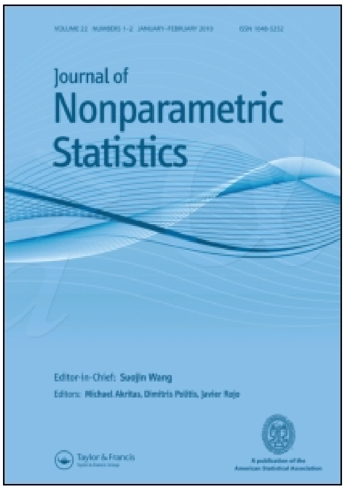

Journal of Nonparametric Statistics
Publication details, including instructions for authors and subscription information:
http://www.informaworld.com/smpp/title~content=t713645758

On a partly linear autoregressive model with moving average errors

Ana Bianco[a]; Graciela Boente[a]
[a] Facultad de Ciencias Exactas y Naturales, Universidad de Buenos Aires and CONICET, Ciudad Universitaria, Buenos Aires, Argentina

First published on: 26 January 2010

To cite this Article Bianco, Ana and Boente, Graciela(2010) 'On a partly linear autoregressive model with moving average errors', Journal of Nonparametric Statistics, 22: 6, 797 — 820, First published on: 26 January 2010 (iFirst)

To link to this Article: DOI: 10.1080/10485250903469744

URL: http://dx.doi.org/10.1080/10485250903469744

PLEASE SCROLL DOWN FOR ARTICLE

Journal of Nonparametric Statistics
Vol. 22, No. 6, August 2010, 797–820

On a partly linear autoregressive model with moving average errors

Ana Bianco and Graciela Boente*

Facultad de Ciencias Exactas y Naturales, Universidad de Buenos Aires and CONICET, Ciudad Universitaria, Pabellón 2, Buenos Aires, C1428EHA, Argentina

(*Received 13 June 2008; final version received 3 November 2009*)

In this paper, we generalise the partly linear autoregression model considered in the literature by including moving average errors when we want to allow a large dependence to the past observations. The strong ergodicity of the process is derived. A consistent procedure to estimate the parametric and nonparametric components is provided together with a test statistic that allows to check the presence of a moving average component in the model. Also, a Monte Carlo study is carried out to check the performance of the given proposals.

Keywords: ergodicity; Fisher-consistency; moving average errors; partly linear autoregression; smoothing techniques

MSC: Primary: 62F35; Secondary: 62H25

1. Introduction

When dealing with time series data, autoregressive models with moving average errors (ARMA models) have been extensively used in applications. They correspond to linear autoregressive models where the errors are described by a moving average process. More precisely, an ARMA (p, q) model is a stationary process $\{y_t : t \geq 1\}$ verifying

$$y_t = \sum_{j=1}^{p} \varphi_j y_{t-j} + \varepsilon_t \ , \tag{1}$$

where $\varepsilon_t = u_t - \sum_{j=1}^{q} \theta_j u_{t-j}$ with u_t independent and identically distributed (i.i.d.) random variables and u_t is independent of $\{y_{t-j}, j \geq 1\}$ with $Eu_t = 0$, $Eu_t^2 < \infty$.

It is well known that, when there is large dependence to the past observations, ARMA models have several advantages with respect to autoregressive models. However, the assumption of a linear autoregression function is quite restrictive. As pointed by Bosq (1996), a nonparametric predictor is ‘in general more efficient and more flexible than the predictor based on Box and

*Corresponding author. Email: gboente@dm.uba.ar

ISSN 1048-5252 print/ISSN 1029-0311 online

DOI: 10.1080/10485250903469744
http://www.informaworld.com

Jenkins method and nearly equivalent if the underlying model is truly linear', see also Carbon and Delecroix (1993) for a comparative study on 17 series. Nevertheless, the nonparametric autoregression model $y_t = m(\mathbf{X}_t) + u_t$, where $\mathbf{X}_t = (y_{t-1}, \dots, y_{t-r})^{\mathrm{T}}$, faces the problem known as the 'curse of dimensionality'. In order to solve the problem of empty neighbourhoods, an approach can be to introduce moving average errors, which reduce the dependence to the past in $\mathbf{X}_t$ obtaining, thus, a smaller dimension r. This approach was followed by Boente and Fraiman (2002) who introduced nonparametric ARMA models that allow the autoregressive part of the model to be nonparametric, while the moving average part remains linear.

As noted by Gao and Yee (2000), another disadvantage of the fully nonparametric autoregressive model is that it neglects a possible linear relationship between y_t and any lag y_{t-k} . To solve the 'curse of dimensionality', following a semiparametric approach, several authors have introduced the 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

$$y_t = \sum_{i=1}^{p_1} \beta_{o,i} y_{t-i} + \sum_{j=1}^{p_2} g_{o,j}(y_{t-p_1-j}) + u_t \ , \tag{2}$$

where $g_{o,j} : \mathbb{R} \to \mathbb{R}$ are smooth functions and u_t are i.i.d. random variables independent of $\{y_{t-j}, j \geq 1\}$, $Eu_t = 0$ and $Eu_t^2 < \infty$. However, these models do not take into account a large dependence on the past unless p_1 and p_2 are large. To reduce the order of the process, we can allow a dependence structure in the errors as in Equation (1). Combining models (1) and (2), one can consider a stationary process $\{y_t : t \geq 1\}$ verifying

$$y_t = \sum_{i=1}^{p_1} \beta_{o,i} y_{t-i} + \sum_{j=1}^{p_2} g_{o,j}(y_{t-p_1-j}) + \varepsilon_t, \quad \varepsilon_t = u_t - \sum_{j=1}^{q} \theta_{o,j} u_{t-j}, \tag{3}$$

with u_t i.i.d. random variables and u_t independent of $\{y_{t-j}, j \geq 1\}$, $Eu_t = 0$ and $Eu_t^2 < \infty$. From now on, we will refer to a stochastic process verifying (3) as a partly linear ARMA (p_1, p_2, q) model and it will be denoted by PARTLIARMA (p_1, p_2, q) model.

This paper is organised as follows. In Section 2, we establish conditions for the strong ergodicity of a PARTLIARMA (p_1, p_2, q) process. From this last statement, under Harris recurrence and aperiodicity of the chain, it follows that the process is also a geometric α-mixing process. As is well known, mixing conditions have shown to be useful to derive asymptotic properties of kernel estimates for nonparametric autoregression models and partially linear time series models (Bosq 1996; Gao 1998 and Gao 2007). Related results for purely nonparametric ARCH time series were given by, for example, Masry and Tjøstheim (1995). In Section 3.1, we discuss several issues regarding how to define a Fisher-consistent functional for g_o, $\boldsymbol{\beta}_o = (\beta_{o,1}, \dots, \beta_{o,p_1})^{\mathrm{T}}$ and $\boldsymbol{\theta}_o = (\theta_{o,1}, \dots, \theta_{o,q})^{\mathrm{T}}$, which will lead to the estimation procedure to be introduced in Section 3.2 and to the algorithm described in Section 3.3. Furthermore, in Section 3.2, the asymptotic behaviour of the proposals is studied. In Section 4, we define a statistic to test $H_0 : \boldsymbol{\theta} = (\theta_1, \dots, \theta_q)^{\mathrm{T}} = \mathbf{0}$. When defining the estimators and the test statistics, for simplicity we will assume that $p_2 = 1$. Finally, in Section 5, we describe the results of a simulation study. Proofs are given in the Appendix.

2. Ergodicity of PARTLIARMA(p_1, p_2, q) models

Let $\mathbf{Y}_t = (y_t, y_{t-1}, \dots, y_{t-p_1-p_2+1})^{\mathrm{T}}$ and $\mathbf{V}_t = (\varepsilon_t, \mathbf{0}_{p_1+p_2-1})^{\mathrm{T}}$ with $\mathbf{0}_p$ the null vector in $\mathbb{R}^p$. Thus, the process $\{y_t\}$, defined by Equation (3), has the following markovian representation:

$$\mathbf{Y}_t = \mathbf{F}(\mathbf{Y}_{t-1}) + \mathbf{V}_t, \tag{4}$$

where $\mathbf{F}(\mathbf{y}) = (F_1(\mathbf{y}), \ldots, F_{p_1+p_2}(\mathbf{y}))^{\mathrm{T}}$, $F_1(\mathbf{y}) = \sum_{i=1}^{p_1} \beta_{o,i}\, y_i + \sum_{j=1}^{p_2} g_{o,j}(y_{p_1+j})$ and $F_j(\mathbf{y}) = y_{j-1}$ for $2 \le j \le p_1 + p_2$, with $\mathbf{y} = (y_1, \ldots, y_{p_1+p_2})^{\mathrm{T}}$.

In this section, we will use similar techniques to those considered in Mokkadem (1987) and Ango Nze (1998) to derive the ergodicity and strong ergodicity of the process defined by Equation (1) using the representation (4).

Denote $\mathbf{u}_{t-1} = (u_{t-1}, u_{t-2}, \ldots, u_{t-q})^{\mathrm{T}}$ and $\tilde{\boldsymbol{\theta}}_o = (\theta_{o,1}, \ldots, \theta_{o,q-1})^{\mathrm{T}}$. Let f and $f_{q,\mathbf{y}}$ be the densities of u_t and $\mathbf{u}_q|\mathbf{Y}_q = \mathbf{y}$, respectively, both with respect to the Lebesgue measure λ. Then, $\varepsilon_t|\mathbf{Y}_{t-1} = \mathbf{y}$ has a density $f_{\varepsilon,\mathbf{y}}$ given by

$$f_{\varepsilon,\mathbf{y}}(e) = \int f(z) \int f_{q,\mathbf{y}}\left(\tilde{\mathbf{w}}, \frac{z - e - \tilde{\boldsymbol{\theta}}_o^{\mathrm{T}}\tilde{\mathbf{w}}}{\theta_{o,q}}\right) \mathrm{d}z\, \mathrm{d}\tilde{\mathbf{w}},$$

where $\tilde{\mathbf{w}} = (w_1, \ldots, w_{q-1})^{\mathrm{T}}$. Let $r_j^+(\mathbf{y}) = E(|u_{t-j}|\,|\mathbf{Y}_{t-1} = \mathbf{y})$ and assume the following conditions:

H1. For all $1 \le j \le p_2$, $g_{o,j}$ is bounded over compact sets.
H2. $\inf_{e\in\mathcal{K}_1,\mathbf{y}\in\mathcal{K}_2} f_{\varepsilon,\mathbf{y}}(e) > b(\mathcal{K}_1, \mathcal{K}_2) > 0$ for any compact sets $\mathcal{K}_1 \subset \mathbb{R}$ and $\mathcal{K}_2 \subset \mathbb{R}^{p_1+p_2}$.
H3. There exist $M > 0$ and $\eta > E(|u_1|)$ such that the following holds:
(i) $|\sum_{i=1}^{p_1} \beta_{o,i}\, y_i + \sum_{j=1}^{p_2} g_{o,j}(y_{p_1+j})| + \sum_{j=1}^{q} |\theta_{o,j}| r_j^+(\mathbf{y}) \le \|\mathbf{y}\| - \eta$, for $\|\mathbf{y}\| > M$.
(ii) $\sup_{\|\mathbf{y}\|\le M} r_j^+(\mathbf{y}) < \infty, \quad 1 \le j \le q$.

Note that **H1** and **H3** (ii) entail that $\sup_{\|\mathbf{y}\|\le M}[|m_o(\mathbf{y})| + \sum_{j=1}^{q} |\theta_{o,j}| r_j^+(\mathbf{y})] < \infty$, where $m_o(\mathbf{y}) = \sum_{i=1}^{p_1} \beta_{o,i}\, y_i + \sum_{j=1}^{p_2} g_{o,j}(y_{p_1+j})$.

Remark 2.1 Let $P^n(\mathbf{y}, \cdot)$ and λ stand for the law of $\mathbf{Y}_t|\mathbf{Y}_{t-n} = \mathbf{y}$ and the Lebesque measure, respectively. It is easy to see that similar arguments to those used in Proposition 1 of Mokkadem (1987) (using the ergodicity criterion given by Tweedie 1975) reduce the problem of proving ergodicity to show the following conditions:

A1. For all Borelian set A with $\lambda(A) \ne 0$ and any compact set $\mathcal{K} \subset \mathbb{R}^{p_1+p_2}$, there exists a positive integer n_0 such that

$$\inf_{\mathbf{y}\in\mathcal{K}} P^{n_0}(\mathbf{y}, A) > 0.$$

A2. There exist $M > 0$, $\eta > 0$ and $s > 0$ such that
(i) $E|\sum_{i=1}^{p_1} \beta_{o,i}\, y_i + \sum_{j=1}^{p_2} g_{o,j}(y_{p_1+j}) + \varepsilon_{\mathbf{y},t}|^s \le \|\mathbf{y}\|^s - \eta, \quad \text{for} \quad \|\mathbf{y}\| > M,$
(ii) $\sup_{\|\mathbf{y}\|\le M} E|\sum_{i=1}^{p_1} \beta_{o,i} y_i + \sum_{j=1}^{p_2} g_{o,j}(y_{p_1+j}) + \varepsilon_{\mathbf{y},t}|^s < \infty$,
where $\varepsilon_{\mathbf{y},t}$ is a random variable with distribution given by the law of $\varepsilon_t|\mathbf{Y}_{t-1} = \mathbf{y}$,

while aperiodicity is implied by condition

A3. There exists $n_1 \in \mathbb{N}$ such that $P^{n_1}(\mathbf{y}, \cdot)$ and λ are equivalent for all $\mathbf{y}$.

Clearly, **H1** and **H3** imply **A2** with $s = 1$. The following Proposition shows that **A1** and **A3** follow from **H1** and **H2**.

Proposition 2.1 *Under **H1** and **H2**, the chain defined by Equation* (4) *satisfies conditions **A1** and **A3**.*

Remark 2.2 Since P^n is absolutely continuous with respect to λ, for $n \ge p_1 + p_2$, under **A1** the chain is strongly irreducible.

Let π be a sub-invariant measure for $\{\mathbf{Y}_t\}$; in the ergodic case, π is the invariant probability. As in Lemma 1 of Mokkadem (1987), we have that, under **H1** and **H2**, for each compact set $\mathcal{K} \subset \mathbb{R}^{p_1+p_2}$, $\lambda(\mathcal{K}) > 0$ implies $0 < \pi(\mathcal{K}) < \infty$. Therefore, we have the following result.

PROPOSITION 2.2 *Under **H1** to **H3** , any* PARTLIARMA(p_1, p_2, q) *is ergodic.*

We recall the following definition and results:

- A Markov chain $\{X_t\}$ is geometrically ergodic if there exists $0 < \rho < 1$ such that $\|P^n(x, \cdot) - \pi\| = O(\rho^n)$ for almost all $x(\pi)$, where $\| \cdot \|$ stands for the total variation norm.
- In Nummelin and Tuominen (1982), it is shown that if $\{X_t\}$ is geometrically ergodic Harris recurrent and aperiodic, then

$$\int \|P^n(x, \cdot) - \pi\|\pi(\mathrm{d}x) = O(\rho^n). \tag{5}$$

- Finally, in Rosenblatt (1971) it is shown that Equation (5) implies that the process $\{X_t\}$ is α-mixing with $\alpha(n) = a^n$, for some $0 < a < 1$ (geometrically α-mixing process).

