Robust Estimation for Nonparametric Generalized Regression

Ana M. Bianco*
Instituto de Cálculo, F.C.E. y N., Universidad de Buenos Aires
and CONICET, Argentina
abianco@dm.uba.ar

Graciela Boente
Instituto de Cálculo and Departamento de Matemáticas, F.C.E. y N.,
Universidad de Buenos Aires and CONICET, Argentina
gboente@dm.uba.ar

and
Susana Sombielle
Instituto de Cálculo, F.C.E. y N., Universidad de Buenos Aires
and Universidad Tecnológica Nacional, Argentina
ssombielle@gmail.com

In this paper, we focus on nonparametric regression estimation for the parameters of a discrete or continuous distribution, such as the Poisson or Gamma distributions, when anomalous data are present. The proposals are nonparametric versions of robust estimators that have been introduced in the parametric setting for generalized linear models. We present two families of estimators and their asymptotic behaviour is studied. Through a Monte Carlo study we compare the performance of the proposed estimators with the classical ones. We also introduce a resistant cross–validation method to choose the smoothing parameter.

Key words and phrases: asymptotic properties; nonparametric generalized regression; robust estimation; smoothing techniques

Running Title: Robust nonparametric generalized regression estimation

1 Introduction

Generalized linear models, introduced by [30], are extensively used in statistical applications due to their flexibility to fit a large variety of regression problems, whenever the response is continuous or discrete. They have been proposed as a natural extension of the classical linear model. Many aspects of these models, such as the estimation of the parameters using iterative procedures and resistant methods, have been studied to a great extent. However, a trend in the last few years has been to determine the underlying model or to check a parametric model, via nonparametric techniques. In the nonparametric situation, classical estimators of the regression function are based on local means and so, they are very sensitive to outliers. The effect of a single outlier depends on how far it lies from the point of interest, that is, only the observations in a neighbourhood of this point need to be considered when studying the sensitivity of the procedure. Thus, robust concepts should be thought in terms of local resistance properties. In this paper, our concern is to estimate robustly and nonparametrically the regression function in the presence of anomalous data.

More precisely, from now on we assume that the response variables $\{Y_i : 1 \leq i \leq n\}$ are independent random variables, related to covariates which may be fixed or random. In the case of fixed or deterministic carriers, we assume that there exists a smooth function $g : \mathbb{R}^d \to \tau \subset \mathbb{R}$ such that Y_i has distribution $F(y, g(\mathbf{x}_i))$, where $\mathbf{x}_i \in A \subset \mathbb{R}^d$, $1 \leq i \leq n$ and A is a compact set. If we observe random carriers $\{\mathbf{X}_i : 1 \leq i \leq n\}$, we assume that $Y_i|\mathbf{X}_i = \mathbf{x}_i$ has distribution $F(y, g(\mathbf{x}_i))$ with $g : \mathbb{R}^d \to \tau \subset \mathbb{R}$. In the generalized lineal model the regression function g depends on a vector of parameters, i.e., $g(\mathbf{x}) = H(\mathbf{x}^{\mathrm{T}}\boldsymbol{\beta})$, where H is known and $\boldsymbol{\beta}$ must be estimated. In our setting, we only assume that g is a smooth function. In most situations, the function g satisfies $g(\mathbf{x}_i) = \mathbb{E}(Y_i)$ in the case of fixed covariates and $g(\mathbf{x}) = \mathbb{E}(Y_i|\mathbf{X}_i = \mathbf{x})$ for random ones.

Many authors have considered the problem of the estimation of the regression function g, for instance [12], [16], [21] and [32]. In the fixed design case, asymptotic properties for linear estimates were obtained by [20] without assuming a regression model. Local polynomial kernel smoothers for one parameter exponential families were introduced by [17]. Their estimators are based on quasi–likelihood functions which are linear in the response variables. They include as a particular case local constants kernel estimators, but their approach turns to be non resistant in most cases. In regression models with fixed carriers, a first resistant proposal was given by [10] who considered a robust locally weighted regression scatterplot (see also, [11]) and asymptotic results for the univariate case were obtained by [23]. Robust methods for estimating $g(\mathbf{x})$ under a nonparametric regression model have been proposed by [22], [24] and [4]. All these methods have been designed in order to include the regression model $Y = g(\mathbf{x}) + \varepsilon\, s(\mathbf{x})$ where ε has a continuous and symmetric distribution function and s is a scale function. In logistic models, the first paper which applied kernel methods is due to [13]. Later on, [18] used this smoothing technique for diagnostics in logistic regression. More recently, [8] have considered nonparametric regression in exponential families using a mean-matching variance stabilizing transformation so as to turn the estimating issue in a standard homoscedastic Gaussian regression problem.

From now on, we focus on nonparametric regression on the parameters of a discrete or continuous distribution, such as the Poisson or Gamma distributions, when anomalous data are present and our aim is to estimate the regression function g in a robust fashion. Our resistant proposals are nonparametric versions of robust estimators that have been proposed in the parametric setting,

such as those considered by [3] for logistic regression, [2] for the Gamma distribution and [26] and [9] for generalized linear models.

In Section 2, two families of estimators are introduced and some applications are described. Their asymptotic behaviour is studied in both Sections 3 and 4. In Section 5, we present the results of a Monte Carlo Study and we introduce a resistant cross–validation method to choose the smoothing parameter. Proofs are relegated to the Appendix.

2 The estimators

2.1 General Definitions

As mentioned in the Introduction our aim is to extend to general nonparametric models robust estimators that have been developed in the parametric setting. We propose two families of local $M-$estimators that can be defined through a loss function ρ or through a score function ψ.

We begin by fixing some notation. From now on, $X_n \xrightarrow{a.s.} X$ means that $X_n \longrightarrow X$ almost surely, while $\xrightarrow{p}$ stands for convergence in probability. Besides, $\mathbb{I}_A$ will denote the indicator function for the set A.

To fix ideas, let us focus on the deterministic design case first. Assume that we observe $\{Y_i : 1 \leq i \leq n\}$ which are independent random variables such that $Y_i \sim F(y, g(\mathbf{x}_i))$ with $g : \mathbb{R}^d \to \tau \subset \mathbb{R}$ and related to fixed carriers $\mathbf{x}_i \in A \subset \mathbb{R}^d$, $1 \leq i \leq n$, where A is a compact set.

Consider a generic random variable Y with distribution $F(y, g(\mathbf{x}))$, for a given continuity point $\mathbf{x} \in A$ of g. In order to introduce a local $M-$estimator, let $\rho : \mathbb{R}^2 \to \mathbb{R}_{>0}$ be a loss function. In some situations, for instance when $\rho(Y,t) = |Y - t|$, $\mathbb{E}(\rho(Y,t))$ may not exist. For that reason, to ensure a proper definition for the target functionals, we include a term $a(Y)$, and so, we define a function $\gamma : \mathbb{R} \to \mathbb{R}$ as $\gamma(t) = \mathbb{E}[\rho(Y,t) - a(Y)]$. Without loss of generality, to define the estimators we will assume that $a(Y) \equiv 0$, since we can absorb it on the function ρ. Assume that

$$\gamma(g(\mathbf{x})) = \min_{t \in \tau} \gamma(t) \, .$$

Now, to estimate the function γ at any point t, we consider a sample version of γ given by

$$\gamma_n(t) = \sum_{i=1}^{n} w_{ni}(\mathbf{x}) \rho(Y_i, t) \, , \tag{1}$$

where $w_{ni}(\mathbf{x}) = w_{ni}(\mathbf{x}, \mathbf{x}_1, \ldots, \mathbf{x}_n)$ is a probability weight function, i.e., $w_{ni}(\mathbf{x}) \geq 0$ and $\sum_{i=1}^{n} w_{ni}(\mathbf{x}) = 1$, as for kernel, $k-$nearest neighbour, nearest neighbour and nearest neighbour with kernel weights. These local weights give more emphasis to those observations closest to $\mathbf{x}$. Then, in order to estimate $g(\mathbf{x})$ it seems natural to minimize (1) over $t \in \tau$. We will call $g_n(\mathbf{x})$ the value where $\gamma_n(t)$ attains its minimum, that is

$$\gamma_n(g_n(\mathbf{x})) = \min_{t \in \tau} \gamma_n(t) \, . \tag{2}$$

More generally, as in [25], one can define estimates of $g(\mathbf{x})$ as any sequence $g_n(\mathbf{x})$ satisfying

$$\gamma_n(g_n(\mathbf{x})) - \inf_{t \in \tau} \gamma_n(t) \xrightarrow{a.s.} 0 \, . \tag{3}$$

