamma,K,F}$ satisfies (4.4) from Theorem 4.1. Consequently, $m^{(\infty)}_{\gamma,K,F}$ is the unique minimizer of the theoretical risk (4.1) with loss L. Thus, the KBR solution for the loss L can be obtained as the limit of a sequence of reweighted LS-KBR estimators with arbitrary initial fit. Note that $|F|_2 < \infty$ is required to find the solution by reweighted LS-KBR. In general, $m^{(\infty)}_{\gamma,K,F}$ might depend on the initial fit and therefore leading to different solutions for different $m^{(0)}_{\gamma,K,F}$. This will be the case if L is non-convex and hence $m^{(\infty)}_{\gamma,K,F}$ can be a local minimum.

Next, the IF of reweighted LS-KBR estimator (4.5) is given by Debruyne *et al.* (2010) for $k \to \infty$.

THEOREM 4.3. *Denote by T_{k+1} the map $T_{k+1}(F) = m^{(k+1)}_{\gamma,K,F}$. Furthermore, let F be a distribution on $\mathcal{X}\times\mathcal{Y}$ with $|F|_2 < \infty$ and $\int_{\mathcal{X}\times\mathcal{Y}} w(y - m^{(\infty)}_{\gamma,K,F}(x))\,dF(x, y) > 0$. Denote by T_∞ the map $T_\infty(F) = m^{(\infty)}_{\gamma,K,F}$. Denote the operators $S_{w,\infty} : \mathcal{H} \to \mathcal{H}$ and $C_{w,\infty} : \mathcal{H} \to \mathcal{H}$ given by*

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

and

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

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

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

The condition $\|S^{-1}_{w,\infty} \circ C_{w,\infty}\| < 1$ 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 $\|\varphi(x)\|^2_{\mathcal{H}} = \langle\varphi(x),\, \varphi(x)\rangle = K(x, x)$, then the $\mathrm{IF}(z; T_\infty, F)$ is bounded if

$$\|w(r)r\varphi(x)\|_{\mathcal{H}} = w(r)|r|\sqrt{K(x, x)} \tag{4.7}$$

is bounded for all $(x, r) \in \mathbb{R}^d \times \mathbb{R}$. From Theorem 4.3 and Definition 4.4, it readily follows that $\|\operatorname{IF}(z; T_\infty, F)\|_{\mathcal{H}}$ bounded implies $\|\operatorname{IF}(z; T_\infty, F)\|_\infty$ bounded for bounded kernels, since for any $m \in \mathcal{H} : \|m\|_\infty \leq \|m\|_{\mathcal{H}} \|K\|_\infty$. If φ is the feature map of a linear kernel, then (4.7) corresponds to the conditions of Dollinger & Staudte (1991) for ordinary linear least squares. In that case, the weight function should decrease with the residual r as well as with x to obtain a bounded influence. This is also true for other unbounded kernels, e.g. the polynomial kernel. This does not hold for the popular RBF and Gaussian kernel. Here, downweighting the residual is the only requirement, as the influence in the x-space is controlled by the kernel. This shows that LS-KBR with bounded kernel is more suited for iterative reweighting than linear least squares regression. Similar conclusions concerning robustness and boundedness of the kernel were obtained in Christmann & Steinwart (2007) for classification and in Christmann & Steinwart (2004) for regression.

Deriving the IF of reweighted LS-KBR is useful from a robustness point of view, but on the other hand establishing conditions for convergence are equally important. Debruyne *et al.* (2010) showed that if the weight function $w(r) = \frac{\psi(r)}{r}, r \in \mathbb{R}$, with ψ the contrast function, satisfies the following conditions

- $(c1)$ $\psi : \mathbb{R} \to \mathbb{R}$ 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, than the reweighted LS-KBR with a bounded kernel converges to a bounded influence even if the initial LS-KBR is not robust. A key issue in proving condition (c4) is the operator norm $\|S_w^{-1} \circ C_w\|$. If the distribution F can be decomposed into an error distribution and a distribution of X, the reweighted LS-KBR is Fisher consistent and hence the ∞-indices in the operator norm can be omitted. Therefore, the operator norm becomes independent of the number of iterations k. Using the spectral theorem and the Fredholm alternative (Steinwart & Christmann, 2008), it can be shown that

$$\|S_w^{-1} \circ C_w\| = \frac{\mathbf{E}_{F_e} \frac{\psi(e)}{e} - \mathbf{E}_{F_e} \psi'(e)}{\mathbf{E}_{F_e} \frac{\psi(e)}{e} + \gamma}, \tag{4.8}$$

implying condition (c4). It is immediately clear that a positive generalization parameter γ improves the convergence of iteratively reweighted LS-KBR. Indeed, since higher values of γ will lead to smoother fits. In this case, the method will be less attracted towards an outlier in the Y-direction, hence leading to better robustness. Further, reweighting is not only helpful when outliers are present in the data, but it also leads to more stable methods especially at heavy tailed distributions. For a detailed discussion on the relation between IF and several stability criteria we refer the reader to Debruyne *et al.* (2010).

