Fractional integration and cointegration in financial time series
Stakenas, P.

Citation for published version (APA):
Stakenas, P. (2012). Fractional integration and cointegration in financial time series Amsterdam: Thela Thesis

General rights
It is not permitted to download or to forward/distribute the text or part of it without the consent of the author(s) and/or copyright holder(s), other than for strictly personal, individual use, unless the work is under an open content license (like Creative Commons).

Disclaimer/Complaints regulations
If you believe that digital publication of certain material infringes any of your rights or (privacy) interests, please let the Library know, stating your reasons. In case of a legitimate complaint, the Library will make the material inaccessible and/or remove it from the website. Please Ask the Library: http://uba.uva.nl/en/contact, or a letter to: Library of the University of Amsterdam, Secretariat, Singel 425, 1012 WP Amsterdam, The Netherlands. You will be contacted as soon as possible.
Chapter 2

Estimation and inference of fractionally integrated time series via autoregressive approximation

2.1 Introduction

In this chapter we study estimation and inference in fractionally integrated time series models. The class of fractionally integrated time series bridges the gap between weakly stationary (or so called \(I(0)\)) and unit root (or: \(I(1)\)) processes, thus constituting an important part of applicable econometric time series models. In our study we focus on univariate type II fractionally integrated processes (for comparison with type I processes, see Marinucci and Robinson (1999)), which is defined as: 

\[ X_t = (1 - L)^{-d} u_t 1_{t>0}, \ t \in \mathbb{N}, \]

for \(d \in \mathbb{R}\), where \(L\) is the lag operator and a process \(u_t\) has a continuous, bounded and non-zero everywhere spectral density. The literature on inference on the parameter \(d\) is vast and still growing and can be divided into parametric and semiparametric methods. The works of Sowell (1990), Beran (1995), Hosoya (1997), Tanaka (1999), Nielsen (2004) among others, belong to the former category. Their inferential methods rely either on Gaussian or Whittle likelihoods and assume correct specification of the model. Semiparametric methods typically are based on approximating the spectral density of time series in a degenerating band around the zero frequency and have been studied in Robinson (1995), Andrews and Sun (2004), Shimotsu and Phillips (2005) and others. The trade-off between estimation methods is as usual: parametric methods offer better estimates in correctly specified models, but suffer from misspecification, while semiparametric methods weaken modeling assumptions offering richer model frameworks. Our approach is based on a finite-order autoregressive approximation of a fractionally integrated autoregressive process.
Estimation and inference of fractionally integrated time series

of (possibly) infinite order and is semiparametric in nature, although it is likelihood-based and relies upon a suitable approximation of the conditional sum of squared residuals. To fix ideas, suppose that an observed time series $X_t$ is generated by the following process:

$$
\phi(L)\Delta^d X_t 1_{t \geq 0} = \varepsilon_t, \ t = 0, 1, \ldots
$$

(2.1)

where $\phi(L)$ is an invertible lag polynomial with absolutely summable coefficients and $\varepsilon_t$ is an i.i.d. sequence. We may define residuals of the model obtained by filtering $X_t$ both with a (truncated) fractional difference filter and a finite-degree autoregressive lag polynomial and form an objective function equal to the conditional sum of squared residuals. We call the resulting estimator the conditional-sum-of-squares (CSS) estimator in line with Robinson (2006), Nielsen (2011). The latter paper also considered the CSS estimator, albeit in a parametric framework, whereas our framework does not require a correct specification of the underlying structure of the DGP. Given the degree of the fitted lag polynomial is not too small and not too big, the approximation is good enough to achieve consistent inference on model parameters. The idea in somewhat different frameworks was exploited in Said and Dickey (1984), Saikkonen (1992) and others. A similar method has been applied to long memory models in Bhansali et al. (2006), where estimation of the long memory parameter in the stationary region was studied by frequency domain methods. Our results extend and complement this study in a few respects: i) we study type II processes in the whole range of fractional values $d$; ii) assumptions on the growth rate of lag length are weaker; iii) the analysis is done in the time domain, which may be preferable for applied researchers.

Our results have applications for (optimal) long memory testing. Optimal testing procedures in parametric models were studied in Robinson (1994) and later contributions include Tanaka (1999), Nielsen (2004) and others. Another strand of literature considers regression-based methods: Wald-type (introduced by Dolado et al. (2002)) and Langrange multiplier (Breitung and Hassler (2002)) tests requiring a known and finite order of autoregression. We extend their results and show that both testing procedures are also applicable for infinite order autoregressive processes using CSS estimates for the unknown parameters.

The rest of this chapter is organized as follows: Section 2.2 discusses the model, assumptions and presents the main results. Section 2.3 considers inference in the long memory parameter. The results of Monte Carlo simulations are presented in Section 2.4, Section 2.5 provides an empirical application, while Section 2.6 concludes. Proofs are given in the Appendix.
2.2 Estimation of fractionally integrated series

2.2.1 Model and assumptions

Suppose, we observe a realization $X_1, \ldots, X_T$ of a time series process satisfying:

**Assumption 2.1.** The observed univariate time series $X_t$ satisfies:

$$\Delta^d_t X_t = u_t, \quad t = 0, 1, \ldots$$ (2.2)

where $d \in \mathbb{R}$ and the expression $\Delta^d_t$ denotes a truncated fractional differencing operator:

$$\Delta^d_t X_t = (1 - L)^d X_{t|t-0} = \sum_{i=0}^{t} \frac{(-d+1) \cdots (-d+t-1)}{t!} X_{t-i}.$$

**Assumption 2.2.** $u_t = \varphi(L)\varepsilon_t$, where $\varphi(z)$ is a power series such that $\sum_i |\varphi_i| < \infty$ and $\sum_i |\phi| < \infty$, where $\phi(z) = \varphi^{-1}(z) = 1 - \sum_{i=1}^{\infty} \phi_i z^i$ and $\varepsilon_t$ is an i.i.d. series with $E\varepsilon_t = 0$, $E\varepsilon_t^2 = \sigma^2$ and $E\varepsilon_t^6 < \infty$.

Assumptions 2.1-2.2 define the data generating process and are relatively weak, requiring six moments of $\varepsilon_t$ and absolute summability of the coefficients of $\phi(z)$ and its inverse.

2.2.2 Parameters and the likelihood function

Denote the parameters of the model by $\kappa = (\sigma^2, \theta')'$, where $\theta' = (d, \phi')$, $\phi' = (\phi_1, \phi_2 \ldots)$ is an infinite vector with coefficients $\phi_i$ of the power series $\phi(z)$: $\phi(z) = 1 - \sum_{i=1}^{\infty} \phi_i z^i$. Throughout the chapter we use subscript “0” for the true values of the model: $\kappa_0 = (\sigma^2_0, \theta'_0)$, $\theta'_0 = (d_0, \phi'_0)$, $\phi'_0 = (\phi_{01}, \phi_{02} \ldots)$. We define a set in the Hilbert space $l^1 = \{(a_1, a_2, \ldots)| \sum_{i=1}^{\infty} |a_i| < \infty\}$ containing the true parameter values: $D_\kappa = D_\sigma \times D_d \times D_\phi(\eta)$, where $D_\sigma = [\eta, \eta^{-1}]$, $D_d = [d_0 - \Delta T, d_0 + \Delta T]$, for some $\Delta T > 0$ depending on $T$ and $D_\phi(\eta) = \{\phi \in l^1 : \sum_i |\phi_i - \phi_0i| \leq \eta^{-1}\}$. Further, we rule out series $\phi(z)$ with unit roots. Define space of power series $\phi(z)$ bounded away from zero on the complex unit circle: $D_\phi(\epsilon) = \{\phi \in l^1 : 0 < \epsilon \leq |\phi(z)| \leq \epsilon^{-1}, \forall z \in \mathbb{C} : |z| = 1\}$ for some $1 > \epsilon > 0$ and consider parameter space $D_\kappa(\eta, \epsilon) = D_\sigma \times \Theta_T = D_\sigma \times D_d \times (D_\phi(\eta) \cap D_\phi(\epsilon))$. Then $D_\kappa(\eta, \epsilon)$ is a closed bounded set and for $\min\{1, \sigma^2_0, \sigma^{-2}_0\} > \eta > 0$: $\kappa_0 \in D_\kappa$.

For $T > k$ we define the approximate conditional Gaussian log-likelihood function\footnote{In the following, we refer to it as simply an “objective function”}. for
the model (2.2):

\[ l_T(\kappa) = -\frac{T}{2} \log(2\pi) - \frac{T}{2} \log(\sigma^2) - \frac{1}{2\sigma^2} \sum_{t=k}^{T} (\Delta^d d X_t - \sum_{i=1}^{k} \phi_i \Delta^d d X_{t-i})^2. \] (2.3)

We may concentrate out the parameter \( \sigma^2 \) and consider the scaled profile objective function:

\[ l_T(\theta) = -\log \hat{\sigma}^2_T(\theta), \] (2.4)

\[ \hat{\sigma}^2_T(\theta) = T^{-1} \sum_{t=k}^{T} (\Delta^d d X_t - \sum_{i=1}^{k} \phi_i \Delta^d d X_{t-i})^2. \] (2.5)

The residuals \( \varepsilon_{t,k}(\theta) = \Delta^d d X_t - \sum_{i=1}^{k} \phi_i \Delta^d d X_{t-i} \) are neither uncorrelated nor Gaussian for any value of \( \theta \), but \( \varepsilon_{t,k}(\theta_0) \) is “almost” uncorrelated and the temporal dependence is asymptotically negligible provided \( k \) grows fast enough with the sample size \( T \). We assume the following on growth rates of lag length \( k \):

**Assumption 2.3.** The growth rate of \( k \) satisfies:

1. \( kT^{-1/4} = o(1) \),

2. \( \sqrt{T} \sum_{|j|>k} |\phi_j| = o(1) \).

Assumption 2.3.1 puts an upper bound on the rate \( k \): it does not increase faster than \( T^{1/4} \), while Assumption 2.3.2 puts a lower bound on the rate. Obviously, the bound is infeasible and depends on the structure of the linear process \( u_t \), but for ARMA-type processes \( u_t \) it translates into: \( k/\log T \to \infty \).

Given the method of estimation we are able to estimate only the first \( k \) components of the infinite vector \( \phi_0 \), hence we define the estimator of the model as:

\[ \hat{\theta} = \arg \max_{\theta \in \Theta^k_T} l_T(\theta), \] (2.6)

where \( \Theta^k_T \) is a restricted parameter space: \( \Theta^k_T = \pi_k \Theta_T \) with \( \pi_k \) being a projection \( l^1 \to \mathbb{R}^{k+1} \) defined as: \( \pi_k((\theta_1, \ldots, \theta_k, \theta_{k+1}, \ldots)) = (\theta_1, \ldots, \theta_{k+1}) \). In the following, it will be implicit that \( \hat{\theta} \) is of dimension \( k + 1 \).

