Adaptive Testing for a Unit Root with Nonstationary Volatility

H. Peter Boswijk
Tinbergen Institute & Amsterdam School of Economics,
Universiteit van Amsterdam

Yang Zu
Department of Economics, City University London

June 5, 2014

Abstract

Recent research has emphasised 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 (2008) show how these size distortions may be resolved using the wild bootstrap. 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 nonparametric estimation of the volatility process leads to the same asymptotic power envelope. Implementation of the resulting test involves cross-validation and the wild bootstrap. A Monte Carlo experiment shows that the asymptotic results are reflected in finite sample properties.

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

Helpful comments on an earlier version from Rob Taylor, Oliver Linton, Peter Phillips, Anders Rahbek, Ulrich Müller are gratefully acknowledged.

Address for correspondence: Amsterdam School of Economics, Universiteit van Amsterdam, Valckenierstraat 65–67, 1018 XE Amsterdam, The Netherlands. E-mail: H.P.Boswijk@uva.nl.
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 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 by Kim et al. (2002) and Cavaliere (2004) for the case of a deterministic volatility function.¹ Cavaliere and Taylor (2008) develop a wild bootstrap version of the standard Dickey-Fuller tests, and show that this leads to tests with a correct asymptotic size. Cavaliere and Taylor (2007) and Beare (2008) provide two alternative solutions, in the form of nonparametric corrections that lead to statistics with the usual asymptotic null distributions.

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.²

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


²An exception is Kim et al. (2002), who consider GLS-based testing for a unit root in case of a single break in the volatility.
Throughout the paper, we use the notation $X_n \xrightarrow{L} X$ to denote convergence in distribution for sequences of random variables or vectors, and $X_n(s) \xrightarrow{L} X(s), s \in [0,1]$ to denote weak convergence in $D[0,1]^k$, the product space of right-continuous functions with finite left limits, under the Skorohod metric. $X_n \xrightarrow{L} P X$ denotes weak convergence in probability, see Giné and Zinn (1990). The notation $\lfloor x \rfloor$ is used for the largest integer less than or equal to $x$.

2 The model and preliminary results

Consider the heteroskedastic first-order autoregressive model

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

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

$$E(\eta_t|\mathcal{F}_{t-1}) = 0, \quad E(\eta_t^2|\mathcal{F}_{t-1}) = 1, \quad \mathcal{F}_{t-1} = \sigma(\eta_{t-j}, \sigma_{t+1-j}, j \geq 1), \quad (3)$$

with $X_0 = 0$. Extensions to models with deterministic components and higher-order autoregressions will be considered in Section 5. The hypothesis of interest is the unit root hypothesis $H_0 : \theta = 0$.

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

If the volatility process satisfies some suitable stationarity condition, then the variation in $\sigma_t^2$ will average out (i.e., $\text{plim}_{n \to \infty} n^{-1} \sum_{t=1}^n \sigma_t^2 = \bar{\sigma}^2$, with $\bar{\sigma}^2$ nonstochastic), and $\varepsilon_t$ 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 $n \to \infty$,

$$\begin{pmatrix} W_n(s) \\ \sigma_n(s) \end{pmatrix} := \begin{pmatrix} n^{-1/2} \sum_{t=1}^{\lfloor sn \rfloor} \eta_t \\ \sigma_{\lfloor sn \rfloor+1} \end{pmatrix} \xrightarrow{L} \begin{pmatrix} W(s) \\ \sigma(s) \end{pmatrix}, \quad s \in [0,1], \quad (4)$$

where $W(\cdot)$ is a standard Brownian motion, and $\sigma(\cdot)$ is a process in $D[0,1]$, independent of $W(\cdot)$ and satisfying $E \left( \int_0^1 \sigma(s)^2 ds \right) < \infty$. 


Remark 1

(a) The invariance principle for \( \eta_t \) would follow if the martingale difference assumption is strengthened to an i.i.d. assumption, or augmented with a Lindeberg condition. Note that joint convergence to \((W(\cdot), \sigma(\cdot))\) is needed.

(b) The assumption that \( \sigma_n(s) \) converges to \( \sigma(s) \) requires that \( \sigma_t \), and hence \( \epsilon_t \) and \( X_t \), are in fact triangular arrays \( \{(X_{nt}, \epsilon_{nt}, \sigma_{nt}), t = 1, \ldots, n; n = 1, 2, \ldots \} \). 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 \( \eta_t = n^{1/2}[W(s_t) - W(s_{t-1})] \sim \) 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 \( n \to \infty \). Note that the case \( \sigma_t = \sigma(t/n) \) with \( \sigma(\cdot) \) a deterministic function in \( D[0,1] \) has been considered, inter alia, by Cavaliere (2004) and Cavaliere and Taylor (2007, 2008).

(d) Assumption 1 is similar in spirit to Hansen (1995)’s analysis, who assumes that \( \sigma_t^2 \) 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 a generalization of these results to the case of a (near-) integrated regressor.

(e) Another instance where Assumption 1 applies is when \( \sigma_t^2 \) follows a GARCH(1,1) specification \( \sigma_t^2 = w + a \sigma_{t-1}^2 + b \eta_{t-1}^2 \), where the true parameter values are sequences satisfying \( w_n = O(n^{-1}) \), \( a_n = O(n^{-1/2}) \) and \( 1 - a_n - b_n = O(n^{-1}) \). As shown by Nelson (1990), this implies (4) 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).

(f) The independence assumption excludes the empirically relevant possibility that future volatilities are affected by the sign of the current shock \( \eta_t \), a phenomenon referred to as (statistical) leverage effects in the GARCH literature. It will be required for inference on \( c \) conditional on \( \sigma(\cdot) \).