Instead of defining $g_n(\mathbf{x})$ as the solution of an optimization problem, one can define it through the relative differentiating equation by considering $\Psi(y,t) = \partial\rho(y,t)/\partial t$ and so $g_n(\mathbf{x})$ satisfies

$$\lambda_n(g_n(\mathbf{x})) = 0, \tag{4}$$

where for $t \in \tau$

$$\lambda_n(t) = \sum_{i=1}^{n} w_{ni}(\mathbf{x})\Psi(Y_i,t) \tag{5}$$

and $w_{ni}(\mathbf{x})$ are as in (1). These two approaches are equivalent if (4) has a unique solution. Besides, for generalized linear models in the parametric setting, $M-$estimators have been defined through several choices of Ψ which are not necessarily obtained by differentiating a loss function. Thus, we will consider a general function $\Psi : \mathbb{R}^2 \to \mathbb{R}$ and the solution $g_n(\mathbf{x})$ of (4), where we suspect that this solution will converge to the solution $g(\mathbf{x})$ of

$$\lambda(t) = \mathbb{E}\left[\Psi(Y,t)\right] = 0\ . \tag{6}$$

In the case of random carriers, the definitions of these two families of estimators follow straightforwardly by considering suitable functions $\gamma(t)$ and $\lambda(t)$ and their sample versions $\gamma_n(t)$ and $\lambda_n(t)$. Indeed, assume that $(\mathbf{X}_1^{\mathrm{T}}, Y_1)\ldots(\mathbf{X}_n^{\mathrm{T}}, Y_n)$ are independent random vectors in $\mathbb{R}^{d+1}$, such that $Y_i|\mathbf{X}_i = \mathbf{x}_i$ has distribution $F(y, g(\mathbf{x}_i))$ with $g : \mathbb{R}^d \to \tau \subset \mathbb{R}$. Consider a generic vector $(\mathbf{X}^{\mathrm{T}}, Y)$ with the same distribution of the sample, that is

$$Y|\mathbf{X} = \mathbf{x} \sim F(y, g(\mathbf{x})). \tag{7}$$

Given a loss function ρ define $\gamma(t) = \mathbb{E}[\rho(Y,t)|\mathbf{X} = \mathbf{x}]$ for $t \in \tau \subset \mathbb{R}$, where $\mathbf{x}$ is a continuity point of g . Assume that $\gamma(g(\mathbf{x})) = \min_{t\in\tau}\gamma(t)$. Then, it is quite natural to estimate $g(\mathbf{x})$ by $g_n(\mathbf{x})$ where $g_n(\mathbf{x})$ minimizes

$$\gamma_n(t) = \sum_{i=1}^{n} W_{ni}(\mathbf{x})\,\rho(Y_i,t)\ , \quad t \in \tau\ , \tag{8}$$

where $W_{ni}(\mathbf{x}) = W_{ni}(\mathbf{x}, \mathbf{X}_1, \ldots, \mathbf{X}_n)$ are local weights as those mentioned above.

On the other hand, for a score function Ψ define

$$\lambda(t) = \mathbb{E}\left[\Psi(Y,t)|\mathbf{X} = \mathbf{x}\right]\ . \tag{9}$$

Therefore, we define $g_n(\mathbf{x})$ as the solution of the equation

$$\lambda_n(g_n(\mathbf{x})) = 0\ , \tag{10}$$

where

$$\lambda_n(t) = \sum_{i=1}^{n} W_{ni}(\mathbf{x})\,\Psi(Y_i,t)\ , \quad t \in \tau\ . \tag{11}$$

2.2 Some applications

In regression models, robust estimates can be obtained by taking $\rho(y,t) = \phi(y-t)$ where ϕ is, for instance, the Huber or the bisquare $\rho-$function. In generalized regression models, typically in order to attain robustness, ρ will be a bounded function performing like the log–likelihood for central values. In this last case, instead of bringing in large observations in the derivative of the log–likelihood function, one option is to smoothly truncate the log–likelihood function and then correct it by an additive term only depending on the parameter in order to obtain both robustness and Fisher–consistency. Thus, in order to estimate $g(\mathbf{x})$ in generalized exponential families, we propose to minimize (1) or (8) using

$$\rho(y,t) = \phi[-\ln\, f(y,t) + H(y)] + G(t)\ ,$$

where ϕ is an odd and bounded non–decreasing function. Typically, ϕ is a function performing like the identity function in a neighbourhood of 0, the function $H(y)$ is used to remove a term from the log–likelihood that is independent of the parameter, while G is a correction term introduced to achieve Fisher–consistency.

When dealing with a one parameter exponential family $\ln f(y,t) = y\,t - b(t) + c(y)$ and $H(y)$ can be taken as $\ln f(y,y)$. Thus, we can choose

$$\rho(y,t) = \phi(\ln f(y,y) - \ln f(y,t)) + G(t)\,, \tag{12}$$

with $G'(t) = \int \varphi\,(\ln f(y,y) - \ln f(y,t))\ f'(y,t)d\mu(y)$ to obtain Fisher–consistency, where φ stands for the derivative of ϕ, $f'(y,t) = \partial/\partial t\, f(y,t)$ and $f(\cdot,t)$ is the density with respect to the measure μ of the distribution function $F(\cdot,t)$. The function ϕ can be chosen as the bisquare function or as in [14]. This approach is in the spirit of [3] where a robustified version of the deviance with a correction term is introduced to achieve Fisher-consistency.

To illustrate with some examples, let us consider first the Poisson regression problem. Assume that Y follows a Poisson distribution, $Y \sim P(t)$. Then, we have that

$$f(y,t) = \left\{ \begin{array}{cc} \exp(-t)t^y/y! & y \in N \cup \{0\} \\ 0 & \text{in other case} \end{array} \right.$$

and thus, $\mathbb{E}(Y) = t$, $\mathbb{V}(Y) = t$. In this case, if ϕ is a loss function, from (12) we get that

$$\rho(y,t) \;\; = \;\; \phi(\ln(f(y,y) - \ln(f(y,t)) + G(t) = \phi(-y + y\ln y + t - y\ln t) + G(t)\,. \tag{13}$$

Moreover, if, as above, ϕ has first derivative φ, we have that

$$\Psi(y,t) = \left\{ \begin{array}{cc} \varphi(-y + y\ln y + t - y\ln t)\dfrac{(t-y)}{t} + G'(t) & \text{if } y > 0 \\ \varphi(t) + G'(t) & \text{if } y = 0 \end{array} \right. ,$$

where

$$G'(t) = -\mathbb{E}\left[\Psi(Y,t)\right] = -\varphi(t)\exp(-t) - \sum_{j=1}^{\infty} \varphi(j\ln j - j + t - j\ln t)\left(\frac{t-y}{t}\right)\exp(-t)t^j/j!\ .$$

Another interesting example is the case of the Gamma regression model. Let us assume that Y follows a Gamma distribution, $Y \sim \Gamma(\alpha, t)$, with parameters α and t. Hence, we have that

$$f(y, \alpha, t) = \frac{\alpha^\alpha y^{\alpha-1}}{t^\alpha \Gamma(\alpha)} \exp(-\alpha y/t) \mathbb{I}_{\{y \geq 0\}} \, .$$

This parametrization implies that $\mathbb{E}(Y) = t$ and $\mathbb{V}(Y) = t^2/\alpha$, where α is assumed to be known. In the case of a continuous response, we have that $G(t) = 0$, see [2]. Thus, $\rho(y, t) = \phi(\ln(f(y, y) - \ln(f(y, t)) = \phi\left(\alpha(y/t - \ln(y/t) - 1)\right)$ which implies that

$$\rho(y,t) = \begin{cases} \phi\left(\alpha(y/t - \ln(y/t) - 1)\right) & \text{if } y > 0 \\ \lim_{y \to 0} \ \phi\left(\alpha(-\ln(y/t) - 1)\right) & \text{if } y = 0 \, , \end{cases}$$

where ϕ is a loss function as above.

On the other hand, if we define the estimator by solving an implicit equation, one can follow the approach introduced by [9] where a robustified quasi-likelihood estimator is developed. The estimator $g_n(\mathbf{x})$ satisfies (4) or (10) using

$$\Psi(y, t) = \psi\left(\frac{y - t}{\sqrt{V(\mu(t))}}\right) \frac{\mu'(t)}{\sqrt{V(\mu(t))}} - \nu(t) \, , \tag{14}$$

where $\nu(t) = \mathbb{E}\left[\psi\left((Y - t)/\sqrt{V(\mu(t))}\right) \mu'(t)/\sqrt{V(\mu(t))}\right]$ for $\mu(t) = \mathbb{E}(Y)$ and $V(\mu(t)) = \mathbb{V}(Y)$. The function $\nu(t)$ is a correction term introduced to obtain Fisher–consistent estimators, while ψ, that may be the Huber's ψ function, controls the Pearson residuals. For instance, in the particular case of the Gamma distribution, we have that the Pearson residuals are of the form $\sqrt{\alpha}\,(y - t)\,/t$, so

$$\Psi(y, t) = \psi\left(\sqrt{\alpha}\,((y - t)/t)\right) \sqrt{\alpha}/t - \nu(t)$$

where $\nu(t) = \mathbb{E}\left[\psi\left(\sqrt{\alpha}\,((Y - t)/t)\right) \sqrt{\alpha}/t\right]$.

In Section 5, we will apply these families of estimators to Poisson and Gamma regression models.

3 Consistency

We will begin by proving consistency results for both families of estimators in the case of fixed carriers. Analogous results were obtained for the case of random carriers in [1]. We will assume the following set of assumptions.

Conditions on the weight function

The conditions stated below are the usual assumptions required to the weights in nonparametric regression (see, for instance, [20]).

W1. (i) $\lim_{n \to \infty} \sum_{i=1}^n w_{ni}(\mathbf{x}) = 1$
(ii) $\lim_{n \to \infty} \sum_{i=1}^n |w_{ni}(\mathbf{x})| \mathbb{I}_{\{\|\mathbf{x}_i - \mathbf{x}\| > a\}} = 0$ for all $a > 0$
(iii) There exists $M > 0$ such that $\sum_{i=1}^n |w_{ni}(\mathbf{x})| \leq M$ for all $n \geq 1$.

W2. $\lim_{n\to\infty} \max_{1\le i\le n} |w_{ni}(\mathbf{x})| = 0$.

W3. $\lim_{n\to\infty} \left\{\max_{1\le i\le n} w_{ni}^2(\mathbf{x})\right\}\, n\log\log n = 0$.

W4. $\lim_{n\to\infty} \{\max_{1\le i\le n} |w_{ni}(\mathbf{x})|\} \log n = 0$.

W5. $w_{ni}(\mathbf{x}) \ge 0$ for all $1 \le i \le n$, $n \ge 1$.

Conditions on the loss function ρ

We will fix some extra notation. Given a compact set $\mathbf{C} \subset \tau$, let $\beta(\mathbf{C}) = \mathbb{E}\left[\inf_{t\notin\mathbf{C}} \rho(Y,t)\right]$ and $\beta_n(\mathbf{C}) = \sum_{i=1}^n w_{ni}(\mathbf{x}) \inf_{t\notin\mathbf{C}} \rho(Y_i,t)$. In the Appendix, it will be shown that Assumptions **W1**, **W2**, **A1**, **A2**, **A5** and **A7** imply that $\beta_n(\mathbf{C}) \xrightarrow{p} \beta(\mathbf{C})$.