Note, that the estimator (2.6) is the same as the minimizer of the conditional sum of squared residuals \( \hat{\sigma}^2_T(\theta) \) and in the following we refer to the estimator (2.6) as the conditional-sum-of-squares (CSS) estimator.
2.2 Estimation of fractionally integrated series

2.2.3 Main theorems

The following theorem is the main result of the section and gives an explicit joint asymptotic distribution of the first \( p + 1 \) coordinates of \( \hat{\theta} \):

**Theorem 2.1.** Fix \( p \) and denote the first \( p + 1 \) coordinates of \( \theta \in \Theta_T \) as \( \theta^p \). Suppose, that Assumptions 2.1-2.3 hold and, in addition, \( \Delta_T = o(k^{-2}) \). Then it holds:

\[
\sqrt{\frac{T}{k}} \left( \hat{\theta}^p - \theta^p_0 \right) \xrightarrow{d} \Gamma z, \tag{2.7}
\]

where \( z \sim N(0, 1) \) is a standard normal random variable and \( \Gamma \) is a deterministic vector: \( \Gamma = (1, \gamma_1, \ldots, \gamma_p)' \) with \( \gamma_i = -\sum_{j=1}^{i} j^{-1} \phi_i - j \). In particular, \( \sqrt{T/k}(\hat{d} - d_0) \xrightarrow{d} N(0, 1) \).

Further, denote \( \Omega = \text{Cov}(U_t) \), where \( U_t = (u_{t-1}, \ldots, u_{t-k})' \), a selection matrix \( S_p = (I_p, 0_{p \times k-p}) \) and a \( p \times (p-1) \) matrix \( \Gamma_\perp \), orthogonal to \( \Gamma \): \( \Gamma_\perp \Gamma = 0 \). Then it holds:

\[
\sqrt{T} \Gamma_\perp' \left( \hat{\theta}^p - \theta^p_0 \right) \xrightarrow{d} N(0, \Gamma_\perp' S_p \Omega^{-1} S_p' \Gamma_\perp). \tag{2.8}
\]

**Remark 2.1.** The additional condition to Assumptions 2.1-2.3 in Theorem 2.1 requires the parameter space to “shrink” around the true value of \( d_0 \) and reflects lack of (global) consistency of \( \hat{d} \). In practice this may be solved using an \( o_p(k^{-2}) \) consistent pre-estimate of \( d \) as the starting value in the optimization procedure. For example, ARMA-type processes \( u_t \) entail \( O_p(T^{-\delta}) \) consistency of the pre-estimate with any \( \delta > 0 \).

**Remark 2.2.** The possible consistency rate of the estimator \( \hat{\theta}^p \) depends on the lower bound on the rate \( k \) in the Assumption 2.3.2, but since for ARMA-type processes \( u_t \) it translates into: \( k/ \log T \to \infty \), thus consistency rate of \( \sqrt{T/\log T} \) for \( \hat{\theta}^p \) may be achieved.

**Remark 2.3.** From Theorem 2.1 it follows that the variance-covariance matrix of \( \hat{\theta}^p \) has a reduced rank one. This is due to the structure of \( \hat{\theta}^p \): it can be decomposed into two parts, one being \( O_p(\sqrt{k/T}) \), while another is \( O_p(T^{-1/2}) \). Thus, linear combinations \( \Gamma_\perp \hat{\theta}^p \) converge to their true values at a rate \( \sqrt{T} \), rather than \( \sqrt{T/k} \).

**Remark 2.4.** Suppose that our observed process is in fact of finite autoregressive order \( k_0 \), i.e. \( \phi_i = 0, i > k_0 \). Then Theorem 2.1 still holds, although optimal inference on the parameter \( \theta \) is achieved by estimating \( \theta \) with a fixed number of lags \( k_0 \). In this case \( \sqrt{T} \)-consistency for the parameter \( \theta \) holds, with the asymptotic variance-covariance matrix of the estimator given by the inverse of Fisher information under Gaussianity:

\[
\Xi = \begin{pmatrix}
\frac{\pi^2}{6} & \xi' \\
\xi & \Omega
\end{pmatrix},
\]
Estimation and inference of fractionally integrated time series

where $\xi' = \text{Cov}(\ln \Delta \varepsilon_t, U_t)$, $U_t = (u_{t-1}, \ldots, u_{t-k_0})'$, $\Omega = \text{Cov}(U_t)$ and $\ln \Delta = \ln(1 - L)$. The reader is referred to Nielsen (2004) for details.

**Remark 2.5.** Suppose that observed time series has some deterministic components, i.e. the observed $X_t$ is such that: $\phi(L) \Delta^d X_t - \gamma' z_t = \varepsilon_t$, where $z_t$ is a deterministic $p \times 1$ vector series. We may concentrate out the parameters $\gamma$ from objective function and the above results will hold under the same assumptions (for details, refer to Nielsen (2004)).

**Remark 2.6.** In practical implementation of the estimator (2.6), an identification issue for the coefficients $\phi$ arises, i.e. the parameter space for $\phi$ has to be restricted so that $\phi(z)$ does not have roots on or inside the unit circle. In practice it may be hard to achieve, especially given that $\hat{\phi}$ are regression estimates of $\Delta^d X_t$ on its $k$ lags. One of the ways to achieve the goal is to consider a penalized objective function:

$$ l_T^P(\theta) = l_T(\theta) + \text{Pen}_T(\theta). \quad (2.9) $$

The idea of penalty term $\text{Pen}_T(\theta)$ is to give an (asymptotically negligible) penalty for the objective function $l_T(\theta)$ which is proportional to the distance of the roots of $\phi(z)$ to the unit circle. One possible penalty could be simply the unconditional scaled Gaussian log-likelihood of the first $k$ observations:

$$ \text{Pen}_T(\theta) = \log L_X(X_1, \ldots, X_k). \quad (2.10) $$

where $L_X(x_1, \ldots, x_k)$ is the likelihood function of the first $k$ observations with Gaussian innovations $\varepsilon_t$ with variance $\hat{\sigma}^2_t(\theta)$. In this case $\phi(L)$ is implicitly forced to have roots outside unit circle. More generally, Theorem 2.1 holds, if the penalty term is a smooth function of data and parameters satisfying:

- $\sup_{\theta \in \Theta_T} |\text{Pen}_T(\theta)| = o_p(1)$,
- $\sup_{\theta \in \Theta_T} |D^i_\theta \text{Pen}_T(\theta)| = o_p(1)$,
- $\sqrt{T} D_\theta \text{Pen}_T(\theta_0) = o_p(1)$.

Here $D^i_\theta \text{Pen}_T(\theta)$ denotes $i$-th derivative of $\text{Pen}_T(\theta)$ w.r.t. $\theta$.

**Remark 2.7.** The model (2.2) is restrictive, since, in particular, it imposes $X_0 = u_0$, what may not be a reasonable assumption in some empirical applications. However, for $d \geq 0$ Theorem 2.1 holds under more general assumption $X_0 = O_p(1)$. We may also consider model (2.2) with untruncated fractional differencing operator: $\Delta X_t = u_t, t = 0, 1, \ldots$, with uniformly bounded initial values: $|X_{-t}| < M, \forall t \in \mathbb{N}$, similarly as in Johansen and Nielsen...
2.3 Hypothesis testing

In this section we consider the testing problem in the model (2.2), where the parameter $d$ is of interest, as it is in most empirical applications. We employ a regression-based approach and extend the Lagrange multiplier (LM) test studied in Breitung and Hassler (2002), and the optimal Wald test considered by Lobato and Velasco (2007), where the test implementation required correct specification of (finite) order of autoregression. We show how CSS estimates (2.6) can be used to construct respective test statistics under more general infinite order autoregressive processes and prove their asymptotic properties. We also show that these two tests are asymptotically equivalent under a local alternative framework.

We begin with the LM test. The test implementation is a two-step procedure requiring short-run noise estimation in the first step and regression with an auxiliary regressor in the second. Consider the null hypothesis: $d = d_0$ and denote $u_t = \Delta^d X_t$. Then restricted estimates of short-run noise parameters of the model (2.2) are simply the OLS estimates $\tilde{\phi}' = (\tilde{\phi}_1, \ldots, \tilde{\phi}_k)$ in the first-step regression:

$$u_t = \sum_{i=1}^{k} \phi_i u_{t-i} + \eta_t. \quad (2.11)$$

We form the auxiliary regressor $z^1_t = \sum_{j=1}^{t} j^{-1}\tilde{\epsilon}_{t-j}$ with $\tilde{\epsilon}_t = (1 - \sum_{i=1}^{k} \tilde{\phi}_i L^i)u_t$ and consider second-step regression:

$$u_t = \rho z^1_t + \sum_{i=1}^{k} \phi_i u_{t-i} + \eta_t. \quad (2.12)$$

If $X_t$ is generated as in (2.2), then under the null it holds: $\rho = 0$ with $\eta_t = \varepsilon_t + \sum_{i=k+1}^{\infty} \phi_i u_{t-i}$, and proper lag selection warrants a consistent procedure enabling to test the null using

---

We do not prove any of these statements, but they are easily verifiable given Remark 2.6: one needs to show that allowing for weaker assumptions on the initial value(s) introduces asymptotically negligible term in the likelihood, which satisfies conditions in the Remark 2.6.
the $t$-statistic for the coefficient $\rho$ in the regression (2.12):

$$t^{LM}_\rho = \frac{\hat{\rho}}{\hat{s}_\rho}.$$  

Here $s^2_\rho$ is the variance of the OLS estimator $\hat{\rho}$ and may be consistently estimated as:

$$\hat{s}^2_\rho = \left[ \hat{\sigma}^2 \sum_{t=k}^T V_t^* V_t'^* \right]^{-1}_{11}$$  

(2.13)

where $[\cdot]_{11}$ denotes the first diagonal element, $V_t^* = (z^1_t, u_{t-1}, \ldots, u_{t-k})'$ and $\hat{\sigma}^2$ is any consistent estimate of $\sigma_0^2$.

If the order of autoregression in (2.2) is finite and known, Breitung and Hassler (2002) showed that $t^{LM}_\rho \xrightarrow{d} N(0,1)$, whereas we show that convergence also holds in case of infinite order autoregressions. A similar idea was explored in Demetrescu et al. (2008), although with a somewhat different auxiliary regressor, namely, $\tilde{z}^1_t = \sum_{j=1}^{t} j^{-1} u_{t-j}$, rather than $z^1_t$ was used in the regression (2.12), which warranted a one-step testing procedure.

In a series of papers Dolado et al. (2002), Lobato and Velasco (2006), Lobato and Velasco (2007) analyzed (optimal) regression-based Wald tests for unit roots and derived their properties for finite (and known) order fractionally integrated autoregressive processes. We derive the properties of the test allowing for infinite order autoregression. Again consider the null hypothesis $d = d_0$ and denote CSS estimates (2.6): $\hat{\theta} = (\hat{d}, \hat{\phi}_1, \ldots, \hat{\phi}_k)$. Then an auxiliary regressor is constructed as (for discussion and motivation see Lobato and Velasco (2007)):

$$z^2_t(\hat{\theta}) = \Delta_{d-d}^d u_t - \sum_{i=1}^k \hat{\phi}_i u_{t-d}.$$  

Again, if $X_t$ is generated by (2.2), then under the null, $\rho = 0$ in the regression:

$$u_t = \rho z^2_t + \sum_{j=1}^k \phi_j u_{t-j} + \eta_t$$  

(2.14)

and we achieve consistent inference on the parameter $d$ considering the $t$-statistic $t^W_\rho$ for the parameter $\rho$ in the regression (2.14) with standard errors (2.13). Under the null, both test statistics $t^W_\rho$, $t^{LM}_\rho$ are asymptotically normally distributed, while the tests are also asymptotically equivalent under local alternatives in an infinite autoregression framework:

**Theorem 2.2.** Suppose that the conditions of Theorem 2.1 hold and consider the null
2.4 Finite sample simulations

hypothesis \( d = d_0 \). Then:

- Under the null, \( t^{LM}_\rho, t^W_\rho \xrightarrow{d} N(0, 1) \).
- Under local (Pitman) alternatives \( d = d_0 + \delta \sqrt{k/T} \), \( t^{LM}_\rho, t^W_\rho \xrightarrow{d} N(\delta, 1) \), if \( kT^{-1/6} = o(1) \).

