doi:10.1111/j.1467-9892.2006.00511.x

ROBUST ESTIMATORS UNDER SEMI-PARAMETRIC PARTLY LINEAR AUTOREGRESSION: ASYMPTOTIC BEHAVIOUR AND BANDWIDTH SELECTION

By Ana Bianco and Graciela Boente

Universidad de Buenos Aires, CONICET and Universidad de Buenos Aires

First Version received November 2004

Abstract. In this article, under a semi-parametric partly linear autoregression model, a family of robust estimators for the autoregression parameter and the autoregression function is studied. The proposed estimators are based on a three-step procedure, in which robust regression estimators and robust smoothing techniques are combined. Asymptotic results on the autoregression estimators 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 to compare the performance of the proposed criteria. Finally, the asymptotic distribution of the autoregression parameter estimator is stated uniformly over the smoothing parameter.

Keywords. Partly linear autoregression; robust estimation; smoothing techniques; cross-validation; rate of convergence; asymptotic properties; filtering; prediction.

MSC. Primary 62F35, Secondary 62H25.

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 *et al*. (1985), Heckman (1986), Engle *et al*. (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 neighbours techniques. An extensive description of the different results obtained in partly linear regression models can be found in Härdle *et al*. (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 (autoregressive conditional heteroscedastic) 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 semi-parametric approach, several authors have introduced partly linear models for autoregressive

0143-9782/07/02 274–306 JOURNAL OF TIME SERIES ANALYSIS Vol. 28, No. 2

models 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

$$y_t = \sum_{i=1}^{p} \beta_{0,i} y_{t-c_i} + \sum_{j=1}^{q} g_j(y_{t-d_j}) + \epsilon_t, \tag{1}$$

where $g_j : \mathbb{R} \to \mathbb{R}$ are smooth functions and ϵ_t are independent and identically distributed (i.i.d.) random variables, with symmetric distribution and independent of $\{y_{t-j}, j \geq 1\}$. In the classical setting, it is usually required that $E\epsilon_t = 0$ and $E\epsilon_t^2 < \infty$, while in this article, as is usual in the robust literature, these moment assumptions will not be required. For simplicity and convenience, we will only consider the case $p = q = 1$, $c_1 = 1$, $d_1 = 2$, which leads to the model

$$y_t = \beta_0 y_{t-1} + g(y_{t-2}) + \epsilon_t, \tag{2}$$

where $-1 < \beta_0 < 1$ is an unknown parameter to be estimated, $g : \mathbb{R} \to \mathbb{R}$ is an unknown smooth function and ϵ_t are as in eqn (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_0, based on a piecewise polynomial approximation, under an α-mixing condition. Gao (1995, 1998) also studied the asymptotic normality and obtained a law of iterated logarithm for the kernel-based estimators, $\tilde{\beta}_{LS}$, while Liang (1996) and Gao and Yee (2000) derived some other results. See also Härdle *et al.* (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_0 the true value of the parameter. In the classical setting, since it is assumed that second moments exist, we have that $g(y) = \phi_2(y) - \beta_0 \phi_1(y)$, 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 estimators of the conditional expectations can be inserted prior to the estimation of the autoregression parameter. Usually, these estimators are linear on the observations and therefore, sensitive to outliers.

Bianco and Boente (2002) considered a different approach which does not require any moments to the errors. Let $\phi_1(y)$ and $\phi_2(y)$ be now any conditional location functionals related to a robust smoother. From now on, we will refer to this kind of functionals as a robust conditional location functional. A typical

example is the conditional median. The definition of a robust location conditional functional, related to a score function ψ, was first introduced in Boente and Fraiman (1989) who noted that it is a natural extension of the conditional expectation. In Thm 2.1 of Boente and Fraiman (1989), it was shown that if the score function ψ is a strictly increasing bounded continuous score function, the robust location conditional functional exists, is unique and measurable. Furthermore, its weak continuity was proved in Thm 2.2 therein. Therefore, by applying this functional to weak consistent estimators 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 estimators 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 $g(y) = \phi_2(y) - \beta_0\phi_1(y)$, we can re-write eqn (2) as $y_t - \phi_2(y_{t-2}) = \beta_0(y_{t-1} - \phi_1(y_{t-2})) + \epsilon_t$. 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 $g(y) = \phi_2(y) - \beta_0\phi_1(y)$ 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 estimators, with a more resistant behaviour, 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, such as local M-type estimators or local medians. Denote $\hat{\phi}_1(y)$ and $\hat{\phi}_2(y)$ the obtained estimators.
- *Step 2*. Estimate the autoregression parameter by applying a robust regression estimator to the residuals $y_t - \hat{\phi}_2(y_{t-2})$ and $y_{t-1} - \hat{\phi}_1(y_{t-2})$. Denote $\tilde{\beta}$ the resulting estimator.
- *Step 3*. Define the estimator of the autoregression function g as $\hat{g}(y) = \hat{\phi}_2(y) - \tilde{\beta}\hat{\phi}_1(y)$.

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 estimators based on M-type smoothing splines for nonparametric and semi-parametric 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 some choices for the estimators considered in Steps 1 and 2.

Consider $\hat{F}_1(z|y_{t-2} = y)$ and $\hat{F}_2(z|y_{t-2} = y)$ the estimators 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

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

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

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

$$w_{tT}(y)=K\left(\frac{y_{t-2}-y}{h_T}\right)\left\{\sum_{t=3}^{T} K\left(\frac{y_{t-2}-y}{h_T}\right)\right\}^{-1}. \tag{3}$$

The function $K:\mathbb{R}\to\mathbb{R}$ is a kernel function, i.e. a non-negative integrable function on $\mathbb{R}$.

As mentioned above, local kernel M-type estimators, $\hat{\phi}_1(y)$ and $\hat{\phi}_2(y)$, defined through a score function ψ, can be considered. These estimators can be viewed as applying the robust M-location conditional functional to the empirical conditional distributions $\hat{F}_1(z|y_{t-2}=y)$ and $\hat{F}_2(z|y_{t-2}=y)$ and so they are the solution of

$$\sum_{t=3}^{T} K\left(\frac{y_{t-2}-y}{h_T}\right)\psi\left(\frac{y_{t-1}-\hat{\phi}_1(y)}{s_1(y)}\right)=0 \tag{4}$$

$$\sum_{t=3}^{T} K\left(\frac{y_{t-2}-y}{h_T}\right)\psi\left(\frac{y_{t}-\hat{\phi}_2(y)}{s_2(y)}\right)=0 \tag{5}$$

with $s_j(y)$ the residual scale. Possible choices of the score function ψ are Huber's function or $\psi(u)=\mathrm{sg}(u)$ that leads to the local medians. For a discussion regarding the choice of the score function leading to the robust location conditional functionals, see He *et al.* (2002). 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 $\psi(u)=sg(u)$ is the natural choice. On the other hand, if the conditional distribution is assumed to be symmetric, the conditional distribution has as natural centre 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 estimators, $\hat{\phi}_1(y)$ and $\hat{\phi}_2(y)$, of $\phi_1(y)$ and $\phi_2(y)$ respectively, the robust estimation of the regression parameter can be perfomed by applying to the residuals $\hat{r}_t=y_t-\hat{\phi}_2(y_{t-2})$ and $\hat{z}_t=y_{t-1}-\hat{\phi}_1(y_{t-2})$, any of the robust methods proposed for linear regression. Among the most popular robust regression estimators, we find GM-estimators (generalized M-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 τ-estimators could be used (Yohai and Zamar, 1988).

In this article, we will focus on the behaviour of any robust autoregression estimator defined as the solution of

$$\sum_{t=3}^{T} \psi_1\left(\frac{\hat{r}_t - \hat{\beta}\hat{z}_t}{s_T}\right) w_2(\hat{z}_t)\hat{z}_t w_3(y_{t-2}) = 0, \tag{6}$$

with $\hat{r}_t = y_t - \hat{\phi}_2(y_{t-2})$, $\hat{z}_t = y_{t-1} - \hat{\phi}_1(y_{t-2})$ and s_T an estimator of the residuals scale σ_0. Besides, ψ_1 is a bounded odd score function, while w_2 and w_3 are weight functions. 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. It looks natural, that, under regularity conditions, the solution $\hat{\beta}$ of eqn (6) will converge in probability to the value $\beta(F)$ solution of

$$E_F\left[\psi_1\left(\frac{r_t - \beta(F)z_t}{\sigma_0}\right)\psi_2(z_t)w_3(y_{t-2})\right] = 0, \tag{7}$$

with $\psi_2(t) = tw_2(t)$, $r_t = y_t - \phi_2(y_{t-2})$ and $z_t = y_{t-1} - \phi_1(y_{t-2})$. Since $g(y) = \phi_2(y) - \beta_0\phi_1(y)$, we have that $y_t - \phi_2(y_{t-2}) = \beta_0(y_{t-1} - \phi_1(y_{t-2})) + \epsilon_t$, and so eqn (7) is equivalent to

$$E_F\left[\psi_1\left(\frac{(\beta_0 - \beta(F))z_t + \epsilon_t}{\sigma_0}\right)\psi_2(z_t)w_3(y_{t-2})\right] = 0.$$

If the score function is stricly increasing, to get Fisher-consistent estimators, i.e. that $\beta(F) = \beta_0$, one only needs to require $E_F[\psi_1(\epsilon_t/\sigma_0)] = 0$, because of the independence between the errors ϵ_t and the past observations. 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). Note that the expectation involved in eqn (7) exists since, from conditions N1, N3 and N6 below, the score functions ψ_1 and ψ_2 and the weight function w_3 are bounded.

In this article, we will study the asymptotic behaviour of the autoregression estimators defined through Steps 1 to 3. We will also propose a procedure to obtain detection residuals and robust predictors of the series. The article is organized as follows. In Section 2, through an example, we illustrate the effect of the outliers on the estimation of β_0 when using the classical estimator and the corresponding robust procedure. In Section 3, a procedure to detect outlying observations is proposed. In Section 4, we derive the asymptotic distribution of the estimators of the autoregression parameter β_0. 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 Section 1, the sensitivity of the least squares estimators 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 behaviour, we have considered the Canadian lynx data which has been widely studied. This data set 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 and considers the variable

$$y_t = \log_{10}(\text{number of lynx trapped in the year}(1822 + t)) - 2.9036,$$

$1 \leq t \leq T = 114$. It was studied, among others, by Campbell and Walker (1977) and Tong (1977) who fitted an AR(11) and an ARMA(3,3) and by 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, and Härdle *et al.* (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 behaviour of the estimated functions both for the least squares and a robust procedure. The results for the original data set are plotted in solid lines while those for the modified one, in dashed lines. The robust procedure is mainly unaffected. As expected, when using the classical estimators, not only the autoregression parameter changes, but also does the shape of the autoregression function, which decreases more slowly. Note that, for the original data, the least squares estimate, $\hat{\beta}_{LS}$, equals 1.354 and the robust estimate, $\hat{\beta}$, takes the value 1.383, so the estimations are quite similar. On the other hand, for the contaminated data, $\hat{\beta}_{\mathrm{LS}}$ changes to 0.543 but $\hat{\beta} = 1.352$, 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. Solid lines correspond to the original data and dashed ones to the modified data. From these plots, it is clear that the least squares predictors are modified not only at time $t = 85$, but also the influence

