# On parameter estimation for locally stationary long-memory processes

Jan Beran
Department of Mathematics and Statistics,
University of Konstanz, Germany

December 2007

**Abstract**

We consider parameter estimation for time-dependent locally stationary long-memory processes. The asymptotic distribution of an estimator based on the local infinite autoregressive representation is derived, and asymptotic formulas for the mean squared error of the estimator, and the asymptotically optimal bandwidth are obtained. In spite of long memory, the optimal bandwidth turns out to be of the order $n^{-1/5}$ and inversely proportional to the square of the second derivative of $d$. In this sense, local estimation of $d$ is comparable to regression smoothing with iid residuals.

Keywords: long memory, fractional ARIMA process, local stationarity, bandwidth selection

## 1 Introduction

The usefulness of stationary long-memory processes for modeling time series has been demonstrated in the literature by numerous examples, including applications in hydrology, geophysics, economics, finance, climatology, physics, biology, medicine, music and telecommunications engineering among others (see e.g Mandelbrot 1977, Beran 1994, 2003, Lowen and Teich 2005). Long

memory of a second order stationary process $X_t$ is characterized by slowly decaying non-summable autocovariances

$$\gamma(k) = cov(X_t, X_{t+k}) \sim c_\gamma \, |k|^{2d-1} \;\; (|k| \to \infty) \tag{1}$$

where $d \in (0, \frac{1}{2})$, and a pole of the spectral density at the origin,

$$f_X(\lambda) = \frac{1}{2\pi} \sum_{k=-\infty}^{\infty} \gamma(k) e^{-ik\lambda} \sim c_f \, |\lambda|^{-2d} \;\; (|\lambda| \to 0). \tag{2}$$

Here " $\sim$ " means that the ratio of both sides tends to one. For some data sets, however, it has been observed that the assumption of stationarity is too restrictive, even after trends in the mean are removed. In particular, the long-memory parameter $d$, as well as other parameters characterizing the spectrum of the process, may change as a function of time. Data examples with time-varying $d$ can be found, for instance, in geophysics, oceanography, meteorology, economics, telecommunication engineering, medicine and other areas of statistical applications (see e.g. Beran et al. 1995, Vesilo and Chan 1996, Whitcher and Jensen 2000, Lavielle and Ludena 2000, Ray and Tsay 2002, Whitcher et al. 2002, Granger and Hyung 2004, Falconer and Fernandez 2007). This motivates introducing locally stationary processes with long-range dependence. For locally stationary processes with short-range dependence see e.g. Subba Rao (1970), Hallin (1978), Priestley (1981), Dahlhaus (1986, 1987), Dahlhaus and Giraitis (1998), Moulines et al. (2005). Jenssen and Whitcher (2000) define locally stationary fractional ARIMA (FARIMA) processes (Granger and Joyeux 1980, Hosking 1981), and estimate parameters using wavelets. Alternatively, given a specific linear model such as a fractional ARIMA, one may consider local estimation based on estimated innovations. This is the approach taken here. For related estimates for stationary long-memory processes, see e.g. Fox and Taqqu (1986), Yajima (1985), Giraitis and Surgailis (1990) and Beran (1995). Modeling time series by locally stationary long-memory processes is closely related to change point detection in the spectral domain. For spectral change point detection in the long memory context, see e.g. Giraitis and Leipus (1990, 1992), Horváth and Shao (1999), Lavielle and Ludena (2000), Ray and Tsay (2002), Ben Hariz et al. (2007), also see Kokoszka and Leipus (2003) for a review. It should also be noted that shifts in the mean can also give rise to long-memory type dependence (see e.g. Granger and Ding 1996, Diebold and

Inoue 2001). Distinguishing nonconstant mean from stationary long memory is possible either under regularity assumptions on a trend function (see e.g. Hall and Hart 1990, Csörgö and Mielniczuk 1995, Ray and Tsay 1997, Beran and Feng 2002a,b) or in the presence of a finite number of change points (see e.g. Horváth and Kokoszka 1997, Kuan and Hsu 1998, Wright 1998, Ray and Tsay 2002, Sibbertsen 2004, Berkes et al. 2006). In this paper, we assume the mean to be constant. The methods proposed here may be extended to situations with nonconstant mean by combining them with suitable algorithms for nonparametric regression smoothing (Beran and Feng 2002b) or change point estimation (Horváth and Kokoszka 1997).

Specifically, we consider a sequence of processes $X_{t,n}$ having a time-varying infinite autoregressive representation

$$X_{t,n} = \sum_{j=1}^{\infty} b_{j,n} X_{t-j,n} + \varepsilon_t \tag{3}$$

where $\varepsilon_t$ are iid zero-mean random variables with finite variance $\sigma_\varepsilon^2 = \sigma_\varepsilon^2(t/n)$ and $b_{j,n} = b_j(\theta(t/n))$. Here $\sigma_\varepsilon^2(u)$ and $\theta(u) = (d(u), \theta_2(u), ..., \theta_k(u))^T$ ($u \in [0,1]$) are sufficiently smooth functions of rescaled time. Moreover, for fixed $u = t/n$, the value of $d(u) \in (0, \frac{1}{2})$ is assumed to be such that

$$0 < \lim_{j\to\infty} j^{d+1} b_j(\theta(u)) = c_b < \infty \tag{4}$$

and

$$0 < \lim_{\lambda\to 0} 2\pi\sigma_\varepsilon^{-2}\lambda^{-2d} \left| 1 - \sum_{j=1}^{\infty} b_j e^{-ij\lambda} \right|^2 = c_f^{-1} < \infty \tag{5}$$

where $c_b$, $c_f$ are positive constants. In the case of a fractional ARIMA$(p,d,q)$ process, we have $c_f = \sigma_\varepsilon^2/(2\pi)$ and for $z \in \mathbb{C}$, with $|z| \le 1$ and $z \neq 1$,

$$1 - \sum_{j=1}^{\infty} b_j(d) z^j = \varphi(z)\psi^{-1}(z)\,(1-z)^d \tag{6}$$

where

$$\varphi(z) = 1 - \varphi_1 z - ... - \varphi_p z^p \neq 0 \ (|z| \le 1), \tag{7}$$

$$\psi(z) = 1 - \psi_1 z - ... - \psi_q z^q \neq 0 \ (|z| \le 1). \tag{8}$$

The time varying parameters are then $\sigma_\varepsilon^2(t/n) = var(\varepsilon_t)$ and $\theta(t/n) = [d(t/n), \varphi_1(t/n), ..., \varphi_p(t/n), \psi_1(t/n), ..., \psi_q(t/n)]^T$. Note that, $d(u) > 0$ means

that locally the process has (approximately) a spectral density with a pole at the origin proportional to $|\lambda|^{-2d(u)}$, and, in the course of time, the rate of divergence of the pole changes slowly.

In this paper, estimation of $\theta(.)$ based on the autoregressive representation (3) is considered. For Gaussian innovations $\varepsilon_t$, this corresponds to an approximate maximum likelihood estimator. Two questions are addressed: 1. asymptotic distribution of $\hat{\theta}(u)$, and 2. the choice of a suitable bandwidth that determines which observations in the neighbourhood of $u$ (or $nu$ on the original time scale) are used for the local estimate. The paper is organized as follows. The asymptotic distribution of $\hat{\theta}$ is derived in section 2. Section 3 addresses the issue of bandwidth choice. In particular, an asymptotic expression for the mean squared error of $\hat{d}$ is obtained. The asymptotically optimal bandwidth turns out to be proportional to $n^{-1/5}$ and inversely proportional to $\{d''\}^2$. In spite of long-range dependence, the formula are similar to results in the context of regression smoothing with iid errors. For the case of short-memory $AR(p)$ processes also see Dahlhaus and Giraitis (1998). Simulations and data examples in section 3 illustrate the approximate validity of the asymptotic results for finite samples. Moreover, a simple iterative plug-in algorithm for data driven bandwidth choice is proposed. General comments in section 4 conclude the paper. Proofs are given in the appendix.