Using methods similar as in Lobato and Velasco (2007) it is not difficult to show that the Wald test is consistent against fixed one-sided (\( d_1 < d_0 \)) alternatives, whereas analysis of the power of the LM test against fixed alternatives is somewhat more difficult, since in this case in the first step estimation of a misspecified model is performed (in other words, the alternative is not contained in the estimated model), but simulations show that the power is non-trivial.

Remark 2.8. It is possible to consider three classical (LR, LM and Wald) tests for the null hypothesis \( d = d_0 \) based on asymptotic standard errors of CSS estimator of \( \hat{d} \); however unreported simulations showed poor finite sample properties of these tests.

Remark 2.9. In the \( t \)-statistics we may also use White standard errors:

\[
\hat{s}^2 = \left[ (\sum_{t=k}^{T} V_t^* V_t^{**})^{-1} (\sum_{t=k}^{T} V_t^* V_t^{**} \eta_t^2) (\sum_{t=k}^{T} V_t^* V_t^{**})^{-1} \right]_{11},
\]

with \( V_t^* = (z_{it}, u_{i,t-1}, \ldots, u_{i,t-k}) \), which would allow for heteroscedasticity-robust inference (see Demetrescu et al. (2008)).

2.4 Finite sample simulations

2.4.1 Monte Carlo setup

We simulate the following fractionally integrated time series: \( X_t = \Delta^{-d}_+ u_t \), where \( u_t \) is ARMA (1, 1) process \( u_t = \phi u_{t-1} + \varepsilon_t + \psi \varepsilon_{t-1} \) and \( \varepsilon_t \) is i.i.d. \((0, 1)\) Gaussian time series. The time series is simulated with the following parameter values: \( d = 0 \), \( \psi = 0.5 \), \( \phi = -0.5, 0, 0.5, 0.8 \) and the sample size \( T = 128, 256 \). The number of simulations is \( R = 10000 \).

\(^3\)Although this thesis does not consider heteroscedasticity-robust inference, our (unreported) simulations show good finite sample performance of considered tests in heteroscedastic framework when using White standard errors.
In estimation we use the following penalized objective function (2.9):

\[ l_T^P(d) = -\frac{T}{2} \log \hat{\sigma}_T^2(d) - \frac{1}{2} \log |R_k(d)|, \]  

\[ \hat{\sigma}_T^2(d) = T^{-1} \sum_{t=k}^T (\Delta_d X_t - \sum_{i=1}^k \hat{\phi}_i \Delta_d X_{t-i})^2, \]  

(2.15)

where \( \sigma^2 R_k(d) \) is \( k \times k \) variance-covariance matrix of \((u_k(d), \ldots, u_1(d))' \) with \( u_t(d) = (1 - \sum_{i=1}^k \hat{\phi}_i L^i)^{-1} \epsilon_t \) and

\[ (\hat{\phi}_1, \ldots, \hat{\phi}_k)' = \left( \sum_{t=k}^T U_t U_t' \right)^{-1} \sum_{t=k}^T U_t \Delta_d^d X_t, \]  

(2.17)

where \( U_t = (\Delta_d^d X_{t-1}, \ldots, \Delta_d^d X_{t-k})' \). Note that the distribution of the objective function \( l_T(d - d_0) \) does not depend on \( d_0 \), hence there is no need to simulate the performance of the CSS estimator for different values of \( d_0 \) other than the one selected.

We study the performance of CSS estimator of \( d \) based on four different lag \( k \) selection rules: Akaike and Bayesian information criteria-based rules and deterministic rules \( pK = [K(T/100)^{0.25}] \) for \( K = 4, 12 \) as suggested by Schwert (1989). We compare CSS estimator to the exact local Whittle estimator (ELWE) due to Shimotsu and Phillips (2005), defined as:

\[ \hat{d} = \arg \min_{d \in [-d_1, d_2]} \left( \log \left( \frac{1}{m} \sum_{j=1}^m I_{\Delta_d^d X}(\lambda_j) \right) - \frac{2d}{m} \sum_{j=1}^m \log \lambda_j \right), \]  

(2.18)

where \( \lambda_j = 2\pi j/T \), \( I_{\Delta_d^d X}(\lambda) \) is the periodogram of \( \Delta_d^d X_t \) at \( \lambda \), \(-\infty < d_1 < d_2 < \infty \) are lower and upper bounds of the admissible values of \( d \) and the number \( m \) is a bandwidth parameter, i.e. the number of periodogram ordinates used in estimation. In the simulations we set \( m = [T^{0.65}] \) and \([d_1, d_2] = [-1.5, 1.5] \).

Further, we study the performance of Wald and LM tests with the above-described simulation designs for the null \( d = d_0 \) under different alternatives. We compare performance of four different lag selection rules in terms of size and both size-adjusted and non-adjusted power.

### 2.4.2 Simulation results

Table 2.1 reports the bias and root mean squared error (RMSE) of the CSS estimator based on four lag selection rules and compares them to those of ELWE. We see that
2.4 Finite sample simulations

deterministic lag selection rule $p4$ dominates $p12$ in terms of RMSE for all values of $\phi$ and $T$ and performs fairly stably in terms of RMSE for different values of $\phi$. It gives worse estimates than IC-based rules for $\phi = -0.5, 0, 0.5$, however performs on par for $\phi = 0.8$. Comparing IC-based rules we see that BIC typically is better than AIC for less persistent error terms (i.e. $\phi = -0.5, 0, 0.5$), but is worse for $\phi = 0.8$. This could be explained by the fact that BIC penalizes additional parameters in a model more severely than AIC and hence selects more parsimonious models, while in case of strong persistence in $u_t$, the CSS estimator with more lags performs relatively better in comparison with conservative models. However, for $\phi = 0.8$, $p4$ performs best, although it uses less lags than $p12$. One explanation could be that it is due to positive bias of CSS estimator, which typically arises in models with strongly persistent error term processes, if large number of lags is used in estimation. Hence our recommendation would be to use conservative criteria for lag selection (like BIC) for CSS estimation with moderate degree of persistence in the error term and more liberal (but not too much), like AIC or $p4$, when the errors are persistent (in practice, strong persistence could be detected by inspecting the roots of estimated autoregressive polynomial or analyzing the impulse response function).

ELWE is generally on par with IC-based CSS estimators, for $\phi = -0.5, 0, 0.5$, but it is significantly worse than any CSS estimator for $\phi = 0.8$. Although IC-based CSSE compares well to ELWE for moderately persistent $u_t$ (for $\phi = 0$ ELWE marginally outperforms CSS since the bias of ELWE effectively diminishes in this case), in case of strong persistence the CSS estimator is a significantly better option than ELWE.

Tables 2.2 and 2.3 report empirical size of LM and Wald tests based on 4 different lag selection rules for 3 different nominal significance levels. Generally, the tests are oversized in nearly all the cases with Wald tests being more oversized than LM tests, which is quite understandable, given that the short-run coefficients are estimated under the null in the LM case, whereas in the Wald case, $d$ is also estimated. The difference between empirical sizes of tests is especially visible with the BIC lag selection rule. Although the BIC-based CSS estimator $\hat{d}$ does indeed perform worst with $\phi = 0.8$, thus distorting the size of the Wald test, another possible explanation for the difference in empirical size of the Wald and LM test is the number of lags used in estimation: on average, BIC is the most conservative rule of the four, followed by AIC, $p4$ and $p12$ (which shows the least difference between Wald and LM tests’ empirical sizes), while the use of more lags “filters out” temporal dependence in $u_t$, reducing the effects of bias in $\hat{d}$.

Deterministic rules display very good size properties of the LM test for all significance levels and values of $\phi$, whereas the Wald test is slightly oversized, with the $p4$ rule marginally more than the $p12$. As noted, IC-based tests are oversized, which is particularly
Table 2.1: Monte Carlo simulation results for estimation of $d$

<table>
<thead>
<tr>
<th>$T$</th>
<th>φ</th>
<th>Bias</th>
<th>RMSE</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>p4</td>
<td>p12</td>
</tr>
<tr>
<td>128</td>
<td>-0.5</td>
<td>-0.09</td>
<td>-0.08</td>
</tr>
<tr>
<td>0</td>
<td></td>
<td>-0.04</td>
<td>-0.03</td>
</tr>
<tr>
<td>0.5</td>
<td></td>
<td>0.03</td>
<td>0.06</td>
</tr>
<tr>
<td>0.8</td>
<td></td>
<td>0.18</td>
<td>0.22</td>
</tr>
<tr>
<td>256</td>
<td>-0.5</td>
<td>-0.07</td>
<td>-0.08</td>
</tr>
<tr>
<td>0</td>
<td></td>
<td>-0.06</td>
<td>-0.04</td>
</tr>
<tr>
<td>0.5</td>
<td></td>
<td>0.03</td>
<td>0.02</td>
</tr>
<tr>
<td>0.8</td>
<td></td>
<td>0.07</td>
<td>0.13</td>
</tr>
</tbody>
</table>

Table 2.2: Empirical size for testing $d$ with LM test

<table>
<thead>
<tr>
<th>$T$</th>
<th>φ</th>
<th>10%</th>
<th>5%</th>
<th>1%</th>
<th>10%</th>
<th>5%</th>
<th>1%</th>
<th>10%</th>
<th>5%</th>
<th>1%</th>
<th>10%</th>
<th>5%</th>
<th>1%</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>p4</td>
<td>p12</td>
<td>AIC</td>
<td>BIC</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>128</td>
<td>-0.5</td>
<td>0.102</td>
<td>0.053</td>
<td>0.011</td>
<td>0.104</td>
<td>0.050</td>
<td>0.010</td>
<td>0.159</td>
<td>0.091</td>
<td>0.019</td>
<td>0.117</td>
<td>0.065</td>
<td>0.014</td>
</tr>
<tr>
<td>0</td>
<td></td>
<td>0.093</td>
<td>0.046</td>
<td>0.010</td>
<td>0.105</td>
<td>0.051</td>
<td>0.011</td>
<td>0.146</td>
<td>0.075</td>
<td>0.022</td>
<td>0.206</td>
<td>0.136</td>
<td>0.071</td>
</tr>
<tr>
<td>0.5</td>
<td></td>
<td>0.089</td>
<td>0.045</td>
<td>0.009</td>
<td>0.108</td>
<td>0.055</td>
<td>0.014</td>
<td>0.151</td>
<td>0.082</td>
<td>0.019</td>
<td>0.158</td>
<td>0.089</td>
<td>0.022</td>
</tr>
<tr>
<td>0.8</td>
<td></td>
<td>0.090</td>
<td>0.044</td>
<td>0.008</td>
<td>0.105</td>
<td>0.052</td>
<td>0.011</td>
<td>0.158</td>
<td>0.095</td>
<td>0.029</td>
<td>0.275</td>
<td>0.197</td>
<td>0.102</td>
</tr>
<tr>
<td>256</td>
<td>-0.5</td>
<td>0.098</td>
<td>0.052</td>
<td>0.010</td>
<td>0.094</td>
<td>0.047</td>
<td>0.009</td>
<td>0.166</td>
<td>0.095</td>
<td>0.022</td>
<td>0.111</td>
<td>0.060</td>
<td>0.015</td>
</tr>
<tr>
<td>0</td>
<td></td>
<td>0.103</td>
<td>0.054</td>
<td>0.011</td>
<td>0.094</td>
<td>0.049</td>
<td>0.009</td>
<td>0.155</td>
<td>0.084</td>
<td>0.019</td>
<td>0.174</td>
<td>0.097</td>
<td>0.024</td>
</tr>
<tr>
<td>0.5</td>
<td></td>
<td>0.103</td>
<td>0.051</td>
<td>0.011</td>
<td>0.095</td>
<td>0.046</td>
<td>0.010</td>
<td>0.151</td>
<td>0.084</td>
<td>0.023</td>
<td>0.210</td>
<td>0.128</td>
<td>0.042</td>
</tr>
<tr>
<td>0.8</td>
<td></td>
<td>0.095</td>
<td>0.049</td>
<td>0.009</td>
<td>0.095</td>
<td>0.047</td>
<td>0.009</td>
<td>0.131</td>
<td>0.075</td>
<td>0.019</td>
<td>0.215</td>
<td>0.135</td>
<td>0.043</td>
</tr>
</tbody>
</table>

