

Adaptive Testing for a Unit Root with Nonstationary Volatility

H. Peter Boswijk

Tinbergen Institute & Department of Quantitative Economics,
Universiteit van Amsterdam*

July 1, 2005

Abstract

Recent research has emphasized that permanent changes in the innovation variance (caused by structural shifts or an integrated volatility process) lead to size distortions in conventional unit root tests. Cavaliere and Taylor (2004) and Beare (2004) propose nonparametrically corrected versions of unit root tests that have the same asymptotic null distribution as the uncorrected versions in case of homoskedasticity. In this paper, we first derive the asymptotic power envelope for the unit root testing problem when the nonstationary volatility process is known. Next, we show that under suitable conditions, adaptation with respect to the volatility process is possible, in the sense that non-parametric estimation of the volatility process leads to the same asymptotic power envelope. Special attention is devoted to the choice of a volatility filter, and to the construction of asymptotically valid critical values or p -values.

1 Introduction

Over the past decade, a large amount of research has been devoted to the effect of heteroskedasticity on unit root tests. When the heteroskedasticity follows a stationary GARCH-type specification, such that the unconditional variance is well-defined and constant, then the invariance principle guarantees that the usual Dickey-Fuller tests remain valid asymptotically. This was illustrated using Monte Carlo simulations by Kim and Schmidt (1993). Subsequent research has indicated, however, that in such cases more powerful tests for a unit root may be obtained from a likelihood analysis of a model with GARCH innovations; see Seo (1999) and Ling *et al.* (2003) (based on Ling and Li (1998)), *inter alia*.

In empirical applications, the assumption that the variation in volatility effectively averages out over the relevant sample is often questionable. In applications involving daily financial prices (interest rates, exchange rates), the degree of mean reversion in the volatility is usually so weak that the volatility process shows persistent deviations from its mean over the relevant time span (often ten years or less). On the other hand, in applications involving macro-economic time series observed at a lower frequency

* Address for correspondence: Department of Quantitative Economics, Universiteit van Amsterdam, Roetersstraat 11, 1018 WB Amsterdam, The Netherlands. E-mail: H.P.Boswijk@uva.nl.

but over a longer time span, one often finds level shifts in the volatility, instead of volatility clustering. Intermediate cases (slowly mean-reverting volatility with changing means) may also occur.

In the presence of such persistent variation in volatility, the invariance principle cannot be expected to apply, such that the null distribution of unit root tests will be affected. The resulting size distortions have been investigated by Boswijk (2001) for the case of a near-integrated GARCH process, and more recently by Cavaliere (2004) for the case of a deterministic volatility function.¹ Cavaliere and Taylor (2004) and Beare (2004) provide two alternative solutions to these size distortions, in the form of non-parametric corrections that lead to statistics with the usual asymptotic null distributions. The approach of Cavaliere and Taylor (2004) is based on time-deformation arguments, whereas Beare (2004) proposes to apply a unit root test to a reweighted cumulative sum of increments of the process.

Although these corrected tests have the same null distribution as the Dickey-Fuller tests under homoskedasticity, they will have a different power function. Furthermore, there is no guarantee that the same correction that delivers the right null distribution will also yield the highest possible power. In particular, one may expect high power from a method that gives the highest weight to observations with the lowest volatility, and this is not the case for the tests discussed above.

The present paper addresses this issue by deriving the asymptotic power envelope, i.e., the maximum possible power against a sequence of local alternatives to the unit root, for a given and known realization of the volatility process. This allows us to evaluate the power loss of various tests, and to construct a class of admissible tests, that have a point of tangency with the envelope. For the empirically more relevant case where the volatility function is not observed, we show that under suitable conditions, adaptation with respect to the volatility process is possible, in the sense that non-parametric estimation of the volatility process leads to the same asymptotic power envelope. The test statistics that come out of this analysis have an asymptotic null distribution that depends on the realization of the volatility process. Therefore, we cannot construct tables with critical values, but the null distribution and hence p -value may be obtained by simulation, conditional on the volatility process.

The plan of the paper is as follows. In Section 2, we present the model, and obtain some preliminary asymptotic results. Section 3 establishes that the model with known volatility has locally asymptotically quadratic (LAQ) likelihood ratios, which enables the power envelope (conditional on the volatility process) to be characterized and simulated. Section 4 discusses nonparametric estimation of the volatility process, and its use in the construction of a class of adaptive tests. The finite-sample behaviour of these tests is investigated in a Monte Carlo experiment in Section 5, and Section 6 contains some concluding remarks. Proofs are given in an appendix.

Throughout the paper, we use the notation $X_n \xrightarrow{\mathcal{L}} X$ to denote convergence in distribution for sequences of random variables or vectors, and $X_n(s) \xrightarrow{\mathcal{L}} X(s)$, $s \in [0, 1]$ to denote weak convergence in $D[0, 1]^k$, the product space of right-continuous functions with finite left limits, under the uniform metric. The notation $\lfloor x \rfloor$ is used for the largest integer $\leq x$, and “ \perp ” denotes stochastic independence.

¹A related analysis of non-stationary volatility in (auto-) regressions with stationary regressors is provided by Hansen (1995) for near-integrated volatility, and by Phillips and Xu (2005) for deterministic volatility.

2 The model and preliminary results

Consider the first-order heteroskedastic autoregression

$$\Delta X_t = \theta X_{t-1} + \varepsilon_t, \quad t = 1, \dots, n, \quad (1)$$

$$\varepsilon_t = \sigma_t \eta_t, \quad (2)$$

$$\eta_t \sim \text{i.i.d. } (0, 1), \quad (3)$$

$$\eta_t \perp\!\!\!\perp \mathcal{F}_{t-1} = \sigma(\eta_{t-j}, \sigma_{t+1-j}, j \geq 1), \quad (4)$$

where the starting value X_0 is considered fixed. The hypothesis of interest is the unit root hypothesis $\mathcal{H}_0 : \theta = 0$. The analysis to follow can be extended to higher-order autoregressions and the inclusion of deterministic components (intercept and linear trend), but we focus on (1) for clarity.

The assumption (4) that η_t is independent of \mathcal{F}_{t-1} and hence σ_t implies that ε_t is a martingale difference sequence relative to \mathcal{F}_t , with conditional variance $E(\varepsilon_t^2 | \mathcal{F}_{t-1}) = \sigma_t^2$, such that σ_t is the volatility (conditional standard deviation) of ε_t . This allows for a deterministic volatility process or a GARCH-type specification, in which case \mathcal{F}_{t-1} reduces to the $\sigma(\eta_{t-j}, j \geq 1)$. However, we also allow for stochastic volatility specifications, where the volatility is driven by its own shocks, provided that σ_t is stochastically independent of the contemporaneous η_t (it may depend on lags of η_t).

If the volatility process satisfies some suitable stationarity condition, then the variation in σ_t^2 will average out (i.e., $\text{plim}_{n \rightarrow \infty} n^{-1} \sum_{t=1}^n \sigma_t^2 = \bar{\sigma}^2$, with $\bar{\sigma}^2$ nonstochastic), and ε_t will satisfy an invariance principle (under appropriate technical conditions). This implies that conventional Dickey-Fuller tests for a unit root will be asymptotically valid, even though more powerful tests may be obtained by explicitly modelling the volatility process, see, e.g., Seo (1999) and Ling *et al.* (2003).

In this paper, we are concerned with cases where the volatility variation does not average out, either because of (deterministic) permanent shifts in the level of the volatility, or because the volatility dynamics is (near-) integrated, or a combination of both. We do not assume a particular parametric specification, but instead require the following:

Assumption 1 *The process $\{(\eta_t, \sigma_t)\}_{t=1}^n$ satisfies, as $n \rightarrow \infty$,*

$$\begin{pmatrix} W_n(s) \\ \sigma_n(s) \end{pmatrix} := \begin{pmatrix} n^{-1/2} \sum_{t=1}^{\lfloor sn \rfloor} \eta_t \\ \sigma_{\lfloor sn \rfloor + 1} \end{pmatrix} \xrightarrow{\mathcal{L}} \begin{pmatrix} W(s) \\ \sigma(s) \end{pmatrix}, \quad s \in [0, 1], \quad (5)$$

where $W(\cdot)$ is a standard Brownian motion, and $\sigma(\cdot)$ is a strictly positive process with continuous sample paths and $E \left[\int_0^1 \sigma(s)^2 ds \right] < \infty$. Furthermore, $X_0 = O_P(1)$.

