Score Permutation Based Finite Sample Inference for Generalized AutoRegressive Conditional Heteroskedasticity (GARCH) Models

Balázs Csanád Csáji
Institute for Computer Science and Control (SZTAKI)
Hungarian Academy of Sciences (MTA)

Abstract

A standard model of (conditional) heteroscedasticity, i.e., the phenomenon that the variance of a process changes over time, is the Generalized AutoRegressive Conditional Heteroskedasticity (GARCH) model, which is especially important for economics and finance. GARCH models are typically estimated by the Quasi-Maximum Likelihood (QML) method, which works under mild statistical assumptions. Here, we suggest a finite sample approach, called ScoPe, to construct distribution-free confidence regions around the QML estimate, which have exact coverage probabilities, despite no additional assumptions about moments are made. ScoPe is inspired by the recently developed Sign-Perturbed Sums (SPS) method, which however cannot be applied in the GARCH case. ScoPe works by perturbing the score function using randomly permuted residuals. This produces alternative samples which lead to exact confidence regions. Experiments on simulated and stock market data are also presented, and ScoPe is compared with the asymptotic theory and bootstrap approaches.

1 INTRODUCTION

Homoscedastic noises, such as i.i.d. variables, are widely used in learning theory (Vapnik, 1998; Hastie et al., 2009), though most real-world phenomena, from social systems and stock markets to telecommunications and ECG signals, can be better described by *heteroscedastic* models as their variances change over time. It is typical that larger disturbances are more likely followed by larger disturbances, while smaller fluctuations tend to be followed by smaller fluctuations. In finance, for example, this phenomenon is called *volatility clustering* and it is a widely-known feature of financial time series (Francq and Zakoian, 2011).

AutoRegressive Conditional Heteroscedasticity (ARCH) processes are standard models of this phenomenon. They were introduced by Robert F. Engle (1982) for which he was awarded the Nobel Prize for Economics in 2003 (jointly with Clive W. J. Granger) "*for methods of analyzing economic time series with time-varying volatility (ARCH)*". The ARCH model was later exetended by Bollerslev (1986) who introduced its *generalized* version, called GARCH (see Section 2). Since then, various other generalizations have also been proposed, but GARCH is still one of the most widely used models (Hansen and Lunde, 2005).

An essential question about GARCH models is how to fit them to available data. Several approaches were proposed for this, but in practice, GARCH models are almost exclusively estimated by the *Quasi-Maximum Likelihood* (QML) method (see Section 3), which maximizes a (conditional) Gaussian log-likelihood function (Berkes et al., 2003). QML works well for a wide range of noise terms and has favorable theoretical properties, e.g., it is *strongly consistent* and, assuming its driving noise has finite 4th moment, *asymptotically normal* (see Theorem 1 in Section 4).

The QML method provides *point estimates*, i.e., single models which best fit the data. Nonetheless, many applications require *confidence regions*, as well, showing how reliable the estimates are. They are fundamental, e.g., for risk management and robust control. The standard way of building confidence sets around a point estimate is to use the level sets of its limiting distribution, which typically leads to *asymptotic* confidence ellipsoids (Ljung, 1999).

This however only provides *approximate* confidence regions for finite datasets. Furthermore, in several cases, the noises are *heavy-tailed* and fail to have 4th moments, which means the asymptotic normality of the Quasi-Maximum Likelihood Estimate (QMLE) is not guaranteed. In case of relatively heavy-tailed innovations, which are common in finance, directly estimating the asymptotic distribution of QMLE becomes very difficult (Hall and Yao, 2003).

Appearing in Proceedings of the 19^{th} International Conference on Artificial Intelligence and Statistics (AISTATS) 2016, Cadiz, Spain. JMLR: W&CP volume 51.

Because of these issues, some authors suggested moving away from the QMLE and using other methods (Chan et al., 2007), like the Hill estimator (Hill, 1975), or applying the ML theory with other specific (non-Gaussian) distributions. The main issues with such approaches are not only that specifying a distribution introduces the risk of misspecification (Spierdijk, 2014), but also that confidence regions are typically built around a selected point estimate, and as QMLE is the most widely applied method in practice, it would be important to build confidence regions for them.

Recently there has been an increase of interest in *bootstrap* (Efron and Tibshirani, 1993) approaches particularly since some of them can create confidence regions around the QMLE even if the innovations are heavy-tailed. One of the most popular bootstrap methods for GARCH processes is the *residual bootstrap* (Pascual et al., 2006; Shimizu, 2009) which is based on resampling (with replacement) the innovations from the *empirical distribution function* of the (standardized) QMLE residuals, simulating alternative trajectories based on which alternative QMLEs can be constructed. Then, typically *asymptotic* statistics are estimated, e.g., based on the sample of bootstrap QMLEs.

Nevertheless, standard bootstrap approaches are generally not consistent if the distribution of the asymptotic statistic is *non-Gaussian* and have inaccurate coverage probabilities if the true innovations are *skewed* (Hall and Yao, 2003).

Alternatives bootstrap variants were also proposed. For example, the *likelihood ratio* (LR) bootstrap (Luger, 2012a) for *stationary* GARCH processes is based on defining a p-value using conventional LR hypothesis tests combined with bootstrap. For a particular parameter it builds alternative bootstrap trajectories and computes bootstrap LRs. These are then compared with the LR of the original parameter. This approach can lead to *finite sample* guarantees, but only for completely *known* noise distributions, moreover, it is computationally demanding as it involves computing several bootsrap ML estimates for each parameter it tests.

Here, we propose a *finite sample* inference technique for GARCH models which can be seen as (i) an *exact* hypothesis test as well as (ii) a way to construct *distribution-free*, exact confidence regions around the QMLE without additional statistical assumptions about moments or stationarity. Its core is a permutation type test (Good, 2005) and it is called *ScoPe* as it applies randomly *permuted* residuals in the *score* function, i.e., the gradient of the log-likelihood (see Section 6). ScoPe was inspired by the recently developed *Sign-Perturbed Sums* (SPS) identification algorithm (Csáji et al., 2012, 2014, 2015), which can build exact, non-asymptotic, distribution-free confidence regions around the prediction error estimate of general linear dynamical systems; however, SPS cannot be applied for GARCH models.

We should note that an exact, distribution-free permutation test in the context of GARCH models was also proposed by Luger (2012b, 2014). However, that test permutes the GARCH process itself (not the residuals in the score function) and it is *only* applicable to test the hypothesis of *conditional homoskedasticity*. Particularly, it cannot be used to build confidence regions for the GARCH parameters.