# 2 Estimation, asymptotic distribution

Denote by $\theta^o(u)$ the true parameter curve. We consider estimation of $\theta^o(u)$ for a fixed rescaled time point $u_o \in (0,1)$. Let $t_o(n) = [nu_o]$, $u_{t,n} = t(n)/n$, and denote by $K : \mathbb{R} \to \mathbb{R}_+$ a nonnegative kernel function with $K(-x) = K(x)$, $K(x) = 0$ ($|x| > 1$) and $\int K(x)dx = 1$. A local estimate of $\theta^o(u_o)$ is defined by minimizing

$$L_n(\theta) = \sum_{t=t_o-[nb]}^{t_o+[nb]} K(\frac{t - t_o(n)}{nb}) e_t^2(\theta) \tag{9}$$

or by solving

$$\dot{L}_n(\hat{\theta}) = \sum_{t=t_o-[nb]}^{t_o+[nb]} K(\frac{t - t_o(n)}{nb}) e_t(\hat{\theta}) \dot{e}_t(\hat{\theta}) = 0 \tag{10}$$

where

$$e_t(\theta) = X_t - \sum_{j=1}^{t-1} b_j(\theta) X_{t-j}, \tag{11}$$

$$\dot{e}_t(\theta) = -\sum_{j=1}^{t-1} \dot{b}_j(\theta) X_{t-j} \tag{12}$$

and

$$\dot{b}_j(x) = \frac{\partial}{\partial x} b_j(x) = \left[ \frac{\partial}{\partial x_1} b_j(x_1), ..., \frac{\partial}{\partial x_k} b_j(x_k) \right]^T. \tag{13}$$

Note that $e_t(\theta)$ and $\dot{e}_t(\theta)$ are approximations of

$$\varepsilon_t(\theta) = X_t - \sum_{j=1}^{\infty} b_j(\theta) X_{t-j} \tag{14}$$

and

$$\dot{\varepsilon}_t(\theta) = -\sum_{j=1}^{\infty} \dot{b}_j(\theta) X_{t-j} \tag{15}$$

Under suitable regularity conditions, $X_{t,n}$ defined by (3) is a locally stationary process (see e.g. Jenssen and Whitcher 2000), i.e. there exist transfer functions $A_{t,n}(e^{-i\lambda}; \theta)$ and $A(e^{-i\lambda}; \theta)$ such that $X_{t,n}$ has a spectral representation

$$X_{t,n} = \sigma_\varepsilon(u_{t,n}) \int_{-\pi}^{\pi} e^{it\lambda} A^o_{t,n}\left(e^{-i\lambda}; \theta(u_{t,n})\right) dZ_\varepsilon(\lambda) \tag{16}$$

and

$$\sup_{\lambda \in [-\pi,\pi], t=1,2,...,n} \left| \sigma_\varepsilon(u_{t,n}) [A_{t,n}(e^{-i\lambda}; \theta(u_{t,n})) - A(e^{-i\lambda}; \theta(u_{t,n}))] \right| \leq C n^{-1} \tag{17}$$

for all $n$ and a constant $C$. In the following we will use the notation $f\{\lambda; \theta(u_{t,n})\} = (2\pi)^{-1} |A(e^{-i\lambda}; \theta(u_{t,n}))|^2$ for the standardized local spectral density. The asymptotic distribution of $\hat{\theta}(u_o)$ is then characterized by

**Theorem 1** *Let $X_{t,n}$ be generated by (3), and $u_o \in (0,1)$. Assume that, as $n$ tends to infinity, $b \to 0$ and $nb^3 \to \infty$. Then, under assumptions (A1)-(A7) given in the appendix, there is a sequence $\hat{\theta}_n$ such that $\dot{L}_n(\hat{\theta}_n) = 0$ and $\hat{\theta}_n \to \theta^o(u_o)$ in probability. Moreover,*

$$\sqrt{nb}(\hat{\theta}_n - E(\hat{\theta}_n)) \to_d N(0, V) \tag{18}$$

*where*

$$V = J^{-1}(\theta^o) \int_{-1}^{1} K^2(x)dx \tag{19}$$

*with*

$$J(\theta^o) = \left[ \frac{1}{4\pi} \int_{-\pi}^{\pi} \frac{\partial}{\partial \theta_r} \log f(\lambda; \theta^o) \frac{\partial}{\partial \theta_s} \log f(\lambda; \theta^o) d\lambda \right]_{r,s=1,\ldots,k} \tag{20}$$

**Remark 1** *The estimate of $\sigma_\varepsilon^2(u_o)$ can be defined similarily by*

$$\hat{\sigma}_\varepsilon^2(u_o) = \sum_{t=t_o-[nb]}^{t_o+[nb]} K(\frac{t - t_o(n)}{nb}) e_t^2(\theta) \tag{21}$$

*As in the stationary case, $\hat{\sigma}_\varepsilon^2(u_o)$ can be shown to be asymptotically independent of $\hat{\theta}_n$. Also note that the asymptotic distribution of $\hat{\theta}$ does not depend on $\sigma_\varepsilon^2$. To simplify presentation, and also since the focus here is on estimation of $\theta$, we will therefore assume that $\sigma_\varepsilon^2(u)$ is known and constant.*

**Remark 2** *Note that in general $V$ depends on $\theta^o$. This property is inherited from the maximum likelihood estimator for the the stationary case (see e.g. Yajima 1985, Fox and Taqqu 1986, Dahlhaus 1989, Giraitis and Surgailis 1990), since $V$ is identical to the corresponding asymptotic covariance matrix of the MLE. An exception is, for instance, the fractional ARIMA$(0, d, 0)$ model (see example 1 below).*

**Remark 3** *FARIMA$(p, d, q)$ processes are very flexible with respect to approximating linear dependence structures (i.e. the underlying spectral density). The difference operator $(1 - B)^d$ incorporates a pole at the origin of the form $c_f |\lambda|^{-2d}$. The ARMA part of $f_X(\lambda)$ $f_{ARMA}(\lambda) = \left|\psi(e^{-i\lambda})/\varphi(e^{-i\lambda})\right|^2$ approximates the bounded part of a spectral density by a trigonometric rational function of degrees $p$ and $q$. This approximation can be made arbitrarily close, uniformly in $\lambda$. In practice, $p$ and $q$ have to be estimated from data. Beran et al. (1998) showed that an appropriate version of the AIC or BIC can be used in spite of the presence of long memory.*

**Remark 4** *While FARIMA$(p, d, q)$ processes provide flexible models of the spectral density whenever second order stationarity can be assumed exactly or in good approximation (for instance locally), time dependence of the parameters increases flexibility in another direction. It allows for structural changes in the dependence structure.*

**Example 1** *For the rectangular kernel* $K(x) = \frac{1}{2}1\{|x| \leq 1\}$*, and a local fractional ARIMA*$(0, p, d)$ *process, the asymptotic variance of* $\sqrt{nb}(\hat{d} - d^o(u_o))$ *is equal to*

$$V = \frac{6}{\pi^2}\frac{1}{2} = \frac{3}{\pi^2} \approx 0.304. \tag{22}$$

*Specifically, for the simulations in section 4.1, we consider* $d(u) = 0.05 + 0.4u^3$ *and* $\varepsilon_t$ *iid* $N(0, \sigma_\varepsilon^2)$*. In this case, assumptions (A1)-(A7) can be verified as follows: (A1), (A3) and (A6) are known from maximum likelihood estimation for the stationary case (see e.g. Fox and Taqqu 1986, Dahlhaus 1989, Giratis and Surgailis 1990); (A2) follows, since* $d(u) \in (0, \frac{1}{2})$ *for all* $u \in [0, 1]$*; (A4) follows, since in this case* $D_n \in \mathbb{R}_+$ *is proportional to* $n$ *so that* $D_n^{-\frac{1}{2}} \ddot{S} D_n^{-\frac{1}{2}} \to_d c > 0$ *follows from the law of large numbers and* $E\left[\partial^2/\partial\theta^2 \varepsilon_t^2|_{\theta=\theta^o}\right] > 0$*; (A5) follows, since* $D_n$ *and hence also* $\lambda_n$ *is proportional to* $n$*; (A7) follows from results in extreme value theory for stationary Gaussian processes (see e.g. Embrechts et al. 1997 and Hüsler et al. 2003).*

# 3 Asymptotic mean squared error and bandwidth choice

An important question that needs to be addressed whenever nonparametric smoothing is applied is the choice of a suitable bandwidth. In theorem 1, no indication is given regarding the bias $E[\hat{d}(u)] - d(u)$, and the general conditions on the bandwidth are not specific enough for practical purposes. The importance of data-driven bandwidth choice is is illustrated by figure 1. A stationary fractional ARIMA$(0, 0.4, 0)$ process of length $n = 250$ is simulated (figure 1a) and $d^o(u) \equiv 0.4$ is estimated by (9) using the bandwidth $b = \frac{1}{4}n^{-1/5}$. The dotted line in figure 1b is the resulting estimate of $d^o(u)$. Obviously, the bandwidth is too small as the estimated curve is mostly far from $d^o = 0.4$ and varies erratically between very weak ($\hat{d}(u) = 0.15$) and very strong ($\hat{d}(u) = 0.45$) long-range dependence. A similar example is given in figure 2. Here, $n = 1000$ and $d(u) = 0.05 - 0.4u^3$. Again, the dotted line representing the estimate with $b = \frac{1}{4}n^{-1/5}$ is far from the true function $d^o(u)$ and fluctuates quite erratically.

To simplify presentation, we restrict attention to the one-dimensional case with $\theta(u) = d^o(u) \in (0, \frac{1}{2})$. An asymptotic formula for the mean squared error of $\hat{d}(u_o)$ is given by

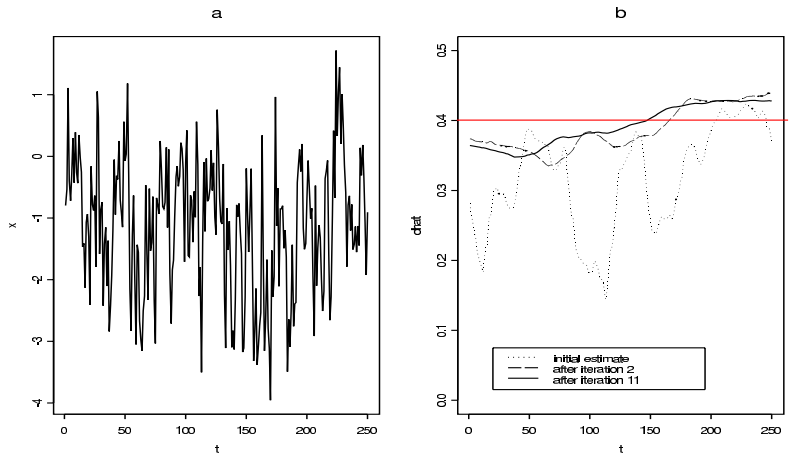


Figure 1: A simulated fractional ARIMA(0,0.4,0) series (figure 1a) and estimates of $d^o(u)$ (figure 1b) using an initial bandwidth $b = 0.25n^{-1/5}$, and two bandwidths obtained after 2 and 11 iterations of the plug-in algorithm defined in section 4.2.

