Sociedad de Estadística e Investigación Operativa
Test (0) *Vol.* **0**, *No.* 0, pp. 111–999

Bootstrap Tests for Nonparametric Comparison of Regression Curves with Dependent Errors

J. M. Vilar-Fernández* **and J. A. Vilar-Fernández**
Departamento de Matemáticas
Universidad de A Coruña, Spain
W. González-Manteiga
Departamento de Estadística e I.O.
Universidad de Santiago de Compostela, Spain

Abstract

In this paper the problem of testing the equality of regression curves with dependent data is studied. Several methods based on nonparametric estimators of the regression function are described. In this setting, the distribution of the test statistic is frequently unknown or difficult to compute, so an approximate test based on the asymptotic distribution of the statistic can be considered. Nevertheless, the asymptotic properties of the methods proposed in this work have been obtained under independence of the observations and just one of these methods was studied in a context of dependence (Vilar-Fernández and González-Manteiga, 2003). In addition, the distribution of these test statistics converges to the limit distribution with convergence rates usually rather slow, so that the approximations obtained for reasonable sample sizes are not satisfactory. For these reasons, many authors have suggested the use of bootstrap algorithms as an alternative approach.
Our main concern is to compare the behavior of three bootstrap procedures that take into account the dependence assumption of the observations when they are used to approximate the distribution of the test statistics considered. A broad simulation study is carried out to observe the finite sample performance of the analyzed bootstrap tests.

Key Words: Hypothesis testing, regression models, nonparametric estimators, dependent data.

AMS subject classification: 62G08, 62G09, 62G10, 62M10.

*Correspondence to: Juan M. Vilar Fernández. Facultad de Informática, Campus de Elviña, 15071, A Coruña. E-mail: eijvilar@udc.es

Research of the authors was supported by the DGICYT Spanish Grant BFM 2002-002665 and BFM2002-03213 (European FEDER support included) and XUGA Grants PGIDT03PXIC10505PN and PGIDT03PXIC20702PN

Received: July 2004; Accepted: February 2005

1 Introduction

The comparison of several regression curves is an important problem of statistical inference. In many cases of practical interest, the objective consists in comparing regression functions of a response variable Y observed in two or more groups on an explanatory variable which is an adjustable parameter, for instance, time. So, if $\{(x_{l,t}, Y_{l,t}) : t = 1, \ldots, n_l;\ l = 1, \ldots, k\}$ denotes the initial sample of observed data, then they are assumed to satisfy the following regression models

$$Y_{l,t} = m_l\left(x_{l,t}\right) + \varepsilon_{l,t}, \quad l = 1, \ldots, k \text{ and } t = 1, \ldots, n_l, \tag{1.1}$$

where $\{\varepsilon_{l,t}\}_{t=1}^{n_l}$ are random errors with distribution function G_l, mean zero and finite variance σ_l^2. The design points $\{x_{l,t}\}$ are fixed and usually rescaled into the unit interval, so $0 \leq x_{l,1} < x_{l,2} < \ldots < x_{l,n_l} \leq 1$.

In this context, we are interested in the following hypothesis test

$$H_0 :\ m_1 = \ldots = m_k \quad \text{versus} \quad H_1 : \text{Exists } (l, j) \text{ such as } m_l \neq m_j. \tag{1.2}$$

In this work nonparametric procedures to test (1.2) are considered. Indeed, classical F-test performs well to test (1.2) when the regression curves m_l follow specific parametric models. For instance, if m_l is modelled according to a generalized linear model, $m(x) = A^t(x)\theta$ and the errors in (1.1) are assumed to be gaussian, then the power of the F-test is higher than those of the nonparametric tests proposed in this paper. However, the nonparametric tests provide a more flexible tool since they only require very general regularity conditions on the regression curves in study. In fact, the problem of testing the equality of k regression functions by using nonparametric techniques has been broadly studied in recent statistics literature. Some relevant papers are Härdle and Marron (1990), King et al. (1991), Hall and Hart (1990), Kulasekera (1995), and Kulasekera and Wang (1997, 1998), among others. Most of these works focus on the case of equal design points for every group and homocedastic errors. In a recent paper of Dette and Neumeyer (2001) the general case of testing the equality of $k \geq 2$ mean functions is studied and the proposed methodology is applicable to the case of different design points in each sample and of heterocedastic errors.

The construction of nonparametric estimators of regression functions requires the previous selection of a smoothing parameter. Some authors have

avoided the selection of the smoothing parameter by using empirical processes, see Delgado (1993), Scheike (2000) and Neumeyer and Dette (2003). Munk and Dette (1998) consider another way of avoiding the bandwidth selection and they propose to directly estimate the difference between the mean curves without using prior nonparametric estimators. Other recent works in this field are those of Hall et al. (1997) and Koul and Schick (1997).

Under independence of the observations, it is well known that the tests based on kernel smoothing methods detect alternatives converging to the null hypothesis at a rate $\left(n\sqrt{h}\right)^{-1/2}$, where h is the bandwidth. On the other hand, the procedures based on empirical processes can distinguish between distant $n^{-1/2}$ regression functions, but the asymptotic results of procedures of this second group are more complex and more linked to Gaussian processes than to the asymptotic normal distribution, typical of tests of the first group. In any case the sampling distribution of the test statistics for checking the equality of regression functions is very difficult to obtain and only an asymptotic approximation is available. Besides, this asymptotic approximation depends on unknown characteristics of the population and the rate of convergence is rather slow. For these reasons, several authors (Hall and Hart, 1990; Neumeyer and Dette, 2003) have proposed resampling procedures to obtain the sampling distribution of the test statistics.

In this paper we consider two regression models in fixed design given by

$$Y_{l,t} = m_l\left(x_t\right) + \varepsilon_{l,t},\ \ l = 1, 2 \text{ and } t = 1, \ldots, n_l, \tag{1.3}$$

with equal design points, that is $x_t = x_{1,t} = x_{2,t}$, for all t. In addition, the processes of random errors $\{\varepsilon_{l,t}\}_{t=1}^{n_l}$, $l = 1, 2$, are assumed to be independent among themselves and to have an ARMA type dependence structure. These regression models often arise by analyzing economical data samples, growth curves, and in general, whenever the observations are sequentially gathered in time. So, essentially, we wish to test the equality of tendencies of two time series. In such a problem it is very important to take into account the existence of correlation among the errors. Ignoring this fact affects the power of the equality test used.

Vilar-Fernández and González-Manteiga (2003) studied the problem of checking the equality of k regression functions with dependent errors in a general context. They used as test statistic a functional distance between nonparametric estimators of the regression functions and obtained

its asymptotic normality. We do not know other works treating this problem under dependence conditions.

The problem of testing the equality of two regression curves under the dependence assumption of the observations is therefore the main concern of this work. Several methods using nonparametric regression estimators are described and three bootstrap algorithms are used to approximate the distribution of the test statistics. The rest of the paper is organized as follows. The methods for testing the equality of regression curves are introduced in Section 2. The resampling algorithms under dependence of the errors are described in Section 3. The results from a broad simulation study performed to compare the proposed tests are reported in Section 4, and finally, the conclusions of our study are presented in Section 5.

2 Testing the equality of regression functions by nonparametric methods

Let us consider the regression models defined in (1.1). Several test statistics for checking the hypothesis of equality of regression curves given in (1.2) are next summarized. In all cases, nonparametric regression estimators are used to evaluate the test statistics.

Test A. The first test statistic considered is computed as the difference between a nonparametric variance estimator of the pooled sample, $\hat{\sigma}_P^2$, and a convex combination of nonparametric variance estimators of the individual samples, $\hat{\sigma}_C^2$. In particular, $\hat{\sigma}_P^2$ is defined by

$$\hat{\sigma}_P^2 = \frac{1}{n} \sum_{l=1}^{k} \sum_{t=1}^{n_l} \left(Y_{l,t} - \hat{m}_{p,g} \left(x_{l,t} \right) \right)^2 , \text{ with } n = \sum_{l=1}^{k} n_l,$$

where $\hat{m}_{p,g}(x)$ is the Nadaraya-Watson estimator of the regression function obtained on the basis of the total combined sample and g is the bandwidth. On the other hand, $\hat{\sigma}_C^2$ is defined by

$$\hat{\sigma}_C^2 = \frac{1}{n} \sum_{l=1}^{k} n_l \hat{\sigma}_l^2,$$

where

$$\hat{\sigma}_l^2 = \frac{1}{n_l} \sum_{t=1}^{n_l} \left(Y_{l,t} - \hat{m}_{l,h_l} \left(x_{l,t} \right) \right)^2$$

is the estimator of the variance of the l-th sample and $\hat{m}_{l,h_l}(x)$ is the individual nonparametric estimator of m_l with bandwidth h_l. This variance estimator was introduced by Hall and Marron (1990). So, the proposed test statistic is

$$\hat{Q}_n^{(A)} = \hat{\sigma}_P^2 - \hat{\sigma}_C^2. \tag{A}$$

The statistic $\hat{Q}_n^{(A)}$ was studied under independence conditions by Dette and Neumeyer (2001).

