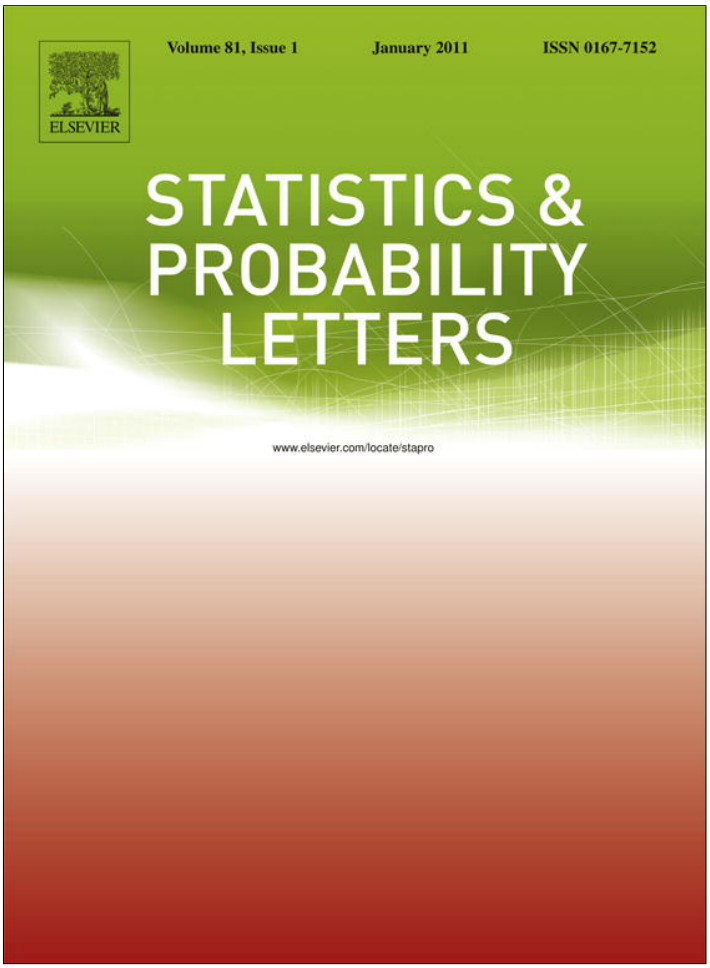

(This is a sample cover image for this issue. The actual cover is not yet available at this time.)

Statistics and Probability Letters 81 (2011) 1986–1994

Contents lists available at SciVerse ScienceDirect

Statistics and Probability Letters

journal homepage: www.elsevier.com/locate/stapro

Robust estimation for nonparametric generalized regression☆

Ana M. Bianco [a,b], Graciela Boente [a,b,*], Susana Sombielle [c,d]

[a] FCEyN, Universidad de Buenos Aires, Argentina
[b] CONICET, Argentina
[c] FCEyN, Universidad de Buenos Aires, Argentina
[d] Universidad Tecnológica Nacional, Argentina

ARTICLE INFO

Article history:
Received 9 May 2011
Received in revised form 11 August 2011
Accepted 15 August 2011
Available online 24 August 2011

Keywords:
Asymptotic properties
Nonparametric generalized regression
Robust estimation
Smoothing techniques

ABSTRACT

This paper focuses 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 proposal is a natural extension of robust methods developed in the setting of parametric generalized linear models. Robust estimators bounding either large values of the deviance or of the Pearson residuals are introduced and their asymptotic behaviour is derived. Through a Monte Carlo study, for the Poisson and Gamma distributions, the finite properties of the proposed procedures are investigated and their performance is compared with that of the classical ones. A resistant cross-validation method to choose the smoothing parameter is also considered.

1. Introduction

Generalized linear models, introduced by Nelder and Wedderburn (1972), 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. A trend in the last few years has been to determine the underlying model or to check a parametric model, via nonparametric techniques. However, most of the work has been developed for the case of homoscedastic additive noise and so these techniques are not applicable to the setting of nonparametric generalized regression.

Outlying observations are of particular concern in the classical problem of estimating location and scale of a family of distribution and in linear regression. In the nonparametric approach outliers are certainly still an important problem. Indeed, classical estimators of the regression function are based on local means and so, they are very sensitive to outliers. An advantage of robust methods over traditional ones is that they still work well under very general assumptions. 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 fixed covariates $\mathbf{x}_i \in A \subset \mathbb{R}^d$, $1 \leq i \leq n$ where A is a compact set. In this case, we assume that there exists a smooth

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

* Correspondence to: Instituto de Cálculo, Ciudad Universitaria, Pabellón 2, Buenos Aires, C1428EHA, Argentina. Fax: +54 11 45763375.
E-mail address: gboente@dm.uba.ar (G. Boente).

0167-7152/$ – see front matter © 2011 Elsevier B.V. All rights reserved.
doi:10.1016/j.spl.2011.08.007

function $g : \mathbb{R}^d \to \tau \subset \mathbb{R}$ such that Y_i has distribution $F(y, g(\mathbf{x}_i))$. Under a generalized linear 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, besides smoothness assumptions on the function g we do not assume any parametric structure on it. In most situations, when first moments exist, g satisfies $g(\mathbf{x}_i) = \mathbb{E}(Y_i)$.

Many authors have considered the problem of the estimation of the regression function g, see for instance Härdle (1990) and Härdle et al. (2004) for a review. In logistic models, the first paper which applied kernel methods is due to Copas (1983). Later on, Fowlkes (1987) used this smoothing technique for diagnostics in logistic regression. More recently, Brown et al. (2010) 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. 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. A first resistant proposal for regression models was given by Cleveland (1979) who considered a robust locally weighted regression scatterplot (see also Cleveland and Devlin, 1988) and asymptotic results for the univariate case were obtained by Härdle and Gasser (1984). Robust methods for estimating $g(\mathbf{x})$ under a nonparametric regression model have been proposed by Härdle (1984), Härdle and Tsybakov (1988) and Boente and Fraiman (1989).

