

CONSISTENT ORDER SELECTION WITH STRONGLY DEPENDENT DATA AND ITS APPLICATION TO EFFICIENT ESTIMATION*

Javier Hidalgo
London School of Economics and Political Science

Contents:

1. Introduction
 2. Statistical Framework and the HI Estimator of the Parameters θ_j
 3. Order Selection of a Distributed Lag Regression Model
 4. Numerical Example
 5. Proof of Theorems 3.1 and 3.2
- References
Tables 4.1 – 4.3

Discussion Paper
No.EM/02/430
February 2002

The Suntory Centre
Suntory and Toyota International Centres for
Economics and Related Disciplines
London School of Economics and Political Science
Houghton Street
London WC2A 2AE
Tel.: 020- 7955 6698

* Research supported by the Economic and Social Research Council (ESRC) reference number R000238212. I thank a referee for helpful comments which led to a much improved and clearer version of the article and to Peter Robinson who suggested the topic to me. All remaining errors are, of course, the sole responsibility of the author.

Abstract

Order selection based on criteria by Akaike (1974), *AIC*, Schwarz (1978), *BIC* or Hannan and Quinn (1979) *HIC* is often applied in empirical examples. They have been used in the context of order selection of weakly dependent *ARMA* models, *AR* models with unit or explosive roots and in the context of regression or distributed lag regression models for weakly dependent data. On the other hand, it has been observed that data exhibits the so-called strong dependence in many areas. Because of the interest in this type of data, our main objective in this paper is to examine order selection for a distributed lag regression model that covers in a unified form weak and strong dependence. To that end, and because of the possible adverse properties of the aforementioned criteria, we propose a criterion function based on the decomposition of the variance of the innovations of the model in terms of their frequency components. Assuming that the order of the model is finite, say p_o , we show that the proposed criterion consistently estimates, p_o . In addition, we show that adaptive estimation for the parameters of the model is possible without knowledge of p_o . Finally, a small Monte-Carlo experiment is included to illustrate the finite sample performance of the proposed criterion.

Keywords: Order selection; distributed lag models; strong dependence.

JEL No.: C22

© Javier Hidalgo. All rights reserved. Short sections of text, not to exceed two paragraphs, may be quoted without special permission provided that full credit, including © notice, is given to the source.

Contact address: Dr Javier Hidalgo, Department of Economics, London School of Economics and Political Science, Houghton Street, London WC2A 2AE. Email: f.j.hidalgo@lse.ac.uk

1. INTRODUCTION

In empirical studies, methods to select the order of a model are commonly invoked and routinely applied. The methods are based on the minimization of a criterion function which involves the estimation of the one-step-prediction error plus some penalty function. Standard procedures are Akaike (1974), *AIC*, Schwarz (1978), *BIC*, or Hannan and Quinn's (1979), *HIC*, information criterion. They only differ on the employed penalty function. These criteria have been used in a context of regression models, see for example Shibata (1981) or Pötscher (1989), in distributed lag regression models, see Geweke and Meese (1981) or in the selection of the order of an Autoregressive or Autoregressive Moving Average model, see Shibata (1976, 1980) and Hannan (1980), among others. See also Pötscher (1991) and George (2000) and references therein for a review.

Although the criteria have been mainly justified in the case of weakly dependent data, some work has been carried out for nonstationary (unit roots) or explosive data in an autoregressive context, see Pötscher (1989), or for *ARFIMA* models, see Beran et al. (1998) and Hidalgo (2001). However, in the context of regression models, and in particular distributed lag regression models, the available work is mainly for uncorrelated data/errors. An exception is Wang (1993) and Kavalieris and Hannan (1994) who allowed the error term of the regression model to be weakly dependent, although it should be noted that Geweke and Meese (1981) conjectured, without proof, that their results should hold when the error term follows a weakly dependent process.

On the other hand, it appears that the situation where the regressors and/or errors are allowed to be strongly dependent has not been examined yet. Due to the interest and immense attention that this type of data has received in recent years, and the possible adverse properties of the aforementioned criteria as we discuss below, the main objective of the paper is to introduce and examine the properties of an alternative criterion $S(p)/S^*(p)$, given in (3.3)/(3.4) below, which covers both weakly and strongly dependent data in a unified framework. The criterion is based on the decomposition of the one-step-prediction error, e.g. the variance of the innovation of the model, in terms of its frequency components. We show that when the dimension of the model, say p_0 , is finite the criterion consistently estimates p_0 .

To be more specific, we consider the following distributed lag regression model

$$y_t = \sum_{j=1}^{p_0} \theta_j x_{t-j+1} + u_t, \quad t = 1, \dots, T \quad (1.1)$$

where both x_t and u_t are scalar processes exhibiting, possibly, strong dependence and where p_0 is known to be finite but with an unknown upper bound. The latter

seems to be of practical interest, since it is unlikely that the upper bound for the number of lags in (1.1) will be known to the practitioner in advanced. It should be mentioned that, although we focus on the case of y_t and x_t scalar, the results follow similarly to the case where y_t and/or x_t are nonscalar, but at the expense of unnecessarily notational and mathematical complication. The main objective of the paper is to show that the value \hat{p} which minimizes the criterion $S(p)$ or $S^*(p)$ is a consistent estimator of p_0 with $g(T)$ chosen as in the *BIC* and/or *HIC* criteria.

We now discuss briefly the motivation to introduce a different criterion instead of the more standard *BIC* and *HIC* criteria and one of their main problem, in particular their inconsistency, when examining the estimator of p_0 with those criteria. We defer to Section 3 to comment on some additional technical difficulties that this criteria have with strongly dependent data. First, recall that the *BIC* and *HIC* are given by

$$\bar{S}^*(p) = \log(\hat{\sigma}_p^2) + \frac{pg(T)}{T}, \quad (1.2)$$

where $g(T)$ is $\log(T)$ and $2c \log(\log(T))$, with $c > 1$, respectively, and where $\hat{\sigma}_p^2$ is the standard estimator of the variance of u_t , that is $T^{-1} \sum_{t=1}^T \hat{u}_{t,p}^2$, with $\hat{u}_{t,p}$ denoting the least squares residuals after a p th order distributed lag regression model was fitted. Alternatively, as in Geweke and Meese (1981), we can employ the criterion

$$\bar{S}(p) = \hat{\sigma}_p^2 + \frac{pg(T)}{T}. \quad (1.3)$$

If the upper bound for p_0 , say P , were known and finite the standard approach to prove the consistency of, say,

$$\tilde{p} = \arg \min_{1 \leq p \leq P} \bar{S}(p) \quad (1.4)$$

would be based on showing that for any $p \neq p_0$ such that $p \leq P$,

$$\lim_{T \rightarrow \infty} \Pr \{ \bar{S}(p_0) - \bar{S}(p) > 0 \} = 0. \quad (1.5)$$

Similarly with the criterion $\bar{S}^*(p)$.

The key ingredient to show (1.5) is based on the assumption that $T(\hat{\sigma}_p^2 - \hat{\sigma}_{p_0}^2) = O_p(1)$. Note that since $\log(\hat{\sigma}_{p_0}^2/\hat{\sigma}_p^2) \leq (\hat{\sigma}_{p_0}^2/\hat{\sigma}_p^2) - 1$ the same applies if the criterion function (1.2) were used. When $T(\hat{\sigma}_p^2 - \hat{\sigma}_{p_0}^2) = O_p(1)$ is the case, (1.5) holds true whenever $g(T)$ increases more slowly than T , that is $g^{-1}(T) + T^{-1}g(T) \rightarrow 0$, being the two more popular choices of $g(T)$ those corresponding to the *BIC* and *HIC* criteria respectively. However, as we argue in Section 2, when the

data exhibits strong dependence, since the least squares estimate (*LSE*) of θ_j is T^α -consistent with $0 < \alpha \leq 1/2$, it will imply that $T(\hat{\sigma}_p^2 - \hat{\sigma}_{p_0}^2) = O_p(T^{1-2\alpha})$, and thus $T(\hat{\sigma}_p^2 - \hat{\sigma}_{p_0}^2)$ will not be bounded in probability if $\alpha < 1/2$. As a consequence, the latter will imply that the left side of (1.5) will be greater than zero if $g(T)$ is chosen as in the *BIC* and/or *HIC*, and thus \tilde{p} will be inconsistent. One possibility for (1.5) to hold true, and hence the consistency of the criteria, is to choose $g(T)$ in such a way that it grows to infinity faster than $\log(T)$, and in particular to choose $g(T) = T^{1-2\alpha} \log(T)$. However, since α can be arbitrarily close to 0, to guarantee that (1.3) is a consistent criterion, it will require $g(T)$ to have a rate of increase to infinity close to T . It is worth mentioning that a similar phenomena occurs in the determination of the order of an *AR* model with all its roots inside the unit circle. In the latter case, the penalty function $g(T)$ should satisfy $\liminf_{T \rightarrow \infty} g(T)/T > 0$ to guarantee that the order of the *AR* polynomial is not overestimated with positive probability, see Pötscher's (1989) Theorem 3.2. But, as Hannan and Quinn (1979) mentioned, the faster the rate of increase of $g(T)$ to infinity it will make the criterion to underestimate the order p_0 more frequently in finite samples. The latter might not be surprising since the bigger $g(T)$ the penalty attached by adding a new regressor can be too big compared with the reduction on $\hat{\sigma}_p^2$.

One possible solution to the problem mentioned above with the *BIC* and *HIC* is to follow an approach similar to that of Wang (1993) or Kavalieris and Hannan (1994). That is, for our model (1.1), we could modify $\bar{S}(p)$ by

$$\log(\hat{\sigma}_p^2) + cp \frac{\log T}{T}$$

or

$$\log(\hat{\sigma}_e^2) + (5p + h) \frac{\log T}{T}$$

where $\hat{\sigma}_e$ is the residual variance of the innovations after a model with p lags and an *AR*(h) for the errors u_t in (1.1), respectively were fitted. However, the former modification requires that $c > (2 \max_\lambda f_{uu}(\lambda)) / \int_{-\pi}^{\pi} f_{uu}(\lambda) d\lambda$, where $f_{uu}(\lambda)$ denotes the spectral density function of u_t , which becomes infinity for strongly dependent processes. On the other hand, for the latter modification, no theory is yet available to know how well an *AR*(h), with $h \rightarrow \infty$, approximates $f_{uu}(\lambda)$, albeit that in our framework u_t needs to be estimated and the number of regressors is allowed to increase with the sample size at a rate $\log T$. Moreover, the asymptotic properties of a *GLS* type estimator for θ_j using an *AR*(h), $h \rightarrow \infty$, as an estimator of $f_{uu}(\lambda)$ remains an open problem, although from the work of Hidalgo and Robinson (2001) it appears far from trivial.

Therefore, in view of the previous comments, it seems to be desirable to implement a procedure with $g(T)$ as in *BIC* and/or *HIC*, for example, independent

of the value of α and at the same time allowing for possible correlation on the errors of unknown form, as a correct parameterization of $f_{uu}(\lambda)$ can be difficult to obtain in applied work. Note that on the contrary to weakly dependent data, with strongly dependent data, an incorrect parameterization of $f_{uu}(\lambda)$ will not guarantee to obtain a $T^{1/2}$ -consistent estimator of the parameters θ_j . See Hidalgo and Robinson (2001).

We now describe our second objective of the paper. As inefficient estimators of θ_j , see (2.6) below, are used when evaluating $S(p)$ or $S^*(p)$ given in (3.3) or (3.4), it will be of interest to illustrate how an adaptive estimator (in the Gauss Markov sense) of the parameters θ_j can be obtained without knowledge of p_0 . That is, the (adaptive) generalized least squares (*GLS*) estimator of the parameters of the model. This is motivated by recent results of Hidalgo and Robinson (2001), who obtained adaptive estimation of θ_j assuming that the number of lags p_0 in the model (1.1) is known a priori.

The remainder of the paper is organized as follows. In the next section, we describe the statistical framework and also the (inefficient) estimator of the parameters θ_j denoted as the *HI* estimator. Section 3 describes and examines the criterion for the order selection of (1.1). Moreover, we indicate how the *GLS* estimator of θ_j is obtained when p_0 is replaced by its estimator. In Section 4, a Monte-Carlo experiment examines the finite sample performance of $S(p)$ and $S^*(p)$, as well as that of $\bar{S}(p)$ and $\bar{S}^*(p)$. Finally, Section 5 gives the proof of our main results.