Test B. The second test statistic is ANOVA type and it is defined by

$$\hat{Q}_n^{(B)} = \frac{1}{n}\sum_{l=1}^{k}\sum_{t=1}^{n_l}\left(\hat{m}_{p,g}(x_{l,t}) - \hat{m}_{l,h_l}(x_{l,t})\right)^2. \tag{B}$$

This test was introduced by Young and Bowman (1995) and is motivated by the classical one-way analysis of variance. Note that there is a strong link between $\hat{Q}_n^{(A)}$ and $\hat{Q}_n^{(B)}$.

Test C. The third test is based on the Crámer-von-Mises type functional distance between individual nonparametric estimators of regression functions. The test statistic is given by

$$\hat{Q}_n^{(C)} = \sum_{l=2}^{k}\sum_{s=1}^{l-1}\int \left(\hat{m}_{l,h_l}(x) - \hat{m}_{s,h_s}(x)\right)^2 \omega_{ls}(x)dx, \tag{C}$$

where $\{\omega_{ls}(x)\}$ are weight functions defined on the support of the design variables, let us say C, a compact set in $\mathbb{R}$. Without loss of generality we can assume that $C = [0,1]$. Essentially, $\hat{Q}_n^{(C)}$ is a consistent estimator of

$$Q = \sum_{l=2}^{k}\sum_{s=1}^{l-1}\int (m_l - m_s)^2 \omega_{ls}.$$

Studies related to statistic $\hat{Q}_n^{(C)}$, in a context of independence, can be seen in King et al. (1991), Kulasekera (1995) and Kulasekera and Wang (1997, 1998) for the case $k = 2$. More recently, Dette and Neumeyer (2001) studied the general case $k \geq 2$ with heterocedasticity and different designs in each group. In this last work, the asymptotic normality of the three test statistics, $\hat{Q}_n^{(A)}$, $\hat{Q}_n^{(B)}$ and $\hat{Q}_n^{(C)}$, was obtained. Also, the consistency of a wild bootstrap version of the tests was established. In a context of

dependence, Vilar-Fernández and González-Manteiga (2003) studied the asymptotic normality of $\hat{Q}_n^{(C)}$ for k regression curves and under general conditions of the design of each group.

In the following methods the case $k = 2$ is considered although the analysis can be easily extended to the comparison of $k > 2$ regression curves.

Tests D and E. Neumeyer and Dette (2003) proposed new tests based on the difference of two marked empirical processes, which are constructed from the residuals obtained under the assumption of equality of the two regression curves. So, the residuals can be generated as

$$\varepsilon_{l,t} = Y_{l,t} - \hat{m}_{p,g}\left(x_{l,t}\right), \; t = 1, \ldots, n_l, \text{ and } l = 1, 2,$$

and the difference between the corresponding marked empirical processes would be given by

$$\hat{R}_n(s) = \frac{1}{n}\sum_{t=1}^{n_1} \varepsilon_{1,t}\, I\left(x_{1,t} \leq s\right) - \frac{1}{n}\sum_{t=1}^{n_2} \varepsilon_{2,t}\, I\left(x_{2,t} \leq s\right),$$

where $s \in [0, 1]$ and $I(\cdot)$ denotes the indicator function.

Neumeyer and Dette (2003) suggested to test the hypothesis of equal regression functions on the basis of real valued functionals of the process $\hat{R}_n(t)$. In particular, the next two test statistics were proposed,

$$\hat{Q}_n^{(D)} = \int_0^1 \hat{R}_n^2(s)\, ds \tag{D}$$

and

$$\hat{Q}_n^{(E)} = \sup_{s \in [0,1]} \hat{R}_n(s). \tag{E}$$

Note that this method is valid in a context of different design points and heterocedasticity. In the case of equal design points, the statistic $\hat{Q}_n^{(E)}$ is essentially the same as the statistic considered in Delgado (1993). The asymptotic behavior under independence of $\hat{Q}_n^{(D)}$ and $\hat{Q}_n^{(E)}$ was studied in Neumeyer and Dette (2003). In particular, the ability of both test statistics to detect alternatives tending to the null at a rate $n^{-1/2}$ was demonstrated and a wild bootstrap version of these methods was analyzed.

It is worthwhile to mention that the results of the five test statistics presented do not depend on the specific smoothing procedure used in the

computation of the test. For example, a local polynomial estimator (Fan and Gijbels, 1996) can be used but with a substantial increase in the mathematical complexity of the proofs of the theoretical results.

Test F. The sixth test considered was proposed by Hall and Hart (1990) and it has nonparametric nature in the sense that no structured functional forms for the regression curves are assumed. In the simple case of equal design and comparing two mean functions, the test statistic is defined as follows.

Define $D_t = Y_{1,t} - Y_{2,t}$ for $1 \leq t \leq n_0$, with $n_0 = n_1 = n_2$, and $D_t = D_{t-n_0}$ for $n_0 + 1 \leq t \leq n_0 + n'$, where n' is the integer part of $n_0 p$, with $0 < p < 1$ fixed. Here n' plays the role of the smoothing parameter. The test statistic is then given by

$$\hat{Q}_n^{(F)} = \left(\sum_{t=0}^{n_0-1} \left(\sum_{i=t+1}^{t+n'} D_t \right)^2 \right) \left(\frac{n_0}{2} \sum_{t=1}^{n_0-1} (D_{t+1} - D_t)^2 \right)^{-1}, \qquad \text{(F)}$$

so that the criterion is to reject the null hypothesis if $\hat{Q}_n^{(F)}$ is too large. The asymptotic behavior of the proposed method was studied in Hall and Hart (1990) under independence of the observations and one bootstrap version was also considered. The test admits several generalizations, including for example, the case of testing the equality of $k \geq 2$ regression curves or of considering different design points.

3 Bootstrap algorithms

In what follows, the regression models of fixed design given in (1.3), with $k = 2$ and $n_1 = n_2 = n_0$, are considered. Without loss of generality, it is assumed that $m(x)$ is defined in $[0, 1]$. The points of the design are taken evenly spaced, that is, $x_t = t/n_0$, for $t = 1, \ldots, n_0$. The processes of random errors $\{\varepsilon_{l,t}\}$ are independent among themselves and each follows an ARMA(p_l, q_l) type dependence structure, i.e.

$$\varepsilon_{l,t} = \sum_{i=1}^{p_l} \phi_{l,i}\, \varepsilon_{l,t-i} + e_{l,t} + \sum_{j=1}^{q_l} \vartheta_{l,j}\, e_{l,t-j}, \text{ with } t \in Z \text{ and } l = 1, 2, \qquad (3.1)$$

where $\{e_{l,t},\, t \in Z\}$ is a sequence of independent random variables with zero mean, finite variance $\sigma_{l,e}^2$ and distribution function $F_{l,e}$. In addition, the series $\{\varepsilon_{l,t}\}$, $l = 1, 2$, are assumed to be stationary and invertible.

In this context, our interest is focused on checking the null hypothesis $H_0 : m_1 = m_2$. For this, any of the test statistics introduced in Section 2, $\hat{Q}_n^{(\bullet)}$, $\bullet =$A, B, C, D, E, F, can be used, so that, in all cases, the criterion is to reject H_0 for large values of $\hat{Q}_n^{(\bullet)}$. In practice, it is obviously necessary to know the distribution of $\hat{Q}_n^{(\bullet)}$ in order to compute the critical values of the test. Unfortunately, it is in general extremely complicated to determine the distribution of these statistics under dependence conditions. In some cases it has been possible to derive the asymptotic distribution, for instance, the asymptotic behavior of $\hat{Q}_n^{(C)}$ was established in Vilar-Fernández and González-Manteiga (2003) for dependent observations. In applications, an alternative procedure for solving this problem is to estimate the unknown parameters of the distribution on the basis of the sample data and then to consider a plug-in version of the test. In any case, it is well known that the convergence rate of the distribution of the test statistic is usually slow and very large sample sizes are necessary to obtain reasonable critical values for the test.

In this work, an alternative simple way of approximating the distribution of the test statistic $\hat{Q}_n^{(\bullet)}$ by means of bootstrapping techniques is proposed. In particular, three different resampling procedures are next studied.

The general idea of the bootstrap methods in time series is that whenever some parametric structure is explicitly stated for the dependence (an ARMA model, for instance), this must be included in the resampling algorithm. This idea is used in the first proposed bootstrap (Bootstrap 1). In the more general case in which a parametric model of dependence cannot be assumed, the bootstrap algorithms are based on replicating the dependence just by resampling a whole block of observations. This is the key point in constructing the other two boostraps considered in this work. The second proposed bootstrap uses blocks of fixed size and the third one uses blocks of random size. Recent reviews on bootstrap methods in time series are those by Li and Maddala (1996), Cao (1999), Berkowitz and Kilian (2000) and Härdle et al. (2003).

The first studied resampling mechanism consists of a simple and direct resampling of the original observations, taking into account that the processes of random errors have an ARMA dependence structure. The algorithm follows the next steps.

Bootstrap 1

Step B1.1 The test statistic $\hat{Q}_n^{(\bullet)}$ is computed from the initial sample given by $\{(x_t, Y_{1,t}, Y_{2,t})\}_{t=1}^{n_0}$.