Remark 1

(a) The invariance principle for η_t follows from the i.i.d. $(0, 1)$ assumption, but is included in Assumption 2 because joint convergence to $(W(\cdot), \sigma(\cdot))$ will be needed.

(b) The assumption that $\sigma_n(s)$ converges to $\sigma(s)$ requires that σ_t , and hence ε_t and X_t , are in fact triangular arrays $\{(X_{nt}, \varepsilon_{nt}, \sigma_{nt}), t = 1, \dots, n; n = 1, 2, \dots\}$. However, we suppress the double index notation for simplicity.

(c) One instance where the assumption arises naturally is in the context of continuous-record asymptotics, where $\{X_t\}_{t=1}^n$ is a (rescaled) discrete-time sample from the continuous-time Itô process $X(s) = \int_0^s \sigma(u) dW(u)$, observed at times $s_t = t/n$. Letting $X_t = n^{1/2} X(s_t)$, an Euler approximation leads to $X_t = X_{t-1} + \sigma_t \eta_t$, with $\sigma_t = \sigma(s_{t-1})$ and $\eta_t = n^{1/2} [W(s_t) - W(s_{t-1})] \sim \text{i.i.d. } N(0, 1)$. However, we do not confine ourselves to this case; the main motivation for Assumption 1 is to preserve persistent changes in the volatility as $n \rightarrow \infty$.

(d) Cavaliere (2004), Cavaliere and Taylor (2004) and Beare (2004) consider a similar assumption, but with σ_t a deterministic function of t . The former two authors do not require $\sigma(\cdot)$ to be continuous, but allow for finitely many discontinuities. In contrast, Beare (2004) requires $\sigma(\cdot)$ to be continuous and twice continuously differentiable, as a necessary assumption for uniform consistency of a kernel estimator of $\sigma(\cdot)$. Assumption 1 represents an intermediate case as far as smoothness of $\sigma(\cdot)$ is concerned, allowing for stochastic volatility specifications with non-differentiable sample paths, but excluding discontinuities to avoid problems with volatility estimation. Note that in practice, volatility shifts in discrete time may be approximated arbitrarily well by an underlying smooth transition function.

(e) Assumption 1 is similar in spirit to Hansen (1995)'s analysis, who assumes that σ_t^2 is a smooth positive transformation of a near-integrated autoregression, converging to an Ornstein-Uhlenbeck process. Hansen considers the effect of such volatility specifications on ordinary least-squares, generalized least-squares and adaptive estimation, when the regressor is a linear process with nonstationary volatility. The analysis in this paper may be interpreted as an attempt to generalize these results to the case of a (near-) integrated regressor.

(f) Another instance where Assumption 1 applies is when σ_t^2 follows a GARCH(1,1) specification $\sigma_t^2 = \omega + \alpha \varepsilon_{t-1}^2 + \beta \sigma_{t-1}^2$, where the true parameter values are sequences satisfying $\omega_n = O(n^{-1})$, $\alpha_n = O(n^{-1/2})$ and $1 - \alpha_n - \beta_n = O(n^{-1})$. As shown by Nelson (1990), this implies (5) with $\sigma(s)$ following a particular diffusion process, independent of $W(\cdot)$. The implications of this for the Dickey-Fuller test and a GARCH-based likelihood ratio test have been analysed by Boswijk (2001).

The following lemma characterizes the limiting behaviour of the process $\{X_t\}$ under a near-integrated parameter sequence $\mathcal{H}_n : \theta_n = c/n$, with $c \in \mathbb{R}$ a fixed constant.

Lemma 1 *In the model (1)–(4), under Assumption 1 and $\theta_n = c/n$,*

$$n^{-1/2} X_{\lfloor sn \rfloor} \xrightarrow{\mathcal{L}} X_c(s) = \int_0^s e^{c(s-u)} \sigma(u) dW(u), \quad s \in [0, 1], \quad (6)$$

jointly with (5), where $X_c(\cdot)$ satisfies $dX_c(s) = cX_c(s)ds + \sigma(s)dW(s)$.

All proofs are given in the appendix. The lemma has direct consequences for the asymptotic properties of the conventional Dickey-Fuller coefficient and t -tests. Let $\hat{\theta}_n$ denote the least-squares estimator of θ in (1), and let $\hat{\tau}_n$ denote the t -statistic for $\theta = 0$. As shown by Cavaliere (2004) (under slightly different conditions), Lemma 1 implies, under the null hypothesis $c = 0$,

$$n\hat{\theta}_n \xrightarrow{\mathcal{L}} \left(\int_0^1 X_0(s)^2 ds \right)^{-1} \int_0^1 X_0(s) \sigma(s) dW(s), \quad (7)$$

$$\widehat{\tau}_n \xrightarrow{\mathcal{L}} \left(\int_0^1 \sigma(s)^2 ds \int_0^1 X_0(s)^2 ds \right)^{-1/2} \int_0^1 X_0(s) \sigma(s) dW(s). \quad (8)$$

The distributions of the right-hand side expressions in (7) and (8) do not coincide with the usual Dickey-Fuller null distributions, unless $\sigma(s) = \sigma$ (constant), such that $X_0(\cdot) = \sigma W(\cdot)$. Thus the Dickey-Fuller tests are not robust to persistent variation in σ_t , leading to a non-constant $\sigma(\cdot)$.

Recently, two different adjustments to the Dickey-Fuller tests have been proposed to solve this lack robustness. Cavaliere and Taylor (2004) use the fact that an Itô process such as $X_0(\cdot)$, with deterministic volatility $\sigma(\cdot)$, can be expressed as a time-deformed Brownian motion. This can be used to define a sampling scheme, where X_t is observed at a lower frequency when the volatility is low, and at a higher frequency when $\sigma(s)$ is high. Applying the Dickey-Fuller (or Phillips-Perron) test to these skip-sampled observations leads to a statistic with the usual asymptotic null distribution (albeit with a different power function than under homoskedasticity). An alternative approach has been developed by Beare (2004), who applies the Dickey-Fuller / Phillips-Perron test to the cumulative sum of reweighted increments of X_t , i.e., to $X_t^* = \sum_{i=1}^t \Delta X_i / \widehat{\sigma}_i$, where $\widehat{\sigma}_t$ is obtained by kernel estimation. This again leads to a test with the same asymptotic null distribution as the Dickey-Fuller test under homoskedasticity.

The purpose of this paper is not to obtain a statistic with the Dickey-Fuller null distribution, but to derive the maximum possible asymptotic power of any test of the unit root null against local alternatives. In the next section, this is done for the (infeasible) case where σ_t (as well as the density of η_t) is known. Next, we show that the asymptotic volatility function $\sigma(\cdot)$ is consistently estimable, and this can be used to construct of a family of tests that reach the asymptotic power envelope. The resulting tests are adaptive, in the sense that there is no loss of asymptotic efficiency or power caused by estimating σ_t .

3 The power envelope

In this section, we derive the asymptotic power envelope for the unit root hypothesis in the model (1)–(4), with $\{\sigma_t\}$ known. The envelope is based on the power of the Neyman-Pearson test in an experiment that provides an asymptotic approximation of the model in a neighbourhood of the null hypothesis. A central role is played by the fact that the log-likelihood ratio of the model is *locally asymptotically quadratic* (LAQ), see, e.g., Jeganathan (1995) and Le Cam and Yang (1990). For this result, we will need to make some assumptions about the density of η_t .

