Chapter 6
Statistical Inference for Nonlinear Processes

In this section, we consider nonlinear processes with long memory. We will mainly focus on volatility models: stochastic volatility (see Definitions 2.3–2.4 and Sect. 4.2.6 for limit theorems), ARCH(∞) processes (see Definition 2.1 and Sect. 4.2.7) and LARCH(∞) models (see (2.47) and (2.48), and Sect. 4.2.8). Statistical inference for traffic models is not well developed yet (see Faÿ et al. 2006, 2007; Hsieh et al. 2007 for some results in this direction).

Volatility models considered in this book have the general form $X_t = \xi_t \sigma_t$, where ξ_t ($t \in \mathbb{Z}$) is an i.i.d. sequence and σ_t depends on the past $(\xi_{t-1}, \xi_{t-2}, \ldots)$ and/or a latent process ζ_t. In particular, in the stochastic volatility model (SV),

$$\sigma_t = \sigma(\zeta_t), \qquad \zeta_t = \sum_{j=1}^{\infty} a_j \varepsilon_{t-j},$$

where (ξ_t, ε_t) ($t \in \mathbb{Z}$) is a sequence of i.i.d. random vectors. If furthermore $\sigma(x) = \exp(x)$ and ζ_t is a long-memory Gaussian sequence independent of the i.i.d. centred sequence ξ_t, then the model is called LMSV.

If

$$\sigma_t = b_0 + \sum_{k=1}^{\infty} b_k X_{t-k}$$

and b_j decay slowly like a constant times j^{d-1} ($d \in (0, 1/2)$), then we obtain a LARCH(∞) model with long memory (recall that σ_t can be expressed explicitly in terms of $\xi_{t-1}, \xi_{t-2}, \ldots$). Finally, if

$$\sigma_t^2 = b_0 + \sum_{k=1}^{\infty} b_k X_{t-k}^2,$$

$\sum_{k=1}^{\infty} |b_k| < \infty$, we obtain a second-order stationary ARCH(∞) sequence. Other models, e.g. FIGARCH, are not discussed in this chapter.

J. Beran et al., *Long-Memory Processes*,
DOI 10.1007/978-3-642-35512-7_6, © Springer-Verlag Berlin Heidelberg 2013

As in Chap. 5, we start our discussion with location estimation. In this case, the stochastic volatility (like LMSV) and LARCH(∞) models follow a similar pattern. The asymptotic distribution of the sample mean is not affected by long memory. The same applies to M-estimators, as long as the function ψ that defines the M-estimator is antisymmetric and the distribution of the noise variables ξ_t is symmetric. Otherwise, asymptotic properties of M-estimators are influenced by long memory. Such results were obtained in Beran (2006) and Beran and Schützner (2008), and are presented in Sects. 6.1.1 and 6.2.1, respectively, for SV and LARCH models. Finally, in Sect. 6.3.1, we discuss location estimation for ARCH(∞) processes. At the moment, a theory for M-estimators is not available.

As for estimation of memory parameters, one may note that long memory appears (if at all) in the squares. It is therefore tempting to apply methods described in Chap. 5 to the squared sequence X_t^2. However, it may be more natural to divide volatility processes into two groups: stochastic volatility-type models (with a possible leverage) and LARCH(∞)-type models.

In the first case, direct maximum likelihood estimation is not always feasible because of the presence of an unobserved latent process. Note, however, that, for instance, for a stochastic volatility model with an exponential volatility function $\sigma(x) = e^x$, one may consider a log-transformation. This approach is taken, among others, in Zaffaroni (2009) using parametric Whittle estimation and in Deo and Hurvich (2001), Hurvich and Soulier (2002), Hurvich et al. (2005b) or Dalla et al. (2006) who consider semiparametric estimation.