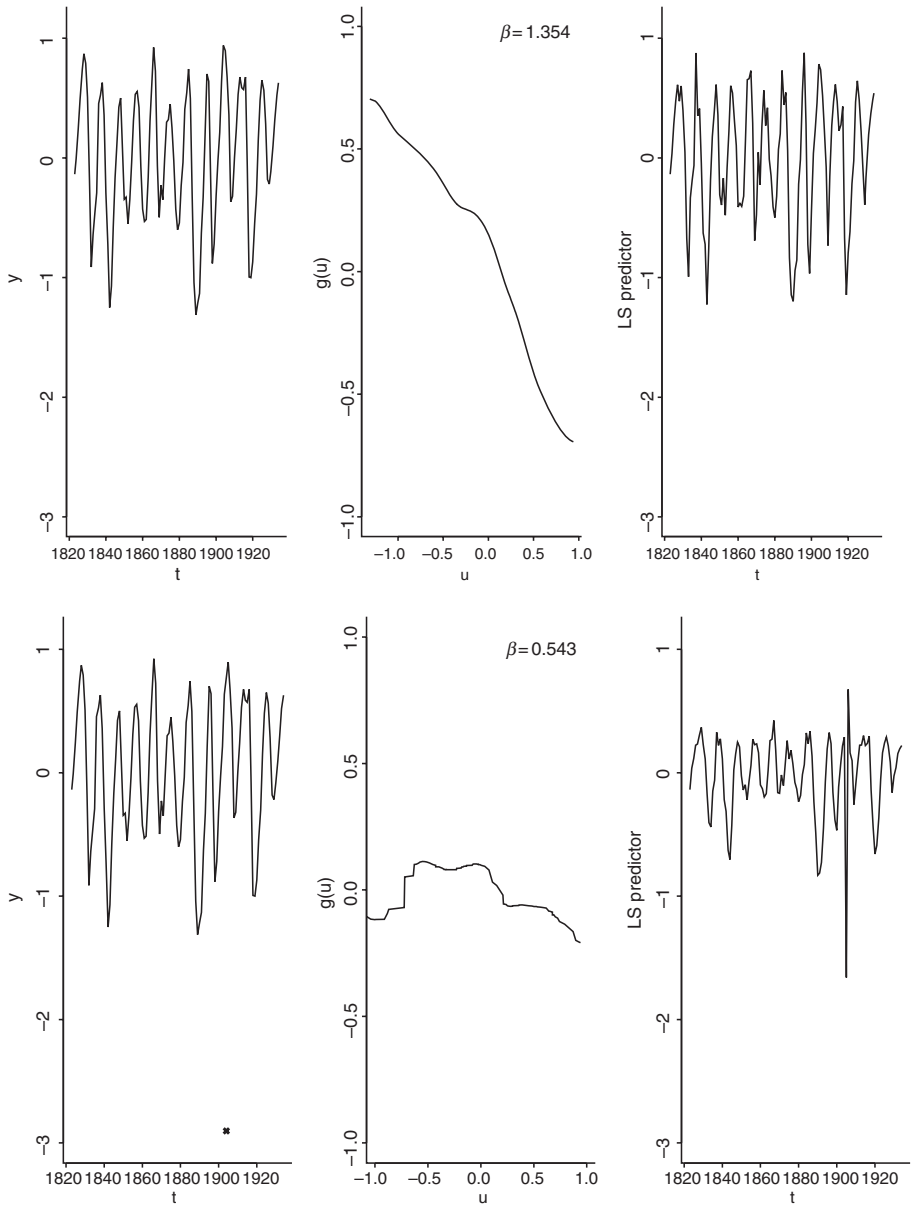

FIGURE 1. Lynx data, estimated g function and predicted values using the classical procedure. Upper plots correspond to original data, while lower ones to the contaminated series.

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.

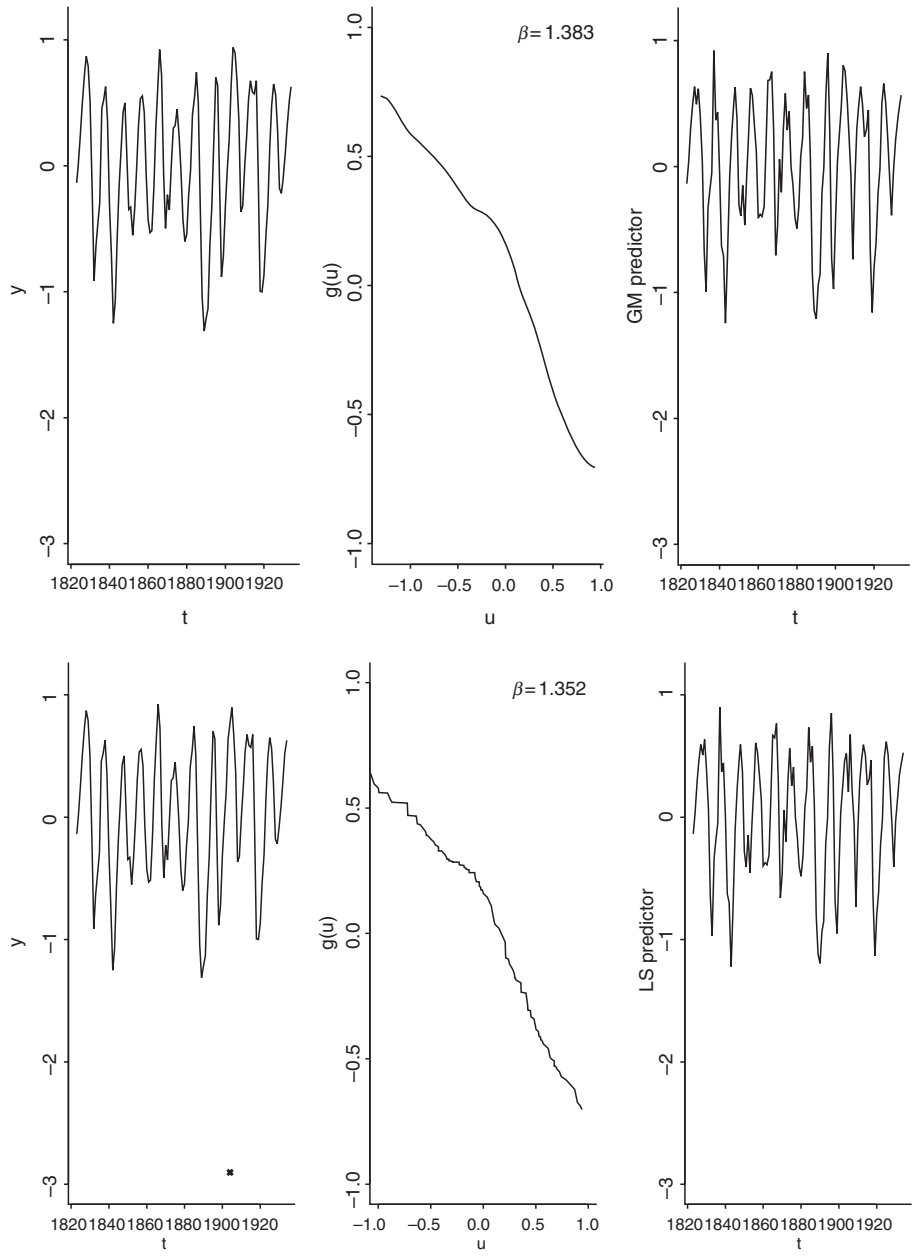

FIGURE 2. Lynx data, estimated g function and predicted values using the GM-estimators. Upper plots correspond to original data, while lower ones to the contaminated series.

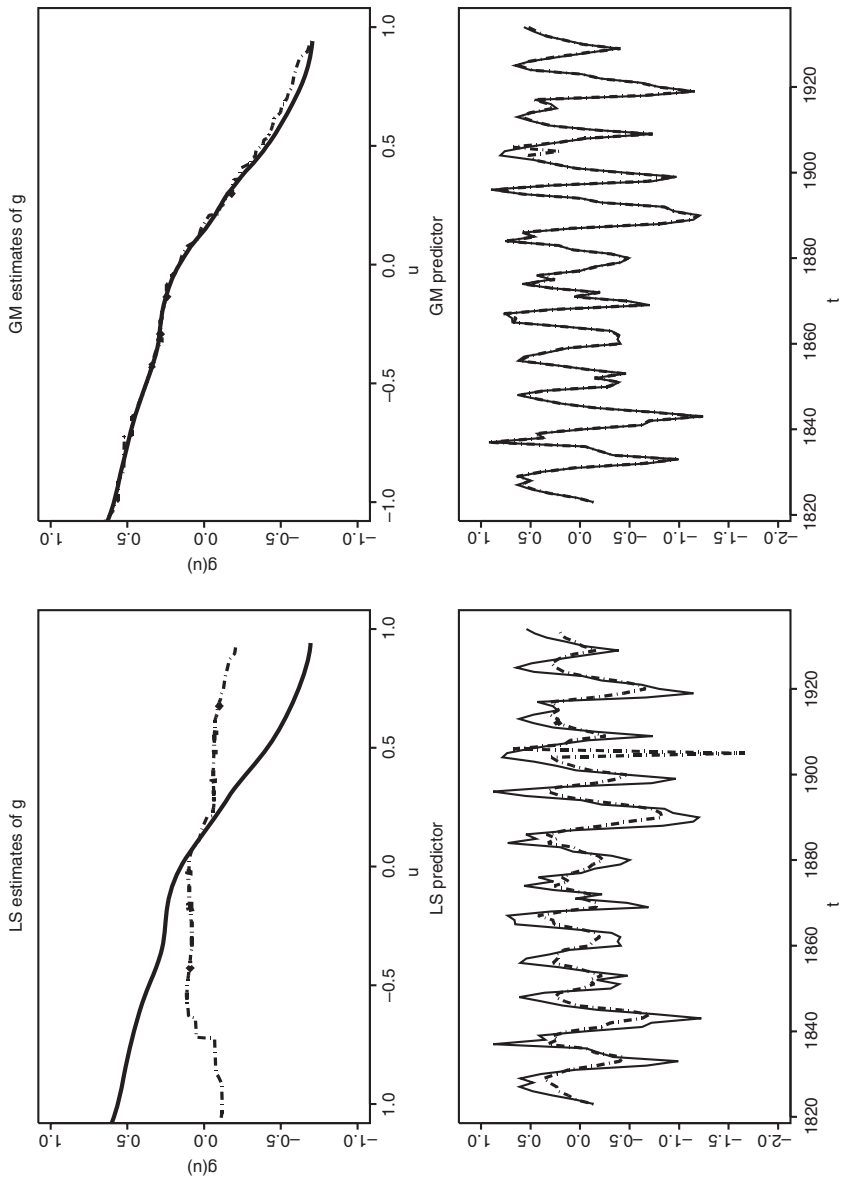

FIGURE 3. Estimated g function (upper plots) and fitted values (lower plots) for lynx data. Solid lines correspond to the original data, while dashed ones (– · – · –) to the modified data.

