Feng, Yuanhua

Working Paper
Optimal Convergence Rates in Nonparametric Regression with Fractional Time Series Errors

CoFE Discussion Paper, No. 02/01

Provided in Cooperation with:
Center of Finance and Econometrics (CoFE), University of Konstanz


This Version is available at:
http://hdl.handle.net/10419/85184

Terms of use:
Documents in EconStor may be saved and copied for your personal and scholarly purposes. You are not to copy documents for public or commercial purposes, to exhibit the documents publicly, to make them publicly available on the internet, or to distribute or otherwise use the documents in public.

If the documents have been made available under an Open Content Licence (especially Creative Commons Licences), you may exercise further usage rights as specified in the indicated licence.
Optimal Convergence Rates in Nonparametric Regression with Fractional Time Series Errors

Yuanhua Feng
Department of Mathematics and Statistics
University of Konstanz

Summary. Consider the estimation of $g^{(\nu)}$, the $\nu$-th derivation of the mean function in a fixed design, nonparametric regression with a linear, invertible, stationary time series error process $\xi_i$. Assume that $g \in \mathcal{C}^k$ and that the spectral density of $\xi_i$ has the form $f(\lambda) \sim c_f |\lambda|^{-\alpha}$ as $\lambda \to 0$ with constants $c_f > 0$ and $\alpha \in (-1, 1)$. Let $r_{\nu} = (1 - \alpha)(k - \nu)/(2k + 1 - \alpha)$. It is shown that the optimal convergence rate for $\hat{g}^{(\nu)}$ is $n^{-r_{\nu}}$. This rate is achieved by local polynomial fitting. It is also shown that the required regular conditions on the innovation distribution in the current context are the same as those in nonparametric regression with iid errors.

Keywords: Nonparametric regression, optimal convergence rate, long memory, antipersistence, inverse process.

1 Introduction

Consider the estimation of the $\nu$-th derivation of the mean function, $g^{(\nu)}$, in the equidistant design nonparametric regression model

\begin{equation}
Y_i = g(x_i) + \xi_i,
\end{equation}

where $x_i = i/n$, $g : [0, 1] \to \mathbb{R}$ is a smooth function and $\xi_i$ is a linear, (second order and strict) stationary process generated by an iid (identically independent distributed) innovation series $\varepsilon_i$ through a linear filter. For the autocovariance function $\gamma(k) = \text{cov}(\xi_i, \xi_{i+k})$, it is assumed that $\gamma(k) \to 0$ as $|k| \to \infty$. Equation (1.1) represents a nonparametric regression model with short memory (including iid $\xi_i$ as a special case), long memory and antipersistence. Here, a stationary process $\xi_i$ is said to have long memory (or long-range dependence), if $\sum \gamma(k) = \infty$. A more strict assumption is that the spectral density
\[ f(\lambda) = (2\pi)^{-1} \sum \gamma(k) \exp(ik\lambda) \] has a pole at the origin of the form

\[ f(\lambda) \sim c_f |\lambda|^{-\alpha} \text{ (as } \lambda \to 0) \]

for some \( \alpha \in (0, 1) \), where \( c_f > 0 \) is a constant and ‘\( \sim \)’ means that the ratio of the left and the right hand sides converges to one (see Beran, 1994, and references therein). Note that, (1.2) implies that \( \gamma(k) \sim c_k |k|^{\alpha-1} \) so that \( \sum \gamma(k) = \infty \). Hence now \( \xi_t \) has long memory. If (1.2) holds with \( \alpha = 0 \), then we have \( 0 < \sum \gamma(k) < \infty \) and \( \xi_t \) is said to have short memory. On the other hand, a stationary process is said to be antipersistent, if (1.2) holds for \( \alpha \in (-1, 0) \) implying that \( \sum \gamma(k) = 0 \).

The aim of this paper is to investigate the minimax optimal convergence rate of a nonparametric estimator of \( g^{(\nu)} \) (see e.g. Farrell, 1972, Stone, 1980, 1982 and Hall and Hart, 1990a for related works). For a summary of the nonparametric minimax theory we refer the reader to Hall (1989). Hall and Hart (1990a) obtained optimal convergence rate for estimating \( g \) in nonparametric regression with Gaussian stationary short- and long-memory errors. In this paper a unified formula for the optimal convergence rate for estimating \( g^{(\nu)} \) in nonparametric regression with short-memory, long-memory and antipersistent errors is given. It is shown that this rate is achieved by local polynomial fitting (Beran and Feng, 2001a). Our finding generalizes in various ways previous results in Stone (1980) and Hall and Hart (1990a). A simple condition under which a sequence \( n^{-r} \) forms a lower bound to the convergence rate is given for nonparametric regression with stationary time series errors at any dependence level. Results in this paper are given for Gaussian and non-Gaussian error processes satisfying some regular conditions.

The estimator and the error process are defined in section 2. Section 3 describes the conditions on the distribution and provides the main results. It turns out that the required regular conditions on the marginal innovation distribution are the same for all \( \alpha \in (-1, 1) \) and hence do not depend on the dependence structure. Some auxiliary results, which can be thought of as part of the proofs, are given in section 4. Detailed proofs are put in the appendix.
2 The estimator and the error process

2.1 The local polynomial fitting

Kernel estimator of $g$ in nonparametric regression with short-memory and long-memory errors was proposed by Hall and Hart (1990a). Beran (1999) extended the kernel estimator to nonparametric regression with antipersistence. However, it is well known that the kernel estimator is affected by the boundary problem. Another attractive nonparametric approach is the local polynomial fitting introduced by Stone (1977) and Cleveland (1979). Beran and Feng (2001a) proposed local polynomial fitting in nonparametric regression with short-memory, long-memory and antipersistent errors. In this paper we will use the proposal in Beran and Feng (2001a) to show the achievability of the optimal convergence rate.

