Finite Sample Performance in Cointegration Analysis of Nonlinear Time Series with Long Memory

Afonso Gonçalves da Silva* and Peter M. Robinson†
Department of Economics, London School of Economics and Political Science,
Houghton Street, London WC2A 2AE, UK

September 14, 2005

Abstract

Nonlinear functions of multivariate financial time series can exhibit long memory and fractional cointegration. However, tools for analysing these phenomena have principally been justified under assumptions that are invalid in this setting. Determination of asymptotic theory under more plausible assumptions can be complicated and lengthy. We discuss these issues and present a Monte Carlo study, showing that asymptotic theory should not necessarily be expected to provide a good approximation to finite-sample performance.

JEL Classification: C32
Keywords: Fractional cointegration; memory estimation; stochastic volatility.

*Research supported by FCT grant SFRH/BD/4783/2001 and ESRC grant R000239936.
†Research supported by ESRC grant R000239936. Corresponding author. Tel.: +44 (0)20 7955 7516; fax: +44 (0)20 7955 6592; e-mail: p.m.robinson@lse.ac.uk.
1 Introduction

Fractional cointegration analysis is increasingly found to be a promising tool for dimensionality reduction in financial time series. On the one hand, series of asset returns may have little autocorrelation, whereas instantaneous nonlinear functions, such as squares, can exhibit evidence of long memory. Considering series on several assets, it is possible that there exists a linear combination of the nonlinear functions that has shorter memory. Then there is said to be fractional cointegration. Note that here, as implied by many stochastic volatility (SV) models, series are supposed to be stationary. By contrast, in analysing macroeconomic time series, levels are typically believed to be nonstationary with a unit root, and cointegration exists when there is a linear combination that is stationary (with short memory).

A variety of tools for analysing fractional cointegration in stationary series is becoming available. The main stress has been on “semiparametric” methods. These avoid full parameterisation of autocorrelation, in favour of a local power law for the spectral density around zero frequency. Estimates of memory parameters can be rendered inconsistent by misspecification of short memory properties. Moreover, when the cointegrating relation is expressed in regression form, with one of the observables on the left hand side, the other observables cannot plausibly be assumed orthogonal to the cointegrating errors. Thus (“full-band”) time domain procedures (in a stationary environment) such as least squares will inconsistently estimate the cointegrating vector. This leads to a focus on methods based on a vanishing neighbourhood of zero in the frequency domain, such that the number, $m$, of Fourier frequencies used increases with sample size $n$, but more slowly. An undesirable consequence of this semiparametric strategy is rates of convergence (in case of both memory parameters and cointegrating vector estimates) that are slower than would be possible in a fully parametric setting. However, parametric estimates of memory parameters and (due to the stationarity) cointegrating vectors can only converge at rate $n^{\frac{1}{2}}$ (there is no super-consistency), and the slower rates of the semiparametric methods (depending on $m$) may be acceptable when $n$ is very large indeed, as is the case with many financial time series.

Asymptotic theory for the semiparametric estimates has been developed mainly under assumptions that are unfortunately implausible in this setting. Usually series have been assumed to be generated by linear filters of conditionally homoscedastic martingale differences. This is justified if, for example, series are Gaussian. Recall, however, that
the long memory property, and the possibility of fractional cointegration, emerges only for certain nonlinear functions. Typically, as in the case of squares, these cannot have linear-in-martingale-difference representations. Models for them can be articulated, in terms of underlying independent and identically distributed (iid) sequences, say, but the nonlinearity makes derivation of asymptotic properties (already a delicate matter in the linear setting) extremely complicated and lengthy. Moreover, due to second order bias that affects some estimates, useful limit distribution theory is unavailable. As a result, relevant asymptotic theory is not well developed, in view of which Monte Carlo simulation here plays a rather larger role than the usual one of investigating relevance of asymptotic theory in finite samples.

In the following section we consider the modelling of cointegration of series that are generated by SV models. Section 3 discusses methods of estimating cointegrating coefficients (the stress being on relatively simple “single equation” methods), and also memory parameters. Section 4 presents Monte Carlo simulations.

2 Long memory, cointegration and stochastic volatility

Consider first a covariance stationary scalar process \( z_t, t = 0, \pm 1, \ldots \), having spectral density \( f_z(\lambda), \lambda \in (-\pi, \pi] \). We say that \( z_t \) is a (fractionally integrated) \( I(d) \) process, for \( d \in (-\frac{1}{2}, \frac{1}{2}) \), if

\[
f_z(\lambda) \sim C|\lambda|^{-2d}, \quad \text{as } \lambda \to 0,
\]

for some \( C \in (0, \infty) \), “\( \sim \)” meaning that the ratio of left- and right-hand sides tends to 1. We call \( d \) the “memory parameter” of \( z_t \). An \( I(0) \) process is said to have short memory, an \( I(d) \) process for \( d < 0 \) is said to have negative memory, and an \( I(d) \) process for \( d > 0 \) is said to have long memory. We will focus on cases \( d \geq 0 \).

Now consider a \( p \times 1 \) column vector \( Z_t = (z_{1t}, \ldots, z_{pt})' \), such that \( z_{it} \) is \( I(d_i), d_i \in [0, \frac{1}{2}), i = 1, \ldots, p > 1 \). In general it is supposed that there is cross-correlation between the \( z_{it} \) but it is not necessary at present to discuss the nature of this, except to note that by the Schwarz inequality the cross-spectral density at frequency \( \lambda \) between \( z_{it} \) and \( z_{jt} \) has modulus of order no greater than \( |\lambda|^{-d_i-d_j} \) as \( \lambda \to 0 \). Now suppose that there exists an unknown nonzero \( p \times 1 \) vector \( \alpha \) (the “cointegrating vector”) such that the unobservable process \( u_t = \alpha'Z_t \) is an \( I(d_u) \) process, for \( d_u < \min_i d_i \). Then \( Z_t \) is said
to be fractionally cointegrated. Notice that if \( p = 2 \) a necessary condition for fractional cointegration is that \( d_1 = d_2 \). Alternative definitions of fractional cointegration are reviewed by Robinson and Yajima (2002), who also discuss the possibility of existence of two or more cointegrating relations, and methods for estimating the number of these.

It is desirable to reconcile these properties of long memory and fractional cointegration with a more fundamental modelling of \( Z_t \), which is plausible in financial series. Consider a jointly strictly stationary \( q \times 1 \) vector process \( \eta_t \), for \( q \geq p \), such that

\[
\eta_t = g_i(\eta_t), \quad i = 1, \ldots, p, \tag{2.2}
\]

where the \( g_i \) are nonlinear functions. As analysed in this kind of general setting by Robinson (2001), if at least one element of \( \eta_t \) has long memory, then, for given \( i \), \( \eta_{it} \) may have long memory, though the existence of long memory, and the actual value of \( d_i \), depends on the nature of \( g_i \) as well as memory parameters of elements of \( \eta_t \). In view of the nonlinearity, theoretical analysis is greatly facilitated if \( \eta_t \) is Gaussian but it is not necessary to stress this possibility here.

It may be possible, further, to infer the cointegrating relation for \( Z_t \) from an underlying structural relation for \( \eta_t \). We consider perhaps the simplest case. We take \( p = 2 \), \( q = 4 \), write \( \eta_t = (\eta_{1t}, \eta_{2t}, \eta_{3t}, \eta_{4t})' \), and assume it is Gaussian. Suppose that the \( \{\eta_{it}\} \) are mutually independent processes, that for \( i = 1, 2, 3 \) the \( \eta_{it} \) are iid with zero mean and variance \( \sigma_i^2 \), and that \( \eta_{4t} \) is an \( I(d_4) \) process, for \( d_4 > 0 \). Suppose that we observe sequences \( x_t, y_t \), generated by

\[
x_t = \theta_1 \zeta_t + \eta_{1t}, \tag{2.3}
\]
\[
y_t = \theta_2 \zeta_t + \eta_{2t}, \tag{2.4}
\]

where \( \theta_1, \theta_2 \neq 0 \) and

\[
\zeta_t = \eta_{3t} h(\eta_{4t}), \tag{2.5}
\]

where \( h \) is a possibly nonlinear function, with \( E\{h(\eta_{4t})^2\} < \infty \). Now \( \zeta_t \) is not an iid sequence but it is a square-integrable martingale difference, and thus uncorrelated, sequence, as therefore are \( x_t \) and \( y_t \). Thus \( x_t \) and \( y_t \) exhibit an ideal property of asset returns, say. Because \( x_t \) and \( y_t \) are therefore \( I(0) \) sequences, and all linear combinations of them are also \( I(0) \), they are not cointegrated. However, we can deduce a cointegrating
relation between the squares $z_{1t} = x_t^2$, $z_{2t} = y_t^2$. We have

\[ z_{2t} = (\theta_2 \zeta_t + \eta_{2t})^2 \]
\[ = \beta z_{1t} + u_t, \tag{2.6} \]

where $\beta = \theta_2^2 / \theta_1^2$ and

\[ u_t = \eta_{2t}^2 + 2\theta_2 \eta_{2t} \zeta_t - 2\beta \theta_1 \eta_{1t} \zeta_t - \beta \eta_{1t}^2. \tag{2.7} \]

Clearly $u_t$ has no autocorrelation, and is thus an $I(0)$ process. We have

\[ z_{1t} = \theta_1^2 \zeta_t^2 + 2\theta_1 \eta_{1t} \zeta_t + \eta_{1t}^2. \tag{2.8} \]

The last two terms on the right are also $I(0)$. However for suitable $h$, the leading term $\theta_1^2 \zeta_t^2$ has long memory, and thence so has $z_{1t}$. For example if $\zeta_t = \eta_{3t} \eta_{4t}$, $z_{1t}$ is $I(2d_4 - \frac{1}{2})$, or if $\zeta_t = \eta_{3t} \eta_{4t}$, $z_{1t}$ is $I(d_4)$. In either case, $z_{2t}$ has the same memory parameter as $z_{1t}$, and $Z_t = (z_{1t}, z_{2t})'$ is fractionally cointegrated, with cointegrating vector $\alpha = (-\beta, 1)'$. A similar conclusion is drawn if, even more simply, $\eta_{1t}$ is missing from (2.3). Notice that $\zeta_t$ is generated by a SV model and plays the role of a common factor. Fractional cointegration can also arise if $\eta_{1t}$ and/or $\eta_{2t}$ are replaced by processes with SV (so that $u_t$ can have long memory), as shown by Gonçalves da Silva and Robinson (2005).