2 GARCH MODELS

Formally, a GARCH(p, q) process, $\{X_t\}$, is defined by the following two equations (Francq and Zakoian, 2011)

$$X_t \triangleq \sigma_t \varepsilon_t, \tag{1a}$$

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

where $\{\varepsilon_t\}$ is a strong white noise, i.e., an i.i.d. sequence of real random variables with zero mean and unit variance; variable σ_t^2 is latent and defines the conditional variance of X_t, given its own past up to $t-1$; and $\omega^* > 0$ as well as $\alpha_i^*, \beta_j^* \geq 0$ are constants, where $1 \leq i \leq p$ and $1 \leq j \leq q$. Integers p and q are called the *orders* of the model. In case $q = 0$, we get back Engle's classical ARCH model.

It is known (Bollerslev, 1986) that there exists a wide-sense stationary solution to (1a)-(1b) if and only if

$$\sum_{i=1}^{p} \alpha_i^* + \sum_{j=1}^{q} \beta_j^* < 1. \tag{2}$$

If $\{X_t\}$ is a wide-sense stationary GARCH process, it is necessarily also strictly stationary. Moreover, it is a (potentially scaled) weak white noise, that is $\mathbb{E}[X_t] = 0$, $\mathbb{E}[X_t X_k] = 0$, and $\mathbb{E}[X_t^2] = \eta$, for all t and $k \neq t$, where $\mathbb{E}[\cdot]$ denotes expectation and η can be calculated by

$$\eta = \frac{\omega^*}{1 - \sum_{i=1}^{q} \alpha_i^* - \sum_{j=1}^{p} \beta_j^*}.$$

Conditions for the *unique* existence of a strictly stationary and causal solution to system (1a)-(1b) can be given in terms of the top Lyapunov exponent of its (Markovian) state space representation (Straumann, 2005).

3 QUASI-MAXIMUM LIKELIHOOD

While there are several approaches to estimate GARCH processes, such as prediction error methods or the Whittle estimator (Straumann, 2005), the most widely used estimators belong to the class of *Quasi-Maximum Likelihood* (QML) methods (Berkes et al., 2003). They are typically applied off-line, while there is also a recursive extension of the QML theory (Gerencsér and Orlovits, 2012). Now, we briefly recall the QML method for GARCH processes.

We do not make further assumptions on the distribution of the noise terms, $\{\varepsilon_t\}$, but accept a "working hypothesis"

that they are Gaussian. This however will not be needed to apply the method: as we will see, the QML method works well under very mild statistical assumptions.

More precisely, if we were to assume that $\{\varepsilon_t\}$ were Gaussian (hence standard normal, since we assumed that $\mathbb{E}[\varepsilon_t] = 0$ and $\mathbb{E}[\varepsilon_t^2] = 1$), then the conditional distribution of X_t/σ_t, given the σ-algebra generated by $\{\varepsilon_k\}_{k<t}$, would also be standard normal. The *quasi maximum likelihood estimate* (QMLE) is derived under this hypothesis.

Assuming known initial values $X_0(\theta), \ldots, X_{1-p}(\theta)$ and $\widehat{\sigma}_0(\theta), \ldots, \widehat{\sigma}_{1-q}(\theta)$, to be discussed below, the *conditional Gaussian quasi-likelihood* function is defined as

$$\mathcal{L}_n(\theta) = \mathcal{L}_n(\theta; x) \triangleq \prod_{t=1}^{n} \frac{1}{\sqrt{2\pi\widehat{\sigma}_t^2(\theta)}} \exp\left(-\frac{X_t^2}{2\widehat{\sigma}_t^2(\theta)}\right),$$

where $x = (X_1, \ldots, X_n)$ is the available sample and

$$\widehat{\sigma}_t^2(\theta) \triangleq \omega + \sum_{i=1}^{p} \alpha_i X_{t-i}^2 + \sum_{j=1}^{q} \beta_j \widehat{\sigma}_{t-j}^2(\theta), \qquad (3)$$

where $\theta \in \mathbb{R}^{p+q+1}$ is a generic vector encoding the parameters, $\theta \triangleq (\omega, \alpha_1, \ldots, \alpha_p, \beta_1, \ldots, \beta_q)$, while the "true" parameter vector is denoted by $\theta^* \triangleq (\omega^*, \alpha_1^*, \ldots, \alpha_p^*, \beta_1^*, \ldots, \beta_q^*)$.

We need initial conditions to calculate $\widehat{\sigma}_t^2(\theta)$ recursively. Standard choices include zero or $X_0^2(\theta) = \cdots = X_{1-p}^2(\theta) = \widehat{\sigma}_0^2(\theta) = \cdots = \widehat{\sigma}_{1-q}^2(\theta) = \omega$, or the unconditional variance w.r.t. θ (Francq and Zakoian, 2011)

$$X_0^2(\theta) = \cdots = X_{1-p}^2(\theta) = \widehat{\sigma}_0^2(\theta) = \cdots = \widehat{\sigma}_{1-q}^2(\theta) = \frac{\omega}{1 - \sum_{i=1}^{q} \alpha_i - \sum_{j=1}^{p} \beta_j}.$$

The QMLE is any measurable solution of the problem

$$\widehat{\theta}_n \triangleq \arg\max_{\theta \in \Theta} \mathcal{L}_n(\theta),$$

where Θ is the set of allowed parameters, for example $\theta \in \Theta$ if $\omega > 0$, $\alpha_i^*, \beta_j^* \geq 0$, for all i, j, and there exists a stationary solution to (1a)-(1b), i.e., property (2) holds.

Taking the natural logarithm, $\log(\cdot)$, of $\mathcal{L}_n(\theta)$ leads to the conditional quasi-log-likelihood function

$$\ell_n^*(\theta) = \ell_n^*(\theta; x) \triangleq \log \mathcal{L}_n(\theta) = \log \mathcal{L}_n(\theta; x),$$

which, under the standard normal hypothesis, simplifies to

$$\ell_n^*(\theta) = -\frac{1}{2} \sum_{t=1}^{n} \left[\log(2\pi) + \log \widehat{\sigma}_t^2(\theta) + \frac{X_t^2}{\widehat{\sigma}_t^2(\theta)} \right].$$

Since the optimal point is not affected by the constants, maximizing $L_n(\theta)$ is equivalent to minimizing $\ell_n(\theta)$,

$$\ell_n(\theta) \triangleq \frac{1}{n} \sum_{t=1}^{n} \left[\log \widehat{\sigma}_t^2(\theta) + \frac{X_t^2}{\widehat{\sigma}_t^2(\theta)} \right],$$

where the $1/n$ term is included for numerical stability. The minimization of $\ell_n(\theta)$ is typically done by an iterative numerical method, such as the Newton-Raphson algorithm.