As mentioned above, our main goal is to estimate nonparametrically the regression function g in a robust fashion, for discrete or continuous distributions, such as the Poisson or Gamma distributions, when anomalous data are present. Our resistant proposals are local M-estimators, which can be thought as nonparametric versions of robust estimators that have been proposed in the parametric setting, such as those considered by Bianco and Yohai (1996) for logistic regression, Bianco et al. (2005) for the Gamma distribution and Künsch et al. (1989) and Cantoni and Ronchetti (2001a) for generalized linear models. We investigate both theoretical and numerical properties of the proposed estimators. As in other nonparametric settings, the estimators are asymptotically normally distributed with a non-zero mean parameter, which is usually called the asymptotic bias. In the situation under study, even when the asymptotic bias seems to depend on the chosen score function, it can be seen that, under regularity conditions, it performs as in the usual nonparametric regression model. The numerical results show that the proposal is very stable under different degrees of contamination and so suggest that the methods are resistant. We also introduce a resistant cross-validation method to choose the smoothing parameter.

The paper is organized as follows. In Section 2, robust proposals bounding either large values of the deviance or of the Pearson residuals are considered, as well as, the particular case of Poisson and Gamma models. In Section 3, we present the results of a Monte Carlo Study which allow to analyse the finite sample properties of the proposed methods and also to compare them with the classical ones based on local means. We also describe a resistant cross-validation method to choose the smoothing parameter. The asymptotic behaviour of the introduced estimators is derived in Section 4. Proofs are relegated to the Appendix.

Most of the robust techniques are developed for fixed carriers. However, they may be directly extended to the case of random covariates, that is when the response variables $\{Y_i : 1 \le i \le n\}$ are related to random covariates $\{\mathbf{X}_i : 1 \le i \le n\}$, in such a way that $Y_i|\mathbf{X}_i = \mathbf{x}_i$ has distribution $F(y, g(\mathbf{x}_i))$ with $g : \mathbb{R}^d \to \tau \subset \mathbb{R}$.

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.

Assume that we observe independent random variables $\{Y_i : 1 \le i \le n\}$ related to fixed carriers $\mathbf{x}_i \in A \subset \mathbb{R}^d$, $1 \le i \le n$ and such that $Y_i \sim F(y, g(\mathbf{x}_i))$ with $g : \mathbb{R}^d \to \tau \subset \mathbb{R}$ and A 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 regression, the classical smoothers are of the form $\widehat{g}(\mathbf{x}) = \sum_{i=1}^n w_{ni}(\mathbf{x})Y_i$, where $w_{ni}(\mathbf{x}) = w_{ni}(\mathbf{x}, \mathbf{x}_1, \ldots, \mathbf{x}_n)$ is a probability weight function, i.e., $w_{ni}(\mathbf{x}) \ge 0$ and $\sum_{i=1}^n w_{ni}(\mathbf{x}) = 1$, as for kernel, k-nearest neighbour and nearest neighbour with kernel weights. These local weights give more emphasis to those observations which are closer to $\mathbf{x}$. The classical smoothers can be seen as the solution of a minimum least squares problem given by $\operatorname{argmin}_{t\in\mathbb{R}} \sum_{i=1}^n w_{ni}(\mathbf{x})(Y_i - t)^2$. As local sample means, these estimators can be severely affected by a single outlier placed in the surroundings of the target point $\mathbf{x}$ and for that reason, more robust estimators, such as local M-smoothers, have been introduced.

In order to introduce a local M-estimator, we replace the quadratic loss function by a more general loss function $\rho : \mathbb{R}^2 \to \mathbb{R}_{>0}$, so as to downweight the effect of an occasional outlier. In order to have a proper definition of the target functionals, we need $\mathbb{E}(\rho(Y, t))$ to exist. For that reason, we eventually need to include a term $a(Y)$ when defining the loss function ρ to ensure that the function $\gamma(t) = \mathbb{E}[\rho(Y, t)]$ is well defined. For instance, when estimating the median, $a(Y) = |Y|$ and so, the loss function is defined as $\rho(Y, t) = |Y - t| - |Y|$. We assume Fisher-consistency, i.e., that $\gamma(g(\mathbf{x})) = \min_{t\in\tau} \gamma(t)$. Now, to estimate $\gamma(t)$, we consider the sample version of γ given by

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

and so, the estimator of $g(\mathbf{x})$ can be obtained minimizing (1) over $t \in \tau$. Then, we define $g_n(\mathbf{x}) = \text{argmin}_{t\in\tau}\gamma_n(t)$. More generally, as in Huber (1967), one can define estimates of $g(\mathbf{x})$ as any sequence $g_n(\mathbf{x})$ satisfying

$$\gamma_n(g_n(\mathbf{x})) - \inf_{t\in\tau}\gamma_n(t) \xrightarrow{a.s.} 0\,. \tag{2}$$

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

$$\lambda_n(g_n(\mathbf{x})) = 0, \tag{3}$$

where for $t \in \tau\lambda_n(t) = \sum_{i=1}^n w_{ni}(\mathbf{x})\Psi(Y_i,t)$ and $w_{ni}(\mathbf{x})$ are as in (1). These two approaches are equivalent if (3) 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 (3), where we suspect that this solution will converge to the solution $g(\mathbf{x})$ of $\lambda(t) = \mathbb{E}[\Psi(Y,t)] = 0$.