Though (2.6) is expressed in the form of a regression model, it does not possess the classical properties. The unobservable sequence $u_t$ actually has nonzero mean (as does $z_{1t}$), but this situation is rectified by introducing an intercept. More important, however, $u_t$ is not orthogonal to the right hand side observable $z_{1t}$:

\[ \text{Cov}(z_{1t}, u_t) = -2\beta \sigma_1^2 \left\{ \sigma_1^2 + 2E(\zeta_t^2) \right\} < 0, \tag{2.9} \]

taking $\theta_1 = 1$ with no loss of generality. For general $p$, after rewriting $\alpha' Z_t = u_t$ in regression form, then even in the absence of an underlying structure like (2.3), (2.4) there is no reason to suppose that orthogonality between cointegrating errors and right-hand side regressors obtains, especially as the designation of left-hand variable is arbitrary.
3 Estimation of cointegrating vector and memory parameters

Assuming the $p$-th element of $\alpha$ is non-zero, adopting an arbitrary normalization, and designating $z_{pt}$ as left-hand side variable, we rewrite the cointegrating relation $\alpha'Z_t = u_t$ as

$$Y_t = \beta'X_t + u_t,$$

(3.1)

where $Y_t = z_{pt}$, $X_t = (z_{1t}, \ldots, z_{p-1,t})'$ and $\beta$ is a $(p-1) \times 1$ vector. It is desired to estimate the unknown $\beta = (\beta_1, \ldots, \beta_{p-1})'$, on the basis of observables $Z_t, t = 1, \ldots, n$.

The most obvious estimate of $\beta$ is ordinary least squares (OLS) with intercept correction (bearing in mind that $u_t$ may have non-zero mean, as the discussion of the previous section suggests). This is

$$\hat{\beta}_O = \left\{ \sum_{t=1}^{n} (X_t - \bar{X})X_t' \right\}^{-1} \sum_{t=1}^{n} (X_t - \bar{X})Y_t,$$

(3.2)

where $\bar{X} = n^{-1} \sum_{t=1}^{n} X_t$. However, the correlation envisaged between $u_t$ and $X_t$ makes $\hat{\beta}_O$ inconsistent for $\beta$, bearing in mind also the stationarity of $Z_t$; this outcome differs from the familiar one in which $Z_t$ has a unit root and $u_t$ is $I(0)$, where the asymptotic dominance of sums of squares of $u_t$ by those of $X_t$ overwhelms the simultaneous equation bias, leading to $n$-consistency of $\hat{\beta}_O$.

A consistent estimate of $\beta$ was proposed by Robinson (1994). For a vector sequence $a_t$, define the discrete Fourier transform

$$w_a(\lambda) = (2\pi n)^{-\frac{1}{2}} \sum_{t=1}^{n} a_t e^{it\lambda},$$

(3.3)

and for a vector sequence $b_t$, possibly the same as $a_t$, define the (cross-) periodogram matrix

$$I_{ab}(\lambda) = w_a(\lambda)w_b'(\lambda).$$

(3.4)

For a sequence $m = m(n)$ such that

$$m \leq \frac{n}{2},$$

(3.5)

$$\frac{1}{m} + \frac{m}{n} \to 0, \quad \text{as } n \to \infty,$$

(3.6)
define the narrow-band least squares estimate of $\beta$,

$$
\hat{\beta}_{NB} = \left( \sum_{j=1}^{m} \text{Re} \{ I_{XX}(\lambda_j) \} \right)^{-1} \sum_{j=1}^{m} \text{Re} \{ I_{XY}(\lambda_j) \},
$$

(3.7)

where $\lambda_j = 2\pi j/n$ and $\text{Re}(\cdot)$ is the real part operator. Note that omission of the frequency $\lambda_0 = 0$ corresponds to a sample mean correction like that in (3.2), while if in contrast to (3.5), (3.6), $m = n - 1$ we have $\hat{\beta}_{NB} = \hat{\beta}_O$. However, the condition (3.6) is crucial to the consistency of $\hat{\beta}_{NB}$. The basic intuition for the consistency is as follows. By the Cauchy inequality, for $i = 1, \ldots, p - 1$,

$$
\left| \sum_{j=1}^{m} \text{Re} \{ I_{z,u}(\lambda_j) \} \right| \leq \left\{ \sum_{j=1}^{m} I_{z,z_i}(\lambda_j) \sum_{j=1}^{m} I_{u,u}(\lambda_j) \right\}^{\frac{1}{2}},
$$

(3.8)

and under suitable conditions this is

$$
O_p \left( n \left\{ \int_0^{\lambda_m} \lambda^{-2d_i} d\lambda \int_0^{\lambda_m} \lambda^{-2d_u} d\lambda \right\}^{\frac{1}{2}} \right) = O_p \left( n \left( \frac{m}{n} \right)^{1-d_i-d_u} \right).
$$

(3.9)

On the other hand, under suitable conditions, for $\Lambda_m = \text{diag} \{ \lambda_{m1}^{d_1}, \ldots, \lambda_{mp}^{d_{p-1}} \}$,

$$
\frac{1}{m} \Lambda_m \sum_{j=1}^{m} \text{Re} \{ I_{XX}(\lambda_j) \} \Lambda_m \rightarrow_p \Omega,
$$

(3.10)

where $\Omega$ is a constant positive definite matrix. It follows that

$$
\hat{\beta}_{NB,i} - \beta_i = O_p \left( \left( \frac{m}{n} \right)^{d_i-d_u} \right), \quad i = 1, \ldots, p - 1,
$$

(3.11)

where $\beta_i$ and $\hat{\beta}_{NB,i}$ are the $i$-th elements of $\beta$ and $\hat{\beta}_{NB}$. Since cointegration entails $d_u < d_i$, $i = 1, \ldots, p - 1$, $\hat{\beta}_{NB}$ is thus consistent for $\beta$. The key is the domination, near zero, of the spectral density of $u_t$ by the spectral densities of $z_{1t}, \ldots, z_{p-1,t}$.

Consistency of $\hat{\beta}_{NB}$ was first shown by Robinson (1994) in case $p = 2$, and then, with the rate in (3.11), by Robinson and Marinucci (2003) for general $p$. The conditions they imposed to deduce the crucial properties (3.9) and (3.10) were that $Z_t$ is generated by a linear moving average in conditionally homoscedastic martingale differences. As previously noted, this is inconsistent with our SV setup, such as illustrated in the previous
section, albeit similar to one for log squared returns for a certain SV model (see e.g. Deo and Hurvich, 2001) and a multiplicative set-up in place of the additive one, typified in (2.3), (2.4). However, Gonçalves da Silva and Robinson (2005) have established (3.11) for \( p = 2 \) under a somewhat more general set-up than that described in connection with (2.3) and (2.4). The proof is exceedingly lengthy, however, and things do not generalise immediately to the case \( p > 2 \).

The estimate \( \hat{\beta}_{NB} \) is desirably computationally simple and has been applied in fractional cointegration analyses of implied and realised volatility by Christensen and Nielsen (2004), Bandi and Perron (2004).

In general the rate in (3.11) is sharp, and indeed under additional conditions it seems that, for each \( i \), \((n/m)^{d_i-d_u}(\hat{\beta}_{NB} - \beta_i)\) converges in distribution not to a non-degenerate random variable, but to a constant. This is due to the presumed coherence between \( X_t \) and \( u_t \) around zero frequency. Without such coherence, asymptotic normality and a faster rate of convergence is possible. Christensen and Nielsen (2004) supposed that the cross-spectral density between \( z_{1t} \) and \( u_t \) is \( o(|\lambda|^{-d_i-d_u}) \), as \( \lambda \rightarrow 0 \), rather than having real part behaving precisely like \( |\lambda|^{-d_i-d_u} \). Assuming also that \( d_i + d_u < \frac{1}{2} \), \( i = 1, \ldots, p - 1 \), they deduced that \( m^{\frac{1}{2}}(m/n)^{d_u} \Lambda_m^{-1}(\hat{\beta}_{NB} - \beta) \) is asymptotically multivariate normal; they assumed \( Z_t \) is linear in homoscedastic martingale differences, as in Robinson (1994), Robinson and Marinucci (2003).

