Nonparametric nonstationarity tests*

Federico M. Bandi* and Valentina Corradi**
*Johns Hopkins University and Edhec-Risk and **University of Warwick

December 2011

Abstract

We propose additive functional-based nonstationarity tests which exploit the different divergence rates of the occupation times of a (possibly nonlinear) process under the null of nonstationarity (stationarity) versus the alternative of stationarity (nonstationarity). We consider both discrete-time series and continuous-time processes. The discrete-time case covers Harris recurrent Markov chains and integrated processes. The continuous-time case focuses on Harris recurrent diffusion processes. The proposed tests are simple to implement and rely on tabutaled critical values. Simulations show that their size and power properties are satisfactory. Our robustness to nonlinear dynamics provides a solution to the typical inconsistency problem between assumed linearity of a time series for the purpose of nonstationarity testing and subsequent nonlinear inference.

Keywords: Nonstationarity testing, Nonlinear Markov processes.

*Paper written for the 2010 International Symposium on Econometric Theory and Applications (SETA 2010) jointly organized by the Center for Financial Econometrics (CoFiE) at the Sim Kee Boon Institute for Financial Economics and the School of Economics, Singapore Management University (Singapore, April 29th - May 1st, 2010). We thank three anonymous referees and the Editor, Peter C.B. Phillips, for very useful comments and suggestions. We are also grateful to David Hendry, Hashem Pesaran, and the participants at the conference "High-Dimensional Econometric Modelling", London, December 3rd and 4th, 2011, for discussions.

1 Introduction

This paper suggests novel nonstationarity tests for possibly nonlinear discrete-time and continuous-time processes. The vast literature on unit-root testing has virtually exclusively focused on linear models, e.g., Phillips and Xiao (1998) for a review. A considerable amount of recent work has, however, been devoted to the use of possibly highly nonlinear specifications to model an array of time series of interest. In continuous-time finance, for example, much attention has been on the use of diffusion structures to model interest rates and stock returns (e.g., Aït-Sahalia, 1996, Conley, Hansen, Luttmer and Scheinkman, 1997, and Pritsker, 1998, among others). A diffusion sampled at discrete time intervals, i.e, the skeleton of a diffusion, is, in general, a *nonlinear* Markov chain. Nonetheless, the common practice is to test for nonstationarity up-front by virtue of methods whose theoretical justification hinges on *linearity*, as in the Dickey-Fuller tradition and its many developments. This issue creates a fundamental inconsistency between nonstationarity testing, which is typically conducted before inference begins, and modelling, in the context of which nonlinear dynamics are now the norm, rather than the exception. To provide a solution to this pervasive inconsistency problem, there is a need for nonstationarity tests which are robust to nonlinear dynamics.

Our aim is to introduce and formalize ideas intended to fill this important, in our view, gap in the literature. We do so for a rather general class of Markov chains. Because the skeleton of a diffusion is a Markov chain, diffusion processes are a sub-case of our broader treatment.

The intuition behind our methods goes as follows. If a process is stationary, the amount of time that the process spends in the local neighborhood of a point diverges to infinity linearly with the number of observations. Under nonstationarity, the returns to open sets are rarer, thereby leading to slower rates of recurrence which depend on the degree of nonstationarity. We employ this fundamental observation to construct nonstationarity tests for processes in the Harris recurrent class.

Formally, let $\{X_t\}_{t\geq 1}$ be a univariate Harris recurrent Markov chain with state space $(\mathbf{E}, \mathcal{E})$ and unique invariant measure π. Denote the number of visits at a point $x \in \mathcal{D} \subset \mathcal{R}$ by

$$L_n(x) = \#\left\{t; 1 \leq t \leq n,\ X_t \in \lim_{\varepsilon \to 0} B_\varepsilon(x)\right\},$$

where $B_\varepsilon(x)$ is an open ball of radius ε centered at x. By recurrence, $L_n(x) \overset{a.s.}{\to} \infty$ as $n \to \infty$. Null recurrent (i.e., nonstationary) and positive recurrent (strictly stationary or stationary in the limit) Markov chains have, however, occupation times $\widehat{L}_n(x)$ which diverge to infinity at different rates. The tests that we propose exploit the different divergence rates of the occupation times of a recurrent Markov chain under the null of nonstationarity (stationarity) versus the alternative of stationarity (nonstationarity).

Estimating occupation times would require selecting a bandwidth parameter to capture locality. Even though, for the class of processes discussed in this paper, the choice of the locality parameter may be conducted as suggested by Bandi, Corradi, and Wilhelm (2011) in recent work, such a choice would add an unnecessary layer of complication to our analysis. Importantly, additive functionals of the type $\sum_{t=1}^{n} f(X_t)$, where f is a non-negative function integrable with respect to the process' invariant measure π, are known to inherit the divergence properties of the corresponding occupation times. The divergence

rates of $\sum_{t=1}^{n} f(X_t)$, under different "degrees" of recurrence, have been established by Chen (1999). We may therefore rely on the divergence rates of additive functionals of the process for the purpose of constructing the tests. The tests combine sample conditioning with a *randomization* procedure. They result in readily tabulated critical values and apply to all Harris recurrent Markov processes. In discrete time, we explicitly cover Harris recurrent Markov chains (as in, e.g., Karlsen and Tjostheim, 2001, Guerre, 2004, and Schienle, 2008) and integrated processes (as in, e.g., Wang and Phillips, 2009a, 2009b). In continuous time, we study the case of Harris recurrent diffusion processes (Bandi and Phillips, 2003, and Bandi and Phillips, 2010, for a review).

Randomized tests have first been suggested in series of papers by Pearson (1950), Stevens (1950), and Tocher (1950) who combine results from independent experiments in the case of discontinuous random variables. The basic idea is to add a uniform $[0, 1]$ random variable to the sample observations. Suppose we have a sample $X_1, ..., X_n$ from a random variable X endowed with a discrete distribution. One can then construct the continuous random variable $Y_i = X_i + U_i$, where, for $i = 1, ..., n$, the U_i's are independent draws from a uniform distribution on $[0, 1]$. Another classical application of randomization is in the context of rank tests, in the presence of ties due, for example, to the discreteness of the underlying distribution, e.g., Hajek and Sidak (Chapter 3, 1967). In this case, one uses a supplementary random experiment so that any possible rank assignment is drawn with equal probability. The rank test statistic is then constructed by drawing one of the possible rank assignments. More recently, Lutkepohl and Burda (1997) have used randomization in the context of Wald tests with asymptotically singular covariance matrices. Specifically, they add a draw from a $N(0, \Sigma)$ random vector to the (function of the) estimated parameters. In all the papers cited above, the limiting distribution is driven by the joint probability law of the sample and that of the added randomness, which is indeed the product of the two, given independence. In this sense, there is no issue of sample conditioning.

A different use of randomization is that involved in the construction of conditional p-values (e.g., Hansen, 1996) or in Monte Carlo tests (e.g., Dufour and Kiviet, 1996). In this case, contrary to the examples above, the actual statistic only depends on the sample of observations. However, the p-value used to decide whether to reject or not the null hypothesis depends on added, simulated, randomness, conditional on the sample. Typically, conditional p-values and Monte Carlo tests are used in situations in which the statistic has a well-defined limiting distribution, though non-standard or dependent on nuisance parameters.

Because of the joint presence of nonstationarity and nonlinearity, it is hardly feasible for our problem to construct a statistic which has, if the null is true, a well-defined limiting distribution under the probability law governing the sample, and which diverges under the alternative. For this reason, we suggest a statistic which, conditional on the sample, and for all samples except a set of zero probability measure, has a well-defined limiting distribution in terms of the law governing the added randomness, and which diverges under the alternative. As explained in detail in the proof of Theorem 1 below, we can decompose the suggested statistic into two terms. The first term, conditional on the sample, converges in distribution under both hypotheses, in terms of the law governing the simulated randomness. The second

term, for all samples under the null, converges to zero, and for all samples under the alternative, diverges. In particular, the speed at which the second term converges to zero, or diverges, depends on the distance between the null and the alternative hypothesis. Related approaches have been employed in other contexts. Corradi and Swansson (2006) use randomized procedures to distinguish between unit-roots in levels and in logs. After defining a (near) rate-optimal bandwidth selection method, Bandi, Corradi, and Wilhelm (2011) employ it to bias-correct (i.e., appropriately center) the asymptotic distribution of kernel estimates of first and second moments in the context of nonlinear autoregressive and cointegrating models. Bandi, Corradi, and Moloche (2009) use it in the nonparametric estimation of continuous-time Markov models to define a *feasible set* in which the bandwidth needed for estimation of a specific infinitesimal moment satisfies all conditions for consistency and asymptotic zero-mean normality.

When dealing with linear unit-root processes, our approach, which relies on less information than classical approaches for linear time series, is bound not to have the theoretical optimality, or near-optimality, properties of autoregressive coefficient-based (or t-ratio based) methods in the literature (see, e.g., Elliott, Rothemberg, and Stock, 1996). However, robustness to nonlinear dynamics makes our procedures particularly appealing when one is unwilling to impose a linear parametric structure on the underlying process of interest. In the case of linear data generating processes, we compare the size and power properties of our tests to that of standard unit-root tests. We do so for samples of moderate magnitude. We find that the size of our test(s) is comparable to that of standard unit-root tests. As expected, our tests are less powerful. However the loss of power, which varies across different configurations, is overall rather mild. In other words, the price paid for robustness to nonlinearities is small.

