# An Overview of Recent Developments in Kernel Based Regression

**Kris De Brabanter**[1], **Jos De Brabanter**[1,2] **and Bart De Moor**[1]

[1] Department of Electrical Engineering (ESAT), Research Division SCD, Katholieke Universiteit Leuven,
Kasteelpark Arenberg 10, B-3001 Leuven, Belgium

[2] Departement Industrieel Ingenieur, KaHo Sint Lieven (Associatie K.U.Leuven), Geb. Desmetstraat 1,
B-9000 Gent, Belgium

### Summary

**We give an overview of recent developments in the area of robust regression, regression in the presence of correlated errors, construction of additive models and density estimation by means of kernel based methods. First, we discuss the difficulties arising in kernel based regression when outliers are present in the data. Iterative reweighting in combination with robust model selection is shown to produce robust kernel based methods. Second, we describe a bandwidth selection procedure based on bimodal kernels which successfully removes the correlation without requiring any prior knowledge about its structure. We also demonstrate the existence of a relation between the bandwidth under correlation and bandwidth under independence. Third, we develop a method to construct additive models with least squares support vector machines without using backfitting. Finally, we establish a connection between density estimation and nonparametric regression via a binning technique.**

*Key words*: nonparametric regression; robustness; iterative reweighting; correlated errors; bandwidth choice; short-range dependence; density estimation; additive models.

## 1 Introduction

Kernel based methods have been developed in the area of statistics (Nadaraya, 1964; Watson, 1964) and became also popular in the fields of machine learning and data mining. These methods often make use of $L_2$ loss functions which result in relatively easy formulations and fast computation techniques. On the other hand they have a potential drawback when outliers are present in the data. In Section 2 we give an overview how one can robustify kernel based methods via iteratively reweighting (De Brabanter *et al.*, 2009; Debruyne *et al.*, 2010) and derive properties of some weight functions. Further, we will demonstrate that in order to obtain a fully robust kernel based method, three requirements have to be fulfilled i.e. (i) robust smoother, (ii) bounded kernel and (iii) a robust model selection procedure.

Most modeling techniques (critically) rely on the i.i.d. (independent and identically distributed) assumptions of the data. In Section 3 we summarize the problems occurring when this assumption is violated. We will illustrate that, for kernel based (nonparametric) regression, classical model selection procedures such as cross-validation will break down in the presence of correlated data and not the chosen smoothing method. Since the latter stays consistent when correlation is present in the data (Kulkarni *et al.*, 2002), it is not necessary to modify or add extra constraints to the smoother. We will theoretically show that by taking a kernel satisfying $K(0) = 0$ the correlation structure is successfully removed without requiring any prior knowledge about its structure (De Brabanter *et al.*, 2011b).

The rate of convergence of kernel (nonparametric) methods have been shown to decrease with increasing dimensionality (Györfi *et al.*, 2002). Several methods have been proposed in order to overcome this curse of dimensionality such as projection pursuit for regression (Friedman & Stuetzle, 1981), additive models (Hastie & Tibshirani, 1990; Opsomer & Ruppert, 1997) and single index models (Amemiya, 1985; Manski, 1988) which are a special case of projection pursuit. In Section 4 we develop a method to construct additive models with least squares support vector machines (Suykens *et al.*, 2002) without using backfitting.

Further, in Section 5 we establish the connection between density estimation on an interval and regression via a binning technique. Via this technique the density estimation problem is converted in to a nonparametric heteroscedastic regression problem. In order the deal with the heteroscedasticity Anscombe's variance stabilizing transformation (Anscombe, 1948) is used to the bin count. Finally, Section 6 states the conclusions of this first part of the paper.

We have used the Matlab toolbox **StatLSSVM** (De Brabanter *et al.*, 2011c) which is freely available for research purposes at `http://www.esat.kuleuven.be/sista/statlssvm/` for every regression analysis.

## 2 Aspects of robustness in kernel smoothers

### 2.1 Robustness: an overview

Regression analysis is an important statistical tool routinely applied in most sciences. However, using least squares techniques, there is an awareness of the dangers posed by the occurrence of outliers present in the data. Not only the response variable can be outlying, but also the explanatory part, leading to leverage points. Both types of outliers may totally spoil an ordinary least squares (LS) analysis. To cope with this problem, statistical techniques have been developed that are not so easily affected by outliers. These methods are called robust or resistant. A *first attempt* was done by Edgeworth. 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. He considered the general gross-error model or $\epsilon$-contamination model

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

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. This contamination model describes the case, where with large probability $(1-\epsilon)$, the data occurs with distribution $F_0$ and with small probability $\epsilon$ outliers occur according to distribution $G$.

**Example 2.1.** *$\epsilon$-contamination model with symmetric contamination*

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

**Example 2.2.** *$\epsilon$-contamination model for the mixture of the Normal and Laplace or double exponential distribution*

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

Huber considered also the class of $M$-estimators of location (also called generalized maximum likelihood estimators) described by some suitable function. The Huber estimator is a minimax solution: it minimizes the maximum asymptotic variance over all $F$ in the gross-error model.

Huber (1965, 1968) and Huber & Strassen (1973, 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. Three main concepts are introduced: (i) qualitative robustness, which is essentially continuity of the estimator viewed as functional in the weak topology; (ii) the Influence Curve (IC) or Influence Function (IF), which describes the first derivative of the estimator, as far as existing; and (iii) the Breakdown Point (BP), a global robustness measure describing how many percent gross errors are still tolerated before the estimator totally breaks down.

Robustness has provided at least two major insights into statistical theory and practice: (i) Relatively small perturbations from nominal models can have very substantial deleterious effects on many commonly used statistical procedures and methods (e.g. estimating the mean, F-test for variances). (ii) Robust methods are needed for detecting or accommodating outliers in the data (Hubert, 2001).

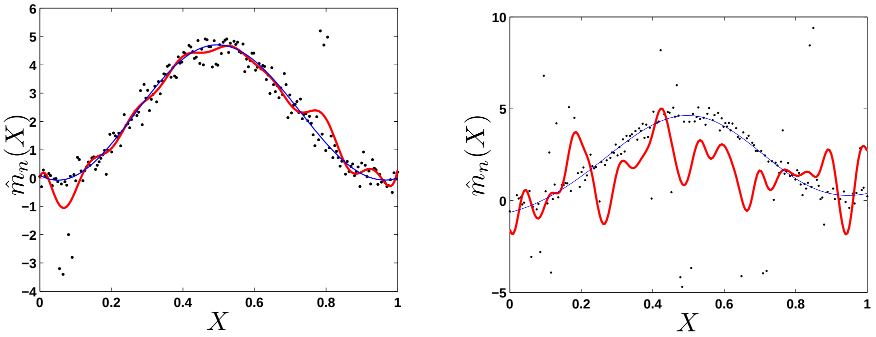
From their work the following methods were developed: $M$-estimators, Generalized $M$-estimators, $R$-estimators, $L$-estimators, $S$-estimators, repeated median estimator, least median of squares, … Detailed information about these estimators as well as methods for robustness measuring can be found 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.

## 2.2 Robustifying least squares kernel based regression

It is known that taking a non-robust loss function, e.g. $L_2$, can totally spoil the LS estimate in the presence of only one outlying observation. In case of Kernel Based Regression (KBR) based on an $L_2$ loss, the estimate is only affected in an influence region if the number of outliers is small. Further, we will demonstrate that even if the initial estimate is non-robust, we can obtain a robust estimate via iteratively reweighting. However, there is another important issue influencing the KBR estimate when outliers are present in the data i.e. the model selection. Next, we will demonstrate how model selection criteria can influence the final result.