PROPOSITION 2.3 *Under **H1**, **H2** and **H3**, the chain* $\{\mathbf{Y}_t\}$ *defined by Equation* (4) *is Harris recurrent and π and λ are equivalent.*

Proposition 3 in Mokkadem (1987) entails that **A1**, **A2** and the following condition:

A4. There exist $s > 0$, $M > 0$ and $0 < \rho < 1$ such that
(i) $E|\sum_{i=1}^{p_1} \beta_{o,i}\, y_i + \sum_{j=1}^{p_2} g_{o,j}(y_{p_1+j}) + \varepsilon_{\mathbf{y},t}|^s \leq \rho\|\mathbf{y}\|^s$, for $\|\mathbf{y}\| > M$,
(ii) $\sup_{\|\mathbf{y}\|\leq M} E|\sum_{i=1}^{p_1} \beta_{o,i}\, y_i + \sum_{j=1}^{p_2} g_{o,j}(y_{p_1+j}) + \varepsilon_{\mathbf{y},t}|^s < \infty$

imply the geometric ergodicity. Moreover, π has a moment of order s.

A4 can be derived from **H1** and **H4** with

H4. (i) There exist $M > 0$ and $0 < \rho < 1$ such that, for $\|\mathbf{y}\| > M$, $|\sum_{i=1}^{p_1} \beta_{o,i}\, y_i + \sum_{j=1}^{p_2} g_{o,j}(y_{p_1+j})| + \sum_{j=1}^{q} |\theta_{o,j}| r_j^+(\mathbf{y}) \leq \rho\|\mathbf{y}\|$.
(ii) $\sup_{\|\mathbf{y}\|\leq M} r_j^+(\mathbf{y}) < \infty$, $1 \leq j \leq q$, for any $M > 0$, with $r_j^+(\mathbf{y}) = E(|u_{t-j}|\, |\mathbf{Y}_{t-1} = \mathbf{y})$.
Putting all together we have the following result.

PROPOSITION 2.4 *Under **H1** to **H4**, any* PARTLIARMA(p_1, p_2, q) *process is a geometrically α-mixing process.*

3. Estimation in PARTLIARMA$(p_1, 1, q)$ models

For simplicity and convenience, from now on, we will focus our attention to the case $p_2 = 1$, which leads to the PARTLIARMA$(p_1, 1, q)$ model

$$y_t = \sum_{i=1}^{p_1} \beta_{o,i} y_{t-i} + g_o(y_{t-p_1-1}) + \varepsilon_t, \quad \varepsilon_t = u_t - \sum_{j=1}^{q} \theta_{o,j} u_{t-j} \tag{6}$$

with u_t i.i.d. and u_t independent of $\{y_{t-j},\ j \geq 1\}$, $Eu_t = 0$ and $Eu_t^2 < \infty$. When $p_2 > 1$, the autoregression components $g_{o,j}$, $1 \leq j \leq p_2$, can be estimated using, for instance, marginal integration under suitable conditions such as $E(g_{o,j}(y_{t-p_1-j})) = 0$.

Denote by $\mathbf{y}_{t-1} = (y_{t-1}, \ldots, y_{t-p_1})^{\mathrm{T}}$, $\boldsymbol{\phi}_1(y) = E(\mathbf{y}_{t-1}|y_{t-p_1-1} = y)$ and $\phi_2(y) = E(y_t|\, y_{t-p_1-1} = y)$.

3.1. *A Fisher-consistent functional*

When $1 \leq q \leq p_1$, model (6) implies that $g_o(y) = \phi_2(y) - \boldsymbol{\beta}_o^{\mathrm{T}}\boldsymbol{\phi}_1(y)$ and so model (6) can be written as $r_t = \boldsymbol{\beta}_o^{\mathrm{T}}\mathbf{z}_t + u_t - \sum_{j=1}^{q}\theta_{o,j}u_{t-j}$ with $r_t = y_t - \phi_2(y_{t-p_1-1})$ and $\mathbf{z}_t = \mathbf{y}_{t-1} - \boldsymbol{\phi}_1(y_{t-p_1-1})$. This implies that the autoregression parameter $\boldsymbol{\beta}_o$ and the autoregression function g_o can be estimated as in the partly linear autoregressive case (i.e. when $\theta_{o,j} = 0$). Finally, the moving average parameter can be estimated by considering the residuals, as in the linear case.

A more interesting situation arises when $q > p_1$. Using that $y_t = \boldsymbol{\beta}_o^{\mathrm{T}}\mathbf{y}_{t-1} + g_o(y_{t-p_1-1}) + u_t - \sum_{j=1}^{q}\theta_{o,j}u_{t-j}$, we get that the function g_o depends not only on the conditional expectations $\boldsymbol{\phi}_1$ and ϕ_2 and the autoregression parameter $\boldsymbol{\beta}_o$, but also on the moving average parameters $\theta_{o,j}$ for $j \geq p_1 + 1$. Indeed, $g_o(y) = \phi_2(y) - \boldsymbol{\beta}_o^{\mathrm{T}}\boldsymbol{\phi}_1(y) + \sum_{j=1}^{q}\theta_{o,j}\eta_j(y)$, where $\eta_j(y) \equiv 0$, $1 \leq j \leq p_1$ and $\eta_j(y) = E(u_{t-j}|y_{t-p_1-1} = y)$ for $j \geq p_1 + 1$. Hence, the unknown parameters and the unknown autoregression function cannot be estimated as easily as in the previous case. However, using that $u_t = (1 - \sum_{j=1}^{q}\theta_{o,j}B^j)^{-1}\epsilon_t$ and $\epsilon_t = y_t - \boldsymbol{\beta}_o^{\mathrm{T}}\mathbf{y}_{t-1} + g_o(y_{t-p_1-1})$, it can easily be seen that

- $g_o(y)$ minimises

$$L_1(a) = E\left[\left(y_t - \boldsymbol{\beta}_o^{\mathrm{T}}\mathbf{y}_{t-1} + \sum_{j=p_1+1}^{q}\theta_{o,j}u_{t-j} - a\right)^2 |y_{t-p_1-1} = y\right];$$

- $\boldsymbol{\beta}_o$ minimises

$$L_2(\mathbf{b}) = E\left[y_t - g_o(y_{t-p_1-1}) + \sum_{j=1}^{q}\theta_{o,j}u_{t-j} - \mathbf{b}^{\mathrm{T}}\,\mathbf{y}_{t-1}\right]^2;$$

- $\boldsymbol{\theta}_o = (\theta_{o,1}, \cdots, \theta_{o,q})^{\mathrm{T}}$ minimises

$$L_3(\vartheta_1, \ldots, \vartheta_q) = E\left[\left(1 - \sum_{j=1}^{q}\vartheta_j B^j\right)^{-1}(y_t - \boldsymbol{\beta}_o^{\mathrm{T}}\mathbf{y}_{t-1} - g_o(y_{t-p_1-1}))\right]^2.$$

This suggests to consider the following system of equations

$$\begin{aligned} g_{\mathbf{b},\boldsymbol{\vartheta},F}(y) &= \operatorname*{argmin}_{a\in\mathbb{R}} E_F\left[\left(y_t - \mathbf{b}^{\mathrm{T}}\mathbf{y}_{t-1} + \sum_{j=p_1+1}^{q}\vartheta_j u_{t-j} - a\right)^2 |y_{t-p_1-1} = y\right], \\ \left(\boldsymbol{\beta}_F^{\mathrm{T}}, \boldsymbol{\theta}_F^{\mathrm{T}}\right)^{\mathrm{T}} &= \operatorname*{argmin}_{(\mathbf{b}^{\mathrm{T}},\boldsymbol{\vartheta}^{\mathrm{T}})^{\mathrm{T}}\in\mathbb{R}^{p_1}\times\Theta} E_F\left[y_t - g_{\mathbf{b},\boldsymbol{\vartheta},F}(y_{t-p_1-1}) + \sum_{j=1}^{q}\vartheta_j u_{t-j} - \mathbf{b}^{\mathrm{T}}\,\mathbf{y}_{t-1}\right]^2 \\ &= \operatorname*{argmin}_{(\mathbf{b}^{\mathrm{T}},\boldsymbol{\vartheta}^{\mathrm{T}})^{\mathrm{T}}\in\mathbb{R}^{p_1}\times\Theta} M(\mathbf{b}, \boldsymbol{\vartheta}), \end{aligned} \tag{7}$$

where F denotes the distribution of the process. The index F will be omitted when the notation does not lead to misunderstanding; in particular, it will be omitted in the conditional expectations. As is well known, $g_{b,\boldsymbol{\vartheta},F}(y) = \phi_2(y) - \mathbf{b}^{\mathrm{T}}\boldsymbol{\phi}_1(y) + \sum_{j=p_1+1}^{q}\vartheta_j\eta_j(y)$. Thus, the solution $(\boldsymbol{\beta}_F^{\mathrm{T}}, \boldsymbol{\theta}_F^{\mathrm{T}})$

of Equation (7) will be a solution of the differentiated equations

$$\begin{aligned} L_0(\mathbf{b}, \boldsymbol{\vartheta}) &= 0, \\ L_\ell(\mathbf{b}, \boldsymbol{\vartheta}) &= 0, \end{aligned} \tag{8}$$

where for $1 \le \ell \le q$,

$$\begin{aligned} L_0(\mathbf{b}, \boldsymbol{\vartheta}) &= E_F\{[y_t - \phi_2(y_{t-p_1-1}) - \mathbf{b}^{\mathrm{T}}(\mathbf{y}_{t-1} - \boldsymbol{\phi}_1(y_{t-p_1-1})) + \boldsymbol{\vartheta}^{\mathrm{T}}\mathbf{s}_t](\mathbf{y}_{t-1} - \boldsymbol{\phi}_1(y_{t-p_1-1}))\}, \\ L_\ell(\mathbf{b}, \boldsymbol{\vartheta}) &= E_F\{[y_t - \phi_2(y_{t-p_1-1}) - \mathbf{b}^{\mathrm{T}}(\mathbf{y}_{t-1} - \boldsymbol{\phi}_1(y_{t-p_1-1})) + \boldsymbol{\vartheta}^{\mathrm{T}}\mathbf{s}_t](u_{t-\ell} - \eta_\ell(y_{t-p_1-1}))\}, \end{aligned}$$

with $\mathbf{s}_t = (u_{t-1} - \eta_1(y_{t-p_1-1}), \ldots, u_{t-q} - \eta_q(y_{t-p_1-1}))^{\mathrm{T}}$ and $\eta_j \equiv 0$ for $j \le p_1$.

The following result states that the unique solution of Equation (7) is $(g_o, \boldsymbol{\beta}_o^{\mathrm{T}}, \boldsymbol{\theta}_o^{\mathrm{T}})$, which entails the Fisher-consistency of the functional. The relevance of considering Fisher-consistent functionals is that Fisher-consistency is the property one usually first derives, since it means that we are estimating the right quantities at the idealised model as it also guarantees uniqueness of solution in the functional equations.

Theorem 3.1.1 *If model* (6) *holds and*

(a) $P(\sum_{i=1}^{p_1} d_i y_{t-i} = h(y_{t-p_1-1}) + \sum_{j=1}^{q} a_j u_{t-j}) < 1, 1 \le s \le p_1$,
(b) $P(u_{t-\ell} + \sum_{j=\ell+1}^{q} a_j u_{t-j} = h(y_{t-p_1-1})) < 1, 1 \le \ell \le q$

for any $\boldsymbol{d} = (d_1, \ldots, d_{p_1})^{\mathrm{T}}$, $\boldsymbol{a} = (a_1, \ldots, a_q)^{\mathrm{T}}$ and any smooth function h that are not simultaneously equal to 0, then $(\boldsymbol{\beta}_o^{\mathrm{T}}, \boldsymbol{\theta}_o^{\mathrm{T}})$ is the unique solution of Equation (7).

Remark 3.1.1 Condition (b) holds if the conditional distribution of $(u_{t-j})_{j=1}^{q}|y_{t-p_1-1} = y$ has a density almost surely. Condition (a) states that model (6) is effectively partly linear and not purely nonparametric, that is, the process cannot be written as $y_t = h_o(y_{t-p_1-1}) + \varepsilon_t$. For instance, when $p_1 = 1$, if $a_j = 0$, Condition (a) prevents y_{t-1} from being a.s. perfectly predictable from y_{t-2} (see Robinson (1988)). On the other hand, if $h \equiv 0$, Condition (a) prevents the nonidentificability of the moving average coefficients.

The following result states that under the same conditions of Theorem 3.1.1, Equation (8) admits as unique solution $(\boldsymbol{\beta}_o^{\mathrm{T}}, \boldsymbol{\theta}_o^{\mathrm{T}})$ and so the differentiated system of equations can be used to define the functional.

Theorem 3.1.2 *If model* (6) *holds and*

(a) $P(\sum_{i=1}^{p_1} d_i y_{t-i} = h(y_{t-p_1-1}) + \sum_{j=1}^{q} a_j u_{t-j}) < 1, 1 \le s \le p_1$,
(b) $P(u_{t-\ell} + \sum_{j=\ell+1}^{q} a_j u_{t-j} = h(y_{t-p_1-1})) < 1, 1 \le \ell \le q$

for any $\boldsymbol{d} = (d_1, \ldots, d_{p_1})^{\mathrm{T}}$, $\boldsymbol{a} = (a_1, \ldots, a_q)^{\mathrm{T}}$ and any smooth function h that are not simultaneously equal to 0, then $(\boldsymbol{\beta}_o^{\mathrm{T}}, \boldsymbol{\theta}_o^{\mathrm{T}})$ is the unique solution of Equation (8).