For the LARCH models, a maximum likelihood approach is feasible in principle because σ_t is an explicit function of past observations (see Beran and Schützner 2009). Up to date there are no theoretical results on semiparametric estimators in the Fourier or wavelet domain. Teyssière and Abry (2006) as well as Jach and Kokoszka (2008) study the numerical performance of wavelet estimators, in particular for LARCH models. For ARCH(∞) processes, σ_t^2 is again a direct function of past observations and MLE-type estimators are not difficult to calculate. In particular, one can show that the MLE is more efficient than Whittle estimation based on the squared observations (which is not really an approximate MLE), see Giraitis and Robinson (2001), Straumann (2004), Berkes and Horváth (2003).

Finally, we consider tail index estimation for heavy-tailed stochastic volatility models. Recall that for linear processes we considered in Sect. 5.15 the tail index M-estimation based on the assumption of stable innovations. Here we consider instead the Hill estimator which is consistent without specifying a particular model. Asymptotic normality of the Hill estimator for SV models was established in Kulik and Soulier (2011) and is presented in Sect. 6.1.3. For LARCH processes, a numerical, although wavelet-based, tail index estimation can be found in Jach and Kokoszka (2008).

6.1 Statistical Inference for Stochastic Volatility Models

In this section, we consider statistical inference for stochastic volatility models of the form

$$X_t = \sigma_t \xi_t \quad (t \in \mathbb{N}), \tag{6.1}$$

where $\sigma_t = \sigma(\zeta_t)$, $\zeta_t = \sum_{j=1}^{\infty} a_j \varepsilon_{t-j}$ and (ξ_t, ε_t) $(t \in \mathbb{Z})$ is a sequence of i.i.d. random vectors. It is assumed that $E(\varepsilon_1) = 0$, however, there is no a priori assumption that the random variables ξ_t are centred.

In Sect. 6.1.1, we consider location estimation in a model $Y_t = \mu + X_t$, where X_t is an SV process. As mentioned in the introduction, the asymptotic distribution of the sample mean is not affected by long memory. The same applies to M-estimators, as long as the function ψ that defines the M-estimator is antisymmetric and the distribution of the noise variables ξ_t is symmetric (Beran and Schützner 2008).

We proceed with estimation of the memory parameter. Consider the volatility model (6.1). We recall that the memory parameter d appears in the asymptotics for the covariance function of the squares (see (2.61)). The graphical methods considered in Sect. 5.4 can be also applied in this case, by replacing X_t there by $Y_t = X_t^2$ here. For example, the R/S statistic can be defined as R_n/S_n, where

$$R_n = \max_{1 \le k \le n} \sum_{t=1}^{k} (Y_t - \bar{y}_n) - \min_{1 \le k \le n} \sum_{t=1}^{k} (Y_t - \bar{y}_n)$$

and $S_n^2 = (n-1)^{-1} \sum_{t=1}^{n} (Y_t - \bar{y}_n)^2$ is the sample variance of $Y_t = X_t^2$. The sample variance S_n^2 converges in probability to $\mathrm{var}(X_1^2)$ (provided it is finite). The same approach can be applied to all other methods considered in Sect. 5.4 (see, e.g. Giraitis et al. 2000b).

However, using the squares may not be appropriate for heavy-tailed data. For instance, the data may have a finite variance, but infinite fourth moments. Then the graphical methods can be quite misleading (see, e.g. Wright 2002).

In general, maximum likelihood estimation is not suitable for SV models because the likelihood function cannot be written in an explicit form (see, e.g. Robinson and Zaffaroni 1997, 1998). Asymptotic normality of the Whittle estimator applied to transformed data was considered explicitly in Breidt et al. (1998) and in case of leverage in Zaffaroni (2009). Some results can be deduced from earlier theory for models with signal and additive noise (Hosoya 1974; Hosoya and Taniguchi 1982). Note, however, that the Whittle approach does not have much to do with maximum likelihood estimation here because the (transformed) data the method is applied to are by definition far from normal.

