Journal of Econometrics 177 (2013) 60–74

Contents lists available at ScienceDirect

Journal of Econometrics

journal homepage: www.elsevier.com/locate/jeconom

# Distribution theory for the studentized mean for long, short, and negative memory time series

Tucker McElroy [a], Dimitris N. Politis [b,*]

[a] Center for Statistical Research and Methodology, U.S. Census Bureau, 4600 Silver Hill Road, Washington, DC 20233-9100, United States

[b] Department of Mathematics, University of California at San Diego, La Jolla, CA 92093-0112, United States

ARTICLE INFO

*Article history:*
Received 10 May 2011
Received in revised form
9 January 2013
Accepted 12 June 2013
Available online 13 July 2013

*Keywords:*
Kernel
Lag-windows
Overdifferencing
Spectral estimation
Subsampling
Tapers
Unit-root problem

ABSTRACT

We consider the problem of estimating the variance of the partial sums of a stationary time series that has either long memory, short memory, negative/intermediate memory, or is the first-difference of such a process. The rate of growth of this variance depends crucially on the type of memory, and we present results on the behavior of tapered sums of sample autocovariances in this context when the bandwidth vanishes asymptotically. We also present asymptotic results for the case that the bandwidth is a fixed proportion of sample size, extending known results to the case of flat-top tapers. We adopt the fixed proportion bandwidth perspective in our empirical section, presenting two methods for estimating the limiting critical values—both the subsampling method and a plug-in approach. Simulation studies compare the size and power of both approaches as applied to hypothesis testing for the mean. Both methods perform well – although the subsampling method appears to be better sized – and provide a viable framework for conducting inference for the mean. In summary, we supply a unified asymptotic theory that covers all different types of memory under a single umbrella.



## 1. Introduction

Consider a sample $Y = \{Y_1, Y_2, \ldots, Y_n\}$ from a strictly stationary time series with mean $EY_t = \mu$, autocovariance function (acf) $\gamma_h = \text{Cov}(Y_t, Y_{t+h})$, and integrable spectral density function $f(\omega) = \sum_h \gamma_h e^{-ih\omega}$. We are interested in studying the distribution of the studentized sample mean where the normalization involves the summation of sample autocovariances weighted by an arbitrary taper, and when the stochastic process exhibits either short or long memory or even when the process is overdifferenced. The latter case is especially tricky – and not well-studied in the literature – because in this case the studentization is achieved by dividing with a quantity that tends to zero. The objective in studentizing/self-normalizing the mean is the generation of a pivotal asymptotic distribution that can serve as the basis for the construction of confidence intervals and hypothesis tests for the unknown mean $\mu$.

In the case that the autocovariances $\gamma_h$ are (absolutely) summable with $\sum_h \gamma_h > 0$, it is well-known – under regularity conditions – that the sample mean $\overline{Y} = n^{-1}\sum_{t=1}^n Y_t$ is asymptotically normal with variance $f(0) = \sum_h \gamma_h$, i.e., $\sqrt{n}(\overline{Y} - \mu) \stackrel{\mathcal{L}}{\Longrightarrow} N(0, f(0))$. (Here $\stackrel{\mathcal{L}}{\Longrightarrow}$ denotes weak convergence.) A consistent estimate of $f(0)$ is given by

$$W_{\Lambda,M} = \sum_{|h|\le M} \Lambda_M(h)\widetilde{\gamma}_h, \qquad (1)$$

where $\Lambda_M$ is an arbitrary taper (described in Section 3), and $\widetilde{\gamma}_h$ are the sample autocovariance estimators defined by

$$\widetilde{\gamma}_k = \frac{1}{n}\sum_{t=1}^{n-|k|}(Y_{t+|k|} - \overline{Y})(Y_t - \overline{Y}) \quad \text{for } |k| < n. \qquad (2)$$

As usual, $M = M(n)$ is a bandwidth parameter tending to $\infty$ as $n \to \infty$ but in such a way that $M$ is smaller than $n$. Define the *bandwidth-fraction* to be $b = b(n) = M(n)/n$. The literature on such taper-based, "lag-window" spectral estimators is extensive, going back over fifty years; see e.g. Hannan (1970), Brillinger (1981), Priestley (1981), Rosenblatt (1985), Brockwell and Davis (1991), and the references therein. Also see Grenander and Rosenblatt (1957), Blackman and Tukey (1959), and Percival and Walden (1993). Much is already known about the classical case where $f(0)$ is bounded above and bounded away from zero, but also about the long memory case where $f(0)$ is infinite—see, e.g., Beran (1994), Robinson (1994), and Palma (2007). Other recent literature includes Sun (2004), Robinson (2005), and Abadir et al. (2009). However, little is yet known in the case that $f(0) = 0$, although this possibility was brought to the forefront early on by Rosenblatt (1961). We attempt to remedy this situation in the paper at hand.

* Corresponding author. Tel.: +1 858 5345861.
*E-mail addresses:* tucker.s.mcelroy@census.gov (T. McElroy), dpolitis@ucsd.edu (D.N. Politis).

0304-4076/$ – see front matter © 2013 Elsevier B.V. All rights reserved.
http://dx.doi.org/10.1016/j.jeconom.2013.06.002

The case $f(0) = 0$ will be referred to henceforth as the superefficient (SE) case, since it implies $\overline{Y} = \mu + o_P(n^{-1/2})$, i.e., superefficient estimation of the mean.

The SE case has received little attention in the literature despite the fact that it is of some importance in applied econometrics. We illustrate this through a brief discussion of the unit root problem that arises from over-differencing. Although most economic time series exhibit obvious trends, much debate rages over whether processes are stationary or $I(1)$; witness the extensive literature on unit-root testing, starting with Dickey and Fuller (1979, 1981). Hamilton (1994) gives an overview; also see Phillips and Perron (1988) and Parker et al. (2006) and the references therein. It is commonly felt that an economic time series is rarely $I(2)$, and yet many such $I(2)$ models are selected by automatic model identification software, such as X-12-ARIMA and TRAMO-SEATS (both of which often select the Box–Jenkins airline model, which is $I(2)$); see the discussion in Findley et al. (1998), and Maravall and Caparello (2004).

Over-differencing relates to over-specification of the order of differencing, i.e., modeling a process as $I(1)$ when it is stationary, or as $I(2)$ when it is only $I(1)$. Naturally, estimation of parameters (such as the mean and other regression effects, but also maximum likelihood estimation of ARMA parameters) is performed on the differenced series, where the nonstationarity has been removed. But if the differencing order has been over-specified, then there will be over-differencing; this results in the spectrum of the differenced data being zero at frequency zero. For example, if $\{Y_t\}$ is stationary with spectral density $f$ but is viewed as $I(1)$, then the spectrum of the data's first difference is $|1 - e^{-i\lambda}|^2 f(\lambda)$. We mention in passing that similar issues exist, at least in theory, for seasonal frequencies, i.e., taking a seasonal difference when the seasonal component of the time series is actually stationary. In this case, the zero in the spectrum takes place at the seasonal frequencies corresponding to the angular portions of the zeros of the seasonal differencing operator. Such zeros offer no impediment to the estimation of the sample mean, but may generate other problems in model estimation; we do not pursue this point further here.

One approach to this problem is to do a pre-test of a possible unit root before differencing the data (Dickey and Fuller, 1979). Also see the treatment in Tanaka (1990, 1996). The power of the Dickey–Fuller statistic is explored in many articles, including Lopez (1997), and the power can be rather small because the statistic is not $\sqrt{n}$ consistent under the alternative. Given that Type II errors will occur some of the time, we advocate in this paper the use of robust studentized sample mean estimates, where the robustness is with respect to the three basic cases for $f(0)$: infinite (long memory), finite and positive (short memory), or zero (the SE case, distinguished in what follows as either negative memory or differential memory). To that end, we study the finite-sample and asymptotic distributional properties of studentized sample means, where the normalization is of the form (1) and the stochastic process satisfies some very general conditions. Note that $W_{\Lambda,M}$ must mirror the properties of the variance of the sample mean under the three scenarios: for long memory, it must diverge at the appropriate rate; for short memory it should converge to the same constant; and for negative/differential memory it should tend to zero at the same rate. Our work determines what conditions on a taper are needed in order to ensure such a robustness against different alternative memory scenarios.

In this paper we focus on the sample mean statistic, because of its central role in nonparametric estimation of location, but there are further applications and implications of our results. The Generalized Method of Moments (see Kiefer and Vogelsang, 2005) uses the partial sums of some function of the time series in order to estimate parameters of interest; for example, the average of squared one-step ahead in-sample forecasts from a given model provide such a partial sum. Partial sum statistics also are featured in unit root tests (Parker et al., 2006) and change point tests (Kirch and Politis, 2011).

Our main goal is to develop a practical procedure for conducting inference for the mean using the studentized sample mean, such that the methodology is valid when long memory is present. To address this goal we derive novel results regarding the joint distributional properties of $\overline{Y}$ and $W_{\Lambda,M}$ for various stochastic processes and various tapers, distinguishing between the case that the bandwidth-fraction $b$ is vanishing and the case that it is a constant proportion. In the former case, we obtain central limit theorems for the sample mean, while the variance estimate tends in probability to a constant when appropriately normalized (Section 3). But when the bandwidth-fraction is constant, the variance estimate tends to a random limit (Section 4). For the SE case these are all novel results. In Section 2 we give precise definitions of the four types of memory (LM, SM, NM, and DM), and also discuss some basic properties, many of which are new. The limit results of Sections 3 and 4 are utilized in Section 5, where the interesting application of testing for the mean is discussed. We propose two methods of estimating limit quantiles, one that utilizes subsampling (Politis et al., 1999), and one that involves estimation of the memory parameter. These procedures are evaluated through size and power simulations, covering many stochastic processes, tapers, and bandwidth-fractions. Section 6 summarizes our findings, and proofs are gathered in an Appendix.

## 2. Types of memory

A primal modeling assumption for many economic time series of interest is that they arise from sampling an $I(d)$ process, where $d$ is integer. Whereas in time past the differenced process was modeled via some short memory device, such as an ARMA model, now more complex covariance structures, including long memory, can be entertained—see Palma (2007) for an overview. So we suppose that $\{X_t\}$ is $I(D)$ with $D$ possibly non-integer, and we difference $d$ times (with $d$ an integer), obtaining $Y_t = (1 - B)^d X_t$.

If $D - d = 0$, then $\{Y_t\}$ is short memory—this corresponds to the classical case alluded to above. If $D - d > 0$, we obtain a long memory process, which is stationary when $D - d < 1/2$. Note that if we differenced insufficiently such that $D - d \geq 1/2$, then $\{Y_t\}$ would be nonstationary, and this might be detected by inspection of the sample autocorrelations. Then $d$ could be incremented through higher integers until $D - d$ falls into the range corresponding to stationarity. One could also proceed more rigorously by using unit root tests.

Over-differencing, whereby $D - d < -1/2$, would be harder to guard against. This is because in practice both the appropriately differenced series and the over-differenced series often have sample autocorrelations consistent with a stationary hypothesis. Unit root testing techniques can be employed to guard against this error, but there is some evidence that they have lower power in the presence of long memory (Diebold and Rudebusch, 1991). This discussion, which applies to many time series being examined for the purposes of forecasting, signal extraction, or other analysis, indicates how long memory, short memory, or SE may arise. When the order of differencing $d$ is selected correctly (i.e., $D - d < 1/2$), then $D \in (d, d + 1/2)$ implies that $\{Y_t\}$ will have (stationary) long memory; if $D = d$ then $\{Y_t\}$ will have short memory. But if $D \in [d - 1/2, d)$, then $\{Y_t\}$ will have negative memory (this is part of the SE case, and is formally described below). Finally, the data is over-differenced when $D - d < -1/2$; this case corresponds to differential memory (the other portion of the SE case—see below), and the sample mean has a radically different behavior. So the SE situation arises either as anti-persistence or from over-differencing due to a false acceptance of the unit root hypothesis. Although anti-persistence may arise naturally in time series, it can also arise at once when nonstationary long memory time series are differenced appropriately.

We now discuss in detail the different memory scenarios. Throughout this paper we use the notation $A_n \sim B_n$ to denote $A_n/B_n \to 1$ as $n \to \infty$. By Short Memory (SM), we refer to the condition that the autocovariances – recall that $\gamma_h = \text{Cov}(Y_t, Y_{t+h})$ – are absolutely summable and their sum $f(0)$ is a nonzero constant. By Long Memory (LM), we mean that the autocovariances are not summable, and the partial sums of them, denoted by $W_n$, satisfy

$$W_n = \sum_{|k|\leq n} \gamma_k = L(n)n^{\beta}. \tag{3}$$

In (3) $L$ is slowly varying at infinity (Embrechts et al., 1997), with a limit that can be zero, $C$, or infinity, where $C$ is a positive constant. Most importantly, $\beta > 0$ (and is less than 1). In the SM case (3) applies with $\beta = 0$ and $L$ tending to $C = \sum_k \gamma_k = f(0)$. The case that $\beta = 0$ but $L$ tends to infinity is also LM (e.g., say $\gamma_k = k^{-1}$ for $k \geq 1$).

We will denote by Negative Memory (NM) the case of an absolutely summable autocovariance sequence such that (3) holds, but with $\beta < 0$ (though the case that $\beta = 0$ and $L$ tends to zero is also considered to be NM). Some authors have used the term "intermediate memory" for this concept (Brockwell and Davis, 1991). Our nomenclature is due to the negative memory exponent, and also the result that most of the autocovariances are negative in this case (see Remark 2 below); the same conditions on $L$ apply here. When the autocovariances are zero past a certain threshold, we obtain an example of Differential Memory (DM). For example, consider an $MA(q)$ data process that is over-differenced; the resulting autocovariances are identically zero for lags exceeding $q$, and $\sum_{|k|<n} \gamma_k = 0$ for $n > q + 1$. These definitions encompass ARFIMA models (Hosking, 1981), FEXP models (Beran, 1993, 1994), and fractional Gaussian Noise models, as well as the case of over-differenced processes. Some authors prefer to parametrize memory in terms of the rate of explosion of $f$ or $1/f$ at frequency zero, but it is more convenient for us to work in the time domain; see Palma (2007) for an overview.

To distinguish between the LM, SM, NM, and DM memory cases, the key determinant in the limit theorems for $S_n = n\overline{Y}$ is the rate of growth of $V_n = \text{Var}(S_n)$, which in turn is related to $\sum_{|k|\leq n} \gamma_k$. In the rest of this paper we study $S_n/\sqrt{\widehat{V}_n}$, where $\widehat{V}_n$ is some estimate of the variance $V_n$ such that $\sqrt{\widehat{V}_n} = O_P(S_n)$. Recall that $W_n = \sum_{|k|\leq n} \gamma_k$ by definition, which in turn has asymptotic behavior given by (3) by assumption; then we have the following identity:

$$V_n = \sum_{k=0}^{n-1} W_k \tag{4}$$

that is proved using summation by parts. Now for LM processes, $W_n$ diverges, whereas in the SM case $W_n$ tends to a nonzero constant. The superefficient case (SE) where $f(0) = 0$ is characterized by $W_n$ tending to zero; however, we distinguish the case that $W_n$ is summable (DM case) versus the case that it is not (NM case).