Though the model constructed in Section 2, (2.6) based on (2.3)-(2.5) and \( z_{1t} = X_t = x_t^2 \), \( z_{2t} = Y_t = y_t^2 \), does not satisfy the linearity assumption of Christensen and Nielsen (2004), it does satisfy a lack-of-coherence assumption that corresponds to theirs. It is easily seen that \( \text{Cov}(z_s, u_t) = 0 \) if \( s \neq t \), so in view of (2.8), the cross-spectral density of \( z_{1t}, u_t \) is finite and constant, and \( o(|\lambda|^{-\delta}) \), where \( \delta > 0 \) represents the memory parameter of \( z_{1t} \). (In the cases discussed after (2.8), the possibilities that \( \delta = d_i \) and \( \delta = 2d_4 - \frac{1}{2} \) emerged.)

Violation of orthogonality represents an important way in which (3.1) disobeys classical regression conditions, but it is not the only one. Though the simple set-up with \( p = 2 \) analysed in the previous section ensured that \( u_t \) has no autocorrelation (see (2.7)), more generally \( u_t \) can be not only autocorrelated but even have long memory, as indicated by Gonçalves da Silva and Robinson (2005). In the absence of simultaneous equations bias, a suitable weighted frequency domain estimate will be more efficient. In (3.1) with short memory \( u_t \) orthogonal to \( X_t \), Hannan (1963) showed that weighting inversely with respect to a nonparametric estimate of \( f_u \) can achieve the same asymptotic efficiency
as generalised least squares based on a correctly specified parametric model for \( f_u \). Hidalgo and Robinson (2002) extended this finding to long memory \( u_t \), with unknown \( d_u \). However, the “full-band” estimates will incur similar simultaneous equations bias to \( \hat{\beta}_O \). Nevertheless, it is worth considering whether some such weighting can improve on \( \hat{\beta}_{NB} \), since \( f_u \) changes even over the interval \([\lambda_1, \lambda_m]\). Smith and Chen (1996) proposed the weighted narrow-band estimate

\[
\hat{\beta}_{W NB} = \tilde{\beta}(\tilde{d}_u),
\]

where

\[
\tilde{\beta}(d) = \left( \sum_{j=1}^{m} \lambda_j^{2d} \text{Re} \{ I_{XX}(\lambda_j) \} \right)^{-1} \sum_{j=1}^{m} \lambda_j^{2d} \text{Re} \{ I_{XY}(\lambda_j) \},
\]

and \( \tilde{d}_u \) is a consistent estimate of \( d_u \) (see below). Note that \( \tilde{\beta}(0) = \hat{\beta}_{NB} \). Smith and Chen (1996) in fact proposed \( \hat{\beta}_{W NB} \) in a more traditional regression setting, with \( u_t \) orthogonal to \( X_t \), and did not establish any asymptotic properties. Recently, Nielsen (2005), under the same kind of incoherence-near-zero assumption as Christensen and Nielsen (2004), established that for given \( d \), which satisfies a suitable constraint relative to \( d_u \) and the \( d_i, m^2 (m/n)^d_u \Lambda_{m}^{-1} (\tilde{\beta}(d) - \beta) \) is asymptotically normal. Nielsen (2005) also discussed the relative efficiency of \( \tilde{\beta}(d) \) and \( \hat{\beta}_{NB} \), noting some circumstances in which \( \tilde{\beta}(d) \) can be the more efficient even when \( d \neq d_u \).

However \( d_u \) is clearly an optimal choice of \( d \), and given that \( d_u \) is unknown it is natural to focus on \( \hat{\beta}_{W NB} \) which, like \( \hat{\beta}_{NB} \), should still be consistent in the presence of coherence between \( u_t \) and \( X_t \), violating Nielsen’s (2005) condition. We have, say,

\[
\left| \sum_{j=1}^{m} \lambda_j^{2\tilde{d}_u} \text{Re} \{ I_{zu}(\lambda_j) \} \right| \leq \left\{ \sum_{j=1}^{m} \lambda_j^{2d_u} I_{z\tilde{u}}(\lambda_j) \right\}^{\frac{1}{2}} \left( \sum_{j=1}^{m} \lambda_j^{2d_u} I_{uu}(\lambda_j) \right)^{\frac{1}{2}}.
\]\n
We assume (as in Robinson, 1994)

\[
(\log n)(\tilde{d}_u - d_u) \rightarrow_p 0, \quad \text{as } n \rightarrow \infty,
\]

which can readily be justified in view of asymptotic theory for various memory parameter estimates. Then

\[
\lambda_j^{2d_u} = \lambda_j^{2d_u} \lambda_j^{2(d_u-d_u)} \leq \lambda_j^{2d_u} n^{o((\log n)^{-1})} \leq \lambda_j^{2d_u} e^{o(1)} \leq 2\lambda_j^{2d_u},
\]
for \( n \) sufficiently large. It is then readily seen that (3.14) is

\[
O_p \left( n \left\{ \int_0^{\lambda m} \lambda^{2(d_u-d_i)} d\lambda \frac{m}{n} \right\}^{\frac{1}{2}} \right) = O_p \left( n \left( \frac{m}{n} \right)^{1+d_u-d_i} \right). \tag{3.17}
\]

Also under (3.15), and similar conditions to those giving (3.10), we can justify the step

\[
\frac{1}{m} \left( \frac{m}{n} \right)^{-2d_u} \Lambda_m \sum_{j=1}^{m} \left( \lambda_j^{2d_u} - \lambda_j^{2d_u} \right) \text{Re} \{ I_{XX}(\lambda_j) \} \Lambda_m \to_p 0, \tag{3.18}
\]

and then that

\[
\frac{1}{m} \left( \frac{m}{n} \right)^{-2d_u} \Lambda_m \sum_{j=1}^{m} \lambda_j^{2d_u} \text{Re} \{ I_{XX}(\lambda_j) \} \Lambda_m \to_p 0, \tag{3.19}
\]

converges in probability to a constant positive definite matrix.

Notice that in the model (2.6) derived from (2.3)-(2.5), \( d_u = 0 \) so we expect no improvement of \( \hat{\beta}_{WNB} \) over \( \hat{\beta}_{NB} \). However, we can extend this model to allow at the same time \( d_u > 0 \), and incoherence at frequency zero between regressors and errors.

Bias and autocorrelation can be corrected simultaneously by more elaborate methods, based on a full system of \( p \) equations that expresses also the long memory properties of the \( z_{it} \), \( i = 1, \ldots, p-1 \). This leads to estimates of \( \beta \) which depend not only on \( \hat{d}_u \), but also on estimates of the \( d_i \), \( i = 1, \ldots, p-1 \). Such estimates are developed by Hualde and Robinson (2004); they are asymptotically normal with the same rate as described for \( \hat{\beta}_{NB} \) and \( \hat{\beta}_{WNB} \) under the incoherence-near-zero assumption, but without imposing that. We focus in our numerical study in the following section only on the “single-equation” estimates (based on (3.1)) we have discussed above, this is partly due to their computational simplicity, but also because incoherence-near-zero can often be justified in a factor model context, as discussed above, whence \( \hat{\beta}_{NB} \) and \( \hat{\beta}_{WNB} \) enjoy a reasonably fast rate of convergence.

Even if simple estimates of \( \beta \) are used, there may be interest in estimation of the \( d_i \), as well as in estimation of \( d_u \), as is required for \( \hat{\beta}_{WNB} \). In particular, such estimates are useful in determining the existence and extent of cointegration, as described by Robinson and Yajima (2002). In this multivariate setting, efficiency gains are possible by estimating memory parameters jointly, especially if prior equality constraints are placed on the \( d_i \). However, joint estimates have principally been developed under the assumption of no cointegration, and if there is cointegration they are liable to be inconsistent. Thus we
describe some leading “semiparametric” estimates. We introduce a generic univariate stationary process \( v_t \) which can represent any of the \( z_t \), or, where estimation of \( d_u \) is concerned, residuals \( y_t - \tilde{\beta}' X_t \), such that \( \tilde{\beta} \) represents one of our consistent estimates of \( \beta \).

