Robust estimators under a semiparametric partly linear autoregression: Asymptotic behavior and bandwidth selection*

Ana Bianco
Universidad de Buenos Aires, Argentina

and
Graciela Boente
Universidad de Buenos Aires and CONICET, Argentina

Author's Address
Instituto de Cálculo
Ciudad Universitaria, Pabellón 2
Buenos Aires, C1428EHA
Argentina

*This research was partially supported by grant 13900-6 from the Fundación Antorchas, grant X094 from University of Buenos Aires, PID 5505 from CONICET and PAV 120 and PICT 21407 from the Agencia Nacional de Promoción Científíca y Tecnológica, Argentina.

Abstract

In this paper, under a semiparametric partly linear autoregression model, a family of robust estimates for the autoregression parameter and the autoregression function is studied. The proposed estimates are based on a three step procedure, in which regression robust estimates and robust smoothing techniques are combined. Asymptotic results on the autoregression estimates are derived. Besides, combining robust procedures with M–smoothers, predicted values for the series and detection residuals, which allow to detect anomalous data, are introduced. Robust cross–validation methods to select the smoothing parameter are presented as an alternative to the classical ones, which are sensitive to outlying observations. A Monte Carlo study is conducted in order to compare the performance of the proposed criteria. Finally, the asymptotic distribution of the autoregression parameter estimate is stated uniformly over the smoothing parameter.

MSC: Primary 62F35, Secondary 62H25.
Key words and phrases: Partly Linear Autoregression, Robust Estimation, Smoothing Techniques, Cross–validation, Rate of Convergence, Asymptotic Properties, Filtering, Prediction.

Abbreviated Title: Behavior of robust semiparametric autoregression estimates

1 Introduction

In the last two decades, partly linear regression models have been extensively studied. Among others we can mention the papers by Ansley and Wecker (1983), Green, Jennison and Seheult (1985), Heckman (1986), Engle, Granger, Rice (1986), Chen (1988), Robinson (1988), Speckman (1988), Chen and Chen (1991), Chen and Shiau (1991, 1994), Gao (1992), Gao and Zhao (1993) and Yee and Wild (1996) who investigated some asymptotic results using smoothing splines, kernel or nearest neighbors techniques. An extensive description of the different results obtained in partly linear regression models can be found in Härdle, Liang and Gao (2000).

When dealing with dependent observations, $\{y_t\}$, autoregressive models have been widely used in applications. Gao and Yee (2000) noticed that, in econometrical problems, one way to solve nonlinearity is to consider non–gaussian ARMA processes, for instance through ARCH models. An alternative could be to use a fully nonparametric autoregressive model which suffers from the "curse of dimensionality"and neglects a possible linear relation between y_t and any lag y_{t-k}. Following a semiparametric approach, several authors have introduced partly linear models for autoregressive models in order to combine the advantages of both parametric and nonparametric methods. A stochastic process $\{y_t\}$, defined over a probability space $(\Omega, \mathcal{A}, \mathcal{P})$, satisfies a partly linear autoregressive model if it can be written as

$$y_t = \sum_{i=1}^{p} \beta_{o,i} y_{t-c_i} + \sum_{j=1}^{q} g_j(y_{t-d_j}) + \epsilon_t \ , \tag{1}$$

where $g_j : I\!R \to I\!R$ are smooth functions and ϵ_t are i.i.d. random variables, independent of $\{y_{t-j}, j \geq 1\}$, $E(\epsilon_t) = 0$ and $E\epsilon_t^2 < \infty$. For simplicity and convenience, we will only consider the case $p = q = 1$, $c_1 = 1$ $d_1 = 2$, which leads to the model

$$y_t = \beta_o y_{t-1} + g(y_{t-2}) + \epsilon_t \ , \tag{2}$$

where $-1 < \beta_o < 1$ is an unknown parameter to be estimated, $g : I\!R \to I\!R$ is an unknown smooth function and ϵ_t are as in (1).

The partly linear autoregressive model (2) is particularly important since it involves not only a linear autoregressive component, but a univariate smoothing which avoids the "curse of dimensionality". Partly linear autoregression models (2) are more flexible than standard linear models since they have a parametric and a nonparametric component. They can be a suitable choice when one suspects that the dependence on the past cannot be adequately explained only through a linear autoregression.

When considering local polynomials, Gao and Liang (1995) established the asymptotic normality of the least squares estimator of β_o, based on a piecewise polynomial approximation, under an $\alpha-$mixing condition. Gao (1995, 1998) also studied the asymptotic normality and obtained a law of iterated logarithm for the kernel-based estimates, $\widehat{\beta}_{LS}$, while Liang (1996) and Gao and Yee (2000) derived some other results (see also Härdle, Liang and Gao (2000) for a review).

It is well known that, both in linear autoregression and in nonparametric autoregression, least squares estimators can be seriously affected by anomalous data. The same statement

holds for partly linear autoregressive models. Let us denote β_o the true value of the parameter. In the classical setting, it is assumed that second moments exists and we have that $g(y) = \phi_2(y) - \beta_o\phi_1(y)$, where $\phi_2(y) = E(y_t|y_{t-2} = y)$ and $\phi_1(y) = E(y_{t-1}|y_{t-2} = y)$. Thus, preliminary estimates of the conditional expectations can be inserted prior to the estimation of the autoregression parameter. Usually, these estimates are linear on the observations and therefore, sensitive to outliers.

Bianco and Boente (2002) considered a different approach which does not require to the errors the existence of moments. Let $\phi_1(y)$ and $\phi_2(y)$ be now any conditional location functionals related to a robust smoother (see Boente and Fraiman, 1988), such as the conditional median. From now on, we will refer to the functional related to a robust smoothing as a robust conditional location functional.

We will briefly remind the definition of robust location conditional functionals introduced in Boente and Fraiman (1989), without requiring any moment conditions.

Let (X, Z) be any random vector and define $s(X)$ any robust measure of the conditional scale, such as the conditional median of the absolute deviations with respect to the conditional median, MAD_{C}, e.g., $s(x) = \text{median}\left(|Z - m(x)| \, |X = x\right)$ where $m(x) = \text{median}(Y|X = x)$ is the median of a regular version $F(z|X = x)$ of the conditional distribution function of $Z|X = x$. For any strictly increasing, bounded continuous score function $\psi : I\!R \to I\!R$, the robust location conditional functional $\phi(X) = E^{\psi}(Z|X)$ defined in Boente and Fraiman (1989) is the essentially unique $\sigma(X)-$measurable function $\phi(X)$ that verifies

$$E\left(h(X)\psi\left(\frac{Z - \phi(X)}{s(X)}\right)\right) = 0 \tag{3}$$

for all integrable function $h(X)$, where $\sigma(X)$ is the $\sigma-$algebra generated by X. If the conditional distribution $F(z|X = x)$ is symmetric around $m(x)$ and ψ is odd, we have that $\phi(x) = m(x)$. Then, in this sense, the robust location conditional functional $\phi(X)$ is a natural extension of the conditional expectation $E(Z|X)$. In Theorem 2.1 of Boente and Fraiman (1989), it was shown that if ψ is an increasing function the solution of (3) exists, is unique and measurable. Furthermore, the weak continuity of the functional defined in this way was proved in Theorem 2.2 therein. Therefore, by applying this functional to weak consistent estimates of the conditional distribution of $y_{t-1}|y_{t-2} = y$ and of $y_t|y_{t-2} = y$, we obtain consistent and asymptotically strongly robust estimates of the robust location conditional functionals $\phi_1(y)$ and $\phi_2(y)$, respectively.

Note that if the functionals ϕ_1 and ϕ_2 satisfy $g(y) = \phi_2(y) - \beta_o\phi_1(y)$, we can re–write model (2) as $y_t - \phi_2(y_{t-2}) = \beta_o(y_{t-1} - \phi_1(y_{t-2})) + \epsilon_t$. For instance, if the errors have a symmetric distribution and $y_{t-1}|y_{t-2} = y$ has a symmetric distribution around $\phi_1(y)$, it is easy to see that $g(y) = \phi_2(y)-\beta_o\phi_1(y)$ holds for local M–functionals with odd score functions, such as the local median.

Using these facts, Bianco and Boente (2002) proposed a class of estimates, with a more resistant behavior, based on a three step procedure under the partly linear autoregressive model (2) which can be described as follows:

- **Step 1:** Estimate $\phi_1(y)$ and $\phi_2(y)$ through a robust smoothing, as local M–type

estimates or local medians. Denote $\widehat{\phi}_1(y)$ and $\widehat{\phi}_2(y)$ the obtained estimates.

- **Step 2:** Estimate the autoregression parameter by applying a robust regression estimate to the residuals $y_t - \widehat{\phi}_2(y_{t-2})$ and $y_{t-1} - \widehat{\phi}_1(y_{t-2})$. Denote $\widehat{\beta}$ the resulting estimator.
- **Step 3:** Define the estimate of the autoregression function g as $\widehat{g}(y) = \widehat{\phi}_2(y) - \widehat{\beta}\widehat{\phi}_1(y)$.

It is worth noticing that this proposal is not but a robust version of the partial autoregression estimators introduced by Gao (1995).

When dealing with independent observations, Gao and Shi (1997) introduced robust estimates based on M-type smoothing splines for nonparametric and semiparametric regression. Their proposal is based on a finite series expansion of the regression function and under a partly linear regression model, asymptotic results for the regression parameter are derived. The three–step proposal defined above follows a different approach since it extends the kernel–based estimators given by Bianco and Boente (2004) for partly linear regression models.

We will briefly discuss the choice of some estimators in Steps 1 and 2.

Consider $\widehat{F}_1(z|y_{t-2} = y)$ and $\widehat{F}_2(z|y_{t-2} = y)$ the estimates of the distribution functions $F_1(z|y_{t-2} = y)$ of $y_{t-1}|y_{t-2} = y$ and $F_2(z|y_{t-2} = y)$ of $y_t|y_{t-2} = y$, defined through

$$\widehat{F}_1(z|y_{t-2} = y) = \sum_{t=3}^{T} w_{tT}(y)\mathbf{1}_{(-\infty,z]}(y_{t-1}) \tag{4}$$

$$\widehat{F}_2(z|y_{t-2} = y) = \sum_{t=3}^{T} w_{tT}(y)\mathbf{1}_{(-\infty,z]}(y_t)\,, \tag{5}$$

where $w_{tT}(y)$ are the kernel weights with bandwidth parameter h_T

$$w_{tT}(y) = \frac{K\left(\dfrac{y_{t-2} - y}{h_T}\right)}{\displaystyle\sum_{t=3}^{T} K\left(\frac{y_{t-2} - y}{h_T}\right)}\,. \tag{6}$$

The function $K : I\!R \to I\!R$ is a kernel function, i.e., a nonnegative integrable function on $I\!R$.

As mentioned above, local kernel M–type estimates, $\widehat{\phi}_1(y) = \widehat{\phi}_{1,\mathrm{M}}(y)$ and $\widehat{\phi}_2(y) = \widehat{\phi}_{2,\mathrm{M}}(y)$, defined through a score function ψ, can be considered. Possible choices of the score function ψ are the Huber or $\psi(u) = \mathrm{sg}(u)$ that leads to the local medians. As noted above, these estimates can be viewed as applying the robust M–location conditional functional to the empirical conditional distributions $\widehat{F}_1(z|y_{t-2} = y)$ and $\widehat{F}_2(z|y_{t-2} = y)$ and so they are the solution of

$$\sum_{t=3}^{T} K\left(\frac{y_{t-2} - y}{h_T}\right)\psi\left(\frac{y_{t-1} - \widehat{\phi}_1(y)}{s_1(y)}\right) = 0 \tag{7}$$

$$\sum_{t=3}^{T} K\left(\frac{y_{t-2} - y}{h_T}\right)\psi\left(\frac{y_{t-2} - \widehat{\phi}_2(y)}{s_2(y)}\right) = 0 \tag{8}$$

with $s_j(y)$ the residual scale. For a discussion regarding the choice of the score function leading to the robust location conditional functionals, see He *et al.* (2002). Furthermore, He *et al.* (2002) comment on how the choice of the score function ψ is directly related to the question of what the practitioner is estimating : *"Without imposing assumptions on the distribution of the errors, we need to understand what an M–estimator estimates. For example, the least squares method estimates the conditional mean ... and the least absolute deviation estimator is the conditional median. ... So, our choice of ψ has to depend on what we are interested in."* As noted by these authors, if we are concerned with a conditional distribution with heavy-tails, the conditional median is generally the summary of choice, in which case, the $\psi(u) = \text{sg}(u)$ is the natural choice. On the other hand, if the conditional distribution is assumed to be symmetric, the conditional distribution has a natural center, the conditional median, so any odd function ψ will give a consistent estimator of the conditional median.

As described in Step 2, once we have obtained robust estimates, $\widehat{\phi}_1(y)$ and $\widehat{\phi}_2(y)$, of $\phi_1(y)$ and $\phi_2(y)$ the robust estimation of the regression parameter can be perfomed by applying to the residuals $\widehat{r}_t = y_t - \widehat{\phi}_2(y_{t-2})$ and $\widehat{z}_t = y_{t-1} - \widehat{\phi}_1(y_{t-2})$, any of the robust methods proposed for linear regression. Note that in these models a difference appears with respect to linear autoregressive models. Model (2) is the counterpart of an AR (2), in which, as is well known, three residuals may be spoiled by an isolated outlier at time t_o. In the estimation procedure described for partly linear autoregression models, the use of a robust smoothing will control the influence of y_{t_o} in $\widehat{\phi}_1(y)$ or $\widehat{\phi}_2(y)$, if more than three points lie at the neighborhood of $y = y_{t_o-1}$ or $y = y_{t_o-2}$, respectively. Thus, only one huge observation appears among all residuals $\widehat{r}_t$ and only another one among $\widehat{z}_t$. Therefore, in our transformed regression model, the point y_{t_o} will yield to the following two outlying data: $(\widehat{r}_{t_o}, \widehat{z}_{t_o})$ and $(\widehat{r}_{t_o+1}, \widehat{z}_{t_o+1})$, except for isolated points. Among the most popular robust regression estimates, we find GM–estimators, which control both high residuals and high leverage points and that have high breakdown point in simple regression. Also, the LMS–estimator (least median of squares) (Rousseeuw and Leroy, 1987), the MM (Yohai, 1987) or $\tau-$estimates could be used (Yohai and Zamar, 1988).

In this paper, we will focus on the behavior of the robust autoregression estimator defined as the solution of

$$\sum_{t=3}^{T} \psi_1 \left(\frac{\widehat{r}_t - \widehat{\beta}\widehat{z}_t}{s_T} \right) w_2\left(\widehat{z}_t\right) \widehat{z}_t \; w_3\left(y_{t-2}\right) = 0, \tag{9}$$

where $\widehat{r}_t = y_t - \widehat{\phi}_2(y_{t-2})$, $\widehat{z}_t = y_{t-1} - \widehat{\phi}_1(y_{t-2})$ and s_T is an estimate of the residuals scale σ_o and with score function ψ_1 and weight functions w_2 and w_3. This estimator is a slight modification of that considered in Bianco and Boente (2002). The weight function w_3 is introduced to prevent from the effect of large values of y_{t-2}, which correspond to isolated points where the estimation of the robust location conditional functionals ϕ_1 and ϕ_2 is a difficult issue, as discussed above. Let F be the joint distribution of (y_t, y_{t-1}, y_{t-2}). The functional $\beta(F)$ related to the estimator defined in (9) is the solution of

$$E_F \left[\psi_1 \left(\frac{r_t - \beta(F) z_t}{\sigma_o} \right) w_2\left(z_t\right) z_t \; w_3\left(y_{t-2}\right) \right] = 0, \tag{10}$$

with $r_t = y_t - \phi_2(y_{t-2})$ and $z_t = y_{t-1} - \phi_1(y_{t-2})$. Note that the expectation involved in (10)

exists since, from conditions **N1**, **N3** and **N6** below, the score functions ψ_1 and $\psi_2(t) = t\, w_2(t)$ and of the weight function w_3 are bounded. Note that, using $g(y) = \phi_2(y) - \beta_o\phi_1(y)$, we have $y_t - \phi_2(y_{t-2}) = \beta_o(y_{t-1} - \phi_1(y_{t-2})) + \epsilon_t$, and so (10) is equivalent to

$$E_F\left[\psi_1\left(\frac{(\beta_o - \beta(F))\, z_t + \epsilon_t}{\sigma_o}\right) w_2\left(z_t\right) z_t\; w_3\left(y_{t-2}\right)\right] = 0.$$

Thus, due to the independence between the errors ϵ_t and the past observations, if the score function is stricly increasing, in order to get Fisher–consistent estimators, i.e., $\beta(F) = \beta_o$, one only needs to require $E_F\left[\psi_1\left(\epsilon_t/\sigma_o\right)\right] = 0$. This is a standard condition when dealing with robust estimators in linear regression and autoregression models. Moreover, when using M–spline estimators in partly linear regression model, this condition is analogous to assumption 2 (ii) of Gao and Shi (1997) and assumption 6 of He *et al.* (2002).