Let $k \geq 2$ be a positive integer. The function class considered in this paper is $\mathcal{C}^k(B)$, the collection of all $k$ times differentiable functions $g$ on $[0,1]$ which satisfy

$$\sup_{0 \leq x \leq 1} \max_{\nu=0,1,\ldots,k} |g^{(\nu)}(x)| \leq B.$$ 

Let $p = k - 1$. Then $g$ can be locally approximated by a polynomial of order $p$ for $x$ in the neighbourhood of a point $x_0$:

$$g(x) = g(x_0) + g'(x_0)(x - x_0) + \cdots + g^{(p)}(x_0)(x - x_0)^p/p! + R_p,$$

where $R_p$ is a remainder term. Let $K$ be a second order kernel (a symmetric density) having compact support $[-1,1]$. Given $n$ observations $Y_1, \ldots, Y_n$, we can obtain an estimator of $g^{(\nu)}$ ($\nu \leq p$) by solving the locally weighted least squares problem

$$Q = \sum_{i=1}^{n} \left\{ Y_i - \sum_{j=0}^{p} \beta_j (x_i - x_0)^j \right\}^2 K \left( \frac{x_i - x_0}{h} \right) \Rightarrow \min,$$

where $h$ is the bandwidth. Let $\hat{\beta} = (\hat{\beta}_0, \hat{\beta}_1, \ldots, \hat{\beta}_p)'$ be the solution of (2.2). Then it is clear from (2.1) that $\hat{g}^{(\nu)}(x_0) := \nu! \hat{\beta}_\nu$ estimates $g^{(\nu)}(x_0)$, $\nu = 0,1,\ldots,p$, which is the local polynomial fitting of $g^{(\nu)}$. Note in particular that $\hat{g}^{(\nu)}$ is the same for nonparametric regression with stationary time series errors at any dependence level.
2.2 The error process

In this paper it is assumed that the spectral density of $\xi$ has the form (1.2). Hence $\xi$ will be called a fractional time series error process. $\xi$ is also assumed to be causal, linear and invertible. That is, $\xi$ can be expressed in two ways:

\begin{equation}
\xi = \psi(B)\varepsilon_i,
\end{equation}

and

\begin{equation}
\varepsilon_i = \varphi(B)\xi_i,
\end{equation}

where the innovations $\varepsilon_i$ are iid mean zero random variables with $\text{var}(\varepsilon_i) = \sigma_\varepsilon^2 < \infty$, $B$ is the backshift operator, and $\psi(B) = \sum_{j=0}^{\infty} a_j B^j$ and $\varphi(B) = \sum_{j=0}^{\infty} b_j B^j$ are the characteristic polynomials of the MA and AR representations of $\xi_i$, respectively, with $a_0 = b_0 = 1$, $\sum a_j^2 < \infty$ and $\sum b_j^2 < \infty$. The causality of $\xi_i$ is made here for convenience.

Some properties of $\xi_i$ can be understood more easily by means of its inverse process. Following Chatfield (1979), the inverse process of $\xi_i$, denote by $\xi^i$, is the process with the same innovations $\varepsilon_i$ and $\varphi(B)$ rep. $\psi(B)$ as its characteristic polynomials for the MA rep. AR representations, which is given by

\begin{equation}
\xi^i = \varphi(B)\varepsilon_i,
\end{equation}

and

\begin{equation}
\varepsilon_i = \psi(B)\xi^i.
\end{equation}

Following Shaman (1975), the spectral density of $\xi^i$, $f^i(\lambda)$ say, is

\begin{equation}
f^i(\lambda) = \sigma_\varepsilon^4 (2\pi)^{-2} (f(\lambda))^{-1} \sim c_j^i |\lambda|^{-\alpha^i}(\text{as } \lambda \to 0),
\end{equation}

where $c_j^i = \sigma_\varepsilon^4 (2\pi)^{-2} (c_j)^{-1}$ and $\alpha^i = -\alpha$. Equation (2.7) implies that: 1. If $\xi$ is a short-memory process, so is $\xi^i$ (in particular, the inverse process of an iid process is the process itself); 2. If $\xi$ is a long-memory process with $0 < \alpha < 1$, then $\xi^i$ is an antipersistent process with $\alpha^i = -\alpha$, and vice versa.

From (2.3) we see that the autocovariances of $\xi$ are $\gamma(k) = \sigma_\varepsilon^2 \sum a_j a_{j+k}$. The inverse autocovariances of $\xi_i$ (Cleveland, 1972 and Chatfield, 1979), i.e. the autocovariances
of $\xi_i$, are given by $\gamma^j(k) = \sigma^2 \sum_{j=0}^{\infty} b_j b_{j+k}$. Hence we have $\sum \gamma(k) = \sigma^2 (\sum a_j)^2$ and $\sum \gamma^j(k) = \sigma^2 (\sum b_j)^2$. This results in $\sum a_j = \infty$, $\sum b_j = 0$ for $\alpha > 0$ and $\sum a_j = 0$, $\sum b_j = \infty$ for $\alpha < 0$. For $\alpha = 0$ we have both, $0 < \sum a_j < \infty$ and $0 < \sum b_j < \infty$.

A class of processes having the property (1.2) is the class of the FARIMA($p, \delta, q$) (fractional ARIMA) processes (Granger and Joyeux, 1980 and Hosking, 1981), where $\delta \in (-0.5, 0.5)$ is the fractional differencing parameter. It is well known that the spectral density of a FARIMA process has the form (1.2) with $\alpha = 2\delta$.

