Volatility models with a time-varying intercept

Preliminary, do not quote or distribute

Niklas Ahlgren[1], Alexander Back[1] & Timo Teräsvirta[2]

[1]Hanken School of Economics
[2]Aarhus University

February 28, 2022

Abstract

We propose a GARCH model augmented by a time-varying intercept. The intercept is parameterized by a logistic transition function with rescaled time as the transition variable, which provides a flexible and simple way of capturing deterministic non-linear changes in the conditional and unconditional variances. By making the intercept a smooth function of time, it is possible to capture changes that occur gradually, rather than abruptly as in regime switching models. It is common for financial time series to exhibit these types of shifts. The time-varying intercept makes the model globally nonstationary but locally stationary. We use the theory of locally stationary processes to derive the asymptotic properties of the quasi maximum likelihood estimator (QMLE) of the parameters of the model. We show that the QMLE is consistent and asymptotically normally distributed. To corroborate the results of the analysis, we provide a small simulation study. An empirical application on stock returns of large US corporations demonstrates the usefulness of the model. We find that the persistence implied by the workhorse GARCH(1,1) parameter estimates is reduced by incorporating a time-varying intercept. In particular, estimates that suggest an integrated volatility model can be reduced to lie within the stationary region.

1 Introduction

Volatility, interpreted as a measure of risk, plays an important role in financial management, and considerable efforts have been devoted to modelling it. The most popular time series models used for forecasting volatility are the Autoregressive conditional heteroskedasticity (ARCH) model by Engle (1982) and its generalisation, the General ARCH or GARCH model by Bollerslev (1986) and Taylor (1986). They are expected to capture the stylized facts in daily or weekly returns of financial assets and thus generate useful (short-run) forecasts for these series.

One such fact is that volatility exhibits a high level of persistence. Many studies have found that conditional volatility modelled using GARCH contains a 'unit root', that is, the unconditional variance implied by the estimated model becomes infinite. This leads to the so-called integrated GARCH (IGARCH) model; see Engle and Bollerslev (1986). Baillie et al. (1996) generalised the IGARCH model to a fractionally integrated GARCH model that can be either weakly stationary or nonstationary.

Diebold (1986), however, argued that time series of conditional volatility may appear integrated because there may be a trend driving the development of the series. Lamoureux and Lastrapes (1990) elaborated on this by arguing that high persistence might be due to time variation in all GARCH parameters and suggested that this can be accounted for by incorporating deterministic shifts in the unconditional variance. Morana (2002) showed that a stochastically shifting variance can also lead to high persistence. In this paper we, following many researchers, take up the lead proposed by these authors. We consider a GARCH model that was briefly mentioned by Teräsvirta (2012). This model, here called the Additive Time-Varying (ATV-) GARCH model, has a flexible, deterministically time-varying intercept that takes care of the nonstationarity regularly present in long daily or weekly asset return series. It is particularly well suited for situations in which volatility of an asset or index is systematically increasing or decreasing over time. Instead of becoming infinite as in IGARCH with a positive intercept, the unconditional variance of ATV-GARCH, under some parameter restrictions, remains finite over time.

The plan of the paper is as follows. In Section 2, we specify and discuss the model and related developments. We continue in Section 3 by considering estimation by maximum likelihood and state two theorems, namely consistency and asymptotic normality of the QMLE. In Section 4, we provide a short simulation study to corroborate the analysis and examine the small sample properties of the model. We conclude in Section 5 with an empirical example and a short discussion. Proofs are given in the appendix.

1.1 Notation

Throughout the paper, we use $\|W\|_p = (\mathbb{E}|W|^p)^{1/p}$ to denote the norm, and when applied to a matrix, we use $|\cdot|$ to denote the maximum norm. Further, let $A^{1/2}$ denote the Cholesky decomposition of a positive definite matrix A. We use $C_1, C_2, \ldots$ as generic constants, not necessarily the same across contexts.

2 The model

We define the model and consider some related developments.

2.1 The additive time-varying GARCH model

We consider the process

$$y_t = \mu_t + X_t,\ t = 1, 2, \ldots, T, \tag{1}$$

where $\mu_t = \mathbb{E}\{y_t|\mathcal{F}_{t-1}\}$, and $\mathcal{F}_t = \sigma(X_t, X_{t-1}, ...X_1)$ is the σ-field generated by $\{X_t, ..., X_1\}$. The error X_t is decomposed as $X_t = \varepsilon_t \sigma_{t,T}$ where $\varepsilon_t \sim$ IID$(0, 1)$ and

$$\sigma^2_{t,T} = g(t/T;\theta) + \sum_{i=1}^{p} \alpha_i X^2_{t-i,T} + \sum_{j=1}^{q} \beta_j \sigma^2_{t-j,T}. \tag{2}$$

The deterministic component $g(t/T;\theta)$ has the following form

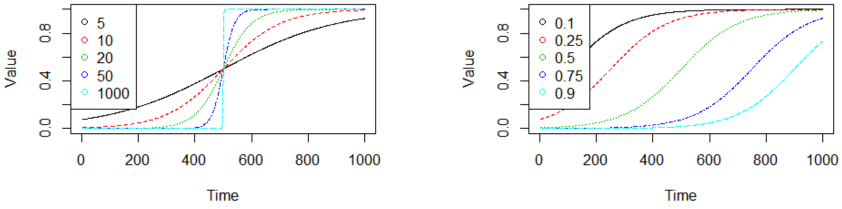
$$g(t/T;\theta) := \alpha_0 + \sum_{l=1}^{L} \alpha_{0l} G_l \left(\frac{t}{T}, \gamma_l, \boldsymbol{c}_l \right), \tag{3}$$

with $g(r;\theta) > 0$ for $r \in [0, 1]$, where

$$G(t/T, \gamma, \boldsymbol{c}) = \left(1 + \exp\left\{ -\gamma \prod_{k=1}^{K} (t/T - c_k) \right\} \right)^{-1}, \tag{4}$$

with $\gamma_l > 0,\ c_{11} < \ldots < c_{1K} < c_{21} < \ldots < c_{LK}$ for all $l = 1, \ldots, L, k = 1, \ldots, K$. Figure 1 shows how the logistic transition function behaves as the shape parameter γ and location parameter c change. In what follows we assume

Figure 1: The logistic transition function



(a) γ_1 varies, c_1 is fixed at 0.5.

(b) c_1 varies, γ_1 is fixed at 10

$\mu_t = 0$. The subscript T is to emphasize that we are working in rescaled time t/T, as is common in the literature of locally stationary processes. The decomposition $X_t = \varepsilon_t \sigma_{t,T}$ together with equations (2), (3) and (4) define the ATV-GARCH model. It is nonlinear, locally stationary and globally nonstationary. Local stationarity will be discussed in the next section. The GARCH equation (2) may easily be made asymmetric. The simplest way to do this is to augment the GARCH equation with an indicator variable taking the value one for negative values of ε_t, which corresponds to the GJR-GARCH model by Glosten et al. (1993). For notational simplicity we retain the form (2).

The model is unidentified if the intercept is constant. It is therefore advisable to test for time-variation in the intercept before attempting to fit the model. The test will be considered in a separate paper.

2.2 Relation to tvGARCH

The decomposition $X_{t,T} = \varepsilon_t \sigma_{t,T}$ together with (2), (3) and (4) define the ATV-GARCH model. The model belongs to the class of time-varying GARCH processes defined by

$$\begin{aligned} X_{t,T} &= \varepsilon_t \sigma_{t,T}, \\ \sigma^2_{t,T} &= \alpha_0\left(\frac{t}{T}\right) + \sum_{i=1}^{p} \alpha_i\left(\frac{t}{T}\right) X^2_{t-i,T} + \sum_{j=1}^{q} \beta_j\left(\frac{t}{T}\right) \sigma^2_{t-j,T}, \end{aligned}$$

where the parameters $\alpha_0(t/T)$, $\alpha_i(t/T)$, $i = 1, \ldots, p$, and $\beta_j(t/T)$, $j = 1 \ldots, q$, are smooth functions of time. The time-varying GARCH process is non-stationary, but under suitable conditions on the nonstationary parameters, $\{X_{t,T}\}$ can locally be approximated by a stationary GARCH process. Theoretical developments have been largely focused on this approach. For a survey of early developments in locally stationary volatility models, see van Bellegem (2012). For contributions in the theory of time-varying ARCH processes, see Dahlhaus and Subba Rao (2006), and for time-varying GARCH processes Subba Rao (2006), Rohan (2013), Chen and Hong (2016), Kristensen and Lee (2019) and Karmakar et al. (2020). In a way, the conditional variances in these models are more general than (4) in the sense that all α_i and β_i are time-varying. The main practical difference between them and ATV-GARCH is that in the former, the coefficients are nonparametric and estimated using kernel or local polynomial estimation, whereas in the latter the time-varying intercept is parametric: the ATV-GARCH model can be considered a parameterization of the more general tvGARCH, with

$$\alpha_0(t/T) := g(t/T),$$

where $g(t/T)$ is defined in (3) and (4), $\alpha_i(t/T) = \alpha_i$, $i = 1, \ldots, p$, and $\beta_j(t/T) = \beta_j$, $j = 1, \ldots, q$.

As the ATV-GARCH augments the GARCH equation by including an additive term, it can also be viewed in the light of a GARCH-X model, as considered by Han and Kristensen (2014). In the ATV-GARCH model, the additive term is a bounded, deterministic trend. In the GARCH-X model it is a stochastic regressor. As a consequence, the model has some similarities with the GARCH-X model with a nonstationary covariate.

The idea behind the model is that a felxible intercept that affects the persistence of the process adequately describes the systematic changes in volatility. We hypothesize that the inclusion of a time-varying intercept helps explain time-variation and structural changes often found in the unconditional volatility of financial time series such as stock, commodity and exchange rate returns.

2.3 Local stationary approximation

The inclusion of a time-varying intercept makes the ATV-GARCH model nonstationary. However, as the parametric intercept is a smooth function, the process has a locally stationary behaviour. As a consequence, there exists a stationary process which locally approximates the ATV-GARCH process in some neighbourhood of a fixed point in time. We will show that the ATV-GARCH model is locally stationary and a general theory for nonlinear locally stationary processes (Dahlhaus, Richter and Wu 2019, henceforth DRW) applies to the model. The theory relies on rescaling time to the unit interval, which enables a meaningful asymptotic analysis.

The concept of local stationarity was introduced already by Dahlhaus (1997), but the contributions of DRW are seminal in that the authors show that the rescaling device can be used to derive several general results, among them a global law of large numbers and a global central limit theorem. We use this theory applied to the ATV-GARCH model. To introduce the notion, following among others Dahlhaus and Subba Rao (2006), a process $X_{t,T}, t = 1, \ldots, T$, is said to be locally stationary if

$$X_{t,T} = \widetilde{X}_t(u) + O_p\left(|t/T - u| + \frac{1}{T}\right), \tag{5}$$

where $u \in [0, 1]$ and $\widetilde{X}_t(u)$ is a stationary approximation at u. The concept relies on using the triangle inequality to decompose the difference between the process and the stationary approximation as

$$\left|X_{t,T} - \widetilde{X}_t(u)\right| \leq \left|X_{t,T} - \widetilde{X}_t(t/T)\right| + \left|\widetilde{X}_t(t/T) - \widetilde{X}_t(u)\right|. \tag{6}$$

From (6) it can be seen that if t/T is close to u, then $X_{t,T}$ and $\widetilde{X}_t(u)$ should be close and the degree of the approximation should depend on the rescaling factor T and the deviation $|t/T - u|$.

The idea in this paper is to use the concept of local stationarity to extend the theory of QML estimation of standard GARCH models to the case where the intercept is parametric and time-varying. We use the framework in Berkes et al. (2003), henceforth BHK, as a baseline for the theory, with the goal of replacing the traditional limit theorems for stationary processes by their locally stationary counterparts in DRW.

To show that the ATV-GARCH model is locally stationary, we use some results derived in Subba Rao (2006) for time-varying GARCH processes. Theorem 2.1 of Subba Rao (2006) gives conditions under which a nonstationary, nonlinear process with time-dependent parameters can be approximated locally by a stationary process. The author uses the theorem to show that under the conditions, the tvGARCH process $X_{t,T}$ can be locally approximated by the stationary GARCH process $\{X_t(u)\}$ given by

$$\begin{aligned} \widetilde{X}_t(u) &= \varepsilon_t \sigma_t(u), \\ \sigma_t^2(u) &= \alpha_0\left(u\right) + \sum_{i=1}^{p} \alpha_i\left(u\right) \widetilde{X}_{t-1}^2(u) + \sum_{j=1}^{q} \beta_j\left(u\right) \sigma_{t-j}^2(u). \end{aligned}$$

In our case, it is given by fixing the intercept at the value that the function g takes at u. The time-varying GARCH process is then said to be locally stationary in the sense that

$$X_{t,T}^2 = \widetilde{X}_t^2(u) + \left(\left| \frac{t}{T} - u \right| + \frac{1}{T} \right) R_{t,T}, \quad \text{where} \quad \sup_{t,T} \mathbb{E}(R_{t,T}) < \infty. \tag{7}$$

The requirements for a tvGARCH to be locally stationary are as follows. **Assumption 1.** Denote $\mu_n = \{\mathbb{E}(\varepsilon_t^{2n})\}$. The parameters satisfy the following properties.

(i) The parameters $\{\alpha_i(\cdot)\}$ and $\{\beta_j(\cdot)\}$ are Lipschitz continuous, i.e. $|a_i(u) - a_i(v)| \leq C\,|u-v|$ and $|\beta_j(u) - \beta_j(v)| \leq C\,|u-v|$, where C is a finite constant.

(ii)

$$\sup_u \left\{ \sum_{i=1}^{p} \alpha_i\left(u\right) + \sum_{j=1}^{q} \beta_j\left(u\right) \right\} < 1 - \eta,$$

(iii) For some $n \in [1, \infty)$,

$$\mu_n \sup_u \left\{ \sum_{i=1}^{p} \alpha_i\left(u\right) + \sum_{j=1}^{q} \beta_j\left(u\right) \right\} < 1 - \eta.$$

Proposition 1. Assume that $\mathbb{E}(\varepsilon_t^2) = 1$, the parameter space $\mathit{\Theta}$ is compact and that $\sum_{i=1}^{p} \alpha_i + \sum_{j=1}^{q} \beta_j < 1$. Then the ATV-GARCH model is locally stationary in the sense (7).

Proof. By taking $n = 1$, (ii) and (iii) in Assumption 1 are immediately fulfilled. We show in the Appendix that the time-varying intercept is Lipschitz continuous under compactness of the parameter space. □

Note that the condition on the GARCH coefficients is a necessary condition for weak stationarity of the standard GARCH(p, q) process. For the sake of estimation by QML, this is a stronger assumption than needed in the standard (strictly) stationary case, where a weaker condition on the coefficients can be derived in terms of the top Lyapunov exponent (BHK, Francq and Zakoïan (2004)).

Further, in the standard case, one does not need to assume that the GARCH process has a fourth moment in order to derive the asymptotic properties of the QMLE. It suffices to make a fourth moment assumption on the errors (BHK, Francq and Zakoïan (2004)). Here, we shall need to make an assumption on the fourth moment of the process, which in practice restricts the sum of the GARCH coefficients further. The assumption is due to the existence of a stationary approximation and will be discussed in the next section.

2.4 Other related developments

The ATV-GARCH model is most closely related to the tvGARCH and GARCH-X models, which we discussed in Section 2.2. In this section, we survey some other related developments.

Other nonlinear GARCH models with smooth transition have been developed in order to take assymmetry of shocks into account. These include the Partially Nonparametric GARCH by Engle and Ng (1993), Smooth Transition GARCH by Hagerud (1997), Gonzalez-Rivera (1998), Anderson et al. (1999) and Lubrano (2001) and Exponential GARCH (Nelson, 1991). Lanne and Saikkonen (2005) considered these types of models from a theoretical standpoint. The authors derived sufficient conditions for the stationarity and existence of moments of various smooth transition GARCH models. As a GARCH$(1,1)$ example they gave

$$\sigma_t^2 = \alpha_0 + \alpha_1 X_{t-1}^2 + \delta_1 g(\sigma_{t-1}^2; \gamma) + \beta_1 \sigma_{t-1}^2, \tag{8}$$

where

$$g(\sigma_{t-1}^2; \gamma) = \int_0^{\sigma_{t-1}^2} \frac{1}{\Gamma(\gamma)} s^{\gamma-1} e^{-s} \mathrm{d}s$$

is the cdf of a standard gamma-distributed random variable, so it obtains values in $[0,1]$. It differs from (2) in that the additive nonlinear component is stochastic and involves σ_{t-1}^2.

The expression (3) also has similarities with the neural network GARCH model by Caulet and Péguin-Feissolle (2000). In their model, the argument in the logistic function in (4) is a linear combination of $\varepsilon_{t-1}, ..., \varepsilon_{t-q}$. Besides, $\alpha_i = 0$, $i = 1, ..., q$, and $\beta_i = 0$, $i = 1, ..., p$, in (2). The idea with logistic transition functions appeared in yet another form in the flexible coefficient GARCH (FCGARCH) model by Medeiros and Veiga (2009). In their model,

$$\sigma_t^2 = \sum_{i=1}^{L} (\alpha_{0i} + \alpha_{1i} X_{t-1}^2 + \beta_{1i} \sigma_{t-1}^2) G_i(s_t; \gamma_i, c_i), \tag{9}$$

where the transition variable s_t is strictly stationary and ergodic, and in deriving asymptotic results it was assumed that $s_t = X_{t-1}$. As the authors note, this precludes using time as the transition variable. Furthermore, compared to (4), $K = 1$. In this very flexible functional form, all linear coefficients are changing over time. When $s_t = X_{t-1}$, (9) is still univariate. Teräsvirta (2012) contains a survey of univariate nonlinear GARCH models.

If the deterministic function $g(t/T;\theta)$ in (3) is replaced by a positive-valued stochastic variable Z_{t-1}, one obtains a so-called GARCH-X model; see Han and Kristensen (2014). These authors provided asymptotic theory for maximum likelihood (ML) estimators for both the case in which Z_{t-1} is stationary and for the one in which it is nonstationary.

The ATV-GARCH model is an example of an *additive* decomposition of the conditional variance in that the deterministic component in (2) is additive. Another popular decomposition is the *multiplicative* one. The idea is to rescale the observations such that the resulting GARCH process is weakly stationary. This is useful when nonstationarity is due to the fact that the amplitude of volatility clusters varies over time. Thus, $X_t = \varepsilon_t \sigma_{t,T} g^{1/2}(t/T)$, where $g(t/T)$ is a positive-valued deterministic function of time, and so $X_t/g^{1/2}(t/T) = \varepsilon_t \sigma_{t,T}$. This decomposition was introduced by Feng (2004) and van Bellegem and von Sachs (2004) who estimated $g(t/T)$ nonparametrically. Amado and Teräsvirta (2008, 2013, 2017) took a parametric approach and used (3) and (4) to define $g(t/T)$. Engle and Rangel (2008) related temporally aggregated $g(t/T)$, estimated by exponential quadratic splines, to macroeconomic fluctuations of a panel of countries. For a recent survey, see Amado et al. (2019). The function $g(\cdot)$ may also be stochastic: $g_t = g(x_t)$, see, for example, Amado and Laakkonen (2013) and Han and Kristensen (2017).