REMARK 4.2. *The choice of the weight function, $w(r) = L'(r)/(2r), r \in \mathbb{R}$, can be seen as a restriction. As far as we are aware, weight functions with $w(r) \neq L'(r)/(2r), r \in \mathbb{R}$ for any convex L is a topic of research for IRLS-KBR. Then, it might be that weighted LSE is not always the best choice. The reason however for*

taking the weight function of the form $L'(r)/(2r)$ is merely a theoretical one. By choosing the weight function of this form, Debruyne et al. (2010) showed that $S_{w,\infty} - C_{w,\infty} = \frac{1}{2}S$ with the operator S as in Theorem 4.2. Also, the IF of T_∞ (Theorem 4.3) equals the expression in Theorem 4.2 Indeed, (4.6) shows that T_∞ with weights $L'(r)/(2r)$ corresponds to KBR with loss function L. Hence their influence functions should coincide as well.

5. ROBUST CROSS VALIDATION AND INFLUENCE FUNCTIONS

When no outliers are present in the data, crossvalidation (CV) (or leave-one-out) has been shown to produce bandwidths that are asymptotically consistent (Härdle *et al.*, 1988, 1992) although convergence can be as slow as $n^{-1/10}$ (Härdle *et al.*, 1988). However, when outliers are present in the data, it has been empirically shown that the use of standard CV can lead to extremely biased bandwidth estimates (Leung, 1993). Standard CV fails, even when applied with a robust smoother, because it no longer produces a reasonable estimate of the prediction error. Therefore, a robust CV method may be superior. In what follows we will relate the leave-one-out criterion to influence functions.

The traditional leave-one-out criterion is given by

$$\text{LOO-CV}(\gamma, K) = \frac{1}{n}\sum_{i=1}^{n} L(y_i - m_{\gamma,K,F_n^{(-i)}}(x_i)), \tag{5.1}$$

where $m_{\gamma,K,F_n^{(-i)}}$ denotes the leave-one-out estimator where point i is left out from the training. Then, the kth order IF of T at F in the point z is defined as

$$\text{IF}_k(z; T, F) = \frac{\partial}{\partial^k \epsilon} T(F_{\epsilon,z}) \mid_{\epsilon=0} .$$

If all influence functions exist, then the following Taylor expansion holds

$$T(F_{\epsilon,z}) = T(F) + \epsilon\, \text{IF}(z; T, F) + \frac{\epsilon^2}{2!}\, \text{IF}_2(z; T, F) + \ldots$$

characterizing the estimate at a contaminated distribution in terms of the estimate at the original distribution and the influence functions. This is a special case of the more general Von Mises expansion. Let F be a distribution on $\mathcal{X} \times \mathcal{Y}$ with finite second moment and let L be a convex loss function such that the third derivative is zero. Then, Debruyne *et al.* (2008) have shown that the $(k+1)$th order IF of $m_{\gamma,K,F}$ exists for all $z := (z_x, z_y) \in \mathcal{X} \times \mathcal{Y}$ and is given by

$$\begin{aligned}\text{IF}_{k+1}(z, T, F) = (k+1) S^{-1} \big(& \mathbf{E}_F \left[\text{IF}_k(z; T, F)(X) L''(Y - m_{\gamma,K,F}(X)) \varphi(X) \right] \\ & - \left[\text{IF}_k(z; T, F)(z_x) L''(z_y - m_{\gamma,K,F}(z_x)) \varphi(z_x) \right] \big) ,\end{aligned}$$

where $S : \mathcal{H} \to \mathcal{H}$ is defined in Theorem 4.2. Let $F_n^{(-i)}$ denote the empirical distribution of a sample without the ith observation $z_i = (x_i, y_i) \in \mathcal{X} \times \mathcal{Y}$, then the following holds

$$m_{\gamma,K,F_n^{(-i)}}(x_i) = m_{\gamma,K,F_n}(x_i) + \sum_{j=1}^{\infty} \left(\frac{-1}{n-1} \right)^j \frac{\text{IF}_j(z_i; T, F_n)(x_i)}{j!}.$$

Let $[IFM_k]$ be the matrix containing $\mathrm{IF}_k(z_j; T, F_n)(x_i)$ at entry i, j. It can be shown that

$$[IFM_{k+1}] = (k+1)(H([IFM_k] \bullet M(1-n))),$$

with M the matrix containing $1/(1-n)$ at the off diagonal and 1 at the diagonal, $H = \left(\frac{\Omega}{n} + \gamma I_w\right)^{-1} \frac{\Omega}{n}$ with I_w a diagonal matrix with $w(y_i - m_{\gamma,K,F_n}(x_i))$ at entry i, i, Ω the kernel matrix with i, jth entry equal to $K(x_i, x_j)$ and $\bullet$ the Hamard product. By constructing a sample version of the operator $S(m)$ (Debruyne *et al.*, 2008) evaluated at F_n, it can be shown that

$$m_{\gamma,K,F_n^{(-i)}} \approx m_{\gamma,K,F_n} + \sum_{j=1}^{k-1} \frac{1}{(1-n)^j j!}[IFM_j]_{i,i} + \frac{1}{(1-n)^k k!}\frac{[IFM_k]_{i,i}}{1-[H]_{i,i}}, \tag{5.2}$$

where $[A]_{i,i}$ denotes the ith diagonal element of the matrix A. Plugging (5.2) in (5.1) yields the leave-one-out criterion based on influence functions. In practice, several choices need to be made w.r.t. k and $\mathcal{L}$. It was empirically shown in Debruyne *et al.* (2008) that setting $k = 5$ yields good results. It is important to note that these results hold for a fixed choice of γ and kernel K. If these parameters are selected in a data driven way, outliers might have a large influence on the selection the parameters. Even if a robust estimator is used, one can expect bad results if wrong choices are made for the parameters due to outliers. The latter was rigorously investigated by Leung (2005) for several linear smoothers. Leung (2005) showed that, if a robust loss function L (e.g. L_1 or Huber loss) and a robust smoother are used together, robust CV differs from the average squared error by a constant shift and a constant multiple; both of which are asymptotically independent of the kernel bandwidth. One of the main reasons (5.2) was proposed is that it provides a fast computational alternative for the standard leave-one-out CV.