2. STATISTICAL FRAMEWORK AND THE *HI* ESTIMATOR OF THE PARAMETERS θ_j

Let $w_t = (y_t, x_t)'$ be a covariance stationary linear process, with absolute continuous spectral matrix distribution, so that its spectral density matrix, denoted $f_{ww}(\lambda)$, is defined as

$$E((w_1 - Ew_1)(w'_{j+1} - Ew'_1)) = \int_{-\pi}^{\pi} f_{ww}(\lambda) e^{-ij\lambda} d\lambda, \quad j = 0, \pm 1, \pm 2, \dots \quad (2.1)$$

In the frequency domain, the lag structure given in (1.1) is described by the frequency response function $\Theta(\lambda) = \sum_{j=0}^{p_0} \theta_j e^{-ij\lambda}$, so that θ_j is interpreted as the j th Fourier coefficient of $\Theta(\lambda) = f_{yx}(\lambda) f_{xx}^{-1}(\lambda)$, that is

$$\theta_j = (2\pi)^{-1} \int_0^{2\pi} \Theta(\lambda) e^{ij\lambda} d\lambda, \quad j = 0, \pm 1, \pm 2, \dots, \quad (2.2)$$

where $f_{yx}(\lambda)$ and $f_{xx}(\lambda)$ are the indicated elements of $f_{ww}(\lambda)$ in (2.1). Due to this interpretation of θ_j , Hannan (1963) proposed to estimate θ_j by the sample

(discrete) analogue of (2.2), that is denoting $\lambda_j = (2\pi j) / T$ for integer j

$$\tilde{\theta}_j = \frac{1}{2M} \sum_{\ell=0}^{2M-1} \hat{\Theta}(\lambda_{2m\ell}) e^{ij\lambda_{2m\ell}}, \quad (2.3)$$

where $\hat{\Theta}(\lambda_{2m\ell}) = \hat{f}_{yx}(\lambda_{2m\ell}) \hat{f}_{xx}^{-1}(\lambda_{2m\ell})$ and $\hat{f}_{yx}(\lambda_{2m\ell})$ and $\hat{f}_{xx}(\lambda_{2m\ell})$ are the indicated elements of the estimator of the spectral matrix $f_{ww}(\lambda)$, $M = \lfloor T/4m \rfloor$, with m a bandwidth parameter which increases more slowly than T , that is $m^{-1} + mT^{-1} \rightarrow 0$ as $T \rightarrow \infty$. In this paper we have chosen to estimate $f_{ww}(\lambda)$ by the average periodogram estimator

$$\hat{f}_{ww}(\lambda) = \frac{1}{2m+1} \sum_{j=-m}^m I_{ww}(\lambda_j + \lambda), \quad (2.4)$$

with

$$I_{ww}(\lambda) = (2\pi T)^{-1} \left(\sum_{t=1}^T w_t e^{it\lambda} \right) \left(\sum_{t=1}^T w'_t e^{-it\lambda} \right) \quad (2.5)$$

as the periodogram of w_t . The estimator $\tilde{\theta}_j$ in (2.3) was coined by Sims (1974) as Hannan's inefficient (*HI*) estimator in contrast to Hannan's (1963) efficient estimator, e.g. the adaptive *GLS* estimator of θ_j .

However, similar to technical problems encountered with many other semi-parametric estimators, since $\hat{f}_{xx}(0)$ tries to estimate $f_{xx}(0)$ which may possibly be infinity in our framework (see Assumptions A.1 and A.3 below), the estimator given in (2.3) is quite difficult to analyze as it stands. Thus, as in Hidalgo (2000), we modify (2.3) by

$$\hat{\theta}_j = \frac{1}{2M} \sum_{\ell=1}^{2M-1'} \hat{f}_{yx,2m\ell} \hat{f}_{xx,2m\ell}^{-1} e^{ij\lambda_{2m\ell}}, \quad (2.6)$$

where $\sum_{p=1}^{2M-1'} a_p e^{ij\lambda_{2mp}}$ means $\sum_{p=1}^{2M-1} a_p e^{ij\lambda_{2mp}} + a_1$ and where henceforth g_j denotes $g(\lambda_j)$ for a generic function $g(\lambda)$. Intuitively, what we have done is to use $\hat{\Theta}_{2m}$ as an estimator of $\Theta(0)$.

The asymptotic properties of the estimator given in (2.6) were first established by Hannan (1967) for a finite, possibly unknown, distributed lag regression model where the spectral density functions of both x_t and u_t were bounded and bounded away from zero, and Brillinger (1981) for the infinite distributed lag regression model. More recently, Hidalgo (2000) has shown that the same holds true under the presence of strong dependence, see Theorem 2.1 below which we present for expositional purposes.

The motivation of the estimator in (2.6) is threefold. First is its ability to estimate the coefficients θ_j irrespective of the number of lags specified in (1.1), which will be relevant when analyzing the properties of $S(p)$ defined in (3.3) below. Second, since there is no gain by exploiting the information of the covariance structure of the errors u_t , as Sims (1974) showed, the *HI* estimator becomes as efficient as the *GLS* estimator. This motivates the *LSE* of θ_j given in Robinson (1979), although under stronger assumptions than those we want to impose in this paper. The third, and perhaps most important motivation comes from the observation that when the joint strong dependence in the regressor x_t and error term u_t is sufficiently strong, that is the product of the spectral density functions of x_t and u_t is not integrable, the *LSE* is no longer root- T nor asymptotically normal, see Robinson (1994).

Motivated by the last observation, Robinson and Hidalgo (1997) showed that a class of frequency-domain weighted *LSE*, including *GLS* (with parametric error spectral density function) as a special case, is root- T consistent, asymptotically normal and Gauss-Markov efficient in model (1.1) when p_0 is known. The intuition why the estimator in Robinson and Hidalgo (1997) is root- T consistent and asymptotically normal is because the weighted function possesses a zero sufficiently strong to compensate for the singularity of the spectral density function induced by the joint strong dependence of x_t and u_t . So, since $f_{xx}^{-1}(\lambda)$ possesses a zero at $\lambda = 0$ when $d > 0$, see Assumption A.2 below, we can expect that $\widehat{f}_{xx,p}^{-1}$ becomes (asymptotically) a weighted function satisfying the conditions of Robinson and Hidalgo (1997). Thus, the modified *HI* estimator given in (2.6) is indeed a desirable estimator.

Let us introduce two definitions:

Definition 1. $\mathcal{K}(d, \alpha)$ is the set of functions, $g(\cdot)$, such that $g(\lambda) > 0$ for all $\lambda \in [0, \pi]$ satisfying

$$g(\lambda) = C\lambda^{-2d} (1 + O(\lambda^\alpha)) \text{ as } \lambda \rightarrow 0+$$

where $C \in (0, \infty)$, $d \in [0, 1/2)$ and $\alpha \in (0, 2]$, and twice continuously differentiable in any open set outside the origin.

Definition 2. \mathcal{L} is the set of, possibly, complex functions $\eta(\cdot)$, such that $|\eta(\lambda)| > 0$ for all $\lambda \in [0, \pi]$, $(\partial/\partial\lambda)|\eta(\lambda)| = O(|\eta(\lambda)|/\lambda)$ as $\lambda \rightarrow 0+$, and twice continuously differentiable in any open set outside the origin.

Let us introduce the following assumptions:

Assumption A.1 (i) $\{w_t\} = \{(y_t, x_t)'\}$ and $\{u_t\}$ are two covariance stationary linear processes defined as

$$w_t = \sum_{j=0}^{\infty} \tau_j \xi_{t-j}, \quad \sum_{j=0}^{\infty} \|\tau_j\|^2 < \infty \quad \text{and} \quad (2.7)$$

$$u_t = \sum_{j=0}^{\infty} \psi_j \varepsilon_{t-j}, \quad \sum_{j=0}^{\infty} |\psi_j|^2 < \infty,$$

where τ_0 is the identity matrix, $\psi_0 = 1$ and $\|D\|$ stands for the norm of the matrix D .

(ii) The functions $\Lambda(\lambda) = \sum_{j=0}^{\infty} \tau_j e^{ij\lambda}$ and $\Psi(\lambda) = \sum_{j=0}^{\infty} \psi_j e^{ij\lambda}$, $\lambda \in [0, \pi]$, belong to \mathcal{L} .

(iii) $\{x_t\}$ and $\{u_t\}$ are two mutually independent sequences of random variables such that $|x_t|^4$ and $|u_t|^4$ are uniformly integrable.

Denote $R(\lambda) = f_{yx}(\lambda) / f_{xx}^{1/2}(\lambda) f_{yy}^{1/2}(\lambda)$ as the coherence between y_t and x_t .

Assumption A.2 (i) $R(\lambda)$ is twice continuously differentiable in any open set outside the origin and for some $\beta \in (1, 2]$,

$$|R(\lambda) - R(0)| = O(\lambda^\beta) \quad \text{as } \lambda \rightarrow 0+.$$

(ii) $f_{xx}(\lambda)$, $f_{yy}(\lambda)$ and $f_{uu}(\lambda)$ belong to $\mathcal{K}(d_x, 2)$, $\mathcal{K}(d_y, 2)$ and $\mathcal{K}(d_u, 2)$, respectively, and where $f_{uu}(\lambda)$ denotes the spectral density function of u_t as defined in (2.1) with u_t replacing w_t .

Assumption A.3 Denote by $\eta_g(\lambda)$ the g th row of the matrix

$$\eta(\lambda) = \text{diag}(\Lambda(\lambda), \Psi(\lambda)),$$

and by $f_{gg}(\lambda)$ the g th diagonal element of

$$f(\lambda) = \text{diag}(\text{diag}(f_{ww}(\lambda)), f_{uu}(\lambda)).$$

Then, for all g , $f_{gg}^{-1/2} |\eta_g(\lambda)|$ is a nonzero finite vector.

Assumption A.4 $\{\varepsilon_t\}$ is a stochastic process, where $E(\varepsilon_t | \mathcal{G}_{t-1}) = 0$, $E(\varepsilon_t^2 | \mathcal{G}_{t-1}) = E(\varepsilon_t^2) = \sigma_\varepsilon^2$ a.s., $E(|\varepsilon_t^3| | \mathcal{G}_{t-1}) = \mu_3 < \infty$ where \mathcal{G}_t is the σ -algebra of events

generated by $\varepsilon_s, s \leq t$, and the joint fourth cumulant of $\varepsilon_{t_i}, i = 1, \dots, 4$ satisfies

$$\text{cum}(\varepsilon_{t_1}, \varepsilon_{t_2}, \varepsilon_{t_3}, \varepsilon_{t_4}) = \begin{cases} \kappa_\varepsilon & t_1 = t_2 = t_3 = t_4 \\ 0 & \text{otherwise,} \end{cases}$$

with $|\kappa_\varepsilon| < \infty$.

Assumption A.5 $\{\xi_t\}$ is a stochastic process, where $E(\xi_t | \mathcal{F}_{t-1}) = 0$, $E(\xi_t \xi_t' | \mathcal{F}_{t-1}) = E(\xi_t \xi_t') = \Xi$ a.s., $E(|\xi_{t_{j_1}} \xi_{t_{j_2}} \xi_{t_{j_3}}| | \mathcal{F}_{t-1}) = \mu_{3, j_1, j_2, j_3} < \infty$ for all $j_1, j_2, j_3 = 1, 2$ where \mathcal{F}_t is the σ -algebra of events generated by $\xi_s, s \leq t$, and the joint fourth cumulant of $\xi_{t_i j_i}, j_i = 1, 2$ and $i = 1, \dots, 4$ satisfies

$$\text{cum}(\xi_{t_1 j_1}, \xi_{t_2 j_2}, \xi_{t_3 j_3}, \xi_{t_4 j_4}) = \begin{cases} \kappa_{j_1, j_2, j_3, j_4} & t_1 = t_2 = t_3 = t_4 \\ 0 & \text{otherwise,} \end{cases}$$

with $\kappa_\xi = \max_{j_i, i=1, \dots, 4} |\kappa_{j_1, j_2, j_3, j_4}| < \infty$.

Assumption A.6 $m^{-2}T + T^{-3}m^4 \rightarrow 0$.