Examples of globally nonstationary but not locally stationary GARCH models comprise GARCH models with a structural break or breaks, see Andreou and Ghysels (2009) for a survey. When $\gamma_l \to \infty$, $l = 1, ..., L$, in (4), the ATV-GARCH model approaches a GARCH model with breaks. These models can be piecewise stationary. It is possible to test for stationary subperiods and fit GARCH models to them; see for example Härdle et al. (2003) and, for surveys Čížek and Spokoiny (2009) and Čížek (2011).

Lastly, we take a look at the time-varying GARCH model by Chen et al. (2014) because it contains not only smooth transitions but also a single discontinuity or break. The structure of the variance component resembles that of Medeiros and Veiga (2009) with some modifications. First, $s_t = t/T$. Second, asymmetry is introduced by thresholds, so the (first-order) variance equation becomes

$$\begin{aligned} \sigma_t^2 &= \alpha_0 + \alpha_1 X_{t-1}^2 + \beta_1 \sigma_{t-1}^2 + \sum_{i=1}^{L} (\alpha_{0i} + \alpha_{1i} X_{t-1}^2 + \beta_{1i} \sigma_{t-1}^2) \\ &\times \; G_i(t/T;\gamma_i, c_i) I(t/T \in S_i) \end{aligned} \tag{10}$$

where $I(\cdot)$ is an indicator variable. Furthermore,

$$G_i(t/T;\gamma_i, c_i) = (1 + \exp\{(-1)^{H-i+1}\gamma(\frac{t}{T} - c_i)\})^{-1}, \ \gamma > 0$$

so the odd-indexed transition functions $G_1, G_3, ...$ are decreasing functions of time, whereas the even ones are increasing. To illustrate, set $L = 2$, and denote

$\sigma_t^2 = h_t + g(t/T)$. This yields $h_t = \alpha_0 + \alpha_1 X_{t-1}^2 + \beta_1 \sigma_{t-1}^2$, and

$$\begin{aligned} g(\frac{t}{T}) &= (\alpha_{01} + \alpha_{11} X_{t-1}^2 + \beta_{11}\sigma_{t-1}^2) G_1(t/T; \gamma_i, c_i) I(\frac{t}{T} \le \frac{t_0}{T}) \\ &+ (\alpha_{02} + \alpha_{12} X_{t-1}^2 + \beta_{12}\sigma_{t-1}^2) G_2(t/T; \gamma_i, c_i) I(\frac{t}{T} > \frac{t_0}{T}). \end{aligned}$$

Note that t_0 is not a 'free' break-point parameter. It is defined as $t_0/T = (\gamma_1 c_1 + \gamma_2 c_2)/(\gamma_1 + \gamma_2)$, where it is assumed that $0 < c_i < 1$, $i = 1, 2$, and $c_1 + c_2 < 1$. Asymmetry is introduced through t_0. Chen et al. (2014) highlighted various identification issues that entail this parameterisation. Model selection and parameter estimation were conducted using the Bayesian MCMC algorithm. A comparison of (10) with the ATV-GARCH model shows that in the former, all parameters are changing over time, whereas in the latter, only the intercept is time-varying.

The conditional variance (2) is simpler than (10), but depending on L, the number of transitions in (3) and K_l and the shape of the transition functions in (4), the intercept can be made a very flexible function of time. Furthermore, L can be determined by sequential testing that includes the possibility that the intercept is constant. Modelling issues of the ATV-GARCH model will be discussed elsewhere. In this paper, the main focus will be on QML estimation and properties of the QML estimators, to which we now turn.

3 Estimation by maximum likelihood

In this section, we discuss some notation and concepts related to the implementation and theory of estimation of the model by QML.

3.1 The log likelihood function

We define the log likelihood function, a truncated version of it and a stationary approximation. We use the recursive definition of the process as given by BHK. Let θ_0 denote the parameter vector at the "true" values. Define

$$\hat{\theta} = \arg\max_{\theta \in \Theta} \frac{1}{T} \sum_{t=1}^{T} l_t(\theta), \tag{11}$$

where $l_t(\theta)$ is the log likelihood function for observation t

$$l_t(\theta) = -\frac{1}{2} \left[\log h_t(\theta) + \frac{X_t^2}{h_t(\theta)} \right], \tag{12}$$

with

$$h_t(\theta) = \frac{\alpha_0}{1 - \sum_{i=1}^{q} \beta_i} + \sum_{j=1}^{\infty} d_j(\theta) g_{t-j+1} + \sum_{j=1}^{\infty} c_j(\theta) X_{t-j}^2, \tag{13}$$

$$:= c_0(\theta) + \sum_{j=1}^{\infty} c_j(\theta) X_{t-j}^2, \tag{14}$$

where the functions $c_j(\cdot)$ and $d_j(\cdot)$ are given in BHK. For the stationary case, BHK (Theorem 2.1) prove that this representation yields σ_t^2 almost surely. The proof can be extended to allow for our case as well. To see this, we consider for simplicity the ATV-GARCH(1,1) model and proceed as in BHK, Theorem 2.1. Some recursive substitution yields

$$\begin{aligned} \alpha_0 + g_t + \alpha_1 X_{t-1}^2 &= \sigma_t^2 - \beta_1 \sigma_{t-1}^2 \\ \alpha_0 + g_t + \alpha_1 X_{t-1}^2 + \beta_1(\alpha_0 + g_{t-1} + \alpha_1 X_{t-2}^2) &= \sigma_t^2 - \beta_1^2 \sigma_{t-2}^2, \end{aligned}$$

and so on. More generally, defining $\Phi_t = \alpha_0 + g_t + \alpha_1 X_{t-1}^2$, we get

$$\Phi_t + \beta_1 \Phi_{t-1} + \ldots + \beta_1^j \Phi_{t-j} = \sigma_t^2 - \beta^j \sigma_{t-j}^2. \tag{15}$$

In the stationary case, BHK Lemma 2.2 can be invoked to show that, under some mild regularity conditions, the LHS converges almost surely as $j \to \infty$. The exponential decay of the second term on the RHS then gives the result. If we define $\Phi_t^* = \alpha_0 + \sup_{[0,1]} g_t + \alpha_1 X_{t-1}^{*2}$, and σ_t^{*2} similarly, we see that since the terms have stationary processes that bound them everywhere, the result continues to hold in our case. Note that $\sigma_t^2 = h_t(\theta_0)$. This produces a natural way of generating the process without the need to initialize it with an arbitrary starting value, but as Francq and Zakoïan (2004) notes, the computational cost of the procedure is of order $O(n^2)$. In practice, any implementation should be optimized for speed before attempting to fit the model to a long time series.

We only observe a finite number of observations. The truncated (”feasible”) estimator is given by:

$$\bar{\theta} = \arg\max_{\theta \in \Theta} \frac{1}{T} \sum_{t=1}^{T} \bar{l}_t(\theta), \tag{16}$$

where $\bar{l}_t(\theta)$ is the log likelihood for observation t

$$\bar{l}_t(\theta) = -\frac{1}{2} \left[\log \bar{h}_t(\theta) + \frac{X_t^2}{\bar{h}_t(\theta)} \right], \tag{17}$$

with

$$\bar{h}_t(\theta) = \bar{c}_0(\theta) + \sum_{j=1}^{t-1} c_j(\theta) X_{t-j}^2. \tag{18}$$

To prove consistency of the truncated estimator $\bar{\theta}_T$ from consistency of the estimator $\widehat{\theta}_T$, it is sufficient to show that the truncated log likelihood function $\bar{L}_T(\theta)$ converges uniformly to the log likelihood function $L_T(\theta)$. In order to motivate how this works, we reproduce some arguments from BHK and discuss how they relate to our model in the Appendix.

We also need the estimator stemming from the stationary approximation. Define

$$\widetilde{\theta} = \arg\max_{\theta \in \Theta} \frac{1}{T} \sum_{t=1}^{T} l_t(u, \theta). \tag{19}$$

where $l_t(u,\theta)$ is the log likelihood for observation t

$$l_t(u,\theta) = -\frac{1}{2}\left[\log \widetilde{h}_t(u,\theta) + \frac{\widetilde{X}_t^2(u)}{\widetilde{h}_t(u,\theta)}\right], \tag{20}$$

with

$$\widetilde{h}_t(u,\theta) = \widetilde{c}_0(\theta) + \sum_{j=1}^{\infty} c_j(\theta)\widetilde{X}_{t-j}^2(u), \tag{21}$$

for some $u \in [0,1]$, where the expression is obtained by approximating $g(t/T)$ by a constant value $g(u)$.

Define

$$L_T(\theta) := \frac{1}{T}\sum_{t=1}^{T} l_t(\theta),$$

$$\bar{L}_T(\theta) := \frac{1}{T}\sum_{t=1}^{T} \bar{l}_t(\theta),$$

$$L_T(u,\theta) := \frac{1}{T}\sum_{t=1}^{T} l_t(u,\theta)$$

and

$$L(u,\theta) := \mathbb{E}[l_t(u,\theta)].$$

3.2 The Score and the Hessian

Denote the score for observation t

$$S_t(\theta) = \frac{\partial l_t(\theta)}{\partial \theta}$$

and the Hessian for observation t

$$H_t(\theta) = \frac{\partial^2 l_t(\theta)}{\partial\theta\partial\theta^T},$$

and $\bar{S}_t(\theta), \widetilde{S}_t(\theta)$ and $\bar{H}_t(\theta), \widetilde{H}_t(\theta)$ similarly. Note that $\widetilde{S}_t(\theta)$ and $\widetilde{H}_t(\theta)$ *are not* derivatives of the stationary approximation of the log likelihood, but rather the stationary approximations of these derivatives. We have that

$$\begin{aligned} S_t(\theta) &= \frac{\partial}{\partial\theta}\left\{\ln h_t(\theta) + \frac{X_t^2}{h_t(\theta)}\right\} \\ &= \left(1 - \frac{X_t^2}{h_t(\theta)}\right)\frac{1}{h_t(\theta)}\frac{\partial h_t(\theta)}{\partial\theta} \end{aligned}$$

and

$$\begin{aligned} H_t(\theta) &= \frac{\partial}{\partial\theta^T}\left\{\left(1-\frac{X_t^2(\theta)}{h_t(\theta)}\right)\frac{\partial\ln h_t(\theta)}{\partial\theta}\right\} \\ &= \left(1-\frac{X_t^2}{h_t(\theta)}\right)\frac{1}{h_t(\theta)}\frac{\partial^2 h_t}{\partial\theta\partial\theta^T}+\left(2\frac{X_t^2}{h_t(\theta)}-1\right)\frac{1}{h_t(\theta)}\frac{\partial h_t(\theta)}{\partial\theta}\frac{1}{h_t(\theta)}\frac{\partial h_t(\theta)}{\partial\theta^T}. \end{aligned}$$

Similarly to the argument for consistency, to prove asymptotic normality of the truncated estimator $\bar{\theta}_T$ from asymptotic normality of the estimator $\widehat{\theta}_T$, it suffices to show that the difference $|\widehat{\theta}_T-\bar{\theta}_T|$ tends to zero faster than the factor $\sqrt{T}$ in the limiting distribution $\sqrt{T}(\widehat{\theta}_T-\theta_0)$. The difference between the score of the log likelihood function evaluated at the estimator $\widehat{\theta}_T$ and the score of the truncated log likelihood function evaluated at the truncated estimator $\bar{\theta}_T$ is

$$S_T(\widehat{\theta}_T)-\bar{S}_T(\bar{\theta}_T).$$

We can replace $\bar{S}_T(\bar{\theta}_T)$ by $S_T(\bar{\theta}_T)$, and the error is

$$S_T(\widehat{\theta}_T)-S_T(\bar{\theta}_T)=S(\widehat{\theta}_T)-\bar{S}_T(\bar{\theta}_T)+O_P\left(\frac{1}{T}\right)=O_P\left(\frac{1}{T}\right). \qquad (22)$$

As in BHK, Theorem 4.4, linearize the difference $\widehat{\theta}_T-\bar{\theta}_T$ by coordinatewise application of the mean value theorem:

$$S(\widehat{\theta}_T)-S(\bar{\theta}_T)=(\widehat{\theta}_T-\bar{\theta}_T)H_T(\theta^*)+O_P\left(\frac{1}{T}\right), \qquad (23)$$

where (for lack of better notation) θ^* lies (coordinate wise) between $\hat{\theta}_T$ and $\bar{\theta}_T$. Here, $H_T(\theta^*)$ is an average evaluated at θ^*. N2 together with N3 in Appendix A imply that

$$H_T(\theta^*)\overset{P}{\to}H(\theta_0).$$

See Francq and Zakoian (2004), (vi) on p. 618 and the proof on p. 626. We can write (23) as

$$S(\widehat{\theta}_T)-S(\bar{\theta}_T)=(\widehat{\theta}_T-\bar{\theta}_T)H(\theta_0)(1+o_P(1)),$$

which then implies

$$|\widehat{\theta}_T-\bar{\theta}_T|=O_P\left(\frac{1}{T}\right).$$

Since the difference between $\widehat{\theta}_T$ and $\bar{\theta}_T$ is $O_P\left(\frac{1}{T}\right)$ and the factor in the limiting distribution is $\sqrt{T}$, it is legitimate to replace $\widehat{\theta}_T$ by $\bar{\theta}_T$ in the limiting distribution.

3.3 On moments and related assumptions

We shall need to make assumptions on the moment structure of the error term, as well as on the coefficients. It is well known that the standard GARCH model has a finite second moment under the assumption (A4) below, coupled with existence of the second moment of the error term. If one instead requires strict stationarity, (A4) can be relaxed considerably. By using the developments of Brandt (1986), who studied a particular stochastic differential equation, Bougerol and Picard (1992) were able to show that strict stationarity entails a weaker condition. McAleer and Ling (2002) used a similar argument to show a necessary and sufficient condition for the existence of the fourth moment of a GARCH(p, q) process. Subba Rao (2006), which contains results that we use extensively, uses the same framework to show existence of moments for a class of locally stationary processes. The class contains in particular time-varying GARCH processes, which is illustrated in an example in the article. In this subsection, we discuss this framework and how it relates to our model. For a detailed exposition of the framework in the stationary case, see Francq and Zakoïan (2019).

Following Subba Rao (2006), the time-varying GARCH(p, q) process $\{X^2_{t,T}\}$ admits the state space representation (assume without loss of generality that $p, q \geq 2$)

$$\mathcal{X}_{t,T} = b_t\left(\frac{t}{T}\right) + A_t\left(\frac{t}{T}\right)\mathcal{X}_{t-1,T}$$

with

$$b_t(u) = \begin{pmatrix} \alpha_0(u) \\ 0 \\ \vdots \\ 0 \end{pmatrix} \in \mathbb{R}^{p+q-1}, \quad \mathcal{X}_{t,T} = \begin{pmatrix} \sigma^2_{t,T} \\ \vdots \\ \sigma^2_{t-q+1,T} \\ X^2_{t-1,T} \\ \vdots \\ X^2_{t-p+1,T} \end{pmatrix}$$

and

$$A_t(u) = \begin{pmatrix} \tau_t(u) & \beta_q(u) & \alpha(u) & \alpha_p(u) \\ \mathbf{I}_{q-1} & \mathbf{0} & \mathbf{0} & \mathbf{0} \\ \mathbf{Z}^2_{t-1} & 0 & \mathbf{0} & 0 \\ \mathbf{0} & \mathbf{0} & \mathbf{I}_{p-2} & \mathbf{0} \end{pmatrix},$$

a $(p+q-1)\times(p+q-1)$ matrix where $\tau_t(u) = (\beta_1(u)+\alpha_1(u)Z^2_{t-1}, \beta_2(u), \ldots, \beta_{q-1}(u))$, $\alpha(u) = (\alpha_2(u), \ldots, \alpha_{p-1}(u))$ and $\mathbf{Z}^2_{t-1} = (\varepsilon^2_{t-1}, 0, \ldots, 0) \in \mathbb{R}^{q-1}$. In the locally stationary framework of Subba Rao (2006), it is further necessary to define quantities that bound these matrices in each interval of rescaled time. Since we only have time variation in the intercept, it suffices to define

$$\widetilde{b}_t = \begin{pmatrix} \sup_{u\in[0,1]} \alpha_0(u) \\ 0 \\ \vdots \\ 0 \end{pmatrix} \in \mathbb{R}^{p+q-1}.$$

Define the stationary approximation of the representation in the obvious way. By Lipschitz continuity of the time-varying intercept, it follows from Subba Rao (2006), Theorem 2.1 that

$$\left|\mathcal{X}_t - \widetilde{\mathcal{X}}_t(u)\right| \leq |t/T - u|\, W_t + \frac{1}{T} V_t, \tag{24}$$

where W_t and V_t are stochastic processes.

In the asymptotic theory, we shall need to show the validity of assumption S1 and S2 in DRW (see the Appendix). We need to show expressions of the type

$$\left\|\mathcal{X}_t - \widetilde{\mathcal{X}}_t(u)\right\|_1 \leq |t/T - u|\, C_1 + \frac{1}{T} C_2 \tag{25}$$

and

$$\left\|\mathcal{X}_t - \widetilde{\mathcal{X}}_t(u)\right\|_2 \leq |t/T - u|\, C_3 + \frac{1}{T} C_4. \tag{26}$$

It is clear from (24) that this means requiring the existence of moments of V_t and W_t. For a matrix A, define $[A]_n = \left(\mathbb{E}\left|A_{i,j}\right|^n\right)^{1/n}$ and let $\lambda_{\text{spec}}[A]$ denote the largest absolute eigenvalue of A. By Subba Rao (2006), Proposition 2.1, if the conditions of Subba Rao (2006) Theorem 2.1 are fulfilled and 1) $\left\|\widetilde{b}_t\right\|_n^n < \infty$ and 2) $\lambda_{\text{spec}}[A]_n < 1 - \delta$ for some (potentially very small) $\delta > 0$, then

$$\sup_{t/T} \|V_t\|_n^n < \infty \tag{27}$$

and

$$\sup_{t/T} \|W_t\|_n^n < \infty. \tag{28}$$

By boundedness of the intercept, 1) is fulfilled. We shall need to assume 2) with $n = 2$. Because we have no time-variation in the parameters contained in the matrix A, this is equivalent to the necessary and sufficient condition for the existence of a fourth moment of the GARCH process in McAleer and Ling (2002). To see this, consider an ATV-GARCH(1,1) model. The state space representation is given by

$$b_t(u) = \begin{pmatrix} \alpha_0 + \alpha_{01} G(u) \\ 0 \\ 0 \end{pmatrix},$$

$$A_t(u) = \begin{pmatrix} \beta_1 + \alpha_1 Z_{t-1}^2 & 0 & 0 \\ 1 & 0 & 0 \\ Z_{t-1}^2 & 0 & 0 \end{pmatrix},$$

and

$$\begin{aligned}
\sup_u |b_t(u)| &\leq \begin{pmatrix} \alpha_0 + \alpha_{01} \\ 0 \\ 0 \end{pmatrix}, \\
([A_t(u)]_2) &= \begin{pmatrix} \{\mathbb{E}(\beta_1 + \alpha_1 Z_{t-1}^2)^2\}^{1/2} & 0 & 0 \\ 1 & 0 & 0 \\ \{\mathbb{E}(Z_{t-1}^2)^2\}^{1/2} & 0 & 0 \end{pmatrix} \\
&= \begin{pmatrix} \{\beta_1^2 + 2\alpha_1\beta_1\mathbb{E}(Z_t^2) + \alpha_1^2\mathbb{E}(Z_t^4)\}^{1/2} & 0 & 0 \\ 1 & 0 & 0 \\ \{\mathbb{E}(Z_t^4)\}^{1/2} & 0 & 0 \end{pmatrix}.
\end{aligned}$$

The condition $\lambda_{\text{spec}}([A_t(u)]_2) < 1 - \delta$ translates into

$$\beta_1^2 + 2\alpha_1\beta_1\mathbb{E}(Z_t^2) + \alpha_1^2\mathbb{E}(Z_t^4) < 1.$$

Figure 2 depicts how restrictive the assumption is in the case of an (ATV)-GARCH(1,1) and normally distributed innovations.