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 ρ-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) 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

$$\rho(y,t) = \phi(\ln f(y,y) - \ln f(y,t)) + G(t), \tag{4}$$

with $G'(t) = \int \varphi(\ln f(y,y) - \ln f(y,t))f'(y,t)\mathrm{d}\mu(y)$ 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 Croux and Haesbroeck (2003). This family of estimators will be called *robust deviance estimators*. This approach is in the spirit of Bianco and Yohai (1996) 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 $f(y,t) = (\exp(-t)t^y/y!)\mathbb{I}_{N\cup\{0\}}(y)$ and thus, $\mathbb{E}(Y) = t$, $\mathbb{V}(Y) = t$. In this case, if ϕ is a loss function, from (4) we get that $\rho(y,t) = \phi(-y + y\ln y + t - y\ln t) + G(t)$. Moreover, if, as above, ϕ has first derivative φ, we have that

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

where $G'(t) = -\varphi(t)\exp(-t) - \sum_{j=1}^{\infty}\varphi(j\ln j - j + t - j\ln t)(1 - y/t)\exp(-t)t^j/j!$.

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 $f(y,\alpha,t) = (\Gamma(\alpha))^{-1}\alpha^\alpha y^{\alpha-1}t^{-\alpha}\exp(-\alpha y/t)\mathbb{I}_{[0,+\infty)}(y)$. 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 Bianco et al. (2005). Thus, $\rho(y,t) = \phi(\ln(f(y,y)) - \ln(f(y,t))) = \phi(\alpha(y/t - \ln(y/t) - 1))$ which implies that

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

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 Cantoni and Ronchetti (2001a) where a robustified quasi-likelihood estimator is developed. Let $\mathbb{E}_t$ stands for the expectation when $Y \sim F(\cdot,t)$. If we choose $\Psi(y,t) = \psi((y-t)/\sqrt{V(t)})(t/\sqrt{V(t)}) - \nu(t)$, with $\nu(t) = \mathbb{E}_t[\psi((Y-t)/\sqrt{V(t)})1/\sqrt{V(t)}]$ for $t = \mathbb{E}_t(Y)$ and $V(t) = \mathbb{V}_t(Y)$, the estimator $g_n(\mathbf{x})$ satisfies (3). The function $\nu(t)$ is a correction term to obtain Fisher-consistent estimators, while ψ, that may be the Huber's ψ function, controls the Pearson residuals. This family of estimators will be called *robust quasi-likelihood estimators*. 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 $\Psi(y,t) = \psi(\sqrt{\alpha}((y-t)/t))\sqrt{\alpha}/t - \nu(t)$, where $\nu(t) = \mathbb{E}_t[\psi(\sqrt{\alpha}((Y-t)/t))\sqrt{\alpha}/t]$.

Table 1
Poisson regression. Values of $\mathrm{EF}(\widehat{g}, \widehat{g}_{\mathrm{CL}})$ for contaminated and non-contaminated samples of size $n = 100$.

m	$y^* = 0$		$y^* = 12$		$y^* = 20$	
	$\widehat{g}_{\mathrm{RD}}$	$\widehat{g}_{\mathrm{RQL}}$	$\widehat{g}_{\mathrm{RD}}$	$\widehat{g}_{\mathrm{RQL}}$	$\widehat{g}_{\mathrm{RD}}$	$\widehat{g}_{\mathrm{RQL}}$
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

In Section 3, we will apply these families of estimators to Poisson and Gamma regression models.

3. Finite sample performance

In a first step, we perform a numerical study to have an insight into the local robustness and the efficiency of our proposals. Secondly, we face the problem of choosing in a robust fashion the bandwidth parameter.

3.1. Simulation conditions

A Monte Carlo study is carried out in order to assess the performance of the proposed robust estimators for finite contaminated and non-contaminated samples. Robust procedures are expected to be less sensitive to outliers than their classical counterparts. In order to measure the local resistance of the proposed procedures we regard different amounts of outliers varying their level of atipicity. We consider a Poisson and a Gamma regression model with one-dimensional covariate x. In both situations, we perform $N = 5000$ replications by generating, 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 consider the weights introduced in Gasser and Müller (1984) using the Epanechnikov kernel with bandwidth h. As is well known, boundary kernels improve the performance of the regression estimators. We study three regression estimators of the regression function g: $\widehat{g}_{\mathrm{CL}}$, the classical local mean estimator; $\widehat{g}_{\mathrm{RD}}$, the robust deviance estimator with ϕ the loss function considered in Croux and Haesbroeck (2003) with tuning constant $d = 2$ and $\widehat{g}_{\mathrm{RQL}}$, the robust quasi-likelihood estimator with $\psi = \psi_c$ the Huber's score function with $c = 1.6$.

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

$$\mathrm{MSE}_j(\widehat{g}, g) = \frac{1}{n}\sum_{i=1}^{n}(\widehat{g}(x_i) - g(x_i))^2 \qquad \mathrm{AMSE}(\widehat{g}, g) = \frac{1}{N}\sum_{j=1}^{N}\mathrm{MSE}_j(\widehat{g}, g), \tag{5}$$

which measure the square error in the j-th replication and the mean square error over replications, respectively. For each robust estimator $\widehat{g}$, we compare its behaviour with that of the classical one by computing as summary measure the efficiency given by $\mathrm{EF}(\widehat{g}, \widehat{g}_{\mathrm{CL}}) = \mathrm{AMSE}(\widehat{g}_{\mathrm{CL}}, g)/\mathrm{AMSE}(\widehat{g}, g)$.