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 will define 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 $\hat{\beta}$ be the estimator of the autoregression parameter introduced in eqn (6), related to a score function ψ_1 and hard-rejection weights w_2 and w_3. Denote $\hat{g}(y)$ the estimator of the autoregresion function g as described in Step 3.
- Compute

$$R_t = \frac{\hat{r}_t - \hat{\beta}\hat{z}_t}{s_T}, \quad \text{where } s_T = \frac{1}{0.6745}\text{median}(|\hat{r}_t - \hat{\beta}\hat{z}_t|).$$

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

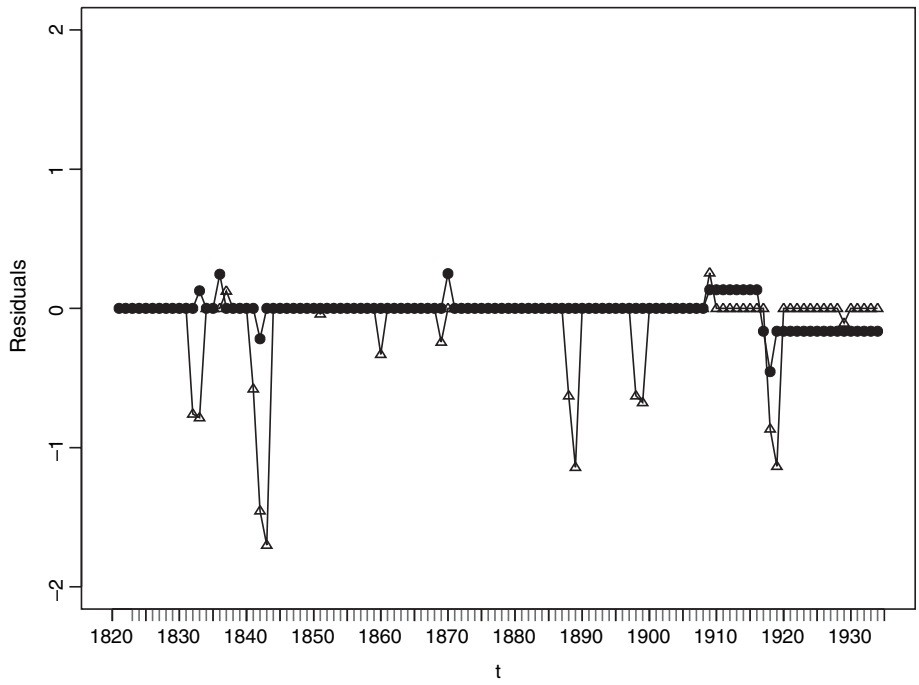

FIGURE 4. Residuals $\tilde{r}_t$. The triangles correspond to the detection residuals defined by eqn (8), while the filled points to the cleaned residuals based on a robust ARMA(3,3) fit.

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

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

$$\tilde{r}_t = \begin{cases} 0, & \text{if } |R_t| < \alpha, w_2(\hat{z}_t) > 0 \text{ and } w_3(y_{t-2}) > 0 \\ y_t - \hat{y}_t, & \text{otherwise.} \end{cases} \tag{8}$$

To compare our procedure with the method proposed by Bianco *et al.* (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 $\tilde{r}_t$ defined in eqn (8), with $\alpha = 0.2$ and as M-smoother the local median. For 'good' data points detection residuals are zero, while suspicious observations correspond to nonzero residuals. The nonzero residuals detected by our procedure indicate nearly the same anomalous data points as those revealed by Brillinger's (1986) plot, while, as shown in 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 article.

4. ASYMPTOTIC DISTRIBUTION

The consistency of the robust procedure defined through Steps 1 to 3, follows under conditions analogous to those stated in Bianco and Boente (2004) using the Ergodic Theorem instead of the Strong Law of Large Numbers.

In this section, we will derive the asymptotic distribution of the autoregression parameter estimators defined as any solution of eqn (6) with $\hat{\phi}_1(y)$ and $\hat{\phi}_2(y)$ consistent estimators of robust location conditional functionals $\phi_1(y)$ and $\phi_2(y)$, satisfying $\phi_2(y) = \beta_0 \phi_1(y) + g(y)$.

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_0, i.e. we will consider the solution $\hat{\beta}$ of

$$\sum_{t=3}^{T} \psi_1 \left(\frac{\hat{r}_t - \hat{\beta} \hat{z}_t}{\sigma_0} \right) w_2(\hat{z}_t) \hat{z}_t w_3(y_{t-2}) = 0. \tag{9}$$

If σ_0 is estimated by s_T, the asymptotic normality can be derived by requiring that $s_T \xrightarrow{p} \sigma_0$, if, in addition, $t^2 \psi_1''(t)$ is bounded.

As in Section 1, denote $r_t = y_t - \phi_2(y_{t-2})$ and $z_t = y_{t-1} - \phi_1(y_{t-2})$ and hence, $\epsilon_t = r_t - \beta_0 z_t$.

We will need the following set of assumptions:

N0. The process $\{y_t : t \geq 3\}$ is a strictly stationary α-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. $E\big(w_2(z_t)z_t^2\big) < \infty$ and

$$A = E\left(\psi_1'\left(\frac{\epsilon_t}{\sigma_0}\right) w_2(z_t) z_t^2 w_3(y_{t-2})\right) = E\left(\psi_1'\left(\frac{\epsilon_t}{\sigma_0}\right)\right) E\big(w_2(z_t)z_t^2 w_3(y_{t-2})\big) > 0.$$

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. $E(\psi_2(z_t)|y_{t-2} = y) = 0$ 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 compact support $\mathcal{K} \subseteq \mathcal{S}$, where $\mathcal{S}$ denotes the support of the marginal distribution of y_t. Moreover, $\|w_3\|_\infty \leq 1$.

REMARK 1.

- With respect to N0, the inclusion of a dependence structure, usually imposing a mixing condition, allows to estimate 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 α-mixing is weaker than that of φ-mixing (uniform strongly mixing), which is a more often studied condition (see Billingsley, 1968). The φ-condition is rather restrictive when we are considering autoregressive models, since for Gaussian stationary processes the φ-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 φ- or α-mixing condition, see, for instance, Bosq (1996), Györfi *et al.* (1989) and Härdle (1990), for a review. Asymptotic normality results for the Nadaraya–Watson estimators, in α-mixing processes, were obtained by Robinson (1983). An α-mixing condition was also required when studying the asymptotic behaviour of local M-estimators 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 α-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 estimators; see also Gao and Yee (2000).

Doukhan (1994, Thm 7, p. 102) gives sufficient conditions on the function g, the autoregression parameter and on the errors distribution that guarantee that the process will be α-mixing. For instance, if g is bounded and the errors ϵ_t have a density and finite first moment, then the condition $|\beta_0| < 1$ entails that the process is geometrically ergodic and thus, α-mixing. When g is unbounded it should be required that there exist some positive constants b and v_0 and some $a \geq 0$, such that $|g(v)| \leq a|v| - b$ for $|v| > v_0$, $\sup_{|v| \leq v_0} |g(v)| < \infty$ and the unique non-negative zero of the polynomial

$$P(z) = z^2 - |\beta_0| z - a, \quad \text{i.e. } \rho = \frac{|\beta_0| + \sqrt{\beta_0^2 + 4a}}{2}, \quad \text{satisfies } \rho < 1.$$

- As noted by Robinson (1988), condition N2 will prevent any element of y_{t-1} from being almost surely perfectly predictable by y_{t-2}.
- It is worth noticing that if, for instance, z_t has a symmetric distribution and ψ_2 is an odd function, condition N4 is fulfilled. Besides, N4 is also satisfied if $\psi_2 = \psi$, the score function used in eqn (4) to compute the local kernel M-type estimators, $\hat{\phi}_1(y)$. Assumption N4 is needed to obtain a uniform result over a class of Lipschitz functions, using the results given in Arcones (1996). For VC-classes (Vapnik-Červonenkis classes) of functions, Andrews and Pollard (1994) and Yu (1994) provided a similar result for strong mixing triangular arrays and for stationary α-mixing sequences respectively.
- The smoothness condition N5 is a standard requirement in classical kernel estimation for semi-parametric models to guarantee asymptotic normality, see, for instance, Robinson (1988) and Severini and Wong (1992).

THEOREM 1. *Let* $\{y_t, j \geq 3\}$ *be a stationary* α*-mixing process satisfying eqn* (2) *with* ϵ_t *having a symmetric distribution and such that* ϵ_t *is independent of* $\{y_{t-j}, j \geq 1\}$. *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* $\phi_2(y) = \beta_0 \phi_1(y) + g(y)$. *Assume that N*0 *to N*4 *and N*6 *hold. Moreover, assume that one of the following two assumptions* (*a*) *or* (*b*) *are satisfied*

(*a*) *N*5 *holds and* $\hat{\phi}_j(\mathrm{y})$ *are robust estimators of* $\phi_j(y)$ *such that, for* $j = 1, 2$, $\hat{\phi}_j(\mathrm{y})$ *is continuously differentiable and*

$$T^{1/4} \sup_{y \in \mathcal{K}} |\hat{\phi}_j(y) - \phi_j(y)| \xrightarrow{p} 0, \tag{10}$$

$$\sup_{y \in \mathcal{K}} |\hat{\phi}'_j(y) - \phi'_j(y)| \xrightarrow{p} 0, \tag{11}$$

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

(b) For $j = 1, 2$, $\hat{\phi}_j(y)$ are robust estimators of $\phi_j(y)$ which admit a linear expansion $\hat{\phi}_j(y) - \phi_j(y) = \hat{\mathcal{L}}_j(y) + \hat{\mathcal{R}}_j(y)$, where $\hat{\mathcal{L}}_j(y) = \sum_{t=3}^{T} w_{tT}(y) v_j\big(y_{t-2+j}, y\big)$, with v_j bounded functions, such that $E(v_j(y_{t-2+j}, y)|y_{t-2} = y) = 0$ almost everywhere. Moreover, assume that for $j = 1, 2$ and for the compact set $\mathcal{K}$ defined in N6

$$T^{1/4} \sup_{y \in \mathcal{K}} |\hat{\mathcal{L}}_j(y)| \xrightarrow{p} 0, \tag{12}$$

$$T^{1/2} \sup_{y \in \mathcal{K}} |\hat{\mathcal{R}}_j(y)| \xrightarrow{p} 0, \tag{13}$$

$$T^{-1/2} \left| \sum_{t=3}^{T} \hat{\mathcal{L}}_j(y_{t-2}) \vartheta_1(\epsilon_t) \vartheta_2(z_t) w_3(y_{t-2}) \right| \xrightarrow{p} 0 \tag{14}$$

hold for bounded functions ϑ_1 and ϑ_2 such that, for almost all y, the product $E(\vartheta_1(\epsilon_t))E(\vartheta_2(z_t)|y_{t-2} = y) = 0$.

Then, under (a) or (b) we have that $T^{1/2}(\hat{\beta} - \beta_0) \xrightarrow{D} N(\mathbf{0}, \sigma^2_{\psi_1, w_2, w_3})$, where

$$\sigma^2_{\psi_1, w_2, w_3} = \sigma_0^2 \frac{E\left(\psi_1^2\left(\frac{\epsilon_t}{\sigma_o}\right)\right)}{\left[E\left(\psi_1'\left(\frac{\epsilon_t}{\sigma_0}\right)\right)\right]^2} \frac{E\big(w_2^2(z_t) z_t^2 w_3^2(y_{t-2})\big)}{\big[E\big(w_2(z_t) z_t^2 w_3(y_{t-2})\big)\big]^2}.$$

Remark 2.

- When $w_2 \equiv 1$ and $w_3 \equiv 1$, the asymptotic efficiency of the autoregression estimators is the same as in the one-dimensional location setting.
- Conditions (10) and (11) are related to the trade-off between the stochastic equicontinuity needed to derive the asymptotic distribution of $\hat{\beta}$ and the smoothness requirements on the estimators $\hat{\phi}_j$, when plugging-in general preliminary estimators of ϕ_j. For a highlighting discussion on this task we refer to Sect. 4.3 in Andrews (1994). It is worth noticing that, even in the independent setting, when dealing with semi-parametric models, derivability of the estimators 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 eqns (10) and (11) are fulfilled when we consider,

in Step 1, local kernel M-type estimators solutions of eqns (4) and (5) when the optimal bandwidth is used. The convergence requirements in (a) are analogous to those required in Cond. (7) in Severini and Staniswalis (1994, p. 510) and are needed to obtain the desired rate of convergence for the autoregression estimators. More precisely, assumption (10) avoids the bias term and ensures that $G_T(\hat{\phi}_1, \hat{\phi}_2)$ will behave asymptotically as $G_T(\phi_1, \phi_2)$, where

$$G_T(\upsilon_1, \upsilon_2) = \frac{1}{\sqrt{T-2}} \sum_{t=3}^{T} \psi_1\left(\frac{r_t(\upsilon_2) - \beta_0 z_t(\upsilon_1)}{\sigma_0}\right) \psi_2(z_t(\upsilon_1)) w_3(y_{t-2}),$$

with $r_t(\upsilon_2) = y_t - \upsilon_2(y_{t-2})$, $z_t(\upsilon_1) = y_{t-1} - \upsilon_1(y_{t-2})$, for any differentiable functions $\upsilon_j : \mathbb{R} \to \mathbb{R}$, $j = 1, 2$.