Noticeable for the Wald test. The AIC-based Wald and LM tests have rather stable empirical sizes for different $\phi$ and fixed $T$, but BIC-based test shows a lot of variation in empirical sizes for different values of $\phi$ and are severely oversized. This again can be attributed to a small number of lags (on average) used in estimation, which has potentially damaging effects as far as the test size is concerned. This difference in performance of deterministic versus IC-based lag selection rules confirms findings in Demetrescu et al. (2008), where deterministic rules performed better in terms of size than IC-based. For a more detailed analysis of the problem of size distortions in post-model-inference, refer to Demetrescu et al. (2009). We conclude that deterministic lag selection rules have far better size properties than IC-based tests and in particular they display both robustness against different short-run dynamics and against small sample sizes.

Next we turn our attention to power comparisons. Figures 2.2-2.7 show both adjusted and unadjusted power curves of LM and Wald tests against fixed alternatives $d_1 = -0.5, \ldots, 0.5$ for values $\phi = -0.5, 0, 0.5$. The Wald test is notably more powerful than LM for the sample size $T = 128$, but the difference becomes smaller for the sample size $T = 256$. With a few exceptions, size-adjusted power for one-sided alternatives $d_1 > d_0$ is highest for the BIC rule, while with alternatives $d_1 < d_0$ AIC and $p4$ have the highest
2.5 Empirical application

Most of established financial models for interest rates assume mean reversion of yields, although the assumption that yields have a unit root is quite common in econometric literature and most recently Iacone (2009) finds support for it (cf. Campbell and Shiller (1987)). In this section we test the hypothesis that interest rates are integrated of order 1. Strictly speaking, interest rates cannot be $I(1)$ processes, because the rates are positive variables, however if statistical properties of observed series match to those of $I(1)$ processes in a large enough finite sample, asymptotic inference is not distorted by the positivity constraints preventing interest rates to be $I(1)$ process, as argued in Hansen (2003).

### Table 2.3: Empirical size for testing $d$ with Wald test

<table>
<thead>
<tr>
<th>T</th>
<th>$\phi$</th>
<th>p4</th>
<th>p12</th>
<th>AIC</th>
<th>BIC</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>10%</td>
<td>5%</td>
<td>1%</td>
<td>10%</td>
<td>5%</td>
</tr>
<tr>
<td>128</td>
<td>-0.5</td>
<td>0.137</td>
<td>0.077</td>
<td>0.020</td>
<td>0.128</td>
</tr>
<tr>
<td></td>
<td>0</td>
<td>0.139</td>
<td>0.079</td>
<td>0.020</td>
<td>0.128</td>
</tr>
<tr>
<td></td>
<td>0.5</td>
<td>0.139</td>
<td>0.076</td>
<td>0.019</td>
<td>0.126</td>
</tr>
<tr>
<td></td>
<td>0.8</td>
<td>0.143</td>
<td>0.079</td>
<td>0.021</td>
<td>0.128</td>
</tr>
<tr>
<td>256</td>
<td>-0.5</td>
<td>0.132</td>
<td>0.070</td>
<td>0.015</td>
<td>0.113</td>
</tr>
<tr>
<td></td>
<td>0</td>
<td>0.127</td>
<td>0.067</td>
<td>0.014</td>
<td>0.111</td>
</tr>
<tr>
<td></td>
<td>0.5</td>
<td>0.118</td>
<td>0.063</td>
<td>0.015</td>
<td>0.112</td>
</tr>
<tr>
<td></td>
<td>0.8</td>
<td>0.110</td>
<td>0.059</td>
<td>0.014</td>
<td>0.111</td>
</tr>
</tbody>
</table>

power. Given good size properties of p4 rule, the rule-of-thumb would be to use tests based on this rule for one-sided alternatives $d_1 < d_0$. Unfortunately, the power of the p4 rule with alternatives $d_1 > d_0$ is disappointingly low, even for a sample size $T = 256$. On the other hand, the power of BIC-based tests is notably highest for one-sided alternatives $d_1 > d_0$, followed by AIC-based tests. However, precisely the IC-based tests showed the biggest size distortions, hence the choice of lag selection rule in this case seems to be a choice of power versus size.

Summarizing the Monte Carlo simulation results we could say that none of the lag selection rules or tests perform best and in practice it may be advisable to use a few different lag selections to get a more robust conclusion of a test based on the following regularities observed in simulations: i) deterministic lag selection rules typically show (very) good size properties, however they could have little power if the chosen number of lags is too big; ii) the empirical size of the Wald test is influenced by the accuracy of the estimate of $\hat{d}$ and the number of lags, and the use of a larger number of them can offset the bias in the estimation of $d$; iii) the Wald test is more powerful than LM in small samples, but the difference diminishes with increasing sample size.
Table 2.4: Results for testing the null $d = 1$ and CSS estimates of integration orders

<table>
<thead>
<tr>
<th>$k_{AIC}$</th>
<th>$W_{AIC}$</th>
<th>$W_{p4}$</th>
<th>$LM_{AIC}$</th>
<th>$LM_{p4}$</th>
<th>$d_{AIC}$</th>
<th>$d_{p4}$</th>
</tr>
</thead>
<tbody>
<tr>
<td>M1</td>
<td>0</td>
<td>0.074(0.786)</td>
<td>0.019(0.890)</td>
<td>0.088(0.766)</td>
<td>0.201(0.654)</td>
<td>0.996</td>
</tr>
<tr>
<td>M3</td>
<td>1</td>
<td>0.762(0.383)</td>
<td>1.813(0.178)</td>
<td>0.126(0.722)</td>
<td>0.669(0.413)</td>
<td>1.008</td>
</tr>
<tr>
<td>M6</td>
<td>1</td>
<td>0.719(0.396)</td>
<td>2.316(0.128)</td>
<td>0.062(0.804)</td>
<td>0.711(0.399)</td>
<td>1.002</td>
</tr>
</tbody>
</table>

Note: p-values for two-sided alternatives are given in the parenthesis.

d = 1, using dataset presented in Section 1.3. Most of previous literature focused on testing pure unit root against stationarity of yields (i.e. $d = 1$ against $d = 0$), however this sharp distinction does not take into consideration possible fractional behavior of time series. We will embed the issue in a fractional unit root framework, in which only the special case $d = 1$ conveys the classical unit root concept, and test the null hypothesis that interest rate series are integrated of order $d = 1$ against a two-sided alternative. We use the above-described Wald and LM tests with lag lengths selected with the Akaike information criteria and the deterministic rule $p_4 = \lfloor 4(T/100)^{0.25} \rfloor = 4$. Estimated values of integration orders and results of Wald and LM tests are reported in the Table 2.4.

As we see from Table 2.4, CSS estimates of integration orders are very close to the hypothesized value $d = 1$ for both lag selections and both LM and Wald tests do not reject the null for all maturities at 5% level (as well as at a more liberal 10% level). Hence, we conclude that the null hypothesis that interest rates are $I(1)$ processes cannot be rejected.

2.6 Conclusions

In this chapter we have studied estimation and inference of fractionally integrated time series via an autoregressive approximation. The estimator of the fractional parameter is defined by minimizing an objective function of the conditional sum of squared residuals (hence: CSS estimator) and is shown to be asymptotically normal provided the number of lags used in estimation grows at an admissible rate. Regression-based LM and Wald tests for the fractional parameter are also studied, extending results of Breitung and Hassler (2002) and Lobato and Velasco (2007) for infinite order autoregressions.

Monte Carlo simulation results show good properties of the CSS estimator in comparison with the exact local Whittle estimator, while the performance of tests gives a mixed picture and in certain cases seems to reflect a choice of power versus size. A few guidelines are given for the choice of number of lags in practical situations. An empirical application concerning testing the order of integration of U.S. interest rates is also given. We conclude
that the null hypothesis that interest rate series are integrated with order $d = 1$ cannot be rejected.
Estimation and inference of fractionally integrated time series

Figure 2.1: \( \phi = -0.5, T = 128 \)
2.6 Conclusions

(a) Wald test, size unadjusted

(b) Wald test, size adjusted

(c) LM test, size unadjusted

(d) LM test, size adjusted

Figure 2.2: $\phi = -0.5$, $T = 256$
26 Estimation and inference of fractionally integrated time series

Figure 2.3: $\phi = 0, T = 128$

(a) Wald test, size unadjusted

(b) Wald test, size adjusted

(c) LM test, size unadjusted

(d) LM test, size adjusted
2.6 Conclusions

(a) Wald test, size unadjusted

(b) Wald test, size adjusted

(c) LM test, size unadjusted

(d) LM test, size adjusted

Figure 2.4: $\phi = 0$, $T = 256$
Figure 2.5. $\phi = 0.5$, $T = 128$.
2.6 Conclusions

(a) Wald test, size unadjusted

(b) Wald test, size adjusted

(c) LM test, size unadjusted

(d) LM test, size adjusted

Figure 2.6: $\phi = 0.5, T = 256$
2.7 Appendix

2.7.1 Preliminaries and notation

We introduce the notation used in the Appendix: \( \xrightarrow{d} \) denotes convergence in distribution and \( \xrightarrow{P} \) denotes convergence in probability. The Euclidian norm of a matrix, vector or scalar \( A \) is denoted as \( \| A \| = \sqrt{\text{tr}(A'A)} \). In addition, we use spectral norm \( \| A \|_1 = \sqrt{\lambda_{\text{max}}(A'A)} \). We will say that an element \( A \) in Euclidian space is \( O_p(1) \), if \( \| A \| = O_p(1) \). The \( i \)-th order derivative of a function \( f \) with respect to \( x \) is denoted as \( D^i_x f \). \( 1_E \) denotes the indicator function for some logical operator \( E \). A positive number arbitrarily close to zero will be denoted as \( \epsilon \), whereas a generic positive constant bounded from below and bounded away from zero as \( C \). The \( ij \)-th element of a matrix \( A \) will be denoted as \( (A)_{ij} \) and the \( i \)-th element of a vector \( a \) as \( (a)_i \).