In this paper, we will study the asymptotic behavior of the autoregression estimates defined through Steps 1 to 3 and we will also propose a procedure to obtain detection residuals and robust predictors of the series. The paper is organized as follows. In Section 2, through an example, we illustrate the effect of the outliers on the estimation of β_o when using the classical estimate and the corresponding robust procedure. In Section 3, a procedure to detect outlying observations is derived. In Section 4, we derive the asymptotic distribution of the estimates of the autoregression parameter β_o. In Section 6, a similar result is stated uniformly over the smoothing parameter after introducing in Section 5 robust alternatives to choose the smoothing parameter. In this latter Section, through a Monte Carlo study, the performance of the different criteria is compared for normal and contaminated samples. Proofs are given in the Appendix.

2 The effect of outliers in the estimation

As mentioned in the Introduction, the sensitivity of the least squares estimates to a small fraction of outliers has been extensively described both in the purely parametric and in the nonparametric setting. For partly linear regression and autoregression models robust methods, less sensitive to wild spike outliers, are desirable. The treatment of outliers is an important task when one explores the main features of a data set, since anomalous observations may affect the recognition of the autoregression function when the estimation is based on a local average procedure. Moreover, outlier detection and robust prediction tools are also necessary.

To illustrate this behavior, we have considered the Canadian lynx data which has been extensively studied. This data set, which is the annual record of the number of Canadian lynx trapped in the Mc Kenzie River district of North–West Canada for the years 1821–1934, considers the variable $y_t = \log_{10}(\text{number of lynx trapped in the year}(1822 + t)) - 2.9036$, $1 \le t \le T = 114$. Several authors have studied this data set. Among others, we can mentioned Campbell and Walker (1977), Tong (1977) who fitted an AR(11) and an ARMA(3,3), Yao and Tong (1994) who selected as regressors y_{t-1}, y_{t-3} and y_{t-6}. Wong and Kohn (1996) used a second order autoregressive additive model, while Härdle, Liang and Gao (2000)

considered a partly linear autoregression model of order one. Moreover, Brillinger (1986) performed a sensitivity analysis, while Martin and Yohai (1986) proposed a filter to detect outliers.

We have artificially contaminated the data set replacing the largest observation y_{84} by -2.9036. Figures 1 and 2 show the behavior of the estimated functions both for the least squares and a robust procedure, for the original data set in black and the contaminated one, in red. The robust procedure is mainly unaffected. As expected, when using the classical estimates, not only the autoregression parameter changes, but also does the shape of the autoregression function, which decreases more slowly. Note that $\widehat{\beta}_{LS} = 1.355$ and $\widehat{\beta} = 1.383$, so the estimations are quite similar for the original data. On the other hand, for the contaminated data, $\widehat{\beta}_{LS} = 0.543$ and $\widehat{\beta} = 1.352$ illustrating the insensitivity to an anomalous observation of the robust procedure. We have also plotted in Figure 3 the estimated function g and the fitted or predicted values. When computing the robust predictors each observation received a weight according to the residuals of the iterative procedure leading to the estimation of the autoregression parameter. Besides, if the observation y_{t-1} received a low weight then, when predicting at time t, y_{t-1} was replaced by its fitted value. Details are given in Section 3. The lower plots in Figure 3, show the predicted values obtained using the least squares or the GM–estimators. Black lines correspond to the original data and red ones to the modified data. From these plots, it is clear that the LS predictors are modified not only at time $t = 85$, but also the influence of this outlying observation propagates all along the future. On the other hand, even though the robust procedure used is slightly sensitive to the outlier, it recovers quite soon the feature of the fitted series.

3 Outlier detection

Outlier detection in time series analysis is an important issue. As mentioned above, Brillinger (1986) performed a sensitivity analysis of lynx data, while Martin and Yohai (1986) proposed a filter to detect outliers. Combining the robust procedures with M–smoothers, we have defined predicted values for the series and detection residuals which allow to detect anomalous data.

For a given cutting point α, the procedure can be described as follows:

- Let $\widehat{\beta}$ be the estimate of the autoregression parameter introduced in (9), related to a score function ψ_1 and hard–rejection weights w_2 and w_3. Denote $\widehat{g}(y)$ the estimate of the autoregresion function g as described in Step 3.

- Compute $R_t = \dfrac{\widehat{r}_t - \widehat{\beta}\widehat{z}_t}{s_T}$, where $s_T = \dfrac{1}{0.6745}\text{median}(|\widehat{r}_t - \widehat{\beta}\widehat{z}_t|)$.

- Define the predicted value $\widehat{y}_t$ as

$$\widehat{y}_t = \begin{cases} \left.\begin{array}{ll} \widehat{\beta} y_{t-1} + \widehat{g}\left(y_{t-2}\right) & \text{if } |R_t| < \alpha \text{ and } w_2(\widehat{z}_t) > 0 \\ \widehat{\beta}\widehat{y}_{t-1} + \widehat{g}\left(y_{t-2}\right) & \text{otherwise} \end{array}\right\} & \text{if } w_3(y_{t-2}) > 0 \\ \\ \left.\begin{array}{ll} \widehat{\beta} y_{t-1} + \widehat{g}\left(\widehat{y}_{t-2}\right) & \text{if } |R_t| < \alpha \text{ and } w_2(\widehat{z}_t) > 0 \\ \widehat{\beta}\widehat{y}_{t-1} + \widehat{g}\left(\widehat{y}_{t-2}\right) & \text{otherwise} \end{array}\right\} & \text{if } w_3(y_{t-2}) = 0 \ . \end{cases}$$

- Define the detection residual value $\widetilde{r}_t$ as

$$\widetilde{r}_t = \begin{cases} 0 & \text{if } |R_t| < \alpha \ , \ w_2(\widehat{z}_t) > 0 \text{ and } w_3(y_{t-2}) > 0 \\ y_t - \widehat{y}_t & \text{otherwise} \ . \end{cases} \tag{11}$$

In order to compare our procedure with the method proposed by Bianco, García Ben, Martínez and Yohai (1996) for ARMA(p,q), we have computed the filtered values using the library rr as implemented in S-Plus with an ARMA(3,3) model. We call this analysis the robust ARMA(3,3). Figure 4 plots the residuals from the robust ARMA(3,3) and the detection residuals $\widetilde{r}_t$ defined in (11), with $\alpha = 0.2$ and as M–smoother the local median. For "good" data points detection residuals are zero, while suspicious observations correspond to non-zero residuals. The nonzero residuals detected by our procedure indicate nearly the same anomalous data points as those revealed by Brillinger (1986)'s plot, while, as shown by Figure 4, the robust ARMA(3,3) analysis detects also a level shift. Note that our procedure shows that most suspicious data are not isolated, revealing that some moving average structure in the errors is necessary in this partly linear autoregression model. The analysis of moving average errors for partly linear autoregression models is beyond the scope of this paper.

4 Asymptotic distribution

Conditions for the consistency of the robust procedure defined through Step 1 to 3, are analogous to those stated in Bianco and Boente (2004) and the strong convergence result follows easily using the Ergodic Theorem instead of the Strong Law of Large Numbers.

In this Section, we will derive the asymptotic distribution of the regression parameter estimates defined as any solution of (9) with $\widehat{\phi}_1(y)$ and $\widehat{\phi}_2(y)$ consistent estimates of robust location conditional functionals $\phi_1(y)$ and $\phi_2(y)$, satisfying $\phi_2(y) = \beta_o \phi_1(y) + g(y)$.

Let ψ_1 be a score function and w_2 and w_3 be weight functions. For the sake of simplicity and without loss of generality, we will assume that the residuals scale is known and equals σ_o, i.e., we will consider the solution $\widehat{\beta}$ of

$$\sum_{t=3}^{T} \psi_1\left(\frac{\widehat{r}_t - \widehat{\beta}\widehat{z}_t}{\sigma_o}\right) w_2\left(\widehat{z}_t\right) \widehat{z}_t \ w_3\left(y_{t-2}\right) = 0 \ . \tag{12}$$

If σ_o is estimated by s_T, asymptotic normality can be derived by requiring that $s_T \xrightarrow{p} \sigma_o$, if, in addition, $t^2\psi_1''(t)$ is bounded.

As in the Introduction, denote $r_t = y_t - \phi_2(y_{t-2})$ and $z_t = y_{t-1} - \phi_1(y_{t-2})$. Thus, $r_t - \beta_o z_t = \epsilon_t$ defined in (2).

We will need the following set of assumptions.

N0. The process $\{y_t : t \geq 3\}$ is a strictly stationary $\alpha-$mixing process with geometric mixing coefficients $\alpha(n)$. (see, Rosenblatt (1956))

N1. ψ_1 is an odd, bounded and twice continuously differentiable function with bounded derivatives ψ_1' and ψ_1'', such that $\varphi_1(t) = t\psi_1'(t)$ and $\varphi_2(t) = t\psi_1''(t)$ are bounded.

N2. $E\left(w_2(z_t)z_t^2\right) < \infty$ and

$$A = E\left(\psi_1'\left(\frac{\epsilon_t}{\sigma_o}\right) w_2\left(z_t\right) z_t^2 \; w_3\left(y_{t-2}\right)\right) = E\left(\psi_1'\left(\frac{\epsilon_t}{\sigma_o}\right)\right) E\left(w_2\left(z_t\right) z_t^2 \; w_3\left(y_{t-2}\right)\right) > 0 \; .$$

N3. $w_2(u) = \psi_2(u)\, u^{-1} > 0$ is a bounded function, Lipschitz of order 1. Moreover, ψ_2 is also a bounded and continuously differentiable function with bounded derivative ψ_2', Lipschitz of order 1 and such that $\lambda_2(t) = t\psi_2'(t)$ is bounded.

N4. $E\left(\psi_2(z_t)|y_{t-2} = y\right) = 0$ for almost all y.

N5. The functions $\phi_j(y)$, $j = 1, 2$ are continuously differentiable.

N6. The function w_3 is a bounded function with $\|w_3\|_\infty \leq 1$ and compact support $\mathcal{K} \subseteq \mathcal{S}$, where $\mathcal{S}$ denotes the support of the marginal distribution of y_t.

Remark 4.1.

- With respect to **N0**, the inclusion of a dependence structure, usually imposing a mixing condition, allows to estimates the autoregression function through a kernel smoother. Roughly speaking, all the mixing conditions say that the dependence between the random variables is weaker the further they are apart. The α-mixing or strong mixing condition introduced by Rosenblatt (1956) is one of the weakest notions where nonparametric inference has been considered. Classical ARMA processes are strongly mixing with geometrical coefficients.

  As it is well known, the concept of $\alpha-$mixing is weaker than that of $\varphi-$mixing (uniform strongly mixing), which is a more often studied condition (see, Billingsley 1968). The $\varphi-$condition is rather restrictive when we are considering autoregressive models, since for Gaussian stationary processes the $\varphi-$mixing condition is equivalent to $m-$dependence (see Ibragimov and Linnik, 1971). When considering a fully nonparametric autoregression model, most of the asymptotic results for the Nadaraya–Watson estimators and predictors have been obtained assuming a $\varphi-$ or $\alpha-$mixing condition, see for instance, Bosq (1996), Györfi, Härdle, Sarda and Vieu (1989) and Härdle (1990), for a review. Asymptotic normality results of the Nadaraya–Watson estimates, for $\alpha-$mixing processes, were obtained by Robinson (1983). An $\alpha-$mixing condition was

also required when studying the asymptotic behavior of local $M-$estimates of the autoregression function (see Robinson, 1984).

The same mixing conditions were considered under partly linear autoregressive models. More precisely, if the process satisfies an $\alpha-$mixing condition, Gao and Liang (1995) derived the asymptotic distribution of the autoregression parameter when considering local polynomials, while Gao (1995) established the asymptotic normality and obtained a law of iterated logarithm for the linear kernel–based estimates; see also Gao and Yee (2000).

Doukhan (1994, Theorem 7, page 102) gives sufficient conditions on the function g, the autoregression parameter and on the errors distribution that guarantee that the process will be $\alpha-$mixing. For instance, if g is bounded and the errors ϵ_t have a density and finite first moment, then the condition $|\beta| < 1$ entails that the process is geometrically ergodic and thus, $\alpha-$mixing. When g is unbounded it should be required that there exist some positive constants b and v_o and some $a \geq 0$, such that $|g(v)| \leq a|v| - b$ for $|v| > v_o$ $\sup_{|v|\leq v_o} |g(v)| < \infty$ and the unique nonnegative zero of the polynomial $P(z) = z^2 - |\beta|z - a$, i.e, $\rho = \dfrac{|\beta| + \sqrt{\beta^2 + 4a}}{2}$, satisfies $\rho < 1$.

- As noted by Robinson (1988), condition **N2** will prevent any element of y_{t-1} from being a.s. perfectly predictable by y_{t-2}. It is worth noticing that if the errors have symmetric distribution and g and ψ_2 are odd functions, condition **N4** is fulfilled. Assumption **N4** is needed in order to obtain a uniform result over a class of Lipschitz functions, using the results given in Arcones (1996). For VC–classes of functions, Andrews and Pollard (1994) and Yu (1994) provided a similar result for strong mixing triangular arrays and for stationary $\alpha-$mixing sequences, respectively.

- The smoothness condition **N5** is a standard requirement in classical kernel estimation in semiparametric models in order to guarantee asymptotic normality, see for instance, Robinson (1988) and Severini and Wong (1992).

It is worthwhile noticing that no moment conditions are required to the errors distribution to derive the asymptotic distribution of the autoregression parameter.

Theorem 4.1. *Let $\{y_t\,,\ j \geq 3\}$ be a stationary $\alpha-$mixing process satisfying (2) with ϵ_t independent of $\{y_{t-j}\,, j \geq 1\}$ with symmetric distribution. Moreover, assume that the mixing coefficients are geometric. Denote $r_t = y_t - \phi_2(y_{t-2})$ and $z_t = y_{t-1} - \phi_1(y_{t-2})$ where $\phi_1(y)$ and $\phi_2(y)$ are robust location conditional functionals satisfying $\phi_2(y) = \beta_o\phi_1(y) + g(y)$. Assume that **N0** to **N4** and **N6** hold. Moreover, assume that one of the following two assumptions a) or b) are satisfied*

*a) **N5** holds and $\widehat{\phi}_j(y)$ are robust estimates of $\phi_j(y)$ such that, for $j = 1, 2$, $\widehat{\phi}_j(y)$ is continuously differentiable and*

$$T^{\frac{1}{4}} \sup_{y\in\mathcal{K}} |\widehat{\phi}_j(y) - \phi_j(y)| \quad \xrightarrow{p} \quad 0\,, \tag{13}$$

$$\sup_{y\in\mathcal{K}} |\widehat{\phi}'_j(y) - \phi'_j(y)| \quad \xrightarrow{p} \quad 0\,, \tag{14}$$

where $\mathcal{K}$ is defined in ***N6****.*

b) $\widehat{\phi}_j(y)$ *are robust estimates of* $\phi_j(y)$ *which admit a linear expansion* $\widehat{\phi}_j(y) - \phi_j(y) = \widehat{\mathcal{L}}_j(y) + \widehat{\mathcal{R}}_j(y)$*, where*

$$\widehat{\mathcal{L}}_j(y) = \sum_{t=3}^{T} w_{tT}(y)\ \nu_j\left(y_{t-2+j}, y\right)\ ,$$

with ν_j *bounded functions, such that* $E\left(\nu_j\left(y_{t-2+j}, y\right)|y_{t-2} = y\right) = 0$ *almost everywhere. Moreover, assume that for* $j = 1, 2$

$$T^{\frac{1}{4}} \sup_{y\in\mathcal{K}} |\widehat{\mathcal{L}}_j(y)| \ \xrightarrow{p}\ 0, \tag{15}$$

$$T^{\frac{1}{2}} \sup_{y\in\mathcal{K}} |\widehat{\mathcal{R}}_j(y)| \ \xrightarrow{p}\ 0, \tag{16}$$

$$T^{-\frac{1}{2}} \left| \sum_{t=3}^{T} \widehat{\mathcal{L}}_j(y_{t-2})\, \vartheta_1\left(\epsilon_t\right)\ \vartheta_2\left(z_t\right)\ w_3\left(y_{t-2}\right) \right| \ \xrightarrow{p}\ 0 \tag{17}$$

hold for bounded functions ϑ_1 *and* ϑ_2 *such that* $E(\vartheta_1(\epsilon_t)) \times E\left(\vartheta_2(z_t)|y_{t-2} = y\right) = 0$*, for almost all* y*, where* $\mathcal{K}$ *is defined in* ***N6****.*

Then,

$$T^{\frac{1}{2}}\left(\widehat{\beta} - \beta_o\right) \xrightarrow{D} N\left(\mathbf{0}, \sigma^2_{\psi_1,w_2,w_3}\right)\ ,$$

where A *is defined in* ***N2*** *and*

$$\begin{aligned}
\sigma^2_{\psi_1,w_2,w_3} &= A^{-2}\sigma_o^2\, E\left(\psi_1^2\left(\frac{\epsilon_t}{\sigma_o}\right)\right)\, E\left(w_2^2\left(z_t\right) z_t^2\ w_3^2\left(y_{t-2}\right)\right) \\
&= \sigma_o^2\, \frac{E\left(\psi_1^2\left(\frac{\epsilon_t}{\sigma_o}\right)\right)}{\left[E\left(\psi_1'\left(\frac{\epsilon_t}{\sigma_o}\right)\right)\right]^2}\, \frac{E\left(w_2^2\left(z_t\right) z_t^2\ w_3^2\left(y_{t-2}\right)\right)}{\left[E\left(w_2\left(z_t\right) z_t^2\ w_3\left(y_{t-2}\right)\right)\right]^2}\ .
\end{aligned}$$

Remark 4.2. When $w_2 \equiv 1$ and $w_3 \equiv 1$, the asymptotic efficiency of the autoregression estimates is the same as in the one dimensional location setting, that is $V(\psi_1) = E\left(\psi_1^2\left(\epsilon_t/\sigma_o\right)\right)\left[E\left(\psi_1'\left(\epsilon_t/\sigma_o\right)\right)\right]^{-2}$.