3.2. *Parameter estimation*

The system of equations (7) suggests that estimators may be obtained by replacing the true distribution F by its empirical version, that is we can define $(\hat{g}, \hat{\boldsymbol{\beta}}^{\mathrm{T}}, \hat{\boldsymbol{\theta}}^{\mathrm{T}})$ as the solution

$$
\begin{aligned}
\hat{g}_{\mathbf{b},\boldsymbol{\vartheta}}(y) &= \operatorname*{argmin}_{a\in\mathbb{R}} \sum_{t=p_1+1}^{T} w_{tT}(y)\left(y_t - \mathbf{b}^{\mathrm{T}}\mathbf{y}_{t-1} + \sum_{j=p_1+1}^{q} \vartheta_j u_{t-j} - a\right)^2 \\
\left(\hat{\boldsymbol{\beta}}^{\mathrm{T}}, \hat{\boldsymbol{\theta}}^{\mathrm{T}}\right)^{\mathrm{T}} &= \operatorname*{argmin}_{(\mathbf{b}^{\mathrm{T}},\boldsymbol{\vartheta}^{\mathrm{T}})^{\mathrm{T}}\in\mathbb{R}^{p_1}\times\Theta} \frac{1}{T}\sum_{t=p_1+2}^{T}\left[y_t - \hat{g}_{\mathbf{b},\boldsymbol{\vartheta}}(y_{t-p_1-1}) + \sum_{j=1}^{q}\vartheta_j u_{t-j} - \mathbf{b}^{\mathrm{T}}\,\mathbf{y}_{t-1}\right]^2 \qquad (9)\\
&= \operatorname*{argmin}_{(\mathbf{b}^{\mathrm{T}},\boldsymbol{\vartheta}^{\mathrm{T}})^{\mathrm{T}}\in\mathbb{R}^{p_1}\times\Theta} M_T(\boldsymbol{\vartheta}, \mathbf{b}),
\end{aligned}
$$

where the local weights w_{tT} may be taken, for instance, as the kernel weights

$$
w_{tT}(y) = K((y - y_{t-p_1-1})/h_T)\left[\sum_{t=p_1+1}^{T} K((y - y_{t-p_1-1})/h_T)\right]^{-1}.
$$

The kernel $K : \mathbb{R} \to \mathbb{R}$ is a density function with 0 mean and finite variance and the bandwidth h_T satisfies $h_T \to 0$, $Th_T \to \infty$ as $T \to \infty$. Note that $\hat{g}_{\mathbf{b},\boldsymbol{\vartheta}}(y) = \hat{\phi}_2(y) - \mathbf{b}^{\mathrm{T}}\hat{\boldsymbol{\phi}}_1(y) + \sum_{j=p_1+1}^{q}\vartheta_j\widehat{\eta}_j(y)$, where

$$
\hat{\boldsymbol{\phi}}_1(y) = \sum_{t=p_1+1}^{T} w_{tT}(y)\mathbf{y}_{t-1}, \quad \widehat{\phi}_2(y) = \sum_{t=p_1+1}^{T} w_{tT}(y)y_t, \quad \widehat{\eta}_j(y) = \sum_{t=p_1+1}^{T} w_{tT}(y)u_{t-j}.
$$

THEOREM 3.2.1 *Let us assume that model (*6*) holds.*

(i) *If, in addition,*
  (a) $\sum_{t=p_1+2}^{T}(\hat{\phi}_2(y_{t-p_1-1}) - \phi_2(y_{t-p_1-1}))^2/T \xrightarrow{p} 0$,
  (b) $\sum_{t=p_1+2}^{T}\|\hat{\boldsymbol{\phi}}_1(y_{t-p_1-1}) - \boldsymbol{\phi}_1(y_{t-p_1-1})\|^2/T \xrightarrow{p} 0$,
  (c) $\sum_{t=p_1+2}^{T}(\hat{\eta}_j(y_{t-p_1-1}) - \eta_j(y_{t-p_1-1}))^2/T \xrightarrow{p} 0, 1 \le j \le q$,
  (d) *the covariance matrix* $\mathbf{C}$ *of* $(\mathbf{z}_t^{\mathrm{T}}, \mathbf{s}_t^{\mathrm{T}})^{\mathrm{T}}$ *is nonsingular, where* $\mathbf{z}_t = \mathbf{y}_{t-1} - \boldsymbol{\phi}_1(y_{t-p_1-1})$, $s_{t,j} = u_{t-j} - \eta_j(y_{t-p_1-1})$ *and* $\mathbf{s}_t = (s_{t,1}, \ldots, s_{t,q})^{\mathrm{T}}$,

  we have that, $\hat{\boldsymbol{\beta}} \xrightarrow{p} \boldsymbol{\beta}_o$, $\hat{\boldsymbol{\theta}} \xrightarrow{p} \boldsymbol{\theta}_o$.

(ii) *Let* $\mathcal{K} \subset \mathbb{R}$ *be a compact set. If* $\hat{\boldsymbol{\beta}} \xrightarrow{p} \boldsymbol{\beta}_o$, $\hat{\boldsymbol{\theta}} \xrightarrow{p} \boldsymbol{\theta}_o$ *and* $\sup_{y\in\mathcal{K}}|\hat{\phi}_2(y) - \phi_2(y)| \xrightarrow{p} 0$ *and* $\sup_{y\in\mathcal{K}}\|\hat{\boldsymbol{\phi}}_1(y) - \boldsymbol{\phi}_1(y)\| \xrightarrow{p} 0$, *we have that* $\sup_{y\in\mathcal{K}}|\hat{g}(y) - g_o(y)| \xrightarrow{p} 0$.

Remark 3.2.1 When considering kernel-based smoothers, conditions ensuring that assumptions (a) to (c) hold can be found in Lemma 6.6.7 in Härdle, Liang and Gao (2000). On the other hand, the uniform convergence conditions required in (ii) imply (a) and (b) if the process y_t is bounded. Besides, the uniform consistency over compact sets can be obtained from Theorem 3.2 in Bosq (1996). When considering the nearest neighbour with kernel weights, uniform consistency results can be found in Collomb (1985).

THEOREM 3.2.2 *Let us assume that model (*6*) holds and that* $\{y_t : t \in \mathbb{Z}\}$ *is a geometric* α*-mixing process. Denote* $\boldsymbol{a}_t = r_t\boldsymbol{x}_t - E(r_t\boldsymbol{x}_t)$, *where* $r_t = y_t - \phi_2(y_{t-p_1-1})$ *and* $\boldsymbol{x}_t = (\boldsymbol{z}_t^T, \boldsymbol{s}_t^T)^T$, *with* $\mathbf{z}_t = \boldsymbol{y}_{t-1} - \boldsymbol{\phi}_1(y_{t-p_1-1})$, $s_{t,j} = u_{t-j} - \eta_j(y_{t-p_1-1})$ *and* $\mathbf{s}_t = (s_{t,1}, \dots, s_{t,q})^T$. *Assume that for some* $\delta > 0$, $E\|\boldsymbol{a}_t\|^{2+\delta} < \infty$. *If, in addition,*

(a) $\sum_{t=p_1+2}^{T} \left(\hat{\phi}_2(y_{t-p_1-1}) - \phi_2(y_{t-p_1-1})\right)^2 / \sqrt{T} \xrightarrow{p} 0,$

(b) $\sum_{t=p_1+2}^{T} \left\|\hat{\boldsymbol{\phi}}_1(y_{t-p_1-1}) - \boldsymbol{\phi}_1(y_{t-p_1-1})\right\|^2 / \sqrt{T} \xrightarrow{p} 0,$

(c) $\sum_{t=p_1+2}^{T} \left(\widehat{\eta}_j(y_{t-p_1-1}) - \eta_j(y_{t-p_1-1})\right)^2 / \sqrt{T} \xrightarrow{p} 0, 1 \le j \le q$

(d) *the matrix* $\boldsymbol{C} = E\boldsymbol{x}_t\boldsymbol{x}_t^{\mathrm{T}}$ *is non–singular,*

(e) *the matrix* $\boldsymbol{D} = \sum_{\ell=-\infty}^{\infty} Cov(\boldsymbol{a}_0, \boldsymbol{a}_\ell)$ *is non–singular,*

we have that

$$\sqrt{T}\begin{pmatrix}\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}_o \\ \hat{\boldsymbol{\theta}} - \boldsymbol{\theta}_o\end{pmatrix} \xrightarrow{\mathcal{D}} N(0, \boldsymbol{\Sigma}),$$

where $\boldsymbol{\Sigma} = \boldsymbol{C}^{-1}\boldsymbol{D}\boldsymbol{C}^{-1}$.

As mentioned above, for kernel–based smoothers, conditions ensuring that assumptions (a) to (c) hold can be found in Lemma 6.6.7 in Härdle *et al.* (2000).

3.3. *Algorithm*

In order to obtain a genuine estimator (i.e. that does not depend on the unknown residuals u_t) it is sufficient to replace in Equation (9) the unknown residuals by predicted ones. The following iterative procedure provides a method to compute these estimators from initial estimates $g^{(0)}(y)$ of $g_o(y)$ and $\boldsymbol{\beta}^{(0)}$ of $\boldsymbol{\beta}_o$. Denote by $\hat{r}_t = y_t - \hat{\phi}_2(y_{t-p_1-1})$ and $\hat{\mathbf{z}}_t = \mathbf{y}_{t-1} - \hat{\boldsymbol{\phi}}_1(y_{t-p_1-1})$. The estimators can be computed through the following procedure:

(i) Denote by

$$\hat{u}_t^{(1)}(\boldsymbol{\vartheta}) = \left(1 - \sum_{j=1}^{q} \vartheta_j B^j\right)^{-1} \left(y_t - \boldsymbol{\beta}^{(0)\mathrm{T}}\mathbf{y}_{t-1} - g^{(0)}\left(y_{t-p_1-1}\right)\right). \tag{10}$$

As in Durbin (1959), the infinite sum can be approximated by a finite sum. Compute $\boldsymbol{\theta}^{(1)}$= $\operatorname{argmin}_{\boldsymbol{\vartheta}\in\Theta} L^{(1)}(\boldsymbol{\vartheta})$, where $L^{(1)}(\boldsymbol{\vartheta}) = 1/T \sum_{t=p_1+2}^{T} (\hat{u}_t^{(1)}(\boldsymbol{\vartheta}))^2$. Define $\hat{u}_t^{(1)} = \hat{u}_t^{(1)}(\boldsymbol{\theta}^{(1)})$.

(ii) Given $\hat{u}_t^{(n)}$, define for $p_1 + 1 \le j \le q$, the smoothers

$$\hat{\eta}_j^{(n)}(y) = \sum_{t=j+1}^{T} w_{tT;j}(y)\hat{u}_{t-j}^{(n)},$$

where

$$w_{tT;j}(y) = \frac{K((y - y_{t-p_1-1})/h_T)}{\sum_{t=j+1}^{T} K((y - y_{t-p_1-1})/h_T)}$$

and $\hat{s}_{t-j}^{(n)} = \hat{u}_{t-j}^{(n)} - \hat{\eta}_j^{(n)}(y_{t-p_1-1})$, for $1 \le j \le q$ with $\hat{\eta}_j^{(n)} \equiv 0$, for $1 \le j \le p_1$.

(iii) Given $\boldsymbol{\theta}^{(n)}$ and $\hat{v}_{t-j}^{(n)}$, define $\boldsymbol{\beta}^{(n)}$ as

$$\boldsymbol{\beta}^{(n)} = \underset{\mathbf{b}\in\mathbb{R}^{p_1}}{\operatorname{argmin}} \sum_{t=p_1+2}^{T} \left(\hat{r}_t - \mathbf{b}^{\mathrm{T}}\hat{\mathbf{z}}_t + \sum_{j=1}^{q} \theta_j^{(n)}\hat{s}_{t-j}^{(n)}\right)^2.$$

(iv) Given $\hat{\eta}_j^{(n)}(y)$, $\boldsymbol{\theta}^{(n)}$ and $\boldsymbol{\beta}^{(n)}$, let

$$g^{(n)}(y) = \hat{\phi}_2(y) - \boldsymbol{\beta}^{(n)T}\hat{\boldsymbol{\phi}}_1(y) + \sum_{j=p_1+1}^{q} \theta_j^{(n)}\hat{\eta}_j^{(n)}(y).$$

(v) Given $\boldsymbol{\beta}^{(n)}$ and $g^{(n)}$, denote by

$$\widehat{u}_t^{(n+1)}(\boldsymbol{\vartheta}) = \left(1 - \sum_{j=1}^{q} \vartheta_j B^j\right)^{-1} \left(y_t - \boldsymbol{\beta}^{(n)\mathrm{T}}\mathbf{y}_{t-1} - g^{(n)}\left(y_{t-p_1-1}\right)\right)$$

and compute $\boldsymbol{\theta}^{(n+1)} = \operatorname{argmin}_{\boldsymbol{\vartheta}\in\Theta} L^{(n+1)}(\boldsymbol{\vartheta})$, where $L^{(n+1)}(\boldsymbol{\vartheta}) = \sum_{t=p_1+2}^{T}(\hat{u}_t^{(n+1)}(\boldsymbol{\vartheta}))^2/T$. Define $\hat{u}_t^{(n+1)} = \hat{u}_t^{(n+1)}(\boldsymbol{\theta}^{(n+1)})$.

Iterate (ii) to (v) until convergence is obtained. Denote by $\hat{g}$, $\hat{\boldsymbol{\beta}}$ and $\hat{\boldsymbol{\theta}}$ the resulting estimates.

Remark 3.2.2 Forecasting is one the most important goals in ARMA models. When moving averages are present, it is typically performed after the parameters have been estimated by using 'estimated residuals'. For our model prediction may be done as follows:

- For $p_1 + 2 \leq \tau \leq t - 1$, define $\widehat{\varepsilon}_\tau = y_\tau - \widehat{\boldsymbol{\beta}}^{\mathrm{T}}\mathbf{y}_{\tau-1} - \hat{g}\left(y_{\tau-p_1-1}\right)$, where $\hat{g}$ and $\widehat{\boldsymbol{\beta}}$ are the estimators of the autoregression function and the autoregression parameters, respectively, and $\widehat{\varepsilon}_\tau = 0$ otherwise.
- Given $\hat{\boldsymbol{\theta}}$ an estimate of the moving average parameters, let $\widehat{u}_\tau = \left(1 - \sum_{j=1}^{q}\hat{\theta}_j B^j\right)^{-1}\widehat{\varepsilon}_\tau$.
- Predict the observation at time t as $\widehat{\boldsymbol{\beta}}^{\mathrm{T}}\mathbf{y}_{t-1} + \hat{g}\left(y_{t-p_1-1}\right) - \sum_{j=1}^{q}\hat{\theta}_j\,\widehat{u}_{t-j}$.

Remark 3.2.3 The initial estimates, $g^{(0)}(y)$ of $g_o(y)$ and $\boldsymbol{\beta}^{(0)}$ of $\boldsymbol{\beta}_o$ can be computed by taking $\theta_j = 0, 1 \leq j \leq q$ (i.e. assuming a partly linear autoregressive model). These estimates are linear kernel-based estimators and are described in Härdle *et al.* (2000), where their asymptotic properties are stated.

3.4. *Data-driven selection of the smoothing parameters*

An important issue in any smoothing procedure is the choice of the smoothing parameter. Under a nonparametric regression model, two commonly used approaches are cross-validation and plug-in. Cross-validation methods have been also extended to the dependent setting. Hart and Wehrly (1986) studied the properties of the asymptotic mean square error of kernel smoothers and found optimal bandwidths in the context of repeated measurements data. Hart and Vieu (1990) studied the behaviour of a cross-validated bandwidth selector for kernel density estimators under an α- mixing condition. They modified the leave-out technique involved in the cross-validation method and they proved that if the leave-out sequence, ℓ_n, does not increase too fast, the bandwidth that minimises the cross–validation criterion is asymptotically optimal. In the case of the autoregression function, Härdle and Vieu (1992) considered the kernel estimator when dealing with a stationary $\alpha-$mixing process and constructed data-driven bandwidths that asymptotically minimise the averaged square error. On the other hand, Hall, Lahiri and Truong (1995) proved that, except for long-range dependent data, under general conditions, the asymptotically optimal bandwidth for the density estimation under independence is still a good choice. They proposed a plug-in rule and through a simulation study they compared their proposal with the leave-out cross-validation bandwidths.

See also, Györfi, Härdle, Sarda and Vieu (1989) and Hart (1996) for a review. In the context of partly linear autoregression models, the selection of the smoothing parameter has been described in Härdle *et al.* (2000).