### 2.2.1 Problems with outliers in nonparametric regression

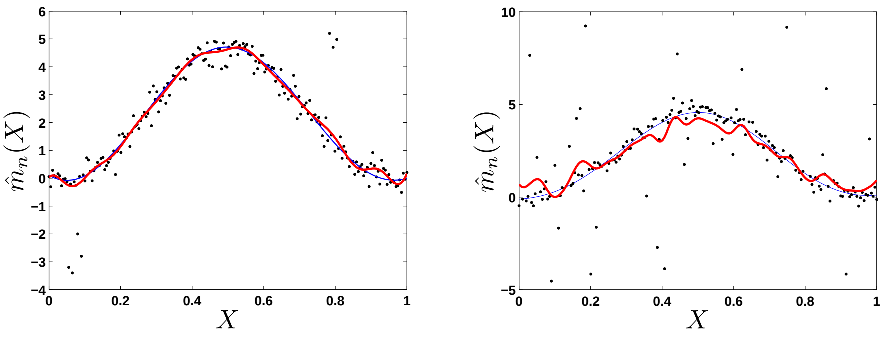
Consider 200 equally spaced observations on the interval $[0,1]$ and a low-order polynomial mean function $m(X) = 300(X^3 - 3X^4 + 3X^5 - X^6)$. Figure 1(a) shows the mean function with normally distributed errors with variance $\sigma^2 = 0.3^2$ and two distinct groups of outliers. Figure 1(b) shows the same mean function, but the errors are generated from the gross error or $\epsilon$-contamination model (1). In this simulation $F_0 \sim N(0, 0.3^2)$, $G \sim N(0, 10^2)$ and $\epsilon = 0.3$. This simple example clearly shows that the estimates based on the $L_2$ norm with classical CV (bold line) are influenced in a certain region or even breakdown (in case of the gross error model) in contrast to estimates based on robust KBR with robust CV (thin line). The fully robust KBR method will be discussed later in this section.



(a) (b)

**Figure 1:** *KBR 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 influenced in a certain region or even breakdown in contrast to estimates based on robust loss functions (thin line).*

Another important issue to obtain robustness in nonparametric regression is the kernel function $K$. Kernels that satisfy $K(u) \to 0$ as $u \to \infty$, for $X \to \infty$ and $X \to -\infty$, are bounded in $\mathbb{R}$. These type of kernels are called decreasing kernels. Using decreasing kernels leads to quite robust methods with respect to outliers in the $X$-direction (leverage points). Common choices of decreasing kernels are: Epanechnikov, Gaussian, Laplace etc. The last issue to acquire a fully robust estimate is the proper type of cross-validation (CV). When no outliers are present in the data, CV has been shown to produce tuning parameters that are asymptotically consistent (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 standard CV using a robust smoother. The reason is that CV no longer produces a reasonable estimate of the prediction error. Therefore, a fully robust CV method is necessary. Figure 2 demonstrates this behavior on the same toy example (see Figure 1). Indeed, it can be clearly seen that CV results in less optimal tuning parameters resulting in a bad estimate. Hence, to obtain a fully robust estimate, every step has to be robust i.e. robust CV with a robust smoother based on a decreasing kernel. An extreme example to show the absolute necessity of a



(a)

(b)

**Figure 2:** *KBR 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.*

robust model selection procedure is given next. Consider 200 observations on the interval $[0, 1]$ and a low-order polynomial mean function $m(X) = 1 - 6X + 36X^2 - 53X^3 + 22X^5$ and $X \sim \mathcal{U}[0, 1]$. The errors are generated from the gross error model 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 (SVM) (Vapnik, 1999), which is known to be robust, 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 is robust, also the model selection procedure has to be robust in order to obtain fully robust estimates.

We have demonstrated that fully robust estimates can only be acquired if (i) the smoother is robust, (ii) decreasing kernels are used and (iii) a robust model selection criterion is applied. In what follows, we will demonstrate how to make an LS kernel based smoother robust.

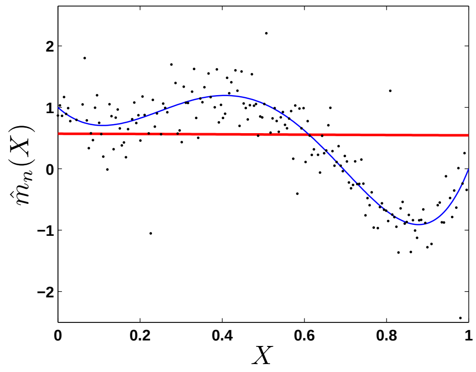


**Figure 3:** *SVM (bold straight line) cannot handle these extreme type of outliers and the estimate becomes useless. The folly robust KBR (thin line) can clearly handle these outliers and does not break down. For visual purposes, not all data is displayed in the figure (the full range of the $Y$-axis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

### 2.2.2 Robustness via iterative reweighting

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 $F$. Let the estimator $T(\hat{F}_n)$ of $T(F)$ be the functional of the sample distribution $F_n$. The influence function (IF) is defined as follows.

**Definition 2.1** (Influence Function (Hampel *et al.*, 1986))**.** *The influence function (IF) of $T$ at $F$ is given by*

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

*in those $x$ where this limit exists. $\Delta_x$ denotes the probability measure which puts mass 1 at the point $x$.*

Hence, the IF reflects the bias caused by adding a few outliers at the point $x$, standardized by the amount $\epsilon$ of contamination. Therefore, a bounded IF leads to robust estimators. In a functional framework, it has been shown that the IF is bounded by using a bounded kernel, e.g. the Gaussian kernel and a loss function with bounded first derivative e.g. $L_1$ loss or Vapnik's $\varepsilon$-insensitive loss (Christmann & Steinwart, 2007). The $L_2$ loss on the other hand leads to an unbounded IF and therefore 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 (QP) problems. In case of least squares support vector machines (LS-SVM) (Suykens *et al.*, 2002) and local linear regression this would mean that the $L_2$ loss should be replaced by e.g. an $L_1$ loss, what immediately would lead to a QP and linear programming problem respectively. In what follows we will study an alternative way of achieving robustness by means of reweighting. This has the advantage of easily computable estimates i.e. solving a weighted least squares problem in every iteration (De Brabanter *et al.*, 2009). First, we need the following definition concerning the weight function $V$.

**Definition 2.2.** *let $V : \mathbb{R} \to [0,1]$ be a weight function depending on the residual $r = Y - \hat{m}_n(X)$. Then the following assumptions will be made on $V$*

- *$V$ is a non-negative bounded Borel measurable function;*
- *$V$ is an even function of $r$;*
- *$V$ is continuous and differentiable with $\frac{dV(r)}{dr} \leq 0$ for $r > 0$.*

The next theorem illustrates that, assuming that the weight function $V$ satisfies Definition 2.2, the influence function 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 (Debruyne *et al.*, 2010).

**Theorem 2.1** (Conditions for convergence)**.** *Define $V(r) = \frac{\psi(r)}{r}$ with $\psi$ the contrast function. Then, reweighted LS-KBR with a bounded kernel converges to a bounded influence, even if the initial LS-KBR is not robust, if*

*(c1) $\psi : \mathbb{R} \to \mathbb{R}$ is a measurable, real, odd function;*

*(c2) $\psi$ is continuous and differentiable;*

*(c3)* $\psi$ *is bounded;*

*(c4)* $\psi$ *is an increasing function.*

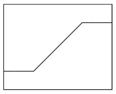
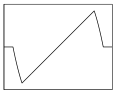
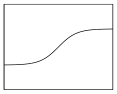
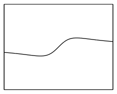
The condition $(c4)$ provides a distribution free condition on the contrast function. However, in KBR where a regularization parameter $\gamma$ is present (e.g. in LS-SVM), this condition can be modified into

$(c4+)$ $\mathbf{E}_{F_e}\, \psi'(e) > -\gamma$ where $F_e$ denotes the distribution of the errors.