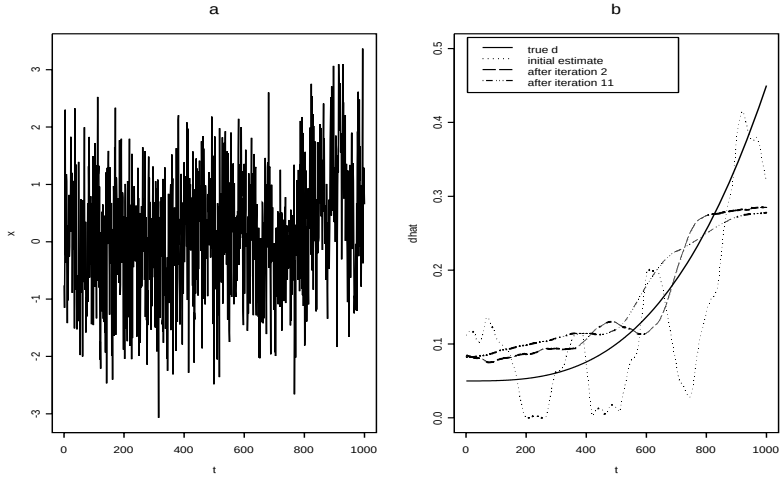


Figure 2: A simulated locally stationary fractional ARIMA(0,d,0) series (figure 2a) of length $n = 1000$ with $d^o(u) = 0.05 + 0.4u^3$, and estimates of $d(u)$ (figure 2b) using an initial bandwidth $b = 0.25n^{-1/5}$, and two bandwidths obtained after 2 and 11 iterations of the plug-in algorithm defined in section 4.2.

**Theorem 2** *Let $d \in C^2[0,1]$ and $d''(u_o) \neq 0$. Then under the assumptions of theorem 1, we have, as $n \to \infty$,*

1. *Bias:*

$$E[\hat{d}(u_o)] - d^o(u_o) = b^2 \frac{1}{2} d''(u_o) \int_{-1}^{1} K(x) x^2 dx + o(b^2) \tag{23}$$

2. *Variance:*

$$var[\hat{d}(u_o)] = (nb)^{-1} J^{-1}(d^o) \int_{-1}^{1} K^2(x) dx + o((nb)^{-1}) \tag{24}$$

3. *Mean squared error:*

$$MSE(d_o) = E[(\hat{d} - d^o)^2] = b^4 C_1 + (nb)^{-1} C_2 + o\left\{\max(b^4, (nb)^{-1})\right\} \tag{25}$$

*with*

$$C_1(u_o) = \left[\frac{1}{2} d''(u_o) \int_{-1}^{1} K(x) x^2 dx\right]^2 \tag{26}$$

*and*

$$C_2(u_o) = J^{-1}(\theta^o) \int_{-1}^{1} K^2(x) dx \tag{27}$$

Theorem 2 implies the following asymptotically optimal bandwidth.

**Corollary 1** *Under the assumptions of theorem 2, the asymptotic mean squared error is minimized by*

$$b_{opt}(u_o) = n^{-1/5} C_3(u_o) \tag{28}$$

*with*

$$C_3(u_o) = \left[\frac{C_2(u_o)}{4 C_1(u_o)}\right]^{1/5} \tag{29}$$

*The resulting MSE is then of the order $O(n^{-4/5})$.*

**Remark 5** *The formulas for $MSE$ and $b_{opt}$ are analogous to results in non-parametric regression with iid errors, as well expressions known for locally stationary $AR(p)$ processes (see e.g. Dahlhaus and Giraitis 1998). This may be surprising at first sight, since we are dealing with long-memory processes and $d(u)$ cannot be estimated directly by kernel or local polynomial regression. The result is in sharp contrast to regression smoothing with long-memory errors. There, the optimal bandwidth depends on $d$ and is of a larger order than $n^{-1/5}$ (Hall and Hart 1990, Ray and Tsay 1997, Beran and Feng 2002).*

**Remark 6** *A globally optimal bandwidth for estimating $d(u)$ in an interval $[\delta, 1-\delta] \subset [0,1]$ may be defined by minimizing the integrated mean squared error*

$$IMSE = b^4 \int_{\delta}^{1-\delta} C_1(u)du + (nb)^{-1} \int_{\delta}^{1-\delta} C_2(u)du. \tag{30}$$

*In general, globally optimal bandwidth choice is easier to implement in practice, since $d''$ can be arbitrarily close to zero, thus leading to highly variable (and possibly infinite) bandwidths (see e.g. Brockmann 1993, for further comments on local bandwidth choice).*

**Remark 7** *An estimated curve $\hat{d}(u)$, obtained from (9), may be smoothed further by applying kernel or local polynomial smoothing directly to $\hat{d}(u)$. This can be done without a noticeable change of the mean squared error, provided that the same bandwidth $b_{opt}$ is used.*

**Remark 8** *Theorem 2 can easily be generalized to FARIMA$(p,d,q)$ processes with $p$ and $q$ arbitrary. The only difference is that the asymptotic variance of $\hat{d}$ is no longer parameter free.*

**Remark 9** *An extension of the results that would be of interest is to consider locally stationary FARIMA models with stable innovations. Since second moments do not exist, this would require another estimation approach. For instance, Stoev and Taqqu (2005), consider wavelet based estimaton (also see Stoev et al. 2002).*

**Remark 10** *Theorems 1 and 2 can be used to obtain pointwise confidence intervals for $d(u)$. Note that, for the optimal bandwidth the squared bias is of the same order as the variance. To construct confidence intervals in this case, an estimate of the bias is required. To obtain simultaneous confidence bands, a functional limit theorem or appropriate computational methods, such as bootstrap, would be needed. For short-memory processes, bootstrap methods in the context of nonparametric regression have been considered for instance in Härdle and Marron (1991) and Hall (1992). Tribouley (2004) considers the same problem using wavelet estimates . In the long-memory context considered here, the question of simultaneous confidence intervals is an open problem. In particular, bootstrap procedures are considerably more complex than under independence or short memory (see e.g. Lahiri 2003).*

| | $b = b_{small}$ | $b = b_{large}$ | $b = b_{opt}$ | asymptotic formula |
|---|---|---|---|---|
| $n^{4/5}\alpha_n(b)$ | | | | |
| $n = 250$ | 0.019 | 0.383 | 0.031 | 0.044 |
| $n = 500$ | 0.007 | 0.500 | 0.054 | 0.044 |
| $n = 1000$ | 0.017 | 0.565 | 0.048 | 0.042 |
| $n^{4/5}\beta_n(b)$ | | | | |
| $n = 250$ | 0.300 | 0.248 | 0.227 | 0.176 |
| $n = 500$ | 0.320 | 0.248 | 0.229 | 0.176 |
| $n = 1000$ | 0.334 | 0.185 | 0.208 | 0.170 |
| $n^{4/5}IMSE^*(b)$ | | | | |
| $n = 250$ | 0.319 | 0.631 | 0.258 | 0.220 |
| $n = 500$ | 0.327 | 0.748 | 0.283 | 0.219 |
| $n = 1000$ | 0.351 | 0.751 | 0.256 | 0.212 |

Table 1: Simulated values of $n^{4/5}$ times the squared bias, variance and $IMSE^*$ for a fractional ARIMA(0,d,0) model with $d(u) = 0.05 + 0.4u^3$. The results are based on one hundred simulations.