Remark 4.3. Conditions (13) and (14) are related to the trade–off between the stochastic equicontinuity needed to derive the asymptotic distribution of $\widehat{\beta}$ and the smoothness requirements on the estimators $\widehat{\phi}_j$, when plugging–in general preliminary estimators of ϕ_j. For a highlighting discussion on this task we refer to Section 4.3 in Andrews (1994). It is worth noticing that, even in the independent setting, when dealing with semiparametric models, derivability of the estimates of the nuisance parameters together with their uniform convergence is usually required, see for instance, Severini and Wong (1992) and Severini and Staniswalis (1994). As it will be discussed below, the uniform convergence rates required in (13) and (14) are fulfilled when we consider, in Step 1, local kernel M–type estimates solutions of (7) and (8) if the optimal bandwidth is used. The convergence requirements in a) are analogous to those required in Condition (7) in Severini and Staniswalis (1994, p. 510)

and are needed in order to obtain the desired rate of convergence for the autoregression estimates. More precisely, assumption (13) avoids the bias term and ensures that $G_T(\widehat{\phi}_1, \widehat{\phi}_2)$ will behave asymptotically as $G_T(\phi_1, \phi_2)$, where for any $\beta \in I\!R$ and any differentiable functions $\upsilon_j : I\!R \to I\!R$, $j = 1, 2$

$$G_T(\upsilon_1, \upsilon_2) = \frac{1}{\sqrt{T-2}} \sum_{t=3}^{T} \psi_1 \left(\frac{y_t - \upsilon_2(y_{t-2}) - \beta_o (y_{t-1} - \upsilon_1(y_{t-2}))}{\sigma_o} \right) \psi_2 (y_{t-1} - \upsilon_1(y_{t-2})) \, w_3(y_{t-2}) \, .$$

Assumption b) avoids equicontinuity arguments by requiring a linear approximation to the estimators of ϕ_1 and ϕ_2 which allows to deal with the reminder terms as in the classical setting, i.e., when using the linear kernel estimators. However, under assumption a) Theorem 4.1 includes other estimators than those based on kernel weights.

Remark 4.3. When, in Step 1, we consider local kernel M–type estimates, $\widehat{\phi}_1(y) = \widehat{\phi}_{1,\mathrm{M}}(y)$ and $\widehat{\phi}_2(y) = \widehat{\phi}_{2,\mathrm{M}}(y)$, solution of (7) and (8), both assumptions a) and b) in Theorem 4.1 are fulfilled, under mild conditions.

To be more precise, if

i) the kernel K is a bounded density function, Lipschitz continuous, such that $|u|^2 K(u)$ is bounded

ii) ψ is an odd, strictly increasing, bounded and continuously differentiable function such that $u\psi'(u) \leq \psi(u)$

iii) the marginal density f of y_t is a bounded function such that $\inf_{y \in \mathcal{K}} f(y) > 0$

iv) the conditional distribution functions $F_1(z|y_{t-2} = y)$ of $y_{t-1}|y_{t-2} = y$ and $F_2(z|y_{t-2} = y)$ of $y_t|y_{t-2} = y$ are uniformly Lipschitz in a neighborhood $\mathcal{K}^\epsilon$ of $\mathcal{K}$, i.e., there exists a positive constant C such that

$$|F_j(z|y_{t-2} = y) - F_j(z|y_{t-2} = v)| < C|y - v|$$

for all $z \in I\!R$, $y, v \in \mathcal{K}^\epsilon$.

v) Moreover, the following equicontinuity condition hold for $j = 1, 2$

$$\forall \epsilon > 0 \; \exists \delta > 0 : |u - z| < \delta \Rightarrow \sup_{y \in \mathcal{K}} |F_j(u|y_{t-2} = y) - F_j(z|y_{t-2} = y)| < \epsilon$$

analogous arguments to those used in Boente and Fraiman (1991, a) allow to show that (13) holds for the optimal bandwidth of order $T^{-\frac{1}{5}}$. Furthermore, if the kernel K and the scale functions s_j have continuous derivatives K' and s_j', respectively and if we denote by $\eta(u) = u \, \psi'(u)$, we have that

$$\widehat{\phi}_1'(y) = \; - \; \frac{\dfrac{s_1(y)}{h_T} \displaystyle\sum_{t=3}^{T} K'\left(\frac{y_{t-2} - y}{h_T}\right) \psi'\left(\frac{y_{t-1} - \widehat{\phi}_1(y)}{s_1(y)}\right) - s_1'(y) \sum_{t=3}^{T} K\left(\frac{y_{t-2} - y}{h_T}\right) \eta\left(\frac{y_{t-1} - \widehat{\phi}_1(y)}{s_1(y)}\right)}{\displaystyle\sum_{t=3}^{T} K\left(\frac{y_{t-2} - y}{h_T}\right) \psi'\left(\frac{y_{t-1} - \widehat{\phi}_1(y)}{s_1(y)}\right)}$$

and similarly for $\widehat{\phi}_2'(y)$. These expressions suggest that if i) to iv) hold, the proof of (14) parallels the proofs given in Härdle and Gasser (1985) and in Boente and Rodriguez (2006) together with the standard arguments used in the $\alpha-$mixing case.

On the other hand, using a Taylor's expansion, it is easy to see that an M–estimator admits the linear expansion given in b), where the remainder term satisfies (16), since, as mentioned above, M–estimators satisfy $T^{\frac{1}{4}} \sup_{y \in \mathcal{K}} |\widehat{\phi}_j(y) - \phi_j(y)| \xrightarrow{p} 0$ when ϕ_j are continuously differentiable functions. Note that this approach avoids the derivability requirements on ϕ_j, $\widehat{\phi}_j$, the kernel K and the scale functions s_j needed to guarantee a).

5 Resistant choice of the smoothing parameter

The sensitivity to outliers of the classical methods for the selection of the smoothing parameter has been widely discussed for independent observations in nonparametric regression. Because it is based on squared residuals, least squares cross–validation is very sensitive to outliers, even when it is used with local M–estimates. As noted by Wang and Scott (1994), in the presence of outliers, the least squares cross–validation function is nearly constant on its whole domain and thus, essentially worthless for the purpose of choosing a bandwidth. Moreover, it can be seen that just one outlier may cause the bandwidth (and so the estimate) to break down, in the sense that it often results in oversmoothing or undersmoothing. Boente and Fraiman (1991, b) pointed out that robust cross–validation methods should be an alternative. Also, Wang and Scott (1994) proposed an L^1 cross–validation method in order to avoid the problems of L^2 cross–validation, while Cantoni and Ronchetti (2001) considered a resistant choice of the smoothing parameter for smoothing splines based on a robust version of C_p and of cross-validation. A similar proposal was suggested by Leung, Marriott and Wu (1993) for kernel M-smoothers. On the other hand, the classical plug–in bandwidth selector also breaks down in the presence of outliers. Boente, Fraiman and Meloche (1997) proposed a robust plug–in bandwidth selection procedure in noparametric regression.

To make explicit the dependence on the bandwidth parameter h, let us denote, from now on, $\widehat{\beta}_h$ and $\widehat{g}_h$ the estimates computed using the kernel weights (6) with smoothing parameter h. As mentioned by Härdle, Liang and Gao (2000), in the setting of partial linear autoregression models, the optimal bandwidth involves functionals of the unknown underlying distribution. These authors considered the average square error as measure of the goodness of the estimates $\widehat{\beta}_h$ and $\widehat{g}_h$. For each bandwidth h they defined

$$\begin{aligned} D_1(h) &= \frac{1}{T-2} \sum_{t=3}^{T} \left(\left\{ \widehat{\beta}_h \, y_{t-1} + \widehat{g}_h \left(y_{t-2} \right) \right\} - \left\{ \beta_o \, y_{t-1} + g \left(y_{t-2} \right) \right\} \right)^2 w \left(y_{t-2} \right) \\ &= \frac{1}{T-2} \sum_{t=3}^{T} u_t^2(h) w \left(y_{t-2} \right) \ , \end{aligned}$$

where the weight function w protects against boundary effects. The cross-validation criterion they have considered to construct an asymptotically optimal data–driven bandwidth and

thus, adaptive data–driven estimates, is defined through

$$C_1(h) = \frac{1}{T-2}\sum_{t=3}^{T}\left(y_t - \left\{\widetilde{\beta}_h\, y_{t-1} + \widehat{g}_{h,t}\,(y_{t-2})\right\}\right)^2 w\,(y_{t-2}) = \frac{1}{T-2}\sum_{t=3}^{T}\widehat{u}_t^2(h) w\,(y_{t-2})\ ,$$

where $\widehat{g}_{h,t}(y) = \widehat{\phi}_{2,t}(y) - \widetilde{\beta}_h\widehat{\phi}_{1,t}(y)$, $\widehat{\phi}_{1,t}(y)$ and $\widehat{\phi}_{2,t}(y)$ are the linear smoothers obtained with all the data except y_{t-2} and $\widetilde{\beta}_h$ is the estimator obtained by least squares considering the residuals $y_t - \widehat{\phi}_{2,t}(y_{t-2})$ and $y_{t-1} - \widehat{\phi}_{1,t}(y_{t-2})$.

A small simulation study was carried on to show that the asympotically optimal bandwidth is very sensitive to outliers. For each value of h, we have computed an estimate of $MSE(h) = E(D_1(h))$, with $w \equiv 1$, by replicating over samples, both for the classical estimator and for the M–smoother combined with a GM–estimator. We have considered a kernel smoother with the gaussian kernel with standard deviation 0.37 such that the interquartile range is 0.5, both for the least squares estimates and for the local M–estimate with bisquare score function. The tuning constant for the local M–estimator is 4.685, which gives a 95% efficiency with respect to its linear relative. Local M–estimates were computed through an iterative procedure with local medians as initial points. After the robust smoothing, GM–estimates with Huber function on the residuals with constant 1.6 and bisquare weights on $y_{t-1} - \widehat{\phi}_1\,(y_{t-2})$ with constant 5.57 were computed. This choice of the tuning constants gives aproximately a numerically computed 95% asymptotic efficiency under normal errors, for the considered model, with respect to the least squares estimate. We performed 50 replications. In order to stabilize the series, we first generate a series of size $N = 1100$ following the model

$$z_t = \beta_o z_{t-1} + 0.25\,\pi\,\sin(\pi z_{t-2}) + \epsilon_t \quad 3 \le t \le N\,,$$

where $\beta_o = 0.25$. As initial values, we took $z_t = \epsilon_t$, for $1 \le t \le 2$. In the case of normal errors, we have chosen $\epsilon_t \sim N(0, \sigma_o^2)$ with $\sigma_o^2 = 0.25$. The data set of size $T = 100$ to be considered consists on the series $\{y_t : 1 \le t \le T\}$, where in the non–contaminated case $y_t = z_{t+1000}$. The contaminated data set corresponds to additive outliers in the series, as follows: ϵ_t, $1 \le t \le 1100$, are i.i.d. $N(0, \sigma_o^2)$ and $y_t = z_{t+1000} + 6\delta_t$ with $\delta_t \sim Bi(1, 0.05)$. The bandwidth h was choosen on a grid of 50 equidistant points between 0.05 and 1.

As can be seen in Figure 5 the shape of the curve is highly influenced by anomalous data and the minimum is highly modified when introducing outliers, since it changes from 0.4 to almost 0.7, both for the least squares and for the GM–estimator.

This suggests that resistant procedures should also be introduced in this context. By analogy with the least median of squares, we can consider the following measures

$$D_2(h) = \underset{3\le t\le T}{\text{median}}\left\{u_t^2(h) w\,(y_{t-2})\right\} \quad \text{and} \quad C_2(h) = \underset{3\le t\le T}{\text{median}}\left\{\widehat{u}_t^2(h) w\,(y_{t-2})\right\}\ .$$

In the right panels of Figure 5, we plot the estimates of $MedSE(h) = E(D_2(h))$ obtained by replicating over samples. These plots show the stability of the criterion, since the minimum value is reached at almost the same value for the GM–estimator, while the least squares estimator is still sensitive. Note that the minimum value of the curve obtained for the classical estimator is shifted to the right, leading to undersmoothing.

Another approach can be to replace the square function in $D_1(h)$ and $C_1(h)$ by a ρ function as Huber or Tukey's function, after scaling the differences, i.e.,

$$D_3(h) = \frac{\sigma_T^2(h)}{T-2}\sum_{t=3}^{T}\rho\left(\frac{u_t(h)}{\sigma_T(h)}\right)w\left(y_{t-2}\right) \quad \text{and} \quad C_3(h) = \frac{\widehat{\sigma}_T^2(h)}{T-2}\sum_{t=3}^{T}\rho\left(\frac{\widehat{u}_t(h)}{\widehat{\sigma}_T(h)}\right)w\left(y_{t-2}\right) \ ,$$

where $\sigma_T(h) = \text{MAD}\left(u_t(h)\right)$ and $\widehat{\sigma}_T(h) = \text{MAD}\left(\widehat{u}_t(h)\right)$. The results obtained with Huber's function were disappointing and this is due to its unboundness. As expected, similar results to those obtained with Tukey's $\rho-$function, were obtained by weighting $u_t(h)$ with a Huber's weight function, which suggests that the measures defined through

$$D_4(h) = \frac{\sigma_T^2(h)}{T-2}\sum_{t=3}^{T}\psi^2\left(\frac{u_t(h)}{\sigma_T(h)}\right)w\left(y_{t-2}\right) \quad \text{and} \quad C_4(h) = \frac{\widehat{\sigma}_T^2(h)}{T-2}\sum_{t=3}^{T}\psi^2\left(\frac{\widehat{u}_t(h)}{\widehat{\sigma}_T(h)}\right)w\left(y_{t-2}\right) \ ,$$

could also be an alternative. Based on the stationarity of the process and taking into account that $D_1(h)$ tries to measure both bias and variance, it would make sense to introduce a new measure that establishes a trade–off between bias and variance. Then, we have defined measures based on a robust estimate of the bias, defined through a location estimate μ_T, and on a robust scale estimator σ_T, as follows,

$$\begin{aligned} D_5(h) &= \mu_T^2\left(u_t(h)w\left(y_{t-2}\right)\right) + \sigma_T^2\left(u_t(h)w\left(y_{t-2}\right)\right) \\ C_5(h) &= \mu_T^2\left(\widehat{u}_t(h)w\left(y_{t-2}\right)\right) + \sigma_T^2\left(\widehat{u}_t(h)w\left(y_{t-2}\right)\right) \ . \end{aligned}$$

We can consider as μ_T the median and as σ_T the bisquare a–scale estimate or the Huber $\tau-$scale estimate. Figure 6 shows the stability of this procedure combined with GM–estimators since, for the $\tau-$scale estimator, the minimum value is attained at the same value for both the contaminated and the normal samples. A similar plot was obtained for the a–scale estimate. The procedures based on a $\psi-$function also show a good performance.

In Table 1 we report the optimal values obtained through the simulation study. For the measures D_3 and D_4, we have considered the Huber's function, while for D_5 the Huber $\tau-$scale estimator as σ_T and the median as μ_T. Similar results were obtained using the Tukey's function and the a–scale. This table shows the advantage of using D_5 over the other procedures.

	Normal Data					Contaminated Data				
	D_1	D_2	D_3	D_4	D_5	D_1	D_2	D_3	D_4	D_5
LS	0.395	0.367	–	–	–	0.709	0.108	–	–	–
GM	0.399	0.380	0.380	0.360	0.399	0.670	0.360	0.418	0.399	0.399

Table 1: Optimal Asymptotic bandwidth for the autoregression function

Based on these results we conducted a simulation study to compare the performance of the five cross–validation criteria. Samples of size $T = 100$ were generated as described above. We have choosen the optimal bandwidth by minimizing $C_j(h)$ over a grid of 50 equidistant points between 0.05 and 1 in the non–contaminated case and in the contaminated one we

took a grid of 256 points between 0.05 and 5, so that the distance among values was the same as in the normal case. The weight function w was selected in two ways: $w \equiv 1$ and

$$w(y_{t-2}) = \begin{cases} 1 & \text{if} \quad \left| \dfrac{y_{t-2} - m_y}{s_y} \right| < 3 \\ 0 & \text{otherwise} , \end{cases}$$

with $m_y = \text{median}_t(y_t)$ and $s_y = \text{MAD}(y_t)$. Moreover, as described by Chu and Marron (1991), Hart and Vieu (1990) and Hart (1996), cross-validation under dependence can show a bias for small samples. For that reason, they modified the leave–out technique involved in the cross–validation method and they proved that, if the leave–out sequence, ℓ_T, does not increase too fast the bandwidth that minimizes the cross–validation criterion is asymptotically optimal. Thus, for all the criteria, we have also compute the cross-validation bandwidth with $\ell_T = 2$. Therefore, for each criterion C_j we obtained 4 values for the bandwidth corresponding to the two choices of w and to $\ell_T = 0$ and $\ell_T = 2$. We plot the results for C_1, when we use the least squares estimate. Since for the GM–estimator, the sensitivity of least squares cross–validation is well known, we have only considered C_2, C_4 and C_5. In C_4, we used as $\psi-$function the Huber function, with constant 1.345, while in C_5 the $\tau-$scale estimator was considered. In the plots, we label the results according to the criterion used to select the bandwidth. Once the bandwidth has been computed, the data–driven estimates of β_o and g were calculated.