### 3 Optimal convergence rates

#### 3.1 Assumptions on the innovation distribution

An important finding of this paper is that the derivation of that a given sequence is a lower bound to the convergence rate in nonparametric regression with error process $\xi_i$ is similar to that for nonparametric regression with the iid errors $\varepsilon_i$. Furthermore, it turns out that the required conditions on the marginal distribution of $\varepsilon_i$ under model (1.1) with any $\alpha \in (-1, 1)$ are the same, i.e. which do not depend on the dependence level. In the following we will adapt the regular conditions in Stone (1980, 1982) to fixed design nonparametric regression. Assume that $Z(g)$ is a real random variable depending on $g \in \mathbb{R}$. It is assumed that the density function $f(z, g)$ is strictly positive and that $f(z, g) = f(z - g, 0)$, where $g$ is the mean function of $Z(g)$, i.e.

$$\int z f(z, g) dz = g$$

for all $g \in \mathbb{R}$. It is further assumed that the equation

$$\int f(z, g) dz = 1$$

can be twice continuously differentiated with respect to $g$ to yield

$$\int f'(z, g) dz = 0$$

and

$$\int f''(z, g) dz = 0.$$
The iid innovations $\varepsilon_i$ are generated by the marginal distribution with density $f(z, 0)$, which will be simply denoted by $f(z)$ in the following. Using this notation the density of $Z(g)$ may be represented as $f(z, g) = f(z - g)$. Set $l(z, g) = \log f(z, g)$.

There are positive constants $\tau_0$ and $C$ and there is a function $M(z, g)$ such that for $g \in \mathbb{R}$

$$|l''(z, g + \tau)| \leq M(z, g) \quad \text{for} \quad |\tau| \leq \tau_0$$

and

$$\int M(z, g) f(z, g) dz \leq C.$$

Note that the last condition is fulfilled, if $l''(z, g)$ is bounded.

**Remark 1.** It is easy to show that all of these conditions are fulfilled, if $Z(g)$ is Gaussian with

$$f(z, g) = \frac{1}{\sqrt{2\pi \sigma_z}} e^{-\frac{(z-g)^2}{2\sigma_z^2}}, \quad -\infty < z, g < \infty.$$

And it is also not hard to show that these conditions are fulfilled, if the marginal distribution of $\varepsilon_i$ is the student $t_m$ distribution with $m \geq 3$, i.e. if $f(z, g)$ is given by

$$f_m(z, g) = \frac{\Gamma((m+1)/2)}{\Gamma(m/2)\sqrt{m\pi}} \left(1 + \frac{(z-g)^2}{m}\right)^{-\frac{(m+1)/2}{2}}, \quad -\infty < z, g < \infty.$$

**Remark 2.** Observe that however other distributions considered by Stone (1980), e.g. the exponential distribution, do not satisfy the regular conditions given above. If $\varepsilon_i$ are iid exponential distributed with $E(\varepsilon_i) = 0$ and $\text{var}(\varepsilon_i) = \lambda$, then density function of $Z(g)$ is given by

$$f(z, g) = \frac{1}{\lambda} e^{-(z+\lambda-g)/\lambda}, \quad -\infty < g < \infty \text{ and } g - \lambda \leq z < \infty$$

and zero otherwise. The support of $f > 0$ for this distribution depends on $g$.

### 3.2 Lower bounds to convergence rates

For the minimax optimal convergence rate we will use the following definition (see e.g. Farrell, 1972, Stone, 1980 and Hall and Hart, 1990a). Let $\nu < k$ be a nonnegative integer and $g_n^{(\nu)}$ denote a generic nonparametric estimator of $g^{(\nu)}$ based on $(Y_1, ..., Y_n)$. Let $r_\nu$ be
a positive number. The sequence \( n^{-r_{\nu}} \) is called a lower bound to the convergence rate at \( x_0 \) if

\[
\lim \inf \sup_{n, g \in \mathcal{G}} P(\left| \hat{g}_n^{(\nu)}(x_0) - g^{(\nu)}(x_0) \right| > c_{\nu} n^{-r_{\nu}}) > 0
\]

for \( c_{\nu} \) sufficiently small. \( n^{-r_{\nu}} \) is called an achievable convergence rate if there is a sequence of estimators \( \hat{g}_n^{(\nu)} \) such that

\[
\lim_{\nu \to \infty} \lim \inf \sup_{n, g \in \mathcal{G}} P(\left| \hat{g}_n^{(\nu)}(x_0) - g^{(\nu)}(x_0) \right| > c_{\nu} n^{-r_{\nu}}) = 0.
\]

Also, the sequence \( n^{-r_{\nu}} \) is called the optimal convergence rate if it is an achievable lower bound to the convergence rate. The optimal convergence rate for a nonparametric regression estimator of \( g^{(\nu)} \) with iid errors is \( n^{-(k-\nu)/(2k+1)} \) (Stone, 1980). In fact, \( n^{-(k-\nu)/(2k+1)} \) is also the optimal convergence rate for estimating \( g^{(\nu)} \) in nonparametric regression with short-memory errors (results for \( \nu = 0 \) may be found in Hall and Hart, 1990a). In the case with \( 0 < \alpha < 1 \), Hall and Hart (1990a) shown that the optimal convergence rate is \( n^{-(1-\alpha)k/(2k+1-\alpha)} \) for estimating \( g \). In this paper we will show that \( n^{-r_{\nu}} \) with \( r_{\nu} = (1-\alpha)(k-\nu)/(2k+1-\alpha) \) is the optimal convergence rate for estimating \( g^{(\nu)} \), uniformly for \( \alpha \in (-1, 1) \). The following theorem shows at first that \( n^{-r_{\nu}} \) is a lower bound to the convergence rate, i.e. \( n^{-r_{\nu}} \) satisfies (3.1).

