An Adaptive Recursive Volatility Prediction Method

Nicklas Werge, Olivier Wintenberger

An Adaptive Recursive Volatility Prediction Method

Nicklas Werge[a], Olivier Wintenberger[a]

[a]*LPSM, Sorbonne Université, 4 place Jussieu, 75005 Paris, France*

Abstract

The Quasi-Maximum Likelihood (QML) procedure is widely used for statistical inference due to its robustness against overdispersion. However, while there are extensive references on non-recursive QML estimation, recursive QML estimation has attracted little attention until recently. In this paper, we investigate the convergence properties of the QML procedure in a general conditionally heteroscedastic time series model, extending the classical offline optimization routines to recursive approximation. We propose an adaptive recursive estimation routine for GARCH models using the technique of Variance Targeting Estimation (VTE) to alleviate the convergence difficulties encountered in the usual QML estimation. Finally, empirical results demonstrate a favorable trade-off between the ability to adapt to time-varying estimates and stability of the estimation routine.

Keywords: volatility models, quasi-likelihood, recursive algorithm, GARCH, prediction method, stock index

1. Introduction

Time series analysis has attracted considerable attention in the last three decades. A crucial issue for time series analysis is modeling heteroscedasticity of the conditional variance e.g. volatility clustering in financial time series. The most known models capturing this feature are the autoregressive conditional heteroscedasticity (ARCH) model and generalized ARCH (GARCH) model introduced by Engle (1982) and Bollerslev (1986), respectively. The success of these models can be explained by many reasons; one of them is that they constitute a stationary time series model with a time-varying conditional variance; another one is that they model time series with heavier tails than the Gaussian one which often occurs in financial time series.

Quasi-Maximum Likelihood (QML) estimation is widely used for statistical inference in GARCH models since it tolerates this overdispersion. In this paper, we study the Quasi-Maximum Likelihood Estimator (QMLE) for the broader class of conditionally heteroscedastic time series models of multiplicative form given by

$$X_t = h_t(\theta_0)Z_t, \quad t \in \mathbb{Z}, \tag{1.1}$$

where θ_0 is the true underlying parameter vector and the (non-negative) volatility process $(h_t)_{t\in\mathbb{Z}}$ is defined as

$$h_t(\theta) = g_\theta(X_{t-1}, \dots, X_{t-p}, h_{t-1}(\theta), \dots, h_{t-q}(\theta)), \quad p, q \geq 0, \tag{1.2}$$

where (Z_t) is a sequence of i.i.d. random variables with $\mathbb{E}[Z_0] = 0$ and $\mathbb{E}[Z_0^2] = 1$. Suppose that the parameter set $\Theta \subset \mathbb{R}^d$ and $\{g_\theta | \theta \in \Theta\}$ denotes the (finite) parametric family of non-negative functions on $\mathbb{R}^p \times [0, \infty)^q$, fulfilling certain regularity conditions. We also require that h_t is $\mathcal{F}_{t-1}$-measurable where for all $t \in \mathbb{Z}$, $\mathcal{F}_t = \sigma(Z_k : k \leq t)$ denotes the σ-field generated by the random variables $\{Z_k : k \leq t\}$.

The stability of model (1.1)-(1.2) is accomplished under the assumption that "g_θ is a contraction". This condition is a random Lipschitz coefficient condition where the Lipschitz coefficient has a negative logarithmic moment. The notion of contractivity is clarified in Straumann and Mikosch (2006) where they study QML inference of general conditionally heteroscedastic models with emphasis on the approximation $(\hat{h}_t)$ of the stochastic volatility (h_t).

QML estimation of the parameters in the class of conditionally heteroscedastic time series models has been studied frequently in recent years, see e.g. Berkes et al. (2003), Francq and Zakoïan (2004), Straumann and Mikosch (2006) and Wintenberger (2013). However, all these references consider non-recursive (offline) estimation where one assemble a batch of data and afterward perform the statistical inference. Nevertheless, recursive estimates made using recursive algorithms are arguably advantageous as one treats only the observations once. Indeed in recursive QML estimation, we update the QMLE at time $t-1$ with the new observations at time t to yield the QMLE of the parameters at time t.

In modern statistics analysis, it is becoming increasingly common to work with streaming data where one observes only a group of observations at a time. Naturally, this has led to an expanded interest in scalable (in time) recursive estimation procedures e.g. see Bottou and Bousquet (2007). However, only a little amount of attention has been given to recursive (online) estimation in conditionally heteroscedastic time series models. Dahlhaus and Subba Rao (2007) presented a recursive method to estimate the parameters in an ARCH process. Under sufficient conditions on the underlying process, Aknouche and Guerbyenne (2006) showed consistency of their recur-

Email addresses: nicklas.werge@upmc.fr (Nicklas Werge), olivier.wintenberger@upmc.fr (Olivier Wintenberger)

sive least squares method for GARCH processes. Kierkegaard et al. (2000) also developed a recursive estimation method for GARCH processes supported by empirical evidence. The authors of Gerencsér et al. (2010) show convergence analysis of recursive QML estimation for GARCH processes based on BMP-theory with the use of a resetting mechanism. A self-weighted recursive estimation algorithm for GARCH models was proposed by Cipra and Hendrych (2018) with a robustification in Hendrych and Cipra (2018). However, none of the above references mention convexity nor estimation issues of small ω parameter values for GARCH models.

In settings with streaming data sets, the difficulty of estimating time-varying parameters of statistical models increases. To sustain computational efficiency and being adaptive one may decrease the number of observations in each iteration in the optimization procedure which may increase the instability of the statistical inference. We propose a natural adaptation of the QML method relying on stochastic approximations with the use of Variance Targeting Estimation (VTE) technique. We give empirical evidence that the procedure achieves a favorable trade-off between adaptation ability and stability.

2. QML Estimation in Conditionally Heteroscedastic Time Series Models

The approximate QMLE $\hat{\theta}_n^*$ is defined as

$$\hat{\theta}_n^* \in \arg\min_{\theta\in\mathcal{K}} \hat{L}_n(\theta), \tag{2.1}$$

where the parameter set $\mathcal{K}$ is a suitable compact subset of the parameter space Θ. The QL function $L_n(\theta)$ and approximate QL function $\hat{L}_n(\theta)$ are, respectively, given by

$$L_n(\theta) = \sum_{t=1}^{n} l_t(\theta) \text{ and } \hat{L}_n(\theta) = \sum_{t=1}^{n} \hat{l}_t(\theta), \tag{2.2}$$

where the QL losses, $l_t(\theta)$ and $\hat{l}_t(\theta)$, are given by

$$l_t(\theta) = \frac{1}{2}\left(\frac{X_t^2}{h_t(\theta)} + \log h_t(\theta)\right) \text{ and } \hat{l}_t(\theta) = \frac{1}{2}\left(\frac{X_t^2}{\hat{h}_t(\theta)} + \log \hat{h}_t(\theta)\right), \tag{2.3}$$

where $(\hat{h}_t)$ is an approximation of (h_t) defined recursively for $t \geq 1$ thanks to Eqn. (1.2) with initialization $\hat{h}_{-q+1} = \cdots = \hat{h}_0 = 0$ or any deterministic constant. Whatever is the initialization the error between $(\hat{h}_t)$ and the true (h_t) will vanish exponentially fast almost surely from (Straumann, 2005, Proposition 5.2.12). Assuming that Z_0 is standard normal distributed, note that X_t is also Gaussian with variance h_t conditionally on $\mathcal{F}_{t-1}$. The QL function $L_n(\cdot)$ in (2.2) is derived under this Gaussian assumption.

The consistency and asymptotic properties of the QMLE $\hat{\theta}_n^*$, combined with the robustness of the QL function with respect to overdispersion (See Patton (2006)) makes the method highly used in practice. Under the conditions in (Straumann and Mikosch, 2006, N.1, N.2, N.3 and N.4) then the QMLE $\hat{\theta}_n^*$ is strongly consistent and asymptotically normal, i.e.

$$\hat{\theta}_n^* \xrightarrow{\text{a.s.}} \theta_0 \text{ and } \sqrt{n}\left(\hat{\theta}_n^* - \theta_0\right) \to \mathcal{N}(0, V_0) \text{ as } n \to \infty, \tag{2.4}$$

with θ_0 as the true parameter vector and V_0 the asymptotic covariance matrix.

Unfortunately, these asymptotic properties in (2.4) come with a drawback on the QL loss; the robustness is achieved thanks to careful domination with logarithms. The concavity of those logarithms makes the criterion insensitive to extreme values but it also implies that the criterion behaves itself as a concave function. As most optimization algorithms are based on convex assumptions then this is striking.

In the next section, we show that the approximate Hessian $\hat{H}_n(\theta) = n^{-1}\nabla^2\hat{L}_n(\theta)$ admits strictly positive eigenvalues for n large enough depending on the model specifications and the underlying data process. Meaning, for sufficiently large batch sizes of observations then the QMLE $\hat{\theta}_n^*$ can be seen as the unique solution of a locally strongly convex optimization problem; The existence and uniqueness of $\hat{\theta}_n^*$ ensure that it can be efficiently approximated by usual (offline) optimization routines for n large enough.

2.1. Asymptotic Convex Properties of the QL Function

In order to establish the asymptotic local convexity of the QL function of model (1.1)-(1.2) we need the following assumptions; Assumption W1, W2 and W3, which naturally emerges by the arguments and properties (Straumann and Mikosch, 2006, N.1, N.2, N.3 and N.4) made in order to have stability of the QL function and QMLE procedure. We will in this paper use two different matrix norms: Let $\|A\|_{op}$ denote the matrix operator norm of matrix $A \in \mathbb{R}^{d\times d}$ with respect to the Euclidean norm i.e. $\|A\|_{op} = \sup_{v\neq 0}|Av|/|v|$. Denote $\|A\|_{\mathcal{K}}$ the norm of the continuous matrix-valued function A on $\mathcal{K}$ i.e. $\|A\|_{\mathcal{K}} = \sup_{x\in\mathcal{K}}\|A(x)\|_{op}$, where $\mathcal{K}$ is a compact set of $\mathbb{R}^d$.

Assumption W1. Assume model (1.1)-(1.2) with $\theta = \theta_0$ admits a unique stationary ergodic solution.

Assumption W2. Assume $\mathcal{K} \subset \Theta$ is a compact set with true parameter vector $\theta_0 \in \mathcal{K}$ in the interior. The random functions fulfill certain conditions, such that $\mathbb{E}[\|l_0\|_{\mathcal{K}}] < \infty$, $\mathbb{E}[\|\nabla^2 l_0\|_{\mathcal{K}}] < \infty$ and further have the following uniform convergences: $\|n^{-1}\hat{L}_n - L_n\|_{\mathcal{K}} \xrightarrow{\text{a.s.}} 0$ and $n^{-1}\|\nabla^2\hat{L}_n - \nabla^2 L_n\|_{\mathcal{K}} \xrightarrow{\text{a.s.}} 0$ for $n \to \infty$.