Step B1.2 Under the null hypothesis, nonparametric residuals $\hat{\varepsilon}_{l,t}$ are obtained by means of

$$\hat{\varepsilon}_{l,t} = Y_{l,t} - \hat{m}_{p,g}(x_t), \text{ for } t = 1, \ldots, n_0 \text{ and } l = 1, 2,$$

where $\hat{m}_{p,g}(x)$ is the nonparametric estimator of the regression function computed from the total combined sample with auxiliary bandwidth g.

Step B1.3 A bootstrap sample of the residuals estimated in **Step B1.2** is drawn as follows.

Step B1.3a Estimates $\left(\hat{\phi}_l, \hat{\vartheta}_l\right)$, $l = 1, 2$, of the parameter vectors associated with the ARMA structure of the errors are constructed on the basis of the residuals estimated $\hat{\varepsilon}_{l,t}$, $l = 1, 2$.

Step B1.3b Since the autoregressive representation of the error processes is invertible, estimates $\{\hat{e}_{l,t},\ t > r_l = \max(p_l, q_l)\}$ of the noise of the ARMA models can be obtained using $\{\hat{\varepsilon}_{l,t}\}$ and $\left(\hat{\phi}_l, \hat{\vartheta}_l\right)$, $l = 1, 2$. The estimated noise series are then centered as $\tilde{e}_{l,t} = \hat{e}_{l,t} - \hat{e}_{l,\cdot}$, for $t > r_l$, where $\hat{e}_{l,\cdot} = \dfrac{1}{n_0 - r_l} \displaystyle\sum_{t=r_l+1}^{n_0} \hat{e}_{l,t}$, for $l = 1, 2$.

Step B1.3c The empirical distribution of $\tilde{e}_{l,t}$ is derived for $l = 1, 2$,

$$\hat{F}_l(x) = \frac{1}{n_0 - r_l} \sum_{t=r_l+1}^{n_0} 1_{\{\tilde{e}_{l,t} \le x\}}.$$

Step B1.3d A sample of independent and identically distributed random variables $\left\{e^*_{l,-M}, \ldots, e^*_{l,-1}, e^*_{l,0}, e^*_{l,1}, \ldots, e^*_{l,n_0}\right\}$, with $M > 0$, is drawn from each $\hat{F}_l$, $l = 1, 2$.
The sequence $\left\{e^\star_{l,t}\right\}_{t=-M}^{n_0}$ is then used together with $(\hat{\phi}_l, \hat{\vartheta}_l)$ to generate a bootstrap sample of the error $\left\{\varepsilon^\star_{l,t}\right\}_{t=1}^{n_0}$, for $l = 1, 2$.

Step B1.4 A bootstrap sample $\{(x_t, Y^*_{1,t}, Y^*_{2,t})\}_{t=1}^{n_0}$ is obtained, making

$$Y^*_{l,t} = \hat{m}_{p,g}(x_t) + \hat{\varepsilon}^*_{l,t},\ t = 1, \ldots, n_0 \text{ and } l = 1, 2.$$

The test statistic $\hat{Q}_n^{(\bullet)*}$ is now computed with this bootstrap sample.

Step B1.5 **Step B1.3d** and **Step B1.4** are repeated a large number of times, say T, so that a sequence $\{\hat{Q}_{n,1}^{(\bullet)*}, \ldots, \hat{Q}_{n,T}^{(\bullet)*}\}$ is obtained. A bootstrap critical region of a significance level α is then given as

$$\hat{Q}_n^{(\bullet)} > \hat{Q}^{(\bullet)*}_{n,([(1-\alpha)T])},$$

where $[\cdot]$ represents the integer part and $\{\hat{Q}^{(\bullet)*}_{n,(i)}\}_{i=1}^{T}$ is the sample $\{\hat{Q}^{(\bullet)*}_{n,i}\}_{i=1}^{T}$ arranged in increasing order of magnitude.

As mentioned before, the other two bootstrap procedures analyzed in this paper do not take into account the autoregressive structure imposed on the error processes. So, there is not an explicit equation to draw the replications of the residuals, and therefore, the resampling mechanism of the estimated residuals $\{\hat{\varepsilon}_{l,t}\}_{t=1}^{n_0}$, $l = 1, 2$, followed in Bootstrap **1** (**Step B1.3**) is not valid in Bootstraps **2** and **3**. In fact, the only modification in the new algorithms concerns the intermediate **Step B1.3** and it is next described for each new bootstrap method.

Bootstrap 2

Step B2.3 The bootstrap sample of the estimated residuals $\{\hat{\varepsilon}_{l,t}\}_{t=1}^{n_0}$, $l = 1, 2$, is generated following the moving blocks bootstrap technique (MBB) (see Künsch (1989) and Li and Maddala (1992)). The method proceeds as follows.

Step B2.3a Fix a positive integer, b, which represents the block size, and take k equal to the smallest integer greater than or equal to n_0/b.

Step B2.3b Define the blocks $B_{l,i} = (\hat{\varepsilon}_{l,i}, \ldots, \hat{\varepsilon}_{l,i+b-1})$, for $i = 1, \ldots, q$, with $q = n_0 - b + 1$, and $l = 1, 2$.

Step B2.3c Draw k blocks, $\epsilon_{l,1}, \ldots, \epsilon_{l,k}$, with equiprobable distribution from the set $\{B_{l,1}, \ldots, B_{l,q}\}$, $l = 1, 2$. Note that every $\epsilon_{l,i}$ is a b-dimensional vector $(\epsilon_{l,i,1}, \ldots, \epsilon_{l,i,b})$.

Step B2.3d The bootstrap version of the estimated residuals is formed with the first n_0 components of

$$(\epsilon_{l,1,1}, \ldots, \epsilon_{l,1,b}, \epsilon_{l,2,1}, \ldots, \epsilon_{l,2,b}, \ldots, \epsilon_{l,k,1}, \ldots, \epsilon_{l,k,b}) .$$

Bootstrap 3

The MBB method is not stationary. To overcome this drawback, Politis and Romano (1994) proposed a stationary bootstrap (SB) which is used in Bootstrap **3** to obtain resamples of $\hat{\varepsilon}_{l,t}$.

Step B3.3 The bootstrap replications of $\{\hat{\varepsilon}_{l,t}\}_{t=1}^{n_0}$, $l = 1, 2$, are now drawn by means of the following stationary bootstrap.

Step B3.3a Fix a positive real number $p \in [0, 1]$.

Step B3.3b The first bootstrap replication, $\hat{\varepsilon}^*_{l,1}$, is directly drawn from the empirical distribution of $\{\hat{\varepsilon}_{l,t}\}_{t=1}^{n_0}$, $l = 1, 2$.

Step B3.3c Once the value $\hat{\varepsilon}^*_{l,i} = \hat{\varepsilon}_{l,j}$, for some $j \in \{1, \ldots, n_0 - 1\}$, has been drawn, with $i < n_0$, then the next bootstrap replication $\hat{\varepsilon}^*_{l,i+1}$ is defined as $\hat{\varepsilon}_{l,j+1}$, with probability $1 - p$, and drawn from the empirical distribution of $\{\hat{\varepsilon}_{l,t}\}_{t=1}^{n_0}$, $l = 1, 2$, with probability p.
In the particular case $j = n_0$, $\hat{\varepsilon}_{l,j+1}$ is replaced by $\hat{\varepsilon}_{l,1}$.

Remark 3.1. *Note that the test statistic $\hat{Q}_n^{(F)}$ differs in its nature from the rest of considered statistics since it is based on the values $D_t = Y_{1,t} - Y_{2,t}$ while all the others obtain previous estimators of the regression functions on the basis of the initial sample $\{(x_t, Y_{1,t}, Y_{2,t})\}_{t=1}^{n}$. As a consequence, the described bootstrap methodologies are not directly applicable when $\hat{Q}_n^{(F)}$ is considered. Under independence conditions, Hall and Hart (1990) proposed a bootstrap approximation to the distribution of $\hat{Q}_n^{(F)}$ based on obtaining bootstrap replications from the sequence $\{D_t\}_{t=1}^{n_0}$. Nevertheless, that approach is not possible under dependence conditions because the explicit dependence structure of the process D_t is not known. For this reason the analysis of the test statistic $\hat{Q}_n^{(F)}$ using Bootstrap **1** is omitted in this work.*

*With regard to Bootstraps **2** and **3** for approximating the distribution of $\hat{Q}_n^{(F)}$, they are run as follows. The centered differences $d_t = (Y_{1,t} - Y_{2,t}) -$*

$\left(\bar{Y}_{1,\cdot} - \bar{Y}_{2,\cdot}\right)$, $t = 1, \ldots, n_0$, *with* $\bar{Y}_{l,\cdot} = \dfrac{1}{n_0}\displaystyle\sum_{t=1}^{n_0} Y_{l,t}$, $l = 1, 2$, *are firstly obtained. Then, a bootstrap resample* $\{d_t^*\}_{t=1}^{n_0}$ *is drawn from* $\{d_t\}_{t=1}^{n_0}$ *by using the MBB technique (Bootstrap 2) or the SB one (Bootstrap 3). Now, a bootstrap version* $\hat{Q}_n^{(F)*}$ *of the statistic* $\hat{Q}_n^{(F)}$ *is computed from each* $\{d_t^*\}_{t=1}^{n_0}$ *and the p-value of the test is finally derived as in* **Step B1.5**.