# 4 Data examples and computational aspects

## 4.1 Simulations

To examine in how far the asymptotic formulae apply to finite samples, a small simulation study is carried out. For $n = 250$, 500 and 1000, one hundred simulations of a locally stationary FARIMA$(0, d, 0)$ with $d(u) = 0.05 + 0.4u^3$ are carried out. Estimates of $d(u)$ are based on (9) with $K(x) = \frac{1}{2}1\{-1 \leq x \leq 1\}$. For each simulated series, $d$ is estimated for $u_j = 0.2 + \Delta \cdot j/n$ where $\Delta = 20$ and $0.2 \leq u_j \leq 0.8$. The optimal bandwidth is defined by minimizing the corresponding discrete approximation of the asymptotic $IMSE$ over the range $[0.2, 0.8]$, given by

$$IMSE_n^*(b) = b^4 \frac{\Delta}{n} \sum_j C_1(u_j) + (nb)^{-1} \frac{\Delta}{n} \sum_j C_2(u_j) \tag{31}$$

$$= \alpha_n(b) + \beta_n(b) \tag{32}$$

For comparison, estimates based on a smaller and a larger bandwidth, namely $b_{small} = \frac{1}{2}b_{opt}$ and $b_{large} = 2b_{opt}$ respectively, are calculated. The following simulated values are listed in table 1: a) the rescaled integrated squared bias,

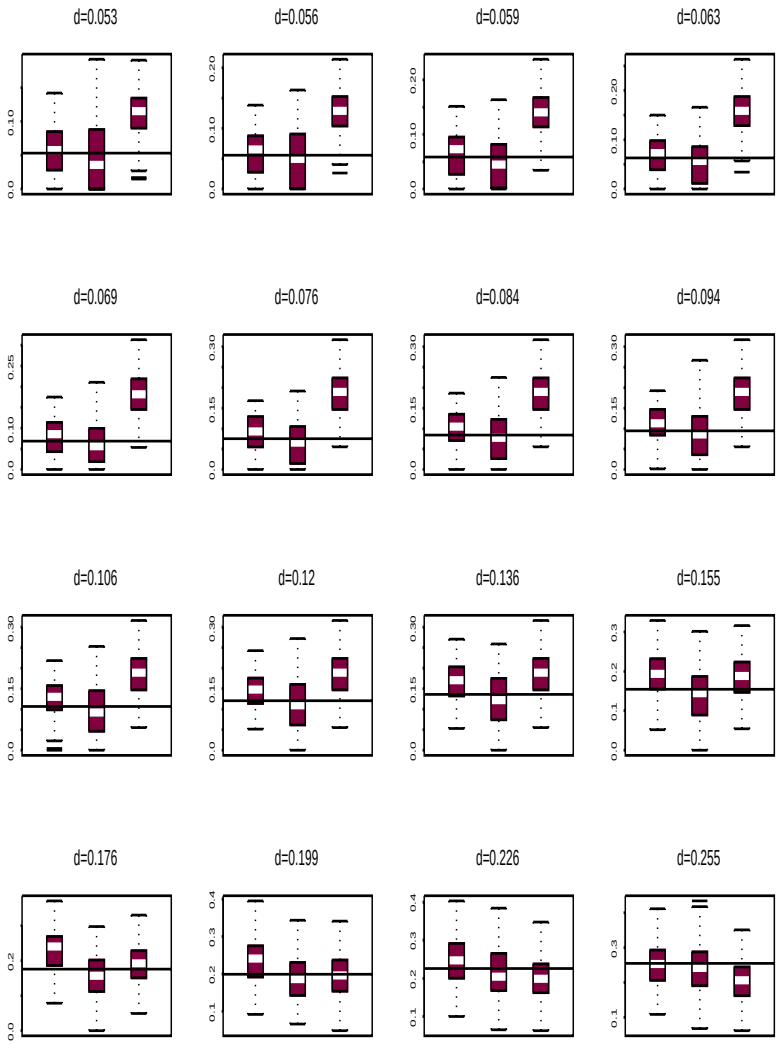


Figure 3: Boxplots of estimates of $d(u)$ for different values of $u$ and the three bandwidths $b_{small}$, $b_{opt}$ and $b_{large}$ respectively. The results are based on one hundred simulations of a locally stationary fractional ARIMA(0,d,0) series of length $n = 500$ and $d(u) = 0.05 + 0.4u^3$.

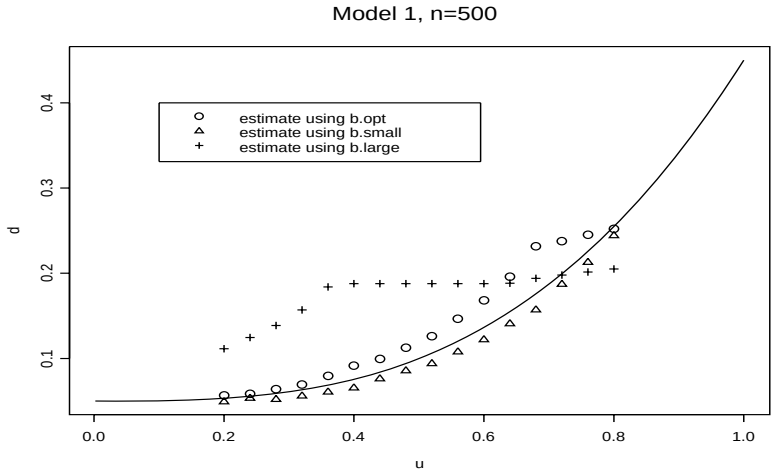


Figure 4: True function $d(u)$ and averages of $\hat{d}(u)$ for the three bandwidths $b_{small}$, $b_{opt}$ and $b_{large}$ respectively. The results are based on one hundred simulations of a locally stationary fractional ARIMA(0,d,0) series of length $n = 500$ and $d(u) = 0.05 + 0.4u^3$.

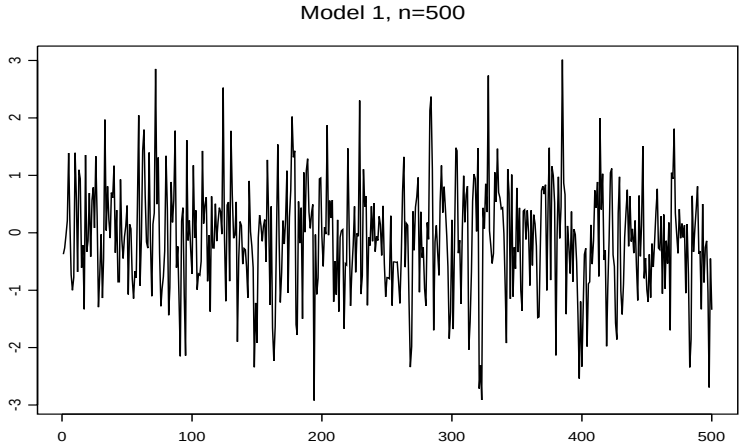


Figure 5: Simulated locally stationary fractional ARIMA(0,d,0) series with $d(u) = 0.05 + 0.4u^3$.

$n^{4/5}\alpha_n(b)$; b) the rescaled integrated variance, $n^{4/5}\beta_n(b)$ and c) $n^{4/5}IMSE_n^*(b)$. For comparison, the theoretical values obtained from theorem 2 are also given. The results indicate that, for $b_{opt}$, the simulated values are reasonably close to the asymptotic approximation, even for moderate sample sizes. The results also show a considerably higher mean squared error for estimates based on $b_{small}$ and $b_{large}$. The reason is a large variance for $b_{small}$ and a large bias for $b_{large}$ respectively. The best results are obtained for $b_{opt}$. This is also illustrated in figure 3 with boxplots of the three estimates for different values of $u_o$. Figure 5 shows a typical simulated series of length $n = 500$. Visually, it seems very difficult to notice any deviations from stationarity. Nevertheless, the simulated averages of $\hat{d}(u)$ (figure 4) are close to the true curve for $b_{small}$ and $b_{opt}$.

## 4.2 Computational issues

For observed time series, the constant $C_{opt}$, and hence $b_{opt}$, are unknown and have to be estimated. In the context of nonparametric regression with iid errors, various data driven methods for bandwidth choice are known (see e.g. Gasser et al. 1991, Herrmann et al. 1992). Similar algorithms may be applied here. For instance, a simple iterative plug-in algorithm can be defined as follows.

**Algorithm 1**

- *Step 1: Set $j = 0$ and set $b_j$ equal to an initial bandwidth.*
- *Step 2: Estimate $d(.)$ using the bandwidth $b_j$.*
- *Step 3: For each $u_o$, fit a local polynomial regression $\beta_o(u_o)+\beta_1(u_o)(u-u_o)+\frac{1}{2}\beta_2(u_o)(u-u_o)^2$ directly to $\hat{d}(u)$ (plotted against $u$) using a suitable bandwidth $b_2$.*
- *Step 4: For each $u_o$, set $\hat{d}''(u_o) = 2\beta_2(u_o)$, and calculate an estimate of $C_{opt}(u_o)$ (or a global value $C_{opt}$ minimizing the integrated mean squared error).*
- *Step 5: Set $j = j+1$ and $b_j = C_{opt}n^{-1/5}$. If $b_j$ and $b_{j-1}$ are very similar (according to a specified criterion), go to step 6. Otherwise go to step 2.*

- *Step 6: Fit a kernel regression with kernel $K$ and bandwidth $b_j$ to $\hat{d}(u)$ directly.*

**Remark 11** *Step 6 is not necessary. The purpose of smoothing the final estimate by direct kernel (or local polynomial) regression is to obtain a somewhat smoother curve, without changing the essential order of the mean squared error.*

**Remark 12** *The algorithm is applicable to arbitrary locally stationary long-memory models, such as for instance FARIMA$(p, d, q)$ with $p$ and $q$ arbitrary. In general, the asymptotic variance of $\hat{d}$ depends on the unknown parameters so that the estimated values of $\hat{\theta}$ parameters have to be plugged in.*

**Remark 13** *The algorithm uses fixed values of $p$ and $q$. To obtain a fully automatic procedure, a data driven model choice criterion would have to be included. Beran et al. (1998) derived a version of the AIC (and BIC) for stationary FARIMA$(p, d, 0)$ models. For locally stationary models, model choice is, to a large extent, an open problem. In a recent study, van Bellegem and Dahlhaus (2006) proposed an AIC-type criterion for short-memory AR(p) processes with time-varying coefficients, under the assumption that $p$ remains constant. An adaptation of their ideas to the long-memory context may be possible, but would require a detailed analysis to avoid artifacts such as overfitting and confusion between $d$ and autoregressive parameters (also see Beran et al. 1998 for comments on the latter problem).*

Experience with simulated and real data sets shows that convergence is reached within a few iterations. To illustrate this, we consider the two simulated examples in figures 1 and 2. The initial bandwidth $b_o = \frac{1}{4}n^{-1/5}$ leads to highly variable estimates. These estimates are misleading, since they suggest extreme local fluctuations in $d$. Considerably improved estimates are obtained already after 2 iterations. These estimates remain almost unchanged by further iterations.