We may consider a cross-validation approach as follows, where to make explicit the dependence on the bandwidth h we introduce the superindex h for the estimators.

- Split the sample into two subsets by selecting a proportion $0 < \alpha < 1$ of the number of observation. Let $\mathcal{I}_\alpha = \{1, \ldots, [\alpha T]\}$ stand for the indexes of these observations and $\mathcal{J}_\alpha$ for the indexes of the remaining ones.
- For each given h, compute the estimates $\widehat{\boldsymbol{\beta}}^{(h)}$, $\hat{g}^{(h)}$ and $\hat{\boldsymbol{\theta}}^{(h)}$ based only on the observations $\{y_t, t \in \mathcal{I}_\alpha\}$.
- Choose

$$\hat{h}_n = \operatorname*{argmin}_h \sum_{t \in \mathcal{J}_\alpha} \left(y_t - \widehat{y}_t^{(h)} \right)^2,$$

where the predicted observation at time t, $\widehat{y}_t^{(h)} = \widehat{\boldsymbol{\beta}}^{(h)\mathrm{T}} \mathbf{y}_{t-1} + \hat{g}^{(h)}(y_{t-p_1-1}) - \sum_{j=1}^q \hat{\theta}_j^{(h)} \hat{u}_{t-j}^{(h)}$, is computed as suggested in Remark 3.2.2 using $\hat{\boldsymbol{\beta}}^{(h)}$, $\hat{g}^{(h)}$, $\hat{\boldsymbol{\theta}}^{(h)}$ and the observations $\{y_t, t \in \mathcal{J}_\alpha\}$. For small sample sizes, one may adapt the time series cross-validation criterion introduced by Hart (1994).

4. An asymptotic test for $H_0 : \theta_j = 0, 1 \leq j \leq q$.

Usually, the moving average component is introduced in order to decrease, in the autoregression function, the dependence of the past. Therefore, it is quite natural to check for a given data set if it is worth to include the MA component. Hence, the aim of this section is to provide a test statistic to test $H_0 : \boldsymbol{\theta} = (\theta_1, \ldots, \theta_q)^{\mathrm{T}} = \mathbf{0}$.

According to the following Lemma, we can consider an equivalent test for the first q coefficients of the inverted MA operator.

LEMMA 4.1 *Let us assume that model* (6) *holds. Then,* $H_0 : \boldsymbol{\theta} = (\theta_1, \ldots, \theta_q)^{\mathrm{T}} = \mathbf{0}$ *is equivalent to* $H_0 : \boldsymbol{\gamma}^{(q)} = (\gamma_1, \ldots, \gamma_q)^{\mathrm{T}} = \mathbf{0}$*, where* γ_j *are the coefficients of* $(1 - \sum_{j=1}^q \theta_j B^j)^{-1}$*, that is,* γ_j *satisfy* $(1 - \sum_{j=1}^q \theta_j B^j)^{-1} v_t = \sum_{r=0}^\infty \gamma_r v_{t-r}$.

Let $\hat{g}^{(0)}(y)$ and $\hat{\boldsymbol{\beta}}^{(0)}$ be the linear kernel estimates of the autoregression function and autoregression parameters computed under the null hypothesis, that is, assuming a partly linear autoregressive model $y_t = \boldsymbol{\beta}_o^{\mathrm{T}} \mathbf{y}_{t-1} + g_o(y_{t-p_1-1}) + u_t$, with u_t independent of $\{y_{t-j}, j \geq 1\}$. Denote $\hat{v}_t = y_t - \hat{\boldsymbol{\beta}}^{(0)\mathrm{T}} \mathbf{y}_{t-1} - \hat{g}^{(0)}(y_{t-p_1-1})$. Following Durbin (1959), given $\vartheta_1, \ldots, \vartheta_q$, we approximate the infinite sum

$$\left(1 - \sum_{j=1}^q \vartheta_j B^j\right)^{-1} \hat{v}_t = \sum_{r=0}^\infty \gamma_r \hat{v}_{t-r}$$

by a finite sum $\sum_{r=0}^N \gamma_r \hat{v}_{t-r}$ with N fairly large. Thus, we will denote $\hat{u}_t^{(1)}(\boldsymbol{\gamma}) = \sum_{r=0}^N \gamma_r \hat{v}_{t-r}$, where $\boldsymbol{\gamma} = (\gamma_0, \ldots, \gamma_N)^{\mathrm{T}}$ with $\gamma_0 = 1$.

In order to test $H_0 : \boldsymbol{\theta} = (\theta_1, \ldots, \theta_q)^{\mathrm{T}} = \mathbf{0}$, we define $\hat{\boldsymbol{\gamma}}$ as

$$\hat{\boldsymbol{\gamma}} = \operatorname*{argmin}_{\boldsymbol{\gamma}} \frac{1}{T} \sum_{t=p_1+2}^{T} \left(\hat{u}_t^{(1)}(\boldsymbol{\gamma})\right)^2.$$

Therefore, $\hat{\boldsymbol{\gamma}}$ solves

$$\sum_{t=p_1+2}^{T} \hat{u}_t^{(1)}(\hat{\boldsymbol{\gamma}})\hat{v}_{t-j} = 0, \qquad \text{for } j = 0, \ldots, N$$

or, equivalently,

$$\sum_{r=0}^{N} \hat{\gamma}_r \widehat{\mathrm{cov}}(\hat{v}_{t-j}, \hat{v}_{t-r}) = 0,$$

where $\widehat{\mathrm{cov}}(\hat{v}_{t-j}, \hat{v}_{t-r}) = \sum_{t=p_1+2}^{T} \hat{v}_{t-j}\hat{v}_{t-r}/T$.

The following theorem shows that, in order to test $H_0 : \boldsymbol{\theta} = (\theta_1, \ldots, \theta_q)^{\mathrm{T}} = \mathbf{0}$, we can reject H_0 with asymptotic α level if for some $1 \le i \le q$,

$$\sqrt{T}|\hat{\gamma}_i| > z_\delta,$$

where $\delta = (1 + (1-\alpha)^{1/q})/2$ with $P(Z \le z_\alpha) = \alpha$ and $Z \sim N(0,1)$.

Theorem 4.1 *Let us assume that model (*6*) holds. Assume that* $\hat{\boldsymbol{\beta}}^{(0)}$ *and* $\hat{g}^{(0)}$ *are consistent estimators of* $\boldsymbol{\beta}_o$ *and* g_o *such that* $\sqrt{T}(\boldsymbol{\beta}_o - \hat{\boldsymbol{\beta}}^{(0)}) = O_p(1)$. *Then, under* $H_0 : \boldsymbol{\theta} = (\theta_1, \ldots, \theta_q)^{\mathrm{T}} = \mathbf{0}$, *we have that*

$$\sqrt{T}\ \hat{\boldsymbol{\gamma}} \xrightarrow{\mathcal{D}} N(\mathbf{0}, \boldsymbol{I}). \tag{11}$$

5. Monte Carlo

5.1. *Estimation*

A simulation study was carried out to study the performance of the proposal given in Section 3.2. We have considered a Gaussian kernel smoother such that its interquartile range is 0.5 and we performed 500 replications. In order to stabilise the series, we first generate a series of size $N = 2000$ following the model

$$z_t = \beta_o z_{t-1} + g_o(z_{t-2}) + u_t - \theta_1 u_{t-1} - \theta_2 u_{t-2}, \quad 3 \le t \le N, \tag{12}$$

where $\beta_o = 0.25$, $\theta_1 = \theta_2 = -0.1$ and $g_o(z) = 0.125\pi \sin(\pi z) + 0.25z$ or $g_o(z) = 0.5z$. With the second choice of g_o, Equation (12) results in an ARMA (2,2) model. As initial values, we took $z_1 = z_2 = 0$. The innovations were i.i.d. normally distributed, $u_t \sim N(0,1)$, such that u_t is independent of $\{z_{t-1}, z_{t-2}, \ldots\}$. The data set of size $T = 1000$ to be considered consists of the series $\{y_t : 1 \le t \le T\}$, where $y_t = z_{t+1000}$.

We considered three different bandwidth values, $h = h_T = 0.4$, 0.8 and 1.2.

We have performed 10 steps with two different initial estimates $g^{(0)}(y)$ and $\beta^{(0)}$ of $g_o(y)$ and β_o, respectively. As suggested in Remark 3.2.3, we computed $g^{(0)}(y)$ and $\beta^{(0)}$ assuming a partly linear autoregressive model (i.e. $\theta_i = 0$, $i = 1, 2$). This will be denoted as Method 1 in all tables and figures. On the other hand, the so-called Method 2 consists in fitting an ARMA(2,2) to the data and then taking $\beta^{(0)} = \zeta_1$ and $g^{(0)}(y) = \zeta_2 y$, where ζ_i are the estimated autoregressive coefficients.

When $g_o(z) = 0.125\pi\ \sin(\pi z) + 0.25z$, Figures 1–3 give the boxplots of $\beta^{(1)}, {\theta_j}^{(1)}$ for $j = 1, 2$ computed in the first step of the iterative procedure and the resulting estimates obtained in the final step $\beta^{(10)}$, ${\theta_j}^{(10)}$ for $j = 1, 2$. Besides, Tables 1 and 2 give the mean square errors of the final parameter estimates and the mean over replications of the estimated mean square errors $M(g^{(1)}, g_o)$ and $M(g^{(10)}, g_o)$, where $M(\hat{g}, g) = \sum_{3\le t\le T}([\hat{g}(y_t) - g(y_t)]^2)/T$, respectively.

On the other hand, when $g_o(z) = 0.5\ z$, the left column of Figures 4–6 gives the boxplots of $\beta^{(1)}, {\theta_j}^{(1)}$ for $j = 1, 2$ and the resulting estimates obtained in final step $\beta^{(10)}, {\theta_j}^{(10)}$ for $j = 1, 2$ computed using Method 1. In order to compare with the optimal estimating procedure for an ARMA(2,2) model, as it is the case, the right column in Figures 4–6, shows the boxplots of the initial estimates $\beta^{(0)}$, ${\theta_j}^{(0)}$ for $j = 1, 2$ computed under the ARMA(2,2) and those of the first and final steps $\beta^{(1)}$, ${\theta_j}^{(1)}$, $j = 1, 2$, and $\beta^{(10)}$, ${\theta_j}^{(10)}$ for $j = 1, 2$, respectively, when using Method 2. Similarly, Tables 3 and 4 give the mean square errors of the final parameter estimates and the mean over replications of $M(g^{(0)}, g_o)$, $M(g^{(1)}, g_o)$ and $M(g^{(10)}, g_o)$, respectively.

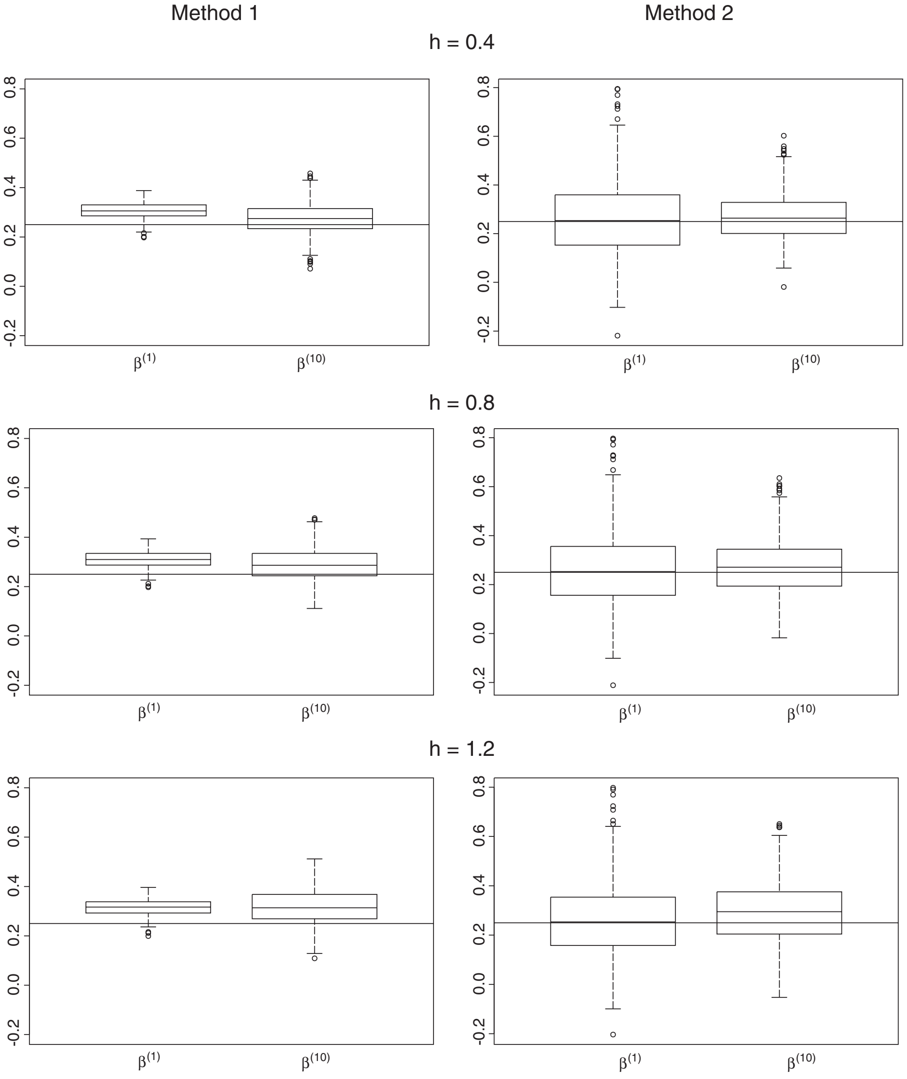

Figure 1. Boxplot of the estimates of the autoregression parameter when $g_o(z) = 0.125\pi\ \sin(\pi z) + 0.25z$.