**Theorem 1** Let model (1.1) hold with \( g \in \mathcal{G}^k \). Let \( x_0 \in (0,1) \) be an interior point of the support of \( g \). Let \( \nu < k \) and \( r_{\nu} = (1-\alpha)(k-\nu)/(2k+1-\alpha) \). Assume that the regular conditions on the marginal innovation distribution as described in Section 3.1 hold. Then \( n^{-r_{\nu}} \) is a lower bound to the convergence rate for estimating \( g^{(\nu)}(x_0) \).

The proof of Theorem 1 is given in the appendix.

Theorem 1 extends previous results as obtained by Stone (1980) and Hall and Hart (1990a) in different ways. The results in Stone (1980) are extended to nonparametric regression with fractional time series errors. Main differences between results of Theorem 1 and those given in Hall and Hart (1990a) are: 1. These results are given for all \( \alpha \in (-1, 1) \) including the antipersistent case and 2. These results are available for non-Gaussian error processes satisfying regular conditions on the marginal innovation distribution. 3. The estimation of derivatives is also considered.
Remark 3. The sequence $n^{-r\nu}$ as defined in Theorem 1 is of course also a lower bound to the convergence rate for the estimation at the two boundary points $x_0 = 0$ or $x_0 = 1$, since the set of all measurable functions of the observations at $x_0 = 0$ (resp. $x_0 = 1$) under the restriction that there are no observations on the left (resp. right) hand side is a subset of all measurable functions.

Remark 4. In the proof of Theorem 1 a two-point discrimination argument is used. It will be shown that the probability on the right hand side of (3.1) can be made arbitrarily close to $\frac{1}{2}$. If a more sophisticated multi-point discrimination argument is used as in Stone (1980), then it can be shown that

$$\lim_{c_n \to 0} \lim_{n \to \infty} \sup_{\theta \in \Theta} P\left(\left|g_n^{(\nu)}(x_0) - g^{(\nu)}(x_0)\right| > c_n n^{-r\nu}\right) = 1.$$ 

Remark 5. Results of Theorem 1 are in general not available for random design nonparametric regression or density estimation with dependent observations, since the effect of dependence in such cases tends to be less profound than in the model to be discussed here (see Hall and Hart, 1990b).

3.3 Achievability

Beran and Feng (2001a) shown that for $g \in C^k$ with $k - \nu$ even, the uniform convergence rate of the local polynomial fitting $\hat{g}^{(\nu)}$ is of order $n^{-r\nu}$ for all $x \in [0, 1]$, if a bandwidth of the optimal order $n^{-(1-\alpha)/(2k+1-\alpha)}$ is used, where $r$ is as defined in Theorem 1 (see Theorem 2 in Beran and Feng, 2001a). Similar results hold for function class $C^k$ with $k - \nu > 0$ odd. This result can be used to show the achievability of the lower bound to the convergence rate as defined in Theorem 1, i.e. (3.2) holds for the local polynomial fitting $\hat{g}^{(\nu)}$ with $n^{-r\nu}$, also at the two boundary points $x_0 = 0$ and $x_0 = 1$. This results in

Theorem 2 Let $x_0 \in [0, 1]$. Under the conditions of Theorem 1 it can be shown that, $n^{-r\nu}$ is the optimal convergence rate for estimating $g^{(\nu)}(x_0)$.

The additional proof of Theorem 2 is straightforward and is omitted to save place.

Remark 6. Indeed, the convergence rate $n^{-r\nu}$ as defined in Theorem 1 may be achieved under much weaker conditions. It is clear that, (3.2) will hold, if $\hat{g}^{(\nu)}$ is asymptotically normal. Some sufficient conditions under which $\hat{g}^{(\nu)}$ is asymptotically normal are given in Beran and Feng (2001b), which are much weaker than those described in Section 3.1.
4 Auxiliary results

4.1 Notations

Note that \( r_\alpha < 1 \) for all \( \alpha \in (-1, 1) \) and that the interpolation error is of order \( n^{-1} \), which is hence negligible. Therefore we may assume without loss of generality that \( x_0 \) is of the form \( i_0/n \). It is notationally convenient to take \( x_0 = i_0/n = 0 \), so we will consider the shifted model

\[
Y_i = g(i/n) + \xi_i, i = -n, ..., -1, 0, 1, ..., n,
\]

and estimate \( g^{(\nu)} \) at the origin. Moreover, we shall assume that both, the infinite past and the infinite future, are given, i.e. we observe

\[
(4.1) \quad Y_i = g(i/n) + \xi_i, -\infty < i < \infty.
\]

Model (4.1) is assumed only for notational convenience, which helps us to save symbols for distinguishing finite and infinite sample paths. It turns out that the extra information is of negligible benefit for the derivation of a lower bound to the convergence rate.

The main idea to prove Theorem 1 is to construct two sequences of functions. If these two sequences are “hard to distinguish”, then the difference of them will form a lower bound to the convergence rate. If they are “far apart” at the same time, then the difference of them will form an achievable convergence rate, hence we will obtain the optimal convergence rate. Following Stone (1980) and Hall and Hart (1990a), let \( \Psi \geq 0 \) be a \( k + 1 \)-differentiable function on \(( -\infty, \infty )\), vanishing outside \((-1, 1)\) and satisfying \( \Psi^{(\nu)}(0) > 0 \) for \( \nu = 0, 1, ..., k \). Put

\[
B' = \sup_{0 \leq \nu \leq 1} \max_{\nu = 0, 1, ..., k} |\psi^{(\nu)}(x)|.
\]

Choose \( a > 0 \) so small that \( aB' < B \). Let \( 0 < s < 1 \) and set \( h = n^{-s} \). Define

\[
(4.2) \quad g_\theta(x) = \theta a h^k \Psi(x/h).
\]

Then \( g_\theta(x) \) for \( \theta \in \{0, 1\} \) are two sequences of functions in \( \Theta^k \).