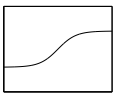
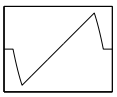
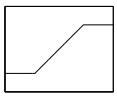
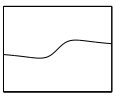
6. WEIGHT FUNCTIONS AND CONVERGENCE

6.1 Weight functions

It is without doubt that the choice of weight function w plays a significant role in the robustness aspects and convergence of the iteratively reweighted LS-KBR. We will demonstrate later that the choice of weight function w has an influence on the speed of convergence, see De Brabanter *et al.* (2009) and Debruyne *et al.* (2010). Table 1 illustrates four weight functions (L is an invariant symmetric convex loss function). For a comprehensive overview about different weight functions and their properties we refer the reader to Huber (1981), Hampel *et al.* (1986), Simpson *et al.* (1992) and Rousseeuw & Leroy (2003). We also show another kind of weight function, called Myriad or Cauchy weight function, that exhibits some remarkable properties. The Myriad, a redescending M-estimator, 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. Stable distributions are a class of probability distributions suitable for modeling heavy tails and skewness (see Appendix C).

Given a set of i.i.d. random variables $X_1, \ldots, X_n \sim X$ and $X \sim C(\zeta, \delta)$, where the location parameter ζ is to be estimated from data with $\delta > 0$ a scaling factor.

TABLE 1
Definitions for the Logistic, Hampel, Huber and Myriad weight functions $w(\cdot)$. The corresponding loss $L(\cdot)$ and score function $\psi(\cdot)$ are also given.

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

The ML principle yields the sample Myriad or Cauchy weight function

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

which is equivalent to

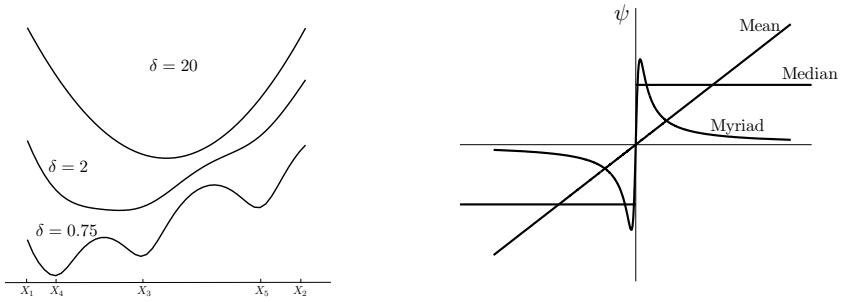
$$\hat{\zeta}_\delta = \arg\min_{\zeta \in \mathbb{R}} \sum_{i=1}^n \log\left[\delta^2 + (X_i - \zeta)^2\right]. \tag{6.1}$$

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 squares and absolute deviations, the sample myriad (approximately) minimizes the sum of logarithmic squared deviations. Some intuition can be gained by plotting the cost function (6.1) for various values of δ. Figure 4(a) 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. According to (6.1), 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-estimator is gained through the definition of the IF. For an M-estimator, the IF is proportional to the score function (Hampel *et al.*, 1986, p. 101). For the Myriad (see also Figure 4(b)), the IF is given by

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



(a) (b)

FIGURE 4. *(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.*

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 $\delta \to \infty$ then the sample Myriad reduces to the sample mean, see e.g. Arce (2005) and De Brabanter *et al.* (2009).

THEOREM 6.1 (Mean Property). *Given a set of samples* $X_1, \ldots, X_n$*. The sample Myriad* $\hat{\zeta}_\delta$ *converges to the sample mean as* $\delta \to \infty$*, i.e.*

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

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 6.2 (Mode Property). *Given a set of samples* $X_1, \ldots, X_n$*. The sample Myriad* $\hat{\zeta}_\delta$ *converges to a mode estimator for* $\delta \to 0$*. Further,*

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

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

REMARK 6.1. *In order to obtain a fully robust LS-KBR estimator via reweighting, the following three settings are required: 1) a weight function for the residuals as given in Table 1; 2) a bounded kernel, for example, the RBF kernel and 3) a cross-validation procedure based on, for example, the* L_1 *norm and not the* L_2 *norm. Other loss functions are also possible (see bottom row of Table 1). We also refer the reader to the simulation section for these settings.*

6.2 Convergence

Equation (4.8) establishes an upperbound on the reduction of the influence function of the initial estimator 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 (4.8), the higher the degree of robustness but the slower the reduction of the influence function of the initial estimator at each step and vice versa. In Table 2 this upper bound is calculated for a Normal distribution, a standard Cauchy and a t-distribution with 5 degrees of freedom 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. Thus when reweighting LS-KBR to obtain L_1-KBR, no fast convergence can be guaranteed, since the upperbound on the reduction factor approaches 1 as $\beta \to 0$. Similar results hold for Myriad weights when $\delta \to 0$. Logistic weights are doing quite well even at heavy tailed distributions such as the Cauchy, the influence of the initial estimator is reduced to 0.32 of the value of the previous step. This means that after k steps, at most 0.32^k is left of the influence of the initial estimator. From these result results it can be seen that Myriad weights offer the most robustness at the expense of a slower convergence. Logistic weights can be viewed as a good tradeoff between robustness and speed of convergence.