**Definition 1.** Define the four types of memory as follows:

- (i) LM: $W_n$ has asymptotics given by (3) with either $\beta \in (0, 1)$ or $\beta = 0$ and $L$ is tending to infinity.
- (ii) SM: $W_n \to C$ with $C > 0$.
- (iii) NM: $W_n$ has asymptotics given by (3) with either $\beta < 0$ or $\beta = 0$ and $L$ is tending to zero. When $\beta = -1$, suppose that $W_n$ is not summable.
- (iv) DM: $W_n$ is summable.

The condition (3) on $W_n$ in case (i) is denoted LM($\beta$), and the same condition on $W_n$ in case (iii) is denoted NM($\beta$).

**Remark 1.** Note that cases (i), (ii), and (iii) are mutually exclusive, since case (ii) essentially corresponds to $\beta = 0$ in (3) with $L$ tending to a nonzero constant. When $\beta = -1$ in (3), it is possible for $W_n$ to be either summable or not summable; the former case belongs to case (iv). If $\beta < -1$ in (3), then necessarily this is case (iv). Proposition 1 below shows that most of the autocovariances are negative in case (iii), and this justifies the nomenclature of "negative memory". In case (iv) $\{Y_t\}$ can be represented as the difference of another stationary process (Proposition 2), and hence the appellation "differential memory". The case of an over-differenced series, as discussed above, is always included in case (iv). Both NM and DM have $f(0) = 0$, so both are part of the SE case.

**Remark 2.** These four cases are mutually exclusive but do not cover all possibilities, since we can consider a function $L$ in (3) that is not slowly-varying (e.g., $L(n) = 1 + \sin(n)$). However, the four cases *are* exhaustive for processes satisfying (3) with $L$ slowly-varying and $\beta < 1$.

The following representation for slowly-varying functions (Theorem A3.3 of Embrechts et al., 1997) will be used below: we can write $L$ in the form

$$L(n) = c(n) \exp \int_z^n \eta(u)/u \, du, \tag{5}$$

with $c(n) \to c$ a positive constant, $z$ some fixed positive constant, and $\eta(u)$ tending to zero as $u \to \infty$. Note that in (5) we can take the variable argument of $L$ to be a continuous variable $x$. The following result gives the behavior of $V_n$ in each of the four cases described in Definition 1, and also discusses the implied asymptotic behavior of $\gamma_k$.

**Proposition 1.** *With $W_n$ and $V_n$ given by (3) and (4) respectively,*

$$V_n \sim \frac{1}{\beta+1} nW_n = \frac{L(n)n^{\beta+1}}{\beta+1} \tag{6}$$

*for cases* (i) *and* (ii), *and for $\beta > -1$ in case* (iii). *When $\beta = -1$ we have $V_n/(nW_n) \to \infty$ (so in a sense (6) holds true). In case* (ii), *we let $\beta = 0$ in the formula and $L(n) \sim f(0)$. In case* (iv), *$V_n \to \sum_{k\geq 0} W_k$. If $\beta \neq 0$ in cases* (i) *and* (iii), *we also have*

$$\gamma_n \sim \frac{\beta}{2} L(n)n^{\beta-1}. \tag{7}$$

*In cases* (i) *and* (iii) *with $\beta = 0$ we have $\gamma_n \sim -0.5 \int_{n-1}^n [\eta(u)/u] \, du = o(n^{-1})$.*

This result shows that $V_n \to \infty$ in cases (i), (ii), and (iii), as long as $\beta > -1$. This will facilitate a fairly standard limit theorem for $\overline{Y}$ under some additional conditions. Case (iv) produces a very different sort of limit theorem; these results are discussed in Section 3.

**Remark 3.** Note that $L(n)$ must be non-negative for all $n$ larger than some $n_0$, say; this follows from (6) and the fact that $V_n > 0$ for all $n$. Hence, for large $n$ all the $\gamma_n$ are *negative* in the NM case and *positive* in the LM case by (7). Essentially, NM is due to heavy negative correlation and LM to heavy positive correlation; this justifies the name "negative memory" for the NM case.

**Example 1.** Let $\beta > 0$ with $\gamma_h = h^{-\beta}$ for $h \geq 1$ and $\gamma_0$ chosen suitably large to guarantee the sequence is positive definite; then this is the acf of a LM process. If we (temporally) difference the process, then the resulting acf is $2\gamma_h - \gamma_{h+1} - \gamma_{h-1}$, which equals

$$2h^{-\beta} - (h+1)^{-\beta} - (h-1)^{-\beta} = -\beta(\beta+1)\, h^{-(\beta+2)} + o(1)$$

when $h \geq 2$. Although this appears at first to be NM (comparing to (7)), in fact it can be shown that $W_n$ is summable so that the differenced process is DM (see Proposition 2).

**Example 2.** Let $\beta < 0$ and $\gamma_h = -h^{\beta-1}$ for $h > 0$, and $\gamma_0 = 2\sum_{h\geq 1} h^{\beta-1}$. Then the discrete Fourier transform of $\{\gamma_h\}$ is $2\sum_{h\geq 1} h^{\beta-1}(1 - \cos \lambda h)$ for $\lambda \in [-\pi, \pi]$, which is always non-negative. Hence $\{\gamma_h\}$ is the acf of a time series process; by (7) it seems to have the form of a NM acf, but we must check the summability of $W_n$. Direct calculation shows that

$$W_n = 2\sum_{h>n} h^{\beta-1} = O(n^{\beta}).$$

Thus if $\beta \geq -1$ the process is NM, but otherwise is DM.

In case (ii) there is no result of the form (7), but in case (iv) the autocovariances have a particular structure if we suppose in addition that $\{Y_t\}$ is purely non-deterministic. This is discussed in the proposition below, whose result is similar to Theorem 8.6 of Bradley (2007).

**Proposition 2.** *Suppose that $W_k$ is summable and $\{Y_t\}$ is purely non-deterministic with mean $\mu$ and acf $\{\gamma_h\}$. Then there exists a strictly stationary process $\{Z_t\}$ with autocovariance sequence $r_k$ such that $Y_t = Z_t - Z_{t-1} + \mu$ and $\gamma_k = 2r_k - r_{k+1} - r_{k-1}$. This decomposition is not unique. Also $W_k = 2(r_k - r_{k+1})$ and $r_0 = \sum_{k\geq 0} W_k/2$. Conversely, given any stationary process $\{Z_t\}$ with acf $r_k$ tending to zero, the differenced process $Z_t - Z_{t-1}$ is DM.*

We note that the process $\{Z_t\}$ might be LM, SM or NM, and might even be DM.

**Example 3.** Let $\{Z_t\}$ be *i.i.d.* and let $Y_t = Z_t - Z_{t-1}$. Then clearly $\{Y_t\}$ is DM with autocorrelation of $-0.5$ at lag one and zero at higher lags. If the innovation variance is unity, then $W_0 = 2$ and $W_k = 0$ for $k \geq 1$; this is clearly a summable sequence.

**Example 4.** A more interesting example is given as follows: let $Y_t = \sum_{j=0}^{\infty} \psi_j \epsilon_{t-j}$ where $\{\epsilon_t\}$ is white noise and $\psi_j = (-1)^j/j^p$ for $p > 0$ and $j \geq 1$, with $\psi_0 = -\sum_{j=1}^{\infty} \psi_j$ (which clearly exists by the alternating series test). Hence $\sum_j \psi_j = 0$, which implies that $\sum_k \gamma_k = 0$. Notice that the variance of $Y_t$ only exists if $p > 1/2$, since

$$\sum_{j=0}^{\infty} \psi_j^2 = \psi_0^2 + \sum_{j=1}^{\infty} j^{-2p} = F^2(1/2, p) + F(1, 2p),$$

where $F(x, s)$ is the periodic-zeta function given by $F(x, s) = \sum_{n\geq 1} e^{2\pi i x n} n^{-s}$. The variance of $\sum_{j=0}^{n} \psi_j \epsilon_{t-j}$ is $\sum_{j=1}^{n} j^{-2p} + \psi_0^2 \sim n^{1-2p}$, which diverges unless $p > 1/2$; we assume this henceforth, as we are not concerned with infinite variance time series in this paper. As in the proof of Proposition 2, define

$$\theta_j = \sum_{k=0}^{j} \psi_k = -\sum_{k=j+1}^{\infty} \frac{(-1)^k}{k^p}.$$

The sum of any two consecutive terms of this series, up to a minus sign, can be written as $k^{-p} - (k+1)^{-p} = ((1+1/k)^p - 1)/(k+1)^{-p}$, which by Taylor series expansion about zero yields an approximation of $p\,k^{-1}(k+1)^{-p}$ plus terms that decay at order $k^{-2-p}$. Thus asymptotically, the sum of consecutive terms in $\theta_j$ is $p\,k^{-p-1}$, plus other terms that decay even faster. So such a sequence is summable, and we find that $\theta_j = O(j^{-p})$ as $j \to \infty$. Since $p > 1/2$, this sequence is square summable. This implies that the time series $Z_t = \sum_{j\geq 0} \theta_j \epsilon_{t-j}$ is well-defined, i.e., is finite almost surely, since it has finite variance. Clearly $Y_t = Z_t - Z_{t-1}$, and the other assertions of Proposition 2 apply; in particular, $\{Y_t\}$ is DM.

**Example 5.** Suppose that an observed time series $\{X_t\}$ is an ARFIMA (0, $D$, 0) – see Palma (2007) – so that $(1-B)^D X_t = \epsilon_t$ is white noise, where $D \in [0, 1]$. If $D = 1$ this is just a random walk, and if $D \in [0.5, 1)$ the process is said to have nonstationary long memory. If $D < 0.5$ the process is stationary, but with long memory if $D > 0$. Of course $D = 0$ corresponds to short memory (white noise). If the observed process is differenced once to produce $Y_t = X_t - X_{t-1}$, it is easy to see that the result is stationary with memory parameter $\beta = 2D - 2$. That is, if $D = 1$ we obtain short memory (white noise); if $D \in [0.5, 1)$ we obtain a negative memory process of parameter $\beta \in [-1, 0)$; if $D < 0.5$ then we obtain a process with differential memory. The borderline case $D = 0.5$ is interesting: we do not get a differential memory process, since the original process is nonstationary—instead we get a negative memory process with $\beta = -1$ and nonsummable $W_k$ sequence.

## 3. Limit theory for the case of vanishing bandwidth-fraction

In the case that $b(n) \to 0$ as $n \to \infty$, we can treat the asymptotics of $S_n$ and $W_{\Lambda,M}$ separately (recall that $W_{\Lambda,M}$ was defined in (1)), because the variance estimate, when appropriately normalized, always converges to a constant. Let us then consider the partial sums first; it is necessary to impose some additional assumptions. Typical assumptions for limit theorems involve either moment and mixing conditions, or linearity of the process involved. Limit theorems have also been derived under the assumption that the given process is a direct function of an underlying Gaussian process. It turns out (see discussion below) that mixing assumptions are not helpful when (3) holds. Instead, one can make fairly strong assumptions on the higher order cumulants of the time series. The $k$th order cumulant of $\{Y_t\}$ is defined by $c_k(u_1, u_2, \ldots, u_{k-1}) = \mathrm{cum}(Y_{t+u_1}, Y_{t+u_2}, \ldots, Y_{t+u_{k-1}}, Y_t)$ for any $t$ and integers $u_1, \ldots, u_{k-1}$, where $k \geq 1$ (Taniguchi and Kakizawa, 2000). Letting $u$ denote the $k-1$-vector of indices, we write $c_k(u)$ for short. Also let $\|\cdot\|$ denote the sup-norm of a vector, so that $\sum_{\|u\|<n} c_k(u)$ is a short-hand for summing the cumulant over all indices such that $|u_j| < n$ for each $j$. Given these notations, the assumptions on our process $\{Y_t\}$ that we will consider are given below:

- P1. $\{Y_t\}$ is a linear process: $Y_t = \sum_j \psi_j \epsilon_{t-j}$ with $\{\psi_j\}$ square summable and $\{\epsilon_t\}$ *i.i.d.* with finite variance. Also, $\{Y_t\}$ is either (i) LM($\beta$) with $\beta \in [0, 1)$; (ii) SM; or (iii) NM($\beta$) with $\beta \in [-1, 0]$.
- P2. $\{Y_t\}$ is a process that is either (i) LM($\beta$) with $\beta \in [0, 1)$; (ii) SM; or (iii) NM($\beta$) with $\beta \in (-1, 0]$. Also assume the $k$th order cumulants exist and are summable over its $k$ indices, for all $k \geq 1$. Moreover, when $\beta < 0$ we also assume that $\sum_{\|u\|<n} |c_k(u)| = O(n^{\beta})$.

**Remark 4.** Yet another type of assumption would require $2+\delta$ moments (for some $\delta > 0$) and a mixing assumption. It is known that the strong mixing condition of Rosenblatt (1956)[1] is not satisfied by long memory Gaussian processes—Gaussian processes are strongly mixing iff their spectrum is bounded away from zero and infinity (Kolmogorov and Rozanov, 1960). If the process is not Gaussian, it is conceivable that a strong mixing condition could be satisfied, although no examples of this have been published. The weak dependence condition of Doukhan and Louhichi (1999) might be used instead, because in Bardet et al. (2008) it is shown to be compatible with long range dependence. Along with a Lindeberg condition on higher moments of the partial sum, a central limit theorem can be established utilizing the Bernstein blocks method described in Rosenblatt (1956) and Doukhan and Louhichi (1999). Unfortunately, such a result is not compatible with (3) when $\beta \neq 0$: examination of Theorem 1 of Rosenblatt (1984) indicates that the rate of convergence in this type of CLT must be $k_n V_{p_n}$, where

[1] A stationary process $\{Y_t\}$ is strongly mixing (Rosenblatt, 1956) if $\alpha_k \to 0$ as $k \to \infty$ where $\alpha_k = \sup_{A\in\mathcal{F}_{-\infty}^{0}, B\in\mathcal{F}_k^{\infty}} |\mathbb{P}(A\cap B) - \mathbb{P}(A)\mathbb{P}(B)|$, and $\mathcal{F}_j^m$ is the $\sigma$-algebra generated by $\{Y_k, j \leq k \leq m\}$.

$k_n$ is the number of big blocks and $p_n$ is the size of the big blocks, and moreover $n \sim k_n p_n$. This is not a paradox: in order to prove a CLT, fairly strong assumptions must be placed upon the mixing coefficients – either they must be strong mixing, or must have rapid decay rate if they are weak dependence coefficients, etc. – which then actually precludes the possibility of (3). For this reason, we do not formulate limit theorems under mixing conditions in this paper; also see the broader discussion in McElroy and Politis (2011).

Each of the above assumptions P1 and P2 provide sufficient conditions for a limit theorem for $S_n$, as shown below. Results under a third condition can be found in McElroy and Politis (2011). Note that these cover only cases of LM, SM, and NM; the DM case must be handled separately in what follows. The $\beta = -1$ case is only handled under P1. Note that in the case of P2 with a Gaussian process, all the cumulant assumptions are automatically satisfied. In any of these scenarios, the assumptions are typically unverifiable, or difficult to verify from the data; these should be viewed as working assumptions.

The following result covers the cases of LM, SM, and NM under the different conditions P1 and P2. We also establish a result for the DM process, restricting to purely non-deterministic processes so that Proposition 2 can be applied.