The conditions for the loss function are quite similar to those given by [25]. However, some of them have been customized to our setting since the observations are not identically distributed.

A1. For each $t \in \tau$, $\rho(y,t)$ is a measurable function and $\rho(y,t)$ is separable in the sense of Doob.

A2. There exists a measurable function $a(y)$ such that $\mathbb{E}|\rho(Y,t) - a(Y)| < \infty$ for all $t \in \tau$. Then, $\gamma(t) = \mathbb{E}(\rho(Y,t) - a(Y))$ is well defined for each t, where $Y \sim F(\cdot, g(\mathbf{x}))$.

A3. The function ρ is almost surely lower semicontinuous in t.

A4. For all $t \ne g(\mathbf{x})$, $\gamma(t) > \gamma(g(\mathbf{x}))$.

A5. For all $t \in \tau$ we have that for some $p \ge 2$. $\sup_i \mathbb{E}|\rho(Y_i,t) - a(Y_i)|^p < M_1 = M_1(t) < \infty$.

A6. For all $t \in \tau$ we have that $\sup_i \mathbb{E}|\rho(Y_i,t) - a(Y_i)|^{2+s} \le M_2 = M_2(t) < \infty$ for some $s > 0$.

A7. For any set $\mathcal{U} \subset \tau$ the function $r(\mathbf{u}) = \int \inf_{t\in\mathcal{U}}[\rho(y,t) - a(y)]\, dF(y, g(\mathbf{u}))$ is continuous at $\mathbf{x}$.

A8. For any sequence of compact sets $\mathbf{C}_n$ converging to τ, $\liminf_{n\to\infty} \beta(\mathbf{C}_n) > \gamma(g(\mathbf{x}))$.

Without loss of generality, as in [25], we can assume $a(y) = 0$ and we have that assumptions **A2**, **A3** and **A4** imply

A4'. $\mathbb{E}\inf_{t\in\mathcal{U}} \rho(Y,t) \to \gamma(g(\mathbf{x}))$ as the neighbourhood $\mathcal{U}$ of t shrinks to $\{g(\mathbf{x})\}$.

Assumption **A5** may be relaxed by requiring $p > 1$, but additional constraints are needed. Besides, assumptions **A5** and **A6** are not restrictive due to the term $a(Y_i)$. For instance, by taking $\rho(y,t) = |y - t|$, $a(y) = |y|$ we include local medians even for Cauchy distributions.

Conditions on the function Ψ.

B1. For each $t \in \tau$, $\Psi(y,t)$ is a measurable function and $\Psi(y,t)$ is separable in the sense of Doob.

B2. The function Ψ is continuous almost everywhere in t.

B3. The expected value $\lambda(t) = \mathbb{E}\Psi(Y,t)$ exists for all $t \in \tau$ and it has a unique zero at $g(\mathbf{x})$. More generally, $\inf\{t : \lambda(t) < 0\} = \sup\{t : \lambda(t) > 0\}$.

B3'. The expected value $\lambda(t) = \mathbb{E}\Psi(Y,t)$ exists for all $t \in \tau$, $\lambda(g(\mathbf{x})) = 0$ and there exists a neighbourhood of $g(\mathbf{x})$ in which $\lambda(t)$ has a unique change of sign.

B4. There exists a continuous function $b : \mathbb{R} \to \mathbb{R}$ bounded away from 0 in τ, i.e., $b(t) \geq b_0 > 0$ for all $t \in \tau$ such that
(i) $\mathbb{E}\sup_{t\in\tau}[|\Psi(Y,t)|/b(t)] < \infty$,
(ii) $\liminf_{t\to\infty}[|\lambda(t)|/b(t)] \geq 1$,
(iii) $\mathbb{E}\limsup_{t\to\infty}[|\Psi(Y,t) - \lambda(t)|/b(t)] < 1$,
where if τ is not compact, ∞ denotes the point at infinity in its one–point compactification.

B5. $\Psi(y,t)$ is a bounded function such that
(i) for each fixed $t \in \tau$, $\Psi(\cdot,t)$ is of bounded variation,
(ii) $\lambda(t)$ is strictly monotone in a neighbourhood of $g(\mathbf{x})$.

B6. There exist $s > 0$ and $M > 0$ such that
(i) $\sup_{1\leq i\leq n}\mathbb{E}|\Psi(Y_i,t)|^{2+s} \leq M$.
(ii) For any compact set $\mathbf{C} \subset \tau$, $\sup_{1\leq i\leq n}\mathbb{E}\left\{\sup_{t\notin\mathbf{C}}(|\Psi(Y_i,t) - \lambda(t)|/b(t))\right\}^{2+s} \leq M$.
(iii) For any neighbourhood $U \subset \tau$ of t and for all $n \geq 1$

$$\sup_{1\leq i\leq n}\mathbb{E}\left[\sup_{t'\in U}|\Psi(Y_i,t) - \Psi(Y_i,t')|\right]^{2+s} \leq M\,.$$

B7. For a compact set $\mathbf{C}$ and a neighbourhood $\mathcal{U}$ of t, the functions $r_1(\mathbf{u}) = \int \sup_{t\notin\mathbf{C}}[|\Psi(y,t) - \lambda(t)|/b(t)]\,dF(y,g(\mathbf{u}))$, $r_2(\mathbf{u},t) = \int \Psi(y,t)dF(y,g(\mathbf{u}))$ and $r_3(\mathbf{u}) = \int \sup_{t'\in\mathcal{U}}|\Psi(y,t) - \Psi(y,t')|dF(y,g(\mathbf{u}))$ are continuous at $\mathbf{x}$.

B8. $F(y,g(\cdot))$ is continuous at $\mathbf{x}$ for each fixed y.

It is worth noticing that when τ is compact, (ii) and (iii) in **B4** are not necessary. As noted by [25] if there is a function b verifying **B4**, one may choose $b(t) = \max\{|\lambda(t)|, b_0\}$. If Ψ is a bounded function, as is usual with $M-$estimates, assumptions **B4** and **B6** are obviously fulfilled. However, we set the assumptions in a more general way in order to include, for instance, regression models with simultaneous scale estimation, or exponential models with location and scale functions. In this case, $\tau \subset \mathbb{R}^2$ and the assumptions must be extended to the multidimensional case in order to obtain the conclusions of the Theorems.

Lemma 3.1. *Let $g_n(\mathbf{x})$ be any value satisfying (3). Assume that **W1**, **W2**, **W5**, **A1**, **A2**, **A7** and **A8** hold. If in addition,*

*i) **W3** and **A6** hold*
or

*ii) **W4** holds and ρ is bounded,*

then, there exists a compact set $\mathbf{K} \subset \tau$ *such that* $\lim_{m\to\infty} \mathbb{P}\left(\bigcap_{n\geq m} g_n(\mathbf{x}) \in \mathbf{K}\right) = 1$.

Note that even if assumption **A8** seems very restrictive, the conclusion of Lemma 3.1 can be directly verified for many families of estimates including, for instance, local medians.

Theorem 3.1. *Under* ***A1*** *to* ***A4***, ***A6***, ***A7***, ***W1*** *to* ***W3***, ***W5*** *and if the conclusion of Lemma 3.1 holds, we have that* $g_n(\mathbf{x}) \xrightarrow{a.s.} g(\mathbf{x})$ *as* $n \to \infty$.

Remark 3.1. If ρ is bounded we can replace **W3** by **W4**, in Theorem 3.1. On the other hand, if we are looking for convergence in probability we can replace **A6** and **W3** by **A5** with $p \geq 2$ and the conclusion of Lemma 3.1 by the assumption that there exists a compact set $\mathbf{K} \subset \tau$ such that $\lim_{n\to\infty} \mathbb{P}(g_n(\mathbf{x}) \in \mathbf{K}) = 1$.

The following Theorem can be proved as in [7] using Proposition 6.1 which can be found in the Appendix.

Theorem 3.2. *Assume* ***W1***, ***W2***, ***W4***, ***B1***, ***B2***, ***B3'***, ***B5*** *and* ***B8***, *then there exists a solution,* $g_n(\mathbf{x})$, *of (4) such that* $g_n(\mathbf{x}) \xrightarrow{a.s.} g(\mathbf{x})$ *as* $n \to \infty$.

Remark 3.2. If $\Psi(y, \cdot)$ is monotone, any solution of (4) will converge to $g(\mathbf{x})$.

We now state consistency results for the case of a unique solution. These results can be extended to the multiparameter situation.

Theorem 3.3. *Under* ***W1***, ***W3***, ***B1*** *to* ***B3***, ***B4***, ***B6*** *and* ***B7*** *we have that* $g_n(\mathbf{x}) \xrightarrow{a.s.} g(\mathbf{x})$ *almost everywhere as* $n \to \infty$.

Remark 3.3. Note that if Ψ is bounded **W3** can be replaced by **W4**.

4 Asymptotic Normality

In this section, we derive the asymptotic distribution of the proposed estimators. In order to study the asymptotic behaviour, we will need some additional assumptions. First, we introduce some notation. Denote by $\Psi'(y,t) = \partial\Psi(y,t)/\partial t$, $\Psi''(y,t) = \partial^2\Psi(y,t)/\partial t^2$, $\lambda_1(t) = \mathbb{E}(\Psi'(Y,t))$ and $c_n = \sum_{i=1}^n w_{ni}^2(\mathbf{x})$. In most situations, $\lambda_1(t) = \partial\lambda(t)/\partial t$.