3.1.1. Poisson regression

As mentioned above, we generate observations Y_i following a Poisson distribution $\mathcal{P}(g(x_i))$ with $g(x) = \exp(2\sin 4\pi x) + 1$. We contaminate the original samples by replacing $m = 1$ and $m = 3$ responses by arbitrary values $y^* = 0$, 12 or 20. These contaminating data are located at fixed positions of the covariate x and when $m = 3$, the outliers lie at successive values of x, so as to obtain a more severe contamination pattern. We choose $h = 0.05$ as smoothing parameter after looking at several choices. An automatic data-driven choice is discussed in Section 3.2.

To illustrate the fit of the considered regression estimators, Fig. 1 presents the different estimators applied to one simulated sample without outliers and when $m = 3$ responses are replaced by $y^* = 20$. The black line corresponds to the true function g, the red one to the classical estimator $\widehat{g}_{\mathrm{CL}}$, while the blue and the green lines to the robust estimators $\widehat{g}_{\mathrm{RD}}$ and $\widehat{g}_{\mathrm{RQL}}$, respectively. The plot on the left side shows 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 throughout the range of the covariate, except for the region where the outlying responses were located. In this crucial interval, the classical estimator $\widehat{g}_{\mathrm{CL}}$ suffers 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 the number of outliers m. The proposed estimators, in particular $\widehat{g}_{\mathrm{RD}}$, are much more stable in presence of anomalous data.

3.1.2. Gamma regression

In this section, we report the results obtained under a Gamma regression model. We generate $n = 100$ observations $Y_i \sim \Gamma(16, 8(\sin 4\pi x_i + 3)^{-1})$, and so $\mathbb{E}(Y_i) = g(x_i) = 2(\sin 4\pi x_i + 3)$, $1 \le i \le n$. We contaminate the original samples

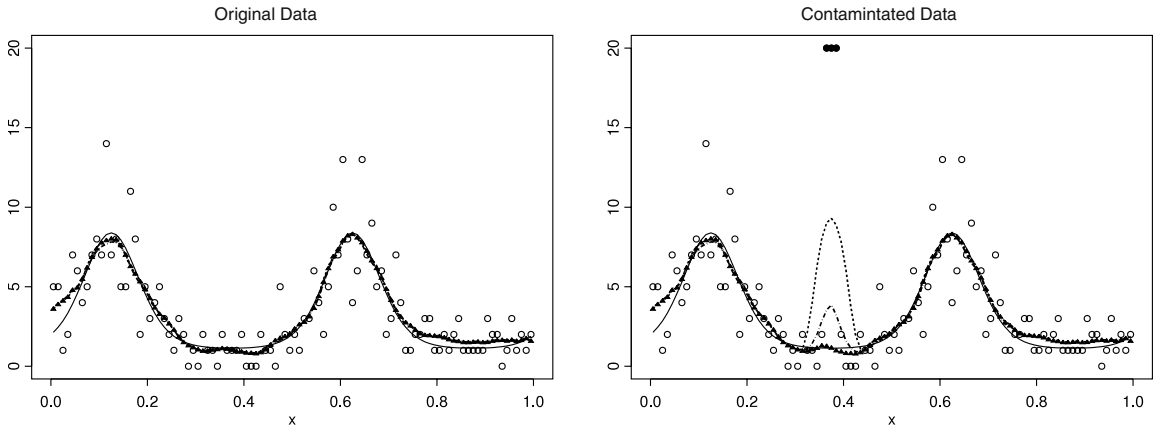

Fig. 1. Simulated data from Poisson regression. The plot on the left corresponds to original samples, while the plot on the right to contaminated ones with $m = 3$ outlying observations with $y^* = 20$. The solid line corresponds to the true function $g(x)$, the dotted one to the classical estimator $\widehat{g}_{\mathrm{CL}}$, while the solid line with filled triangles and the dashed line $(- \cdot -)$ to the robust estimators $\widehat{g}_{\mathrm{RD}}$ and $\widehat{g}_{\mathrm{RQL}}$, respectively. Filled circles represent the outliers.

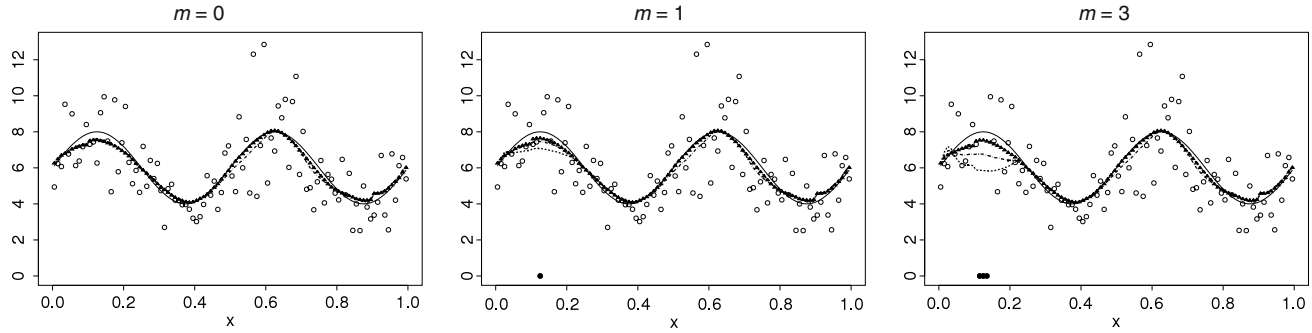

Fig. 2. Simulated data from Gamma regression. Samples with m outlying observations with $y^* = 0$. The solid line corresponds to the true function $g(x)$, the dotted one to the classical estimator $\widehat{g}_{\mathrm{CL}}$, while the solid line with filled triangles and the dashed line $(- \cdot -)$ to the robust estimators $\widehat{g}_{\mathrm{RD}}$ and $\widehat{g}_{\mathrm{RQL}}$, respectively. Filled circles represent the outliers.