Assumption 2 *The distribution of η_1 has absolutely continuous Lebesgue density $p(\eta) > 0$, which is twice continuously differentiable. The score function $\psi(\eta) = -\partial \log p(\eta) / d\eta$ satisfies*

$$\text{var} [\psi(\eta_1)] =: \mathcal{I} < \infty, \quad (9)$$

and its derivative $\psi'(\eta) = \partial \psi(\eta) / \partial \eta = -\partial^2 \log p(\eta) / \partial \eta^2$, satisfies

$$\sup_{\eta \in \mathbb{R}} |\psi'(\eta)| < M < \infty, \quad (10)$$

$$E[\psi'(\eta_1) \eta_1] = 0. \quad (11)$$

Remark 2

(a) The definition of the score function as the derivative of a log-density with respect to its argument stems from a location model $X_t = \theta + \eta_t$, with log-likelihood contributions $\ell_t(\theta) = \log p(X_t - \theta)$, and hence score contributions $\partial \ell_t / \partial \theta = \partial \log p(X_t - \theta) / \partial \theta = \psi(X_t - \theta) = \psi(\eta_t)$. This clarifies that \mathcal{I} in (9) is in fact the Fisher information in this location model, and similarly (10) imposes a bound on the Hessian contributions $\partial^2 \ell_t(\theta) / \partial^2 \theta = -\partial \psi(\eta_t) / \partial \eta_t$. It can be shown that (9) implies the usual regularity conditions

$$E[\psi(\eta_1)] = 0, \quad E[\psi'(\eta_1)] = E[\psi(\eta_1)^2] = \mathcal{I}, \quad (12)$$

as well as the following result:

$$E[\psi(\eta_1)\eta_1] = \text{cov}[\psi(\eta_1), \eta_1] = 1. \quad (13)$$

(b) When the distribution of η_1 is standard Gaussian, then $\psi(\eta) = \eta$, and consequently $\mathcal{I} = \text{var}(\eta_1) = 1$. In that case the scores $\psi(\eta_t)$ and the innovations η_t are perfectly correlated. For any other distribution, (9) and (13) imply $\text{corr}[\eta_1, \psi(\eta_1)] = \mathcal{I}^{-1/2}$, which shows that $\mathcal{I} \geq 1$, with equality only holding in the Gaussian case. For example, the standardized Student's $t(\nu)$ distribution with $\nu > 2$ has $\mathcal{I} = \nu(\nu + 1) / [(\nu - 2)(\nu + 3)]$, which increases rapidly as ν approaches the infinite-variance bound $\nu = 2$.

(c) The assumption (11) is implied by symmetry of the density $p(\eta)$. Its importance will be clarified in the next section.

(d) In the statistics literature on LAQ and local asymptotic normality (LAN), the assumption on second derivatives of the log-likelihood is usually replaced by the weaker requirement of *differentiability in quadratic mean* of the square root of the density, see, e.g., van der Vaart (1998), Section 7.2. However, we make the somewhat stronger Assumption 2, because it is easier to interpret and corresponds to the classical Cramér conditions for asymptotic normality of the maximum likelihood estimator.

The following lemma contains some necessary ingredients for the first main result of the paper.

Lemma 2 Consider the model (1)–(4), under Assumptions 1 and 2, and define $Z_{t-1} = \sigma_t^{-1} X_{t-1}$. Under $\theta_n = c/n$, and as $n \rightarrow \infty$,

$$\begin{pmatrix} n^{-1/2} \sum_{t=1}^{\lfloor sn \rfloor} \eta_t \\ (n\mathcal{I})^{-1/2} \sum_{t=1}^{\lfloor sn \rfloor} \psi(\eta_t) \\ n^{-1/2} Z_{\lfloor sn \rfloor} \end{pmatrix} \xrightarrow{\mathcal{L}} \begin{pmatrix} W(s) \\ B(s) \\ Z_c(s) \end{pmatrix}, \quad s \in [0, 1], \quad (14)$$

where $(W(\cdot), B(\cdot))$ is a bivariate Brownian motion process with unit variances and correlation $\mathcal{I}^{-1/2}$, and where

$$Z_c(s) = \sigma(s)^{-1} X_c(s) = \int_0^s e^{c(s-u)} \frac{\sigma(u)}{\sigma(s)} dW(u). \quad (15)$$

Furthermore,

$$n^{-1} \sum_{t=1}^n Z_{t-1} \psi(\eta_t) \xrightarrow{\mathcal{L}} \mathcal{I}^{1/2} \int_0^1 Z_c(s) dB(s), \quad (16)$$

jointly with (14).

Because $\{X_t, \sigma_t, t = 0, \dots, n\}$ are observed, and σ_t is possibly stochastic, it would seem natural to define the likelihood by the joint density of $\{X_t, \sigma_t, t = 1, \dots, n\}$, conditional on starting values. Letting $\mathbf{X}_{t-1} = (X_0, \dots, X_{t-1})$ and $\boldsymbol{\sigma}_{t-1} = (\sigma_0, \dots, \sigma_{t-1})$, this joint density may be factorized as

$$\begin{aligned} f((X_1, \sigma_1), \dots, (X_n, \sigma_n) | (X_0, \sigma_0)) &= \prod_{t=1}^n f(X_t, \sigma_t | \mathbf{X}_{t-1}, \boldsymbol{\sigma}_{t-1}) \\ &= \prod_{t=1}^n f(X_t | \sigma_t, \mathbf{X}_{t-1}, \boldsymbol{\sigma}_{t-1}) \prod_{t=1}^n f(\sigma_t | \mathbf{X}_{t-1}, \boldsymbol{\sigma}_{t-1}). \end{aligned} \quad (17)$$

As long as we do not specify an explicit model for σ_t given the past, the second factor is unknown. We will define the log-likelihood function as the logarithm of the first factor, which leads to

$$\ell_{(n)}(\theta) = \sum_{t=1}^n \ell_t(\theta) = \sum_{t=1}^n \left\{ -\log \sigma_t + \log p \left(\frac{\Delta X_t - \theta X_{t-1}}{\sigma_t} \right) \right\}. \quad (18)$$

Ignoring the second factor, related to $f(\sigma_t | \mathbf{X}_{t-1}, \boldsymbol{\sigma}_{t-1})$, is essentially a weak exogeneity condition in the sense of Engle *et al.* (1983); i.e., we assume that any parameter vector that might characterize this density is variation independent of θ , such that it may be ignored for likelihood inference on θ .

Define the log-likelihood ratio of $\theta_n = c/n$ relative to $\theta = 0$:

$$\Lambda_n(c) = \log \frac{dP_{\theta_n, n}}{dP_{0, n}} = \ell_{(n)}(\theta_n) - \ell_{(n)}(0), \quad (19)$$

where $P_{\theta, n}$ is the distribution of the observables implied by the model. Let

$$S_n = \frac{\partial \Lambda_n}{\partial c}(0) = \frac{1}{n} \frac{\partial \ell_{(n)}}{\partial \theta}(0) = \frac{1}{n} \sum_{t=1}^n Z_{t-1} \psi \left(\frac{\Delta X_t}{\sigma_t} \right), \quad (20)$$

$$J_n = -\frac{\partial^2 \Lambda_n}{\partial c^2}(0) = -\frac{1}{n^2} \frac{\partial^2 \ell_{(n)}}{\partial \theta^2}(0) = \frac{1}{n^2} \sum_{t=1}^n Z_{t-1}^2 \psi' \left(\frac{\Delta X_t}{\sigma_t} \right). \quad (21)$$

Theorem 1 establishes that in a local neighbourhood of the unit-root value $\theta = 0$, the log-likelihood ratio is locally asymptotically quadratic. The formulation of the result, and its proof, is based on Jegannathan (1995).

Theorem 1 Consider the model (1)–(4), under Assumptions 1 and 2. Under $P_{0, n}$, we have as $n \rightarrow \infty$,

$$\Lambda_n(c) = cS_n - \frac{1}{2}c^2J_n + o_P(1), \quad (22)$$

with

$$\begin{pmatrix} S_n \\ J_n \end{pmatrix} \xrightarrow{\mathcal{L}} \begin{pmatrix} S \\ J \end{pmatrix} = \begin{pmatrix} \mathcal{I}^{1/2} \int_0^1 Z_0(s) dB(s) \\ \mathcal{I} \int_0^1 Z_0(s)^2 ds \end{pmatrix}. \quad (23)$$

Thus

$$\Lambda_n(c) \xrightarrow{\mathcal{L}} \Lambda(c) = cS - \frac{1}{2}c^2J. \quad (24)$$

An interpretation of this result is that the experiment $\mathcal{E}_n = (\mathbb{R}^n, \mathcal{A}, \{P_{\theta,n}\}_{\theta \in \mathbb{R}})$ is locally approximated, for $\theta_n = c/n$, by the limit experiment $\mathcal{G} = (\mathbb{R}^2, \mathcal{B}, \{Q_c\}_{c \in \mathbb{R}})$, where \mathcal{A} and \mathcal{B} are the relevant Borel σ -fields, and where Q_c is the limit distribution of (S_n, J_n) under $P_{\theta_n,n}$, characterized by

$$\mathcal{L} \left(\begin{array}{c} S_n \\ J_n \end{array} \middle| P_{\theta_n,n} \right) \longrightarrow \mathcal{L} \left(\begin{array}{c} \mathcal{I}^{1/2} \int_0^1 Z_c(s) dB(s) + c\mathcal{I} \int_0^1 Z_c(s)^2 ds \\ \mathcal{I} \int_0^1 Z_c(s)^2 ds \end{array} \right) = Q_c, \quad (25)$$

with log-likelihood ratio $\Lambda(c) = \log \frac{dQ_c}{dQ_0}$. The limit experiment \mathcal{G} is a curved exponential model with one parameter c and two sufficient statistics (S, J) . Note that the information J is not ancillary, since its distribution under Q_c depends on c . This implies that the log-likelihood ratio is not locally asymptotically mixed normal (LAMN), but locally asymptotically Brownian functional (LABF); see Jeganathan (1995).