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 1 includes other estimators than those based on kernel weights. In Theorem 2, a stronger version of assumption (b) will be required in order to derive a uniform result over the smoothing parameter.

- Assumptions (a) and (b) in Theorem 1 are fulfilled, under mild conditions, when we consider local kernel M-type estimators, $\hat{\phi}_1(y)$ and $\hat{\phi}_2(y)$, solutions of eqns (4) and (5). To be more precise, assume that:
  - (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 $F_1(z|y_{t-2} = y)$ and $F_2(z|y_{t-2} = y)$ are uniformly Lipschitz in a neighborhood $\mathcal{U}$ 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 $z \in \mathbb{R}$, $y, v \in \mathcal{U}$;
  - (v) Moreover, the following equicontinuity condition holds, for $j = 1, 2$

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

Then, arguments analogous to those considered in Boente and Fraiman (1991a) allow to show that eqn (10) 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', we have that $\hat{\phi}_1'(y) = -A_T^{-1}(y)[s_1(y)B_T(y) - s_1'(y)C_T(y)]$, where

$$A_T(y) = \sum_{t=3}^{T} K\left(\frac{y_{t-2} - y}{h_T}\right)\psi'\left(\frac{y_{t-1} - \hat{\phi}_1(y)}{s_1(y)}\right)$$

$$B_T(y) = \frac{1}{h_T}\sum_{t=3}^{T} K'\left(\frac{y_{t-2} - y}{h_T}\right)\psi\left(\frac{y_{t-1} - \hat{\phi}_1(y)}{s_1(y)}\right)$$

$$C_T(y) = \sum_{t=3}^{T} K\left(\frac{y_{t-2} - y}{h_T}\right)\eta\left(\frac{y_{t-1} - \hat{\phi}_1(y)}{s_1(y)}\right)$$

with $\eta(u) = u\psi'(u)$. A similar expression can be obtained for $\hat{\phi}_2'(y)$. These expressions suggest that if (i)–(iv) hold, the proof of eqn (11) parallels the proofs given in Härdle and Gasser (1985) and in Boente and Rodriguez (2006) together with the standard arguments used in the α-mixing case.

On the other hand, using Taylor's expansion, it is easy to see that a kernel M-estimator admits the linear expansion given in (b), where the remainder term satisfies eqn (13), since, as mentioned above, M-estimators satisfy eqn (10), when ϕ_j are continuously differentiable functions. Note that this approach avoids the derivability requirements on $\hat{\phi}_j$, 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-estimators. 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 (1991b) pointed out that robust cross-validation methods should be an alternative. Also, Wang and Scott (1994) proposed an L^1 cross-validation method 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 *et al.* (1993) for kernel M-smoothers. On the other hand, the classical plug-in bandwidth selector also breaks down in the presence of outliers. Boente *et al.* (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, $\hat{\beta}(h)$ and $\hat{g}(\cdot, h)$ the estimators computed using the kernel weights (3) with smoothing parameter h. As mentioned by Härdle *et al.* (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 a measure of the goodness of the estimators $\hat{\beta}(h)$ and $\hat{g}(\cdot, h)$. For each bandwidth h they defined

$$D_1(h) = \frac{1}{T-2}\sum_{t=3}^{T}\left(\left\{\hat{\beta}(h)y_{t-1} + \hat{g}(y_{t-2}, h)\right\} - \{\beta_0 y_{t-1} + g(y_{t-2})\}\right)^2 w(y_{t-2})$$
$$= \frac{1}{T-2}\sum_{t=3}^{T} u_t^2(h) w(y_{t-2}),$$

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 estimators, is defined through

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

where $\hat{g}_t(y, h) = \hat{\phi}_{2,t}(y, h) - \tilde{\beta}(h)\hat{\phi}_{1,t}(y, h)$, with $\hat{\phi}_{1,t}(y, h)$ and $\hat{\phi}_{2,t}(y, h)$ the linear smoothers computed with bandwidth h and obtained with all the data except y_{t-2} and $\tilde{\beta}(h)$ is the least squares estimator considering the residuals $y_t - \hat{\phi}_{2,t}(y_{t-2}, h)$ and $y_{t-1} - \hat{\phi}_{1,t}(y_{t-2}, h)$.

A small simulation study was carried out to show that the asympotically optimal bandwidth is very sensitive to outliers. For each value of h, we have computed an estimator of $\mathrm{MSE}(h) = E(D_1(h))$, 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 estimators and for the local M-estimator 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-estimators were computed through an iterative procedure with local medians as initial points. After the robust smoothing, GM-estimators with Huber's function on the residuals with constant 1.6 and bisquare weights on $y_{t-1} - \hat{\phi}_1(y_{t-2})$ 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 estimator. We performed 50 replications. To stabilize the series, we first generate a series of size $N = 1100$ following the model

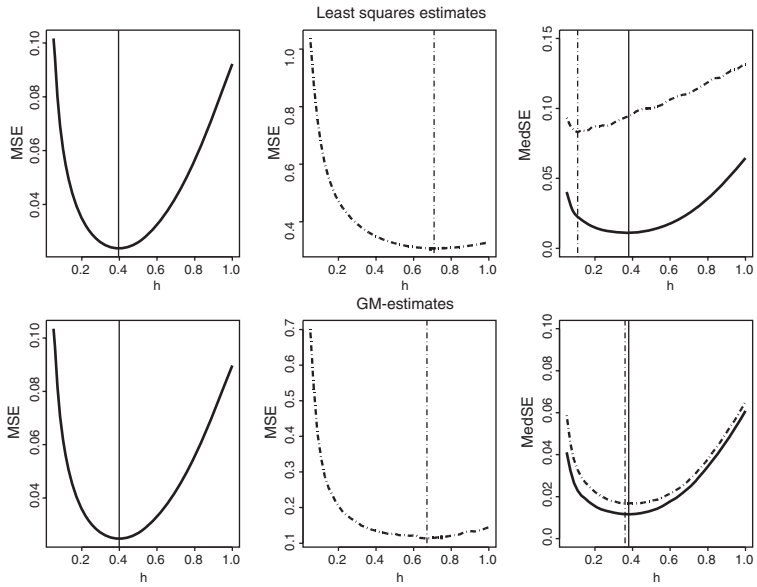

FIGURE 5. MSE(h) on the left and middle panels and MedSE(h) on the right ones. The upper plots correspond to the classical estimate, while the lower ones to the robust estimator. In dashed lines ($- \cdot - \cdot -$) the results are plotted over contaminated samples. The vertical lines show the point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

$$z_t = \beta_0 z_{t-1} + 0.25\pi \sin(\pi z_{t-2}) + \epsilon_t, \quad 3 \le t \le N,$$

where $\beta_0 = 0.25$. As initial values, we took $z_t = \epsilon_t$, for $1 \le t \le 2$. In the case of normal errors, we have chosen $\epsilon_t \sim N(0, \sigma_0^2)$ with $\sigma_0^2 = 0.25$. The data set of size $T = 100$ to be considered consists of 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. $N(0, \sigma_0^2)$ and $y_t = z_{t+1000} + 6\delta_t$ with $\delta_t \sim Bi(1, 0.05)$. The bandwidth h was chosen 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

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

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's or Tukey's function, after scaling the differences, i.e.

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

where $\sigma_T(h) = \mathrm{MAD}(u_t(h))$ and $\hat{\sigma}_T(h) = \mathrm{MAD}(\hat{u}_t(h))$ where MAD stands for the median of the absolute deviations from the median. The results obtained with Huber's function were disappointing and this is due to its unboundness. As expected, results similar to those obtained with Tukey's ρ-function, were obtained by weighting $u_t(h)$ with Huber's weight function, which suggests that the measures defined through

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

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 estimator of the bias, defined through a location estimator μ_T, and on a robust scale estimator σ_T, as follows,

$$D_5(h) = \mu_T^2(u_t(h)w(y_{t-2})) + \sigma_T^2(u_t(h)w(y_{t-2}))$$
$$C_5(h) = \mu_T^2(\hat{u}_t(h)w(y_{t-2})) + \sigma_T^2(\hat{u}_t(h)w(y_{t-2})).$$

We can consider as μ_T the median and as σ_T the bisquare a-scale estimator or the Huber τ-scale estimator. Figure 6 shows the stability of this procedure combined with *GM*-estimators since, for the τ-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 estimator. The procedures based on a ψ-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 τ-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.

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

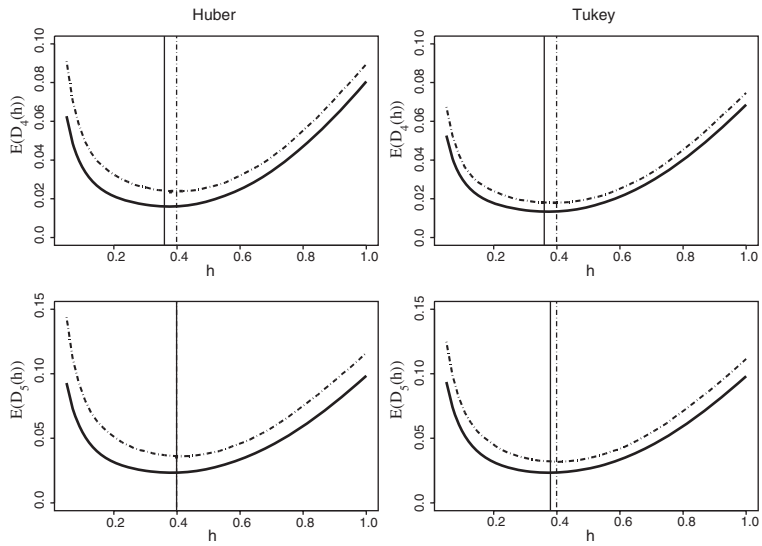

FIGURE 6. The upper plots correspond to the estimates of $E(D_4(h))$, while the lower one to those of $E(D_5(h))$. The solid lines correspond to normal errors and in dashed lines (– · – · –) the results are plotted over contaminated samples. The vertical lines show the point where the minimum value is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

TABLE I

OPTIMAL ASYMPTOTIC BANDWIDTH FOR THE AUTOREGRESSION FUNCTION

	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

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

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

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. Based on this fact, for all the criteria, we have also computed the cross-validation bandwidth with $\ell_T = 2$. Therefore, for each criterion C_j, we obtained four values for the bandwidth corresponding to the two choices of w and to $\ell_T = 0$ and 2. When

using the least squares estimator, we only plot the results for C_1. On the other hand, 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 ψ-function the Huber function, with constant 1.345, while in C_5 the τ-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 estimators of β_0 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 with $\ell_T = 2$ than $\ell_T = 0$, under normal errors. For this particular model, by weighting we obtain a smaller dispersion and the classical procedures performs better even under contamination. Leaving out one data or taking $\ell_T = 2$, produces similar results. More research should be done in this direction to find a way to select the leaving sequence. The best criterion is, in all cases, C_5 in the sense it produces less spread bandwidths with median around the optimal value $h = 0.3989$.