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 t-distribution with five degrees of freedom. The bold values represent an upper bound for the reduction of the influence function at each step.

Weight function	Parameter settings	$N(0,1)$			$C(0,1)$			t_5		
		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**	0.31	0.67	**0.46**
	$\beta = 1$	0.22	0.91	**0.25**	0.22	0.72	**0.27**	0.23	0.87	**0.27**
	$\beta = 2$	0.04	0.99	**0.04**	0.14	0.85	**0.17**	0.08	0.98	**0.08**
Logistic		0.22	0.82	**0.26**	0.21	0.66	**0.32**	0.22	0.79	**0.28**
Hampel	$b_1 = 2.5$ $b_2 = 3$	0.05	0.99	**0.05**	0.20	0.77	**0.26**	0.13	0.95	**0.14**
Myriad	$\delta = 0.1$	0.11	0.12	**0.92**	0.083	0.091	**0.91**	0.10	0.11	**0.92**
	$\delta = 0.6475$	0.31	0.53	**0.60**	0.24	0.39	**0.61**	0.30	0.49	**0.61**
	$\delta = 1$	0.31	0.66	**0.47**	0.25	0.50	**0.50**	0.30	0.62	**0.49**

REMARK 6.2. *Iteratively reweighted least squares (IRLS) has been well studied under conditions of uncorrupted data or when the noise is assumed to be Gaussian. Convergence results for these settings are dependent on p for L_p loss functions i.e. global linear for $p = 1$ and local super-linear for $0 < p < 1$ (Ba et al., 2014). However, in case of robust loss functions (as the ones mentioned in the paper for example), guaranteeing convergence results is absolutely not trivial. Typically, if one is working with convex loss functions, Osborne (1985) and Bissantz et al. (2009) demonstrated that IRLS approaches small loss function values. In case*

of non-convex loss functions, usually monotonicity is guaranteed but no global convergence (Aftab & Hartley, 2015). Recently, Mukhoty et al. *(2019) provide a stronger analysis based on a truncated variant of IRLS guaranteeing global linear convergence (including the Huber and L_1 loss) under mild conditions. Their method is based on extending basic notions of convexity and smoothness to the corresponding weighted versions. Further rigorous analysis of IRLS can be found in the technical paper of Sigl (2016).*

7. SIMULATIONS

All simulations have been performed using the freely available software `StatLSSVM` (De Brabanter *et al.*, 2013).

7.1 Empirical Maxbias Curve

We compute the empirical version of the maxbias curve (1.3) for a LS-KBR method and its robust counterpart iteratively reweighted LS-KBR on a test point. Given 150 "good" equispaced observations according to the relation Wahba (1990, Chapter 4, p. 45)

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

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

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

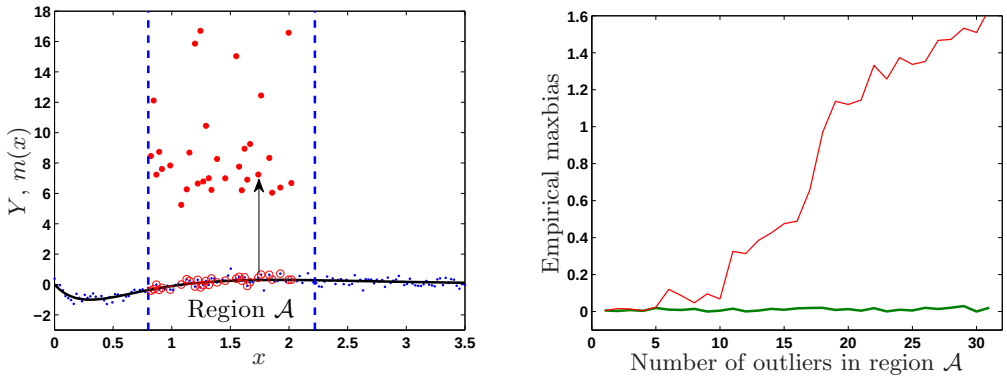
Let $\mathcal{A} = \{x : 0.8 \leq x \leq 2.22\}$ denote a particular region (consisting of 60 data points) and let $x = 1.5$ be a test point in that region. In each step, we start to contaminate the region $\mathcal{A}$ by deleting one "good" observation and replacing it by a "bad" point (x_k, Y_k^b), see Figure 5(a). 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 5(b) where the values of the LS-KBR estimate (non-robust) and the robust IRLS-KBR estimate 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 (5.2), based on an L_1 loss and Myriad weights, for IRLS-KBR. The maxbias curve increases only slightly with an increasing 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 which has a breakdown point equal to zero.

7.2 Outliers and leverage points

In this section we provide examples to illustrate the following concepts

- The IRLS-KBR is more robust than classical KBR based on the least squares loss function when outliers are present in the Y direction.
- In general, there is no hope in obtaining robust predictions with IRLS-KBR if x belongs to a subset of the design space that is sparse, i.e., x is a leverage point.

In the simulation below the sample size is $n = 101$. We consider the linear model $Y_i = x_i + e_i$, where the variable x_i ranges from -5 to 5 and the error e_i is a random variable from a standard normal distribution. All hyperparameters are tuned via robust cross validation and we consider the RBF and linear kernel for IRLS-KBR.