**Theorem 1.** *Suppose that $\{Y_t\}$ is a strictly stationary process with finite variance that satisfies (3). If the process satisfies one of* P1 *or* P2, *then $\frac{S_n - n\mu}{\sqrt{V_n}} \stackrel{\mathcal{L}}{\Longrightarrow} B$ as $n \to \infty$, where $B$ is standard normal. If $\{Y_t\}$ is purely non-deterministic with summable $W_k$, and if the process $\{Z_t\}$ of Proposition 2 is either linear or strong mixing, then $\frac{S_n - n\mu}{\sqrt{V_n}} \stackrel{\mathcal{L}}{\Longrightarrow} (Z_* - Z_0)/\sqrt{2r_0}$ as $n \to \infty$, where $Z_*$ is a random variable equal in distribution to $Z_0$, but independent of it.*

To utilize Theorem 1 it is important to know $V_n$, or to estimate it. In view of (3), the taper-based estimate given by $W_{\Lambda,M}$ in (1) is a nonparametric estimate, and can be related to estimation of $V_n$ via (4) and (6). For example, with an SM process $\widehat{V}_n = nW_{\Lambda,M}$ is often used as an estimator.

The asymptotic behavior of $W_{\Lambda,M}$ depends on the type of taper as well as the type of memory of the process. The tapers that we consider are very general: $\Lambda_M$ is a piecewise smooth (i.e., piecewise differentiable), even function on the integers such that $\Lambda_M(h) = 0$ for $|h| \geq M$. Letting $U_M$ denote the maximum value of $\Lambda_M(h)$ for all $h$, we suppose that $U_M$ does not grow too fast as $M \to \infty$. Classical tapers are bounded, in which case $U_M$ can be taken constant. The triangular (Bartlett) kernel, the trapezoidal (Politis and Romano, 1995), and the more general flat-top kernels (Politis, 2001; Politis, 2005) all satisfy these conditions.

Our next main result is that under some conditions the estimator $W_{\Lambda,M}$ is asymptotic to the deterministic sequence $\widetilde{W}_M$ as $M \to \infty$ (recall that $M/n \to 0$, though), which is defined via

$$\widetilde{W}_M = \sum_h \Lambda_M(h) \left(1 - \frac{|h|}{n}\right) \gamma_h. \tag{8}$$

We require the following condition, which parallels Assumption A of Andrews (1991), to establish our result:

**Assumption B.** $\mathbb{E}|Y_t^4| < \infty$ and the fourth order cumulant $c_4$ is absolutely summable.

The above assumption is compatible with the process conditions P1 and P2, for the LM, SM, and NM cases. Assumption B is also compatible with the DM case. As discussed in Andrews (1991), linear processes with absolutely summable coefficients and finite fourth moments satisfy Assumption B, even if the process is LM. As Lemma 1 of Andrews (1991) shows, Assumption B is also implied by a strong mixing plus moments condition. Of course, a long memory Gaussian process trivially satisfies Assumption B, because its fourth order cumulants are zero.

**Proposition 3.** *Suppose that $\{Y_t\}$ is a strictly stationary process satisfying Assumption B, which is either LM, SM, NM, or DM. Also suppose that $b(n) + 1/M(n) \to 0$ as $n \to \infty$. Then $W_{\Lambda,M} = \widetilde{W}_M(1 + o_P(1))$.*

Next we consider the asymptotics of the deterministic sequence $\widetilde{W}_M$ (8), which depends upon the memory assumptions. Together with Proposition 3, this will show that $W_{\Lambda,M}$ is asymptotic to a constant times $W_M$. This constant will be denoted by $\zeta$, and depends upon $\beta$ and the taper used. This type of result is pertinent, as discussed above, to the estimation of $V_n$. We consider tapers of the form $\Lambda_M(h) = \Lambda(h/M)$ for a fixed function $\Lambda(x)$. We assume $\Lambda$ is an even, piecewise smooth function, that is real-analytic on every such interval; by $\dot{\Lambda}_+(x)$ for $x \geq 0$, we denote the derivative from the right (and $\Lambda_+^{(j)}(x)$ for higher order derivatives). An example is a "flat-top" taper where there exists an interval $[0, c]$ (with $c > 0$) for which $\Lambda$ equals unity—see Politis (2001). Then define $\zeta = 1$ when $\Lambda$ is the truncation taper (i.e., $\Lambda = 1_{[-1,1]}$), and otherwise

$$\zeta = -\int_0^1 \dot{\Lambda}_+(x) x^\beta \, dx. \tag{9}$$

Then we have the following result.

**Theorem 2.** *Let $\Lambda(x)$ be an even, piecewise differentiable function supported on $[-1, 1]$, with $\Lambda_M(h) = \Lambda(h/M)$. Suppose that $\{Y_t\}$ is a strictly stationary process satisfying Assumption B, which is either LM, SM, NM, or DM. Also suppose that $b(n) + 1/M(n) \to 0$ as $n \to \infty$. Then if the process is LM, SM, or NM with $\beta \in (-1, 1)$,*

$$\frac{\widetilde{W}_M}{W_M} \to \zeta \quad \text{as } M \to \infty. \tag{10}$$

*If the process is DM, then*

$$\widetilde{W}_M \sim -2\, r_0\, \dot{\Lambda}_+(0)\, M^{-1} \left(1_{\{c=0\}} + o(1)\right).$$

In the SM case recall that $C \sim \sum_k \gamma_k$, so that (9) with $\beta = 0$ yields $\zeta = \Lambda(0)$; this equals unity as long as $\Lambda(0) = 1$, which is commonly assumed. Thus $W_{\Lambda,M} \sim C\Lambda(0)$ and $V_n \sim nC$ can be estimated by $nW_{\Lambda,M}$ when $\Lambda(0) = 1$. In the DM case for non-flat-top kernels, the overall error is controlled by the first derivative $\dot{\Lambda}_+(0)$. This is because $c = 0$; but when $c > 0$, the rate of decay is even faster, and is hard to describe in a general result. For example, with the Bartlett taper $\dot{\Lambda}_+(0) = -1$, and $\widetilde{W}_M \sim 2r_0 M^{-1}$. However, for the Parzen and Trapezoidal tapers $\dot{\Lambda}_+(0) = 0$. The quantity $\zeta$ is asymptotic to one plus the relative bias $(\widetilde{W}_M - W_M)/W_M$, measuring the asymptotic discrepancy between our variance estimate $W_{\Lambda,M}$ and the sequence $W_M$. We refer to $\zeta$ as the *quotient bias* hereafter. Note that $\zeta$ is well-defined, since the derivative of $\Lambda$ exists almost everywhere.

**Remark 5.** It follows from Proposition 3, Theorem 2, and Proposition 1 – when their respective conditions are satisfied – that in probability

$$MW_{\Lambda,M} \sim (\beta + 1)\zeta\, V_M \tag{11}$$

$$MW_{\Lambda,M} \sim -2\, r_0\, \dot{\Lambda}_+(0) \left(1_{\{c=0\}} + o(1)\right)$$

for the cases of LM/SM/NM and DM respectively. Although $V_M = \mathrm{Var}(S_M)$ is of some interest, we really need to obtain a quantity asymptotic to $V_n$. Unfortunately, since $b(n) = M(n)/n \to 0$, we have $V_M/V_n \to 0$ except in the DM case. So we cannot normalize $S_n$ by $\sqrt{MW_{\Lambda,M}}$ in the LM/SM/NM case, since the normalization rate is not correct (also there is the matter of the unknown $(\beta + 1)\zeta$ factor). In the DM case, if we use a non-flat-top taper, we can indeed utilize $S_n \left(-MW_{\Lambda,M}/\dot{\Lambda}_+(0)\right)^{-1/2}$ since $2r_0 = \sum_{k \geq 0} W_k = V_\infty$; by Theorem 1, this studentized statistic converges to $Z_* - Z_0$.

Now if $\beta > 0$, then $\zeta$ can be rewritten as $\beta \int_0^1 \Lambda(x) x^{\beta-1}\, dx$ using integration by parts piecewise. In this case, we proceed to calculate $\zeta$ for some commonly used tapers. The simplest flat-top taper is the trapezoidal taper of Politis and Romano (1995) given by

$$\Lambda^{T,c}(x) = \begin{cases} 1 & \text{if } |x| \leq c \\ \dfrac{|x|-1}{c-1} & \text{if } c < |x| \leq 1 \\ 0 & \text{else} \end{cases}$$

with $c \in (0, 1]$. Then it follows that

$$\zeta^{T,c} = -\int_0^1 \dot{\Lambda}^{T,c}(x) x^\beta \, dx = \frac{1 - c^{\beta+1}}{(1-c)(1+\beta)}.$$

Also in the DM case, it can be shown rather easily that $M\widetilde{W}_M \sim \frac{2}{1-c}(r_{[cM]} - r_M)$, which tends to zero at a rate that depends upon the autocovariance sequence $\{r_k\}$ and the truncation $c$. The triangular (Bartlett's) taper is obtained as the limiting case of $\Lambda^{T,c}$ as $c \to 0$; in this case, $\zeta^{Bar} = 1/(\beta+1)$. Interestingly, the factor $(1+\beta)\zeta$ appearing in (11) is then unity for any $\beta \in (-1, 1)$. Also since $\dot{\Lambda}_+(0) = -1$ for the Bartlett, we have $MW_{\Lambda,M} \sim V_M$ for all cases, including DM.

The asymptotic quotient bias $\zeta$ is also easily computed for the Parzen taper, given by

$$\Lambda^{Par}(x) = \begin{cases} 1 - 6|x|^2 + 6|x|^3 & \text{if } |x| \leq 1/2 \\ 2(1-|x|)^2 & \text{if } 1/2 < |x| \leq 1 \\ 0 & \text{if } |x| > 1. \end{cases}$$

Then $\zeta^{Par} = (2 - (1/2)^\beta)(3/(\beta+3) - 6/(\beta+2) + 3/(\beta+1))$.

## 4. Limit theory for the case of fixed bandwidth-fraction

The vanishing bandwidth-fraction results of Theorem 2 and Remark 5 indicate a difficulty with using tapers when LM or NM is present, because it is difficult to capture the correct rate for all types of memory. In fact, there is an asymptotic distortion equal to the quotient bias $\zeta$ that depends on the unknown $\beta$ parameter. In addition, there is the presence of the slowly-varying function $L$. These problems can be resolved by using a fixed bandwidth-fraction approach, which is described in this section. Because the results for the DM case are quite different and do not allow for all of the same statistical methods, we do not pursue this case further here (though it is treated in McElroy and Politis (2011)).

As in Kiefer and Vogelsang (2005), let the bandwidth $M$ be proportional to sample size $n$, i.e., $M = bn$ with $b \in (0, 1]$. We stress that in this section $b$ is a fixed number, and does not depend on $n$ as in the previous section. Let $\widehat{S}_i = \sum_{t=1}^{i}(Y_t - \overline{Y})$ (so that $\widehat{S}_n = 0$). A derivative of $\Lambda$ from the left is denoted $\dot{\Lambda}_-$, whereas the second derivative is $\ddot{\Lambda}$. The greatest integer function is denoted by $[\cdot]$. We consider tapers from the following family:

$$\begin{aligned} &\{\Lambda \text{ is even and equals unity for } |x| \leq c,\ c \in [0,1). \\ &\quad \text{Furthermore, } \Lambda \text{ is supported on } [-1,1], \\ &\quad \text{is continuous, and is twice continuously differentiable on} \\ &\quad (c,1) \cup (-1,-c)\}. \end{aligned} \tag{12}$$

This assumption is slightly less restrictive than the conditions of Theorem 2, because we only require two continuous derivatives. This family of tapers includes the family of "flat-top" kernels of Politis (2005) where $c > 0$, as well as the Bartlett kernel (letting $c = 0$ and a linear decay of $\Lambda$), and other kernels considered in Kiefer and Vogelsang (2005). The following result was proved in McElroy and Politis (2012), and is restated here for convenience.

**Proposition 4** (*McElroy and Politis, 2012*). *Let $\Lambda$ be a kernel from family (12), and let the bandwidth be $M = bn$. Let $b \in (0, 1]$ be a constant bandwidth-fraction. Then*

$$\begin{aligned} nW_{\Lambda,M} &= \sum_{i,j=1}^{n} \widehat{S}_i \widehat{S}_j \left( 2\Lambda\left(\frac{i-j}{M}\right) - \Lambda\left(\frac{i-j+1}{M}\right) - \Lambda\left(\frac{i-j-1}{M}\right)\right) \\ &= -\frac{2}{bn} \sum_{i=1}^{n-[cbn]} \widehat{S}_i \widehat{S}_{i+[cbn]} \left( \dot{\Lambda}_+(c) + \frac{1}{2bn}\ddot{\Lambda}(c) + O(n^{-2}) \right) \\ &\quad - \frac{1}{b^2 n^2} \sum_{[cbn] < |i-j| < [bn]} \widehat{S}_i \widehat{S}_j \left( \ddot{\Lambda}\left(\frac{|i-j|}{bn}\right) + O(n^{-1}) \right) \\ &\quad + \frac{2}{bn} \sum_{i=1}^{n-[bn]} \widehat{S}_i \widehat{S}_{i+[bn]} \left( \dot{\Lambda}_-(1) + O(n^{-1}) \right). \end{aligned}$$

**Remark 6.** In case the taper is continuously differentiable at $c$, $\dot{\Lambda}_+(c) = 0$ and the second derivative becomes dominant in the first term, which can then be recombined with the second term to yield

$$-\frac{1}{b^2 n^2} \sum_{[cbn] \leq |i-j| < [bn]} \widehat{S}_i \widehat{S}_j \left( \ddot{\Lambda}\left(\frac{|i-j|}{bn}\right) + O(n^{-1}) \right).$$

Likewise, if there is no kink at $|x| = 1$, then $\dot{\Lambda}_-(1) = 0$ and the third term vanishes completely.

In order to apply this result, we need functional limit theorems for the partial sums, since $\widehat{S}_i = S_i - i/n\, S_n$. For the LM, SM, and NM cases such limit theorems can be proved which extend Theorem 1 under more restrictive conditions.

Functional limit theorems are often formulated in the Skorohod space, denoted $D[0, 1]$—see Karatzas and Shreve (1991). Because it is more convenient to prove tightness in $C[0, 1]$, the space of continuous functions, we will construct a linearly-interpolated process for the partial sums, and prove its convergence to Fractional Brownian Motion (FBM), which is defined in Samorodnitsky and Taqqu (1994). This provides a result of independent interest (Marinucci and Robinson (2000) work in $D[0, 1]$), and also facilitates our main application, given in Theorem 4 below, under fairly simple conditions. One additional stricture, which seems unavoidable, is the requirement for higher moments in the NM case; this is because the sample paths are less smooth in the NM case than the LM case, so that the tightness criterion is satisfied only by requiring higher moments. (The same problem affects results in the Skorohod space, as shown in Marinucci and Robinson (2000).)

So we consider the step function sum process $S_n(t) = S_{[nt]}$, and its linear interpolant $\xi_n(t) = S_{[nt]} + (nt - [nt])Y_{[nt]+1}$. The step function process is mean-centered at $[nt]\mu$, while the latter is mean-centered by $nt\mu$; both will be normalized by the sequence $\sqrt{V_n}$. It is immediate that $S_n \in D[0, 1]$ and $\xi_n \in C[0, 1]$.