Figure 9 shows the boxplots of the data-driven estimators of β_0. 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

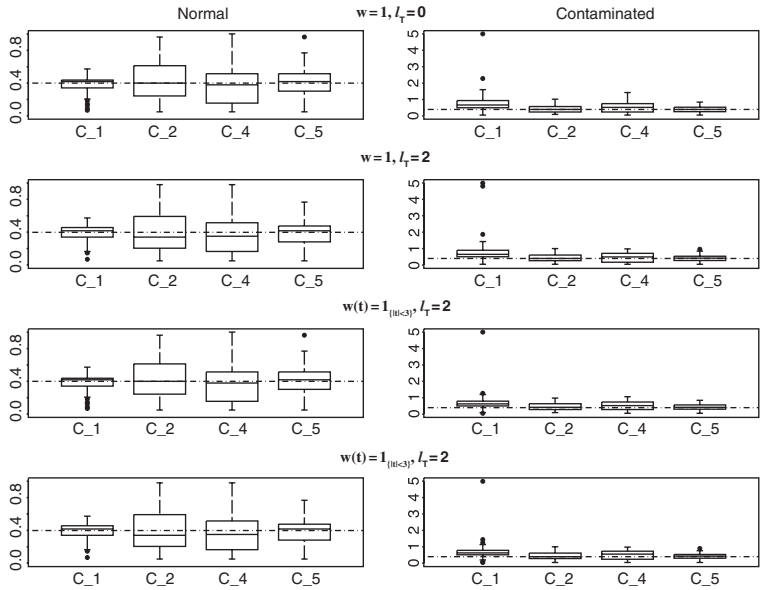

FIGURE 7. Boxplots of the bandwidth obtained through cross-validation. The dashed line (– · – · –) corresponds to $h = 0.3989$.

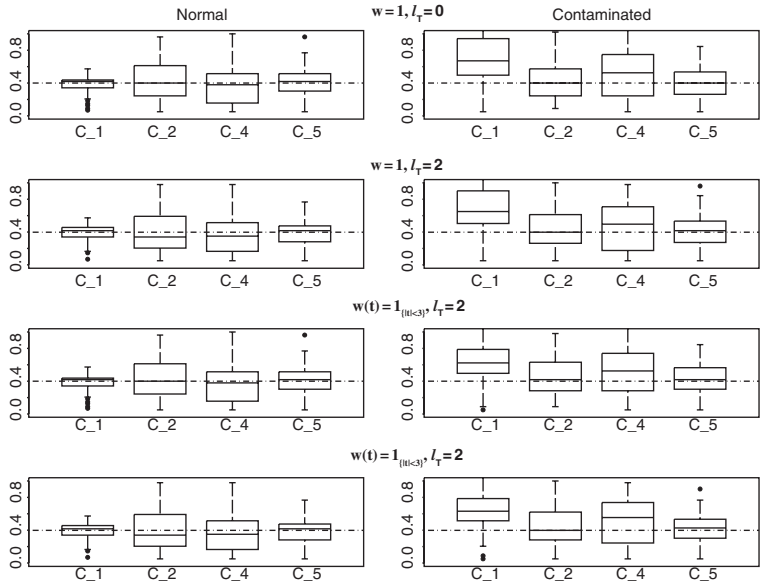

FIGURE 8. Boxplots of the bandwidth obtained through cross-validation. The dashed line (– · – · –) corresponds to $h = 0.3989$.

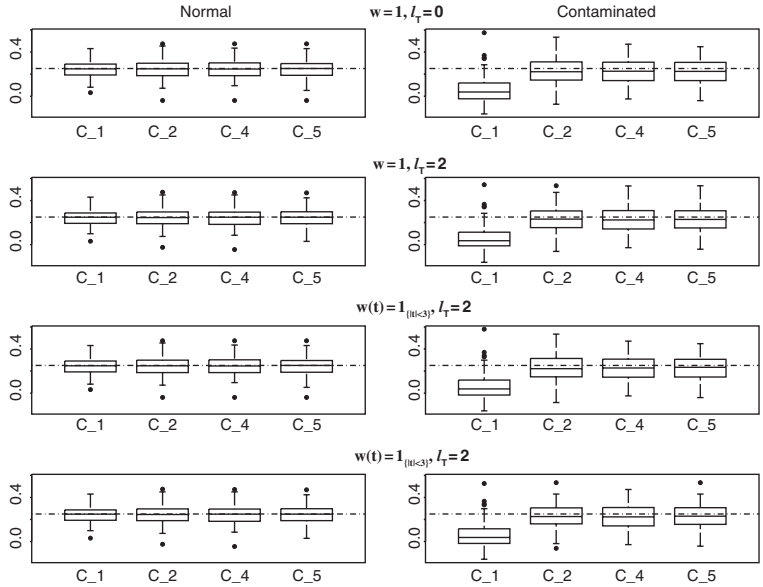

FIGURE 9. Boxplots of the data-driven estimates of β_0. The dashed line (– · – · –) corresponds to the true value.

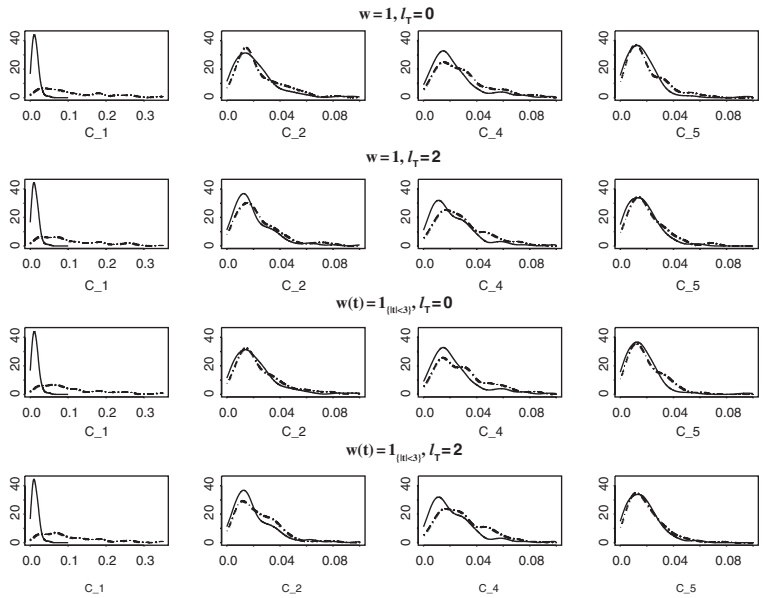

FIGURE 10. Density estimator of $M(\hat{g}, g)$. The solid lines correspond to normal errors, while the dashed ones (– · – · –) to the contaminated samples.

produces smaller data-driven bandwidths in this situation. Again, the best performance is obtained by the criterion based on the τ-scale estimator.

Regarding the behaviour of the estimators of the function g, its performance was evaluated computing at each replication

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

Figure 10 shows the estimates of the density of $M(\hat{g}, g)$. A density kernel estimator with bandwidth 0.02 was computed in all cases, apart from the classical estimators under contamination, where because of a different range of values, we took 0.05. Again, the data-driven estimator based on the τ-scale criterion shows its advantage over the other cross-validation rules.

6. UNIFORM ASYMPTOTIC DISTRIBUTION

In most practical situations, data-driven estimators of β_0 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 eqn (3). In the classical setting, the optimal bandwidth, which minimizes $D_1(h)$, has order $T^{-1/5}$. If we denote by $\hat{h}_1 = \text{argmin}_{h \in \mathcal{H}_T} C_1(h)$, where $\mathcal{H}_T = [aT^{-1/5-c},\ bT^{-1/5+c}]$ with $0 < a < b < \infty$ and $0 < c < 1/20$, it has been shown (see, for instance,

Härdle *et al.*, 2000) that the bandwidth minimizing $C_1(h)$ is asymptotically optimal, in the sense that

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

This property suggests that results regarding the asymptotic behaviour of the estimator $\hat{\beta}(\hat{h})$ are needed, where $\hat{h}$ denotes a bandwidth selector and $\hat{\beta}(h)$ is the estimator of the parameter β_0 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 $\hat{h}_j = \text{argmin}_{h\in\mathcal{H}_T} C_j(h)$, $2 \le j \le 5$, then

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

That's why, in this section, we will focus our attention to derive results regarding the asymptotic distribution of $T^{-1/2}(\hat{\beta}(h) - \beta_0)$, uniformly over $h \in \mathcal{H}_T$, which will imply that, for $2 \le j \le 5$, the data-driven estimators $\hat{\beta}(\hat{h}_j)$ of β_0 will be asymptotically normally distributed.

THEOREM 2. *Let $\{y_t, j \ge 3\}$ be a stationary α-mixing process satisfying eqn* (2) *with ϵ_t having a symmetric distribution and such that ϵ_t is independent of $\{y_{t-j}, j \ge 1\}$. Moreover, assume that the mixing coefficients are geometric. Let $\mathcal{H}_T = [aT^{-(1/5)-c}, bT^{-(1/5)+c}]$ with $0 < a < b < \infty$ and $0 < c < 1/20$. 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 $\phi_2(y) = \beta_0\phi_1(y) + g(y)$. Consider robust estimators of $\phi_j(y)$, $\hat{\phi}_j(y,h)$, based on the kernel weights $w_{tT}(y)$ defined in eqn* (3), *which admit a linear expansion $\hat{\phi}_j(y,h) - \phi_j(y) = \hat{\mathcal{L}}_j(y,h) + \hat{\mathcal{R}}_j(y,h)$, where*

$$\hat{\mathcal{L}}_j(y,h) = \sum_{t=3}^{T} w_{tT}(y) v_j\left(y_{t-2+j}, y\right), \tag{15}$$

with v_j bounded functions, such that $E(v_j(y_{t-2+j}, y)|y_{t-2} = y) = 0$ almost everywhere. Moreover, assume that for $j = 1, 2$ and for the compact set $\mathcal{K}$ defined in N6

$$T^{1/4} \sup_{h\in\mathcal{H}_T} \sup_{y\in\mathcal{K}} |\hat{\mathcal{L}}_j(y,h)| \xrightarrow{p} 0, \tag{16}$$

$$T^{1/2} \sup_{h\in\mathcal{H}_T} \sup_{y\in\mathcal{K}} |\hat{R}_j(y,h)| \xrightarrow{p} 0, \tag{17}$$

$$T^{-1/2} \sup_{h\in\mathcal{H}_T} \left| \sum_{t=3}^{T} \hat{\mathcal{L}}_j(y_{t-2},h)\vartheta_1(\epsilon_t)\vartheta_2(z_t)w_3(y_{t-2}) \right| \xrightarrow{p} 0 \tag{18}$$

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

$$T^{1/2}\left(\hat{\beta}(h) - \beta_0\right) \xrightarrow{D} N\left(\mathbf{0}, \sigma^2_{\psi_1,w_2,w_3}\right),$$

where $\sigma^2_{\psi_1,w_2,w_3}$ *is defined in Theorem* 1.

REMARK 3. Note that eqns (16) and (17) entail that

$$T^{1/4} \sup_{h\in\mathcal{H}_T} \sup_{y\in[0,1]} |\hat{\phi}_j(y,h) - \phi_j(y)| \xrightarrow{p} 0.$$

As discussed in Remark 2, using Taylor's expansion, it is easy to see that an M-estimator can be written $\hat{\phi}_j(y) = \phi_j(y) + \hat{\mathcal{L}}_j(y) + \hat{\mathcal{R}}_j(y)$, where the remainder term satisfies eqn (17), since M-estimators satisfy

$$T^{1/4} \sup_{h\in\mathcal{H}_T} \sup_{y\in\mathcal{K}} |\hat{\phi}_j(y) - \phi_j(y)| \xrightarrow{p} 0,$$

when ϕ_j are continuously differentiable functions. This last result and eqn (16) hold if the kernel is of bounded variation and can be derived using arguments similar to those considered in Boente and Fraiman (1991a) and a bound for the covering number of the family $h^{-1}K(\cdot/h)$. Conditions to guarantee eqn (18) can be found in Lemma 6.6.7 in Härdle *et al.* (2000).