Denote by \( d \) the unknown memory parameter of \( v_t \). Geweke and Porter-Hudak (1983) proposed a log-periodogram estimate, a simplified version of which, due to Robinson (1995a), is

\[
\hat{d}_{LP} = \left\{ \sum_{j=1}^{m} \left( \ln j - \frac{1}{m} \sum_{i=1}^{m} \ln i \right)^2 \right\}^{-1} \sum_{j=1}^{m} \left( \ln j - \frac{1}{m} \sum_{i=1}^{m} \ln i \right) \ln I_{vv}(\lambda_j). \tag{3.20}
\]

Assuming that \( m \) satisfies at least (3.5) and (3.6), Robinson (1995a), Hurvich, Deo, and Brodsky (1998) showed that

\[
m^{\frac{1}{2}}(\hat{d}_{LP} - d) \rightarrow_d N(0, \pi^2/6), \quad \text{as } n \rightarrow \infty. \tag{3.21}
\]

An efficiency improvement is possible, for the same \( m \) sequence, via the local Whittle estimate (Künsch, 1987),

\[
\hat{d}_{LW} = \arg \min_{d \in D} \left\{ \ln \left( \sum_{j=1}^{m} j^{-2d} I_{vv}(\lambda_j) \right) - d \frac{1}{m} \sum_{j=1}^{m} \ln j \right\}, \tag{3.22}
\]

where \( D \) is a compact subset of \(( -\frac{1}{2}, \frac{1}{2} ) \). This was shown by Robinson (1995b) to satisfy

\[
m^{\frac{1}{2}}(\hat{d}_{LW} - d) \rightarrow_d N(0, \frac{1}{4}), \quad \text{as } n \rightarrow \infty. \tag{3.23}
\]

Note that the conditions imposed to deduce (3.21) and (3.23) do not cover the SV setup described in the previous section, but see e.g. Deo and Hurvich (2001). Various modifications, in particular bias corrections, have been introduced. Hurvich, Moulines, and Soulier (2005) allow for a more refined approximation to \( f_v(\lambda) \) than \( C \lambda^{-2d} \), there is something of a signal-plus-noise character to the model (2.3), (2.4), so we might consider the estimate

\[
(\hat{d}_{MLW}, \hat{\theta}) = \arg \min_{(d, \theta) \in D \times \Theta} \left\{ \ln \left( \sum_{j=1}^{m} I_{vv}(\lambda_j) \right) - \theta \frac{1}{m} \sum_{j=1}^{m} \ln(j^{-2d} + \theta) \right\}, \tag{3.24}
\]
where $\Theta$ is a compact subset of the positive real line. Hurvich, Moulines, and Soulier (2005) justify asymptotic normality of $d_{MLW}$, but with a different asymptotic variance from that in (3.23).

4 Simulations

We now present a Monte Carlo study of finite-sample performance. For linear processes, Robinson and Marinucci (2003) reported simulation experiments of NBLS with $I(1)$ observables and $I(0)$ cointegrating errors, while Marinucci and Robinson (2001) explored different cases of fractional cointegration with nonstationary observables and stationary errors. Bandi and Perron (2004) examined NBLS for the regression between realised and implied volatility, generating the data from a discretised continuous time SV model. Gonçalves da Silva and Robinson (2005) reported experiments of NBLS in a SV framework similar to ours. We present results for two settings, one linear, and the other generalising (2.2)-(2.6). Under the linear model, we generate (see (2.3), (2.4))

$$z_1t = \zeta_t + \delta_t,$$

$$z_2t = \beta \zeta_t + \varepsilon_t,$$

where we use the abbreviated notation $\zeta_t = \eta_{1t}$, $\delta_t = \eta_{2t}$, $\varepsilon_t = \eta_{3t}$, and for $i = 1, 2, 3$, $\{\eta_{it}\}$ is a zero mean Gaussian ARFIMA$(0, d_i, 0)$ process with variance $\sigma_i^2$. In the nonlinear case, we use (see (2.3)-(2.6))

$$z_1t = (\zeta_t + \delta_t)^2,$$

$$z_2t = (\theta_2 \zeta_t + \varepsilon_t)^2,$$

where $\zeta_t = \xi_{1t} h(\eta_{1t})$, $\delta_t = \xi_{2t} h(\eta_{2t})$, $\varepsilon_t = \xi_{3t} h(\eta_{3t})$, and for $i = 1, 2, 3$, $\{\eta_{it}\}$ is an independent standard Gaussian sequence, and $\{\eta_{it}\}$ a zero mean Gaussian ARFIMA$(0, d_i, 0)$ with variance $\sigma_i^2$. In both models, the basic processes $\{\xi_{it}\}$ and $\{\eta_{it}\}$, $i = 1, 2, 3$, are all generated independently of each other, and we will denote the variances of $\zeta_t$, $\delta_t$, $\varepsilon_t$ by $\sigma_{\zeta}^2$, $\sigma_{\delta}^2$, $\sigma_{\varepsilon}^2$ respectively.

Under each model, we employ 1,000 replications of series of length $n = 2048$ and estimate $\beta$ by narrow-band regressions of $z_2t$ on $z_1t$, where $\beta = \theta_2^2$ in the nonlinear
setting. Note that both models can be written as (2.6), with

\[ u_t = \varepsilon_t - \beta \delta_t \]  \hspace{1cm} (4.5)

in the linear setting and

\[ u_t = \varepsilon_t^2 - \beta \delta_t^2 + 2\zeta_t (\theta_2 \varepsilon_t - \beta \delta_t) \]  \hspace{1cm} (4.6)

in the nonlinear setting. We present bias, standard deviation (SD) and root mean squared error (RMSE) of \( \tilde{\beta}(d) \) given by (3.13), for various values of \( d \), both fixed and estimated. All are evaluated at the bandwidth, \( m^* \), that minimises RMSE.

**Asymptotic theory**

We first examine the performance of Nielsen’s (2005) asymptotic theory under the linear model, when \( \delta_t \) is absent in (4.1). We set \( \beta = 1 \), \( d_1 = 0.4 \), \( d_3 = 0.2 \), \( \sigma^2_{\varepsilon} = 4 \) and \( \sigma^2_{\xi} = 2 \). This simulation is comparable to his model A, although we focus on full-band estimates, i.e. \( m = n/2 \). (Given the independence between \( u_t \) and \( z_{1t} \), this choice dominates any other value of \( m \).) Table 1 reports asymptotic (Asy.) and Monte Carlo (MC) SD for different values of \( d \). Monte Carlo bias is negligible in this setting and therefore omitted. Note that Nielsen’s (2005) theory requires

\[ (2d_1 + 2d_3 - 1)/4 < d \leq d_3, \]  \hspace{1cm} (4.7)

which in this case is equivalent to \( 0.05 < d \leq 0.2 \), but we compute his asymptotic SD also for \( d > 0.2 \). Here we find that Monte Carlo SD is almost always over 30% larger than the asymptotic one, so the asymptotic theory is not a good approximation even when \( n = 2048 \). The inclusion of measurement error (ME), while still compatible with Nielsen’s (2005) assumptions, would increase the Monte Carlo SD even further without changing the asymptotic value (as long as \( d_2 < d_3 \)), thereby making the gap much larger. Figure 1 plots the theoretical and Monte Carlo SD of \( \tilde{\beta}(d) \) relative to that of \( \tilde{\beta}(d_3) \), for different values of \( d \). Although the asymptotic and Monte Carlo levels in Table 1 substantially differ, their ratios across \( d \) are comparable, and \( d = d_3 \) is the optimal choice in both.
<table>
<thead>
<tr>
<th>$d$</th>
<th>Asy. SD</th>
<th>MC SD</th>
<th>Ratio</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.10</td>
<td>0.0176</td>
<td>0.0213</td>
<td>1.211</td>
</tr>
<tr>
<td>0.15</td>
<td>0.0155</td>
<td>0.0203</td>
<td>1.307</td>
</tr>
<tr>
<td>0.20</td>
<td>0.0152</td>
<td>0.0201</td>
<td>1.323</td>
</tr>
<tr>
<td>0.25</td>
<td>0.0154</td>
<td>0.0204</td>
<td>1.328</td>
</tr>
<tr>
<td>0.30</td>
<td>0.0157</td>
<td>0.0209</td>
<td>1.332</td>
</tr>
<tr>
<td>0.35</td>
<td>0.0161</td>
<td>0.0216</td>
<td>1.337</td>
</tr>
<tr>
<td>0.40</td>
<td>0.0166</td>
<td>0.0223</td>
<td>1.341</td>
</tr>
<tr>
<td>0.45</td>
<td>0.0171</td>
<td>0.0230</td>
<td>1.344</td>
</tr>
</tbody>
</table>

Table 1: Asymptotic and Monte Carlo SD of WNBLS, for varying $d$; linear setting with $\delta_t$ absent.

**Variation in measurement error**

We present results for different types of ME, namely: no ME, i.e. $\delta_t$ absent in (4.1) or (4.3); antipersistent ME ($d_2 = -0.2$); iid ME ($d_2 = 0$); and long memory ME ($d_2 = 0.2$). In the nonlinear model, the antipersistent case would still generate $I(0)$ ME in (4.3) and is therefore omitted. In both settings we use $\beta = 1$, $d_1 = 0.4$, $d_3 = 0.2$, $\sigma_\xi^2 = 4$, $\sigma_\delta^2 = 2$, and $h(x) = \exp(x)$ as the volatility function for the nonlinear setting.