Suppose we have a lag polynomial \( A(L) = \sum_{i=0}^{\infty} A_i L^i \), then we denote:
\[
A(L) = \sum_{i=0}^{\infty} A_i L^i = \sum_{i=0}^{k} A_i L^i + \sum_{i=k+1}^{\infty} A_i L^i = A^k_+(L) + A^k_-(L).
\]
In addition, we denote the following lag polynomials for a generic time series process \( u_t \):
\[
\ln \Delta^d u_t = \sum_{j=1}^{\infty} j^{-1} u_{t-j} \quad \text{and} \quad \ln \Delta^d+ u_t = \sum_{j=1}^{\infty} j^{-1} u_{t-j}.
\]

The objective function for the model is defined in (2.4). Note, that \( l_T(\theta) \) and \( \hat{\sigma}_T^2(\theta) \) are related by a strictly monotonic transformation and hence:
\[
\arg\min_{\theta \in \Theta_T} \hat{\sigma}_T^2(\theta)/2 = \arg\max_{\theta \in \Theta_T} l_T(\theta).
\]
In the following, due to technical reasons\(^5\) we analyze \( \bar{l}_T(\theta) = \hat{\sigma}_T^2(\theta)/2 \):
\[
\bar{l}_T(\theta) = \hat{\sigma}_T^2(\theta)/2 = \frac{1}{2T} \sum_{t=k}^{T} \epsilon_{t,k}^2(\theta),
\]
\[
\epsilon_{t,k}(\theta) = \phi(L)^{[k]}_+ \Delta^d X_t = \Delta^d_{+} \phi(L)^{[k]}_+ \varphi(L) \epsilon_t.
\]

For \( \theta \in \Theta_T \) we define the gradient vector of \( \bar{l}_T(\theta) \) w.r.t. the components of the vector \( \theta \):
\[
\nabla_T(\theta) = \left( \frac{\partial}{\partial \theta_1} \bar{l}_T(\theta), \ldots, \frac{\partial}{\partial \theta_{k+1}} \bar{l}_T(\theta) \right).
\]
where \( \theta_1 = d, \theta_{i+1} = \phi_i, i > 1 \). Similarly we define the Hessian matrix of second derivatives:
\[
\mathcal{H}_T(\theta) = \left( \frac{\partial^2}{\partial \theta_i \partial \theta_j} \bar{l}_T(\theta) \right)_{i,j=1,\ldots,k+1}.
\]

\(^5\)Likelihood derivatives of \( \hat{\sigma}_T^2(\theta) \) have one term less.
Then:

$$\nabla_T(\theta) = \frac{1}{T} \sum_{t=k}^{T} D_{\theta} \varepsilon_{t,k}(\theta) \varepsilon_{t,k}(\theta),$$

$$\mathcal{H}_T(\theta) = \frac{1}{T} \sum_{t=k}^{T} (D_{\theta} \varepsilon_{t,k}(\theta) D_{\theta} \varepsilon'_{t,k}(\theta) + \varepsilon_{t,k}(\theta) \otimes D_{\theta}^2 \varepsilon_{t,k}(\theta))$$

where:

$$D_{\theta} \varepsilon_{t,k}(\theta) = (\ln \Delta_+ \varepsilon_{t,k}(\theta), \Delta_+^d X_{t-1}, \ldots, \Delta_+^d X_{t-k})',$n

$$(D_{\theta}^2 \varepsilon_{t,k}(\theta))_{ij} = \begin{cases} 
\ln \Delta_+ \Delta_+^d X_{t-\min\{i,j\}}, & \text{if } i > j = 1 \text{ or } j > i = 1 \\
\ln^2 \Delta_+ \varepsilon_{t,k}(\theta), & \text{if } i = j = 1 \\
0, & \text{otherwise.}
\end{cases}$$

We also define the following quantities:

$$\varepsilon_t(\theta) = \Delta^{d-d_0} \phi(L) T \varepsilon_t,$$

$$\dot{\varepsilon}_t(\theta) = (\ln \Delta \varepsilon_t, \Delta^{d-d_0} \phi(L) \varepsilon_{t-1}, \ldots, \Delta^{d-d_0} \phi(L) \varepsilon_{t-k})',$$

$$\dot{\mathcal{H}}_T(\theta) = \frac{1}{T} \sum_{t=k}^{T} \dot{\varepsilon}_t(\theta) \dot{\varepsilon}'_t(\theta).$$

Finally, we identify the probabilistic limit of $\dot{\mathcal{H}}_T(\theta)$ and $\mathcal{H}_T(\theta)$ at the point $\theta = \theta_0$: $\mathcal{H}_\infty = \mathcal{H}_\infty(\theta_0) = \text{Cov}(\dot{\varepsilon}_t(\theta_0)).$

### 2.7.2 Additional lemmas

Although not mentioned explicitly, all lemmas in the Appendix are proved under the assumptions of Theorem 2.1.

**Lemma 2.1.** Consider

$$\mathcal{H}_\infty = \text{Cov}(\dot{\varepsilon}_t(\theta_0)) = \begin{pmatrix} \omega & \zeta' \\ \zeta & \Omega \end{pmatrix}.$$  \hspace{1cm} (2.21)

\footnote{In an element-wise sense.}
Then for the sequence of \((k + 1) \times 1\) vectors \(\tau_1 = (1, 0, \ldots, 0) \in \mathbb{R}^{k+1}\), it holds:

\[
(\tau'_1 H^{-1}_\infty \tau_1)^{-1} = \sigma_0^2 k^{-1} + o(k^{-1}), \tag{2.22}
\]

\[
(\Omega^{-1} \zeta)_i = \sum_{j=1}^{i} j^{-1} \phi_{i-j} + o(1). \tag{2.23}
\]

**Proof.** We proceed similarly as in [Bhansali et al. (2006)]. Denote: \(\zeta' = (\zeta_1, \ldots, \zeta_k)\). We have: \(\Delta^d_+ X_t = u_t = \varphi(L) \varepsilon_t\) and \(\phi(L) u_t = \varepsilon_t\), thus the following relationships hold:

\[
\omega = \text{Var}(\ln \Delta \varepsilon_t) = \sigma_0^2 \sum_{i=1}^{\infty} i^{-2},
\]

\[
\zeta_i = \text{Cov}(\ln \Delta \varepsilon_t, u_{t-i}) = \sigma_0^2 \sum_{j=0}^{\infty} (i + j)^{-1} \varphi_j < Ci^{-1},
\]

\[
\sigma_0^2 i^{-1} = \text{Cov}(\ln \Delta \varepsilon_t, \varepsilon_{t-i}) = \text{Cov}(\ln \Delta \varepsilon_t, \phi(L) u_{t-i}) = \sigma_0^2 \sum_{j=0}^{\infty} \phi_j \zeta_{i+j}.
\]

As is known in prediction theory (cf. Kailath et al. (1978)): \(\Omega^{-1} = \sigma_k^{-2}(A' A - B' B)\), where

\[
A = \begin{pmatrix}
    a_0 & a_{1,k} & a_{2,k} & \ldots & a_{k-1,k} \\
    0 & a_0 & a_{1,k} & \ldots & a_{k-2,k} \\
    \ldots & \ldots & \ldots & \ldots & \ldots \\
    0 & \ldots & \ldots & a_0 & a_{1,k} \\
    0 & 0 & \ldots & 0 & a_0
\end{pmatrix},
\]

\[
B = \begin{pmatrix}
    a_{k,k} & a_{k-1,k} & a_{k-2,k} & \ldots & a_{1,k} \\
    0 & a_{k,k} & a_{k-1,k} & \ldots & a_{2,k} \\
    \ldots & \ldots & \ldots & \ldots & \ldots \\
    0 & \ldots & \ldots & a_{k,k} & a_{k-1,k} \\
    0 & 0 & \ldots & 0 & a_{k,k}
\end{pmatrix}
\]

with:

\[
(a_{1,i}, \ldots, a_{i,i}) = \arg \min_{(b_1, \ldots, b_i) \in \mathbb{R}^i} E \left( u_t + \sum_{j=1}^{i} b_j u_{t-j} \right)^2, \quad \sigma_i^2 = E \left( u_t + \sum_{j=1}^{i} a_{j,i} u_{t-j} \right)^2
\]

and \(a_0 = 1\). Denote

\[
C = \begin{pmatrix}
    \phi_0 & \phi_1 & \phi_2 & \ldots & \phi_k \\
    0 & \phi_0 & \phi_1 & \ldots & \phi_{k-1} \\
    \ldots & \ldots & \ldots & \ldots & \ldots \\
    0 & 0 & \ldots & \phi_0 & \phi_1 \\
    0 & 0 & 0 & \ldots & \phi_0
\end{pmatrix}.
\]

Then we apply the relationship between \(\zeta\) and \(C\): \(C\zeta = (1, 2^{-1}, \ldots, k^{-1})\) and Baxter’s
inequality:

\[ \sum_{i=0}^{j} |a_{i,j} - \phi_i| = O \left( \sum_{s>j} |\phi_s| \right) \]

to get the bounds:

\[ ||A - C||_1 \leq \sum_{i=1}^{k} |a_{i,k} - \phi_i| = O \left( \sum_{s>k} |\phi_s| \right) = O(k^{-2}), \]

\[ ||\zeta||^2 = \sigma_0^2 \sum_{i=1}^{k} (\sum_{j=0}^{\infty} (i+j)^{-1} \varphi_j)^2 \leq \sigma_0^2 \sum_{i=1}^{k} i^{-2} \left( \sum_{j=0}^{\infty} |\varphi_j| \right)^2 < C, \]

\[ ||C\zeta||^2 = \sum_{i=1}^{k} i^{-2} < C, \]

\[ ||B\zeta||^2 = \sum_{i=1}^{k} (\zeta_i a_{k-i,k})^2 \leq C \sum_{i=1}^{[k/2]} a_{k-i,k}^2 + C \sum_{i=[k/2]}^{k} a_{k-i,k}^2 i^{-2} = O(k^{-4}) + O(k^{-2}) = O(k^{-2}), \]

\[ |\sigma_k^2 - \sigma_0^2| \leq \sigma_0^2 \sum_{j=1}^{k} |a_{j,k} - \phi_j|^2 + \sigma_0^2 \sum_{j=k}^{\infty} \phi_j^2 \leq O(k^{-4}) + C (\sum_{j=k}^{\infty} |\phi_j|)^2 = O(k^{-4}). \]

The above bounds and inequality \( ||AB|| \leq ||A|| ||B|| \) give:

\[ ||\zeta'(A - C)'\sigma_k^{-2} C\zeta|| \leq \sigma_k^{-2} ||\zeta|| ||C\zeta|| ||A - C||_1 = O(k^{-2}), \]

\[ ||\zeta'(A - C)'(\sigma_k^{-2} - \sigma_0^{-2}) C\zeta|| = (\sigma_k^{-2} - \sigma_0^{-2}) \sum_{i=1}^{k} i^{-2} \leq C(\sigma_k^{-2} - \sigma_0^{-2}) \sum_{i=1}^{k} i^{-2} = O(k^{-4}), \]

\[ ||\zeta'(B\sigma_k^{-2}) B\zeta|| \leq \sigma_k^{-2} ||B\zeta||^2 = O(k^{-2}). \]

We use the above bounds to approximate \( A'B - B'B \) by \( C'C \):

\[ \zeta'\Omega^{-1}\zeta = \zeta'C'\sigma_0^{-2}C\zeta + \zeta'C'(\sigma_k^{-2} - \sigma_0^{-2})C\zeta + \zeta'(A - C)'\sigma_k^{-2} A\zeta + \zeta'C'\sigma_k^{-2}(A - C)\zeta \]

\[ - \zeta'B\sigma_k^{-2} B\zeta = \zeta'\Omega^{-1}\zeta = (C\zeta)'\sigma_0^{-2}C\zeta + O(k^{-4}) + O(k^{-2}) = \sigma_0^{-2} \sum_{i=1}^{k} i^{-2} + o(k^{-1}). \]
Finally, for a sequence $\tau_1 = (1, 0, \ldots, 0)' \in \mathbb{R}^{k+1}$, we have:

\[
(\tau_1' \mathcal{H}_\infty^{-1} \tau_1)^{-1} = \omega - \zeta' \Omega^{-1} \zeta = \sigma_0^2 \sum_{i=1}^{\infty} i^{-2} - \zeta' C' \sigma_0^{-2} C \zeta + o(k^{-1}) = \sigma_0^2 \sum_{i=k}^{\infty} i^{-2} + o(k^{-1})
\]

\[
= \sigma_0^2 k^{-1} + o(k^{-1}).
\]

In addition, \[||(\Omega^{-1} - C' C \sigma_0^{-2}) \zeta|| = o(1)\] implies:

\[
(\Omega^{-1} \zeta)_i = (C'(1, 2^{-1}, 3^{-1}, \ldots, k^{-1}'))_i + o(1) = \sum_{j=1}^{i} j^{-1} \phi_{i-j} + o(1).
\]

\[\square\]

**Lemma 2.2.** Define: $e_{1t} = \varepsilon_{t,k}(\theta_0) - \varepsilon_t(\theta_0)$, $e_{2t} = D_\theta \varepsilon_{t,k}(\theta_0) - \dot{\varepsilon}_t(\theta_0)$. Then uniformly in $t$: $E||e_{1t}||^2 = o(T^{-1})$, $E||e_{2t}||^2 = O(t^{-1})$.