Figures 7 and 8 show the boxplots of the obtained values of h. In the second one, the range of values in the vertical axis was truncated to make comparisons easier. As expected the L^2-criterion, C_1, is very sensitive to outliers. The cross–validation criterion based on the median, i.e., C_2, tends to provide smaller bandwidths than the classical cross–validation under normal errors when $w \equiv 1$ and $\ell_T = 0$. The best criterion is, in all cases, C_5. For this particular model, by weighting we obtain a smaller dispersion and the classical procedure performs better even under contamination. Leaving–out one data or taking $\ell_T = 2$, produce similar results. More research should be done in this direction to find a way to select the leaving sequence.

Figure 9 shows the boxplots of the data–driven estimators of β_o. These plots show the GM–estimators obtained using C_2 and $\ell_T = 2$ perform better than those computed with $\ell_T = 0$. This can be explained by the dependence structure that produces smaller data–driven bandwidths in this situation. Again, the best performance is obtained by the criterion based on the $\tau-$scale estimator. The behavior of the estimators of the g function was evaluated computing at each replication

$$M\left(\widehat{g}, g\right) = \underset{3 \le t \le T}{\text{median}} \left(\left[\widehat{g}(y_t) - g(y_t) \right]^2 \right) .$$

Figure 10 shows the estimates of the density of $M\left(\widehat{g}, g\right)$. A density kernel estimate with bandwidth 0.02 was computed in all cases, except for the classical estimates under contamination, where due to a different range of values, we took 0.05. Again, the $\tau-$scale estimator shows its advantage over the other criteria.

6 Uniform Asymptotic Distribution

In most practical situations data–driven estimators of β_o are computed. In this Section, we will consider the case where the robust smoothers computed in Step 1 are obtained using the kernel weigths defined in (6). In the classical setting, the optimal bandwidth, which minimizes $D_1(h)$, has order $T^{-\frac{1}{5}}$. If we denote by $\widehat{h}_1 = \underset{h\in\mathcal{H}_T}{\operatorname{argmin}} C_1(h)$, where $\mathcal{H}_T = [aT^{-\frac{1}{5}-c}, bT^{-\frac{1}{5}+c}]$ with $0 < a < b < \infty$ and $0 < c < \frac{1}{20}$, it has been shown (see, for instance, Härdle, Liang and Gao (2000)) that the bandwidth minimizing $C_1(h)$ is asymptotically optimal, in the sense that

$$\frac{D_1\left(\widehat{h}_1\right)}{\inf\limits_{h\in\mathcal{H}_T} D_1(h)} \xrightarrow{p} 1 \ .$$

This property suggests that results regarding the asymptotic behavior of the estimator $\widehat{\beta}\left(\widehat{h}\right)$ are needed, where $\widehat{h}$ denotes a bandwidth selector and $\widehat{\beta}\left(h\right)$ is the estimate of the parameter β_o obtained when the bandwidth h is used in the M-smoothing procedure. Several data–driven methods for choosing the bandwidth were discussed in Section 5 and we conjecture that, beyond their resistance to anomalous observations, they will lead to optimal bandwidths in the sense that if $\widehat{h}_j = \underset{h\in\mathcal{H}_T}{\operatorname{argmin}} C_j(h)$, $2 \leq j \leq 5$, then

$$\frac{D_j\left(\widehat{h}_j\right)}{\inf\limits_{h\in\mathcal{H}_T} D_j(h)} \xrightarrow{p} 1 \ .$$

That's why, in this Section, we will focus our attention to derive results regarding the asymptotic distribution of $T^{-\frac{1}{2}}\left(\widehat{\beta}\left(h\right) - \beta_o\right)$, uniformly over $h \in \mathcal{H}_T$, which will imply that, for $2 \leq j \leq 5$, the data–driven estimators $\widehat{\beta}(\widehat{h}_j)$ of β_o will be asymptotically normally distributed.

Theorem 6.1. *Let $\{y_t\,,\ j \geq 3\}$ be a stationary $\alpha-$mixing process satisfying (2) with ϵ_t independent of $\{y_{t-j}\,, j \geq 1\}$ with symmetric distribution. Moreover, assume that the mixing coefficients are geometric. Let $\mathcal{H}_T = [aT^{-\frac{1}{5}-c}, bT^{-\frac{1}{5}+c}]$ with $0 < a < b < \infty$ and $0 < c < \frac{1}{20}$. Denote $r_t = y_t - \phi_2(y_{t-2})$ and $z_t = y_{t-1} - \phi_1(y_{t-2})$ where $\phi_1(y)$ and $\phi_2(y)$ are robust location conditional functionals satisfying $\phi_2(y) = \beta_o\phi_1(y) + g(y)$. Let $w_{tT}(y)$ be the kernel weights defined in (6). Consider robust estimates of $\phi_j(y)$, $\widehat{\phi}_j(y)$, which admit a linear expansion $\widehat{\phi}_j(y) - \phi_j(y) = \widehat{\mathcal{L}}_j(y) + \widehat{\mathcal{R}}_j(y)$, where*

$$\widehat{\mathcal{L}}_j(y) = \sum_{t=3}^{T} w_{tT}(y)\ \nu_j\left(y_{t-2+j}, y\right)\ , \tag{18}$$

with ν_j bounded functions, such that $E\left(\nu_j\left(y_{t-2+j}, y\right)|y_{t-2} = y\right) = 0$ almost everywhere. Moreover, assume that for $j = 1, 2$

$$T^{\frac{1}{4}} \sup_{h\in\mathcal{H}_T} \sup_{y\in\mathcal{K}} |\widehat{\mathcal{L}}_j(y)| \ \xrightarrow{p}\ 0, \tag{19}$$

$$T^{\frac{1}{2}} \sup_{h\in\mathcal{H}_T} \sup_{y\in\mathcal{K}} |\widehat{\mathcal{R}}_j(y)| \xrightarrow{p} 0, \tag{20}$$

$$T^{-\frac{1}{2}} \sup_{h\in\mathcal{H}_T} \left| \sum_{t=3}^{T} \widehat{\mathcal{L}}_j(y_{t-2})\, \vartheta_1(\epsilon_t)\, \vartheta_2(z_t)\, w_3(y_{t-2}) \right| \xrightarrow{p} 0 \tag{21}$$

hold for bounded functions ϑ_1 *and* ϑ_2 *such that* $E(\vartheta_1(\epsilon_t)) \times E(\vartheta_2(z_t)|y_{t-2}=y) = 0$, *for almost all* y, *where* $\mathcal{K}$ *is defined in* ***N6****. Then, under* ***N0*** *to* ***N4*** *and* ***N6****, the following assertion holds uniformly over* $h \in \mathcal{H}_T$

$$T^{\frac{1}{2}} \left(\widehat{\beta}(h) - \beta_o \right) \xrightarrow{D} N\left(\mathbf{0}, \sigma^2_{\psi_1,w_2,w_3} \right),$$

where $\sigma^2_{\psi_1,w_2,w_3}$ *is defined in Theorem 4.1.*

Remark 6.1. Note that (19) and (20) entail that $T^{\frac{1}{4}} \sup_{h\in\mathcal{H}_T} \sup_{y\in[0,1]} |\widehat{\phi}_j(y) - \phi_j(y)| \xrightarrow{p} 0$, for $j = 1, 2$.

Using a Taylor's expansion, it is easy to see that an M–estimator can be written $\widehat{\phi}_j(y) = \phi_j(y) + \widehat{\mathcal{L}}_j(y) + \widehat{\mathcal{R}}_j(y)$, where the remainder term satisfies (20), since M–estimators satisfy $T^{\frac{1}{4}} \sup_{h\in\mathcal{H}_T} \sup_{y\in\mathcal{K}} |\widehat{\phi}_j(y) - \phi_j(y)| \xrightarrow{p} 0$ when ϕ_j are continuously differentiable functions. This last result and (19) hold if the kernel is of bounded variation and can be derived using similar arguments to those considered in Boente and Fraiman (1991, a) and a bound for the covering number of the family $h^{-1}K(\cdot/h)$. Conditions to guarantee (21) can be found in Lemma 6.6.7 in Härdle, Liang and Gao (2000).

A Appendix

From now on, C_χ will denote the Lipschitz constant for a Lipschitz function χ.

In the following Lemma, we get a consistent sequence of estimates of the matrix A given in **N2**.

Lemma A.1. *Let* $\{y_t\}$, $t \geq 3$ *be a stationary and ergodic process satisfying (2) with* ϵ_t *independent of* $\{y_{t-j}\,,\ j \geq 1\}$. *Denote* $r_t = y_t - \phi_2(y_{t-2})$ *and* $z_t = y_{t-1} - \phi_1(y_{t-2})$. *Assume* ***N1*** *to* ***N3*** *and* ***N6*** *and that* $\widetilde{\beta}$ *is a sequence of estimates such that* $\widetilde{\beta} \xrightarrow{p} \beta_o$. *Let* $\widehat{\phi}_j(y)$, $j = 1, 2$ *be robust estimates of* $\phi_j(y)$ *such that*

$$\sup_{y\in\mathcal{K}} |\widehat{\phi}_j(y) - \phi_j(y)| \xrightarrow{p} 0, \quad j = 1, 2\,,$$

where $\mathcal{K}$ *is defined in* ***N6****. Then,* $A_T \xrightarrow{p} A$, *where* A *is given in* ***N2*** *and*

$$A_T = \frac{1}{T-2} \sum_{t=3}^{T} \psi_1' \left(\frac{\widehat{r}_t - \widehat{z}_t\widetilde{\beta}}{\sigma_o} \right) w_2(\widehat{z}_t)\, \widehat{z}_t^2\, w_3(y_{t-2})\ .$$

Proof. Denote ξ_t intermediate points between $r_t - z_t\widetilde{\beta}$ and $\widehat{r}_t - \widehat{z}_t\widetilde{\beta}$ and $\widehat{\eta}_j(y) = \widehat{\phi}_j(y) - \phi_j(y)$ for $j = 1, 2$. A first order Taylor's expansion and some algebra lead us to $A_T = A_T^1 + A_T^2 +$

$A_T^3 + A_T^4$, where

$$\begin{aligned}
A_T^1 &= \frac{1}{T-2}\sum_{t=3}^{T}\psi_1'\left(\frac{r_t - z_t\widetilde{\beta}}{\sigma_o}\right) w_2\left(z_t\right) z_t^2\; w_3\left(y_{t-2}\right) \\
A_T^2 &= -\frac{1}{T-2}\sum_{t=3}^{T}\psi_1'\left(\frac{\widehat{r}_t - \widehat{z}_t\widetilde{\beta}}{\sigma_o}\right) w_2\left(\widehat{z}_t\right)\left[\widehat{\eta}_1(y_{t-2})z_t + \widehat{z}_t\,\widehat{\eta}_1(y_{t-2})\right]\; w_3\left(y_{t-2}\right) \\
A_T^3 &= -\frac{1}{T-2}\sum_{t=3}^{T}\psi_1''\left(\frac{\xi_t}{\sigma_o}\right)\left(\frac{\widehat{\eta}_2(y_{t-2}) - \widehat{\eta}_1(y_{t-2})\widetilde{\beta}}{\sigma_o}\right) w_2\left(z_t\right) z_t^2\; w_3\left(y_{t-2}\right) \\
A_T^4 &= \frac{1}{T-2}\sum_{t=3}^{T}\psi_1'\left(\frac{\widehat{r}_t - \widehat{z}_t\widetilde{\beta}}{\sigma_o}\right)\left[w_2\left(\widehat{z}_t\right) - w_2\left(z_t\right)\right] z_t^2\; w_3\left(y_{t-2}\right)\,.
\end{aligned}$$

Analogous arguments to those used in Lemma 1 in Bianco and Boente (2001) allow us to show that $A_T^1 \xrightarrow{\;p\;} A$, since Theorem 2 in Pollard (1984) holds under stationarity and ergodicity.

From **N3**, it is easy to see that

$$z_t^2\left|w_2\left(\widehat{z}_t\right) - w_2\left(z_t\right)\right| \leq \left|\widehat{\eta}_1(y_{t-2})\right|\Big(\left\|\psi_2\right\|_\infty + \left|\widehat{\eta}_1(y_{t-2})\right|\left(\left\|w_2\right\|_\infty + \left\|\psi_2'\right\|_\infty\right) + \left\|\lambda_2\right\|_\infty\Big)\,.$$

Now, the result follows from **N2**, the consistency of $\widetilde{\beta}$, the Ergodic Theorem and the fact that $\max\limits_{1\leq j\leq 2}\sup\limits_{y\in\mathcal{K}}\left|\widehat{\eta}_j(y)\right| \xrightarrow{\;p\;} 0$ and $\left\|w_3\right\|_\infty \leq 1$, since

$$\begin{aligned}
\left|A_T^2\right| &\leq \left\|\psi_1'\right\|_\infty \max_{1\leq j\leq 2}\sup_{y\in\mathcal{K}}\left|\widehat{\eta}_j(y)\right|\left(2\left\|\psi_2\right\|_\infty + \left\|w_2\right\|_\infty \max_{1\leq j\leq 2}\sup_{y\in\mathcal{K}}\left|\widehat{\eta}_j(y)\right|\right) \\
\left|A_T^3\right| &\leq \left\|\psi_1''\right\|_\infty \max_{1\leq j\leq 2}\sup_{y\in\mathcal{K}}\left|\widehat{\eta}_j(y)\right|\left(\frac{1+\left|\widetilde{\beta}\right|}{\sigma_o}\right)\frac{1}{T-2}\sum_{t=3}^{T} w_2\left(z_t\right) z_t^2 \\
\left|A_T^4\right| &\leq \left\|\psi_1'\right\|_\infty \sup_{y\in\mathcal{K}}\left|\widehat{\eta}_1(y)\right|\left(\left\|\psi_2\right\|_\infty + \sup_{y\in\mathcal{K}}\left|\widehat{\eta}_1(y)\right|\left(\left\|w_2\right\|_\infty + \left\|\psi_2'\right\|_\infty\right) + \left\|\lambda_2\right\|_\infty\right)\,. \;\square
\end{aligned}$$

Proof of Theorem 4.1. Denote

$$\begin{aligned}
L_T\left(\beta\right) &= \frac{\sigma_o}{T-2}\sum_{t=3}^{T}\psi_1\left(\frac{r_t - z_t\beta}{\sigma_o}\right) w_2\left(z_t\right) z_t\; w_3\left(y_{t-2}\right) \\
\widehat{L}_T\left(\beta\right) &= \frac{\sigma_o}{T-2}\sum_{t=3}^{T}\psi_1\left(\frac{\widehat{r}_t - \widehat{z}_t\beta}{\sigma_o}\right) w_2\left(\widehat{z}_t\right)\widehat{z}_t\; w_3\left(y_{t-2}\right)\,.
\end{aligned}$$

Using a first order Taylor's expansion around $\widehat{\beta}_T$, we get

$$\begin{aligned}
\widehat{L}_T\left(\beta_o\right) &= \frac{\sigma_o}{T-2}\sum_{t=3}^{T}\psi_1\left(\frac{\widehat{r}_t - \widehat{z}_t\widehat{\beta}_T}{\sigma_o}\right) w_2\left(\widehat{z}_t\right)\widehat{z}_t\; w_3\left(y_{t-2}\right) + \\
&+ \left(\widehat{\beta}_T - \beta_o\right)\frac{1}{T-2}\sum_{t=3}^{T}\psi_1'\left(\frac{\widehat{r}_t - \widehat{z}_t\widetilde{\beta}}{\sigma_o}\right) w_2\left(\widehat{z}_t\right)\widehat{z}_t^2\; w_3\left(y_{t-2}\right)\,,
\end{aligned}$$

with $\widetilde{\beta}$ an intermediate point between $\widehat{\beta}_T$ and β_o. This implies that

$$\widehat{L}_T(\beta_o) = 0 + (\widehat{\beta}_T - \beta_o)\frac{1}{T-2}\sum_{t=3}^{T}\psi_1'\left(\frac{\widehat{r}_t - \widehat{z}_t\widetilde{\beta}}{\sigma_o}\right) w_2(\widehat{z}_t)\ \widehat{z}_t^2\ w_3(y_{t-2})$$

and so, we get that $(\widehat{\beta}_T - \beta_o) = A_T^{-1}\widehat{L}_T(\beta_o)$ with A_T defined in Lemma A.1. From the consistency of $\widehat{\beta}_T$, Lemma A.1 implies that $A_T \xrightarrow{p} A$ and therefore, from **N2** it will be enough to show that

a) $T^{\frac{1}{2}}L_T(\beta_o) \xrightarrow{D} N(\mathbf{0}, \sigma^2)$ with $\sigma^2 = \sigma_o^2\, E\left(\psi_1^2\left(\frac{\epsilon_t}{\sigma_o}\right)\right) E\left(w_2^2(z_t)\, z_t^2\ w_3^2(y_{t-2})\right)$,

b) $T^{\frac{1}{2}}\left[\widehat{L}_T(\beta_o) - L_T(\beta_o)\right] \xrightarrow{p} 0$.