N1. Ψ is twice continuously differentiable in t and for any neighbourhood $\mathcal{U}$ of $g(\mathbf{x})$, there exists $c > 0$ such that
i) $\sup_{t\in\mathcal{U}} |\Psi''(y,t)| \leq c$ for all y
or
ii) $\sup_{1\leq i\leq n} \mathbb{E}\sup_{t\in\mathcal{U}} |\Psi''(Y_i,t)| \leq c$ for all $n \geq 1$

N2. $\lambda_1(g(\mathbf{x})) \neq 0$.

N3. For some positive constant c, $\sup_{1\leq i\leq n}\mathbb{E}|\Psi'(Y_i,g(\mathbf{x}))|^2\leq c$ for all $n\geq 1$.

N4. Ψ' satisfies one of the following conditions
i) $r_4(\mathbf{u})=\int\Psi'(y,t)dF(y,g(\mathbf{u}))$ is continuous at $\mathbf{x}$ for each fixed $t\in\tau$.
ii) For each $t\in\tau$, $\Psi'(\cdot,t)$ is of bounded variation.

Theorem 4.1. *Let r_2 be the function defined in **B7**. Assume that $g_n(\mathbf{x})\xrightarrow{p}g(\mathbf{x})$ and **W1**, **W2**, **N1** to **N4** hold. If in addition, $\lim_{n\to\infty}c_n^{-1/2}\max|w_{ni}(\mathbf{x})|=0$ and $\lim_{n\to\infty}c_n^{-1/2}\sum_{i=1}^n w_{ni}(\mathbf{x})$ $r_2(\mathbf{x}_i,g(\mathbf{x}_i))=\beta$, we have that*

$$c_n^{-1/2}(g_n(\mathbf{x})-g(\mathbf{x}))\xrightarrow{w}N(\beta_1,\sigma_1^2(\mathbf{x}))\ ,$$

where $\beta_1=\beta/\lambda_1(g(\mathbf{x}))$ and $\sigma_1^2(\mathbf{x})=\sigma^2(\mathbf{x})/[\lambda_1(g(\mathbf{x}))]^2$ with $\sigma^2(\mathbf{u})=\int\Psi^2(y,g(\mathbf{x}))dF(y,g(\mathbf{u}))$ and $\xrightarrow{w}$ stands for convergence in law.

Remark 4.1. Even though the asymptotic bias seems to depend on the Ψ function, it can be seen that, under **N1** and **N4(i)**, it performs as in the usual nonparametric regression model, if $\lim_{n\to\infty}c_n^{-1/2}\sum_{i=1}^n w_{ni}(\mathbf{x})\mathbb{I}_{\{\|\mathbf{x}_i-\mathbf{x}\|>a\}}=0$. That is, the asymptotic bias equals $\lim_{n\to\infty}c_n^{-1/2}\sum_{i=1}^n w_{ni}(\mathbf{x})[g(\mathbf{x}_i)-g(\mathbf{x})]$. Assumption **N1** requires differentiability of the Ψ function which in many situations is not satisfied (for instance, if we take Ψ as the Huber or the sign function). This assumption can be relaxed, as is usual for $M-$estimators, by requiring that the finite discontinuities of the derivative are continuity points of $F(\cdot,g(\mathbf{x}))$. The proof is essentially the same with some technical modifications.

We will now give an additional result of asymptotic normality which holds only if $\tau\subset\mathbb{R}$, but does not require smoothness conditions on Ψ. We will obtain it under the following assumptions

W6. $\sum_{i=1}^n w_{ni}(\mathbf{x})=1$.

W7. There exists a constant $M>0$ such that $c_n^{-1/2}\sum_{i=1}^n|w_{ni}(\mathbf{x})|\,|g(\mathbf{x}_i)-g(\mathbf{x})|\leq M$ for all $n\geq 1$.

N5. $\lambda'(g(\mathbf{x}))=\partial\lambda(t)/\partial t|_{t=g(\mathbf{x})}\neq 0$.

N6. $F(y,t)$ is Lipschitz as a function of t uniformly in y, i.e., there exists $L>0$ such that $|F(y,t)-F(y,t')|\leq L|t-t'|$ for all $y\in\mathbb{R}$, $t,t'\in\tau$.

Remark 4.2. The Lipschitz condition required in **N6** is easily verified for the most of the standard distributional families.

Theorem 4.2. *Under **B2**, **B5(i)**, **W6**, **W7**, **N5**, **N6** we have that $c_n^{-1/2}(g_n(\mathbf{x})-g(\mathbf{x}))$ has the same asymptotic distribution as $c_n^{-1/2}\sum_{i=1}^n w_{ni}(\mathbf{x})\Psi(Y_i,g(\mathbf{x}))/\lambda'(g(\mathbf{x}))$, provided that $\lim_{t\to g(\mathbf{x})}\|\Psi(\cdot,t)-\Psi(\cdot,g(\mathbf{x}))\|_V=0$, where $\|.\|_V$ stands for the variation norm.*

5 Monte Carlo Study

5.1 Simulation

A Monte Carlo study was carried out in order to assess the performance of the proposed robust estimators for finite contaminated and non–contaminated samples. We considered a Poisson and a Gamma regression model with one-dimensional covariate x. In both situations, we performed $N = 5000$ replications and we generated, at each replication, $n = 100$ observations, $Y_1, \ldots, Y_n$, such that $Y_i \sim F(\cdot, g(x_i))$ with $x_i = (i - 0.5)/100$. To avoid boundary effects, we considered the weights introduced in [19] using the Epanechnikov kernel with bandwidth h. Boundary kernels were considered to improve the performance of the regression estimators. We considered three regression estimators of the regression function g

$\widehat{g}_{\mathrm{CL}}$: the classical estimator,

$\widehat{g}_{\mathrm{RD}}$: the robustified version of the deviance estimator obtained by minimizing (1) using (12), with ϕ the loss function considered in [14] with tuning constant $d = 2$,

$\widehat{g}_{\mathrm{RQL}}$: the robustified quasi-likelihood obtained as solution of (4) using (14) with $\psi = \psi_c$ the Huber' s score function with $c = 1.6$.

For each generic estimator $\widehat{g}$ we computed the following two summary measures

$$\mathrm{MSE}_j(\widehat{g}, g) = \frac{1}{n}\sum_{i=1}^{n}\left(\widehat{g}(x_i) - g(x_i)\right)^2 \tag{15}$$

$$\mathrm{AMSE}(\widehat{g}, g) = \frac{1}{N}\sum_{j=1}^{N}\mathrm{MSE}_j(\widehat{g}, g)\,, \tag{16}$$

which measure the square error in the $j-$th replication and the square error along all the replications, respectively. For each robust estimators $\widehat{g}$, we compared the behaviour of the estimator with respect of the classical one by computing a summary measure of the efficiency given by

$$\mathrm{EF}(\widehat{g}, \widehat{g}_{\mathrm{CL}}) = \frac{\mathrm{AMSE}(\widehat{g}_{\mathrm{CL}}, g)}{\mathrm{AMSE}(\widehat{g}, g)}\,.$$

5.1.1 Poisson regression

As mentioned above, we generated observations Y_i following a Poisson distribution $\mathcal{P}(g(x_i))$ with $g(x) = \exp(2\sin 4\pi x) + 1$. We contaminated the original samples by replacing $m = 1$ and $m = 3$ responses by arbitrary values $y^* = 0$, 12 or 20. These contaminating data were located at fixed positions of the covariate x and when $m = 3$ the outliers were located at successive values of x, so as to obtain a more severe contamination pattern. We chose $h = 0.05$ as smoothing parameter.

To illustrate the fit of the considered regression estimators, Figure 1 presents the different estimators applied to one simulated sample in its original version and when $m = 3$ responses are

replaced by $y^* = 20$. The black line corresponds to the true function g, while the blue and the red lines correspond to the classical and robust estimators, respectively. The plots on the right side show that the robust and the classical estimators overlap in the non–contaminated sample. However, under contamination, the proposed estimators are very close to the classical estimator in all the range of the covariate, except for the region where the outlying responses were located. In this crucial area the classical estimator $\widehat{g}_{\mathrm{CL}}$ suffers from the effect of the outliers leading to a fit far from the expected curve g. Indeed, in presence of the three outliers, $\widehat{g}_{\mathrm{CL}}$ is pulled up by the anomalous observations deviating from the true curve g. Instead, the proposed estimators remain more stable, especially $\widehat{g}_{\mathrm{RD}}$ which looks much more insensitive than $\widehat{g}_{\mathrm{RQL}}$ to the severe outlying points. These conclusions can be extended to the whole simulation study whose results are summarized in Table 1. In fact, when there are no outliers ($m = 0$) the proposed estimators achieve a 95% efficiency with respect to the classical estimator. On the contrary, for the contaminated samples the AMSE of $\widehat{g}_{\mathrm{CL}}$ increases according to the size of the outlier y^* and also to m. The proposed estimators, in particular $\widehat{g}_{\mathrm{RD}}$, are much more stable in presence of anomalous data. This behaviour is also observed in Figure 2 that shows the density estimates of the mean square error MSE of the computed estimators.