(a) (b)

FIGURE 5. *(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* $\hat{m}_{n,rob}(x)$ *(bold line) in a test point* $x = 1.5$.

In case of no outliers, classical KBR (based on L_2 loss) and IRLS-KBR (both based on the RBF kernel) almost result in the same fitted curves, see Figure 6(a). In contrast, Figure 6(b) illustrates that IRLS-KBR is much less influenced by the two outliers (two points moved to $(x, y) = (-2.5, 30)$ and $(x, y) = (-2, 30)$) in the Y direction compared to classical KBR. Both fitted curves are based on the RBF kernel. Next, we sequentially add three samples to the original data set all equal to $(x, y) = (100, 0)$. These three points are leverage points w.r.t. to a linear regression model. The number of leverage points clearly has a large impact on IRLS-KBR with linear kernel while the predictions of IRLS-KBR with RBF kernel are stable but nonlinear, see Figure 6(c). Finally, we consider the impact of the prediction of IRLS-KBR by adding two data points ($(x, y) = (100, 0)$ and $(x, y) = (100, 100)$) to the original data set. None of the two regression models is able to fit all data points well since the x components of both added points is equal by construction. The latter shows that there is no hope in obtaining robust predictions with IRLS-KBR if x belongs to a subset of the design space that is sparse.

7.3 Real Data Sets

Consider two real life data sets frequently used in robust statistics. The octane data (Hubert *et al.*, 2005) consist of NIR absorbance spectra over 226 wavelengths ranging from 1102 to 1552 nm. For each of the 39 production gasoline samples the octane number Y was measured. It is well known that the octane data set contains six outliers to which alcohol was added. Table 3 shows the result (medians and mean absolute deviations) of a Monte Carlo simulation (200 times) of the iteratively reweighted LS-KBR with different weight functions 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 3 for randomly chosen test sets of size 15. To illustrate the tradeoff between the degree of robustness and speed of convergence, the number of iterations $k_{\max}$ are also given in

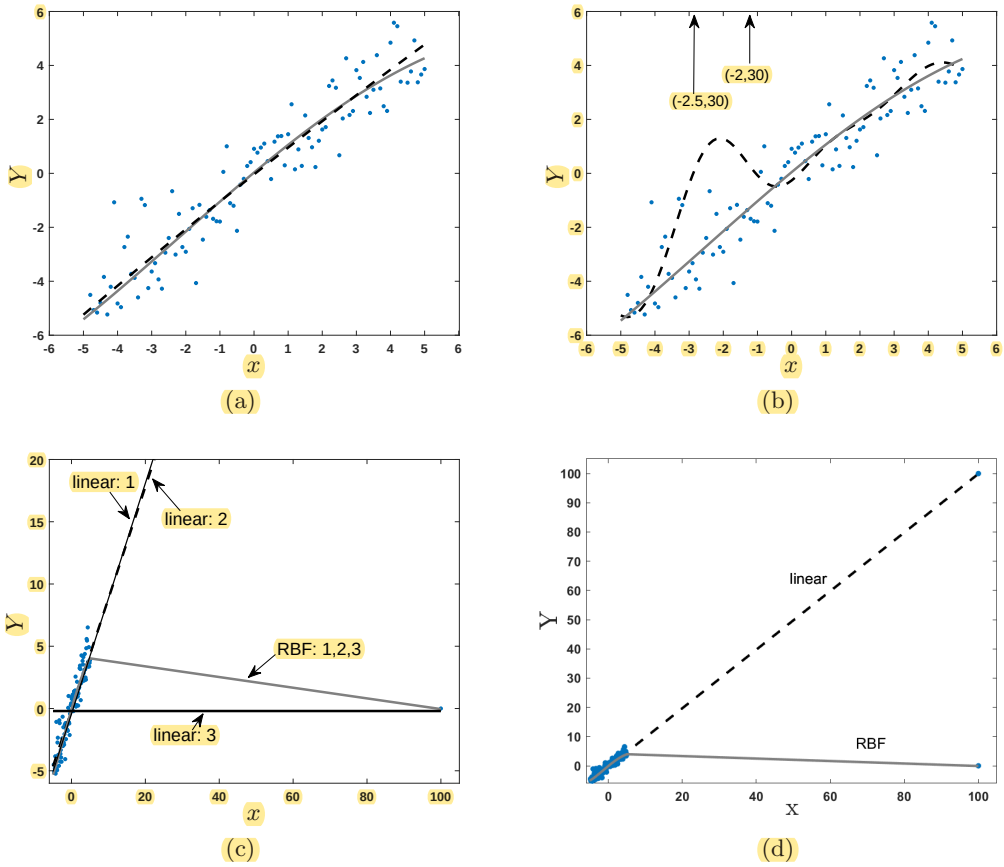


FIGURE 6. *Results for the simulated data sets with or without outliers. (a) no outliers: classical KBR (dashed) and IRLS-KBR (full) almost result in identical fitted curves. Both are based on the RBF kernel. (b) Linear relationship with two outliers in the Y direction at* $(x, y) = (-2.5, 30)$ *and* $(x, y) = (-2, 30)$. *IRLS-KBR with RBF kernel (solid) is more robust than classical KBR (dashed). (c) Linear relationship with additional 1, 2 and 3 points in* $(x, y) = (100, 0)$. *IRLS-KBR with linear kernel (dashed) and RBF kernel (solid). (d) Linear relationship with two additional data points* $(x, y) = (100, 0)$ *and* $(x, y) = (100, 100)$ *to the original data set. IRLS-KBR with linear kernel (dashed) and RBF kernel (solid).*

Table 3. The number of iterations, needed by each weight function, are confirmed by the results in Table 2.