a) Follows inmediately from the Central Limit Theorem for geometrically $\alpha-$mixing process, since $r_t - z_t\beta_o = \epsilon_t$ is independent of $\{y_s : s \leq t\}$ (see for instance, Theorem 1.7 in Bosq (1996)).
b) Denote ξ_t intermediate points between $r_t - z_t\beta_o$ and $\widehat{r}_t - \widehat{z}_t\beta_o$ and $\widehat{\eta}_j(y) = \widehat{\phi}_j(y) - \phi_j(y)$ for $j = 1, 2$. Using a second order Taylor' s expansion, we have that $\widehat{L}_T(\beta_o) = L_T(\beta_o) + \widehat{L}_{T,1} + \widehat{L}_{T,2} + \widehat{L}_{T,3} + \widehat{L}_{T,4} + \widehat{L}_{T,5}$, where

$$\begin{aligned}
\widehat{L}_{T,1} &= \frac{1}{T-2}\sum_{t=3}^{T}\psi_1'\left(\frac{r_t - z_t\beta_o}{\sigma_o}\right)\left[\widehat{\eta}_1(y_{t-2})\beta_o - \widehat{\eta}_2(y_{t-2})\right] w_2(z_t)\ z_t\ w_3(y_{t-2}) \\
\widehat{L}_{T,2} &= \frac{\sigma_o}{T-2}\sum_{t=3}^{T}\psi_1\left(\frac{r_t - z_t\beta_o}{\sigma_o}\right)\left[w_2(\widehat{z}_t)\,\widehat{z}_t - w_2(z_t)\ z_t\right] w_3(y_{t-2}) \\
\widehat{L}_{T,3} &= \frac{\sigma_o}{T-2}\sum_{t=3}^{T}\left[\psi_1\left(\frac{\widehat{r}_t - \widehat{z}_t\beta_o}{\sigma_o}\right) - \psi_1\left(\frac{r_t - z_t\beta_o}{\sigma_o}\right)\right] w_2(\widehat{z}_t)\ (\widehat{z}_t - z_t)\ w_3(y_{t-2}) \\
\widehat{L}_{T,4} &= \frac{1}{2\,\sigma_o}\frac{1}{T-2}\sum_{t=3}^{T}\psi_1''\left(\frac{\xi_t}{\sigma_o}\right)\left[\widehat{\eta}_1(y_{t-2})\beta_o - \widehat{\eta}_2(y_{t-2})\right]^2 w_2(\widehat{z}_t)\ z_t\ w_3(y_{t-2}) \\
\widehat{L}_{T,5} &= \frac{1}{T-2}\sum_{t=3}^{T}\psi_1'\left(\frac{r_t - z_t\beta_o}{\sigma_o}\right)\left[\widehat{\eta}_1(y_{t-2})\beta_o - \widehat{\eta}_2(y_{t-2})\right]\left[w_2(\widehat{z}_t) - w_2(z_t)\right] z_t\ w_3(y_{t-2})\ .
\end{aligned}$$

Since, $\|w_3\|_\infty \leq 1$ and **N3** entails $|w_2(\widehat{z}_t) - w_2(z_t)| \leq C\frac{|\widehat{\eta}_1(y_{t-2})|}{|z_t|}$, where $C = \|w_2\|_\infty + C_{\psi_2}$, we get

$$\begin{aligned}
T^{\frac{1}{2}}\|\widehat{L}_{T,3}\| &\leq p\,\|w_2\|_\infty\|\psi_1'\|_\infty T^{\frac{1}{2}}\left[\sup_{y\in\mathcal{K}}|\widehat{\eta}_1(y)|\right]^2(1 + |\beta_o|) \\
T^{\frac{1}{2}}\|\widehat{L}_{T,4}\| &\leq \frac{1}{2}\frac{1}{\sigma_o}\|\psi_1''\|_\infty T^{\frac{1}{2}}\left[\sup_{y\in\mathcal{K}}\max_{1\leq j\leq 2}|\widehat{\eta}_j(y)|\right]^2(1+|\beta_o|)^2\left(\|\psi_2\|_\infty + \|w_2\|_\infty\sup_{y\in\mathcal{K}}|\widehat{\eta}_1(y)|\right) \\
T^{\frac{1}{2}}\|\widehat{L}_{T,5}\| &\leq p\,C\|\psi_1'\|_\infty(1+|\beta_o|)\ T^{\frac{1}{2}}\left[\max_{1\leq j\leq 2}\sup_{y\in\mathcal{K}}|\widehat{\eta}_j(y)|\right]^2\ ,
\end{aligned}$$

which together with (13), implies that, for $3 \leq j \leq 5$, $T^{\frac{1}{2}}\|\widehat{L}_{T,j}\| \xrightarrow{p} 0$. Note that

$$\begin{aligned}
\widehat{L}_{T,2} &= \frac{\sigma_o}{T-2}\sum_{t=3}^{T}\psi_1\left(\frac{r_t - z_t\beta_o}{\sigma_o}\right)\left[w_2\left(\widehat{z}_t\right)\widehat{z}_t - w_2\left(z_t\right)\, z_t\right] w_3\left(y_{t-2}\right) \\
&= \frac{\sigma_o}{T-2}\sum_{t=3}^{T}\psi_1\left(\frac{\epsilon_t}{\sigma_o}\right)\left[\psi_2\left(\widehat{z}_t\right) - \psi_2\left(z_t\right)\right] w_3\left(y_{t-2}\right) \\
&= \frac{\sigma_o}{T-2}\sum_{t=3}^{T}\psi_1\left(\frac{\epsilon_t}{\sigma_o}\right)\psi_2'\left(z_t\right)\widehat{\eta}_1\left(y_{t-2}\right) w_3\left(y_{t-2}\right) + \sum_{t=3}^{T}\psi_1\left(\frac{\epsilon_t}{\sigma_o}\right)\left[\psi_2'\left(\xi_t\right) - \psi_2'\left(z_t\right)\right]\widehat{\eta}_1\left(y_{t-2}\right) w_3\left(y_{t-2}\right) \\
&= \widehat{L}_{T,2}^{(1)} + \widehat{L}_{T,2}^{(2)}
\end{aligned}$$

where ξ_t denotes intermediate points between $\widehat{z}_t$ and z_t. Using that ψ_2' is Lispchitz of order 1, with constant $C_{\psi_2'}$, we get

$$T^{\frac{1}{2}}|\widehat{L}_{T,2}^{(2)}| \leq C_{\psi_2'}\|\psi_1\|_\infty\, T^{\frac{1}{2}}\sup_{y\in\mathcal{K}}|\widehat{\eta}_1\left(y\right)|^2$$

which together with (13) and (16) entail that $T^{\frac{1}{2}}|\widehat{L}_{T,2}^{(2)}\| \xrightarrow{p} 0$.

It remains to show that $T^{\frac{1}{2}}\widehat{L}_{T,1} \xrightarrow{p} 0$ and $T^{\frac{1}{2}}\|\widehat{L}_{T,2}^{(1)}\| \xrightarrow{p} 0$, that is,

$$\widehat{R}_{T,j} = T^{-\frac{1}{2}}\sum_{t=3}^{T}\psi_1'\left(\frac{\epsilon_t}{\sigma_o}\right)\widehat{\eta}_j(y_{t-2})\, w_2\left(z_t\right)\, z_t\, w_3\left(y_{t-2}\right) \xrightarrow{p} 0\,, \quad j = 1,2 \tag{A.1}$$

$$\widehat{R}_{T,3} = T^{-\frac{1}{2}}\sum_{t=3}^{T}\psi_1\left(\frac{\epsilon_t}{\sigma_o}\right)\psi_2'\left(z_t\right)\widehat{\eta}_1\left(y_{t-2}\right)\, w_3\left(y_{t-2}\right) \xrightarrow{p} 0\,, \tag{A.2}$$

since $\epsilon_t = r_t - z_t\beta_o$.

We begin by proving the desired result when a) holds.

Note that proving (A.1) is equivalent to show that, for $j = 1, 2$

$$\widehat{R}_{T,j,1} = T^{-\frac{1}{2}}\sum_{t=3}^{T}\left[\psi_1'\left(\frac{\epsilon_t}{\sigma_o}\right) - E\left(\psi_1'\left(\frac{\epsilon_t}{\sigma_o}\right)\right)\right]\widehat{\eta}_j(y_{t-2})\, w_2\left(z_t\right)\, z_t\, w_3\left(y_{t-2}\right) \xrightarrow{p} 0\,, \tag{A.3}$$

$$\widehat{R}_{T,j,2} = T^{-\frac{1}{2}}\sum_{t=3}^{T}\widehat{\eta}_j(y_{t-2})\, w_2\left(z_t\right)\, z_t\, w_3\left(y_{t-2}\right) \xrightarrow{p} 0\,. \tag{A.4}$$

For any function υ with domain on the compact support of the weight function w_3, $\mathcal{K}$, we define

$$\begin{aligned}
J_{T,1}\left(\upsilon\right) &= T^{-\frac{1}{2}}\sum_{t=3}^{T}\left[\psi_1'\left(\frac{\epsilon_t}{\sigma_o}\right) - E\left(\psi_1'\left(\frac{\epsilon_t}{\sigma_o}\right)\right)\right]\upsilon(y_{t-2})\,\psi_2\left(z_t\right)\, w_3\left(y_{t-2}\right) \\
J_{T,2}\left(\upsilon\right) &= T^{-\frac{1}{2}}\sum_{t=3}^{T}\psi_1\left(\frac{\epsilon_t}{\sigma_o}\right)\psi_2'\left(z_t\right)\upsilon(y_{t-2})\, w_3\left(y_{t-2}\right) \\
J_{T,3}\left(\upsilon\right) &= T^{-\frac{1}{2}}\sum_{t=3}^{T}\upsilon(y_{t-2})\,\psi_2\left(z_t\right)\, w_3\left(y_{t-2}\right)\,.
\end{aligned}$$

Let $\mathcal{L} = \{\upsilon \in \mathcal{C}^1(\mathcal{K}) : \|\upsilon\|_{1,\infty} = \|\upsilon\|_\infty + \|\upsilon'\|_\infty \leq 1\}$. Note that, from Theorem 2.7.1 in van der Vaart and Wellner (1996) the covering number $N(\epsilon, \mathcal{L}, \|\cdot\|_\infty) \leq \exp(K\epsilon^{-1})$ for every $\epsilon > 0$, where K is a constant.

For any $\epsilon > 0$, let $0 < \delta < 1$. Using that (13) and (14) entail, for $j = 1, 2$

$$T^{\frac{1}{4}} \sup_{t\in\mathcal{K}} |\widehat{\eta}_j(t)| = T^{\frac{1}{4}} \sup_{t\in\mathcal{K}} |\widehat{\phi}_j(t) - \phi_j(t)| \xrightarrow{p} 0,$$
$$\sup_{t\in\mathcal{K}} |\widehat{\eta}'_j(t)| = \sup_{t\in\mathcal{K}} |\widehat{\phi}'_j(t) - \phi'_j(t)| \xrightarrow{p} 0\,,$$

we have that, for T large enough, $P\left(\widehat{\eta}_j \in \mathcal{L} \text{ and } \|\widehat{\eta}_j\|_\infty < \delta T^{-\frac{1}{4}}\right) > 1 - \delta$, for $j = 1, 2$.

Denote by $A_1 = 2\|\psi'_1\|_\infty\|\psi_2\|_\infty$, $A_2 = \|\psi_1\|_\infty\|\psi'_2\|_\infty$, $A_3 = \|\psi_2\|_\infty$ and $A = \max_{1\leq i\leq 3} A_i$ and by $\mathcal{L}_\delta = \{\upsilon \in \mathcal{L} : \|\upsilon\|_{1,\infty} < \delta \text{ and } \|\upsilon\|_\infty < \delta_T = \delta T^{-\frac{1}{4}}\}$. Let $\alpha_T = \dfrac{\epsilon}{2\,A} T^{-\frac{1}{2}}$ and $N_T = N\left(\dfrac{\alpha_T}{2}, \mathcal{L}_\delta, \|\cdot\|_\infty\right) \leq N(\vartheta_T, \mathcal{L}, \|\cdot\|_\infty)$ with $\vartheta_T = (2\,\delta)^{-1}\alpha_T = \dfrac{\epsilon}{4A\,\delta} T^{-\frac{1}{2}}$. Note that, if $\upsilon_\ell \in \mathcal{L}_\delta$ and $\upsilon \in \mathcal{L}_\delta$ satisfy $\|\upsilon_\ell - \upsilon\|_\infty < \alpha_T$, then, from **N6**, $|J_{T,i}(\upsilon_\ell) - J_{T,i}(\upsilon)| \leq \dfrac{\epsilon}{2}$, for $1 \leq i \leq 3$. For any $\upsilon \in \mathcal{L}_\delta$, denote $\mathcal{V}(\upsilon) = \{u \in \mathcal{L}_\delta : \|u - \upsilon\|_\infty < \alpha_T\}$. Note that given $\upsilon \in \mathcal{L}_\delta$ there exists $1 \leq \ell \leq N_T$ and $\upsilon_\ell \in \mathcal{L}_\delta$ such that $\upsilon \in \mathcal{V}(\upsilon_\ell)$ and so, $\|\upsilon_\ell\|_\infty \leq \delta T^{-\frac{1}{4}}$. Thus, for $j = 1, 2$, we have that for $i = 1, 2$

$$\begin{aligned}
P(|J_{T,i}(\widehat{\eta}_j)| > \epsilon) &\leq P\left(|J_{T,i}(\widehat{\eta}_j)| > \epsilon\,, \widehat{\eta}_j \in \mathcal{L} \text{ and } \|\widehat{\eta}_j\|_\infty < \delta T^{-\frac{1}{4}}\right) + \delta \\
&\leq P\left(\sup_{\upsilon\in\mathcal{L}_\delta} |J_{T,i}(\upsilon)| > \epsilon\right) + \delta \\
&\leq P\left(\max_{1\leq\ell\leq N_T} \sup_{\upsilon\in\mathcal{V}(\upsilon_\ell)} |J_{T,i}(\upsilon)| > \epsilon\right) + \delta \\
&\leq P\left(\max_{1\leq\ell\leq N_T} \sup_{\upsilon\in\mathcal{V}(\upsilon_\ell)} \{|J_{T,i}(\upsilon) - J_{T,i}(\upsilon_\ell)| + |J_{T,i}(\upsilon_\ell)|\} > \epsilon\right) + \delta \\
&\leq P\left(\max_{1\leq\ell\leq N_T} |J_{T,i}(\upsilon_\ell)| > \epsilon/2\right) + \delta \\
&\leq N_T \max_{1\leq\ell\leq N_T} P(|J_{T,i}(\upsilon_\ell)| > \epsilon/2) + \delta\,,
\end{aligned}$$

Let us consider $\mathcal{F}_t$ be the $\sigma-$ field generated by $\{y_j : j \leq t-1\}$, which forms an increasing family of $\sigma-$fields. Note that $X_{t,1}(\upsilon) = \left[\psi'_1\left(\dfrac{\epsilon_t}{\sigma_o}\right) - E\left(\psi'_1\left(\dfrac{\epsilon_1}{\sigma_o}\right)\right)\right] \upsilon(y_{t-2})\,\psi_2(z_t)\; w_3(y_{t-2})$ is bounded and a martingale difference with respect to $\mathcal{F}_t$ since the independence between ϵ_t and $\{y_{t-j} : \; j \geq 1\}$ entail $E(X_{t,1}(\upsilon)|\mathcal{F}_t) = 0$. Similarly, since $E\psi_1\left(\dfrac{\epsilon_t}{\sigma_o}\right) = 0$, $X_{t,2}(\upsilon) = \psi_1\left(\dfrac{\epsilon_t}{\sigma_o}\right)\psi'_2(z_t)\,\upsilon(y_{t-2})\; w_3(y_{t-2})$ is also a martingale difference with respect to $\mathcal{F}_t$. Using Theorem 2.3.1 in Györfi, Härdle, Sarda and Vieu (1989) and using that $X_{t,i}(\upsilon_\ell) \leq K_T = A\delta T^{-\frac{1}{4}}$, $1 \leq i \leq 3$, we obtain that for $i = 1, 2$

$$P(|J_{T,i}(\upsilon_\ell)| > \epsilon) \leq 2\exp\left\{-\frac{\epsilon^2 T}{2\,T\,K_T^2}\right\} = 2\exp\left\{-\frac{\epsilon^2 T^{\frac{1}{2}}}{2\,A^2\delta^2}\right\}$$

Therefore, if $B_1 = \epsilon^2(2\,A^2\,\delta^2)^{-1}$ and $B_2 = 4\,K\,A\,\delta\epsilon^{-1}$, we have that

$$\begin{aligned} N_T \max_{1\leq\ell\leq N_T} P\left(|J_{T,i}\left(\upsilon_\ell\right)| > \epsilon\right) &\leq 2\exp\left(K\vartheta_T^{-1}\right)\exp\left\{-\frac{\epsilon^2 T^{\frac{1}{2}}}{2\,A^2\delta^2}\right\} \\ &\leq 2\exp\left\{-B_1 T^{\frac{1}{2}} + B_2 T^{\frac{1}{2}}\right\} = 2\exp\left\{-T^{\frac{1}{2}}(B_1 - B_2)\right\}\ . \end{aligned}$$

Choosing $\delta < \dfrac{\epsilon}{2\,A(K)^{\frac{1}{3}}}$, we get that $B_1 - B_2 > 0$ which entails that

$$\limsup_{T\to\infty} P\left(|J_{T,i}\left(\widehat{\eta}_j\right)| > \epsilon\right) \leq \delta\ . \quad i = 1,2$$

which entails that (A.3) and (A.2).