Assumption W3. Assume the components of the vector $\nabla_\theta g_\theta(X_0, h_0)$ from (1.2) with $\theta = \theta_0$ are linearly independent random variables.

The following Theorem 2.1 is an extension of Ip et al. (2006) which established similar results for the likelihood function of GARCH models under the assumption that (X_t) is strictly stationary and strongly mixing with geometric rate and (Z_t) is Gaussian. Solving the QML estimation problem (2.1) for $\hat{\theta}_n^*$ is known to be computationally heavy since one has to find the solution of the non-linear equation (2.2). Nonetheless, Theorem 2.1 ensures the existence of a N such that we have a unique global QMLE $\hat{\theta}_n^*$.

Theorem 2.1. *Under Assumption W1, W2 and W3 there exist positive constants $C, \delta > 0$ and a random positive integer $N \in \mathbb{N}_+$ such that we have*

$$g^T \hat{H}_n(\theta) g > C g^T g, \ \forall n \geq N, \ g \in \mathbb{R}^d \setminus \{0\}, \quad a.s., \tag{2.5}$$

for all $\theta \in B(\theta_0, \delta)$.

The above results shows local strongly convexity of the QL function $\hat{L}_n$. The following corollary arises from the proof of Theorem 2.1:

Corollary 2.1. *Under Assumption W1, W2 and W3, the QMLE $\hat{\theta}_n^*$ exists and is unique, namely*

$$\hat{\theta}_n^* = \underset{\theta \in \mathcal{K}}{\arg\min} \, \hat{L}_n(\theta).$$

Local strong convexity is crucial in order to have convergence of an optimization algorithm. Thus, Theorem 2.1 is an essential result to compute the QMLE $\hat{\theta}_n^*$ for the parameters in model (1.1)-(1.2). But to guarantee the property in (2.5), we need a sufficiently large (and maybe unbounded) random N which depends on the true parameter vector, the parameter estimates and the observations. In practice, one often has a fixed size of observations, which is why the iterative algorithm may not converge. To our experience, this phenomena will occur when the true parameter vector is near the boundary of $\mathcal{K}$ or if the initial values are far away from the true parameters.

2.2. QML Estimation of GARCH(p, q) Parameters

The general class of conditionally heteroscedastic time series models includes the very popular ARCH and GARCH models. These models have for more than three decades, since their introduction, attracted considerable amounts of attention in the literature. A process (X_t) is called a GARCH(p, q) process with parameter vector $\theta = (\omega, \alpha_1, \ldots, \alpha_p, \beta_1, \ldots, \beta_q)^T$ if it satisfies

$$\begin{cases} X_t = \sigma_t Z_t, \\ \sigma_t^2 = \omega + \sum_{i=1}^p \alpha_i X_{t-i}^2 + \sum_{j=1}^q \beta_j \sigma_{t-j}^2, \end{cases} \tag{2.6}$$

where ω, α_i and β_j for $1 \leq i \leq p$ and $1 \leq j \leq q$ are non-negative parameters ensuring the non-negativity of the conditional variance process (σ_t^2). The innovations (Z_t) is a sequence of i.i.d. random variables with $\mathbb{E}[Z_0] = 0$ and $\mathbb{E}[Z_0^2] = 1$. Similarly, one can define an ARCH(p) process by setting $\beta_j = 0$ for $1 \leq j \leq q$ in (2.6). A GARCH(p, q) process (X_t) given in (2.6) has QL losses given by $\hat{l}_t(\theta) = 2^{-1}(X_t^2/\hat{\sigma}_t^2(\theta) + \log \hat{\sigma}_t^2(\theta))$ with first derivative

$$\nabla \hat{l}_t(\theta) = \nabla \hat{\sigma}_t^2(\theta) \left(\frac{\hat{\sigma}_t^2(\theta) - X_t^2}{2\hat{\sigma}_t^4(\theta)} \right), \tag{2.7}$$

and second derivate

$$\nabla^2 \hat{l}_t(\theta) = \nabla \hat{\sigma}_t^2(\theta)^T \nabla \hat{\sigma}_t^2(\theta) \left(\frac{2X_t^2 - \hat{\sigma}_t^2(\theta)}{2\hat{\sigma}_t^6(\theta)} \right) + \nabla^2 \hat{\sigma}_t^2(\theta) \left(\frac{\hat{\sigma}_t^2(\theta) - X_t^2}{2\hat{\sigma}_t^4(\theta)} \right), \tag{2.8}$$

where $\nabla \hat{\sigma}_t^2(\theta) = \vartheta_t(\theta) + \sum_{j=1}^q \beta_j \nabla \hat{\sigma}_{t-j}^2(\theta)$ with $\vartheta_t(\theta) = (1, X_{t-1}^2, \ldots, X_{t-p}^2, \hat{\sigma}_{t-1}^2(\theta), \ldots, \hat{\sigma}_{t-q}^2(\theta))^T \in \mathbb{R}^{p+q+1}$ and Hessian $\hat{H}_n(\theta) = n^{-1} \sum_{t=1}^n \nabla^2 \hat{l}_t(\theta)$.

The equation (2.6) creates a complicated probabilistic structure that is not easily understood although it looks rather simple. A solution for the problem of finding conditions for the existence and uniqueness of a stationary solution to the equations (2.6) for GARCH$(1, 1)$ was provided by Nelson (1990) while Bougerol and Picard (1992) showed it for the GARCH(p, q) model. Bougerol and Picard (1992) make use of the fact that GARCH(p, q) can be embedded in a Random Iterated Lipschitz Map (RILM). See Bougerol (1993) for a formal definition of RILMs.

We can illustrate the RILM method on the GARCH$(1, 1)$ model with parameter vector $\theta = (\omega, \alpha_1, \beta_1)^T$. The RILM for σ_t^2 is then given by $\sigma_t^2 = A_t \sigma_{t-1}^2 + B_t$ with $t \in \mathbb{Z}$ where $A_t = \alpha_1 Z_{t-1}^2 + \beta_1$ and $B_t = \omega$. Remark that $((A_t, B_t))$ constitutes an i.i.d. sequence. From the literature on RILMs, it is well known that the conditions $\mathbb{E}[\log |A_0|] < 0$ and $\mathbb{E}[\log^+ |B_0|] < \infty$ guarantee the existence and uniqueness of a strictly stationary solution of the RILM $Y_t = A_t Y_{t-1} + B_t$ for $t \in \mathbb{Z}$, provided $((A_t, B_t))$ is a stationary ergodic sequence. Applying this to the GARCH$(1, 1)$ model we have $\mathbb{E}[\log(\alpha_1 Z_0^2 + \beta_1)] < 0$ which is known as the sufficient condition for the existence of a stationary solution. This also implies $\beta_1 < 1$ since $\log(\beta_1) \leq \mathbb{E}[\log(\alpha_1 Z_0^2 + \beta_1)] < 0$. In the same way, the ARCH(1) process ($\beta_1 = 0$) then require $\mathbb{E}[\log(\alpha_1 Z_0^2)] < 0$ which is the same as $\alpha < 2e^{\epsilon} \approx 3.56$ with Z_0 Gaussian. Thus, the condition for stationary is much weaker than the second order stationary condition where $\alpha_1 < 1$ is demanded.

The statistical inference leads to further nontrivial problems since the exact distribution of (Z_t) remains unspecified and thus one usually determines the likelihoods under the hypothesis of standard Gaussian innovations. Moreover, the volatility (σ_t) is an unobserved quantity that is approximated by mimicking the recursion (2.6) with an initialization $X_{-p+1} = \cdots = X_0 = 0$ and $\sigma_{-q+1}^2 = \cdots = \sigma_0^2 = 0$ (for example). Berkes et al. (2003) showed under minimal assumptions that the QMLE is strongly consistent and asymptotically normal.

Furthermore, under Assumption W1-W3 we have asymptotic local strong convexity of the QL function in GARCH(p, q) models by Theorem 2.1. However, the number of observations needed to guarantee local strong convexity vary. This can easily be seen by looking at the simplest case, namely, where (X_t) is an ARCH(1) process with parameter vector $\theta = (\omega, \alpha_1)^T$. The volatility process $\sigma_t^2(\theta)$ is given as $\omega + \alpha_1 X_{t-1}^2$. The non-negativity of $\nabla^2 l_t(\theta)$ given by $(1 + X_{t-1}^4)(2X_t^2 - \sigma_t^2(\theta))\sigma_t^{-6}(\theta)$, would ensure convexity at iteration t in our QML procedure. However, the probability of having convexity at each iteration is unlikely as $\mathbb{P}(\cap_{t=1}^n \nabla^2 l_t(\theta) \geq 0) = \mathbb{P}(\cap_{t=1}^n Z_t^2 \geq 1/2) = \mathbb{P}(Z_0^2 \geq 1/2)^n$ is approximately 0.52^n with i.i.d. Gaussian innovations (Z_t) i.e. (Z_t) is chi-squared distributed with 1 degree of freedom, $Z_0^2 \sim \chi_1^2$. On the opposite side, increasing the number of observations used at each iteration would increase the probability of having local strong convexity.

3. Adaptive Recursive QML Estimation

Our recursive QML method relies on stochastic approximations introduced by Robbins and Monro (1951) which only requires the previous parameter estimate at each iterate to update the parameter estimate using the new observation. We perform the first-order stochastic gradient method defined as

$$\hat{\theta}_t = \hat{\theta}_{t-1} - \eta_{t-1}\nabla\hat{l}_t(\hat{\theta}_{t-1}), \tag{3.1}$$

where $\eta_{t-1} > 0$ is the step-size at the $t-1$ step and $\nabla\hat{l}_t(\hat{\theta}_{t-1})$ is the gradient using the X_t observation and the QMLE estimate $\hat{\theta}_{t-1}$. Depending on the amount of observations, we have a trade-off between the accuracy of the QML estimates and the time it takes to perform a parameter update (See Bottou and Bousquet (2007)).

According to Robbins and Monro (1951) we must schedule the step-size such that $\sum_{t=1}^{\infty}\eta_t = \infty$ and $\sum_{t=1}^{\infty}\eta_t^2 < \infty$. But these bounds does not make the choice of an appropriate step-size η_t easier in practice. A more suitable approach is an adaptive learning rate which update the step-size in (3.1) on the fly pursuant to the gradient $\nabla\hat{l}_t(\cdot)$. Thus, our choice of step-size η_t have less impact on performance, making convergence more robust and lower the demand for manual fine-tuning. This is often used in streaming settings where generic methods are preferred. Adaptive learning rates and a separate learning rate for each parameter was proposed by Duchi et al. (2011) in their AdaGrad procedure. This speeds up convergence in situations where the appropriate learning rates vary across parameters. Other well-known examples of adaptive learning rates could be AdaDelta by Zeiler (2012), RMSProp by Tieleman and Hinton (2012) and ADAM by Kingma and Ba (2015). As we may expect a lack of convexity then we select the AdaGrad algorithm since it has shown promising result in non-convex optimization (See Ward et al. (2018)). The AdaGrad procedure is given by the updates

$$\hat{\theta}_t = \hat{\theta}_{t-1} - \frac{\eta}{\sqrt{\sum_{i=1}^{t}\nabla\hat{l}_i(\hat{\theta}_{i-1})^2 + \epsilon}}\nabla\hat{l}_t(\hat{\theta}_{t-1}), \tag{3.2}$$

(thought element-wise) where $\eta > 0$ is a constant learning rate and $\epsilon > 0$ a small number ensuring positivity. Note $\nabla\hat{l}_i(\hat{\theta}_{i-1})^2$ indicates the element-wise square $\nabla\hat{l}_i(\hat{\theta}_{i-1}) \odot \nabla\hat{l}_i(\hat{\theta}_{i-1})$.