TABLE 3

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). k_{max} *denotes the number of iterations for each weight functions. The best results are bold faced.*

	Octane				Demographic			
weights	L_1	L_2	L_∞	$k_{\max}$	L_1	L_2	L_∞	$k_{\max}$
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

8. CONCLUSIONS

Outliers are a common occurrence in practical applications. In nonparametric regression, these outliers can have important consequences on the statistical properties of the estimator and on the selection of the smoothing parameter using data-driven methods. We have reviewed the existing literature on the effect of outliers in case of iterative reweighted least squares kernel based regression. We have illustrated, by means of influence functions, that robust least squares kernel based regression estimates can be obtained by iterative reweighting. Even if the initial fit is not robust, robustness can be guaranteed by simple conditions on the weight function.

The techniques reviewed in this article were shown to be able to handle the outliers. When outliers are present in the data, data-driven smoothing parameter selection methods have to be adequately adapted.

Currently, optimization based results regarding this topic are more extensive than those we have investigated in this paper. The reason for this is that one can use or design loss functions with bounded first derivative and hence the problem can be often reformulated as a convex optimization problem. Perhaps one of the pressing areas for future research is the implementation of many of the existing theoretical and methodological results into fast algorithms and software. Although, we have shown that one can obtain robust solutions by simply solving several weighted least squares problems, computational complexity quickly grows when dealing with large data sets.

APPENDIX A: DEFINITION OF REPRODUCING KERNEL HILBERT SPACE

DEFINITION A.1. *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 $K : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ 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 $m(x) = \langle m, K(\cdot, x)\rangle_{\mathcal{H}}$ 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 $\delta_x : \mathcal{H} \to \mathbb{R}$ defined by*

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

is continuous.

APPENDIX B: LOSS FUNCTIONS FOR ROBUST ESTIMATORS

Huber (1964) developed a theory that allows finding the best strategy for choosing the loss function only using general information about the noise model.

THEOREM B.1. *Let $-\log f_0 \in C^2$, then the class of densities $\{f : f(x) = (1-\epsilon)f_0(x)+\epsilon g(x), 0 \leq \epsilon \leq 1\}$, as defined in (1.1), possesses the following robust density*

$$f_{robust}(x) = \begin{cases} (1-\epsilon)f_0(a)\exp\{-c(a-x)\}, & x < a; \\ (1-\epsilon)f_0(x), & a \leq x < b; \\ (1-\epsilon)f_0(b)\exp\{-c(x-b)\}, & x \geq b, \end{cases} \tag{B.1}$$

where $a, b \in \mathbb{R}$ are endpoints of the interval $[a, b]$ on which the monotonic function

$$-\frac{d \log f_0(x)}{dx} = -\frac{f_0'(x)}{f_0(x)}$$

is bounded in absolute value by a constant $c \in \mathbb{R}$ determined by the normalization condition

$$(1-\epsilon)\left(\int_a^b f_0(x)\,dx + \frac{f_0(a)+f_0(b)}{c}\right) = 1.$$

Based on the robust density (B.1) it is possible to construct a robust regression estimator that minimizes the empirical risk $-\sum_{i=1}^n -\log f_{robust}(\zeta_i)$ where $\zeta_i = Y_i - m(X_i)$. For example, let f_0 be the normal density and consider the above class of densities, then the loss functions derived from this robust density is the Huber loss function (see Table 1).

APPENDIX C: STABLE DISTRIBUTIONS

Stable distributions are a class of probability distributions suitable for modeling heavy tails and skewness. A univariate stable distribution uses the following parameters in Table 4. We denote stable distributions by $S_\alpha(\gamma, \beta, \delta)$. In

TABLE 4
Parameters of stable distributions.

Parameter	Description	Domain	Remarks
α	index of stability	$0 < \alpha \leq 2$	This parameter determines the probability in the extreme tails. A normal distribution has $\alpha = 2$. Distributions below that number have heavier tails. The Cauchy distribution has $\alpha = 1$.
β	skewness parameter	$-1 \leq \beta \leq 1$	For $\beta = 0$, then the distribution is symmetric.
γ	scale parameter	$0 < \gamma < \infty$	A measure of dispersion. For the normal distribution, γ equals half of the population variance.
δ	location parameter	$-\infty < \delta < \infty$	This parameter equals the median. When $\alpha > 1$, it also equals the mean. For a normal distribution, the sample mean can be used as an estimate for δ

the literature, univariate stable distributions are defined in several equivalent ways (Samorodnitsky, 1994). In this overview paper we give a definition in the context of the central limit theorem.

DEFINITION C.1. *A random variable X is said to have a stable distribution if for any $n \geq 1$, there exists constants $a_n > 0$ and b_n and independent random variables $\varsigma_1, \ldots, \varsigma_n$, distributed like X such that*

$$a_n T + b_n \stackrel{d}{=} \varsigma_1 + \ldots + \varsigma_n.$$

The probability densities of α-stable random variables exist and are continuous but, with a few exceptions, they are not known in closed form (Zolotarev, 1986). The exceptions are: Gaussian ($S_2(\gamma, 0, \delta)$), Cauchy ($S_1(\gamma, 0, \delta)$) and Lévy ($S_{0.5}(\gamma, 1, \delta)$) distribution.