As for semiparametric methods, asymptotic results in the SV case are a relatively simple generalization of the theory for linear processes considered in Chap. 5. Specifically, if $X_t = \xi_t \exp(\sum_{j=1}^{\infty} a_j \varepsilon_{t-j})$, then one can apply the log-transformation to X_t^2 and the resulting model has the form of a linear long-memory

process corrupted by i.i.d. noise. Asymptotic properties of semiparametric estimators in SV models were considered in Deo and Hurvich (2001), Hurvich and Soulier (2002), Hurvich et al. (2005b).

Finally, we discuss tail index estimation. In Sect. 5.2.3, we considered M-estimation for heavy-tailed long-memory processes. Such an approach requires strong assumptions on an innovation sequence of the linear process. Rates of convergence and the asymptotic distribution is affected by long memory and tail behaviour. In the present context, based on results on M-estimators in Sect. 6.1.1 below, one may be expected that the asymptotic behaviour of an M-estimator of the tail index is not affected by long memory. However, such results are not known at present. Instead, we consider the so-called Hill estimator (see, e.g. Embrechts et al. 1997). Its asymptotic properties are built upon results for the tail empirical process considered in Sect. 4.8.5. It is proven (see Kulik and Soulier 2011) that long memory does not affect the rate of convergence. This is confirmed in Jach et al. (2012) and Luo (2011), both in theory and numerical studies.

6.1.1 Location Estimation

Consider a time series $Y_t = \mu + X_t$ ($t \in \mathbb{N}$) such that the residuals X_t are generated by a stochastic volatility model (6.1). Furthermore, assume that the random variables ξ_t that appear in the model definition (6.1) are centred. Hence $E(X_t) = 0$. In Sect. 4.2.6, we found out that under appropriate moment assumptions,

$$n^{-1/2} \sum_{t=1}^{[nu]} X_t \Rightarrow vB(u),$$

where $v^2 = \mathrm{var}(X_1)$ and $B(u)$ ($u \in [0,1]$) is a standard Brownian motion. In other words, long memory in volatility does not affect rates of convergence for the sample mean.

More generally, if ψ is a deterministic function such that $E[\psi(X_1)|\mathscr{G}_0] = 0$, where $\mathscr{G}_t$ is the sigma field generated by $(\xi_t, \varepsilon_t, \xi_{t-1}, \varepsilon_{t-1}, \ldots)$, then the central limit theorem above still holds with $v^2 = \mathrm{var}(\psi(X_1))$.

The condition $E[\psi(X_1)|\mathscr{G}_0] = 0$ is equivalent to

$$\int \psi\big(s\sigma(\zeta_1)\big)\, dF_\xi(s) = 0,$$

where F_ξ is the distribution function of ξ_1. If, for example, $\psi(x) = \mathrm{sign}(x)$, bearing in mind that $\sigma(\cdot) > 0$, this integral has the form

$$-\int_{-\infty}^{0} dF_\xi(s) + \int_0^{\infty} dF_\xi(s).$$

Thus, if the random variable ξ_1 is symmetric, then this expression vanishes. Recalling from Sect. 5.2.3 that the sign function yields the sample median (written down as an M-estimator), we can expect that in the particular case of symmetric random variables ξ_t and antisymmetric functions ψ, the asymptotic theory for M-estimators is the same as for i.i.d. data. To be more specific, if $\hat{\mu}$ is a solution of $\sum_{t=1}^{n} \psi(Y_t - \mu) = 0$, then

$$\sqrt{n}(\hat{\mu} - \mu) \to_d N\big(0, \sigma_\psi^2\big), \tag{6.2}$$

where $\sigma_\psi^2 = E[\psi^2(X_1)]/E^2[\psi'(X_1)]$. A general result was obtained in Beran and Schützner (2008) (cf. also Theorem 6.2 in Sect. 4.2.6). In particular, if

(A1) The random variables ξ_t are symmetric,
(A2) σ_t is a second-order stationary process with a finite variance such that ξ_t is independent of σ_s, $s \le t$ (but the sequences ξ_t and σ_t are not necessary independent),
(A3) The function $\psi(\cdot)$ is measurable and antisymmetric, that is, $\psi(x) = -\psi(-x)$, and $E[\psi^2(X_1)] < \infty$,

then (6.2) holds.