As the QL loss is only defined for $\hat{\theta}_n \in \mathcal{K}$, we will require that the recursive algorithm always lies in $\mathcal{K}$. As suggested by Zinkevich (2003) we project our approximation $\hat{\theta}_n$ onto $\mathcal{K}$ preventing large jumps and enforcing our stochastic gradient method to converge. This is implemented on (3.2) and our method is given by

$$\hat{\theta}_t = \text{Projection}_{\mathcal{K}}\left[\hat{\theta}_{t-1} - \frac{\eta}{\sqrt{\sum_{i=1}^{t}\nabla\hat{l}_i(\hat{\theta}_{i-1})^2 + \epsilon}}\nabla\hat{l}_t(\hat{\theta}_{t-1})\right]. \tag{3.3}$$

3.1. Adaptive Recursive QML Estimation for GARCH Models

The parameters in a GARCH process (X_t) are numerically difficult to estimate in empirical applications. The numerical optimization algorithms can easily fail or converge to irregular solutions (See Zumbach (2000)). Therefore, the approximation of the QMLE $\hat{\theta}_n^*$ must be examined with a healthy dose of skepticism. A well-discussed problem for the GARCH(p,q) models is that the QMLE performs badly for numerically small (but still positive) ω values. The parameter ω is a vital and often tricky parameter to estimate. Stabilizing the estimation of ω would not only improve the ω estimate but also have a positive impact on the other model parameters.

On way to overcome small ω values for the GARCH(p,q) model is by scaling (X_t) with some factor $\lambda > 0$ since we have homogeneity[1]. However, we wish to avoid this form of inference in our recursive algorithm as one then need to come up with a scaling parameter which have to be estimated beforehand. Instead, we comprehend this issue by introducing a concept called Variance Targeting Estimation (VTE) (see Francq et al. (2011)). We apply VTE for estimating ω by use of γ^2 which is the unconditional variance estimated by the sample variance as seen in (3.4). Thus we have a two-step estimator where we estimate γ^2 recursively as the sample variance and the remaining parameters are estimated by the QML method. Pseudo-code of our proposed adaptive recursive algorithm is presented in Algorithm 1. The reparametrization is obtained by defining

$$\omega = \gamma^2\left(1 - \sum_{i=1}^{p}\alpha_i - \sum_{j=1}^{q}\beta_j\right). \tag{3.4}$$

The volatility process in the GARCH(p,q) process can then be rewritten as

$$(\sigma_t^2 - \gamma^2) = \sum_{i=1}^{p}\alpha_i(X_{t-i}^2 - \gamma^2) + \sum_{j=1}^{q}\beta_j(\sigma_{t-j}^2 - \gamma^2). \tag{3.5}$$

Similarly, one can define an ARCH(p) process by setting $\beta_j = 0$ for $1 \le j \le q$. The GARCH(p,q) process (X_t) in (3.5) has similar QL losses as before except from $\nabla\hat{\sigma}_t^2(\theta)$ in (2.7) and (2.8) where $\vartheta_t(\theta)$ is given as $(X_{t-1}^2 - \gamma^2, \ldots, X_{t-p}^2 - \gamma^2, \hat{\sigma}_{t-1}^2(\theta) - \gamma^2, \ldots, \hat{\sigma}_{t-q}^2(\theta) - \gamma^2)^T \in \mathbb{R}^{p+q}$ and the corresponding $\mathcal{K} = \{(\alpha_1, \cdots, \alpha_p, \beta_1, \ldots, \beta_q) \in \mathbb{R}_+^{p+q} | \sum_{i=1}^{p}\alpha_i + \sum_{j=1}^{q}\beta_j \le 1\}$.

An advantage with VTE is that we ensure a consistent estimate of the long-run variance even if the model is misspecified. Also, given γ is well estimated, we reduce the dimension of the parameter space and increase the speed of convergence of the optimization routines. Moreover, the nice geometry of the new set of optimization $\mathcal{K}$ lets the projection step in (3.3) being efficiently implemented following Duchi et al. (2008). However, VTE requires stronger assumptions for the existence of the variance and is likely to suffer from efficiency loss. Francq et al. (2011) also showed that VTE will never be asymptotically more accurate than the QMLE. Another drawback of using VTE is that one needs a finite fourth moment of the process

[1] Let (X_t) follow a GARCH(p,q) process with parameter vector $\theta = (\omega, \alpha_1, \ldots, \alpha_p, \beta_1, \ldots, \beta_q)^T$ and innovations (Z_t). Then for any $\lambda > 0$ the process $(\sqrt{\lambda}X_t)$ is a GARCH(p,q) process with parameter vector $\theta = (\lambda\omega, \alpha_1, \ldots, \alpha_p, \beta_1, \ldots, \beta_q)^T$ and identical innovations (Z_t).

Algorithm 1: Adaptive recursive QML estimation for GARCH(p, q) models using the technique of VTE applied on $(X_t)_{t\geq 1}$ with an initialization $X_{-p+1} = \cdots = X_0 = 0$ and $\hat{\sigma}^2_{-q+1} = \cdots = \hat{\sigma}^2_0 = 0$. The projection onto $\mathcal{K}$ is indicated by $\mathrm{P}_{\mathcal{K}}[\cdot]$. All vector operations are element-wise e.g. $\hat{g}_t^2$ denotes the element-wise square $\hat{g}_t \odot \hat{g}_t$. Good default settings are $\eta = 0.1$ and $\epsilon = 10^{-8}$.

input : $\hat{\theta}_0$ (initial parameter vector)
begin
  initialize: $\hat{\sigma}^2_1 = X^2_1$, $\hat{\mu}_0 = 0$, $\hat{\gamma}^2_0 = 0$, $\hat{G}_0 = 0$ and $t = 0$
  while $\hat{\theta}_t$ **not converged do**
    $t = t + 1$
    $\hat{\mu}_t = t(t+1)^{-1}\hat{\mu}_{t-1} + (t+1)^{-1}X_t$
    $\hat{\gamma}^2_t = (t-1)t^{-1}\hat{\gamma}^2_{t-1} + t^{-1}\,(X_t - \hat{\mu}_t)^2$
    $\hat{g}_t = \nabla \hat{l}_t(\hat{\theta}_{t-1})$
    $\hat{G}_t = \hat{G}_{t-1} + \hat{g}^2_t$
    $\hat{\theta}_t = \mathrm{P}_{\mathcal{K}}\left[\hat{\theta}_{t-1} - \eta(\hat{G}_t + \epsilon)^{-1/2}\hat{g}_t\right]$
    $\hat{\sigma}^2_{t+1} = \hat{\gamma}^2_t + \sum_{i=1}^{p}\hat{\alpha}_i^{(t)}(X^2_{t-i} - \hat{\gamma}^2_t) + \sum_{j=1}^{q}\hat{\beta}_j^{(t)}(\hat{\sigma}^2_{t-j} - \hat{\gamma}^2_t)$
  end
end
Result: $\hat{\theta}_t$ (resulting estimates)

(X_t). Meaning, one would need $\alpha_1 < 0.57$ for an ARCH(1) model using standard Gaussian noise as $EX^4_t < \infty$ if and only if $\alpha^2_1 + (EZ^4_0 - 1)\alpha^2_1 < 1$. For a GARCH(1, 1) model we should have $(\alpha_1 + \beta_1)^2 + (EZ^4_0 - 1)\alpha^2_1 < 1$. These parameter bounds restrict the usefulness and range of applications for our method.

4. Applications

In this section, we apply our recursive method on simulated and real-life data. Our implementation of Algorithm 1 is provided in a repository at Werge (2019). We compare our approach to the non-recursive QMLE approximation $\tilde{\theta}_n$ which is estimated at every two thousand incremental using all observations up to this point i.e. $(\tilde{\theta}_t)_{(k-2000)+1\leq t\leq k}$ is estimated using $(X_t)_{1\leq t\leq k}$ for $k = 2000, 4000, \ldots$. As suggested by Ip et al. (2006), we use the *L-BFGS-B* algorithm to solve the nonlinear optimization problem in (2.1) for $\tilde{\theta}_n$ with initial guess $\tilde{\theta}_0 \in \mathcal{K}$. Our two-step recursive QMLE approximation $\hat{\theta}_n$ is described in Section 3.1 for the GARCH(p, q) model. It takes our initial value $\hat{\theta}_0 \in \mathcal{K}$, learning rate $\eta = 0.1$ and $\epsilon = 10^{-8}$ as input[2]. At last, for a fair comparison, we always use the same initial guess for both methods, namely $\hat{\theta}_0 = \tilde{\theta}_0 \in \mathcal{K}$.

[2]Algorithm 1 can be tuned by changing the learning parameter η e.g. by choosing the best performing learning rate η using the first part of the observations. The choice of learning rates is arduous as a too large learning rate may cause the algorithm to diverge away from the parameter estimate. Contrarily, a too small learning rate may result in slow convergence. However, a small learning rate can be preferred if one only wishes to keep track of minor changes in the parameter estimates.

4.1. Simulations

All simulations are performed by the use of twenty thousand observations ($n = 20000$), and the simulated data (X_t) is always generated using Gaussian innovations with zero mean and unit variance.

4.1.1. ARCH Models

As discussed before, the non-recursive QMLE approximation $\tilde{\theta}_n$ performs badly for numerically small $\omega > 0$ values which is a situation often encountered in financial time series. However, before moving to the case of small ω parameter values then in Figure 1 we have the trajectories of both QMLE approximations using an ARCH(1) process with true parameter vector and initial values given by

$$\theta_0 = \begin{pmatrix}\omega \\ \alpha_1\end{pmatrix} = \begin{pmatrix}2.0 \\ 0.6\end{pmatrix} \text{ and } \hat{\theta}_0 = \tilde{\theta}_0 = \begin{pmatrix}1.5 \\ 0.4\end{pmatrix}. \tag{4.1}$$

Figure 1 shows a very reasonable convergence of both estimators, $\hat{\theta}_n = (\hat{\omega}^{(n)}, \hat{\alpha}_1^{(n)})^T$ and $\tilde{\theta}_n = (\tilde{\omega}^{(n)}, \tilde{\alpha}_1^{(n)})^T$, when the true parameter $\omega = 2.0$. Not surprisingly, our method experience some fluctuations in the beginning but as the learning rate decrease the fluctuation evaporates.