Some discussion about our assumptions is in order. Part (i) of Assumption A.1 is restrictive in the linearity it imposes, but not otherwise. In particular A.1(i) combined with A.4 and A.5, e.g. $E(\varepsilon_t | \mathcal{G}_{t-1}) = 0$, is equivalent to the assertion that the best linear predictor is the best predictor, in the least squares sense. Part (ii) of Assumption A.1 and Assumption A.2 deal with the need for smoothness on both $f_{ww}(\lambda)$ and $f_{uu}(\lambda)$. For frequencies $\lambda \rightarrow 0+$, these are the same used elsewhere by, for example, Robinson (1995) and thus, the same comments apply here, while for frequencies λ outside any close set containing the origin are standard. The requirement of independence between x_t and u_t in the first part of Assumption A.1 (iii), as in Robinson and Hidalgo (1997), is necessary for the proof of the asymptotic normality of (2.6), c.f. Theorem 2.1 below. We believe that it might be possible to relax this assumption to some extent, but that will complicate enormously the otherwise already technical proof given in Robinson and Hidalgo (1997). This certainly remains an open question. Furthermore, it should also be noted that a sufficient condition for the last part of Assumption A.1 (iii) is $\sup_t E(|x_t|^{4+\delta} + |u_t|^{4+\delta}) < \infty$ for some $\delta > 0$. Assumption A.3 is not strong, see for instance the comments made after (2.8) below, once λ^{d_g} is identified there as $f_{gg}^{-1/2}$ up to constants. Finally, Assumption A.6 gives upper and lower bounds for the rate of increase of m to infinity.

Examples of processes satisfying A.1-A.5 are as follows. Let ξ_t be a v -dimensional unobservable covariance stationary process with a continuous and bounded away

from zero spectral density matrix, and consider the filter

$$z_t = \sum_{j=0}^{\infty} G(j) \xi_{t-j} \quad (2.8)$$

where $G(j)$ are $v \times v$ matrices. Denote by $G_g(\lambda)$ the g th row of the matrix $G(\lambda) = \sum_{j=0}^{\infty} G(j) e^{ij\lambda}$, such that $|G_g(\lambda)| \lambda^{d_g}$, with $d_g \in [0, \frac{1}{2})$ for $g = 1, \dots, v$, tends to a non-zero finite vector as $\lambda \rightarrow 0+$, $g = 1, \dots, v$, then A.1-A.3 hold. As an example, let ξ_t be a stationary invertible autoregressive moving average (ARMA) process and let each z_{tg} be formed by separate fractional integration of the corresponding ξ_t element, so that

$$G(\lambda) = \text{diag} \left((1 - e^{i\lambda})^{-d_1}, \dots, (1 - e^{i\lambda})^{-d_v} \right).$$

In particular, when z_t is scalar, the spectral density function of z_t is

$$f(\lambda) = \frac{\sigma^2}{2\pi} |1 - e^{i\lambda}|^{-2d} \left| \frac{\Theta(e^{i\lambda})}{\Phi(e^{i\lambda})} \right|^2, \quad -\pi < \lambda \leq \pi, \quad (2.9)$$

with $0 \leq d < 1/2$, and where $\Theta(\lambda)$ and $\Phi(\lambda)$ are the MA and AR polynomials respectively, having no zeroes in or on the unit circle. This is the familiar ARFIMA model, see for instance Granger and Joyeux (1980) or Hosking (1981).

Another process which exhibits strong dependence is the fractional gaussian noise process introduced by Mandelbrot and Van Ness (1968), whose spectral density function, see Sinai (1976), is

$$f(\lambda) = \frac{4\sigma^2 \Gamma(2d)}{(2\pi)^{3+2d}} \cos(\pi d) \sin^2(\lambda/2) \sum_{j=-\infty}^{\infty} \left| j + \frac{\lambda}{2\pi} \right|^{-2-2d}, \quad (2.10)$$

where $\sigma^2 = E(z_t - E(z_t))^2 < \infty$ and $\Gamma(\cdot)$ denotes the gamma function. From (2.9) and (2.10) we observe that they do not represent the same model, although as $\lambda \rightarrow 0+$, their spectral density functions behave as $K\lambda^{-2d}$ with $K \in (0, \infty)$. For a review of these models, see for instance Beran's (1994) monograph.

Theorem 2.1. *Assuming A.1-A.6, for any finite collection j_1, \dots, j_q , as $T \rightarrow \infty$,*

$$(i) \quad T^{1/2} \left(\widehat{\theta}_{j_1} - \theta_{j_1}, \dots, \widehat{\theta}_{j_q} - \theta_{j_q} \right) \xrightarrow{d} N \left(0, \Omega = (\Omega_{j_k, j_\ell})_{k, \ell=1, \dots, q} \right) \text{ where}$$

$$\Omega_{j_k, j_\ell} = (2\pi)^{-1} \int_{-\pi}^{\pi} f_{xx}^{-1}(\lambda) f_{uu}(\lambda) e^{i(j_k - j_\ell)\lambda} d\lambda, \quad k, \ell = 1, \dots, q,$$

denotes the indicated element of Ω , which corresponds to the asymptotic covariance matrix between $\widehat{\theta}_{j_k}$ and $\widehat{\theta}_{j_\ell}$.

(ii) Let $\widehat{f}_{uu,2ms} = \widehat{f}_{yy,2ms} - \widehat{f}_{xx,2ms}^{-1} \left| \widehat{f}_{yx,2ms} \right|^2$. A consistent estimator of Ω_{j_k, j_ℓ} , $k, \ell = 1, \dots, q$, is

$$\widehat{\Omega}_{j_k, j_\ell} = \frac{1}{2M} \sum_{s=1}^{2M-1} \widehat{f}_{xx,2ms}^{-1} \widehat{f}_{uu,2ms} e^{i(j_k - j_\ell)\lambda_{2ms}}.$$

Proof. See Theorem 1 of Hidalgo (2000). □

3. ORDER SELECTION OF A DISTRIBUTED LAG REGRESSION MODEL

As was mentioned in the introduction, the estimation of p_0 in the model (1.1) is by no means new. This problem was examined by Geweke and Meese (1981) when both x_t and u_t were sequences of independent and identically distributed (*iid*) random variables and by Pötscher (1989) who assumed that x_t followed a stationary (weakly dependent) sequence and u_t a martingale difference sequence. However the assumptions on x_t and u_t are stronger than those we impose in this paper, see Assumptions A.1-A.5. In particular, we want to allow both x_t and u_t to exhibit, possibly, strong dependence. As we discussed in Section 2 and in the introduction, one consequence of strong dependence is that it may induce some undesirable properties on the *LSE* of θ_j and thus on the criteria *BIC/HIC*, e.g. their inconsistency.

Moreover, when p is allowed to increase with T , as is our case, (1.3) has some additional technical problems. In this situation, the proof of (1.5), and the route adopted in Geweke and Meese (1981), is based on Theorem 4.1 of Spitzer (1956). In particular that $\widehat{\sigma}_p^2 - \widehat{\sigma}_{p_0}^2$ can be decomposed as partial sums of quadratic forms, say $\sum_{t=1}^T \widehat{u}_{t,p}^2 - \sum_{t=1}^T \widehat{u}_{t,p-1}^2$, which are (asymptotically) independent identically distributed χ^2 random variables. Unfortunately, even if $T^{1/2}$ -consistency were achieved for the *LSE* $\widehat{\theta}_j$ of θ_j , see Robinson and Hidalgo (1997), and thus the inconsistency of the *BIC/HIC* mentioned in the introduction when $p \leq P$ may disappear, Spitzer's (1956) result could not straightforwardly be used in our framework either. The reason is because under strong dependence, it is well known, see for example Taqqu (1975), that $\sum_{t=1}^T \widehat{u}_{t,p}^2 - \sum_{t=1}^T \widehat{u}_{t,p-1}^2$ may not be independent neither distributed as χ^2 random variables. In general the (asymptotic) distribution is unknown.

Thus, we propose to estimate p_0 in the model (1.1) by the value \widehat{p} or \widehat{p}^* defined

as

$$\widehat{p} = \arg \min_{1 \leq p \leq [\log(T)]} S(p) \quad (3.1)$$

or

$$\widehat{p}^* = \arg \min_{1 \leq p \leq [\log(T)]} S^*(p) \quad (3.2)$$

where

$$S(p) = \frac{1}{2M} \sum_{j=1}^{2M-1} \left\{ \frac{\widehat{f}_{uu,2mj}(p)}{\widehat{f}_{uu,2mj}} - 1 \right\} + \frac{pg(T)}{T}, \quad (3.3)$$

and

$$S^*(p) = \log \left(\frac{1}{2M} \sum_{j=1}^{2M-1} \frac{\widehat{f}_{uu,2mj}(p)}{\widehat{f}_{uu,2mj}} \right) + \frac{pg(T)}{T}, \quad (3.4)$$

with

$$\begin{aligned} \widehat{f}_{uu,2mj}(p) &= \frac{1}{2m+1} \sum_{\ell=-m}^m I_{\widehat{u}\widehat{u}}(\lambda_\ell + \lambda_{2mj}) \quad \text{and} \\ \widehat{f}_{uu,2mj} &= \widehat{f}_{uu,2mj}(\min([\log(T)], M)). \end{aligned} \quad (3.5)$$

Here $I_{\widehat{u}\widehat{u}}(\lambda)$ is as defined in (2.5) but using the residuals $\widehat{u}_t = y_t - \sum_{\ell=1}^p \widehat{\theta}_\ell x_{t-\ell+1}$ instead of w_t , and where $\widehat{\theta}_\ell$ is given in (2.6). Observe that $S(p)$ becomes more in the spirit of (1.3) whereas $S^*(p)$ is closer to the *BIC* or *HIC* criteria given in (1.2).

It is worth giving some intuition behind the criterion $S(p)$. Since by Proposition 3 and Lemma 1 of Hidalgo (2000), $\left(f_{uu,2mj}^{-1} \widehat{f}_{uu,2mj} - (1 + Kj^{-2}) \right) = o_p(1)$ uniformly in j , where henceforth K is some finite positive constant, we can expect that for $p < p_0$, $S(p) - S(p_0)$ will asymptotically be equivalent to

$$\frac{1}{2M} \sum_{j=1}^{2M-1} (1 + Kj^{-2})^{-1} \left\{ \frac{\widehat{f}_{uu,2mj}(p)}{f_{uu,2mj}} - \frac{\widehat{f}_{uu,2mj}(p_0)}{f_{uu,2mj}} \right\} + \frac{(p - p_0)g(T)}{T}.$$

But, Theorem 2.1 implies that $\widehat{\theta}_\ell - \theta_\ell = O_p(T^{-1/2})$ irrespective of the number of lags p included in the model (1.1). So, the last displayed expression is approximated by

$$\begin{aligned} & \frac{1}{2M} \sum_{j=1}^{2M-1} (1 + Kj^{-2})^{-1} \sum_{\ell_1, \ell_2=p+1}^{p_0} \theta_{\ell_1} \theta_{\ell_2} e^{i(\ell_1 - \ell_2)\lambda_{2mj}} \frac{f_{xx,2mj}}{f_{uu,2mj}} \\ & + \frac{(p - p_0)g(T)}{T} + o_p(1) \\ & = \sum_{\ell_1, \ell_2=p+1}^{p_0} \theta_{\ell_1} \widehat{\Upsilon}(|\ell_1 - \ell_2|) \theta_{\ell_2} + \frac{(p - p_0)g(T)}{T} + o_p(1), \end{aligned}$$

where

$$\widehat{\Upsilon}(\ell) = \frac{1}{2M} \sum_{j=1}^{2M-1} e^{i\ell\lambda_{2mj}} \frac{f_{xx,2mj}}{f_{uu,2mj}} (1 + Kj^{-2})^{-1}.$$

However the right side of the last displayed equation, (see the proof of Theorem 3.1 in Section 5), converges uniformly in ℓ to the ℓ th Fourier coefficient of $f_{uu}^{-1}(\lambda) f_{xx}(\lambda)$, say $\Upsilon(\ell)$. Thus, for T large enough, $S(p) - S(p_0) > 0$ which implies that the minimum of (3.3) cannot be achieved for any value $p < p_0$. In Theorem 3.1 below we show that the minimum is not achieved either for any value $p > p_0$ as $T \rightarrow \infty$. The same arguments apply to $S^*(p)$ since $\log(z)$ is a monotonic increasing function.