The following lemma characterizes the limiting behaviour of the stochastic part 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** Consider the model (1)–(3) under Assumption 1. Under \( \mathcal{H}_n : \theta_n = c/n \), and as \( n \to \infty \),

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

jointly with (4), where \( X_c(\cdot) \) satisfies

\[
dx_c(s) = cX_c(s)ds + \sigma(s)dW(s). \tag{5}
\]
All proofs are given in the appendix. The lemma has direct consequences for the asymptotic properties of the conventional Dickey-Fuller tests. In particular, let \( \hat{\tau}_n \) denote the \( t \)-statistic for \( \theta = 0 \) in the first-order autoregression \( \Delta X_t = \theta X_{t-1} + \varepsilon_t \). As shown by Cavaliere (2004) (under slightly different conditions), Lemma 1 implies, under the null hypothesis \( c = 0 \),

\[
\hat{\tau}_n \xrightarrow{L} (\int_0^1 \sigma(s)^2 ds \int_0^1 X_0(s)^2 ds)^{-1/2} \int_0^1 X_0(s) \sigma(s) dW(s). \tag{6}
\]

The distribution of the right-hand side expression in (6) does not coincide with the usual Dickey-Fuller null distribution, 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 \( \sigma_t \), leading to a non-constant \( \sigma(\cdot) \). As shown by Cavaliere and Taylor (2008), this problem may be resolved by the use of the so-called wild bootstrap. Alternatively, Cavaliere and Taylor (2007) 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). Yet another approach was developed by Beare (2008), who applies the Dickey-Fuller / Phillips-Perron test to the cumulative sum of reweighted increments of \( X_t \), i.e., to \( X_t^* = \sum_{i=1}^t \Delta X_i / \hat{\sigma}_i \), where \( \hat{\sigma}_i \) 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 \( \sigma_t \) is observed, and \( \eta_t \) is an i.d.d. \( N(0, 1) \) sequence. Next, we show that the asymptotic volatility function \( \sigma(\cdot) \) is consistently estimable, and this can be used to construct a family of tests that reach the Gaussian asymptotic power envelope. The resulting tests are adaptive, in the sense that there is no loss of asymptotic efficiency or power caused by estimating \( \sigma_t \).

### 3 The Gaussian power envelope

In this section, we derive the Gaussian asymptotic power envelope for the unit root hypothesis in the model (1)–(3), 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 \( \text{locally asymptotically quadratic} \) (LAQ), see, e.g., Jeganathan (1995) and Le Cam and Yang (1990).

Because \( \{X_t, \sigma_t, t = 1, \ldots, n\} \) are observed, and \( \sigma_t \) is possibly stochastic, it would seem natural to define the likelihood by the joint density of \( \{X_t, \sigma_t, t = 1, \ldots, n\} \). Letting \( X_{t-1} = \)
(X_1, \ldots, X_{t-1}) and \sigma_{t-1} = (\sigma_1, \ldots, \sigma_{t-1}), this joint density may be factorized as
\[ f((X_1, \sigma_1), \ldots, (X_n, \sigma_n)) = f(X_1, \sigma_1) \prod_{t=2}^{n} f(X_t|X_{t-1}, \sigma_{t-1}) \]
\[ = f(X_1|\sigma_1) \prod_{t=2}^{n} f(X_t|\sigma_t, X_{t-1}, \sigma_{t-1}) \times f(\sigma_1) \prod_{t=2}^{n} f(\sigma_t|X_{t-1}, \sigma_{t-1}) \].

As long as we do not specify an explicit model for \sigma_t given the past, the second factor is unknown. We will define the log-likelihood function as the logarithm of the first factor, which under Gaussianity and \( X_0 = 0 \) leads to
\[ \ell_n(\theta) = \sum_{t=1}^{n} -\frac{1}{2} \left\{ \log 2\pi\sigma_t^2 + \frac{(\Delta X_t - \theta X_{t-1})^2}{\sigma_t^2} \right\} . \]

Ignoring the second factor, related to \( f(\sigma_t|X_{t-1}, \sigma_{t-1}) \), 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 \( \theta \), such that it may be ignored for likelihood inference on \( \theta \).

Define the log-likelihood ratio of \( \theta_n = c/n \) relative to \( \theta = 0 \):
\[ \Lambda_n(c) = \log \frac{dP_{\theta_n,n}}{dP_{0,n}} = \ell_n(\theta_n) - \ell_n(0), \]
where \( P_{\theta,n} \) is the distribution of the observables implied by the model. Let \( Z_{t-1} = X_{t-1}/\sigma_t \), and
\[ S_n = \frac{1}{n} \sum_{t=1}^{n} Z_{t-1} \Delta X_t / \sigma_t = \frac{d\Lambda_n}{dc} (0), \]
\[ J_n = \frac{1}{n^2} \sum_{t=1}^{n} Z_{t-1}^2 = -\frac{d^2\Lambda_n}{dc^2} (0). \]

Because of the Gaussianity assumption, the log-likelihood ratio is exactly quadratic, i.e., \( \Lambda_n(c) = cS_n - \frac{1}{2}c^2J_n \). Theorem 1 gives its limiting behaviour under the null hypothesis and local alternatives, and characterises the log-likelihood as locally asymptotically quadratic, see Jeganathan (1995).

**Theorem 1** Consider the model (1)–(3), under Assumption 1. Let
\[ Z_c(s) = \sigma(s)^{-1}X_c(s) = \int_0^s e^{c(s-u)} \frac{\sigma(u)}{\sigma(s)} dW(u). \]
Under \( H_n : \theta_n = c/n \), we have as \( n \to \infty \),
\[ \begin{pmatrix} S_n \\ J_n \end{pmatrix} \xrightarrow{L^2} \begin{pmatrix} S_c \\ J_c \end{pmatrix} = \begin{pmatrix} \int_0^1 Z_c(s) dW(s) + c \int_0^1 Z_c(s)^2 ds \\ \int_0^1 Z_c(s)^2 ds \end{pmatrix}, \tag{7} \]
and hence, for fixed \( \bar{c} \in \mathbb{R} \),
\[ \Lambda_n(\bar{c}) \xrightarrow{L^2} \Lambda_c(\bar{c}) = \bar{c}S_c - \frac{1}{2}\bar{c}^2J_c. \tag{8} \]
Remark 2

(a) As usual in a likelihood analysis, the asymptotic distributions and hence power functions derived below will continue to hold when the Gaussianity assumption is violated, as long as \{η_t\} satisfies an invariance principle. However, the optimality claims in the results to follow critically depend on its validity: if the actual density η_t differs from the Gaussian density, then more powerful tests may be constructed from a likelihood function derived from the actual density. In an earlier working paper version of this paper, we considered the power envelope for an arbitrary but known density p(η); see Boswijk (2005).

(b) Note that in (7) and (8), c refers to the true data-generating process (the probability measure \( P_{\theta_n} \), with \( \theta_n = c/n \)), whereas \( \bar{c} \) characterizes a chosen local alternative. Therefore, setting \( c = 0 \) in gives the asymptotic null distribution of the Neyman-Pearson test statistic for \( H_0 : \theta = 0 \) against \( H_n : \theta_n = \bar{c}/n \), whereas setting \( c = \bar{c} \) gives the asymptotic distribution under local alternatives, and hence may be used to evaluate local power.