Table 2
Gamma regression. Values of $\mathrm{EF}(\widehat{g}, \widehat{g}_{\mathrm{CL}})$ for contaminated and non-contaminated samples of size $n = 100$.

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

by replacing $m = 1$ and $m = 3$ responses by arbitrary values $y^* = 0$. As before, these outliers are located at fixed positions of covariate x and when $m = 3$ they correspond to successive values of x. As in Section 3.1.1, after a preliminary inspection, we choose $h = 0.1$ as the smoothing parameter.

As an example, Fig. 2 pictures the fitted curves obtained from the classical estimator $\widehat{g}_{\mathrm{CL}}$ and the robust estimators $\widehat{g}_{\mathrm{RD}}$ and $\widehat{g}_{\mathrm{RQL}}$ when they are applied to one simulated sample without outliers and with $m = 1$ and $m = 3$ outlying observations with $y^* = 0$. As above, we plot in black the true regression function $g(x)$, in red the classical estimator, while in blue and green the estimators $\widehat{g}_{\mathrm{RD}}$ and $\widehat{g}_{\mathrm{RQL}}$, 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 from the true regression function g in presence of just one outlying observation. Instead, the fit obtained from $\widehat{g}_{\mathrm{RD}}$ looks very stable throughout 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 1.61 times the $\mathrm{AMSE}(\widehat{g}_{\mathrm{RD}}, g)$.

Remark 3.1. In the situations considered for the Poisson and Gamma regression models, both robust estimators were calibrated so as to achieve the same asymptotic efficiency. Under this condition, the robust deviance estimators outperform the robust quasi-likelihood ones. This behaviour was also observed in the literature, when considering generalized linear

models where the influence of some outliers is better controlled bounding the deviance than bounding the Pearson residuals. Note that, in our simulation there are no leverage points and hence, one can guess that the redescending nature of the score function used in the numerical study is enough to handle the outlying points in all the cases considered.

3.2. 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 Leung et al. (1993), Leung (2005), Wang and Scott (1994), Boente et al. (1997) and Cantoni and Ronchetti (2001b). 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. Boente and Fraiman (1991) pointed out that robust cross-validation methods should be an alternative. In the following, we describe a resistant cross-validation procedure (RCV) based on robustified deviances

1. For each given bandwidth h compute $\widehat{y}_h^{-i} = \operatorname{argmin}_t \sum_{j\neq i}^n w_{nj}(x_i, h)\rho(y_j, t)$, 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 $\widehat{h}_{n,\mathrm{R}} = \operatorname{argmin}_h \sum_{i=1}^n \rho(y_i, \widehat{y}_h^{-i})$.

To study the performance of the RCV procedure, we carry 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 (CV) based on the deviance. As in Section 3.1.2, we generate $n = 100$ 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 follow the contamination scheme with $m = 3$ outliers described in Section 3.1.2. For each non-contaminated sample, we compute the classical bandwidth $\widehat{h}_n$ and the resistant one $\widehat{h}_{n,\mathrm{R}}$. For the contaminated samples, we perform the same computations obtaining classical and resistant bandwidths denoted $\widehat{h}_n^c$ and $\widehat{h}_{n,\mathrm{R}}^c$, respectively. We replicate $N = 500$ times. Fig. 3(a) 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 compute the mean square error defined in (5) for each estimator. Fig. 3(b) 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.

4. Asymptotic behaviour

In this section we derive the asymptotic behaviour of the proposed estimators. In Bianco et al. (2011) these results are stated under weaker conditions on the loss function ρ.

4.1. Consistency

We will assume the following set of assumptions. The conditions on the weight function stated below are the usual assumptions required to the weights in nonparametric regression (see, for instance Georgiev, 1988).

W1. $\lim_{n\to\infty} \sum_{i=1}^n w_{ni}(\mathbf{x}) = 1$, $\lim_{n\to\infty} \sum_{i=1}^n |w_{ni}(\mathbf{x})| \mathbb{I}_{\{\|\mathbf{x}_i - \mathbf{x}\| > a\}} = 0$ for all $a > 0$ and there exists $M > 0$ such that $\sum_{i=1}^n |w_{ni}(\mathbf{x})| \leq M$ for all $n \geq 1$. Moreover, $\lim_{n\to\infty} \max_{1\leq i\leq n} |w_{ni}(\mathbf{x})| = 0$.
W2. $\lim_{n\to\infty}\{\max_{1\leq i\leq n} |w_{ni}(\mathbf{x})|\} \log n = 0$.
W3. $w_{ni}(\mathbf{x}) \geq 0$ for all $1 \leq i \leq n, n \geq 1$.

In order to obtain consistency results for the estimators defined through the minimization procedure (2), we need the following conditions for the loss function which are similar to those given by Huber (1967). However, some of them have been customized to our setting since the observations are not identically distributed. We will fix some extra notation. Given a compact set $\mathbf{C} \subset \tau$, let $\beta(\mathbf{C}) = \mathbb{E}[\inf_{t\notin\mathbf{C}} \rho(Y, t)]$ and $\beta_n(\mathbf{C}) = \sum_{i=1}^n w_{ni}(\mathbf{x}) \inf_{t\notin\mathbf{C}} \rho(Y_i, t)$. Moreover, for any set $\mathcal{U} \subset \tau$, let $\gamma_n(\mathcal{U}) = \sum_{i=1}^n w_{ni}(\mathbf{x}) \inf_{t\in\mathcal{U}} \rho(Y_i, t)$ and $\gamma(\mathcal{U}) = \mathbb{E}(\inf_{t\in\mathcal{U}} \rho(Y, t))$. Note that, $\beta(\mathbf{C}) = \gamma(\tau - \mathbf{C})$. In the Appendix, it is shown that W1, A1 and A4 below imply that $\beta_n(\mathbf{C}) \xrightarrow{p} \beta(\mathbf{C})$.