## 4.3 Data examples

### 4.3.1 Nile river minima

The yearly minimal water levels of the Nile River (622-1284 AD, Tousson 1925), measured at the Roda Gauge near Cairo, are one of the prime examples of long-memory processes. The periodogram (in log-log-coordinates, figure 6b) shows a typical negative slope for all frequencies. It has been noted by

some authors, however, that the series may not be completely homogeneous (Beran and Terrin 1994, Beran 1994, Whitcher et al. 2002, Ray and Tsay 2002). In particular, about the first one hundred observations seem to follow a slightly different pattern. Beran and Terrin (1994) consider, for instance, the following simple heuristic test of the null hypothesis that $d$ is constant (also see Beran 1994). A FARIMA$(0, d, 0)$ model is fitted to six disjoint blocks of 100 observations (first block: $t = 1, ..., 100$; last block: $t = 501, ..., 600$). Under the null hypothesis, the six estimates $\hat{d}_1, ..., \hat{d}_6$ are approximately independent $N(0, v)-$distributed with $v = (100)^{-1}6/\pi^2 \approx 0.00608$ so that the test statistic $T = \sum \left(\hat{d} - \bar{d}\right)^2 / v$ with $\bar{d} = 6^{-1} \sum \hat{d}_i$ is approximately $\chi^2_5-$distributed. The observed value is $T = 22.8$ leading to a p-value of $P(\chi^2_5 > 22.8) = 0.0004$. A local FARIMA$(0, d, 0)$ fit based on the iterative plug-in algorithm defined in the previous section (and the integrated mean squared error as criterion), confirms this finding (figure 6c). Visually, the change can be seen by comparing the log-log-periodogram plots of the first one hundred observations (figure 6d) with the plots for observations 101 through 200 (figure 6e) and 201 through 300 (figure 6f) respectively. Since the impression is that of a rather abrupt change, local bandwidth choice may be more appropriate for this data. We therefore also applied the iterative algorithm using locally optimal bandwidths $C_{opt}(u_o)n^{-1/5}$. The result in figure 7 does indeed point in favour of an abrupt change. Similar findings based on statistical tests were obtained e.g. in Beran (1994) and Whitcher et al. (2002). Whitcher et al. (2002) conjecture that the change in $d$ may be related to the construction of a new measuring device around 715 AD.

### 4.3.2 Tree ring widths

Figure 7a shows a tree ring width series (chronology) of bristlecone pine at Mammoth Creek, Utah, USA (D. A. Graybill, //ftp.ncdc.noaa.gov). The periodogram in log-log-coordinates (figure 7b) shows a clear negative slope near the origin, indicating strong long memory. The estimated function $\hat{d}(u)$ in figure 7c is essentially monotonically decreasing. The decrease in $d$ is illustrated in figures 7d through e, with log-log-periodgrams for the years 1 to 400, 901 to 1300 and 1501 to 1900 respectively.

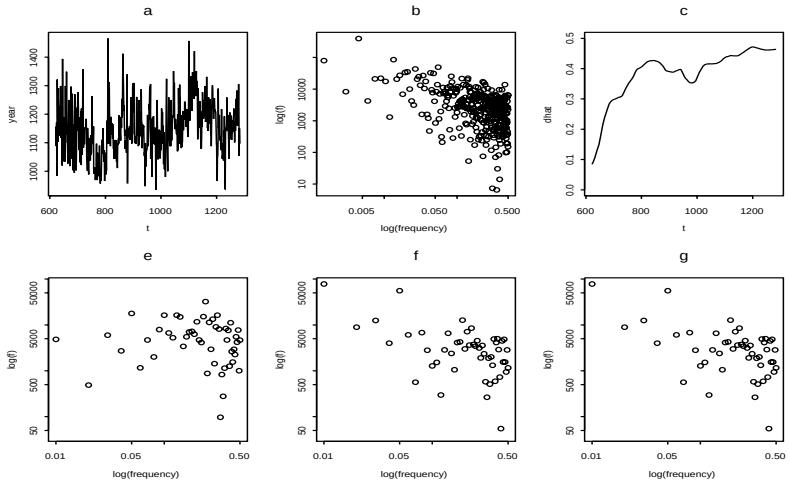


Figure 6: Nile River minima (622-1284 AD, Tousson 1925), measured at the Roda Gauge near Cairo: a) observed series, b) log-log-periodogram, c) estimate of $d(u)$ plotted against year, d)-f) log-log-periodograms for observations 1 to 100, 101 to 200 and 201 to 300 respectively.

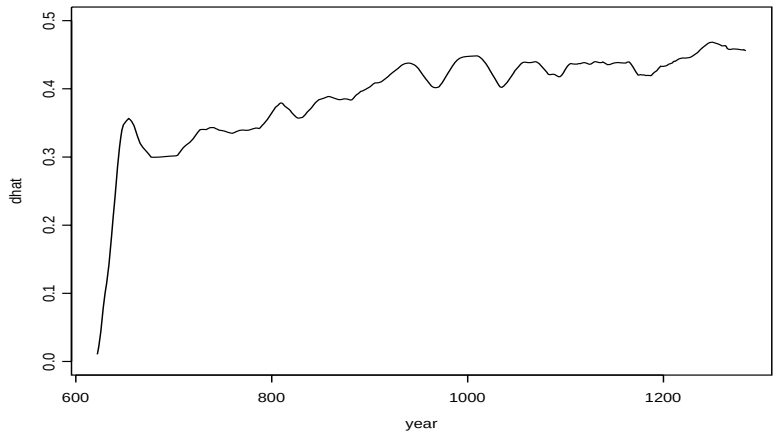


Figure 7: Nile River minima (622-1284 AD): Estimate of $d(u)$ based on the plug-in algorithm with local bandwidth choice.

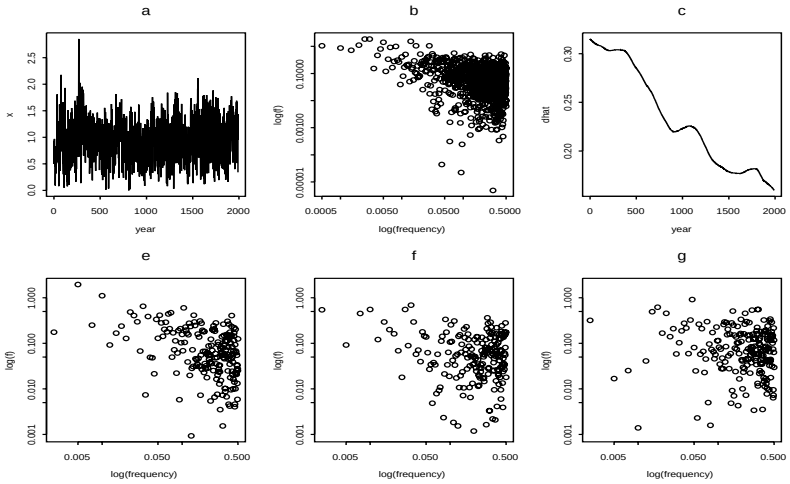


Figure 8: Tree ring data (Mammoth Creek, Utah): a) observed series, b) log-log-periodogram, c) estimate of $d(u)$ plotted against year, d)-f) log-log-periodograms for observations 1 to 400, 901 to 1300 and 1501 to 1900 respectively.

# 5 Final remarks

In this paper, some basic issues regarding parameter estimation for locally stationary long-memory models were addressed. A number of practically relevant open questions remain. These include simultaneous nonparametric trend estimation, alternative smoothing techniques and boundary problems.

# Appendix

## 5.1 Assumptions

Let

$$\ddot{S}_n(\theta) = \sum_{t=t_o-[nb]}^{t_o+[nb]} K(\frac{t-t_o(n)}{nb}) \frac{\partial^2}{\partial \theta^2} \varepsilon_t^2(\theta) = \sum_{t=t_o-[nb]}^{t_o+[nb]} \ddot{S}_{t,n}(\theta).$$

and

$$D_n = E[\ddot{S}_n(\theta)]$$
$$= \left\{ \sum_{t=t_o-[nb]}^{t_o+[nb]} K(\frac{t-t_o(n)}{nb}) E\left[\frac{\partial^2}{\partial\theta_r\partial\theta_s}\varepsilon_t^2(\theta)\right] \right\}_{r,s=1,\ldots,k}.$$

We will use the notation $\theta^o(u)$ ($u \in [0,1]$) for the true parameter curve and $\theta^o = \theta^o(u_o)$ for the value of $\theta^o(u)$ at a specific (rescaled) time point $u_o \in [0,1]$.