(c) An interpretation of Theorem 1 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 distribution of \( (S_c, J_c) \), with log-likelihood ratio \( \Lambda_c(\bar{c}) = \log dQ_c/dQ_0 \). An interpretation of this limit experiment is that we observe \( X_c(s), s \in [0, 1] \), generated by (5), to make inference on \( c \). The limit experiment is a curved exponential model with one parameter \( c \) and two sufficient statistics \( (S_c, J_c) \). Note that the information \( J_c \) 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).

(d) If the volatility process \( \sigma(\cdot) \) is stochastic, then the weak exogeneity assumption mentioned above implies that its distribution under \( Q_c \) does not vary with \( c \), so that \( \sigma(\cdot) \) is ancillary in the limit experiment. As long as we do not specify the distribution of \( \sigma(\cdot) \), we do not have a complete characterization of the distribution of \( (S_c, J_c) \), and hence of \( \Lambda_c(c) \). In what follows we will focus on the conditional distribution given \( \sigma(\cdot) \), which is justified by the same exogeneity (and hence ancillarity) assumption.

The power of the Neyman-Pearson test for \( c = 0 \) against \( c = \bar{c} \), which rejects for large values of \( \Lambda_n(\bar{c}) \), defines the asymptotic power envelope (conditional on \( \sigma(\cdot) \)) for testing \( H_0 : \theta = 0 \) against \( H_n : \theta_n = \bar{c}/n \). We evaluate this power envelope by Monte Carlo simulation, for \( -\bar{c} \in \{0, \ldots, 30\} \), and for four different volatility functions, inspired by the simulations in Cavaliere (2004):

1. \( \sigma_1(s) = 1_{[0,0.9]}(s) + 5 \cdot 1_{[0.9,1]}(s) \); this represents a level shift in the volatility from 1 to 5 at time \( t = \frac{1}{10}n \) (i.e., late in the sample).
2. \( \sigma_2(s) = 1_{[0,0.1]}(s) + 5 \cdot 1_{[0.1,1]}(s) \); an early level shift from 1 to 5.
3. \( \sigma_3(s) = \exp(\frac{1}{2}V(s)) \), 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.
Figure 1: Realization of volatility processes $\sigma_1(s)$ through $\sigma_4(s)$.

$$
\sigma_4(s) = \exp\left(\frac{1}{\alpha}V(s)\right), \quad \text{where} \quad V(s) = 5\tilde{W}(s);
$$
a realization of a stochastic volatility process with no mean-reversion and a lower volatility-of-volatility.

The volatility paths $\sigma_1(\cdot)$ through $\sigma_4(\cdot)$ that we use in our simulations are depicted in Figure 1. The two stochastic volatilities $\sigma_3$ and $\sigma_4$ have been obtained by discretizing the relevant Brownian motions and integrals over 5000 equidistant time points in the unit interval. All computations have been performed in Ox, see Doornik (2013). 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.

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 realizations of $\sigma_3(\cdot)$ and $\sigma_4(\cdot)$ are 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 $c = \bar{c}$.

Figure 2 depicts the power envelopes for the four volatility functions, as well as the asymptotic power curves of the likelihood-based $t$-test that rejects for small values of $J_n^{-1/2}S_n$ (labelled “ML”) and of the Dickey-Fuller $t$-test (“DF”).

From these figures, we observe that the power of the likelihood-based $t$-test is close to the envelope, but not equal to it, especially in case of stochastic volatility. Furthermore, in most cases (with the exception of $\sigma_2$), the power of the Dickey-Fuller test is substantially less than that of the MLE test (and hence the envelope). Therefore, reweighting observations indeed has
an important effect on the power of unit root tests.

It should be emphasized once more that we have chosen 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.

4 Volatility filtering and adaptive bootstrap testing

In the previous section we have studied the power of procedures that assume that \( \{\sigma_t\}_{t=1}^n \) is known and observed. In practice this is not the case, and \( \sigma_t \) will have to be estimated. One option is to specify a parametric model for \( \sigma_t \), 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, inspired by Hansen (1995), we consider non-parametric estimation of \( \{\sigma_t\}_{t=1}^n \). Let \( k: [-1, 1] \rightarrow [0, 1] \) be a kernel satisfying \( \int_{-1}^{1} k(x)dx > 0 \), and consider the kernel estimator:

\[
\hat{\sigma}_n(s) = \hat{\sigma}_{[sn]},
\]

where

\[
\hat{\sigma}_t^2 = \frac{\sum_{j=-N}^{N} k \left( \frac{j}{N} \right) 1_{\{1 \leq t-j \leq n\}} \hat{\varepsilon}_t^2 - \hat{\tau}_j^2}{\sum_{j=-N}^{N} k \left( \frac{j}{N} \right) 1_{\{1 \leq t-j \leq n\}}}, \quad t = 1, \ldots, n.
\]
Here $N$ is a window width, and $\tilde{\varepsilon}_t = \Delta X_t - \hat{\theta}_n X_{t-1}$, with $\hat{\theta}_n$ the least-squares estimator. The estimator is a weighted average of leads and lags of $\tilde{\varepsilon}_t$, with weights summing to 1. For $t < N$, the estimator is determined by leads more than by lags, and for $t > n - N$, the relative weight of the lags is larger. The estimator proposed by Hansen (1995) involves only lags, and hence may be interpreted as a filtered volatility, whereas the double-sided version considered here may be seen as the smoothed volatility. Preliminary Monte Carlo experiments have revealed that the use of the double-sided weighted average leads to better finite-sample behaviour of the adaptive test considered below.

To prove consistency of $\hat{\sigma}_n(s)$, we need the following assumptions.

**Assumption 2** $\sigma(\cdot)$ is strictly positive (with probability one) and has continuous sample paths.

**Assumption 3** For some $r > 2$, $E[|\eta_t|^{2r}] < \infty$.

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

**Lemma 2** Consider the model (1)–(3), under Assumptions 1–3. If $N = an^b$ for some $a$ and $b$ satisfying $0 < a < \infty$ and $b \in (2/r, 1)$, then

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

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 continuity of $\sigma(\cdot)$. Hence we exclude level shifts in $\sigma(\cdot)$, such as considered in some of the examples in the previous section. Such level shifts can be approximated arbitrarily well by a smooth transition function, such as the logistic function; but it is expected that the non-parametric estimator will perform relatively badly around the change point.