It should be noted that if u_t were a white noise process, then

$$\begin{aligned} S(p) &\approx \frac{1}{2M} \sum_{j=1}^{2M-1} \left\{ \frac{\widehat{f}_{uu,2mj}(p)}{(2\pi)^{-1} \sigma_u^2} - 1 \right\} + \frac{pg(T)}{T} \\ &= \sigma_u^{-2} \left(\frac{1}{2M} \sum_{j=1}^{2M-1} \left\{ (2\pi) \widehat{f}_{uu,2mj}(p) - \sigma_u^2 \right\} \right) + \frac{pg(T)}{T} \\ &= \sigma_u^{-2} \left(\frac{1}{2M} \sum_{j=1}^{2M-1} (2\pi) \widehat{f}_{uu,2mj}(p) + \frac{pg(T) \sigma_u^2}{T} \right) + K. \end{aligned}$$

But $(2M)^{-1} (2\pi) \sum_{j=1}^{2M-1} \widehat{f}_{uu,2mj}(p) \approx \int_0^{2\pi} \widehat{f}_{uu}(\lambda; p) d\lambda$, which is an estimator of $\sigma_u^2 = \int_0^{2\pi} f_{uu}(\lambda) d\lambda$. Thus, when u_t is a white noise sequence, the criterion $S(p)$ is just a reformulation of that given in (1.3). Proceeding in the same form, if $S^*(p)$ given in (3.4) then it would be just a reformulation of the more familiar *BIC* or *HIC* criteria given in (1.2).

Theorem 3.1. *Assuming A.1-A.6 and the model (1.1), if $|g^{-1}(T) \log(T)| < K$, as $T \rightarrow \infty$, \widehat{p} given in (3.1) satisfies*

$$\widehat{p} \xrightarrow{P} p_0.$$

We now elaborate on the results of Theorem 3.1. First, it is worth noting that unlike Geweke and Meese (1981) or Pötscher (1989, 1991), the upper bound for p_0 is not assumed to be finite, but the upper bound of admissible values for p_0 is allowed to increase to infinity at the rate $\log T$, as in An et al. (1982). Second, a closer inspection of the proofs indicates that the results of the theorem hold the same if $\log T$ is replaced by $\log^{1+\delta} T$ for any $0 \leq \delta < 1$. Finally, the results of the theorem also hold true if $\log T$ were replaced by $\log \log T$ in the definition of $S(p)$ but in this case p would have to be restricted to $1 \leq p \leq [\log \log T] + 1$.

We now examine \widehat{p}^* in (3.2).

Theorem 3.2. *Assuming A.1-A.6 and the model (1.1), if $|g^{-1}(T) \log(T)| < K$, as $T \rightarrow \infty$, \widehat{p} given in (3.2) satisfies*

$$\widehat{p}^* \xrightarrow{P} p_0.$$

As we mentioned in the introduction, we finish this section by describing how an efficient (adaptive) estimator of θ_j can be obtained. To that end, assume first, that p_0 is known. Then, following Hannan (1963), the (adaptive) *GLS* estimator of $\beta = (\theta_1, \dots, \theta_{p_0})'$ is given by

$$\begin{aligned} \widetilde{\beta} &= \left\{ \frac{1}{2M} \sum_{k=0}^{2M-1} \widehat{f}_{\widehat{u}\widehat{u},2mk}^{-1} \widehat{f}_{\widetilde{x}\widetilde{x},2mk} \right\}^{-1} \left\{ \frac{1}{2M} \sum_{k=0}^{2M-1} \widehat{f}_{\widehat{u}\widehat{u},2mk}^{-1} \widehat{f}_{\widetilde{x}y,2mk} \right\} \\ &= \beta + \left\{ \frac{1}{2M} \sum_{k=0}^{2M-1} \widehat{f}_{\widehat{u}\widehat{u},2mk}^{-1} \widehat{f}_{\widetilde{x}\widetilde{x},2mk} \right\}^{-1} \left\{ \frac{1}{2M} \sum_{k=0}^{2M-1} \widehat{f}_{\widehat{u}\widehat{u},2mk}^{-1} \widehat{f}_{\widetilde{x}u,2mk} \right\} \end{aligned} \quad (3.6)$$

where $\widehat{u}_t = y_t - \sum_{j=1}^{p_0} x_{t-j+1} \widehat{\theta}_j$, $\widehat{\theta}_j$ as defined in (2.6) and $\widehat{f}_{\widetilde{x}\widetilde{x},2mk}$ is the estimator of the spectral matrix function of $\widetilde{x} = (x_t, \dots, x_{t-p_0+1})'$, $f_{\widetilde{x}\widetilde{x},2mk}$, given in (2.4) but with \widetilde{x}_t replacing w_t . The reason why we have chosen $\widehat{\theta}_j$ instead of the *LSE* of β is due to our comments in Section 2. Alternatively, following Robinson and Hidalgo (1997), we could have estimated β by the weighted least squares estimator in the frequency domain

$$\widehat{\beta}_\phi = \left(\sum_{j=1}^{n-1} I_{\widetilde{x}\widetilde{x},j} \phi_j \right)^{-1} \left(\sum_{j=1}^{n-1} I_{\widetilde{x}y,j} \phi_j \right),$$

where $\phi(\lambda)$ is real-valued, even and periodic of period 2π function such that $\phi(\lambda) \sim K\lambda$ as $\lambda \rightarrow 0+$ and " \sim " means that the ratio of the left- and right hand side tends to one. An example of a function $\phi(\lambda)$ satisfying the aforementioned properties is $\phi(\lambda) = |1 - e^{i\lambda}|$. However since both $\widehat{\beta}_\phi$ and $\widehat{\theta}_j$ in (2.6) are $T^{1/2}$ -consistent and the latter has already been computed for the estimation of p_0 , we have decided to use the latter estimator. We then have the following result,

Theorem 3.3. *Assuming A.1-A.6, as $T \rightarrow \infty$*

$$T^{1/2} \left(\tilde{\beta} - \beta \right) \xrightarrow{d} \mathcal{N} \left(0, \Omega = (\Omega_{j,k})_{j,k=1,\dots,p_0} \right)$$

where $\tilde{\beta} = \left(\tilde{\theta}_1, \dots, \tilde{\theta}_{p_0} \right)'$ is given in (3.6) and

$$\Omega_{j,k} = (2\pi)^{-1} \int_{-\pi}^{\pi} f_{uu}^{-1}(\lambda) f_{xx}(\lambda) e^{i(j-k)\lambda} d\lambda, \quad j, k = 1, \dots, p_0,$$

denotes the indicated element of Ω , which corresponds to the asymptotic covariance matrix between $\hat{\theta}_j$ and $\hat{\theta}_k$.

Proof. The proof follows immediately from Hidalgo and Robinson's (2001) Theorem 2.1, so it is omitted. \square

We now define the "adaptive" estimator. Let \hat{p} be the estimator of p_0 in (3.1). Then we define the "adaptive" estimator by

$$\check{\beta} = \left\{ \frac{1}{2M} \sum_{k=0}^{2M-1} \hat{f}_{\hat{u}\hat{u},2mk}^{-1} \hat{f}_{\tilde{x}\tilde{x},2mk} \right\}^{-1} \left\{ \frac{1}{2M} \sum_{k=0}^{2M-1} \hat{f}_{\hat{u}\hat{u},2mk}^{-1} \hat{f}_{\tilde{x}y,2mk} \right\}, \quad (3.7)$$

where now $\beta = (\theta_1, \dots, \theta_{\hat{p}})'$ and $\tilde{x}_t = (x_t, \dots, x_{t-\hat{p}+1})'$. With the results of the previous theorem and Pötscher (1991), we achieve the following

Theorem 3.4. *Assuming A.1-A.6, for any finite collection $j_1, \dots, j_q \in (1, \dots, \hat{p})$, as $T \rightarrow \infty$*

$$T^{1/2} \left(\check{\theta}_{j_1} - \theta_{j_1}, \dots, \check{\theta}_{j_q} - \theta_{j_q} \right) \xrightarrow{d} \mathcal{N} \left(0, \Omega = (\Omega_{j_k, j_\ell})_{k,\ell=1,\dots,q} \right)$$

where $\check{\theta}_j$ is the j th element of $\check{\beta}$ in (3.7), and

$$\Omega_{j_k, j_\ell} = (2\pi)^{-1} \int_{-\pi}^{\pi} f_{uu}^{-1}(\lambda) f_{xx}(\lambda) e^{i(j_k - j_\ell)\lambda} d\lambda, \quad k, \ell = 1, \dots, q$$

denotes the indicated element of Ω , which corresponds to the asymptotic covariance matrix between $\check{\theta}_{j_k}$ and $\check{\theta}_{j_\ell}$.

Proof. The proof follows as a consequence of Theorem 3.3 and Lemma 1 of Pötscher (1991). \square

So the previous theorem indicates that indeed the asymptotic distribution of our estimates of the parameters $\theta_j, \check{\theta}_j$, for $j \leq p_0$, in (3.7) are not affected by the estimation of p_0 , e.g. they are efficient in a Gauss-Markov sense. In addition, it is obvious that the results of Theorem 3.4 hold true if \hat{p} were replaced by \hat{p}^* .

Remark 1. It should be indicated that the results of the last two theorems hold true if $\tilde{\beta}$ and $\check{\beta}$ were replaced, respectively, by

$$\widehat{\beta}_{\widehat{f}} = \left(\sum_{j=1}^{n-1} I_{\widehat{x}\widehat{x},j} \widehat{f}_{\widehat{u}\widehat{u},j}^{-1} \right)^{-1} \left(\sum_{j=1}^{n-1} I_{\widehat{x}y,j} \widehat{f}_{\widehat{u}\widehat{u},j}^{-1} \right)$$

where $\beta = (\theta_1, \dots, \theta_{p_0})'$ and $\tilde{x}_t = (x_t, \dots, x_{t-p_0+1})'$ and

$$\check{\beta}_{\widehat{f}} = \left\{ \frac{1}{n} \sum_{j=1}^{n-1} I_{\check{x}\check{x},j} \widehat{f}_{\widehat{u}\widehat{u},j}^{-1} \right\}^{-1} \left\{ \sum_{j=1}^{n-1} I_{\check{x}y,j} \widehat{f}_{\widehat{u}\widehat{u},j}^{-1} \right\},$$

where now $\beta = (\theta_1, \dots, \theta_{\widehat{p}})'$ and $\tilde{x}_t = (x_t, \dots, x_{t-\widehat{p}+1})'$, or $\beta = (\theta_1, \dots, \theta_{\widehat{p}^*})'$ and $\tilde{x}_t = (x_t, \dots, x_{t-\widehat{p}^*+1})'$.

4. NUMERICAL EXAMPLE

In order to investigate how well \widehat{p} and \widehat{p}^* given in (3.1) and (3.2) perform in finite samples and its relative performance compared with \tilde{p} provided in (1.4) with P replaced by $[\log(T)]$ there and $\widehat{p}^* = \arg \min_{1 \leq p \leq [\log(T)]} \overline{S}^*(p)$, a small Monte-Carlo study was carried out. The penalty functions $g(T)$ were $g(T) = \log T$ and $g(T) = 2.0 \log(\log T)$, corresponding to those employed in the criteria *BIC* and *HIC* respectively. In Tables 4.1-4.3 below, *METHOD 1* and *2* correspond to the *BEC* type of criterion, e.g. (1.3) / (3.3), using $g(T) = \log T$ and $g(T) = 2.0 \log(\log T)$ respectively, whereas *METHOD 3* and *4* employed the more standard *BIC* and *HIC* criteria. Note that we have used $c = 1$ for the *HIC*. This is so, as Hannan and Quinn (1979) mentioned, it would seem pedantic for the values of T used in the simulations. All throughout the simulations, we have generated the model

$$y_t = x_t + 0.5x_{t-1} + 0.25x_{t-2} + u_t \quad t = 1, \dots, T$$

which corresponds to $p_0 = 3$ in (1.1), for sample sizes $T = 64, 128$ and 256 , and where x_t and u_t were generated as mutually independent Gaussian *ARFIMA* models (2.9) with $\Theta(L) = \Phi(L) = 1$ and $d = d_x$ and $d = d_u$ respectively. We considered three different values of d_x and d_u , specifically $d_x = 0.15, 0.25$ and 0.35 and $d_u = 0.15, 0.25$ and 0.35 .

The results of the experiment for the nine possible combinations of d_x and d_u are given in Tables 4.1 to 4.3, where PAR denotes the estimator of p_0 when $\overline{S}^*(p)$ and $\overline{S}(p)$ in (1.2) and (1.3), respectively, was used, and NP1 and NP2, the estimator of p_0 when the criterion function used was $S(p)$ or $S^*(p)$, given in (3.3) or (3.4), with $M = T/8$ and $M = T/4$, respectively. Finally, for each sample size and different combination of d_x and d_u , we performed 1000 Monte Carlo samples.