It is also important to notice that there exist at least two ways for using Bootstrap 1 with $\hat{Q}_n^{(F)}$. *The first of these would consist in drawing bootstrap resamples of the initial sample* $\{(x_t, Y_{1,t}, Y_{2,t})\}_{t=1}^{n_0}$ *and then computing* $\hat{Q}_n^{(F)*}$ *from each resample. The second one would be to fit an ARMA model to the process* $\{d_t\}_{t=1}^{n_0}$ *and then to follow* **Step B1.3-Step B1.5** *to obtain naive resamples of* $\{d_t\}_{t=1}^{n_0}$.

Remark 3.2. *The described bootstrap algorithms admit several variations. For instance, in Bootstrap 1, the bootstrap resample of centered estimated errors,* $\tilde{e}_{l,t}$, *is obtained from the empirical distribution (naive bootstrap), but one alternative would be to obtain this resample by wild bootstrap (see Härdle and Mammen, 1993). In a similar problem (testing linear regression model) Vilar-Fernández and González-Manteiga (2000) showed that both methods have similar behavior.*

4 Simulation study

This section shows some of the results of a broad simulation study performed to compare the different tests based on the statistics $\hat{Q}_n^{(\bullet)}$ using the three bootstrap algorithms described.

Samples $\{(x_t, Y_{1,t}, Y_{2,t})\}_{t=1}^{n_0}$, of size $n_0 = 100$, were simulated by following the regression model given in (1.3), with $k = 2$ and $x_t = t/n_0$, $t = 1, \ldots, n_0$. The error processes were designed to follow the same AR(1) model

$$\varepsilon_{l,t} = \phi\varepsilon_{l,t-1} + e_{l,t},\ t \in Z \text{ and } l = 1, 2,$$

where $\varepsilon_{1,t}$ and $\varepsilon_{2,t}$ have the same distribution function, $N(0, \sigma^2)$, $\sigma^2 = 0.5$, and $\phi = 0.50$ or 0.80.

In the first study the regression functions $m_1(x) = m_2(x) = \cos(\pi x)$ were considered under the null hypothesis. A total of 500 trials were carried out. Each one consisted in obtaining an initial sample and computing the

values of the statistics $\hat{Q}_n^{(\bullet)}$, $(\bullet) =$ A, B, C, D, E, F. Then, a bootstrap resample of size $T = 500$ was obtained for each statistic by using the three bootstrap procedures (although bootstrap 1 was omitted for $\hat{Q}_n^{(F)}$ as already mentioned). So, bootstrap critical regions of several significance levels α were determined and the corresponding associated p-values were approximated.

The smoothing parameter used in the computation of the statistics $\hat{Q}_n^{(\bullet)}$, $(\bullet) =$ A, B, C, D, E, was $h = 0.30$ and the value of the parameter p to obtain $\hat{Q}_n^{(F)}$ was $p = 0.75$. Other auxiliary parameter values were empirically chosen.

Smoothing parameter selection indeed plays an important role here and its influence will be analyzed further on. In any case, our initial choice $h = 0.30$ allowed us to obtain satisfactory results for all the tests considered under the null hypothesis, so that $h = 0.30$ was a reasonable choice to compare the relative merits of the competing tests. Nevertheless, in general, an automatic selector of the smoothing parameter aimed to increase the test power would be a more adequate approach.

The results for $\phi = 0.50$ are summarized in Table 1. In particular, the simulated rejection probabilities of the proposed bootstrap tests with level 10%, 5% and 2.5% are shown in Table 1 together with the average and standard deviation of the set of p-values obtained in the 500 trials.

A simple inspection of Table 1 allows us to conclude that the tests based on Bootstrap **1** present the best performance with independence of the statistic considered. In fact, the differences due to the kind of statistic are practically negligible. In contrast, relevant differences are found when the results are compared according to the bootstrap procedure used. Bootstraps **2** and **3** are clearly worse, providing rejection levels significantly greater than the theoretical ones. Furthermore, note that Bootstrap **3** leads to more satisfactory results than Bootstrap **2** for any of the statistics, A, C, D and E, while both methods present similar results with the statistics B and F. In particular, the results for $\hat{Q}_n^{(A)}$, $\hat{Q}_n^{(C)}$ and $\hat{Q}_n^{(D)}$ using Bootstrap **3** are somewhat acceptable.

The same numerical study was next carried out with a stronger dependence level for the observations. In particular, $\phi = 0.80$ was chosen and the results obtained are shown in Table 2.

Table 1: *Mean and standard deviation of the critical values of the tests analyzed and simulated rejection probabilities for three levels* $\alpha = 0.10$, 0.05 *and* 0.025 *with* $\phi = 0.50$.

Bootstrap 1	$\alpha = 0.10$	$\alpha = 0.05$	$\alpha = 0.025$	Mean p-values	St. Dev. p-values
Test A	0.1020	0.0520	0.0160	0.5322	0.2863
Test B	0.0980	0.0520	0.0320	0.5185	0.2868
Test C	0.0960	0.0500	0.0240	0.5228	0.2867
Test D	0.1080	0.0480	0.0300	0.5020	0.2862
Test E	0.1080	0.0560	0.0240	0.5118	0.2891
Bootstrap 2					
Test A	0.2180	0.1420	0.0960	0.4343	0.3115
Test B	0.1860	0.1140	0.0880	0.4395	0.3044
Test C	0.1760	0.1120	0.0760	0.4480	0.3033
Test D	0.1960	0.1540	0.1020	0.4327	0.3072
Test E	0.2500	0.1640	0.1200	0.3841	0.2949
Test F	0.3000	0.2260	0.1680	0.3751	0.3142
Bootstrap 3					
Test A	0.0880	0.0620	0.0420	0.6491	0.3281
Test B	0.1500	0.1240	0.0880	0.5598	0.3521
Test C	0.0720	0.0560	0.0380	0.6788	0.3144
Test D	0.1300	0.0840	0.0580	0.6101	0.3443
Test E	0.1720	0.1180	0.0900	0.5568	0.3482
Test F	0.3080	0.2380	0.1780	0.4610	0.3723

Results in Table 2 are uniformly worse than those in Table 1. This was expected since it is well known that to increase the dependence level means to lose sample information for a given sample size. Once this is made clear, the conclusions obtained from Table 1 can be extended to the results in Table 2 in spite of the increasing dependence. So, all the test statistics provided reasonable rejection levels (although slightly higher than the theoretical ones) when Bootstrap **1** was used, while both Bootstrap **2** and Bootstrap **3** led to quite poor results. In fact, the empirical levels obtained with these two bootstrap procedures are quite far from the nominal ones. Hence, it is interesting to explore how much the sample size, n_0, must be increased to approximate well the theoretical levels. For this purpose the numerical study was again run with sample sizes $n_0 = 300$ and $n_0 = 500$.

Table 2: *Mean and standard deviation of the critical values of the tests analyzed and simulated rejection probabilities for three levels* $\alpha = 0.10$, 0.05 *and* 0.025 *with* $\phi = 0.80$.

Bootstrap 1	$\alpha = 0.10$	$\alpha = 0.05$	$\alpha = 0.025$	Mean p-values	St. Dev. p-values
Test A	0.1240	0.0880	0.0480	0.4918	0.2943
Test B	0.1200	0.0660	0.0480	0.4907	0.2937
Test C	0.1220	0.0680	0.0480	0.4891	0.2931
Test D	0.1240	0.0700	0.0500	0.4866	0.2972
Test E	0.1400	0.0760	0.0500	0.4799	0.2912
Bootstrap 2					
Test A	0.3280	0.2440	0.1940	0.3404	0.3078
Test B	0.3000	0.2200	0.1840	0.3443	0.3014
Test C	0.2660	0.1940	0.1440	0.3727	0.3046
Test D	0.3320	0.2660	0.2020	0.3568	0.3204
Test E	0.4040	0.3060	0.2500	0.2715	0.2784
Test F	0.4700	0.4060	0.3540	0.2581	0.2914
Bootstrap 3					
Test A	0.2560	0.2060	0.1820	0.4932	0.3783
Test B	0.3880	0.3180	0.2780	0.3800	0.3789
Test C	0.1880	0.1520	0.1280	0.5669	0.3700
Test D	0.3100	0.2720	0.2420	0.4692	0.3897
Test E	0.3960	0.3400	0.3020	0.3629	0.3609
Test F	0.5920	0.5480	0.5100	0.2343	0.3314

Table 3 shows the rejection probabilities pertaining to the theoretical significance level $\alpha = 0.05$ and the autocorrelation value $\phi = 0.50$. In this case a smaller bandwidth $h = 0.20$ was used.

Table 3 confirms that the good behavior of Bootstrap **1** becomes already apparent with $n_0 = 100$. However, Bootstrap **3** needs to work with at least $n_0 = 300$ to obtain acceptable results and Bootstrap **2** even requires sample sizes larger than $n_0 = 500$ to reach reasonable approximations to the theoretical significance level.