We start off with preliminary technical notions (Section 2). Section 3 discusses additive functional-based nonstationarity testing for Harris recurrent Markov chains. Section 4 covers the classical linear unit-root case. Section 5 focuses on recurrent diffusion processes. Size and power properties are examined in Section 6. Some final remarks are in Section 7. Section 8 concludes. All proofs are in the Appendix.

2 Preliminary technical notions

We begin with formal assumptions on the underlying Markov process.

Assumption A. Let $\{X_t\}_{t\geq 1}$ be a $p-$regular, $\phi-$irreducible Markov chain on a general state space $(\mathbf{E}, \mathcal{E})$ with transition probability $P(x, A)$ and invariant measure π. Let $p \in (0, 1]$.[1]

We now introduce two results from Chen (1999) which will be employed, in what follows, to derive our tests.

Proposition 1 (Chen, 1999, Theorem 2.3.) *Let* $\{X_t\}$, $t \geq 1$, *be a* $p-$*regular Harris recurrent chain. For every nonnegative function* $f \in \mathcal{L}^1(\mathbf{E}, \mathcal{E}, \pi)$, *the additive functional* $\sum_{j=1}^{n} f(X_j)$, *when standardized*

[1] As said, the case $p = 1$, with the addition of some innocuous regularity conditions, corresponds to the case of positive recurrent (or strictly stationary) chains.

by $\alpha(n) = L(n)n^p$ *with* $L(n)$ *slowly-varying at infinity and* $0 \leq p \leq 1$, *satisfies*

$$\frac{\sum_{j=1}^{n} f(X_j)}{\alpha(n)} \Rightarrow (ml_p) \int f(x)\pi(dx),$$

where ml_p *is the Mittag-Leffler density with the same parameter* p.

Proposition 2 (Chen, 1999, Theorem 2.4.) *Let* $\{X_t\}$, $t \geq 1$, *be a* $p-$*regular Harris recurrent chain. Define* $L_2\lambda = \log\log\max\{\lambda, e^e\}$ *with* $\lambda \geq 0$. *For every nonnegative function* $f \in \mathcal{L}^1(\mathbf{E}, \mathcal{E}, \pi)$, *the additive functional* $\sum_{j=1}^{n} f(X_j)$, *when standardized by* $\alpha\left(\frac{n}{L_2 a(n)}\right) L_2 a(n)$ *with* $\alpha(n) = L(n)n^p$, $L(n)$ *slowly-varying at infinity and* $0 \leq p \leq 1$, *satisfies*

$$\limsup_{n\to\infty} \frac{\sum_{j=1}^{n} f(X_j)}{\alpha\left(\frac{n}{L_2\alpha(n)}\right) L_2\alpha(n)} = \frac{\Gamma(p+1)}{p^p(1-p)^{1-p}} \int f(x)\pi(dx) \qquad a.s.,$$

where one should interpret $p^p = (1-p)^{1-p} = 1$ *if* $p = 0$ *or* 1.

Proposition 1 provides a weak convergence result for additive functionals of recurrent Markov chains. As $n \to \infty$, the standardized additive functional $\sum_{j=1}^{n} f(X_j)$ converges to a re-scaled Mittag-Leffler random variable with parameter p consistent with the regularity of the underlying process. If $p = 0$, the Mittag-Leffler density reduces to the exponential density and the limit distribution of the additive functional is that of an exponential random variable with parameter $\int f(x)\pi(dx)$. If $p = 1$, the Mittag-Leffler density is degenerate and $\frac{\sum_{j=1}^{n} f(X_j)}{\alpha(n)} = \frac{\sum_{j=1}^{n} f(X_j)}{n} \xrightarrow{p} \int f(x)p(dx)$. As is well-known, this convergence is also with probability one. Proposition 2 provides *strong* increasing rates for additive functionals. Naturally, the number of times that the process $\{X_t\}_{t\geq 1}$ visits a given set $A \in \mathcal{E}$ with $0 < \pi(A) < \infty$ can be obtained by replacing f with 1_A, the indicator function of the set A. Thus, Proposition 1 and 2 also provide the weak and strong rate of divergence of the occupation times of positive-recurrent ($p = 1$) and null-recurrent ($p < 1$) chains. The class of $p-$ regular Markov chains is rather broad. For example, the $\beta-$ recurrent Markov chains studied by Karlsen and Tjostheim (2001) are indeed $p-$ regular with $p = \beta$. Similarly, the skeleton of a nonlinear diffusion process is, in general, a $p-$ regular chain.

3 Additive functionals-based nonstationarity tests

Propositions 1 and 2 will be used below to justify novel nonstationarity tests. They readily imply that, in the positive recurrent case $p = 1$, $\frac{1}{n}\sum_{j=1}^{n} f(X_j) \xrightarrow{a.s.} \mathrm{E}(f(X)) > 0$ as $n \to \infty$, whereas in the null recurrent case $p < 1$, $\frac{1}{n}\sum_{j=1}^{n} f(X_j) \xrightarrow{a.s.} 0$ as $n \to \infty$.

Of course, one cannot distinguish between $p = 1$ and $p < 1$ for any fixed sample size n. Any testing argument should therefore hinge on asymptotic statements. This is indeed the same situation occurring in the linear case when the goal is to discriminate between $I(0)$ processes and $I(1)$ processes using the fact that partial sums of $I(0)$ processes satisfy a functional central limit theorem (FCLT). Kwiatkowski, Phillips, Schmidt and Shin (1992), KPSS henceforth, for example, test the null of $I(0)$

versus the alternative of $I(1)$. Breitung (2002) tests the null of $I(1)$ versus the alternative of $I(0)$. Müller (2008) discusses the difference between setting a null of $I(0)$ versus a null of $I(1)$, or vice versa.

It will be clear in what follows that we can choose the null as being stationarity (as in KPSS, 1992) or nonstationarity (as in Breitung, 2002). Similarly to the KPSS test statistic, but differently from the Breitung statistic which converges to zero under the alternative, the proposed statistic will converge in distribution under the null and will diverge under the alternative.

Because of nonlinearity, we have considerably less information than in the approaches mentioned above. In particular, we only know that $\frac{1}{n}\sum_{j=1}^{n} f(X_j)$ has a strictly positive almost-sure limit under positive recurrence and has a zero almost-sure limit under null recurrence. Thus, we cannot rely on a FCLT and derive well-defined limiting distributions under the probability law governing the sample. To overcome this issue, which is really a by-product of the mild assumptions that we impose on the dynamics, we rely on a testing procedure based on the joint use of sample conditioning and randomization. While the approach relates to those in Corradi and Swansson (2006), Bandi, Corradi and Moloche (2009) and Bandi, Corradi and Wilhelm (2011) in other contexts, our focus on nonstationarity testing leads to different randomized statistics to which we now turn.

3.1 Null of nonstationarity

We wish to test the null hypothesis

$$H_0 : p \leq \overline{p} < 1$$

against the alternative

$$H_A : p = 1.$$

It is immediate to see that our null is "larger" than the usual null of a unit root, which may be stated as $p = 1/2$. Under some additional regularity assumption, p can be estimated. However, its estimator would only converge at a logarithmic rate (see Remark 3.7 in Karlsen and Tjostheim, 2001). Furthermore, no limiting distribution result for the estimated p has been established so far. Hence, a t-ratio based test is currently not viable.

We suggest the following randomized statistic:

$$V_{R,n} = \frac{2}{\sqrt{R}} \sum_{j=1}^{R} \left(1\left\{ \eta_j \leq \lambda\left(\frac{\sum_{j=1}^{n} f(X_j)}{n} \right) \right\} - \frac{1}{2} \right), \tag{1}$$

where f is non-negative, $\pi-$integrable function on $\mathbf{E}$, $\lambda(x)$ is a positive monotonic function such that $\lambda(x) \to 0$ as $x \to 0$ and the η_js are a set of standard normal draws ($1 \leq j \leq R$). The sample size of the simulated series, R, is chosen in such a way as to guarantee that $\sqrt{R}\lambda\left(\frac{b_{\overline{p}}(n)}{n}\right) \to 0$ with $b_p(n) = \left(\frac{n}{\log\log(L(n)n^p)}\right)^p L\left(\frac{n}{\log\log(L(n)n^p)}\right)\log\log(L(n)n^p)$ and $L(n)$ is a slowly-varying function at infinity. It is important to note that the upper bound of the value of p under the null, i.e. $\overline{p}$, plays no role in the construction of the statistic.[2] Nevertheless, it plays a role in determining the rate at which the sample

[2] This is contrast with fractional Dickey-Fuller tests, in which the statistic depend on both the fractional differencing parameter under the null and under the alternative, see Dolado, Gonzalo and Mayoral (2002).

size of the simulated randomness R can grow relative to n. The further $\overline{p}$ is from 1, and so the further the null and the alternative are, the faster R can grow. Intuitively, given a sample size n, the more distant the null and the alternative, the more we are able to discriminate between the two hypotheses.

In what follows, the symbols P^* and d^* denote convergence in probability and in distribution under P^*, which is the probability law governing the simulated random draws η_j, conditional on the sample. Also, E^* and Var^* denote the mean and variance operators under P^*. Furthermore, the notation $a.s.-P$ is used to mean "for all samples but a set of measure 0."

