

---

# On Robustness in Kernel Based Regression

---

## **Kris De Brabanter**

Dep. of Electrical Engineering ESAT-SCD  
Katholieke Universiteit Leuven  
Kasteelpark Arenberg 10, B-3001 Leuven  
kris.debrabanter@esat.kuleuven.be

## **Jos De Brabanter**

Departement I.I. - E&A  
KaHo Sint-Lieven (Associatie K.U.Leuven)  
G. Desmetstraat 1, B-9000 Gent  
jos.debrabanter@kahosl.be

## **Johan A.K. Suykens**

Dep. of Electrical Engineering ESAT-SCD  
Katholieke Universiteit Leuven  
Kasteelpark Arenberg 10, B-3001 Leuven  
johan.suykens@esat.kuleuven.be

## **Peter Karsmakers**

Departement IBW  
K.H.Kempen (Associatie K.U.Leuven)  
Kleinhoefstraat 4, B-2440 Geel  
peter.karsmakers@khk.be

## **Kristiaan Pelckmans**

Department of Information Technology  
Uppsala University  
Box 337 SE-751 05 Uppsala, Sweden  
kristiaan.pelckmans@it.uu.se

## **Bart De Moor**

Dep. of Electrical Engineering ESAT-SCD  
Katholieke Universiteit Leuven  
Kasteelpark Arenberg 10, B-3001 Leuven  
bart.demoor@esat.kuleuven.be

## **Abstract**

It is well-known 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 and optimization problems. In classical statistics, robustness is improved by reweighting the original estimate. We study reweighting the KBR estimate using four different weight functions. In addition, we show that both the smoother as well as the cross-validation procedure have to be robust in order to obtain a fully robust procedure.

## **1 Introduction**

An important statistical tool routinely applied in most sciences is regression analysis. Since Edgeworth first argued that outliers have a very large influence on Least Squares (LS) many robust techniques have been developed [7, 11]. These involve  $L_1$  regression,  $M$ -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 [8, 12, 14]. Also other type of methods called adaptive regression techniques [6] have also been used to obtain robustness. In these techniques an adaptive combination of estimators is made in order to obtain robustness. However, all the techniques above were originally proposed for parametric regression.

The evaluation of a statistical estimator is to determine how close it is to the true parameter. In case of nonparametric regression popular criteria are integrated squared error, mean integrated squared error, mean integrated absolute deviation, . . . Any of these criteria can be used in practice as they are asymptotically quite similar [10]. In the nonparametric regression setting the choice of bandwidth (and regularization parameter) is crucial. In what follows we will denote bandwidth and/or regularization parameter as tuning parameter(s). These tuning parameters are chosen to minimize the sum of squares of the prediction errors from all observations. Cross-validation (CV) [2] is probably one of the most popular data-driven methods of tuning parameter(s) selection methods. We show, in order to obtain a fully robust procedure, that both the smoother as well as the CV procedure have to be robust.

The rest of the paper is organized as follows. Section 2 explains the practical difficulties associated with estimating the underlying function in the presence of outliers. In Section 3 we review the basic principles of iteratively reweighted least squares support vector machines and discuss and derive the properties of the Myriad. Section 4 states the conclusions.

## 2 Problems with Outliers in Kernel Based Regression

Some quite fundamental problems occur when regression techniques are attempted in the presence of outliers. In [15] a comprehensive study about this topic is given for parametric techniques. In case of nonparametric regression e.g. Nadaraya-Watson kernel estimator, local polynomial regression, least squares support vector machines (LS-SVM) the  $L_2$  risk is commonly used. However, the  $L_2$  norm is extremely sensitive to outliers. The breakdown of kernel nonparametric regression based on the  $L_2$  norm, as well as a possible solution to it, is illustrated by means of a simple toy example in Figure 1. In all examples LS-SVM (see Section 3) is used as smoother. Consider 200 equally spaced observations on the interval  $[0, 1]$  and a low-order polynomial mean function  $f(x) = 300(x^3 - 3x^4 + 3x^5 - x^6)$ . 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  $\epsilon$ -contamination model  $\mathcal{U}(F_0, G, \epsilon)$  [11]. This model is defined as follows

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

where  $F_0$  is some given distribution (the ideal nominal model),  $G$  is an arbitrary continuous distribution and  $\epsilon$  is the first parameter of contamination. 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 (bold line) are less stable or even breakdown in contrast to estimates based on robust loss functions (thin line). Another important issue to obtain robustness in nonparametric regression