4 ASYMPTOTICS OF QMLE

It is known that QMLE is strongly consistent and asymptotically normal. In order to make these claims more precise, let us introduce some assumptions (Straumann, 2005). When we talk about the marginal distribution of a stationary process $\{Y_t\}$, a generic element will be denoted by Y_0.

(Q1) *The noise is nondegenerate: the distribution of ε_0 is not concentrated in two points.*

(Q2) *The process is identifiable, that is $(\alpha_p^*, \beta_q^*) \neq (0, 0)$, $\omega > 0$, $\exists\, i : \alpha_i^* > 0$, and the polynomials $p(z) \triangleq \alpha_1^* z + \cdots + \alpha_p^* z^p$ and $q(z) \triangleq \beta_1^* z + \cdots + \beta_q^* z^q$ do not have any common zeros.*

(Q3) *The true parameter, θ^*, is in the interior of Θ.*

(Q4) $\exists\, \mu > 0$, such that $\mathbb{P}(|\varepsilon_0| \leq t) = o(t^\mu)$ as $t \downarrow 0$.

Then, using the four assumptions above, it can be proven that (Berkes et al., 2003; Straumann, 2005)

Theorem 1 *Let $\{X_t\}$ be a stationary GARCH(p, q) process with true parameter $\theta^* \in \Theta$. Then, assuming Q1 and Q2, QMLE is strongly consistent, that is*

$$\widehat{\theta}_n \xrightarrow{as} \theta^* \quad as \quad n \to \infty.$$

If additionally $\mathbb{E}[\varepsilon_0^4] < \infty$ and Q3, Q4 hold, the QMLE is also asymptotically normally distributed, i.e.

$$\sqrt{n}(\widehat{\theta}_n - \theta^*) \xrightarrow{d} \mathcal{N}(0, F_0^{-1} G_0 F_0^{-1}) \quad as \quad n \to \infty,$$

where F_0, G_0 are $(p \times q \times 1) \times (p \times q \times 1)$ matrices

$$F_0 = -\frac{2}{\mathbb{E}\,[\varepsilon_0^4 - 1]}\, G_0,$$

$$G_0 = \frac{\mathbb{E}\,[\varepsilon_0^4 - 1]}{4}\, \mathbb{E}\left[\frac{1}{\sigma_0^4} \nabla_\theta \widehat{\sigma}_0^2(\theta^*) \nabla_\theta \widehat{\sigma}_0^2(\theta^*)^{\mathrm{T}} \right],$$

where $\mathcal{N}(m, C)$ denotes the (multivariate) Gaussian distribution with mean vector m and covariance matrix C, while $\nabla_\theta f(\cdot)$ is the gradient vector of $f(\cdot)$ with respect to θ.

In many applications not just a point estimate, like QMLE, but a confidence region is also needed. The standard approach in practice is to use the (level sets of the) asymptotic distribution of the estimate to build a confidence region (Ljung, 1999; Söderström and Stoica, 1989).

To be more specific, assume we have an estimate Γ_n of the true covariance matrix $F_0^{-1} G_0 F_0^{-1}$, based on n data points. Then, an asymptotic confidence ellipsoid can be built by

$$\widetilde{\Theta}_n(s) \triangleq \left\{ \theta \in \mathbb{R}^d : (\theta - \widehat{\theta}_n)^{\mathrm{T}} \, \Gamma_n^{-1} \, (\theta - \widehat{\theta}_n) \, \leq \, \frac{s}{n} \right\}, \quad (4)$$

where $d \triangleq p+q+1$ and the probability that the true parameter is covered, i.e., $\theta^* \in \widetilde{\Theta}_n$, is *approximately* $F_{\chi^2(d)}(s)$, where $F_{\chi^2(d)}$ is the cumulative distribution function (CDF) of the standard χ^2 distribution with d degrees of freedom.

The main problems with such an approach are that (i) even if the covariance matrix of the asymptotic distribution was known exactly, the levels sets for a finite sample would still be only *approximately* correct, as they rely on a result which is only guaranteed in the limit. Moreover, (ii) the asymptotic normality of the estimation error requires finite 4th moment from the noise, which is often not the case in practice, e.g., for heavy-tailed distributions. Hence, such ellipsoids can only be used as *heuristics* in a finite sample setup, as their confidence levels are not guaranteed.

5 FINITE SAMPLE, DISTRIBUTION-FREE CONFIDENCE REGIONS

Now we turn our attention to finite sample, distribution-free results to overcome the issues mentioned above. In the next section, the ScoPe method is introduced to construct non-asymptotic, distribution-free confidence regions around the QML estimate for GARCH processes, which have *exact* confidence probabilities. The main motivation of the suggested approach comes from the *Sign-Perturbed Sums* (SPS) algorithm (Csáji et al., 2012, 2014, 2015). Though, SPS cannot be used for GARCH processes, for reasons discussed below, we briefly present it here to motivate our permutation-rank based method.

Let us consider the following scalar general linear dynamical system (Ljung, 1999; Box et al., 2008):

$$Y_t \, \triangleq \, G(z^{-1}; \theta^*) \, U_t + H(z^{-1}; \theta^*) \, N_t,$$

where t denotes (discrete) time, Y_t is output, U_t is an input, N_t is a noise, G, H are (causal) rational transfer functions, and z^{-1} is the backward shift operator. As previously, θ^* denotes the unknown true parameter of the system. The assumptions on the system are as follows (Csáji et al., 2012)

(S1) *The "true" system is in the model class which has polynomials with known orders.*

(S2) $H(\, z^{-1}; \theta \,)$ *has a stable inverse,* $G(\, 0 \,; \theta \,) \, = \, 0$ *and* $H(\, 0 \,; \theta \,) = 1$*, for* $\theta \in \Theta$*.*

(S3) *The noise* $\{N_t\}$ *is independent, and each* N_t *has a symmetric distribution about zero.*

(S4) *The system operates in open-loop, i.e., the inputs* $\{U_t\}$ *are independent of* $\{N_t\}$*.*

(S5) *The initialization of the system is known; for simplicity, we use* $Y_t = N_t = U_t = 0$*, for* $t \leq 0$*.*

Under these assumptions, the noise terms can be reconstructed, given a particular θ, by

$$\widehat{N}_t(\theta) \, \triangleq \, H^{-1}(z^{-1}; \theta)(Y_t - G(z^{-1}; \theta) \, U_t),$$

which are called *residuals* or *prediction errors*. It is important to note that $\widehat{N}_t(\theta^*) = N_t$, for all t.