The logic underlying the statistic in Eq. (1) is as follows. We can decompose $V_{R,n}$ into two terms:

$$\begin{aligned} V_{R,n} &= \frac{2}{\sqrt{R}}\sum_{j=1}^{R}\left(1\left\{\eta_j \leq \lambda\left(\frac{\sum_{j=1}^{n} f\left(X_j\right)}{n}\right)\right\} - \mathrm{E}^*\left(1\left\{\eta_j \leq \lambda\left(\frac{\sum_{j=1}^{n} f\left(X_j\right)}{n}\right)\right\}\right)\right) \\ &\quad +2\sqrt{R}\left(\mathrm{E}^*\left(1\left\{\eta_j \leq \lambda\left(\frac{\sum_{j=1}^{n} f\left(X_j\right)}{n}\right)\right\}\right) - \frac{1}{2}\right). \end{aligned} \tag{2}$$

The first term on the right-hand side of Eq. (2) converges in distribution to a normal random variable under P^* regardless of which hypothesis is satisfied. Specifically, it converges to a standard normal random variable under the null. Under H_0, $\lambda\left(\frac{\sum_{j=1}^{n} f(X_j)}{n}\right) \overset{a.s.}{\to} 0$, at speed $\lambda\left(\frac{b_p(n)}{n}\right)$ where, up to a slowly-varying term, $\frac{b_p(n)}{n} \approx n^{p-1}$. Thus, for all samples, under the null, the second term is $O_{a.s.}\left(\sqrt{R}\lambda\left(\frac{b_p(n)}{n}\right)\right) = o_{a.s.}(1)$ for all $p \leq \overline{p} < 1$, provided $\sqrt{R}\lambda\left(\frac{b_{\overline{p}}(n)}{n}\right) \to 0$. Under H_A, $\lambda\left(\frac{\sum_{j=1}^{n} f(X_j)}{n}\right) \overset{a.s.}{\to} \lambda\left(\mathrm{E}(f(X))\right) > 0$, hence, for all samples, under the alternative, $\left(\mathrm{E}^*\left(1\left\{\eta_j \leq \lambda\left(\frac{\sum_{j=1}^{n} f(X_j)}{n}\right)\right\}\right) - \frac{1}{2}\right) > 0$ and the second term on the right-hand side of Eq. (2) diverges at rate $\sqrt{R}$. In light of these observations, it is clear that the optimal choice of number of random draws R is to let it grow at rate $\lambda^{-2(1+\varepsilon)}\left(\frac{b_{\overline{p}}(n)}{n}\right)$ with $\varepsilon > 0$ arbitrarily small. When doing so, however, if $\overline{p} < p < 1$, then the second term diverges, leading to the wrong conclusion that the chain is positive recurrent.

The following theorem establishes the limiting behavior of $V_{R,n}$.

Theorem 1. *Let Assumption A hold, f be non-negative and such that $f \in \mathcal{L}^1(\mathbf{E}, \mathcal{E}, \pi)$, and let $\lambda(x)$ be monotonically-decreasing to zero as $x \to 0$. Also, let $b_p(n) = \left(\frac{n}{\log\log(L(n)n^p)}\right)^p L\left(\frac{n}{\log\log(L(n)n^p)}\right)\log\log(L(n)n^p)$, with $L(n)$ slowly-varying at infinity. If, as $R, n \to \infty$, $\sqrt{R}\lambda\left(\frac{b_{\overline{p}}(n)}{n}\right) \to 0$,*
(i) Under H_0,

$$V_{R,n} \overset{d^*}{\to} N(0,1) \ a.s.-P.$$

(ii) Under H_A, there are constants $c_1, c_2 > 0$ so that

$$P^*\left(R^{-1/2+c_1}V_{R,n} > c_2\right) \to 1 \ a.s.-P.$$

Noting that the second term on the right-hand side of Eq. (2) cannot be negative, we should perform a one-sided test, rejecting at level $\alpha\%$, whenever $V_{R,n}$ is larger than the $(1-\alpha)$-percentile of the standard normal random variable. Contrary to classical nonstationarity tests of the Dickey-Fuller type, the critical values of the test are readily tabulated being those of a standard normal random variable.

The implementation of the test requires a choice of $\lambda(.)$ and $f(.)$. The choice of the function $\lambda(.)$ determines a finite sample trade-off between size and power. The faster $\lambda(x)$ decreases to zero as $x \to 0$, the better the finite sample size, the worse the finite sample power. In practice, as illustrated in the Monte Carlo section, the natural choice is a power function. Needless to say, the larger the sample size, the less important the choice of $\lambda(.)$. The choice of the non-negative function $f(.)$ depends on the sub-class of processes being considered. It has to be such that integrability with respect to the invariant density of the process is satisfied. The indicator function of a compact set surely satisfies the positivity and the integrability requirement. Though, in practice this is not the best choice, as it leaves with the selection of a compact set to use. In the case of random walks (more on this in Section 4), any non-negative function which is integrable with respect to the Lebesgue measure may, in principle, be employed. In finite samples, however, different integrable (with respect to π) functions may perform differently, thereby requiring care for implementation. In Section 6, we further discuss these issues.

Finally, it is worthwhile to point out the analogies and the differences between the wild bootstrap and our joint use of randomization and sample conditioning. Wild bootstrap statistics are constructed using sample observations as well as simulated randomness. By drawing B simulated samples of the same size as the actual sample size, one may construct B wild bootstrap statistics and their empirical distribution. The (possible) rejection of the null hypothesis at level $\alpha\%$ is then based on the comparison of the actual statistic and the $(1-\alpha)$-percentile of the wild bootstrap empirical distribution. In our case, instead, we draw only one random sample of size R. We then construct one statistic based on the R random draws and on the n sample observations. The statistic is then compared to the critical value of a standard normal. The wild bootstrap is used in situations in which the statistic has a well-defined limiting distribution in terms of the probability law governing the sample. This is not our case. Wild bootstrap critical values are used either to deal with the presence of nuisance parameters (as in Hansen, 1996) or to obtain higher order refinements over asymptotic critical values (as in Davidson and Flachaire, 2008, and Gonçalves and Meddahi, 2009).

3.2 Null of stationarity

By switching the hypotheses in Section 3.1, we may also test the null of positive recurrence

$$H_0' : p = 1$$

against the alternative of null recurrence

$$H_A' : p \leq \overline{p} < 1.$$

Let, again, f be a non-negative, $\pi-$integrable function on $\mathbf{E}$. We suggest the following statistic

$$\widetilde{V}_{R,n}(\overline{p}) = \frac{2}{\sqrt{R}} \sum_{j=1}^{R} \left(1 \left\{ \eta_j \leq \lambda \left(\frac{b_{\overline{p}}(n)}{\sum_{j=1}^{n} f(X_j)} \right) \right\} - \frac{1}{2} \right), \tag{3}$$

where $b_p(n)$ and $\lambda(x)$ are defined as in the previous subsection, and the η_j's is a set of standard normal draw ($1 \leq j \leq R$). The sample size of the simulated series, R, is chosen in such a way as to guarantee that $\sqrt{R}\lambda\left(\frac{b_{\overline{p}}(n)}{n}\right) \to 0$, as before.

It is important to note that, contrary to our earlier results, the parameter $\overline{p}$ controlling the distance between the null and the alternative hypothesis is now used in the construction of the statistic as well as to determine the rate of growth of R. As in the case of $V_{R,n}$, $\widetilde{V}_{R,n}$ can also be decomposed into two terms. The first term converges in distribution under P^*, regardless of whether H_0' or H_A' is true. The second term, which depends only on sample observations, converges to zero at rate $\sqrt{R}\lambda\left(\frac{b_{\overline{p}}(n)}{n}\right)$, for any sample generated under the null, and diverges at rate $\sqrt{R}$ for any sample generated under the alternative. Not surprisingly, the speed at which the second term approaches zero under H_0', or diverges under H_A', increases the further the two hypotheses are. As for the first term, which depends on both simulated randomness and sample observations, it converges to a normal random variable under P^* for any sample. Finally, the test has power against closer alternative, i.e., $\overline{p} < p < 1$, provided R is such that $\sqrt{R}\lambda\left(\frac{b_{\overline{p}}(n)}{b_p(n)}\right) \to \infty$.

Theorem 2. *Let Assumption A hold, f be non-negative and such that $f \in \mathcal{L}^1(\mathbf{E}, \mathcal{E}, \pi)$, and let $\lambda(x)$ be monotonically-decreasing to zero as $x \to 0$. Also, let $b_p(n) = \left(\frac{n}{\log\log(L(n)n^p)}\right)^p L\left(\frac{n}{\log\log(L(n)n^p)}\right)\log\log(L(n)n^p)$, with $L(n)$ slowly-varying at infinity. If, as $R, n \to \infty$, $\sqrt{R}\lambda\left(\frac{b_{\overline{p}}(n)}{n}\right) \to 0$,*
(i) Under H_0,

$$\widetilde{V}_{R,n} \xrightarrow{d^*} N(0,1) \ a.s. - P.$$

(ii) Under H_A, there are constants $c_1, c_2 > 0$ so that

$$P^*\left(R^{-1/2+c_1}\widetilde{V}_{R,n} > c_2\right) \to 1 \ a.s. - P.$$

Again, we reject the null at $\alpha\%$ if $\widetilde{V}_{R,n}$ is larger that the $(1-\alpha)-$percentile of the standard normal distribution.