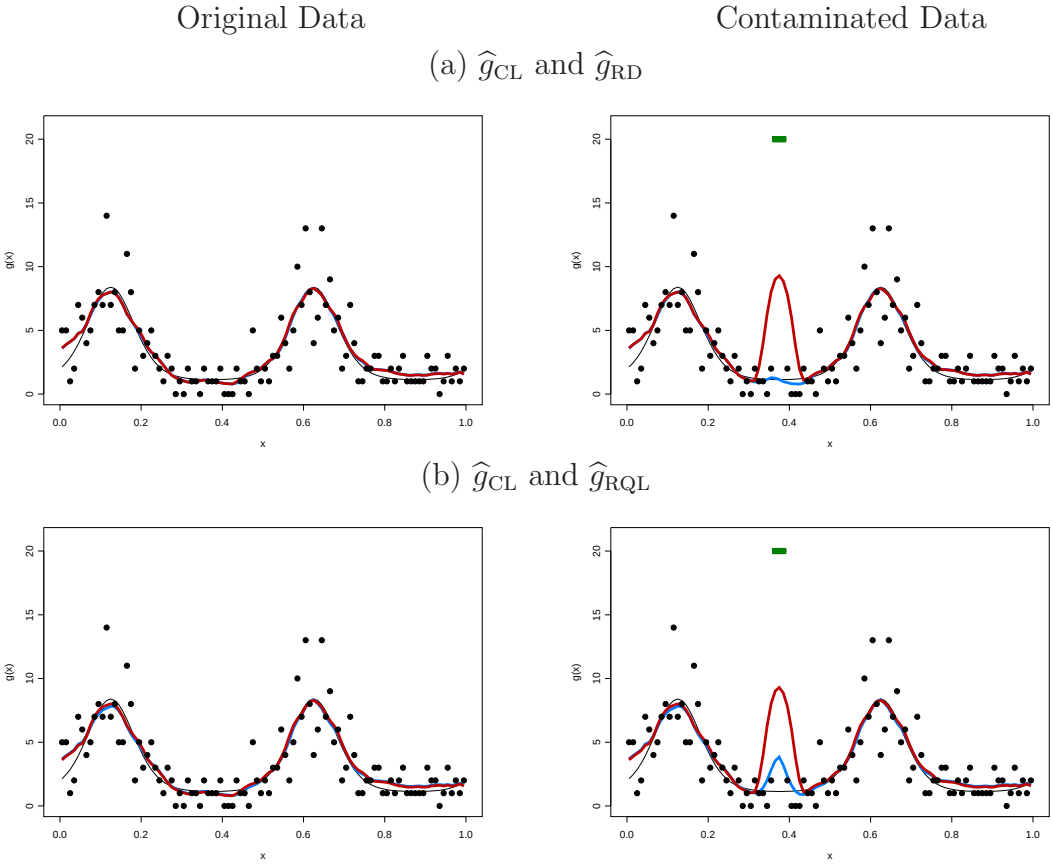

Figure 1: Simulated data from Poisson regression. The plots on the left correspond to original samples, while plots on the right to contaminated ones with $m = 3$ outlying observations with $y^* = 20$. The black lines correspond to the true function $g(x)$, the red one to the classical estimator $\widehat{g}_{\mathrm{CL}}$ while the blue one to the robust estimators $\widehat{g}_{\mathrm{RD}}$ and $\widehat{g}_{\mathrm{RQL}}$ in (a) and (b), respectively. Green points represent the outliers.

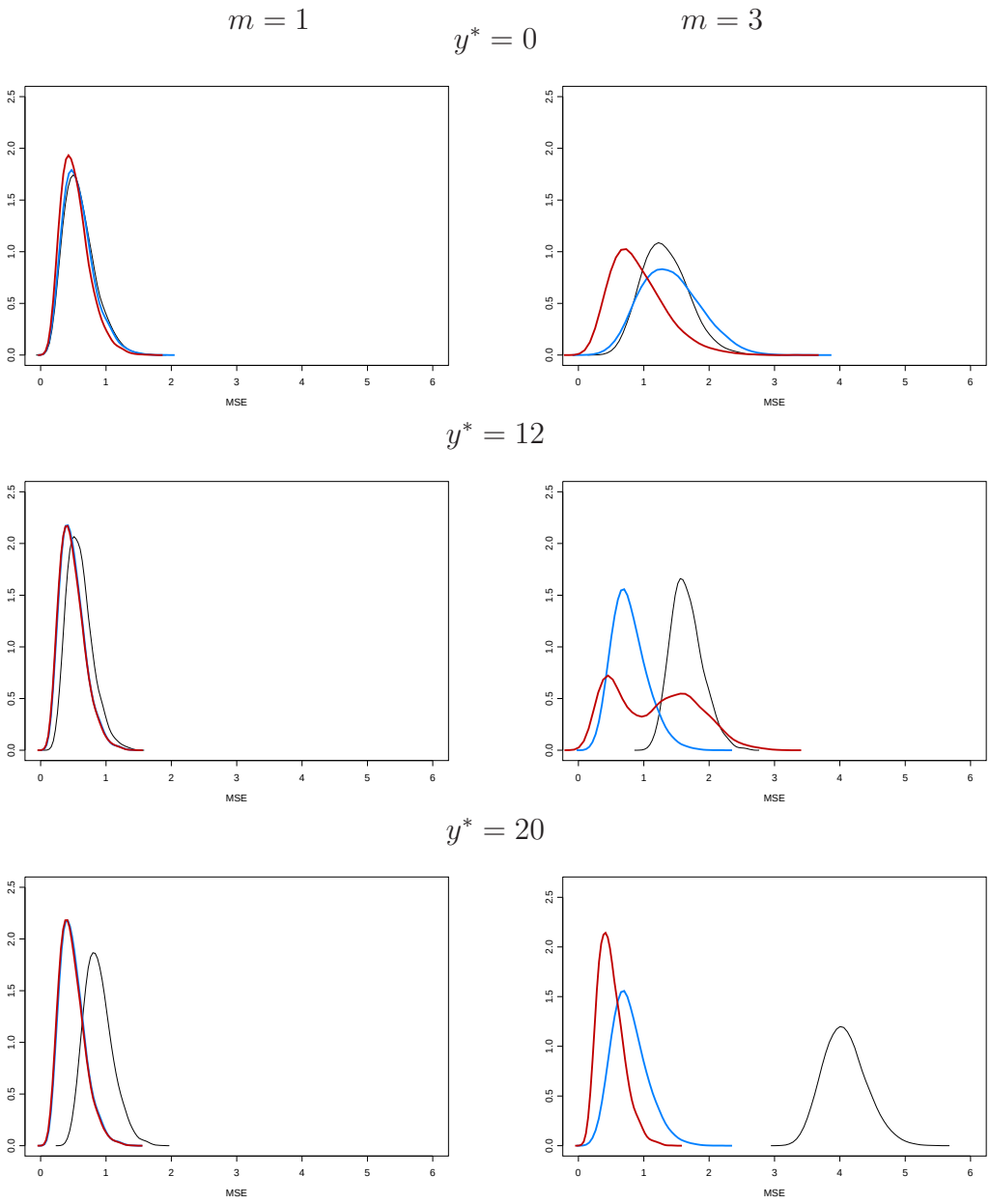

Figure 2: Density estimates of the mean square error MSE over $N = 5000$ Poisson samples with $m = 1$ and $m = 3$ outliers. The black line correspond to the classical estimators $\widehat{g}_{\mathrm{CL}}$, while the blue and red ones to $\widehat{g}_{\mathrm{RD}}$ and $\widehat{g}_{\mathrm{RQL}}$, respectively.

	$y^* = 0$		$y^* = 12$		$y^* = 20$	
m	$\widehat{g}_{\mathrm{RD}}$	$\widehat{g}_{\mathrm{RQL}}$	$\widehat{g}_{\mathrm{RD}}$	$\widehat{g}_{\mathrm{RQL}}$	$\widehat{g}_{\mathrm{RD}}$	$\widehat{g}_{\mathrm{RQL}}$
0	0.954	0.954	0.954	0.954	0.954	0.954
1	1.123	1.032	1.232	1.216	1.831	1.794
3	1.447	0.944	1.426	2.115	8.229	5.198

Table 1: Poisson regression. Values of $\mathrm{EF}(\widehat{g}, \widehat{g}_{\mathrm{CL}})$ for contaminated and non–contaminated samples

5.1.2 Gamma regression

In this section, we report the results obtained under Gamma regression model. We generated $n = 100$ observations $Y_i \sim \Gamma(16, 8(\sin 4\pi x_i + 3)^{-1})$, and so $\mathbb{E}(Y_i) = g(x_i) = 2(\sin 4\pi x_i + 3)$, $1 \leq i \leq n$. We contaminated the original samples by replacing $m = 1$ and $m = 3$ responses by arbitrary values $y^* = 0$. As before, these outliers were located at fixed positions of covariate x and when $m = 3$ they were located at successive values of x. We choose $h = 0.1$ as smoothing parameter.

As an example, in Figure 3 we plot the classical estimator $\widehat{g}_{\mathrm{CL}}$ and the robust estimator $\widehat{g}_{\mathrm{RD}}$ when they are applied to one simulated sample without outliers and with $m = 1$ and $m = 3$ outliers. As above, the lines in black correspond to the true regression function $g(x)$, while those in red and blue to the classical estimator $\widehat{g}_{\mathrm{CL}}$ and to the robust estimator $\widehat{g}_{\mathrm{RD}}$, respectively. The robust and the classical estimators again overlap in the non–contaminated case, but the classical estimator seems to be affected even by just one outlier since it deviates form the true regression function g in presence of just one outlying observation. Instead, the fit obtained from $\widehat{g}_{\mathrm{RD}}$ is very stable in all the range in the three cases considered.

Table 2 summarizes the results over the $N = 5000$ samples. In fact, when there are no outliers ($m = 0$) the proposed estimators are very efficient with respect to the classical estimator. Under contamination the AMSE of $\widehat{g}_{\mathrm{CL}}$ increases with m, while the proposed estimators resist the presence of outliers. Besides, we conclude that when three values of the response variable are replaced by outliers, $\widehat{g}_{\mathrm{RD}}$ is more stable than $\widehat{g}_{\mathrm{RQL}}$, since $\mathrm{AMSE}(\widehat{g}_{\mathrm{RQL}}, g)$ is around 1.6 times the $\mathrm{AMSE}(\widehat{g}_{\mathrm{RD}}, g)$.