### 2.2.3 Weight functions

It is without doubt that the choice of weight function $V$ plays a significant role in the robustness aspects of the smoother. We will show later that the choice of weight function also has an influence on the speed of convergence. First, consider the following four different weight functions illustrated in Table 1. The first three are well-known in the field of robust statistics, while the last one is less or not known. We introduce some of the properties of the last weight function i.e. the Myriad. The Myriad is derived from the Maximum Likelihood (ML) estimation of a Cauchy distribution with scaling factor $\delta$ and can be used as a robust location estimator in stable noise environments. 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 $\zeta$ is to be estimated from data i.e. $\hat{\zeta}$ and $\delta > 0$ is a scaling factor. The ML principle yields the sample Myriad

**Table 1:** *Definitions for the Huber ($\beta \in \mathbb{R}_0$), Hampel ($b_1, b_2 \in \mathbb{R}_0$), Logistic and Myriad ($\delta \in \mathbb{R}^+$) weight functions $V(\cdot)$. The corresponding loss $L(\cdot)$ and score function $\psi(\cdot)$ are also given.*

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

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

which is equivalent to

$$\hat{\zeta}_\delta = \underset{\zeta \in \mathbb{R}}{\arg\min} \sum_{i=1}^{n} \log\left[\delta^2 + (X_i - \zeta)^2\right]. \tag{3}$$

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 $|X_i - \zeta| \gg \delta$, the cost associated with this sample is approximately

$\log(X_i - \zeta)^2$ i.e. the log of squared deviation. Thus, much as the sample mean and sample median respectively minimize the sum of square and absolute deviations, the sample myriad (approximately) minimizes the sum of logarithmic squared deviations. Some intuition can be gained by plotting the cost function (3) for various values of $\delta$. 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 $\zeta$ 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 (3), the cost function associated with the sample Myriad is given by

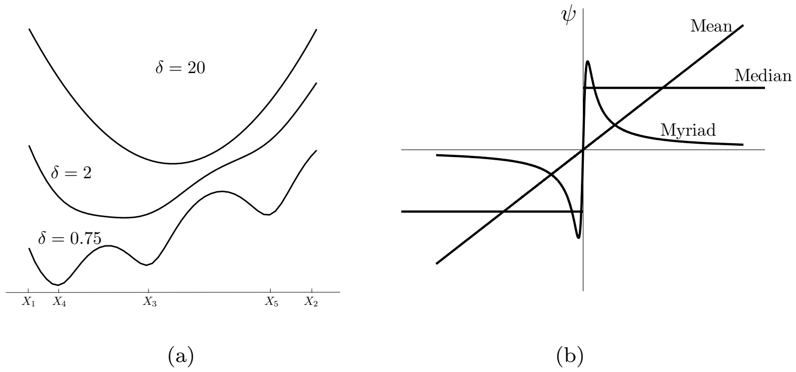


**Figure 4:** *(a) Myriad cost functions for the observation samples $X_1 = -3, X_2 = 8, X_3 = 1, X_4 = -2, X_5 = 5$ for $\delta = 20, 2, 0.2$; (b) Influence function for the mean, median and Myriad.*

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

When using the Myriad as a location estimator, it can be shown that the Myriad offers a rich class of operation modes that can be controlled by varying the parameter $\delta$. When the noise is Gaussian, large values of $\delta$ 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 $\delta$. Also, the Myriad has a linearity property i.e. when $\delta \to \infty$ then the sample Myriad reduces to the sample mean.

**Theorem 2.2** (Linearity Property (De Brabanter, 2011))**.** *Given a set of samples $X_1, \dots, X_n$. 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 $\delta$) to lower values of $\delta$, the estimator becomes more resistant to outliers. When $\delta$ tends to zero, the myriad approaches the mode of the sample.

**Theorem 2.3** (Mode Property (De Brabanter, 2011)). *Given a set of samples $X_1, \dots, X_n$. 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.*

### 2.2.4 Kernel based regression: iterative reweighting

Next, we will describe an algorithm that results into robust estimates of any KBR smoother based on an $L_2$ loss e.g. NW, local polynomial regression, LS-SVM,... As an example, we illustrate the algorithm for LS-SVM. A first attempt to robustify LS-SVM was introduced by (Suykens *et al.*, 2002b). Their approach is based on weighting the residuals once from the unweighted LS-SVM. However, based on theoretical results of (Debruyne *et al.*, 2010), we know that reweighting only once does not guarantee that the IF is bounded. Examples confirming the break down of reweighting only once are given in (De Brabanter *et al.*, 2009). In order to bound the IF, we have to repeat the weighting procedure a number of times. Given a data set $\mathcal{D}_n = \{(X_1, Y_1), \dots, (X_n, Y_n)\}$, the weighted LS-SVM is formulated as follows

$$\begin{aligned} \min_{w,b,e} \mathcal{J}(w, e) &= \tfrac{1}{2} w^T w + \tfrac{\gamma}{2} \sum_{k=1}^{n} v_k e_k^2 \\ \text{s.t.} \quad Y_k &= w^T \varphi(X_k) + b + e_k, \quad k = 1, \dots, n, \end{aligned} \tag{4}$$

where $\varphi : \mathbb{R}^d \to \mathbb{R}^{n_h}$ is the feature map to the high dimensional feature space, $n_h$ is the dimension of the feature space (can be infinite dimensional), $w \in \mathbb{R}^{n_h}$, $b \in \mathbb{R}$ and $v_k$ denotes the weight of the $k^{\text{th}}$ residual. By using Lagrange multipliers, the solution to (4) in the dual variables $\alpha$ is given by solving the linear system

$$\left( \begin{array}{c|c} 0 & 1_n^T \\ \hline 1_n & \Omega + D_\gamma \end{array} \right) \left( \begin{array}{c} b \\ \hline \alpha \end{array} \right) = \left( \begin{array}{c} 0 \\ \hline Y \end{array} \right), \tag{5}$$

with $D_\gamma = \operatorname{diag}\left\{ \frac{1}{\gamma v_1}, \dots, \frac{1}{\gamma v_n} \right\}$, $Y = (Y_1, \dots, Y_n)^T$, $1_n = (1, \dots, 1)^T$, $\alpha = (\alpha_1, \dots, \alpha_n)^T$ and $\Omega_{kl} = \varphi(X_k)^T \varphi(X_l) = K(X_k, X_l)$ for $k, l = 1, \dots, n$ and $K$ a positive definite bounded kernel. Based on the previous LS-SVM solutions, using an iteratively reweighting approach, a robust estimate can be obtained. In the $i^{\text{th}}$ iteration, one weighs the error variables $\hat{e}_k^{(i)} = \hat{\alpha}_k^{(i)} / \gamma$ for $k = 1, \dots, n$ with weighting factors $v^{(i)} = (v_1^{(i)}, \dots, v_n^{(i)})^T \in \mathbb{R}^n$, determined by a weight function $V$. The final estimate yields

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

Hence, one obtains an iterative algorithm (Algorithm 1) to obtain a robust estimate.