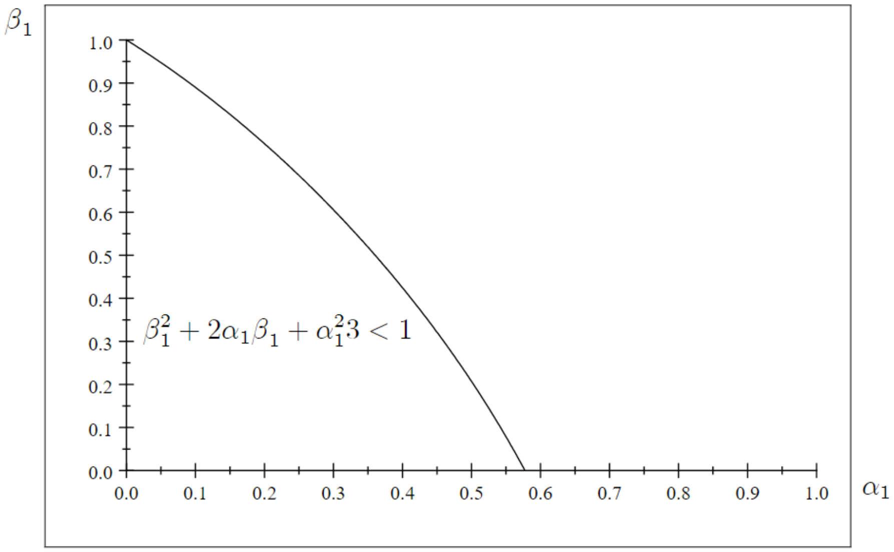


Figure 2: The fourth moment assumption for an (ATV)-GARCH(1,1) with $N(0,1)$ errors.

3.4 Assumptions

We use the framework developed by BHK as a baseline for our theory. BHK develop a theory for the consistency and asymptotic normality of the GARCH(p, q) QMLE. Consequently, we largely inherit their assumptions. Let $\theta \in \Theta$ be a vector containing the parameters in (2), that is, $\theta = (\alpha_0, \boldsymbol{\alpha}', \boldsymbol{\beta}', \boldsymbol{\delta}', \boldsymbol{\gamma}', \boldsymbol{c}')'$, where $\boldsymbol{\alpha} = (\alpha_1, \ldots, \alpha_p)'$, $\boldsymbol{\beta} = (\beta_1, \ldots, \beta_q)'$, $\boldsymbol{\delta} = (\alpha_{01}, \ldots, \alpha_{0L})'$, $\boldsymbol{\gamma} = (\gamma_1, \ldots, \gamma_L)'$ and $\boldsymbol{c} = (c_{11}, \ldots, c_{1L}, c_{21}, \ldots, c_{KL})'$. We make the following assumptions.

(A1) The random variables $\varepsilon_1, \ldots, \varepsilon_T$ are IID with $\mathbb{E}(\varepsilon_0) = 0$ and $\mathbb{E}(\varepsilon_0^2) = 1$, $\mathbb{E}\left|\varepsilon_0^2\right|^d < \infty$ for some $d > 0$ and ε_0^2 is non-degenerate. Further, $\lim\limits_{t \to 0} t^{-\mu}\mathbb{P}\{\varepsilon_0^2 \leq t\} = 0$, for some $\mu > 0$.

(A2) The parameter space Θ is compact and $\theta_0 \in \text{int}(\Theta)$.

(A3) $\alpha_0 > 0, \alpha_0 + \sum\limits_{l=1}^{L} \alpha_{0l} G_l\left(u, \gamma_l, \boldsymbol{c}_l\right) > \inf_{\theta \in \Theta} \alpha_0 \ \forall u \in [0, 1]$. The functions G_l are non-constant for each l, that is, $\forall l$ and $\forall k$ $\gamma_l, c_{lk}, |\alpha_{0l}| > 0$.

(A4) $\sum_{i=1}^{p} \alpha_i + \sum_{j=1}^{q} \beta_j < 1$, $\alpha_1, \alpha_2, \ldots, \alpha_p \geq 0$, $\beta_1, \beta_2, \ldots, \beta_q \geq 0$. Moreover, $\lambda_{\text{spec}}[A_t]_2 < 1 - \delta$ for some $\delta > 0$, where A_t is given in Section 3.3.

(A5) The polynomials $\mathcal{A} = \alpha_1 x + \alpha_2 x^2 + \ldots + \alpha_p x^p$ and $\mathcal{B} = \beta_1 x + \beta_2 x^2 + \ldots + \beta_q x^q$ are coprime on the set of polynomials with real coefficients.

(A6) $\mathbb{E}|\varepsilon_t|^{4+s} < \infty$ for some $s > 0$.

Remark 1. *(A1) and (A5) are directly inherited from BHK and (A2) is a standard assumption for proving consistency and asymptotic normality. (A3) is a modified nonnegativity condition. (A4) is discussed in the previous subchapter. (A6) is a moment condition on the innovations.*

4 Asymptotic theory

As a part of a general theory DRW discuss estimation by QML of non-linear, locally stationary processes. The estimation framework in DRW is non-parametric. To characterize the processes that can be estimated as locally stationary, the results in DRW rely on imposing smoothness conditions on the curves that model parameter change. In a non-parametric estimation procedure, these smoothness conditions are assumed. Any parameterization naturally necessitates verifying the conditions, rather than leaving them as high-level assumptions. In practice, this means verifying that Lipschitz or Hölder-type conditions hold for the parameter curves in the conditional variance equation, as well as for necessary transformations of them. The log likelihood function and its derivatives are such necessary transformations.

We use results from Subba Rao (2006) to show that our model and the related transformations are locally stationary. As mentioned in Section 3, we use the framework in BHK as a baseline for the theory, with the goal of motivating a substitution of the traditional limit theorems for stationary processes to their locally stationary counterparts in DRW.

4.1 Consistency

Theorem 1. *Under (A1-A6)*

$$\bar{\theta}_T \xrightarrow{p} \theta_0 \text{ as } T \to \infty. \tag{29}$$

Proof. The proof is given in Appendix A. □

Remark 2. *In order to show the result, we need to apply a law of large numbers for locally stationary processes to the log likelihood function, namely Theorem 2.7(i) in DRW. Due to an application of the Cauchy-Schwarz inequality, we require Assumptions (A4) and (A7) already in the proof of consistency.*

4.2 Asymptotic normality

Theorem 2. *Under (A1-A6)*

$$\sqrt{T}\left(\bar{\theta}_T - \theta_0\right) \xrightarrow{D} N\left(B^{-1}AB^{-1}\right) \text{ as } T \to \infty, \tag{30}$$

where the matrices A and B are given in the Appendix.

Proof. The proof is given in Appendix A. □

5 Practical considerations and simulation study

We discuss the implementation of the model and some practical consideration regarding estimation. We conduct a simulation study to corroborate the theoretical analysis and demonstrate the small sample properties of the test. In what follows, we shall consider an ATV-GARCH(1,1)-model with one transition function and refer to the parameters α_{01}, α_1 and β_1 as the "GARCH parameters", and γ_1, c_1 and α_{01} as the "G parameters".

5.1 Considerations

When conducting simulations, the "true" DGP is given by an ATV-GARCH model, so we would expect the parameters to be consistently estimated and asymptotically normal. In empirical applications, however, it is implicit that the hypothesized DGP is a GARCH model with some non-linear structure in the intercept. The logistic transition function approximates this non-linearity. It is therefore perhaps less important to look at the exact values and t-statistics of the G parameters, and more important to look at the shape of the transition that they imply.

Especial care need to be taken when estimating the speed of the transition, i.e. the γ parameter in an ATV-GARCH model. As γ increases, its marginal effect on the shape of the logistic transition function becomes smaller. Therefore, the optimization routine cannot distinguish between large values of γ, which in practice seems to lead to the parameter increasing towards infinity. By requiring that the parameter space is compact, or equivalently (for a metric space) closed and bounded, we have assumed that each parameter is bounded from above and below by some value.

Further, When γ increases, the derivatives of the logistic transition function increase around c, making the required Lipschitz constant larger. For the theory to be valid, γ needs to be a finite value. Besides, one might argue that large values of γ_1 defeats the purpose of the model. The transitions are meant to be smooth, which is to say easily distinguishable from a step function.

For the other parameters, we have imposed the necessary conditions in our assumptions, as well as upper and lower bounds that are very unlikely to be binding in any practicable application. This is necessary as we have assumed that the parameter space is compact.

To overcome these issues, we note that we can follow Ekner and Nejstgaard (2013) who propose rescaling γ to the unit interval. This can be achieved by replacing γ with $\gamma = \delta/(1-\delta)$, where δ is the parameter to be estimated. For identification, the case $\delta = 0$ has to be ruled out, and for preservation of Lipschitz continuity, the case $\delta = 1$ is prohibited, so $\delta \in (0,1)$. As in the LSTAR-case in Ekner and Nejstgaard (2013), we have found that this small reparameterization is conducive to stable estimation of the speed of transition.

In general, we expect there to be a trade-off between the speed of convergence of the estimators and the numerical accuracy of the estimation. We hypothesize that large, aggressive transitions are easier to estimate because the parameters that govern them might affect the log-likelihood more if the transition is a distinct feature of the data. However, large transitions imply large deviations from stationarity, which might make the convergence slower.

The parameter c governs the locations of the inflection point of the transition. If c is either small or large, ceteris paribus a larger part of the transition will be outside of the observed data, which of course makes it more difficult to estimate. If γ is large, this might not be a problem, as the transition will then happen rapidly enough to be mostly observed. We note that if a researcher is dealing with a dataset where it it suspected that c is close to 0 or 1 with a small γ, it might be wise to collect more data before attempting to fit the model.

Numerical optimization of GARCH-type log likelihood equations is subject to sensitivity to starting values. Many solvers can accommodate this by allowing the user to specify a sequence of starting values, or selecting the starting values based on some randomization around a vector of the user's choice. It is therefore reasonable to expect that, in applications, the user can benefit from knowledge of the data at hand and the solver's options to select starting values that come fairly close to the parameters that generated the time series. In a practical application, it might be a good idea to plot profile likelihoods and try at least a few sets of starting values before settling on a final model.

5.2 Simulation study

We conduct a simulation study to investigate the small sample properties of the model and support the results of the asymptotic analysis in the previous section. Using the statistical programming language `R`, we simulate time series of lengths $T = 1000$, $T = 2500$, $T = 5000$, fit the ATV-GARCH model and use the (numerically computed) Hessian to calculate standard errors and t-statistics for the parameter estimates. We use 2000 Monte Carlo repetitions and the standard random number generator with the seed set to 8493.

In estimating the model, we have used the solver `solnp` from the R package `Rsolnp`. As we are estimating a non-linear function in the conditional variance of the process, that is to say a numerically difficult optimization problem, some of the results might not be due to the asymptotic properties in the theory, but rather due to the solver converging to a local maxima. This shortcoming is likely less prominent for large values of T, but serves as a sober reminder to applied researchers that relying on the "canned" solution to a difficult log-likelihood problem might cause sub-optimal outcomes. To make sure that this effect is as small as possible, we have supplied the solver with the "correct" starting values, as well as adjusted the default tolerance downwards.

We consider three data generating processes (DGPs). DGP 1 features a rapid and relatively large transition. Because the transition is a pronounced part of the process, we expect it to be well estimated in small samples. However, it makes the process highly non-stationary, so the convergence of the GARCH parameters might be slow. DGP 3 is a slow and small transition, so we expect the opposite effect. DPG 2 is a middle-ground.

DGP 1 is given by

$$X_{t,T} = \sigma_{t,T}\varepsilon_t,$$

where ε_t is $NID(0,1)$ for all t,

$$\sigma^2_{t,T} = g(t/T) + 0.1X^2_{t-i,T} + 0.8\sigma^2_{t-j,T}$$

with

$$g(t/T) = 0.05 + 0.15G_1\left(\frac{t}{T}, 20, 0.5\right).$$

Figure 3 illustrates a typical realization of DGP 1.

DGP 2 is given by

$$X_{t,T} = \sigma_{t,T}\varepsilon_t,$$

where ε_t is $NID(0,1)$ for all t,

$$\sigma^2_{t,T} = g(t/T) + 0.1X^2_{t-i,T} + 0.8\sigma^2_{t-j,T}$$

with

$$g(t/T) = 0.05 + 0.05G_1\left(\frac{t}{T}, 20, 0.5\right).$$

Figure 4 illustrates a typical realization of DGP 2.

Figure 3: A typical realization of DGP 1.

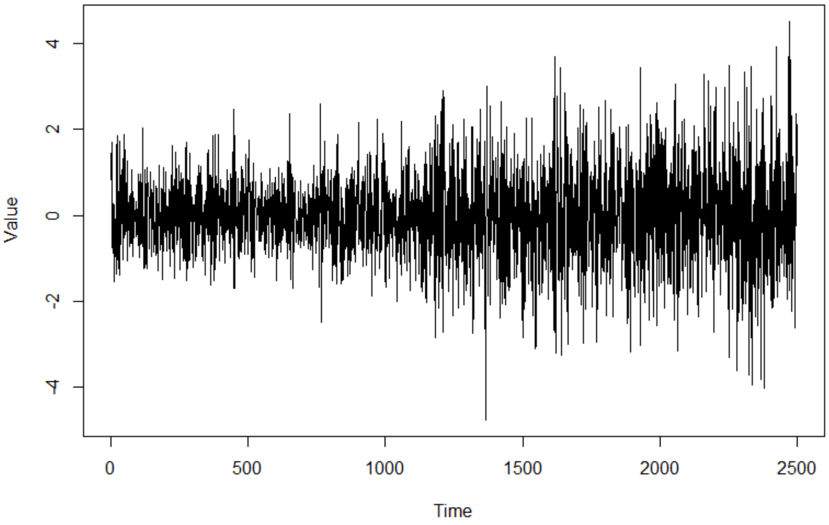
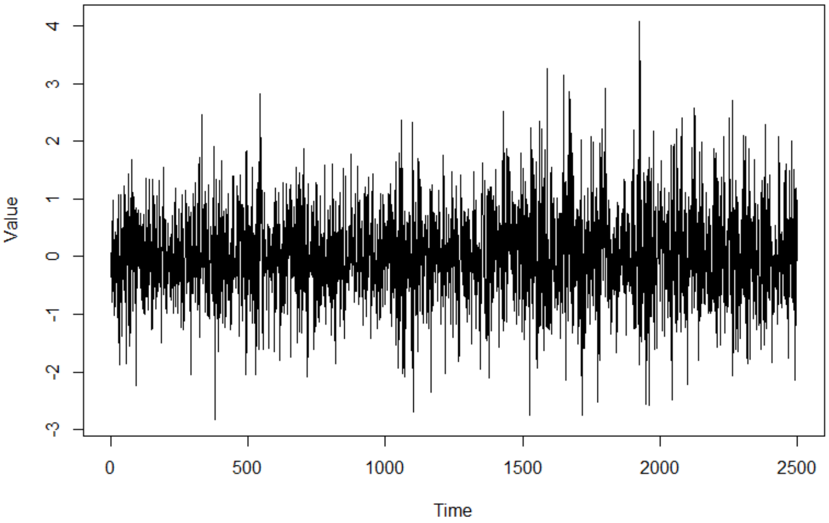


Figure 4: A typical realization of DGP 2.



DGP 3 is given by

$$X_{t,T} = \sigma_{t,T}\varepsilon_t,$$

where ε_t is $NID(0,1)$ for all t,

$$\sigma^2_{t,T} = g(t/T) + 0.1X^2_{t-i,T} + 0.8\sigma^2_{t-j,T}$$

with

$$g(t/T) = 0.05 + 0.05G_1\left(\frac{t}{T}, 10, 0.5\right).$$

Figure 5 illustrates a typical realization of DGP 3.

To avoid initialization issues, we use a burn-in period of 500 observations from the corresponding GARCH process with the transition set to zero.

The results of the study are reported in Table 1 and illustrated by the figures in the Appendix. Table 1 shows that as the number of time series observations increase, the mean of the estimated parameters tend to their true values. For $T = 5000$, they are reasonably close to the values that generates the process.

We find that the standardized parameter estimates, or t-statistics, converge toward their limiting distributions. This effect is illustrated by the figures in the Appendix. For the smallest sample size $T = 1000$, we observe some rather significant deviations. The purpose of the model is to capture behaviour that happens over long periods of time, so empirically this shold not be an issue. As T increases to 2500, the estimates and the shape of the distribution improves quite dramatically. Using for example daily return data, the common assumption of 252 trading days in a year suggests that 10 years of data would yield roughly 2500 observations.

The empirical distributions have some probability mass distributed from the shoulders into their centre and tails. This effect is less pronounced for the GARCH parameters than the nonlinear parameters in the g function. The empirical distributions become closer to the standard normal distribution as T increases.

We mention that we have noticed that for DGPs with small values of γ_1 and α_{01} (slow, small transitions) and a few number of observations, there might be problems with the invertibility of the Hessian. It also seems that for small transitions, the normal distribution approximation of the G parameters needs many observations to be cogent. We conjecture that as a particularly slow transition is difficult to distinguish from no transition, especially if the transition is small, some parameters appear roughly superfluous over short periods of time in some realizations, which might lead to approximate linear dependence in the column space of the Hessian. Further, the normal approximation tends to be sub-optimal because the G parameters skew right. This is likely because the model is unidentified if one or all of the G parameters are zero, so they err on the side of a rapid, aggressive transition. Especially in DGP 3, where the transition is slow and small, we see a clear tendency to overestimate rather than to underestimate: the convergence in Table 1 is approaching from the right rather than from the left for the α_{01} parameter, and the corresponding distributions skew right. This tendency to overestimate in turn causes β_1 to skew left: as β_1 determines how fast the effect of past values of the time-varying intercept decays, an overestimation of α_{01} leads to an underestimation of β_1.

Table 1: True values and empirical means of fitted coefficients in the simulation study.