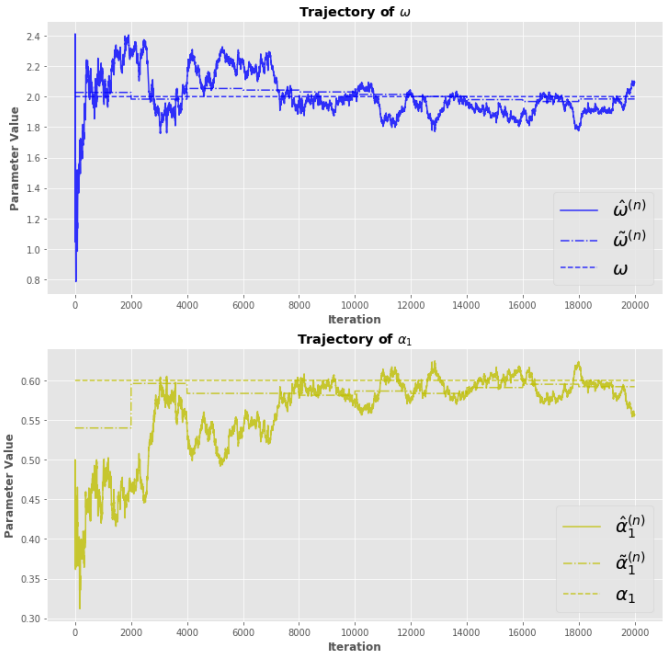

Figure 1: Trajectory of the recursive $\hat{\theta}_n$ (solid line) and non-recursive $\tilde{\theta}_n$ (semi-dotted line) for an ARCH(1) process with true parameter vector (dotted line) and initial guess given in (4.1).

Likewise, in Figure 2 we have the trajectories of the QMLE approximations for an ARCH(1) process but now with true parameter vector and initial guess given as

$$\theta_0 = \begin{pmatrix}1 \cdot 10^{-8} \\ 0.6\end{pmatrix} \text{ and } \hat{\theta}_0 = \tilde{\theta}_0 = \begin{pmatrix}5 \cdot 10^{-8} \\ 0.4\end{pmatrix}. \tag{4.2}$$

Figure 2 indicates a modest convergence of $\hat{\theta}_n$ but shows slow convergence of $\tilde{\alpha}_n$ towards the true α_1 parameter, in addition,

$\tilde{\alpha}_n$ seems bias with respect to the initial value $\tilde{\alpha}_0 = 0.4$ as it processes almost half of the observations before moving closer to the true $\alpha_1 = 0.6$.

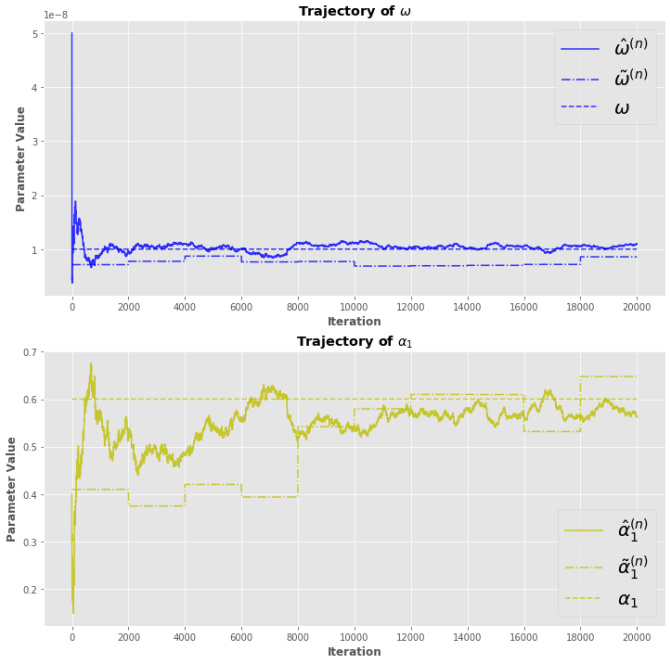

Figure 2: Trajectory of the recursive $\hat{\theta}_n$ (solid line) and non-recursive $\tilde{\theta}_n$ (semi-dotted line) for an ARCH(1) process with true parameter vector (dotted line) and initial guess given in (4.2).

A way of demonstrating the variation in performance of $\hat{\theta}_n$ and $\tilde{\theta}_n$ for small ω values is presented in Figure 3 and Figure 4, respectively, where we have the average trajectory of one hundred trajectories with their corresponding boxplots showing the distribution of these one hundred trajectories. Here, in Figure 3, we can see that our recursive algorithm converges to the true parameter values with low sensitivity respect to the initial values. But one should still expect some likelihood to end up with an irregular solution where the algorithm fails to converge. However, in Figure 4 we see somehow the opposite in which $\tilde{\theta}_n$ have convergence issues; it is consistently underestimating the ω parameter and for some iterations also the α parameter.

As we observe the true volatility process (σ_t) in this section then we can evaluate the accuracy of our recursive $(\hat{\sigma}_t)$ and the non-recursive $(\tilde{\sigma}_t)$ volatility process. We do this by use of the Mean Percentage Errors (MPE), given as

$$\hat{\sigma}_{\text{MPE}} = \frac{1}{n}\sum_{t=1}^{n}\frac{\sigma_t - \hat{\sigma}_t}{\sigma_t} \text{ and } \tilde{\sigma}_{\text{MPE}} = \frac{1}{n}\sum_{t=1}^{n}\frac{\sigma_t - \tilde{\sigma}_t}{\sigma_t}, \qquad (4.3)$$

and the Mean Absolute Percentage Errors (MAPE), given by

$$\hat{\sigma}_{\text{MAPE}} = \frac{1}{n}\sum_{t=1}^{n}\frac{|\sigma_t - \hat{\sigma}_t|}{\sigma_t} \text{ and } \tilde{\sigma}_{\text{MAPE}} = \frac{1}{n}\sum_{t=1}^{n}\frac{|\sigma_t - \tilde{\sigma}_t|}{\sigma_t}. \qquad (4.4)$$

Boxplots of one hundred accuracy scores, MPE in (4.3) and MAPE in (4.4), can be found in Figure 5. To avoid bias due to the choice of the true parameter vector θ_0 and initial values

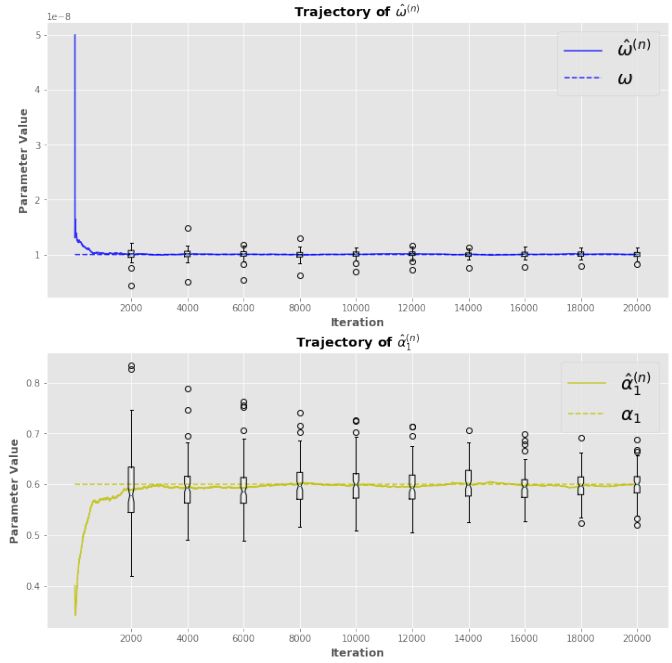

Figure 3: Average trajectory (solid line) of one hundred recursive $\hat{\theta}_n$'s for an ARCH(1) process with true parameter vector (dotted line) and initial guess from (4.2). The boxplots shows the distribution of the one hundred trajectories.

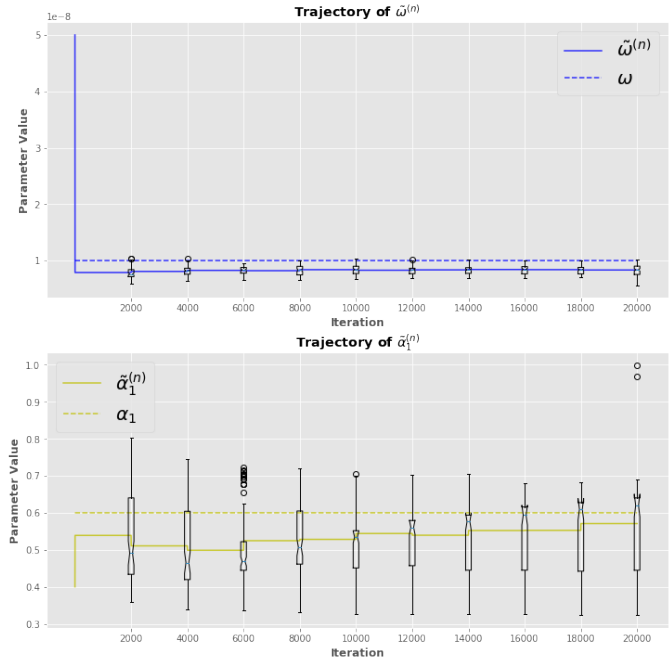

Figure 4: Average trajectory (solid line) of one hundred non-recursive $\tilde{\theta}_n$'s for an ARCH(1) process with true parameter vector (dotted line) and initial guess from (4.2). The boxplots shows the distribution of the one hundred trajectories.

$\hat{\theta}_0, \tilde{\theta}_0$, we then have the accuracy scores with a random parameter vector $\theta_0 \in \mathcal{K}$ and random initial guesses $\hat{\theta}_0, \tilde{\theta}_0 \in \mathcal{K}$ in Figure 5. In the top graph of Figure 5, one can observe that the MPE for both methods is symmetric around zero but $\tilde{\sigma}_{\text{MPE}}$ has a negative tail (meaning the non-recursive method may overes-

timate the volatility in some cases). Also, the spread of $\tilde{\sigma}_{\text{MPE}}$ is higher than the $\hat{\sigma}_{\text{MPE}}$, which is clearly seen by looking at $\tilde{\sigma}_{\text{MAPE}}$ in the bottom graph of Figure 5.

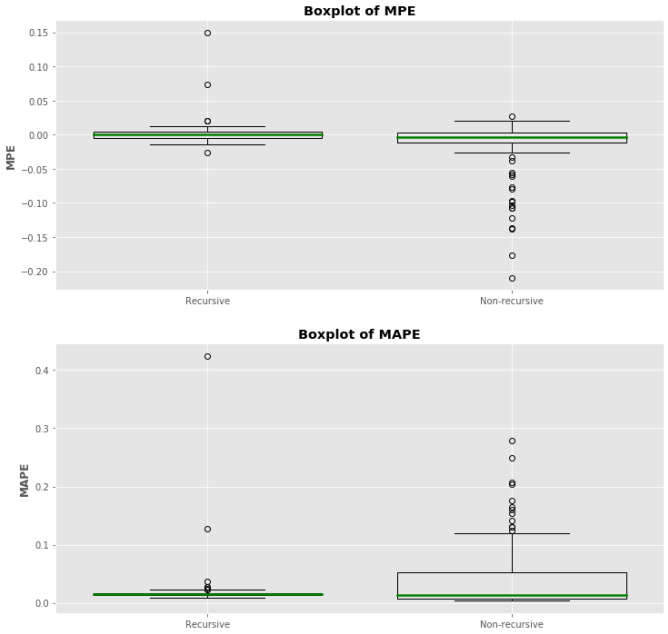

Figure 5: Boxplots of accuracy scores, MPE from (4.3) and MAPE from (4.4), for one hundred trajectories of $\hat{\theta}_n$ and $\tilde{\theta}_n$ using an ARCH(1) process with random true parameter vector and initial guess in $\mathcal{K}$.