**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 $\hat{s} = 1.483\,\mathrm{MAD}(e_k^{(i)})$ from the $e_k^{(i)}$ distribution
4:     Choose a weight function $V$ (Table 1) and determine weights $v_k^{(i)}$ based on $r^{(i)} = e_k^{(i)}/\hat{s}$;
5:     Solve (5) with $D_\gamma = \mathrm{diag}\left\{1/(\gamma v_1^{(i)}), \ldots, 1/(\gamma v_n^{(i)})\right\}$,
6:     Set $i = i + 1$
7: until consecutive estimates $\alpha_k^{(i-1)}$ and $\alpha_k^{(i)}$ are close to each other: $\max_k(|\alpha_k^{(i-1)} - \alpha_k^{(i)}|) \leq 10^{-4}$
```

### 2.2.5 Speed of convergence-robustness tradeoff

Define

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

then $c/d$ establishes an upper bound on the reduction of the influence function at each step (Debruyne *et al.*, 2010). 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 $\beta$, $\delta \to 0$, indicating a high degree of robustness but slow convergence rate. Therefore, logistic weights offer a good tradeoff between speed of convergence and degree of robustness. Also notice the small ratio for the Hampel weights indicating a low degree of robustness. The highest degree of robustness is achieved by using Myriad weights.

**Table 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 | **0.46** | 0.26 | 0.55 | **0.47** |
| | $\beta = 1$ | 0.22 | 0.91 | **0.25** | 0.22 | 0.72 | **0.31** |
| Logistic | | 0.22 | 0.82 | **0.26** | 0.21 | 0.66 | **0.32** |
| Hampel | $b_1 = 2.5$<br>$b_2 = 3$ | 0.006 | 0.99 | **0.006** | 0.02 | 0.78 | **0.025** |
| Myriad | $\delta = 0.1$ | 0.11 | 0.12 | **0.92** | 0.083 | 0.091 | **0.91** |
| | $\delta = 1$ | 0.31 | 0.66 | **0.47** | 0.25 | 0.50 | **0.50** |

### 2.2.6 Robust selection of tuning parameters

It is shown in Figure 2 that also the model selection procedure plays a significant role in obtaining fully robust estimates. It is theoretically shown that a robust CV procedure differs from the Mean

Asymptotic Squared Error (MASE) by a constant shift and a constant multiple (Leung, 2005). Neither of these are dependent on the bandwidth. Further, it is shown that this multiple depends on the score function and therefore, also on the weight function. To obtain a fully robust procedure for LS-KBR one also needs, besides a robust smoother and bounded kernel, a robust model selection criterion. Consider for example the robust leave-one-out CV (RLOO-CV) given by

$$\text{RLOO-CV}(\theta) = \frac{1}{n} \sum_{i=1}^{n} L\left(Y_i, \hat{m}_{n,\text{rob}}^{(-i)}(X_i;\theta)\right), \tag{6}$$

where $L$ is a robust loss function e.g. $L_1$, Huber loss, Myriad loss, $\hat{m}_{n,\text{rob}}$ is a robust smoother and $\hat{m}_{n,\text{rob}}^{(-i)}(X_i;\theta)$ denotes the leave-one-out estimator where point $i$ is left out from the training and $\theta$ denotes the tuning parameter vector. A similar principle can be used in robust $v$-fold CV. For robust counterparts of GCV and complexity criteria see e.g. Lukas (2008) and references therein.

## 2.3 Examples

### 2.3.1 Empirical maxbias curve

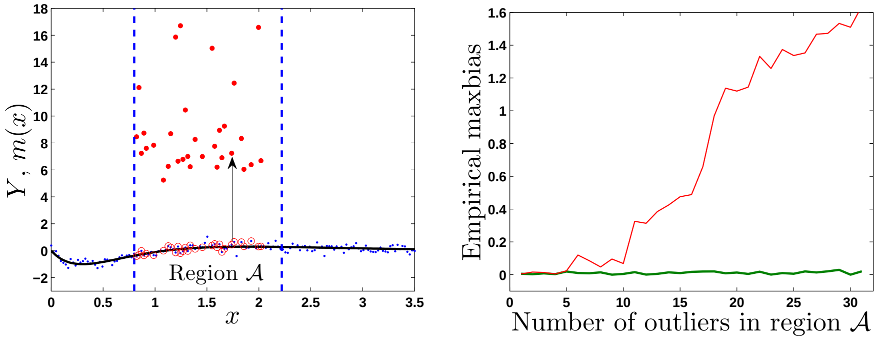
We compute the empirical maxbias curve for both LS-SVM and its robust counterpart iteratively reweighted (IRLS-SVM) on a test point. The maxbias curve gives the maximal bias that an estimator can suffer from when a fraction of the data come from a contaminated distribution. By letting the fraction vary between zero and the breakdown value a curve is obtained. Given 150 “good” equispaced observations according to the relation

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

where $e_k \sim \mathcal{N}(0, 0.1^2)$ and $m(x_k) = 4.26(\exp(-x_k) - 4\exp(-2x_k) + 3\exp(-3x_k))$ (Wahba, 1990). 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 non-robust LS-SVM estimate $\hat{m}_n(x)$ and the robust IRLS-SVM estimate $\hat{m}_{n,\text{rob}}(x)$ are drawn as a function of the number of outliers in region $\mathcal{A}$. The tuning parameters are tuned with $L_2$ LOO-CV for LS-SVM and RLOO-CV (6), based on an $L_1$ loss and Myriad weights, for IRLS-SVM. The maxbias curve of $\hat{m}_{n,\text{rob}}(x)$ increases very slightly with the number of outliers in region $\mathcal{A}$ and stays bounded right up to the breakdown point. This is in strong contrast with the non-robust LS-SVM estimate $\hat{m}_n(x)$ which has a breakdown point equal to zero.

### 2.3.2 Real life data sets

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 was measured. It is well known that the octane data set contains six outliers to which alcohol was added.



(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 non-robust LS-SVM estimator* $\hat{m}_n(x)$ *(thine line) and IRLS-SVM estimator* $\hat{m}_{n,rob}(x)$ *(bold line) in a test point* $x = 1.5$.

Table 3 shows the result (median and median absolute deviation for each method are reported) of a Monte Carlo simulation (200 runs) of the iteratively reweighted LS-SVM (IRLS-SVM)and SVM in different norms on a randomly chosen test set of size 10. Model selection was performed using robust LOO-CV (6).

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. The results of the simulations show that by using reweighting schemes the performance can be improved over weighted LS-SVM and SVM. To illustrate the trade-off between the degree of robustness and speed of convergence, the number of iterations $i_{\max}$ are also given in Table 3. The stopping criterion was taken identically to the one in Algorithm 1. The number of iterations, needed by each weight function, confirms the results in Table 2.

**Table 3:** *Results on the Octane and Demographic data sets. For 200 simulations the medians and median absolute deviations (between brackets) of three norms are given (on test data).* $i_{max}$ *denotes the number of iterations needed to satisfy the stopping criterion in Algorithm 1. The best results are bold faced.*

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

# 3 Kernel regression with correlated Errors

## 3.1 Some basic assumptions

Nonparametric regression is a very popular tool for data analysis because these techniques impose few assumptions about the shape of the mean function and the errors $e$ i.e. $\mathbf{E}[e] = 0, \mathbf{Var}[e] < \infty$ and the errors are assumed to be i.i.d. However, when the latter assumption is violated model selection procedures break down rendering the analysis useless (Altman, 1990; Opsomer *et al.*, 2001; De Brabanter *et al.*, 2011b). Given the bivariate data $(x_1, Y_1), \ldots, (x_n, Y_n)$ where $x_i \equiv i/n$ and $x \in [0, 1]$. Then, the data can be written as

$$Y_i = m(x_i) + e_i, \qquad i = 1, \ldots, n, \tag{7}$$

where $e_i = Y_i - m(x_i)$ satisfying $\mathbf{E}[e] = 0$. Further, we assume that the sequence $\{e_i\}$ is a covariance stationary process.

**Definition 3.1** (Covariance Stationarity). *The sequence $\{e_i\}$ is covariance stationary if*

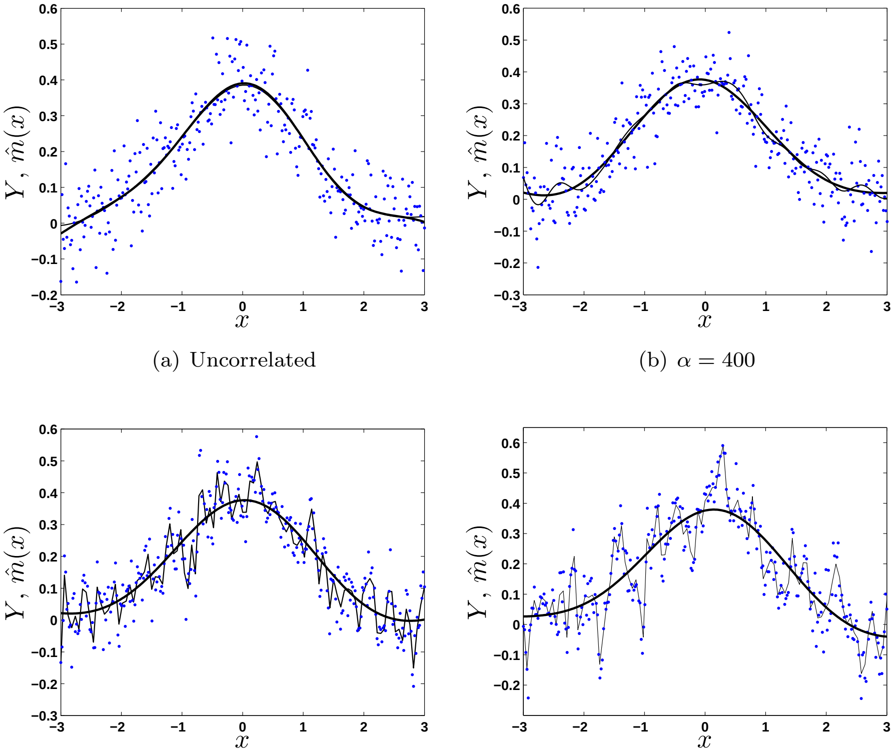
- $\mathbf{E}[e_i] = \mu$ *for all $i$*
- $\mathbf{Cov}[e_i, e_{i-j}] = \mathbf{E}[(e_i - \mu)(e_{i-j} - \mu)] = \gamma_j$ *for all $i$ and any $j$.*