(A1) $A(e^{-i\lambda};\theta_1) \equiv A(e^{-i\lambda};\theta_2)$ (a.s. with respect to the Lebesgue measure) implies $\theta_1 = \theta_2$

(A2) $\theta^o \in \Theta^o \subset \Theta$ where $\Theta^o$ is an open set;

(A3) $A_{t,n}(e^{-i\lambda};\theta)$, $A(e^{-i\lambda};\theta)$, $b_j(\theta)$ are three times continously differentiable w.r.t. $\theta$

(A4) Define the $\delta-$neigbourhood $N_n(\theta^o,\delta) = \{\theta : (\theta-\theta^o)^T D_n(\theta^o)(\theta-\theta^o) \leq \delta^2\}$ for some fixed $\delta \geq 1$. Then $D_n^{-\frac{1}{2}}(\theta)\ddot{S}_n(\theta)D_n^{-\frac{1}{2}}(\theta)$ converges in probability to the $k \times k$ identity matrix $I$ uniformly in $N_n$, with respect to the Matrix norm $\|x\| = \sum_{i,j}|x_{ij}|$.

(A5) Let $\lambda_{\min}(\theta^o,n)$ be the smallest eigenvalue of $D_n(\theta^o)$. Then there exists a constant $c_\lambda > 0$ such that

$$\liminf_{n\to\infty} n^{-1}\lambda_{\min}(\theta^o,n) > c_\lambda$$

(A6)

$$\frac{\partial}{\partial\theta_j}E[\varepsilon_t^2(\theta)] = E\left[\frac{\partial}{\partial\theta_j}\varepsilon_t^2(\theta)\right],$$
$$\frac{\partial^2}{\partial\theta_j\partial\theta_k}E[\varepsilon_t^2(\theta)] = E\left[\frac{\partial}{\partial\theta_j\partial\theta_k}\varepsilon_t^2(\theta)\right]$$
$$\frac{\partial^3}{\partial\theta_j\partial\theta_k\partial\theta_l}E[\varepsilon_t^2(\theta)] = E\left[\frac{\partial}{\partial\theta_j\partial\theta_k\partial\theta_l}\varepsilon_t^2(\theta)\right]$$

(A7) For any $\vartheta^o \in \Theta^o$ let

$$Y_t(\vartheta^o) = \int e^{it\lambda} A(e^{-i\lambda}; \vartheta^o) dZ_\varepsilon(\lambda),$$

$$\varepsilon_{t,Y}(\vartheta^o) = \sum_{j=0}^{\infty} b_j(\vartheta^o) X_{t-j}$$

and

$$\dot{\varepsilon}_{t,Y}(\vartheta^o) = \frac{\partial}{\partial\theta} \varepsilon_{t,Y}(\theta)|_{\theta=\vartheta^o}$$

Then

$$n^{-1} \max_{1\le t\le n} \left[\varepsilon_{t,Y}(\vartheta^o)\dot{\varepsilon}_{t,Y}(\vartheta^o)\right]^2 = o_p(1)$$

and

$$\lim_{n\to\infty} E\left\{ \max_{1\le t\le n} \left[\varepsilon_{t,Y}(\vartheta^o)\dot{\varepsilon}_{t,Y}(\vartheta^o)\right]^2 \right\} = 0$$

**Remark 14** *The meaning of assumptions (A1)-(A7) is as follows (see also remarks in section 2, example 1): (A1) is an identifiability condition; (A2), (A4), (A5) and (A6) are standard conditions in the context of maximum likelihood estimation for FARIMA processes; (A3) is needed to carry over asymptotic results obtained under stationarity to the locally stationary case; (A7) is needed for applying a limit theorem for martingale differences (Hall and Heyde 1980).*

## 5.2 Proof of theorem 1

**Step 1 - consistency:** Since data are observed for $t = 1, 2, ..., n$, estimation is based on observations for $t = t_o - [nb], ..., t_o + [nb]$, and $nb^3 \to \infty$, we may replace $L_n(\theta)$ by

$$S_n(\theta) = (nb)^{-1} \sum K(\frac{t - t_o(n)}{nb}) \varepsilon_t^2(\theta).$$

Define the $k \times k$ matrix

$$D_n(\theta) = E[\ddot{S}_n(\theta)],$$

For $\delta \geq 1$, denote by $\partial N_n = \{\theta : (\theta - \theta^o)^T D_n(\theta^o)(\theta - \theta^o) = \delta^2\}$ the border of $N_n(\theta^o, \delta)$, and by $N_n^o = N_n \setminus \partial N_n$ its interior. Then, for $\theta_n \in \partial N_n(\theta^o, \delta)$,

$$P(S_n(\theta_n) > S_n(\theta^o)) = P(\dot{S}_n(\theta^o)(\theta_n - \theta^o) + \frac{1}{2}(\theta_n - \theta^o)^T \ddot{S}(\theta^*)(\theta_n - \theta^o) > 0)$$
$$\geq P((\theta_n - \theta^o)^T \ddot{S}(\theta^*)(\theta_n - \theta^o) > \frac{\delta}{2}) - P(\dot{S}(\theta^o)(\theta_n - \theta^o) \leq -\frac{\delta}{4})$$

where $\theta^* = a\theta_0 + (1-a)\theta_n$ $(0 \leq a \leq 1)$ is a vector between $\theta_0$ and $\theta_n$. Since

$$\sup_{\theta \in N_n} D_n^{-1/2}(\theta) \ddot{S}_n(\theta_n) D_n^{-1/2}(\theta_n) \to I$$

in the norm $\|x\| = \sum_{i,j} |x_{ij}|$, and $\theta_n^* \to \theta^o$, we may approximate the first probability by

$$P((\theta_n - \theta^o)^T D(\theta^o)(\theta_n - \theta^o) > \frac{\delta}{2}) \to 1$$

(see e.g. Fahrmeier and Kaufmann 1985). The second probability converges to zero since $(nb)^{-1/2}\dot{S}(\theta^o)$ converges in distribution to a zero mean normal variable and $\theta_n - \theta^o$ is of the order $(nb)^{-1/2}$. Thus, $\lim P(S_n(\theta_n) > S_n(\theta^o)) = 1$ so that with probability approaching to 1, $S_n(\theta)$ $(\theta \in N_n)$ assumes its minimum in $N_n^o$. (By analogous arguments it follows that the minimum is not attained for $\theta \notin N_n$, with probability converging to one.) Since $N_n^o$ is a shrinking neighborhood of $\theta^o$, consistency follows. Note also that, because of convexity of $S_n(\theta)$ for large $n$, the minimum is unique (unless $S_n$ is constant in an interval) and thus coincides with $\hat{\theta}_n$.

**Step 2 - asymptotic normality:** Without loss of generality we will assume $\sigma_\varepsilon^2 \equiv 1$. By Taylor expansion we have

$$0 = \dot{S}_n(\hat{\theta}) = \dot{S}_n(\theta^o) + \ddot{S}_n(\theta^o)(\hat{\theta} - \theta^o) + \dddot{S}_n(\theta^*)(\hat{\theta} - \theta^o)^2$$

with $\theta^* = (1-a)\theta^o + a\hat{\theta}$ for some $a \in [0,1]$, and hence

$$(nb)^{1/2}(\hat{\theta} - \theta^o) = M^{-1}(\theta^o)(nb)^{-1/2}\dot{S}_n(\theta^o) + o_p((nb)^{-1/2})$$

with $M_{ij} = E[\dot{\varepsilon}(\theta^o)\dot{\varepsilon}^T(\theta^o)]$. Now $(nb)^{-1}\sum K\left\{(t - t_o(n))/nb\right\}\varepsilon_t(\theta^o)\dot{\varepsilon}_t(\theta^o)$ may be approximated by

$$\tilde{S}_n = (nb)^{-1}\sum K(\frac{t - t_o(n)}{nb})\left\{\varepsilon_t(\theta(\frac{t}{n}))\dot{\varepsilon}_t(\theta(\frac{t}{n})) + E\left[\varepsilon_t(\theta^o)\dot{\varepsilon}_t(\theta^o)\right]\right\}$$

Since $\varepsilon_t(\theta(\frac{t}{n}))\dot{\varepsilon}_t(\theta(\frac{t}{n}))$ is a martingale difference, (A7) together with theorem 3.2 in Halle and Heyde (1980) implies that $\sqrt{nb}(\tilde{S}_n - E\left[\varepsilon_t(\theta^o)\dot{\varepsilon}_t(\theta^o)\right])$ converges in distribution to a zero mean normal variable with covariance matrix

$$\begin{aligned} W &= \lim_{n\to\infty} \frac{1}{nb} \sum_{t=t_o-[nb]}^{t_o+[nb]} K^2(\frac{t-t_o}{nb}) E\left[\dot{\varepsilon}_t(\theta(\frac{t}{n}))\dot{\varepsilon}_t^T(\theta(\frac{t}{n}))\right] \\ &= \int_{-1}^{1} K^2(x)dx \cdot J(\theta^o) \end{aligned}$$

with

$$J(\theta^o) = \left[\frac{1}{4\pi}\int_{-\pi}^{\pi} \frac{\partial}{\partial\theta_r}\log f(\lambda;\theta^o)\frac{\partial}{\partial\theta_s}\log f(\lambda;\theta^o)d\lambda\right]_{r,s=1,\ldots,l}$$