A1. For each $t \in \tau$, $\rho(y, t)$ is a bounded measurable function and $\rho(y, t)$ is separable in the sense of Doob.
A2. The function ρ is almost surely continuous in t.

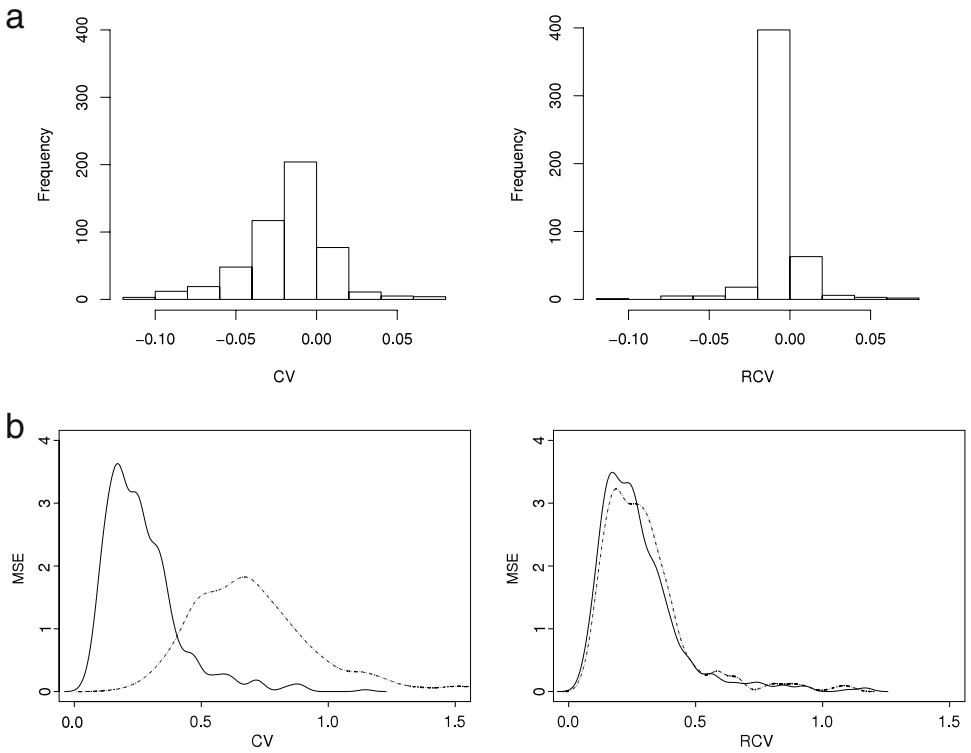

Fig. 3. (a) Histograms of the differences $\widehat{h}_n - \widehat{h}_n^c$ on the left and $\widehat{h}_{n,\mathrm{R}} - \widehat{h}_{n,\mathrm{R}}^c$ on the right. (b) Density estimator of the mean squared errors for the original samples using the cross validation bandwidths. The solid line correspond to those computed with the original samples, while the dashed line to those obtained for the contaminated samples. The sample size was $n = 100$.

A3. For all $t \neq g(\mathbf{x})$, $\gamma(t) > \gamma(g(\mathbf{x}))$.
A4. For any set $\mathcal{U} \subset \tau$ the function $r(\mathbf{u}) = \int \inf_{t\in\mathcal{U}}[\rho(y,t) - a(y)]\,\mathrm{d}F(y, g(\mathbf{u}))$ is continuous at $\mathbf{x}$.
A5. For any sequence of compact sets $\mathbf{C}_n$ converging to τ, $\liminf_{n\to\infty} \beta(\mathbf{C}_n) > \gamma(g(\mathbf{x}))$.

We have that assumptions A1–A3 imply that $\gamma(\mathcal{U}) \to \gamma(g(\mathbf{x}))$ as the neighbourhood $\mathcal{U}$ of t shrinks to $\{g(\mathbf{x})\}$.

When ρ is not bounded, weaker conditions can be required by assuming that $\mathbb{E}|\rho(Y,t)|^p < \infty$ for all $t \in \tau$, for some $p \geq 2$ (see Bianco et al., 2011).

The following result states that the estimators defined through (2) are consistent.

Lemma 4.1.1. *Let $g_n(\mathbf{x})$ be any value satisfying (2). Assume that* W1–W3, A1, A4 *and* A5 *hold. 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 A5 seems very restrictive, the conclusion of Lemma 4.1.1 can be directly verified for many families of estimates including, for instance, local medians.

Theorem 4.1.1. *Under* A1–A4, W1–W3 *and if the conclusion of Lemma 4.1.1 holds, we have that $g_n(\mathbf{x}) \xrightarrow{a.s.} g(\mathbf{x})$ as $n \to \infty$.*

Remark 4.1.1. If ρ is not bounded we can replace, in Theorem 4.1.1, W2 by $\lim_{n\to\infty}\{\max_{1\leq i\leq n} w_{ni}^2(\mathbf{x})\}\, n \log\log n = 0$ by requiring that $\sup_i \mathbb{E}|\rho(Y_i,t) - a(Y_i)|^{2+s} \leq M_2(t) < \infty$ for some $s > 0$ and for all $t \in \tau$ (see Bianco et al., 2011).