A useful tool to discern between the proposed tests is the empirical distribution function of the simulated p-values, say $\hat{F}_p^{(\bullet)}$, $(\bullet) =$ A, B, C, D, E, F. In particular, it is of interest to check graphically and numerically how

Table 3: *Simulated rejection probabilities for $\alpha = 0.05$, with $\phi = 0.50$ and three different initial sample sizes.*

$n_0 = 100$	Test A	Test B	Test C	Test D	Test E	Test F
Bootstrap 1	0.0880	0.0660	0.0680	0.0700	0.0760	
Bootstrap 2	0.2440	0.2200	0.1940	0.2660	0.3060	0.4060
Bootstrap 3	0.2060	0.3180	0.1520	0.2720	0.3400	0.5480
$n_0 = 300$	**Test A**	**Test B**	**Test C**	**Test D**	**Test E**	**Test F**
Bootstrap 1	0.0620	0.0460	0.0640	0.0580	0.0580	
Bootstrap 2	0.1300	0.1220	0.0940	0.1100	0.1200	0.1600
Bootstrap 3	0.0680	0.1280	0.0440	0.0600	0.0700	0.1580
$n_0 = 500$	**Test A**	**Test B**	**Test C**	**Test D**	**Test E**	**Test F**
Bootstrap 1	0.0417	0.0617	0.0583	0.0600	0.0650	
Bootstrap 2	0.1000	0.0900	0.0883	0.0983	0.1083	0.1350
Bootstrap 3	0.0283	0.0750	0.0217	0.0517	0.0533	0.1333

close $\hat{F}_p^{(\bullet)}$ is to the distribution of a uniform $[0,1]$ random variable, F_U. For example, with regard to the graphical procedures, a very convenient plot for this purpose is the p-value discrepancy plot (see Davidson and MacKinnon, 1986), consisting in plotting the pairs $(p_s, \hat{F}_p^{(\bullet)}(p_s) - p_s)$ where s indexes the simulated samples. Figure 1 shows the discrepancy plots associated with tests C and D for each bootstrap procedure with $n_0 = 100$ and $\phi = 0.50$.

Figure 1 corroborates the above comments in the sense that Bootstrap **1** shows the best performance of the test statistics. Note that although the conduct of both the $\hat{Q}_n^{(D)}$-test and the $\hat{Q}_n^{(C)}$-test is quite close, $\hat{Q}_n^{(D)}$ presents slightly better behavior. In any case, Bootstraps **2** and **3** are clearly worse and the best performance corresponds to the $\hat{Q}_n^{(C)}$-test with Bootstrap **2**.

To illustrate numerically the differences between distribution functions $\hat{F}_p^{(\bullet)}$ and F_U, Delicado and Placencia (2001) proposed to use the following distances:

$$d_{KS} = \sqrt{S} \sup_{p\in[0;1]} \left| \hat{F}_{p,S}^{(\bullet)}(p) - p \right|,$$

$$d_{L_r} = \sqrt{S} \left(\int \left| \hat{F}_{p,S}^{(\bullet)}(p) - p \right|^r dp \right)^{1/r}, \quad r = 1, 2,$$

where S is the number of samples (here $S = 500$). In hypothesis test context, the above distances are specially sensitive to deviations of $\hat{F}_p^{(\bullet)}$

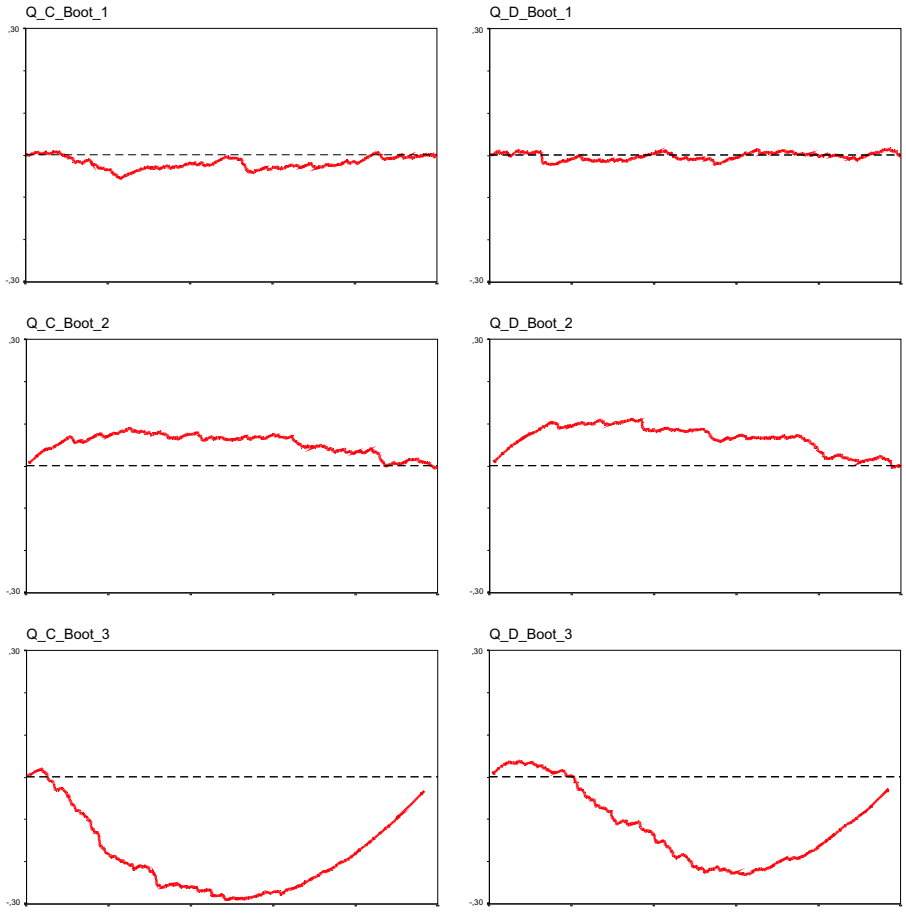

Figure 1: *The $p-$value discrepancy plots associated with tests C and D for each bootstrap procedure used with $n_0 = 100$ and $\phi = 0.50$. Axis $OY = [-0.30, 0.30]$.*

from the diagonal at low values of the nominal size α. So, it is interesting to introduce a weight function ω in the distance definitions to correct this effect and to pay more attention to that range of values. Delicado and Placencia (2001) proposed to use as weight function the density of a beta distribution $\beta(a = 2, b = 8)$, that is,

$$\omega(p) = 72p(1-p)^7, \quad 0 \leq p \leq 1.$$

So, the distances are reformulated to be

$$d_{KS}^{\omega} = \sqrt{S} \sup_{p\in[0;1]} \left|\hat{F}_{p,S}^{(\bullet)}(p) - p\right| \omega(p),$$

$$d_{L_r}^{\omega} = \sqrt{S} \left(\int \left|\hat{F}_{p,S}^{(\bullet)}(p) - p\right|^r \omega(p)\, dp \right)^{1/r}, \quad r = 1, 2.$$

The six distances introduced to measure the discrepancy between the distribution functions $\hat{F}_p^{(\bullet)}$ and F_U were computed for all statistics and for the three bootstrap procedures. The results for the case $n_0 = 100$ and $\phi = 0.50$ are shown in Table 4. For the case $\phi = 0.80$ only the results obtained for the last three distances are shown in Table 5.

Table 4: *Distances between the empirical function of the simulated p-values of the proposed tests and the uniform distribution with* $n_0 = 100$ *and* $\phi = 0.50$.

		Test A	**Test B**	**Test C**	**Test D**	**Test E**	**Test F**
$\mathbf{d}_{KS}$	**B.1**	1.431	1.029	1.252	0.492	0.939	
	B.2	2.728	2.236	2.057	2.504	3.578	4.651
	B.3	5.814	3.757	6.484	5.143	3.622	4.830
$\mathbf{d}_{KS}^{\omega}$	**B.1**	3.695	3.011	3.328	1.725	2.364	
	B.2	7.707	6.328	5.601	8.076	8.980	10.953
	B.3	8.260	6.006	12.537	4.088	5.995	11.398
$\mathbf{d}_{L_1}$	**B.1**	0.705	0.497	0.421	0.169	0.298	
	B.2	1.493	1.376	1.185	1.526	2.614	2.816
	B.3	3.359	1.856	4.004	2.647	1.866	2.087
$\mathbf{d}_{L_1}^{\omega}$	**B.1**	0.718	0.523	0.488	0.214	0.318	
	B.2	1.992	1.683	1.479	1.944	2.505	3.113
	B.3	1.995	0.999	3.117	1.151	1.203	3.007
$\mathbf{d}_{L_2}$	**B.1**	0.816	0.583	0.510	0.200	0.386	
	B.2	1.686	1.507	1.303	1.674	2.797	3.105
	B.3	3.885	2.201	4.507	3.151	2.109	2.457
$\mathbf{d}_{L_2}^{\omega}$	**B.1**	0.870	0.623	0.587	0.242	0.396	
	B.2	2.069	1.735	1.519	2.004	2.640	3.319
	B.3	2.672	1.125	3.695	1.584	1.317	3.246

Tables 4 and 5 allow us to conclude that, according to the minimum distance criterion between $\hat{F}_p^{(\bullet)}$ and F_U, Bootstrap **1** definitely behaves much better than the other two bootstrap methods. The best performance of

Table 5: *Distances between the empirical function of the simulated p-values of the proposed tests and the uniform distribution with* $n_0 = 100$ *and* $\phi = 0.80$.