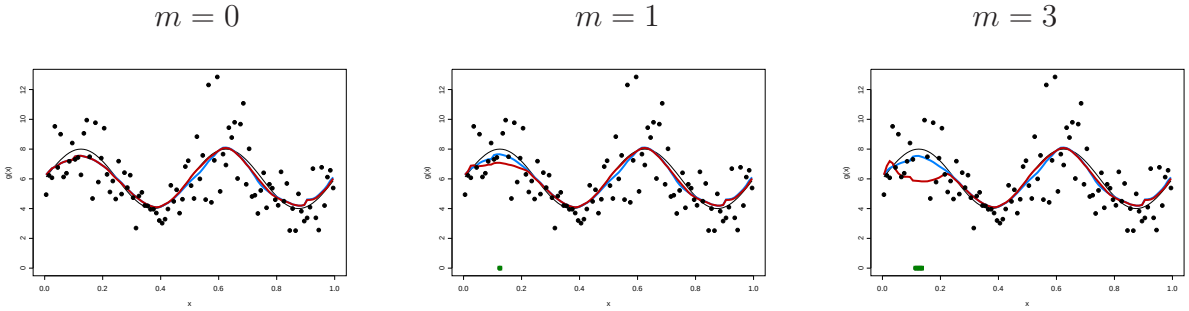

Figure 3: Simulated data from Gamma regression. Samples with m outyling responses. The black line correspond to the true function $g(x)$, while the red and blue ones to the classical, $\widehat{g}_{\mathrm{CL}}$, and robust, $\widehat{g}_{\mathrm{RD}}$, estimators. Green points represent the outliers.

y^*	0	
m	$\widehat{g}_{\mathrm{RD}}$	$\widehat{g}_{\mathrm{RQL}}$
0	0.976	0.976
1	1.272	1.147
3	2.666	1.652

Table 2: Gamma regression. Values of $\mathrm{EF}(\widehat{g}, \widehat{g}_{\mathrm{CL}})$ for contaminated and non–contaminated samples.

5.2 Resistant choice of the smoothing parameter

As in any smoothing procedure the selection of the bandwidth parameter is an important task. Classical procedures for the choice of the smoothing parameter, such as cross–validation or plug–in methods, may be very sensitive to the presence of outliers. This sensitivity has been discussed in the literature of nonparametric regression; among others, we can mention [28], [27], [33], [6] and [9]. Least squares cross-validation method may be severely affected by outliers, even when the nonparametric regression estimators are based on local $M-$estimators and this is due to the fact that it is based on an L^2-norm. One outlier may cause the bandwidth to break down, in the sense that it may result in oversmoothing or undersmoothing. When a small bandwidth is considered, few outlying responses with similar covariates x_i could damage the estimate seriously. [5] pointed out that robust cross-validation methods should be an alternative. In the following, we describe a resistant cross–validation procedure based on robustified deviances

1. For each given bandwidth h compute

$$\widehat{y}_h^{-i} = \underset{t}{\operatorname{argmin}} \sum_{j \neq i}^{n} w_{nj}(x_i, h) \rho(y_j, t)\,,$$

   where ρ is taken as in (1) and the weights are given by

$$w_{nj}(x, h) = \left\{ \sum_{\ell \neq i} K(x - x_\ell/h) \right\}^{-1} K(x - x_j/h)\,.$$

2. Choose the robust bandwidth as

$$\widehat{h}_{n,\mathrm{R}} = \underset{h}{\operatorname{argmin}} \sum_{i=1}^{n} \rho(y_i, \widehat{y}_h^{-i})\,.$$

To study the performance of the CV procedure, we carried out a Monte Carlo study for the case of the Gamma regression model. We also intend to compare it with the classical cross–validation method based on the deviance. As in Section 5.1.2, we generated $n = 100$ observations $Y_i \sim \Gamma(16, 8(\sin 4\pi x_i + 3)^{-1})$ and so the regression function is $g(x_i) = 2(\sin 4\pi x_i + 3)$ in the non–contaminated samples. We followed the contamination scheme with $m = 3$ outliers described in Section 5.1.2. For each non–contaminated sample, we computed the classical bandwidth $\widehat{h}_n$ and the

resistant one $\widehat{h}_{n,\mathrm{R}}$. For the contaminated samples, we performed the same computations obtaining classical and resistant bandwidths denoted $\widehat{h}_n^c$ and $\widehat{h}_{n,\mathrm{R}}^c$, respectively. We replicated $N = 500$ times. Figure 4 displays the histograms for the differences $\widehat{h}_n - \widehat{h}_n^c$ and $\widehat{h}_{n,\mathrm{R}} - \widehat{h}_{n,\mathrm{R}}^c$. These histograms show that in most cases the windows achieved by the resistant method in contaminated samples are very similar to those obtained by the same method in the corresponding non–contaminated samples. On the contrary, the classical selection procedure seems to be much more unstable.

In order to assess the performance of the regression estimates, we computed the mean square error defined in (15) for each estimator. Figure 5 shows, with a solid line, the density estimator of the mean squared errors for the original samples using the cross validation bandwidths and with a dashed line those obtained for the contaminated samples. As we can see, the mean squared errors achieved using the resistant bandwidth parameters for contaminated and non–contaminated samples are comparable, while those obtained using the classical cross–validation method are larger when the samples have outliers.

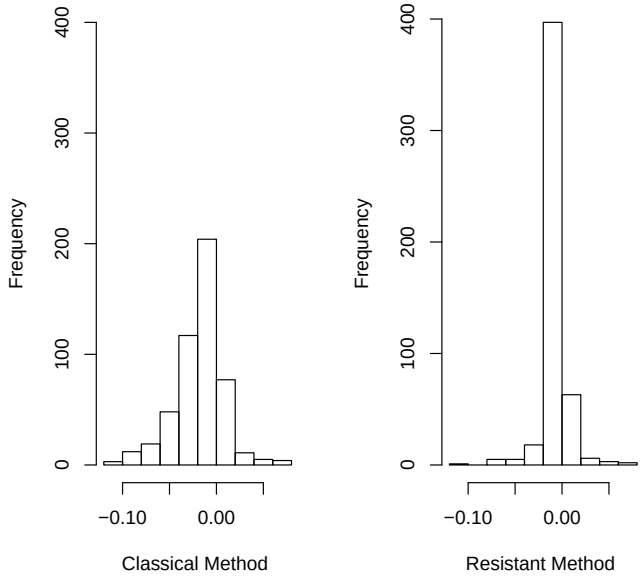

Figure 4: Histograms of the differences $\widehat{h}_n - \widehat{h}_n^c$ on the left and $\widehat{h}_{n,\mathrm{R}} - \widehat{h}_{n,\mathrm{R}}^c$ on the right.

6 Appendix

6.1 Proofs of the results in Section 3.

In order to prove Lemma 3.1 we will first establish some auxiliary results.

Lemma 6.1. *Let* $\{Z_j : j \geq 1\}$ *be independent random variables such that* $Z_j \sim F(z, \mathbf{x}_j)$

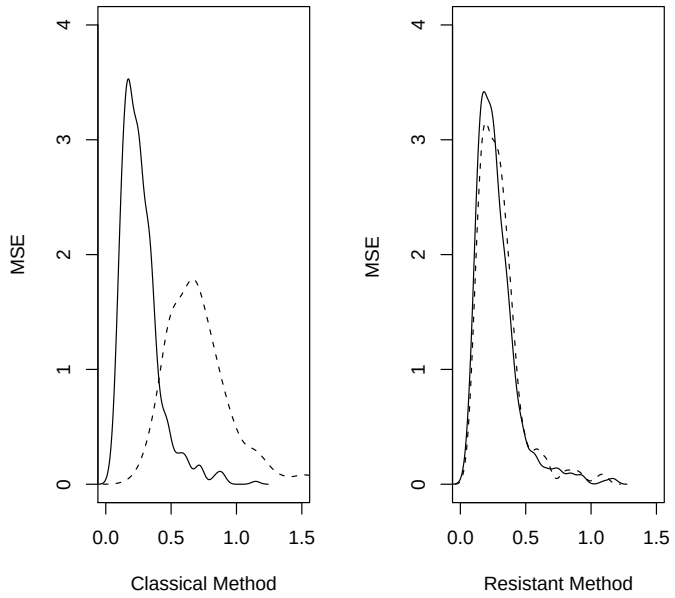

Figure 5: Density estimator of the mean squared errors for the original samples using the cross validation bandwidths. The solid line correspond to those computed with the original samples while the dashed line to those obtained for the contaminated samples.

and $\mathbb{E}(Z_j) = r(\mathbf{x}_j)$. *Assume that* r *is continuous at* $\mathbf{x}$ *and that there exists* $\mathbf{C} > 0$ *such that* $\sup_{1\leq i\leq n} \mathbb{E}|Z_j|^p \leq \mathbf{C}$ *for some* $p \geq 2$. *Let* $r_n(\mathbf{x}) = \sum_{i=1}^n w_{ni}(\mathbf{x}) Z_i$. *Then, if* ***W1*** *and* ***W2*** *hold, we have that* $\lim_{n\to\infty} \mathbb{E}\left[|r_n(\mathbf{x}) - r(\mathbf{x})|^p\right] = 0$.

Proof. By Minkowsky's inequality, we have that

$$\mathbb{E}|r_n(\mathbf{x}) - r(\mathbf{x})|^p \leq \left\{ \left[\mathbb{E}|r_n(\mathbf{x}) - \mathbb{E}r_n(\mathbf{x})|^p\right]^{1/p} + |\mathbb{E}r_n(\mathbf{x}) - r(\mathbf{x})| \right\}^p .$$

Thus, using that $\mathbb{E}r_n(\mathbf{x})$ converges to $r(\mathbf{x})$, it is enough to show that $\mathbb{E}|r_n(\mathbf{x}) - \mathbb{E}r_n(\mathbf{x})|^p$ converges to 0.