The *prediction error estimate*, $\tilde{\theta}_n$, is defined as the minimizer of the squared prediction errors (Ljung, 1999),

$$\tilde{\theta}_n \, \triangleq \, \arg\min_{\theta \in \Theta} \sum_{t=1}^{n} \widehat{N}_t^2(\theta),$$

which can be found be solving the *normal equation*,

$$\sum_{t=1}^{n} \widehat{N}_t(\tilde{\theta}_n) \nabla_\theta \widehat{N}_t(\tilde{\theta}_n) \, = \, 0.$$

where Θ contains the allowed models, e.g., stable systems.

SPS builds its confidence region by perturbing the normal equation: given a θ, it builds $m-1$ alternative output trajectories using perturbed versions of the estimated residuals,

$$\bar{Y}_t(\theta, \alpha_i) \, \triangleq \, G(z^{-1}; \theta) \, U_t + H(z^{-1}; \theta) \, (\alpha_{i,t} \widehat{N}_t(\theta)),$$

where $\{\alpha_{i,t}\}$ are $(m-1) \times n$ i.i.d. random signs, that is random variables which take values ± 1 with probability $1/2$ each; and α_i denotes the vector $(\alpha_{i,1}, \ldots, \alpha_{i,n})$. Note that n is the sample size of the residuals we can reconstruct from $\{Y_t\}$, and m is a user-chosen design parameter.

Let us denote $\nabla_\theta \widehat{N}_t(\theta)$ by $\psi_t(\theta)$. Then, $\psi_t(\theta)$ can be treated as a linear filter on $\{Y_t\}$ and $\{U_t\}$,

$$\psi_t(\theta) \, = \, W_0(z^{-1}; \theta) \, Y_t + W_1(z^{-1}; \theta) \, U_t,$$

where W_0 and W_1 are vector-valued linear filters (Ljung, 1999). We produce perturbed versions of $\psi_t(\theta)$ by

$$\bar{\psi}_t(\theta, \alpha_i) \, \triangleq \, W_0(z^{-1}; \theta) \, \bar{Y}_t(\theta, \alpha_i) + W_1(z^{-1}; \theta) \, U_t,$$

where $i \in \{1, \ldots, m\}$, and define a reference function, S_0, and $m-1$ sign-perturbed functions, $\{S_i\}$, as

$$S_0(\theta) \, \triangleq \, \Psi_n^{-\frac{1}{2}}(\theta) \sum_{t=1}^{n} \psi_t(\theta) \widehat{N}_t(\theta),$$

$$S_i(\theta) \, \triangleq \, \bar{\Psi}_n^{-\frac{1}{2}}(\theta, \alpha_i) \sum_{t=1}^{n} \alpha_{i,t} \, \bar{\psi}_t(\theta, \alpha_i) \widehat{N}_t(\theta),$$

where Ψ_n and $\bar{\Psi}_n(\theta, \alpha_i)$ are covariance estimates, only used to shape the confidence region (Csáji et al., 2012).

Let use denote by $\mathcal{R}^0_m(\theta)$ the position of $\|S_0(\theta)\|^2$ in the ordering of variables $\{\|S_i(\theta)\|^2\}$, where ties are broken randomly. Therefore, $\mathcal{R}^0_m(\theta) = 1$ if $\|S_0(\theta)\|^2$ is the smallest in the ordering, $\mathcal{R}^0_m(\theta) = 2$ if it is the second smallest and so on. Then, the SPS confidence region is built by

$$\widetilde{\Theta}_n(m, r) \triangleq \left\{ \theta \in \Theta \,:\, \mathcal{R}^0_m(\theta) \leq m - r \right\},$$

where $m > r > 0$ are user-chosen integers. The SPS region has exact confidence probability (Csáji et al., 2012):

Theorem 2 *Under assumptions S1, S2, S3, S4, S5,*

$$\mathbb{P}\big(\theta^* \in \widetilde{\Theta}_n(m, r) \big) \,=\, 1 - \frac{r}{m}.$$

Since $\|S_0(\tilde{\theta}_n)\|^2 = 0$, i.e., the prediction error estimate satisfies the normal equation, it is always in the confidence region, assuming it is non-empty. In the special case of linear regression problems in which the regressors are independent of the noise, for example, generalized finite impulse response systems, it can be proved, as well, that the SPS confidence region is strongly consistent (Csáji et al., 2014) and it is also star convex with the least-squares estimate as a star center (Csáji et al., 2015). Nevertheless, the main strength of SPS lies in the fact that its confidence probability is *exact*, i.e., its confidence sets are non-conservative.

6 SCORE PERMUTATION

In this section, inspired by the core ideas underlying SPS, we present the *Score Permutation* (ScoPe) method, which can construct *exact*, *finite sample*, *distribution-free* confidence regions around the QMLE of GARCH processes.

Unfortunately, SPS cannot be applied to build such confidence regions, since, e.g., (i) SPS is defined for linear systems, while GARCH processes are nonlinear; (ii) it is built for a quadratic cost criterion, not for the QMLE; moreover, (iii) the GARCH residuals appear squared in the dynamics of the conditional variances, see (1b), thus, perturbing their signs does not produce alternative variance trajectories. Instead of sign-perturbations, ScoPe uses random permutations in the spirit of statistical permutation-rank tests (Good, 2005). A random permutation matrix based alternative to SPS was also analyzed by Kolumbán et al. (2015) for linear regression problems with deterministic regressors.

Recall that the QMLE satisfies the *likelihood equation*,

$$\nabla_\theta \, \ell_n(\widehat{\theta}_n) = 0,$$

and the gradient of the (conditional) log-likelihood function, the *score* function, can be written as

$$\nabla_\theta \, \ell_n(\theta) = \frac{1}{n} \sum_{t=1}^{n} \left(1 - \frac{X_t^2}{\widehat{\sigma}_t^2(\theta)} \right) \frac{1}{\widehat{\sigma}_t^2(\theta)} \nabla_\theta \, \widehat{\sigma}_t^2(\theta) =$$

$$\frac{1}{n} \sum_{t=1}^{n} \big(1 - \widehat{\varepsilon}_t^2(\theta) \big) \, \frac{1}{\widehat{\sigma}_t^2(\theta)} \nabla_\theta \, \widehat{\sigma}_t^2(\theta),$$

where $\widehat{\varepsilon}_t(\theta) \triangleq X_t / \widehat{\sigma}_t(\theta)$ is a reconstructed residual (innovation) for time t assuming a particular parameter θ, and $\widehat{\sigma}_t(\theta)$ is an estimate of σ_t, which can be calculated recursively using (the square root of) formula (3).