**Proof.** We have:

\[
e_{1t} = \varepsilon_{t,k}(\theta_0) - \varepsilon_t(\theta_0) = (\phi_0(L)_+^{[k]} \varphi(L) - 1) \varepsilon_t = -\phi_0(L)_+^{[k]} \varphi(L) \varepsilon_t = - \sum_{i=k}^{\infty} \psi_{i,k} \varepsilon_{t-i},
\]

where $\psi_{i,k} = \sum_{j=k}^{i} \phi_{0j} \varphi_{i-j}$. Hence:

\[
Ee_{1t}^2 = \sigma_0^2 \sum_{i=k}^{\infty} \psi_{i,k}^2 \leq \sigma_0^2 \left( \sum_{i=k}^{\infty} \sum_{j=k}^{i} |\phi_{0j}| |\varphi_{i-j}| \right)^2 \leq \sigma_0^2 \left( \sum_{j=k}^{\infty} |\phi_{0j}| \right)^2 \left( \sum_{j=0}^{\infty} |\varphi_j| \right)^2 = o(T^{-1}).
\]

The last bound follows from Assumption 2.2 and absolute summability of $\varphi_j$. Now consider the first component of $e_{2t}$:

\[
e_{21t} = \ln \Delta_+ \phi_0(L)_+^{[k]} \varphi(L) \varepsilon_t - \ln \Delta \varepsilon_t = \ln \Delta_+ \varepsilon_t - \ln \Delta_+ \phi_0(L)_-^{[k]} \varphi(L) \varepsilon_t = \ln \Delta_+ \varepsilon_t + \ln \Delta_+ e_{1t}.
\]
2.7 Appendix

From here:

\[ E|\varepsilon_{2t}||^2 = E|\varepsilon_{21t}||^2 = E|\ln \Delta_\varepsilon t||^2 + E|\ln \Delta_\varepsilon 41t||^2 \leq \sigma_0^2 \sum_{t=1}^\infty t^{-2} \]

\[ + \sum_{h=0}^t \sum_{j=1}^{t-h} E\varepsilon_{1t-j}e_{1t-j-h}j^{-1}(j + h)^{-1} \leq Ct^{-1} + \sum_{h=0}^t \sum_{j=1}^{t-h} E\varepsilon_{1t-j}e_{1t-j-h}j^{-2} \]

\[ \leq Ct^{-1} + C \sum_{h=0}^t \sum_{j=1}^{t-h} \left( \sum_{i=k}^\infty |\psi_{i,k}\psi_{i+h,k}| \right) j^{-2} \leq Ct^{-1} + C \sum_{h=0}^t \left( \sum_{i=k}^\infty |\psi_{i,k}\psi_{i+h,k}| \right) \]

\[ \leq Ct^{-1} + C \sum_{i=k}^\infty \psi_{i,k} \sum_{h=0}^t |\psi_{i+h,k}| \leq Ct^{-1} + C \left( \sum_{j=k}^\infty |\psi_{j,k}| \right)^2 = O(t^{-1}) + O(T^{-1}). \]

\[ \square \]

Lemma 2.3. \[ ||\mathcal{H}_T(\theta_0) - \mathcal{H}_T(\theta_0)|| = o_p(k^{-1}). \]

Proof. We have:

\[ k||\mathcal{H}_T(\theta_0) - \mathcal{H}_T(\theta_0)|| \leq \sum_{t=k}^T ||D_\theta \varepsilon_{t,k}(\theta_0)D_\theta \varepsilon_{t,k}(\theta_0)' - \dot{\varepsilon}_t(\theta_0)\dot{\varepsilon}_t(\theta_0)'|| + ||k T^{-1} \sum_{t=k}^T \varepsilon_{t,k}(\theta_0) \otimes D_\theta^2 \varepsilon_{t,k}(\theta_0)||. \]

Consider the first term:

\[ \left| \left| D_\theta \varepsilon_{t,k}(\theta_0)D_\theta \varepsilon_{t,k}(\theta_0)' - \dot{\varepsilon}_t(\theta_0)\dot{\varepsilon}_t(\theta_0)' \right| \right| \leq \left| \left| (D_\theta \varepsilon_{t,k}(\theta_0) - \dot{\varepsilon}_t(\theta_0))\dot{\varepsilon}_t(\theta_0) \right| \right| \]

\[ + \left| \left| D_\theta \varepsilon_{t,k}(\theta_0) (D_\theta \varepsilon_{t,k}(\theta_0) - \dot{\varepsilon}_t(\theta_0))' \right| \right| \leq 2||\varepsilon_{2t}|| \left( ||D_\theta \varepsilon_{t,k}(\theta_0)|| + ||\dot{\varepsilon}_t(\theta_0)|| \right) \leq 2||\varepsilon_{2t}||^2 \]

\[ + 4||\varepsilon_{2t}|| ||\dot{\varepsilon}_t(\theta_0)||. \]

Since the first term is \( O_p(t^{-1/2}) \) uniformly in \( t \), we have:

\[ \sum_{t=k}^T E||D_\theta \varepsilon_{t,k}(\theta_0)D_\theta \varepsilon_{t,k}(\theta_0)' - \dot{\varepsilon}_t(\theta_0)\dot{\varepsilon}_t(\theta_0)'|| \leq C E||\dot{\varepsilon}_t(\theta_0)|| \left| \left| \sum_{t=k}^T t^{-1/2} = CKT^{-1/2} = o(1) \right. \right. \]
Now consider the second term:

\[
\left| \left| \frac{1}{T} \sum_{t=k}^{T} \varepsilon_{t,k}(\theta_0) \otimes D_{\theta}^2 \varepsilon_{t,k}(\theta_0) \right| \right| = \left| \left| \frac{1}{T} \sum_{t=k}^{T} \varepsilon_{1t} \otimes D_{\theta}^2 \varepsilon_{t,k}(\theta_0) + \varepsilon_{t} \otimes D_{\theta}^2 \varepsilon_{t,k}(\theta_0) \right| \right| 
\]
\[
\leq \frac{4k}{T} \sum_{t=k}^{T} \varepsilon_{t} \ln \Delta + \sum_{i=1}^{k} \varepsilon_{t} \ln \Delta + \varepsilon_{t} \ln \Delta + \varepsilon_{t} \ln \Delta 
\]
\[
+ \left| \left| \frac{1}{T} \sum_{t=k}^{T} \varepsilon_{t} \ln \Delta \right| \right|. 
\]

Applying the Cauchy-Schwarz inequality for the first two terms we get that they are of orders \(o_p(k^2/\sqrt{T})\) and \(o_p(k/\sqrt{T})\), whereas the bound for the second moment of the third and fourth terms is:

\[
\frac{k^2}{T^2} \sum_{t=k}^{T} \sum_{i=1}^{k} E(\varepsilon_t \ln \Delta) + \frac{k^2}{T^2} \sum_{t=k}^{T} E(\varepsilon_t \ln \Delta) = O(k^3/T) + O(k^2/T) = o(1), 
\]

and the lemma is proved.

**Lemma 2.4.** For \(\theta_T \xrightarrow{P} \theta_0\), it holds: \(\left| \left| \mathcal{H}_T(\theta_T) - \mathcal{H}_T(\theta_0) \right| \right| = \left| \left| \theta_T - \theta_0 \right| \right| O_p(k)\).

**Proof.** Applying mean value theorem for \(\mathcal{H}_T(\theta)\) we have:

\[
\left| \left| \mathcal{H}_T(\theta_T) - \mathcal{H}_T(\theta_0) \right| \right|^2 = \sum_{i,j=1}^{k+1} \left| \left| (\theta_T - \theta_0)'D_{\theta}(\mathcal{H}_T(\theta_T))_{ij} \right| \right|^2 \leq \left| \left| \theta_T - \theta_0 \right| \right|^2 \sum_{i,j=1}^{k+1} \left| \left| \frac{\partial^3}{\partial \theta_i \partial \theta_j \partial \theta_s} \tilde{l}_T(\tilde{\theta}) \right| \right|^2, 
\]

for some \(\left| \left| \tilde{\theta} - \theta_0 \right| \right| < \left| \left| \theta_T - \theta_0 \right| \right|\). \(\frac{\partial^3}{\partial \theta_i \partial \theta_j \partial \theta_s} \tilde{l}_T(\theta)\) is non-zero only if at least one of the indices \(i, j, s\) is equal to 1, hence there are \(O(k^2)\) non-zero terms in the sum. Given the structure of partial derivatives of \(\tilde{l}_T(\theta)\), it is enough to show uniform boundedness of sample moments in some local neighbourhood of \(\theta_0\) of the following terms:

\[
\frac{\partial^2}{\partial \theta_i \partial \theta_j} \varepsilon_{t,k}(\theta) \frac{\partial}{\partial \theta_s} \varepsilon_{t,k}(\theta), \quad \frac{\partial^3}{\partial \theta_i \partial \theta_j \partial \theta_s} \varepsilon_{t,k}(\theta) \varepsilon_{t,k}(\theta). \quad (2.24)
\]

Denote a partial derivative of order \(i \leq 3\) of \(\Delta_X^4 X_t\) w.r.t. \(d\): \(u_t^i(d) = \frac{\partial^i}{\partial d^i} \Delta_X^4 X_t\). We may apply Theorem 2.1 (a2) in Findley et al. (2001) and show that sample moments of \(u_t^i(d)\).
converge uniformly in \( \Theta_T \) to their deterministic limits, if the following hold:

\[
\text{Var}(u^*_t(d)) < C,
\]
\[
\left| \frac{1}{T} \sum_{t=k}^T u^*_t(d)u^*_r(d) \right| = O_p(1),
\]
\[
u^*_t(d) = o_p(1),
\]

uniformly in \( d \in D \) for \( s, r = 0, \ldots, 3 \). These uniform bounds follow from Lemmas A.4 and C.4 in Johansen and Nielsen (2010). The limits itself are uniformly bounded in \( \Theta_T \), as follows again from Lemma C.4 in Johansen and Nielsen (2010) and hence we have:

\[
||H_T(\theta_T) - H_T(\theta_0)||^2 = ||\theta_T - \theta_0||^2 O_p(k^2).
\]

\[\square\]