REFERENCES

Aftab K. & Hartley R. (2014). *Convergence of iteratively re-weighted least squares to robust M-estimators*. IEEE Winter Conference on Applications of Computer Vision (WACV), p. 480–487.

Arce G.R. (2005). *Nonlinear Signal Processing: A Statistical Approach*, Wiley, New York. MR2501508

Ba D., Babadi B., Purdon P. L. & Brown E. N. (2014). Convergence and stability of iteratively re-weighted least squares algorithms. *IEEE Trans. Signal Process.*, **1**, 183—195. MR3187987

Bissantz N., Dümbgen L., Munk A., & Stratmann B. (2009). Convergence analysis of generalized iteratively reweighted least squares algorithms on convex function spaces. *SIAM J. Optim.*, **19**, 1828—1845, 2009. MR2486052

Christmann A. & Steinwart I. (2004). On robustness properties of convex risk minimization methods for pattern recognition. *J. Mach. Learn. Res.*, **5**, 1007—1034. MR2248007

Christmann A. & Van Messem A. (2008). Bouligand derivatives and robustness of support vector machines for regression. *J. Mach. Learn. Res.*, **9**, 915—936. MR2417258

Christmann A. & Steinwart I. (2007). Consistency and robustness of kernel-based regression in convex risk minimization. *Bernoulli*, **13**, 799–819. MR2348751

Croux C., Rousseeuw P. J. & Hössjer O. (1994). Generalized *S*-estimators. *J. Amer. Statist. Assoc.*, **89**, 1271–1281. MR1310221

De Brabanter K., Pelckmans K., De Brabanter J., Debruyne M., Suykens J.A.K., Hubert M. & De Moor B. (2009). Robustness of kernel based regression: a comparison of iterative weighting schemes. *Proc. of the 19th International Conference on Artificial Neural Networks (ICANN)*, pp. 100–110.

De Brabanter K., Suykens J.A.K. & De Moor B. (2013). Nonparametric Regression via StatLSSVM. *Journal of Statistical Software*, vol. 55, no. 2, p. 1–23. https://www.jstatsoft.org/article/view/v055i02

Debruyne M., Hubert M. & Suykens J.A.K. (2008). Model selection in kernel based regression using the influence function. *J. Mach. Learn. Res.*, **9**, 2377–2400. MR2452631

Debruyne M., Christmann A., Hubert M. & Suykens J.A.K. (2010). Robustness of reweighted least squares kernel based regression. *J. Multivariate Anal.*, **101**, 447–463. MR2564353

Devito E., Rosasco L., Caponnetto A., Piana M. & Verri A. (2004). Some properties of regularized kernel methods. *J. Mach. Learn. Res.*, **5**, 1363—1390. MR2248020

Dollinger M.B. & Staudte R.G. (1991). Influence functions of iteratively reweighted least squares estimators. *J. Amer. Statist. Assoc.*, **86**, 709–716. MR1147096

F.Y. Edgeworth (1887). On observations relating to several quantities. *Hermathena*, **6**, 279-–285.

T. Evgeniou, M. Pontil & T. Poggio (2000). Regularization networks and support vector machines. *Adv. Comput. Math.*, **13**, 1-–50. MR1759187

Fox J. (2016). *Applied Regression and Generalized Linear Models, 3rd Ed.*. Sage, Los Angeles.

Friedman J.H. (1991). Multivariate adaptive regression splines. *Ann. Statist.*, **19**, 1–67. MR1091842

Girosi F. (1998). An equivalence between sparse approximation and support vector machines. *Neural Comput.*, **10**, 1455–1480.

Hampel F.R. (1974). The influence curve and its role in robust estimation. *J. Amer. Statist. Assoc.*, **69**, 383–393. MR0362657

Hampel F.R., Ronchetti E.M., Rousseeuw P.J. & Stahel W.A. (1986). *Robust Statistics: The Approach Based On Influence Functions*. Wiley, New York.

Härdle W., Hall P. & Marron J.S. (1988). How far are automatically chosen regression smoothing parameters from their optimum? *J. Amer. Statist. Assoc.*, **83**, 86–95. MR0941001

Härdle W., Hall P. & Marron J.S. (1992). Regression smoothing parameters that are not far from their optimum. *J. Amer. Statist. Assoc.*, **87**, 227–233.

Hettmansperger T.P. & McKean J.W. *Robust Nonparametric Statistical Methods*. Chapman & Hall/CRC. MR2779026

Huber P.J. (1964). Robust estimation of a location parameter. *Ann. Math. Statist.*, **35**, 73–101. MR0161415

Huber P.J. (1965). A robust version of the probability ratio test. *Ann. Math. Statist.*, **36**, 1753–1758. MR0185747

Huber P.J. (1968). Robust confidence limits. *Probab. Theory Related Fields*, **10**, 269–278. MR0242330

P.J. Huber (1981). *Robust Statistics*, Wiley, New York.