**Theorem 3.** *Let $\kappa = 2 \wedge [2/(1+\beta)]$ and suppose that $\{Y_t\}$ is a strictly stationary process satisfying (3), and with moments of order $\kappa + \delta$ for some $\delta > 0$. Moreover, suppose that $\mathbb{E}[|S_n - n\mu|^{\kappa+\delta}] = O(V_n^{(\kappa+\delta)/2})$. Furthermore, assume that the process satisfies either* P1 *or* P2*, with $\beta \in (-1, 1)$. Then as $n \to \infty$*

$$V_n^{-1/2}\left(\xi_{[n\cdot]} - n\mu\cdot\right) \stackrel{\mathcal{L}}{\Longrightarrow} B \tag{13}$$

*in the sense that the corresponding probability measures on $C[0, 1]$ converge weakly. $B$ is a FBM process of parameter $\beta$.*

Letting $\widehat{S}_n(t) = S_n(t) - \frac{[tn]}{n} S_n$ and $\widehat{\xi}_n(t) = \xi_n(t) - \frac{[tn]}{n}\xi_n(1)$, it follows from Theorem 3 that $\widehat{S}_{[rn]}/\sqrt{V_n}$ converges weakly to the process $\widetilde{B}(r) = B(r) - rB(1)$, which is a Fractional Brownian Bridge

(FBB). Then putting Proposition 4 and Theorem 3 together – with the fact that $\widehat{S_n}$ and $\widehat{\xi_n}$ are equivalent stochastic processes – yields the following result.

**Theorem 4.** *Let $\Lambda$ be a kernel from family (12), and let the bandwidth $M = bn$. Let $b \in (0, 1]$ be a constant bandwidth-fraction. Let $\kappa = 2 \wedge [2/(1+\beta)]$ and suppose that $\{Y_t\}$ is a strictly stationary process satisfying (3), and with moments of order $\kappa + \delta$ for some $\delta > 0$. Moreover, suppose that $\mathbb{E}[|S_n - n\mu|^{\kappa+\delta}] = O(V_n^{(\kappa+\delta)/2})$. Furthermore, assume that the process satisfies either* P1 *or* P2 *with $\beta \in (-1, 1)$. Then as $n \to \infty$*

$$\frac{S_n - n\mu}{\sqrt{nW_{\Lambda,M}}} \overset{\mathcal{L}}{\Longrightarrow} \frac{B(1)}{\sqrt{Q(b)}}, \tag{14}$$

*where (recall that $c$ is defined in (12))*

$$\begin{aligned} Q(b) = & -\frac{2}{b}\dot{\Lambda}_+(c)\int_0^{1-cb}\widetilde{B}(r)\widetilde{B}(r+cb)\,dr \\ & -\frac{1}{b^2}\int_{cb<|r-s|<b}\widetilde{B}(r)\widetilde{B}(s)\ddot{\Lambda}\left(\frac{|r-s|}{b}\right)drds \\ & +\frac{2}{b}\dot{\Lambda}_-(1)\int_0^{1-b}\widetilde{B}(r)\widetilde{B}(r+b)\,dr. \end{aligned} \tag{15}$$

**Remark 7.** The first result (14) was derived in a preliminary calculation in McElroy and Politis (2009), and the distribution (15) has been tabulated. The joint distribution of $B(1)$ and $Q(b)$ was also explored in McElroy and Politis (2012) through the device of the joint Fourier–Laplace Transform.

Unlike the special case studied by Kiefer et al. (2000), the numerator $B(1)$ of (14) is not independent of the denominator $Q(b)$ if $\beta \neq 0$. To elaborate, Kiefer et al. (2000) considered the case $b = 1$ and $c = 0$, the kernel is the Bartlett, and $\beta = 0$ (although later work by Kiefer and Vogelsang (2002, 2005) generalizes to $b < 1$). Then $Q(1) = 2\int_0^1 \widetilde{B}^2(r)\,dr$, and the authors note that $B(1)$ is independent of $\widetilde{B}(r)$. As shown in McElroy and Politis (2012), this is true for other kernels as long as $\beta = 0$; however, if $\beta \neq 0$, then $B(1)$ and $Q(b)$ are dependent. Fortunately, it is a simple matter to determine the limiting distribution numerically for any given value of $\beta$, and any choice of taper and bandwidth fraction $b$.

## 5. Applications and numerical studies

The preceding two sections give two different perspectives on the asymptotic behavior of taper-normalized sample means. If we normalize $S_n - n\mu$ by $\sqrt{nW_{\Lambda,M}}$, the studentized statistic does not converge – except in the SM case – when adopting the vanishing bandwidth-fraction perspective (see Remark 5 above[2]). However, $(S_n - n\mu)/\sqrt{nW_{\Lambda,M}}$ converges to a nondegenerate distribution by Theorem 4 when adopting the fixed bandwidth-fraction approach. Since bandwidth choice is up to the practitioner, it appears that the fixed bandwidth-fraction viewpoint might be preferable in our attempt towards a unified treatment of inference for the mean that is valid in all kinds of scenarios.

However, the limit distribution of the studentized sample mean will generally depend on $\beta$. Either one must estimate the nuisance parameters – including $\beta$ – or a nonparametric technique such as the bootstrap or subsampling (Politis et al., 1999) must be utilized to get the limit quantiles. The parametric bootstrap is not feasible here (since no model is specified for the data in our context) and the block bootstrap tends to perform badly when autocorrelation dies gradually (Lahiri, 2003). However, given that the limit distribution has been tabulated for some values of $\beta \in (-1, 1)$ and some popular tapers (cf. McElroy and Politis, 2009), one can utilize a plug-in estimator of $\beta$ instead; this is similar in spirit to the approach advocated in Robinson (2005).

These two techniques are described in more detail below, along with statistical justification. A new estimator of $\beta$, based on the rate estimation ideas of McElroy and Politis (2007), is also discussed. Then in the following subsection, both methods are applied to the study of size and power for testing the null hypothesis that $\mu = 0$. We look at Gaussian processes exhibiting LM, SM, or NM, at a variety of sample sizes and choices of taper.

### 5.1. Subsampling methodology for obtaining critical values

Firstly, we consider the subsampling method applied to the statistic $S_n/\sqrt{nW_{\Lambda,M}}$, or equivalently, $\overline{Y}/\sqrt{W_{\Lambda,M}/n}$. Letting $M = bn$ as usual with $b$ fixed and constant, under the assumptions of Theorem 4 we obtain the nondegenerate limit distribution $B(1)/\sqrt{Q(b)}$ for the LM/SM/NM case. The subsampling distribution estimator (sde) will be consistent for this limit distribution under a mixing condition, such as weak dependence (see discussion in Section 3). In Jach et al. (2012) the weak dependence condition is shown to be sufficient to establish consistency of subsampling distribution estimators for studentized statistics; Ango-Nze et al. (2003) was an earlier work on subsampling under weak dependence.

As for the sde itself, we first select a subsampling block size $an$, where $a$ is the subsampling-fraction; as usual, this is assumed to be vanishing, i.e., $a = a(n) \to 0$, though $an \to \infty$. Then $n - an + 1 = (1-a)n + 1$ contiguous overlapping blocks of the time series are constructed, and the statistics $\overline{Y}$ and $W_{\Lambda,bn}$ are evaluated on each block. This means a sample mean over $an$ random variables, and the corresponding tapered variance estimate based on this subsample, so that the bandwidth is actually $abn$ rather than $bn$. As a practical matter, unless $b$ and $a$ are taken fairly large, the bandwidth $abn$ becomes unmanageably small. The subsampled statistics can be collected into a set:

$$\left\{\frac{\overline{Y}_{an,i} - \overline{Y}_n}{\sqrt{W_{\Lambda,abn,i}/(an)}} \text{ for } i = 1, \ldots, n - an + 1\right\}.$$

Now taking the order statistics on this collection produces the quantiles of the sde. Further details of the construction can be found in Politis et al. (1999), but we here sketch the remaining theoretical details. First we note a general result that follows from Theorem 2 and Remark 5: if $M/n \to 0$ as $n \to \infty$, then

$$\frac{V_n/n^2}{\widetilde{W}_M/M} \to 0. \tag{16}$$

This is true for the LM/SM/NM case, since by Proposition 1 the limit is $O([M/n]^{1-\beta})$. Therefore, the probability that the $i$th subsample statistic exceeds a given $x$ is

$$\begin{aligned} &\mathbb{P}\left[\sqrt{an}\,\frac{\overline{Y}_{an,i} - \overline{Y}_n}{\sqrt{W_{\Lambda,abn,i}}} > x\right] \\ &= \mathbb{P}\left[\frac{S_{an} - an\,\mu}{\sqrt{anW_{\Lambda,abn}}} - b^{-1/2}\frac{S_n - n\mu}{\sqrt{V_n}}\sqrt{\frac{V_n/n^2}{W_{\Lambda,abn}/(abn)}} > x\right]. \end{aligned}$$

[2] Remark 5 shows that the normalization $MW_{\Lambda,M}$ fails for the LM/SM/NM cases under the vanishing bandwidth fraction asymptotics. In the SM case, one can use $nW_{\Lambda,M}$ as normalization instead, because then $W_{\Lambda,M} \sim \zeta f(0)$ and $V_n \sim nf(0)$, so that $nW_{\Lambda,M} \sim \zeta V_n$ (and $\zeta = \Lambda(0) - \Lambda(1)$, which is typically equal to 1). However, $nW_{\Lambda,M}$ still does not work for the LM/NM cases, because $nW_{\Lambda,M} \sim n\zeta L(M)M^{\beta}$ and $V_n \sim L(n)n^{\beta+1}/(\beta+1)$.

Using Theorems 1 and 4, as well as Proposition 3 and (16), the second term in the probability will tend to zero. Hence

$$\mathbb{P}\left[\frac{S_{an} - an\,\mu}{\sqrt{anW_{\Lambda,abn}}} > x\right] \to \mathbb{P}\left[\frac{B(1)}{\sqrt{Q(b)}} > x\right]$$

as $n \to \infty$, noting that $an \to \infty$ by assumption (and $b$ is fixed). Now, as long as the $i$th subsample statistic is approximately independent to the $j$th one when $|i-j|$ is large – which is implied by the weak dependence condition – the sde is consistent for the target limit distribution, and subsampling is valid. Note that we assume the fixed bandwidth-fraction condition for this result, but end up utilizing some of the vanishing bandwidth-fraction results, since the actual bandwidth-fraction for the subsampled tapered variance estimate is the vanishing quantity $ab$.

### 5.2. Plug-in methodology for obtaining critical values

Alternatively, one can use a plug-in approach as in McElroy and Politis (2012). This plug-in philosophy has an extensive history in econometrics, as summarized in Robinson (2005). The basic idea of that paper is to take Theorem 1 together with Eq. (6), and estimate the rate explicitly via a plug-in estimator of the memory parameter $\beta$; cf. (3.13) of Robinson (2005) and the Memory Autocorrelation Consistent (MAC) estimator. Because a vanishing bandwidth fraction approach is used in the MAC theory, the approach yields a normal limit as in our Theorem 1. This can be contrasted to the fixed bandwidth fraction approach, together with a tapered variance estimate to normalize instead by $n^{\widehat{\beta}+1}$ (as in the MAC case), which results in a more complicated limit distribution (Theorem 4) that depends on the true $\beta$. Either approach – that of MAC or that of McElroy and Politis (2012) – requires estimation of $\beta$, ultimately. The MAC approach does not allow for slowly-varying functions in Eq. (3) – these functions would need to be modeled separately – while the plug-in approach here requires prior calculation of limiting critical values by simulation (see below). So there are some pros and cons to each approach (this discussion is not exhaustive; e.g., there are subsampling approaches to rate estimation as well).

Adopting the fixed bandwidth-fraction asymptotics (so assume $b$ is constant throughout), and assuming that $\beta \in (-1, 1)$, we proceed to estimate the quantiles of the limit distribution $B(1)/\sqrt{Q(b)}$ via first estimating $\beta$ from the data, and then utilizing $x_\alpha(\widehat{\beta})$, where $x_\alpha(\beta)$ is the upper right $\alpha$ quantile of $B(1)/\sqrt{Q(b)}$. That is,

$$\mathbb{P}\left[\frac{B(1)}{\sqrt{Q(b)}} > x_\alpha(\beta)\right] = \alpha.$$

These quantiles have been tabulated for $\beta \in \mathbb{B} = \{-0.8, -0.6, -0.4, -0.2, 0, 0.2, 0.4, 0.6, 0.8\}$, three values of $\alpha$, several commonly used tapers, and all values of $b$ (via regression)—see the tables in McElroy and Politis (2009). Since the distribution is continuous in $\beta$, any consistent estimate $\widehat{\beta}$ can be utilized. Then one finds the member of $\mathbb{B}$ closest to the given $\widehat{\beta}$ – call this $\widetilde{\beta}$ – and utilizes $x_\alpha(\widetilde{\beta})$. This will be called the empirical plug-in method. Clearly, a finer mesh of simulation values for $\mathbb{B}$ would improve the procedure, but we may yet expect to obtain results superior to just using $\beta = 0$ in ignorance of the true memory. This latter approach, which essentially assumes that only short memory is present, will be referred to as the default plug-in method, and will be utilized as a benchmark for the empirical plug-in method.

Now many nonparametric estimators of $\beta$ can be utilized. Here we consider the simple estimator proposed in McElroy and Politis (2012), namely:

$$\widehat{\beta} = \frac{\log W_{\Lambda,bn}}{\log n}. \tag{17}$$

This is consistent for $\beta$ under the assumptions common to this paper.

**Table 1**
Empirical size for two-sided test with Type I error rate $\alpha = 0.05$ based on sample size $n = 250$, using the Bartlett kernel. In each cell, the first row is for sampling fraction $a = 0.2$, the second row for $a = 0.1$, and the third row for $a = 0.04$. Various values of the memory parameter $\beta$ are considered, as well as block sizes $b = 0.5, 1$.

| Empirical size for Bartlett taper | Beta | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|
| | −0.8 | −0.6 | −0.4 | −0.2 | 0 | 0.2 | 0.4 | 0.6 | 0.8 |
| Bartlett, $b = 0.5$ | | | | | | | | | |
| $a = 0.2$ | 90.6 | 89.8 | 83.2 | 85.3 | 84.7 | 82.6 | 76.8 | 76.6 | 66.8 |
| $a = 0.1$ | 95.2 | 92.9 | 90.5 | 92.3 | 89.9 | 90.9 | 88.2 | 84.8 | 81.2 |
| $a = 0.04$ | 100 | 99.1 | 98.7 | 97.1 | 94.7 | 96.2 | 95.0 | 90.4 | 81.9 |
| Bartlett, $b = 1$ | | | | | | | | | |
| $a = 0.2$ | 88.6 | 90.8 | 87.5 | 86.0 | 83.4 | 81.8 | 79.2 | 75.9 | 67.2 |
| $a = 0.1$ | 94.1 | 93.3 | 94.0 | 90.4 | 87.9 | 92.8 | 88.5 | 87.7 | 80.6 |
| $a = 0.04$ | 97.4 | 96.5 | 95.1 | 93.8 | 92.0 | 94.8 | 95.7 | 91.1 | 88.6 |

**Proposition 5.** *Assume that $b \in (0, 1]$ is fixed, as well as the hypotheses of Theorem 4. Then $\widehat{\beta}$ defined by (17) converges in probability to $\beta$.*

This estimator is similar in spirit to the tail index estimator of Meerschaert and Scheffler (1998), since it is based upon a convergence rate. The performance of $\widehat{\beta}$ can be poor in finite sample when long memory is present, but it is quite versatile and simple to implement.