## 3.2 Problems With Correlation

For all nonparametric regression techniques, the shape and the smoothness of the estimated function depends on a large extent on the specific value(s) chosen for the kernel bandwidth (and/or regularization parameter). In order to avoid selecting values for these parameters by trial and error, several data-driven methods are developed. However, the presence of correlation between the errors, if ignored, causes breakdown of commonly used automatic tuning parameter selection methods such as cross-validation (CV) or plug-in. The breakdown of automated methods, as well as a suitable solution, is illustrated by means of a simple example shown in Figure 6. For 400 equally spaced observations and a mean function $m(x) = (1/\sqrt{2\pi}) \exp(-x^2/2)$, four progressively more correlated sets of errors were generated from the same vector of independent noise and added to the mean function. The errors are normally distributed with variance $\sigma^2 = 0.01$ and correlation following an Auto Regressive process of order 1, denoted by AR(1), $\text{corr}(e_i, e_j) = \exp(-\alpha|x_i - x_j|)$ (Fan & Yao, 2003). Figure 6 shows four local linear regression estimates for these data sets. For each data set, two bandwidth selection methods were used: leave-one-out CV (LOO-CV) and a correlation-corrected CV (CC-CV) which is discussed at the end of this section. Table 4 summarizes the bandwidths selected for the four data sets under both methods. Table 4 and Figure 6 clearly show that when correlation increases, the bandwidth selected by CV becomes small, and the estimates become more undersmoothed. The bandwidths selected by CC-CV, a method that accounts for the presence of correlation, are much more stable and tends to select larger bandwidths with increasing correlation. This effect will later be theoretically confirmed.

**Table 4:** *Summary of bandwidth selection for simulated data in Figure 6*

| Correlation level | Autocorrelation | CV | CC-CV |
|---|---|---|---|
| Independent | 0 | 1.02 | 1.05 |
| $\alpha = 400$ | 0.37 | 0.62 | 1.07 |
| $\alpha = 200$ | 0.61 | 0.04 | 1.17 |
| $\alpha = 100$ | 0.79 | 0.03 | 1.51 |



(a) Uncorrelated

(b) $\alpha = 400$

(c) $\alpha = 200$

(d) $\alpha = 100$

**Figure 6:** *Simulated data with four levels of AR(1) correlation, estimated with least squares support vector machines; thin line represents estimate obtained with bandwidth selected by LOO-CV; bold line represents estimate obtained with bandwidth selected by our method.*

## 3.3 Kernel Regression with Correlated Errors Revised

In this section we will show that the form of the kernel is very important when errors are correlated. This is in contrast with the i.i.d. case where the choice between the various kernels is not very crucial (Härdle, 1999). In what follows, the kernel $K$ is assumed to be an isotropic kernel.

To estimate the unknown regression function $m$, consider the Nadaraya-Watson (NW) kernel estimator defined as

$$\hat{m}_n(x) = \sum_{i=1}^{n} \frac{K(\frac{x-x_i}{h})Y_i}{\sum_{j=1}^{n} K(\frac{x-x_j}{h})},$$

where $h$ is the bandwidth of the kernel $K$. This kernel can be one of the following kernels: Epanechnikov, Gaussian, triangular, spline,... An optimal $h$ can for example be found by minimizing the leave-one-out cross-validation (LOO-CV) score function

$$\text{LOO–CV}(h) = \frac{1}{n}\sum_{i=1}^{n}\left(Y_i - \hat{m}_n^{(-i)}(x_i;h)\right)^2, \tag{8}$$

where $\hat{m}_n^{(-i)}(x_i;h)$ denotes the leave-one-out estimator where point $i$ is left out from the training. For notational ease, the dependence on the bandwidth $h$ will be suppressed. In what follows we will use the following notation

$$k(u) = \int_{-\infty}^{\infty} K(y)e^{-iuy}\,dy$$

for the Fourier transform of the kernel function $K$. De Brabanter *et al.* (2011b) proved the following result:

**Theorem 3.1** (Bimodal kernel theorem)**.** *Assume uniform equally spaced design,* $x \in [0,1]$*,* $\mathbf{E}[e] = 0$*,* $\mathbf{Cov}[e_i, e_{i+k}] = \mathbf{E}[e_ie_{i+k}] = \gamma_k$ *and* $\gamma_k \sim k^{-a}$ *for some* $a > 2$*. Assume that*

*(C1)* $K$ *is Lipschitz continuous at* $x = 0$*;*

*(C2)* $\int K(u)\,du = 1, \lim_{|u|\to\infty} |uK(u)| = 0, \int |K(u)|\,du < \infty, \sup_u |K(u)| < \infty$*;*

*(C3)* $\int |k(u)|\,du < \infty$ *and* $K$ *is symmetric.*

*Assume further that boundary effects are ignored and that* $h \to 0$ *as* $n \to \infty$ *such that* $nh^2 \to \infty$*, then for the NW smoother it follows that*

$$\mathbf{E}[\text{LOO–CV}(h)] = \frac{1}{n}\mathbf{E}\left[\sum_{i=1}^{n}\left(m(x_i) - \hat{m}_n^{(-i)}(x_i)\right)^2\right] + \sigma^2 - \frac{4K(0)}{nh - K(0)}\sum_{k=1}^{\infty}\gamma_k + o(n^{-1}h^{-1}). \tag{9}$$

From this result it is clear that, by taking a kernel satisfying the condition $K(0) = 0$, the correlation structure is removed without requiring any prior information about its structure and (9) reduces to

$$\mathbf{E}[\text{LOO–CV}(h)] = \frac{1}{n}\mathbf{E}\left[\sum_{i=1}^{n}\left(m(x_i) - \hat{m}_n^{(-i)}(x_i)\right)^2\right] + \sigma^2 + o(n^{-1}h^{-1}). \tag{10}$$

Therefore, it is natural to use a bandwidth selection criterion based on a kernel satisfying $K(0) = 0$, defined by

$$\hat{h}_b = \underset{h\in\mathcal{Q}_n}{\arg\min}\,\text{LOO–CV}(h),$$

where $\mathcal{Q}_n$ is a finite set of parameters. Notice that if $K$ is a symmetric probability density function, then $K(0) = 0$ implies that $K$ is not unimodal. Hence, it is obvious to use bimodal kernels. A major advantage of using a bandwidth selection criterion based on bimodal kernels is the fact that is more efficient in removing the correlation than leave-$(2l+1)$-out CV (Chu & Marron, 1991).

**Definition 3.2** (Leave-$(2l+1)$-out CV)**.** *Leave-*$(2l+1)$*-out CV or modified CV (MCV) is defined as*

$$\text{MCV}(h) = \frac{1}{n}\sum_{i=1}^{n}\left(Y_i - \hat{m}_n^{(-i)}(x_i)\right)^2, \tag{11}$$

*where $\hat{m}_n^{(-i)}(x_i)$ is the leave-$(2l+1)$-out version of $m(x_i)$, i.e. the observations $(x_{i+j}, Y_{i+j})$ for $-l \leq j \leq l$ are left out to estimate $\hat{m}_n(x_i)$.*

Taking a bimodal kernel satisfying $K(0) = 0$ results in (10) while leave-$(2l+1)$-out CV with unimodal kernel $K$, under the conditions of Theorem 3.1, yields

$$\mathbf{E}[\mathrm{MCV}(h)] = \frac{1}{n}\,\mathbf{E}\left[\sum_{i=1}^{n}\left(m(x_i) - \hat{m}_n^{(-i)}(x_i)\right)^2\right] + \sigma^2 - \frac{4K(0)}{nh - K(0)}\sum_{k=l+1}^{\infty}\gamma_k + o(n^{-1}h^{-1}).$$

The result above clearly shows that leave-$(2l+1)$-out CV with unimodal kernel $K$ partially removes the correlation structure i.e. only the first $l$ elements of the correlation are removed.