		Test A	**Test B**	**Test C**	**Test D**	**Test E**	**Test F**
$\mathbf{d}^{\omega}_{KS}$	**B.1**	2.845	2.656	2.135	2.366	3.152	
	B.2	11.756	11.101	9.938	11.929	12.789	13.615
	B.3	11.345	13.444	8.216	12.901	13.460	13.689
$\mathbf{d}^{\omega}_{L_1}$	**B.1**	0.444	0.333	0.328	0.472	0.592	
	B.2	3.435	3.193	2.825	3.434	3.849	4.212
	B.3	2.587	3.870	1.516	3.244	3.922	4.427
$\mathbf{d}^{\omega}_{L_2}$	**B.1**	0.497	0.369	0.386	0.482	0.633	
	B.2	3.718	3.438	2.998	3.704	4.258	4.757
	B.3	2.762	4.257	1.683	3.526	4.329	5.123

Bootstrap **1** is indeed justified since it takes advantage of knowing the particular error autoregressive structure. The other bootstrap mechanisms considered replicate the dependence without assuming an explicit error structure, so that their generality involves a loss of efficiency. In any case both Bootstrap **2** and Bootstrap **3** behave worse than expected in terms of efficiency and hence large sample sizes are required when they are used. In this sense, note that in contrast with the results in Table 3, where Bootstrap **3** presents advantage over Bootstrap **2** for $\alpha = 0.05$, the distances observed between $\hat{F}_p^{(\bullet)}$ and F_U are significantly smaller for Bootstrap **2** than for Bootstrap **3**. So, it can be concluded that, in general, Bootstrap **2** exhibits better performance than Bootstrap **3**. Similar conclusions were also derived from our simulation study when different autocorrelation values were considered.

As Bootstrap **1** procedure requires to assume a parametric dependence structure, it is appropriate to examine its robustness to misspecification problems. For this purpose, the simulation design was modified to generate errors following arbitrary ARMA models but keeping the Bootstrap **1** algorithm without changes. This means assuming always an AR(1) error structure. Some of the results obtained with this new simulation plan are shown in Table 6. Specifically, Table 6 provides the rejection probabilities simulated with the three bootstrap procedures for a theoretical level $\alpha = 0.05$, a sample size $n_0 = 500$ and the following dependence models for the error process:

$$\text{Model 1 (ARMA(1,1))}: \quad \varepsilon_t = 0.8\varepsilon_{t-1} + e_t + 0.3e_{t-1} \tag{1}$$

$$\text{Model 2 (ARMA(1,1))}: \quad \varepsilon_t = 0.5\varepsilon_{t-1} + e_t + 0.8e_{t-1} \tag{2}$$

$$\text{Model 3 (MA(2))}: \quad \varepsilon_t = e_t + 1.4e_{t-1} + 0.4e_{t-2} \tag{3}$$

$$\text{Model 4 (AR(2))}: \quad \varepsilon_t = 0.4\varepsilon_{t-1} + 0.5\varepsilon_{t-2} + e_t \tag{4}$$

Table 6: *Simulated rejection probabilities with the three bootstrap procedures for* $\alpha = 0.05$, $n_0 = 500$ *and four different dependence models for the error process. Bootstrap* **1** *was constructed as if the error was AR(1).*

Bootstrap 1	**Test A**	**Test B**	**Test C**	**Test D**	**Test E**	**Test F**
Model 1	0.0100	0.0150	0.0200	0.0225	0.0125	
Model 2	0.0000	0.0025	0.0100	0.0100	0.0025	
Model 3	0.0050	0.0050	0.0075	0.0050	0.0050	
Model 4	0.6500	0.5475	0.4975	0.3375	0.4500	
Bootstrap 2						
Model 1	0.0725	0.1425	0.0900	0.0875	0.1825	0.3400
Model 2	0.0200	0.0375	0.0325	0.0625	0.0625	0.0150
Model 3	0.0175	0.0200	0.0100	0.0300	0.0200	0.0500
Model 4	0.7700	0.6900	0.6150	0.4750	0.6525	0.6825
Bootstrap 3						
Model 1	0.1075	0.3675	0.0725	0.0875	0.1350	0.5500
Model 2	0.0250	0.1250	0.0200	0.0275	0.0275	0.2475
Model 3	0.0020	0.0275	0.0025	0.0075	0.0125	0.1400
Model 4	0.5825	0.7375	0.3800	0.3575	0.4900	0.7375

Results in Table 6 do not yield definitive conclusions on the robustness of the Bootstrap **1** procedure. In fact, when Bootstrap **1** was used, Models 1, 2 and 3 led to conservative tests and Model 4 provided very poor results. This is because of the amount of departure from the AR(1) structure. Thus, while a certain similarity with the AR(1) covariance structure exists for the first three models, Model 4 is very different from an AR(1). In general, according to the results in Tables 1, 2 and 6, it is reasonable to expect that if the covariance structure used to perform Bootstrap **1** is similar to the true dependence structure, then Bootstrap **1** will provide acceptable results. In this sense, when the theoretical dependence model is not clear from data, it may be advisable to consider an AR(p) structure, with p large enough, to construct the Bootstrap **1** algorithm.

Except for Model 4, results in Table 6 for Bootstrap **2** and Bootstrap **3** are similar to those previously achieved under an AR(1) error structure (see Table 3). Note that, in general, both Bootstrap **2** and Bootstrap **3** provided less conservative tests than Bootstrap **1**. Results for Model 4 were again very poor but here this performance is not expected since Bootstraps **2** and **3** are constructed by resampling blocks of observations without assuming a particular model for the error. This unexpected behavior is likely due to an unsuitable selection of two parameter values required in the resampling algorithms. These parameter values are the block size b, in the case of Bootstrap **2** algorithm (see **Step B2.3a** in Section 3), and the probability value p, in the case of Bootstrap **3** algorithm (see **Step B3.3a** in Section 3). Each of these values has a heavy influence on the behavior of the corresponding bootstrap algorithm and, in our simulation study, both b and p were kept fixed for the four models. We think that a more suitable choice in the case of Model 4 should improve the results significantly. However, the selection of b and p from the sample data is still an open problem, as well as it is to establish the consistency of Bootstraps **2** and **3**.

Next, our numerical study was extended to analyze the influence of the smoothing parameter on the tests. In particular, the statistic $\hat{Q}_n^{(C)}$ was chosen to explore the effect of the bandwidth h used in its computation (assuming $h_1 = h_2 = h$). The simulation was performed as follows. A total of 300 random samples of size $n = 100$ were drawn under null hypothesis with autoregressive parameter $\phi = 0.50$. The statistic $\hat{Q}_n^{(C)}$ was then computed from each sample and for different values of h ranging from $h_{\min} = 0.010$ to $h_{\max} = 0.600$ in 0.010 steps. So, 300 trials were run for each h and the corresponding 300 p-values obtained by using each bootstrap procedure were averaged. Let $AC_i(h)$ denote the function assigning to each h the average of the p-values of the test statistic $\hat{Q}_n^{(C)}$ using bootstrap i, $i = 1, 2, 3$. In addition, let $Q^{(C)}(h)$ be the function defined by $Q^{(C)}(h) = 10 \times \overline{\hat{Q}}_n^{(C)}(h)$, where $\overline{\hat{Q}}_n^{(C)}(h)$ denotes the average of the 300 values obtained for the statistic when the bandwidth h was used. The graphs derived from the simulation study for $AC_1(h)$, $AC_2(h)$, $AC_3(h)$ and $Q^{(C)}(h)$ are jointly shown in Figure 2.

Figure 2 shows that, with independence of the selected smoothing parameter, Bootstrap **1** leads to average p-values always close to the mean of the uniform on $[0, 1]$. This fact confirms the very good performance of Bootstrap **1** under the null hypothesis, and simultaneously, allows us to

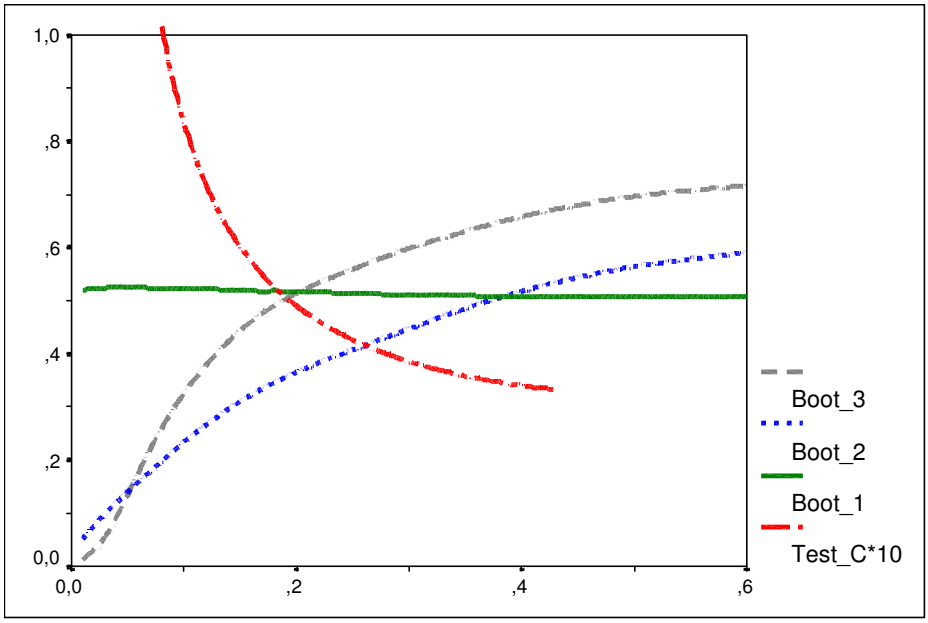

Figure 2: *Graphs of* $10 \times \overline{\hat{Q}}_n^{(C)}(h)$ *and of the averaged p-values for test C using the three bootstraps as a function of* h. *Sample size is* $n = 100$ *and the autoregressive coefficient is* $\phi = 0.50$