Another way of measuring the accuracy can be made by studying the conditional quantiles using the recursive ($\hat{\sigma}_t$) and non-recursive ($\tilde{\sigma}_t$) predicted volatility processes (See Biau and Patra (2011)). Under the assumption of standard Gaussian innovations then X_t is Gaussian with zero mean and variance σ_t^2. Thus, for any $\alpha \in (0,1)$, the α-quantile of a Gaussian distribution $\mathcal{N}(0,\sigma_t^2)$ is $\sigma_t\Phi^{-1}(\alpha)$, where $\Phi^{-1}(\alpha)$ is the α-quantile of the standard Gaussian one. We use the so-called α-quantile loss function proposed by Koenker and Bassett (1978): The α-quantile loss function ρ_α using the volatility process σ_t is defined as

$$\rho_\alpha(X_t,\sigma_t) = \begin{cases} \alpha\left(X_t - \Phi^{-1}(\alpha)\sigma_t\right), & \text{for } X_t > \Phi^{-1}(\alpha)\sigma_t \\ (1-\alpha)\left(\Phi^{-1}(\alpha)\sigma_t - X_t\right), & \text{for } X_t \le \Phi^{-1}(\alpha)\sigma_t \end{cases} \tag{4.5}$$

with tilting parameter $\alpha \in (0,1)$. The idea of the α-quantile loss function is to penalize quantiles of low probability more for overestimation than for underestimation (and contrariwise in the case of high probability quantiles). We evaluate across the α-quantile scores ρ_α of (σ_t) by the (normalized) cumulative α-quantile scoring function QS_α:

$$QS_\alpha(X_n,\sigma_n) = \frac{1}{n}\sum_{t=1}^{n}\sum_{m=1}^{M}\rho_{\alpha_m}(X_t,\sigma_t), \tag{4.6}$$

with M as the number of quantiles $\alpha = \{\alpha_1,\dots,\alpha_M\}$. The best ability of volatility forecast is indicated by the lowest QS_α score. The findings of one hundred $QS_\alpha(X_n,\hat{\sigma}_n)$ and $QS_\alpha(X_n,\tilde{\sigma}_n)$ scores, with $\alpha = \{0.01, 0.02, \dots, 0.99\}$ and random true parameter vector and random initialization in $\mathcal{K}$, is presented in Figure 6. The QS_α scores in Figure 6 are indistinguishable.

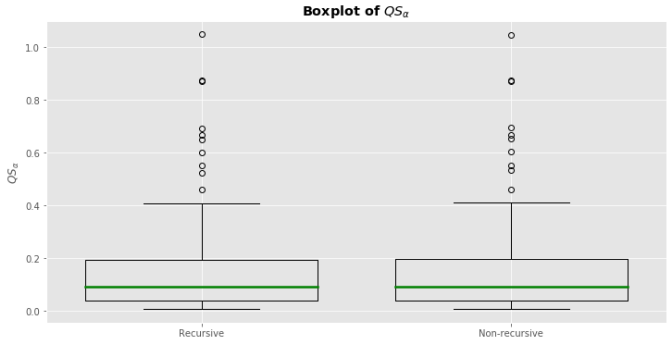

Figure 6: Boxplots of one hundred QS_α scores using the recursive $\hat{\sigma}_t$ and non-recursive $\tilde{\sigma}_t$ volatility process, respectively, for $\alpha = \{0.01, 0.02, \dots, 0.99\}$, for an ARCH(1) model with random true parameter vector and initial value in $\mathcal{K}$.

4.1.2. GARCH Models

We have similar findings for the GARCH(1, 1) model presented in Figure 7 for the recursive $\hat{\theta}_n$ and Figure 8 for the non-recursive $\tilde{\theta}_n$ with true parameter vector and initial guess given by

$$\theta_0 = \begin{pmatrix}\omega\\ \alpha_1\\ \beta_1\end{pmatrix} = \begin{pmatrix}1\cdot 10^{-8}\\ 0.2\\ 0.7\end{pmatrix} \text{ and } \hat{\theta}_0 = \tilde{\theta}_0 = \begin{pmatrix}5\cdot 10^{-8}\\ 0.1\\ 0.8\end{pmatrix}. \tag{4.7}$$

That said, the non-recursive $\tilde{\theta}_n$ is consistently overestimating the β parameter. It is worth mentioning that even if all initial values are chosen in the stationary region i.e. $\hat{\theta}_0 = \tilde{\theta}_0 = \theta_0$, we still have a proper amount of fluctuation in our estimates trajectories. As discussed before, this may partially be due to the volatility the stochastic gradient descent introduce and the flatness of the QL loss (See Zumbach (2000)).

The accuracy scores, namely MPE in (4.3) and MAPE in (4.4), can be found in Figure 9 for the GARCH(1, 1) model using both random true parameter vector and random initial values in $\mathcal{K}$. As in the ARCH(1) case, we obtain lower spread for $\hat{\sigma}_{\text{MPE}}$ than $\tilde{\sigma}_{\text{MPE}}$.

Figure 10 present the results of one hundred QS_α scores with random true parameter vector and initial value in $\mathcal{K}$. Again, the QS_α scores are indistinguishable (even when the non-recursive method is forward looking).

4.2. Real-life Observations

We demonstrate our method on real-life observations, showing how our technique works in practice. Table 1 shows an overview of the used stock market indices. All empirical studies are made using the GARCH(1, 1) model but parameters of higher-order may yield a better fit. As the observation period spans over a long time then it is unlikely that the log-return series is stationary. To exhibit our method ability to adapt to time-varying estimates then we begin by considering the S&P500 In-

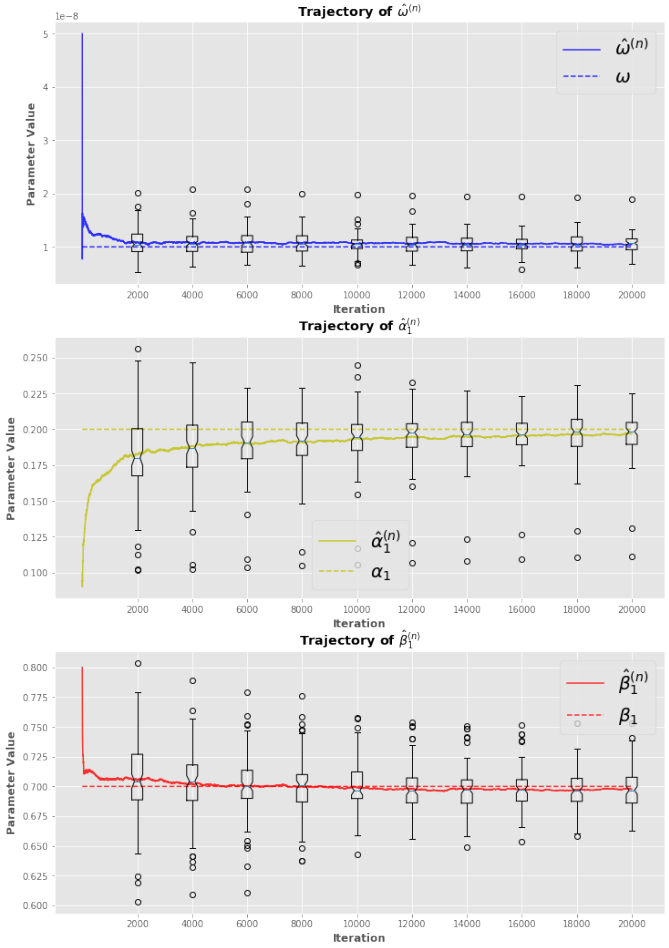

Figure 7: Average trajectory (solid line) of one hundred $\hat{\theta}_n$'s for a GARCH(1, 1) process with true parameter vector (dotted line) and initial guess given in (4.7). The boxplots shows the distribution of the one hundred trajectories.

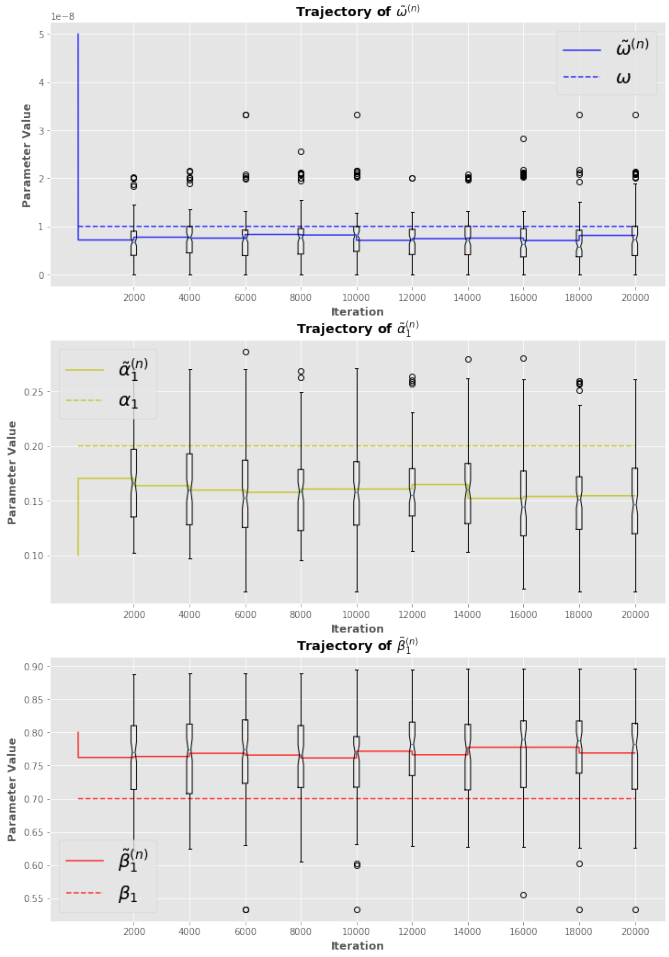

Figure 8: Average trajectory (solid line) of one hundred $\tilde{\theta}_n$'s for a GARCH(1, 1) process with true parameter vector (dotted line) and initial guess given in (4.7). The boxplots shows the distribution of the one hundred trajectories.

dex in Section 4.2.1. Thereafter, in Section 4.2.2, we investigate the remaining six stock market indices presented in Table 1.