4 Unit roots

We now turn to the most classical modelling approach in the literature, namely linear integrated processes. In the case of martingale difference series errors, linear integrated process are, in fact, $\frac{1}{2}-$regular recurrent Markov chains. Hence, the statements of Theorem 1 and 2 immediately apply with $p = \overline{p} = \frac{1}{2}$. On the other hand, in the linear case, we can dispense with the Markov assumption and can still apply the test outlined in the previous section under Assumption B below.

Assumption B. Let $\{X_t\}_{t\geq 1}$ satisfy $X_t = \rho X_{t-1} + \xi_t$ where ξ_t is $\alpha-$mixing with size $-(4(4+\gamma))/\gamma$, $\gamma > 0$, and $\mathrm{E}\left(|\xi_t|^{2(4+\gamma)}\right) \leq C_1 < \infty$. Also, there exists $0 < \omega_0^2 < \infty$ so that $\left|T^{-1}\mathrm{E}\left(\left(\sum_{k=m+1}^{m+T}\xi_k\right)^2\right) - \omega_0^2\right| \leq C_2 T^{-\psi}$ with $\psi > 0$ and C_2 independent of m.

Assumption B is rather standard. It controls the degree of memory and heterogeneity of the innovation sequence. The null and the alternative hypothesis may also be cast in a familiar framework. We test for nonstationarity

$$H_0'' : \rho = 1$$

versus stationarity

$$H_A'' : |\rho| < 1.$$

Theorem 3. *Let Assumption B hold, f be non-negative and such that $f \in \mathcal{L}^1(\mathbf{E}, \mathcal{E}, \pi)$, and $\lambda(x)$ be monotonically decreasing to zero as $x \to 0$. Also, let $R, n \to \infty$ and $\sqrt{R}\lambda\left(\frac{\sqrt{n \log\log n}}{n}\right) \to 0$.*
(i) Under H_0'',

$$V_{R,n} \xrightarrow{d^*} N(0,1) \quad a.s. - P.$$

(ii) Under H_A'', there are constants $c_1, c_2 > 0$ so that

$$P^*\left(R^{-1/2+c_1} V_{R,n} > c_2\right) \to 1 \quad a.s. - P,$$

where $V_{R,n}$ is defined as in Eq. (1).

One may again switch the hypotheses above and perform a test of stationarity versus nonstationarity under Assumption B, and a linear data-generating process, using the statistics $\widetilde{V}_{R,n}\left(\frac{1}{2}\right)$ defined in Eq. (3) provided $\sqrt{R}\lambda\left(\frac{\sqrt{n \log\log n}}{n}\right) \to 0$.

As discussed above, because of their reliance on more limited structure, in the case of linear data-generating processes, our tests do not share the optimality against n-local alternatives which standard tests (such as the Dickey-Fuller test or Phillips' Z test) have. In Section 6, we show that the actual power loss can be minimal in practise.

5 Diffusion processes

The skeleton of a diffusion, i.e. a diffusion sampled at discrete time intervals, inherits the recurrence properties of the underlying continuous-time process (Meyn and Tweedie, 1993). Hence, the tests outlined in Sections 3 and 4 should, in principle, be applicable to widely-used continuous-time processes sampled discretely. However, if high-frequency observations on the process are available, one may wish to use them, rather than just resort to a low-frequency skeleton. In this section, we formalize this intuition.

Consider a diffusion process $\{X_t : t \geq 0\}$ defined as the unique, strong solution to $dX_t = \mu(X_t)dt + \sigma(X_t)dB_t$ on $\mathcal{A} = (l, u)$, where $\{B_t : t \geq 0\}$ is a standard Brownian motion.

Define $t_x^* = \inf\{t \geq 0 | X_t \in \lim_{\varepsilon \to 0} B_\varepsilon(x)\}$, the first crossing time of the level x. It is known that, if $P(t_x^* < \infty | X_0 = a) = 1$, for all a and x in $\mathcal{A}$, the process is recurrent. Specifically, it is null recurrent if $\mathrm{E}(t_x^* |\ X_0 = a) = \infty$ for all a and x in $\mathcal{A}$. Alternatively, if $\mathrm{E}(t_x^* |\ X_0 = a) < \infty$, the process is positive recurrent.

We assume recurrence. In terms of the shape of the drift and diffusion function $\mu(.)$ and $\sigma(.)$, the process is recurrent if, and only if, $\lim_{b \to l} S(b) = -\infty$ and $\lim_{b \to u} S(b) = \infty$, where $S(b) =$

$\int_c^b \exp\left\{\int_c^x \left[-\frac{2\mu(s)}{\sigma^2(s)}\right] ds\right\} dx$ (with $c \in \mathcal{A}$) is the so-called scale function. Positive recurrence requires the speed (or invariant) measure $m(dx) = \frac{2dx}{S'(x)\sigma^2(x)} = \pi(dx)$ to be integrable over $\mathcal{A}$, i.e., $m(\mathcal{A}) = \int_{\mathcal{A}} m(x)dx < \infty$. In this case, the stationary density of the process is $p(x) = \frac{m(x)}{m(\mathcal{D})}$ for x in $\mathcal{A}$. We refer the reader to Bandi and Phillips (2010) for further discussions.

Assume the process X_t is observed at discrete points $\{t_1, t_2, .., t_n\}$ in the time interval $[0, T]$ with $T \geq T_0$, where T_0 and T are positive constants. Also, assume the data is equispaced.
Then, $\{X_{\Delta_{n,T}}, X_{2\Delta_{n,T}}, X_{3\Delta_{n,T}}, ..., X_{n\Delta_{n,T}}\}$ are n observations, i.e., the diffusion's skeleton, at $\{t_1 = \Delta_{n,T}, t_2 = 2\Delta_{n,T}, t_3 = 3\Delta_{n,T}, ..., t_n = n\Delta_{n,T}\}$ with $\Delta_{n,T} = T/n$. In the limit, let $n \to \infty$, $T \to \infty$, and $\Delta_{n,T} = T/n \to 0$.

As in the previous section, we work with additive functionals. For a $\pi-$integrable, non-negative function $f(.)$, we have

$$\Delta_{n,T} \sum_{i=1}^{n} f\left(X_{i\Delta_{n,T}}\right) \overset{a.s.}{\sim} \int_0^T f(X_s)ds,$$

uniformly in T as $\Delta_{n,T} \to 0$. Further, Theorem 3.1 in Löcherbach and Loukianova (2009) implies that

$$\limsup_{T\to\infty} \frac{\int_0^T f(X_s)ds}{v\left(\frac{T}{L_2(v(T))}\right) L_2(v(T))} = C_X \int_{-\infty}^{\infty} f(X_s)\pi(ds)$$

where $C_X > 0$ is a process-specific constant, $v(T) = \mathrm{E}_\varphi\left(\int_0^T f(X_s)ds\right) \sim T^p \log(T)$ for any initial measure φ, and $L_2(v(T)) = L_2\lambda = \log\log\max\{\lambda, e^e\}$ with $\lambda \geq 0$. Thus,

$$\limsup_{T,n\to\infty} \frac{\Delta_{n,T}\sum_{i=1}^{n} f\left(X_{i\Delta_{n,T}}\right)}{v\left(\frac{T}{L_2(v(T))}\right) L_2(v(T))} = C_X \int_{-\infty}^{\infty} f(X_s)\pi(ds)$$

with $\Delta_{n,T} \to 0$.

We can now proceed as earlier. Under null recurrence ($H_0 : p < 1$):

$$\frac{\Delta_{n,T}\sum_{i=1}^{n} f\left(X_{i\Delta_{n,T}}\right)}{T} = O_{a.s.}\left(\frac{b_p(T)}{T}\right) = o_{a.s.}(1).$$

where $b_p(T) = v\left(\frac{T}{L_2(v(T))}\right) L_2(v(T))$. Under positive recurrence ($H_A : p = 1$),

$$\frac{\Delta_{n,T}\sum_{i=1}^{n} f\left(X_{i\Delta_{n,T}}\right)}{T} = O_{a.s.}(1).$$

Define now the statistics

$$V_{R,n,T} = \frac{2}{\sqrt{R}} \sum_{j=1}^{R} \left(1\left\{\eta_j \leq \lambda\left(\frac{\Delta_{n,T}\sum_{i=1}^{n} f\left(X_{i\Delta_{n,T}}\right)}{T}\right)\right\} - \frac{1}{2}\right),$$

where the η_js are, as earlier, R standard normal draws. We have the following:

Theorem 4. *Let $X_t, t \in \mathcal{R}^+$ be a p-null recurrent diffusion process. Let f be non-negative and such that $f \in \mathcal{L}^1(\mathbf{E}, \mathcal{E}, \pi)$ and $\lambda(x)$ be a monotonically-decreasing to zero as $x \to 0$. Assume $R, n, T \to \infty$, $\Delta_{n,T} = T/n \to 0$. Also, assume $\sqrt{R}\lambda\left(\frac{b_{\overline{p}}(T)}{T}\right) \to 0$ with $b_p(T) = \left(\frac{T}{\log\log(L(T)T^p)}\right)^p L\left(\frac{T}{\log\log(L(T)T^p)}\right)\log\log(L(T)T^p)$ and $L(T)$ slowly-varying at infinity. Then,*
(i) Under $H_0 : p \leq \overline{p} < 1$,

$$V_{R,n,T} \xrightarrow{d^*} N(0,1) \quad a.s.-P.$$

(ii) Under $H_A : p = 1$, there are constants $c_1, c_2 > 0$ so that

$$P^*\left(R^{-1/2+c_1}V_{R,n,T} > c_2\right) \to 1 \quad a.s.-P.$$

Note that the admissible divergence rate of the number of random draws R should now depend on the time span T rather than on the number of observations (n) in the sample.