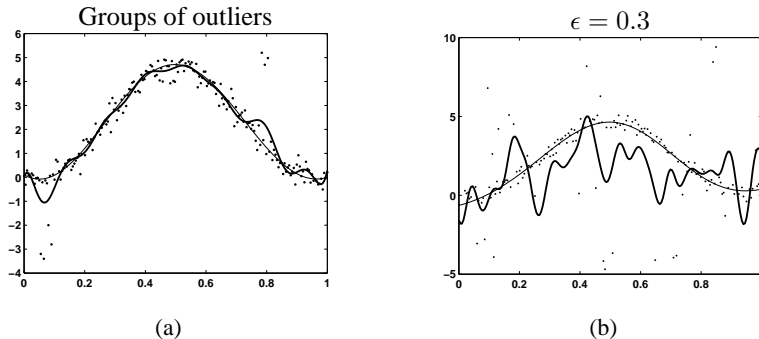


Figure 1: LS-SVM estimates with (a) normal distributed errors and two groups of outliers; (b) the  $\epsilon$ -contamination model. This clearly shows that the estimates based on the  $L_2$  norm (bold line) are less stable or even breakdown in contrast to estimates based on robust loss functions (thin line).

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

The last issue to acquire a 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 [9]. In [17] it is shown, under some regularity conditions, that 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 [13] 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 as before. 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.

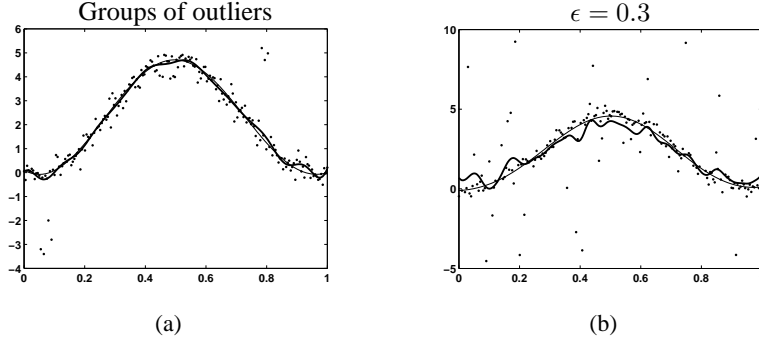


Figure 2: LS-SVM estimates and type of errors as in Figure 1. The bold line represents the estimate based on classical  $L_2$  CV and a robust smoother. The thin line represents estimates based on a fully robust procedure.

### 3 Robust Approaches to Kernel Based Regression

In this section we discuss possible strategies to robustify classical smoothers. In particular we focus on Least Squares Support Vector Machines (LS-SVM), but the described procedures can be generally applied to other smoothers (Nadaraya-Watson, Priestley-Chao, local polynomial regression,...).

#### 3.1 Robustness via Iteratively Reweighting

In order to obtain a robust estimate, one can replace the  $L_2$  loss function in the LS-SVM formulation by e.g.  $L_1$  or Huber's loss function. This would lead to a Quadratic Programming (QP) problem and hence increasing the computational load. Instead of using robust cost functions, one can obtain a robust estimate based upon the previous LS-SVM solution. Given a training set defined as  $\mathcal{D}_n = \{(X_k, Y_k) : X_k \in \mathbb{R}^d, Y_k \in \mathbb{R}; k = 1, \dots, n\}$  of size  $n$  drawn i.i.d. from an unknown distribution  $F_{XY}$  according to  $Y = m(X) + e$ , where  $e \in \mathbb{R}$  are assumed to be i.i.d. random errors with  $E[e|X] = 0$ ,  $\text{Var}[e] = \sigma^2 < \infty$ ,  $m \in C^z(\mathbb{R})$  with  $z \geq 2$ , is an unknown real-valued smooth function and  $E[Y|X] = m(X)$ . The optimization problem of finding the vector  $w$  and  $b \in \mathbb{R}$  for regression can be formulated as follows [16]

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

where the error variables from the unweighted LS-SVM  $\hat{e}_k = \hat{\alpha}_k / \gamma$  (case  $v_k = 1, \forall k$ ) are weighted by weighting factors  $v_k$  and  $\varphi : \mathbb{R}^d \rightarrow \mathbb{R}^{n_h}$  is the feature map. By using Lagrange multipliers, the solution of (1) can be obtained by taking the Karush-Kuhn-Tucker (KKT) conditions for optimality. The result is given by the following linear system in the dual variables  $\alpha$