We can observe that $\widehat{\varepsilon}_t(\theta^*) = \varepsilon_t$, for all t, assuming the initial conditions are known, more precisely

(P1) *The "true" system is in the model class, i.e., upper bounds on the orders p, q are known and θ^* is in Θ.*

(P2) *The initial conditions $\widehat{\sigma}_0^2(\theta), \ldots, \widehat{\sigma}_{1-q}^2(\theta)$ of the conditional variances are known, i.e., $\widehat{\sigma}_t^2(\theta^*) = \sigma_t^2$.*

Initial conditions for $\{X_t^2\}$ are not needed, as system (1b) is autoregressive with finite order in the X_t^2 variables and $\{X_t\}$ is observed. Henceforth, for notational simplicity, we assume (w.l.o.g.) that $X_0, \ldots, X_{1-p}$ are available.

Note that P1 is a standard assumption and it is even needed to define the concept of confidence regions (namely, a subset of parameters which contains the "true" parameter with at least a given probability). Assumption P2 is also typical, especially for methods aiming at finite sample guarantees. It is mild, as it can be simply omitted if the system is ARCH (which was Engle's original model). Even if it is GARCH, the autocorrelation of conditional variances decays exponentially, therefore, it is expected that the effect of violating this assumption vanishes as we have more and more data. This is also supported by our experiments (Section 7).

Since $\{\varepsilon_t\}$ is i.i.d., its every permutation results in a sequence having the same distribution, namely

$$\{\varepsilon_t\} \stackrel{d}{=} \{\varepsilon_{\pi(t)}\}$$

where $\pi(\cdot)$ is an arbitrary permutation on the indices, i.e., a bijection of $\{1, \ldots, n\}$ onto itself.

Given a parameter θ, the main idea is to first "invert" the system to get the residuals $\{\widehat{\varepsilon}_t(\theta)\}$ and then generate alternative trajectories by applying random permutations on their indices. More precisely, we must generate $m - 1$ random permutations $\pi_1, \ldots, \pi_{m-1}$, where each permutation has the same probability $1/n!$ to be selected. Then, we can define alternative residuals for parameter θ by

$$\widehat{\varepsilon}_{\pi_i(1)}(\theta), \ldots, \widehat{\varepsilon}_{\pi_i(n)}(\theta),$$

for all $i \in \{1, \ldots, m-1\}$, where m is a user-chosen parameter as before. For simplicity, we denote the identity permutation by π_0, i.e., $\pi_0(t) = t$, for all t. In some cases it is useful to first standardize the residuals, by subtracting their sample mean and dividing with their standard deviation. Using this notation, the original and the perturbed

score function (the gradient of the log-likelihood) is

$$B(\theta,\pi_i) \triangleq \frac{1}{n}\sum_{t=1}^{n}\frac{\left(1-\widehat{\varepsilon}^2_{\pi_i(t)}(\theta)\right)}{\bar{\sigma}^2_t(\theta,\pi_i)}\nabla_\theta\,\bar{\sigma}^2_t(\theta,\pi_i),$$

where the perturbed variances $\bar{\sigma}^2_t(\theta,\pi_i)$ are defined as

$$\bar{\sigma}^2_t(\theta,\pi_i) \triangleq \omega+\sum_{k=1}^{p}\alpha_k\bar{X}^2_{t-k}(\theta,\pi_i)+\sum_{j=1}^{q}\beta_j\bar{\sigma}^2_{t-j}(\theta,\pi_i),$$

with the same initial values for all generated permutations,

$$\bar{\sigma}^2_0(\theta,\pi_i) \triangleq \widehat{\sigma}_0(\theta),\ldots,\bar{\sigma}^2_{1-q}(\theta,\pi_i) \triangleq \widehat{\sigma}_{1-q}(\theta).$$

This gives rise to an alternative output trajectory with

$$\bar{X}_t(\theta,\pi_i) \triangleq \bar{\sigma}_t(\theta,\pi_i)\,\widehat{\varepsilon}_{\pi_i(t)}(\theta). \tag{5}$$

Observe that $B(\theta,\pi_0)=\nabla_\theta\,\ell_n(\theta)$ as π_0 is the identity permutation. We use $\|B(\theta,\pi_0)\|^2$ as a reference and compute its rank in the ordering of $\{\|B(\theta,\pi_i)\|^2\}$ variables. There is a chance two such functions take on the same value, e.g., if the noise is discrete. In order to handle this, we use a tie-breaking, namely, with the help of another random permutation ν. This one is on $\{0,\ldots m-1\}$. Given m real numbers $Z_0,\ldots,Z_{m-1}$ we define a strict total order $\succ_\nu$ as

$$Z_k \succ_\nu Z_j \qquad \text{if and only if}$$

$$(\,Z_k > Z_j\,) \quad\text{or}\quad (\,Z_k = Z_j \ \text{ and } \ \nu(k)>\nu(j)\,).$$

The *rank* of $\|B(\theta,\pi_0)\|^2$ w.r.t. the ordering $\succ_\nu$ is then

$$\mathcal{R}_m(\theta) \triangleq 1+\sum_{i=1}^{m-1}\mathbb{I}\big(\|B(\theta,\pi_0)\|^2 \succ_\nu \|B(\theta,\pi_i)\|^2\big),$$

where $\mathbb{I}(\cdot)$ is an indicator function: it is 1 if its argument is true and 0 otherwise. The ScoPe confidence set is

$$\widehat{\Theta}_n(m,r) \triangleq \{\,\theta\in\Theta\,:\,\mathcal{R}_m(\theta)\le m-r\,\},$$

where $m>r>0$ are user-chosen integers affecting the coverage probability of the confidence region. The main theoretical claim of the paper is the following theorem.

Theorem 3 *Assuming P1 and P2, we have*

$$\mathbb{P}\big(\,\theta^*\in\widehat{\Theta}_n(m,r)\,\big) = 1-\frac{r}{m}.$$

Proof of Theorem 3

The main idea is to show that $\{\|B(\theta^*,\pi_i)\|^2\}$ are *conditionally* i.i.d. (for the case of the true parameter, θ^*), therefore exchangeable, which leads to the fact that each ordering of them has the same probability, namely, $1/m!$.