**Lemma 2.5.** For any sequence of \((k+1) \times 1\) vectors: \( \tau \in \mathbb{R}^{k+1} \), the following convergence holds:

\[
\sqrt{T} \nabla_T(\theta_0) = \frac{1}{\sqrt{T}} \sum_{t=k}^T D_{\theta} \varepsilon_{t,k}(\theta_0) \varepsilon_{t,k}(\theta_0) = \frac{1}{\sqrt{T}} \sum_{t=k}^T (\dot{\varepsilon}_t(\theta_0) \varepsilon_t + \dot{\varepsilon}_t(\theta_0) \varepsilon_{1t} + e_{2t} \varepsilon_t + e_{2t} \varepsilon_{1t}).
\]

We prove that the last three terms are of order \( o_p(1) \). Consider the first term:

\[
|| \frac{1}{\sqrt{T}} \sum_{t=k}^T \dot{\varepsilon}_t(\theta_0) \varepsilon_{1t} ||^2 = || \frac{1}{\sqrt{T}} \sum_{t=k}^T \ln \Delta \varepsilon_{1t} ||^2 + \sum_{i=1}^k || \frac{1}{\sqrt{T}} \sum_{t=k}^T u_{t-i} \varepsilon_{1t} ||^2.
\]

(2.26)

For the first summand it holds:

\[
E|| \frac{1}{\sqrt{T}} \sum_{t=k}^T \ln \Delta \varepsilon_{1t} || \leq \frac{1}{\sqrt{T}} \sum_{t=k}^T E|| \ln \Delta \varepsilon_{1t} || \leq \frac{1}{\sqrt{T}} \sum_{t=k}^T (E|| \ln \Delta \varepsilon_{1t} ||^2 E\varepsilon_{1t}^2)^{1/2} \]
\[
\leq \frac{C}{\sqrt{T}} \sum_{t=k}^T (E\varepsilon_{1t}^2)^{1/2} = o(1),
\]
where the bound for $E\varepsilon_t^2$ from Lemma 2.2 was used. For the second summand:

$$E \frac{1}{\sqrt{T}} \sum_{t=k}^{T} u_{t-i} \varepsilon_{1t}^2 = \frac{1}{T} \sum_{i=1}^{k} \sum_{t_1, t_2 = k}^{T} \sum_{j_1, j_2 = k}^{T} \phi_{j_1, j_2} E u_{t_1-i} u_{t_2-i} u_{t_1-j_1} u_{t_2-j_2}$$

$$\leq \sum_{i=1}^{k} C \frac{T}{T} \sum_{t_1, t_2 = k}^{T} \sum_{j_1, j_2 = k}^{T} \phi_{j_1, j_2} \leq kC \left( \sum_{j > k}^{\infty} |\phi_j| \right)^2 = o(1),$$

since cumulants of fourth-order stationary linear processes are summable and thus $\sum_{i_1, i_2, i_3} E u_{i_1} u_{i_2} u_{i_3} u_{i_3} < C$. Hence the whole term in (2.26) is $o_p(1)$. Applying the Cauchy-Schwarz inequality and Lemma 2.2, we find:

$$E \left| \frac{1}{\sqrt{T}} \sum_{t=k}^{T} e_{2t} \varepsilon_{1t} \right| = o(1). \quad (2.27)$$

Finally, uncorrelatedness of $\varepsilon_t$ and $e_{2t-i}$ for $i > 0$ gives:

$$E \left| \frac{1}{\sqrt{T}} \sum_{t=1}^{T} e_{2t} \varepsilon_t \right|^2 = E \left( \frac{1}{\sqrt{T}} \sum_{t=1}^{T} e_{2t} \varepsilon_t \right)^2 = \frac{1}{T} \sum_{t=1}^{T} E e_{2t}^2 \varepsilon_t^2 = \frac{\sigma_0^2}{T} \sum_{t=1}^{T} E e_{2t}^2 \leq \frac{C}{T} \sum_{t=1}^{T} t^{-1} = o(1).$$

Hence we have:

$$\sqrt{\frac{T}{\sigma_0^2 \tau' \mathcal{H}_\infty^1 \tau}} \mathcal{H}_\infty^1 H_T(\theta_0) = \frac{1}{\sqrt{T}} \sum_{t=1}^{T} V_t \varepsilon_t + o_p(1), \quad (2.28)$$

where $V_t = \sigma_V^{-1} \tau' \mathcal{H}_\infty^1 \varepsilon_t(\theta_0)$, with $\sigma_V^2 = \sigma_0^2 \tau' \mathcal{H}_\infty^1 \tau$. We prove that (2.28) converges to a standard normal random variable, which will prove the lemma. We apply the central limit theorem for martingale difference series to $V_t \varepsilon_t$ (Corrolary 3.1, Hall and Heyde (1980)), proving the following:

$$T^{-1} \sum_{t=k}^{T} E(V_t^2 \varepsilon_t^2 | \mathcal{F}_{t-1}) \overset{P}{\to} 1, \quad (2.29)$$

$$T^{-1} \sum_{t=k}^{T} E \left( V_t^2 \varepsilon_t^2 1 \{ |V_t \varepsilon_t| > \eta^2 \sqrt{T} \} | \mathcal{F}_{t-1} \right) \overset{P}{\to} 0, \forall \eta > 0, \quad (2.30)$$

where $\mathcal{F}_t$ is a $\sigma$-field generated by $\{ \varepsilon_i, i \leq t \}$. Note, that $V_t$ is adapted to $\mathcal{F}_{t-1}$ and thus the first equality follows from $E(V_t^2 \varepsilon_t^2 | \mathcal{F}_{t-1}) = V_t^2 E(\varepsilon_t^2 | \mathcal{F}_{t-1}) = \sigma_0^2 V_t^2$ and $E\sigma_0^2 V_t^2 = ...
We show that $E^2.7$ Appendix 39

$\sigma_\varepsilon^{-2}\sigma_0^2\tau'\mathcal{H}_\infty^{-1}Cov(\dot{\varepsilon}_t)\mathcal{H}_\infty^{-1}\tau = 1$. For the second equation, we have:

$$T^{-1} \sum_{t=k}^{T} E[V_t^2\varepsilon_t^21(|V_t\varepsilon_t| > \eta^2\sqrt{T})|\mathcal{F}_{t-1}] \leq T^{-1} \sum_{t=k}^{T} V_t^2E[\varepsilon_t^21(|V_t\varepsilon_t| > \eta^2\sqrt{T})|\mathcal{F}_{t-1}]$$

$$\leq T^{-1} \sum_{t=k}^{T} V_t^2E[\varepsilon_t^2(1(|V_t| > \eta T^{1/4}) + 1(|\varepsilon_t| > \eta T^{1/4}))|\mathcal{F}_{t-1}]$$

$$\leq T^{-1} \sum_{t=k}^{T} V_t^2E[\varepsilon_t^21(|\varepsilon_t| > \eta T^{1/4})] + T^{-1} \sum_{t=k}^{T} V_t^2E[\varepsilon_t^21(|V_t| > \eta T^{1/4})|\mathcal{F}_{t-1}] = A_1 + A_2.$$  

We show that $E[A_1] = o(1)$, $E[A_2] = o(1)$, which will prove $A_1 + A_2 = o_P(1)$. Note:

$$E[A_1] = (T - k)T^{-1}EV_t^2E(\varepsilon_t^21(|\varepsilon_t| > \eta T^{1/4})) = o(1),$$

due to stationarity of $\varepsilon_t$. Applying the Chebyshev inequality gives:

$$E[A_2] = \sigma_0^2T^{-1} \sum_{t=k}^{T} EV_t^2E(1(|V_t| > \eta T^{1/4})) = T^{-1} \sum_{t=k}^{T} Pr(|V_t| > \sqrt{\eta T^{1/4}}) \leq T^{-1} \sum_{t=k}^{T} \frac{EV_t^2}{\eta^2\sqrt{T}} = o(1).$$

Hence $E[A_2] = o(1)$ and the convergence follows.

\[\square\]

### 2.7.3 Proofs of main theorems

**Proof of Theorem 2.1.** A Taylor expansion for the gradient $\nabla_T(\theta)$ gives:

$$0 = \nabla_T(\hat{\theta}) = \nabla_T(\theta_0) + \mathcal{H}_T(\hat{\theta})(\hat{\theta} - \theta_0) = \nabla_T(\theta_0) + \mathcal{H}_\infty(\theta_0)(\hat{\theta} - \theta_0) \quad (2.31)$$

$$+ (\mathcal{H}_T(\hat{\theta}) - \mathcal{H}_\infty(\theta_0))(\hat{\theta} - \theta_0), \quad (2.32)$$

for some $||\hat{\theta} - \theta_0|| \leq ||\theta - \theta_0||$. From here:

$$\hat{\theta} - \theta_0 = -\mathcal{H}_\infty^{-1}(\theta_0)\nabla_T(\theta_0) - \mathcal{H}_\infty^{-1}(\theta_0) (\mathcal{H}_T(\hat{\theta}) - \mathcal{H}_\infty(\theta_0)) (\hat{\theta} - \theta_0). \quad (2.33)$$

Note:

$$||\mathcal{H}_T(\hat{\theta}) - \mathcal{H}_\infty(\theta_0)|| \leq ||\mathcal{H}_T(\hat{\theta}) - \mathcal{H}_T(\theta_0)|| + ||\mathcal{H}_T(\theta_0) - \mathcal{H}_T(\theta_0)|| + ||\mathcal{H}_T(\theta_0) - \mathcal{H}_\infty(\theta_0)||.$$
By Lemma 2.4, the first term is of order $||\bar{\theta} - \theta_0||O_p(k)$, the second term is $o_p(k^{-1})$ by Lemma 2.3 and the last term is of order $O_p(kT^{-1/2})$ as follows from Hannan (1974). Hence, provided $k = o(T^{-1/4})$:

$$||H_T(\bar{\theta}) - H_\infty(\theta_0)|| = \left(||\bar{\theta} - \theta_0||O_p(k) + o_p(k^{-1}) + O_p(kT^{-1/2})\right) = o_p(k^{-1}).$$

On the other hand, $||H_T^{-1}(\theta_0)||_1 = O(k)$ (cf. Lemma 4.1 in Bhansali et al. (2006)) implies:

$$||H_T^{-1}(\theta_0) (H_T(\bar{\theta}) - H_\infty(\theta_0)) (\hat{\theta} - \theta_0)|| \leq ||H_T^{-1}(\theta_0)||_1 ||H_T(\bar{\theta}) - H_\infty(\theta_0)|| ||\hat{\theta} - \theta_0|| = o_p(||\hat{\theta} - \theta_0||).$$

Hence we have:

$$\hat{\theta} - \theta_0 = -H_T^{-1}(\theta_0)\nabla T(\theta_0) + o_p(||\hat{\theta} - \theta_0||). \quad (2.34)$$

Using definition of $H_\infty(\theta_0)$ in Lemma 2.1 we proceed further as in Bhansali et al. (2006) and decompose the term $H_T^{-1}(\theta_0)\nabla T(\theta_0)$ as follows:

$$H_T^{-1}(\theta_0)\nabla T(\theta_0) = \begin{pmatrix} 1 \\ -\Omega^{-1}\zeta \end{pmatrix} V + \begin{pmatrix} 0 & 0 \\ 0 & \Omega^{-1} \end{pmatrix} \nabla T(\theta_0), \quad (2.35)$$

where $V$ is the first coordinate of the vector $H_T^{-1}(\theta_0)\nabla T(\theta_0)$. The decomposition is justified by the block inverse matrix formula. Then from Lemma 2.5 with $\tau_1 = (1, 0, \ldots, 0)' \in \mathbb{R}^{k+1}$ we have:

$$\sqrt{T} \frac{1}{\sigma_0^2 \tau_1' H_\infty^{-1} \tau_1} \tau_1' H_\infty^{-1} \nabla T(\theta_0) = \sqrt{T} \frac{1}{\sigma_0^2 \tau_1' H_\infty^{-1} \tau_1} V \overset{d}{\rightarrow} N(0, 1).$$

On the other hand, from Lemma 2.1 we have $(\tau_1' H_\infty^{-1} \tau_1)^{-1} = k^{-1} \sigma_0^2 + o(k^{-1})$ and hence:

$$\sqrt{T} \frac{1}{\sigma_0^2 \tau_1' H_\infty^{-1} \tau_1} V = \sqrt{\frac{T}{k}} V(1 + o(1)).$$

Thus: $\sqrt{\frac{T}{k}} V \overset{d}{\rightarrow} N(0, 1)$.