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

with  $D_\gamma = \text{diag} \left\{ \frac{1}{\gamma v_1}, \dots, \frac{1}{\gamma v_n} \right\}$ . The weights  $v_k$  are based upon  $\hat{e}_k = \hat{\alpha}_k / \gamma$  from the (unweighted) LS-SVM ( $D_\gamma = I_n / \gamma$ ),  $Y = (Y_1, \dots, Y_n)^T$ ,  $1_n = (1, \dots, 1)^T$ ,  $\alpha = (\alpha_1, \dots, \alpha_n)^T$  and  $\Omega_{kl} = \varphi(X_k)^T \varphi(X_l) = K(X_k, X_l)$  for  $k, l = 1, \dots, n$ , with  $K$  a positive definite kernel e.g. the Gaussian density with bandwidth  $h$ . The resulting weighted LS-SVM model for function estimation becomes

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

Instead of weighting only once [16], one can use a weighting scheme from Table 1 and iteratively solve (2) a number of times [4]. This idea is summarized in Algorithm 1.

#### 3.2 Some Properties of the Myriad

It is without doubt that the choice of weight function  $V$  plays a significant role in the robustness aspects of the smoother. We consider four different weight function illustrated in Table 1. The first

---

**Algorithm 1** Iteratively Reweighted LS-SVM
 

---

- 1: Compute the residuals  $\hat{e}_k = \hat{\alpha}_k/\gamma$  from the unweighted LS-SVM ( $v_k = 1, \forall k$ )
  - 2: **repeat**
  - 3:   Compute  $\hat{s} = 1.483 \text{MAD}(e_k^{(i)})$  from the  $e_k^{(i)}$  distribution
  - 4:   Determine weights  $v_k^{(i)}$  based upon  $r^{(i)} = e_k^{(i)}/\hat{s}$ ; choose weight function  $V$  (see Table 1)
  - 5:   Solve (2) with  $D_\gamma = \text{diag} \left\{ 1/(\gamma v_1^{(i)}), \dots, 1/(\gamma v_n^{(i)}) \right\}$ ,
  - 6:   Set  $i := i + 1$
  - 7: **until** consecutive estimates  $\alpha_k^{(i-1)}$  and  $\alpha_k^{(i)}$  are sufficiently close to each other
- 

three are well-known in the field of robust statistics, the last one however is less or not known. We will study some of the properties of the last weight function i.e. the Myriad [1]. The Myriad is derived from the Maximum Likelihood (ML) estimation of a Cauchy distribution with scaling factor  $\delta$  (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  $X \sim C(\beta, \delta)$ , where the location parameter  $\beta$  is to be estimated from data i.e.  $\hat{\beta}$  and  $\delta > 0$  is a scaling factor. The ML principle yields the sample Myriad

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

which is equivalent to

$$\hat{\beta} = \arg \min_{\beta} \sum_{i=1}^n \log [\delta^2 + (X_i - \beta)^2]. \quad (3)$$

Note that, unlike the sample mean or median, the definition of the sample Myriad involves the free parameter  $\delta$ . We will refer to  $\delta$  as the linearity parameter of the Myriad. The behavior of the Myriad estimator is markedly dependent on the value of its linearity parameter  $\delta$ . Tuning the linearity parameter  $\delta$  adapts the behavior of the myriad from impulse-resistant mode-type estimators (small  $\delta$ ) to the Gaussian-efficient sample mean (large  $\delta$ ). If an observation in the set of input samples has a large magnitude such that  $|X_i - \beta| \gg \delta$ , the cost associated with this sample is approximately  $\log(X_i - \beta)^2$  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 in (3) for various values of  $\delta$ . Figure 3a depicts the different cost function characteristics obtained for  $\delta = 20, 2, 0.75$  for a sample set of size 5. For the a set

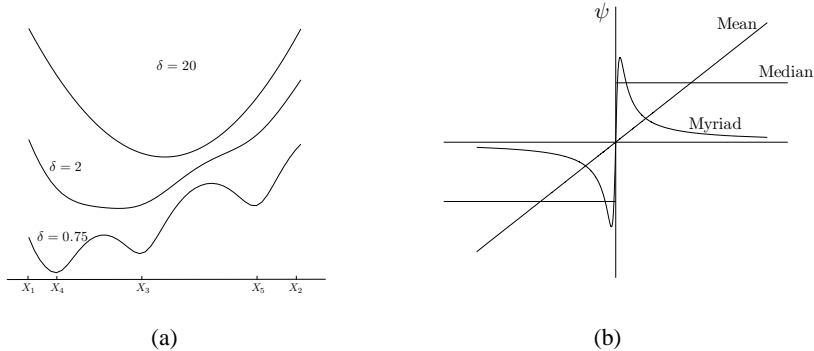


Figure 3: (a) Myriad cost functions for the observation samples  $X_1 = -3, X_2 = 8, X_3 = 1, X_4 = -2, X_5 = 5$  for  $\delta = 20, 2, 0.2$ ; (b) Influence function for the mean, median and Myriad.