The use of bimodal kernels looks promising and easy to use in case of Nadaraya-Watson kernel regression, local polynomial regression etc. in the presence of correlated errors. This is indeed an appealing property, however several other type of kernel smoothers such as support vector machines (SVM) and least squares support vector machines (LS-SVM) cannot directly use such type of kernels because these methods require a positive (semi-) definite kernel. Unfortunately, a kernel satisfying $K(0) = 0$ is never positive (semi-) definite (De Brabanter *et al.*, 2011b). Therefore we need to work in a two-step procedure in order to make this work for SVM and LS-SVM. First, calculate the residuals $\hat{e}$ based on a bimodal kernel via a Nadaraya-Watson kernel smoother or local linear regression. Second, choose $l$ for modified CV (11) to be the smallest $q \geq 1$ such that

$$|r_q| = \left|\frac{\sum_{i=1}^{n-q}\hat{e}_i\hat{e}_{i+q}}{\sum_{i=1}^{n}\hat{e}_i^2}\right| \leq \frac{2}{\sqrt{n}}.$$

Third, use SVM or LS-SVM with modified CV to tune the regularization parameter and kernel bandwidth. We conclude this section with the following theorem describing the relation between the asymptotic mean integrated squared error bandwidth under correlation $\hat{h}_{\mathrm{AMISE}}$ and the bandwidth under independent errors $\hat{h}_0$.

**Theorem 3.2** (De Brabanter *et al.*, 2011b)**.** *Let* (7) *hold and assume the regression function $m$ has two continuous derivatives. Assume also that $\mathbf{Cov}[e_i, e_{i+k}] = \gamma_k$ for all $k$, where $\gamma_0 = \sigma^2 < \infty$ and $\sum_{k=1}^{\infty} k|\gamma_k| < \infty$. Now, as $n \to \infty$, then*

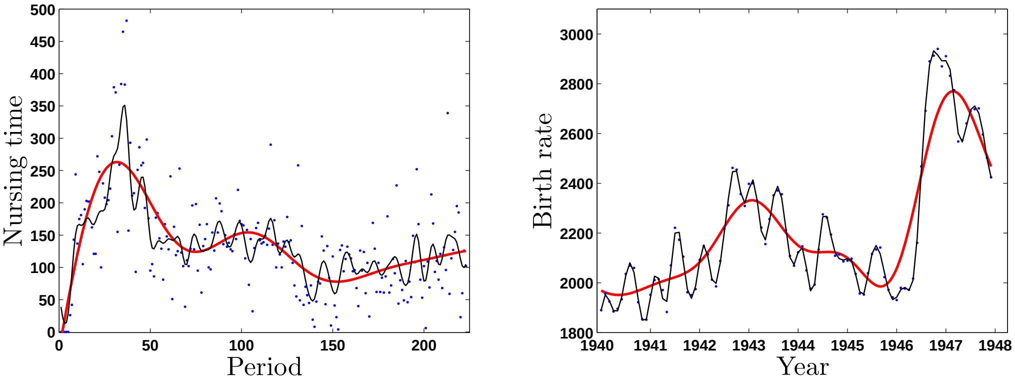
$$\hat{h}_{\mathrm{AMISE}} = \left[1 + 2\sum_{k=1}^{\infty}\rho(k)\right]^{1/5}\hat{h}_0,$$

*where $\rho(k)$ denotes the autocorrelation function at lag $k$, i.e. $\rho(k) = \gamma_k/\sigma^2 = \mathbf{E}[e_i e_{i+k}]/\sigma^2$.*

Thus, if the data are positively autocorrelated ($\rho(k) \geq 0 \quad \forall k$), the optimal bandwidth under correlation is larger than that for independent data. The behavior of increasing bandwidth with increasing correlation has been clearly observed in Table 4.

## 3.4 Simulations

Consider the nursing time of the beluga whale and birth rate data set (Simonoff, 1996). Figure 7(a) shows the scatter plot that relates the nursing time (in seconds) of a newborn beluga whale calf Hudson to the time after birth, where time is measured is six-hour time periods. Figure 7(b) shows the the U.S. monthly birth rate for the period from January 1940 through December 1947. Both results show that the estimate based on classical leave-one-out CV (assumption of no correlation) is very rough while the proposed CC-CV method produces a smooth regression fit.



(a) (b)

**Figure 7:** Typical results of the LS-SVM regression estimates for the (a) nursing time of the beluga whale and (b) birth rate data set. The thin line represents the estimate with tuning parameters determined by LOO-CV and the bold line is the estimate based on the CC-CV tuning parameters.

# 4 Additive least squares support vector machines

In general, direct estimation of high dimensional nonlinear functions using a nonparametric technique without imposing restrictions faces the problem of the curse of dimensionality. Several attempts were made to overcome this obstacle, for example projection pursuit (Friedman & Stuetzle, 1981) and additive models (Hastie & Tibshirani, 1990). Especially additive models and their extensions have become a widely used modeling technique as they offer a compromise between flexibility, dimensionality and interpretability. Traditionally splines are often used for additive models. Estimation of the nonlinear components is usually performed with backfitting (Hastie & Tibshirani, 1990) or a marginal integration based estimator (Linton & Nielsen, 1995; Mammen *et al.*, 1999). Next, we show an alternative to construct additive models based on LS-SVM.

## 4.1 Formulation of LS-SVM additive models

Given a data set $(X, Y)$ where $X \in \mathbb{R}^d$ and $Y \in \mathbb{R}$. Let $X^{(j)} \in \mathbb{R}$ denote the $j$-th component of an input vector $X \in \mathbb{R}^d$. Consider the following additive model

$$Y = b + \sum_{j=1}^{d} m_j(X^{(j)}) + e = m(X) + e,$$

where the errors $e$ are independent of the $X^{(j)}$s, $\mathbf{E}[e|X] = 0$, $\mathbf{Var}[e|X] = \sigma^2 < \infty$ and the $m_j$s are arbitrary univariate functions, one for each dimension (predictor). Consider a model classes of the form

$$\mathcal{F}^{\star}_{n,\Psi} = \left\{ f : f(X) = \sum_{j=1}^{d} w_j^T \varphi_j(X^{(j)}) + b, \varphi_j : \mathbb{R} \to \Psi, w \in \mathbb{R}^{n_f}, b \in \mathbb{R} \right\}.$$

The optimization problem for LS-SVM (see also (4)) can be written w.r.t. the model class as follows

$$\begin{aligned} &\min_{w,b,e} \mathcal{J}_P(w,e) = \tfrac{1}{2} \textstyle\sum_{j=1}^{d} w_j^T w_j + \tfrac{\gamma}{2} \displaystyle\sum_{i=1}^{n} e_i^2 \\ &\text{s.t.} \quad Y_i = \textstyle\sum_{j=1}^{d} w_j^T \varphi_j(X_i^{(j)}) + b + e_i, \quad i = 1, \ldots, n, \end{aligned}$$

By using Lagrange multipliers, the solution is given by

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

where $\Omega^{\star} = \sum_{j=1}^{d} \Omega^{(j)}$ and $\Omega_{kl}^{(j)} = K^{(j)}(X_k^{(j)}, X_l^{(j)})$ for all $k, l = 1, \ldots, n$ (sum of univariate kernels). The resulting additive LS-SVM is given by

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

**Remark 4.1.** Fitting an additive model using backfitting (Hastie & Tibshirani, 1990) was also employed. Since this resulted in similar estimates we did not display the results. Although the training of the explained algorithm can be done in one shot in contrast to backfitting, the model selection process can become quite tedious if one chooses to use one bandwidth per dimension especially when the number of dimensions is relatively high. This can be solved by taking only one bandwidth for all dimensions. Generally, less good results can be expected in the latter case.

### 4.2 Simulation

Consider the U.S. temperature data (Peixoto, 1990; Ruppert *et al.*, 2003) containing 56 observations on the temperature and location of 56 U.S. cities. In general, the scales on the vertical axes are only meaningful in a relative sense; they have no absolute interpretation. Since we have the freedom to choose the vertical positionings, we should try to make them meaningful in the absolute sense. A reasonable solution, is to plot, for each predictor, the profile of the response surface with each of the other predictors set at their average. This is shown in Figure 8.

## 5 Density estimation

### 5.1 Rosenblatt-Parzen kernel density estimator

Probably one of the most popular and known methods to estimate a density function is the Rosenblatt-Parzen estimator (Rosenblatt, 1956; Parzen, 1962). This kernel estimator with kernel $K$ is defined

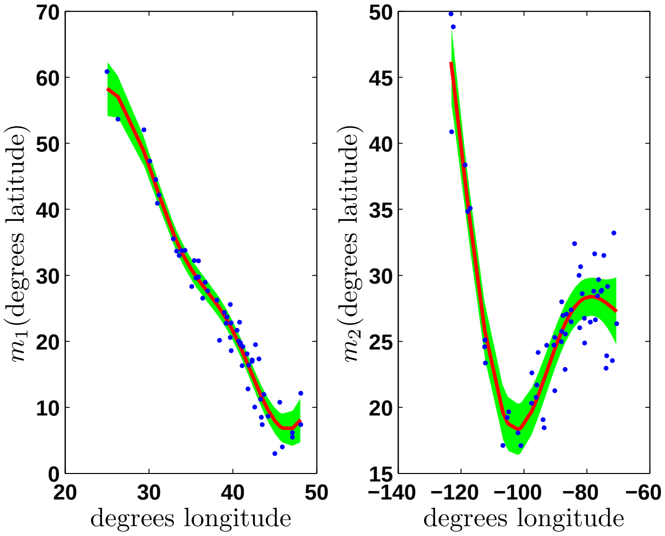


**Figure 8:** Fitted functions for the additive LS-SVM model for the U.S. temperature data. The shaded area represents twice the pointwise standard errors of the estimated curve. The points plotted are the partial residuals: the fitted values for each function plus the overall residuals from the additive model.

by

$$\hat{f}(x) = \frac{1}{nh^d} \sum_{i=1}^{n} K\left(\frac{x - X_i}{h}\right), \tag{12}$$