6 Size and power

We consider 5% level tests and simulate three data generating processes.

Model I A classical autoregressive process, viz.

$$X_t = \rho X_{t-1} + u_t.$$

We set $x_0 = 0$ and let u_t be i.i.d. $N(0, \sigma^2)$ with three values of σ, namely 1, 100, and 0.01. Under $H_0 : \rho = 1$ the invariant measure of the process $\pi(dx) \sim dx$.

Model II An affine diffusion process with $\mu(x) = \kappa(\theta - x)$ and $\sigma(x) = \sigma$, viz.

$$dX_t = \kappa(\phi - X_t)dt + \sigma dW_t.$$

We set $\phi = 0$, $x_0 = 0$, and $\sigma = \sqrt{0.008742}$. The process is simulated after discretization using a classical Milshtein scheme. The case $\kappa = 0$ gives null recurrence of the unit-root type. Under $H_0 : \kappa = 0$, the invariant measure is, again, $\pi(dx) \sim dx$.

Model III A "natural scale" diffusion with $\mu(x) = 0$ and $\sigma(x) = \sigma(1 + x^2)^\gamma$, viz.

$$dX_t = \sigma(1 + X_t^2)^\gamma dW_t.$$

We set $\sigma = 1$. Again, the process is simulated after discretization using a Milshtein scheme. For $\gamma \leq \frac{1}{2}$ the process is null-recurrent. For $\gamma > \frac{1}{2}$ the process is positive-recurrent. The invariant measure is $\pi(dx) \sim \frac{dx}{(1+x^2)^{2\gamma}}$.

In order to preserve the conditioning on the sample, we simulate a specific sample and calculate $1,000$ statistics (conditional on that sample) based on $1,000$ draws of an R-vector of standard normal

draws. This procedure gives us one rejection frequency, conditional on the sample. The same method is implemented multiple times (100 times) before averaging the rejection frequencies across the 100 samples. In the case of Model I, in agreement with much existing work on unit-root testing, results are based on samples of moderate length. We set n equal to 500 and increase the sample size to $n = 1,000$ to evaluate the impact of this increase. In the case of Model II and Model III, we set the sample size equal to $n = 5,000$. This larger sample size is typical of the continuous-time finance literature in which the proposed models have been estimated. It corresponds to 40 years of daily data. The quantity R is set equal to $1,000$ but is sometimes extended to $10,000$ to assess the gain in power, and the corresponding loss in size, of an increase in the number of random draws. The functions $\lambda(x)$ and $f(x)$ are set equal to x^θ, for some $\theta > 0$, and $\frac{2}{1+x^2}$, respectively. The choice of $f(x)$ guarantees π-integrability in all three cases. We focus on a nonstationary null. As discussed, the test is immediate to code up and hinges on tabulated critical values, i.e., those of the standard normal distribution. We compare it to the classical Dickey-Fuller test as well as to Phillips' Z test (Phillips, 1987). The latter is computed using a Parzen kernel and an AR(1) filter to estimate the spectrum.

6.1 Results

Even though the asymptotic properties of the test are not affected by the choice of $\lambda(.)$ and $f(.)$, provided these functions satisfy the conditions listed in the theorems, finite sample performance is naturally influenced by these choices and requires care. While a complete discussion of these issues is beyond the scopes of the present paper, we intend to give the reader general principles about how to implement the test in practise.

We begin with Model I (Table 1 and 2). We set $\theta = 5$, thereby obtaining $\lambda(x) = x^5$. It is intuitive that a small σ^2 may easily translate into an oversized test. Similarly, a large σ^2 will likely translate into an undersized test. The reason for this is that a small σ^2 will result in observations which do not move away from 0 fast enough in a small sample, thereby yielding values of $\frac{2}{1+x^2}$ which remain in a neighborhood of about 2. This means that $\sum_{j=1}^{n} \frac{1}{1+X_j^2}$ may grow roughly with the sample size and lead to rejections of the null, even if $\rho = 1$. Conversely, a large σ^2 will make the process drift away from 0 quickly even in a small sample, thereby yielding small values of $\frac{2}{1+x^2}$ and, hence, excessively "nonstationary" dynamics in a finite sample, even when $|\rho| < 1$. To this extent, in order to eliminate the finite sample impact of the shocks' variance, we first standardize the data by the (estimated) standard deviation of the shocks. This is going to lead to $\frac{2}{1+x^2}$ values which, in light of the unit variance properties of the standardized data, will be in the vicinity of 1 when the data is stationary and will be closer to zero under the null. As we show below, the proposed correction achieves a finite sample invariance to the shocks' variance which mirrors the asymptotic invariance of the proposed tests as well as that of more classical tests for unit roots.

Table 1 reports size and power for alternative choices of σ^2. Size is very satisfactory. As expected in light of the superior efficiency of classical unit root tests in the context of linear processes, power is a bit smaller than for the existing tests. Increases in the number of random draws R (from $1,000$ to

10,000, in our case) will, however, yield slight size distortions but substantial power increases leading to an overall performance which is comparable to that of extant, popular alternatives (Table 2). As expected, increasing the number of observations leads, in general, to superior performance across the board. The obvious size improvements might, however, be accompanied by slight deteriorations in power for very close alternatives (see Table 2).

We now turn to Model II (Table 3). We only report the case $\sigma^2 = 0.008742$, which is typical of the literature on short-term interest rate estimation using daily data in continuous time (see, e.g., Pritsker, 1998). As done in the case of Model I, in order to improve finite sample performance, relying on the linearity of the data generating process, we standardize the data by the estimated shocks' standard deviation. Alternative values of σ^2 could therefore be handled similarly and would yield, as for Model I, identical results. The integrable function $f(x)$ is set equal to $\frac{2}{1+x^2}$, as earlier. We employ $n = 5,000$, a typical sample size in the continuous-time literature, and set, as before, $R = 1,000$. The coefficient θ is now chosen equal to 3, rather than 5. The reason for this modification has to do with the larger sample size. If the sample is large, the function $\lambda(.)$ will play less of a role. Asymptotically, in fact, one could even dispense with $\lambda(.)$ or, equivalently, set $\theta = 1$ if assuming that $\lambda(.)$ is a power function. Said differently, the smaller the sample size, the faster $\lambda(.)$ has to go to zero to aid the asymptotics. The larger the sample size, the more ineffective the function $\lambda(.)$ has to be in order to avoid undersizing and power losses. Said differently, if assuming a power function, we advocate decreasing the size of θ as n increases. In the case of Model II, we vary κ to assess size and power. The implied, given choices of κ, autoregressive parameters are reported in the last line of Table 3. The results are analogous to those derived from Model I. The test is properly sized, but is less powerful, than classical alternatives in the literature. Both the Dickey-Fuller test and Phillips' Z test have very high power for local alternatives ($\kappa = 2$) given the assumed sample size. Needless to say, an increase in the number of random draws R would increase the power of the proposed test, as earlier, while determining some size deterioration.

Table 4 reports results for a nonlinear alternative, i.e., Model III. The parameter γ controls, in this case, the stationarity properties of the process. If $0 < \gamma \leq 0.5$, the process is null recurrent. It is positive recurrent if $\gamma > 0.5$. This is a case of volatility-induced stationarity, a specification introduced in the context of interest modelling in continuous time (Conley, Hansen, Luttmer and Scheinkman, 1997). We assess size by setting γ equal to 0.1 and 0.2 and power by setting γ equal to 0.6, 0.7, and 0.8. In agreement, again, with the continuous-time literature and Model II, the sample size is 5,000 observations. The number of random draws is 1,000. The parameter θ is, again, equal to 3. We find that traditional tests have very little power in this case. This is true across the board, not only for local alternatives ($\gamma = 0.6$). Consistent with this observation, the autoregressive parameter is always estimated at values that are extremely close to 1. Conversely, the additive-functional based test is only slightly oversized but has extremely high power. This result is striking and points to the inability of traditional coefficient-based tests to adapt to nonlinear structures in the data. We find, for instance, that with ten times as many observations (namely, with a sample size of 50,000 observations) the local power of the Dickey-Fuller test would still be around 30%. This is in sharp contrast with the 63.2%

rejection probability of the test that we propose for a more realistic sample size of $5,000$ observations.

7 Final remarks

As we emphasize above, the tests are *asymptotically* invariant to the magnitude of the process' shocks. They are, however, not invariant in finite samples since the scale of the function $f(.)$ depends on the variability of X_t. While in the nonlinear case one does not have, in general, a clean way to standardize the data using the estimated variance of the process' shocks, it may still help to re-scale X_t by a nonparametric estimator of its conditional variance suitably averaged over the evaluation points in order not to alter the regularity properties of the chain. The conditional variance may be identified along the lines of Bandi, Corradi, and Wilhelm (2011) who, for classes of discrete-time models analogous to the ones covered in this paper, discussed consistency and asymptotic normality of a nonparametric conditional variance estimator without requiring assumptions on the degree of recurrence (for the continuous-time case, we refer to Bandi and Phillips, 2003).