Let $\varepsilon_i = Z_i - \mathbb{E}Z_i$ then $\sup_i \mathbb{E}|\varepsilon_i|^p \leq 2^{p-1}\mathbf{C} = M_1$. The inequality given in [29] (see also [15]) implies that there exists a constant $A > 0$ depending only upon p such that

$$\begin{aligned} \mathbb{E}|r_n(\mathbf{x}) - \mathbb{E}r_n(\mathbf{x})|^p &\leq A\,\mathbb{E}\left|\sum_{i=1}^n w_{ni}^2(\mathbf{x})\varepsilon_i^2\right|^{p/2} \leq A\left[\max|w_{ni}(\mathbf{x})|\right]^{p/2}\left[\sum_{i=1}^n w_{ni}(\mathbf{x})\left(\mathbb{E}|\varepsilon_i|^p\right)^{2/p}\right]^{p/2} \\ &\leq M_1\, AM^{p/2}\left(\max|w_{ni}(\mathbf{x})|\right)^{p/2} \end{aligned}$$

from Minkowsky's inequality and **W1**. Then,**W2** entails now the desired result. □

Denote by $\gamma_n(\mathcal{U}) = \sum_{i=1}^n w_{ni}(\mathbf{x}) \inf_{t\in\mathcal{U}} \rho(Y_i, t)$ and $\gamma(\mathcal{U}) = \mathbb{E}\left(\inf_{t\in\mathcal{U}} \rho(Y, t)\right)$ where $\mathcal{U} \subset \tau$.

Lemma 6.2. *Under* ***W1***, ***W2***, ***A1***, ***A2*** *and* ***A7*** *we have that for any* $t \in \tau$ *and any set* $\mathcal{U} \subset \tau$

(i) $\gamma_n(t) \xrightarrow{p} \gamma(t)$ *and* $\gamma_n(\mathcal{U}) \xrightarrow{p} \gamma(\mathcal{U})$, *if, in addition,* ***A5*** *holds.*

(ii) $\gamma_n(t) \xrightarrow{a.s.} \gamma(t)$ *and* $\gamma_n(\mathcal{U}) \xrightarrow{a.s.} \gamma(\mathcal{U})$, *if* ***A6*** *and* ***W3*** *hold.*

PROOF. We will just prove the desired results for a given set $\mathcal{U}$. Denote by $Z_i = \inf_{t\in\mathcal{U}} \rho(Y_i, t)$, then we have

$$\gamma_n(\mathcal{U}) - \gamma(\mathcal{U}) = \sum_{i=1}^n w_{ni}(\mathbf{x})[Z_i - \mathbb{E}(Z_i)] + \sum_{i=1}^n w_{ni}(\mathbf{x})[\mathbb{E}(Z_i) - \gamma(\mathcal{U})] + \left[\sum_{i=1}^n w_{ni}(\mathbf{x}) - 1\right]\gamma(\mathcal{U}) \ . \quad (17)$$

(i) Since $\mathbb{E}(Z_i) = r(\mathbf{x}_i)$ and $\gamma(\mathcal{U}) = r(\mathbf{x})$ where r, given in **A7**, is continuous, **A5** and **W1** imply that the last two terms in (17) converge to 0. Thus, from Lemma 6.1, we get the desired result.

(ii) follows easily from Theorem 4 of [20]. □

PROOF OF LEMMA 3.1. The proof is quite similar to that given in Lemma 1 of [25]. Indeed, assumption **A8** implies that given $\varepsilon > 0$ there exists a compact set $\mathbf{C}$ such that $\beta(\mathbf{C}) > \gamma(g(\mathbf{x})) + \varepsilon$.

Denote by $A_n = \{\sum_{i=1}^n w_{ni}(\mathbf{x}) \inf_{t\notin\mathbf{C}} \rho(Y_i, t) > \beta(\mathbf{C}) - \varepsilon/2 > \gamma(g(\mathbf{x})) + \varepsilon/2\}$. Then, from Lemma 6.2 $\lim_{m\to\infty} \mathbb{P}\left(\bigcap_{n\geq m} A_n\right) = 1$. Thus, we can find $n_0 \in \mathbb{N}$ such that $\mathbb{P}\left(\bigcap_{n\geq n_0} A_n\right) > 1-\varepsilon/2$, which implies, together with **A5**, that

$$\mathbb{P}\left(\bigcap_{n\geq n_0} \inf_{t\notin\mathbf{C}} \gamma_n(t) > \gamma(g(\mathbf{x})) + \frac{\varepsilon}{2}\right) > 1 - \frac{\varepsilon}{2} \ .$$

Now the proof follows as in [25] using Lemma 6.2 instead of the strong law of large numbers. □

PROOF OF THEOREM 3.1. The proof is similar to that of Theorem 1 in [25] using again Lemma 6.2. □

Proposition 6.1. *Under* ***B1***, ***B2***, ***B5***, ***B8***, ***W1***, ***W2*** *and* ***W4***, *we have that for all* $t \in \tau$, $\lambda_n(t) \xrightarrow{a.s.} \lambda(t)$.

PROOF. **B5(i)** entails that $|\lambda_n(t) - \lambda(t)| \leq \|\Psi(\cdot, t)\|_V \, \|F_n - F\|_\infty$. On the other hand, by **B8**, Theorem 6 of [20] and a similar argument to the proof of Glivenko Cantelli's Theorem we get that $\|F_n - F\|_\infty = \sup_y |F_n(y) - F(y, g(\mathbf{x}))|$ converges to 0 almost surely as $n \to \infty$ where $F_n(y) = \sum_{i=1}^n w_{ni}(\mathbf{x}) \mathbb{I}_{(-\infty,y]}(Y_i)$. This concludes the proof. □

In order to obtain consistency results for a multidimensional parameter or without requiring **B5**, we will need the following Lemma.

Lemma 6.3. *Under* ***B1*** *to* ***B3***, ***B6*** *and* ***B7***, ***W1*** *and* ***W3*** , *given* $t_0 \in \tau$, *a compact set* $\mathbf{C}$ *and an open set* $\mathcal{U}$, *we have that* $\lambda_n(t_0) \xrightarrow{a.s.} \lambda(t_0)$ *and besides,*

$$\sum_{i=1}^n w_{ni}(\mathbf{x}) \sup_{t\notin\mathbf{C}} [|\Psi(Y_i, t) - \lambda(t)|/b(t)] \xrightarrow{a.s.} \mathbb{E}\left\{\sup_{t\notin\mathbf{C}} [|\Psi(Y, t) - \lambda(t)|/b(t)]\right\}$$

$$\sum_{i=1}^n w_{ni}(\mathbf{x}) \sup_{t\in\mathcal{U}} |\Psi(Y_i, t) - \Psi(Y_i, t_0)| \xrightarrow{a.s.} \mathbb{E}\left[\sup_{t\in\mathcal{U}} |\Psi(Y, t) - \Psi(Y, t_0)|\right] \ .$$

PROOF. Follows easily from Theorem 4 of [20].

Proposition 6.2. *If **W1**, **W3**, **B1** to **B4**, **B6** and **B7** hold, we have that there exists a compact set* $\mathbf{K} \subset \tau$ *such that*

$$\lim_{m\to\infty} \mathbb{P}\left(\bigcap_{n\geq m} g_n(\mathbf{x}) \in \mathbf{K} \right) = 1 \, .$$

PROOF. As in [25] denote by $v(y, \mathbf{C}) = \sup_{t\notin\mathbf{C}} |\Psi(y,t) - \lambda(t)|/b(t)$. Given $0 < \varepsilon < (2M)^{-1}$, where M is given in **W1**, **B4** and the dominated convergence theorem entail that there exists a compact set $\mathbf{C} \subset \tau$ such that

$$\mathbb{E}v(Y, \mathbf{C}) < 1 - 3\varepsilon \quad \text{and} \quad \inf_{t\notin\mathbf{C}} \frac{|\lambda(t)|}{b(t)} > 1 - \varepsilon \, . \tag{18}$$

Denote by $\widetilde{w}_{ni}(\mathbf{x}) = |w_{ni}(\mathbf{x})|\{\sum_{i=1}^n |w_{ni}(\mathbf{x})|\}^{-1}$. Then $\widetilde{w}_{ni}(\mathbf{x})$ are probability weight functions satisfying **W1** and **W3**.

From Lemma 6.3 we have that $\sum_{i=1}^n \widetilde{w}_{ni}(\mathbf{x}) v(Y_i, \mathbf{C})$ converges to $\mathbb{E}[v(Y, \mathbf{C})]$ almost surely as $n \to \infty$. Therefore, with probability 1, for n large enough we obtain that

$$\sup_{t\notin\mathbf{C}} \sum_{i=1}^n \widetilde{w}_{ni}(\mathbf{x}) \frac{|\Psi(Y_i,t) - \lambda(t)|}{b(t)} \leq 1 - 2\,\varepsilon$$

which together with (18) implies that for all $t \notin \mathbf{C}$, $\sum_{i=1}^n \widetilde{w}_{ni}(\mathbf{x})|\Psi(Y_i,t) - \lambda(t)| \leq (1-\varepsilon)|\lambda(t)|$.

From **W1(i)**, for n large enough we have that $1 - \varepsilon/2 \leq \sum_{i=1}^n w_{ni}(\mathbf{x})$. Thus, for all $t \notin \mathbf{C}$

$$|\lambda_n(t)| \geq \Big| \sum_{i=1}^n w_{ni}(\mathbf{x}) \Big| \, |\lambda(t)| - M \sum_{i=1}^n \widetilde{w}_{ni}(\mathbf{x}) |\Psi(Y_i,t) - \lambda(t)| > \left(1 - \frac{\varepsilon}{2}\right) |\lambda(t)| > 0 \, ,$$

which concludes the proof. □

PROOF OF THEOREM 3.3. This proof is a slight modification of Theorem 2 of [25] using Lemma 6.3 and working with $\widetilde{w}_{ni}(\mathbf{x}) = |w_{ni}(\mathbf{x})|\{\sum_{i=1}^n |w_{ni}(\mathbf{x})|\}^{-1}$ as in the proof of Proposition 6.2.