APPENDIX

From now on, C_χ will denote the Lipschitz constant for a Lipschitz function χ. In Lemma A1, we get a consistent sequence of estimators of the matrix A given in N2.

LEMMA A1. *Let* $\{y_t\}$, $t \geq 3$ *be a stationary and ergodic process satisfying eqn* (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* $\tilde{\beta}$ *is a sequence of estimators such that* $\tilde{\beta} \xrightarrow{p} \beta_0$. *Let* $\hat{\phi}_j(y), j = 1, 2$ *be robust estimators of* $\phi_j(y)$ *such that*

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

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

$$A_T = \frac{1}{T-2}\sum_{t=3}^{T} \psi_1'\left(\frac{\hat{r}_t - z_t\tilde{\beta}}{\sigma_0}\right) w_2(\hat{z}_t)\hat{z}_t^2 w_3(y_{t-2}).$$

PROOF. Denote ξ_t intermediate points between $r_t - z_t\tilde{\beta}$ and $\hat{r}_t - \hat{z}_t\tilde{\beta}$ and $\hat{\eta}_j(y) = \hat{\phi}_j(y) - \phi_j(y)$ for $j = 1, 2$. A first-order Taylor expansion and some algebra lead us to $A_T = A_T^1 + A_T^2 + A_T^3 + A_T^4$, where

$$A_T^1 = \frac{1}{T-2}\sum_{t=3}^{T}\psi_1'\left(\frac{r_t - z_t\tilde{\beta}}{\sigma_0}\right)w_2(z_t)z_t^2 w_3(y_{t-2})$$

$$A_T^2 = -\frac{1}{T-2}\sum_{t=3}^{T}\psi_1'\left(\frac{\hat{r}_t - \hat{z}_t\tilde{\beta}}{\sigma_0}\right)w_2(\hat{z}_t)[\hat{\eta}_1(y_{t-2})z_t + \hat{z}_t\hat{\eta}_1(y_{t-2})]w_3(y_{t-2})$$

$$A_T^3 = -\frac{1}{T-2}\sum_{t=3}^{T}\psi_1''\left(\frac{\xi_t}{\sigma_0}\right)\left(\frac{\hat{\eta}_2(y_{t-2}) - \hat{\eta}_1(y_{t-2})\tilde{\beta}}{\sigma_0}\right)w_2(z_t)z_t^2 w_3(y_{t-2})$$

$$A_T^4 = \frac{1}{T-2}\sum_{t=3}^{T}\psi_1'\left(\frac{\hat{r}_t - \hat{z}_t\tilde{\beta}}{\sigma_0}\right)[w_2(\hat{z}_t) - w_2(z_t)]z_t^2 w_3(y_{t-2}).$$

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

From N3, it is easy to see that

$$z_t^2|w_2(\hat{z}_t) - w_2(z_t)| \le |\hat{\eta}_1(y_{t-2})|\Big(\|\psi_2\|_\infty + |\hat{\eta}_1(y_{t-2})|\big(\|w_2\|_\infty + \|\psi_2'\|_\infty\big) + \|\lambda_2\|_\infty\Big).$$

Now, the result follows from N2, the consistency of $\tilde{\beta}$, the Ergodic Theorem and the fact that

$$\max_{1\le j\le 2}\sup_{y\in\mathcal{K}}|\hat{\eta}_j(y)| \xrightarrow{p} 0 \quad \text{and} \quad \|w_3\|_\infty \le 1,$$

since

$$|A_T^2| \le \|\psi_1'\|_\infty \max_{1\le j\le 2}\sup_{y\in\mathcal{K}}|\hat{\eta}_j(y)|\left(2\|\psi_2\|_\infty + \|w_2\|_\infty \max_{1\le j\le 2}\sup_{y\in\mathcal{K}}|\hat{\eta}_j(y)|\right)$$

$$|A_T^3| \le \|\psi_1''\|_\infty \max_{1\le j\le 2}\sup_{y\in\mathcal{K}}|\hat{\eta}_j(y)|\left(\frac{1+|\tilde{\beta}|}{\sigma_0}\right)\frac{1}{T-2}\sum_{t=3}^{T}w_2(z_t)z_t^2$$

$$|A_T^4| \le \|\psi_1'\|_\infty \sup_{y\in\mathcal{K}}|\hat{\eta}_1(y)|\left(\|\psi_2\|_\infty + \sup_{y\in\mathcal{K}}|\hat{\eta}_1(y)|\big(\|w_2\|_\infty + \|\psi_2'\|_\infty\big) + \|\lambda_2\|_\infty\right). \quad \square$$

PROOF OF THEOREM 1. Denote

$$L_T(\beta) = \frac{\sigma_0}{T-2}\sum_{t=3}^{T}\psi_1\left(\frac{r_t - z_t\beta}{\sigma_0}\right)w_2(z_t)z_t w_3(y_{t-2})$$

$$\hat{L}_T(\beta) = \frac{\sigma_0}{T-2}\sum_{t=3}^{T}\psi_1\left(\frac{\hat{r}_t - \hat{z}_t\beta}{\sigma_0}\right)w_2(\hat{z}_t)\hat{z}_t w_3(y_{t-2}).$$

Using first-order Taylor's expansion around $\hat{\beta}$, we get

$$\hat{L}_T(\beta_0) = \frac{\sigma_0}{T-2}\sum_{t=3}^{T}\psi_1\left(\frac{\hat{r}_t - \hat{z}_t\hat{\beta}}{\sigma_0}\right)w_2(\hat{z}_t)\hat{z}_t w_3(y_{t-2}) + (\hat{\beta}-\beta_0)\frac{1}{T-2}\sum_{t=3}^{T}\psi_1'\left(\frac{\hat{r}_t - \hat{z}_t\tilde{\beta}}{\sigma_0}\right)w_2(\hat{z}_t)\hat{z}_t^2 w_3(y_{t-2}),$$

with $\tilde{\beta}$ an intermediate point between $\hat{\beta}$ and β_0. This implies that

$$\hat{L}_T(\beta_0) = 0 + (\hat{\beta}-\beta_0)\frac{1}{T-2}\sum_{t=3}^{T}\psi_1'\left(\frac{\hat{r}_t - \hat{z}_t\tilde{\beta}}{\sigma_0}\right)w_2(\hat{z}_t)\hat{z}_t^2 w_3(y_{t-2})$$

and so, we get that $(\hat{\beta}-\beta_0) = A_T^{-1}\hat{L}_T(\beta_0)$ with A_T defined in Lemma A1. From the consistency of $\hat{\beta}$, Lemma A1 implies that $A_T \xrightarrow{p} A$ and therefore, from N2 it will be enough to show that

(a) $T^{1/2}L_T(\beta_0) \xrightarrow{D} N(\mathbf{0}, \sigma^2)$ with $\sigma^2 = \sigma_0^2 E\left(\psi_1^2\left(\frac{\epsilon_t}{\sigma_0}\right)\right)E\left(w_2^2(z_t)z_t^2 w_3^2(y_{t-2})\right)$,

(b) $T^{1/2}\left[\hat{L}_T(\beta_0) - L_T(\beta_0)\right] \xrightarrow{p} 0$.

(a) Follows inmediately from the Central Limit Theorem for geometrically α-mixing process, since $r_t - z_t\beta_0 = \epsilon_t$ is independent of $\{y_s : s \le t\}$ (see, for instance, Thm 1.7 in Bosq, 1996).

(b) Denote ξ_t intermediate points between $r_t - z_t\beta_0$ and $\hat{r}_t - \hat{z}_t\beta_0$ and $\hat{\eta}_j(y) = \hat{\phi}_j(y) - \phi_j(y)$ for $j = 1, 2$. Using second-order Taylor's expansion, we have that $\hat{L}_T(\beta_0) = L_T(\beta_0) + \hat{L}_{T,1} + \hat{L}_{T,2} + \hat{L}_{T,3} + \hat{L}_{T,4} + \hat{L}_{T,5}$, where

$$\hat{L}_{T,1} = \frac{1}{T-2}\sum_{t=3}^{T}\psi_1'\left(\frac{r_t - z_t\beta_0}{\sigma_0}\right)[\hat{\eta}_1(y_{t-2})\beta_0 - \hat{\eta}_2(y_{t-2})]w_2(z_t)z_t w_3(y_{t-2})$$

$$\hat{L}_{T,2} = \frac{\sigma_0}{T-2}\sum_{t=3}^{T}\psi_1\left(\frac{r_t - z_t\beta_0}{\sigma_0}\right)[w_2(\hat{z}_t)\hat{z}_t - w_2(z_t)z_t]w_3(y_{t-2})$$

$$\hat{L}_{T,3} = \frac{\sigma_0}{T-2}\sum_{t=3}^{T}\left[\psi_1\left(\frac{\hat{r}_t - \hat{z}_t\beta_0}{\sigma_0}\right) - \psi_1\left(\frac{r_t - z_t\beta_0}{\sigma_0}\right)\right]w_2(\hat{z}_t)(\hat{z}_t - z_t)w_3(y_{t-2})$$

$$\hat{L}_{T,4} = \frac{1}{2\sigma_0}\frac{1}{T-2}\sum_{t=3}^{T}\psi_1''\left(\frac{\xi_t}{\sigma_0}\right)[\hat{\eta}_1(y_{t-2})\beta_0 - \hat{\eta}_2(y_{t-2})]^2 w_2(\hat{z}_t)z_t w_3(y_{t-2})$$

$$\hat{L}_{T,5} = \frac{1}{T-2}\sum_{t=3}^{T}\psi_1'\left(\frac{r_t - z_t\beta_0}{\sigma_0}\right)[\hat{\eta}_1(y_{t-2})\beta_0 - \hat{\eta}_2(y_{t-2})][w_2(\hat{z}_t) - w_2(z_t)]z_t w_3(y_{t-2}).$$

First, note that when (b) holds, eqns (12) and (13) entail that eqn (10) holds. Using that $\|w_3\|_\infty \le 1$ and that N3 entails $|w_2(\hat{z}_t) - w_2(z_t)| \le C|\hat{\eta}_1(y_{t-2})|/|z_t|$, with $C = \|w_2\|_\infty + C_{\psi_2}$, we get

$$T^{1/2}\|\hat{L}_{T,3}\| \le p\|w_2\|_\infty\|\psi_1'\|_\infty T^{1/2}\left[\sup_{y\in\mathcal{K}}|\hat{\eta}_1(y)|\right]^2(1+|\beta_0|)$$

$$T^{1/2}\|\hat{L}_{T,4}\| \le \frac{1}{2}\frac{1}{\sigma_0}\|\psi_1''\|_\infty T^{1/2}\left[\sup_{y\in\mathcal{K}}\max_{1\le j\le 2}|\hat{\eta}_j(y)|\right]^2(1+|\beta_0|)^2\left(\|\psi_2\|_\infty + \|w_2\|_\infty\sup_{y\in\mathcal{K}}|\hat{\eta}_1(y)|\right)$$

$$T^{1/2}\|\hat{L}_{T,5}\| \le pC\|\psi_1'\|_\infty(1+|\beta_0|)T^{1/2}\left[\max_{1\le j\le 2}\sup_{y\in\mathcal{K}}|\hat{\eta}_j(y)|\right]^2,$$

which together with eqn (10), implies that, for $3 \le j \le 5$, $T^{1/2}\|\hat{L}_{T,j}\| \xrightarrow{p} 0$. Note that $\hat{L}_{T,2} = \hat{L}_{T,2}^{(1)} + \hat{L}_{T,2}^{(2)}$ with

$$\hat{L}_{T,2}^{(1)} = \frac{\sigma_0}{T-2}\sum_{t=3}^{T}\psi_1\left(\frac{\epsilon_t}{\sigma_0}\right)\psi_2'(z_t)\hat{\eta}_1(y_{t-2})w_3(y_{t-2})$$

$$\hat{L}_{T,2}^{(2)} = \frac{\sigma_0}{T-2}\sum_{t=3}^{T}\psi_1\left(\frac{\epsilon_t}{\sigma_0}\right)\left[\psi_2'(\xi_t) - \psi_2'(z_t)\right]\hat{\eta}_1(y_{t-2})w_3(y_{t-2}),$$

where ξ_t denotes intermediate points between $\hat{z}_t$ and z_t. Using that ψ_2' is Lispchitz of order 1, we get $T^{1/2}|\hat{L}_{T,2}^{(2)}| \le C_{\psi_2'}\|\psi_1\|_\infty T^{1/2}\sup_{y\in\mathcal{K}}|\hat{\eta}_1(y)|^2$. Using that eqn (10) holds, under (a) or (b), we get that $T^{1/2}|\hat{L}_{T,2}^{(2)}\| \xrightarrow{p} 0$.