There are alternative ways in which randomized nonstationarity tests can be constructed. The issue of finite sample invariance should have implications for the construction of the tests in the presence of alternative, possible, test specifications. It should also influence empirical implementations for any chosen specification.

We start with the former, i.e., test construction. As pointed out by a referee, whom we thank, a statistic having a normal limiting distribution under the null, and diverging under the alternative, conditionally on the sample, could, for instance, also be defined as

$$\overline{V}_{R,n} = \xi + \sqrt{R}\lambda\left(\frac{1}{n}\sum_{j=1}^{n} f(X_j)\right),$$

where ξ is a simulated $\mathrm{N}(0,1)$ draw. Because $\sqrt{R}\lambda\left(\frac{1}{n}\sum_{j=1}^{n} f(X_j)\right)$ is almost surely zero under H_0, provided $R = o\left(\lambda^{-2}\left(\frac{b_{\overline{p}}(n)}{n}\right)\right)$, and diverges almost surely under H_A, $\overline{V}_{R,n}$ has the same asymptotic properties as $V_{R,n}$. The advantage of $\overline{V}_{R,n}$ is that ξ is exactly normal, rather than asymptotically normal as the first term in Eq. (2). Such a statistic, which is logically identical to the one we propose, is easy to compute, provides additional intuition for the identical conditions on R and $\lambda(.)$ illustrated in the theorems, and complements our proposed $V_{R,n}$. However, due to the fact that both size and power depend on the magnitude of $\lambda\left(\frac{1}{n}\sum_{j=1}^{n} f(X_j)\right)$ for a finite n, we believe that the finite sample scale of $f(.)$ will affect $\overline{V}_{R,n}$ more severely than $V_{R,n}$. Simulations, not reported here for conciseness, show that - for the same choices of R, $\lambda(.)$, and $f(.)$ - $\overline{V}_{R,n}$ is oversized as compared to $V_{R,n}$. The reason for this outcome is that the relative impact of the magnitude of $\lambda\left(\frac{1}{n}\sum_{j=1}^{n} f(X_j)\right)$ on $V_{R,n}$ is attenuated by the use of the indicator function. The component which multiplies $\sqrt{R}$ in $V_{R,n}$ is, in fact, between 0 and $\frac{1}{2}$, whereas the component multiplying $\sqrt{R}$ in $\overline{V}_{R,n}$ is also positive and, in theory, arbitrarily large.[3] In this

[3] In the unit-root case (Model I) above, for example, $\lambda\left(\frac{1}{\sqrt{n}}\sum_{j=1}^{n} f(X_j)\right) \sim \left(\int_{-\infty}^{\infty} f(x)\pi(x)dx\right)^{\theta} \sim (C\pi)^{\theta}$, for an unrestricted $C > 0$, where π in the last expression denotes the number π, if $f(x) \sim \frac{1}{1+x^2}$ and $\lambda(x) \sim x^{\theta}$, since $\pi(dx) \sim dx$.

sense, we conjecture that $\overline{V}_{R,n}$ is, in general, more sensitive than $V_{R,n}$ to scaling issues and the related selection of $\lambda(.)$ and $f(.)$. Hence, it is less preferable in practise.

We now turn to implementation. The choice of $\lambda(.)$, $f(.)$ and R is important and non trivial. It is a price to pay to handle nonlinear dynamics. As outlined above, one may set $\lambda(x) = x^{\theta}$, where θ ranges between 2 and 5, say, with a preference for a smaller θ the larger the sample. Provided π-integrability is guaranteed, the choice of $f(x)$ may not be limited to the class of functions $\frac{a}{1+x^2}$, with $a > 0$, used in our Monte Carlo exercise. As emphasized above, re-scaling the data and selecting an appropriate $f(.)$, so as to attenuate finite sample scaling issues for a smaller sample size, appear important. As for R, one needs $\sqrt{R}\lambda\left(\frac{b_{\overline{p}}(n)}{n}\right) \to 0$, asymptotically, for correct sizing. The larger R, the higher power is. Hence, in principle, one should select $R \sim \lambda^{-(2-\varepsilon)}\left(\frac{b_{\overline{p}}(n)}{n}\right)$, with $\varepsilon > 0$ as small as possible. The focus of this paper is on laying out ideas and providing preliminary recommendations for implementation. The design of adaptive rules to select $\lambda(.)$, $f(.)$ and R is important and will be the subject of future work.

8 Conclusions

A great deal of work in econometrics, particularly in financial econometrics, has been focusing on nonlinear models. Stationarity is often tested up-front, and subsequently invoked if supported by classical tests, as a way to justify inferential procedures which rely on it either for identification or to derive limiting results. This sequential approach is pragmatic and defensible. However, it generates a theoretical inconsistency between the use of classical stationarity/nonstationarity tests, which assume linearity before inference begins, and subsequent nonlinear inference. To address this issue, this paper introduces, and formalizes, initial ideas for nonstationarity testing based on sample conditioning and randomization. We show how randomization and conditional inference can be jointly put to work to derive nonstationarity tests which are robust to nonlinearities of unknown form. In particular, we show how one may handle situations in which well-defined parameter-based nonstationarity tests, as in the unit-root tradition, can not be derived.

While randomization has some history in statistics, its use for occupation density-based nonstationarity testing is, to the best of our knowledge, novel. We use it here to evaluate relative "magnitudes," namely the magnitude of sums of integrable functions of the data as compared to the magnitude of the sample size itself. We show that, when properly conducted, this comparison will give us information, under mild assumptions, about stationary/nonstationarity behavior irrespective of the linearity properties of the underlying data-generating process. Much remains to be done. While the class of processes which we evaluated is wider than that covered by classical unit-root tests, it now seems important to broaden the scope of application further.

9 Appendix

Proof of Theorem 1. Given Assumption A, by Proposition 2,

$$\limsup_{n\to\infty} \frac{\sum_{j=1}^{n} f\left(X_j\right)}{b_p(n)} = \frac{\Gamma(p+1)}{p^p(1-p)^{1-p}} \int f(x)\pi(dx) \qquad a.s.,$$

where $b_p(n) = \left(\frac{n}{\log\log(L(n)n^p)}\right)^p L\left(\frac{n}{\log\log(L(n)n^p)}\right)\log\log(L(n)n^p)$. Hence, under the null of $p<1$, $\frac{\sum_{j=1}^{n} f(X_j)}{n} = O_{a.s.}\left(\frac{b_p(n)}{n}\right) = o_{a.s.}(1)$. First, note that for all j, conditional on the sample, $v_{j,n} = \frac{\eta_j}{\lambda\left(\frac{\sum_{j=1}^{n} f(X_j)}{n}\right)} \overset{d}{\sim} N\left(0, \frac{1}{\lambda^2\left(\frac{\sum_{j=1}^{n} f(X_j)}{n}\right)}\right)$. Let

$$\Omega_n = \left\{\omega : \lambda^{-1}\left(\frac{\sum_{j=1}^{n} f\left(X_j\right)}{n}\right) > \varepsilon > 0\right\}$$

so that, under H_0, $P\left(\lim_{n\to\infty}\Omega_n\right) = 1$. We shall proceed conditional on $\omega\in\Omega_n$. We obtain

$$V_{R,n} = \frac{2}{\sqrt{R}}\sum_{j=1}^{R}\left(1\left\{v_{j,n}\le 1\right\} - \mathrm{E}^*\left(1\left\{v_{j,n}\le 1\right\}\right)\right) + \frac{2}{\sqrt{R}}\sum_{j=1}^{R}\left(\mathrm{E}^*\left(1\left\{v_{j,n}\le 1\right\}\right) - \frac{1}{2}\right),$$

where $\mathrm{E}^*\left(1\left\{v_{j,n}\le 1\right\}\right) = 1/2 + P^*\left(0\le v_{j,n}\le 1\right)$. Now,

$$\begin{aligned}
& P^*\left(0\le v_{j,n}\le 1\right) \\
= & \frac{1}{\left(2\pi\lambda^{-2}\left(\frac{\sum_{j=1}^{n} f(X_j)}{n}\right)\right)^{1/2}}\int_0^1 \exp\left(-\frac{x^2}{2\lambda^{-2}\left(\frac{\sum_{j=1}^{n} f(X_j)}{n}\right)}\right)dx \\
= & O\left(\lambda\left(\frac{\sum_{j=1}^{n} f\left(X_j\right)}{n}\right)\right) \\
= & O\left(\lambda\left(\frac{b_p(n)}{n}\right)\right), \qquad (4)
\end{aligned}$$