conclude the negligible influence of the bandwidth over the results of the test when the bandwidth is chosen in a reasonable range of values. In contrast to Bootstrap **1**, an undersmoothing of the regression functions leads to incorrect rejection of the null hypothesis of equality when Bootstrap **2** or Bootstrap **3** are used. On the other hand, the three bootstrap procedures yield similar results for large values of h. This was expected since an oversmoothing provides estimates tending to the averages of the response variables. So, if $\bar{Y}_{1\cdot} \approx \bar{Y}_{2\cdot}$, then $\hat{Q}_n^{(\cdot)} \approx 0$, and therefore, the null hypothesis is accepted and the resulting test presents low power. Therefore it can be concluded that the bandwidth selection problem is indeed crucial to obtain a satisfactory power for the test, but its influence is greater with Bootstrap **2** and Bootstrap **3** than with Bootstrap **1**.

Similar behavior was observed with a larger number of samples. Table 7 shows the average and standard deviation of the p-values of test C using the three bootstrap methods for several values of h on the basis of 1000 samples.

Table 7: *Mean and standard deviation of critical values of the test C and simulated rejection probabilities for level* $\alpha = 5\%$, *varying the smoothing parameter* h, *on the basis of* 1000 *samples of size* $n_0 = 100$ *and autoregressive coefficient* $\phi = 0.50$.

Test C		$h = 0.10$	$h = 0.20$	$h = 0.30$	$h = 0.40$
Bootstrap 1	Mean p-value	0.5246	0.5176	0.5133	0.5105
	St. Dev. p-value	0.2902	0.2929	0.2899	0.2886
	$\alpha = 0.05$	0.0370	0.0500	0.0560	0.0560
Bootstrap 2	Mean p-value	0.2419	0.3629	0.4475	0.5198
	St. Dev. p-value	0.2827	0.3072	0.3027	0.2924
	$\alpha = 0.05$	0.3870	0.2020	0.1060	0.0690
Bootstrap 3	Mean p-value	0.3127	0.5031	0.5972	0.6628
	St. Dev. p-value	0.3438	0.3497	0.3425	0.3321
	$\alpha = 0.05$	0.3510	0.1370	0.0870	0.0710

The simulation study was next driven to investigate the power of the proposed tests. The same model described above, with $\phi = 0.50$, was simulated but now under alternative hypotheses. Specifically, the simulated regression functions were $m_1(x) = \cos(\pi x)$ and $m_2(x) = m_1(x) + \Delta(x)$, with $\Delta(x)$ taking several forms described in Table 8, so that the differences $m_1 - m_2$ represent a variety of alternatives.

Table 8: *Simulated rejection probabilities of the proposed tests under the alternative hypothesis* $m_2(x) = m_1(x) + \Delta(x)$ *using bootstrap 1 for level* $\alpha = 0.05$ *with* $\phi = 0.50$ *and* $n_0 = 100$.

$\Delta(x)$	**Test A**	**Test B**	**Test C**	**Test D**	**Test E**
0.25	0.2160	0.2680	0.2760	0.2740	0.2560
0.50	0.7260	0.7740	0.7700	0.8200	0.7920
0.75	0.9760	0.9820	0.9820	0.9880	0.9840
$0.50\,x$	0.2840	0.2900	0.2900	0.3040	0.2480
$0.75\,x$	0.6040	0.6060	0.5920	0.5800	0.5200
$1.00\,x$	0.8520	0.8460	0.8440	0.8300	0.7940
$0.5\sin(2\pi x)$	0.1780	0.0820	0.0620	0.0360	0.0600
$\sin(2\pi x)$	0.5720	0.2020	0.1160	0.0100	0.1580

First, sample size $n_0 = 100$ was considered. Bootstraps **2** and **3** do not lead to acceptable results of the tests under the null hypothesis with $n_0 = 100$ (see Table 1) and so it makes no sense to investigate their power for that sample size. Thus, just Bootstrap **1** technique was initially included in the simulation. Simulated rejection probabilities for level $\alpha = 0.05$ of the proposed tests with Bootstrap **1** are shown in Table 8.

Table 9: *Simulated rejection probabilities of the proposed tests under the alternative hypothesis* $m_2(x) = m_1(x) + \Delta(x)$ *using the three bootstrap algorithms for level* $\alpha = 0.05$ *with* $\phi = 0.50$ *and* $n_0 = 500$.

$\Delta(x)$		**Test A**	**Test B**	**Test C**	**Test D**	**Test E**	**Test F**
0.10	**B1**	0.2017	0.2417	0.2383	0.3017	0.2717	
	B2	0.2867	0.3217	0.2250	0.3633	0.3600	0.4067
	B3	0.0917	0.2483	0.0733	0.1600	0.1750	0.3683
0.25	**B1**	0.8300	0.8683	0.8667	0.9000	0.8867	
	B2	0.8633	0.9033	0.8417	0.9267	0.9217	0.9333
	B3	0.4683	0.7300	0.4917	0.6433	0.5683	0.8700
0.50	**B1**	0.7260	0.7740	0.7700	0.8200	0.7920	
	B2	0.8240	0.8560	0.8440	0.8940	0.8880	0.9280
	B3	0.5520	0.7340	0.5780	0.7020	0.7120	0.8860
$0.25\,x$	**B1**	0.9760	0.9820	0.9820	0.9880	0.9840	
	B2	0.8900	0.8880	0.8860	0.9980	0.9920	0.9980
	B3	0.8580	0.9460	0.8440	0.9220	0.9200	0.9920
$0.50\,x$	**B1**	1.000	1.000	1.000	1.000	1.000	
	B2	1.000	1.000	1.000	1.000	1.000	1.000
	B3	0.9575	1.000	0.9600	0.9825	0.9925	1.000
$0.25\,x$	**B1**	0.3500	0.4083	0.3917	0.3967	0.3400	
	B2	0.4183	0.4967	0.4583	0.4850	0.4350	0.5067
	B3	0.1800	0.3617	0.1750	0.2017	0.1967	0.4467
$0.50\,x$	**B1**	0.9100	0.9300	0.9200	0.9100	0.8675	
	B2	0.9325	0.9575	0.9475	0.9350	0.9150	0.9375
	B3	0.5800	0.8000	0.5450	0.5975	0.5725	0.8900
$0.25\,\sin(2\pi x)$	**B1**	0.3750	0.3225	0.2850	0.0525	0.1300	
	B2	0.4500	0.4225	0.3500	0.0925	0.2275	0.1975
	B3	0.1600	0.2725	0.0925	0.0550	0.1075	0.1875
$0.50\,\sin(2\pi x)$	**B1**	0.9600	0.9400	0.9225	0.0475	0.5125	
	B2	0.9700	0.9550	0.9425	0.00825	0.6425	0.3850
	B3	0.6450	0.7550	0.4275	0.0300	0.2650	0.5650

The inspection of any row in Table 8 allows us to conclude that there are no significant differences among the several tests considered since all of them provided similar rejection probabilities. However note the worse performance of test D in the case of the oscillating alternative ($\Delta(x) = \sin(2\pi x)$), which has been previously observed in Neumeyer and Dette (2003) under independence conditions.

Next, the sample size was increased to compare the bootstrap procedures in terms of power of the tests. Table 9 includes the results generated with $n_0 = 500$.

Valuable information can be derived from the results in Table 9. First, regardless of the alternative considered, the tests based on Bootstrap **3** were clearly the worst in terms of power. In general, the highest rejection probabilities were attained with Bootstrap **2** although the results generated with Bootstrap **1** were only a little bit worse. Therefore, Bootstrap **1** can be placed in the first position of the ranking since it is competitive in terms of power and, in contrast to others, permits to achieve acceptance rates very close to the nominal ones under the null hypothesis. Concerning the test statistics, note that test D again showed the worst performance in the case of the sinusoidal alternative. In addition, also test E and, to a less extent, test F presented a worse performance than the rest for this kind of alternative.

Additional simulation studies were carried out and similar conclusions were derived in all the cases. So, for example, Table 10 collects the results of some of these studies where different regression functions m_1 and m_2 were used. The regression model and the parameter values are the same as in the previous study.

Table 10: *Simulated rejection probabilities of the proposed tests under different alternative hypotheses* $m_2(x)$. *Here,* $m_1(x) = 0$, $\alpha = 0.05$, $\phi = 0.50$ *and* $n_0 = 500$.