Parameter	$\hat{\alpha}_0$	$\hat{\alpha}_1$	$\hat{\beta}_1$	$\hat{\gamma}_1$	$\hat{c}_1$	$\hat{\alpha}_{01}$
Panel A: DGP 1						
True value	**0.05**	**0.1**	**0.8**	**20**	**0.5**	**0.15**
Mean 1000	0.073	0.102	0.748	20.459	0.509	0.24
Mean 2500	0.058	0.1	0.781	20.598	0.5	0.178
Mean 5000	0.054	0.1	0.791	20.413	0.5	0.163
Panel B: DGP 2						
True value	**0.05**	**0.1**	**0.8**	**20**	**0.5**	**0.05**
Mean 1000	0.071	0.102	0.748	22.365	0.514	0.091
Mean 2500	0.057	0.101	0.781	18.380	0.509	0.066
Mean 5000	0.054	0.1	0.791	20.142	0.503	0.056
Panel C: DGP 3						
True value	**0.05**	**0.1**	**0.8**	**10**	**0.5**	**0.05**
Mean 1000	0.07	0.102	0.747	12.175	0.534	0.104
Mean 2500	0.055	0.101	0.781	9.707	0.528	0.075
Mean 5000	0.052	0.1	0.791	9.811	0.515	0.062

We used 2000 Monte Carlo repetitions to calculate the means.

We have rounded the results to three significant digits.

6 Empirical application

We motivate the use of the model by a short empirical example. We consider Intel corporation (ticker INTC) returns during the period $2000 - 2008$. The return and price time series are plotted in Figure 6. The period starts with a high degree of turbulence, possibly due to the tech bubble. After a period of relative calm towards the middle of the time series, the volatility seems to pick up as the financial crisis approaches. Summary statistics of the returns can be found in the first panel of Table 2. The returns have been multiplied by 10 in order to make the intercept larger in magnitude, which helps numerically in the optimization routine and with ease of reporting of coefficient estimates.

We begin by testing for additive misspecification using a Lagrange mutliplier test and a sequential procedure; we fit the model that is correct under the null hypothesis, initially a traditional GARCH(1,1) model, and calculate the corresponding test statistic. If the test rejects, we fit an ATV-GARCH(1,1) model with one transition and continue to test for misspecification. We add transition functions until the test statistic no longer exceeds its critical value. The asymptotic distribution of the test statistic under the null hypothesis is chi-square with three degrees of freedom. The critical value is calculated using the 5% level of significance, which corresponds to a value of $\chi^2(3) \approx 7.815$. We assume that the errors are normally distributed, and account for distributional misspecification by reporting a robust version of the test statistic. The test is going to be expanded on in a subsequent article.

The results can be found in the second panel of Table 2. The test rejects the null hypothesis of a standard GARCH$(1,1)$ model ($\chi^2(3) = 11.838$, Robust $\chi^2(3) = 12.332$). Moreover, the sum (0.994) of the $\hat{\alpha}_1$ (0.066) and $\hat{\beta}_1$ (0.928) coefficients is close to unity, indicating a very high level of persistence.

After fitting a model with one transition function, the test no longer rejects the null hypothesis of no remaining misspecification ($\chi^2(3) = 7.077$, Robust $\chi^2(3) = 6.227$). The sum of the GARCH coefficients (0.963) is reduced. We choose an ATV-GARCH(1,1) as our final model. Coefficients estimates, t-statistics and their corresponding p-values can be found in the last panel of Table 2.

Table 2: Summary of the empirical application

Statistics							
Series	Mean	Sd	Med	Min	Max	Skew	Kurt
INTC	-0.005	0.296	0.000	-2.489	1.833	-0.474	9.435
Testing							
Null hyp.	$\chi^2(3)$	R $\chi^2(3)$	p-val	R p-val	$\widehat{\alpha}_1$	$\widehat{\beta}_1$	$\widehat{\alpha}_1 + \widehat{\beta}_1$
0 tr.	11.838	12.332	0.008***	0.006***	0.066	0.928	0.994
1 tr.	7.077	6.227	0.069*	0.101	0.062	0.901	0.963
Final model							
Coefficient	Value	t-stat	R t-stat	p-val	R p-val		
$\widehat{\alpha}_{00}$	0.006	3.812	2.633	0.000***	0.008***		
$\widehat{\alpha}_1$	0.062	6.074	4.301	0.000***	0.000***		
$\widehat{\beta}_1$	0.901	54.365	44.174	0.000***	0.000***		
$\widehat{\gamma}_1$	55.034	1.799	2.206	0.072*	0.027**		
$\widehat{c}_1$	0.340	17.625	14.516	0.000***	0.000***		
$\widehat{\alpha}_{01}$	-0.005	-3.564	-2.423	0.000***	0.015**		

Note: All values are rounded to three decimals.
”Mean” is the average value of the series over the time period, ”SD” is the standard deviation and ”Med” is the median.
”Min” and ”Max” are the minimum and maximum values, respectively.
”Skew” is the skewness and ”Kurt” is the kurtosis.
”Null hyp.” is the null hypothesis of the misspecification test.
”0 tr.” and ”1 tr.” are the number of transitions in the model under the null.
”R” is an abbreviation for ”robust”.
***,** and * indicate significance on the 1%, 5% and 10% levels, respectively.

Figure 5: A typical realization of DGP 3.

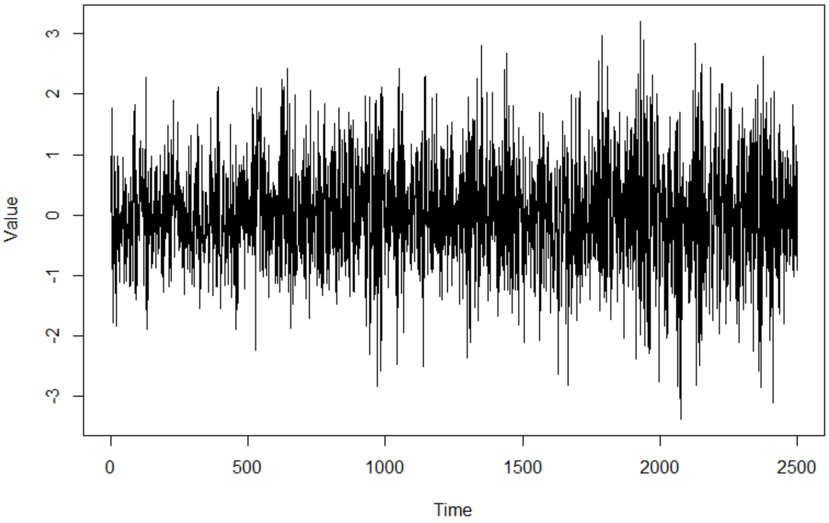
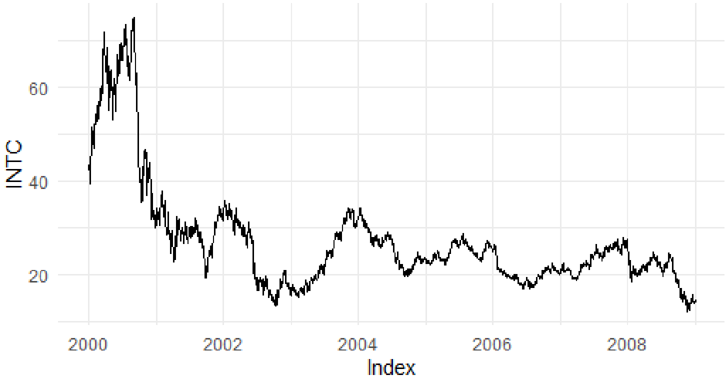
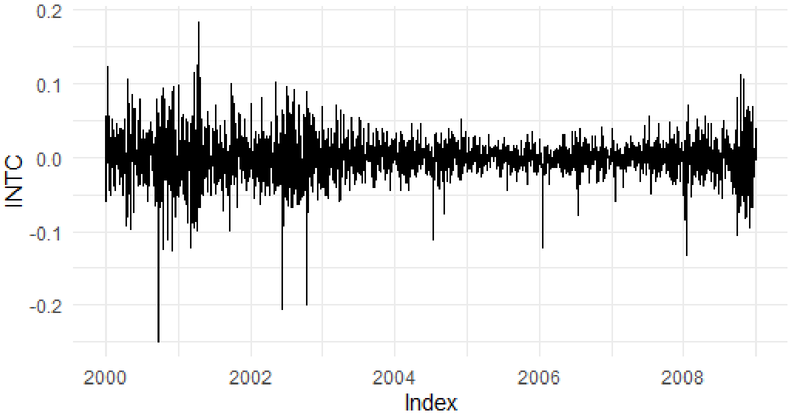


Figure 6: Intel prices and returns 2000 – 2008.

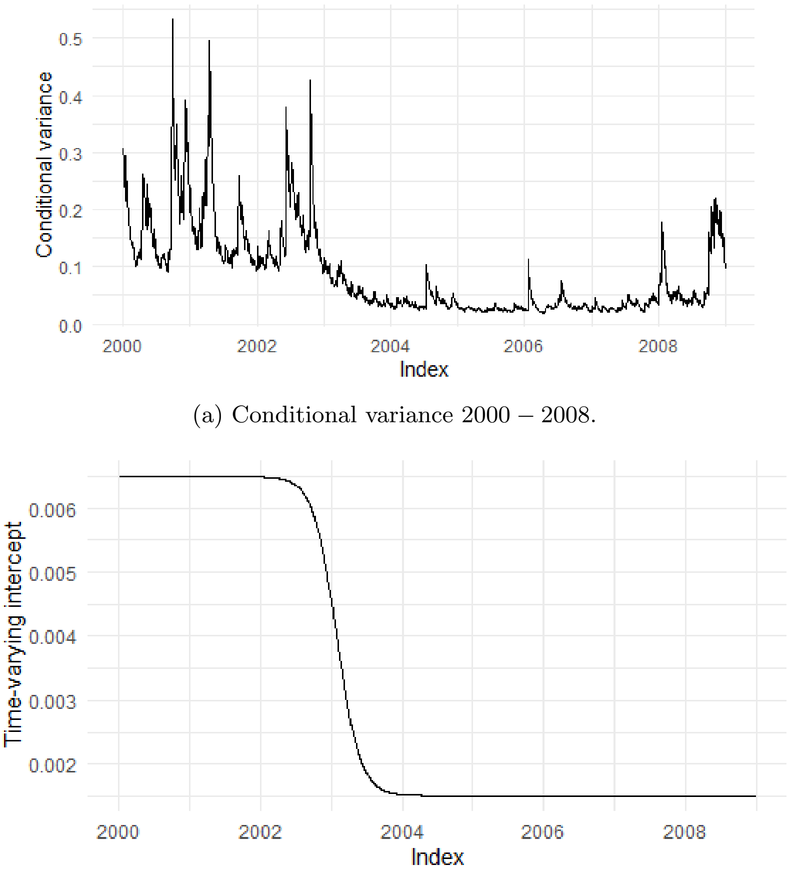


(a) Intel prices 2000 – 2008.



(b) Intel returns 2000 – 2008.

Figure 7: ATV-GARCH(1,1) with one transition function fitted to $2000-2008$ Intel returns.



(a) Conditional variance $2000-2008$.

(b) Time varying intercept $2000-2008$.

7 Conclusion

In this paper we have proposed an additive time-varying GARCH model where the intercept is allowed to have a smoothly time-varying structure. The ATV-GARCH model has a representation as a tvGARCH process. By rescaling the parameters to the unit interval, an asymptotic theory for locally stationary processes becomes available for asymptotic analysis. We use this theory applied to the ATV-GARCH model. We derive the asymptotic properties of the QMLE of the parameters of the model. We prove consistency and asymptotic normality of the QMLE. A simulation study supports the theoretical findings. We provide an illustration of the ATV-GARCH model fitted to stock returns.

Overall, the results indicate that our model provides a simple, flexible and effective way of accounting for time-variation in the intercept. One does not have to specify an exogenous, stochastic transition variable, but can use rescaled time instead. A time-varying intercept indeed reduces persistence, which is in line with previous empirical work. We conjecture that the reduction in the persistence could impact results in academic studies where a GARCH model has been used over long time horizons to obtain estimates of volatility.

Appendix: Proofs

We provide proofs of consistency and asymptotic normality of the QMLE. We use results from DRW for processes that can be locally approximated by stationary processes. Following Dahlhaus (1997), such processes are called locally stationary process. As shown by Subba Rao (2006), the time-varying GARCH process is included in the class of locally stationary processes. For consistency, we mainly need the global law of large numbers in Theorem 2.7 of DRW, which requires showing Assumption S1 in DRW is satisfied. To apply the global central limit theorem in Theorem 2.9 of DRW, we need Assumptions S1, S2 and MI in DRW. We start by stating these assumptions. We continue with a short section containing some useful Lemmata and then proceed to the proofs.

Assumptions from DRW

Our proofs require verification of Assumptions S1, S2 and M1 from DRW. We state them here.

Assumptions S1 and S2 are concerned with the existence of a stationary approximation. Let $p > 0$. Let $X_{t,T}$ be a triangular array of stochastic processes. For each $u \in [0,1]$, let $\widetilde{X}_t(u)$ be a stationary and ergodic process such that the following holds.

Assumption (DRW S1)**.** $\sup_{u\in[0,1]} \left\|\widetilde{X_t(u)}\right\|_p < \infty$. *There exists* $1 \geq \alpha > 0$, $C > 0$ *s.t. uniformly in* $t = 1, \dots, T$ *and* $u, v \in [0,1]$,

$$\left\|\widetilde{X_t}(u) - \widetilde{X_t}(v)\right\|_p \leq C\,|u-v|^{\alpha},$$

$$\left\|X_{t,T} - \widetilde{X}_t(t/T)\right\|_p \leq CT^{-\alpha}.$$

Assumption (DRW S2)**.** $u \mapsto \widetilde{X}_t(u)$ *is almost surely continuous for all* $t \in \mathbb{Z}$ *and* $\left\|\sup_{u\in[0,1]}\left|\widetilde{X_t}(u)\right|\right\|_p < \infty$.

Assumption M1 is a mixing condition on the stationary approximation $\widetilde{X}_t(u)$. Let ε_t, $t \in \mathbb{Z}$, be a sequence of independent and identically distributed random variables. For $t \geq 0$, define the σ-fields

$$\mathcal{F}_t = \sigma(\dots, \varepsilon_{t-1}, \varepsilon_t)$$

and

$$\mathcal{F}_t^{e0} = \sigma(\dots, \varepsilon_{-1}, \varepsilon_0^e, \varepsilon_1, \dots, \varepsilon_{t-1}, \varepsilon_t),$$

where ε_0^e has the same distribution as ε_1 and is independent of all ε_t, $t \in \mathbb{Z}$. The (uniform) functional dependence measure for the stationary process $\widetilde{X}_t$ is defined as (see Wu 2011, p. 2)

$$\delta_{0,p}(t) = \left\|\widetilde{X}_t - \widetilde{X}_t^e\right\|_p,$$

where $\widetilde{X}_t^e$ is a coupled version of $\widetilde{X}_t$ with ε_0 in the latter being replaced by ε_0^e. The process is said to be p-stable if (see Wu 2011, p. 3)

$$\sum_{t=1}^{\infty} \delta_{0,p}(t) < \infty.$$

Assumption (DRW M1)**.** *For each $u \in [0,1]$ there exists a measurable function $H(u,\cdot)$ s.t. $\widetilde{X}_t(u) = H(u,\mathcal{F}_t)$ and $\delta_{0,p}^{\widetilde{X}}(t) := \sup_{u\in[0,1]} \delta_{0,p}^{\widetilde{X}(u)}(t)$ fulfills $\sum_{t=0}^{\infty} \delta_{0,p}^{\widetilde{X}}(t) < \infty$.*

We continue by establishing some intermediate results.

Lemmata

Lemma 1. *The logistic function in (4) is Lipschitz continuous in $t/T \in [0,1]$, and its first three partial derivatives with respect to the parameters c and γ, $|c|, |\gamma| < \infty$, are bounded and Lipschitz in t/T.*

Proof. The first partial derivatives are

$$\begin{aligned}
\frac{\partial G}{\partial \gamma} &= (t/T - c)G(1-G),\\
\frac{\partial G}{\partial c} &= -\gamma G(1-G).
\end{aligned}$$

The second partial derivatives are

$$\begin{aligned}
\frac{\partial^2 G}{\partial \gamma^2} &= \frac{\partial}{\partial \gamma}\left\{(t/T-c)G(1-G)\right\} = (t/T-c)^2 G(1-G)(1-2G),\\
\frac{\partial^2 G}{\partial \gamma \partial c} &= \frac{\partial}{\partial c}\left\{(t/T-c)G(1-G)\right\} = (t/T-c)(-\gamma)G(1-G)(1-2G) - G(1-G),\\
\frac{\partial^2 G}{\partial c^2} &= \frac{\partial}{\partial c}\left\{-\gamma G(1-G)\right\} = (-\gamma)^2 G(1-G)(1-2G).
\end{aligned}$$

The third partial derivatives are

$$\begin{aligned}
\frac{\partial^3 G}{\partial \gamma^3} &= \frac{\partial}{\partial \gamma}\left\{(t/T-c)^2 G(1-G)(1-2G)\right\} = (t/T-c)^3 G(1-G)(1-6G+6G^2),\\
\frac{\partial^3 G}{\partial \gamma^2 \partial c} &= \frac{\partial}{\partial c}\left\{(t/T-c)^2 G(1-G)(1-2G)\right\} = (t/T-c)^2(-\gamma)G(1-G)(1-6G+6G^2)\\
&\quad - 2(t/T-c)G(1-G)(1-2G),\\
\frac{\partial^3 G}{\partial \gamma \partial c^2} &= \frac{\partial}{\partial \gamma}\left\{(-\gamma)^2 G(1-G)(1-2G)\right\} = (t/T-c)(-\gamma)^2 G(1-G)(1-6G+6G^2)\\
&\quad - 2\gamma G(1-G)(1-2G),\\
\frac{\partial^3 G}{\partial c^3} &= \frac{\partial}{\partial c}\left\{(-\gamma)^2 G(1-G)(1-2G)\right\} = (-\gamma)^3 G(1-G)(1-6G+6G^2).
\end{aligned}$$

Since $G \in [0,1]$, these derivatives are all bounded. The partial derivative of G with respect to t/T is given by

$$\frac{\partial G(t/T,\theta)}{\partial(t/T)} = \gamma G(t/T,\theta)\left(1 - G(t/T,\theta)\right),$$

which achieves its maximum $\frac{\gamma}{4}$ when $G(t/T) = 1/2$. Therefore, by the mean value theorem, for any points $a, b \in [0,1]$,

$$G(a,\theta) - G(b,\theta) \leq \frac{\gamma}{4}(a-b).$$

The derivatives consist of products of Lipschitz continuous, bounded functions, so they are Lipschitz. □

For later use, it will be useful to consider a stationary stochastic process that bounds the locally stationary process $\left\{X_{t,T}^2\right\}$. To do so, define

$$\sigma_t^{*2} = g^* + \sum_{i=1}^{p} \alpha_i X_{t-i}^{*2} + \sum_{j=1}^{q} \beta_j \sigma_{t-j}^{*2}, \tag{31}$$

where $g^* := \sup_{u\in[0,1]} g(u)$. This is a finite value as the parameter space is compact and (4) is bounded from above by 1. Note that by definition, $\sigma_{t,T}^{*2} \geq \sigma_{t,T}^2$. The process $\sigma_{t,T}^{*2}$ is the variance equation of a standard, stationary GARCH(p,q) process with an intercept that bounds some time-varying deterministic function. Denote the corresponding process by $X_{t,T}^{*2}$.