Thus, for all $\omega\in\Omega_n$,

$$V_{R,n} = \frac{2}{\sqrt{R}}\sum_{j=1}^{R}\left(1\left\{v_{j,n}\le 1\right\} - \mathrm{E}^*\left(1\left\{v_{j,n}\le 1\right\}\right)\right) + O\left(\sqrt{R}\lambda\left(\frac{b_p(n)}{n}\right)\right),$$

where the last term is $o(1)$ since, for all $p\le\overline{p}$, $\sqrt{R}\lambda\left(\frac{b_p(n)}{n}\right)\to 0$ as $n, R\to\infty$. Given Eq. (4), and recalling that $\mathbf{E}^*\left(v_{j,n}v_{s,n}\right) = 0$ for $s\neq j$ conditionally on the sample,

$$\begin{aligned}
& \mathrm{Var}^*\left(\frac{1}{\sqrt{R}}\sum_{j=1}^{R}\left(1\left\{v_{j,n}\le 1\right\} - \mathbf{E}^*\left(1\left\{v_{j,n}\le 1\right\}\right)\right)\right) \\
= & \frac{1}{R}\sum_{j=1}^{R}\left(\mathrm{E}^*\left(1\left\{v_{j,n}\le 1\right\} - \mathrm{E}^*\left(1\left\{v_{j,n}\le 1\right\}\right)\right)^2\right) \\
= & \frac{1}{R}\sum_{j=1}^{R}\left(\mathrm{E}^*\left(1\left\{v_{j,n}\le 1\right\} - P^*\left(v_{j,n}\le 1\right)\right)^2\right) \\
= & P^*\left(v_{j,n}\le 1\right)\left(1 - P^*\left(v_{j,n}\le 1\right)\right) \\
= & \left(1/2 + O\left(\lambda\left(\frac{b_p(n)}{n}\right)\right)\right)\left(1/2 + O\left(\lambda\left(\frac{b_p(n)}{n}\right)\right)\right) \\
= & 1/4 + O\left(\lambda^2\left(\frac{b_p(n)}{n}\right)\right).
\end{aligned}$$

Thus, $V_{R,n}$ is correctly standardized for a classical central limit theory for iid sequences to apply and $V_{R,n} \overset{d^*}{\to} N(0,1)$. Now, let

$$\Omega_n^+ = \left\{ \omega : \lambda^{-1} \left(\frac{\sum_{j=1}^n f(X_j)}{n} \right) < \Delta,\ 0 < \Delta < \infty \right\}$$

so that, under H_A, $P(\lim_{n\to\infty} \Omega_n^+) = 1$. For $\omega \in \Omega_n^+$, $\lambda^{-1}\left(\frac{\sum_{j=1}^n f(X_j)}{n}\right) \overset{a.s.}{\to} M$. Hence, $v_{j,n} \overset{d^*}{\to} N(0, M^2)$. As in the null case, the statistic writes as

$$\begin{aligned} & \frac{2}{\sqrt{R}} \sum_{i=1}^{R} \left(1\{v_{j,n} \leq 1\} - \frac{1}{2} \right) \\ = & \frac{2}{\sqrt{R}} \sum_{i=1}^{R} \left(1\{v_{j,n} \leq 1\} - \mathrm{E}^*\left(1\{v_{j,n} \leq 1\}\right) \right) + 2\sqrt{R} \left(\mathrm{E}^*\left(1\{v_{j,n} \leq 1\}\right) - \frac{1}{2} \right), \end{aligned} \tag{5}$$

where, again, $\mathrm{E}^*\left(1\{v_{j,n} \leq 1\}\right) = 1/2 + P^*(0 \leq v_{j,n} \leq 1)$ with $P^*(0 \leq v_{j,n} \leq 1)$ as in Eq. (4). Now, for any $\omega \in \Omega_n^+$, the first term on the right-hand side of Eq. (5) converges in distribution to a (non-standard) zero-mean normal random variable. However, $P^*(0 \leq v_{j,n} \leq 1) > 0$ and, thus, the second term diverges at rate $\sqrt{R}$. ■

Proof of Theorem 2. Let $v_{j,n} = \frac{\eta_j}{\lambda\left(\frac{b_{\overline{p}}(n)}{\sum_{j=1}^n f(X_j)}\right)} \overset{d}{\sim} N\left(0, \lambda^{-2}\left(\frac{b_{\overline{p}}(n)}{\sum_{j=1}^n f(X_j)}\right)\right)$. Now, for any sample, under the null, by Proposition 2, $\lambda\left(\frac{b_{\overline{p}}(n)}{\sum_{j=1}^n f(X_j)}\right) = O\left(\lambda\left(\frac{b_{\overline{p}}(n)}{n}\right)\right)$. On the other hand, for any sample, under the alternative, $\lambda\left(\frac{b_{\overline{p}}(n)}{\sum_{j=1}^n f(X_j)}\right) = O\left(\lambda\left(\frac{b_{\overline{p}}(n)}{b_p(n)}\right)\right)$. The statement, then, follows by the same argument used in the proof of Theorem 1. ■

Proof of Theorem 3. We solely have to prove that $\frac{1}{n}\sum_{t=1}^n f(X_t) = O_{a.s.}\left(\sqrt{\frac{\log\log n}{n}}\right)$. Again, the statement of the theorem will then follow from the same arguments leading to Theorem 1. To this extent, we show the result for the case $f(X_t) = 1\{a \leq X_t \leq b\}$. Because the indicator function of a compact set is dense in the class of bounded functions, the proof is without loss of generality. Let $B_t = \omega_0 W_t$ with W_t a standard Brownian motion. Let $A = [a, b]$ and, with an abuse of notation, define $A/\sqrt{n} = [a/\sqrt{n}, b/\sqrt{n}]$ and

$$\frac{A}{\sqrt{n}} - \left(\frac{B_t}{\sqrt{n}} - \frac{X_t}{\sqrt{n}}\right) = \left[\frac{a}{\sqrt{n}} - \left(\frac{B_t}{\sqrt{n}} - \frac{X_t}{\sqrt{n}}\right), \frac{b}{\sqrt{n}} - \left(\frac{B_t}{\sqrt{n}} - \frac{X_t}{\sqrt{n}}\right)\right].$$

Finally, let $\Phi\left(\frac{A}{\sqrt{n}}\right) = \Pr\left(\frac{a}{\sqrt{n}} \leq \omega_0 Z \leq \frac{b}{\sqrt{n}}\right)$, with Z denoting a standard normal random variable. The function $\Phi\left(\frac{A}{\sqrt{n}} - \left(\frac{B_t}{\sqrt{n}} - \frac{X_t}{\sqrt{n}}\right)\right)$ is defined analogously. Let, also, ϕ be the density function associated with Φ. We have,

$$\begin{aligned} & \frac{1}{n}\sum_{t=1}^n f(X_t) = \frac{1}{n}\sum_{t=1}^n 1\left\{\frac{B_t}{\sqrt{n}} \in \frac{A}{\sqrt{n}}\right\} \\ & + \frac{1}{n}\sum_{t=1}^n \left(1\left\{\frac{B_t}{\sqrt{n}} \in \left(\frac{A}{\sqrt{n}} - \left(\frac{B_t}{\sqrt{n}} - \frac{X_t}{\sqrt{n}}\right)\right)\right\} - 1\left\{\frac{B_t}{\sqrt{n}} \in \frac{A}{\sqrt{n}}\right\}\right) \end{aligned}$$

and

$$\begin{aligned}
\frac{1}{n}\sum_{t=1}^{n} f\left(X_t\right) &= \frac{1}{n}\sum_{t=1}^{n}\left(1\left\{\frac{B_t}{\sqrt{n}} \in \frac{A}{\sqrt{n}}\right\} - \Phi\left(\frac{A}{\sqrt{n}}\right)\right) + \Phi\left(\frac{A}{\sqrt{n}}\right) \\
&\quad + \left[\frac{1}{n}\sum_{t=1}^{n}\left(1\left\{\frac{B_t}{\sqrt{n}} \in \left(\frac{A}{\sqrt{n}} - \left(\frac{B_t}{\sqrt{n}} - \frac{X_t}{\sqrt{n}}\right)\right)\right\} - \Phi\left(\frac{A}{\sqrt{n}} - \left(\frac{B_t}{\sqrt{n}} - \frac{X_t}{\sqrt{n}}\right)\right)\right)\right. \\
&\quad \left. - \frac{1}{n}\sum_{t=1}^{n}\left(1\left\{\frac{B_t}{\sqrt{n}} \in \frac{A}{\sqrt{n}}\right\} - \Phi\left(\frac{A}{\sqrt{n}}\right)\right)\right] \\
&\quad + \frac{1}{n}\sum_{t=1}^{n}\left(\Phi\left(\frac{A}{\sqrt{n}} - \left(\frac{B_t}{\sqrt{n}} - \frac{X_t}{\sqrt{n}}\right)\right) - \Phi\left(\frac{A}{\sqrt{n}}\right)\right) \\
&= I_n + II_n + III_n + IV_n.
\end{aligned}$$

The strong invariance principle for the Brownian motion ensures that $I_n = O_{a.s.}\left(\sqrt{\frac{\log\log n}{n}}\right)$. It is immediate to see that $II_n = O\left(\frac{1}{\sqrt{n}}\right)$. Given Assumption B, because of the strong stochastic equicontinuity of the indicator function, $III_n = O_{a.s.}\left(\frac{1}{\sqrt{n}}\right)$. Finally, letting $d_n \in \left(\frac{a}{\sqrt{n}} - \left(\frac{B_t}{\sqrt{n}} - \frac{X_t}{\sqrt{n}}\right), \frac{b}{\sqrt{n}} - \left(\frac{B_t}{\sqrt{n}} - \frac{X_t}{\sqrt{n}}\right)\right)$, in light of Assumption B,

$$IV_n = \frac{1}{n}\sum_{t=1}^{n}\phi\left(d_n\right)\left(\frac{B_t}{\sqrt{n}} - \frac{X_t}{\sqrt{n}}\right) = O_{a.s.}\left(\sqrt{\frac{\log\log n}{n}}\right),$$

because of the functional law of the iterated logarithm for strong mixing processes (e.g., Eberlein, 1986). Thus,

$$\frac{1}{n}\sum_{t=1}^{n} f\left(X_t\right) = O_{a.s.}\left(\sqrt{\frac{\log\log n}{n}}\right).$$

■

Proof of Theorem 4. Given Theorem 3.1. in Löcherbach and Loukianova (2009), it follows from the same argument as that of the proof of Theorem 1. In this case, however, the rate of growth of the occupation measure depends on the time span T rather than on the sample size n. ■

References

1. AÏT-SAHALIA, Y., (1996). Testing Continuous Time Models of the Spot Interest Rate. *Review of Financial Studies* 9, 385-426.