Finally, under a) assumption **N4** entail that $X_{t,3}(\upsilon) = \upsilon(y_{t-2})\,\psi_2\left(z_t\right)\,w_3\left(y_{t-2}\right)$ satisfies $E\left(X_{t,3}(\upsilon)\right) = 0$ for any bounded function υ and $J_{T,3}\left(\upsilon\right) = T^{-\frac{1}{2}}\sum_{t=3}^{T} X_{t,3}(\upsilon)$. Using that the process is geometrically $\alpha-$mixing, Theorem 1.5 in Bosq (1996) and Theorem 1 in Doukhan *et al.* (1994) we have that the finite dimensional distributions of $\{J_{T,3}\left(\upsilon\right) : \upsilon \in \mathcal{L}\}$ converge to the finite dimensional distributions of an eventually degenerate Gaussian Process $\{J_3\left(\upsilon\right) : \upsilon \in \mathcal{L}\}$ with covariance given by

$$\begin{aligned} &E\left(J_3\left(\upsilon_1\right)J_3\left(\upsilon_2\right)\right) = E\left(\upsilon_1(y_1)\upsilon_2(y_1)\,\psi_2^2\left(y_2 - \phi(y_1)\right)\,w_3^2\left(y_1\right)\right) \\ +\ &\sum_{j=1}^{\infty} E\left(\left[\upsilon_1(y_1)\upsilon_2(y_{j+1}) + \upsilon_1(y_{j+1})\upsilon_2(y_1)\right]\,\psi_2\left(y_2 - \phi(y_1)\right)\psi_2\left(y_{2+j} - \phi(y_{1+j})\right)\,w_3\left(y_1\right)w_3\left(y_{j+1}\right)\right) \end{aligned}$$

On the other hand, using Jensen's inequality, Theorem 1.2 in Rio (1993) and the fact that the mixing coefficients are geometric, we have that for some finite constant C depending on the mixing coefficients, and any $p > 2$

$$\begin{aligned} E\left|J_{T,3}\left(\upsilon\right)\right| &\leq \left[EJ_{T,3}^2\left(\upsilon\right)\right]^{\frac{1}{2}} \\ &\leq C\left[E\left|\upsilon(y_{t-2})\,\psi_2\left(z_t\right)\,w_3\left(y_{t-2}\right)\right|^p\right]^{\frac{1}{p}} \\ &\leq C\|\psi_2\|_\infty\left[E\left|\upsilon(y_{t-2})\right|^p\right]^{\frac{1}{p}}\ . \end{aligned}$$

Therefore, Theorem 2.1 in Arcones (1996) implies the weak convergence of $\{J_{T,3}\left(\upsilon\right) : \upsilon \in \mathcal{L}\}$ to the a Gaussian Process $\{J_3\left(\upsilon\right) : \upsilon \in \mathcal{L}\}$ which entails that the process $\{J_{T,3}\left(\upsilon\right) : \upsilon \in \mathcal{L}\}$ is stochastically equicontinuous. As noted by Andrews (1994), the stochastic equicontinuity of the process and the fact that from (13) $\sup_{y\in\mathcal{K}}|\widehat{\eta}_j(t)| = \sup_{y\in\mathcal{K}}|\widehat{\phi}_j(y) - \phi_j(y)| \xrightarrow{p} 0$, we obtain that $J_{T,3}\left(\widehat{\eta}_j\right) \xrightarrow{p}$ and so (A.4) holds, concluding the proof when a) holds.

Assume now that b) holds.

Using the linear expansion for $\widehat{\eta}_j(y)$, we get that for $j = 1,2$

$$\begin{aligned} T^{\frac{1}{2}}\widehat{R}_{T,j} = \sum_{t=3}^{T}\psi_1'\left(\frac{\epsilon_t}{\sigma_o}\right)\,\widehat{\eta}_j(y_{t-2})\,\psi_2\left(z_t\right)\,w_3\left(y_{t-2}\right) &= \sum_{t=3}^{T}\psi_1'\left(\frac{\epsilon_t}{\sigma_o}\right)\,\widehat{\mathcal{L}}_j(y_{t-2})\,\psi_2\left(z_t\right)\,w_3\left(y_{t-2}\right) + \\ &+ \sum_{t=3}^{T}\psi_1'\left(\frac{\epsilon_t}{\sigma_o}\right)\,\widehat{\mathcal{R}}_j(y_{t-2})\,\psi_2\left(z_t\right)\,w_3\left(y_{t-2}\right)\ , \end{aligned}$$

which implies that

$$|\widehat{R}_{T,j}| \leq \left| \frac{1}{\sqrt{T}} \sum_{t=3}^{T} \psi_1' \left(\frac{\epsilon_t}{\sigma_o} \right) \widehat{\mathcal{L}}_j(y_{t-2}) \, \psi_2(z_t) \; w_3(y_{t-2}) \right| + \|\psi_1'\|_\infty \|\psi_2\|_\infty \, T^{\frac{1}{2}} \sup_{y \in \mathcal{K}} |\widehat{\mathcal{R}}_j(y)| \, .$$

Now (A.1) follows from **N4**, (16) and (17) with $\vartheta_1(t) = \psi_1'(t/\sigma_o)$ and $\vartheta_2 \equiv \psi_2$.

Finally, (A.2) follows using similar arguments to those used to deal with (A.1), applying (17) to $\vartheta_1(t) = \psi_1(t/\sigma_o)$ and $\vartheta_2 \equiv \psi_2'$ and (16) which concludes the proof. □

The following Lemma states a result analogous to that given in Lemma A.1, uniformly on the bandwidth parameter.

Lemma A.2. *Let $\{y_t\}$, $t \geq 3$ be a stationary and ergodic process satisfying (2) with ϵ_t independent of $\{y_{t-j}\,,\ j \geq 1\}$. Let $\mathcal{H}_T = [aT^{-\frac{1}{5}-c}, bT^{-\frac{1}{5}+c}]$ with $0 < a < b < \infty$ and $0 < c < \frac{1}{20}$. Denote $r_t = y_t - \phi_2(y_{t-2})$ and $z_t = y_{t-1} - \phi_1(y_{t-2})$. Let $\widehat{\phi}_j(y)$, $j = 1, 2$ be robust estimates of $\phi_j(y)$ such that*

$$\sup_{y \in \mathcal{K}} |\widehat{\phi}_j(y) - \phi_j(y)| \xrightarrow{p} 0, \quad j = 1, 2$$

*uniformly for $h \in \mathcal{H}_T$ and assume that $\widetilde{\beta} \xrightarrow{p} \beta_o$ also uniformly for $h \in \mathcal{H}_T$. Then, under **N1** to **N3** and **N6**, $A_T \xrightarrow{p} A$ uniformly for $h \in \mathcal{H}_T$, where A is given in **N2** and A_T is defined in Lemma 4.1.*

Proof. As in Lemma A.1, we have that $A_T = A_T^1 + A_T^2 + A_T^3 + A_T^4$. The bounds obtained for A_T^j, for $j = 2, 3, 4$ hold uniformly for $h \in \mathcal{H}_T$. On the other hand, defining $\mathcal{A}_T(\beta) = \dfrac{1}{T-2} \displaystyle\sum_{t=3}^{T} \psi_1' \left(\frac{r_t - z_t\beta}{\sigma_o} \right) w_2(z_t) \, z_t^2 \; w_3(y_{t-2})$ and using analogous arguments to those considered in Lemma 1 in Bianco and Boente (2001), we get that, for any $\delta > 0$, $\sup_{|\beta - \beta_o| < \delta} |\mathcal{A}_T(\beta) - E(\mathcal{A}_T(\beta))| \xrightarrow{p} 0$, which together with the uniform convergence of $\widetilde{\beta}$ to β_o entails the desired result. □

Remark A.1. It is worth noticing that the conclusion of Lemmas A.1 and A.2 still holds without requiring **N6**. In that case, we need to assume that $\sup_{y \in \mathcal{K}} |\widehat{\phi}_j(y) - \phi_j(y)| \xrightarrow{p} 0$, $j = 1, 2$, for any compact set $\mathcal{K} \subset I\!R$.

Proof of Theorem 6.1. Follows using analogous arguments as those considered in the proof of Theorem 4.1. First notice that $T^{\frac{1}{2}} L_T(\beta_o) \xrightarrow{D} N(\mathbf{0}, \sigma^2)$ holds uniformly for $h \in \mathcal{H}_T$, since $L_T(\beta_o)$ does not depend on the smoothing parameter. Therefore, it remains to show that $T^{\frac{1}{2}} \left[\widehat{L}_T(\beta_o) - L_T(\beta_o) \right] \xrightarrow{p} 0$ uniformly in $\mathcal{H}_T$.

As in Theorem 4.1, using a second order Taylor' s expansion, we have that $\widehat{L}_T(\beta_o) = L_T(\beta_o) + \widehat{L}_{T,1} + \widehat{L}_{T,2}^{(1)} + \widehat{L}_{T,2}^{(2)} + \widehat{L}_{T,3} + \widehat{L}_{T,4} + \widehat{L}_{T,5}$, where $\widehat{L}_{T,j}$ are defined in Theorem 4.1. Moreover, with the bounds obtained therein it is easy to see that, for $3 \leq j \leq 5$, $T^{\frac{1}{2}} \sup_{h \in \mathcal{H}_T} |\widehat{L}_{T,j}| \xrightarrow{p} 0$ and that $T^{\frac{1}{2}} \sup_{h \in \mathcal{H}_T} |\widehat{L}_{T,2}^{(2)}| \xrightarrow{p} 0$.

We need to show that $T^{\frac{1}{2}}\widehat{L}_{T,1} \xrightarrow{p} 0$ and $T^{\frac{1}{2}}\widehat{L}^{(1)}_{T,2} \xrightarrow{p} 0$ uniformly for $h \in \mathcal{H}_T$, that is, for $j = 1, 2$

$$\sup_{h\in\mathcal{H}_T} |\widehat{R}_{T,j}| = \sup_{h\in\mathcal{H}_T} T^{-\frac{1}{2}} \left| \sum_{t=3}^{T} \psi_1' \left(\frac{\epsilon_t}{\sigma_o}\right) \widehat{\eta}_j(y_{t-2})\, w_2\left(z_t\right)\, z_t\, w_3\left(y_{t-2}\right) \right| \xrightarrow{p} 0 \qquad \text{(A.5)}$$

$$\sup_{h\in\mathcal{H}_T} |\widehat{R}_{T,3}| = \sup_{h\in\mathcal{H}_T} T^{-\frac{1}{2}} \left| \sum_{t=3}^{T} \psi_1 \left(\frac{\epsilon_t}{\sigma_o}\right) \psi_2'\left(z_t\right) \widehat{\eta}_1(y_{t-2})\, w_3\left(y_{t-2}\right) \right| \xrightarrow{p} 0\,. \qquad \text{(A.6)}$$

We begin by proving (A.5). Note that using the linear expansion for $\widehat{\eta}_j(y)$, we get

$$\begin{aligned} \sum_{t=3}^{T} \psi_1' \left(\frac{\epsilon_t}{\sigma_o}\right) \widehat{\eta}_j(y_{t-2})\, \psi_2\left(z_t\right)\, w_3\left(y_{t-2}\right) &= \sum_{t=3}^{T} \psi_1' \left(\frac{\epsilon_t}{\sigma_o}\right) \widehat{\mathcal{L}}_j(y_{t-2})\, \psi_2\left(z_t\right)\, w_3\left(y_{t-2}\right) + \\ &+ \sum_{t=3}^{T} \psi_1' \left(\frac{\epsilon_t}{\sigma_o}\right) \widehat{\mathcal{R}}_j(y_{t-2})\, \psi_2\left(z_t\right)\, w_3\left(y_{t-2}\right)\,, \end{aligned}$$

which implies that

$$|\widehat{R}_{T,j}| \le \left| \frac{1}{\sqrt{T}} \sum_{t=3}^{T} \psi_1' \left(\frac{\epsilon_t}{\sigma_o}\right) \widehat{\mathcal{L}}_j(y_{t-2})\, \psi_2\left(z_t\right)\, w_3\left(y_{t-2}\right) \right| + \|\psi_1'\|_\infty \|\psi_2\|_\infty T^{\frac{1}{2}} \sup_{y\in\mathcal{K}} |\widehat{\mathcal{R}}_j(y)|\,.$$

Now (A.5) follows from **N4**, (20) and (21) with $\vartheta_1(t) = \psi_1'(t/\sigma_o)$ and $\vartheta_2 \equiv \psi_2$.

On the other hand, (A.6) follows using similar arguments to those used to deal with (A.5), applying (20) and (21) to $\vartheta_1(t) = \psi_1(t/\sigma_o)$ $\vartheta_2 \equiv \psi_2'$ which concludes the proof. □

B References

Andrews, D. and Pollard, D. (1994). An introduction to functional central limit theorems for dependent stochastic processes. *Int. Statist. Rev.*, **62**, 119-132.

Andrews, D. (1994). Asymptotics for semiparametric econometric models via stochastic equicontinuity. *Econometrica*, **62**, 43-72.

Ansley, C. and Wecker, W. (1983). Extension and examples of the signal extraction approach to regression. *In Applied Time Series Analysis of Economic Data,* 181-192.

Arcones, M. (1996). Weak convergence of stochastic processes indexed by smooth functions. *Stochastic Process. Appl.*, **62**, 11-138.

Bianco, A. and Boente, G. (2001). On the asymptotic behavior of one-step estimates in heteroscedastic regression models. *Statistics and Probability Letters*, **60**, 33-47.

Bianco, A. and Boente, G. (2002). A robust approach to partly linear autoregressive models. *Estadística*, **54**, 249-287.

Bianco, A. and Boente, G. (2004). Robust estimators in semiparametric partly linear regression models. *J. Statist. Planning and Inference*, **122**, 229-252.

Bianco, A., Garcia Ben, M., Martinez, E. and Yohai, V. (1996). Robust procedures for regression models with ARIMA errors. In: *COMPSTAT 96*, Albert Prat (ed.). *Proceedings in Computational Statistics*, Physica-Verlag, Heidelberg, pp. 27-38.

Billingsley, P. (1968). *Convergence of Probability Measures* , Wiley, New York.

Boente, G. and Fraiman, R. (1989). Robust nonparametric regression estimation. *J. Multiv. Anal.* **29**, 180-198.

Boente, G. and Fraiman, R. (1991, a). Strong uniform convergence rates for some robust equivariant nonparametric regression estimates for mixing processes. *Int. Statist. Rev.*, **59**, 355-372.

Boente, G. and Fraiman, R. (1991, b). A functional approach to robust nonparametric regression. In: *Directions in robust statistics and diagnostics*, Werner Stahel and Sanford Weisberg (ed.). *Proceedings of the IMA Institute*, USA, **33**, Part I, pp. 35-46.

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

Boente, G. and Rodriguez, D. (2006). Robust estimators of high order derivatives of regression functions. In press in *Statistics and Probability Letters*. Available at http://www.ic.fcen.uba.ar/preprints/boenterodriguez2005.pdf

Bosq, D.(1996). *Nonparametric Statistics for Stochastic Processes. Estimation and Prediction.* Springer–Verlag, New York.

Brillinger, D. R.(1986). Discussion of "Influence functionals for time series" by Martin R. D. and Yohai, V. J. *Ann. Statist.* **14**, 819-822.

Campbell, M. J. and Walker, A. M. (1977). A survey of statistical work on the Mackenzie river series of annual Canadian lynx trapping on the years 1821-1934 and a new analysis. *J. Roy. Statist. Soc.*, Ser. A, **140**, 411-431.

Cantoni, E. and Ronchetti, E. (2001). Resistant selection of the smoothing parameter for smoothing splines. *Statistics and Computing*, **11**, 141-146.

Chen, H. (1988). Convergence rates for parametric components in a partly linear model. *Ann. Statist.*, **16**, 136-146.

Chen, H. and Chen, K. (1991). Selection of the splined variables and convergence rates in a partial spline model. *Canad. J. Statist.*, **19**, 323-339.

Chen, H. and Shiau, J. (1991). A two-stage spline smoothing method for partially linear models. *J. Statist. Planning and Inference*, **25**, 187-201.

Chen, H. and Shiau, J. (1994). Data-driven efficient estimates for partially linear models. *Ann. Statist.*, **22**, 211-237.

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

Doukhan, P. (1994). *Mixing: Properties and Examples. Lecture Notes in Statistics*, **85**, Springer–Verlag, New York.

Doukhan, P., Massart, P. and Rio, E. (1994). The central limit theorem for strongly mixing processes. *Ann. Inst. H. Poincaré Probab. Statist.*, **30**, 63-82.