$m_2(x)$		**Test A**	**Test B**	**Test C**	**Test D**	**Test E**	**Test F**
0	**B1**	0.0800	0.0680	0.0680	0.0580	0.0620	
	B2	0.1680	0.1640	0.1480	0.1520	0.1940	0.2560
	B3	0.0720	0.1580	0.0540	0.0540	0.0680	0.1880
$x^2/2$	**B1**	0.5340	0.5200	0.4500	0.3020	0.3540	
	B2	0.6580	0.6940	0.5560	0.5760	0.5380	0.5160
	B3	0.3640	0.5280	0.2740	0.2500	0.2400	0.4740
x^2	**B1**	0.9800	0.9760	0.9520	0.9180	0.8920	
	B2	0.9840	0.9840	0.9740	0.9500	0.9420	0.9420
	B3	0.8500	0.9280	0.7220	0.7420	0.7340	0.8640
$\sqrt{x}/4$	**B1**	0.3700	0.3920	0.3920	0.3940	0.3680	
	B2	0.4920	0.5280	0.4180	0.5320	0.5060	0.5700
	B3	0.2680	0.4360	0.2480	0.2620	0.2620	0.5280

(*Continued on next page*)

(*Table 10. Continued from previous page*)

$m_2(x)$		**Test A**	**Test B**	**Test C**	**Test D**	**Test E**	**Test F**
$\sqrt{x}/2$	**B1**	0.8900	0.9060	0.9160	0.9120	0.8880	
	B2	0.9440	0.9500	0.9420	0.9540	0.9400	0.9420
	B3	0.6680	0.8500	0.6760	0.7420	0.7360	0.9040
$\sin(2\pi x)/4$	**B1**	0.3060	0.2540	0.2260	0.0880	0.1400	
	B2	0.4560	0.4020	0.3500	0.1500	0.3080	0.2340
	B3	0.2040	0.3380	0.1300	0.0500	0.1060	0.2580
$\sin(2\pi x)/2$	**B1**	0.8040	0.7520	0.7180	0.0700	0.3820	
	B2	0.8700	0.8360	0.8120	0.1480	0.6160	0.4000
	B3	0.5860	0.7000	0.4120	0.0420	0.25400	0.3500

5 Conclusions

In this work several methods of testing the equality of regression curves by using regression nonparametric estimators have been analyzed in a context of dependence. In particular, three bootstrap algorithms were used to approximate the distribution of the statistic tests considered. One of the three bootstrap procedures, Bootstrap **1**, is based on replicating the dependence structure which is assumed to be known in this case. The other two procedures, named Bootstraps **2** and **3**, are more general since they are constructed without assuming an explicit structure for the error. The broad simulation study carried out allows us to conclude the superiority of Bootstrap **1** when compared to the other two bootstrap algorithms, which was expected since Bootstrap **1** takes advantage of knowing the dependence model. It is remarkable that, under the null hypothesis and with moderate sample sizes, both Bootstrap **2** and Bootstrap **3** have provided substantially greater rejection rates than the theoretical ones, and in fact, this poor performance was uniformly presented over all test statistics analyzed. At the same time, Bootstrap **1** also performed well in terms of power under different alternative hypotheses and with large sample sizes. In this sense, Bootstrap **1** and Bootstrap **2** performed similarly and much better than Bootstrap **3**. On the other hand, in the present fixed design regression context, the dependence structure for the error can be frequently well adjusted by means of an ARMA model. The suitable parameter values for the ARMA structure would be obtained on the basis of the residuals derived from a previous nonparametric estimation of the regression function. In any case, our simulation study also showed that Bootstrap **1** is reasonably

robust to small deviations from the true dependence structure. Indeed, this is an additional reason in favor of using Bootstrap **1** algorithm. In summary, the previous arguments suggest that it is better to first estimate the error dependence structure and then to use Bootstrap **1** than to use Bootstrap **2** or Bootstrap **3** directly.

Concerning to the analyzed test statistics, some conclusions can be also made from the simulation study. First, regardless of the resampling technique used, the six statistics have shown similar performance under the null hypothesis. None of them exhibited a significant advantage over the rest. There was not a clear winner in terms of power under different alternatives either. However, in this setting, it is remarkable to note that statistics D and E presented the worst results with the oscillating alternatives, while the behavior of statistic D was very poor in these cases. As in every nonparametric procedure, it is necessary to take into account the smoothing parameter selected in order to assess the behavior of the different statistics. Our simulation study allowed us to observe that the influence of the bandwidth is not as strong as in the case of having to fit a regression curve. Nevertheless, it was made clear that too small bandwidths lead to incorrectly reject the null hypothesis while too large bandwidths frequently lead to accept the null hypothesis when it is false. Therefore, the bandwidth selection problem is also very important in this context. The first strategy to tackle this problem would consist in adjusting the methods proposed in Kulasekera and Wang (1997, 1998) to dependence setting. In any case, an empirical solution would be to compute the p-values associated with the test statistic for a grid of the bandwidth values and to make a decision on the basis of the results achieved.

Acknowledgement

The authors thank the associate editor and the referee for their helpful comments that substantially improved this paper.

References

Berkowitz, J. and Kilian, L. (2000). Recent developments in bootstrapping time series. *Econometrics Reviews*, 19:1–48.

Cao, R. (1999). An overview of bootstrap methods for estimating and predicting in time series. *Test*, 8(1):95–116.

Davidson, R. and MacKinnon, J. G. (1986). Graphical methods for investigating the size and power of hypothesis tests. *The Manchester School*, 66:1–26.

Delgado, M. A. (1993). Testing the equality of non-parametric regression curves. *Statistics & Probability Letters*, 17:199–204.

Delicado, P. and Placencia, I. (2001). Comparing empirical distributions of p-values from simulations. *Communications in Statistics. Simulation and Computing*, 30(2):403–422.

Dette, H. and Neumeyer, N. (2001). Nonparametric analysis of covariance. *The Annals of Statistics*, 29(5):1361–1400.

Fan, J. and Gijbels, I. (1996). *Local Polynomial Modelling and Its Applications*. Chapman and Hall, London.

Hall, P. and Hart, J. D. (1990). Bootstrap test for difference between means in nonparametric regression. *Journal of American Statistical Association*, 412(85):1039–1049.

Hall, P., Huber, C., and Speckman, P. L. (1997). Covariate-matched one-sided tests for the difference between functional means. *Journal of American Statistical Association*, 439(92):1074–1083.

Hall, P. and Marron, J. S. (1990). On variance estimation in nonparametric regression. *Biometrika*, 77:415–419.

Härdle, W., Horowitz, J., and Kreiss, J. P. (2003). Bootstrap methods for time series. *International Statisitical Review*, 71(2):435–460.

Härdle, W. and Mammen, E. (1993). Comparing nonparametric versus parametric regression fits. *The Annals of Statistics*, 21:1926–1947.

Härdle, W. and Marron, J. S. (1990). Semiparametric comparison of regression curves. *The Annals of Statistics*, 18:63–89.

King, E. C., Hart, J. D., and Wehrly, T. E. (1991). Testing the equality of two regression curves using linear smoothers. *Statistics & Probability Letters*, 12(3):239–247.

Koul, H. L. and Schick, A. (1997). Testing for the equality of two nonparametric regression curves. *Journal of Statistical, Planning and Inference*, 65:293–314.

Kulasekera, K. B. (1995). Comparison of regression curves using quasi-residuals. *Journal of American Statistical Association*, 431(90):1085–1093.

Kulasekera, K. B. and Wang, J. (1997). Smoothing parameter selection for power optimality in testing of regression curves. *Journal of American Statistical Association*, 438(92):500–511.

Kulasekera, K. B. and Wang, J. (1998). Bandwidth selection for power optimality in a test of equality of regression curves. *Statistics & Probability Letters*, 37:287–293.

Künsch, H. R. (1989). The jackknife and the bootstrap for general stationary observations. *The Annals of Statistics*, 17:1217–1241.

Li, H. and Maddala, G. (1992). Moving blocks jackknife and bootstrap capture weak dependence. In R. LePage and L. Billard, eds., *Exploring the limits of bootstrap*, pp. 225–248. John Wiley & Sons, New York.

Li, H. and Maddala, G. (1996). Bootstrapping time series models. *Econometric Theory*, 15:115–195.

Munk, A. and Dette, H. (1998). Nonparametric comparison of several regression functions: exact and asymptotic theory. *The Annals of Statistics*, 26(6):2339–2368.

Neumeyer, N. and Dette, H. (2003). Nonparametric comparison of regression curves: an empirical process approach. *The Annals of Statistics*, 31(3):880–920.

Politis, D. N. and Romano, J. R. (1994). The stationary bootstrap. *Journal of American Statistical Association*, 89:1303–1313.

Scheike, T. H. (2000). Comparison of nonparametric regression functions through their cumulatives. *Statistics & Probability Letters*, 46:21–32.

Vilar-Fernández, J. M. and González-Manteiga, W. (2000). Resampling for checking linear regression models via non-parametric regression estimation. *Computational Statistics & Data Analysis*, 35:211–231.

Vilar-Fernández, J. M. and González-Manteiga, W. (2003). Nonparametric comparison of curves with dependent errors. *Statistics*, 38(2):81–99.

Young, S. G. and Bowman, A. W. (1995). Nonparametric analysis of covariance. *Biometrics*, 51:920–931.