Table 2 reports Monte Carlo optimal bandwidth, bias and RMSE, under the linear setting, for various regression estimates of $\beta$: unweighted NBLS, $\tilde{\beta}(0)$; the theoretically optimal but infeasible weighted estimate, $\tilde{\beta}(d_3)$; and feasible versions of it, $\tilde{\beta}(\hat{d}_3)$, where $\hat{d}_3$ is a consistent estimate of $d_3$. In these cases, $d_3$ is estimated using LP (3.20), LW (3.22), or MLW (3.24) based on the regression residuals from a first step unweighted NBLS regression; the same $m$ is used in the first and second steps. Due to the modified spectral approximation in (3.24), when using MLW we compute WNBLS as

$$
\tilde{\beta}(\hat{d}, \hat{\theta}) = \left( \sum_{j=1}^{m} \frac{\text{Re} \{I_{XX}(\lambda_j)\}}{j^{-2\hat{d} + \hat{\theta}}} \right)^{-1} \sum_{j=1}^{m} \frac{\text{Re} \{I_{XY}(\lambda_j)\}}{j^{-2\hat{d} + \hat{\theta}}},
$$

instead of (3.13). Table 3 reports bias and RMSE for these preliminary estimates of $d_3$.

In the model without ME all regression estimates have, as expected, virtually no bias and perform best in the full-band case. Here, $\tilde{\beta}(d_3)$ clearly exhibits an efficiency gain over $\tilde{\beta}(0)$, which is equivalent to OLS (3.2). However, as progressively more persistent ME is introduced, both estimates have increasing bias, and the RMSE of $\tilde{\beta}(d_3)$ grows much faster than that of $\tilde{\beta}(0)$. Indeed, in the presence of ME, simple NBLS always
Figure 1: Asymptotic and Monte Carlo relative SD of $\tilde{\beta}(d)$ versus $\tilde{\beta}(d_3)$, for varying $d$; linear setting with $\delta_t$ absent.

outperforms the weighted estimate. Here and throughout all experiments, estimates are biased towards zero, due to the negative correlation between $z_{1t}$ and $u_t$ caused by ME. The feasible versions of WNBLS seem to closely match the infeasible one in both RMSE and bias, in many cases even appearing slightly better. This behaviour arises because whenever ME is present, the optimal weighting is actually obtained for $d < d_3$, so the negative bias of LP and LW, seen in Table 3, can actually work to their advantage. Although MLW actually displays positive bias, the weights in (4.8) do not depend on $\hat{d}_2$ alone but also on $\hat{\theta}$ in (3.24), allowing it to still outperform the infeasible estimate for $d_2 = -0.2$. The optimal bandwidths for each estimate are lower the more persistent the ME is, since frequencies closer to zero become more contaminated with the correlation between $z_{1t}$ and $u_t$.

Table 3 shows that both LP and LW perform relatively well throughout. The small biases are insufficient for the bias reduction properties of MLW to make a difference; in fact, this estimate displays larger bias than LP and LW in three of the four cases. As expected, the much lower SD of LW makes it the best in RMSE. Although some of the bias can be attributed to estimation error, most of it surely comes from the “signal-plus-
Tables 4 and 5 present results for the nonlinear setting. Here it can be seen that the weighted estimate is always outperformed by NBLS, with ME causing much more significant bias. Even in the absence of ME, the optimal bandwidth is slightly below the full-band case, possibly as a consequence of $u_t$ being orthogonal to but not independent of $z_{1t}$, as can be seen by setting $d = 0$ in (4.6). All feasible weighted estimates outperform the infeasible one, which can again be explained by the negative biases found in Table 5. Biases are stronger here than in the linear setting, partly because of the estimation error and the nonlinear setting, but also because of the signal-plus-noise structure. Note that in this setting the $I(0)$ noise in (4.6) does not vanish even if $d$ is absent. For both LP and LW, bias is the main component of RMSE. Therefore, the bias reduction provided by MLW allows it to dominate the other estimates in the presence of ME. Again, the inferior performance of the weighted estimate relative to simple NBLS demonstrates that $d = d_3$ is not the optimal choice in this setting.

Naturally, if $d_3$ is no longer the optimal choice for $d$, the usefulness of estimating it from the data can be questioned. This is verified in Figures 2 and 3, which show...
Table 4: Monte Carlo bias and RMSE of regression estimates, for different types of measurement error; nonlinear setting.

<table>
<thead>
<tr>
<th>( \hat{\beta} )</th>
<th>( d_1 ) absent</th>
<th>( d_2 = 0 )</th>
<th>( d_2 = 0.2 )</th>
</tr>
</thead>
<tbody>
<tr>
<td>( \beta )</td>
<td>( m^* ) Bias RMSE</td>
<td>( m^* ) Bias RMSE</td>
<td>( m^* ) Bias RMSE</td>
</tr>
<tr>
<td>NBLS</td>
<td>973 -0.0042 0.0840</td>
<td>8 -0.1495 0.2717</td>
<td>8 -0.1944 0.3210</td>
</tr>
<tr>
<td>True ( d_3 )</td>
<td>973 -0.0042 0.0855</td>
<td>8 -0.1589 0.2829</td>
<td>8 -0.2020 0.3290</td>
</tr>
<tr>
<td>LP</td>
<td>973 -0.0042 0.0840</td>
<td>8 -0.1516 0.2753</td>
<td>8 -0.1969 0.3243</td>
</tr>
<tr>
<td>LW</td>
<td>973 -0.0043 0.0842</td>
<td>8 -0.1514 0.2747</td>
<td>8 -0.1966 0.3237</td>
</tr>
<tr>
<td>MLW</td>
<td>973 -0.0043 0.0847</td>
<td>8 -0.1532 0.2776</td>
<td>8 -0.1987 0.3266</td>
</tr>
</tbody>
</table>

Table 5: Monte Carlo bias and RMSE of residual memory estimates, for different types of measurement error; nonlinear setting.

<table>
<thead>
<tr>
<th>( \hat{d_3} )</th>
<th>( m )</th>
<th>( \delta_1 ) absent</th>
<th>( d_2 = 0 )</th>
<th>( d_2 = 0.2 )</th>
</tr>
</thead>
<tbody>
<tr>
<td>( \hat{d_3} )</td>
<td>( m ) Bias RMSE</td>
<td>( d_2 = 0 ) Bias RMSE</td>
<td>( d_2 = 0.2 ) Bias RMSE</td>
<td></td>
</tr>
<tr>
<td>LP</td>
<td>80    -0.1512 0.1704</td>
<td>-0.1827 0.2016</td>
<td>-0.1640 0.1861</td>
<td></td>
</tr>
<tr>
<td>LW</td>
<td>80    -0.1562 0.1690</td>
<td>-0.1847 0.1969</td>
<td>-0.1666 0.1810</td>
<td></td>
</tr>
<tr>
<td>MLW</td>
<td>200   -0.0411 0.1827</td>
<td>-0.0986 0.1840</td>
<td>-0.0697 0.1788</td>
<td></td>
</tr>
</tbody>
</table>

The RMSE of \( \hat{\beta}(d) \) relative to that of \( \hat{\beta}(d_3) \), for different values of \( d \), in the linear and nonlinear settings. Only in the linear case without ME is \( d = d_3 \) optimal; in all other cases, the optimal value is smaller, and it is reduced the more persistent the ME is. In the nonlinear case the optimal values for \( d \) are always negative, and in a region excluded by (4.7). It should also be noted that, in the absence of information on the optimal \( d \), NBLS should be chosen over \( \hat{\beta}(d_3) \) (or its feasible versions). Tables 6 and 7 report optimal bandwidth, bias and RMSE for \( \hat{\beta}(d) \), with \( d = 0, 0.2 \) and the values of \( d \) that minimise RMSE in each case (indicated in bold-face), in the linear and nonlinear settings. The degradation in performance with more persistent ME can still be seen here, and bias is often slightly smaller for the optimal \( d \). However, the variation in bias across \( d \) is relatively small, and most of the variation in RMSE can be explained by variations in SD.

The minimization of RMSE at values different from \( d = d_3 \) is surprising since it does not conform to the asymptotic theory. A frequency domain generalised least squares approach will weigh the contribution of each frequency by the inverse of their approximate SD, thereby “whitening” the observations. A possible explanation for the discrepancy lies in the approximation error in (2.1), which in the limit theory is made irrelevant by assuming enough smoothness in the spectral density, but can play a major role in finite
Figure 2: Relative RMSE of $\hat{\beta}(d)$ versus $\hat{\beta}(d_3)$, for varying $d$ and $d_2$; linear setting.

Figure 3: Relative RMSE of $\hat{\beta}(d)$ versus $\hat{\beta}(d_3)$, for varying $d$ and $d_2$; nonlinear setting.
Table 6: Monte Carlo bias and RMSE of \( \hat{\beta}(d) \), for varying \( d \) and different types of measurement error; linear setting. The minimum RMSE choice of \( d \) is indicated in bold-face.