2. BANDI, F.M., and P.C.B. PHILLIPS (2003). Fully Nonparametric Estimation of Scalar Diffusion Models. *Econometrica* 71, 241-283.

3. BANDI, F.M., and P.C.B. PHILLIPS (2010). Nonstationary Continuous-Time Processes. *Handbook of Financial Econometrics, Elsevier.*

4. BANDI, F.M., V. CORRADI and G. MOLOCHE (2009). Bandwidth Selection for Continuous Time Markov Processes. *Working paper.*

5. BANDI, F.M., V. CORRADI and D. WILHELM (2011). Data-driven Bandwidth Selection for Nonparametric Nonstationary Regressions. *Working paper.*

6. BREITUNG, J. (2002). Nonparametric Tests for Unit Roots and Cointegration. *Journal of Econometrics* 108, 343-363.

7. CHEN, X. (1999). How Often Does a Harris Recurrent Markov Chain Recur? *Annals of Probability* 27, 1324-1346.

8. CONLEY, T.G., L.P. HANSEN, E.G.J. LUTTMER and J.A. SCHEINKMAN (1997). Short-term Interest Rates as Subordinated Diffusions. *Review of Financial Studies* 10, 525-577.

9. CORRADI, V., and N. SWANSON (2006). The Effects of Data Transformation on Common Cycle, Cointegration, and Unit Root Tests: Monte Carlo and a Simple Test. *Journal of Econometrics* 132, 195-229.

10. DOLADO, J.J., J. GONZALO and L. MAYORAL (2002). A Fractional Dickey-Fuller Test for Unit Roots. *Econometrica* 70, 1963-2006.

11. DUFOUR, J.M. and J.F. KIVIET (1996). Exact Tests for Structural Changes in First-Order Dynamic Models. *Journal of Econometrics* 70, 39-68.

12. EBERLEIN, E (1986). On String Invariance Principles Under Dependence Assumptions. *Annals of Probability* 14, 260-270.

13. ELLIOTT, G., T.G. ROTHENBERG and J.H. STOCK (1996). Efficient Tests for an Autoregressive Unit Root. *Econometrica* 64, 813-836.

14. DAVIDSON, R. and E. FLACHAIRE (2008). The Wild-Bootstrap Tamed at Last. *Journal of Econometrics* 146, 162-169.

15. GONCALVES, S. and N. MEDDAHI (2009). Bootstrapping Realized Volatility. *Econometrica* 77, 283-306.

16. GUERRE, E. (2004). Design-Adaptive Pointwise Nonparametric Regression Estimation for Recurrent Markov Time Series. *Working Paper.*

17. HAJEK, J. and SIDAK Z. (1967), *Theory of Rank Tests.* Academic Press, New York.

18. HANSEN, B.E. (1996). Inference when a Parameter is Not Identified under the Null Hypothesis. *Econometrica* 64, 413-430.

19. KARLSEN, H.A. and V. TJOSTHEIM (2001). Nonparametric Estimation in Null Recurrent Time Series. *Annals of Statistics* 29, 372-416.

20. KWIATKOWSKI, D., P.C.B. PHILLIPS, P. SCHMIDT, and Y. SHIN (1992). Testing the Null Hypothesis of Stationarity Against the Alternative of a Unit Root: How Sure Are We That Economic Time Series Have a Unit Root? *Journal of Econometrics* 54, 159-178.

21. LÖCHERBACH, E. and D. LOUKIANOVA (2009). The Law of the Iterated Logarithm for Additive Functionals and Martingale Additive Functionals of Harris Recurrent Markov Processes. *Stochastic Processes and Their Applications* 119, 2312-2335.

22. LUTKEPOHL, H. and M.M. BURDA (1997). Modified Wald Tests under Nonregular Conditions. *Journal of Econometrics* 78, 315-332.

23. MEYN, S.P., and R.L. TWEEDIE (1993). Stability of Markovian Processes 2: Continuous Time Processes and Sampled Chains. *Advances in Applied Probability* 25, 487-517.

24. MÜLLER, U.K. (2008) The Impossibility of Consistent Discrimination Between $I(0)$ and $I(1)$ Processes. *Econometric Theory,* 24, 616.

25. PEARSON, E.S. (1950). On Questions Raised by the Combination of Tests Based on Discontinuous Distributions. *Biometrika* 37, 383-398.

26. PHILLIPS, P.C.B. (1987). Time Series Regressions with a Unit Root. *Econometrica* 55, 277-301.

27. PHILLIPS, P.C.B., and Z. XIAO (1998). A Primer on Unit Root Testing, *Journal of Economic Surveys* 12, 423-469.

28. PRITSKER, M., (1998), Nonparametric Density Estimators and Tests of Continuous Time Interest Rate Models, *Review of Financial Studies* 11, 449-487.

29. SCHIENLE, M. (2008). Nonparametric Nonstationary Regression with Many Covariates. Working paper.

30. STEVENS, W.L. (1950). Fiducial Limits of the Parameters of a Discontinuous Distribution. *Biometrika* 37, 117-129.

31. TOCHER, K.D. (1950). Extension of the Neyman-Pearson Theory of Tests to Discontnuous Variates. *Biometrika* 37, 130-144.

32. WANG, Q. and P.C.B. PHILLIPS (2009a). Asymptotic Theory for Local Time Density Estimation and Nonparametric Cointegrating Regression, *Econometric Theory* 25, 710-738.

33. WANG, Q. and P.C.B. PHILLIPS (2009b). Structural Nonparametric Cointegrating Regression. *Econometrica* 77, 1901-1948.

Table 1.

	Size	**Power**				
	ρ=1	**ρ=0.99**	**0.98**	**0.97**	**0.96**	**0.95**
BC (σ = 1)	6.1%	11.7%	21.4%	38.2%	51.2%	70.3%
BC (σ = 100)	6.2%	11.8%	20.5%	34.4%	48.6%	68.5%
BC (σ = 0.01)	5.5%	9.6%	19.7%	32.8%	59.1%	76.6%
DF (σ = 1)	4.75%	11.3%	30%	58.7%	84.3%	96.3%
Z (σ = 1)	5.6%	21%	48%	79.1%	95%	99.3%

Model I. The number of simulated samples is 100. The number of simulations per sample is 1,000. The number of data points (n) is 500. The number of normal draws (R) is equal to 1,000. The starting point is zero. DF stands for Dickey-Fuller. Z stands for Phillips' Z test.

Table 2. (Larger number of draws and observations)

	Size	**Power**				
	ρ=1	**ρ=0.99**	**0.98**	**0.97**	**0.96**	**0.95**
BC (σ = 1) n=500, R=10,000	7.8%	22%	44.8%	69.3%	87%	95.2%
BC (σ = 1) n=1,000, R=10,000	5.2%	11.5%	44.9%	78.1%	91%	99.73%

Model I. The number of simulated samples is 100. The number of simulations per sample is 1,000. The number of data points (n) is 500 or 1,000. The number of normal draws (R) is equal to 10,000. The starting point is zero.

.

Table 3.

	Size	**Power**		
	κ=0	**κ=2**	**κ=6**	**κ=8**
BC (σ = $\sqrt{0.008742}$ **)**	4.8%	16.0%	78.0%	95.2%
DF (σ = $\sqrt{0.008742}$ **)**	4.7%	85%	100%	100%
Z (σ = $\sqrt{0.008742}$ **)**	5.2%	90%	100%	100%
ρ	1	0.992	0.976	0.968

Model II. The number of simulated samples is 100. The number of simulations per sample is 1,000. The number of data points (n) is 5,000. The number of normal draws (R) is equal to 1,000. The starting point is zero. ρ is the autoregressive parameter implied by the choice of κ for daily data (dt = 1/252). DF stands for Dickey-Fuller. Z stands for Phillips' Z test.

Table 4.

	Size		**Power**		
	γ=0.1	**γ=0.2**	**γ=0.6**	**γ=0.7**	**γ=0.8**
BC (σ=1)	5.6%	7.4%	63.2%	93.2%	100%
DF (σ =1)	4.2%	4.9%	7.0%	8.8%	12.2%
Z (σ =1)	5.3%	7.0%	12.3%	17.2%	24.0%
$\widehat{\rho}$	0.9998	0.9987	0.9971	0.9966	0.9964

Model III. The number of simulated samples is 100. The number of simulations per sample is 1,000. The number of data points (n) is 5,000. The number of normal draws (R) is equal to 1,000. The starting point is zero. $\widehat{\rho}$ is the estimated autoregressive parameter. DF stands for Dickey-Fuller. Z stands for Phillips' Z test.