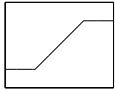
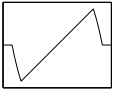
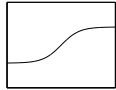
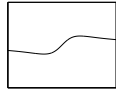
of samples defined as above, an M-estimator of location is defined as the parameter  $\beta$  minimizing a sum of the form  $\sum_{i=1}^n \rho(X_i - \beta)$ , where  $\rho$  is the cost or loss function. In general, when  $\rho(x) = -\log f(x)$ , with  $f$  a density, the M-estimate  $\hat{\beta}$  corresponds to the ML estimator associated with  $f$ . According to (3), the cost function associated with the sample Myriad is given by

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

Some insight into the operation of M-estimates is gained through the definition of the influence function (IF) [8]. For an M-estimate the IF is proportional to the score function. For the Myriad (see also Figure 3b), the IF is given by

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

Table 1: Definitions for the Huber, Hampel, Logistic and Myriad weight functions  $V(\cdot)$ . The corresponding loss  $\rho(\cdot)$  and score function  $\psi(\cdot)$  are also given.

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

### 3.3 Robust Selection of Tuning Parameters

It is shown in Figure 2 that also the CV procedure plays an significant role in the robustness properties of used method. Leung (2005) [13] theoretically shows that a robust CV procedure differs from the Mean Asymptotic Squared Error (MASE) by a constant shift and a constant multiple. Neither of these is 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 KBR one needs (i) a robust smoother and (ii) a robust CV (RCV) procedure based on the robust smoother or more formal

$$RCV(\theta) = \frac{1}{n} \sum_{i=1}^n L(Y_i - \hat{m}_{-i}(X_i; \theta)),$$

where  $L(\cdot)$  is a robust loss function e.g.  $L_1$ , Huber loss, Myriad loss,  $\hat{m}$  is a robust smoother and  $\hat{m}_{-i}(X_i; h, \gamma)$  denotes the leave-one-out estimator where point  $i$  is left out from the training and  $\theta$  denotes the parameter vector e.g. when using the Myriad weights  $\theta = (h, \gamma, \delta)$ .

### 3.4 Speed of Convergence-Robustness Trade-off

In a functional analysis setting it has been shown in [3] and [5] that the influence function [7] of reweighted Least Squares Kernel Based Regression (LS-KBR) with a bounded kernel converges to bounded influence function, even when the initial LS-KBR is not robust, if (i)  $\psi : \mathbb{R} \rightarrow \mathbb{R}$  is a measurable, real, odd function, (ii)  $\psi$  is continuous and differentiable, (iii)  $\psi$  is bounded and (iv)  $E_{P_e} \psi'(e) > 0$  where  $P_e$  denotes the distribution of the errors. This condition can be relaxed into  $\psi$  is increasing. Define

$$d = E_{P_e} \frac{\psi(e)}{e} \quad \text{and} \quad c = d - E_{P_e} \psi'(e),$$

then it can be shown [5] that  $c/d$  establishes an upper bound on the reduction of the influence function at each step. The upper bound represents a trade-off between the reduction of the influence function (speed of convergence) and the degree of robustness. The higher the ratio  $c/d$  the higher the degree of robustness but the slower the reduction of the influence function at each step and vice versa. In Table 2 this upper bound is calculated 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  $\beta$  respectively  $\delta$  increases. This behavior is to be expected, since the larger  $\beta$  respectively  $\delta$ , 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. A good choice between convergence and robustness is therefore Logistic weights. Also notice the small ratio for the Hampel weights indicating a low degree of robustness.

Table 2: 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	$N(0, 1)$			$C(0, 1)$		
		$c$	$d$	$c/d$	$c$	$d$	$c/d$
Huber	$\beta = 0.5$	0.32	0.71	<b>0.46</b>	0.26	0.55	<b>0.47</b>
	$\beta = 1$	0.22	0.91	<b>0.25</b>	0.22	0.72	<b>0.31</b>
Logistic		0.22	0.82	<b>0.26</b>	0.21	0.66	<b>0.32</b>
Hampel	$b_1 = 2.5$ $b_2 = 3$	0.006	0.99	<b>0.006</b>	0.02	0.78	<b>0.025</b>
Myriad	$\delta = 0.1$	0.11	0.12	<b>0.92</b>	0.083	0.091	<b>0.91</b>
	$\delta = 1$	0.31	0.66	<b>0.47</b>	0.25	0.50	<b>0.50</b>

## 4 Conclusions

In this paper we have compared four different type of weight functions and their use in iterative reweighted LS-SVM. By using an upper bound for the reduction of the influence function we have demonstrated the existence of a trade-off between speed of convergence and the degree of robustness. The Myriad weight function is highly robust against (extreme) outliers but has a slow speed of convergence. A good compromise between speed of convergence and robustness can be achieved by using Logistic weights. To obtain a fully robust solution, we showed that the smoother needs to be robust as well as the CV procedure.