If the volatility process $\sigma(\cdot)$ is stochastic, then as long as its distribution under Q_c is not specified, we do not have a complete characterization of the distribution of (S, J) , and hence of $\Lambda(c)$. In what follows we will focus on the conditional distribution given $\sigma(\cdot)$, hence we require the following:

Assumption 3 *The distribution of $\sigma(\cdot)$ under Q_c does not vary with c , and the bivariate Brownian motion $(W(\cdot), B(\cdot))$ is independent of $\sigma(\cdot)$.*

The assumption implies that $\sigma(\cdot)$ is locally asymptotically ancillary, hence conditioning on it does not entail a loss of information on the parameter of interest. The independence assumption then guarantees that $(W(\cdot), B(\cdot))$ is still a bivariate Brownian motion, conditional on $\sigma(\cdot)$, and therefore allows us to completely characterize the conditional distribution of (S, J) and hence $\Lambda(c)$. It should be emphasized that the independence of $W(\cdot)$ and $\sigma(\cdot)$ excludes the empirically relevant possibility that future volatilities are affected by the sign of the current shock η_t , a phenomenon referred to as leverage in the GARCH literature.

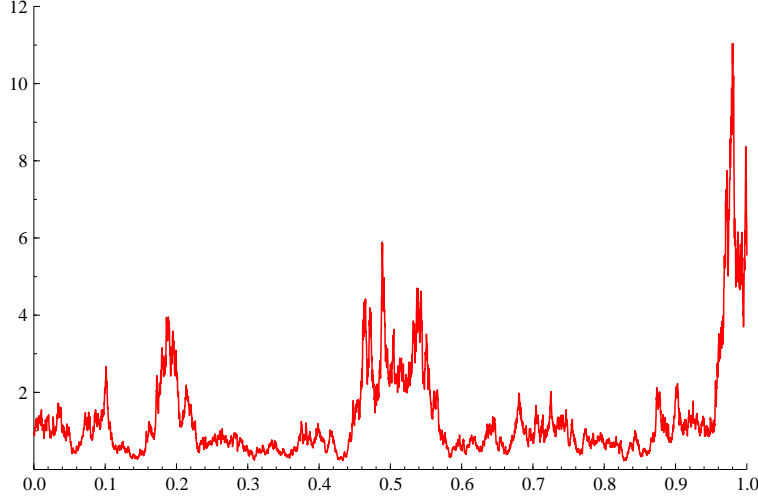
The power of the point-optimal Neyman-Pearson test for $c = 0$ against $c = \bar{c}$, which rejects for large values of $\Lambda(\bar{c})$, defines the asymptotic power envelope (conditional on $\sigma(\cdot)$) for testing $\mathcal{H}_0 : \theta = 0$ against $\mathcal{H}_n : \theta_n = \bar{c}/n$. We evaluate this power envelope by Monte Carlo simulation, for $c \in \{-20, \dots, 0\}$, with Gaussian innovations $\{\eta_t\}$, and for two volatility functions, inspired by the simulations in Cavaliere and Taylor (2004):

1. $\sigma_1(s) = \mathbf{1}_{[0,0.8)}(s) + 5 \cdot \mathbf{1}_{[0.8,1]}(s)$; this represents a level shift in the volatility from 1 to 5 at time $t = \frac{4}{5}n$.
2. $\sigma_2(s) = \exp(\frac{1}{2}V(s))$, where $dV(s) = -10V(s)ds + 10d\tilde{W}(s)$, with $\tilde{W}(\cdot)$ a standard Brownian motion, independent of $W(\cdot)$; this represents a realization of a stochastic volatility process, with a low degree of mean-reversion and a fairly high volatility-of-volatility.

The realization of the stochastic volatility process $\sigma_2(\cdot)$ that we use in our simulations is depicted in Figure 1. It may be noted that these two examples are deliberately chosen to generate a larger amount of

variation in the volatility than what may be considered empirically relevant; the purpose of this is that power differences between various procedures is most evident.

Figure 1: Realization of stochastic volatility process $\sigma_2(s)$.



The power envelopes are based on Monte Carlo simulation of $\Lambda(\bar{c})$ under Q_c , with $c \in \{0, \bar{c}\}$, where the same realization of $\sigma_2(\cdot)$ is used for all replications. The simulations of $\Lambda(\bar{c})$ under Q_0 provides 5% critical values for the test, and the rejection frequencies under $Q_{\bar{c}}$ then indicate the maximum possible power against $c = \bar{c}$.

Figures 2 and 3 depict the power envelopes for the two volatility functions, as well as the asymptotic power curves of a number of alternative unit root tests:

- *MLE*: the test that rejects for small values of the normalized Gaussian maximum likelihood estimator $n\tilde{\theta}_n = J_n^{-1}S_n$;
- *Dickey-Fuller*: the Dickey-Fuller “coefficient” test, rejecting for small values of the normalized least-squares estimator $n\hat{\theta}_n$;
- *Cavaliere-Taylor*: the Dickey-Fuller coefficient test applied to $\{\tilde{X}_t\} = \{X_{[g(t/n)]}\}$, where $g(s)$ the inverse function of the variance profile $\left(\int_0^1 \sigma_u^2 du\right)^{-1} \int_0^s \sigma_u^2 du$;
- *Beare*: the Dickey-Fuller coefficient test applied to $\{X_t^* = \sum_{i=1}^t \Delta X_i / \sigma_i\}$.

Note that the Gaussian MLE $\tilde{\theta}_n$ is just the weighted least-squares estimator of θ , using σ_t^{-2} as weights. For the Dickey-Fuller test, we note that asymptotic critical values are obtained by simulation, such that the size-corrected asymptotic power is depicted. The power function for the Cavaliere-Taylor test is based on rejection frequency of

$$\frac{\frac{1}{2} (X_c(1))^2 - 1}{\int_0^1 X_c(g(s))^2 ds},$$

whereas the power of the Beare test is the rejection frequency of

$$\frac{\frac{1}{2} (X_c^*(1))^2 - 1}{\int_0^1 X_c^*(s)^2 ds}, \quad X_c^*(s) = \int_0^s \sigma(u)^{-1} dX_c(u).$$

Figure 2: Asymptotic power curves for $\sigma_1(s)$ (volatility level shift).

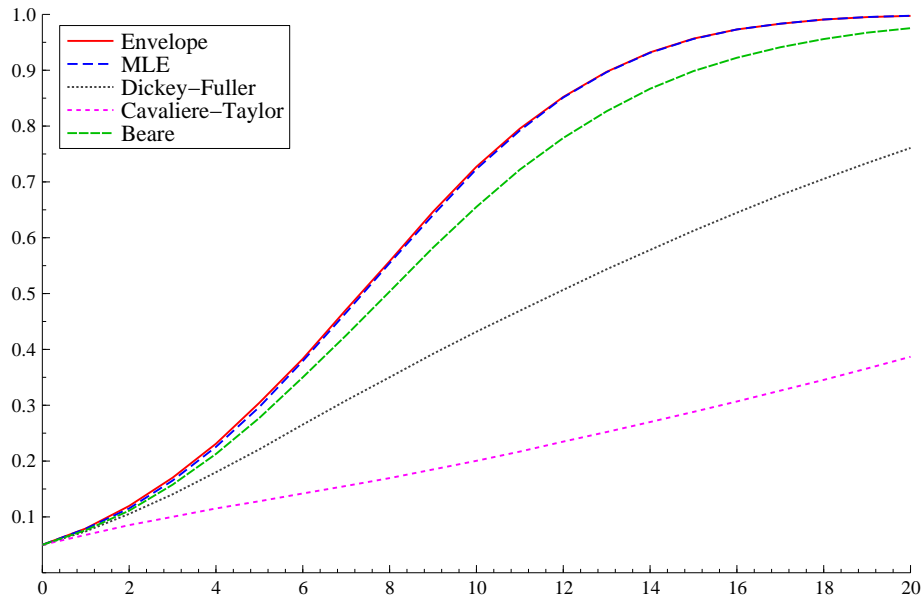
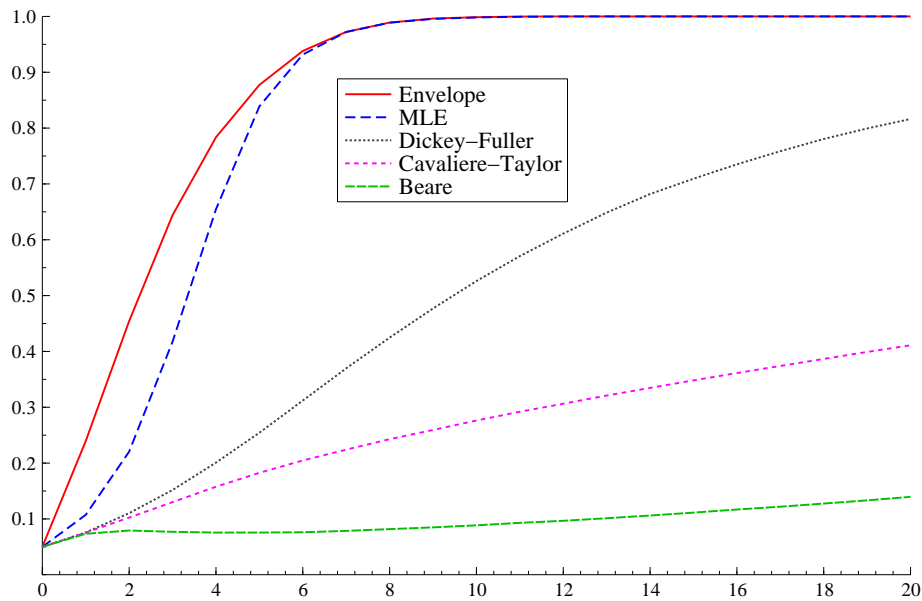