### 5.3. Size and power of methods

We next evaluate the two methodologies – subsampling and empirical plug-in – through simulations. We adopt the perspective of testing a null hypothesis of $\mu = \mu_0$, since it is easier to evaluate the finite-sample properties through size and power, as opposed to the confidence interval perspective. In the simulations we take $\mu_0 = 0$. For each method, we consider values of $\mu$ between 0 and 1, with $\mu = 0$ corresponding to the null hypothesis. We compute the statistic $S_n/\sqrt{nW_{\Lambda,bn}}$ for a few different values of $b$, and five different tapers. The critical values of the limit distribution can be approximated using either of the two methods described above. Then we record the proportion of times that the statistic exceeds these critical values, using a two-sided test. Note that when $\mu > 0$, this assessment is interpreted as empirical power, but when $\mu = 0$ we obtain the empirical size.

The size of the plug-in approach has been partially addressed in McElroy and Politis (2009, 2012). However, the subsampling method's performance has not been previously studied, so we provide some additional material regarding its size. Table 1 displays size results for Type I error rates for a two-sided test with $\alpha = 0.05$. The sampling fraction for the subsampling method was selected at values $a = 0.2, 0.1, 0.04$, and the Bartlett taper with bandwidth fraction $b = 0.5$ and $b = 1$ was utilized. The sample size is fixed at $n = 250$. Results are for the empirical coverage, and so the target for the columns is the values 0.95. The results displayed in Table 1 were fairly typical for various tapers, block sizes, values of $\beta$, and differing $\alpha$—see McElroy and Politis (2011) for full results. Also increasing the sample size improved size slightly.

For data generation processes we focus on the Gaussian distribution, and consider simple white noise for the SM case (since serially correlated processes have been considered in previous literature, it suffices to take the simple white noise case); we consider four NM processes with $\beta = -0.2, -0.4, -0.6, -0.8$, where the autocovariance function is determined from Example 2 of Section 2. Also there are four LM processes, with $\beta = 0.2, 0.4, 0.6, 0.8$ and autocovariance function given in Example 1 of Section 2. We generated 1000 simulations of each specification.

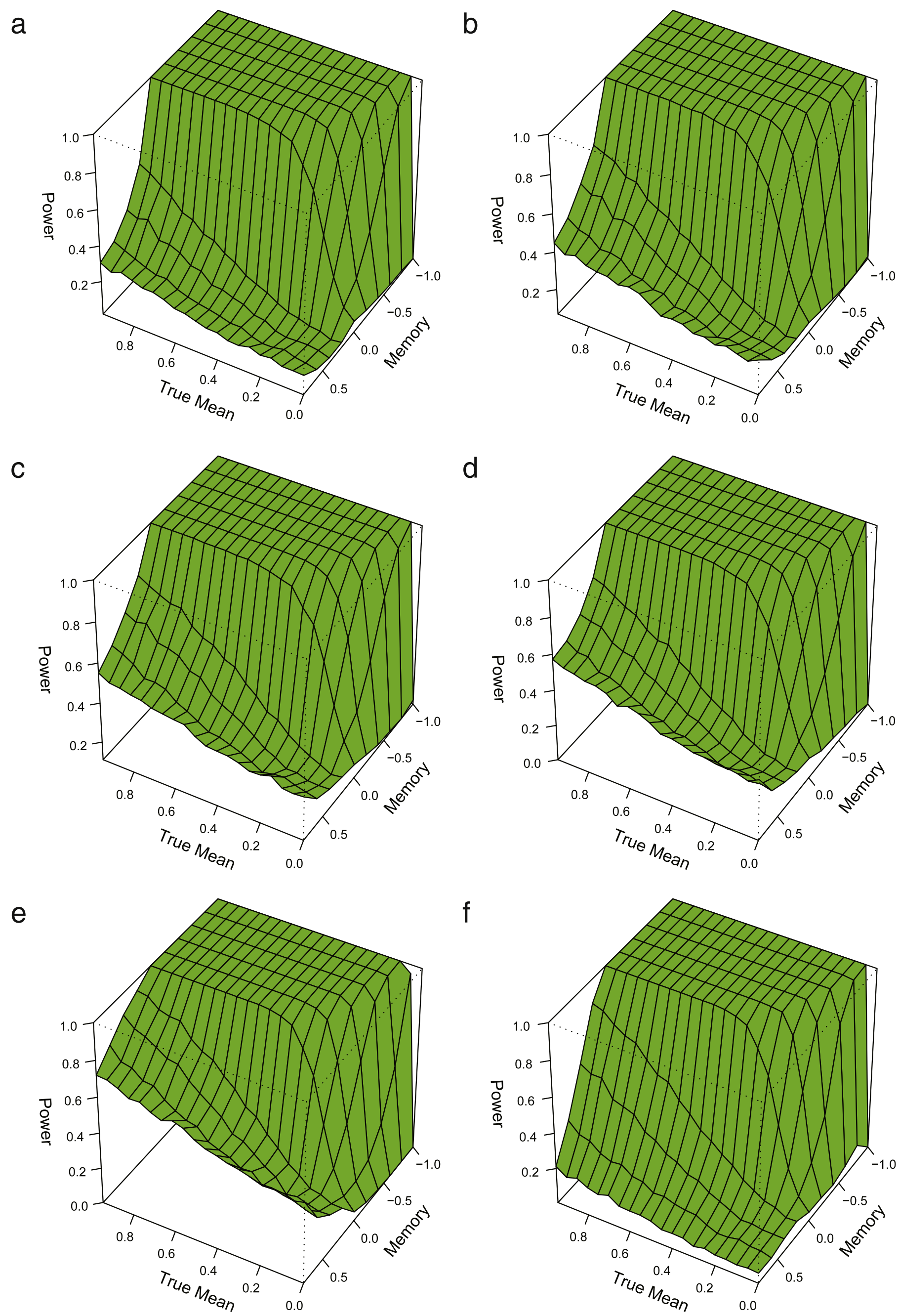


**Fig. 1.** Power surfaces by memory parameter $\beta \in \{-1, -0.8, -0.6, -0.4, -0.2, 0, 0.2, 0.4, 0.6., 0.8\}$ and true mean $\mu \in [0, 1)$. These power surfaces are for the Bartlett taper with bandwidth fraction $b = 0.5$. Panel (a) corresponds to subsampling with sampling fraction $a = 0.04$; panel (b) corresponds to subsampling with sampling fraction $a = 0.12$; panel (c) corresponds to subsampling with sampling fraction $a = 0.2$. Panel (d) corresponds to using the plug-in estimate of $\beta$; panel (e) corresponds to using the plug-in method with $\beta = 0$; panel (f) corresponds to using the true (unknown) value of $\beta$.

The power surfaces are organized a bit differently. Again we focus on one sample size $n = 250$ and the Type I error rate $\alpha = 0.05$ (for a two-sided test for the mean, so we use the upper one-sided critical value at 0.975 from the subsampling distributions and the tabulated values). Restricting to this $\alpha$ value gives the general behavior, and a reasonable sense of the power can be gleaned from the $n = 250$ case—higher sample sizes tend to shift the contours upwards (not dramatically for high $\beta$), but the overall shape is the same. We consider the Bartlett taper with bandwidth fractions $b = 0.5, 1$ and the sampling fractions $a = 0.04, 0.12, 0.2$. These choices are convenient, as it is always guaranteed that $abn$ is an integer. The range of $\mu$ was chosen so as to capture the main qualitative features of the power surface across all data processes: $\mu \in \{j/20\}_{j=0}^{19}$ proved to yield power close to 50% for the long memory processes, while being small enough to allow visual discrimination of cases.

The six methods are placed in Figs. 1 and 2 as sub-panels. Moving from top left to bottom right, the first three methods correspond to subsampling with various sampling fractions. Then we have the empirical plug-in method, followed by the default plug-in (which uses $\beta = 0$ critical values throughout) method. The final panel is an omniscient plug-in, based on knowing the true value of $\beta$ (so it is not a practicable method, but is helpful for understanding power). The discrepancy in power between this and the empirical plug-in method (middle right panels) is mainly due to error in our estimator of $\beta$.

Now we discuss these numerical results. The size results for the SM case are fairly standard, being adequate at sample size $n = 250$

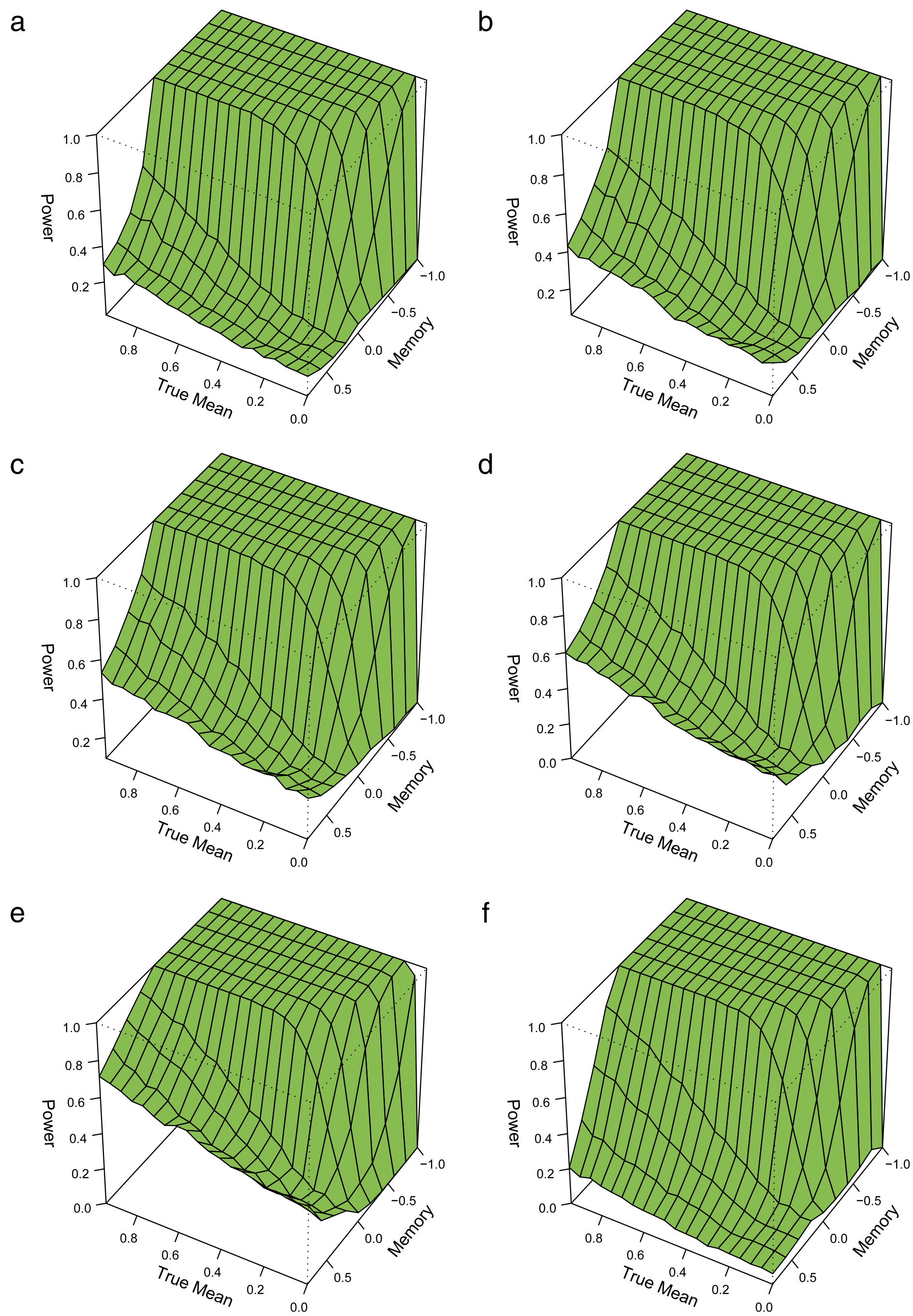


**Fig. 2.** Power surfaces by memory parameter $\beta \in \{-1, -0.8, -0.6, -0.4, -0.2, 0, 0.2, 0.4, 0.6., 0.8\}$ and true mean $\mu \in [0, 1)$. These power surfaces are for the Bartlett taper with bandwidth fraction $b = 1$. Panel (a) corresponds to subsampling with sampling fraction $a = 0.04$; panel (b) corresponds to subsampling with sampling fraction $a = 0.12$; panel (c) corresponds to subsampling with sampling fraction $a = 0.2$. Panel (d) corresponds to using the plug-in estimate of $\beta$; panel (e) corresponds to using the plug-in method with $\beta = 0$; panel (f) corresponds to using the true (unknown) value of $\beta$.

and greater; the higher bandwidth fractions gave slightly better results. For NM the coverage improved for higher values of $\beta$, while the results for the LM case were much worse, with poor coverage at $\beta = 0.2$; higher bandwidths seemed to improve performance slightly. In summary, if there is NM or SM, then one should use a small or moderate size bandwidth fraction, such as $b = 0.1$ or $b = 0.05$. If LM is present, a larger bandwidth is preferable.

If a statistic rejects too often under the null hypothesis – as is seen to happen in the tables for the subsampling methods – then it is liable to have higher power than otherwise. The fact that all methods tend to be mis-sized is evident in the surface plots in Figs. 1 and 2 by examination of the $\mu = 0$ cross-sectional curve towards the right side of the surface. But as $\mu$ increases, the NM processes generate high power relatively quickly, giving a mesa shape to the surfaces. For SM and weaker LM, the rise to full power is slower. An ironic feature is that when $\mu$ is quite low, the power for strong LM is better than for weaker LM, essentially due to the methods being over-sized. This is seen in the "ruffle" feature of the curves along the $\beta = 0.8$ cross-section.[3] Since power approaches 50% for all processes as $\mu$ increases to unity, there is the question

[3] Some authors prefer to investigate what the power would be were these statistics to be adjusted to be correctly sized; however, in practice such a procedure is impossible to implement on real data. We have chosen to display the power that would occur were a practitioner to utilize any of the methods on real data.