When $g_o(z) = 0.125\pi\,\sin(\pi z) + 0.25z$, the best results are obtained for $h = 0.4$, the other two bandwiths seem to oversmooth the autoregression function (see Table 2). The initial estimates of the parameters and those obtained after 10 steps are more stable and less biased using Method 2 than Method 1, for all the considered bandwidths. Some benefit in variability is obtained after 10 steps of the iterative procedure in all cases, but the final estimates are more biased for larger bandwidths. Besides, using Method 2, the final estimates of β_o are more spread, while with Method 1 a larger bias is observed for the estimates of β_o. Similar comments hold for the estimates of θ_1 and θ_2, even when the effect on the bias is more evident for the estimates of these parameters, especially for greater values of the bandwidth. However, when balancing bias and variability, better mean square errors are obtained using Method 1, in particular, when estimating the autoregression parameter (Table 1). Table 2 shows that there is a clear improvement in the fit of g when we iterate Method 2, while Methods 1 and 2 give mean square errors of the same order after 10 steps. On

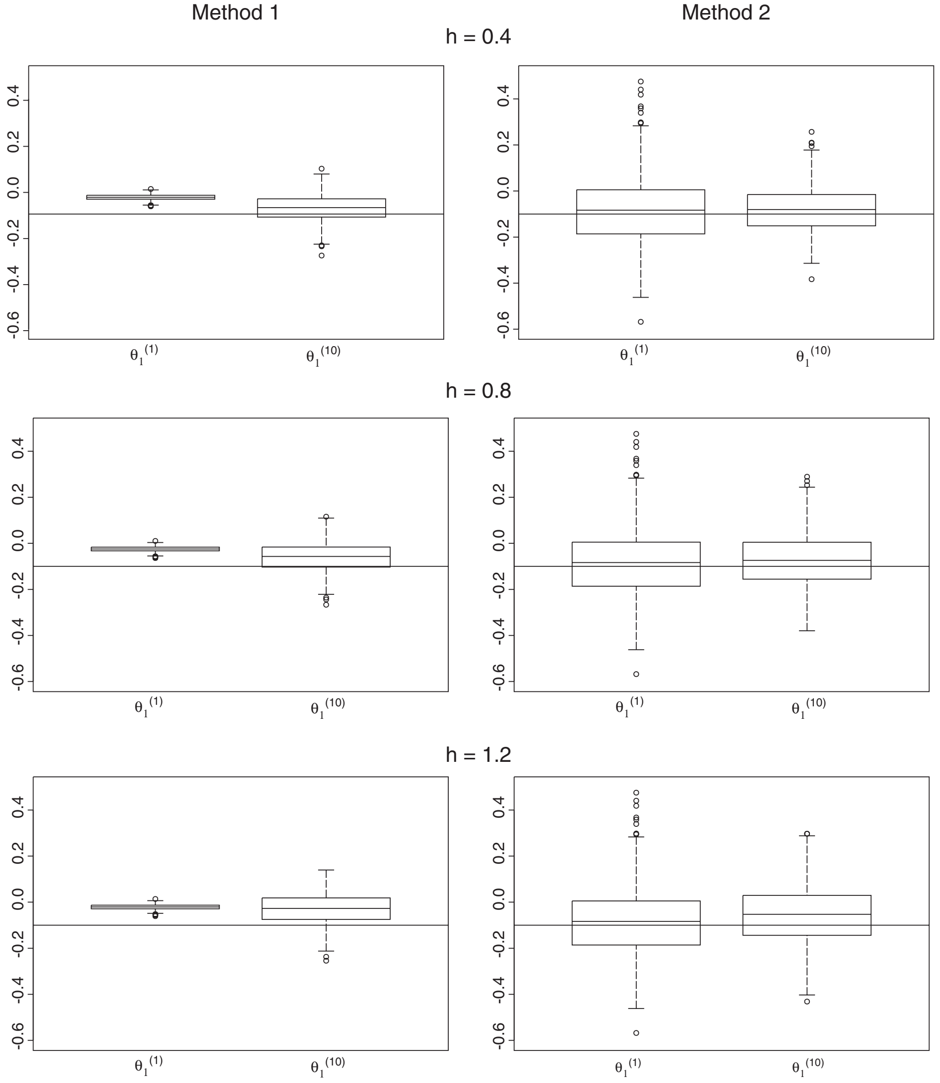

Figure 2. Boxplot of the estimates of the moving average parameter θ_1 when $g_o(z) = 0.125\pi\,\sin(\pi z) + 0.25z$.

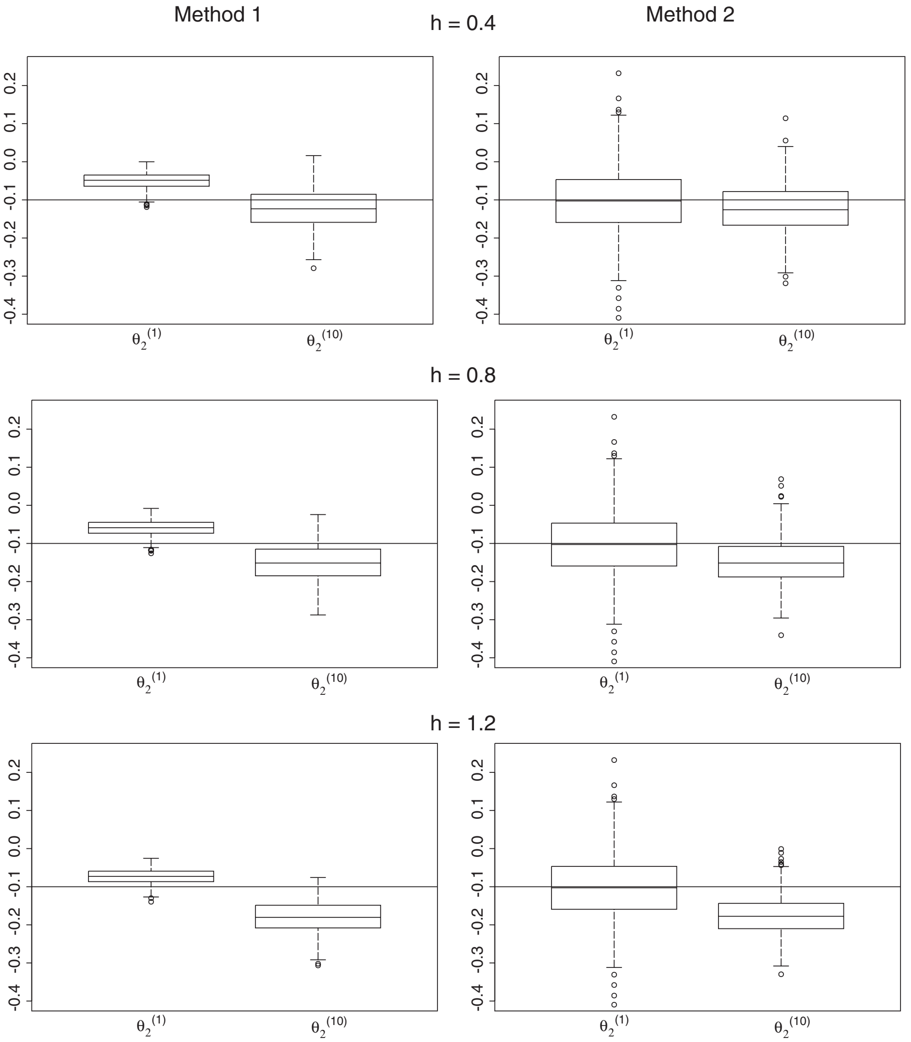

Figure 3. Boxplot of the estimates of the moving average parameter θ_2 when $g_o(z) = 0.125\pi \sin(\pi z) + 0.25z$.

Table 1. Mean square error for the autoregressive and moving average parameter estimators for $g_o(z) = 0.125\,\pi \sin(\pi z) + 0.25z$.

	$\beta^{(10)}$		$\theta_1^{(10)}$		$\theta_2^{(10)}$	
	Method 1	Method 2	Method 1	Method 2	Method 1	Method 2
$h = 0.4$	0.0046	0.0098	0.0043	0.0103	0.0035	0.0049
$h = 0.8$	0.0060	0.0142	0.0056	0.0150	0.0051	0.0063
$h = 1.2$	0.0096	0.0180	0.0094	0.0189	0.0085	0.0087
Data–driven	0.0051	0.0100	0.0047	0.0104	0.0039	0.0051

the other hand, when using Method 1, quite surprisingly, the first step estimators provide lower mean square errors than those obtained after 10 iterations.

When $g_o(z) = 0.5\,z$, the choice $h = 0.8$ gives the lower mean square errors $M(g^{(1)}, g_o)$ and $M(g^{(10)}, g_o)$ both for Methods 1 and 2, while $h = 0.4$ seems to be the best choice when estimating

Table 2. Mean of $M(g^{(1)}, g_o)$ and $M(g^{(10)}, g_o)$ for $g_o(z) = 0.125\,\pi\,\sin(\pi z) + 0.25z$.

	$h = 0.4$		$h = 0.8$		$h = 1.2$		Data–driven	
	Method 1	Method 2	Method 1	Method 2	Method 1	Method 2	Method 1	Method 2
$M(g^{(1)}, g_o)$	0.01503	0.03818	0.01774	0.03823	0.03614	0.05170	0.01653	0.04152
$M(g^{(10)}, g_o)$	0.02002	0.02445	0.02674	0.03114	0.05093	0.05241	0.02275	0.02788

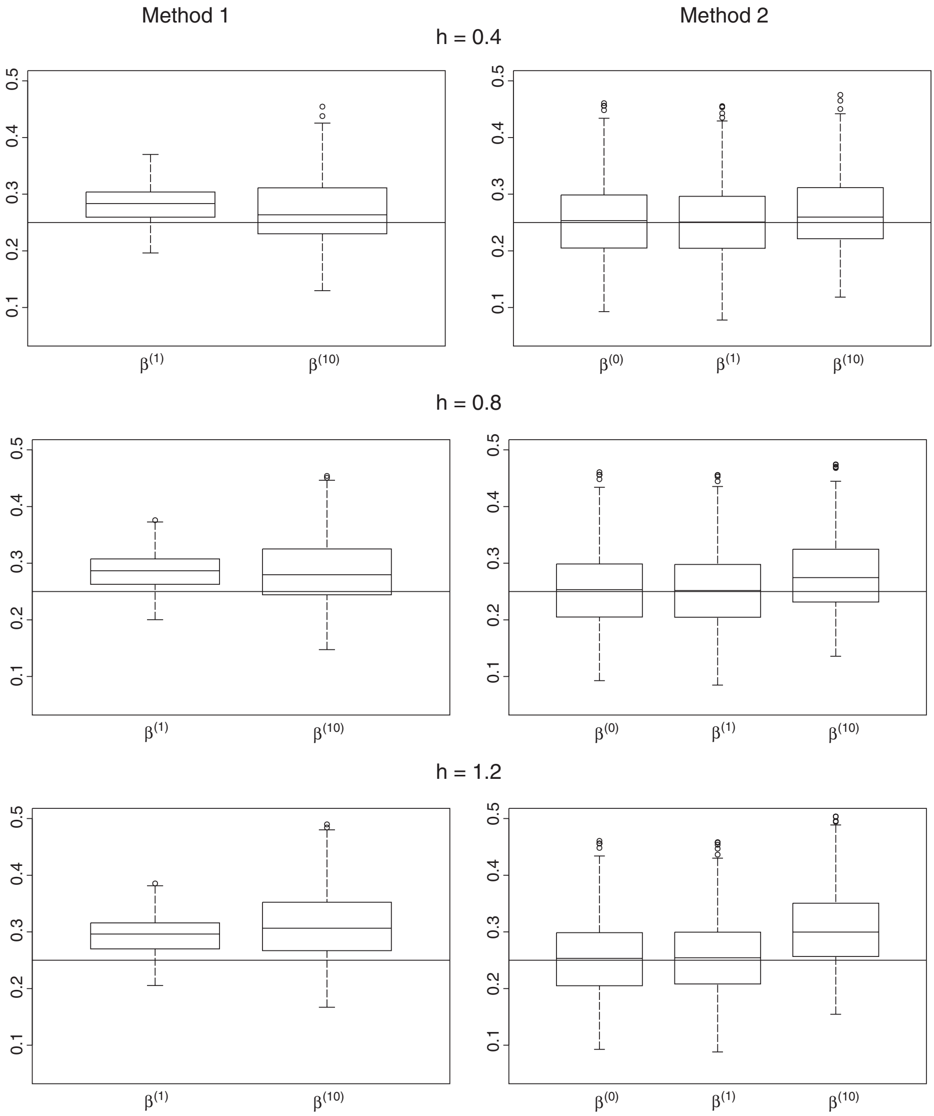

Figure 4. Boxplot of the estimates of the autoregression parameter when $g_o(z) = 0.5z$.

the moving average parameters (see Table 3). It is woth noticing that for this model, the first step estimator $g^{(1)}$ using Method 1 gives smaller mean square errors than those obtained with Method 2. Moreover, the mean square error of $g^{(1)}$, when using Method 1, is quite closer to the optimal one $M(g^{(0)}, g_o)$ that corresponds to an ARMA(2,2) fit.

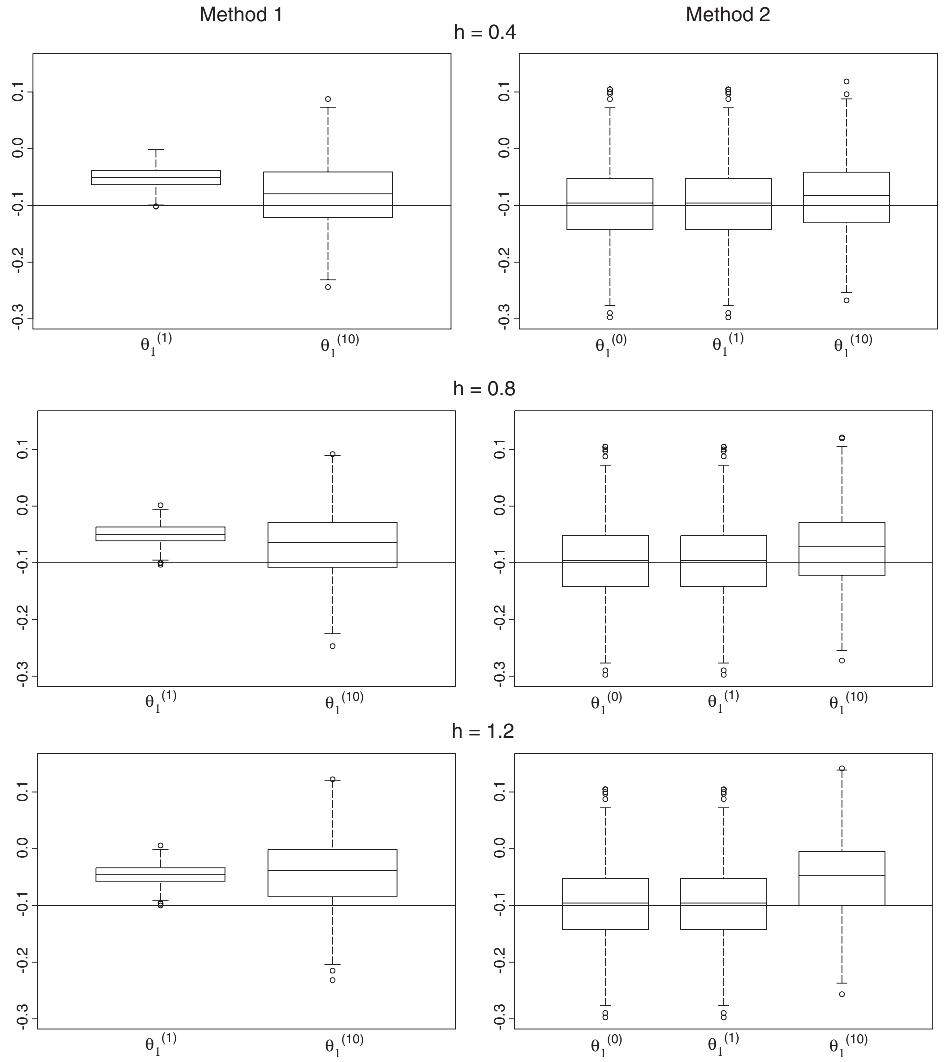

Figure 5. Boxplot of the estimates of the moving average parameter θ_1 when $g_o(z) = 0.5z$.