Figure 3: Asymptotic power curves for $\sigma_2(s)$ (stochastic volatility).



The main conclusions from both figures are:

- The power difference between the various procedures is substantial.
- The power of the MLE test is close to the envelope, but not equal to it, especially in case of stochastic volatility.
- In both cases, the power of the Dickey-Fuller test is substantially less than that of the MLE test. Hence reweighting observations indeed has an important effect on the power of unit root tests.

- The skip-sample procedure of Cavaliere and Taylor leads to a test that has less power than the Dickey-Fuller test. An intuitive explanation for this is that the test skips observations in times of low volatility, whereas these are exactly the most informative observations.
- The Beare test has very little power in case of stochastic volatility, but is fairly close to the power envelope for a level shift. A possible explanation is that the transformed series X_t^* has a time-varying autoregressive coefficient under the alternative, and this will bias the least-squares estimator of this coefficient towards unity. Apparently this bias is much more substantial for the stochastic volatility process.

It should be emphasized once more that we have chosen two fairly extreme volatility functions; for more realistic volatility paths, the power differences will be much smaller. However, in such cases the size distortions of the Dickey-Fuller test will also be rather small. Note also that we have considered just one realization of a stochastic volatility process; further experiments have revealed that power ordering of the different tests tend to remain the same for other realizations of the process for the same parameter values, but this need not be the case for other parameter values or specifications.

4 Volatility filtering and adaptive testing

In the previous section we have studied the power of procedures that assume that $\{\sigma_t\}$ is known and observed. In practice this is not the case, and σ_t will have to be estimated. One option is to specify a parametric model for σ_t , such as a GARCH model, and then consider maximum likelihood estimation of that model. However, it is desirable to have a testing procedure that is not too sensitive to deviations from such assumption, and that will also work well, e.g., in case of (gradual) changes in the level of the volatility.

Therefore, following Hansen (1995), we consider non-parametric estimation of $\{\sigma_t\}$. Let $k : [0, 1] \rightarrow [0, 1]$ be a kernel satisfying $\int_0^1 k(x)dx > 0$, and consider the (one-sided) kernel estimator:

$$\hat{\sigma}_n(s) = \hat{\sigma}_{\lfloor sn \rfloor + 1}, \quad (26)$$

where

$$\hat{\sigma}_t^2 = \frac{\sum_{j=0}^{N-1} k\left(\frac{j}{N}\right) \hat{\varepsilon}_{t-1-j}^2}{\sum_{j=0}^{N-1} k\left(\frac{j}{N}\right)}, \quad t > N, \quad (27)$$

and $\hat{\sigma}_t^2 = \hat{\sigma}_{N+1}^2$ for $1 \leq t \leq N$. Here N is a window width, and $\hat{\varepsilon}_t$ is either ΔX_t or $\Delta X_t - \hat{\theta}_n X_{t-1}$. To prove consistency of $\hat{\sigma}_n(s)$, we need the following assumption.

Assumption 4 For some $r > 2$, $E[|\eta_t|^{2r}] < \infty$.

The following theorem is adapted from Hansen (1995), Theorem 2:

Theorem 2 Consider the model (1)–(4), under Assumptions 1 and 4. If $N = an^b$ for some a and b satisfying $0 < a < \infty$ and $b \in (2/r, 1)$, then

$$\sup_{s \in [0,1]} |\hat{\sigma}_n(s) - \sigma(s)| \xrightarrow{P} 0. \quad (28)$$

Note that the theorem involves a trade-off between existence of moments and window width; for distributions with relatively fat tails, such that extreme observations occur with some frequency, more smoothing is needed to obtain consistency. It is important to emphasize that the theorem requires Assumption 1, and in particular, continuity of $\sigma(\cdot)$. Hence we exclude level shifts in $\sigma(\cdot)$.

A simple example of an implementation of the kernel estimator is given by exponential smoothing. Take $k(x) = e^{-5x}$, where the coefficient 5 is chosen such that $k(1) \approx 0$. Then, letting $\lambda_N = k(1/N) = e^{-5/N}$, we have $k(j/N) = \lambda_N^j$, and $\sum_{j=0}^{N-1} k\left(\frac{j}{N}\right) \approx (1 - \lambda_N)^{-1}$, such that $\hat{\sigma}_t^2 \approx (1 - \lambda_N) \sum_{j=0}^{N-1} \lambda_N^j \hat{\varepsilon}_{t-1-j}^2$. For $N = 100$, this corresponds to a smoothing parameter of $\lambda_N \approx 0.95$. As the sample size increases, λ_N would have to converge to 1 to guarantee consistency, at the rate determined by Theorem 2.

The consistency of the kernel estimator $\hat{\sigma}_n(\cdot)$ may be used for constructing tests for a unit root as follows. First, we may estimate the asymptotic score S and information J by

$$\hat{S}_n = \frac{1}{n} \sum_{t=1}^n \frac{1}{\hat{\sigma}_t} X_{t-1} \psi\left(\frac{\Delta X_t}{\hat{\sigma}_t}\right), \quad \hat{J}_n = \frac{1}{n^2} \sum_{t=1}^n \frac{1}{\hat{\sigma}_t^2} X_{t-1}^2 \mathcal{I}. \quad (29)$$

These may be used to construct approximate point-optimal test statistics $\hat{\Lambda}_n(\bar{c}) = \bar{c} \hat{S}_n - \frac{1}{2} \bar{c}^2 \hat{J}_n$, or coefficient and t -type statistics $\hat{J}_n^{-1} \hat{S}_n$ and $\hat{J}_n^{-1/2} \hat{S}_n$. At the same time, the estimator of $\sigma(\cdot)$ can be used to obtain p -values for such tests (as well as for the Dickey-Fuller test), by Monte Carlo simulation of S and J under Assumption 3, replacing $\sigma(\cdot)$ by $\hat{\sigma}_n(\cdot)$.

Consistency of p -values based on $\hat{\sigma}_n(\cdot)$ follows directly from Theorem 2. Consistency of (\hat{S}_n, \hat{J}_n) is considered in the next theorem.

Theorem 3 Consider the model (1)–(4), under Assumptions 1–4. Under $P_{\theta,n}$, we have as $n \rightarrow \infty$,

$$\begin{pmatrix} \hat{S}_n \\ \hat{J}_n \end{pmatrix} \xrightarrow{\mathcal{L}} \begin{pmatrix} S \\ J \end{pmatrix} = \begin{pmatrix} \mathcal{I}^{1/2} \int_0^1 Z_c(s) dB(s) + c \mathcal{I} \int_0^1 Z_c(s)^2 ds \\ \mathcal{I} \int_0^1 Z_c(s)^2 ds \end{pmatrix}. \quad (30)$$

Note that S and J reduce to the expressions in (23) when $c = 0$, which is also covered by the theorem. Note also that all assumptions are required; in particular the symmetry condition (11) is needed, as is clear from the proof of the theorem.