By definition, $\theta^*\in\widehat{\Theta}_n(m,r)$ if and only if $\mathcal{R}_m(\theta^*)\le m-r$, i.e., if $\|B(\theta^*,\pi_0)\|^2$ takes one of the positions $1,\ldots,m-r$ in the ordering of $\{\|B(\theta^*,\pi_i)\|^2\}$ variables, with respect to $\succ_\nu$. Our main aim will be to prove that $\{\|B(\theta^*,\pi_i)\|^2\}$ are *uniformly ordered*, that is to show that

$$\mathbb{P}\left(\|B(\theta^*,\pi_{\gamma(0)})\|^2\succ_\nu\ldots\succ_\nu\|B(\theta^*,\pi_{\gamma(m-1)})\|^2\right)=\frac{1}{m!},$$

for all possible permutation γ on $\{0,\ldots,m-1\}$. From this, the theorem follows immediately, since then $\|B(\theta^*,\pi_0)\|^2$ takes each position in the ordering with probability exactly $1/m$, thus $\mathbb{P}(\mathcal{R}_m(\theta^*)=i)=1/m$ for $i\in\{0,\ldots,m-1\}$, from which $\mathbb{P}\big(\theta^*\in\widehat{\Theta}_n(m,r)\big)=1-r/m$.

Before we show the uniform ordering of $\{\|B(\theta^*,\pi_i)\|^2\}$, we introduce some notations and state some useful facts about random permutations.

If γ is a permutation on $\{1,\ldots,n\}$ and $Z=(Z_1,\ldots,Z_n)$ is a vector of dimension n, then let

$$\gamma(Z) \triangleq (Z_{\gamma(1)},\ldots,Z_{\gamma(n)}).$$

The inverse of a permutation γ is denoted by γ^{-1}, that is $\gamma^{-1}(\gamma(Z))=Z$. If γ and π are permutations, their composition is denoted by $\gamma\circ\pi$, that is $(\gamma\circ\pi)(Z)=\gamma(\pi(Z))$.

It can be proven that if $Z=(Z_1,\ldots,Z_n)$ is an i.i.d. random vector and γ is a random permutation which is uniformly chosen from all possible permutations of $\{1,\ldots,n\}$, with γ being independent of Z, then we can conclude that γ and $\gamma^{-1}(Z)$ are independent, as well.

Also if $\gamma,\pi_1,\ldots,\pi_k$ are $k+1$ i.i.d. uniformly chosen random permutations, then $\gamma,\pi_1\circ\gamma^{-1},\ldots,\pi_k\circ\gamma^{-1}$ are also $k+1$ i.i.d. random permutations (also uniform).

Finally, if $Z_0,\ldots,Z_{m-1}$ are i.i.d. random variables, then they are uniformly ordered w.r.t. $\succ_\nu$ (Csáji et al., 2015, Lemma 3). Note that this is even the case for discrete random variables, since $\succ_\nu$ takes care of the tie-breaking; recall that ν is a random permutation on $\{0,\ldots,m-1\}$.

Now, we proceed with the proof by showing the uniform ordering property of $\{\|B(\theta^*,\pi_i)\|^2\}$ variables.

In order to simplify the notations, let us introduce

$$f(\varepsilon,\pi_i) \triangleq \|B(\theta^*,\pi_i)\|^2,$$

for indices $i\in\{0,\ldots,m-1\}$, where $\varepsilon=(\varepsilon_1,\ldots,\varepsilon_n)=(\widehat{\varepsilon}_1(\theta^*),\ldots,\widehat{\varepsilon}_n(\theta^*))$. Note that here we used our assumptions P1 and P2, i.e., that we could reconstruct the "true" noise sequence ε, in case we knew the true parameter θ^*.

Let us introduce a new (uniform) random permutation μ on $\{1,\ldots,n\}$, generated independently of $\pi_1,\ldots,\pi_{m-1}$. We can "inject" the new permutation μ into our system by

$$f(\mu(\varepsilon),\pi_i\circ\mu^{-1}) = f(\varepsilon,\pi_i),$$

for all $i\in\{0,\ldots,m-1\}$, since we simply undo the effect of permutation μ by composing π_i with its inverse.

Now, let us fix a realization of $\mu(\varepsilon)$, denoted by r, i.e., from now on we condition on this realization. Then, the only random element in the variables $W_i \triangleq f(r, \pi_i \circ \mu^{-1})$ are $\pi_i \circ \mu^{-1}$. Now, we know that π_0 is the identity permutation, but with injecting μ we have managed to "re-randomize" it without chaining the method. By using the previously mentioned fact, we know that $\mu^{-1}, \pi_1 \circ \mu^{-1}, \ldots, \pi_{m-1} \circ \mu^{-1}$ are i.i.d. random elements. Since applying the same function to elements of an i.i.d. collection results in an i.i.d. collection, it follows that $W_0, \ldots, W_{m-1}$ are i.i.d. Hence, they are uniformly ordered w.r.t. $\succ_\nu$, *conditionally on* r.

Until now, we have showed that $\{\|B(\theta^*, \pi_i)\|^2\}$ are uniformly ordered given a realization of $\mu(\varepsilon)$. To get rid of this conditioning, we can observe that (i) the ordering distribution we got is independent of the actual realization of $\mu(\varepsilon)$, and (ii) vector $\mu(\varepsilon)$ is independent of $\mu^{-1}, \pi_1 \circ \mu^{-1}, \ldots, \pi_{m-1} \circ \mu^{-1}$. In this case, we know (Csáji et al., 2015, Lemma 2) that the uniform ordering also holds without conditioning on a realization. □

Now, let us make some remarks on ScoPe:

(i) The confidence probability is *exact* for any *finite sample*, thus no conservativism is introduced.

(ii) Parameters m and r are user-chosen, hence, the confidence probability is *under our control*.

(iii) The applied statistical assumptions are very mild, e.g., we do not assume knowing the particular distribution of the noise, i.e., it is a *distribution-free* method.

(iv) Unlike the standard asymptotic ellipsoids or bootstrap approaches, the confidence probability of ScoPe is exact even for *heavy-tailed* and *skewed* distributions.

(v) Scope neither needs assumptions on stationary, therefore, it can work for *nonstationary* processes.

(vi) Since the QMLE satisfies the likelihood equation, i.e., $\nabla_\theta \ell_n(\widehat{\theta}_n) = 0$, we have $\|B(\widehat{\theta}_n, \pi_0)\|^2 = 0$. Hence, the *QMLE is always included* in the confidence set, assuming it is non-empty. In other words, ScoPe builds its confidence regions around the QML estimate.

(vii) Finally, if we evaluate $\mathcal{R}_m(\theta)$ on a set of θ values, their ranks indicate which confidence levels can be associated to those parameters, therefore, these "rank fields" contain information about the distribution of the estimation error. On the other hand, since the confidence region also contains potential other roots of the score, it is not a direct estimate of this distribution.