The good performance of the first step estimates is quite surprising since, even when there is nonlinearity and dependence, with Method 2 iterating is almost as good as a one-step method with respect to the estimator of g. It also shows that Method 2 gives better results than Method 1. In conclusion, our recommendation is to use a ten-step iteration procedure combined with initial estimators computed assuming no MA structure.

5.2. *Bandwidth selection*

As in any situation in which we deal with nonparametric estimators, we have to face the decision of how much smoothing is necessary. In fact, in our simulation study we have computed the estimators using Methods 1 and 2 for different values of the bandwidth parameter and the results show that the estimators may be affected by the choice of the bandwidth.

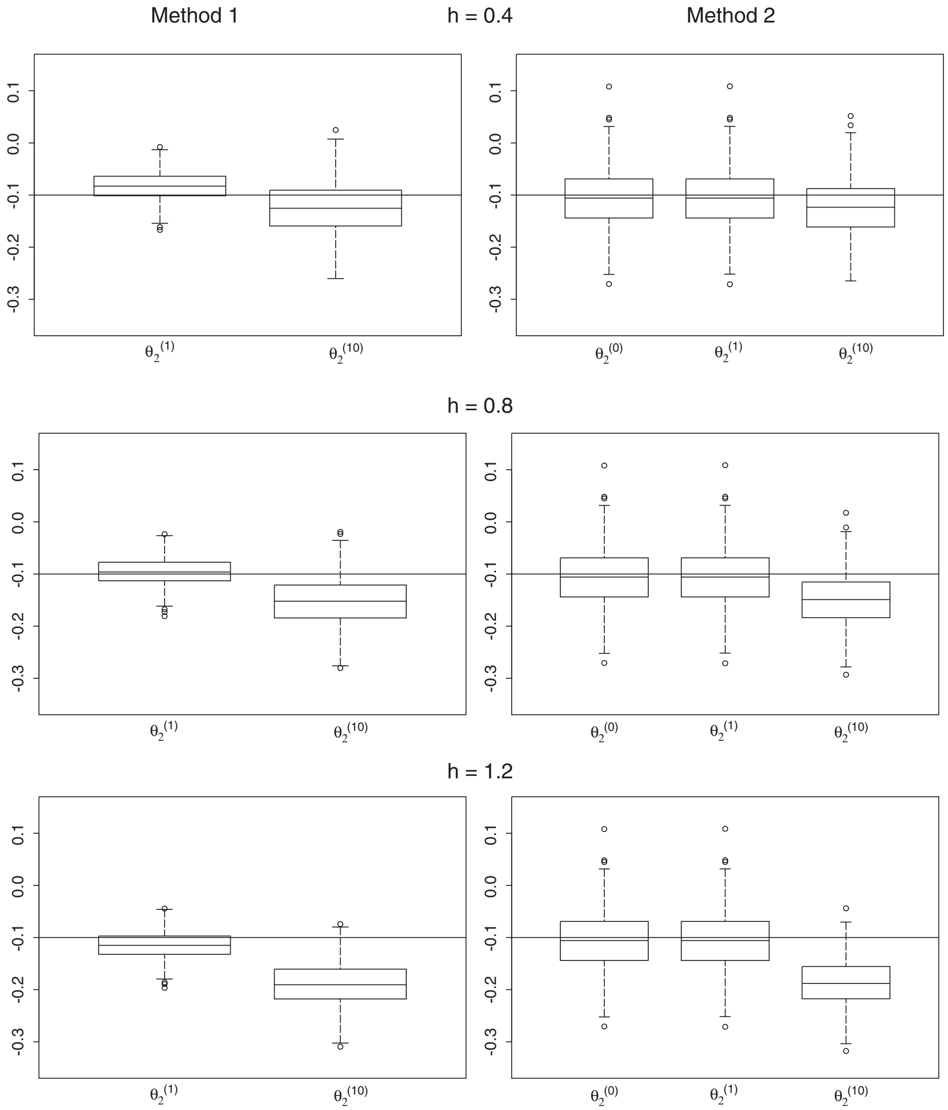

Figure 6. Boxplot of the estimates of the moving average parameter θ_2 when $g_o(z) = 0.5z$.

Table 3. Mean square error for the autoregressive and moving average parameter estimators for $g_o(z) = 0.5z$.

	$\beta^{(10)}$		$\theta_1^{(10)}$		$\theta_2^{(10)}$	
h	Method 1	Method 2	Method 1	Method 2	Method 1	Method 2
0.4	0.0035	0.0042	0.0034	0.0043	0.0029	0.0033
0.8	0.0045	0.0050	0.0043	0.0051	0.0047	0.0048
1.2	0.0072	0.0073	0.0069	0.0071	0.0099	0.0096

In order to select the smoothing parameter h, we consider the cross-validation criterion described in Section 3.4, in which the estimates of g_o, β_o, $\boldsymbol{\theta}_o = (\theta_1, \theta_2)^{\mathrm{T}}$ are computed just using the first half of the sample (i.e. $\alpha = 0.5$) and the smoothing parameter is chosen as the value that minimises a global error computed from the second half of the sample. We have carried

Table 4. Mean of $M(g^{(0)}, g_o)$, $M(g^{(1)}, g_o)$ and $M(g^{(10)}, g_o)$ for $g_o(z) = 0.5z$.

	$h = 0.4$		$h = 0.8$		$h = 1.2$	
	Method 1	Method 2	Method 1	Method 2	Method 1	Method 2
$M(g^{(0)}, g_o)$		0.00874		0.00874		0.00874
$M(g^{(1)}, g_o)$	0.01859	0.02459	0.01218	0.01611	0.01615	0.01524
$M(g^{(10)}, g_o)$	0.02429	0.02536	0.02147	0.02173	0.03521	0.03398

out a simulation study to assess the performance of the method. As in the previous section, we performed 500 replications generating a series of size $T = 1000$ following model (12), where $\beta_o = 0.25$, $\theta_1 = \theta_2 = -0.1$ and $g_o(z) = 0.125\pi \sin(\pi z) + 0.25z$.

For each bandwidth h, we computed the estimates $\hat{\beta}^{(h)}$, $\hat{g}^{(h)}$, $\hat{\theta}_1^{(h)}$ and $\hat{\theta}_2^{(h)}$ using only $y_1, \ldots, y_{[T/2]}$ (i.e. the first half of the sample). We define the cross-validation criterion

$$C(h) = \frac{1}{[T/2]} \sum_{t=[T/2]+1}^{T} (y_t - \hat{y}_t^{(h)})^2,$$

where the predicted observation at time t, $\hat{y}_t^{(h)} = \hat{\beta}^{(h)} y_{t-1} + \hat{g}^{(h)}(y_{t-2}) - \hat{\theta}_2^{(h)} \hat{u}_{t-2}^{(h)}$, is calculated using $\hat{\beta}^{(h)}$, $\hat{g}^{(h)}$, $\hat{\theta}_1^{(h)}$, $\hat{\theta}_2^{(h)}$ and the second half of the sample, as described in Section 3.4. The data-driven bandwidth selector is obtained as the value minimising $C(h)$ over a grid of 50 points in the interval [0.01, 1.5]. A refined search was performed if the minimum is attained at 1.5. The final estimates of β_o, g_o, θ_1 and θ_2 are computed with the resulting bandwidth. Figure 7 shows the boxplots of the optimal bandwiths using Methods 1 and 2. Figure 8 shows the boxplots of the data-driven estimates of the parameters obtained in the first and last steps of the iterative procedures. As shown in Figure 8, the bandwidth selector tends to choose bandwidths around 0.5 for both methods. Besides, the bandwidth selectors when using Method 2 are slightly more spread than the corresponding ones for Method 1. With respect to the estimation of the finite-dimensional parameters, the data-driven estimates performed better than those with fixed bandwidths. Method 1 shows its advantage when combined with cross-validation both with respect to the estimation of the finite-dimensional parameters and with respect to the estimation of the autoregression function g.

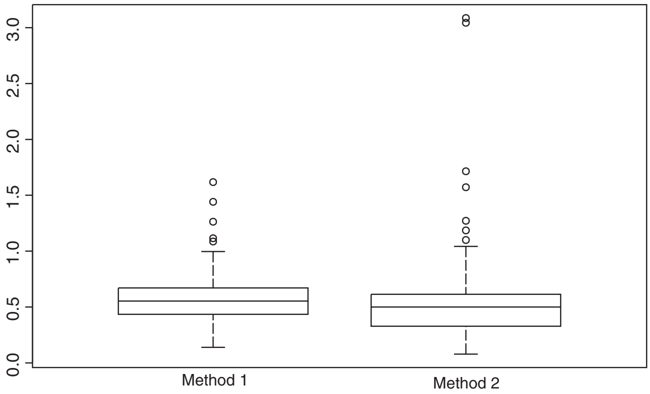

Figure 7. Boxplot of the optimal data-driven bandwidths $\hat{h}$ obtained using Methods 1 and 2.

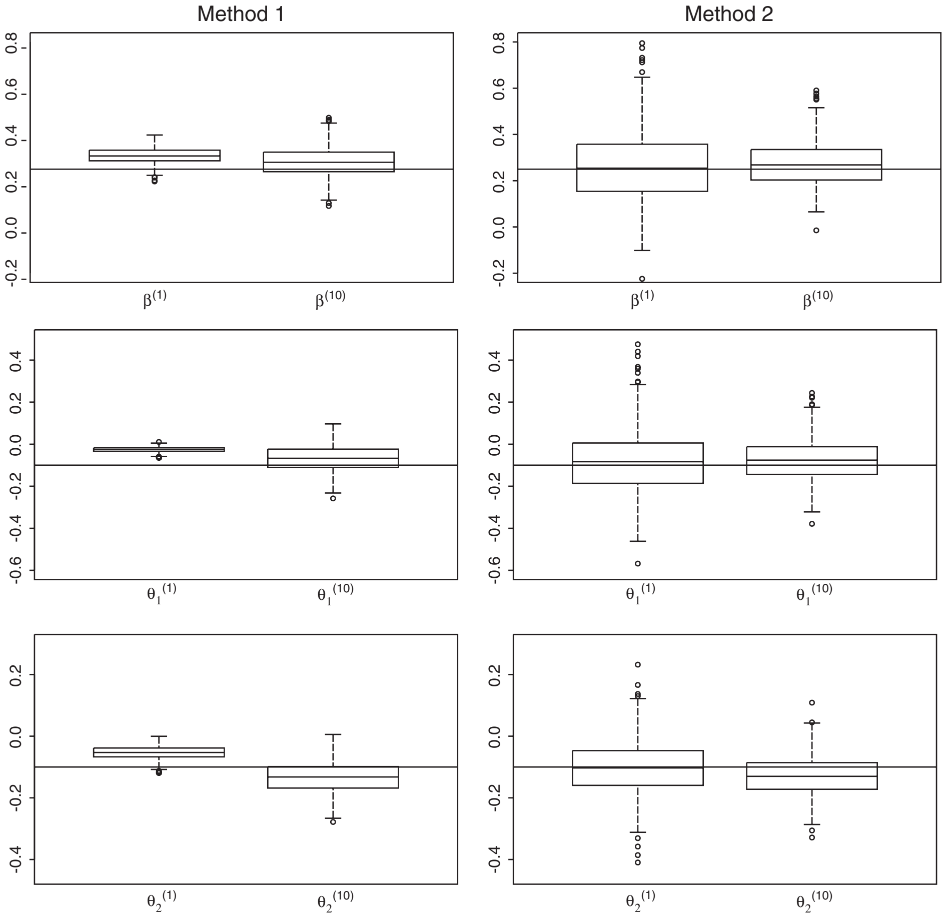

Figure 8. Boxplot of the estimates of the autoregression parameter β_o, the moving average parameters θ_1 and θ_2 when $g_o(z) = 0.125\pi\ \sin(\pi z) + 0.25z$ and the cross-validation bandwidth is used.

5.3. *Test*

We conduct a simulation study in order to assess the performance of the test proposed in Section 4. We consider the simulation scheme given in the previous section and we take $g_o(z) = 0.125\pi\ \sin(\pi z) + 0.25z$ and $g_o(z) = 0.5z$. We perform 5000 replications of the series taking θ_1 and θ_2 in the set $\{-0.1, -0.05, 0\}$. According to the results obtained in the previous section, we select the bandwidth $h_T = 0.4$. We choose the nominal level of the tests as $\alpha = 0.05$.

Table 5. Observed frequencies of rejection under the null hypothesis.

		$g_o(z) = 0.125\pi\ \sin(\pi z) + 0.25z$			$g_o(z) = 0.5z$		
		θ_1			θ_1		
		0	−0.05	−0.10	0	−0.05	−0.10
θ_2	0	0	0	0.0020	0.0030	0.0086	0.0320
	−0.05	0.0030	0.0078	0.0196	0.0410	0.1078	0.2318
	−0.10	0.0338	0.0846	0.1588	0.2822	0.4518	0.6434

For each of the nine combinations, we compute the observed frequency of rejection of the null hypothesis

$$H_0 : \theta_1 = \theta_2 = 0.$$

Table 5 summarises the results obtained for each model. For both models, we can see that the test is conservative. On the other hand, as expected, in both cases the observed frequencies of rejection increase as long as the true parameters become far away from the null hypothesis, showing the power of the test to reject H_0 when the values of the parameters lie in the alternative region.

Acknowledgements

The authors would like to thank an anonymous referee for his valuable comments and suggestions that lead to improve the paper. This research was partially supported by Grants X018 from the Universidad of Buenos Aires, PID 5505 and 112-200801-00216 from CONICET and PICT 21407 and 821 from ANPCYT, Argentina.

References

Anderson, T.W. (1994), *The Statistical Analysis of Time Series*, New York: John Wiley and Sons.

Ango Nze, P. (1998), 'Critères d'Ergodicité Géométrique ou Arithmétique de Modèles linéaires Perturbés à Représentation Markovienne', *Comptes Rendus del'Academic des Sciences, series I (Paris)*, 326, pp. 371–376.

Boente, G., and Fraiman, R. (2002), 'Ergodicity, Geometric Ergodicity and Mixing Conditions for Nonparametric ARMA Processes', *Bulletin of the Brazilian Mathematical Society*, 33, 13–23.

Bosq, D. (1996), *Nonparametric Statistics for Stochastic Processes: Estimation and Prediction* (Vol. 110), Lectures Notes in Statistics, Berlin: Springer-Verlag.

Carbon, M., and Delecroix, M. (1993), 'Nonparametric Forecasting in Time Series, a Computational Point of View', *Applied Stochastic Models and Data Analysis*, 9, 215–229.

Collomb, G. (1985), 'Non Parametric Time Series Analysis and Prediction: Uniform Almost Sure Convergence of the Window and K-NN Autoregression Estimates', *Statistics*, 16, pp. 297–307.

Durbin, J. (1959), 'Efficient Estimation of Parameters in Moving-Average Models', *Biometrika*, 46, pp. 306–316.

Gao, J. (1998), 'Semiparametric Regression Smoothing of Nonlinear Time Series', *Scandinavian Journal of Statistics*, 25, pp. 521–539.