<table>
<thead>
<tr>
<th>( d )</th>
<th>( \delta_t ) absent</th>
<th>( d_2 = -0.2 )</th>
<th>( d_2 = 0 )</th>
<th>( d_2 = 0.2 )</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>( m^* ) Bias RMSE</td>
<td>( m^* ) Bias RMSE</td>
<td>( m^* ) Bias RMSE</td>
<td>( m^* ) Bias RMSE</td>
</tr>
<tr>
<td>-0.05</td>
<td>1024 -0.0005 0.0331</td>
<td>67 -0.0303 0.0663</td>
<td>33 -0.0519 0.0902</td>
<td>12 <strong>-0.1269</strong> 0.1750</td>
</tr>
<tr>
<td>0.00</td>
<td>1024 -0.0005 0.0273</td>
<td>55 -0.0279 0.0642</td>
<td>25 -0.0470 0.0897</td>
<td>12 -0.1301 0.1752</td>
</tr>
<tr>
<td>0.05</td>
<td>1024 -0.0004 0.0235</td>
<td><strong>52 -0.0294 0.0632</strong></td>
<td><strong>25 -0.0502 0.0896</strong></td>
<td>11 -0.1298 0.1757</td>
</tr>
<tr>
<td>0.20</td>
<td><strong>1024 -0.0001 0.0201</strong></td>
<td>39 -0.0283 0.0652</td>
<td>23 -0.0555 0.0933</td>
<td>10 -0.1326 0.1789</td>
</tr>
</tbody>
</table>

Table 7: Monte Carlo bias and RMSE of \( \hat{\beta}(d) \), for varying \( d \) and different types of measurement error; nonlinear setting. The minimum RMSE choice of \( d \) is indicated in bold-face.

<table>
<thead>
<tr>
<th>( d )</th>
<th>( \delta_t ) absent</th>
<th>( d_2 = 0 )</th>
<th>( d_2 = 0.2 )</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>( m^* ) Bias RMSE</td>
<td>( m^* ) Bias RMSE</td>
<td>( m^* ) Bias RMSE</td>
</tr>
<tr>
<td>-0.30</td>
<td>1022 -0.0025 0.0946</td>
<td>14 -0.1476 0.2674</td>
<td>14 <strong>-0.1901</strong> 0.3148</td>
</tr>
<tr>
<td>-0.20</td>
<td>976 -0.0033 0.0845</td>
<td><strong>8 -0.1386 0.2666</strong></td>
<td>14 -0.1963 0.3154</td>
</tr>
<tr>
<td>-0.10</td>
<td><strong>973 -0.0039 0.0830</strong></td>
<td>8 -0.1441 0.2681</td>
<td>14 -0.2022 0.3178</td>
</tr>
<tr>
<td>0.00</td>
<td>973 -0.0042 0.0840</td>
<td>8 -0.1495 0.2717</td>
<td>8 -0.1944 0.3210</td>
</tr>
<tr>
<td>0.20</td>
<td>973 -0.0042 0.0855</td>
<td>8 -0.1589 0.2829</td>
<td>8 -0.2020 0.3290</td>
</tr>
</tbody>
</table>

samples. The whitening approach will give low weight to the frequencies closer to zero, where variance is higher but (2.1) is a more accurate approximation, and will boost the impact of more distant frequencies where the approximation is not so accurate. Another relevant factor is the coherence between \( z_{1t} \) and \( u_t \), here generated by \( \delta_t \), which is the leading source of bias. Being of smaller order than the spectral pole, it will be irrelevant asymptotically, but this also means higher frequencies are more contaminated than lower ones. Again, decreasing the weight of the lowest frequencies is likely to worsen the estimation. Both these factors lead to an optimal \( d \) that will tend to be lower than \( d_2 \); in some circumstances they can outweigh the heteroskedasticity in the periodogram, and the optimal \( d \) will be negative, as can be seen in Figures 2 and 3 and Tables 6 and 7.
Table 8: Monte Carlo bias and RMSE of $\tilde{\beta}(d)$, for varying $d$ and $n$; linear setting. The minimum RMSE choice of $d$ is indicated in bold-face.

<table>
<thead>
<tr>
<th>$n$</th>
<th>$d$</th>
<th>512</th>
<th></th>
<th>2048</th>
<th></th>
<th>8192</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>$m^*$</td>
<td>Bias</td>
<td>RMSE</td>
<td>$m^*$</td>
<td>Bias</td>
<td>RMSE</td>
</tr>
<tr>
<td></td>
<td>0.00</td>
<td>15</td>
<td>-0.0950</td>
<td>0.1555</td>
<td>25</td>
<td>-0.0470</td>
<td>0.0897</td>
</tr>
<tr>
<td></td>
<td>0.05</td>
<td>13</td>
<td>-0.0965</td>
<td>0.1560</td>
<td>25</td>
<td>-0.0502</td>
<td>0.0896</td>
</tr>
<tr>
<td></td>
<td>0.20</td>
<td>12</td>
<td>-0.0977</td>
<td>0.1601</td>
<td>23</td>
<td>-0.0555</td>
<td>0.0933</td>
</tr>
</tbody>
</table>

Table 9: Monte Carlo bias and RMSE of $\tilde{\beta}(d)$, for varying $d$ and $n$; nonlinear setting. The minimum RMSE choice of $d$ is indicated in bold-face.

<table>
<thead>
<tr>
<th>$n$</th>
<th>$d$</th>
<th>512</th>
<th></th>
<th>2048</th>
<th></th>
<th>8192</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>$m^*$</td>
<td>Bias</td>
<td>RMSE</td>
<td>$m^*$</td>
<td>Bias</td>
<td>RMSE</td>
</tr>
<tr>
<td></td>
<td>-0.30</td>
<td>15</td>
<td>-0.3060</td>
<td>0.4520</td>
<td>14</td>
<td>-0.1476</td>
<td>0.2674</td>
</tr>
<tr>
<td></td>
<td>-0.20</td>
<td>15</td>
<td>-0.3145</td>
<td>0.4528</td>
<td>8</td>
<td>-0.1386</td>
<td>0.2666</td>
</tr>
<tr>
<td></td>
<td>-0.05</td>
<td>15</td>
<td>-0.3261</td>
<td>0.4572</td>
<td>8</td>
<td>-0.1468</td>
<td>0.2697</td>
</tr>
<tr>
<td></td>
<td>0.00</td>
<td>15</td>
<td>-0.3296</td>
<td>0.4591</td>
<td>8</td>
<td>-0.1495</td>
<td>0.2717</td>
</tr>
<tr>
<td></td>
<td>0.20</td>
<td>8</td>
<td>-0.3096</td>
<td>0.4669</td>
<td>8</td>
<td>-0.1589</td>
<td>0.2829</td>
</tr>
</tbody>
</table>

**Variation in sample size**

Failure of asymptotic theory to provide a good approximation in finite samples is further explored by changing the sample size. Figures 4 and 5 and Tables 8 and 9 present similar results to Figures 2 and 3 and Tables 6 and 7, for $n = 512, 2048, 8192$. We set $\beta = 1, d_1 = 0.4, d_2 = 0, d_3 = 0.2, \sigma^2_z = 4, \sigma^2_s = \sigma^2_\delta = 2$, and use $h(x) = \exp(x)$ as the volatility function for the nonlinear setting. In both the linear and nonlinear settings, the optimal value for $d$ increases with $n$, but not dramatically. Even for $n = 8192$, the optimal $d$ is not only below $d_3$, but also outside the parameter range in (4.7). For all values of $d$, there is a strong improvement in both bias and RMSE as $n$ increases. While in the linear case the optimal bandwidth for each $d$ increases with $n$, in the nonlinear setting it is often higher for $n = 512$ than for $n = 8192$. Bandwidths for $n = 2048$ are the lowest of the three sample sizes, suggesting a “U-shaped” bandwidth profile that will continue diverging to infinity as the theory requires.

Figures 6 and 7 illustrate the distributional properties of NBLS by plotting kernel density estimates for varying $n$, under the linear and nonlinear setting. Density estimates
Figure 4: Relative RMSE of $\tilde{\beta}(d)$ versus $\tilde{\beta}(d_3)$, for varying $d$ and $n$; linear setting.

Figure 5: Relative RMSE of $\tilde{\beta}(d)$ versus $\tilde{\beta}(d_3)$, for varying $d$ and $n$; nonlinear setting.
are computed for a sequence of \( s = 50,000 \) NBLS estimates \( b_i, i = 1, \ldots, s \), using

\[
\hat{f}(b) = \frac{1}{sh} \sum_{i=1}^{s} \phi \left( \frac{b_i - b}{h} \right),
\]

where \( \phi(\cdot) \) is the standard Gaussian density function and the bandwidth \( h \) is chosen using (3.31) of Silverman (1986),

\[
h = 0.9s^{-1/5} \min(\text{SD}, IQR/1.34),
\]

where \( SD \) and \( IQR \) are the sample standard deviation and interquartile range of the \( b_i \). Estimates for other values of \( d \) yield very similar shapes and are thus omitted, but available from the authors upon request. However, unlike for other values of \( d \), NBLS is not covered by (4.7). Still, in the linear case all curves in Figure 6 seem to be fairly close in shape to that of a normal density. On the contrary, densities in Figure 7 are all highly skewed to the left, even for \( n = 8192 \), suggesting that the asymptotic distribution under the nonlinear setting might not be normal. In both settings, bias and SD seem to be decaying at the same rate, which is natural given our minimum RMSE bandwidth choice.