A simple example of an implementation of the kernel estimator is that of an exponentially weighted (double-sided) moving average. Take $k(x) = e^{-5|x|}$, where the coefficient 5 is chosen such that $k(1) = k(-1) \approx 0$. Then, letting $\lambda_N = k(1/N) = e^{-5/N}$, we have $k(j/N) = \lambda_N^j$, and $\sum_{j=-N}^N k\left(\frac{j}{N}\right) \approx (1 + \lambda_N)/(1 - \lambda_N)$, such that $\hat{\sigma}_t^2 \approx (1 + \lambda_N)^{-1}(1 - \lambda_N)\sum_{j=-N}^N \lambda_N^j \tilde{\varepsilon}_{t-j}^2$. For $N = 100$, this corresponds to a smoothing parameter of $\lambda_N \approx 0.95$. As the sample size increases, $\lambda_N$ would have to converge to 1 to guarantee consistency, at the rate determined by Theorem 2. In practice, the window width may be chosen by a leave-one-out cross-validation procedure, which involves minimizing

$$\sum_{t=1}^n \frac{(\tilde{\varepsilon}_t^2 - \hat{\sigma}_t^2(N))^2}{1 - w_{tt}(N)}, \quad w_{tt}(N) = \frac{k(0)}{\sum_{j=-N}^{N} k\left(\frac{j}{N}\right) 1_{\{1 \leq t-j \leq n\}}}$$

over $N$; see Wasserman (2006).

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

$$\hat{S}_n = \frac{1}{n} \sum_{t=1}^n \frac{X_{t-1} \Delta X_t}{\hat{\sigma}_t^2}, \quad \hat{J}_n = \frac{1}{n^2} \sum_{t=1}^n \frac{X_{t-1}^2}{\hat{\sigma}_t^2}.$$
These may be used to construct approximate point-optimal test statistics \( \hat{\Lambda}_n(c) = \hat{c} \hat{S}_n - \frac{1}{2} \hat{c}^2 \hat{J}_n \), or coefficient and \( t \)-type statistics \( \hat{S}_n^{-1} \hat{S}_n \) and \( \hat{S}_n^{-1/2} \hat{S}_n \). Consistency of \((\hat{S}_n, \hat{J}_n)\) is considered in the next theorem.

**Theorem 2** Consider the model (1)–(3), under Assumptions 1–3. Under \( H_n : \theta_n = c/n \), we have as \( n \to \infty \),

\[
\left( \frac{\hat{S}_n}{\hat{J}_n} \right) \xrightarrow{\mathcal{L}} \left( \frac{S_c}{J_c} \right).
\]

This theorem implies that we may asymptotically recover the likelihood ratio \( \Lambda(c) \) by non-parametric 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) \( \{ P_{\theta, \phi, n} \}_{\theta \in \mathbb{R}, \phi \in \Phi} \), where \( \phi \in \Phi \) is a parameter vector characterizing \( \{ \sigma_t \} \). Adaptivity requires block-diagonality of the information matrix in this model, which is guaranteed by the Gaussianity assumption. To see this, note that the log-likelihood of the model now becomes

\[
\ell_n(\theta, \phi) = \sum_{t=1}^{n} \frac{1}{2} \left\{ \log 2\pi \sigma_t^2(\phi) + (\Delta X_t - \theta X_{t-1})^2 / \sigma_t^2(\phi) \right\},
\]

such that

\[
\frac{\partial^2 \ell_n}{\partial \theta \partial \phi}(\theta, \phi) = - \sum_{t=1}^{n} \frac{X_{t-1} (\Delta X_t - \theta X_{t-1}) \partial \sigma_t^2(\phi)}{\sigma_t^2(\phi)} \frac{\partial \sigma_t^2(\phi)}{\partial \phi},
\]

and this will have mean zero when evaluated at the true value.

Theorem 3 implies that the limiting null distribution of adaptive tests will be affected by nuisance parameters, just like the Dickey-Fuller test, see (6). This means that critical values for such tests cannot be tabulated, and should be generated on a case-by-case basis. One possibility is to simulate the asymptotic distribution of \((S_c, J_c)\), with \( \sigma(\cdot) \) replaced by \( \hat{\sigma}_n(s) \). This would result in a bootstrap testing procedure, where bootstrap samples are generated as \( X^b_t = X_0 + \sum_{i=1}^{t} \hat{\sigma}_{t} \hat{z}_t \), with \( \{ z_t \}_{t=1}^{n} \) an i.i.d. sequence with zero mean and unit variance, independent of the data.

An alternative is given by the wild bootstrap, see Liu (1988) and Mammen (1993). The particular implementation we propose is to generate bootstrap samples from

\[
X^b_t = X_0 + \sum_{i=1}^{t} \varepsilon^b_t = X_0 + \sum_{i=1}^{t} \hat{\varepsilon}_t \hat{z}_t,
\]

where \( z_t \sim \text{i.i.d. } N(0,1) \), independent of the data. Next, the bootstrap test statistics are constructed from

\[
\hat{\varepsilon}^b_n = \frac{1}{n} \sum_{t=1}^{n} \frac{X^b_{t-1} \Delta X^b_t}{\sigma^2_t}, \quad \hat{J}^b_n = \frac{1}{n^2} \sum_{i=1}^{n} \frac{(X^b_{t-1})^2}{\sigma^2_t}.
\]

Note that we do not propose to re-estimate \( \sigma^2_t \) for each bootstrap replication, for the simple reason that \((\varepsilon^b_t)^2 = \hat{\varepsilon}^2_t + \varepsilon^2_t (z_t^2 - 1)\), so that a smoother applied to \((\varepsilon^b_t)^2\) would have an additional noise term relative to \( \hat{\sigma}^2_t \).
Let \( \hat{\tau}_n = \hat{J}_n^{-1/2} \hat{S}_n \), the adaptive \( t \)-test statistic, and let \( \hat{\tau}_n^b = (\hat{J}_n^b)^{-1/2} \hat{S}_n^b \), its bootstrap version. In the following theorem, \( W^b(\cdot) \) is a standard Brownian motion, independent of \((W(\cdot), \sigma(\cdot))\), and

\[
X_0^b(s) = \int_0^s e^{c(s-u)} \sigma(u) dW^b(u), \quad Z_0^b(s) = \sigma(s)^{-1} X_0^b(s).
\]