where $h$ is called the bandwidth or smoothing parameter of the kernel and $X \in \mathbb{R}^d$. Also, the kernel $K$ satisfies the following conditions

$$\int_{\mathbb{R}^d} K(u)\,du = 1, \quad \int_{\mathbb{R}^d} uK(u)\,du = 0, \quad 0 < \int_{\mathbb{R}^d} u^2 K(u)\,du < \infty.$$

The estimator (12) can be considered as a sum of "bumps" placed above each observation. The resulting estimator is then the sum of all these bumps. The kernel function $K$ determines the shape of the bumps while the bandwidth $h$ determines their width.

## 5.2 Regression view of density estimation

We will establish a connection between density estimation and nonparametric regression (Fan & Gijbels, 1996; Wasserman, 2006). This connection can be seen by using a binning technique. Suppose we are interested in estimating the density function $f$ on an interval $[a, b]$. Partition the interval $[a, b]$ into $N$ subintervals $\{I_k, k = 1, \ldots, N\}$ of equal length $\Delta = (b - a)/N$. Let $x_k$ be the center of $I_k$ and $y_k$ be the proportion of the data $\{X_i, i = 1, \ldots, n\}$ falling in the interval $I_k$, divided by the bin length $\Delta$. The number of subintervals can be determined by $N = \lceil (b - a)/3.49\,\mathrm{MAD}(X) n^{-1/3} \rceil$ (Silverman, 1986), where $\lceil \cdot \rceil$ denotes the largest integer and MAD denotes the median absolute deviation, or by least squares crossvalidation (Rudemo, 1982) etc. It is clear that the bin counts $n\Delta y_k$ i.e. the number of sample points falling in the $k$th bin, have a binomial distribution (Johnson *et al.*, 1997)

$$n\Delta y_k \sim \mathrm{Bin}(n, p_k) \quad \text{with} \quad p_k = \int_{I_k} f(x)dx, \tag{13}$$

with $p_k$ the probability content of the $k$th bin. For a fine enough partition i.e. $N \to \infty$, it can be calculated from (13), by using a Taylor series expansion of the density in the $k$th bin around the midpoint of the bin, that

$$\mathbf{E}[y_k] \approx f(x_k), \quad \mathbf{Var}[y_k] \approx \frac{f(x_k)}{n\Delta}. \tag{14}$$

Consequently, we could regard the density estimation problem as a heteroscedastic nonparametric regression problem based on the data $\{(x_k, y_k) : k = 1, \ldots, N\}$ which are approximately independent (Fan, 1996). The nonparametric regression problem is defined as follows

$$y_k = m(x_k) + \varepsilon_k, \quad \varepsilon_k = e_k\eta(m(x_k), x_k),$$

where $e_k$ are independent and identically distributed. The function $\eta$ expresses the heteroscedasticity and $m$ is an unknown real-valued smooth function that we want to estimate. Often in practice homoscedastic data are preferred. The homoscedasticity can be accomplished via Anscombe's variance stabilizing transformation (Anscombe, 1948) to the bin count $c_k = n\Delta y_k$, i.e.

$$y_k^\star = 2\sqrt{c_k + \frac{3}{8}}.$$

The density estimator is then obtained by applying a nonparametric smoother to the transformed data set $\{(x_k, y_k^\star) : k = 1, \ldots, N\}$, and taking the inverse of Anscombe's transformation. Let $\hat{m}_n^\star(x)$ be a regression smoother based on the transformed data. Then the density estimator is defined by

$$\hat{f}(x) = \mathcal{C}\left[\frac{\hat{m}_n^\star(x)^2}{4} - \frac{3}{8}\right]_+, \tag{15}$$

where $\mathcal{C}$ is a normalization constant such that $\hat{f}(x)$ integrates to 1 and $[z]_+ = \max(z, 0)$. Then, $\hat{m}_n^\star$ can be estimated by means of any nonparametric method.

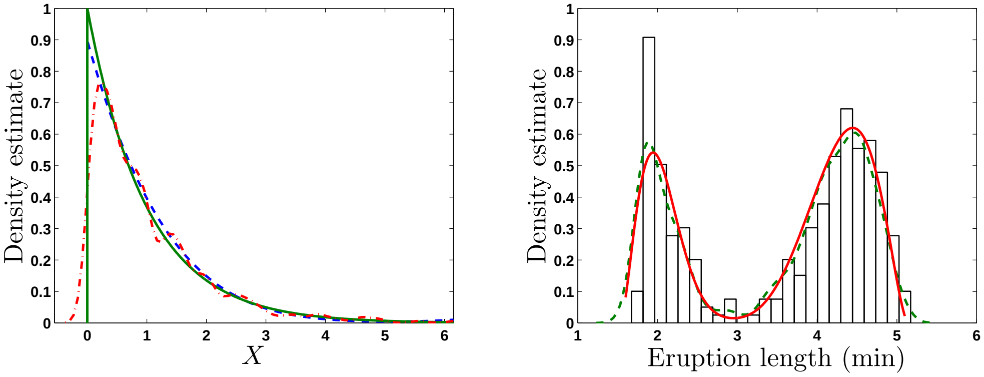
**Remark 5.1.** To determine the number of subintervals $N$ we have used a reference rule. The literature describes better procedures to find the number of subintervals, see e.g. (Park & Marron, 1990) for an overview, (Wand, 1997) for a plug-in type and (Devroye & Lugosi, 2004) for $L_1$ optimal bin width selection. Such methods can also be applied for the $d$-dimensional case.

**Remark 5.2.** Histogram based density estimators have been frequently studied in literature. The authors in (Beirlant *et al.*, 1999) considered estimating consistently an unknown probability density function in $L_1$ from a sample of i.i.d. random variables by means of histogram-based estimators. They showed that their proposed estimator is universally $L_1$ consistent and attains the kernel estimate rate in the expected $L_1$ error i.e. $n^{-2/5}$ in the univariate case. The disadvantage of their density estimate is that it can take on negative values with probability close to 1. In order to overcome this, the idea of (Beirlant *et al.*, 1999) was extended to nonnegative piecewise linear histograms (Berlinet & Vajda, 2001) and to generalized piecewise linear histograms (Berlinet *et al.*, 2002a,b).

### 5.3 Simulations

In the first example we generate a data set $X_i \sim \text{Exp}(1)$ for $i = 1, \ldots, 500$. Then, we estimate the exponential density by the Rosenblatt-Parzen kernel density estimator (12) and by (15). The bandwidth of (12) is determined by the Sheather-Jones solve-the-equation plug-in method (Sheather & Jones, 1991). The results are shown in Figure 9(a). From this simple example, it is clear that the Rosenblatt-Parzen kernel density estimator (dash dotted line) can fail when the region of definition of the data at hand is not unbounded. However, this drawback can be overcome by adaptive bandwidths (Silverman, 1986) or boundary kernels. On the other hand, the proposed density estimator (15) (dashed line) can deal with this difficulty.

In a second example, we consider the Old Faithful geyser data set. A histogram of the data as well as the Rosenblatt-Parzen kernel density estimate (dashed line) and the proposed estimate (full line) is shown in Figure 9(b). There is little difference between the two estimates.



(a)

(b)

**Figure 9:** (a) Exponential density (full line) estimated with the Rosenblatt-Parzen kernel density estimator (dash dotted line) and the proposed estimator (dashed line);(b) Rosenblatt-Parzen kernel density estimator (dashed line) compared with the proposed estimator (full line) on the Old Faithful geyser data set.

## 6 Conclusions

In this paper we gave an overview of some recent developments of kernel based modeling in the area of robustness, correlated data analysis, construction of additive models and density estimation. First, we illustrated that in order to obtain a full robust procedure three requirements have to be fulfilled: (i) robust smoother, (ii) bounded kernel and (iii) robust model selection criterion. Second, we advocated the use of kernels $K$ satisfying the condition $K(0) = 0$ in model selection criteria to automatically remove correlation without requiring any prior knowledge about its structure. Third, we have formulated an alternative approach for backfitting in the case of least squares support vector machine. Although model selection can be tedious, the training algorithm is done in one shot in contrast to the backfitting algorithm where an iterative approach has to be used. Finally, we have established a link between density estimation and nonparametric regression via binning.