Huber P.J. & Strassen V. (1973). Minimax tests and the Neyman-Pearson lemma for capacities. *Ann. Statist.*, **1**, 251–263. MR0356306

Huber P.J. & Strassen V. (1974). Minimax tests and the Neyman-Pearson lemma for capacities (Correction of Proof 4.1). *Ann. Statist.*, **2**, 223–224. MR0362587

Hubert M. (2001). Multivariate outlier detection and robust covariance matrix estimation - discussion. *Technometrics*, **43**, 303–306.

Hubert M., Rousseeuw P.J. & Vanden Branden K. (2005). ROBPCA: A new approach to robust principal components analysis. *Technometrics*, **47**, 64–79. MR2135793

Jurečková J. & Picek J. (2006). *Robust Statistical Methods with R*. Chapman & Hall (Taylor & Francis Group). MR2191689

Kutner M. H., Nachtsheim C. J., Neter J. & Li W. (2005). *Applied Linear Statistical Models, 5th Ed.*. McGraw-Hill.

Leung D., Marriott F. & Wu E. (1993). Bandwidth selection in robust smoothing. *J. Nonparametr. Stat.*, **2**, 333–339. MR1256384

Leung D.H-Y. (2005). Cross-validation in nonparametric regression with outliers. *Ann. Statist.*, **33**, 2291–2310. MR2211087

Lukas M.A. (2008). Strong robust generalized cross-validation for choosing the regularization parameter. *Inverse Problems*, **24**(3): 034006. MR2421943

Maronna R., Martin D. & Yohai V. (2006). *Robust Statistics: Theory and Methods*. Wiley.

S. Mukherjee, P. Niyogi, T. Poggio & R. Rifkin (2006). Learning theory: stability is sufficient for generalization and necessary and sufficient for consistency of empirical risk minimization. *Adv. Comput. Math.*, **25**, 161–193. MR2231700

Mukhoty B., Gopakumar G., Jain P. & Kar P. (2019). *Globally-convergent iteratively reweighted least squares for robust regression problems*. Proceedings of Machine Learning Research 89 (AISTATS 2019), p. 313-322, 2019.

Osborne M. R. (1985). *Finite Algorithms in Optimization and Data Analysis*. Wiley & Sons.

T. Poggio & S. Smale (2003). The mathematics of learning: Dealing with data. *Notices of the American Mathematical Society*, **50**, 537-–544. MR1968413

Riazoshams H., Midi H. & Ghilagaber G. (2018). *Robust Nonlinear Regression with Applications using R*. Wiley & Sons.

Rousseeuw P.J. & Leroy A.M. (2003). *Robust Regression and Outlier Detection*. Wiley & Sons. MR0914792

Samorodnitsky G. (1994). *Stable Non-Gaussian Random Processes: Stochastic Models with Infinite Variance*. Chapman & Hall/CRC . MR1280932

Schölkopf B., Herbrich R. & Smola A. (2001). *A generalized representer theorem*. in: D. Helmblod, B. Williamson (Eds.), Neural Networks and Computational Learning Theory, Springer, Berlin, pp. 416-–426.

Schölkopf B. & Smola A. (2002). *Learning with Kernels: Support Vector Machines, Regularization, Optimization, and Beyond*. MIT Press.

Sigl J. (2016). Nonlinear residual minimization by iteratively reweighted least squares. *Comput. Optim. Appl.*, **64**, 755-–792. MR3506232

Simpson D.G., Ruppert D. & Carroll R.J. (1992). On one-step GM-estimates and stability of inferences in linear regression. *J. Amer. Statist. Assoc.*, **87**, 439-–450. MR1173809

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

Steinwart I. (2003). Sparseness of support vector machines. *J. Mach. Learn. Res.*, **4**, 1071–1105. MR2125346

Steinwart I. & Christmann A. *Support Vector Machines*. Springer, 2008.

Tikhonov A.N. & Arsenin V.Y. (1997). *Solutions of Ill Posed Problems*. W.H. Winston, Washington D.C.

Tukey J.W. (1960). *Contributions to Probability and Statistics*, chapter A survey of sampling from contaminated distributions, (Ed.) I. Olkin, pages 448–485. Stanford University Press.

Vapnik V.N. (1995). *The Nature of Statistical Learning Theory*. Springer.

Vapnik V.N. (1999). *Statistical Learning Theory*. John Wiley & Sons.

Wahba G. (1990). *Spline Models for Observational Data*. SIAM, Philidelphia, PA. MR1045442

Wahba G. (1999). *Support vector machines, reproducing kernel Hilbert spaces and the randomized GACV*, in: B. Schölkopf, C. Burges, A. Smola (Eds.), Advances in Kernel Methods - Support Vector Learning, Cambridge, MA, pp. 69-–88.

Wasserman L .A. (2005). *All of Nonparametric Statistics*. Springer. MR2172729

Wilcox R.R. (1996). *Br. J. Math. Stat. Psychol.*, **49**, 253–274 MR1438832

Wilcox R.R. (2012). *Introduction to Robust Estimation and Hypothesis Testing*. Academic Press. MR3642283

Yang Y. (2007). Consistency of cross validation for comparing regression procedures. *Ann. Statist.*, **35**, 2450–2473. MR2382654

Yu C. & Yai W. (2017). Robust linear regression: A review and comparison. *Comm. Statist. Simulation Comput.*, **46**, 6261–6282. MR2382654

Zolotarev V.M. (1986). *One-dimensional stable distributions*, Translated from the Russian by H.H. McFaden. Translation edited by Ben Silver. Translations of Mathematical Monographs, 65. American Mathematical Society, Providence, RI. MR0854867