A quick inspection of Tables 4.1- 4.3 illustrates that the choice of the penalty function $g(T) = \log T$ underestimate the true order more often than when we employed $g(T) = 2.0 \log(\log T)$ across all sample sizes, for all different combinations of d_x and d_u . This is somehow not surprising, as we mentioned in the introduction, since the motivation to use the *HIC* criterion was precisely for this reason. On the other hand, it seems that the number of times that the criteria estimate the true order is not affected by the penalty function $g(T)$ employed, except when the criterion given in (1.2) is employed. A second general feature is that NP2 performs better than NP1, although that difference becomes smaller as the sample size increases. In particular, for $T = 256$, the difference is negligible.

Regarding the relative finite sample performance of \hat{p} and \hat{p}^* to \tilde{p} and \tilde{p}^* respectively, it appears to be that our proposed criterion function enjoys better finite sample properties in two aspects. First, the criteria (3.3) / (3.4) tend to overestimate the true value p_0 to a lesser degree than (1.3) / (1.2), and second, the number of times that the criteria (3.3) / (3.4) correctly estimate p_0 , that is, $\hat{p}_0 = 3$, is greater, especially when T is small and $M = T/4$. This may indicate that the finite sample distribution of \hat{p} or \hat{p}^* is more concentrated around p_0 than that of \tilde{p} or \tilde{p}^* . This difference, however, appears to be greater when $d_x + d_u > 0.5$, which is the situation where the criteria $S(p)$ or $S^*(p)$ may have some adverse properties, as was mentioned in the introduction. However, when comparing \hat{p}^* with \tilde{p}^* , the former underestimate the true order less often than the latter, whereas when \hat{p} is compared with \tilde{p} , it becomes the other way round for moderate sample sizes. Overall, Tables 4.1 to 4.3 induce us to think that the proposed criteria functions (3.3) / (3.4) possess very good finite sample properties, and more importantly, that even for moderate sample sizes, say $T = 128$, the choice of M does not appear to play a crucial or important role, so far the estimation of p_0 is concerned.

5. PROOF OF THEOREMS 3.1 AND 3.2

5.1. Proof of Theorem 3.1

Because $\hat{p} = \arg \min_{1 \leq p \leq \lfloor \log(T) \rfloor} S(p)$, it implies that $S(\hat{p}) - S(p_0) < 0$, or equivalently $S(p_0) - S(\hat{p}) > 0$. On the other hand, from the definition of $S(p)$, we obtain that

$$S(p_0) - S(p) = \frac{1}{2M} \sum_{q=1}^{2M-1} \left\{ \frac{\hat{f}_{uu,2mq}(p_0)}{\hat{f}_{uu,2mq}} - \frac{\hat{f}_{uu,2mq}(p)}{\hat{f}_{uu,2mq}} \right\} + \frac{(p_0 - p) g(T)}{T}$$

$$= \frac{1}{2M} \sum_{q=1}^{2M-1} \frac{f_{uu,2mq}}{\widehat{f}_{uu,2mq}} \left\{ \frac{\widehat{f}_{uu,2mq}(p_0) - \widehat{f}_{uu,2mq}(p)}{f_{uu,2mq}} \right\} + \frac{(p_0 - p)g(T)}{T}.$$

We begin showing that $\Pr \{\widehat{p} > p_0\} \rightarrow 0$. Denote $\widehat{p} = p_0 + j$ for some $j > 0$. Using the inequality

$$\Pr \{\widehat{p} > p_0\} \leq \Pr \left\{ \sup_{1 \leq j \leq [\log(T)] - p_0} S(p_0) - S(p_0 + j) > 0 \right\},$$

it suffices to show that the right side of the last displayed inequality converges to zero. First, observing that by Proposition 3 of Hidalgo (2000),

$$\sup_{q=1, \dots, M} \left| f_{uu,2mq}^{-1} \widehat{f}_{uu,2mq} - (1 + Kq^{-2}) \right| = O_p(MT^{-1/2}) \quad (5.1)$$

it implies that it suffices to show that

$$\Pr \left\{ \sup_{1 \leq j \leq [\log(T)] - p_0} \widetilde{S}(p_0, j) > 0 \right\} \rightarrow 0, \quad (5.2)$$

where

$$\widetilde{S}(p_0, j) = \frac{1}{2M} \sum_{q=1}^{2M-1} \left(1 - \frac{K}{K + q^2} \right) \left\{ \frac{\widehat{f}_{uu,2mq}(p_0) - \widehat{f}_{uu,2mq}(p_0 + j)}{f_{uu,2mq}} \right\} - \frac{jg(T)}{T}.$$

Let us introduce

$$\widehat{b}(k) = (2M)^{-1} \sum_{q=1}^{2M-1} \left(1 - \frac{K}{K + q^2} \right) f_{uu,2mq}^{-1} \widehat{f}_{uu,2mq} e^{ik\lambda_{2mq}} \quad (5.3)$$

and

$$\widehat{\Upsilon}(k) = (2M)^{-1} \sum_{q=1}^{2M-1} \left(1 - \frac{K}{K + q^2} \right) f_{uu,2mq}^{-1} \widehat{f}_{xx,2mq} e^{ik\lambda_{2mq}}.$$

It is worth observing that since (5.1) also holds true with u_t replaced by x_t , and

$$(2M)^{-1} \sum_{q=1}^{2M-1} f_{uu,2mq}^{-1} f_{xx,2mq} e^{ik\lambda_{2mq}} \rightarrow \frac{1}{2\pi} \int_{-\pi}^{\pi} f_{uu}^{-1}(\lambda) f_{xx}(\lambda) e^{ik\lambda} d\lambda = \Upsilon(k), \quad (5.4)$$

it implies that as $M \rightarrow \infty$, $\widehat{\Upsilon}(k) \xrightarrow{p} \Upsilon(k)$ uniformly in k .

Because $\widehat{\theta}_j$ given in (2.6) is the same irrespective of the number of lags specified in (1.1), after some straightforward algebra, we conclude that

$$\widetilde{S}(p_0, j) = \Psi(j) - \frac{jg(T)}{T}$$

where

$$\begin{aligned} \Psi(j) &= -2 \sum_{k=p_0+1}^{p_0+j} \widehat{\theta}_k \widehat{b}(k-p_0) - 2 \sum_{k_1=1}^{p_0} \left(\widehat{\theta}_{k_1} - \theta_{k_1} \right) \sum_{k_2=p_0+1}^{p_0+j} \widehat{\Upsilon}(|k_1 - k_2|) \widehat{\theta}_{k_2} \\ &\quad - \sum_{k_1, k_2=p_0+1}^{p_0+j} \widehat{\theta}_{k_1} \widehat{\Upsilon}(|k_1 - k_2|) \widehat{\theta}_{k_2} \end{aligned}$$

But as the third term on the right of the last displayed equation is a quadratic form, then it is less than or equal to zero with probability 1, and thus (5.2) holds true if

$$\Pr \left\{ \sup_{1 \leq j \leq [\log(T)] - p_0} \sum_{k=p_0+1}^{p_0+j} \widehat{\theta}_k \widehat{b}(k-p_0) - \frac{jg(T)}{T} > 0 \right\} \rightarrow 0 \quad (5.5)$$

and

$$\Pr \left\{ \sum_{k_1=1}^{p_0} \left(\widehat{\theta}_{k_1} - \theta_{k_1} \right) \sup_{1 \leq j \leq [\log(T)] - p_0} \sum_{k_2=p_0+1}^{p_0+j} \widehat{\Upsilon}(|k_1 - k_2|) \widehat{\theta}_{k_2} - \frac{jg(T)}{T} > 0 \right\} \rightarrow 0. \quad (5.6)$$

We begin showing (5.5), whose left side is upper bounded by

$$\Pr \left\{ \sup_{1+p_0 \leq k \leq [\log(T)]} T \widehat{\theta}_k \widehat{b}(k-p_0) > g(T) \right\}. \quad (5.7)$$

First, from the proof of Theorem 2.1, see Hidalgo (2000), for $M \geq k > p_0$,

$$\begin{aligned} T^{1/2} \widehat{\theta}_k &= \frac{T^{1/2}}{2M} \sum_{q=1}^{2M-1} f_{xx,2mq}^{-1} \widehat{f}_{ux,2mq} e^{ik\lambda_{2mq}} + O_p \left(\frac{\log M}{M^{1/2}} \right) \\ &= \bar{\theta}_k + O_p \left(\frac{\log M}{M^{1/2}} \right), \end{aligned} \quad (5.8)$$

where the second term on the right of (5.8) is uniform in k . (Recall that the statistical properties of $\widehat{\theta}_k$ are independent of k and/or the number of lags assumed in (1.1).) So, we obtain that (5.7) is upper bounded by

$$\Pr \left\{ \sup_{1+p_0 \leq k \leq [\log(T)]} \bar{\theta}_k T^{1/2} \widehat{b}(k-p_0) > g(T) \right\}. \quad (5.9)$$

Next, from the definitions of $\widehat{b}(k)$ and $\bar{\theta}_k$ given in (5.3) and (5.8) respectively, it is easily shown that

$$E \left(\bar{\theta}_k T^{1/2} \widehat{b}(k - p_0) \right) \rightarrow \frac{1}{2M} \sum_{q=1}^{2M-1} e^{ip_0 \lambda_{2mq}} = 0,$$

and since both $\bar{\theta}_k$ and $T^{1/2} \widehat{b}(k - p_0)$ converge to a normal random variable, Assumption A.1 and Serfling's (1980, p.14) Theorem A imply that

$$E \left(\bar{\theta}_k T^{1/2} \widehat{b}(k - p_0) \right)^2 \leq K,$$

as we now show. Indeed, the latter is true since the second moments of $\bar{\theta}_k T^{1/2} \widehat{b}(k - p_0)$ converge to those of the limiting distribution by Serfling's (1980, p.14) Theorem A provided that the sequence of random variables is uniform integrable which is the case since by A.1 x_t^4 and u_t^4 are uniformly integrable. From here, using the inequality $\sup_k |c_k|^2 \leq \sum_k |c_k|^2$ and Markov inequality, (5.9) is upper bounded by $K \sum_{k=1}^{\lfloor \log(T) \rfloor} g^{-2}(T)$ which converges to 0 since from the assumption on $g(T)$, $|g^{-1}(T) \log T| < K$.

On the other hand, since by Theorem 2.1, $\sum_{k=1}^{p_0} T^{1/2} (\widehat{\theta}_k - \theta_k) = O_p(1)$, it implies that the left side of (5.6) is bounded by

$$\Pr \left\{ \sup_{1+p_0 \leq k \leq \lfloor \log(T) \rfloor} T^{1/2} \widehat{\theta}_k > g(T) \right\} \leq K \sum_{k=1}^{\lfloor \log(T) \rfloor} g^{-2}(T) \rightarrow 0$$

proceeding similarly as with the proof of (5.5). Thus, we have completed the proof that $\Pr \{\widehat{p} > p_0\} \rightarrow 0$.