The following intermediate result is often required in our subsequent derivations. It is a special case of Subba Rao (2006), Theorem 2.1 and Proposition 2.1.

Lemma 2. *Under (A1)-(A6), for $p = 1, 2$, uniformly in t and $u, v \in [0,1]$,*

$$\left\| X_{t,T}^2 - \widetilde{X}_t^2(u) \right\|_p \leq \frac{C_1}{T} + C_2 \left| t/T - u \right|, \tag{32}$$

$$\left\| X_{t,T}^2 - \widetilde{X}_t^2(t/T) \right\|_p = \frac{C_1}{T}, \tag{33}$$

and

$$\left\| \widetilde{X}_t^2(u) - \widetilde{X}_t^2(v) \right\|_p \leq C_3 \left| u - v \right|. \tag{34}$$

It also holds that

$$\left\| h_t(\theta) - \widetilde{h}_t(u,\theta) \right\|_p \leq \frac{C_4}{T} + C_5 \left| t/T - u \right|, \tag{35}$$

$$\left\| h_t(\theta) - \widetilde{h}_t(t/T,\theta) \right\|_p \leq \frac{C_4}{T} \tag{36}$$

and

$$\left\| \widetilde{h}_t(u,\theta) - \widetilde{h}_t(v,\theta) \right\|_p \leq C_5 \left| u - v \right|. \tag{37}$$

Proof. It follows from Subba Rao (2006, Section 5.2) that for a TV-GARCH(p,q) process to be locally stationary, it is sufficient that the parameter curves $\alpha_i(u), i = 0,\dots,p$, and $\beta_j(u), j = 1,\dots,q$, are Lipschitz continuous, and that

$$\sup_u \left(\sum_{i=1}^{p} \alpha_i(u) + \sum_{j=1}^{q} \beta_j(u) \right) < 1 - \eta, \tag{38}$$

for some $\eta > 0$ and $u \in [0,1]$. We have effectively assumed (38) in (A4). It remains to show that the time-varying intercept is Lipschitz continuous. We need to verify that

$$|\alpha_0(u) - \alpha_0(v)| \leq K\,|u - v|\,, \tag{39}$$

for some constant K. It follows from the mean value theorem that it is sufficient to show that the curve $\alpha_0(u)$ has bounded first derivatives. Therefore, our Lemma 1, in conjunction with compactness of the parameter space, shows that the curve is Lipschitz continuous. By the discussion in Subsection 1.3 and assumption (A4), the first three results now follow from Subba Rao (2006), Theorem 2.1 and Proposition 2.1. For (35), note that BHK, Lemma 3.1 gives

$$c_i(\theta) \leq C\rho_0^{i/q}$$

for some $0 < \rho_0 < 1$. We have that

$$h_t(\theta) = c_0(\theta) + \sum_{i=1}^{\infty} c_i(\theta) X_{t-i}^2$$

and

$$\widetilde{h}_t(\theta) = \widetilde{c}_0(\theta) + \sum_{i=1}^{\infty} c_i(\theta) \widetilde{X}_{t-i}^2$$

Thus, we can write

$$\begin{aligned} \left| h_t(\theta) - \widetilde{h}_t(u,\theta) \right| &\leq |c_0(\theta) - \widetilde{c}_0(\theta)| \\ &\quad + \left| C \left(\sum_{i=1}^{\infty} \rho_0^{i/q} \left(X_{t-i}^2 - \widetilde{X}_{t-i}^2(u) \right) \right) \right|. \end{aligned}$$

Similarly to an argument in the proof of Kristensen and Lee (2019), Theorem 6, by an application of the triangle inequality, we get that at each index

$$\begin{aligned} \left| X_{t-i}^2 - \widetilde{X}_{t-i}^2(u) \right| &\leq \left| X_{t-i}^2 - \widetilde{X}_{t-i}^2 \left(\frac{t-i}{T} \right) \right| + \left| \widetilde{X}_{t-i}^2(t/T) - \widetilde{X}_{t-i}^2 \left(\frac{t-i}{T} \right) \right| \\ &\quad + \left| \widetilde{X}_{t-i}^2(t/T) - \widetilde{X}_{t-i}^2(u) \right|. \end{aligned}$$

Taking norms and using (32) we get

$$\left\| X_{t-i}^2 - \widetilde{X}_{t-i}^2(u) \right\|_p \le \frac{C_1}{T} + C_3 \frac{i}{T} + C_3 \left| t/T - u \right| .$$

It follows that

$$\begin{aligned} \left\| h_t(\theta) - \widetilde{h}_t(u,\theta) \right\|_p &\le \frac{C_4}{T} + C_5 \left| t/T - u \right| \\ &\quad + C \sum_{i=1}^{\infty} \rho_0^{i/q} \left(\frac{C_1}{T} + C_3 \frac{i}{T} + C_3 \left| t/T - u \right| \right) \\ &\le \frac{C_6}{T} + C_7 \left| t/T - u \right| . \end{aligned}$$

where we have used the fact that $\rho_0 < 1$, so the arithmetic-geometric series $\sum_{i=1}^{\infty} \rho_0^{i/q} i < \infty$. □

In our proofs, we shall require a result similar to BHK, Lemma 5.1. Some explanation of how this is related to our assumptions is appropriate here. The Lemma states that for a stationary, standard GARCH (p,q) process, under assumptions (A1) and (A2), which include the moment condition on the innovations,

$$\mathbb{E}|\varepsilon_0^2|^{\gamma} < \infty,$$

for some $\gamma > 0$, then for any $0 < \upsilon < \gamma$, it holds

$$\mathbb{E} \left\{ \sup_{\theta \in \Theta} \frac{\widetilde{h}_t(\theta_0)}{\widetilde{h}_t(\theta)} \right\}^{\upsilon} < \infty.$$

The lemma holds for the stationary process $\widetilde{X}_t(u)$ with variance function $\widetilde{h}_t(u,\theta)$. Moreover, the lemma continues to hold with the variance function $h_t(\theta)$ replaced by $h_t^*(\theta)$. We note that under assumption (A1), the result is true for $\upsilon = 2$. Note further that

$$\begin{aligned} \mathbb{E} \left\{ \sup_{\theta \in \Theta} \frac{X_t^2}{h_t(\theta)} \right\}^{\upsilon} &= \mathbb{E} \left\{ \varepsilon_t^2 \sup_{\theta \in \Theta} \frac{h_t(\theta_0)}{h_t(\theta)} \right\}^{\upsilon} && (40) \\ &\le \mathbb{E} \left\{ \varepsilon_t^2 \sup_{\theta \in \Theta} \frac{h_t^*(\theta_0)}{h_t(\theta)} \right\}^{\upsilon} && (41) \\ &< \mathbb{E} \left\{ \varepsilon_t^2 \sup_{\theta \in \Theta} \frac{h_t^*(\theta_0)}{\widetilde{h}_t(\theta)} \right\}^{\upsilon} < \infty, && (42) \end{aligned}$$

where the last line is true because by (A3), $\alpha_0 + \sum_{l=1}^{L} \alpha_{0l} G_l \left(u, \gamma_l, \boldsymbol{c}_l \right) > \inf_{\theta \in \Theta} \alpha_0$, so $h_t(\theta)$ is increasing in $g(t/T)$ at the lower bound of α_0. As an increasing $g(t/T)$ makes the denominator larger, the supremum will be reached when $g(t/T)$ is as small as possible close to this bound, that is, nearly constant at $\inf_{\theta \in \Theta} \alpha_0$.

In other words, the smallest stationary process is bounding the smallest locally stationary process from below. For identification reasons, we cannot allow a constant intercept. Therefore the last line is an inequality. The last line consists entirely of terms that fall under the jurisdiction of BHK Lemma 5.1, so we can continue to apply it here.

Lemma 3. *Under our assumptions, we have*

$$\mathbb{E}\sup_{\theta\in\Theta}\left|\frac{1}{h_t(\theta)}\frac{\partial h_t(\theta)}{\partial\theta}\right|^{\upsilon}<\infty, \tag{43}$$

$$\mathbb{E}\sup_{\theta\in\Theta}\left|\frac{1}{h_t(\theta)}\frac{\partial^2 h_t(\theta)}{\partial\theta\partial\theta'}\right|^{\upsilon}<\infty \tag{44}$$

and

$$\mathbb{E}\sup_{\theta\in\Theta}\left|\frac{1}{h_t(\theta)}\frac{\partial^3 h_t(\theta)}{\partial\theta_i\partial\theta_j\partial\theta_k}\right|^{\upsilon}<\infty \tag{45}$$

for any $\upsilon>0$.

Proof. As in Boussama (2000) and Francq and Zakoian (2004), we can exploit the inequality $x/(1+x)\leq x^s$ for $s\in(0,1)$. Using Lemma 1 and BHK, Lemma 3.2, we obtain

$$\begin{aligned}
\mathbb{E}\sup_{\theta\in\Theta}\left|\frac{1}{h_t(\theta)}\frac{\partial h_t(\theta)}{\partial\theta}\right| &\leq \mathbb{E}\left(\sup_{\theta\in\Theta}\frac{|\partial c_0(\theta)/\partial\theta|+\sum_{i=1}^{\infty}ic_i(\theta)X_{t-i}^2}{c_0(\theta)+\sum_{i=1}^{\infty}c_i(\theta)X_{t-i}^2}\right) && (46)\\
&\leq \mathbb{E}\left(\sup_{\theta\in\Theta}\frac{C_2}{c_0(\theta)+\sum_{i=1}^{\infty}c_i(\theta)X_{t-i}^2}\right) && (47)\\
&+\mathbb{E}\left(\sup_{\theta\in\Theta}\frac{\sum_{i=1}^{\infty}ic_i(\theta)X_{t-i}^2}{c_0(\theta)+\sum_{i=1}^{\infty}c_i(\theta)X_{t-i}^2}\right) && (48)\\
&\leq C_3+\mathbb{E}\left(\sup_{\theta\in\Theta}\sum_{i=1}^{\infty}ic_i^s(\theta)X_{t-i}^{2s}\right) && (49)\\
&\leq C_3+\mathbb{E}\left(\sup_{\theta\in\Theta}\sum_{i=1}^{\infty}i\rho_0^{is/q}X_{t-i}^{2s}\right) && (50)\\
&\leq C_3+\mathbb{E}\left(\sup_{\theta\in\Theta}\sum_{i=1}^{\infty}i\rho_0^{is/q}X_{t-i}^{2*s}\right) && (51)\\
&<\infty, && (52)
\end{aligned}$$

for all $s\in(0,1]$ by weak stationarity, and similarly for (44) and (45) using the boundedness of the derivatives in Lemma 1. Note now that for large but finite values of υ, we can pick s close to zero to offset their impact on the expectation. Hence the result. □

Motivating the feasible estimator

We reproduce some arguments from BHK in order to motivate the truncated, feasible estimator. We can use similar arguments as long as the unobserved, truncated part of the recursion is bounded by a stationary process. By the triangle inequality, we obtain

$$\sup_{\theta\in\Theta}\left|L_T(\theta)-\bar{L}_T(\theta)\right| \leq \sup_{\theta\in\Theta}\frac{1}{T}\sum_{t=1}^{T}\left|\frac{X_t^2}{h_t(\theta)}-\frac{X_t^2}{\bar{h}_t(\theta)}\right| + \sup_{\theta\in\Theta}\frac{1}{T}\sum_{t=1}^{T}\left|\ln h_t(\theta)-\ln\bar{h}_t(\theta)\right|$$

Consider the second term. We have

$$\begin{aligned}\sup_{\theta\in\Theta}\frac{1}{T}\sum_{t=1}^{T}\left|\ln h_t(\theta)-\ln\bar{h}_t(\theta)\right| &\leq \sup_{\theta\in\Theta}\frac{1}{T}\sum_{t=1}^{T}\frac{1}{C_1}\left|\sum_{i=1}^{\infty}c_i(\theta)X_{t-i}^2-\sum_{i=1}^{t-1}c_i(\theta)X_{t-i}^2\right| &(53)\\ &\leq \sup_{\theta\in\Theta}\frac{1}{T}\sum_{t=1}^{T}\left|\frac{C_2}{C_1}\sum_{i=t}^{\infty}\rho_0^{i/q}X_{t-i}^2\right| &(54)\\ &= \sup_{\theta\in\Theta}\frac{1}{T}\sum_{t=1}^{T}\left|\frac{C_2}{C_1}\rho_0^{t/q}\sum_{j=0}^{\infty}\rho_0^{j/q}X_{-j}^2\right|. &(55)\end{aligned}$$

Note that the sum in (54) starts at $t-t=0$. Under our assumptions, the term that causes the global nonstationarity is constant until $t=0$, so the difference between the infinite and the feasible estimator stems from a stationary process. The sum in (55) is always smaller than a corresponding sum of a stationary process obtained by fixing the intercept of X_j^2 at $\sup_{u<0}g(u)$. BHK Lemma 2.2 and Lemma 2.3 gives the result: the term in (55) is $o(1)$. For the first term we similarly get

$$\begin{aligned}\sup_{\theta\in\Theta}\frac{1}{T}\sum_{t=1}^{T}\left|\frac{X_t^2}{h_t(\theta)}-\frac{X_t^2}{\bar{h}_t(\theta)}\right| &= \frac{1}{T}\sum_{t=1}^{T}\sup_{\theta\in\Theta}\frac{X_t^2}{h_t(\theta)}\left|\frac{h_t(\theta)-\bar{h}_t(\theta)}{\bar{h}_t(\theta)}\right| &(56)\\ &\leq \frac{1}{T}\frac{C_2}{C_1}\sum_{j=0}^{\infty}\rho_0^{j/q}X_{-j}^2\sup_{\theta\in\Theta}\sum_{t=1}^{T}\frac{X_t^2}{h_t(\theta)}\rho_0^{t/q} &(57)\\ &\leq \frac{1}{T}\frac{C_2}{C_1}\sum_{j=0}^{\infty}\rho_0^{j/q}X_{-j}^2\sup_{\theta\in\Theta}\sum_{t=1}^{T}\frac{X_t^{*2}}{\widetilde{h}_t(\theta)}\rho_0^{t/q}, &(58)\end{aligned}$$

which is $o_p(1)$ by the same argument as above and the arguments in BHK Lemma 5.9. The last line is necessary because the term $X_t^2/h_t(\theta)$ is not stationary, so the arguments in BHK are not directly applicable. We solve this by bounding it by the stationary process $X_t^{*2}/\widetilde{h}_t(\theta)$. By similar slight adaptations of the arguments in the second parts of BHK Lemma 5.8 and 5.9, we can show that

$$\sup_{\theta\in\Theta}\left|S_t(\theta)-\bar{S}_t(\theta)\right| \tag{59}$$

is $o_p(1)$. This validates the use of the truncated, feasible estimator.

Proof of Theorem 1 (Consistency)

Since we shall need to apply a law of large numbers to the log likelihood function, namely Theorem 2.7(i) in DRW, we have to verify the assumptions (DRW Assumption S1) that are needed to do so. The invariance principle, Theorem 2.5 in the same article, details when a transformation of such a process preserves local stationarity. However, the transformation is a function with the process as the only argument, which does not immediately translate to a log likelihood function. Later in the article, the authors make a structural assumption on a log likelihood function of a recursively defined autoregressive process in order to derive the asymptotic properties of a QML estimator. It is, however, not immediately clear how to deal with transformations in the form of log likelihood functions of GARCH type models. This difficulty is remedied by Kristensen and Lee (2019, Theorem 1), who state conditions under which a transformation involving a parameter vector, an IID process and a parameter dependent locally stationary process preserves the local stationarity. Instead of relying on an invariance principle, however, we have chosen to provide a direct proof.

We use the following consistency theorem

Theorem 3 (Amemiya, 1985)**.** *If*

C1 The parameter space Θ is a compact subset of $\mathbb{R}^k$,

C2 The objective function $Q_n(\theta)$ is a measurable function of the data for all $\theta \in \Theta$, and $Q_n(\theta)$ is continuous in $\theta \in \Theta$,

C3 $Q_n(\theta)$ converges uniformly in probability to a non-stochastic function $Q_0(\theta)$, and $Q_0(\theta)$ attains a unique global maximum at θ_0,

then

$$\hat{\theta} \xrightarrow{p} \theta_0.$$

The proof will follow by considering the following argument. The condition C1 is satisfied by (A2). The log likelihood is clearly a measurable function of the data. Moreover, it is Lipschitz continuous in θ, which implies that it is continuous in θ, so C2 is satisfied. For C3, we use the law of large numbers in DRW to show that the objective function converges to its expectation in probability. Uniform convergence follows from stochastic equicontinuity. As in BHK and Francq and Zakoïan (2004) and others, since

$$L(u, \theta_0) - L(u, \theta) = -\frac{1}{2} + \frac{1}{2}\mathbb{E}\left(\frac{\widetilde{h}_t(\theta_0)}{\widetilde{h}_t(\theta)} - \log\frac{\widetilde{h}_t(\theta_0)}{\widetilde{h}_t(\theta)}\right),$$

we can use that $x - \log(x)$ is positive for $x > 0$ and reaches its minimum when $x = 1$ to deduce that for each u, $L(u, \theta)$ has an absolute maximum at θ_0. BHK Lemma 5.5 shows that for each u, if $\widetilde{h}_t(\theta_1) = \widetilde{h}_t(\theta_2)$, then $\theta_1 = \theta_2$. Therefore, the log likelihood function is uniquely maximized at θ_0. We now turn to our detailed argument for convergence requirement in C3.

7.1 Verifying C3

As in Kristensen and Lee (2019), we observe that the log likelihood function $l\left(X_t,\theta\right)$ can be written as $l\left(h_t(\theta),\varepsilon_t,\theta\right)$. The following Lemma shows that the log likelihood function is locally stationary and Lipschitz continuous in the parameters, from which it follows that it is also continuous in the parameters.