**Theorem 3** Consider the model (1)–(3), under Assumptions 1–3. Under \( P_{\theta_n,n} \), we have as \( n \to \infty \),

\[
\left( \frac{\hat{S}_n^b}{\hat{J}_n^b} \right) \xrightarrow{L} P \left( \left[ \int_0^1 Z_0^b(s) dW^b(s) \right] \int_0^1 Z_0^b(s)^2 ds \right)
\]

so that

\[
\hat{\tau}_n^b \xrightarrow{L} P \left( \left( \int_0^1 Z_0^b(s)^2 ds \right)^{-1/2} \int_0^1 Z_0^b(s) dW^b(s) \right).
\]

The theorem implies that the wild bootstrap is asymptotically valid, in the sense that the bootstrap \( p \)-value is asymptotically uniform on the unit interval under the null hypothesis. Because has \( \hat{\tau}_n^b \) has the same limiting null distribution under \( P_{c/n,n} \), with \( c \neq 0 \) as under \( P_{0,n} \), it follows that the wild bootstrap test has the same asymptotic power function as the test based on the true but unknown critical values.

## 5 Deterministic components and dependent errors

The first-order autoregression with a known zero mean and a zero starting value is too restrictive in many empirical applications. Therefore, this section discusses how the adaptive test derived in the previous section may be extended in these directions.

Suppose first that the observed data are

\[
Y_t = \mu d_t + X_t, \quad t = 1, \ldots, n,
\]

where \( X_t \) satisfies the same assumptions as in the previous sections, and \( d_t \) is a vector of deterministic functions of \( t \), with \( \mu \) is a conformable parameter vector. As usual in the unit root literature, we focus on the cases \( d_t = 1 \) (constant mean \( \mu \)) and \( d_t = (1, t)' \) (linear trend \( \mu d_t = \mu_1 + \mu_2 t \)). Maintaining the assumption that \( X_0 = 0 \), this implies that the point-optimal invariant test for \( \theta = 0 \) against \( \theta_n = \bar{c}/n \), with observed \( \{\sigma_t\}_{t=1}^n \), follows as a straightforward extension of the analysis of Elliott et al. (1996). In particular, let \( \hat{\mu}(\bar{c}) \) be the OLS estimator of \( \mu \) in the regression

\[
\frac{Y_1}{\sigma_1} = \frac{\mu d_1}{\sigma_1} + \eta_1,
\]

\[
\frac{\Delta Y_t - (\bar{c}/n) Y_{t-1}}{\sigma_t} = \frac{\mu \Delta d_t - (\bar{c}/n) d_{t-1}}{\sigma_t} + \eta_t, \quad t = 2, \ldots, n.
\]

Using the notational convention \( Y_0 = 0 \) and \( d_0 = 0 \), so that \( \Delta Y_1 = Y_1 \) and \( \Delta d_1 = d_1 \), we have

\[
\hat{\mu}(\bar{c}) = \left[ \sum_{t=1}^n \frac{1}{\sigma_t^2} (\Delta d_t - \bar{c} n d_{t-1}) (\Delta Y_t - \bar{c} Y_{t-1}) \right]^{-1} \sum_{t=1}^n \frac{1}{\sigma_t^2} (\Delta d_t - \bar{c} n d_{t-1}) (\Delta Y_t - \bar{c} Y_{t-1}).
\]

12
Next, let $Y_t^d(c) = Y_t - \hat{\mu}(c)'d_t$. Then the point-optimal invariant test rejects for large values of

$$\Lambda_n^d(\bar{c}) = -\frac{1}{2} \sum_{t=1}^{n} \left( \frac{(\Delta Y_t^d(\bar{c}) - (\hat{\mu}/n)Y_{t-1}^d(\bar{c}))^2}{\sigma_t^2} \right) + \frac{1}{2} \sum_{t=1}^{n} \frac{\Delta Y_t^d(0)^2}{\sigma_t^2} \sigma_t^2$$

$$= \hat{c}S_n^d(\bar{c}) - \frac{1}{2} \hat{c}^2 J_n^d(\bar{c}) - \frac{1}{2} A_n^d(\bar{c}),$$

where, defining $Z_{t-1}(\bar{c}) = Y_{t-1}^d(\bar{c})/\sigma_t$,

$$S_n^d(\bar{c}) = \frac{1}{n} \sum_{t=1}^{n} Z_{t-1}(\bar{c}) \frac{\Delta Y_t^d(\bar{c})}{\sigma_t},$$

$$J_n^d(\bar{c}) = \frac{1}{n^2} \sum_{t=1}^{n} Z_t^d(\bar{c})^2,$$

$$A_n^d(\bar{c}) = (\hat{\mu}(\bar{c}) - \hat{\mu}(0))' \sum_{t=1}^{n} \frac{\Delta d_t \Delta d_t'}{\sigma_t^2} (\hat{\mu}(\bar{c}) - \hat{\mu}(0)).$$

The asymptotic distribution of $\Lambda_n^d(\bar{c})$ is given next.

**Theorem 4** Consider the model defined by (1)–(3) and (9), under Assumptions 1–2.

1. If $d_t = 1$ (constant mean), then under $\mathcal{H}_n : \theta_n = c/n$, as $n \to \infty$,

$$\begin{pmatrix} S_n^d(\bar{c}) \\ J_n^d(\bar{c}) \\ A_n^d(\bar{c}) \end{pmatrix} \overset{\mathcal{L}}{\to} \begin{pmatrix} S_c^d(\bar{c}) \\ J_c^d(\bar{c}) \\ A_c^d(\bar{c}) \end{pmatrix} = \begin{pmatrix} S_c \\ J_c \\ 0 \end{pmatrix},$$

where $(S_c, J_c)$ are as in Theorem 1.