To finish the proof of the theorem, it remains to show that $\Pr \{\widehat{p} < p_0\} \rightarrow 0$. Because $p_0 \leq K < \infty$, it suffices to show that for all $1 \leq j < p_0$,

$$\Pr \{S(p_0 - j) - S(p_0) < 0\} \rightarrow 0.$$

The left side of the last displayed expression is

$$\Pr \left\{ \widehat{A}_1(j) + \widehat{A}_2(j) + \widehat{A}_3(j) < \frac{jg(T)}{T^{1/2}} \right\}, \quad (5.10)$$

where

$$\widehat{A}_1(j) = T^{1/2} \sum_{k_1, k_2 = p_0 - j + 1}^{p_0} \left\{ \widehat{\theta}_{k_1} \widehat{\Upsilon}(|k_1 - k_2|) \widehat{\theta}_{k_2} \right\},$$

$$\begin{aligned}\widehat{A}_2(j) &= 2 \sum_{k=p_0-j+1}^{p_0} \widehat{\theta}_k T^{1/2} \widehat{b}(k+j-p_0) \quad \text{and} \\ \widehat{A}_3(j) &= 2 \sum_{k_1=1}^{p_0-1} T^{1/2} \left(\widehat{\theta}_{k_1} - \theta_{k_1} \right) \sum_{k_2=q_0-j}^{q_0} \widehat{\Upsilon}(|k_1-k_2|) \widehat{\theta}_{k_2}.\end{aligned}$$

But

$$\begin{aligned}T^{1/2} \left(\widehat{\theta}_{\ell_1} - \theta_{\ell_1}, \dots, \widehat{\theta}_{\ell_r} - \theta_{\ell_r} \right)' &\xrightarrow{d} (\omega(\ell_1), \dots, \omega(\ell_r))' \cong N(0, \Omega), \\ T^{1/2} \left(\widehat{b}(\ell_1), \dots, \widehat{b}(\ell_r) \right)' &\xrightarrow{d} (\tau(\ell_1), \dots, \tau(\ell_r))' \cong N(0, \Upsilon), \text{ and} \\ \sum_{k_1, k_2=p_0-j}^{p_0} \left\{ \widehat{\theta}_{k_1} \widehat{\Upsilon}(|k_1-k_2|) \widehat{\theta}_{k_2} \right\} &\xrightarrow{P} \psi > 0\end{aligned}$$

because $\widehat{\Upsilon}(|k_1-k_2|) \xrightarrow{P} \Upsilon(|k_1-k_2|)$, $\Upsilon = (\Upsilon(|k_1-k_2|))_{k_1, k_2=p_0-j, \dots, p_0}$ is a positive definite matrix, c.f. (5.4), and $\left(\widehat{\theta}_{p_0-j}, \dots, \widehat{\theta}_{p_0} \right)$ converges in probability to a non-zero vector. Hence

$$\lim_{T \rightarrow \infty} \Pr \left\{ T^{-1/2} \left(\widehat{A}_1(j) + \widehat{A}_2(j) + \widehat{A}_3(j) \right) > K \right\} \rightarrow 1,$$

which implies that (5.10) converges to 0 since $T^{-1}g(T) \rightarrow 0$. So, $\Pr \{ \widehat{p} < p_0 \} \rightarrow 0$ and the proof of the theorem is completed. \square

5.2. Proof of Theorem 3.2

Because $\widehat{p}^* = \arg \min_{1 \leq p \leq \lfloor \log(T) \rfloor} S^*(p)$, it implies that $S^*(\widehat{p}) - S^*(p_0) < 0$, or equivalently $S^*(p_0) - S^*(\widehat{p}) > 0$.

We begin showing that $\Pr \{ \widehat{p}^* > p_0 \} \rightarrow 0$. Denote $\widehat{p}^* = p_0 + j$ for some $j > 0$. Using the arguments of Theorem 3.1, it suffices to show that

$$\Pr \{ \widehat{p}^* > p_0 \} \leq \Pr \left\{ \sup_{1 \leq j \leq \lfloor \log(T) \rfloor - p_0} S^*(p_0) - S^*(p_0 + j) > 0 \right\}. \quad (5.11)$$

Denoting $h(p) = (2M)^{-1} \sum_{j=1}^{2M-1} \widehat{f}_{uu, 2mj}^{-1} \widehat{f}_{uu, 2mj}(p)$,

$$\begin{aligned}S^*(p_0) - S^*(p_0 + j) &= \log \left(\frac{h(p_0)}{h(p_0 + j)} \right) - \frac{jg(T)}{T} \\ &\leq \frac{h(p_0) - h(p_0 + j)}{h(p_0 + j)} - \frac{jg(T)}{T}\end{aligned}$$

because $\log(x) \leq x - 1$. So that the right side of (5.11) is bounded by

$$\Pr \left\{ \sup_{1 \leq j \leq [\log(T)] - p_0} \frac{h(p_0) - h(p_0 + j)}{h(p_0 + j)} - \frac{jg(T)}{T} > 0 \right\}. \quad (5.12)$$

On the other hand, we have that

$$\begin{aligned} h(p_0 + j) &= h(p_0) - (h(p_0) - h(p_0 + j)) \\ &= h(p_0) - \frac{jg(T)}{T} - \left(h(p_0) - h(p_0 + j) - \frac{jg(T)}{T} \right). \end{aligned}$$

But, with probability approaching one

$$\begin{aligned} \inf_j h(p_0 + j) &= h(p_0) - \frac{jg(T)}{T} - \sup_j \left(h(p_0) - h(p_0 + j) - \frac{jg(T)}{T} \right) \\ &\geq h(p_0) - \frac{jg(T)}{T} \end{aligned}$$

since by Theorem 3.1., $\Pr \left\{ h(p_0) - h(p_0 + j) - \frac{jg(T)}{T} > 0 \right\} \rightarrow 0$. Since $h(p_0) \xrightarrow{P} 1$, we have that $\inf_j h(p_0 + j) > 1/2$. In addition, by standard arguments, we have that $h(p_0 + j) \xrightarrow{P} 1$. Hence with probability approaching one, $1/2 < \inf_j h(p_0 + j) < 2$, which implies that the behaviour of (5.12) is that of

$$\Pr \left\{ \sup_{1 \leq j \leq [\log(T)] - p_0} h(p_0) - h(p_0 + j) - \frac{jg(T)}{T} > 0 \right\},$$

which converges to zero by Theorem 3.1. Thus, $\Pr \{\widehat{p}^* > p_0\} \rightarrow 0$.

The proof that $\Pr \{\widehat{p}^* < p_0\} \rightarrow 0$, follows by similar steps to those of Theorem 3.1, and it is omitted. \square

References

- [1] AKAIKE, H. (1974): "A new look at the Statistical Model Identification," *IEEE Transactions on Automatic Control*, AC-19, 716-723.
- [2] AN, H.Z., CHEN, Z.G. AND HANNAN, E.J. (1982): "Autocorrelation, Autoregression and Autoregressive Approximation," *Annals of Statistics*, 10, 926-936.
- [3] BERAN, J. (1994): *Statistics for Long-Memory Processes*. Chapman and Hall. London.
- [4] BERAN, J., BHANSALI, R.J. AND OCKER, D. (1998): "On Unified Model Selection for Stationary and Nonstationary Short- and Long-Memory Autoregressive Processes," *Biometrika*, 85, 921-934.
- [5] BRILLINGER, D.R. (1981): *Time Series, Data Analysis and Theory*. San Francisco: Holden-Day.
- [6] GEORGE, E.I. (2000): "The Variable Selection Problem," *Journal of the American Statistical Association*, 95, 1304-1308.
- [7] GEWEKE, J. AND R. MEESE (1981): "Estimating Regression Models of Finite but Unknown Order," *International Economic Review*, 22, 55-70.
- [8] GRANGER, C.W.J. AND R. JOYEUX (1980): "An Introduction to Long Memory Time Series and Fractional Differencing," *Journal of Time Series Analysis*, 1, 15-30.
- [9] HANNAN, E.J. (1963): "Regression for Time Series," In M. Rosenblatt (ed.), *Time Series Analysis*, 17-37. New York: John Wiley.
- [10] HANNAN, E.J. (1967): "The Estimation of a Lagged Regression Relation," *Biometrika*, 54, 409-418.
- [11] HANNAN, E.J. (1980): "The Estimation of the Order of an ARMA Process," *Annals of Statistics*, 8, 1071-1081.
- [12] HANNAN, E.J. AND B.G. QUINN (1979): "The Determination of the Order of an Autoregression," *Journal of the Royal Statistical Society, Ser. B*, 41, 190-195.
- [13] HIDALGO, J. (2000): "Nonparametric Test for Causality with Long-Range Dependence," *Econometrica*, 68, 1465-1490.

- [14] HIDALGO, J. (2001): "Order Selection with Strongly Dependent Data," Preprint.
- [15] HIDALGO, J. AND P.M. ROBINSON (2001): "Adapting to Unknown Error Autocorrelation in Regression with Long Memory," Forthcoming in *Econometrica*.
- [16] HOSKING, J. (1981): "Fractional Differencing," *Biometrika*, 68, 165-176.
- [17] KAVALIERIS, L. AND HANNAN, E.J. (1994): "Determining the Number of Terms in a Trigonometric Regression," *Journal of Time Series Analysis*, 15, 613-625.
- [18] MANDELBROT, B.B. AND J.W. VAN NESS (1968): "Fractional Brownian Motions, Fractional Noises and Applications," *SIAM Review*, 10, 422-437.
- [19] PÖTSCHER, B.M. (1989): "Model Selection under Nonstationarity: Autoregressive Models and Stochastic Linear Regression Models," *Annals of Statistics*, 17, 1257-1274.
- [20] PÖTSCHER, B.M. (1991): "Effects of Model Selection on Inference," *Econometric Theory*, 7, 163-185.
- [21] ROBINSON, P.M. (1979): "Distributed Lag Approximation to Linear Time-Invariant Systems," *Annals of Statistics*, 7, 507-515.
- [22] ROBINSON, P.M. (1994): "Time Series with Strong Dependence," In C.A. Sims, ed., *Advances in Econometrics: Sixth World Congress*, Vol.1, 47-95. Cambridge: Cambridge University Press.
- [23] ROBINSON, P.M. (1995): "Gaussian Semiparametric Estimation of Long-Range Dependence," *Annals of Statistics*, 23, 1630-1661.
- [24] ROBINSON, P.M. AND J. HIDALGO (1997): "Time Series Regression with Long Range Dependence," *Annals of Statistics*, 25, 77-104.
- [25] SCHWARZ, G. (1978): "Estimating the Dimension of a Model," *Annals of Statistics*, 6, 461-464.
- [26] SHIBATA, R. (1976): "Selection of the Order of an Autoregressive Model by Akaike's Information Criterion," *Biometrika*, 63, 117-126.
- [27] SHIBATA, R. (1980): "Asymptotic Efficiency Selection of the Order of the Model for Estimating Parameters of a Linear Process," *Annals of Statistics*, 8, 147-164.

- [28] SHIBATA, R. (1981): "An Optimal Selection of Regression Variables," *Biometrika*, 68, 45-54.
- [29] SIMS, C.A. (1974): "Distributed Lags," *In Frontiers of Quantitative Economics*, Vol. II (M.D. Intrilligator and D.A. Kendrick, eds.). Amsterdam: North Holland.
- [30] SINAI, Y.G. (1976): "Self-Similar Probability Distribution," *Theory of Probability and its Applications*, 21, 64-80.
- [31] SPITZER, F. (1956): "A Combinatorial Lemma and its Application to Probability Theory," *Transactions of the American Mathematical Society*, 82, 323-339.
- [32] TAQQU, M.S. (1975): "Weak Convergence to Fractional Brownian Motion and the Rosenblatt Process," *Zeitschrift für Wahrscheinlichkeitstheorie und Verwandte Gebiete*, 31, 287-302.
- [33] WANG, X. (1993): "An AIC Type Estimator for the Number of Cosinusoids," *Journal of Time Series Analysis*, 14, 431-440.