Lemma 4. *Under (A1-A6), uniformly in t and $u \in [0,1]$,*

$$\left\| l_t(h_t(\theta),\varepsilon_t,\theta) - l_t(\widetilde{h}_t(u,\theta),\varepsilon_t,\theta) \right\|_1 \leq \frac{C_4}{T} + C_5\left|t/T - u\right| \tag{60}$$

Proof. The condition (60) is required in order to preserve local stationarity. Write

$$\begin{aligned}\left\| l_t(h_t,\varepsilon_t,\theta) - l_t(h_t',\varepsilon_t,\theta)\right\|_1 &\leq \left\|\ln h_t - \ln h_t'\right\|_1 + \left\| \frac{X_t^2}{h_t} - \frac{X_t^{2\prime}}{h_t'} \right\|_1 \\ &:= A + B,\end{aligned}$$

say. Consider A and the following argument. Assume that y is bounded from below by some $\delta > 0$. Use the inequality $\ln(1+u) \leq u$ for all $u > -1$. Then,

$$\begin{aligned}\left|\ln x - \ln y\right| &= \left|\ln\frac{x}{y}\right| = \left|\ln\left(1 + \left(\frac{x}{y} - 1\right)\right)\right| \\ &\leq \left|\frac{x}{y} - 1\right| = \frac{1}{y}\left|x - y\right| \\ &\leq \frac{1}{C}\left|x-y\right|.\end{aligned}$$

Therefore,

$$\left\|\ln h_t - \ln h_t'\right\|_1 \leq \left\|\frac{1}{C}\left(h_t - h_t'\right)\right\|_1,$$

where we have used the fact that the conditional variance is bounded from below by some positive constant. To the best of our knowledge, Francq and Zakoian (2004) were the first ones to use this argument for a GARCH log likelihood.

Now consider B. As in the proof of Chen and Hong (2016), Theorem A.2, by adding and subtracting $X_t'^2/h_t$ and using the triangle inequality, we obtain

$$\left\|\frac{X_t^2}{h_t} - \frac{X_t'^2}{h_t'}\right\|_1 \leq \left\|\frac{1}{h_t}\left(X_t^2 - X_t'^2\right)\right\|_1 + \left\|\frac{X_t'^2}{h_t h_t'}\left(h_t - h_t'\right)\right\|_1.$$

Now, let $h_t = h_t(\theta)$ and $h_t' = \widetilde{h}_t(u,\theta)$. We obtain

$$\left\|\ln h_t(\theta) - \ln\widetilde{h}_t(u,\theta)\right\|_1 \leq \left\|\frac{1}{C}\left(h_t(\theta) - \widetilde{h}_t(u,\theta)\right)\right\|_1, \tag{61}$$

and

$$\left\|\frac{X_t^2}{h_t(\theta)} - \frac{\widetilde{X}_t^2(u)}{\widetilde{h}_t(u,\theta)}\right\|_1 \leq \left\|\frac{1}{h_t(\theta)}\left(X_t^2 - \widetilde{X}_t^2(u)\right)\right\|_1 + \left\|\frac{\widetilde{X}_t^2(u)}{\widetilde{h}_t(u,\theta)}\frac{\left(h_t(\theta) - \widetilde{h}_t(u,\theta)\right)}{h_t(\theta)}\right\|_1. \tag{62}$$

By BHK, Lemma 3.1 and Lemma 5.1, and (A6)

$$\mathbb{E}\left(\sup_{\theta\in\Theta}\frac{\widetilde{X}_t^2(u)}{\widetilde{h}_t(u,\theta)}\right)^2 < \infty, \tag{63}$$

and by boundedness from below and positivity of the conditional variance, we have

$$\frac{1}{h_t(\theta)} \leq \frac{1}{\inf_{u\in[0,1]} g(u)} < \infty. \tag{64}$$

An application of the Cauchy-Schwarz inequality yields

$$\left\|\frac{X_t^2}{h_t(\theta)} - \frac{\widetilde{X}_t^2(u)}{\widetilde{h}_t(u,\theta)}\right\|_1 \leq \left\|\frac{1}{C_1}\left(X_t^2 - \widetilde{X}_t^2(u)\right)\right\|_1 + \left\|\frac{\widetilde{X}_t^2(u)}{\widetilde{h}_t(u,\theta)}\right\|_2 \left\|\frac{\left(h_t(\theta) - \widetilde{h}_t(u,\theta)\right)}{h_t(\theta)}\right\|_2. \tag{65}$$

Since the argument remains valid if we in the first step instead add and subtract X_t^2/h_t', we can without loss of generality assume that $h_t(\theta) \geq \widetilde{h}_t(u,\theta)$. Then, since $\left(h_t(\theta) - \widetilde{h}_t(u,\theta)\right)/h_t(\theta) \in (0,1]$, we obtain

$$\frac{h_t(\theta) - \widetilde{h}_t(u,\theta)}{h_t(\theta)} \leq \left(\frac{h_t(\theta) - \widetilde{h}_t(u,\theta)}{h_t(\theta)}\right)^{1/2}.$$

The result follows by using (63) and (64) and applying Lemma 3 to (61) and (65). □

Lemma 5. *Under (A1-A6)* $\sup_{u\in[0,1]}\|l(u,\theta)\|_1 < \infty$.

Proof. As this is the log likelihood of the stationary approximation, the result follows for each u from BHK, Lemma 5.3, and therefore for the supremum in particular. □

Lemmas 5 and 6 enables us to apply DRW, Theorem 2.7(i) to the log likelihood function and we obtain pointwise

$$L_T(\theta) \overset{P}{\to} \int_0^1 L(u,\theta)\,\mathrm{d}u = \mathbb{E}(L_T(\theta)). \tag{66}$$

To complete the verification of C3, we need the convergence to be uniform. A sufficient condition for uniform convergence is a compact parameter space, pointwise convergence and stochastic equicontinuity (SE) of the sequence of estimators (see e.g. Andrews (1992)). Therefore, we consider the following Lemma.

Lemma 6. *The function $L_T(\theta) := \frac{1}{T}\sum_{t=1}^{T} l_t(\theta)$ is stochastically equicontinuous.*

Proof. We make use of the sufficient conditions for SE in Andrews (1992), Lemma 2. In order to proceed as in BHK, we would have to argue that

$$\sup_{\theta\in\Theta} |L_T(\theta_1) - \text{Ł}_T(\theta_2)| \leq C\,|\theta_1 - \theta_2|$$

almost surely. In BHK, this was done by an application of the ergodic theorem. Since we consider a process that is only locally stationary, rather than ergodic stationary, this is not viable.

Consider instead the following argument. Similarly to BHK, Lemma 5.3, by an application of the mean value theorem we get

$$|l_t(\theta_1) - l_t(\theta_2)| \leq K_t|\theta_1 - \theta_2|,$$

where

$$K_t = \left|\frac{1}{h_t(\theta^m)}\frac{\partial h_t(\theta^m)}{\partial\theta}\left(1 + \frac{X_t^2}{h_t(\theta^m)}\right)\right|, \tag{67}$$

where θ^m lies between θ_1 and θ_2.

Now we could proceed by arguing that the terms in K_t satisfy S1 and S2 in DRW, and apply the law of large numbers from the same article to obtain $\frac{1}{T}\sum_{t=1}^{T} K_t \overset{P}{\to} K < \infty$. This turns out to be more arduous than worthwhile. In BHK, the authors set out to prove strong consistency, i.e. almost sure convergence of the sequence of estimators. This would require that

$$\frac{1}{t}\sum_{t\geq 1} K_t = O(1) \text{ almost surely,}$$

which in Andrews (1992) is a condition required for *strong* SE. The limit theorem from DRW only holds in L_1, however, and in this article we are only interested in weak consistency, i.e. convergence *in probability* of the sequence of estimators. This means that we only require *weak* SE. It follows from Andrews (1992), Lemma 2 (a) that in order to obtain weak SE of our sequence of estimators, it is enough to show

$$\sup_{t\geq 1}\frac{1}{t}\sum_{t\geq 1}\mathbb{E}(K_t) < \infty. \tag{68}$$

Now it is easy to see that this is fulfilled, since Cauchy-Schwarz inequality yields

$$\left\|\frac{X_t^2}{h_t(\theta^m)}\frac{1}{h_t(u,\theta^m)}\frac{\partial h_t(u,\theta^m)}{\partial\theta}\right\|_1 \leq \left\|\frac{X_t^2}{h_t(\theta^m)}\right\|_2\left\|\frac{1}{h_t(u,\theta^m)}\frac{\partial h_t(u,\theta^m)}{\partial\theta}\right\|_2 < \infty$$

by (63) and (43). Now, by Amemiya (1985) Theorem 4.1.1, we obtain the result

$$\hat{\theta} \overset{P}{\to} \theta_0.$$

□

This completes the proof.

Proof of Theorem 2 (Asymptotic normality)

We shall prove the results for the estimator (11) based on the infinite representation of the process, but as discussed in the beginning of the Appendix, the results also hold for the truncated, feasible estimator (16).

Proving asymptotic normality of the QMLE can be done in a variety of ways. The decision to be made when approaching the proof is whether one wants to rely on a third order Taylor expansion of the log likelihood (examples with GARCH include Comte and Lieberman (2003) and Francq and Zakoïan (2004)) or on a second order expansion (see BHK). Here, we present our results using a third order expansion.

7.1.1 Sufficient results for asymptotic normality

Similarly to Comte and Lieberman (2003), we rely on the sufficient conditions for normality of the MLE in Basawa et al. (1976). These are

(N1)

$$\frac{1}{\sqrt{T}}\sum_{t=1}^{T}\frac{\partial l_t\left(\theta_0\right)}{\partial\theta}\overset{D}{\to}N\left(0,A\right)$$

when $T\to\infty$ for a nonrandom A.

(N2)

$$-\frac{1}{T}\sum_{t=1}^{T}\frac{\partial^2 l_t\left(\theta_0\right)}{\partial\theta\partial\theta^T}\overset{P}{\to}B$$

as $T\to\infty$ for a nonrandom positive-definite matrix B.

(N3)

$$\mathbb{E}\left(\sup_{\|\theta-\theta_0\|\le\delta}\left|\frac{\partial^3 l_t(\theta)}{\partial\theta_i\partial\theta_j\partial\theta_k}\right|\right)$$

is bounded for all i,j,k and all $\delta>0$. We continue by verifying the conditions.

7.1.2 Verifying N1

The score evaluated at θ_0 is

$$S(X_t,\theta_0)=(1-\varepsilon_t^2)\left(\frac{1}{h_t(\theta_0)}\frac{\partial h_t(\theta_0)}{\partial\theta}\right) \tag{69}$$

We want to apply the global central limit theorem from DRW. In order to do so, we shall need to verify that the assumptions S1, S2 and M1 hold with $p=2$. The first term in (69) is independent of the second term, so in order to show S1 we only need to consider differences of the type

$$\left\|\frac{1}{h_t(\theta)}\frac{\partial h_t(\theta)}{\partial\theta}-\frac{1}{h_t(\theta)'}\frac{\partial h_t(\theta)'}{\partial\theta}\right\|_2 \tag{70}$$

Consider for a moment only the derivative of $h_t(\theta)$ with respect to the parameters. By BHK, Lemma 3.2 we can write

$$\left\|\frac{\partial h_t(\theta)}{\partial\theta} - \frac{\partial h_t(\theta)'}{\partial\theta}\right\|_2 \leq \left\|\frac{\partial\alpha_0(t/T,\theta)}{\partial\theta} - \frac{\partial\alpha_0(t/T,\theta)'}{\partial\theta}\right\|_2 + \left\|C\sum_{i=1}^{\infty} i\rho_0^{i/q}\left(X_{t-i}^2 - X_{t-i}^{2\prime}\right)\right\|_2. \tag{71}$$

for a generic $h_t(\theta)' \neq h_t(\theta)$. When replacing

$$\frac{\partial\alpha_0(t/T,\theta)'}{\partial\theta}$$

by the stationary approximation, the first term goes to zero as $T \to \infty$ by Lipschitz continuity of the derivatives of the logistic transition function. The infinite sum

$$\sum_{i=1}^{\infty} i\rho_0^{i/q}$$

is an arithmetic-geometric series and converges. Therefore, by Lemma 3, we can write

$$\left\|\frac{\partial h_t}{\partial\theta} - \frac{\partial h_t(u)}{\partial\theta}\right\|_2 \leq C\left(\frac{1}{T} + |t/T - u|\right). \tag{72}$$

By adding and subtracting

$$\frac{1}{h_t(\theta)}\frac{\partial h_t'}{\partial\theta},$$

rearranging terms and using the triangle inequality, we can write

$$\begin{aligned}\left|\frac{1}{h_t(\theta_0)}\frac{\partial h_t}{\partial\theta} - \frac{1}{h_t(\theta_0)'}\frac{\partial h_t'}{\partial\theta}\right| &\leq \left|\frac{1}{h_t(\theta_0)}\right|\left|\frac{\partial h_t}{\partial\theta} - \frac{\partial h_t'}{\partial\theta}\right| \\ &\quad + \left|\frac{1}{h_t(\theta_0)'}\frac{\partial h_t(\theta)'}{\partial\theta}\right|\left|\frac{h_t(\theta_0) - h_t(\theta_0)'}{h_t(\theta_0)}\right| \\ &\leq \left|\frac{1}{C_1}\right|\left|\frac{\partial h_t}{\partial\theta} - \frac{\partial h_t'}{\partial\theta}\right| \\ &\quad + \left|\frac{1}{h_t(\theta_0)'}\frac{\partial h_t(\theta_0)'}{\partial\theta}\right|\left|\frac{h_t(\theta_0) - h_t(\theta_0)'}{h_t(\theta_0)}\right|,\end{aligned}$$

where we have emphasized dependence on θ_0 where it is important. Note that, by (43),

$$\|D\|_p := \left\|\frac{1}{h_t(\theta_0)'}\frac{\partial h_t(\theta_0)'}{\partial\theta}\right\|_p \leq \infty$$

for *any* p. Since the argument is still valid if we instead add and subtract

$$\frac{1}{h_t(\theta)'}\frac{\partial h_t}{\partial\theta},$$

we can without loss of generality assume that $h_t(\theta_0) \geq h_t(\theta_0)'$, so for $s \in (0,1)$,

$$\left\| D\left(\frac{h_t(\theta_0) - h_t(\theta_0)'}{h_t(\theta_0)}\right)\right\|_2 \leq \left\| D\left(\frac{h_t(\theta_0) - h_t(\theta_0)'}{h_t(\theta_0)}\right)^s\right\|_2$$

since

$$\frac{h_t(\theta_0) - h_t(\theta_0)'}{h_t(\theta_0)} = 1 - \frac{h_t(\theta_0)'}{h_t(\theta_0)} \in [0,1).$$

Pick for example $s = 1/2$. Define

$$\widetilde{D} = \frac{1}{\widetilde{h}_t(u,\theta_0)}\frac{\partial \widetilde{h}_t(u,\theta_0)}{\partial\theta}.$$

Now the Cauchy-schwarz inequality gives

$$\begin{aligned}\left\|\widetilde{D}\left(\frac{h_t(\theta_0) - \widetilde{h}_t(u,\theta_0)}{h_t(\theta_0)}\right)\right\|_2 &\leq \left\|\widetilde{D}\right\|_4 \left\|\left(\frac{h_t(\theta_0) - \widetilde{h}_t(u,\theta_0)}{h_t(\theta_0)}\right)^{1/2}\right\|_4 \\ &\leq C_6 \left\|\left(h_t(\theta_0) - \widetilde{h}_t(u,\theta_0)\right)\right\|_2 \\ &\leq C_7\left(\frac{1}{T} + |t/T - u|\right).\end{aligned}$$

From (72) we get

$$\left\|\left(\frac{1}{C_1}\right)\left(\frac{\partial h_t}{\partial\theta} - \frac{\partial\widetilde{h}_t(u)}{\partial\theta}\right)\right\|_2 \leq C_8\left(\frac{1}{T} + |t/T - u|\right). \tag{73}$$

Now S1 is verified.

To obtain S2, consider that by independence and as BHK Lemma 5.2 gives in particular

$$\mathbb{E}\left|\sup_{u\in[0,1]}\frac{1}{\widetilde{h}_t(\theta_0)}\frac{\partial\widetilde{h}_t(u,\theta_0)}{\partial\theta}\right|^2 < \infty, \tag{74}$$

we obtain

$$\left\|\sup_{u\in[0,1]}\widetilde{S}(\widetilde{X}_t,u,\theta_0)\right\|_2 = \left\|(1-\varepsilon_t^2)\right\|_2 \left\|\sup_{u\in[0,1]}\left(\frac{1}{\widetilde{h}_t(u,\theta_0)}\frac{\partial\widetilde{h}_t(u,\theta_0)}{\partial\theta}\right)\right\|_2 \leq \infty. \tag{75}$$

This verifies S2 for $p = 2$.

Finally, we need assumption (M1) in DRW to be fulfilled with $p = 2$. The assumption entails a mixing condition on the stationary approximation of the score. Suppressing dependence on the data, u and θ_0, we specifically need

$$\sup_{u\in[0,1]}\sum_{t=0}^{\infty}\left\|\left(\widetilde{S}_t - \widetilde{S}_t^e\right)\right\|_p < \infty, \tag{76}$$

for some $p \geq 2$, where as described in Section 7, $\widetilde{S}_t^e$ is a coupled version of the stationary approximation of the score where the error term has been replaced at index $t = 0$ in the information set. As Wu (2011) writes, the condition ensures that the process "forgets the history geometrically quickly".

By BHK and previous inequalities, for $p \leq 2$,

$$\begin{align}
\sup_{u\in[0,1]} \left\| \left(\widetilde{S}_t - \widetilde{S}_t^e \right) \right\|_p &\leq C_1 \left\| \sup_{u\in[0,1]} \left| \frac{\partial \widetilde{h}_t}{\partial \theta} - \frac{\partial \widetilde{h}_t^e}{\partial \theta} \right| \right\|_p \tag{77} \\
&+ C_2 \left\| \sup_{u\in[0,1]} \left| \widetilde{h}_t - \widetilde{h}_t^e \right| \right\|_p \tag{78} \\
&\leq C_1 \sum_{i=1}^{\infty} i \rho_0^{i/q} \mathbb{E} \left| \sup_{u\in[0,1]} \left(\widetilde{X}_{t-i}^2 - \widetilde{X}_{t-i}^{2e} \right) \right|^p \tag{79} \\
&+ C_2 \sum_{i=1}^{\infty} \rho_0^{i/q} \mathbb{E} \left| \sup_{u\in[0,1]} \left(\widetilde{X}_{t-i}^2 - \widetilde{X}_{t-i}^{2e} \right) \right|^p \tag{80}
\end{align}$$

Here, the terms are very similar. It suffices to consider the first one and the second will follow similarly. Since for $x, y \geq 0$ we have $(x^2 - y^2) = (x + y)(x - y) \leq C\,|x + y|\,|x - y|$ for some $C \geq 1$, the function $f(x) = x^2$ satisfies the invariance property of proposition (2.5) in DRW with $M = 1$. The principle states that if some assumptions, among them (M1), are fulfilled for a process with $\widetilde{p} = p(M + 1)$, then the same assumptions are fulfilled for a function of the process that satisfies the invariance property. In our case, since $M = 1$ and we need $p = 2$, we have $\widetilde{p} = 4$, which translates to us needing

$$C \sum_{i=1}^{\infty} i \rho_0^{i/q} \mathbb{E} \left| \left(\widetilde{X}_{t-i}^* - \widetilde{X}_{t-i}^{*e} \right) \right|^4 < \infty. \tag{81}$$

Since the expectation in (81) involves a stationary GARCH process, Proposition 3 in Wu and Min (2005) can be invoked, and we obtain

$$C \sum_{i=1}^{\infty} i \rho_0^{i/q} \mathbb{E} \left| \left(\widetilde{X}_{t-i}^* - \widetilde{X}_{t-i}^{*e} \right) \right|^4 \leq C \sum_{i=1}^{\infty} i \rho_0^{i/q} r^{t-i}, \tag{82}$$

where $r \in (0, 1)$ for $t - i \geq 0$, $r = 0$ otherwise. Now choose $a \in (0, 1)$ such that $a > r$, $a > \rho_0$ for $t - i \geq 0$, $a = 0$ otherwise. Then

$$G := C \sum_{i=1}^{\infty} i a^{i/q + t - i} \geq C \sum_{i=1}^{\infty} i \rho_0^{i/q} r^{t-i}.$$

To illustrate, we have

$$G = \begin{cases} C \left(a^{1/q + t - 1} + \ldots + (t - 1) a^{(t-1)/q + t - (t-1)} \right), & \text{if } t > i \\ 0, & \text{o.w.} \end{cases} \tag{83}$$

Since G is increasing in i for $q > 0$ up to $i = t-1$, we have that it is less than the term at index $i = t-1$ multiplied by $(t-1)$, which yields the inequality

$$\sum_{t=0}^{\infty} G < \sum_{t=0}^{\infty} (t-1)^2 a^{(t-1)/q+1} \tag{84}$$

$$\leq C_2 + C \sum_{t=1}^{\infty} t^2 a^{t/q+1} < \infty \tag{85}$$

since $|a| < 1$. The result follows.