**Table 2**
Optimal bandwidth fractions $b$ as a function of taper and memory parameter $\beta$. Each row corresponds to either the (upper one-sided) 0.90,0.95, 0.975, or 0.99 quantile.

| Optimal Bandwidth fraction | Beta | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|
| | −0.8 | −0.6 | −0.4 | −0.2 | 0 | 0.2 | 0.4 | 0.6 | 0.8 |
| Bartlett | | | | | | | | | |
| 0.90 | 0.02 | 0.02 | 0.02 | 0.02 | 0.02 | 0.10 | 0.14 | 0.14 | 0.18 |
| 0.95 | 0.02 | 0.02 | 0.02 | 0.02 | 0.02 | 0.06 | 0.10 | 0.14 | 0.18 |
| 0.975 | 0.02 | 0.02 | 0.02 | 0.02 | 0.02 | 0.06 | 0.10 | 0.14 | 0.14 |
| 0.99 | 0.02 | 0.02 | 0.02 | 0.02 | 0.02 | 0.04 | 0.10 | 0.14 | 0.14 |
| Trapezoid (0.25) | | | | | | | | | |
| 0.90 | 0.02 | 0.02 | 0.02 | 0.02 | 0.02 | 0.06 | 0.12 | 0.14 | 0.14 |
| 0.95 | 0.02 | 0.02 | 0.02 | 0.02 | 0.02 | 0.06 | 0.10 | 0.12 | 0.14 |
| 0.975 | 0.02 | 0.02 | 0.02 | 0.02 | 0.02 | 0.04 | 0.08 | 0.14 | 0.14 |
| 0.99 | 0.02 | 0.02 | 0.02 | 0.02 | 0.02 | 0.04 | 0.08 | 0.08 | 0.10 |
| Trapezoid (0.50) | | | | | | | | | |
| 0.90 | 0.02 | 0.02 | 0.02 | 0.02 | 0.02 | 0.04 | 0.10 | 0.12 | 0.16 |
| 0.95 | 0.02 | 0.02 | 0.02 | 0.02 | 0.02 | 0.04 | 0.10 | 0.10 | 0.12 |
| 0.975 | 0.02 | 0.02 | 0.02 | 0.02 | 0.02 | 0.04 | 0.06 | 0.10 | 0.12 |
| 0.99 | 0.02 | 0.02 | 0.02 | 0.02 | 0.02 | 0.02 | 0.04 | 0.06 | 0.06 |
| Parzen | | | | | | | | | |
| 0.90 | 0.02 | 0.02 | 0.02 | 0.02 | 0.02 | 0.12 | 0.20 | 0.22 | 0.28 |
| 0.95 | 0.02 | 0.02 | 0.02 | 0.02 | 0.02 | 0.06 | 0.16 | 0.20 | 0.20 |
| 0.975 | 0.02 | 0.02 | 0.02 | 0.02 | 0.02 | 0.06 | 0.10 | 0.22 | 0.20 |
| 0.99 | 0.02 | 0.02 | 0.02 | 0.02 | 0.02 | 0.06 | 0.08 | 0.14 | 0.20 |
| Daniell | | | | | | | | | |
| 0.90 | 0.02 | 0.02 | 0.02 | 0.02 | 0.02 | 0.06 | 0.10 | 0.16 | 0.18 |
| 0.95 | 0.02 | 0.02 | 0.02 | 0.02 | 0.02 | 0.06 | 0.10 | 0.16 | 0.14 |
| 0.975 | 0.02 | 0.02 | 0.02 | 0.02 | 0.02 | 0.06 | 0.10 | 0.08 | 0.14 |
| 0.99 | 0.02 | 0.02 | 0.02 | 0.02 | 0.02 | 0.04 | 0.06 | 0.08 | 0.10 |

of how this is meaningful relative to the variation in the process. All were constructed with $\gamma_0 = 1$ (an alternative way to normalize is to set the innovation variances equal, by in each case dividing through by the square root of the integral of the log spectrum), so the coefficient of variation is $1/\mu$ for all processes. Primarily, we view these figures as a way to contrast the power of methods and tapers.

For a given taper, there is moreover the issue of selecting the best possible bandwidth fraction. If there is only a minor discrepancy between the true mean and the null mean, then higher power will be associated with smaller (upper one-sided) critical values. Leaving aside the issue of size, we can maximize power in the worst possible scenario (i.e., where the true mean $\mu$ is approximately equal to the null mean) by finding $b$ such that the quantile is minimized. Table 2 provides such bandwidth fractions for each given taper (we consider the Bartlett, Parzen, Daniell, and two Trapezoidal tapers), as a function of the true $\beta$, for a variety of $\alpha$ values (0.10, 0.05, 0.025, and 0.01). For NM and SM a small bandwidth is best, but larger bandwidths – between 0.02 and 0.28 – provide better asymptotic power for the LM case. This result also tends to parallel the size results, in that large bandwidths should be avoided in the NM and SM cases, while being more appropriate for the LM case.

## 6. Discussion

This paper sets out a thorough study of self-normalized mean estimation when long memory or negative memory is present. The main statistic of interest is the sample mean of a stationary time series (appropriately differenced beforehand), normalized by a tapered sum of sample autocovariances. The behavior of autocovariances changes greatly depending on whether a time series has long range dependence, anti-persistence, or short memory. This in turn has a large impact on the convergence rates of sample mean and tapered autocovariances. We provide a unified treatment of the various types of memory, including the important SuperEfficient (SE) case, wherein the partial sums are $o_P(\sqrt{n})$. The SE scenario is important, since it can easily arise from over-differencing of a time series suspected of having trend nonstationarity.

Several novel results on the memory of a time series are presented, which – together with examples – furnish some intuition for the qualitative behavior. The properties of Differential Memory (DM) processes are elucidated, and shown to be distinct from the behavior of Negative Memory (NM) processes – together, the DM and NM cases partition the important SE case. But our main interest is in the asymptotics of sample mean and tapered autocovariances, and we treat these topics through several theorems. For the asymptotic results we consider both the vanishing bandwidth-fraction case (a more classical approach, going back to Parzen (1957)) and the fixed bandwidth-fraction case (a more recent approach espoused in Kiefer et al. (2000)). We both summarize known results, and prove new ones, examining two broad classes of data process that exhibit the various types of memory described herein.

In order to make use of the asymptotic results, it is still necessary to get the critical values of the limiting distributions, which in the fixed bandwidth-fraction case are functionals of the Fractional Brownian Bridge. We propose two methodologies: subsampling, which avoids explicit estimation of the memory parameter $\beta$ but requires selection of a sampling fraction $a$; and the plug-in approach, which requires an estimate of $\beta$ and a look-up table of critical values (computed ahead of time via simulation) for the limiting distributions.[4] These methods are compared through extensive finite-sample size and power simulations, which are succinctly summarized here.[5] While power tends to deteriorate with greater memory, Type II error is comparatively quite small with anti-persistent processes. There are size problems with the plug-in method, whereas in contrast the subsampling method tends to have superior coverage (i.e., empirical size is closer to the nominal level).

One outstanding problem in the literature on this topic is the question of bandwidth selection. If one seeks an optimal bandwidth, it must be selected to minimize a pertinent criterion. Recent work by Sun et al. (2008) provides an attractive approach based on examining both type I and type II errors. However, their paper is focused on the SM case; any attempt to mimic their procedure makes the bandwidth dependent on $\beta$, which then must be estimated. Also, higher-level assumptions are typically required for such an analysis, which is at war with the fairly generic assumptions of this paper.

What is really needed in practice is a prescription for applied statisticians. Choice of taper, bandwidth, and subsampling fraction are all to be determined by the user, and the optimal combination of such depends on the unknown data process. Moreover, any such optimality is typically derived using asymptotic criteria, whose impact on finite samples is less clear; also see Table 2. However, some of the lessons from our simulations can be repeated here: asymptotic critical values vary more widely with respect to $\beta$ when $b$ is low; flat-top tapers and smooth tapers (e.g., the Parzen taper) perform adequately in terms of size and power, except in the LM case; size and power can depend substantially on the subsampling fraction. Note that the subsampling fraction might be selected using the technique of Bickel and Sakov (2008), as described in Jach et al. (2012). Or one might utilize the plug-in approach described in Section 5, which requires simulation of quantiles ahead of time.

4 While the MAC approach of Robinson (2005) is also applicable, we omit to study it here because it is not based upon self-normalization, which is the focus of our paper.

5 All code and results are available from the first author; see McElroy and Politis (2011) for full numerical results.

For any of the methods discussed, neither computer programming nor computation time is burdensome (unless quantiles are being simulated). Therefore, one can generate results for a multitude of tapers and bandwidths, and include all outcomes as alternative explanations of the data. However, if for some reason the practitioner is unable to produce a spectrum of results, we recommend utilizing a small bandwidth fraction (say $b$ close to 0.02) when it is judged that only NM or SM is present, but a bandwidth fraction closer to 0.1 or 0.2 if LM is suspected. To obtain decent performance over a range of data processes, either the Parzen or a trapezoidal taper is recommended.

In summary, we provide a viable framework for conducting inference for the mean, supplying a unified asymptotic theory that covers all different types of memory under a single umbrella. This framework is robust against different memory specifications, obviating the need to do extensive modeling. Future work should examine the bandwidth selection problem, using these theoretical results as groundwork, as well as documenting the empirical performances of competing methodologies.

## Acknowledgments

The authors thank the associate editor and two anonymous referees for helpful comments.

The second author's research was partially supported by NSF grants DMS-07-06732 and DMS-10-07513.

*Disclaimer*. This report is released to inform interested parties of research and to encourage discussion. The views expressed on statistical issues are those of the authors and not necessarily those of the U.S. Census Bureau.

## Appendix

**Proof of Proposition 1.** First consider cases (i), (ii), and (iii) with $\beta \in (-1, 1)$. By Bingham et al. (1987), a slowly varying function satisfies, for any $\delta > 0$, $|L(k)/L(n)| \leq \tilde{C}(|k/n|^{\delta} + |n/k|^{\delta})$ for $1 \leq k \leq n$ and some constant $\tilde{C} > 0$. Then by the Dominated Convergence Theorem

$$\frac{V_n}{nW_n} = n^{-1}\sum_{k=0}^{n-1}\frac{W_k}{W_n} = n^{-1}\sum_{k=0}^{n-1}\frac{L(k)}{L(n)}(k/n)^{\beta} \to \int_0^1 x^{\beta}\,dx = \frac{1}{\beta+1}.$$

In case (iii) when $\beta = -1$, we have $V_n/(nW_n) \sim \sum_{h<n} h^{-1}L(h)/L(n) \to \infty$ by Theorem A3.6 of Embrechts et al. (1997). Case (iv) is immediate from its definition. For (7), in cases (i) and (iii) with $\beta \in [-1, 1) \setminus \{0\}$ we have

$$\gamma_k = \frac{1}{2}(W_k - W_{k-1}) = \frac{1}{2}\left[(L(k) - L(k-1))k^{\beta}\right] + \frac{1}{2}\left[L(k-1)(k^{\beta} - (k-1)^{\beta})\right].$$

Dividing through by $L(k)k^{\beta-1}/2$ yields

$$2\,\frac{\gamma_k}{L(k)k^{\beta-1}} = [(1 - L(k-1)/L(k))k] + \left[\frac{kL(k-1)}{L(k)}(1-(1-1/k)^{\beta})\right].$$

The second expression on the right hand side is asymptotic to $\beta$ via Taylor series. The first expression on the right hand side is asymptotic to $-kg(k)$, where $g(x) = \int_{x-1}^{x}\eta(u)/u\,du$ by expanding the exponential term in the slowly-varying function $L$. But $kg(k) = o(1)$ as $k \to \infty$. So when $\beta \neq 0$, the acf is asymptotic to $\beta L(k)k^{\beta-1}/2$. But when $\beta = 0$ we have instead

$$\frac{\gamma_k}{L(k)} = \frac{1}{2}(1 - L(k-1)/L(k)) \sim -g_k/2$$

by the previous analysis, which is $o(k^{-1})$. □

**Proof of Proposition 2.** Without loss of generality, suppose that $\mu = 0$. By the Wold decomposition (see Brockwell and Davis, 1991) there exists an uncorrelated sequence $\{\epsilon_t\}$ and square summable coefficients $\{\psi_j\}$ such that $Y_t = \sum_{j\geq 0}\psi_j\epsilon_{t-j}$. Let $\sigma^2 = \mathrm{Var}(\epsilon_t)$ be equal to one for simplicity. We have $0 = \lim_{k\to\infty} W_k = \sum_h \gamma_h = (\sum_j \psi_j)^2\sigma^2$, so $\sum_j \psi_j = 0$. It follows that $1 - z$ should divide $\Psi(z) = \sum_{j\geq 0}\psi_j z^j$, though we must establish that $\Theta(z) = \Psi(z)/(1-z)$ converges. Extending $\psi_j$ to be zero if $j < 0$, we define $\theta_j = \sum_{k=-\infty}^{j}\psi_k$ for any integer $j$. Note that $\theta_j \to 0$ as $j \to \infty$ and is zero if $j < 0$. We will show that the $\theta_j$ sequence is square summable, so that $\sum_j \theta_j\epsilon_{t-j}$ is finite with probability one; then this will define $Z_t$, from which $Y_t = Z_t - Z_{t-1}$ follows at once. Note that $\theta_j = \sum_{k\geq 0}\psi_{j-k}$; let $\theta_{j,m} = \sum_{k=0}^{m}\psi_{j-k}$. Then $\sum_j \theta_{j,m}^2 = \sum_{i,k=0}^{m}\gamma_{i-k} = \sum_{k=0}^{m} W_k$ for each $m$. Then by Fatou's lemma

$$\sum_j \theta_j^2 \leq \lim_{n\to\infty}\sum_j \theta_{j,m}^2 = \sum_{k\geq 0} W_k < \infty.$$

This establishes the existence of a series $\{Z_t\}$ with the required properties. However, the series given by $\sum_j \theta_j\epsilon_{t-j} + A$, where $A$ is uncorrelated with the series $\{\epsilon_t\}$, also has the requisite properties, as the temporal difference will be the same as that of $Z_t - Z_{t-1}$. Although an additional term of $\mathrm{Var}(A)$ is added on to the acf, this will cancel out in $\gamma_k = 2r_k - r_{k+1} - r_{k-1}$. So the series $\{Z_t\}$ cannot be determined uniquely.

Next, we note that the formula for $W_k$ follows from summing the $\gamma_k$ via (3); the formula for $r_0$ is obtained by summing the formula for $W_k$, and using the property of telescoping sums. The assertions of the converse are now immediate. □

**Proof of Theorem 1.** From Proposition 1 we can utilize (6). Also without loss of generality we can take $\mu = 0$ in the proof. Let us first consider that the process satisfies P1, so that $Y_t = \sum_j \psi_j\epsilon_{t-j}$ for an *i.i.d.* sequence $\{\epsilon_t\}$, and by assumption $\gamma_0 \propto \sum_j \psi_j^2$ exists. So Theorem 5.2.3 of Taniguchi and Kakizawa (2000) (which is due to Ibragimov and Linnik (1971, p. 359)) gives the result. Note that by Hosoya (1996), we could relax the independence assumption on $\{\epsilon_t\}$ to a type of mixing condition when $\beta = 0$.

Next, consider P2. The $k$-fold cumulant of $S_n$ is

$$\mathrm{cum}\,(S_n, \ldots, S_n) = \sum_{t_k=1}^{n}\sum_{t_1=1}^{n}\cdots\sum_{t_{k-1}=1}^{n} c_k(t_1 - t_k, t_2 - t_k, \ldots, t_{k-1} - t_k),$$

which is $O(n \cdot n^{\beta\wedge 0})$ when $k > 2$. Letting $\kappa_{k,n}$ denote the $k$th order cumulant of $S_nV_n^{-1/2}$, we obtain $\kappa_{k,n} = O(n^{(\beta+1)\wedge 1}\cdot V_n^{-k/2})$, which tends to zero as $n \to \infty$ if $k > 2$. Of course $\kappa_{1,n} = 0$ and $\kappa_{2,n} = 1$, and the limit theorem is proved by expanding the characteristic function of $S_nV_n^{-1/2}$ as in Eq. (2.19) of Hall (1992).

Finally, we turn to the DM case. Applying Proposition 2, we have $S_n = Z_n - Z_0$. If $\{Z_t\}$ is strongly mixing, let $\tilde{Z}_n$ be equal in distribution to $Z_n$ for every $n$, but independent of $Z_0$. Then

$$\left|\mathbb{E}\exp\{i\nu(Z_n - Z_0)\} - \mathbb{E}\exp\{i\nu(\tilde{Z}_n - Z_0)\}\right| \leq 16\,\alpha_n$$

by Lemma B.0.6 of Politis et al. (1999), which is due to Ibragimov (1962). Now $\phi_{\tilde{Z}_n - Z_0}(\nu) = \phi_Z(\nu)\cdot\phi_Z(-\nu)$ where $\phi$ denotes the characteristic function, and $Z$ has the common distribution of the $\{Z_t\}$ process.