Strong consistency of the estimators defined through the differentiated equations, i.e., those defined through (3) can be derived under mild conditions. When $\tau \subset \mathbb{R}$, two different situations may be distinguished:

(a) $\lambda(t)$ is strictly monotone in a neighbourhood of $g(\mathbf{x})$ which includes the case when $\Psi(y,\cdot)$ is monotone
(b) $\lambda(t)$ has a unique solution $g(\mathbf{x})$.

Both cases are considered in Bianco et al. (2011).

4.2. Asymptotic normality

In order to study the asymptotic behaviour, we will need some additional assumptions. First, we introduce some notation assuming differentiability of the function Ψ. 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, $\lambda_1(t) = \partial\lambda(t)/\partial t$. Theorem 4.2.2 states the asymptotic distribution under weaker conditions than differentiability of Ψ.

N1. Ψ is twice continuously differentiable in t and for any neighbourhood $\mathcal{U}$ of $g(\mathbf{x})$, there exists $c > 0$ such that one of the following conditions holds
  (i) $\sup_{t\in\mathcal{U}} |\Psi''(y,t)| \le c$ for all y
  (ii) $\sup_{1\le i\le n} \mathbb{E}\sup_{t\in\mathcal{U}} |\Psi''(Y_i,t)| \le c$ for all $n \ge 1$
N2. $\lambda_1(g(\mathbf{x})) \neq 0$.
N3. For some positive constant c, $\sup_{1\le i\le n} \mathbb{E}|\Psi'(Y_i, g(\mathbf{x}))|^2 \le c$ for all $n \ge 1$.
N4. Ψ' satisfies one of the following conditions
  (i) $\widetilde{r}(\mathbf{u}) = \int \Psi'(y,t)\mathrm{d}F(y, g(\mathbf{u}))$ 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 4.2.1. *Let* $r^\star(\mathbf{u},t) = \int \Psi(y,t)\mathrm{d}F(y,g(\mathbf{u}))$. *Assume that* $g_n(\mathbf{x}) \xrightarrow{p} g(\mathbf{x})$ *and* W1, N1–N4 *hold. If in addition,* $\lim_{n\to\infty} c_n^{-1/2} \max|w_{ni}(\mathbf{x})| = 0$ *and* $\lim_{n\to\infty} c_n^{-1/2}\sum_{i=1}^n w_{ni}(\mathbf{x}) r^\star(\mathbf{x}_i, g(\mathbf{x}_i)) = \beta$, *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* $\sigma^2(\mathbf{u}) = \int \Psi^2(y, g(\mathbf{x}))\mathrm{d}F(y, g(\mathbf{u}))$.

Remark 4.2.1. 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

W4. $\sum_{i=1}^n w_{ni}(\mathbf{x}) = 1$.
W5. There exists a constant $M > 0$ such that $c_n^{-1/2}\sum_{i=1}^n |w_{ni}(\mathbf{x})|\,|g(\mathbf{x}_i) - g(\mathbf{x})| \le M$ for all $n \ge 1$.
N5. $\Psi(y,t)$ is a bounded continuous function such that for each fixed $t \in \tau$, $\Psi(\cdot,t)$ is of bounded variation.
N6. $\lambda'(g(\mathbf{x})) = \partial\lambda(t)/\partial t|_{t=g(\mathbf{x})} \neq 0$.
N7. $F(y,t)$ is Lipschitz as a function of t uniformly in y, i.e., there exists $L > 0$ such that $|F(y,t_1) - F(y,t_2)| \le L|t_1 - t_2|$ for all $y \in \mathbb{R}$, $t_1, t_2 \in \tau$.

Remark 4.2.2. The Lipschitz condition required in N7 is easily verified for most of the standard distribution families.

Theorem 4.2.2. *Under* W4, W5, N5–N7, *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.*

Acknowledgment

We want to thank an anonymous referee for constructive comments on an earlier version of the paper.

Appendix

In order to prove Lemma 4.1.1, we need the following auxiliary result whose proof follows using standard arguments and can be found in Bianco et al. (2011).

Lemma A. *Under* W1, A1 *and* A4 *we have that for any* $t \in \tau$ *and any set* $\mathcal{U} \subset \tau$

(i) $\gamma_n(t) \xrightarrow{p} \gamma(t)$ *and* $\gamma_n(\mathcal{U}) \xrightarrow{p} \gamma(\mathcal{U})$
(ii) $\gamma_n(t) \xrightarrow{a.s.} \gamma(t)$ *and* $\gamma_n(\mathcal{U}) \xrightarrow{a.s.} \gamma(\mathcal{U})$ *if* W2 *holds.*

Proof of Lemma 4.1.1. The boundedness of ρ implies that given $\varepsilon > 0$ there exists a compact set $\mathbf{C}$ such that $\beta(\mathbf{C}) > \gamma(g(\mathbf{x})) + \varepsilon$.

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

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

The proof follows now using Lemma A. □

Proof of Theorem 4.1.1. The proof is similar to that of Theorem 1 in Huber (1967) using again Lemma A. □

Proposition A.1. (i) *Assume that* $g_n(\mathbf{x})$ *converges to* $g(\mathbf{x})$ *in probability as* $n \to \infty$. *Then, under* W1, N1–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* $r^\star(\mathbf{u}, t) = \int \Psi(y, t)\mathrm{d}F(y, g(\mathbf{u}))$, *if* $\lim_{n\to\infty} c_n^{-1/2}\sum_{i=1}^n w_{ni}(\mathbf{x})r^\star(\mathbf{x}_i, g(\mathbf{x}_i)) = \beta$ *and* $\lim_{n\to\infty} c_n^{-1/2}\max|w_{ni}(\mathbf{x})| = 0$, *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* $\sigma^2(u) = \int \Psi^2(y, g(\mathbf{x}))\mathrm{d}F(y, g(u))$.