Now fix $m$ and define a selection matrix $S_m = (I_m, 0_{m \times k - m})$ for $k \geq m$. Then (2.34) and (2.35) gives:

$$\sqrt{\frac{T}{k}} (\hat{\theta}^p - \theta_0^p) = \sqrt{\frac{T}{k}} S_{p+1}(\hat{\theta} - \theta) = \Gamma \sqrt{\frac{T}{k}} V + \sqrt{\frac{T}{k}} S_p \Omega^{-1} \nabla T(\theta_0) + o_p(1),$$

where $\Gamma$ is the matrix of the block inverse.
where $\nabla_T^\dag(\theta_0) = (\frac{\partial}{\partial \rho_2} I_T, \ldots, \frac{\partial}{\partial \rho_{k+1}} I_T)'$. Since $\sqrt{T}S_p\Omega^{-1}\nabla_T^\dag(\theta_0) \xrightarrow{d} N(0, S_p\Omega^{-1}S_p')$ (cf. Berk (1974), Theorem 4), for a fixed $p$ it holds:

\[
\sqrt{T/k}(\hat{\theta}^p - \theta_0^p) \xrightarrow{d} \Gamma z,
\]

\[
\sqrt{T\Gamma'_\perp}(\hat{\theta}^p - \theta_0^p) \xrightarrow{d} N(0, \Gamma'_\perp S_p\Omega^{-1}S_p' \Gamma'_\perp),
\]

and the theorem is proved. \(\square\)

**Proof of Theorem 2.2.** Consider the regression under the local alternative framework $d_0 = d + \delta \sqrt{k/T}$:

\[
y_t = \rho z_t^3 + \sum_{j=1}^k \phi_j y_{t-j} + \eta_t, \quad (2.36)
\]

where $y_t = \Delta^d_+ X_t$. A Taylor expansion gives:

\[
\Delta^{d+\delta \sqrt{k/T}} = \Delta^d_+ + \ln \Delta^d_+ \Delta^{d+\delta \sqrt{k/T}} + \ln^2 \Delta^d_+ \Delta^{d+\delta \sqrt{k/T}} \delta^2 k/T, \quad (2.37)
\]

for some $|\tilde{\delta}| < |\delta|$. Hence:

\[
\epsilon_{t,k}(\theta_0) = \phi_0(L)[k]_+^\dag \Delta^d_+ X_t = \phi_0(L)[k]_+^\dag \Delta^{d+\delta \sqrt{k/T}} X_t = \phi_0(L)[k]_+^\dag \Delta^d_+ X_t + \delta \sqrt{k/T} \ln \Delta^d_+ \phi_0(L)[k]_+^\dag \Delta^d_+ X_t + e_{3t},
\]

where $e_{3t} = \delta^2 k/T \ln^2 \Delta^d_+ \Delta^{d+\delta \sqrt{k/T}} \epsilon_t$. Denote the regression coefficients as $\beta' = (\rho, \Phi) = (\rho, \phi_1, \ldots, \phi_k)$ and their corresponding true values as: $\beta_0' = (\rho_0, \Phi_0) = (\delta \sqrt{k/T}, \phi_{01}, \ldots, \phi_{0k})$. Then:

\[
\hat{\beta} - \beta_0 = \left( \sum_{t=k}^T V_t^* V_t'^* \right)^{-1} \sum_{t=k}^T V_t^* \eta_t,
\]

where the error term in the regression under local alternatives is $\eta_t = \epsilon_{t,k}(\theta_0) - e_{3t} - e_{4t}$, with $e_{4t} = \delta \sqrt{k/T} \left( \ln \Delta^d_+ \Delta^{d+\delta \sqrt{k/T}} \phi_0(L)[k]_+^\dag X_t - z_t^3 \right)$, $V_t^* = (z_t^3, y_{t-1}, \ldots, y_{t-k})'$ and $z_t^3$ is an auxiliary regressor in either LM ($i = 1$) or Wald ($i = 2$) test regression. Denote: $\tilde{H}_T = T^{-1} \sum_t V_t^* V_t'^*$, then we have:

\[
\beta_0 - \hat{\beta} = \tilde{H}_T^{-1}(\theta_0) \nabla_T(\theta_0) + (\tilde{H}_T^{-1} - \tilde{H}_T^{-1}(\theta_0)) \nabla_T(\theta_0) + \tilde{H}_T^{-1} \left( \frac{1}{T} \sum_{t=k}^T V_t^* \eta_t - \nabla_T(\theta_0) \right).
\]
We prove the following using additional condition on the rate $k^a/T = o(1)$:

\[
\sqrt{k} ||\tilde{H}^{-1}_T - \tilde{H}^{-1}_T(\theta_0)|| = o_p(1), \tag{2.38}
\]

\[
k\sqrt{T} ||\frac{1}{T} \sum_i V_i^* \eta_t - \nabla_T(\theta_0)|| = o_p(1). \tag{2.39}
\]

Note, that the bound:

\[
||V_t^* - D_\theta \varepsilon_{t,k}(\theta_0)|| = O_p(\sqrt{k/T}) \tag{2.40}
\]

implies $||\tilde{H}_T - \tilde{H}_T(\theta_0)|| = O_p(\sqrt{k/T})$ and hence (2.38). We prove (2.38):

\[
||V_t^* - D_\theta \varepsilon_{t,k}(\theta_0)||^2 = (z_t^i - \ln \Delta_+ \varepsilon_{t,k}(\theta_0))^2 + \sum_{i=1}^k (y_{t-i} - u_{t-i})^2. \tag{2.41}
\]

Then: $y_{t-i} - u_{t-i} = \delta \sqrt{k/T} \ln \Delta_+ \Delta_+^{\frac{T}{2}} \varepsilon_t = O_p(\sqrt{k/T})$. Since $z_t^i = \ln \Delta_+ \tilde{\phi}(L) \Delta_+^{\frac{T}{2}} X_t = \ln \Delta_+ \varepsilon_{t,k}(\tilde{\theta})$, where $||\tilde{\theta} - \theta_0|| = O_p(\sqrt{k/T})$ in both Wald and LM cases, we have:

\[
||z_t^i - \ln \Delta_+ \varepsilon_{t,k}(\theta_0)||^2 = ||\ln \Delta_+(\tilde{\theta} - \theta_0) D_\theta \varepsilon_{t,k}(\tilde{\theta})||^2 \leq ||\tilde{\theta} - \theta_0||^2 ||\ln \Delta_+ D_\theta \varepsilon_{t,k}(\tilde{\theta})||^2 = O_p(k^2/T).
\]

and (2.40) and (2.38) follows. For (2.39), we need to show:

\[
||\frac{k}{\sqrt{T}} \sum_{t=k}^T (V_t^* - D_\theta \varepsilon_{t,k}(\theta_0)) (\varepsilon_{t,k}(\theta_0) + \varepsilon_{3t} + \varepsilon_{4t})|| = o_p(1), \tag{2.42}
\]

\[
||\frac{k}{\sqrt{T}} \sum_{t=k}^T D_\theta \varepsilon_{t,k}(\theta_0) (\varepsilon_{3t} + \varepsilon_{4t})|| = o_p(1). \tag{2.43}
\]

We find the bound:

\[
||\varepsilon_{4t}|| \leq \delta \sqrt{k/T} ||\ln \Delta_+ \Delta_+^{d_0} \phi_0(L)_{+}^{[k]} X_t - z_t^i|| + \delta^2 k/T ||\ln^2 \Delta_+ \Delta_+^{d_0+k/2} \phi_0(L)_{+}^{[k]} X_t|| \tag{2.44}
\]

\[
= O_p(k^{3/2}T^{-1}). \tag{2.45}
\]

Then applying the norm inequality together with (2.40) and bounds $||\varepsilon_{t,k}(\theta_0)|| = O_p(1)$, $||\varepsilon_{1t}|| = O_p(T^{-1/2})$, $||\varepsilon_{3t}|| = O_p(kT^{-1})$, $||\varepsilon_{4t}|| = O_p(k^{3/2}T^{-1})$, $||D_\theta \varepsilon_{t,k}(\theta_0)|| = O_p(\sqrt{k})$, give (2.42) and (2.43), if we prove:

\[
||\frac{k}{\sqrt{T}} \sum_{t=k}^T (V_t^* - D_\theta \varepsilon_{t,k}(\theta_0)) \varepsilon_t|| = o_p(1). \tag{2.46}
\]

This follows from uncorrelatedness of $V_t^* - D_\theta \varepsilon_{t,k}(\theta_0)$ with $\varepsilon_t$, (2.40) and the Cauchy-
Schwarz inequality.

Hence (2.38), (2.39) hold and since \( \sqrt{T}||\nabla T(\theta_0)|| = O_p(\sqrt{k}) \) and \( ||\tau_1'\mathcal{H}^{-1}_T||_1 = O_p(k) \) with \( \tau_1' = (1, 0, \ldots) \in \mathbb{R}^{k+1} \), we have: \( \sqrt{T}\tau_1'(\hat{\beta} - \beta_0) = \sqrt{T}\tau_1'\mathcal{H}^{-1}_T(\theta_0)\nabla T(\theta_0) + o_p(1) \). Then we may proceed as in the proof of Theorem 2.2, which implies \( \sqrt{T}\tau_1'(\hat{\beta} - \beta_0) \xrightarrow{d} N(0, \sigma_0^2\tau_1'\mathcal{H}^{-1}_\infty\tau_1) \).

On the other hand, (2.38) implies \( T\hat{s}_\rho^2 - \sigma_0^2\tau_1'\mathcal{H}^{-1}_\infty\tau_1 = o_p(\tau_1'\mathcal{H}^{-1}_\infty\tau_1) = O_p(k) \) and that gives:

\[
\frac{\hat{\rho} - \rho_0}{\hat{s}_\rho} \xrightarrow{d} N(0, 1). \tag{2.47}
\]

This proves convergence of the test statistic under local alternatives. For convergence under the null simply take \( \delta = 0 \) and note that the error term in the regression (2.36) is \( \eta_t = \varepsilon_{t,k}(\theta) \) and additional condition on the rate is not needed, since \( e_{3t} = e_{4t} = 0 \). 

\[\square\]