Since $M = J^{-1}$, we have

$$\sqrt{nb}(\hat{\theta} - E[\hat{\theta}]) \to_d N(0, J^{-1}(\theta^o)\int K^2(x)dx)$$

## 5.3 Proof of theorem 2:

We will use the notation $\theta_{t,n} = \theta(t/n)$ and $\theta_o = \theta^o(u_o)$. Consider

$$0 = \dot{S}_n(\hat{\theta}) = (nb)^{-1} \sum_{t=t_o-[nb]}^{t_o+[nb]} K(\frac{t-t_o}{nb})\varepsilon_t(\hat{\theta})\dot{\varepsilon}_t(\hat{\theta})$$

$$= (nb)^{-1} \sum_{t=t_o-[nb]}^{t_o+[nb]} K(\frac{t-t_o}{nb})\varepsilon_t(\theta_{t,n})\dot{\varepsilon}_t(\theta_{t,n})$$

$$+ (nb)^{-1} \sum_{t=t_o-[nb]}^{t_o+[nb]} K(\frac{t-t_o}{nb})\{\dot{\varepsilon}_t^2(\theta_{t,n}) + \varepsilon_t(\theta_{t,n})\ddot{\varepsilon}_t(\theta_{t,n})\}(\hat{\theta} - \theta_{t,n}) + o_p\{(nb)^{-1/2}\}$$

$$= (nb)^{-1} \sum_{t=t_o-[nb]}^{t_o+[nb]} K(\frac{t-t_o}{nb})\varepsilon_t(\theta_{t,n})\dot{\varepsilon}_t(\theta_{t,n})$$

$$+ (nb)^{-1} \sum_{t=t_o-[nb]}^{t_o+[nb]} K(\frac{t-t_o}{nb})\{\dot{\varepsilon}_t^2(\theta_{t,n}) + \varepsilon_t(\theta_{t,n})\ddot{\varepsilon}_t(\theta_{t,n})\}(\hat{\theta} - \theta_o)$$

$$+ (nb)^{-1} \sum_{t=t_o-[nb]}^{t_o+[nb]} K(\frac{t-t_o}{nb})\{\dot{\varepsilon}_t^2(\theta_{t,n}) + \varepsilon_t(\theta_{t,n})\ddot{\varepsilon}_t(\theta_{t,n})\}(\theta_o - \theta_{t,n}) + o_p\{(nb)^{-1/2}\}.$$

Now, $E[\varepsilon_t(\theta_{t,n})\dot{\varepsilon}_t(\theta_{t,n})] = E[\varepsilon_{t_o}(\theta_o)\dot{\varepsilon}_{t_o}(\theta_o)] = 0$ and $\varepsilon_t(\theta_{t,n})\dot{\varepsilon}_t(\theta_{t,n})$ is a martingale difference, so that

$$(nb)^{-1} \sum_{t=t_o-[nb]}^{t_o+[nb]} K(\frac{t-t_o}{nb})\varepsilon_t(\theta_{t,n})\dot{\varepsilon}_t(\theta_{t,n}) = (nb)^{-1/2}Z + o_p\{(nb)^{-1/2}\}$$

where $Z$ is a zero mean normal variable. Also, $E[\varepsilon_t(\theta_{t,n})\ddot{\varepsilon}_t(\theta_{t,n})] = 0$ and $\varepsilon_t(\theta_{t,n})\ddot{\varepsilon}_t(\theta_{t,n})$ is a martingale difference so that the same approximation applies to

$$(nb)^{-1} \sum_{t=t_o-[nb]}^{t_o+[nb]} K(\frac{t-t_o}{nb})\varepsilon_t(\theta_{t,n})\ddot{\varepsilon}_t(\theta_{t,n})(\hat{\theta} - \theta_o).$$

Furthermore, $E[\dot{\varepsilon}_t^2(\theta_{t,n})] = E[\dot{\varepsilon}_{t_o}^2(\theta_o)] = J(\theta_o)$ and

$$(nb)^{-1} \sum_{t=t_o-[nb]}^{t_o+[nb]} K(\frac{t-t_o}{nb})\dot{\varepsilon}_t^2(\theta_{t,n})(\hat{\theta} - \theta_o) = J(\theta_o)(\hat{\theta} - \theta_o) + o_p(1)$$

For the other terms we have

$$(nb)^{-1}\sum_{t=t_o-[nb]}^{t_o+[nb]}K(\frac{t-t_o}{nb})\dot{\varepsilon}_t^2(\theta_{t,n})(\theta_o-\theta_{t,n})$$

$$=-(nb)^{-1}\sum_{t=t_o-[nb]}^{t_o+[nb]}K(\frac{t-t_o}{nb})\dot{\varepsilon}_t^2(\theta_{t,n})(\theta^{'}(u_o)\frac{t-t_o}{n}+\frac{1}{2}\theta^{''}(u_o)\left(\frac{t-t_o}{n}\right)^2+...)$$

$$=-\theta^{'}(u_o)J(\theta_o)(nb)^{-1}\sum_{t=t_o-[nb]}^{t_o+[nb]}K(\frac{t-t_o}{nb})\frac{t-t_o}{n}$$

$$-\frac{1}{2}\theta^{''}(u_o)J(\theta_o)(nb)^{-1}\sum_{t=t_o-[nb]}^{t_o+[nb]}K(\frac{t-t_o}{nb})\left(\frac{t-t_o}{n}\right)^2+o_p(b^2)$$

$$=-b^2\frac{1}{2}\theta^{''}(u_o)J(\theta_o)\int_{-1}^{1}K(x)x^2dx+o_p(b^2)$$

and

$$-(nb)^{-1}\sum_{t=t_o-[nb]}^{t_o+[nb]}K(\frac{t-t_o}{nb})\varepsilon_t(\theta_{t,n})\ddot{\varepsilon}_t(\theta_{t,n})(\theta_o-\theta_{t,n})$$

$$=(nb)^{-1}\sum_{t=t_o-[nb]}^{t_o+[nb]}K(\frac{t-t_o}{nb})\varepsilon_t(\theta_{t,n})\ddot{\varepsilon}_t(\theta_{t,n})(\theta^{'}(u_o)\frac{t-t_o}{n}+\frac{1}{2}\theta^{''}(u_o)\left(\frac{t-t_o}{n}\right)^2+...)$$

$$=o_p(b^2)$$

Then

$$J(\theta_o)(\hat{\theta}-\theta_o)-b^2\frac{1}{2}\theta^{''}(u_o)J(\theta_o)\int_{-1}^{1}K(x)x^2dx+o_p(b^2)=0$$

so that

$$\hat{\theta}-\theta_o=b^2\frac{1}{2}\theta^{''}(u_o)\int_{-1}^{1}K(x)x^2dx+o_p(b^2)$$

## 5.4 Proof of corollary 1:

The asymptotic mean squared error is approximated by

$$\begin{aligned}MSE(\hat{\theta})&=Bias^2+Variance\\&=b^4C_1+(nb)^{-1}C_2+o\left\{\max(b^4,(nb)^{-1})\right\}\end{aligned}$$

where

$$C_1 = \left[\frac{1}{2}\theta''(u_o)\int_{-1}^{1} K(x)x^2dx\right]^2$$

and

$$C_2 = J^{-1}\int K^2(x)dx$$

Minimizing w.r.t. $b$ yields

$$b_{opt} = n^{-1/5}C_3$$

with

$$C_3 = \left[\frac{C_2}{4C_1}\right]^{1/5}$$

The resulting MSE is then of the order $MSE(\hat{\theta}) = O(n^{-4/5})$.

**Acknowledgement 1** *I would like to thank the referees for their very useful constructive comments.*

# References

[1] Ben Hariz, S., Wylie, J.J. and Zhang, Q. (2007). Optimal rate of convergence for nonparametric change-point estimators for nonstationary sequences. Ann. Statist., Vol. 35, No. 4, 1802–1826.

[2] Beran, J. (1994). Statistics for long-memory processes. Chapman & Hall, London.

[3] Beran, J. (1995). Maximum likelihood estimation of the differencing parameter for invertible short- and long-memory ARIMA models. J. Roy. Statist. Soc., Series B, 57 (4), 672-695.

[4] Beran, J. (2003). Statistics in musicology. Chapman & Hall (CRC Press), New York.

[5] Beran, J., Bhansali, R.J. and Ocker, D. (1998). On unified model selection for stationary and nonstationary short- and long-memory autoregressive processes. Biometrika, Vol. 85, No. 4, 921-934.

[6] Beran, J. and Feng, Y. (2002a). SEMIFAR models - A semiparametric framework for modelling trends, long-range dependence and nonstationarity. Computat. Statist. Data Anal., 40, 393-419.