## Acknowledgments

Research supported by Research Council KUL: GOA AMBioRICS, GOA MaNet, CoE EF/05/006 Optimization in Engineering(OPTEC), IOF-SCORES4CHEM, several PhD/post-doc & fellow grants; Flemish Government: FWO: PhD/postdoc grants, projects G.0452.04 (new quantum algorithms), G.0499.04 (Statistics), G.0211.05 (Nonlinear), G.0226.06 (cooperative systems and optimization), G.0321.06 (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 (ICCoS, ANMMM, MLDM); G.0377.09 (Mechatronics MPC), IWT: PhD Grants, McKnow-E, Eureka-Flite+, SBO LeCoPro, SBO Climaqs, POM, Belgian Federal Science Policy Office: IUAP P6/04 (DYSCO, Dynamical systems, control and optimization, 2007-2011); EU: ERNSI; FP7-HD-MPC (INFSO-ICT-223854), COST intelliCIS, EMBOCOM, Contract Research: AMINAL, Other: Helmholtz, viCERP, ACCM, Bauknecht, Hoerbiger. BDM is a full professor at the Katholieke Universiteit Leuven, Belgium. JS is a professor at the Katholieke Universiteit Leuven, Belgium.

## References

- [1] Arce, G. R. (2005) *Nonlinear Signal Processing: A Statistical Approach* Wiley
- [2] Burman, P. (1989) A comparative study of ordinary cross-validation, v-fold cross-validation and the repeated learning-testing methods. *Biometrika* **76**(3):503–514
- [3] Christmann, A., Steinwart, I. (2004) Consistency and Robustness of Kernel Based Regression in Convex Risk Minimization. *Bernoulli* **13**(3):799–819
- [4] 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. in *Proc. of the 19th International Conference on Artificial Neural Networks (ICANN)* pp. 100-110.
- [5] Debruyne, M., Christmann, A., Hubert, M., Suykens, J.A.K. (2010) Robustness of reweighted Least Squares Kernel Based Regression. *Journal of Multivariate Analysis* **101**(2):447–643
- [6] Dodge, Y., Jurečková, J. (2000) *Adaptive Regression*. Springer
- [7] Hampel, F. R. (1971) A General Definition of Qualitative Robustness. *Ann. Math. Stat* **42**:1887–1896
- [8] Hampel, F.R., Ronchetti, E.M., Rousseeuw, P.J., Stahel, W.A. (1986) *Robust Statistics: The Approach Based on Influence Functions*. Wiley
- [9] Härdle, W., Hall, P. and Marron, J. S. (1988) How far are automatically chosen regression smoothing parameters from their optimum? (with discussion). *J. Amer. Statist. Assoc.* **83**:86101
- [10] Härdle, W. (1990) *Applied Nonparametric Regression*. Cambridge University Press
- [11] Huber, P. J. (1964) Robust Estimation of a Location Parameter. *Ann. Math. Stat* **35**:73–101
- [12] Huber, P. J., Ronchetti, E. M. (2009) *Robust Statistics* (2nd ed.) Wiley
- [13] Leung, D. H-Y. (2005) Cross-Validation in Nonparametric Regression with Outliers. *Ann. Statist* **33**(5):2291–2310
- [14] Rousseeuw, P. J., Leroy, A. M. (2003) *Robust Regression and Outlier Detection*. Wiley
- [15] Rousseeuw, P. J., Debruyne, M., Engelen, S., Hubert, M. (2006) Robustness and outlier detection in chemometrics. *Critical Reviews in Analytical Chemistry* **36**:221-242
- [16] Suykens J.A.K., De Brabanter J., Lukas L., Vandewalle J. (2002) Weighted Least Squares Support Vector Machines : Robustness and Sparse Approximation *Neurocomputing* **48**(1-4): 85–105.
- [17] Yang, Y. (2007) Consistency of Cross Validation for Comparing Regression Procedures. *Ann. Statist.* **35**(6):2450–2473