Stock Market Index		Period (years)
CAC 40	(CAC)	1991-2020
DAX 30	(DAX)	1988-2020
Dow Jones Industrial Average	(DJIA)	1986-2020
NASDAQ Composite	(NDAQ)	1971-2020
Nikkei 225	(NKY)	1965-2020
Russell 2000	(RUT)	1988-2020
Standard & Poor's 500	(S&P500)	1950-2020

Table 1: Overview of considered stock market indices including their observation periods. The observations consist of daily log-returns which are defined as log differences of the closing prices of the index between two consecutive days.

4.2.1. Application to the S&P500 Index

We apply our method on the S&P500 Index for the years 1950 to 2020 (consisting of $n = 17505$ observations) to test the performance on real-life data. We use the GARCH(1, 1) model with initial values:

$$\hat{\theta}_0 = \tilde{\theta}_0 = \begin{pmatrix} 5 \cdot 10^{-5} \\ 0.05 \\ 0.9 \end{pmatrix}. \tag{4.8}$$

The QML trajectories can been seen in Figure 11: Our recursive approximation $\hat{\theta}_n = (\hat{\omega}^{(n)}, \hat{\alpha}_1^{(n)}, \hat{\beta}_1^{(n)})^T$ fluctuates more than the QMLE approximation $\tilde{\theta}_n = (\tilde{\omega}^{(n)}, \tilde{\alpha}_1^{(n)}, \tilde{\beta}_1^{(n)})^T$ which is estimated incremental for every two thousand observation. Remember, the fluctuations we experience in the recursive method can be reduced by lowering the learning rate η. It is remarkable that the QMLE approximation $\tilde{\theta}_n$ estimate $\tilde{\beta}_1^{(n)}$ is so low in some years between 1990 and 2000 even when it is estimated using over half of the observations.

In Figure 12, we have the log-returns r_t of the S&P500 Index and the confidence intervals $\bar{r} \pm 1.96\hat{\sigma}_t$ and $\bar{r} \pm 1.96\tilde{\sigma}_t$ using the recursive $\hat{\sigma}_t$ and non-recursive $\tilde{\sigma}_t$ predicted volatilities, respectively, where $\bar{r}$ is the mean of the log-returns r_t. It seems that the recursive method $\hat{\sigma}_t$ adapts more rapidly than the non-recursive one $\tilde{\sigma}_t$ to changes in the S&P500 Index observations r_t.

The efficiency of our recursive ($\hat{\sigma}_t$) and the non-recursive ($\tilde{\sigma}_t$) volatility can be appraised with use of the squared log-

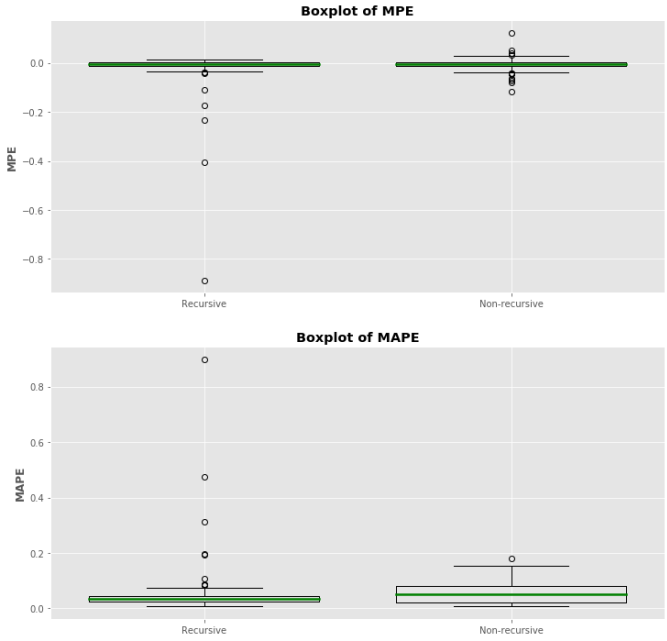

Figure 9: Boxplots of accuracy scores, MPE from (4.3) and MAPE from (4.4), for one hundred trajectories of $\hat{\theta}_n$ and $\tilde{\theta}_n$ using a GARCH(1, 1) process with true parameter vector and random initial guess in $\mathcal{K}$.

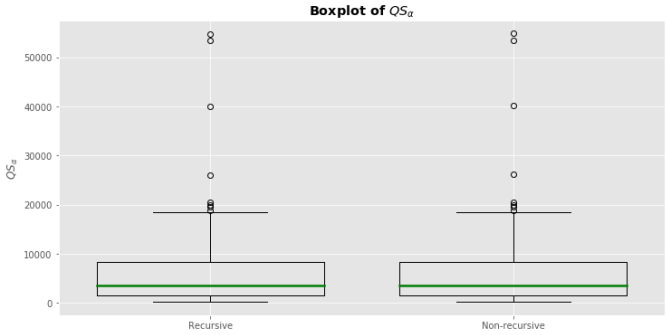

Figure 10: Boxplots of one hundred QS_α scores using the recursive $\hat{\sigma}_t$ and non-recursive $\tilde{\sigma}_t$ volatility process, respectively, for $\alpha = \{0.01, 0.02, \ldots, 0.99\}$, using the GARCH(1, 1) model with random true parameter vector and initial value in $\mathcal{K}$.

returns (r_t^2) in absence of the true (unobserved) variance process (σ_t^2). In Table 2, we have the Mean Absolute Errors (MAE) defined by

$$\hat{\sigma}^2_{\text{MAE}} = \frac{1}{n}\sum_{t=1}^{n}|r_t^2 - \hat{\sigma}_t^2| \text{ and } \tilde{\sigma}^2_{\text{MAE}} = \frac{1}{n}\sum_{t=1}^{n}|r_t^2 - \tilde{\sigma}_t^2|, \quad (4.9)$$

for the same periods used in Figure 12, including for the full dataset. The results in Table 2 confirm our conclusions about Figure 12; the recursive method tracks the volatility better than the non-recursive method.

Figure 13 contain the results of one hundred QS_α scores using the recursive ($\hat{\sigma}_t$) and non-recursive ($\tilde{\sigma}_t$) volatility process, respectively, with random initial values in $\mathcal{K}$. It is remarkable that the recursive method outperforms the non-recursive method although the latter uses future information i.e. $(\tilde{\theta}_t)_{(k-2000)+1\le t\le k}$ is estimated using $(r_t)_{1\le t\le k}$ for $k = 2000, 4000, \ldots, 16000, 17505$. Thus, indicating one could obtain proper performance using the recursive method which predicts the volatility only by use of the previous estimate.

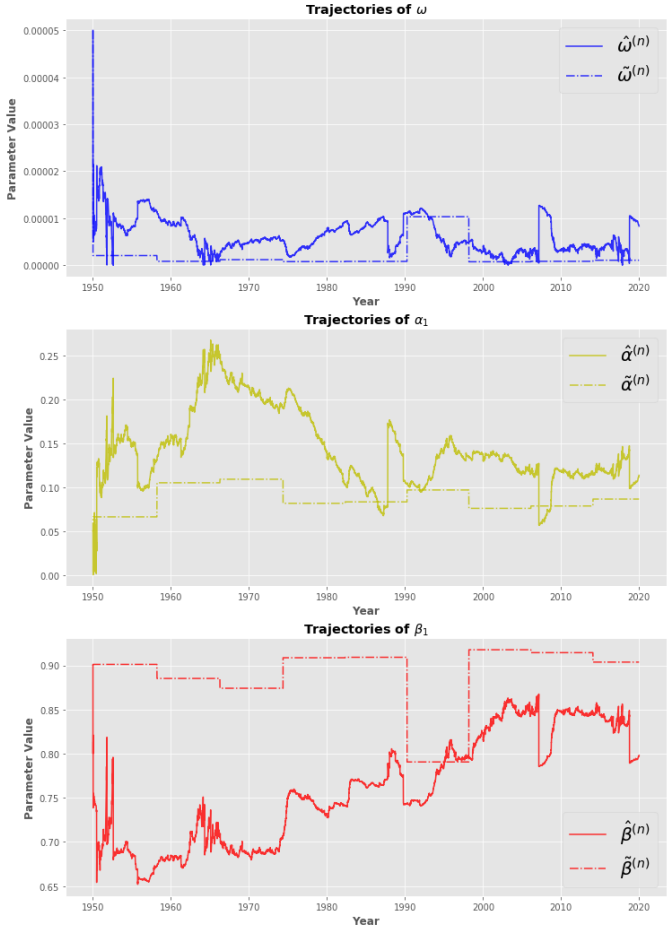

Figure 11: Trajectory of recursive QML estimates $\hat{\theta}_n$ (solid line) and non-recursive $\tilde{\theta}_n$ (semi-dotted line) using a GARCH(1, 1) model on S&P500 Index log-returns from year 1950 to 2020. Both methods use initial value from (4.8).

Periods (years)	Recursive $\hat{\sigma}^2_{\text{MAE}}$	Non-recursive $\tilde{\sigma}^2_{\text{MAE}}$
1950-1952	7.9177	8.5628
1985-1987	7.1374	7.4949
2018-2020	9.6666	9.7905
1950-2020	9.7360	10.1409

Table 2: MAEs (4.9) using log-returns r_t of S&P500 Index with the recursive $\hat{\sigma}_t$ and non-recursive $\tilde{\sigma}_t$ predicted volatilities, respectively. Both methods has initial value given in (4.8). The $\hat{\sigma}^2_{\text{MAE}}$ and $\tilde{\sigma}^2_{\text{MAE}}$ numbers are scaled by 10^{-5}.

4.2.2. *Other Stock Market Indices*

The results of one hundred QS_α scores using the recursive ($\hat{\sigma}_t$) and non-recursive ($\tilde{\sigma}_t$) volatility process, respectively, with random initial values in $\mathcal{K}$ is presented in Figure 14 for the remaining six stock market indices in Table 1 (i.e. the CAC, DAX, DJIA, NDAQ, NKY and RUT index). As for the S&P500 Index (See Figure 13), these findings indicate that the recursive approach estimate the quantiles better than the non-recursive method, both on average but also with a lower spread.