Theorem 3 implies that we may asymptotically recover the likelihood ratio $\Lambda(c)$ by nonparametric estimation of the infinite-dimensional nuisance parameter $\sigma(\cdot)$, meaning that *adaptive* estimation and testing is possible. A formal analysis of adaptivity involves finding a so-called least-favourable parametric sub-model (see van der Vaart (1998), Chapter 25) $\{P_{\theta,\phi,n}\}_{\theta \in \mathbf{R}, \phi \in \Phi}$, where $\phi \in \Phi$ is a parameter vector characterizing $\{\sigma_t\}$. Adaptivity requires block-diagonality of the information matrix in

this model, and this in turn requires the symmetry condition (11). To see this, note that the log-likelihood of the model now becomes

$$\ell_{(n)}(\theta, \phi) = \sum_{t=1}^n \left\{ -\log \sigma_t(\phi) + \log p \left(\frac{\Delta X_t - \theta X_{t-1}}{\sigma_t(\phi)} \right) \right\}, \quad (31)$$

such that

$$\begin{aligned} \frac{\partial^2 \ell_{(n)}}{\partial \theta \partial \phi}(\theta, \phi) &= - \sum_{t=1}^n \frac{X_{t-1}}{\sigma_t^2(\phi)} \frac{\partial \sigma_t(\phi)}{\partial \phi} \psi \left(\frac{\Delta X_t - \theta X_{t-1}}{\sigma_t(\phi)} \right) \\ &\quad - \sum_{t=1}^n \frac{X_{t-1}}{\sigma_t(\phi)} \psi' \left(\frac{\Delta X_t - \theta X_{t-1}}{\sigma_t(\phi)} \right) \frac{\Delta X_t - \theta X_{t-1}}{\sigma_t^2(\phi)} \frac{\partial \sigma_t(\phi)}{\partial \phi}, \end{aligned} \quad (32)$$

and this will have mean zero, when evaluated at the true value, if and only if $E[\psi'(\eta_1)\eta_1] = 0$.

If obtaining p -values by Monte Carlo simulation is considered too time-consuming, the following approximation, inspired by asymptotics for stationary volatility, may be used for the t -statistic $\widehat{J}_n^{-1/2} \widehat{S}_n$. First, approximate $Z_0(s) = \sigma(s)^{-1} \int_0^s \sigma(u) dW(u)$ by a Brownian motion, correlated with $B(s)$, with correlation coefficient

$$\rho = \frac{\langle Z_0, B \rangle(1)}{\langle Z_0 \rangle(1)^{1/2} \langle B \rangle(1)^{1/2}} = \frac{\int_0^1 \sigma(s) ds}{\left(\mathcal{I} \int_0^1 \sigma(s)^2 ds \right)^{1/2}}, \quad (33)$$

where $\langle X \rangle(s)$ is the quadratic variation process of X , and $\langle X, Y \rangle(s)$ is the covariation of X and Y . Next, use the $N(-0.48\rho, 1 - 0.2\rho^2)$ approximation, derived by Abadir and Lucas (2000), of the distribution of $\left(\int_0^1 Z(s)^2 ds \right)^{-1/2} \int_0^1 Z(s) dB(s)$, where $Z(s)$ is standard Brownian motion with $\text{corr}[Z(1), B(1)] = \rho$. It should be emphasized that the approximation error involved in this procedure is hard to characterize or bound analytically. Therefore, its adequacy can only be investigated by Monte Carlo simulation.

5 Monte Carlo results

In this section we compare the finite-sample behaviour of an adaptive t -test for a unit root with that of the Dickey-Fuller t -test in a small-scale Monte Carlo experiment. We consider two data-generating processes, corresponding to the two volatility functions $\sigma_1(\cdot)$ and $\sigma_2(\cdot)$ considered in Section 3, with standard Gaussian innovations $\{\eta_t\}$. The sample size is taken to be $n = 1000$, and we use the exponential kernel $k(x) = e^{-5x}$, with window widths $N \in \{25, 50, 100, 500\}$, corresponding to exponential smoothing parameters λ varying from approximately 0.8 to 0.99. The volatility filter is based on restricted residuals $\widehat{\varepsilon}_t = \Delta X_t$.

We first consider the size of the tests, using conventional critical values for the Dickey-Fuller test, and using p -values for the adaptive test obtained either by simulation (1000 replications; labelled p_1) or by the approximation considered at the end of the previous section (labelled p_2). Table 1 lists the empirical rejection frequencies under the null hypothesis of the tests, using a 5% nominal level, and based on 10,000 replications.

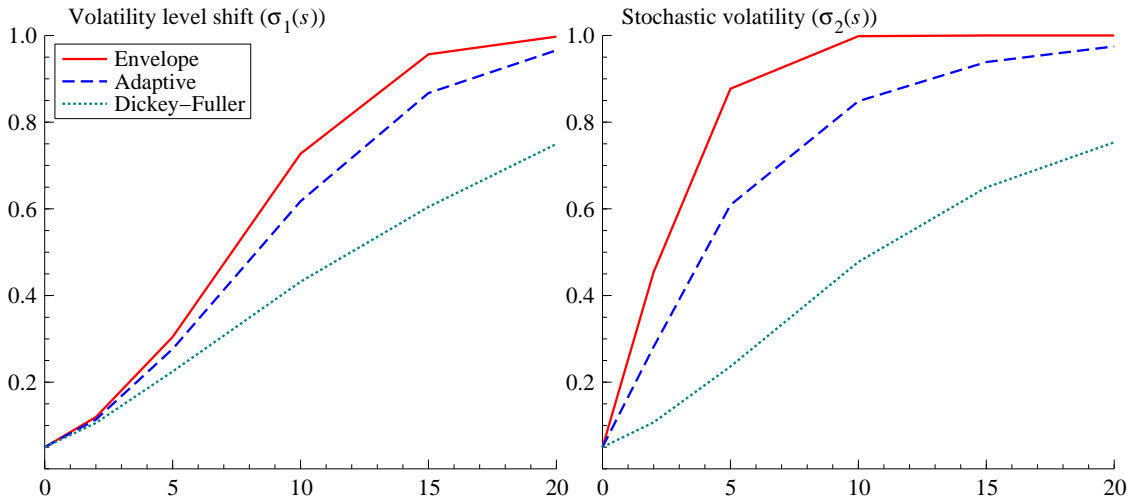
Table 1: Empirical size of adaptive t -test and Dickey-Fuller test, 5% nominal level.

N	Volatility level shift ($\sigma_1(s)$)			Stochastic volatility ($\sigma_2(s)$)		
	Adaptive, p_1	Adaptive, p_2	Dickey-Fuller	Adaptive, p_1	Adaptive, p_2	Dickey-Fuller
25	0.09	0.11	0.15	0.10	0.10	0.09
50	0.06	0.08	0.15	0.07	0.08	0.09
100	0.05	0.07	0.15	0.06	0.07	0.09
500	0.07	0.09	0.15	0.06	0.08	0.09

Note that the Dickey-Fuller size does not vary with the window width N . The size distortions for this test are in line with the results of Cavaliere (2004) and Cavaliere and Taylor (2004). For the adaptive test, we see that in both cases, under-smoothing leads to the most severe size distortions. In the present set-up, it appears that the best results are obtained for $N = 100$, corresponding to an exponential smoothing parameter around 0.95. The resulting volatility estimate is very close to the RiskMetrics volatility filter for daily financial returns, which is an exponentially weighted moving average of squared returns with $\lambda = 0.94$. The simulation-based p -values p_1 lead to empirical sizes that are closest to their nominal counterparts, although the approximate p -values also seem to perform reasonably well.

Next, we consider the size-corrected power of the two tests, in comparison with the power envelope obtained in Section 3, for $\theta_n = c/n$ with $n = 1000$ and $-c \in \{2, 5, 10, 15, 20\}$. Here we fix the window width at $N = 100$; further simulations indicate that slightly better power results are obtained by undersmoothing ($N = 25$ or 50), but as indicated above, this leads to more serious size distortions.

Figure 4: Size-corrected power of the adaptive and Dickey-Fuller tests, and power envelope.



From Figure 4, we observe that the power of the adaptive test is substantially larger than that of the Dickey-Fuller test, but that it does not reach the power envelope. The distance from the power envelope is most substantial in the stochastic volatility case. This difference is caused by estimation errors in $\sigma(s)$, which will cause the standardized errors $\varepsilon_t/\hat{\sigma}_t$ to display some mild heteroskedasticity, and probably also some unconditional excess kurtosis. Therefore, we suspect that the power might be

further improved by fitting a Student's t -based likelihood instead of a Gaussian likelihood.