The main idea of the proof is that $\{\|B(\theta, \pi_i)\|^2\}$ are uniformly ordered in case they are evaluated at the true parameter θ^*. The other intuition behind this construction is, similarly to SPS, that as we get farther away from the "true" parameter, θ^*, our reference element $\|B(\theta, \pi_0)\|^2$ should dominate the ordering of $\{\|B(\theta, \pi_i)\|^2\}$: for $i \neq 0$, $\|B(\theta, \pi_0)\|^2 \succ_\nu \|B(\theta, \pi_i)\|^2$, with "high probability" as θ gets "sufficiently far away" from θ^*, for example, $\|\theta - \theta^*\|$ is "large enough". However, this property is hard to formulate and prove rigorously, therefore, in Section 7 we continue with investigating ScoPe experimentally.

7 EXPERIMENTAL RESULTS

Now, we evaluate ScoPe through numerical experiments on simulated data as well as on major stock market indices. ScoPe is compared with standard asymptotic ellipsoids, residual- and likelihood ratio bootstrap constructions.

Here we focus on GARCH(1, 1) processes that constitute a very important special case, as they are by far the most widely used GARCH models in industry (Ruppert, 2011) as well as typical reference models in empirical comparisons (Francq and Zakoian, 2011). An explanation of this was given by Hansen and Lunde (2005), who compared the forecasting potential of 330 volatility models on historical exchange rates and found no statistical evidence that more sophisticated models could outperform GARCH(1, 1).

In our first simulated experiment the driving noise $\{\varepsilon_t\}$ of the GARCH(1, 1) model had logistic distribution with zero mean and scale $\sqrt{3}/\pi$, to ensure unit variance. Thus,

$$\begin{aligned} X_t &\triangleq \sigma_t \varepsilon_t, \\ \sigma_t^2 &\triangleq \omega^* + \alpha^* X_{t-1}^2 + \beta^* \sigma_{t-1}^2, \end{aligned}$$

where the true parameter vector was $\theta^* = [\alpha^*, \beta^*, \omega^*]$, where $\alpha^* = 0.44$, $\beta^* = 0.33$ and, since we assumed a system with unit variance, i.e., weak white noise, $\omega^* = 1 - \alpha^* - \beta^* = 0.23$. Because of this, it is enough to build a confidence region for (α^*, β^*), as they determine ω^*.

In order to test the method, 100 observations were generated. The rank of $\|B_0(\theta)\|^2$, $\mathcal{R}_m(\theta)$, was then calculated for parameters in $[0, 1] \times [0, 1]$. The resulting "rank field" is shown in Figure 1 with its 90 % confidence region, in which parameters leading to only non-stationary processes were also eliminated, namely, the ones with $\alpha + \beta \geq 1$. Note that this region has two connected components.

In our next simulated experiment, illustrated in Figure 2, the process was generated using Laplacian innovations. Now, the ω parameter was also estimated (90 % confidence was targeted), and the true parameter was $\theta^* = [0.44, 0.33, 0.22]$. As we can observe from the image, the QMLE and the true parameter are situated in different rank-valleys, which explains disconnected confidence regions.

Table 1 compares 90 % confidence sets of the asymptotic approach (4), the residual (Pascual et al., 2006) and (Gaussian) likelihood ratio (Luger, 2012a) bootstraps and ScoPe. In this experiment Gaussian and Logistic driving noises

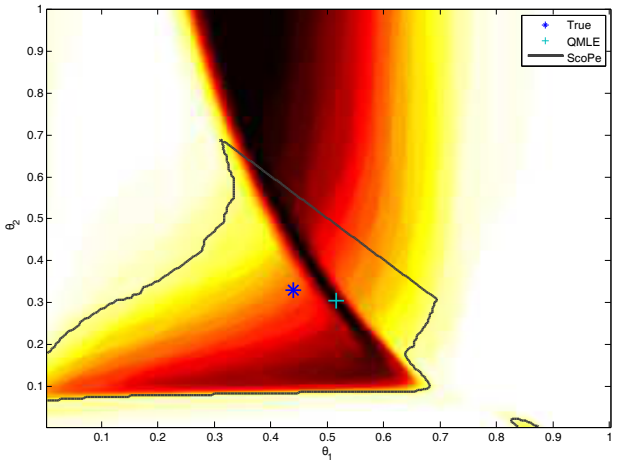

Figure 1: Logistic noise, $n = 100$, $m = 100$, $r = 10$; Exact 90% ScoPe confidence set for a GARCH$(1, 1)$ process and its rank field. Darker color indicates smaller rank.

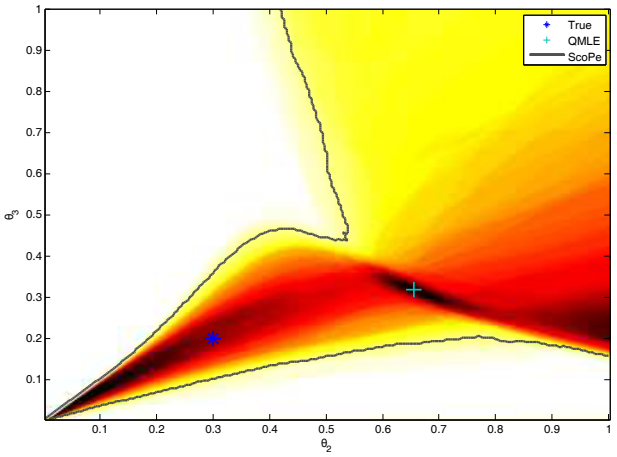

Figure 2: Laplacian noise, $n = 1000$, $m = 100$, $r = 10$; The rank of $\|B((\theta), \pi_0)\|^2$ is shown as a function of (β, ω); α is fixed at its QMLE. Darker color indicates smaller rank.

were applied. Both the empirical coverage of the true parameter and the relative size of the confidence sets compared to the whole class of allowed models (weak white noise assumption) were evaluated by 1000 Monte Carlo trials. Random noises were generated and it was tested whether θ^* is in the confidence set (empirical coverage). Next several random parameter values were selected and the relative size of the confidence regions were estimated as the ratio of those parameters which were fallen into the set (relative area). A 1000 long "burn in" simulation was used for initialization, thus assumption P2 was violated. Nevertheless, ScoPe provided close to exact coverage probabilities combined with relatively small confidence regions.