In the following we will denote the limits \( \lim_{n \to \infty} \prod_{i = n}^n \) and \( \lim_{n \to \infty} \sum_{i = n}^n \) by \( \prod \) and \( \sum \) for simplicity. For \(-\infty < i < \infty\), define the doubly infinitive column vectors \( \xi = (\xi_i) \), \( \varepsilon = (\varepsilon_i) \) and \( g = (g_i(i/n)) \). Define the doubly infinite matrices \( \Sigma = (\gamma(i - j)) \) and
Γ = (γ^j(i − j)). Let Λ = (b_{i−j}) as given in (A.6) in the appendix, where b_k = 0 for i < 0. Let Ω = (a_{i−j}) be the as Λ but with b_{i−j} being replaced by a_{i−j}. Then we have ξ = Ωε and ε = Λξ. Let Y = (Y_i). We have Y_θ = θg + ξ. Define X_θ = ΛY_θ = θη + ε, where η = Λg. Note that X_0 = ε and Y_0 = ξ. Furthermore, we see that X_1 is a sequence of independent random variables.

4.2 The likelihood functions and the error probabilities

Let L_0 and L_0 denote the likelihood functions of X_0 = ε and Y_0 = ξ, respectively. Observe that L_0(x) = \prod f(x_i), where x = (..., x_{−1}, x_0, x_1, ...)’ is a doubly infinite vector and f is the marginal density function of \varepsilon_i. The following lemma gives the relationship between these two likelihood functions.

Lemma 1 For the fractional time series process defined by (2.3) and (2.4), and a doubly infinite real vector y we have

\begin{equation}
L_0(y) = L_0(x) = \prod_{i=−\infty}^{\infty} f(x_i),
\end{equation}

where x = Λy with x_i = \sum_{j=−\infty}^{\infty} b_{ij}y_{i−j}, −∞ < i < ∞, and f is the marginal density function of \varepsilon_i.

The proof of Lemma 1 is given in the appendix. Lemma 1 shows that L is uniquely determined by L. Note that, inversely, L is also uniquely determined by L. Following Lemma 1 the estimation of the likelihood function of an invertible stationary time series is equivalent to that of the corresponding iid innovations. The idea behind this lemma plays a very important role for the derivation of asymptotic results in nonparametric regression with dependent errors, which shows that discussions on asymptotic results in this case may often be reduced to those for models with iid errors after a suitable transformation. Note that Lemma 1 only holds for causal processes.

Let L_1 and L_1 denote the likelihood functions of X_1 = ε + η and Y_1 = ξ + g, respectively. To prove Theorem 1 we need to estimate P(L_0 < L_1|θ = 0) and P(L_0 > L_1|θ = 1). The following corollary of lemma 1 reduces the estimation of these error probabilities to that of the independent sequences X_θ.
**Corollary 1** Let $X_\theta$ and $Y_\theta$ are defined above. Let $y$ is a doubly infinite real vector. Then, under the assumptions of Lemma 1, we have

$$P(L_0(y) < L_1(y) | \theta = 0) = P(L_0(x) < L_1(x) | \theta = 0)$$

and

$$P(L_0(y) > L_1(y) | \theta = 1) = P(L_0(x) > L_1(x) | \theta = 1),$$

where $x = \Lambda y$.

The proof of Corollary 1 is given in the appendix. Following Corollary 1, a method for estimating the error probability developed for nonparametric regression with iid errors could be adapted to the current case. In this paper we will use the methodology proposed by Stone (1980). Note that $\eta$, the deterministic part of $X_1$, does not necessarily have the same smooth properties as $g$, the deterministic part of $Y_1$. However, this does not affect the estimation of the error probability.

### 4.3 A sufficient condition

Let $Y_n = \frac{1}{2} g_1(0) = \frac{1}{2} a \Psi(0) h^k = c_0 h^k$, where $c_0 = \frac{1}{2} a \Psi(0)$. Let $T_n = c_v h^{(k-v)}$, where $c_v = \frac{v}{2} a \Psi^{(v)}(0)$ for $v < k$. If $\xi_i$ in model (1.1) are iid, then, following Stone (1980), it can be shown that a sufficient condition, under which $T_n$ is a lower rate of convergence for estimating $g^{(v)}$, is that there is an $M > 0$ such that $\sum g^2(i/n) < M$ (see equation (2.1) in Stone, 1980). The following lemma gives a simple extension of this result to the case when $\xi_i$ are fractional stationary time series errors defined by (2.3) and (2.4).

**Lemma 2** Let $\xi_i$ be defined by (2.3) and (2.4). Consider the estimation of $g^{(v)}$. Then $T_n$ is a lower rate of convergence, if there is an $M > 0$ such that

$$\sum_{i=-\infty}^{\infty} \eta_i^2 = g' \Lambda' \Lambda g < M,$$

where $\eta_i$ are the elements of $\eta = \Lambda g$.

The proof of Lemma 2 is given in the appendix. Note that $g_0 \equiv 0$ and hence $g$ is the difference sequence between the two functions $g_0$ and $g_1$. Lemma 2 shows that this
sequence will form a lower rate of convergence for estimating $g$, if the transferred difference sequence $\eta$ is squared summable. From Lemma 2 we can also see that, if $\gamma_n$ is a lower rate of convergence for estimating $g$, then $\gamma_n$, the sequence of the $\nu$-th derivative $\gamma_n$, is a lower rate of convergence for estimating $g^{(\nu)}$ providing $\Psi^{(\nu)}(0) > 0$.

It is easy to show that condition (4.4) is equivalent to

$$g^T g < \sigma^2 \epsilon M$$

and further equivalent to

$$g^T \Sigma^{-1} g < \sigma^{-2} \epsilon M.$$

Proofs of (4.5) and (4.6) are given in the appendix. These two representations are easy to understand. Equation (4.6) directly shows the change in this sufficient condition caused by the dependence structure. The following remarks clarify the above results.