TABLE 4.1

Distribution of \hat{q} in 1000 Monte Carlo samples with the criterion function (1.3) and (3.3) with $M = T/8$ and $T/4$ respectively

	$d_x = 0.15, d_u = 0.15$						$d_x = 0.25, d_u = 0.15$						$d_x = 0.35, d_u = 0.15$									
	METH 1			METH 2			METH 1			METH 2			METH 1			METH 2						
	PAR	NP1	NP2	PAR	NP1	NP2	PAR	NP1	NP2	PAR	NP1	NP2	PAR	NP1	NP2	PAR	NP1	NP2	PAR	NP1	NP2	
$T = 64$	$\hat{q} < 3$	14.5	40.9	50.7	7.1	30.2	30.3	17.2	45.4	55.6	9.1	33.4	42.2	24.1	53.2	63.6	13.2	40.8	51.6			
	$\hat{q} = 3$	42.8	46.0	44.3	28.7	43.9	43.9	41.1	42.0	40.0	26.2	41.6	47.9	34.3	35.1	32.2	24.2	35.1	38.5			
	$\hat{q} = 4$	13.5	8.1	4.4	13.2	12.5	12.5	13.3	7.3	3.9	12.6	11.9	6.8	12.6	6.6	3.7	12.4	10.8	7.4			
	$\hat{q} = 5$	7.4	2.8	0.6	10.4	5.6	5.6	7.7	3.0	0.5	10.8	5.6	2.6	8.4	2.6	0.4	10.3	6.1	2.1			
	$\hat{q} = 6$	7.4	1.0	0.0	10.9	4.5	4.5	7.6	1.3	0.0	10.7	4.5	0.4	7.8	1.4	0.1	10.6	4.4	0.2			
	$\hat{q} = 7$	14.4	1.2	0.0	29.7	3.2	3.2	13.1	1.0	0.0	30.6	3.0	0.1	12.8	1.1	0.0	29.3	2.8	0.1			
$T = 128$	$\hat{q} < 3$	4.6	24.7	26.6	11.6	13.2	14.8	5.9	29.6	32.3	2.7	17.9	19.0	12.3	40.0	44.0	5.3	26.9	30.1			
	$\hat{q} = 3$	56.4	69.1	71.1	37.6	70.0	77.0	54.3	63.8	65.2	37.7	65.6	72.1	50.4	53.7	53.2	34.5	57.1	61.6			
	$\hat{q} = 4$	12.5	5.2	2.3	13.6	10.3	6.8	12.8	5.4	2.5	13.0	10.2	7.3	17.9	4.9	2.7	12.7	9.7	6.8			
	$\hat{q} = 5$	7.5	0.6	0.0	11.2	3.8	0.8	8.5	0.7	0.0	11.3	3.5	0.8	8.0	0.9	0.1	11.2	3.4	0.6			
	$\hat{q} = 6$	8.2	0.1	0.0	11.3	1.6	0.5	8.1	0.2	0.0	11.3	1.9	0.7	7.9	0.2	0.0	11.3	1.9	0.8			
	$\hat{q} = 7$	10.8	0.3	0.0	24.7	1.1	0.1	10.4	0.3	0.0	24.5	0.9	0.1	9.5	0.3	0.0	25.0	1.0	0.1			
$T = 256$	$\hat{q} < 3$	0.6	4.4	5.6	0.4	1.8	2.1	0.9	8.3	9.1	0.5	3.0	3.5	2.6	18.8	19.1	1.2	8.2	8.8			
	$\hat{q} = 3$	68.0	90.8	91.9	41.9	86.2	90.0	67.8	86.9	88.3	41.9	84.7	88.6	65.6	96.4	78.9	40.2	78.2	82.7			
	$\hat{q} = 4$	11.9	4.4	2.5	13.9	8.0	6.6	11.6	4.2	2.6	12.8	8.1	6.5	12.8	4.1	2.0	13.0	9.3	7.0			
	$\hat{q} = 5$	6.7	0.2	0.0	10.4	2.0	0.7	7.2	0.4	0.0	10.5	2.5	0.8	7.1	0.5	0.0	10.8	2.5	1.0			
	$\hat{q} = 6$	5.4	0.2	0.0	10.5	1.4	0.6	5.8	0.2	0.0	10.7	1.2	0.6	5.5	0.2	0.0	10.4	1.4	0.5			
	$\hat{q} = 7$	7.4	0.0	0.0	22.9	0.3	0.0	6.7	0.0	0.0	23.6	0.5	0.0	6.4	0.0	0.0	24.4	0.4	0.0			

TABLE 4.1(Cont)

Distribution of \hat{q} in 1000 Monte Carlo samples with the criterion function (1.2) and (3.4) with $M = T/8$ and $T/4$ respectively

	$d_x = 0.15, d_u = 0.15$							$d_x = 0.25, d_u = 0.15$							$d_x = 0.35, d_u = 0.15$						
	METH 3			METH 4				METH 3			METH 4				METH 3			METH 4			
	PAR	NP1	NP2	PAR	NP1	NP2		PAR	NP1	NP2	PAR	NP1	NP2		PAR	NP1	NP2	PAR	NP1	NP2	
$T = 64$	$\hat{q} < 3$	51.0	44.7	51.6	35.0	33.4	37.4	56.5	48.9	55.7	39.4	35.8	42.5	65.9	56.6	63.2	29.2	43.3	52.0		
	$\hat{q} = 3$	44.2	45.5	44.1	49.6	44.1	52.1	39.1	41.0	39.9	45.8	42.0	47.7	30.1	34.5	32.5	36.8	35.2	38.9		
	$\hat{q} = 4$	3.5	6.5	3.8	8.3	11.6	8.1	2.9	6.3	13.9	8.1	11.3	7.1	2.9	5.1	3.8	7.0	10.0	6.8		
	$\hat{q} = 5$	1.0	2.1	0.5	2.9	4.8	2.1	1.1	2.4	0.5	2.7	5.0	2.2	0.6	2.2	0.4	3.1	5.5	1.9		
	$\hat{q} = 6$	0.2	0.5	0.0	2.0	3.6	0.3	0.2	0.8	0.6	2.0	3.5	0.4	0.3	0.8	0.1	2.1	3.9	0.3		
	$\hat{q} = 7$	0.1	0.7	0.0	2.2	2.5	0.0	0.2	0.6	0.0	2.0	2.4	0.1	0.0	0.8	0.0	1.8	2.1	0.1		
$T = 128$	$\hat{q} < 3$	27.2	26.6	26.5	15.0	14.2	14.9	33.4	31.6	32.1	18.7	19.3	18.9	48.0	42.4	44.7	29.9	28.3	30.2		
	$\hat{q} = 3$	69.8	67.8	71.4	73.4	69.6	77.0	63.5	62.5	65.7	69.8	65.0	72.5	49.2	51.9	52.8	59.7	56.4	61.7		
	$\hat{q} = 4$	2.5	4.6	2.1	7.3	10.0	6.9	2.5	4.9	2.2	7.2	9.7	7.3	2.3	4.4	2.5	6.2	9.5	7.0		
	$\hat{q} = 5$	0.1	0.5	0.0	1.8	3.7	0.8	0.1	0.6	0.0	1.7	3.5	0.6	0.1	0.8	0.0	1.8	3.1	0.4		
	$\hat{q} = 6$	0.1	0.1	0.0	1.5	1.5	0.4	0.3	0.1	0.0	1.4	1.7	0.6	0.2	0.2	0.0	1.3	1.8	0.6		
	$\hat{q} = 7$	0.3	0.3	0.0	1.0	1.0	0.0	0.2	0.3	0.0	1.2	0.8	0.1	0.2	0.3	0.0	1.1	0.9	0.1		
$T = 256$	$\hat{q} < 3$	5.8	4.7	5.6	1.5	2.0	2.1	9.7	8.9	9.1	3.0	3.3	3.3	23.9	19.6	19.4	9.6	8.5	9.0		
	$\hat{q} = 3$	91.0	90.9	92.0	87.4	86.5	89.9	87.0	86.8	88.3	85.8	88.9	88.9	73.3	75.8	77.8	79.8	78.1	82.6		
	$\hat{q} = 4$	2.9	4.1	2.4	7.2	7.6	6.8	3.0	3.8	2.6	7.0	6.4	6.4	2.6	3.9	2.8	6.4	9.2	6.9		
	$\hat{q} = 5$	0.3	0.1	0.0	2.0	2.3	0.6	0.3	0.3	0.0	2.3	0.8	0.8	0.1	0.5	0.0	2.4	2.5	0.9		
	$\hat{q} = 6$	0.0	0.2	0.0	1.3	1.3	0.6	0.0	0.2	0.0	1.4	0.6	0.6	0.1	0.2	0.0	1.4	1.3	0.6		
	$\hat{q} = 7$	0.0	0.0	0.0	0.6	0.3	0.0	0.0	0.0	0.0	0.5	0.0	0.0	0.0	0.0	0.0	0.4	0.4	0.0		

TABLE 4.2

Distribution of \hat{q} in 1000 Monte Carlo samples with the criterion function (1.3) and (3.3) with $M = T/8$ and $T/4$ respectively

	$d_x = 0.15, d_u = 0.25$							$d_x = 0.25, d_u = 0.25$							$d_x = 0.35, d_u = 0.15$						
	METH 1			METH 2				METH 1			METH 2				METH 1			METH 2			
	PAR	NP1	NP2	PAR	NP1	NP2		PAR	NP1	NP2	PAR	NP1	NP2		PAR	NP1	NP2	PAR	NP1	NP2	
$T = 64$	$\hat{q} < 3$	14.7	37.4	43.6	7.4	26.3	30.9	17.8	41.2	51.2	9.5	31.3	36.9	24.4	50.7	60.5	14.4	38.6	48.1		
	$\hat{q} = 3$	46.5	49.2	51.5	33.8	48.9	56.9	44.3	45.4	44.1	31.4	44.0	52.0	37.5	37.5	35.2	27.7	37.6	42.0		
	$\hat{q} = 4$	12.4	8.3	4.3	11.8	12.6	9.0	11.9	8.3	4.2	11.9	11.5	8.0	12.3	6.4	3.9	12.0	11.0	7.2		
	$\hat{q} = 5$	7.0	3.1	0.6	10.4	5.2	2.5	6.8	3.2	0.5	10.6	5.8	2.5	7.4	3.3	0.4	10.8	5.0	2.2		
	$\hat{q} = 6$	6.7	1.1	0.0	10.7	4.1	0.6	6.8	1.2	0.0	10.3	4.1	0.5	6.3	1.4	0.0	10.4	4.3	0.4		
	$\hat{q} = 7$	12.7	0.9	0.0	25.9	2.9	0.1	12.4	0.7	0.0	26.3	3.1	0.1	12.1	0.7	0.0	24.7	2.7	0.1		
$T = 128$	$\hat{q} < 3$	3.5	19.2	19.9	1.5	8.6	10.9	5.7	24.8	26.3	2.1	13.7	15.8	11.4	36.5	40.1	4.7	22.9	26.2		
	$\hat{q} = 3$	59.9	74.7	77.0	42.3	73.9	80.5	58.4	68.6	71.4	42.0	69.2	76.3	53.6	57.2	57.8	39.5	60.6	65.0		
	$\hat{q} = 4$	12.1	5.5	3.1	13.9	10.5	6.9	11.9	5.1	2.3	13.3	10.4	6.8	11.9	5.1	2.1	12.6	9.8	6.8		
	$\hat{q} = 5$	7.8	1.1	0.0	10.8	3.9	1.0	7.5	1.1	0.0	10.5	3.8	1.2	7.1	0.7	0.0	10.4	3.9	0.9		
	$\hat{q} = 6$	7.0	0.1	0.0	11.0	1.9	0.7	7.1	0.1	0.0	10.4	1.7	0.8	6.7	0.2	0.0	10.4	1.7	0.9		
	$\hat{q} = 7$	9.7	0.3	0.0	21.5	1.2	0.0	9.4	0.3	0.0	21.7	1.2	0.1	9.3	0.3	0.0	22.4	1.0	0.2		
$T = 256$	$\hat{q} < 3$	0.6	2.6	3.7	0.2	1.1	1.2	0.7	4.5	5.8	0.4	1.8	2.5	2.2	13.9	15.1	1.0	6.1	6.8		
	$\hat{q} = 3$	70.9	92.2	93.8	46.9	86.7	89.7	69.6	90.5	91.6	45.6	86.1	89.2	67.9	81.2	82.1	44.4	81.5	84.6		
	$\hat{q} = 4$	11.1	4.8	2.5	13.6	7.9	7.4	11.9	4.5	2.6	12.2	7.8	6.8	11.7	4.1	2.7	12.5	8.0	6.8		
	$\hat{q} = 5$	6.0	0.2	0.0	8.5	2.3	1.1	6.7	0.3	0.0	9.5	2.4	0.9	6.8	0.6	0.1	9.4	2.5	1.2		
	$\hat{q} = 6$	4.5	0.2	0.0	10.2	1.6	0.6	4.5	0.2	0.0	9.4	1.5	0.6	4.4	0.2	0.0	9.0	1.5	0.6		
	$\hat{q} = 7$	6.9	0.0	0.0	20.6	0.4	0.0	6.6	0.0	0.0	22.9	0.4	0.0	7.0	0.0	0.0	23.7	0.4	0.0		

TABLE 4.2(Cont)

Distribution of \hat{q} in 1000 Monte Carlo samples with the criterion function (1.2) and (3.4) with $M = T/8$ and $T/4$ respectively