**Acknowledgements**

Research supported by Onderzoeksfonds K.U.Leuven/Research Council KUL: GOA/11/05 Ambiorics, GOA/10/09 MaNet, CoE EF/05/006 Optimization in Engineering (OPTEC) en PFV/10/002 (OPTEC), IOF-SCORES4CHEM, several PhD/post-doc & fellow grants; Flemish Government: FWO: FWO: PhD/postdoc grants, projects: G0226.06 (cooperative systems and optimization), G0321.06 (Tensors), G.0302.07 (SVM/Kernel), G.0320.08 (convex MPC), G.0558.08 (Robust MHE), G.0557.08 (Glycemia2), G.0588.09 (Brain-machine) research communities (WOG: ICCoS, ANMMM, MLDM); G.0377.09 (Mechatronics MPC), IWT: PhD Grants, Eureka-Flite+, SBO LeCoPro, SBO Climaqs, SBO POM, O&O-Dsquare, Belgian Federal Science Policy Office: IUAP P6/04 (DYSCO, Dynamical systems, control and optimization, 2007-2011), IBBT, EU: ERNSI; FP7-HD-MPC (INFSO-ICT-223854), COST intelliCIS, FP7-EMBOCON (ICT-248940), Contract Research: AMINAL, Other: Helmholtz, viCERP, ACCM. BDM is a full professor at the Katholieke Universiteit Leuven.

# References

Altman, N.S. (1990). Kernel smoothing of data with correlated errors. *J. Amer. Statist. Assoc.*, **85**, 749–759.

Amemiya T. (1985). *Advanced Econometrics.* Harvard University Press, Cambridge, MA.

Anscombe F.J. (1948). The transformation of Poisson, binomial and negative-binomial Data. *Biometrika*, **35**, 246–254.

Beirlant J., Berlinet A. & Györfi L. (1999). On piecewise linear density estimators. *Stat. Neerl.*, **53**, 287–308.

Berlinet A. & Vajda I. (2001). Nonnegative piecewise linear histograms. *Statistics*, **35**, 295–317.

Berlinet A., Hobza T. & Vajda I. (2002). Generalized piecewise linear histograms. *Stat. Neerl.*, **56**, 301–313.

Berlinet A., Hobza T. & Vajda I. (2002). Asymptotics for generalized piecewise linear histograms. *Ann. I.S.U.P.*, **46**, 3–19.

Chu C.K. & Marron J.S. (1991). Comparison of two bandwidth selectors with dependent errors. *Ann. Statist.*, **19**, 1906–1918.

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

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. (2011). *Least Squares Support Vector Regression with Applications to Large-Scale Data: a Statistical Approach.* PhD thesis, Katholieke Universiteit Leuven, Belgium.

De Brabanter K., De Brabanter J., Suykens J.A.K. & De Moor B. (2011b). Kernel regression in the presence of correlated errors. *J. Mach. Learn. Res.*, **12**, 1955-1976.

De Brabanter K., Suykens J.A.K. & De Moor B. (2011c). Nonparametric regression via StatLSSVM. Technical Report 11-134 ESAT-SISTA, K.U.Leuven (Leuven, Belgium).

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.

Devroye L. & Lugosi G. (2004). Bin width selection in multivariate histograms by the combinatorial method. *Test*, **13**, 129–145.

Fan, J. (1996). Test of significance based on wavelet thresholding and Neyman's truncation. *J. Amer. Statist. Assoc.*, **91**, 674–688.

Fan J. & Gijbels I. (1996). *Local Polynomial Modelling and Its Applications*. Wiley.

Fan J. & Yao Q. (2003). *Nonlinear Time Series: Nonparametric and Parametric Methods*. Springer.

Friedman J.H. & Tukey J.W. (1974). A projection pursuit algorithm for exploratory data analysis. *IEEE Trans. Comput.*, **C-23**, 881–890.

Friedman J.H. & Stuetzle W. (1981). Projection pursuit regression. *J. Amer. Statist. Assoc.*, **76**, 817–823.

Györfi L., Kohler M., Krzyżak A. & Walk H. (2002). *A Distribution-Free Theory of Nonparametric Regression*. Springer.

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

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.

Härdle W. (1999). *Applied Nonparametric Regression* (reprinted). Cambridge University Press.

Hastie T. & Tibshirani R. (1990). *Generalized addidive models*. London: Chapman and Hall

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

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

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

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

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.

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.

Johnson N., Kotz S. & Balakrishnan, N. (1997). *Discrete Multivariate Distributions.* Wiley & Sons.

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

Kulkarni S.R., Posner S.E. & Sandilya S. (2002). Data-dependent $k_n$-NN and kernel estimators consistent for arbitrary processes. *IEEE Trans. Inform. Theory*, **48**, 2785–2788.

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

Linton O.B. & Nielsen J.P. (1995). A kernel method for estimating structured nonparameteric regression based on marginal integration. *Biometrika*, **82**, 93–100.

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

Mammen E., Linton O. & Nielsen J.P. (1999). The existence and asymptotic properties of a backfitting projection algorithm under weak conditions. *Ann. Statist.*, **27**, 1443–1490.

Manski C.F. (1988). Identification of binary response models. *J. Amer. Statist. Assoc.*, **83**, 729-738.

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

Nadaraya E.A. (1964). On estimating regression. *Theory Probab. Appl.*, **9**, 141–142.

Opsomer J. & Ruppert D. (1997). Fitting a bivariate additive model by local polynomial regression, *Ann. Statist.*, **25**, 186–211.

Opsomer J., Wang Y. & Yang, Y. (2001). Nonparametric regression with correlated errors. *Statist. Sci.*, **16**, 134–153.

Park B.U. & Marron J.S. (1990). Comparison of data-driven bandwidth selectors. *J. Amer. Statist. Assoc.*, **85**, 66–72.

Parzen E. (1962). On estimation of a probability density function and mode. *Ann. Statist.*, **33**, 1065–1076.

Peixoto J.L. (1990). A property of well-formulated polynomial regression models. *Amer. Statist.*, **44**, 26–30.

Rosenblatt M. (1956). Remarks on some nonparametric estimates of a density function. *Ann. Statist.*, **27**, 832–837.

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

Rudemo M. (1982). Empirical choice of histograms and kernel density estimators. *Scand. J. Stat.*,**9**, 65–78.

Ruppert D., Wand M.P. & Carroll R.J. (2003). *Semiparametric Regression.* Cambridge University Press.

Sheather S.J. & Jones M.C. (1991). A reliable data-based bandwidth selection method for kernel density estimation. *J. R. Stat. Soc. Ser. B Stat. Methodol.*, **53**, 683–690.

Silverman B.W. (1986). *Density Estimation for Statistics and Data Analysis.* Chapman & Hall.

Simonoff J.S. (1996). *Smoothing Methods in Statistics.* Springer.

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

Suykens J.A.K., De Brabanter J., Lukas L. & Vandewalle J. (2002b). Weighted least squares support vector machines: robustness and sparse approximation. *Neurocomputing*, **48**, 85–105.

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

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

Wand M.P. (1997). Data-based choice of histogram bin width. *Amer. Statist.*, **51**, 59–64.

Wasserman L. (2006). *All of Nonparametric Statistics.* Springer.

Watson G.S. (1964). Smooth regression analysis. *Sankhyā*, **26**, 359–372.

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