2. If $d_t = (1,t)'$ (linear trend), then under $\mathcal{H}_n : \theta_n = c/n$, as $n \to \infty$,

$$\begin{pmatrix} S_n^d(\bar{c}) \\ J_n^d(\bar{c}) \\ A_n^d(\bar{c}) \end{pmatrix} \overset{\mathcal{L}}{\to} \begin{pmatrix} S_c^d(\bar{c}) \\ J_c^d(\bar{c}) \\ A_c^d(\bar{c}) \end{pmatrix} = \begin{pmatrix} \int_0^1 Z_t^d(s)dW(s) + c \int_0^1 Z_t^d(s)^2 ds \\ \int_0^1 Z_t^d(s)^2 ds \\ \ldots \end{pmatrix},$$

where . . . .
6 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 four data-generating processes, corresponding to the volatility functions $\sigma_1(\cdot)$ through $\sigma_4(\cdot)$ considered in Section 3, with standard Gaussian innovations $\{\eta_t\}$. The sample sizes considered are $n \in \{250, 1000\}$, and we use the exponential kernel $k(x) = e^{-5x}$, with window widths $N$ corresponding to exponential smoothing parameters $\lambda \in \{0.8, 0.9, 0.95, 0.99\}$; we also consider selecting the smoothing parameter by least-squares. The volatility filter is based on restricted residuals $\tilde{\varepsilon}_t = \Delta X_t$.

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 by simulation (1000 replications). We have also investigated the approximation considered at the end of the previous section, but to save space we do not report these results explicitly (the size distortions resulting from this approximation are slightly larger than those obtained by simulation). 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.

<table>
<thead>
<tr>
<th>$n$</th>
<th>$\sigma_1(s)$</th>
<th>$\sigma_2(s)$</th>
<th>$\sigma_3(s)$</th>
<th>$\sigma_4(s)$</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>250 1000</td>
<td>250 1000</td>
<td>250 1000</td>
<td>250 1000</td>
</tr>
<tr>
<td>Dickey-Fuller</td>
<td>0.13 0.13</td>
<td>0.06 0.06</td>
<td>0.08 0.09</td>
<td>0.05 0.06</td>
</tr>
<tr>
<td>Adaptive, $\lambda = 0.8$</td>
<td>0.13 0.10</td>
<td>0.11 0.10</td>
<td>0.14 0.11</td>
<td>0.12 0.11</td>
</tr>
<tr>
<td>Adaptive, $\lambda = 0.9$</td>
<td>0.10 0.07</td>
<td>0.08 0.07</td>
<td>0.11 0.07</td>
<td>0.08 0.07</td>
</tr>
<tr>
<td>Adaptive, $\lambda = 0.95$</td>
<td>0.09 0.05</td>
<td>0.09 0.05</td>
<td>0.10 0.06</td>
<td>0.07 0.06</td>
</tr>
<tr>
<td>Adaptive, $\lambda = 0.99$</td>
<td>0.16 0.07</td>
<td>0.19 0.07</td>
<td>0.16 0.06</td>
<td>0.09 0.05</td>
</tr>
<tr>
<td>Adaptive, $\lambda = \hat{\lambda}$</td>
<td>0.11 0.06</td>
<td>0.09 0.05</td>
<td>0.15 0.09</td>
<td>0.10 0.07</td>
</tr>
<tr>
<td>Average $\lambda$</td>
<td>0.85 0.93</td>
<td>0.95 0.98</td>
<td>0.80 0.86</td>
<td>0.85 0.91</td>
</tr>
</tbody>
</table>

The size distortions for the Dickey-Fuller test are in line with the results of Cavaliere (2004) and ?. For the adaptive test, we see that in all cases, under-smoothing leads to the most severe size distortions. In the present set-up, it appears that the best results are obtained for a window width 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 $\lambda = 0.94$. Even for this optimal window width, the size distortions of the adaptive test are still substantial for $n = 250$; as the sample size increases, the empirical size for $\lambda = 0.95$ does appear to converge to the nominal size. Unfortunately, the data-based estimate of $\lambda$ is often considerably lower, especially for $\sigma_3(\cdot)$.

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 $n \in \{250, 1000\}$ and $-c \in \{2, 5, 10, 15, 20\}$. Here we fix the smoothing parameter at $\lambda = 0.95$; further simulations indicate that slightly
better power results are obtained by undersmoothing, but as indicated above, this leads to more serious size distortions.

From Figure 6, we observe that the power of the adaptive test is substantially larger than that of the Dickey-Fuller test only for the first volatility path ($\sigma_1(\cdot)$, late level shift): when the volatility shifts early in the sample, the power is actually worse, and in the two cases of stochastic volatility the power is about the same. This indicates that for the volatility functions $\sigma_2(\cdot)$ through $\sigma_4(\cdot)$, a sample size of 250 is not sufficient to estimate the volatility function precisely enough.

When we increase the sample size to $n = 1000$, the results clearly improve. In Figure 7, the power of the adaptive test is larger than that of the Dickey-Fuller test in all cases except $\sigma_2(\cdot)$, although it still does not quite reach the power envelope. This is caused by estimation errors in the volatility function, which will cause the standardized errors $\varepsilon_t/\hat{\sigma}_t$ to display some mild heteroskedasticity, and probably also some unconditional excess kurtosis. Therefore, it is quite possible that the power could be further improved by fitting, e.g., a GARCH-$t$-based likelihood instead of a Gaussian likelihood.

The conclusion from the Monte Carlo experiments in this section is that the nonparametric nature of the adaptive test requires rather large sample sizes to become fully effective, such that this procedure may be recommended in particular for high-frequency (financial) data-sets. It should be stressed however, that this conclusion is partly due to the rather extreme nature of the volatility functions considered here. For more modest variations in the volatility, we may
Figure 4: Size-corrected power of the adaptive and Dickey-Fuller tests, and power envelope, $n = 1000$.

expect the adaptive procedure to be effective for smaller sample sizes than those considered here, even though in such cases, the possibilities for power gains are also smaller.

7 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. The Monte Carlo experiment in this paper suggests that the procedure may perform reasonably well for level shifts in the volatility, as long as they do not occur too early in the sample (this asymmetry is related to the fact that we consider only one-sided filtering of the volatility; a two-sided smoother may yield better results in this respect).

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, in order to apply the test in practice, it needs to be extended to allow for deterministic components (constant and trend), and for higher-order dynamics. We have not considered these extensions explicitly in this paper, but we suspect that they would not lead to additional theoretical complications, although it is well known that higher-order dynamics in particular may lead to larger size distortions in finite samples.