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

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

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, N3 and N4 imply that $\lambda_{1n}(g(\mathbf{x}))$ converges to $\lambda_1(g(\mathbf{x}))$ in probability while N1 and W1 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 Lindberg's central limit theorem. □

Proof of Theorem 4.2.1. It is an immediate consequence of Proposition A.1. □

Proposition A.2. *Under* W4, W5 *and* N7 *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 N7, we have that

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

Thus, using W5, it is enough to show that $c_n^{-1/2}\sup_y|\sum_{i=1}^n w_{ni}(\mathbf{x})[\mathbb{I}_{(-\infty,y]}(Y_i) - F(y, g(\mathbf{x}_i))]| = O_p(1)$, which follows easily using the transformation given in Shorack and Wellner (1986, pp. 102–103) and the Marcus and Zinn inequality (see Shorack and Wellner, 1986, pp. 820). □

Proof of Theorem 4.2.2. Follows as in Boos and Serfling (1980) using Proposition A.2 □

References

Bianco, A.M., Boente, G., Sombielle, S., 2011. Robust nonparametric generalized regression estimation. Available at: http://www.ic.fcen.uba.ar/preprints.html.

Bianco, A., García Ben, M., Yohai, V., 2005. Robust estimation for linear regression with asymmetric errors. Canad. J. Statist. 33, 511–528.

Bianco, A., Yohai, V., 1996. Robust Estimation in the Logistic Regression Model. In: Lecture Notes in Statist., vol. 109. Springer-Verlag, New York, pp. 17–34.

Boente, G., Fraiman, R., 1989. Robust nonparametric regression estimation. J. Multivariate Anal. 29, 180–198.

Boente, G., Fraiman, R., 1991. A functional approach to robust nonparametric regression. In: Stahel, Werner, Weisberg, Sanford (Eds.), Directions in Robust Statistics and Diagnostics. In: The IMA Volumes in Mathematics and its Applications, vol. 33. pp. 35–46.

Boente, G., Fraiman, R., Meloche, J., 1997. Robust plug-in bandwidth estimators in nonparametric regression. J. Statist. Plann. Inference 57, 109–142.

Boos, D., 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.

Brown, L.D., Cai, T., Zhou, H., 2010. Nonparametric regression in exponential families. Ann. Statist. 38, 2005–2046.

Cantoni, E., Ronchetti, E., 2001a. Robust inference for generalized linear models. J. Amer. Statist. Assoc. 96, 1022–1030.

Cantoni, E., Ronchetti, E., 2001b. Resistant selection of the smoothing parameter for smoothing splines. Stat. Comput. 11, 141–146.

Cleveland, W., 1979. Robust locally weighted regression and smoothing scatterplots. J. Amer. Statist. Assoc. 74, 829–836.

Cleveland, W., Devlin, S., 1988. Locally weighted regression: an approach to regression analysis by local fitting. J. Amer. Statist. Assoc. 83, 596–610.

Copas, J.B., 1983. Plotting *p* against *x*. J. Appl. Stat. 32, 25–31.

Croux, C., Haesbroeck, G., 2003. Implementing the Bianco and Yohai estimator for logistic regression. Comput. Statist. Data Anal. 44, 273–295.

Fowlkes, E.B., 1987. Some diagnostics for binary logistic regression via smoothing. Biometrika 74, 503–516.

Gasser, T., Müller, H.G., 1984. Estimating regression functions and their derivatives by the kernel method. Scand. J. Statist. 11, 171–185.

Georgiev, A., 1988. Consistent nonparametric multiple regression: the fixed design case. J. Multivariate Anal. 25, 100–110.

Härdle, W., 1984. Robust regression function estimation. J. Multivariate Anal. 14, 169–180.

Härdle, W., 1990. Applied Nonparametric Regression. In: Econometric Society Monographs, Cambridge University Press.

Härdle, W., Gasser, Th., 1984. Robust nonparametric function fitting. J. Roy. Statist. Soc. Ser. B 46, 42–51.

Härdle, W., Müler, M., Sperlich, S., Werwatz, A., 2004. Nonparametric and Semiparametric Models. In: Springer Series in Statistics, Springer Press, Germany.

Härdle, W., Tsybakov, A., 1988. Robust nonparametric regression with simultaneous scale curve estimation. Ann. Statist. 16, 120–135.

Huber, P., 1967. The behaviour of maximum likelihood estimates under nonstandard conditions, in: Proc. 6th Berkeley Symp., vol. 1, pp. 221–233.

Künsch, H., Stefanski, L., 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.

Leung, D., 2005. Cross-validation in nonparametric regression with outliers. Ann. Statist. 3, 2291–2310.

Leung, D., Marrot, F., Wu, E., 1993. Bandwidth selection in robust smoothing. J. Nonparametr. Stat. 4, 333–339.

Nelder, J.A., Wedderburn, R., 1972. Generalized linear models. J. Roy. Statist. Soc. Ser. A 135, 370–384.

Shorack, G., Wellner, J., 1986. Empirical Processes with Applications to Statistics. Wiley, New York.

Wang, F., Scott, D., 1994. The L1 method for robust nonparametric regression. J. Amer. Statist. Assoc. 89, 65–76.