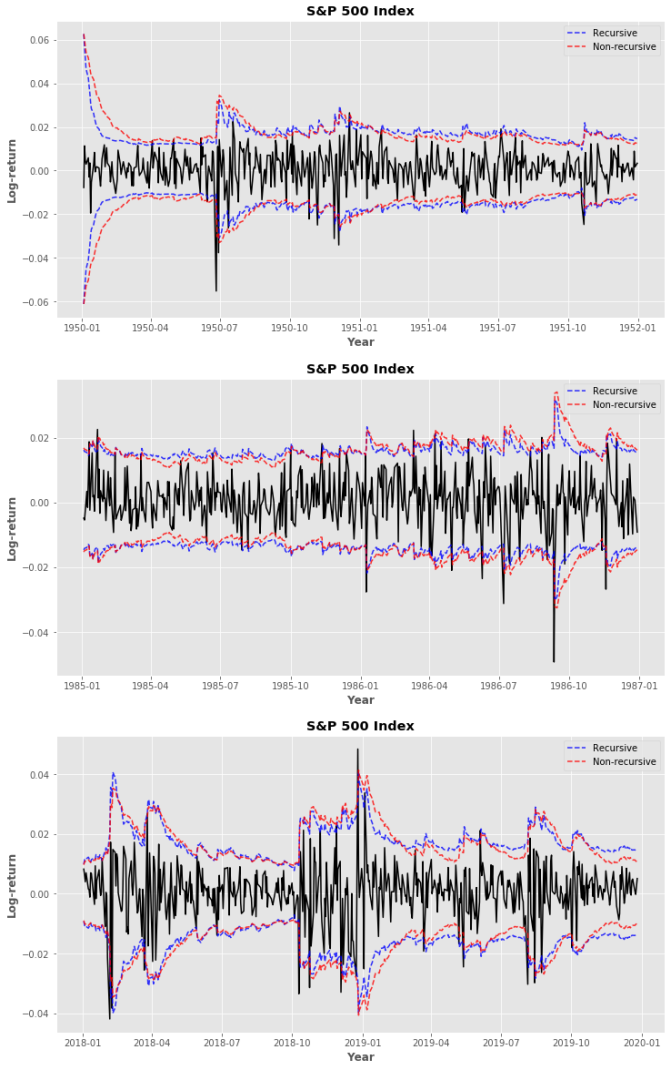

Figure 12: Log-returns r_t of S&P500 Index (solid lines) and confidence intervals $\bar{r} \pm 1.96\hat{\sigma}_t$ and $\bar{r} \pm 1.96\tilde{\sigma}_t$ (dotted lines) using the recursive $\hat{\sigma}_t$ (blue) and non-recursive $\tilde{\sigma}_t$ (red) predicted volatilities, respectively, where $\bar{r}$ is the mean of the log-returns r_t. From top to bottom we have the periods: year 1950 to 1952, 1985 to 1987 and 2018 to 2020.

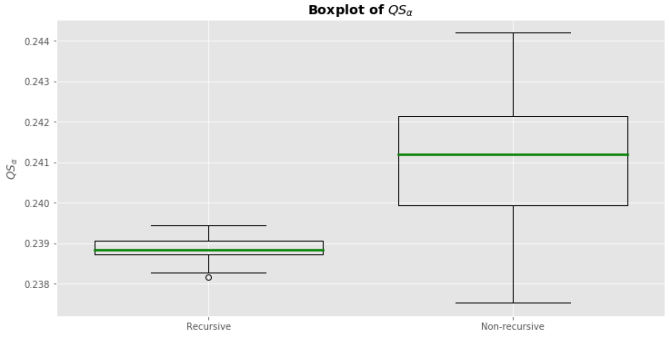

Figure 13: Boxplots of one hundred QS_α scores with use of the recursive $\hat{\sigma}_t$ and non-recursive $\tilde{\sigma}_t$ volatility process, respectively, for $\alpha = \{0.01, 0.02, \ldots, 0.99\}$, using the GARCH(1, 1) model on the log-returns r_t of S&P500 Index with random initial value in $\mathcal{K}$.

The assumption of having an underlying data generation process with constant "true" parameters may not hold in real-life examples. Thus, our recursive method seems to have an advantage compared to the non-recursive method, as it estimates the parameters step-by-step whereas the non-recursive method always has to estimate the parameters using all observations over a large period of time.

5. Discussion

We proved asymptotic local convexity of the QL function in a general conditionally heteroscedastic time series model of multiplicative form. An interesting question arises: can one prove Theorem 2.1 for a bounded set of N observations? Expressed differently, can one find a N bounded, such that we have convergence/convexity of recursive algorithms e.g. for the GARCH, EGARCH and AGARCH models. To our knowledge, this is not been proved yet.

We proposed an adaptive approach to recursively estimate the parameters of GARCH models in an online setting with use of the VTE technique (See Algorithm 1). We obtain a more stable, reliable and adaptive method. We know that the stability of using our recursive approach to solve the QML problem could be improved by using a mini-batch approach. This would be lowering the volatility in each incremental as one use more observations per iteration to update the QML estimate.

Furthermore, applying a mini-batch method does not require much more computational power compared to the stochastic gradient descent, and by using more observations we could get more consistency and smoothness in the convergence of the estimation procedure. The size of the mini-batch to use is left to future research work.

Appendix A.

Proof of Theorem 2.1. To prove local strong convexity for the approximate QL function $\hat{L}_n$ using the approximate QMLE $\hat{\theta}_n^*$ we first list some bounds for the Hessians: Under the regularity conditions on the derivatives of h_t, then by use of (2.3) we can write

$$\nabla l_t(\theta) = \frac{1}{2}\frac{\nabla h_t(\theta)}{h_t(\theta)}\left(1 - \frac{X_t^2}{h_t(\theta)}\right),$$
$$\nabla^2 l_t(\theta) = \frac{1}{2h_t^2(\theta)}\left(\nabla h_t(\theta)^T \nabla h_t(\theta)\left(\frac{2X_t^2}{h_t(\theta)} - 1\right) + \nabla^2 h_t(\theta)\left(h_t(\theta) - X_t^2\right)\right),$$

where the Hessian $H_n(\theta)$ is defined as $n^{-1}\nabla^2 L_n(\theta) = n^{-1}\sum_{t=1}^n \nabla^2 l_t(\theta)$. Similarly, for $\nabla\hat{l}_t(\theta)$, $\nabla^2\hat{l}_t(\theta)$ and $\hat{H}_n(\theta)$ we replace $h_t(\theta), \nabla h_t(\theta)$ and $\nabla^2 h_t(\theta)$ by $\hat{h}_t(\theta), \nabla\hat{h}_t(\theta)$ and $\nabla^2\hat{h}_t(\theta)$, respectively. From Assumption W2, we know $n^{-1}\|\nabla^2\hat{L}_n - \nabla^2 L_n\|_{\mathcal{K}} \xrightarrow{\text{a.s.}} 0$ for $n \to \infty$. Hence, for some random N_1 large enough there exists $\epsilon > 0$ such that $n^{-1}\|\nabla^2\hat{L}_n - \nabla^2 L_n\|_{\mathcal{K}} < \epsilon$ for all $n \geq N_1$ a.s. As a consequence, we get

$$\|\hat{H}_n - H_n\|_{\mathcal{K}} < \epsilon, \quad \text{a.s.}, \tag{A.1}$$

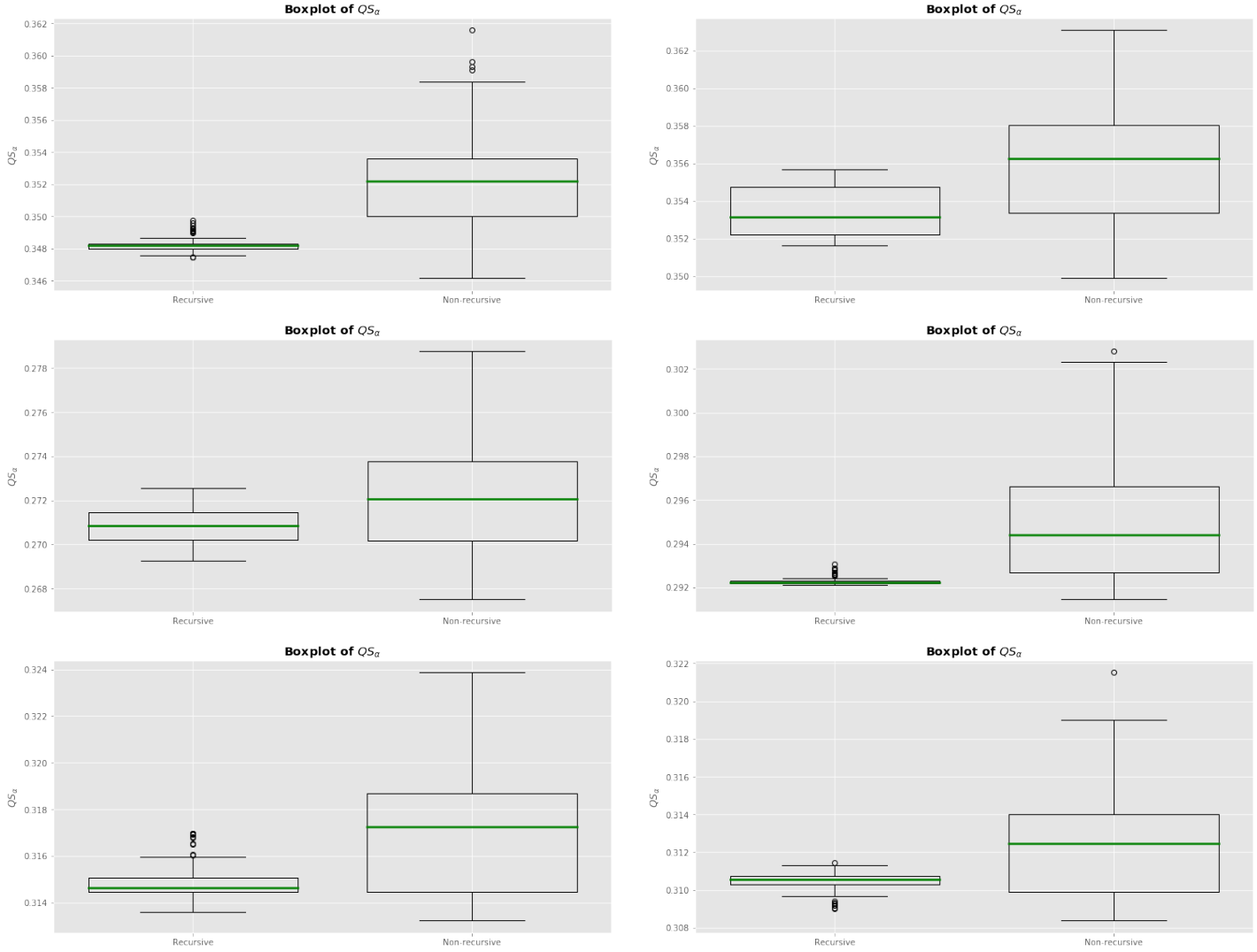

Figure 14: Boxplots of one hundred QS_α scores with use of the recursive $\hat{\sigma}_t$ and non-recursive $\tilde{\sigma}_t$ volatility process, respectively, for $\alpha = \{0.01, 0.02, \ldots, 0.99\}$, using the GARCH(1, 1) model on the log-returns r_t of the CAC (top left), DAX (top right), DJIA (mid left), NDAQ (mid right), NKY (bottom left) and RUT (bottom right) index with random initial value in $\mathcal{K}$.

for all $n \geq N_1$. Similarly, applying the ergodic theorem on the integrable sequence (uniformly over $\mathcal{K}$) $(\nabla^2 l_t)$ of continuous functions over the compact set $\mathcal{K}$, we obtain $\|n^{-1}\sum_{t=1}^{n} \nabla^2 l_t - \mathbb{E}[\nabla^2 l_0]\|_{\mathcal{K}} \xrightarrow{\text{a.s.}} 0$ for $n \to \infty$. Then there exists N_2 such that

$$\|H_n - H_0\|_{\mathcal{K}} < \epsilon, \quad \text{a.s.}, \tag{A.2}$$

for all $n \geq N_2$. Thus, by equation (A.1) and (A.2) we know there exists $N = \max(N_1, N_2)$ such that for all $n \geq N$ we have

$$\|\hat{H}_n - H_0\|_{\mathcal{K}} \leq \|\hat{H}_n - H_n\|_{\mathcal{K}} + \|H_n - H_0\|_{\mathcal{K}} < 2\epsilon, \quad \text{a.s.}$$

Especially, as $\|\hat{H}_n - H_0\|_{\mathcal{K}}$ is defined as $\sup_{\theta\in\mathcal{K}}\|\hat{H}_n(\theta) - H_0(\theta)\|_{op}$ then

$$\|\hat{H}_n(\theta) - H_0(\theta)\|_{op} < 2\epsilon, \tag{A.3}$$

for all $\theta \in \mathcal{K}$.