Engle, R., Granger, C., Rice, J. and Weiss, A. (1986). Semiparametric estimates of the relation between weather and electricity sales. *J. Amer. Statist. Assoc.*, **81**, 310-320.

Gao, J. (1992). *A Large Sample Theory in Semiparametric Regression Models*. Phd. Thesis, University of Science and Technology of China, Hefei, China.

Gao, J. (1995). Asymptotic theory for partly linear models. *Comm. Statist., Theory & Methods*, **24**, 1985-2010.

Gao, J. (1998). Semiparametric regression smoothing of nonlinear time series. *Scandinavian J. Statist.*, **25**, 521-539.

Gao, J. and Liang, H. (1995). Asymptotic normality of pseudo-LS estimator for partly linear autoregression models. *Statist. and Prob. Letters*, **23**, 27-34.

Gao, J. and Shi, P. (1997). M–type smoothing splines in nonparametric and semiparametric regression models. *Statistica Sinica*, **7**, 1155-1169.

Gao, J. and Zhao, L. (1993). Adaptive estimation in partly linear regression models. *Science in China*, Ser. A, **1**, 14-27.

Gao, J. and Yee, T. (2000). Adaptive estimation in partly linear autoregressive models. *Canad. J. Statist.*, **28**, 571–586.

Green, P., Jennison, C. and Seheult, A. (1985). Analysis of field experiments by least squares smoothing. *J. Roy. Statist. Soc.*, Ser. B, **47**, 299-315.

Györfi, L., Härdle, W., Sarda, P. and Vieu, P.(1989). *Nonparametric Curve Estimation from Time Series. Lecture Notes in Statistics*, **60**, Springer-Verlag.

Härdle, W. (1990). *Applied Nonparametric Regression*, Cambridge University Press.

Härdle, W., Liang, H. and Gao, J. (2000). *Partially Linear Models*. Physica–Verlag, Heidelberg.

Härdle, W. and Gasser, T. (1985). On robust kernel estimation of derivatives of regression functions. *Scand. J. Statist.*, **12**, 233-240.

Hart, J.(1996). Some automated methods of smoothing time–dependent data. *Nonparametric Statistics* , **6**, 115-142.

Hart, P. and Vieu, P. (1990). Data–driven bandwidth choice for density estimation based on dependent data. *Ann. Statist.* , **18**, 873-890.

He, X., Zhu, Z. and Fung, W. (2002). Estimation in a semiparametric model for longitudinal data with unspecified dependence structure. *Biometrika* **89**, 579-590.

Heckman, N. (1986). Spline smoothing in a partly linear model. *J. Roy. Statist. Soc.*,

Ser. B, **48**, 244-248.

Ibragimov, I. and Linnik, Y. (1971). *Independent and Stationary Sequences of Random Variables*, Wolters– Noerdhoff, Groningen.

Leung, D. H. Y., Marriott, F. H. C. and Wu, E. K. H. (1993). Bandwidth selection in robust smoothing. *Journal of Nonparametric Satistics*, **2**, 333-339.

Liang, H. (1996). Asymptotically efficient estimators in a partly linear autoregressive model. *System Sciences and Matematical Sciences*, **9**, 164-170.

Martin, R. D. and Yohai, V. J. (1986). Influence functionals for time series. *Ann. Statist.* **14**, 781-818.

Pollard, D. (1984). *Convergence of Stochastic Processes.* Springer–Verlag, New York.

Rio, E. (1993). Covariance inequalities for strongly mixing processes. *Ann. Inst. H. Poincaré Probab. Statist.*, **29**, 587-597.

Robinson, P. M. (1983). Nonparametric estimators for time series. *J. Time Ser. Anal.*, **4**, 185-206.

Robinson, P. M. (1984). Robust nonparametric autoregression, in *Robust and Nonlinear Time Series Analysis*, J. Franke, W. Härdle and D. Martin (eds.), *Lecture Notes in Statistics*, **4**, Springer-Verlag, 185-206.

Robinson, P. (1988). Root-n-consistent Semiparametric regression. *Econometrica*, **56**, 931-954.

Rousseeuw, P. and Leroy, A. (1987). *Robust Regression and Outlier Detection*, Wiley, New York.

Rosenblatt, M. (1956). A central limit theorem and a strong mixing condition. *Proc. Nat. Acad. Sci. USA*, **42**, 43-47.

Severini, T. and Wong, W. (1992). Profile likelihood and conditionally parametric models. *Ann. Statist.* **20**, 1768-1802.

Severini, T. and Staniswalis, J. (1994). Quasi-likelihood estimation in semiparametric models. *J. Amer. Statist. Assoc.* **89** 501-511.

Speckman, P. (1988). Kernel smoothing in partial linear models. *J. Roy. Statist. Soc.*, Ser. B, **50**, 413-436.

Tong, H. (1977). Some comments on the Canadian lynx data (with discussion). *J. Roy. Statist. Soc.*, Ser. A, **140**, 432-436.

Van Der Vaart, A. and Wellner, J. (1996). *Weak Convergence and Empirical Processes: With Applications to Statistics.* New York: Springer.

Yao, Q. and Tong, H. (1994). On the subset selection in nonparametric stochastic regression. *Statistica Sinica* , **4**, 51-70.

Yee, T. and Wild, C. (1996). Vector generalized additive models. *J. Roy. Statist. Soc.*, Ser. B, **58**, 481-493.

YOHAI, V. (1987). High breakdown point and high efficiency robust estimates for regression. *Ann. Statist.*, **15**, 642-656.

YOHAI, V. and ZAMAR, R. (1988). High breakdown estimates of regression by means of the minimization of an efficient scale. *J. Amer. Statist. Assoc.*, **83**, 406-413.

YU, B. (1994). Rates of convergence of empirical processes for stationary mixing sequences. *Ann. Prob.*, **22**,94-116.

WANG, F. and SCOTT, D. (1994). The L_1 method for robust nonparametric regression. *J. Amer. Stat. Assoc.*, **89**, 65-76.

WONG, C. M. and KOHN, R. (1996). A Bayesian approach to estimating and forecasting additive nonparametric autoregressive models. *Journal of Time Series Analysis* **17**, 203-220.

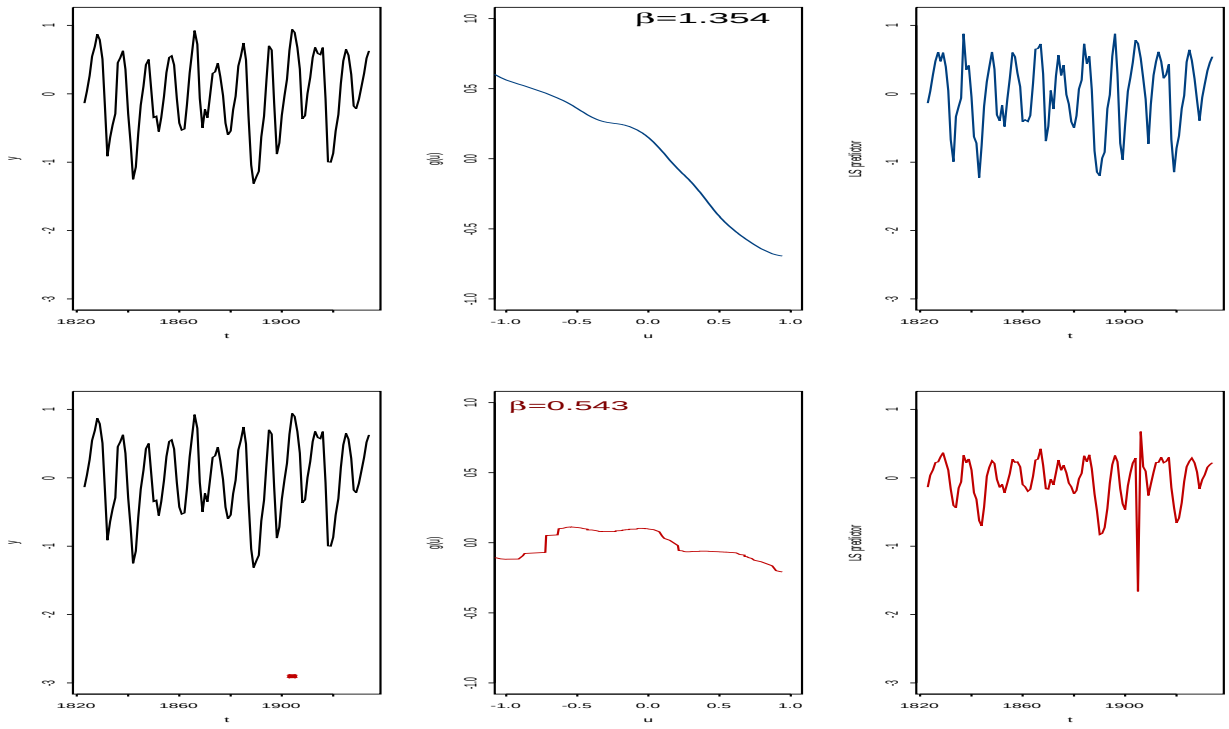

Figure 1: Lynx data, Estimated g function and Predicted Values using the classical procedure. Upper plots correspond to original data, while lower ones to the contaminated series.

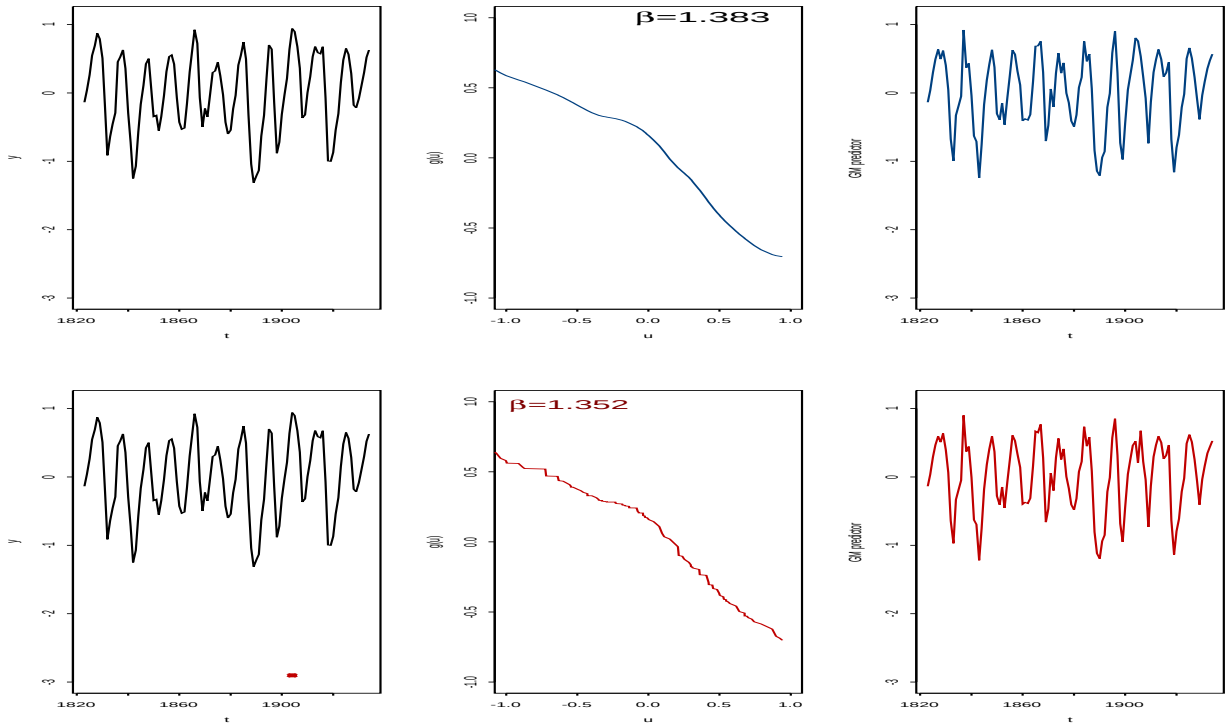

Figure 2: Lynx data, Estimated g function and Predicted Values using the GM–estimators. Upper plots correspond to original data, while lower ones to the contaminated series.

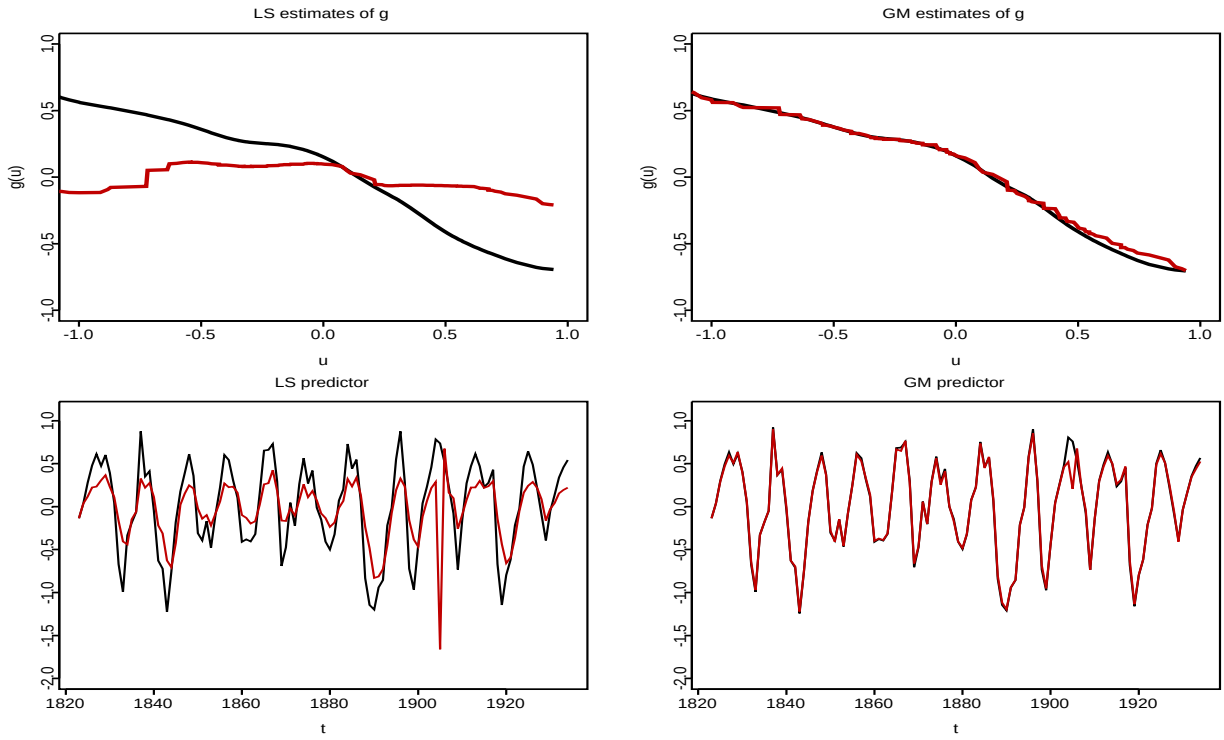

Figure 3: Estimated g function (upper plots) and fitted values (lower plost) for lynx data. Black lines correspond to the original data, while red ones to the modified data.

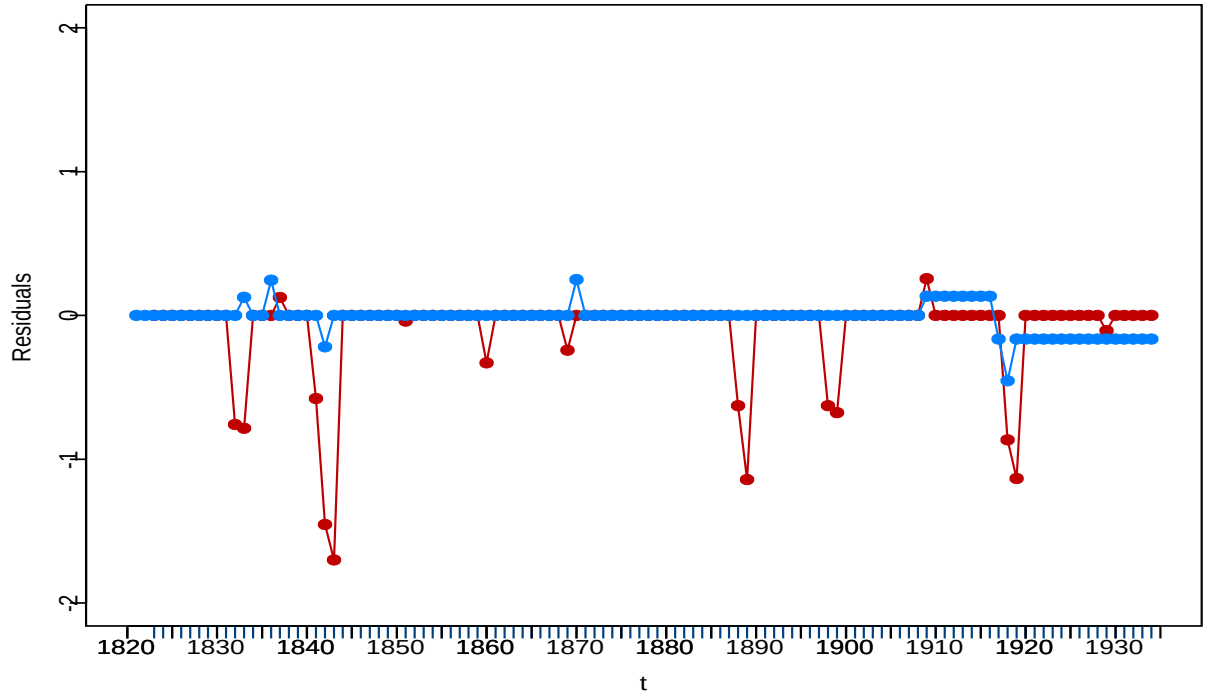

Figure 4: Residuals $\widetilde{r}_t$. The red points corresponds to the detection residuals defined by (11), while the blue ones to the cleaned residuals based on a robust ARMA(3,3) fit.