	$d_x = 0.15, d_u = 0.25$							$d_x = 0.25, d_u = 0.25$							$d_x = 0.35, d_u = 0.25$						
	METH 3			METH 4				METH 3			METH 4				METH 3			METH 4			
	PAR	NP1	NP2	PAR	NP1	NP2		PAR	NP1	NP2	PAR	NP1	NP2		PAR	NP1	NP2	PAR	NP1	NP2	
$T = 64$	$\hat{q} < 3$	46.1	40.8	44.2	30.8	28.1	31.3	52.6	44.6	51.8	35.5	33.9	37.2	62.9	53.8	61.4	45.9	40.6	48.5		
	$\hat{q} = 3$	47.9	48.7	50.8	52.1	49.4	57.0	42.2	45.2	43.7	48.1	44.4	51.7	32.2	37.6	34.3	38.9	38.4	41.9		
	$\hat{q} = 4$	3.9	7.2	4.4	8.1	11.7	9.0	3.0	6.5	4.1	7.1	10.7	8.3	2.8	5.1	3.8	6.3	10.5	7.1		
	$\hat{q} = 5$	1.4	2.3	0.6	3.8	5.2	2.3	1.3	2.4	0.4	3.7	5.5	2.3	0.9	2.2	0.4	3.3	5.0	2.0		
	$\hat{q} = 6$	0.4	0.3	0.0	2.2	3.5	0.4	0.5	0.7	0.0	2.4	3.3	0.4	0.7	0.7	0.1	2.3	3.3	0.4		
	$\hat{q} = 7$	0.3	0.7	0.0	3.2	2.1	0.0	0.4	0.6	0.0	2.3	2.2	0.1	0.5	0.6	0.0	3.3	2.2	0.1		
$T = 128$	$\hat{q} < 3$	23.5	19.5	20.5	12.0	9.3	11.2	29.6	26.7	26.9	16.8	14.7	15.1	44.2	38.3	40.6	26.8	24.8	26.3		
	$\hat{q} = 3$	73.4	74.4	76.6	75.7	74.4	79.8	67.2	67.8	70.8	70.5	68.8	76.3	52.8	56.1	57.4	61.8	60.2	65.4		
	$\hat{q} = 4$	2.4	5.2	2.9	6.5	10.2	7.4	2.6	4.6	2.3	60.3	10.3	6.8	2.3	4.8	2.0	5.3	9.1	6.8		
	$\hat{q} = 5$	0.2	0.5	0.0	1.8	3.6	1.0	0.1	0.5	0.0	2.1	3.7	0.9	0.2	0.5	0.0	2.2	3.6	0.7		
	$\hat{q} = 6$	0.2	0.1	0.0	1.8	1.6	0.6	0.2	0.1	0.0	1.9	1.5	0.8	0.3	0.1	0.0	1.8	1.5	0.7		
	$\hat{q} = 7$	0.3	0.2	0.0	2.2	0.9	0.0	0.3	0.2	0.0	2.4	1.0	0.1	0.2	0.2	0.0	2.1	0.8	0.1		
$T = 256$	$\hat{q} < 3$	4.9	2.7	3.6	1.4	1.1	1.2	8.2	5.3	6.4	2.2	1.8	2.5	21.8	14.8	15.6	8.1	6.3	6.7		
	$\hat{q} = 3$	91.2	92.7	93.8	87.5	86.8	89.9	88.0	89.9	91.0	86.3	86.5	89.4	74.9	80.5	81.7	80.7	81.6	84.8		
	$\hat{q} = 4$	3.2	4.4	2.6	6.6	7.8	7.3	2.8	4.4	2.6	6.8	7.6	6.7	2.5	4.1	2.7	6.3	7.9	6.7		
	$\hat{q} = 5$	0.7	0.1	0.0	2.5	2.4	1.1	0.8	0.2	0.0	2.5	2.3	0.8	0.4	0.4	0.0	2.3	2.4	1.2		
	$\hat{q} = 6$	0.0	0.1	0.0	1.0	1.5	0.5	0.1	0.2	0.0	1.4	1.4	0.6	0.2	0.2	0.0	1.7	1.4	0.6		
	$\hat{q} = 7$	0.0	0.0	0.0	1.0	0.4	0.0	0.1	0.0	0.0	0.8	0.4	0.0	0.2	0.0	0.0	0.9	0.4	0.0		

TABLE 4.3

Distribution of \hat{q} in 1000 Monte Carlo samples with the criterion function (1.3) and (3.3) with $M = T/8$ and $T/4$ respectively

	$d_x = 0.15, d_u = 0.35$							$d_x = 0.25, d_u = 0.35$							$d_x = 0.35, d_u = 0.35$						
	METH 1			METH 2				METH 1			METH 2				METH 1			METH 2			
	PAR	NP1	NP2	PAR	NP1	NP2		PAR	NP1	NP2	PAR	NP1	NP2		PAR	NP1	NP2	PAR	NP1	NP2	
$T = 64$	$\hat{q} < 3$	13.9	25.9	31.3	8.2	28.2	21.6	17.6	31.4	38.3	9.8	22.4	26.6	25.2	41.6	51.8	16.6	31.9	38.0		
	$\hat{q} = 3$	55.1	59.5	63.1	46.1	55.0	65.6	51.5	54.9	56.3	44.5	51.6	61.3	45.5	45.5	43.8	38.6	44.2	51.3		
	$\hat{q} = 4$	10.6	9.3	5.0	11.6	13.4	9.3	10.2	8.5	4.7	12.0	12.8	8.9	9.0	7.9	3.8	11.3	11.5	7.6		
	$\hat{q} = 5$	5.5	3.2	0.6	7.6	6.4	2.8	6.0	3.2	0.7	7.6	6.0	2.5	6.7	3.0	0.6	7.5	5.4	2.4		
	$\hat{q} = 6$	5.0	1.3	0.0	7.2	4.1	0.5	5.0	1.3	0.0	7.3	4.1	0.5	4.8	1.4	0.0	7.8	4.1	0.5		
	$\hat{q} = 7$	9.9	0.8	0.0	19.3	2.9	0.2	9.5	0.7	0.0	18.8	3.1	0.2	8.8	0.6	0.0	18.2	2.9	0.2		
$T = 128$	$\hat{q} < 3$	2.5	7.0	8.4	1.0	3.6	4.2	4.7	11.6	13.7	1.6	6.0	6.8	9.9	24.2	27.6	4.8	14.3	14.7		
	$\hat{q} = 3$	69.7	84.9	88.1	52.1	77.8	86.1	67.0	80.9	83.2	50.4	76.2	83.5	62.4	69.0	70.0	48.1	68.4	75.8		
	$\hat{q} = 4$	9.4	6.0	3.5	12.4	11.4	7.5	9.8	5.8	3.1	12.8	10.7	7.2	9.1	5.3	2.4	12.6	10.2	6.9		
	$\hat{q} = 5$	5.8	1.5	0.0	8.8	3.7	1.4	5.9	1.7	0.0	8.7	4.0	1.4	5.9	1.2	0.0	8.6	4.3	1.4		
	$\hat{q} = 6$	4.4	0.2	0.0	7.6	2.2	0.8	4.6	0.1	0.0	7.6	1.9	0.1	4.7	0.1	0.0	8.2	1.6	0.9		
	$\hat{q} = 7$	8.2	0.4	0.0	18.1	1.3	0.0	8.0	0.4	0.0	18.9	1.2	0.2	8.0	0.2	0.0	17.7	1.2	0.3		
$T = 256$	$\hat{q} < 3$	0.4	0.7	0.8	0.0	0.4	0.2	0.5	1.2	2.2	0.3	0.6	1.0	1.8	5.1	6.4	0.5	1.9	2.9		
	$\hat{q} = 3$	76.0	94.0	96.0	55.8	86.8	89.7	74.8	93.3	94.8	54.8	86.9	89.5	72.6	89.7	90.8	53.2	85.5	88.4		
	$\hat{q} = 4$	9.6	4.8	2.9	11.7	8.4	8.3	9.9	4.8	2.7	11.0	8.0	7.8	10.8	4.6	2.6	11.7	7.8	7.1		
	$\hat{q} = 5$	4.7	0.3	0.3	7.8	2.5	1.2	4.9	0.4	0.3	8.0	2.4	1.2	4.9	0.4	0.2	7.9	2.8	1.2		
	$\hat{q} = 6$	3.7	0.2	0.0	7.1	1.6	0.6	3.9	0.3	0.0	7.0	1.7	0.5	3.3	0.2	0.0	6.8	1.6	0.4		
	$\hat{q} = 7$	5.6	0.0	0.0	17.6	0.3	0.3	6.0	0.0	0.0	18.9	0.4	0.0	6.6	0.0	0.0	19.9	0.4	0.0		

TABLE 4.3(Cont)

Distribution of \hat{q} in 1000 Monte Carlo samples with the criterion function (1.2) and (3.4) with $M = T/8$ and $T/4$ respectively

	$d_x = 0.15, d_u = 0.35$							$d_x = 0.25, d_u = 0.35$							$d_x = 0.35, d_u = 0.35$						
	METH 3			METH 4				METH 3			METH 4				METH 3			METH 4			
	PAR	NP1	NP2	PAR	NP1	NP2		PAR	NP1	NP2	PAR	NP1	NP2		PAR	NP1	NP2	PAR	NP1	NP2	
$T = 64$	$\hat{q} < 3$	36.4	28.3	31.5	22.8	19.6	20.7	43.1	34.5	38.8	28.3	23.8	27.0	55.0	45.3	52.5	38.7	34.1	37.5		
	$\hat{q} = 3$	56.2	60.0	62.6	58.5	56.4	66.0	49.6	54.1	55.9	53.2	52.9	60.6	38.3	44.2	43.5	43.9	44.3	51.6		
	$\hat{q} = 4$	3.9	8.2	5.2	7.1	13.1	9.7	3.5	7.4	4.6	6.8	12.4	9.3	3.2	6.4	3.5	6.0	11.1	8.0		
	$\hat{q} = 5$	1.8	2.4	0.7	3.9	5.2	2.7	1.8	2.7	0.7	4.3	5.5	2.5	1.5	2.9	0.5	4.3	5.2	2.3		
	$\hat{q} = 6$	0.6	0.5	0.0	2.4	3.3	0.7	1.0	0.7	0.0	2.4	3.2	0.4	1.0	0.6	0.0	2.5	3.3	0.5		
	$\hat{q} = 7$	0.9	0.6	0.0	5.3	2.4	0.2	1.0	0.6	0.0	5.0	2.2	0.2	1.0	0.6	0.0	4.6	2.0	0.1		
$T = 128$	$\hat{q} < 3$	17.3	7.5	8.8	7.7	3.7	4.2	22.0	12.8	14.0	11.2	6.3	6.8	35.1	26.3	28.1	20.0	14.9	15.1		
	$\hat{q} = 3$	78.8	84.9	87.7	78.5	78.8	86.1	74.3	80.6	83.1	74.5	76.5	83.9	61.1	68.3	69.6	67.1	68.8	75.6		
	$\hat{q} = 4$	2.6	5.9	3.5	6.0	10.9	7.7	2.4	5.5	2.9	5.8	10.6	7.1	2.2	4.4	2.3	4.8	10.0	7.0		
	$\hat{q} = 5$	0.3	1.2	0.0	2.3	3.7	1.3	0.2	0.8	0.0	2.5	3.9	1.3	0.5	0.7	0.0	2.3	4.2	1.3		
	$\hat{q} = 6$	0.3	0.2	0.0	1.9	1.9	0.7	0.3	0.1	0.0	2.4	1.8	0.9	0.3	0.1	0.0	2.2	1.2	0.9		
	$\hat{q} = 7$	0.7	0.8	0.0	3.6	1.0	0.0	0.8	0.2	0.0	3.6	0.9	0.0	0.8	0.2	0.0	3.6	0.9	0.1		
$T = 256$	$\hat{q} < 3$	2.4	0.7	0.9	0.8	0.4	0.2	5.3	1.2	2.2	1.4	0.6	1.0	16.7	5.7	6.4	5.2	1.9	2.9		
	$\hat{q} = 3$	92.9	94.2	96.1	87.2	87.0	89.7	89.8	93.9	94.8	85.8	87.2	89.7	79.0	89.3	90.8	82.4	85.9	88.4		
	$\hat{q} = 4$	3.2	4.6	2.7	6.4	8.4	8.4	3.1	4.4	2.7	7.0	8.0	7.8	2.9	4.5	2.6	6.1	7.6	7.1		
	$\hat{q} = 5$	1.0	0.3	0.3	2.8	2.4	1.1	1.0	0.4	0.3	2.8	2.4	1.1	0.8	0.4	0.2	2.6	2.6	1.2		
	$\hat{q} = 6$	0.3	0.2	0.0	0.9	1.5	0.6	0.5	0.1	0.0	1.1	1.5	0.4	0.3	0.1	0.0	1.7	1.6	0.4		
	$\hat{q} = 7$	0.2	0.0	0.0	1.9	0.3	0.0	0.3	0.0	0.0	1.9	0.3	0.0	0.3	0.0	0.0	2.0	0.4	0.0		