**Remark 7.** For iid errors $\xi_i = \xi_j$ we have $A = I$, $\Gamma = \sigma^2 I$ and $\Sigma^{-1} = \sigma^{-2} I$, where $I$ denote the doubly infinite identity matrix. In this case we have simply $\sum g^2(i/n) < M$. Note that $D = \sqrt{\sum g^2(i/m)}$ is the $L^2$-norm of $g$. Lemma 2 implies that any method of deciding between $\theta = 0$ and $\theta = 1$, i.e. of deciding between the vector $g$ and the zero vector must have overall positive error probability, if the norm of $g$ is bounded.

**Remark 8.** Assume that $\epsilon_i$ are normal. Following Hall and Hart (1990a) it can be shown that, the overall error probability of any estimator of $\theta$ based on $Y$ is at least

$$P_a = 1 - \Phi \left( \left( g^T \Sigma^{-1} g \right)^{1/2} \right),$$

where $\Phi$ is the standard normal distribution function. The error probability $P_a$ will be positive, if $g^T \Sigma^{-1} g$ is finite. $P_a$ in (4.7) can be made arbitrarily close to $\frac{1}{2}$ by choosing the constant $a$ in (4.2) so that $a \rightarrow 0$ and hence $g^T \Sigma^{-1} g \rightarrow 0$.

## 5 Acknowledgements

This work was finished under the advice of Prof. Jan Beran, Chair of the Department of Mathematics and Statistics, University of Konstanz, Germany, and was financially supported by the Center of Finance and Econometrics (CoFE) at the University of Konstanz. The author gratefully acknowledges Prof. Jan Beran for his useful advice and comments, which lead to improve the quality of this paper.
Appendix: Proofs

**Proof of Lemma 1.** It is well known that, under common conditions, the likelihood functions of two random vectors forming a reciprocal one-to-one mapping are uniquely determined by each other (see e.g. Theorem 2 of Section 4.4 in Rohatgi and Saleh, 2001, pp. 127ff). Note that this result can be extended to doubly infinite random vectors. The proof of Lemma 1 remains to check that all of the conditions of this theorem hold. At first, \( \varepsilon = \Lambda \xi \) form a doubly infinite dimensional reciprocal one-to-one-mapping with the inverse transformation \( \xi = \Omega \varepsilon \), where both, the original function and the inverse transformation are linear. Hence, conditions (a) to (c) of Theorem 2 of Section 4.4 in Rohatgi and Saleh (2001) hold. Furthermore, \( \Lambda \) is also the matrix of the partial derivatives of \( \varepsilon \) with respect to \( \xi \). And the Jacobian \( J \) of the inverse transformation is the determinant \( |\Lambda| = 1 \), since \( \Lambda \) is a (doubly infinite) lower triangle matrix, whose diagonal elements are identically one. The relationship between \( \mathcal{L}_0 \) and \( L_0 \) as given in Lemma 1 holds. \( \diamond \)

**Proof of Corollary 1.** Observe that \( X_1 = X_0 + \eta \) and \( Y_1 = Y_0 + g \). Hence we have, 
\[
L_1(x) = L_0(x - \eta) \quad \text{and} \quad \mathcal{L}_1(y) = \mathcal{L}_0(y - g).
\]

It follows from Lemma 1, for any doubly infinite dimensional real vectors \( y \) and \( g \),
\[
\mathcal{L}_1(y) = \mathcal{L}_0(y - g) = L_0(x - \eta) = L_1(x) = \prod_{i=-\infty}^{\infty} f(x_i - \eta_i),
\]
where \( x = \Lambda y, \eta = \Lambda g \) and \( f \) is the marginal density function of \( \varepsilon_i \). Equations (4.3) and (A.1) together show that \( \mathcal{L}_0(y) < \mathcal{L}_1(y) \) (or \( \mathcal{L}_0(y) > \mathcal{L}_1(y) \)), or \( \mathcal{L}_0(y) = \mathcal{L}_1(y) \), if and only if \( L_0(x) < L_1(x) \) (or \( L_0(x) > L_1(x) \)), or \( L_0(x) = L_1(x) \), where \( x = \Lambda y \). Corollary 1 follows from this fact. \( \diamond \)

The proofs given in the following are related to those in Stone (1980) and Hall and Hart (1990a). Hence some details will be omitted to save place. To this end we refer the reader to the proofs in these works. We also refer the reader to read Theorem 1 in Hall (1989) and its proof. Note that the symbol \( \alpha \) in this paper is differently defined as that used in Hall and Hart (1990a).

**Proof of Lemma 2.** Let \( \Upsilon_n^\nu \) is as defined in Lemma 2. Note that
\[
\sup_{g \in \mathbb{C}^n} P_g \{ |\hat{g}_n^{(\nu)}(0) - g^{(\nu)}(0)| \geq \Upsilon_n^\nu \} \geq \max_{\theta = 0, 1} P_\theta \{ |\hat{g}^{(\nu)}(0) - g^{(\nu)}(0)| \geq \Upsilon_n^\nu \}.
\]
Let $\tilde{\theta}_n = 0$ or 1 minimizes $|\tilde{g}^{(v)}(0) - g^{(v)}(0)|$. Then $\tilde{\theta}_n \neq \theta$ implies $|\tilde{g}^{(v)}(0) - g^{(v)}(0)| \geq \gamma_n^{(v)}$, and hence

$$
\max_{\theta = 0, 1} P_0(\tilde{g}^{(v)}(0) - g^{(v)}(0) \geq \gamma_n^{(v)}) \geq \frac{1}{2} \{ P_0(\tilde{\theta} = 1) + P_1(\tilde{\theta} = 0) \}
$$

(A.2)

where $\tilde{\theta}$ is the maximum likelihood estimator of $\theta$ (or the likelihood ratio discriminator) in the two-parameter problem. The last inequality follows from the Neyman-Pearson lemma.