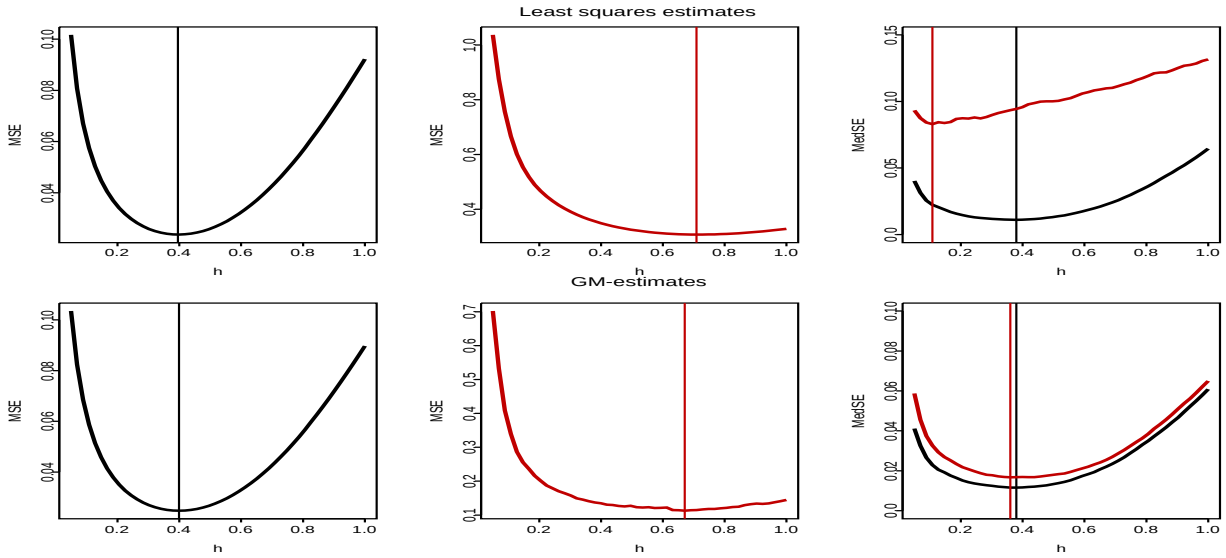

Figure 5: $MSE(h)$ on the left and middle panels and $MedSE(h)$ on the right ones. The upper plots correspond to the classical estimate, while the lower ones to the robust estimator. In red are plotted the results over contaminated samples. The vertical lines show the point where the minimum value is attained

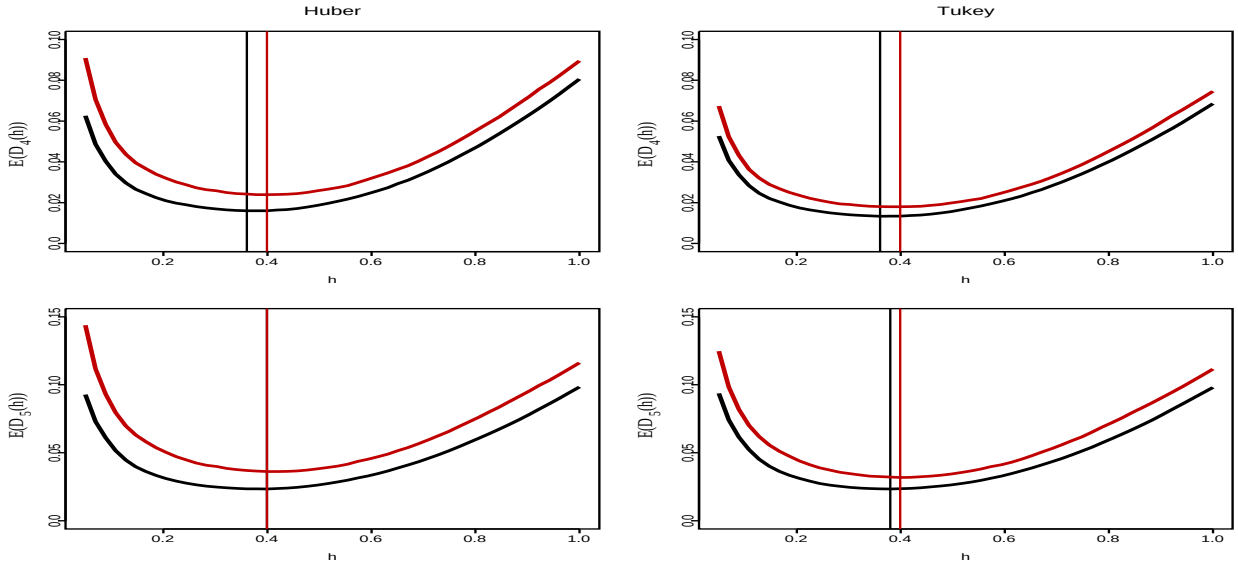

Figure 6: The upper plots correspond to the estimates of $E(D_4(h))$, while the lower one to those of $E(D_5(h))$. The black lines correspond to normal errors and in red are plotted the results over contaminated samples. The vertical lines show the point where the minimum value is attained

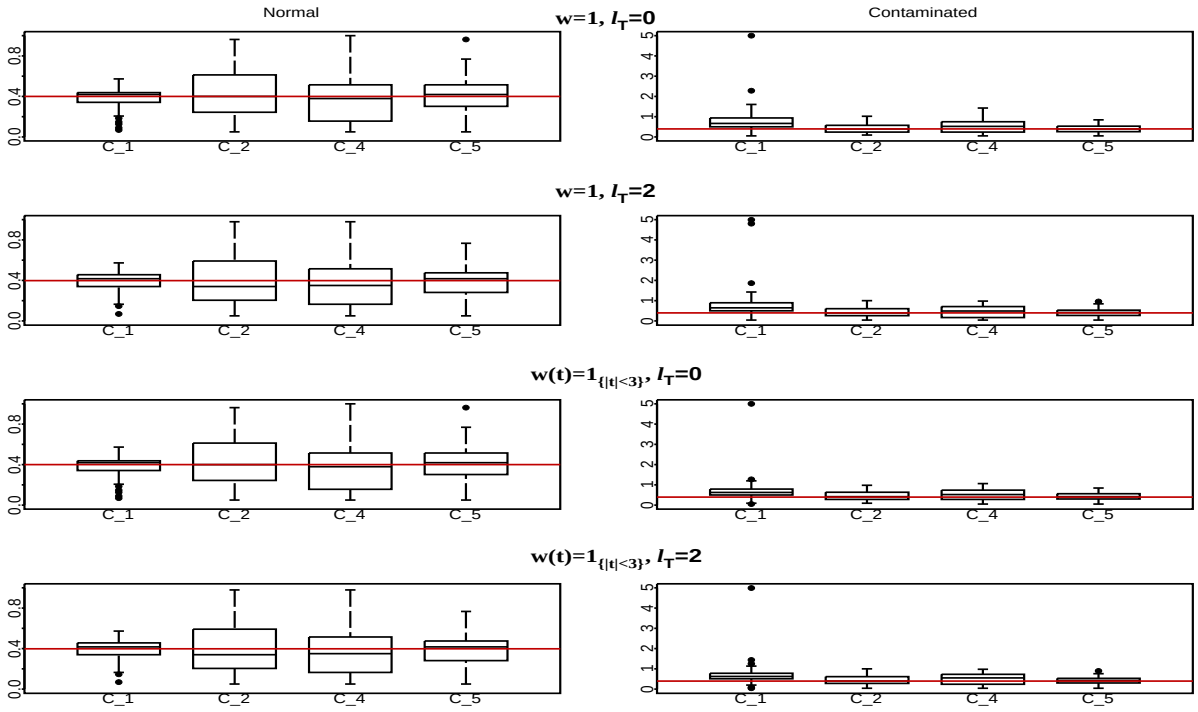

Figure 7: Boxplots of the bandwidth obtained through cross–validation. The red line corresponds to $h = 0.3989$.

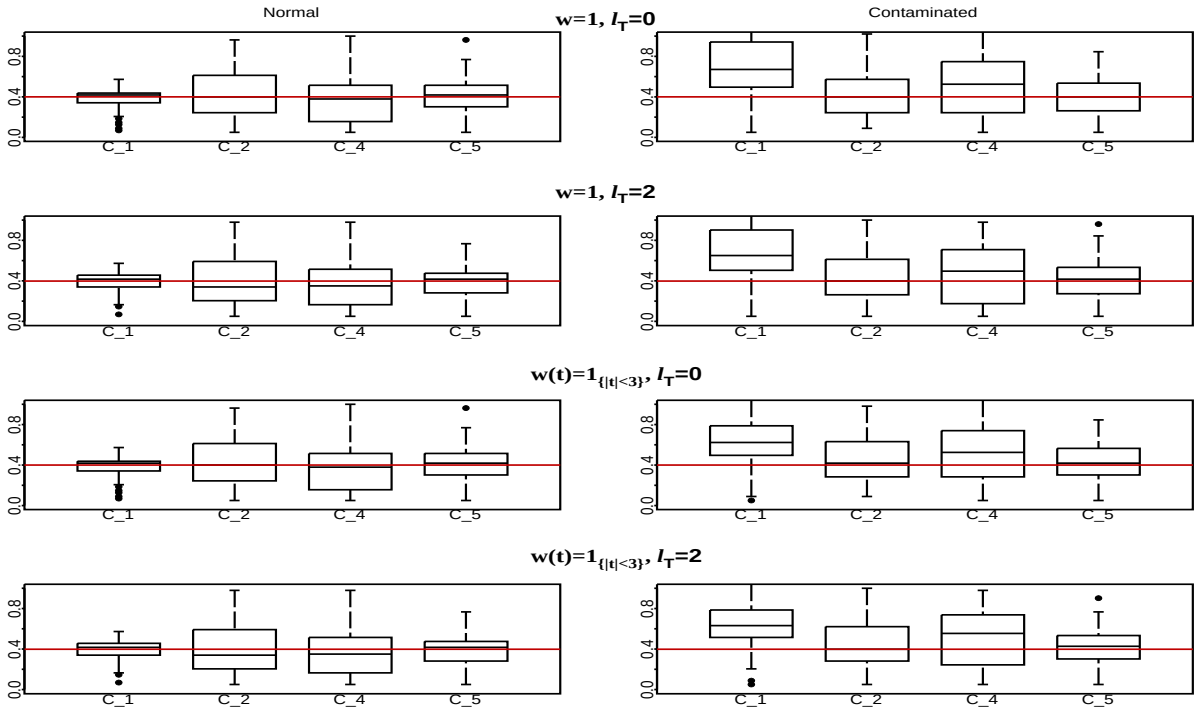

Figure 8: Boxplots of the bandwidth obtained through cross–validation. The red line corresponds to $h = 0.3989$.

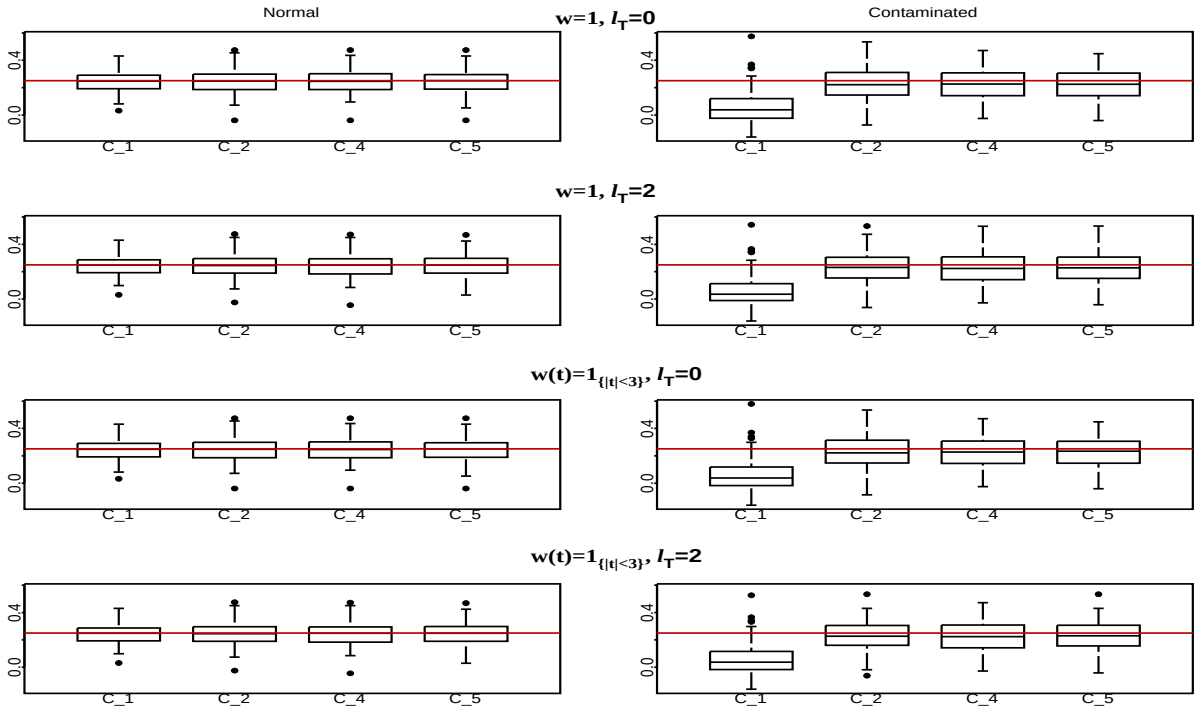

Figure 9: Boxplots of the data–driven estimates of β_o. The red line corresponds to the true value.

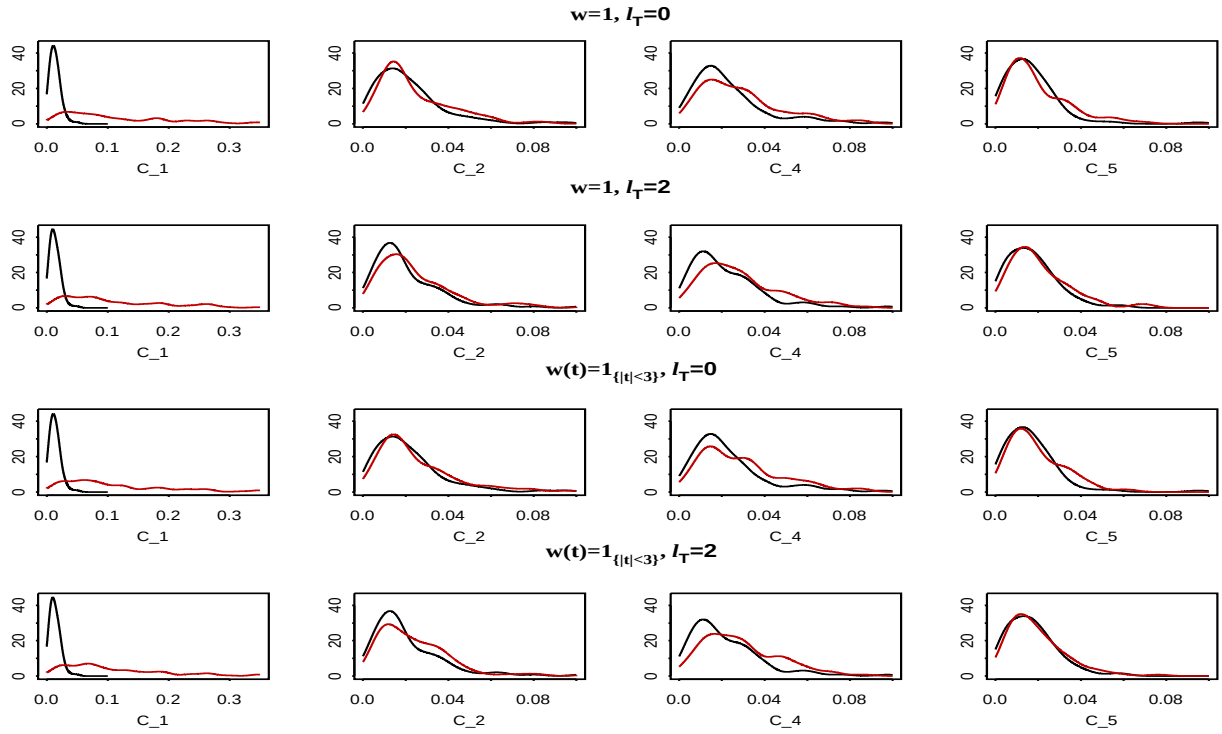

Figure 10: Density estimator of $M(\widehat{g}, g)$. The black lines correspond to normal errors, while the red ones to the contaminated samples.