6.2 Proofs of the results in Section 4.

Proposition 6.3. *(i) Assume that $g_n(\mathbf{x})$ converges to $g(\mathbf{x})$ in probability as $n \to \infty$. Then, under **W1**, **W2**, **N1** to **N4** we have that $c_n^{-1/2}(g_n(\mathbf{x}) - g(\mathbf{x}))$ has the same asymptotic distribution as $c_n^{-1/2} \sum_{i=1}^n w_{ni}(\mathbf{x}) \Psi(Y_i, g(\mathbf{x}))/\lambda_1(g(\mathbf{x}))$.*
*(ii) Let r_2 be the function defined in **B7**, if*

a) $\lim_{n\to\infty} c_n^{-1/2} \sum_{i=1}^n w_{ni}(\mathbf{x}) r_2(\mathbf{x}_i, g(\mathbf{x}_i)) = \beta$ *and*

b) $\lim_{n\to\infty} c_n^{-1/2} \max |w_{ni}(\mathbf{x})| = 0$

we have that $c_n^{-1/2} \sum_{i=1}^n w_{ni}(\mathbf{x})\Psi(Y_i, g(\mathbf{x})) \xrightarrow{w} N(\beta, \sigma^2(\mathbf{x}))$, *where* $\sigma^2(u) = \int \Psi^2(y, g(\mathbf{x}))dF(y, g(u))$.

PROOF. (i) Since $\lambda_n(g_n(\mathbf{x})) = 0$, a second order Taylor's expansion leads to

$$0 = c_n^{-1/2}\lambda_n(g(\mathbf{x})) + c_n^{-1/2}\left(g_n(\mathbf{x}) - g(\mathbf{x})\right)\left[\lambda_{1n}(g(\mathbf{x})) + \left(g_n(\mathbf{x}) - g(\mathbf{x})\right)\lambda_{2n}(\xi_n)\right] ,$$

where $\lambda_{1n}(t) = \sum_{i=1}^n w_{ni}(\mathbf{x})\Psi'(Y_i, t)$, $\lambda_{2n}(t) = \sum_{i=1}^n w_{ni}(\mathbf{x})\Psi''(Y_i, t)$ and $\xi_n = \theta_n g_n(\mathbf{x}) + (1-\theta_n)g(\mathbf{x})$ is an intermediate point.

W1, **W2**, **N3** and **N4** imply that $\lambda_{1n}(g(\mathbf{x}))$ converges to $\lambda_1(g(\mathbf{x}))$ in probability while **N1**, **W1** and **W2** ensure that $\lambda_{2n}(\xi_n)$ is bounded in probability since ξ_n converges to $g(\mathbf{x})$ in probability. Thus, the conclusion is easily derived.

(ii) follows by applying Lindeberg's central limit theorem. □

PROOF OF THEOREM 4.1. It is an immediate consequence of Proposition 6.3.

Proposition 6.4. *Under* ***W6***, ***W7*** *and* ***N6*** *we have that* $c_n^{-1/2} \sup_y |F_n(y) - F(y, g(\mathbf{x}))| = O_p(1)$, *where* $F_n(y) = \sum_{i=1}^n w_{ni}(\mathbf{x})\mathbb{I}_{(-\infty,y]}(Y_i)$.

PROOF. From **N6**, we have that

$$\begin{aligned} c_n^{-1/2}\|F_n - F\|_\infty \quad &\leq \quad c_n^{-1/2} \sup_y \left| \sum_{i=1}^n w_{ni}(\mathbf{x})[\mathbb{I}_{(-\infty,y]}(Y_i) - F(y, g(\mathbf{x}_i)] \right| \\ &\quad + c_n^{-1/2} L \sum_{i=1}^n |w_{ni}(\mathbf{x})| \; |g(\mathbf{x}_i) - g(\mathbf{x})|. \end{aligned}$$

Thus, using **W7**, it is enough to show that $c_n^{-1/2} \sup_y \left| \sum_{i=1}^n w_{ni}(\mathbf{x}) \left[\mathbb{I}_{(-\infty,y]}(Y_i) - F(y, g(\mathbf{x}_i))\right] \right| = O_p(1)$, which follows easily using the transformation given in [31] (pp. 102-103) and the Marcus and Zinn inequality (see [31] pp.820). □

PROOF OF THEOREM 4.2. Follows as in [7] using Proposition 6.4. □

Acknowledgements. This research was partially supported by Grants PID 00821 from CONICET, PICT 0216 from ANPCYT and X-018 from the Universidad de Buenos Aires at Buenos Aires, Argentina.

References

[1] Bianco, A. and Boente, G. (1996). Robust nonparametric generalized regression estimation. *Impresiones previas del Departamento de Matemática*. FCEN, Nro 93.

[2] Bianco, A., García Ben, M. and Yohai, V. (2005). Robust estimation for linear regression with asymmetric errors. *Canad. J. Statist.*, **33**, 511–528.

[3] Bianco, A. and Yohai, V. (1996). Robust estimation in the logistic regression model. *Lecture Notes in Statist.*, **109**, 17–34. Springer–Verlag, New York.

[4] Boente, G. and Fraiman, R. (1989). Robust nonparametric regression estimation. *J. Multivariate Anal.*, **29**, 180–198.

[5] Boente G. and Fraiman R. (1991). A functional approach to robust nonparametric regression. The IMA Volumes in Mathematics and its applications, Vol. 33. *Directions in Robust Statistics and Diagnostics*, eds. Werner Stahel and Sanford Weisberg, 35–46.

[6] Boente, G., Fraiman, R. and Meloche, J. (1997). Robust plug-in bandwidth estimators in nonparametric regression. *J. Statist. Plann. Inference*, **57**, 109-142.

[7] Boos, D. and Serfling, J. (1980). A note on differentials and the CLT and LIL for statistical functions, with applications to M–estimates. *Ann. Statist.*, **8**, 618–624.

[8] Brown, L. D., Cai, T. and Zhou, H. (2010). Nonparametric Regression in Exponential Families. *Ann. Statist.*, **38**, 2005–2046.

[9] Cantoni, E. and Ronchetti, E. (2001). Resistant selection of the smoothing pa- rameter for smoothing splines. *Statist. Comput.*, **11**, 141-146.

[10] Cleveland, W. (1979). Robust locally weighted regression and smoothing scatterplots. *J. Amer. Statist. Assoc.*, **74**, 829–836.

[11] Cleveland, W. and Devlin, S. (1988). Locally weighted regression: an approach to regression analysis by local fitting. *J. Amer. Statist. Assoc.*, **83**, 596–610.

[12] Collomb, G. (1981). Estimation nonparamétrique de la regression: Revue bibliographique. *Internat. Statist. Rev.*, **49**, 75–93.

[13] Copas, J.B. (1983). Plotting p against x. *Appl. Statist.*, **32**, 25–31.

[14] Croux, C and Haesbroeck, G (2003). Implementing the Bianco and Yohai estimator for logistic regression. *Comp. Statist. Data Anal.*, **44**, 273–295.

[15] Devroye, L. (1981). On the almost everywhere convergence of nonparametric regression function estimates. *Ann. Statist.*, **9**, 1310–1319.

[16] Devroye, L. (1982). Necessary and sufficient conditions for the pointwise convergence of nearest neighbor regression function estimates. *Z. Wahrsch. Verw. Gebiete*, **51**, 15–25.

[17] Fan, J., Heckman, N. and Wand, P. (1995). Local polynomial kernel regression for generalized linear models and quasi–likelihood functions. *J. Amer. Statist. Assoc.*, **90**, 141–160.

[18] Fowlkes, E.B. (1987). Some diagnostics for binary logistic regression via smoothing. *Biometrika*, **74**, 503–516.

[19] Gasser, T. and Müller, H. G. (1984). Estimating regression functions and their derivatives by the kernel method. *Scand. J. Statist.*, **11**, 171-185.

[20] Georgiev, A. (1988). Consistent nonparametric multiple regression: The fixed design case. *J. Multivariate Anal.*, **25**, 100–110.

[21] Greblicki, N., Krzyźak, A. and Pawlak, M. (1984). Distribution-free pointwise consistency of kernel regression estimate. *Ann. Statist.*, **12**, 1570–1575.

[22] Härdle, W. (1984). Robust regression function estimation. *J. Multivariate Anal.*, **14**, 169-180.

[23] Härdle, W. and Gasser, Th. (1984). Robust nonparametric function fitting. *J. Roy. Statist. Soc.*, Ser. B, **46**, 42–51.

[24] Härdle, W. and Tsybakov, A. (1988). Robust nonparametric regression with simultaneous scale curve estimation. *Ann. Statist.*, **16**, 120–135.

[25] Huber, P. (1967). The behavior of maximum likelihood estimates under nonstandard conditions. *Proc. 6th Berkeley Symp.*, **1**, 221–233.

[26] Künsch, H., Stefanski, L. and Carroll, R. (1989). Conditionally unbiased bounded-influence estimation in general regression models with applications to generalized linear models. *J. Amer. Statist. Assoc.*, **84**, 460–466.

[27] Leung, D. (2005). Cross–validation in nonparametric regression with outliers. *Ann. Statist.*, **3**, 2291-2310

[28] Leung, D., Marrot, F. and Wu, E. (1993). Bandwidth selection in robust smoothing. *J. Nonparametr. Stat.*, **4**, 333-339.

[29] Marcinkiewicz, J. and Zygmund, A. (1937). Sur les fonctions indépendantes. *Fund. Math.*, **29**, 60–90.

[30] Nelder, J.A. and Wedderburn, R. (1972). Generalized linear models. *J. Roy. Statist. Soc.*, Ser. A, **135**, 370–384.

[31] Shorack, G. and Wellner, J. (1986). *Empirical Processes with Applications to Statistics*. Wiley: New York.

[32] Stute, W. (1984). Asymptotic Normality of Nearest Neighbor Regression Function Estimates. *Ann. Statist.*, **12**, 917–926.

[33] Wang, F. and Scott, D. (1994). The L1 method for robust nonparametric regression. *J. Amer. Statist. Assoc.*, **89**, 65-76.