Another possible extension is to 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.** Consider first the autoregressive process \( \phi(L)U_t = \varepsilon_t \). Because all roots of \( \phi(z) = 0 \) are outside the unit circle, \( U_t \) is a linear process \( \omega(L)\varepsilon_t \), where the power series \( \omega(z) = \sum_{j=0}^{\infty} \omega_j z^j \) has exponentially decaying weights \( \omega_j \). Using the Beveridge-Nelson decomposition, see Phillips and Solo (1992), the partial sum \( S_t = \sum_{i=1}^{t} U_i \) may be written as

\[
S_t = \omega \sum_{i=1}^{t} \varepsilon_i + V_t - V_0, \tag{A.1}
\]

where \( V_t = \sum_{j=0}^{\infty} \gamma_j \varepsilon_{t-j} \), with \( \gamma_j = -\sum_{i=j+1}^{\infty} \omega_j \), a linear process with exponentially decaying weights, and where \( \omega = \omega(1) = 1/\phi(1) \). The assumptions imply that \( n^{-1/2}V_{[sn]} = O_P(n^{-1/2}) \)
for all \( s \in [0, 1] \), such that, using \( \varepsilon_t = \sigma_t \eta_t \),

\[
U_n(s) := n^{-1/2} S_{[sn]} = \omega n^{-1/2} \sum_{i=1}^{[sn]} \sigma_i \eta_i + o_P(1) = \omega \int_0^s \sigma(u) dW_n(u) + o_P(1)
\]

\[
\xrightarrow{\mathcal{L}} \omega \int_0^s \sigma(u) dW(u) =: U(s),
\]

say, where the convergence follows from Hansen (1992)’s Theorem 2.1, using the fact that \( \{(\sigma_{t+1}, \eta_t)\}_{t \geq 1} \) is adapted to \( \{F_t\}_{t \geq 1} \), and \( \{\eta_t\}_{t \geq 1} \) is a martingale difference sequence with respect to \( \{F_t\}_{t \geq 0} \), with \( \sup_n n^{-1} \sum_{t=1}^n E(\eta_t^2) = 1 \).

Next, \( X_t = \beta' z_t + U_t \), such that \( X_t = \beta' z_t + \sum_{i=0}^{t-1} \alpha_i U_{t-i} \), and hence

\[
n^{-1/2}(X - \beta' z)_{[sn]} = f_n(s)n^{-1/2}(X_0 - \beta' z_0) + f_n(s) \int_0^s h_n(u) dU_n(u),
\]

where \( f_n(s) = \alpha_n^{-[sn]} \) and \( h_n(s) = \alpha_n^{-[sn]-1} \). It follows that \( f_n(s) = (1 + c/n)^{[sn]} \rightarrow e^{cs} \), and similarly \( h_n(s) \rightarrow e^{-cs} \). The first right-hand side term of (A.4) converges to zero, because \( X_0 = O_P(1) \) by assumption and \( z_0 = O_P(1) \) by definition. The required result \( n^{-1/2}(X - \beta' z)_{[sn]} \xrightarrow{\mathcal{L}} \int_0^s e^{(s-u)} dU(u) \) then follows from the continuous mapping theorem, noting that the right-hand side integral of (A.4) may be rewritten (using integration by parts) as a Riemann-Stieltjes integral (with respect to a process \( h_n \) of bounded variation).

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

\[
dX_c(s) = ce^{cs} Y_c(s) ds + e^{cs} dY_c(s) = cX_c(s) ds + \sigma(s) dW(s).
\]

Proof of Lemma 2. If \( \mathcal{I} > 1 \), 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 \( n^{-1/2} Z_{[sn]} \) to \( Z_c(s) \) follows from Lemma 1, together with Assumption 1 and the continuous mapping theorem.

Next, (??) follows from Hansen (1992)’s Theorem 2.1, using the fact that \( \{(Z_t, \psi(\eta_t))\}_{t \geq 1} \) is adapted to \( \{F_t\}_{t \geq 1} \), and \( \{\psi(\eta_t)\}_{t \geq 1} \) is a martingale difference sequence with respect to \( \{F_t\}_{t \geq 0} \) with \( \sup_n n^{-1} \sum_{t=1}^n E[\psi(\eta_t)^2] < \infty. \)
Proof of Theorem 1. define
\[ F_i(\theta) = \frac{\partial \varepsilon_i(\theta)}{\partial \theta_1} = \frac{1}{\sigma_t} \phi(L) \left( X_{t-1} - \beta' z_{t-1} \right), \quad (A.6) \]
\[ G_i(\theta) = \frac{\partial \varepsilon_i(\theta)}{\partial \phi} = \frac{1}{\sigma_t} \left( U_{t-1}(\theta) \right). \quad (A.7) \]

Next, define the normalized score vector and observed information matrix, evaluated at \( \theta_0 \):
\[ S_{1n} = \frac{\partial \Lambda_n}{\partial h_1}(0) = D_{1n} \frac{\partial \ell_n}{\partial \theta_1}(\theta_0) = D_{1n} \sum_{t=1}^{n} F_i(\theta_0) \psi \left( \frac{\varepsilon_i(\theta_0)}{\sigma_t} \right), \quad (A.8) \]
\[ S_{2n} = \frac{\partial \Lambda_n}{\partial f}(0) = n^{-1/2} \frac{\partial \ell_n}{\partial \phi}(\theta_0) = n^{-1/2} \sum_{t=1}^{n} G_i(\theta_0) \psi \left( \frac{\varepsilon_i(\theta_0)}{\sigma_t} \right), \quad (A.9) \]
and
\[ J_{1n} = -\frac{\partial^2 \Lambda_n}{\partial h_1 \partial h_1}(0) = D_{1n} \sum_{t=1}^{n} \left[ F_i(\theta_0) F_i(\theta_0) \psi' \left( \frac{\varepsilon_i(\theta_0)}{\sigma_t} \right) + \frac{\partial F_i}{\partial \phi}(\theta_0) \psi \left( \frac{\varepsilon_i(\theta_0)}{\sigma_t} \right) \right] D_{1n}, \quad (A.10) \]
\[ J_{2n} = -\frac{\partial^2 \Lambda_n}{\partial f \partial f}(0) = \frac{1}{n} \sum_{t=1}^{n} G_i(\theta_0) G_i(\theta_0) \psi' \left( \frac{\varepsilon_i(\theta_0)}{\sigma_t} \right), \quad (A.11) \]
\[ J_{12,n} = -\frac{\partial^2 \Lambda_n}{\partial h_1 \partial f}(0) = n^{-1/2} D_{1n} \sum_{t=1}^{n} \left[ F_i(\theta_0) G_i(\theta_0) \psi' \left( \frac{\varepsilon_i(\theta_0)}{\sigma_t} \right) + \frac{\partial F_i}{\partial \phi}(\theta_0) \psi \left( \frac{\varepsilon_i(\theta_0)}{\sigma_t} \right) \right] \quad (A.12) \]