6 Discussion

This paper has demonstrated that substantial power differences of unit root tests may arise in models with nonstationary volatility. Next, we have shown that it is possible to construct a class of tests that have a point of tangency with the power envelope. The tests are based on nonparametric volatility filtering, and therefore do not require very specific assumptions on the parametric form of the volatility process. However, we do need some assumptions that may be violated in practice.

First, for consistency of the nonparametric volatility filter, the volatility process needs to have continuous sample paths. This means that sudden level shifts are excluded. In practice, one might argue that these may be approximated arbitrarily well by smooth transition functions, but we may expect the kernel estimator to perform relatively poorly around the point where this (sudden or smooth) change occurs. Further Monte Carlo evidence is needed to investigate how sensitive the adaptive procedure is to such volatility estimation errors.

Secondly, the proposed method to calculate p -values by Monte Carlo simulation involves inference conditional on the realization of the volatility process. This in turn requires independence of that process and the Brownian motions defined from the standardized innovations, and hence excludes volatility processes with leverage. This is a serious limitation, which may be violated in many possible applications in finance. One could adapt the procedure to the more general case by making explicit the type of dependence between the volatility process and the Brownian motions, but this seems impossible without a parametric volatility model such as exponential GARCH.

In addition, the existence of finite fourth moments is required, but this assumption is less likely to be violated in practice.

The analysis of this paper could be extended in various directions. First, one could estimate the density $p(\cdot)$ nonparametrically as well, instead of assuming that it is known. However, it is not obvious that this additional flexibility is worth the effort. In related work on non-Gaussian unit root and cointegration analysis, it appears that for increasing power, the main condition is that both the true density of the innovations and the assumed density used for constructing the likelihood function have fat tails. Hence it may be reasonable to simply assume, e.g., a Student's $t(\nu)$ density in applications involving fat tails.

A more promising extension of this analysis is to the multivariate case. The volatility filter considered here has a very obvious extension to an estimator of a time-varying covariance matrix; as long as the same kernel and window width is used for all variances and covariances, the resulting estimator will be positive semi-definite by construction. This may be used to construct more efficient cointegration tests or adaptive estimators of cointegrating vectors in the presence of nonstationary volatility. We intend to explore this possibility in future work.

Appendix

Proof of Lemma 1. Let $\alpha_n = 1 + \theta_n$, and note that $X_t = \alpha_n X_{t-1} + \sigma_t \eta_t$, such that $X_t = \alpha_n^t X_0 + \sum_{i=0}^{t-1} \alpha_n^i \sigma_{t-i} \eta_{t-i}$ and hence

$$n^{-1/2} X_{\lfloor sn \rfloor} = f_n(s) n^{-1/2} X_0 + f_n(s) \int_0^s H_n(u) dW_n(u), \quad (\text{A.1})$$

where $f_n(s) = \alpha_n^{\lfloor sn \rfloor}$ and $H_n(s) = \alpha_n^{-\lfloor sn \rfloor - 1} \sigma_n(s)$. It follows that $f_n(s) = (1 + c/n)^{\lfloor sn \rfloor} \xrightarrow{\mathcal{L}} e^{cs}$. Furthermore, Assumption 1 implies, by the continuous mapping theorem, $(H_n(s), W_n(s)) \xrightarrow{\mathcal{L}} (e^{-cs} \sigma(s), W(s))$. The first right-hand side term of (A.1) converges to zero, because $X_0 = O_P(1)$. The required result $n^{-1/2} X_{\lfloor sn \rfloor} \xrightarrow{\mathcal{L}} \int_0^s e^{c(s-u)} \sigma(u) dW(u)$ then follows from Hansen (1992)'s Theorem 2.1, using the fact that $\{(\alpha_n^{-t-1} \sigma_{t+1}, \eta_t)\}_{t \geq 1}$ is adapted to $\{\mathcal{F}_t\}_{t \geq 1}$, and $\{\eta_t\}_{t \geq 1}$ is a martingale difference sequence with respect to $\{\mathcal{F}_t\}_{t \geq 0}$, with $\sup_n n^{-1} \sum_{t=1}^n E(\eta_t^2) < \infty$.

The stochastic differential equation for $X_c(s)$ follows from the fact that $Y_c(s) = e^{-cs} X_c(s)$ satisfies $dY_c(s) = e^{-cs} \sigma(s) dW(s)$, and applying Itô's lemma to $X_c(s) = e^{cs} Y_c(s) = f(s, Y_c(s))$, leading to

$$dX_c(s) = ce^{cs} Y_c(s) ds + e^{cs} dY_c(s) = cX_c(s) ds + \sigma(s) dW(s). \quad (\text{A.2})$$

□

Proof of Lemma 2. If $\mathcal{I} > 0$, joint weak convergence of the partial sum process of $(\eta_t, \mathcal{I}^{-1/2} \psi(\eta_t))$ to $(W(\cdot), B(\cdot))$ follows from the invariance principle for i.i.d. vectors with finite and positive definite variance matrix

$$\begin{bmatrix} 1 & \mathcal{I}^{-1/2} \\ \mathcal{I}^{-1/2} & 1 \end{bmatrix}.$$

If $\mathcal{I} = 1$, then the variance matrix becomes singular, but then $\psi(\eta_t) = \eta_t$, such that the joint convergence to the bivariate Brownian motion still applies. Weak convergence of $n^{-1/2} Z_{\lfloor sn \rfloor}$ to $Z_c(s)$ follows from Lemma 1, together with Assumption 1 and the continuous mapping theorem.

Next, (16) follows from Hansen (1992)'s Theorem 2.1, using the fact that $\{(Z_t, \psi(\eta_t))\}_{t \geq 1}$ is adapted to $\{\mathcal{F}_t\}_{t \geq 1}$, and $\{\psi(\eta_t)\}_{t \geq 1}$ is a martingale difference sequence with respect to $\{\mathcal{F}_t\}_{t \geq 0}$ with $\sup_n n^{-1} \sum_{t=1}^n E[\psi(\eta_t)^2] < \infty$. □

Proof of Theorem 1. The theorem follows from Jeganathan (1995), Theorem 13, where less stringent assumptions are made on $\psi(\cdot)$. Under Assumption 2, the result follows more directly from a second-order Taylor series expansion of $\Lambda_n(c)$, leading to

$$\Lambda_n(c) = cS_n - \frac{1}{2} c^2 J_n^*, \quad (\text{A.3})$$

where $J_n^* = n^{-2} \sum_{t=1}^n Z_{t-1}^2 \psi'(\sigma_t^{-1} [\Delta X_t - (c^*/n) X_{t-1}])$, with c^* between 0 and c . Note that under $P_{0,n}$, $\Delta X_t = \varepsilon_t$. Continuity and boundedness of $\psi'(\cdot)$ implies that

$$\frac{1}{n} \sum_{t=1}^{\lfloor sn \rfloor} \psi' \left(\frac{\varepsilon_t - (c^*/n) X_{t-1}}{\sigma_t} \right) = \frac{1}{n} \sum_{t=1}^{\lfloor sn \rfloor} \psi' \left(\eta_t - \frac{c^*}{n} Z_{t-1} \right) \xrightarrow{\mathcal{L}} s\mathcal{I}, \quad (\text{A.4})$$

and this can be used to prove

$$J_n^* = \frac{1}{n^2} \sum_{t=1}^n Z_{t-1}^2 \psi' \left(\eta_t - \frac{c^*}{n} Z_{t-1} \right) \xrightarrow{\mathcal{L}} \mathcal{I} \int_0^1 Z_0(s)^2 ds. \quad (\text{A.5})$$

Analogously, it follows that $J_n^* = J_n + o_P(1)$. The limit of S_n follows directly from Lemma 2. \square

Proof of Theorem 2. The proof is adapted from Hansen (1995), Theorem 2. Continuity of $\sigma(s)$ implies that it is sufficient to prove

$$\max_{1 \leq t \leq n} |\hat{\sigma}_t^2 - \sigma_t^2| \xrightarrow{P} 0. \quad (\text{A.6})$$