**Variation in the signal-to-noise ratio**

Figures 8 through 11 and Tables 10 through 13 can be interpreted in the same way as Figures 2 and 3 and Tables 6 and 7, for the linear and nonlinear settings, where we first change the variance of the ME, then the variance of the signal. In both experiments we start with \( \beta = 1, d_1 = 0.4, d_2 = 0, d_3 = 0.2, \sigma^2_\zeta = 4, \sigma^2_x = \sigma^2_\delta = 2 \), and \( h(x) = \exp(x) \) as the volatility function for the nonlinear setting. The variance of the ME in the first experiment is then set to \( \sigma^2_\zeta = 1/2, 2, 8 \), by varying \( \sigma^2_\zeta \) in the linear setting, and by using \( h_k(x) = k \exp(x) \), with \( k = 1/2, 1, 2 \), as the volatility function for \( \delta_t \), while keeping \( \sigma^2_\zeta \) constant, in the nonlinear setting. The resulting sequences \( \delta_t \) are consequently the same, up to a multiplicative factor, for each value of \( \sigma^2_\zeta \). In the second experiment, the variance of the signal is changed by choosing \( \sigma^2_1 \) so that \( \sigma^2_\zeta = 2, 4, 8 \).

These parameters affect the accuracy of the estimates by influencing the relative variance of \( z_{1t} \) and \( u_t \) in (4.2), which can be interpreted as a signal-to-noise ratio, and the covariance between \( z_{1t} \) and \( u_t \), which can be seen in (4.5) and (4.6) to depend
Figure 6: Kernel density estimates of NBLS for varying $n$; linear setting.

Figure 7: Kernel density estimates of NBLS for varying $n$; nonlinear setting.
Table 10: Monte Carlo bias and RMSE of $\tilde{\beta}(d)$, for varying $d$ and $\sigma_3^2$; linear setting. The minimum RMSE choice of $d$ is indicated in bold-face.

<table>
<thead>
<tr>
<th>$d$</th>
<th>$\sigma_3^2$</th>
<th>$m^*$</th>
<th>Bias</th>
<th>RMSE</th>
<th>$m^*$</th>
<th>Bias</th>
<th>RMSE</th>
<th>$m^*$</th>
<th>Bias</th>
<th>RMSE</th>
</tr>
</thead>
<tbody>
<tr>
<td>-0.10</td>
<td>142</td>
<td>-0.0297</td>
<td>0.0629</td>
<td>39</td>
<td>-0.0524</td>
<td>0.0913</td>
<td>10</td>
<td>-0.0897</td>
<td>0.1516</td>
<td></td>
</tr>
<tr>
<td>0.00</td>
<td>112</td>
<td>-0.0329</td>
<td>0.0584</td>
<td>25</td>
<td>-0.0470</td>
<td>0.0897</td>
<td>10</td>
<td>-0.0969</td>
<td>0.1527</td>
<td></td>
</tr>
<tr>
<td>0.05</td>
<td>82</td>
<td>-0.0301</td>
<td>0.0579</td>
<td>25</td>
<td>-0.0502</td>
<td>0.0896</td>
<td>10</td>
<td>-0.1005</td>
<td>0.1541</td>
<td></td>
</tr>
<tr>
<td>0.10</td>
<td>81</td>
<td>-0.0326</td>
<td>0.0578</td>
<td>25</td>
<td>-0.0533</td>
<td>0.0904</td>
<td>10</td>
<td>-0.1041</td>
<td>0.1560</td>
<td></td>
</tr>
<tr>
<td>0.20</td>
<td>55</td>
<td>-0.0280</td>
<td>0.0596</td>
<td>23</td>
<td>-0.0555</td>
<td>0.0933</td>
<td>7</td>
<td>-0.0893</td>
<td>0.1595</td>
<td></td>
</tr>
</tbody>
</table>

Table 11: Monte Carlo bias and RMSE of $\tilde{\beta}(d)$, for varying $d$ and $\sigma_3^2$; nonlinear setting. The minimum RMSE choice of $d$ is indicated in bold-face.

<table>
<thead>
<tr>
<th>$d$</th>
<th>$\sigma_3^2$</th>
<th>$m^*$</th>
<th>Bias</th>
<th>RMSE</th>
<th>$m^*$</th>
<th>Bias</th>
<th>RMSE</th>
<th>$m^*$</th>
<th>Bias</th>
<th>RMSE</th>
</tr>
</thead>
<tbody>
<tr>
<td>-0.40</td>
<td>1022</td>
<td>-0.0477</td>
<td>0.1333</td>
<td>18</td>
<td>-0.1454</td>
<td>0.2682</td>
<td>4</td>
<td>-0.4607</td>
<td>0.5858</td>
<td></td>
</tr>
<tr>
<td>-0.20</td>
<td>98</td>
<td>-0.0444</td>
<td>0.1233</td>
<td>8</td>
<td>-0.1386</td>
<td>0.2666</td>
<td>4</td>
<td>-0.4703</td>
<td>0.5880</td>
<td></td>
</tr>
<tr>
<td>-0.10</td>
<td>68</td>
<td>-0.0447</td>
<td>0.1216</td>
<td>8</td>
<td>-0.1441</td>
<td>0.2681</td>
<td>4</td>
<td>-0.4751</td>
<td>0.5903</td>
<td></td>
</tr>
<tr>
<td>0.00</td>
<td>66</td>
<td>-0.0491</td>
<td>0.1228</td>
<td>8</td>
<td>-0.1495</td>
<td>0.2717</td>
<td>4</td>
<td>-0.4798</td>
<td>0.5933</td>
<td></td>
</tr>
<tr>
<td>0.20</td>
<td>63</td>
<td>-0.0555</td>
<td>0.1308</td>
<td>8</td>
<td>-0.1589</td>
<td>0.2829</td>
<td>4</td>
<td>-0.4889</td>
<td>0.6007</td>
<td></td>
</tr>
</tbody>
</table>

Figures 8 and 9 and Tables 10 and 11 show that both $m^*$ and the optimal $d$ decrease rather heavily as $\sigma_3^2$ increases, especially in the nonlinear setting. For large values of $\sigma_3^2$, the common component in $z_{1t}$ and $u_t$ becomes very important, influencing even frequencies relatively close to zero. As a result, both the bandwidth and the weights should adjust so that only the lowest frequencies (where the spectral pole still dominates) have significant influence. Tables 10 and 11 display a strong degradation in both bias and RMSE, caused by the increased coherence between regressor and residuals.

While increasing $\sigma_3^2$ influences both $u_t$ and $z_{1t}$, scaling up the common component in both, increasing the cointegrating parameter $\beta$ boosts the weight of the common component in $u_t$ alone, keeping $z_{1t}$ constant. Still, this provokes a comparable increase in correlation, causing very similar effects to those reported for $\sigma_3^2$. Monte Carlo results for this case are omitted but available upon request.

Figures 10 and 11 and Tables 12 and 13 display the effect of the strength of the signal $\zeta_t$. In the linear case, this scales up the signal in $z_{1t}$ without affecting $u_t$. In the nonlinear case, both are affected, but since the SV model used generates heavily
Figure 8: Relative RMSE of $\tilde{\beta}(d)$ versus $\tilde{\beta}(d_3)$, for varying $d$ and $\sigma_\delta^2$; linear setting.

Figure 9: Relative RMSE of $\tilde{\beta}(d)$ versus $\tilde{\beta}(d_3)$, for varying $d$ and $\sigma_\delta^2$; nonlinear setting.
Table 12: Monte Carlo bias and RMSE of $\tilde{\beta}(d)$, for varying $d$ and $\sigma^2_{\zeta}$; linear setting. The minimum RMSE choice of $d$ is indicated in bold-face.

<table>
<thead>
<tr>
<th>$d$</th>
<th>$m^*$</th>
<th>Bias</th>
<th>RMSE</th>
<th>2</th>
<th>$m^*$</th>
<th>Bias</th>
<th>RMSE</th>
<th>4</th>
<th>$m^*$</th>
<th>Bias</th>
<th>RMSE</th>
<th>8</th>
</tr>
</thead>
<tbody>
<tr>
<td>-0.05</td>
<td>25</td>
<td>-0.0843</td>
<td>0.1380</td>
<td>33</td>
<td>-0.0519</td>
<td>0.0902</td>
<td>42</td>
<td>-0.0310</td>
<td>0.0585</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>0.00</td>
<td>23</td>
<td>-0.0852</td>
<td>0.1382</td>
<td>25</td>
<td>-0.0470</td>
<td>0.0897</td>
<td>40</td>
<td>-0.0327</td>
<td>0.0579</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>0.05</td>
<td>22</td>
<td>-0.0883</td>
<td>0.1392</td>
<td>25</td>
<td>-0.0502</td>
<td>0.0896</td>
<td>39</td>
<td>-0.0346</td>
<td>0.0580</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>0.20</td>
<td>19</td>
<td>-0.0946</td>
<td>0.1455</td>
<td>23</td>
<td>-0.0555</td>
<td>0.0933</td>
<td>25</td>
<td>-0.0305</td>
<td>0.0603</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Table 13: Monte Carlo bias and RMSE of $\tilde{\beta}(d)$, for varying $d$ and $\sigma^2_{\zeta}$; nonlinear setting. The minimum RMSE choice of $d$ is indicated in bold-face.