Gao, J. (2007), *Nonlinear Time Series: Semiparametric and Nonparametric Methods*, London: Chapman & Hall/CRC.

Gao, J., and Yee, T. (2000), 'Adaptive Estimation in Partly Linear Autoregressive Models', *The Canadian Journal of Statistics*, 28, 571–586.

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

Hall, P., Lahiri, S.N., and Truong, Y.K. (1995), 'On Bandwidth Choice for Density Estimation With Dependent Data', *Annals of Statistics*, 23, pp. 2241–2263.

Härdle, W., and Vieu, P. (1992), 'Kernel Regression Smoothing of Time Series', *Journal of Time Series Analysis*, 13 pp. 209–232.

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

Hart, J.D. (1994), 'Automated Kernel Smoothing of Dependent Data by Using Time Series Cross-validation', *Journal of the Royal Statistical Society, Series B*, 56, pp. 529–542.

Hart, J.D. (1996), 'Some Automated Methods of Smoothing Time-Dependent Data', *Journal of Nonparametric Statistics*, 6, pp. 115–142.

Hart, J.D., and Wehrly, T.E. (1986), 'Kernel Regression Estimation Using Repeated Measurements Data', *Journal of American Statistical Association*, 81, pp. 1080–1088.

Hart, J.D., and Vieu, P. (1990), 'Data-driven Bandwidth Choice for Density Estimation Based on Dependent Data', *Annals of Statistics*, 18, pp. 873–890.

Masry, E., and Tjøstheim, D. (1995), 'Nonparametric Estimation and Identification of Nonlinear ARCH Time Series', *Econometric Theory*, 11, pp. 258–289.

Mokkadem, A. (1987), 'Sur un Modèle Autorégressif Non linéaire, Ergodicité et Ergodicité Géométrique', *Journal of Time Series Analysis*, 2, pp. 195–204.

Nummelin, E., and Tuominen, P. (1982), 'Geometric Ergodicity of Harris Recurrent Markov Chains With Applications to Renewal Theory', *Stochastics Processes and their Application*, 2, pp. 187–202.

Robinson, P. (1988), 'Root-n-Consistent Semiparametric Regression', *Econometrica*, 56, pp. 931–954.

Rosenblatt, M. (1971), *Markov Processes: Structure and Asymptotic Behaviour*, Berlin: Springer-Verlag.

Tweedie, R.L. (1975), 'Sufficient Conditions for Ergodicity and Recurrence of Markov Chains on a General State Space', *Stochastics Processes and their Application*, 3, pp. 385–403.

Tweedie, R.L. (1976), 'Criteria for Classifying General Markov Chains', *Advances in Applied Probability*, 8, pp. 737–771.

Appendix

Proof of Proposition 2.1 As in Ango Nze (1998), in order to prove **A1** it is enough to show that **A1** holds for any measurable set $A = A_1 \times A_2 \times \cdots \times A_{p_1+p_2}$, with A_i measurables.

To fix ideas, we will begin with the simplest case $p_1 = p_2 = 1$. Let $A = A_1 \times A_2$ be a Borelian set such that $\lambda(A) > 0$ and $\mathcal{K} \subset \mathbb{R}^2$ a compact set. Then, the transition probability satisfies the following

$$P(\mathbf{y}, A) = I_{A_2}(y_1) \int_{A_1} f_{\varepsilon,\mathbf{y}}(v - m_o(\mathbf{y}))\mathrm{d}v.$$

Since there exist bounded sets $B_i \subset A_i$ such that $\lambda(B) > 0$, where $B = B_1 \times B_2$, we get

$$\begin{aligned} P^2(\mathbf{y}, A) &= \int P(\mathbf{v}, A)P(\mathbf{y}, d\mathbf{v}) = \int P((v_1, y_1), A) f_{\varepsilon,\mathbf{y}}(v_1 - m_o(\mathbf{y}))\mathrm{d}v_1 \\ &= \int_{A_2} \left[\int_{A_1} f_{\varepsilon,(v_1,y_1)}(z - m_o(v_1, y_1))\mathrm{d}z \right] f_{\varepsilon,\mathbf{y}}(v_1 - m_o(\mathbf{y}))\mathrm{d}v_1 \\ &\geq \int_{B_2} \left[\int_{B_1} f_{\varepsilon,(v_1,y_1)}(z - m_o(v_1, y_1))\mathrm{d}z \right] f_{\varepsilon,\mathbf{y}}(v_1 - m_o(\mathbf{y}))\mathrm{d}v_1. \end{aligned}$$

Let $C_1 = \bigcup_{\mathbf{y}\in\mathcal{K}}(B_2 - m_o(\mathbf{y})) \subset \overline{\bigcup_{\mathbf{y}\in\mathcal{K}}(B_2 - m_o(\mathbf{y}))} = \mathcal{K}_1$. Note that $\mathcal{K}_1 \subset \mathbb{R}$ is a compact set, since m_o is bounded on $\mathcal{K}$ by **H1**. Similarly, define the compact sets

- $\mathcal{K}_1^\star = \overline{\bigcup_{\mathbf{y}\in\mathcal{K}} \mathcal{K}_1 + m_o(\mathbf{y})} \subset \mathbb{R}$,
- $\mathcal{K}_2^\star = \mathcal{K}_1^\star \times \mathrm{proj}_2(\mathcal{K})$, where $\mathrm{proj}_2(\mathcal{K})$ is the projection over the second component of the set $\mathcal{K}$,
- $\mathcal{K}_1^{\star\star} = \overline{\bigcup_{\mathbf{y}\in\mathcal{K}} \bigcup_{v\in\mathcal{K}_1} B_1 - m_o(v - m_o(\mathbf{y}), y_1)}$.

Therefore,

$$P^2(\mathbf{y}, A) \geq b(\mathcal{K}_1, \mathcal{K})b(\mathcal{K}_1^{\star\star}, \mathcal{K}_2^\star)\lambda(B_1)\lambda(B_2) > 0$$

and **A1** holds with $n_0 = 2$.

We will show that **A3** holds for $n_1 = 2$. Indeed, if $\lambda(A) = 0$ then $P^2(\mathbf{y}, A) = \int_A P(\mathbf{y}, \mathbf{v})\mathrm{d}\mathbf{v} = 0$ for all $\mathbf{y}$. On the other hand, since $f_{\varepsilon,\mathbf{y}}(u) > 0$ from **H2**, if $P^2(\mathbf{y}, A) = 0$ for all $\mathbf{y}$, we have that $\lambda(A) = 0$.

In the general situation, we have that the transition probability satisfies the following:

$$P(\mathbf{y}, A) = \prod_{j=1}^{p_1+p_2-1} I_{A_{j+1}}(y_j) \int_{A_1} f_{\varepsilon,\mathbf{y}}(v - m_o(\mathbf{y}))\mathrm{d}v.$$

Therefore, if $\lambda(A) > 0$, there exist bounded sets $B_i \subset A_i$ such that $\lambda(B) > 0$, where $B = B_1 \times \cdots \times B_{p_1+p_2}$, which entails that for some constant C

$$P^{p_1+p_2}(\mathbf{y}, A) \geq C \prod_{j=1}^{p_1+p_2} \lambda(B_j) > 0$$

and **A1** holds with $n_0 = p_1 + p_2$. Arguing as above, it is easy to see that **A3** holds for $n_1 = p_1 + p_2$. ■

Proof of Proposition 2.3 Since **H1** and **H2** entail **A1**, the process $\{\mathbf{Y}_t\}$ is strongly irreducible (Tweedie 1976). On the other hand, Proposition 2.2, **H1**, **H2** and **H3** imply the ergodicity of $\{\mathbf{Y}_t\}$ and, therefore, the conclusion follows from Tweedie (1976). ■

Proof of Theorem 3.1.1 Using that $g_{\mathbf{b},\boldsymbol{\vartheta}}(y) = \phi_2(y) - \mathbf{b}^{\mathrm{T}}\boldsymbol{\phi}_1(y) + \sum_{j=1}^{q} \vartheta_j \eta_j(y)$ and

$$y_t = \boldsymbol{\beta}_o^{\mathrm{T}}(\mathbf{y}_{t-1} - \boldsymbol{\phi}_1(y_{t-p_1-1})) + \phi_2(y_{t-p_1-1}) + u_t - \sum_{j=1}^{q} \theta_{o,j}(u_{t-j} - \eta_j(y_{t-p_1-1})), \tag{A1}$$

we get that

$$y_t - g_{\mathbf{b},\boldsymbol{\vartheta}}(y_{t-p_1-1}) + \sum_{j=1}^{q} \vartheta_j u_{t-j} - \mathbf{b}^{\mathrm{T}}\mathbf{y}_{t-1} = (\boldsymbol{\beta}_o - \mathbf{b})^{\mathrm{T}}\mathbf{z}_t + u_t + \sum_{j=1}^{q}(\vartheta_j - \theta_{o,j})s_{t,j},$$

where $\mathbf{z}_t = \mathbf{y}_{t-1} - \boldsymbol{\phi}_1(y_{t-p_1-1})$ and $s_{t,j} = u_{t-j} - \eta_j(y_{t-p_1-1})$. Therefore, since $E_F\left(u_t\ s_{t,j}\right) = 0$ for $1 \le j \le q$ and $E_F\ (u_t\ \mathbf{z}_t) = 0$, we obtain that

$$M(\mathbf{b}, \boldsymbol{\vartheta}) = E_F \left[(\boldsymbol{\beta}_o - \mathbf{b})^{\mathrm{T}} \mathbf{z}_t + \sum_{j=1}^{q} (\vartheta_j - \theta_{o,j}) s_{t,j} \right]^2 + E_F u_t^2,$$

which implies that $M(\mathbf{b}, \boldsymbol{\vartheta}) \ge M(\boldsymbol{\beta}_o, \boldsymbol{\theta}_o)$. The equality holds if and only if

$$P\left((\boldsymbol{\beta}_o - \mathbf{b})^{\mathrm{T}} \mathbf{z}_t + \sum_{j=1}^{q} (\vartheta_j - \theta_{o,j}) s_{t,j} = 0 \right) = 1. \tag{A2}$$

If $\mathbf{b} \ne \boldsymbol{\beta}_o$, Equation (A2) holds if and only if $P(\mathbf{d}^{\mathrm{T}} \mathbf{z}_t = \sum_{j=1}^{q} a_j s_{t,j}) = 1$ with $a_j = (\vartheta_j - \theta_{o,j})$, $\mathbf{d} = \boldsymbol{\beta}_o - \mathbf{b}$ and $\|\mathbf{d}\| \ne 0$, which is equivalent to

$$P\left(\mathbf{d}^{\mathrm{T}} \mathbf{y}_{t-1} = h(y_{t-p_1-1}) + \sum_{j=1}^{q} a_j u_{t-j} \right) = 1,$$

where $h(y) = \mathbf{d}^{\mathrm{T}} \boldsymbol{\phi}_1(y) - \sum_{j=1}^{q} a_j \eta_j(y)$, which contradicts assumption (a). Thus, $\mathbf{b} = \boldsymbol{\beta}_o$ and so Equation (A2) can be written as

$$P\left(\sum_{j=1}^{q} (\vartheta_j - \theta_{o,j}) s_{t,j} = 0 \right) = 1. \tag{A3}$$

If $\vartheta_1 \ne \theta_{o,1}$, dividing by $\vartheta_1 - \theta_{o,1}$, we get

$$P\left(u_{t-1} = \eta_1(y_{t-p_1-1}) + \sum_{j=p_1+1}^{q} a_j s_{t,j} \right) = 1$$

for some constants a_j. Using that u_{t-1} is independent of $\{y_{t-p_1-1}, u_{t-j}, j \ge 2\}$, we get again a contradiction with assumption (b) which allows to conclude that $\vartheta_1 = \theta_{o,1}$. The proof follows iteratively using assumption (b). ■

Proof of Theorem 3.1.2 Denote $\mathbf{z}_t = \mathbf{y}_{t-1} - \boldsymbol{\phi}_1(y_{t-p_1-1})$, $s_{t,j} = u_{t-j} - \eta_j(y_{t-p_1-1})$ and $\mathbf{s}_t = (s_{t,1}, \ldots, s_{t,q})^{\mathrm{T}}$. Using Equation (A1), since $E_F(u_t(u_{t-j} - \eta_j(y_{t-p_1-1}))) = 0$ for $1 \le j \le q$ and $E_F(u_t(\mathbf{y}_{t-1} - \boldsymbol{\phi}_1(y_{t-p_1-1}))) = 0$, we get that Equation (8) is equivalent to the linear system

$$L_0(\mathbf{b}, \boldsymbol{\theta}) = (\boldsymbol{\beta}_o - \mathbf{b})^{\mathrm{T}} E_F \mathbf{z}_t \mathbf{z}_t^{\mathrm{T}} + \sum_{j=1}^{q} (\theta_{o,j} - \theta_j) E_F s_{t,j} \mathbf{z}_t = 0,$$

$$L_\ell(\mathbf{b}, \boldsymbol{\theta}) = (\boldsymbol{\beta}_o - \mathbf{b})^{\mathrm{T}} E_F \mathbf{z}_t\, s_{t,\ell} + \sum_{j=1}^{q} (\theta_{o,j} - \theta_j) E_F s_{t,j} s_{t,\ell} = 0.$$

This system of equations can be written as $\mathbf{Cd} = 0$, where $\mathbf{d} = (\mathbf{b}^{\mathrm{T}}, \boldsymbol{\theta}^{\mathrm{T}})^{\mathrm{T}}$ and $\mathbf{C}$ is the covariance matrix of $(\mathbf{z}_t^{\mathrm{T}}, \mathbf{s}_t^{\mathrm{T}})$. Since $L_0(\boldsymbol{\beta}_o, \boldsymbol{\theta}_{o,j}) = 0$ and $L_\ell(\boldsymbol{\beta}_o, \boldsymbol{\theta}_{o,j}) = 0$, it will be enough to show that $\mathbf{C}$ is non-singular, which follows from the required assumptions. ■

Proof of Theorem 3.2.1.(i) Note that

$$y_t - \hat{g}_{\mathbf{b},\boldsymbol{\vartheta}}(y_{t-p_1-1}) + \sum_{j=1}^{q} \vartheta_j u_{t-j} - \mathbf{b}^{\mathrm{T}}\, \mathbf{y}_{t-1}$$

$$= (y_t - \hat{\phi}_2(_{t-p_1-1})) - \mathbf{b}^{\mathrm{T}}\, (\mathbf{y}_{t-1} - \hat{\boldsymbol{\phi}}_1(_{t-p_1-1})) + \sum_{j=1}^{q} \vartheta_j (u_{t-j} - \hat{\eta}_j(_{t-p_1-1})).$$