If $\{Z_t\}$ is linear, then $\{\epsilon_t\}$ are *i.i.d.* in the representation $Z_t = \sum_{j\geq 0}\theta_j\epsilon_{t-j}$, with the convention that $\theta_j = 0$ if $j < 0$. Then

$$S_n = \sum_j (\theta_{n-j} - \theta_{-j})\epsilon_j = \sum_{j=1}^{n} \theta_{n-j}\epsilon_j + \sum_{j\leq 0}(\theta_{n-j} - \theta_{-j})\epsilon_j.$$

These two terms are independent, and the first is equal in distribution to $\sum_{j=0}^{n-1}\theta_j\epsilon_j$, which tends in probability (and thus weakly) to $\sum_{j\geq 0}\theta_j\epsilon_j$ by Theorem 22.6 of Billingsley (1995). This limit is equal in distribution to $Z$, so $\sum_{j=1}^{n}\theta_{n-j}\epsilon_j \stackrel{\mathcal{L}}{\Longrightarrow} Z$. For the second term, we have $\sum_{j\leq 0}\theta_{n-j}\epsilon_j \stackrel{P}{\longrightarrow} 0$, since its variance is equal to $\sum_{j\geq n}\theta_j^2\sigma^2$. What is left is $-\sum_{j\leq 0}\theta_{-j}\epsilon_j$, which is equal to $-Z$ in distribution. Since the two terms are independent, we have that $S_n$ converges weakly to the difference of two independent random variables, each with the distribution of $Z$. □

**Proof of Proposition 3.** The expression $\widetilde{\gamma}_h - \gamma_h$ is unchanged if we replace $Y_t$ by $Y_t - \mu$, so without loss of generality suppose that $\mu = 0$. We begin by decomposing $W_{\Lambda,M}$ into

$$W_{\Lambda,M} = \widetilde{W}_M + E_M^{(1)} + E_M^{(2)} \tag{A.1}$$
$$E_M^{(1)} = \sum_h \Lambda_M(h)\left(\widetilde{\gamma}_h - \overline{\gamma}_h\right)$$
$$E_M^{(2)} = \sum_h \Lambda_M(h)\left(\overline{\gamma}_h - \gamma_h\right).$$

Here $\overline{\gamma}_k = \frac{1}{n}\sum_{t=1}^{n-|k|}(Y_t - \mu)(Y_{t+k} - \mu)$ for $k \geq 0$, and recall $\widetilde{\gamma}_k$ is given by (2). The terms $E_M^{(1)}$ and $E_M^{(2)}$ are stochastic errors, whereas $\widetilde{W}_M$ is deterministic and serves as an approximation to $W_M$. We will show that the error terms $E_M^{(1)}$ and $E_M^{(2)}$ of (A.1) satisfy $E_M^{(1)} = O_P(U_M MV_n n^{-2})$ and $E_M^{(2)} = O_P(U_M Mn^{-\eta}K(n))$ as $n \to \infty$, where $K$ is slowly-varying and $\eta = 1/2$ if $\beta < 1/2$ and $\eta = 1-\beta$ if $\beta \geq 1/2$ (in the LM case); then the statement of the proposition follows. For each $h \geq 0$

$$\begin{aligned}\widetilde{\gamma}_h &= \gamma_h + (\overline{\gamma}_h - \gamma_h) + (\tilde{\gamma}_h - \overline{\gamma}_h)\\ &= \left(1 - \frac{h}{n}\right)\gamma_h + \frac{1}{n}\sum_{t=1}^{n-h}(Y_tY_{t+h} - \gamma_h)\\ &\quad + \overline{Y}\left(\overline{Y} - \overline{Y}_{1:n-h} - \overline{Y}_{h+1:n}\right),\end{aligned}$$

where $\overline{Y}_{1:n-h} = \sum_{t=1}^{n-h} Y_t/n$ and $\overline{Y}_{h+1:n} = \sum_{t=h+1}^{n} Y_t/n$. Let $c_n = V_n n^{-2}$. Then $\overline{Y} = O_P(n^{-1}V_n^{-1/2})$, so the third term above is $O_P(c_n)$ uniformly in $h$ (this is true since $|h| \leq M = o(n)$). Now $c_n = n^{\beta-1}L(n)$ in the LM and NM cases, with $\beta \in [-1, 1)$, but $c_n = n^{-1}$ for the SM case and $c_n = n^{-2}$ for the DM case. Using the $U_M$ bound on $\Lambda_M(h)$, we obtain $E_M^{(1)} = O_P(U_M Mc_n)$ as claimed. For the error term $E_M^{(2)}$, we compute

$$\begin{aligned}\mathrm{Var}(E_M^{(2)}) &\leq n^{-2}\sum_{h,k}\Lambda_M(h)\Lambda_M(k)\sum_{t,s=1}^{n}\mathrm{Cov}(Y_tY_{t+h}, Y_sY_{s+k})\\ &\leq n^{-2}\sum_{h,k}\Lambda_M(h)\Lambda_M(k)\sum_{t,s=1}^{n}\big(\mathrm{cum}(Y_t, Y_{t+h}, Y_s, Y_{s+k})\\ &\quad + \gamma_{t-s}\gamma_{t-s+h-k} + \gamma_{t-s-k}\gamma_{t+h-s}\big)\\ &\leq n^{-1}\sum_{h,k}\Lambda_M(h)\Lambda_M(k)\sum_{|l|\leq n}(1 - |l|/n)\\ &\quad \times\big(\mathrm{cum}(Y_0, Y_h, Y_l, Y_{l+k}) + \gamma_l\gamma_{l-h+k} + \gamma_{l+k}\gamma_{h-l}\big).\end{aligned}$$

The inequality used here only concerns $|h|$ terms, which is negligible with respect to $n$. The sum of the cumulant function is bounded using Assumption B, resulting in an overall bounds of $U_M^2/n$. For the sum over $\gamma_l\gamma_{l-h+k}$, we can use the Cauchy–Schwarz inequality to obtain the bound $O(M^2U_M^2 n^{\xi-1}K(n))$, where $K$ is a slowly-varying function and $n^{\xi}K(n)$ represents the order of $\sum_{|l|\leq n}\gamma_l^2$; using Proposition 1 we have $\xi = 0$ if $\beta < 1/2$ (as this results in a square summable sequence) or $\xi = 2\beta - 1$ if $\beta > 1/2$. When $\beta = 1/2$ and $L(n) \equiv 1$, we take $\xi = 0$ and $K(n) = \log n$. The analysis of the sum over $\gamma_{l+k}\gamma_{h-l}$ yields a similar order. Taking square roots, we learn that $E_M^{(2)} = O_P(MU_M n^{-\eta}K^{1/2}(n))$ as asserted. □

**Proof of Theorem 2.** First note that we can remove the term $(1 - |h|/n)$ since $M/n \to 0$. The case of the truncation filter is trivial. Case (iv) is treated differently, so we first consider cases (i), (ii), and (iii). We proceed to break the sum over $h$ up according to the intervals of smoothness for $\Lambda$. The first such interval is $[-c, c]$, which corresponds to the flat-top interval; if there is no flat-top region, then $c = 0$. In general, consider an interval $(r, s]$ such that the restriction of $\Lambda$ is continuously differentiable there. Then

$$\begin{aligned}&\sum_{[rM]<|h|\leq[sM]}\Lambda\left(\frac{h}{M}\right)\gamma_h\\ &= \Lambda\left(\frac{[sM]}{M}\right)W_{[sM]} - \Lambda\left(\frac{[rM]}{M}\right)W_{[rM]}\\ &\quad + \sum_{h=[rM]}^{[sM]-1}\left[\Lambda\left(\frac{h}{M}\right) - \Lambda\left(\frac{h+1}{M}\right)\right]W_h\end{aligned} \tag{A.2}$$

via summation by parts. The first two terms on the right hand side will cancel with other like terms for the other intervals, leaving the last term on the right hand side. The terms in the square brackets consist of values of $\Lambda$ restricted to $(r, s]$, excepting only the first term $h = [rM]$; however, since $\Lambda([rM]/M) - \Lambda(([rM]+1)/M) \approxeq -M^{-1}\dot{\Lambda}_+(r)$ by continuity, the first term's analysis is the same as the others. In general, $\Lambda(h/M) - \Lambda(h+1/M) = -\dot{\Lambda}(h/M)M^{-1} + O(M^{-2})$. The case of $r > 0$ for an interval $(r, s]$ has a different analysis from the $r = 0$ case, which we consider later. So long as $r > 0$, we have $h \to \infty$ as $M \to \infty$ in the above summation. Thus in case (ii), $W_h \to W_\infty$ and this convergence occurs uniformly in $h \in (rM, sM)$ as $M \to \infty$. Using the boundedness of $\dot{\Lambda}(x)$ and the limit of a Cesaro sum, we obtain a limit of $-W_\infty\int_r^s \dot{\Lambda}(x)\,dx$. The argument can be extended to cases (i) and (iii) as follows:

$$\begin{aligned}&\frac{\sum_{h=[rM]}^{[sM]-1}\left[\Lambda\left(\frac{h}{M}\right) - \Lambda\left(\frac{h+1}{M}\right)\right]W_h}{CM^{\beta}L(M)}\\ &= \sum_{h=[rM]}^{[sM]-1}\left\{\Lambda\left(\frac{h}{M}\right) - \Lambda\left(\frac{h+1}{M}\right)\right\}(h/M)^{\beta}\left[W_h h^{-\beta}/CL(M)\right],\end{aligned}$$

and the expression in square brackets equals one plus error tending to zero as $M \to \infty$, uniformly in $h \in ([rM], [sM]]$. This is because

$$\left|\frac{W_h}{Ch^{\beta}L(M)} - 1\right| \leq \left|\frac{W_h}{Ch^{\beta}L(h)} - 1\right| + \left|\frac{W_h}{Ch^{\beta}L(h)}\left(\frac{L(h)}{L(M)} - 1\right)\right|.$$

The first term tends to zero uniformly in $h$ as $M \to \infty$. For the second, we have $L(h)/L(M) \to 1$ uniformly in $h$ as $M \to \infty$ as well, which is seen by using the representation (5). Note that these arguments hold for $\beta = 0$. Using the boundedness of $x^{\beta}$ for $x \in (r, s]$ and $r > 0$, we obtain a limit of $-\int_r^s \dot{\Lambda}(x)x^{\beta}\,dx$ as $M \to \infty$. This argument works for any $\beta \in (-1, 1)$.

The first interval must be treated differently—unless it is flat-top, i.e., $c > 0$, in which case it is trivially given by $W_{[cM]}$, which cancels with boundary term in the next intervals. More generally, the first interval has the form $\sum_{0\leq|h|\leq[sM]}\Lambda(h/M)\gamma_h$, which tends to $\Lambda(0)W_\infty$ in case (ii). Otherwise in cases (i) and (iii) we have

$$\sum_{|h|=0}^{[sM]}\Lambda\left(\frac{h}{M}\right)\gamma_h = \Lambda(0)\,W_{[sM]} + 2\sum_{h=1}^{[sM]}\left[\Lambda\left(\frac{h}{M}\right) - \Lambda(0)\right]\gamma_h.$$

Note that $\beta > -1$ by assumption. The expression in brackets can be expanded in the Taylor series $\sum_{j\geq 1} \Lambda^{(j)}(0) h^j M^{-j}/j!$. Noting that $\sum_{h=1}^{M} h^j \gamma_h$ is divergent and asymptotic to $\beta M^j W_M / 2(\beta + j)$ (proved by l'Hopital's rule and (7)) for any $j \geq 1$, we can interchange summations to obtain

$$W_{[sM]} + \beta\, W_{[sM]} \sum_{j\geq 1} \Lambda^{(j)}(0) \frac{s^j}{\beta + j}$$
$$\sim W_M \left( s^\beta + \beta \int_0^s (\Lambda(x) - 1) x^{\beta - 1}\, dx \right)$$

after some algebra. Now piecing all intervals together, taking into account cancellations and integration by parts, we arrive at (10).

Now we turn to case (iv). Letting the partition $0 = s_0 < s_1 < \cdots s_T < s_{T+1} = 1$ denote the points of non-differentiability in $\Lambda$, we can write

$$\sum_{|h|=0}^{[sM]} \Lambda\left(\frac{h}{M}\right) \gamma_h = \sum_{j=0}^{T} \sum_{h=[s_j M]}^{[s_{j+1}M]-1} \left[ \Lambda\left(\frac{h}{M}\right) - \Lambda\left(\frac{h+1}{M}\right) \right] W_h.$$

The first term is identically zero for a flat-top kernel; otherwise when $c = 0$, we can use a Taylor series expansion at $h/M$ again. Since $\{W_h\}$ is a summable sequence, we can apply the Dominated Convergence Theorem and obtain $2r_0$ times the Taylor series at zero, as stated in the theorem. As for the other intervals, note that $\sum_{h=[s_j M]}^{[s_{j+1}M]-1} W_h = 2(r_{[s_j M]} - r_{[s_{j+1}M]})$ is tending to zero as $M \to \infty$, since $j \geq 1$. Thus these other terms decay even faster than the first term. Hence for flat-top tapers, the rate of convergence is $o(M^{-1})$. □

**Proof of Theorem 3.** The proof proceeds by first showing that the finite-dimensional distributions of $\xi_n$ converge to those of FBM. Then we show tightness. We will now show that $S_n$ and $\xi_n$ are asymptotically equivalent processes, i.e., any linear combination over any set of times of their difference tends to zero in probability. Because $\xi_n(t) - S_n(t) = (nt - [nt]) Y_{[nt]+1}$, clearly for every $\epsilon > 0$ and any collection of times $t_1, \ldots, t_k$ and constants $\alpha_1, \ldots, \alpha_k$ (for any $k \geq 1$),

$$\mathbb{P}\left[ \left| \sum_{j=1}^{k} \alpha_j \left( \xi_n(t_j) - S_n(t_j) \right) V_n^{-1/2} \right| > \epsilon \right] \to 0$$

as $n \to \infty$. This follows from (6). Hence it suffices to show that the finite-dimensional distributions of $S_n$ converge to those of FBM. We may as well assume $\mu = 0$ henceforth. We proceed to show this for the cases of P1 and P2 in turn.