From Corollary 1 we have

$$
\max_{\theta = 0, 1} P_\theta(\tilde{g}^{(v)}(0) - g^{(v)}(0) \geq \gamma_n^{(v)}) \geq \frac{1}{2} \{ P_0(L_0 < L_1) + P_1(L_1 < L_0) \}
$$

(A.3)

Let $L_R$ denote the likelihood ratio $L_1/L_0$. By calculations similar to those given on pages 1352 - 1353 of Stone (1980), it can be shown under the regular conditions on the marginal distribution of $\varepsilon_i$ as given in Section 3.1, that there is a positive constant $M_1$ such that

$$
E_0|\log(L_R)| < M_1
$$

(A.4)

and

$$
\lim_{a \to 0} E_0|\log(L_R)| = 0.
$$

(A.5)

Similar formulas as given in (A.4) and (A.5) hold for the expectation under $\theta = 1$ with another positive constant $M_2$. Let $M_0 = \max(M_1, M_2)$. Then we can find an integer $K \geq 2$ and $0 < \tau < \frac{1}{2}$ such that if $L_R > (1 - \tau)/\tau$ or $L_R < \tau/(1 - \tau)$, then $|\log(L_R)| \geq K M_0$.

Following the Markov inequality

$$
P_0 \left( \frac{\tau}{1 - \tau} \leq L_R \leq \frac{1 - \tau}{\tau} \right) > \frac{K - 1}{K}
$$

$$
P_1 \left( \frac{\tau}{1 - \tau} \leq L_R \leq \frac{1 - \tau}{\tau} \right) > \frac{K - 1}{K}.
$$

Put prior probabilities 1/2 each on $\theta = 0$ and $\theta = 1$. Then

$$
P(\theta = 1|Y) = \frac{\frac{1}{2} L_1}{\frac{1}{2} L_1 + \frac{1}{2} L_0} = \frac{L_R}{L_R + 1}
$$

14
and
\[
P(\tau \leq P(\theta = 1|Y) \leq 1 - \tau) = P\left(\frac{L_R}{L_R + 1} \leq 1 - \tau\right) = P\left(\frac{\tau}{1 - \tau} \leq \frac{L_R}{1 - \tau}\right) = \frac{1}{2} P_0\left(\frac{\tau}{1 - \tau} \leq \frac{L_R}{1 - \tau}\right) + \frac{1}{2} P_1\left(\frac{\tau}{1 - \tau} \leq \frac{L_R}{1 - \tau}\right) \geq \frac{K - 1}{K}.
\]

That is, the error probability of \( \hat{\theta} \) is at least \( K-1 \).\( \hat{\theta} \).

Note that \( \frac{K-1}{K} \) can be made arbitrarily close to \( \frac{1}{2} \) as \( \delta \to 0 \) by choosing \( K \) sufficiently large and \( \tau \) sufficiently close to \( \frac{1}{2} \) at the same time. 

\begin{proof}
Proof of equations (4.5) and (4.6). The matrix \( \Lambda \) is given by
\[
\Lambda = \begin{pmatrix}
\cdots & \cdots & \cdots & \cdots & \cdots & \cdots & \cdots \\
\cdots & 1 & 0 & 0 & \cdots & 0 & 0 \\
\cdots & b_1 & 1 & 0 & \cdots & 0 & 0 \\
\cdots & b_2 & b_1 & 1 & \cdots & 0 & 0 \\
\cdots & \vdots & \vdots & \vdots & \ddots & \vdots & \vdots \\
\cdots & b_{n-1} & b_{n-2} & b_{n-3} & \cdots & 1 & 0 \\
\cdots & b_n & b_{n-1} & b_{n-2} & \cdots & b_1 & 1 \\
\cdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\
\end{pmatrix}.
\]

Following the definition of \( \gamma^j(i-j) \) we have \( \Gamma = \sigma^2 \Lambda \Lambda' \). Furthermore, it can be shown that \( \Lambda \Lambda' = \Lambda' \Lambda \). The equivalence between (4.4) and (4.5) follows from this fact. The equivalence between the two conditions (4.5) and (4.6) is due to the fact that \( \Sigma^{-1} = \sigma^{-4}_\varepsilon \Gamma \) in the sense that \( \Sigma \Gamma / \sigma^{-4}_\varepsilon = I \) (see e.g. Shaman (1975) and Beran 1994, pp. 109 ff.). 
\end{proof}
Let $m = \lfloor nh \rfloor$ be the integer part of $nh$. Let $v_i = \Psi(i/m)$, $-\infty < i < \infty$, and let $\mathbf{v} = (v_i)$ denote the corresponding doubly infinite vector. Then we have

$$g' \Lambda' \Lambda g = \frac{1}{4} h^{2k} \Psi^2(0) \mathbf{v}' \Lambda' \Lambda \mathbf{v}.$$

Observe that $v_i = 0$ for $i < -m$ or $i > m$. We have

$$\mathbf{v}' \Lambda' \Lambda \mathbf{v} = \sum_{j=-\infty}^{\infty} \left( \sum_{i=-m}^{m} v_i b_{i+j} \right)^2$$

(A.7)

$$= (2m + 1) \sum_{k=-2m}^{2m} \gamma(k) \frac{1}{2m + 1} \sum_{j=-m}^{m} \Psi(j/m) \Psi((j + k)/m).$$

Equation (A.7) can also be obtained by directly analyzing $\mathbf{v}' \Lambda' \mathbf{v}$.