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
\[ \Lambda_n(c) = c S_n - \frac{1}{2} c^2 J^*_n, \quad (A.13) \]
where \( J^*_n = n^{-2} \sum_{t=1}^{n} Z_{t-1}^2 \psi' \left( \sigma_t^{-1} [\Delta X_t - (c^*/n) X_{t-1}] \right) \), 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
\[ \frac{1}{n} \sum_{t=1}^{n} \psi' \left( \frac{\varepsilon_t - (c^*/n) X_{t-1}}{\sigma_t} \right) \xrightarrow{\mathcal{L}} 0, \quad \| \psi' \| \leq 1, \quad (A.14) \]
and this can be used to prove
\[ J^*_n = \frac{1}{n^2} \sum_{t=1}^{n} Z_{t-1}^2 \psi' \left( \frac{\eta_t - c^*}{\sigma_t} Z_{t-1} \right) \xrightarrow{\mathcal{L}} \mathcal{I} \int_0^1 Z_0(s)^2 ds. \quad (A.15) \]

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
\[ \max_{1 \leq t \leq n} \left| \tilde{\sigma}_t^2 - \sigma_t^2 \right| \xrightarrow{P} 0. \quad (A.16) \]
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
\[
\hat{\sigma}_t^2 - \sigma_t^2 = R_t^a + \sigma_t^2 R_t^b + R_t^c + R_t^d,
\]
where
\[
R_t^a = \sum_{j=0}^{N-1} w_{jN} (\sigma_{t-1-j}^2 - \sigma_t^2), \quad R_t^b = \sum_{j=0}^{N-1} w_{jN} (\eta_{t-1-j}^2 - 1),
\]
\[
R_t^c = \sum_{j=0}^{N-1} w_{jN} (\sigma_{t-1-j}^2 - \sigma_t^2)(\eta_{t-1-j}^2 - 1), \quad R_t^d = \sum_{j=0}^{N-1} w_{jN} (\hat{\varepsilon}_{t-1-j}^2 - \varepsilon_{t-1-j}^2).
\]
Hansen’s proof that $\max_{N<t \leq n} |R_t^d| \overset{P}{\rightarrow} 0$, $\max_{N<t \leq n} |\sigma_t^2 R_t^b| \overset{P}{\rightarrow} 0$ and $\max_{N<t \leq n} |R_t^c| \overset{P}{\rightarrow} 0$ is directly applicable here. For the fourth term, we note that $\hat{\varepsilon}_t = \varepsilon_t + (c/n) X_{t-1}$, 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,
\[
\left| \sum_{j=0}^{N-1} w_{jN} (\hat{\varepsilon}_{t-1-j}^2 - \varepsilon_{t-1-j}^2) \right| \leq 2 \left[ \frac{1}{n} \sum_{j=0}^{N-1} w_{jN} \varepsilon_{t-1-j} X_{t-2-j} \right] c_n + \left[ \frac{1}{n^2} \sum_{j=0}^{N-1} w_{jN} X_{t-2-j}^2 \right] c_n^2. \tag{A.18}
\]
Analogous to Hansen (1995), p. 1130, it follows that
\[
\max_{N<t \leq n} \left| \frac{1}{n} \sum_{j=0}^{N-1} w_{jN} \varepsilon_{t-1-j} X_{t-2-j} \right| \overset{P}{\rightarrow} 0, \tag{A.19}
\]
\[
\max_{N<t \leq n} \left[ \frac{1}{n^2} \sum_{j=0}^{N-1} w_{jN} X_{t-2-j}^2 \right] = O_P \left( \frac{N^2}{n^2} \right) \overset{P}{\rightarrow} 0, \tag{A.20}
\]
such that $\max_{N<t \leq n} |R_t^d| \overset{P}{\rightarrow} 0$. This proves (A.16) 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.

\begin{proof}[Proof of Theorem 3] Consistency of $\hat{S}_n$ follows directly from Lemma 2, Theorem 2, and the continuous mapping theorem. For $\hat{S}_n$, we use $\Delta X_t = \varepsilon_t + (c/n) X_{t-1}$, and a first-order Taylor series expansion of the function $g(\varepsilon, \sigma) = (\psi(\varepsilon) / \sigma)$ about the point $(\varepsilon, \sigma) = (\hat{\varepsilon}_t, \hat{\sigma}_t)$:
\[
\psi \left( \frac{\Delta X_t}{\hat{\sigma}_t} \right) - \psi \left( \frac{\hat{\varepsilon}_t}{\hat{\sigma}_t} \right) \approx \psi' \left( \frac{\hat{\varepsilon}_t}{\hat{\sigma}_t} \right) \frac{c}{n \hat{\sigma}_t} \varepsilon_{t-1} - \psi' \left( \frac{\hat{\varepsilon}_t}{\hat{\sigma}_t} \right) \frac{\hat{\varepsilon}_t}{\hat{\sigma}_t^2} (\hat{\sigma}_t - \sigma_t), \tag{A.21}
\]
where the approximation error is of lower order in probability than the right-hand side terms. This leads to
\[
\hat{S}_n = \frac{1}{n} \sum_{t=1}^n \frac{1}{\hat{\sigma}_t} X_{t-1} \psi'(\eta_t) + \frac{c}{n^2} \sum_{t=1}^n \frac{1}{\hat{\sigma}_t \sigma_t} X_{t-1} \psi'(\eta_t) - \frac{1}{n} \sum_{t=1}^n \frac{1}{\hat{\sigma}_t} X_{t-1} \psi'(\eta_t) \eta_t \frac{\hat{\sigma}_t - \sigma_t}{\sigma_t} + o_P(1). \tag{A.22}
\]
The first two right-hand-side terms together converge to $S = T^{1/2} \int_0^1 Z_c(s) dB(s) + cT \int_0^1 Z_c(s)^2 ds$, 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 (??) is not satisfied, then the third term does not vanish, and hence the effect of estimating $\sigma_t$ is no longer negligible.
\end{proof}
References


Ling, S., W. K. Li and M. McAleer (2003), “Estimation and testing for unit root processes with