From (Straumann and Mikosch, 2006, Lemma 7.2), the asymptotic Hessian $H_0(\theta_0) = \mathbb{E}[\nabla^2 l_0(\theta_0)]$ is a symmetric positive definite matrix a.s. under Assumption W3. As $H_0(\theta)$ is the limit of the continuous matrix-valued function $H_n(\theta)$ then it is itself a continuous matrix-valued function. Thus, the eigenvalue function $\lambda_0^i(\theta)$ for $1 \leq i \leq d$ of $H_0(\theta)$ is also continuous. The eigenvalues $\lambda_0^i(\theta_0)$ are positive real numbers with the smallest one $\lambda_0^{\min}(\theta_0)$ denoted by

$$\lambda_0^{\min}(\theta_0) = \min_{1\leq i\leq d} \lambda_0^i(\theta_0) > 0,$$

satisfying $g^T H_0(\theta_0) g \geq \lambda_0^{\min}(\theta_0) g^T g$ for all $g \in \mathbb{R}^d \setminus \{0\}$.

To shorten the notation we write with no ambiguity $H_0(\theta_0) \geq \lambda_0^{\min}(\theta_0) I_d$ where I_d denote the d-dimensional identity matrix. Note that by continuity $\lambda_0^{\min}(\theta)$ is also positive on a neighborhood $B(\theta_0, \delta)$ so that $\exists \epsilon > 0$ satisfying $\lambda_0^{\min}(\theta_0) - \epsilon > 0$, meaning

$$H_0(\theta) \geq (\lambda_0^{\min}(\theta_0) - \epsilon) I_d,$$

for $\theta \in B(\theta_0, \delta)$. Hence, for $\theta \in B(\theta_0, \delta)$ and $g \in \mathbb{R}^d \setminus \{0\}$, we have

$$\begin{aligned}\frac{g^T \hat{H}_n(\theta)}{g^T g} &= \frac{g^T H_0(\theta) g}{g^T g} + \frac{g^T \left(\hat{H}_n(\theta) - H_0(\theta)\right) g}{g^T g} \\ &\geq \lambda_{\min} - \epsilon - \frac{g^T \|\hat{H}_n(\theta) - H_0(\theta)\|_{op} g}{g^T g} \\ &> \lambda_{\min} - 3\epsilon \\ &> C, \quad \text{a.s.},\end{aligned}$$

with use of (A.3) for all $n \geq N$ by taking $0 < \epsilon < 6^{-1}\lambda_{\min}$ and letting $C = 2^{-1}\lambda_{\min}$. Then we have the desired inequality (2.5). □

Proof of Corollary 2.1. The uniqueness of the QMLE $\hat{\theta}_n^*$ follows from a Pfanzagl argument (See Pfanzagl (1969)). By Theorem 2.1, we know there exists N such that

$$\inf_{\theta \in B(\theta_0, \delta_0)} g^T \hat{H}_n(\theta) g > C g^T g, \quad \text{a.s.},$$

for all $n \geq N$ where $B(\theta_0, \delta_0)$ denotes the open ball around θ_0 with radius $\delta_0 > 0$. For each element $\theta_i \in \mathcal{K}$ we make an open ball $B(\theta_i, \delta_i)$ for $\delta_i > 0$ such that the union of $B(\theta_i, \delta_i)$ for all i only contains θ_0 once, i.e. $\theta_0 \notin B(\theta_i, \delta_i)$ for $i \neq 0$. As $\mathcal{K}$ is compact and contained in the union of all $B(\theta_i, \delta_i)$ then there is a finite covering of $\mathcal{K}$, i.e. $\mathcal{K} \subseteq \bigcup_{i=0}^{k} B(\theta_i, \delta_i)$. Let $\mathcal{K}' = \mathcal{K} \setminus B(\theta_0, \delta_0)$. As $\mathcal{K}'$ is compact then the minimum of the continuous QL function $\mathbb{E}[l_0]$ exists. Moreover, as $\mathbb{E}[l_0]$ is a unique minimum at θ_0 under Assumption W1, we get

$$\inf_{\theta \in \mathcal{K}'} \mathbb{E}[l_0(\theta)] > \mathbb{E}[l_0(\theta_0)] \quad \text{a.s.}$$

From Assumption W2, we know that $\|n^{-1}\hat{L}_n - L_0\|_{\mathcal{K}'} \xrightarrow{\text{a.s.}} 0$ as $n \to \infty$. Hence, we have

$$\inf_{\theta \in \mathcal{K}'} n^{-1}\hat{L}_n(\theta) \xrightarrow{\text{a.s.}} \inf_{\theta \in \mathcal{K}'} L_0(\theta),$$

where $\inf_{\theta \in \mathcal{K}'} L_0(\theta) > \mathbb{E}[l_0(\theta_0)]$. Thus, the $B(\theta_0, \delta_0)$ gives us a unique global minimum of the QL function $\hat{L}_n$, i.e.

$$\inf_{\theta \in \mathcal{K}} n^{-1}\hat{L}_n(\theta) \geq \mathbb{E}[l_0(\theta_0)], \quad \text{a.s.},$$

where equality is only attained when $\theta = \theta_0$. □

References

Aknouche, A., Guerbyenne, H., 2006. Recursive estimation of garch models. Communications in Statistics - Simulation and Computation 35, 925–938.

Berkes, I., Horváth, L., Kokoszka, P., 2003. GARCH processes: structure and estimation. Bernoulli 9(2), 201–227.

Biau, G., Patra, B., 2011. Sequential quantile prediction of time series. Information Theory, IEEE Transactions on 57, 1664 – 1674.

Bollerslev, T., 1986. Generalized autoregressive conditional heteroscedasticity. Journal of Econometrics 31(3), 307–327.

Bottou, L., Bousquet, O., 2007. The tradeoffs of large scale learning. Advances in Neural Information Processing Systems (NIPS) 20, 161–168.

Bougerol, P., 1993. Kalman filtering with random coefficients and contractions. SIAM Journal on Control and Optimization 31(4), 942–959.

Bougerol, P., Picard, N., 1992. Stationarity of GARCH processes and of some nonnegative time series. Journal of Econometrics 52(1-2), 115–127.

Cipra, T., Hendrych, R., 2018. Robust recursive estimation of garch models. Kybernetika -Praha 54, 1138–1155.

Dahlhaus, R., Subba Rao, S., 2007. A recursive online algorithm for the estimation of time-varying arch parameters. Bernoulli 13, 389–422.

Duchi, J., Hazan, E., Singer, Y., 2011. Adaptive subgradient methods for online learning and stochastic optimization. Journal of Machine Learning Research 12, 2121–2159.

Duchi, J., Shalev-Shwartz, S., Singer, Y., Chandra, T., 2008. Efficient projections onto the l1-ball for learning in high dimensions. Proceedings of the 25th International Conference on Machine Learning , 272–279.

Engle, R., 1982. Autoregressive conditional heteroscedasticity with estimates of the variance of the united kingdom inflation. Econometrica 50(4), 987–1008.

Francq, C., Zakoïan, J.M., 2004. Maximum likelihood estimation of pure garch and arma-garch processes. Bernoulli 10, 605–637.

Francq, C., Zakoïan, J.M., Horvath, L., 2011. Merits and drawbacks of variance targeting in garch models. Journal of Financial Econometrics 9, 619–656.

Gerencsér, L., Orlovits, Z., Torma, B., 2010. Recursive estimation of garch processes, in: The 19th International Symposium on Mathematical Theory of Networks and Systems, (MTNS 2010), Budapest, Hungary, forthcoming, pp. 2415–2422.

Hendrych, R., Cipra, T., 2018. Self-weighted recursive estimation of garch models. Communications in Statistics - Simulation and Computation 47, 315–328.

Ip, W.C., Wong, H., Pan, J., Li, D., 2006. The asymptotic convexity of the negative likelihood function of garch models. Computational Statistics & Data Analysis 50, 311–331.

Kierkegaard, J., Jensen, L., Madsen, H., 2000. Estimating garch models using recursive methods.

Kingma, D., Ba, J., 2015. Adam: A method for stochastic optimization. International Conference on Learning Representations (ICLR) .

Koenker, R.W., Bassett, G., 1978. Regression quantiles. Econometrica 46, 33–50.

Nelson, D., 1990. Stationarity and persistence in the garch(1,1) model. Econometric Theory 6, 318–334.

Patton, A., 2006. Volatility forecast comparison using imperfect volatility proxies. Journal of Econometrics 160, 246–256.

Pfanzagl, J., 1969. On the measurability and consistency of minimum contrast estimates. Metrika 14, 249–272.

Robbins, H., Monro, S., 1951. A stochastic approximation method. Annals of Mathematical Statistics 22, 400–407.

Straumann, D., 2005. Maximum Likelihood Estimation in Conditionally Heteroscedastic Time Series Models. chapter 5. pp. 85–140.

Straumann, D., Mikosch, T., 2006. Quasi-maximum-likelihood estimation in conditionally heteroscedastic time series: A stochastic recurrence equations approach. Annals of Statistics 34(5), 2449–2495.

Tieleman, T., Hinton, G., 2012. Lecture 6.5-rmsprop, coursera: Neural networks for machine learning. University of Toronto, Technical Report.

Ward, R., Wu, X., Bottou, L., 2018. Adagrad stepsizes: Sharp convergence over nonconvex landscapes, from any initialization. `arXiv:1806.01811`.

Werge, N., 2019. Adavol. GitHub repository URL: `https://github.com/nhwerge/AdaVol.git`.

Wintenberger, O., 2013. Continuous invertibility and stable qml estimation of the egarch(1,1) model. Scandinavian Journal of Statistics 40, 846–867.

Zeiler, M.D., 2012. Adadelta: An adaptive learning rate method. `arXiv:1212.5701`.

Zinkevich, M., 2003. Online convex programming and generalized infinitesimal gradient ascent. In Proceedings of the Twentieth International Conference on Machine Learning 2, 928–936.

Zumbach, G., 2000. The pitfalls in fitting garch (1, 1) processes, in: Advances in Quantitative Asset Management. Springer, pp. 179–200.