Now, indexing the elements of the score by $s = 1, 2, \dots,$ by an application of DRW Theorem 2.9,

$$\frac{1}{\sqrt{T}} \sum_{t=1}^{T} \frac{\partial l_t\left(\theta_0\right)}{\partial \theta_s} \overset{D}{\to} \left\{ \int_0^1 A_s(v)^{\frac{1}{2}} dW(v) \right\}, \tag{86}$$

where W is a Brownian motion and $A_s(v) = \sum_{k \in \mathbb{Z}} \operatorname{cov}(\tilde{S}_{0s}(v), \tilde{S}_{ks}(v))$. By the Ito Isometry we can conclude that

$$\mathbb{E}\left\{ \int_0^1 A_s(v)^{\frac{1}{2}} \mathrm{d}W(v) \right\} = 0$$

and

$$\operatorname{Var}\left\{ \int_0^1 A_s(v)^{\frac{1}{2}} \mathrm{d}W(v) \right\} = \int_0^1 \mathbb{E}A_s(v) \mathrm{d}v := A_s.$$

It is well known that integrals of the form of the distributional limit in (86) follow a normal distribution. By an application of the Cramér-Wold device, we can write

$$\frac{1}{\sqrt{T}} \sum_{t=1}^{T} \frac{\partial l_t\left(\theta_0\right)}{\partial \theta} \overset{D}{\to} N(0, A). \tag{87}$$

We emphasize that the matrix A is nonrandom. Therefore, N1 is fulfilled.

7.1.3 Verifying N2

We continue by verifying N2.

Lemma 7. *Under (A1-A7), the Hessian evaluated at θ_0 fulfills the assumptions of DRW, Theorem 2.7(i).*

Proof. The Hessian at θ_0 is

$$H_t(X_t, \theta_0) = \left(1 - \varepsilon_t^2\right) \frac{1}{h_t(\theta_0)} \frac{\partial^2 h_t(\theta_0)}{\partial \theta \partial \theta^T} \tag{88}$$

$$+ \left(2\varepsilon_t^2 - 1\right) \frac{1}{h_t(\theta_0)} \frac{\partial h_t(\theta_0)}{\partial \theta} \frac{1}{h_t(\theta_0)} \frac{\partial h_t(\theta_0)}{\partial \theta^T}. \tag{89}$$

The first term has expectation 0 when evaluated at θ_0, so it suffices to consider the second term. Recall the definition D from the verification of N1 and consider (89). By adding and subtracting a suitable term, we can write

$$\begin{aligned}
\left\| DD^T - \widetilde{D}\widetilde{D}^T \right\|_1 &= \left\| D\left(D^T - \widetilde{D}^T\right) + \left(D - \widetilde{D}\right)\widetilde{D}^T \right\|_1 && (90)\\
&\leq \left\| D\left(D^T - \widetilde{D}^T\right) \right\|_1 && (91)\\
&+ \left\| \left(D - \widetilde{D}\right)\widetilde{D}^T \right\|_1 && (92)\\
&\leq \|D\|_2 \left\| \left(D^T - \widetilde{D}^T\right) \right\|_2 && (93)\\
&+ \left\| \left(D - \widetilde{D}\right) \right\|_2 \left\| \widetilde{D}^T \right\|_2. && (94)
\end{aligned}$$

Now use the same argument as in the verification of S1 to see that this term is locally stationary. By using Lemma 4 and an argument similar to the one for the score, we get that

$$\left\| H_t(X_t, \theta_0) - \widetilde{H}_t(\widetilde{X}_t, u, \theta_0) \right\|_1 \leq C\left(\frac{1}{T} + |t/T - u|\right). \tag{95}$$

By BHK Lemma 5.6, we have

$$\mathbb{E}\left| \sup_{\theta\in\Theta} \frac{\partial^2 \widetilde{L}_t(u,\theta)}{\partial\theta\partial\theta'} \right| < \infty, \tag{96}$$

so

$$\sup_{u\in[0,1]} \mathbb{E}\left| \frac{\partial^2 \widetilde{L}_t(u,\theta)}{\partial\theta\partial\theta'} \right| \leq \mathbb{E}\left| \sup_{u\in[0,1]} \frac{\partial^2 \widetilde{L}_t(u,\theta)}{\partial\theta\partial\theta'} \right| < \infty. \tag{97}$$

□

Now, we can apply the law of large numbers, Theorem 2.7(i) in DRW (component-wise) to the Hessian. This yields

$$-\frac{1}{T}\sum_{t=1}^{T} \frac{\partial^2 l_t(\theta_0)}{\partial\theta\partial\theta'} \xrightarrow{P} \int_0^1 \mathbb{E}\left(\widetilde{H}(u,\theta_0)\right) \, \mathrm{d}u := B.$$

We emphasize that the matrix B is nonrandom. It remains to show that B is positive definite. It is clear that B is positive semi-definite. We need to show that it is not singular. As in Francq and Zakoïan (2004), it suffices to show that for any vector λ with the same number of elements as a column in the Hessian,

$$\lambda^T \left(\int_0^1 \mathbb{E}\left(\frac{1}{\widetilde{h}_t(u,\theta)^2} \frac{\partial \widetilde{h}_t(u,\theta)}{\partial\theta} \frac{\partial \widetilde{h}_t(u,\theta)}{\partial\theta^T} \right) \mathrm{d}u \right) \lambda \tag{98}$$

$$= \left(\int_0^1 \mathbb{E}\left(\frac{1}{\widetilde{h}_t(u,\theta)^2} \left(\lambda^T \frac{\partial \widetilde{h}_t(u,\theta)}{\partial\theta} \right)^2 \right) \mathrm{d}u \right) = 0 \tag{99}$$

implies that λ is the zero vector. Define the vector

$$\phi_t := \frac{\partial \widetilde{h}_t(u,\theta)}{\partial \theta}.$$

If we consider a standard GARCH(p,q), it is shown in BHK, Lemma 5.7, that if there is such a vector such that $\lambda^T \phi_t = 0$, $\lambda \neq 0$, (we mean the equality in the sense *almost surely*, but will henceforth use it without this qualifier) then by stationarity it holds for all t, which contradicts the minimality of the model. As argued by Han and Kristensen (2014) in a slightly different but comparable case, in the GARCH(1,1) case it implies that the distribution of ε_t is degenerate. This is ruled out by (A1). We can follow this logic exactly for the GARCH part of the model, but need to adjust it to account for the parameters concerning the transition function.

Consider for simplicity the GARCH(1,1) case and one transition. The vector ϕ_t takes the form

$$\begin{aligned}\phi_t &= (1, \widetilde{X}_t^2(u), \sigma_t^2(u), g_\gamma(u), g_c(u), g_{\alpha_{01}}(u))^T + \beta_1 \phi_{t-1} &(100)\\ &= w_t(u)^T + \beta_1 \phi_{t-1}, &(101)\end{aligned}$$

where the partial derivatives $g_\gamma, g_c, g_{\alpha_0 1}$ are given in Lemma 1. If we consider all terms in $w_t(u)$ we note that, as we are considering the stationary approximation, the three last derivatives are constant over time. Naturally for each u, there will exist a λ_u depending on u such that $\lambda_u^T w_t(u) = 0$. The dependence on u of this λ_u is what makes the integral positive definite. If λ was independent of u, then $\widetilde{B}$ would be singular. Viewing the integral in (99) as a Riemann sum, and taking two points $u_1, u_2 \in [0,1]$ from that sum, we see that

$$\lambda^T \left(\widetilde{H}(u_1,\theta_0) + \widetilde{H}(u_2,\theta_0) \right) \lambda = \lambda^T \widetilde{H}(u_1,\theta_0)\lambda + \lambda^T \widetilde{H}(u_2,\theta_0)\lambda = 0$$

for $\lambda \neq 0$ implies $\widetilde{H}(u_1,\theta_0) = \widetilde{H}(u_2,\theta_0)$, which will only happen if there is no variation in G over time, which is ruled out by (A3). Since for each u it holds that $\widetilde{H}(u,\theta_0)$ is at least positive semi-definite, and there exist no $\lambda \neq 0$ that makes all the terms in the Riemann sum 0, it holds

$$\lambda^T \left(\int_0^1 \mathbb{E} \left(\frac{1}{\widetilde{h}_t(u,\theta_0)^2} \frac{\partial \widetilde{h}_t(u,\theta_0)}{\partial \theta} \frac{\partial \widetilde{h}_t(u,\theta_0)}{\partial \theta^T} \right) \mathrm{d}u \right) \lambda > 0$$

for $\lambda \neq 0$. This gives N2.

7.1.4 Verifying N3

Let θ_i, θ_j and θ_k be elements of θ. The third derivatives of the log likelihood function with respect to the parameters are given by

$$\frac{\partial^3 l_t(\theta)}{\partial\theta_i\partial\theta_j\partial\theta_k} = \left\{1 - \frac{X_t^2}{h_t(\theta)}\right\}\left\{\frac{1}{h_t(\theta)}\frac{\partial^3 h_t(\theta)}{\partial\theta_i\partial\theta_j\partial\theta_k}\right\} \tag{102}$$

$$+ \left\{2\frac{X_t^2}{h_t(\theta)} - 1\right\}\left\{\frac{1}{h_t(\theta)}\frac{\partial h_t(\theta)}{\partial\theta_i}\right\}\left\{\frac{1}{h_t(\theta)}\frac{\partial^2 h_t(\theta)}{\partial\theta_j\partial\theta_k}\right\} \tag{103}$$

$$+ \left\{2\frac{X_t^2}{h_t(\theta)} - 1\right\}\left\{\frac{1}{h_t(\theta)}\frac{\partial h_t(\theta)}{\partial\theta_j}\right\}\left\{\frac{1}{h_t(\theta)}\frac{\partial^2 h_t(\theta)}{\partial\theta_i\partial\theta_k}\right\} \tag{104}$$

$$+ \left\{2\frac{X_t^2}{h_t(\theta)} - 1\right\}\left\{\frac{1}{h_t(\theta)}\frac{\partial h_t(\theta)}{\partial\theta_k}\right\}\left\{\frac{1}{h_t(\theta)}\frac{\partial^2 h_t(\theta)}{\partial\theta_i\partial\theta_j}\right\} \tag{105}$$

$$+ \left\{2 - 6\frac{X_t^2}{h_t(\theta)}\right\}\left\{\frac{1}{h_t(\theta)}\frac{\partial h_t(\theta)}{\partial\theta_i}\right\}\left\{\frac{1}{h_t(\theta)}\frac{\partial h_t(\theta)}{\partial\theta_j}\right\}\left\{\frac{1}{h_t(\theta)}\frac{\partial h_t(\theta)}{\partial\theta_k}\right\}. \tag{106}$$

By repeated use of the Cauchy-Schwarz inequality, Francq and Zakoïan (2004) (see p. 622-626) proved that in the standard (strictly stationary) case, N3 is true in some neighbourhood of θ_0. The authors use weaker assumptions than ours. The restriction to *some* neighbourhood, rather than all of Θ, was necessary in their case because the first term in each row is not uniformly integrable under their assumptions, but *is* integrable in some neighbourhood of θ_0.

Note however, in our case, by (A6), (42) and independence,

$$\left\|\sup_{\theta\in\Theta}\frac{X_t^2}{h_t(\theta)}\right\|_2 = \left\|\varepsilon_0^2\right\|_2 \left\|\sup_{\theta\in\Theta}\frac{h_t(\theta_0)}{h_t(\theta)}\right\|_2 < \infty,$$

and the subsequent terms admit moments of any order by Lemma 4. The proof now follows similarly to Francq and Zakoïan (2004) by applying the Cauchy-Schwarz and Hoelder inequalities to the terms in the derivatives, i.e

$$\left\|\left\{1 - \frac{X_t^2}{h_t(\theta)}\right\}\left\{\frac{1}{h_t(\theta)}\frac{\partial^3 h_t(\theta)}{\partial\theta_i\partial\theta_j\partial\theta_k}\right\}\right\|_1 \le \left\|\left\{1 - \frac{X_t^2}{h_t(\theta)}\right\}\right\|_2 \left\|\left\{\frac{1}{h_t(\theta)}\frac{\partial^3 h_t(\theta)}{\partial\theta_i\partial\theta_j\partial\theta_k}\right\}\right\|_2 < \infty,$$

$$\begin{aligned}&\left\|\left\{2\frac{X_t^2}{h_t(\theta)} - 1\right\}\left\{\frac{1}{h_t(\theta)}\frac{\partial h_t(\theta)}{\partial\theta_i}\right\}\left\{\frac{1}{h_t(\theta)}\frac{\partial^2 h_t(\theta)}{\partial\theta_j\partial\theta_k}\right\}\right\|_1 \\ &\le \left\|\left\{2\frac{X_t^2}{h_t(\theta)} - 1\right\}\right\|_2 \left\|\left\{\frac{1}{h_t(\theta)}\frac{\partial h_t(\theta)}{\partial\theta_i}\right\}\right\|_p \left\|\left\{\frac{1}{h_t(\theta)}\frac{\partial^2 h_t(\theta)}{\partial\theta_j\partial\theta_k}\right\}\right\|_q < \infty\end{aligned}$$

for some $1/p + 1/q = 1/2$ and similarly for the two subsequent terms. As in

Francq and Zakoïan (2004), we can deal with the last term by writing

$$\begin{aligned}
&\left\| \left\{ 2 - 6\frac{X_t^2}{h_t(\theta)} \right\} \left\{ \frac{1}{h_t(\theta)} \frac{\partial h_t(\theta)}{\partial \theta_i} \right\} \left\{ \frac{1}{h_t(\theta)} \frac{\partial h_t(\theta)}{\partial \theta_j} \right\} \left\{ \frac{1}{h_t(\theta)} \frac{\partial h_t(\theta)}{\partial \theta_k} \right\} \right\|_1 \\
&\leq \left\| \left\{ 2 - 6\frac{X_t^2}{h_t(\theta)} \right\} \right\|_2 \max_i \left\| \left\{ \frac{1}{h_t(\theta)} \frac{\partial h_t(\theta)}{\partial \theta_i} \right\} \right\|_6^3 < \infty.
\end{aligned}$$