Based on (A.7) we can obtained results for the cases with $\alpha = 0$, $0 < \alpha < 1$ and $-1 < \alpha < 0$, separately. Note that the methodology used in the proof of Theorem 3.1 in Hall and Hart (1990a) for the case with $0 < \alpha < 1$ is based on the assumption $b_{-i} = b_i$ for $i = 1, 2, ..., $ and is hence not suitable for the causal error process in this paper, since now we have $b_k = 0$ for $i < 0$. The methodology used in the following is developed based on the property (1.2) of a fractional time series, which does not involve the exact structure of $b_i$.

Assume that $\alpha = 0$. Note that in this case $\sum \gamma(k) > 0$ and $\sum |\gamma(k)| < \infty$. From (A.7) we have

$$\mathbf{v}' \Lambda' \Lambda \mathbf{v} = (2m + 1) \left( \sum \gamma(k) \right) \int_{-1}^{1} \Psi^2(u) \text{d}u.$$

Note that $h = n^{-1/(2k+1)}$ and $m = nh = n^{2k/(2k+1)} = h^{-2k}$ for $\alpha = 0$, whence

$$\frac{1}{4} h^{2k} \Psi^2(0) \mathbf{v}' \Lambda' \Lambda \mathbf{v} < \infty.$$

In the case with $0 < \alpha < 1$ the inverse process $\xi^l$ is an antipersistent process with the parameter $-1 < \alpha^l = -\alpha < 0$ in (2.7) and hence for $|k|$ sufficiently large we have $\gamma^l(k) \sim c_{\gamma} |k|^{-\alpha - 1}$, where $c_{\gamma} = 2 c_{\gamma} \Gamma(1 - \alpha) \sin(\pi \alpha^l / 2) < 0$ (see Beran, 1994 and Beran and Feng, 2001a), which implies that $\gamma^l(k)$ are ultimately negative for $|k|$ sufficiently large. Furthermore, we have $\sum \gamma^l(k) = 0$ and hence $\sum_{k=-m}^{m} \gamma^l(k) = -2 \sum_{k>m} \gamma^l(k) = O(m^{-\alpha})$. 

16
It follows from (A.7)

\[ \mathbf{v}' \mathbf{A}' \mathbf{A} \mathbf{v} = (2m + 1) \sum_{k=-2m}^{2m} \gamma^1(k) \frac{1}{2m + 1} \sum_{j=-m}^{m} \Psi(j/m) \Psi((j + k)/m) \]

\[ \leq (2m + 1) \sum_{k=-m}^{m} \gamma^1(k) \frac{1}{2m + 1} \sum_{j=-m}^{m} \Psi(j/m) \Psi((j + k)/m) \]

\[ = (2m + 1) O \left( \sum_{k=-m}^{m} \gamma^1(k) \right) = O(m^{1-\alpha}). \]

Now we have \( h = n^{-(1-\alpha)/(2k+1-\alpha)} \) and \( m = nh = n^{2k/(2k+1-\alpha)} \). This results in \( m^{1-\alpha} = h^{-2k} \), so that

\[ \frac{1}{4} h^{2k} \Psi^2(0) \mathbf{v}' \mathbf{A}' \mathbf{A} \mathbf{v} = h^{2k} O(h^{-2k}) < \infty. \]

If \(-1 < \alpha < 0\), the inverse process \( \xi^1 \) is a long-memory process with the parameter \( 0 < \alpha^1 = -\alpha < 1 \) in (2.7) and hence, for \(|k|\) sufficiently large, \( \gamma^1(k) \sim c_{\gamma^1} |k|^{-\alpha-1} \), where \( c_{\gamma^1} = 2 c_{\gamma} \Gamma(1 - \alpha^1) \sin(\pi \alpha^1 / 2) > 0 \), so that \( \gamma^1(k) > 0 \) for \(|k|\) sufficiently large. Furthermore, we have \( \sum_{k=-2m}^{2m} \gamma^1(k) = \infty \) with \( \sum_{j=-m}^{m} \gamma^1(k) = O(m^{-\alpha}) \). Note that \( \Psi \) can be chosen so that, for large \( k \), \( \sum_{j=-m}^{m} \Psi(j/m) \Psi((j + k)/m) \leq \sum_{j=-m}^{m} \Psi^2(j/m) \). Hence we have

\[ \mathbf{v}' \mathbf{A}' \mathbf{A} \mathbf{v} = (2m + 1) \sum_{k=-2m}^{2m} \gamma^1(k) \frac{1}{2m + 1} \sum_{j=-m}^{m} \Psi(j/m) \Psi((j + k)/m) \]

\[ \leq (2m + 1) \sum_{k=-2m}^{2m} \gamma^1(k) \frac{1}{2m + 1} \sum_{j=-m}^{m} \Psi^2(j/m) \]

\[ = (2m + 1) \sum_{k=-2m}^{2m} \gamma^1(k) \int_{-1}^{1} \Psi^2(u) du \]

\[ = O(m^{1-\alpha}). \]

In fact, we have

\[ \mathbf{v}' \mathbf{A}' \mathbf{A} \mathbf{v} = O(m^{1-\alpha}) \]

uniformly for \( \alpha \in (-1, 1) \). However, the derivation for this result is a little different in the three cases. Now, note that \( h = n^{-(1-\alpha)/(2k+1-\alpha)} \), whence, as before, \( m^{1-\alpha} = h^{-2k} \), so that

\[ \frac{1}{4} h^{2k} \Psi^2(0) \mathbf{v}' \mathbf{A}' \mathbf{A} \mathbf{v} = h^{2k} O(h^{-2k}) < \infty. \]

Theorem 1 is proved. 
References


Beran, J. (1999), SEMIFAR models – A semiparametric framework for modelling trends, long range dependence and nonstationarity, Discussion paper No. 99/16, Center of Finance and Econometrics, University of Konstanz.