So, if we denote $\hat{r}_t = y_t - \hat{\phi}_2(y_{t-p_1-1})$, $\hat{\mathbf{z}}_t = \mathbf{y}_{t-1} - \hat{\boldsymbol{\phi}}_1(y_{t-p_1-1})$, $\hat{s}_{t,j} = u_{t-j} - \hat{\eta}_j(_{t-p_1-1})$ and $\hat{\mathbf{s}}_t = (\hat{s}_{t,1}, \ldots, \hat{s}_{t,q})^{\mathrm{T}}$, $\hat{\mathbf{x}}_t = (\hat{\mathbf{z}}_t^{\mathrm{T}}, \hat{\mathbf{s}}_t^{\mathrm{T}})^{\mathrm{T}}$, $\boldsymbol{\Delta} = (\mathbf{b}^{\mathrm{T}}, \boldsymbol{\vartheta}^{\mathrm{T}})^{\mathrm{T}}$, $\hat{\boldsymbol{\Delta}} = (\hat{\boldsymbol{\beta}}^{\mathrm{T}}, \hat{\boldsymbol{\theta}}^{\mathrm{T}})^{\mathrm{T}}$, we obtain that

$$M_n(\mathbf{b}, \boldsymbol{\vartheta}) = \frac{1}{T} \sum_{t=p_1+2}^{T} \left(\hat{r}_t - \mathbf{b}^{\mathrm{T}} \hat{\mathbf{z}}_t + \boldsymbol{\vartheta}^{\mathrm{T}} \hat{\mathbf{s}}_t \right)^2 = \frac{1}{T} \sum_{t=p_1+2}^{T} \left(\hat{r}_t - \boldsymbol{\Delta}^{\mathrm{T}} \hat{\mathbf{x}}_t \right)^2$$

and $\hat{\boldsymbol{\Delta}} = \operatorname{argmin}_{\boldsymbol{\Delta}\in\mathbb{R}^{p_1}\times\Theta} M_n(\mathbf{b}, \boldsymbol{\vartheta})$, which implies that

$$\frac{1}{T}\sum_{t=p_1+2}^{T} \hat{\mathbf{x}}_t\hat{\mathbf{x}}_t^{\mathrm{T}}\,\hat{\boldsymbol{\Delta}} = \frac{1}{T}\sum_{t=p_1+2}^{T} \hat{r}_t\hat{\mathbf{x}}_t.$$

Now, recall that $r_t = y_t - \phi_2(y_{t-p_1-1})$, $\mathbf{z}_t = \mathbf{y}_{t-1} - \boldsymbol{\phi}_1(y_{t-p_1-1})$, $s_{t,j} = u_{t-j} - \eta_j(y_{t-p_1-1})$ and $\mathbf{s}_t = (s_{t,1},\ldots,s_{t,q})^{\mathrm{T}}$ and denote by $\mathbf{x}_t = (\mathbf{z}_t^{\mathrm{T}}, \mathbf{s}_t^{\mathrm{T}})^{\mathrm{T}}$. Using that from (d), $\mathbf{C} = E(\mathbf{x}_t\mathbf{x}_t^{\mathrm{T}})$ is non-singular and that $\boldsymbol{\Delta}_o = (\boldsymbol{\beta}_o^{\mathrm{T}}, \boldsymbol{\theta}_o^{\mathrm{T}})^{\mathrm{T}}$ solves Equation (8), we get that $\boldsymbol{\Delta}_o = \mathbf{C}^{-1}E(r_t\mathbf{x}_t)$ and so it is enough to show that

$$\frac{1}{T}\sum_{t=p_1+2}^{T}\left(r_t\mathbf{x}_t - \hat{r}_t\hat{\mathbf{x}}_t\right) \xrightarrow{p} 0, \tag{A4}$$

$$\frac{1}{T}\sum_{t=p_1+2}^{T}\left(\mathbf{x}_t\mathbf{x}_t^{\mathrm{T}} - \hat{\mathbf{x}}_t\hat{\mathbf{x}}_t^{\mathrm{T}}\right) \xrightarrow{p} 0. \tag{A5}$$

Note that $\sum_{t=p_1+2}^{T}\left(r_t\mathbf{x}_t - \hat{r}_t\hat{\mathbf{x}}_t\right)/T = S_{1,T} + S_{2,T}$, where

$$S_{1,T} = \frac{1}{T}\sum_{t=p_1+2}^{T} r_t\left(\mathbf{x}_t - \hat{\mathbf{x}}_t\right), \tag{A6}$$

$$S_{2,T} = \frac{1}{T}\sum_{t=p_1+2}^{T}\left(r_t - \hat{r}_t\right)\hat{\mathbf{x}}_t. \tag{A7}$$

Using Cauchy–Schwartz inequality, we get that

$$\begin{aligned}\|S_{1,T}\|^2 &\le \frac{1}{T}\sum_{t=p_1+2}^{T} r_t^2\,\frac{1}{T}\sum_{t=p_1+2}^{T}\left\|\mathbf{x}_t - \hat{\mathbf{x}}_t\right\|^2\\ &\le \frac{1}{T}\sum_{t=p_1+2}^{T} r_t^2\left(\frac{1}{T}\sum_{t=p_1+2}^{T}\left\|\hat{\boldsymbol{\phi}}_1(y_{t-p_1-1}) - \boldsymbol{\phi}_1(y_{t-p_1-1})\right\|^2 + \frac{1}{T}\sum_{j=1}^{q}\sum_{t=p_1+2}^{T}\left(\hat{\eta}_j(y_{t-p_1-1}) - \eta_j(y_{t-p_1-1})\right)^2\right)\end{aligned} \tag{A8}$$

and

$$\begin{aligned}\|S_{2,T}\|^2 &\le \frac{1}{T}\sum_{t=p_1+2}^{T}\left(r_t - \hat{r}_t\right)^2\frac{1}{T}\sum_{t=p_1+2}^{T}\|\hat{\mathbf{x}}_t\|^2\\ &\le \frac{1}{T}\sum_{t=p_1+2}^{T}\left(\hat{\phi}_2(y_{t-p_1-1}) - \phi_2(y_{t-p_1-1})\right)^2\left\{\frac{1}{T}\sum_{t=p_1+2}^{T}\|\mathbf{x}_t\|^2 + \frac{1}{T}\sum_{t=p_1+2}^{T}\|\mathbf{x}_t - \hat{\mathbf{x}}_t\|^2\right\},\end{aligned} \tag{A9}$$

which together with assumptions (a) to (c) concludes the proof of Equation (A4). Similar arguments lead to Equation (A5), concluding the proof of (i). The proof of (ii) is immediate. ■

Proof of Theorem 3.2.2.(i) As in Theorem 3.2.1, if $\hat{\boldsymbol{\Delta}} = (\hat{\boldsymbol{\beta}}^{\mathrm{T}}, \hat{\boldsymbol{\theta}}^{\mathrm{T}})^{\mathrm{T}}$, we have that

$$\frac{1}{T}\sum_{t=p_1+2}^{T} \hat{\mathbf{x}}_t\hat{\mathbf{x}}_t^{\mathrm{T}}\,\hat{\boldsymbol{\Delta}} = \frac{1}{T}\sum_{t=p_1+2}^{T} \hat{r}_t\hat{\mathbf{x}}_t.$$

In the proof of Theorem 3.2.1, we have shown that $\sum_{t=p_1+2}^{T}\hat{\mathbf{x}}_t\hat{\mathbf{x}}_t^{\mathrm{T}}/T \xrightarrow{p} \mathbf{C}$, thus it remains to obtain the asymptotic distribution of

$$\sqrt{T}\hat{S}_T = \frac{1}{\sqrt{T}}\sum_{t=p_1+2}^{T}\left(\hat{r}_t\hat{\mathbf{x}}_t - E(r_t\mathbf{x}_t)\right).$$

Note that $\hat{S}_T = S_T - (S_{1,T} + S_{2,T})$, where $S_{1,T}$ and $S_{2,T}$ are defined in Equations (A6) and (A7), respectively, and $S_T = \sum_{t=p_1+2}^{T}(r_t\mathbf{x}_t - E(r_t\mathbf{x}_t))/T$. Using the bounds given in Equations (A8) and (A9) together with assumptions (a) to (c), it follows immediately that $\|S_{1,T}\| + \|S_{2,T}\| \xrightarrow{p} 0$. On the other hand, using Theorem 1.7 in Bosq (1996), we obtain that $\sqrt{T} S_T \xrightarrow{\mathcal{D}} N(0, \mathbf{D})$, which concludes the proof. ■

Proof of Lemma 4.1 Denote I the identity operator. Note that since

$$I = \left(1 - \sum_{r=1}^{q} \theta_r B^r\right)\left(\sum_{j=0}^{\infty} \gamma_j B^j\right) = \sum_{j=0}^{\infty} \gamma_j B^j - \theta_1 \sum_{j=0}^{\infty} \gamma_j B^{j+1} - \cdots - \theta_q \sum_{j=0}^{\infty} \gamma_j B^{j+q}$$

$$\begin{aligned} I &= \gamma_0 + (\gamma_1 - \theta_1\gamma_0)B + (\gamma_2 - \theta_1\gamma_1 - \theta_2\gamma_0)B^2 + \cdots + (\gamma_q - \theta_1\gamma_{q-1} - \theta_2\gamma_{q-2} - \cdots - \theta_q\gamma_0)B^q \\ &\quad + \sum_{j=q+1}^{\infty} (\gamma_j - \theta_1\gamma_{j-1} - \theta_2\gamma_{j-2} - \cdots - \theta_q\gamma_{j-q})B^j, \end{aligned}$$

we have that $\gamma_0 = 1$ and $\gamma_j = \sum_{r=1}^{h_j} \theta_r \gamma_{j-r}$ with $h_j = \min(j, q)$. Therefore, $\theta_1 = \theta_2 \cdots = \theta_q = 0$ is equivalent to $\gamma_1 = \gamma_2 \cdots = \gamma_q = 0$. ■

Proof of Theorem 4.1. Denoting $\hat{c}_{j-r} = \sum_{t=p_1+2}^{T} \hat{v}_{t-j}\hat{v}_{t-r}/T$, we have that $\hat{\boldsymbol{\gamma}}$ satisfies $\hat{c}_j + \sum_{r=1}^{N} \hat{\gamma}_r \hat{c}_{j-r} = 0$ and so $\sum_{r=1}^{N} \hat{\gamma}_r \hat{c}_{j-r} = -\hat{c}_j$, $j = 1, \ldots, N$.

In order to derive the asymptotic distribution under the null hypothesis, we will use the relationship among $\hat{\boldsymbol{\gamma}}$ and the estimated covariances $\hat{c}_{j-r}$. Since $y_t = \boldsymbol{\beta}_o^{\mathrm{T}} \mathbf{y}_{t-1} + g_o(y_{t-p_1-1}) + u_t$, under H_0, we have that $\hat{v}_t = (\boldsymbol{\beta}_o - \hat{\boldsymbol{\beta}}^{(0)})^{\mathrm{T}} \mathbf{y}_{t-1} + g_o(y_{t-p_1-1}) - \hat{g}^{(0)}(y_{t-p_1-1}) + u_t$. Let $\Delta_{\boldsymbol{\beta}} = (\boldsymbol{\beta}_o - \hat{\boldsymbol{\beta}}^{(0)})$ and $\Delta_g(y) = g_o(y) - \hat{g}^{(0)}(y)$, then we get that

$$\begin{aligned} \hat{v}_{t-j}\hat{v}_{t-r} &= \Delta_{\boldsymbol{\beta}}^{\mathrm{T}} \left\{\mathbf{y}_{t-j-1}u_{t-r} + \mathbf{y}_{t-r-1}u_{t-j}\right\} + \Delta_{\boldsymbol{\beta}}^{\mathrm{T}} \mathbf{y}_{t-j-1}\mathbf{y}_{t-r-1}^{\mathrm{T}} \Delta_{\boldsymbol{\beta}} \\ &\quad + \Delta_{\boldsymbol{\beta}}^{\mathrm{T}} \left\{\mathbf{y}_{t-j-1}\Delta_g(y_{t-r-p_1-1}) + \Delta_g(y_{t-j-p_1-1})\mathbf{y}_{t-r-1}\right\} + \Delta_g(y_{t-r-p_1-1})\Delta_g(y_{t-j-p_1-1}) \\ &\quad + \Delta_g(y_{t-j-p_1-1})u_{t-r} + \Delta_g(y_{t-r-p_1-1})u_{t-j} + u_{t-r}u_{t-j} \end{aligned}$$

and so we can write $\hat{c}_{j-r} = \sum_{\ell=1}^{7} S_\ell$, corresponding to the seven terms of the above expression for $\hat{v}_{t-j}\hat{v}_{t-r}$.

Using analogous arguments to those considered in Lemma 6.6.7 of Härdle *et al.* (2000), it is easy to see that $\sqrt{T}\ S_j \xrightarrow{p} 0$, $3 \le j \le 6$. On the other hand, since $\sqrt{T}(\boldsymbol{\beta}_o - \hat{\boldsymbol{\beta}}^{(0)}) = O_p(1)$ and u_t is independent of $\{y_{t-j}\}_{j \ge 1}$, we get that $\sqrt{T}(S_1 + S_2) \xrightarrow{p} 0$. Thus, the asymptotic behaviour of $\hat{c}_{j-r}$ is that of $c_{j-r} = \sum_{t=p_1+2}^{T} u_{t-r}u_{t-j}/T$.

Besides, $\hat{\boldsymbol{\gamma}} = (\hat{\gamma}_1, \ldots, \hat{\gamma}_N)^{\mathrm{T}}$ solves $\hat{\mathbf{C}}\hat{\boldsymbol{\gamma}} = -\hat{\mathbf{c}}$, where

$$\hat{\mathbf{C}} = \begin{pmatrix} \hat{c}_0 & \hat{c}_1 & \cdots & \hat{c}_{N-1} \\ \hat{c}_1 & \hat{c}_2 & \cdots & \hat{c}_{N-2} \\ \cdots & \cdots & \cdots & \cdots \\ \cdots & \cdots & \cdots & \cdots \end{pmatrix} \xrightarrow{p} \sigma^2 \mathbf{I}$$

with $\sigma^2 = Var(u_t)$ and $\hat{\mathbf{c}} = (\hat{c}_1, \ldots, \hat{c}_N)^{\mathrm{T}}$. As mentioned above, the asymptotic behaviour of $\hat{\mathbf{c}}$ coincides with that of $\mathbf{c} = (c_1, \ldots, c_N)^{\mathrm{T}}$ which entails

$$\begin{aligned} \sqrt{T}\hat{\boldsymbol{\gamma}} &= -\hat{\mathbf{C}}^{-1}\hat{\mathbf{c}}\sqrt{T} \\ &= -\hat{\mathbf{C}}^{-1}\mathbf{c}\sqrt{T} + o_p(1) \\ &= -\hat{\mathbf{C}}^{-1}\hat{\mathbf{D}}\sqrt{T}\hat{\boldsymbol{\rho}} + o_p(1), \end{aligned}$$

where $\hat{\mathbf{D}} = \operatorname{diag}(c_0, \ldots, c_0) = c_0\mathbf{I}$ and $\hat{\boldsymbol{\rho}} = (\hat{\rho}_1, \ldots, \hat{\rho}_N)^{\mathrm{T}}$ is the vector of estimated correlations. From Theorem 5.7.1 of Anderson (1994), we have that $\sqrt{T}\hat{\boldsymbol{\rho}} \xrightarrow{\mathcal{D}} N(\mathbf{0}, \mathbf{I})$ and so Equation (11) follows. ■