[7] Beran, J. and Feng, Y. (2002b). Iterative plug-in algorithms for SEMIFAR models - definition, convergence and asymptotic properties. J. Computat. Graph. Statist., 11, 690-713.

[8] Beran, J., Sherman, R., Taqqu, M.S., and Willinger, W. (1995). Long-range dependence in variable-bit-rate video traffic. Communications, IEEE Transactions on Communications. Vo. 43, Issue 234, 1566-1579.

[9] Beran, J., and Terrin, N. (1994). Estimation of the long-memory parameter,based on a multivariate central limit theorem. J. Time Series Analysis, 15, 269-284.

[10] Berkes, I., Horváth, L., Kokoszka, P. and Shao. Q. (2006). On discriminating between long-range dependence and changes in mean.Ann. Statist., Vol. 34, No. 3, 1140-1165.

[11] Brockmann, M. (1993). Locally adaptive bandwidth choice for kernel regression estimators. J. Amer. Statist. Assoc., 88, 1302-1309.

[12] Csörgö, S., Mielniczuk, J. (1995). Nonparametric regression under long-range dependent normal errors. Ann. Statist. 23, 1000–1014.

[13] Dahlhaus, R. (1989). Efficient parameter estimation for self-similar processes. Ann. Statist., 17, 1749-1766.

[14] Dahlhaus, R. (1996). On the Kullback-Leibler information divergence of locally stationary processes. Stoch. Proc. Appl., 62, 139-162.

[15] Dahlhaus, R. (1997). Fitting time series models to nonstationary processes. Annals of Statistics, 25, 1-37.

[16] Dahlhaus, R. and Giraitis, L. (1998). On the optimal segment length for estimates for locally stationary time series. J. Time Ser. Anal., 19, 629–636

[17] Diebold, F.X. and Inoue, A. (2001). Long memory and regime switching. J. Econometrics, 105, 131-159.

[18] Embrechts, P., Klüppelberg, C. and Mikosch, T. (1997). Modelling Extremal Events, Springer, New York.

[19] Fahrmeir, L. and Kaufmann, H. (1985). Consistency and asymptotic normality of the maximum likelihood estimator in generalized linear models. The Annals of Statistics, Vol. 13, 342-368.

[20] Falconer, K. and Fernandez, C. (2007). Inference on fractal processes using multiresolution approximation. Biometrika, Vol. 94, No. 2, 313-334.

[21] Fox, R. and Taqqu, M. S. (1986). Large sample properties of parameter estimates for strongly dependent stationary Gaussian time series. Ann. Statist., 14, 517-532.

[22] Gasser, T., Kneip, A., and Köhler, W. (1991). A flexible and fast method for automatic smoothing. J. Amer. Statist. Assoc., 86, 643-652.

[23] Giraitis, L. and Leipus, R. (1990). A function CLT for nonparametric estimates of spectra and change-point problem for spectral function. Lietuvos Matematikos Rinkinys, 30, 674-497.

[24] Giraitis, L. and Leipus, R. (1992). Testing and estimating in the change point problem of the spectral function. A function CLT for nonparametric estimates of spectra and change-point problem for spectral function. Lietuvos Matematikos Rinkinys, 32, 20-38.

[25] Giraitis, L., Surgailis, D. (1990). A central limit theorem for quadratic forms in strongly dependent linear variables and its application to the asymptotic normality of Whittle's estimate. Probab. Th. Rel. Fields, 86, 87-104.

[26] Granger, C.W. and Ding, Z. (1996). Varieties of long memory models. J. Econometrics, Vol. 73, No. 1, 61-77.

[27] Granger, C.W. and Hyung, N. (2004). Occasional structural breaks and long memory with an application to the S&P 500 absolute stock returns. J. Emp. Finance, Vol. 11, No. 3, 399-421

[28] Granger, C.W.J., Joyeux, R. (1980). An introduction to long-range time series models and fractional differencing. J. Time Series Anal. 1, 15-30.

[29] Hall, P. (1992). On bootstrap confidence intervals in nonparametric regression.Ann. Statist., Vol. 20, No. 2, 695-711.

[30] Hall, P. and Hart, J. (1990). Nonparametric regression with long-range dependence. Stoch. Proc. Appl., 36, 339-351.

[31] Hallin, M. (1978). Mixed autoregressive-moving average multivariate processes with time-dependent coefficients. J. Multivariate Anal., 8, 567–572.

[32] Härdle, W. and Marron, J.S. (1991). Bootstrap simultaneous error bars for nonparametric regression. Ann. Statist., 19, 778-796.

[33] Herrmann, E., Gasser, T. and Kneip, A. (1992). Choice of bandwidth for kernel regression when residuals are correlated. Biometrika, 79, 783-795.

[34] Horváth, L. and Shao, Qi-Man (1999). Limit theorems for quadratic forms with applications to Whittle's estimate.

[35] Horváth, L. and Kokoszka, P. (1997). The effect of long-range dependence on changepoint estimators. J. Statist. Plann. Inf., Vol. 64, No. 1, 57-81.

[36] Hosking, J.R.M. (1981). Fractional differencing. Biometrika, 68, 165-176.

[37] Hüsler, J., Piterbarg, V. and Seleznjev, O. (2003). On convergence of the uniform norms for Gaussian processes and linear approximation problems. The Annals of Applied Probability, 13(4), 1615–53.

[38] Jensen, M.J. and Whitcher, B. (2000). Time-varying long memory in volatiliy: detection and estimation with wavelets. Technical Report, EURANDOM.

[39] Kokoszka and Leipus (2003). Detection and estimation of changes in regime. In: Long-Range Dependence, Paul Doukhan, Murad S. Taqqu, Georges Oppenheim (eds.). Birkhäuser, Basel, pp. 325-337.

[40] Kuan, C.M., Hsu, C.C. (1998). Change-Point Estimation of Fractionally Integrated Processes. J. Time Series Analysis, Vol. 19, No. 6, 693-708.

[41] Lahiri, S.N. (2003). Resampling Methods for Dependent Data. Springer, New York.

[42] Lavielle, M. and Ludena, C. (2000). The multiple change-points problem for the spectral distribution. Bernoulli, Vol. 6, No.5, 845-869.

[43] Lowen, S.B. and Teich, M.C. (2005). Fractal-based point processes. Wiley, New York.

[44] Mandelbrot, B.B. (1977). Fractals. Form, chance and dimension. Freeman, San Francisco.

[45] Moulines, E., Priouret, P. and Roueff, F. (2005). On recursive estimation for time varying autoregressive processes. Ann. Statist. 33, no. 6 (2005), 2610–2654

[46] Priestley, M.B. (1981). Spectral Analysis and Time Series. Academic Press, London.

[47] Ray, B.K. and Tsay, R.S. (1997). Bandwidth selection for kernel regression with long-range dependence. Biometrika, 84, 791-802.

[48] Ray, B.K., and Tsay, R.S. (2002). Bayesian methods for change-point detection in long-range dependent processes. Journal of Time Series Analysis, 23, No. 6, 687–705.

[49] Sibbertsen, P. (2004). Long memory versus structural breaks: an overview. Statistical Papers, 45, 465-515.

[50] Stoev, S., Pipiras, V. and Taqqu, M.S. (2002). Estimation of the self-similarity parameter in linear fractional stable motion. Signal Processing, Vol. 82, Issue 12, 1873-1901.

[51] Stoev, S. and Taqqu, M.S. (2005). Asymptotic self-similarity and wavelet estimation for long-range dependent fractional autoregressive integrated moving average time series with stable innovations. J. Time Series Analysis, Vol. 26, No. 2, 211–249.

[52] Subba Rao, T. (1970). The fitting of non-stationary time series models with time-dependent parameters. J. Roy. Stat. Soc., Series B, 32, 312 - 322.

[53] Tousson, O. (1925). Mémoire sur l'histoire du Nil. In: Mémoires de l'Institut d'Egypte, Vol. 18, pp. 366-404.

[54] Tribouley, K. (2004). Adaptive simultaneous confidence intervals in non-parametric estimation. Statistics & Probability Letters, Vol. 69, Issue 1, 37-51.

[55] van Bellegen, S. and Dahlhaus, R. (2006). Semiparametric estimation by model selection for locally stationary processes. Journal Roy. Statist. Soc. B, 68, 721-764.

[56] Vesilo, R.A. and Chan, A. (1996). Detecting change points in long range dependency traffic. In: Proceedings of the Australian Telecommunication Networks & Applications Conference, Melbourne, 3-6 December 1996, pp. 567 - 572.

[57] Whitcher, B., Byers, S. D., Guttorp, P. and Percival, D.B. (2002). Testing for Homogeneity of Variance in Time Series: Long Memory, Wavelets and the Nile River. Water Resources Research, 38, no. 5, 1000-1029.

[58] Whitcher, B. and Jensen, M.J. (2000). Wavelet estimation of a local long memory parameter. Exploration Geophysics, 31, 94-103.

[59] Wright, J.H. (1998). Testing for a structural break at unknown date with long-mmory disturbances. J. Time Seies Analysis, 19, 369-376.

[60] Yajima, Y. (1985). On estimation of long-memory time series models. Austral. J. Statist., 27, 303-320.