For the linear case P1, we note that historically Davydov (1970) and Gorodetskii (1977) provide a proof of the result requiring higher moments. Marinucci and Robinson (2000) relax the requirement to $2 + \delta$ moments for some $\delta > 0$. We will adapt the argument used in Theorem 5.2.3 of Taniguchi and Kakizawa (2000). The linear representation allows us to write $Y_t = \sum_j \epsilon_j \psi_{t-j}$ as in the proof of Theorem 1. We have $\sum_{j=1}^{m} \alpha_j S_{[r_j n]} = \sum_j \epsilon_j b_{j,n}$, with $b_{j,n} = \sum_{i=1}^{m} \alpha_i \sum_{t=1}^{[r_i n]} \psi_{t-j}$. Since $m < \infty$, the same types of bounds used in the proof of Theorem 5.2.3 of Taniguchi and Kakizawa (2000) still apply. Hence $\sum_{j=1}^{m} \alpha_j S_{[r_j n]}$ is asymptotically standard normal when normalized by the square root of $\sum_j b_{j,n}^2 \sigma^2$, which by algebra equals the variance of $\sum_{j=1}^{m} \alpha_j S_{[r_j n]}$. Expanding this expression yields

$$\sum_{i_1, i_2 = 1}^{m} \alpha_{i_1} \alpha_{i_2} \sum_{t_1=1}^{[r_{i_1} n]} \sum_{t_2=1}^{[r_{i_2} n]} \gamma_{t_1 - t_2}$$
$$= \sum_{i_1=1}^{m} \alpha_{i_1}^2 \sum_{t_1, t_2 = 1}^{[r_{i_1} n]} \gamma_{t_1 - t_2} + 2 \sum_{i_1 < i_2} \alpha_{i_1} \alpha_{i_2} \sum_{t_1=1}^{[r_{i_1} n]} \sum_{t_2=1}^{[r_{i_2} n]} \gamma_{t_1 - t_2}$$
$$= \sum_{i_1=1}^{m} \alpha_{i_1}^2 V_{[r_{i_1} n]} + \sum_{i_1 < i_2} \alpha_{i_1} \alpha_{i_2}$$
$$\times \left( V_{[r_{i_1} n]} + V_{[r_{i_2} n]} - V_{[(r_{i_2} - r_{i_1}) n]} \right). \tag{A.3}$$

This final breakdown of the variance is true for any variance of a sum, and is generic. Dividing (A.3) through by $V_n$ gives the asymptotic expression

$$\sim \sum_{i_1=1}^{m} \alpha_{i_1}^2 r_{i_1}^{\beta+1} + \sum_{i_1 < i_2} \alpha_{i_1} \alpha_{i_2} \left( r_{i_1}^{\beta+1} + r_{i_2}^{\beta+1} - (r_{i_2} - r_{i_1})^{\beta+1} \right),$$

which is the same as the variance of $\sum_{i=1}^{m} \alpha_i B(r_i)$. In other words the limit of $V_n^{-1/2} \sum_{j=1}^{m} \alpha_j S_{[r_j n]}$ has the same distribution as $\sum_{i=1}^{m} \alpha_i B(r_i)$.

For the P2 case the $k$th order cumulant of $\sum_{i=1}^{m} \alpha_i S_{[r_i n]}$ can be expanded into a $k$-fold sum

$$\sum_{i_1=1}^{m} \sum_{i_2=1}^{m} \cdots \sum_{i_k=1}^{m} \alpha_{i_1} \alpha_{i_2} \cdots \alpha_{i_k} \operatorname{cum}\left( Z_{i_1}, Z_{i_2}, \ldots, Z_{i_k} \right),$$

where $Z_{i_j} = \sum_{t=[r_{i_j} n]+1}^{[r_{i_{j+1}} n]} Y_t$. The same overall bounds can then be obtained as in the proof of Theorem 1, showing that all higher order cumulants with $k > 2$ of the normalized sum will tend to zero. Again (A.3) shows that the variance converges to the variance of FBM, and so the result follows.

Finally, we establish tightness. The criterion we use is given by Problem 4.11 of Karatzas and Shreve (1991), which is appropriate for $C[0, 1]$. Letting $\gamma = (\kappa + \delta)/2$, and taking any times $s < t$ and any $n$,

$$\mathbb{E}\left[ |\, (\xi_n(t) - nt\mu) - (\xi_n(s) - ns\mu)\, |^{2\gamma} V_n^{-\gamma} \right]$$
$$= V_n^{-\gamma} \mathbb{E}\left[ \left| \sum_{j=[ns]+1}^{[nt]} Y_j + (nt - [nt]) Y_{[nt]+1} \right.\right.$$
$$\left.\left. - (ns - [ns]) Y_{[ns]+1} - n\mu(t - s) \right|^{2\gamma} \right]$$
$$\sim V_n^{-\gamma} \mathbb{E}\left[ |S_{[nt]-[ns]} - ([nt] - [ns])\mu|^{2\gamma} \right]$$
$$= V_n^{-\gamma}\, O(V_{[n(t-s)]}^{\gamma}) = O((t - s)^{(\beta+1)\gamma})$$

by Proposition 1. Because $\gamma > 1/(\beta + 1)$, tightness is assured. This completes the proof. □

**Proof of Theorem 4.** We wish to apply Theorem 3, which is in terms of $\xi_n(t)$, to the result of Proposition 4, which is in terms of $S_n(t)$. We can do this because $\widehat{\xi}_n(t) = \widehat{S}_n(t) + (nt - [nt]) Y_{[nt]+1}$, so that $(\widehat{\xi}_n(t) - nt\mu)/\sqrt{V_n}$ is asymptotically equivalent to $(\widehat{S}_n(t) - [nt]\mu)/\sqrt{V_n}$. Then all linear or quadratic integral expressions of such processes will also be asymptotically the same. So note that $\widehat{S}_i = S_{[rn]} - r S_n$ with $r = i/n$, and recognize the summations in the expression for $W_{\Lambda,M}$ in Proposition 4 as Riemann sums; the result now follows at once from Theorem 3. □

**Proof of Proposition 5.** By Theorem 4 we know that $Q_n := n W_{\Lambda, bn}/V_n \overset{\mathcal{L}}{\Longrightarrow} Q(b)$, and by Proposition 1 we have $V_n/n \sim n^\beta C L(n)/(\beta + 1)$ in the LM/SM/NM case. Thus it follows that (17) satisfies

$$\widehat{\beta} \sim \beta + \frac{\log Q_n}{\log n} + \frac{\log\left( CL(n)/(\beta + 1) \right)}{\log n}.$$

Since $Q_n = O_P(1)$ and $\log L(n)/\log n \to 0$ for slowly-varying functions $L$, the estimator will be consistent. □

## References


Abadir, K., Distaso, W., Giraitis, L., 2009. Two estimators of the long-run variance: beyond short memory. Journal of Econometrics 150, 56–70.
Andrews, D., 1991. Heteroskedasticity and autocorrelation consistent covariance matrix estimation. Econometrica 59, 817–858.
Ango-Nze, P., Dupoiron, S., Rios, R., 2003. Subsampling under weak dependence conditions. Technical Report from CREST, DT2003-42.
Bardet, J., Doukhan, P., Lang, G., Ragache, N., 2008. Dependent Lindeberg central limit theorem and some applications. Probability and Statistics 12, 154–172.
Beran, J., 1993. Fitting long-memory models by generalized linear regression. Biometrika 80, 817–822.
Beran, J., 1994. Statistics for Long Memory Processes. Chapman and Hall, New York.
Bickel, P., Sakov, A., 2008. On the choice of $m$ in the $m$ out of $n$ bootstrap and confidence bounds for extrema. Statistica Sinica 18, 967–985.
Billingsley, P., 1995. Probability and Measure. John Wiley and Sons, New York.
Bingham, N., Goldie, C., Teugels, J., 1987. Regular Variation. Cambridge University Press, Cambridge.
Blackman, R.B., Tukey, J.W., 1959. The Measurement of Power Spectra from the Point of View of Communications Engineering. Dover, New York.
Bradley, R., 2007. Introduction to Strong Mixing Conditions, Vol. I. Kendrick Press, Heber City, Utah.
Brillinger, D.R., 1981. Time Series: Data Analysis and Theory. Holden-Day, New York.
Brockwell, P., Davis, R., 1991. Time Series: Theory and Methods, second ed. Springer, New York.
Davydov, Y., 1970. The invariance principle for stationary processes. Theory of Probability and its Applications 14, 487–498.
Dickey, D., Fuller, W., 1979. Distribution of the estimators for autoregressive time series with a unit root. Journal of the American Statistical Association 74, 427–431.
Dickey, D., Fuller, W., 1981. Likelihood ratio statistics for autoregressive time series with a unit root. Econometrica 49, 1057–1072.
Diebold, F., Rudebusch, G., 1991. On the power of the Dickey–Fuller tests against fractional alternatives. Economics Letters 35, 155–160.
Doukhan, P., Louhichi, S., 1999. A new weak dependence and applications to moment inequalities. Stochastic Processes and their Applications 84, 313–343.
Embrechts, P., Klüppelberg, C., Mikosch, T., 1997. Modeling Extremal Events for Insurance and Finance. Springer-Verlag, Berlin.
Findley, D.F., Monsell, B.C., Bell, W.R., Otto, M.C., Chen, B.C., 1998. New capabilities and methods of the X-12-ARIMA seasonal adjustment program. Journal of Business and Economic Statistics 16, 127–177 (with discussion).
Gorodetskii, V., 1977. On convergence to semi-stable Gaussian processes. Theory of Probability and its Applications 22, 498–508.
Grenander, U., Rosenblatt, M., 1957. Statistical Analysis of Stationary Time Series. Wiley, New York.
Hall, P., 1992. The Bootstrap and Edgeworth Expansion. Springer-Verlag, New York.
Hamilton, J.D., 1994. Time Series Analysis. Princeton University Press, Princeton.
Hannan, E.J., 1970. Multiple Time Series. John Wiley, New York.
Hosking, J.R.M., 1981. Fractional differencing. Biometrika 68, 165–176.
Hosoya, Y., 1996. Some limit theorems on stationary processes with long-range dependence. In: Robinson, P.M., Rosenblatt, R. (Eds.), Athens Conference on Applied Probability and Time Series, Vol. 2. Springer, pp. 233–245.
Ibragimov, I., 1962. Some limit theorems for stationary processes. Theory of Probability and its Applications 7, 349–382.
Ibragimov, I., Linnik, Y., 1971. Independent and Stationary Sequences of Random Variables. Wolters-Noordhof Publishing, Groningen.
Jach, A., McElroy, T., Politis, D., 2012. Subsampling inference for the mean of heavy-tailed long memory time series. Journal of Time Series Analysis 33, 96–111.
Karatzas, I., Shreve, S., 1991. Brownian Motion and Stochastic Calculus. Springer-Verlag, New York.
Kiefer, N., Vogelsang, T., 2002. Heteroscedastic-autocorrelation robust standard errors using the Bartlett kernel without truncation. Econometrica 70, 2093–2095.
Kiefer, N., Vogelsang, T., 2005. A new asymptotic theory for heteroskedasticity-autocorrelation robust tests. Econometric Theory 21, 1130–1164.
Kiefer, N., Vogelsang, T., Bunzel, H., 2000. Simple robust testing of regression hypotheses. Econometrica 68, 695–714.
Kirch, C., Politis, D., 2011. TFT-bootstrap: resampling time series in the frequency domain to obtain replicates in the time domain. Annals of Statistics 39, 1427–1470.
Kolmogorov, A., Rozanov, J., 1960. On a strong mixing condition for a stationary Gaussian process. Theory of Probability and its Applications 5, 204–208.
Lahiri, S., 2003. Resampling Methods for Dependent Data. Springer-Verlag, New York.
Lopez, J., 1997. The power of the ADF test. Economics Letters 57, 5–10.
Maravall, A., Caparello, G., 2004. Program TSW: revised reference manual. Working Paper 2004, Research Department, Bank of Spain. http://www.bde.es.
Marinucci, D., Robinson, P., 2000. Weak convergence of multivariate fractional processes. Stochastic Processes and their Applications 86, 103–120.
McElroy, T., Politis, D., 2007. Computer-intensive rate estimation, diverging statistics, and scanning. Annals of Statistics 35, 1827–1848.
McElroy, T., Politis, D., 2009. Fixed-b asymptotics for the studentized mean from time series with short, long or negative memory. Discussion Paper, Department of Economics, UC San Diego. http://escholarship.org/uc/item/70c4x0sq.
McElroy, T., Politis, D., 2011. Distribution theory for the studentized mean for long, short, and negative memory time series. Discussion Paper, Department of Economics, UC San Diego. http://escholarship.org/uc/item/35c7r55c.
McElroy, T., Politis, D., 2012. Fixed-b asymptotics for the studentized mean for long and negative memory time series. Econometric Theory 28, 471–481.
Meerschaert, M., Scheffler, H., 1998. A simple robust estimator for the thickness of heavy tails. Journal of Statistical Planning and Inference 71, 19–34.
Palma, W., 2007. Long-Memory Time Series. John Wiley and Sons, New York.
Parker, C., Paparoditis, E., Politis, D., 2006. Unit root testing via the stationary bootstrap. Journal of Econometrics 133, 601–638.
Parzen, E., 1957. On consistent estimates of the spectrum of a stationary time series. Annals of Mathematical Statistics 28, 329–348.
Percival, D.B., Walden, A.T., 1993. Spectral Analysis for Physical Applications. Multitaper and Conventional Univariate Techniques. Cambridge University Press, Cambridge.
Phillips, P., Perron, P., 1988. Testing for a unit root in time series regression. Biometrika 75, 335–346.
Politis, D., 2001. On nonparametric function estimation with infinite-order flat-top kernels. In: Charalambides, Ch., et al. (Eds.), Probability and Statistical Models with Applications. Chapman and Hall/CRC, Boca Raton, pp. 469–483.
Politis, D., 2005. Higher-order accurate, positive semi-definite estimation of large-sample covariance and spectral density matrices, UCSD Dept. of Economics Discussion Paper 2005-03.
Politis, D., Romano, J., 1995. Bias-corrected nonparametric spectral estimation. Journal of Time Series Analysis 16, 67–103.
Politis, D., Romano, J., Wolf, M., 1999. Subsampling. Springer-Verlag, New York.
Priestley, M.B., 1981. Spectral Analysis and Time Series. Adademic Press, New York.
Robinson, P., 1994. Semiparametric analysis of long memory time series. Annals of Statistics 22, 515–539.
Robinson, P., 2005. Robust covariance matrix estimation: HAC estimates with long memory/antipersistence correction. Econometric Theory 21, 171–180.
Rosenblatt, M., 1956. A central limit theorem and a strong mixing condition. Proceedings of the National Academy of Sciences 42, 43–47.
Rosenblatt, M., 1961. Independence and dependence. In: Proc. Fourth Berkeley Symp. on Math. Statist. and Prob., Vol. 2. University of California Press, pp. 431–443.
Rosenblatt, M., 1984. Stochastic processes with short-range and long-range dependence. In: David, H.A., David, H.T. (Eds.), Statistics: An Appraisal, Proceedings of the 50th Anniversary Conference. The Iowa State University Press, Ames, IA, pp. 509–520.
Rosenblatt, M., 1985. Stationary Sequences and Random Fields. Birkhäuser, Boston.
Samorodnitsky, G., Taqqu, M., 1994. Stable Non-Gaussian Random Processes. Chapman & Hall, New York City, New York.
Sun, Y., 2004. A convergent $t$-statistic in spurious regressions. Econometric Theory 20, 943–962.
Sun, Y., Phillips, P., Jin, S., 2008. Optimal bandwidth selection in heteroskedasticity-autocorrelation robust testing. Econometrica 76, 175–194.
Tanaka, K., 1990. Testing for a moving average unit root. Econometric Theory 6, 433–444.
Tanaka, K., 1996. Time Series Analysis: Nonstationary and Noninvertible Distribution Theory. John Wiley and Sons, New York City, New York.
Taniguchi, M., Kakizawa, Y., 2000. Asymptotic Theory of Statistical Inference for Time Series. Springer-Verlag, New York City, New York.