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

$$\hat{R}_{T,j} = T^{-1/2}\sum_{t=3}^{T}\psi_1'\left(\frac{\epsilon_t}{\sigma_0}\right)\hat{\eta}_j(y_{t-2})w_2(z_t)z_t w_3(y_{t-2}) \xrightarrow{p} 0, \quad j = 1,2 \tag{A.1}$$

$$\hat{R}_{T,3} = T^{-1/2}\sum_{t=3}^{T}\psi_1\left(\frac{\epsilon_t}{\sigma_0}\right)\psi_2'(z_t)\hat{\eta}_1(y_{t-2})w_3(y_{t-2}) \xrightarrow{p} 0. \tag{A.2}$$

We begin by proving the desired result when (a) holds.
Note that proving eqn (A.1) is equivalent to show that, for $j = 1, 2$

$$\hat{R}_{T,j,1} = T^{-1/2}\sum_{t=3}^{T}\left[\psi_1'\left(\frac{\epsilon_t}{\sigma_0}\right) - E\left(\psi_1'\left(\frac{\epsilon}{\sigma_0}\right)\right)\right]\hat{\eta}_j(y_{t-2})w_2(z_t)z_t w_3(y_{t-2}) \xrightarrow{p} 0, \tag{A.3}$$

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

For any function υ with domain on $\mathcal{K}$, we define

$$\begin{aligned}
J_{T,1}(\upsilon) &= T^{-1/2}\sum_{t=3}^{T}\left[\psi_1'\left(\frac{\epsilon_t}{\sigma_0}\right) - E\left(\psi_1'\left(\frac{\epsilon}{\sigma_0}\right)\right)\right]\upsilon(y_{t-2})\psi_2(z_t)w_3(y_{t-2}) \\
&= T^{-1/2}\sum_{t=3}^{T}X_{t,1}(\upsilon) \\
J_{T,2}(\upsilon) &= T^{-1/2}\sum_{t=3}^{T}\psi_1\left(\frac{\epsilon_t}{\sigma_0}\right)\psi_2'(z_t)\upsilon(y_{t-2})w_3(y_{t-2}) = T^{-1/2}\sum_{t=3}^{T}X_{t,2}(\upsilon) \\
J_{T,3}(\upsilon) &= T^{-1/2}\sum_{t=3}^{T}\upsilon(y_{t-2})\psi_2(z_t)w_3(y_{t-2}) = T^{-\frac{1}{2}}\sum_{t=3}^{T}X_{t,3}(\upsilon).
\end{aligned}$$

Let $\mathcal{L} = \{\upsilon \in \mathcal{C}^1(\mathcal{K}) : \|\upsilon\|_{1,\infty} = \|\upsilon\|_\infty + \|\upsilon'\|_\infty \le 1\}$. Assumption N1 and the fact that ϵ_t has a symmetric distribution entail that $E(X_{t,2}(\upsilon)) = 0$. Besides, N4 entails that $E(X_{t,3}(\upsilon)) = 0$ for any bounded function υ. Using that the process $\{y_t : t \ge 3\}$ is geometrically α-mixing, and since $\epsilon_t = y_t - \beta_0 y_{t-1} - g(y_{t-2})$ and $z_t = y_{t-1} - \phi_1(y_{t-2})$, $\{X_{t,\ell}(\upsilon) : t \ge 3\}$ is still a geometrically α-mixing process, for any bounded function υ. Therefore, Thm 1.5 in Bosq (1996) and Thm 1 in Doukhan *et al.* (1994) entail that the finite-dimensional distributions of $\{J_{T,\ell}(\upsilon) : \upsilon \in \mathcal{L}\}$ converge to the finite-dimensional distributions of an eventually degenerate Gaussian process $\{J_\ell(\upsilon) : \upsilon \in \mathcal{L}\}$. Besides, from Jensen's inequality, Thm 1.2 in Rio (1993), and the fact that the mixing coefficients are

geometric, we get that for some finite constant C depending on the mixing coefficients, and any $p > 2$

$$E\big|J_{T,3}(\upsilon)\big| \leq \Big[EJ_{T,3}^2(\upsilon)\Big]^{1/2} \leq C[E|\upsilon(y_{t-2})\psi_2(z_t)w_3(y_{t-2})|^p]^{1/p} \leq C\|\psi_2\|_\infty [E|\upsilon(y_{t-2})|^p]^{1/p}.$$

Similarly, we get that

$$\begin{aligned} E|J_{T,1}(\upsilon)| &\leq 2C\|\psi_1'\|_\infty \|\psi_2\|_\infty [E\upsilon(y_{t-2})|^p]^{1/p} \\ E|J_{T,2}(\upsilon)| &\leq C\|\psi_1\|_\infty \|\psi_2'\|_\infty [E|\upsilon(y_{t-2})|^p]^{1/p}. \end{aligned}$$

Therefore, Thm 2.1 in Arcones (1996) implies the weak convergence of $\{J_{T,\ell}(\upsilon) : \upsilon \in \mathcal{L}\}$ to a Gaussian process $\{J_\ell(\upsilon) : \upsilon \in \mathcal{L}\}$. Hence, the process $\{J_{T,\ell}(\upsilon) : \upsilon \in \mathcal{L}\}$ is stochastically equicontinuous. As noted by Andrews (1994), the stochastic equicontinuity of the process and the fact that from eqn (10)

$$\sup_{y\in\mathcal{K}} |\hat{\eta}_j(t)| = \sup_{y\in\mathcal{K}} |\hat{\phi}_j(y) - \phi_j(y)| \xrightarrow{p} 0,$$

we obtain that for $\ell = 1, 2, 3$, $J_{T,\ell}(\hat{\eta}_j) \xrightarrow{p} 0$ and so eqns (A.2)–(A.4) hold, concluding the proof when (a) holds.

Assume now that (b) holds.

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

$$T^{1/2}\hat{R}_{T,j} = \sum_{t=3}^{T} \psi_1'\left(\frac{\epsilon_t}{\sigma_0}\right)\hat{\mathcal{L}}_j(y_{t-2})\psi_2(z_t)w_3(y_{t-2}) + \sum_{t=3}^{T} \psi_1'\left(\frac{\epsilon_t}{\sigma_0}\right)\hat{R}_j(y_{t-2})\psi_2(z_t)w_3(y_{t-2}),$$

which implies that

$$|\hat{R}_{T,j}| \leq \left|\frac{1}{\sqrt{T}}\sum_{t=3}^{T} \psi_1'\left(\frac{\epsilon_t}{\sigma_0}\right)\hat{\mathcal{L}}_j(y_{t-2})\psi_2(z_t)w_3(y_{t-2})\right| + \|\psi_1'\|_\infty \|\psi_2\|_\infty T^{\frac{1}{2}} \sup_{y\in\mathcal{K}} |\hat{R}_j(y)|.$$

Now eqn (A.1) follows from N4, eqn (13) and (14) with $\vartheta_1(t) = \psi_1'(t/\sigma_0)$ and $\vartheta_2 \equiv \psi_2$.

Finally, eqn (A.2) follows using similar arguments to those used to deal with eqn (A.1), applying eqn (14) to $\vartheta_1(t) = \psi_1(t/\sigma_0)$ and $\vartheta_2 \equiv \psi_2'$ and eqn (13) which concludes the proof. □

Lemma A2 states a result analogous to that given in Lemma A.1, uniformly on the bandwidth parameter.

Lemma A2. *Let $\{y_t\}$, $t \geq 3$ be a stationary and ergodic process satisfying eqn (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 $\hat{\phi}_j(y,h)$, $j = 1, 2$, be robust estimators of $\phi_j(y)$, computed with the kernel weights defined in eqn (3), such that*

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

uniformly for $h \in \mathcal{H}_T$ and assume that $\tilde{\beta}(h) \xrightarrow{p} \beta_0$ 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 A1.

PROOF. As in Lemma A1, 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

$$\mathcal{A}_T(\beta) = \frac{1}{T-2}\sum_{t=3}^{T} \psi_1'\left(\frac{r_t - z_t\beta}{\sigma_0}\right) w_2(z_t) z_t^2 \; w_3(y_{t-2})$$

and using arguments analogous to those considered in Lemma 1 in Bianco and Boente (2001), we get that, for any $\delta > 0$,

$$\sup_{|\beta-\beta_0|<\delta} |\mathcal{A}_T(\beta) - E(\mathcal{A}_T(\beta))| \xrightarrow{p} 0,$$

which together with the uniform convergence of $\tilde{\beta}(h)$ to β_0 entails the desired result. □

REMARK A1. It is worth noticing that the conclusion of Lemmas A1 and A2 still holds without requiring N6. In that case, we need to assume that the uniform convergence of the estimates of $\phi_j(y)$ holds for any compact set $\mathcal{K} \subset \mathbb{R}$.

PROOF OF THEOREM 2. Follows using arguments analogous as those considered in the proof of Theorem 1. First notice that

$$T^{1/2} L_T(\beta_0) \xrightarrow{D} N\left(\mathbf{0}, \sigma^2\right)$$

holds uniformly for $h \in \mathcal{H}_T$, since $L_T(\beta_0)$ does not depend on the smoothing parameter. Therefore, it remains to show that

$$T^{1/2}\left[\hat{L}_T(\beta_0) - L_T(\beta_0)\right] \xrightarrow{p} 0$$

uniformly in $\mathcal{H}_T$. Denote $\hat{\eta}_j(y,h) = \hat{\phi}_j(y,h) - \phi_j(y)$.