Appendix: Simulation study

Figure 8: Simulation study DGP 1, 1000 time series observations

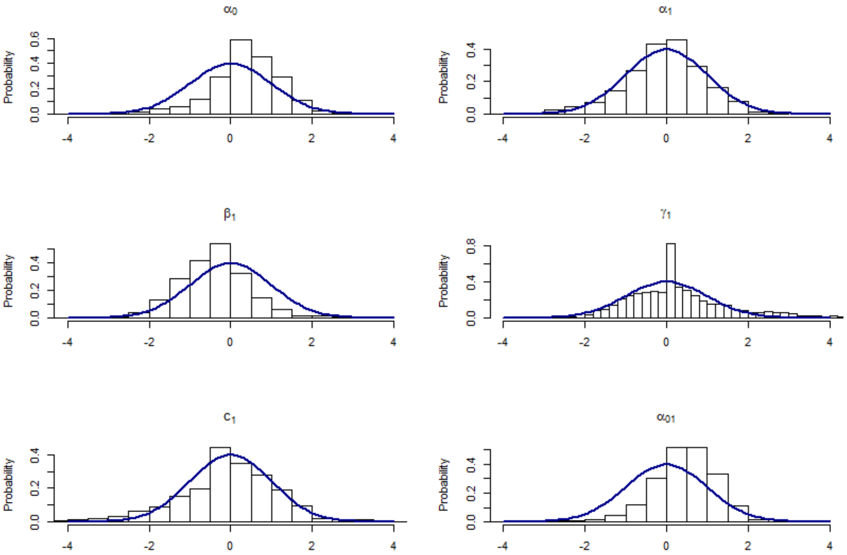


Figure 9: Simulation study DGP 1, 2500 time series observations

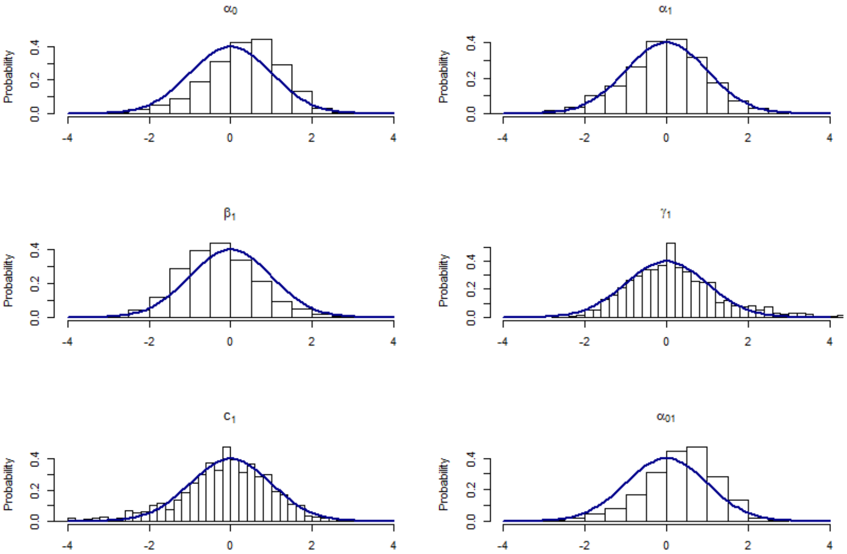


Figure 10: Simulation study DGP 1, 5000 time series observations

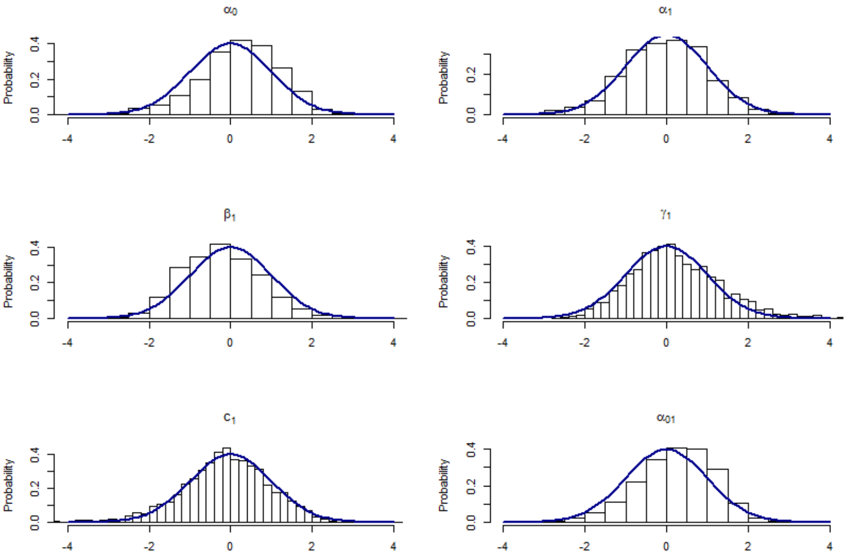


Figure 11: Simulation study DGP 2, 1000 time series observations

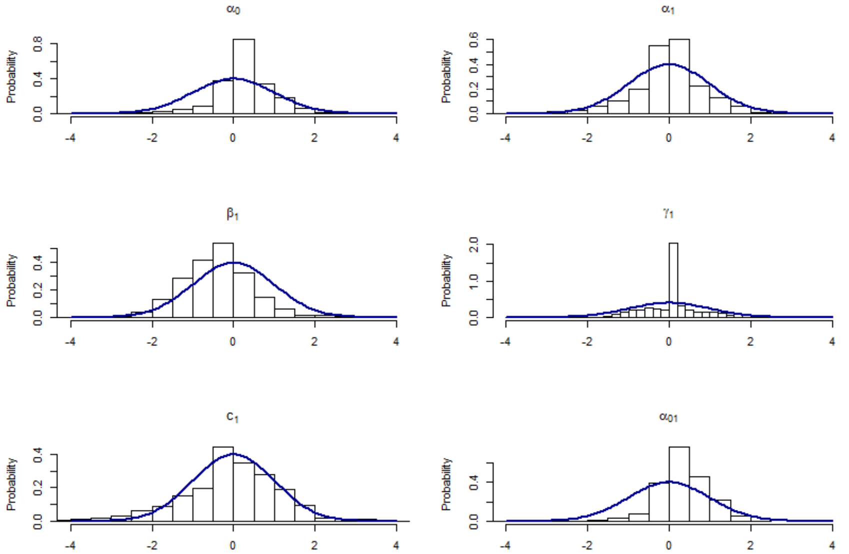


Figure 12: Simulation study DGP 2, 2500 time series observations

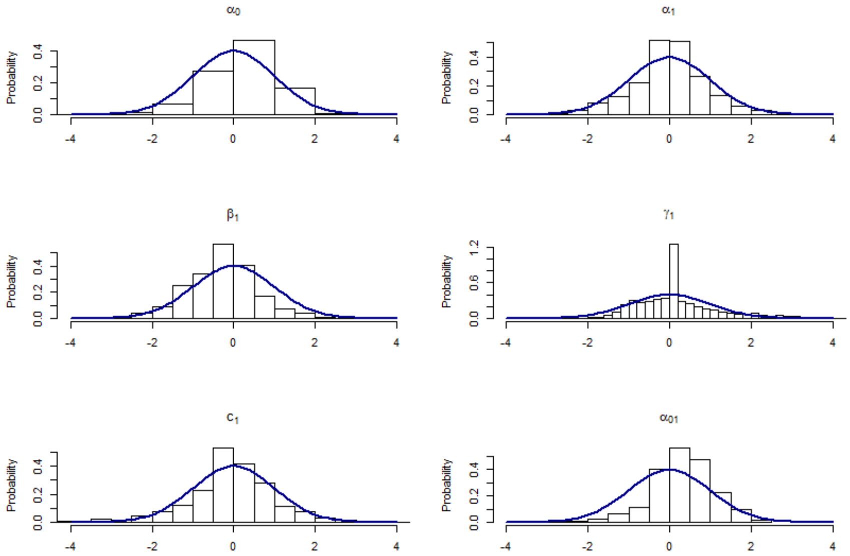


Figure 13: Simulation study DGP 2, 5000 time series observations

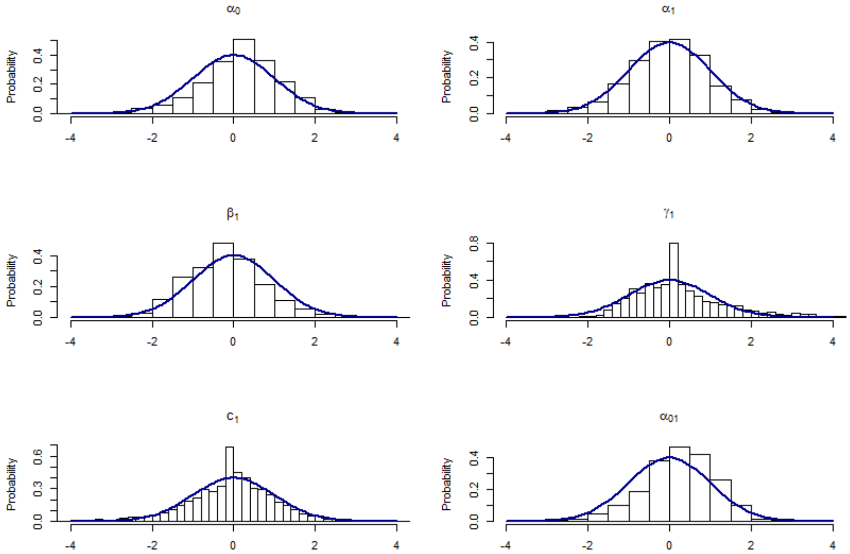


Figure 14: Simulation study DGP 1, 1000 time series observations

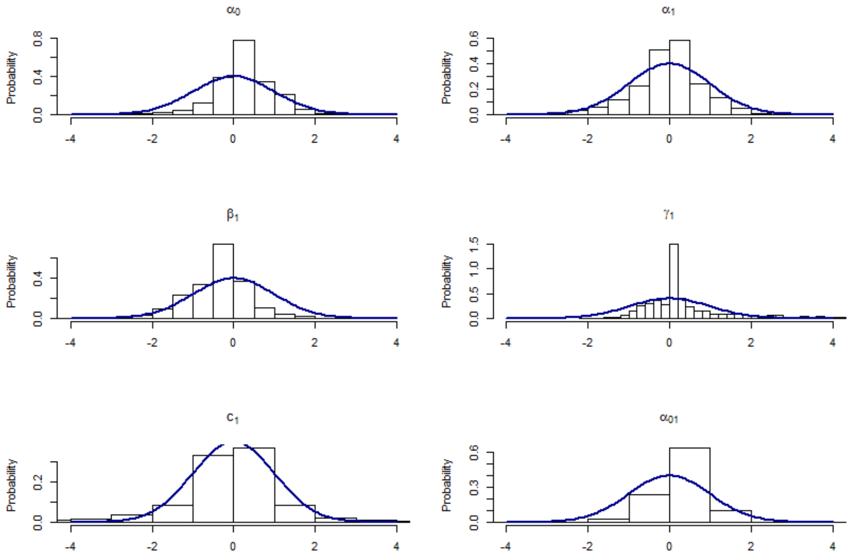


Figure 15: Simulation study DGP 3, 2500 time series observations

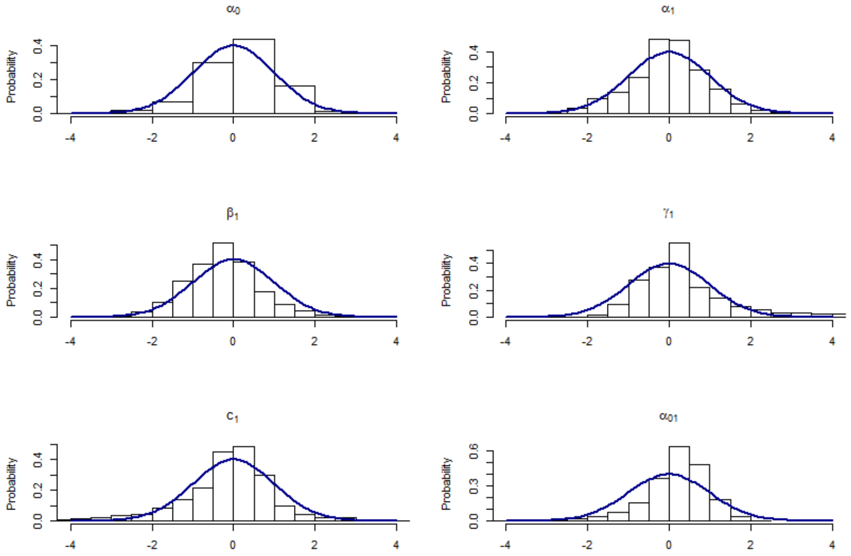
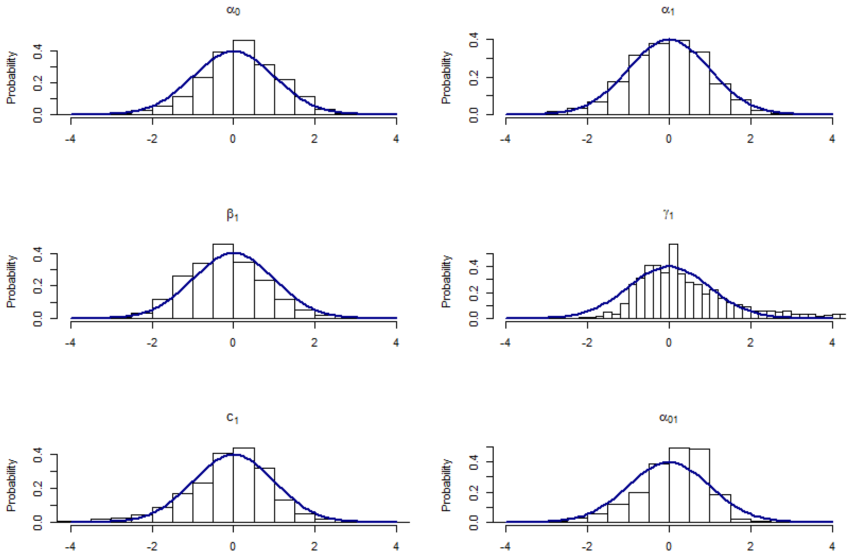


Figure 16: Simulation study DGP 3, 5000 time series observations



References

Amado, C., & Laakkonen, H. (2013). Modelling time-varying volatility in financial returns: Evidence from bond markets. In N. Haldrup, M. Meitz, & P. Saikkonen (Eds.), *Essays in nonlinear time series econometrics* (pp. 139–160). Oxford University Press.

Amado, C., Silvennoinen, A., & Teräsvirta, T. (2019). Models of multiplicative decomposition of conditional variances and correlations. In J. Chevallier, S. Goutte, D. Guerreiro, S. Saglio, & B. Sanhaji (Eds.), *Financial mathematics, volatility and covariance modelling* (pp. 217–260). Routledge.

Amado, C., & Teräsvirta, T. (2008). *Modelling conditional and unconditional heteroskedasticity with smoothly time-varying structure* (SSE/EFI Working Paper Series in Economics and Finance No. 691). Stockholm School of Economics.

Amado, C., & Teräsvirta, T. (2013). Modelling volatility by variance decomposition. *Journal of Econometrics*, *175*, 153–165.

Amado, C., & Teräsvirta, T. (2017). Specification and testing of multiplicative time-varying GARCH models with applications. *Econometric Reviews*, *36*, 421–446.

Amemiya, A., Takeshi. (1985). *Advanced econometrics*. Harvard University Press.

Anderson, H. M., Nam, K., & Vahid, F. (1999). Asymmetric nonlinear smooth transition GARCH models. In P. Rothman (Ed.), *Nonlinear time series analysis of economic and financial data* (pp. 191–207). Kluwer.

Andreou, E., & Ghysels, E. (2009). Structural breaks in financial time series. In T. G. Andersen, R. A. Davis, J.-P. Kreiss, & T. Mikosch (Eds.), *Handbook of financial time series* (pp. 839–870). Springer.

Andrews, D. W. (1992). Generic uniform convergence. *Econometric theory*, *8*(2), 241–257.

Baillie, R. T., Bollerslev, T., & Mikkelsen, H. O. (1996). Fractionally integrated generalized autoregressive conditional heteroskedasticity. *Journal of Econometrics*, *74*, 3–30.

Basawa, I. V., Feigin, P. D., & Heyde, C. C. (1976). Asymptotic properties of maximum likelihood estimators for stochastic processes. *Sankhyā: The Indian Journal of Statistics, Series A (1961-2002)*, *38*(3), 259–270. http://www.jstor.org/stable/25050051

Berkes, I., Horváth, L., & Kokoszka, P. (2003). Garch processes: Structure and estimation. *Bernoulli*, *9*(2), 201–227. http://www.jstor.org/stable/3318937

Bollerslev, T. (1986). Generalized autoregressive conditional heteroskedasticity. *Journal of Econometrics*, *31*, 307–327.

Bougerol, P., & Picard, N. (1992). Stationarity of garch processes and of some nonnegative time series. *Journal of Econometrics*, *52*(1), 115–127. https://doi.org/https://doi.org/10.1016/0304-4076(92)90067-2

Brandt, A. (1986). The stochastic equation $Y_{n+1} = A_n Y_n + B_n$ with stationary coefficients. *Advances in Applied Probability*, *18*(1), 211–220. http://www.jstor.org/stable/1427243

Caulet, R., & Péguin-Feissolle, A. (2000). Un test d'hétéroscédasticité conditionnelle inspiré de la modélisation en termes de réseaux neuronaux artificiels. *Annales d'économie et de statistique*, *59*, 178–197.

Chen, C. W. S., Gerlach, R., & Lin, E. M. H. (2014). Bayesian estimation of smoothly mixing time-varying parameter GARCH models. *Computational Statistics and Data Analysis*, *76*, 194–209.

Čížek, P. (2011). Modelling conditional heteroscedasticity in nonstationary series. In P. Čížek, W. Härdle, & R. Weron (Eds.), *Statistical tools for finance and insurance* (pp. 101–132). Springer.

Čížek, P., & Spokoiny, V. (2009). Varying coefficient GARCH models. In T. G. Andersen, R. A. Davis, J.-P. Kreiss, & T. Mikosch (Eds.), *Handbook of financial time series* (pp. 169–185). Springer.

Comte, F., & Lieberman, O. (2003). Asymptotic theory for multivariate garch processes. *Journal of Multivariate Analysis*, *84*(1), 61–84. https://doi.org/https://doi.org/10.1016/S0047-259X(02)00009-X

Diebold, F. X. (1986). Modeling persistence in conditional variances: A comment. *Econometric Reviews*, *5*, 51–56.

Engle, R. F. (1982). Autoregressive conditional heteroscedasticity with estimates of the variance of United Kingdom inflation. *Econometrica*, *50*, 987–1007.

Engle, R. F., & Bollerslev, T. (1986). Modeling persistence in conditional variances. *Econometric Reviews*, *5*, 1–50.

Engle, R. F., & Ng, V. K. (1993). Measuring and testing the impact of news on volatility. *Journal of Finance*, *48*, 1749–1777.

Engle, R. F., & Rangel, J. G. (2008). The spline-GARCH model for low-frequency volatility and its global macroeconomic causes. *Review of Financial Studies*, *21*, 1187–1222.

Feng, Y. (2004). Simultaneously modeling conditional heteroskedasticity and scale change. *Econometric Theory*, *20*, 563–596.

Francq, C., & Zakoïan, J.-M. (2004). Maximum likelihood estimation of pure garch and arma-garch processes. *Bernoulli*, *10*(4), 605–637. http://www.jstor.org/stable/3318818

Francq, C., & Zakoïan, J.-M. (2019). Garch(p, q) processes. *Garch models* (pp. 16–57). John Wiley & Sons, Ltd. https://doi.org/https://doi.org/10.1002/9781119313472.ch2

Glosten, L., Jagannathan, R., & Runkle, D. (1993). On the relation between expected value and the volatility of the nominal excess return on stocks. *Journal of Finance*, *48*, 1779–1801.

Gonzalez-Rivera, G. (1998). Smooth transition GARCH models. *Studies in Nonlinear Dynamics and Econometrics*, *3*, 161–178.

Hagerud, G. (1997). *A new non-linear GARCH model*. EFI Economic Research Institute.

Han, H., & Kristensen, D. (2014). Asymptotic theory for the QMLE in GARCH-X models with stationary and nonstationary covariates. *Journal of Business & Economic Statistics*, *32*, 416–429.

Han, H., & Kristensen, D. (2017). *Semiparametric multiplicative GARCH-X model: Adopting economic variables to explain volatility* (Working paper).

Härdle, W., Herwartz, H., & Spokoiny, V. (2003). Time inhomogeneous multiple volatility modeling. *Journal of Financial Econometrics*, *1*, 55–95.

Lamoureux, C. G., & Lastrapes, W. G. (1990). Persistence in variance, structural change and the GARCH model. *Journal of Business and Economic Statistics*, *8*, 225–234.

Lanne, M., & Saikkonen, P. (2005). Nonlinear GARCH models for highly persistent volatility. *Econometrics Journal*, *8*, 251–276.

Lubrano, M. (2001). Smooth transition GARCH models: A Bayesian perspective. *Recherches économiques de Louvain*, *67*, 257–287.

McAleer, M., & Ling, S. (2002). Necessary and sufficient moment conditions for the garch(r, s) and asymmetric power garch(r, s) models. *Econometric Theory*, *18*, 722–729. https://doi.org/10.1017/S0266466602183071

Medeiros, M. C., & Veiga, A. (2009). Modeling multiple regimes in financial volatility with a flexible coefficient GARCH(1,1) model. *Econometric Theory*, *25*, 117–161.

Morana, C. (2002). IGARCH effects: An interpretation. *Applied Economics Letters*, *9*, 745–748.

Nelson, D. B. (1991). Conditional heteroskedasticity in asset returns: A new approach. *Econometrica*, *59*, 347–370.

Taylor, S. J. (1986). *Modelling financial time series*. Wiley.

Teräsvirta, T. (2012). Nonlinear models for autoregressive conditional heteroskedasticity. In L. Bauwens, C. Hafner, & S. Laurent (Eds.), *Handbook of volatility models and their applications* (pp. 49–69). Wiley.

van Bellegem, S., & von Sachs, R. (2004). Forecasting economic time series with unconditional time-varying variance. *International Journal of Forecasting*, *20*, 611–627.

van Bellegem, S. (2012). Locally stationary volatility modeling. In L. Bauwens, C. Hafner, & S. Laurent (Eds.), *Handbook of volatility models and their applications* (pp. 249–268). Wiley.

Wu, W. B. (2011). Asymptotic theory for stationary processes. *Statistics and its Interface*, *4*(2), 207–226.

Wu, W. B., & Min, W. (2005). On linear processes with dependent innovations. *Stochastic Processes and their Applications*, *115*(6), 939–958.