<table>
<thead>
<tr>
<th>$d$</th>
<th>$m^*$</th>
<th>Bias</th>
<th>RMSE</th>
<th>2</th>
<th>$m^*$</th>
<th>Bias</th>
<th>RMSE</th>
<th>4</th>
<th>$m^*$</th>
<th>Bias</th>
<th>RMSE</th>
<th>8</th>
</tr>
</thead>
<tbody>
<tr>
<td>-0.30</td>
<td>8</td>
<td>-0.3350</td>
<td>0.4722</td>
<td>14</td>
<td>-0.1476</td>
<td>0.2674</td>
<td>18</td>
<td>-0.0702</td>
<td>0.1696</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>-0.20</td>
<td>8</td>
<td>-0.3453</td>
<td>0.4733</td>
<td>8</td>
<td>-0.1386</td>
<td>0.2666</td>
<td>8</td>
<td>-0.0628</td>
<td>0.1695</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>-0.15</td>
<td>8</td>
<td>-0.3504</td>
<td>0.4750</td>
<td>8</td>
<td>-0.1414</td>
<td>0.2671</td>
<td>8</td>
<td>-0.0642</td>
<td>0.1691</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>0.00</td>
<td>8</td>
<td>-0.3649</td>
<td>0.4833</td>
<td>8</td>
<td>-0.1495</td>
<td>0.2717</td>
<td>8</td>
<td>-0.0680</td>
<td>0.1705</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>0.20</td>
<td>7</td>
<td>-0.3714</td>
<td>0.4980</td>
<td>8</td>
<td>-0.1589</td>
<td>0.2829</td>
<td>8</td>
<td>-0.0726</td>
<td>0.1764</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Leptokurtic processes (implying that the variance of $\zeta_t^2$ is the major contribution to the variance of $z_{1t}$) and $\zeta_t$ only affects $u_t$ through a white noise component (thus having a bounded contribution to the spectrum around the zero frequency), the impact on $u_t$ will be minimal compared to that on $z_{1t}$. In both models, increasing $\sigma^2_{\zeta}$ will have the double effect of increasing the variance of $z_{1t}$, thereby making the observables more correlated at all frequencies, and scaling up the spectral pole caused by the memory in $\eta_{1t}$, improving the local signal-to-noise ratio. While both effects will have a clearly positive influence on the accuracy of the estimates, as seen in Tables 12 and 13, the effect on $m^*$ and on the optimal $d$ is not clear, as even frequencies distant from zero become less contaminated by the dependence between $z_{1t}$ and $u_t$. As a result, Figures 10 and 11 show very little variation on relative RMSE with $\sigma^2_{\zeta}$.

**Distributional properties of residual memory estimates**

While the previous experiments show that estimates of residual memory are not necessarily useful for choosing $d$ in (3.13), they might still be relevant for other purposes, namely to verify if a cointegrating relationship exists at all. The use of the LP and
Figure 10: Relative RMSE of $\hat{\beta}(d)$ versus $\tilde{\beta}(d_3)$, for varying $d$ and $\sigma^2_\zeta$; linear setting.

Figure 11: Relative RMSE of $\hat{\beta}(d)$ versus $\tilde{\beta}(d_3)$, for varying $d$ and $\sigma^2_\zeta$; nonlinear setting.
LW estimates is well established by now, and their finite-sample properties have been examined in various settings (see e.g. Robinson and Henry, 1999; Nielsen and Frederiksen, 2005). In finite samples, LW is generally found to have bias of similar magnitude but lower variance than LP, to conform with (3.21) and (3.23). However, the recent MLW estimate has not yet been directly compared to LW. The findings of Hurvich, Moulines, and Soulier (2005), Hurvich and Ray (2003), and Table 5, indicate that, even for moderate sample sizes, MLW can successfully reduce bias in the presence of a “signal-plus-noise” structure, but at the cost of a substantially higher SE than LW. We now present a short comparison of finite-sample distributional properties of LW and MLW in the context of residual memory estimation, for \( n = 512, 2048, 8192 \). Residuals are obtained from \( s = 1,000 \) replications of NBLS regression in the linear and nonlinear settings, with \( \beta = 1, d_1 = 0.4, d_2 = 0, d_3 = 0.2, \sigma_c^2 = 4, \sigma_v^2 = \sigma_b^2 = 2, \) and \( h(x) = \exp(x) \) as the volatility function for the nonlinear setting. The minimum RMSE bandwidths reported in Tables 8 and 9 are used in this step. Then, LW and MLW estimates are constructed from the residuals for a grid of bandwidths (from 10 to \( n/2 \), with increments of 10), allowing us to approximately locate the minimum RMSE bandwidth for each memory estimate. Figures 12 through 15 show kernel density estimates (see (4.9), (4.10)) of LW and MLW, under the linear and nonlinear settings, using the approximately optimal bandwidths given in Table 14.

Table 14 shows that while LW works best with a narrow-band approach, MLW has optimal bandwidth rather close to \( n/2 \). This is possible because, unlike LW, MLW corrects for the presence of iid noise, and thus its spectral approximation is relatively accurate throughout all frequencies considered. However, for higher frequencies to be informative, the absence of short memory dynamics is crucial; the inclusion of, say, ARMA dynamics in any of the \( \{ \eta_{it} \} \) would undoubtedly require MLW bandwidths to be much lower.

All curves in Figures 12 and 13 suggest that the finite-sample density of LW is fairly

<table>
<thead>
<tr>
<th>n</th>
<th>Linear setting</th>
<th>Nonlinear setting</th>
</tr>
</thead>
<tbody>
<tr>
<td>512</td>
<td>140 270 360</td>
<td>240 320 300</td>
</tr>
<tr>
<td>2048</td>
<td>240 940 3720</td>
<td>240 680 4010</td>
</tr>
<tr>
<td>8192</td>
<td>512 2048 8192</td>
<td></td>
</tr>
</tbody>
</table>

Table 14: Approximate minimum RMSE bandwidths of \( \hat{d}_{LW} \) and \( \hat{d}_{MLW} \), for varying \( n \); linear and nonlinear settings.
Figure 12: Kernel density estimates of LW for varying $n$; linear setting.

Figure 13: Kernel density estimates of LW for varying $n$; nonlinear setting.
Figure 14: Kernel density estimates of MLW for varying $n$; linear setting.

Figure 15: Kernel density estimates of MLW for varying $n$; nonlinear setting.
close in shape to that of a normal density, but heavily biased downwards. While in the linear setting both bias and SD are substantially reduced when \( n \) increases, estimation in the nonlinear one seems surprisingly insensitive to sample size; even for \( n = 8192 \) the mean is much closer to 0 than to 0.2. Figures 14 and 15 highlight a potential problem of MLW in finite samples. In several cases, the distribution of MLW is bimodal, with peaks close to 0 and 1/2, the boundaries of the parameter space. In the nonlinear setting, this behaviour is apparent even for \( n = 8192 \), with a small mode close to the true parameter value being barely distinguishable. Performance in the linear setting is more encouraging: for \( n = 8192 \), the “boundary” modes disappear and are replaced by an essentially unbiased unimodal density. Still, it is worth noting that the SD in this case is roughly twice that of LW, and that the tails of the density are still moderately asymmetric. The findings of bimodality and higher SD in MLW are maintained in alternative (unreported) experiment designs, suggesting that they are linked to the additive noise structure itself, not to first step estimation error or nonlinearity. Estimation error in the first step regression actually contaminates the true errors (4.5), (4.6) with a higher memory component (in this case, of memory \( d_1 = 0.4 \)), which should induce a positive contribution to both bias (thereby reducing the LW bias) and SD.

**Concluding remarks**

The results presented indicate that asymptotic theory should not necessarily be expected to provide a good approximation to finite-sample performance.

We first showed that, even in a standard setting, where error and regressor are independent Gaussian processes, Monte Carlo SD deviates substantially from its asymptotic counterpart. While in this setting \( d = d_3 \) is the optimal choice for WNBLS, further results demonstrate that the introduction of nonlinearity or ME makes this choice suboptimal, and indeed dominated by simple NBLS. Furthermore, the nonlinear setting always yields a negative optimal \( d \), even in the absence of ME. Although optimal bandwidths somewhat vary, they appear to be lower than those implied by commonly used feasible rules. For instance, Nielsen (2005) uses \( m = [n^{0.4}] \) and \( m = [n^{0.5}] \), yielding \( m = 21, 45 \) for \( n = 2048 \), which would be clearly too high for most of the nonlinear settings considered. While in the linear setting the RMSE profiles seem to be relatively sensitive to the choice of \( d \), in the nonlinear one a wide range of values for \( d \) perform comparably; this is possibly a consequence of the lower bandwidths used. The optimal
choice of $d$ seems to be sensitive to most parameters in the model, so a feasible rule would undoubtedly require preliminary estimation of these.

A brief comparison of residual memory estimates was also presented. It seems that, while MLW is found to dominate LW in RMSE for large enough $n$, due to the large negative bias of the latter, it displays high dispersion and bimodality, which can be especially misleading in cointegration analysis, where the focus is often on the difference between memory estimates obtained from observables and residuals. On the contrary, LW, being biased downwards in both cases, might yield more accurate inference on the existence and degree of fractional cointegration. Evaluation of these issues is left for future research.

References