Let $w_{jN} = \left(\sum_{j=0}^{N-1} k(j/N) \right)^{-1} k(j/N)$, such that $\hat{\sigma}_t^2 = \sum_{j=0}^{N-1} w_{jN} \hat{\varepsilon}_{t-1-j}^2$ for $t > N$, with $\sum_{j=0}^{N-1} w_{jN} = 1$. For $t > N$, we have

$$\hat{\sigma}_t^2 - \sigma_t^2 = R_t^a + \sigma_t^2 R_t^b + R_t^c + R_t^d, \quad (\text{A.7})$$

where

$$\begin{aligned} R_t^a &= \sum_{j=0}^{N-1} w_{jN} (\sigma_{t-1-j}^2 - \sigma_t^2), & R_t^b &= \sum_{j=0}^{N-1} w_{jN} (\eta_{t-1-j}^2 - 1), \\ R_t^c &= \sum_{j=0}^{N-1} w_{jN} (\sigma_{t-1-j}^2 - \sigma_t^2) (\eta_{t-1-j}^2 - 1), & R_t^d &= \sum_{j=0}^{N-1} w_{jN} (\hat{\varepsilon}_{t-1-j}^2 - \varepsilon_{t-1-j}^2). \end{aligned}$$

Hansen's proof that $\max_{N < t \leq n} |R_t^a| \xrightarrow{P} 0$, $\max_{N < t \leq n} |\sigma_t^2 R_t^b| \xrightarrow{P} 0$ and $\max_{N < t \leq n} |R_t^c| \xrightarrow{P} 0$ is directly applicable here. For the fourth term, we note that $\hat{\varepsilon}_t = \varepsilon_t + (c_n/n)X_{t-1}$, where c_n is given by c if $\hat{\varepsilon}_t = \Delta X_t$ (restricted residuals), and by $c - n\hat{\theta}_n$ if $\hat{\varepsilon}_t = \Delta X_t - \hat{\theta}_n X_{t-1}$ (unrestricted residuals). In both cases, $c_n = O_P(1)$. Therefore,

$$\left| \sum_{j=0}^{N-1} w_{jN} (\hat{\varepsilon}_{t-1-j}^2 - \varepsilon_{t-1-j}^2) \right| \leq 2 \left| \frac{1}{n} \sum_{j=0}^{N-1} w_{jN} \varepsilon_{t-1-j} X_{t-2-j} \right| c_n + \left| \frac{1}{n^2} \sum_{j=0}^{N-1} w_{jN} X_{t-2-j}^2 \right| c_n^2. \quad (\text{A.8})$$

Analogous to Hansen (1995), p. 1130, it follows that

$$\max_{N < t \leq n} \left| \frac{1}{n} \sum_{j=0}^{N-1} w_{jN} \varepsilon_{t-1-j} X_{t-2-j} \right| \xrightarrow{P} 0, \quad (\text{A.9})$$

$$\max_{N < t \leq n} \left| \frac{1}{n^2} \sum_{j=0}^{N-1} w_{jN} X_{t-2-j}^2 \right| = O_P \left(\frac{N^2}{n^2} \right) \xrightarrow{P} 0, \quad (\text{A.10})$$

such that $\max_{N < t \leq n} |R_t^d| \xrightarrow{P} 0$. This proves (A.6) for $N < t \leq n$. The extension to $1 \leq t \leq N$ follows from continuity of $\sigma(s)^2$, using Hansen (1995)'s Lemma A.1. \square

Proof of Theorem 3. Consistency of \hat{J}_n follows directly from Lemma 2, Theorem 2, and the continuous mapping theorem. For \hat{S}_n , we use

$$\hat{S}_n = \frac{1}{n} \sum_{t=1}^n \frac{1}{\hat{\sigma}_t} X_{t-1} \psi \left(\frac{\Delta X_t}{\sigma_t} \right) + \frac{1}{n} \sum_{t=1}^n \frac{1}{\hat{\sigma}_t} X_{t-1} \left\{ \psi \left(\frac{\Delta X_t}{\hat{\sigma}_t} \right) - \psi \left(\frac{\Delta X_t}{\sigma_t} \right) \right\}. \quad (\text{A.11})$$

Using $\Delta X_t = \varepsilon_t + (c/n)X_{t-1}$, a first-order Taylor series expansion yields

$$\psi\left(\frac{\Delta X_t}{\widehat{\sigma}_t}\right) - \psi\left(\frac{\varepsilon_t}{\sigma_t}\right) \approx \psi'\left(\frac{\varepsilon_t}{\sigma_t}\right) \frac{c}{n} X_{t-1} - \psi'\left(\frac{\varepsilon_t}{\sigma_t}\right) \frac{\varepsilon_t}{\sigma_t^2} (\widehat{\sigma}_t - \sigma_t), \quad (\text{A.12})$$

where the approximation error is of lower order in probability than the right-hand side terms. This leads to

$$\widehat{S}_n = \frac{1}{n} \sum_{t=1}^n \frac{1}{\widehat{\sigma}_t} X_{t-1} \psi(\eta_t) + \frac{1}{n^2} \sum_{t=1}^n \frac{c}{\widehat{\sigma}_t} X_{t-1}^2 \psi'(\eta_t) - \frac{1}{n} \sum_{t=1}^n \frac{1}{\widehat{\sigma}_t} X_{t-1} \psi'(\eta_t) \eta_t \frac{\widehat{\sigma}_t - \sigma_t}{\sigma_t} + o_P(1). \quad (\text{A.13})$$

The first two terms together converge to $S = \mathcal{I}^{1/2} \int_0^1 Z_c(s) dB(s) + c\mathcal{I} \int_0^1 Z_c(s)^2 ds$, using Lemma 2, Theorem 2 and the continuous mapping theorem. The third term converges to zero, because $\psi'(\eta_t)\eta_t$ is a mean-zero innovation. Note that if condition (11) is not satisfied, then the third term does not vanish, and hence the effect of estimating σ_t is no longer negligible. \square

References

- Abadir, K. and A. Lucas (2000), “Quantiles for t -Statistics Based on M -Estimators of Unit Roots”, *Economics Letters*, 67, 131–137.
- Beare, B. K. (2004), “Robustifying Unit Root Tests to Permanent Changes in Innovation Variance”, Working paper, Yale University.
- Boswijk, H. P. (2001), “Testing for a Unit Root with Near-Integrated Volatility”, Tinbergen Institute Discussion Paper # 01-077/4, <http://www.tinbergen.nl/discussionpapers/01077.pdf>.
- Cavaliere, G. (2004), “Unit Root Tests Under Time-Varying Variances”, *Econometric Reviews*, 23, 259–292.
- Cavaliere, G. and A. M. R. Taylor (2004), “Testing for Unit Roots in Time Series Models with Nonstationary Volatility”, Working paper, University of Bologna.
- Engle, R. F., D. F. Hendry and J.-F. Richard (1983), “Exogeneity”, *Econometrica*, 51, 277–304.
- Hansen, B. E. (1992), “Convergence to Stochastic Integrals for Dependent Heterogeneous Processes”, *Econometric Theory*, 8, 489–500.
- Hansen, B. E. (1995), “Regression with Nonstationary Volatility”, *Econometrica*, 63, 1113–1132.
- Jeganathan, P. (1995), “Some Aspects of Asymptotic Theory with Applications to Time Series Models”, *Econometric Theory*, 11, 818–887.
- Kim, K., and P. Schmidt (1993), “Unit Root Tests with Conditional Heteroskedasticity”, *Journal of Econometrics*, 59, 287–300.
- Le Cam, L. and G. L. Yang (1990), *Asymptotics in Statistics. Some Basic Concepts*. New York: Springer-Verlag.
- Ling, S. and W. K. Li (1998), “Limiting Distributions of Maximum Likelihood Estimators for Unstable Autoregressive Moving-Average Time Series with General Autoregressive Heteroscedastic Errors”, *Annals of Statistics*, 26, 84–125.
- Ling, S., W. K. Li and M. McAleer (2003), “Estimation and Testing for Unit Root Processes with GARCH(1,1) Errors: Theory and Monte Carlo Evidence”, *Econometric Reviews*, 22, 179–202.
- Nelson, D. B. (1990), “ARCH Models as Diffusion Approximations”, *Journal of Econometrics*, 45, 7–38.
- Phillips, P. C. B. and K.-L. Xu (2005), “Inference in Autoregression under Heteroskedasticity”, Working paper, Yale University.
- Seo, B. (1999), “Distribution Theory for Unit Root Tests with Conditional Heteroskedasticity”, *Journal of Econometrics*, 91, 113–144.
- van der Vaart, A. W. (1998), *Asymptotic Statistics*. Cambridge: Cambridge University Press.