As in Theorem 1, using a second-order Taylor's expansion, we have that

$$\hat{L}_T(\beta_0) = L_T(\beta_0) + \hat{L}_{T,1} + \hat{L}_{T,2}^{(1)} + \hat{L}_{T,2}^{(2)} + \hat{L}_{T,3} + \hat{L}_{T,4} + \hat{L}_{T,5},$$

where $\hat{L}_{T,j}$ and $\hat{L}_{T,2}^{(j)}$ are defined in Theorem 1. Moreover, with the bounds obtained therein, since

$$T^{1/4} \sup_{h\in\mathcal{H}_T} \sup_{y\in[0,1]} |\hat{\eta}_j(y,h)| \xrightarrow{p} 0,$$

it is easy to see that, for $3 \le j \le 5$,

$$T^{1/2} \sup_{h\in\mathcal{H}_T} |\hat{L}_{T,j}| \xrightarrow{p} 0 \quad \text{and that} \quad T^{1/2} \sup_{h\in\mathcal{H}_T} |\hat{L}_{T,2}^{(2)}| \xrightarrow{p} 0.$$

The proof will be concluded if we show that

$$T^{1/2}\hat{L}_{T,1} \xrightarrow{p} 0 \quad \text{and} \quad T^{1/2}\hat{L}_{T,2}^{(1)} \xrightarrow{p} 0$$

uniformly for $h \in \mathcal{H}_T$, that is, for $j = 1, 2$

$$\sup_{h\in\mathcal{H}_T}|\hat{R}_{T,j}| = \sup_{h\in\mathcal{H}_T} T^{-1/2}\left|\sum_{t=3}^{T}\psi_1'\left(\frac{\epsilon_t}{\sigma_0}\right)\hat{\eta}_j(y_{t-2},h)w_2(z_t)z_t w_3(y_{t-2})\right| \xrightarrow{p} 0 \qquad \text{(A.5)}$$

$$\sup_{h\in\mathcal{H}_T}|\hat{R}_{T,3}| = \sup_{h\in\mathcal{H}_T} T^{-1/2}\left|\sum_{t=3}^{T}\psi_1\left(\frac{\epsilon_t}{\sigma_0}\right)\psi_2'(z_t)\hat{\eta}_1(y_{t-2},h)w_3(y_{t-2})\right| \xrightarrow{p} 0. \qquad \text{(A.6)}$$

Using the linear expansion for $\hat{\eta}_j(y,h)$ we get eqn (A.5) from N4, together with eqns (17) and (18) with $\vartheta_1(t) = \psi_1'(t/\sigma_0)$ and $\vartheta_2 \equiv \psi_2$, as in the proof of Theorem 1 under (b). On the other hand, eqn (A.6) follows using similar arguments, applying eqns (17) and (18) to $\vartheta_1(t) = \psi_1(t/\sigma_0)$ and $\vartheta_2 \equiv \psi_2'$. □

ACKNOWLEDGMENTS

We thank the Associate Editor and Referee for detailed and constructive comments on an earlier version of the article that substantially improved it. This research was partially supported by Grants 13900-6 from the Fundación Antorchas, X094 from University of Buenos Aires, PID 5505 from CONICET and PAV 120 and PICT 21407 from the Agencia Nacional de Promoción Científica y Tecnológica, Argentina.

NOTE

Corresponding author: Ana Bianco, Instituto de Cálculo, Ciudad Universitaria, Pabellón 2, Buenos Aires. C1428 EHA, Argentina. E-mail: abianco@dm.uba.ar

REFERENCES

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

ANDREWS, D. and POLLARD, D. (1994) An introduction to functional central limit theorems for dependent stochastic processes. *International Statistical Review* 62, 119–32.

ANSLEY, C. and WECKER, W. (1983) Extension and examples of the signal extraction approach to regression. In *Applied Time Series Analysis of Economic Data* (ed. ARNOLD ZELLNER). Economic Research Report ER-5. Washington, DC: U.S. Bureau of the Census, 181–92.

ARCONES, M. (1996) Weak convergence of stochastic processes indexed by smooth functions. *Stochastic Processes and their Application* 62, 11–138.

BIANCO, A. and BOENTE, G. (2001) On the asymptotic behaviour 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, 249–87.

BIANCO, A. and BOENTE, G. (2004) Robust estimators in semi-parametric partly linear regression models. *Journal of Statistical Planning and Inference* 122, 229–52.

BIANCO, A., GARCIA BEN, M., MARTINEZ, E. and YOHAI, V. (1996) Robust procedures for regression models with ARIMA errors. In *COMPSTAT 96* (ed. ALBERT PRAT). *Proceedings in Computational Statistics*. Heidelberg: Physica-Verlag, 27–38.

BILLINGSLEY, P. (1968) *Convergence of Probability Measures*. New York: Wiley.
BOENTE, G. and FRAIMAN, R. (1989) Robust nonparametric regression estimation. *Journal Multivariate Analysis* 29, 180–98.
BOENTE, G. and FRAIMAN, R. (1991a) Strong uniform convergence rates for some robust equivariant nonparametric regression estimates for mixing processes. *International Statistical Review* 59, 355–72.
BOENTE, G. and FRAIMAN, R. (1991b) A functional approach to robust nonparametric regression. In *Directions in Robust Statistics and Diagnostics* (eds W. STAHEL and S. WEISBERG). *Proceedings of the IMA Institute*, USA 33 (Part I), 35–46.
BOENTE, G. and RODRIGUEZ, D. (2006) Robust estimators of high order derivatives of regression functions. *Statistics and Probability Letters* 76, 1335–44.
BOENTE, G., FRAIMAN, R. and MELOCHE, J. (1997) Robust plug-in bandwidth estimators in nonparametric regression. *Journal of Statistical Planning and Inference* 57, 109–42.
BOSQ, D. (1996) *Nonparametric Statistics for Stochastic Processes. Estimation and Prediction*. New York: Springer-Verlag.
BRILLINGER, D. R. (1986) Discussion of 'Influence functionals for time series' by Martin R. D. and Yohai, V. J. *Annals of Statistical* 14, 819–22.
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. *Journal of the Royal Statistical Society Series* A 140, 411–31.
CANTONI, E. and RONCHETTI, E. (2001) Resistant selection of the smoothing parameter for smoothing splines. *Statistics and Computing* 11, 141–6.
CHEN, H. (1988) Convergence rates for parametric components in a partly linear model. *Annals of Statistics* 16, 136–46.
CHEN, H. and CHEN, K. (1991) Selection of the splined variables and convergence rates in a partial spline model. *Canadian Journal of Statistics* 19, 323–39.
CHEN, H. and SHIAU, J. (1991) A two-stage spline smoothing method for partially linear models. *Journal of Statistical Planning and Inference* 25, 187–201.
CHEN, H. and SHIAU, J. (1994) Data-driven efficient estimates for partially linear models. *Annals of Statistics* 22, 211–37.
CHU, C. K. and MARRON, S. (1991) Comparison of two bandwidth selectors with dependent errors. *Annals of Statistics* 19, 1906–18.
DOUKHAN, P. (1994) *Mixing Properties and Examples. Lecture Notes in Statistics*, 85. New York: Springer-Verlag.
DOUKHAN, P., MASSART, P. and RIO, E. (1994) The central limit theorem for strongly mixing processes. *Annales de l' Institut Henri Poincaré (B) Probabilités et Statistiques* 30, 63–82.
ENGLE, R., GRANGER, C., RICE, J. and WEISS, A. (1986) semi-parametric estimates of the relation between weather and electricity sales. *Journal of the American Statistical Association* 81, 310–20.
GAO, J. (1992) *A Large Sample Theory in semi-parametric Regression Models*. PhD Thesis, University of Science and Technology of China, Hefei, China.
GAO, J. (1995) Asymptotic theory for partly linear models. *Communications in Statistics, Theory & Methods* 24, 1985–2010.
GAO, J. (1998) Semi-parametric regression smoothing of nonlinear time series. *Scandinavian Journal of Statistics* 25, 521–39.
GAO, J. and LIANG, H. (1995) Asymptotic normality of pseudo-LS estimator for partly linear autoregression models. *Statistics and Probability Letters* 23, 27–34.
GAO, J. and SHI, P. (1997) M-type smoothing splines in nonparametric and semi-parametric regression models. *Statistica Sinica* 7, 1155–69.
GAO, J. and YEE, T. (2000) Adaptive estimation in partly linear autoregressive models. *Canadian Journal of Statistics* 28, 571–86.
GAO, J. and ZHAO, L. (1993) Adaptive estimation in partly linear regression models. *Science in China*, Ser. A 1, 14–27.
GREEN, P., JENNISON, C. and SEHEULT, A. (1985) Analysis of field experiments by least squares smoothing. *Journal of the Royal Statistical Society Series* 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. Berlin: Springer-Verlag.
HÄRDLE, W. (1990) *Applied Nonparametric Regression*. Cambridege: Cambridge University Press.
HÄRDLE, W. and GASSER, T. (1985) On robust kernel estimation of derivatives of regression functions. *Scandinanian Journal Statistics* 12, 233–40.
HÄRDLE, W., LIANG, H. and GAO, J. (2000) *Partially Linear Models*. Heidelberg: Physica-Verlag.

HART, J. (1996) Some automated methods of smoothing time-dependent data. *Nonparametric Statistics* 6, 115–42.
HART, P. and VIEU, P. (1990) Data-driven bandwidth choice for density estimation based on dependent data. *Annals of Statistics* 18, 873–90.
HE, X., ZHU, Z. and FUNG, W. (2002) Estimation in a semiparametric model for longitudinal data with unspecified dependence structure. *Biometrika* 89, 579–90.
HECKMAN, N. (1986) Spline smoothing in a partly linear model. *Journal of the Royal Statistical Society Series* B 48, 244–8.
IBRAGIMOV, I. and LINNIK, Y. (1971) *Independent and Stationary Sequences of Random Variables*. Groningen: Wolters-Noerdhoff.
LEUNG, D. H. Y., MARRIOTT, F. H. C. and WU, E. K. H. (1993) Bandwidth selection in robust smoothing. *Journal of Nonparametric Statistics* 2, 333–9.
LIANG, H. (1996) Asymptotically efficient estimators in a partly linear autoregressive model. *System Sciences and Mathematical Sciences*, 9, 164–70.
MARTIN, R. D. and YOHAI, V. J. (1986) Influence functionals for time series. *Annals of Statistics* 14, 781–818.
POLLARD, D. (1984) *Convergence of Stochastic Processes*. New York: Springer-Verlag.
RIO, E. (1993) Covariance inequalities for strongly mixing processes. *Annales de l' Institut Henri Poincaré (B) Probabilités et Statistiques* 29, 587–97.
ROBINSON, P. M. (1983) Nonparametric estimators for time series. *Journal of Time Series Analysis* 4, 185–206.
ROBINSON, P. M. (1984) Robust nonparametric autoregression. In *Robust and Nonlinear Time Series Analysis* (eds J. FRANKE, W. HÄRDLE and D. MARTIN). *Lecture Notes in Statistics*, 4. Berlin: Springer-Verlag, 185–206.
ROBINSON, P. (1988) Root-n-consistent semi-parametric regression. *Econometrica* 56, 931–54.
ROSENBLATT, M. (1956) A central limit theorem and a strong mixing condition. *Proceedings of the National Academy of Sciences USA* 42, 43–7.
ROUSSEEUW, P. and LEROY, A. (1987) *Robust Regression and Outlier Detection*. New York: Wiley.
SEVERINI, T. and STANISWALIS, J. (1994) Quasi-likelihood estimation in semi-parametric models. *Journal of the American Statistical Association* 89, 501–11.
SEVERINI, T. and WONG, W. (1992) Profile likelihood and conditionally parametric models. *Annals of Statistics* 20, 1768–802.
SPECKMAN, P. (1988) Kernel smoothing in partial linear models. *Journal of the Royal Statistical Socity, Series* B 50, 413–36.
TONG, H. (1977) Some comments on the Canadian lynx data (with discussion). *Journal of the Royal Statistical Society Series* A 140, 432–6.
WANG, F. and SCOTT, D. (1994) The L_1 method for robust nonparametric regression. *Journal of the American Statistical Association* 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, 203–20.
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. *Journal of the Royal Statistical Society Series* B 58, 481–93.
YOHAI, V. (1987) High breakdown point and high efficiency robust estimates for regression. *Annals of Statistics* 15, 642–56.
YOHAI, V. and ZAMAR, R. (1988) High breakdown estimates of regression by means of the minimization of an efficient scale. *Journal of American Statistical Association* 83, 406–13.
YU, B. (1994) Rates of convergence of empirical processes for stationary mixing sequences. *Annals of Probability* 22, 94–116.