In our last experiment we evaluated the methods on major stock market indices. More precisely, the daily closing prices of Nasdaq 100, S&P 500 and FTSE 100 were used from the entire period of 2014 (which means 252 observations for each dataset). First, the *compound returns* (Francq and Zakoian, 2011) were calculated from the data, i.e., for each sequence $\{P_t\}$, the data were transformed by $R_t = \log(P_t/P_{t-1})$. They were then standardized, after which GARCH(1, 1) models were estimated using QML.

Table 1: Empirical Coverages and Areas on Simulated Data

	Gaussian Noise		Logistic Noise	
Method	**Emp. Cov.**	**Rel. Area**	**Emp. Cov.**	**Rel. Area**
Asym.Ell.	0.8656	0.4715	0.8264	0.5446
Res.Boots.	0.8567	0.7051	0.8152	0.6139
LR.Boots.	0.9655	0.6623	0.9762	0.7681
ScoPe	0.8961	0.5324	0.9147	0.6727

Table 2: Relative Areas on Stock Market Indices (2014)

Method	**Nasdaq 100**	**S&P 500**	**FTSE 100**
Asym.Ell.	0.3426	0.1679	0.1535
Res.Boots.	0.3791	0.2549	0.2850
LR.Boots.	0.8150	0.7919	0.8326
ScoPe	0.3801	0.2862	0.2412

Only the relative areas of 90 % confidence regions, approximated using 1000 Monte Carlo trials, are shown in Table 2 as "true" parameters were not available. The size (but not the shape) of ScoPe confidence sets were about the same as the ones obtained by residual bootstrap, indicating the promising practical applicability of ScoPe, especially, since it has stronger theoretical guarantees than bootstrap.

8 CONCLUSIONS

GARCH processes are widespread models of (conditional) heteroscedasticity, and are archetypically estimated by the QML mehtod, which provides point estimates. Often confidence regions are also needed, but unfortunately the standard approach (based on limiting distributions) fails in case the driving noise is heavy-tailed. Alternative approaches, such as bootstrap methods, may also fail for skewed distributions or require knowledge about the noise terms.

In this paper the ScoPe method was proposed which is based on permuting the score function. At the best of our knowledge, it is the first approach that can construct (i) exact, (ii) non-asymptotic, (iii) distribution-free confidence regions (iv) around the QMLE, (v) without additional assumptions about moments or stationarity. Its exact coverage probability was proved and numerical experiments on simulated as well as stock market data were also presented.

Acknowledgments

The work of B. Cs. Csáji was supported by the Hung. Sci. Res. Fund (OTKA), projects 113038 and 111797, and by the János Bolyai Research Fellowship, BO / 00683 / 12 / 6.

References

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

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

G. E. P. Box, G. M. Jenkins, and G. C. Reinsel. *Time Series Analysis: Forecasting and Control*. Prentice-Hall, 4 edition, 2008.

N. H. Chan, S.-J. Deng, L. Peng, and Z. Xia. Interval estimation of value-at-risk based on GARCH models with heavy-tailed innovations. *Journal of Econometrics*, 137 (2):556–576, 2007.

B. Cs. Csáji, M.C. Campi, and E. Weyer. Sign-Perturbed Sums (SPS): A method for constructing exact finite-sample confidence regions for general linear systems. In *Proceedings of the 51st IEEE Conference on Decision and Control, Maui, Hawaii*, pages 7321–7326, 2012.

B. Cs. Csáji, M. C. Campi, and E. Weyer. Strong consistency of the Sign-Perturbed Sums method. In *Proceedings of the 53rd IEEE Conference on Decision and Control, Los Angeles, California*, pages 3352–3357, 2014.

B. Cs. Csáji, M. C. Campi, and E. Weyer. Sign-Perturbed Sums: A new system identification approach for constructing exact non-asymptotic confidence regions in linear regression models. *IEEE Transactions on Signal Processing*, 63(1):169–181, 2015.

B. Efron and R. J. Tibshirani. *An Introduction to the Bootstrap*. Chapman & Hall, New York, 1993.

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

C. Francq and J. M. Zakoian. *GARCH Models: Structure, Statistical Inference and Financial Applications*. John Wiley & Sons, 2011.

L. Gerencsér and Zs. Orlovits. Real time estimation of stochastic volatility processes. *Annals of Operations Research*, 200(1):223–246, 2012.

P. Good. *Permutation, Parametric, and Bootstrap Tests of Hypotheses*. Springer, 3 edition, 2005.

P. Hall and Q. Yao. Inference in ARCH and GARCH models with heavy–tailed errors. *Econometrica*, 71(1):285–317, 2003.

P. R. Hansen and A. Lunde. A forecast comparison of volatility models: does anything beat a GARCH(1,1)? *Journal of Applied Econometrics*, 20(7):873–889, 2005.

T. Hastie, R. Tibshirani, and J. Friedman. *The elements of statistical learning*. Springer, New York, 2 edition, 2009.

B. M. Hill. A simple general approach to inference about the tail of a distribution. *The Annals of Statistics*, 3(5): 1163–1174, 1975.

S. Kolumbán, I. Vajk, and J. Schoukens. Perturbed datasets methods for hypothesis testing and structure of corresponding confidence sets. *Automatica*, 51:326–331, 2015.

L. Ljung. *System Identification: Theory for the User*. Prentice-Hall, 2nd edition, 1999.

R. Luger. Finite-sample bootstrap inference in GARCH models with heavy-tailed innovations. *Computational Statistics & Data Analysis*, 56(11):3198–3211, 2012a.

R. Luger. Testing for GARCH effects: an exact procedure based on quasi-likelihood ratios. Technical report, Georgia State University, 2012b.

R. Luger. Testing for GARCH effects with quasilikelihood ratios. *The Journal of Risk*, 16(4):23, 2014.

L. Pascual, J. Romo, and E. Ruiz. Bootstrap prediction for returns and volatilities in GARCH models. *Computational Statistics & Data Analysis*, 50(9):2293–2312, 2006.

D. Ruppert. *Statistics and Data analysis for Financial Engineering*. Springer, 2011.

K. Shimizu. *Bootstrapping Stationary ARMA-GARCH Models*. Springer, 2009.

T. Söderström and P. Stoica. *System Identification*. Prentice Hall International, Hertfordshire, UK, 1989.

L. Spierdijk. Confidence intervals for ARMA–GARCH value-at-risk: The case of heavy tails and skewness. *Computational Statistics & Data Analysis*, 2014.

D. Straumann. *Estimation in Conditionally Heteroscedastic Time Series Models*. Springer, 2005.

V. N. Vapnik. *Statistical Learning Theory*. Wiley-Interscience, New York, 1 edition, 1998.