Theorem 6.1 *Consider the stochastic volatility model defined in* (6.1). *Assume that* (A1)–(A3) *above hold. Under additional regularity conditions*, (6.2) *holds*.

We note that "additional regularity conditions" refer to assumptions (A4)–(A8) in Beran and Schützner (2008).

Proof The proof differs from the proof of the central limit theorem for M-estimators based on linear processes with long-range dependence; see the proof of Theorem 5.1. The reason is that in the proof of that theorem we were looking for the asymptotic equivalence between an M-estimator and the sample mean.

To proceed, we expand

$$0 = \sum_{t=1}^{n} \psi(Y_t - \hat{\mu}) = \sum_{t=1}^{n} \psi(Y_t - \mu) + (\hat{\mu} - \mu) \sum_{t=1}^{n} \psi'\big(Y_t - \mu^*\big),$$

where $|\mu^* - \mu| \le |\hat{\mu} - \mu|$. Under appropriate differentiability properties of ψ, $|\hat{\mu} - \mu| < \delta$ implies $|\psi'(Y_t - \hat{\mu}) - \psi'(Y_t - \mu)| < k_1(\delta)$, where k_1 is a constant that depends on δ only. Hence, recalling that $Y_t - \mu = X_t$,

$$\sqrt{n}(\hat{\mu} - \mu) \approx \frac{n^{-1/2} \sum_{t=1}^{n} \psi(X_t)}{n^{-1} \sum_{t=1}^{n} \psi'(X_t)}.$$

One can argue that the denominator converges in probability to $E[\psi'(X_1)]$. Furthermore, a martingale central limit theorem yields asymptotic normality of the nominator. Hence, the result follows. For further details, we refer to Beran and Schützner (2008). □

The most general statement is given in Beran and Schützner (2008). The Gaussian assumption used in the statement of Theorem 4.10 is replaced by

(A1) The random variables ξ_t are symmetric;
(A2) σ_t is a second-order stationary process with a finite variance such that ξ_t is independent of σ_s, $s \le t$ (but the sequences ξ_t and σ_t are not necessary independent).

Furthermore, as in Theorem 4.10, it is assumed that

(A3) The function $\psi(\cdot)$ is measurable and antisymmetric, that is, $\psi(x) = -\psi(-x)$, and $E[\psi^2(X_1)] < \infty$.

Finally, there is an additional assumption on extremal behaviour of the sequence $\psi(X_t)$, as well as further technical conditions on function ψ, see (A4)–(A5) and (A6)–(A8) in Beran and Schützner (2008).

Theorem 6.2 *Consider the stochastic volatility model defined above. Assume that* (A1)–(A3) *above as well as* (A4)–(A5) *and* (A6)–(A8) *in Beran and Schützner* (2008). *Then* (4.68) *holds.*

6.1.2 Estimation of Dependence Parameters

As mentioned in the introduction to this section, maximum likelihood estimation does not seem to be feasible for models of the form (6.1). To be more specific, let us consider the LMSV model,

$$X_t = \xi_t \ \exp\left(\sum_{j=1}^{\infty} a_k \varepsilon_{t-j}\right), \tag{6.3}$$

where the sequences ξ_t $(t \in \mathbb{Z})$ and ε_t $(t \in \mathbb{Z})$ are mutually independent. Furthermore, we shall assume that all random variables are standard normal and $\sum_{j=1}^{\infty} a_j^2 = 1$. Then the density p_X of X_t is

$$p_X(x) = \int_0^{\infty} \phi\big(\log(x/y)\big)\phi(y)\, dy \quad (x > 0),$$

where ϕ is the standard normal density. An analogous formula is valid for $x < 0$. Furthermore, the joint density of $(X_1, \ldots, X_n)$ can be written as an n-fold integral with respect to $\phi(y_1) \cdots \phi(y_n) dy_1 \cdots dy_n$. Consequently, finding the maximum likelihood estimator is extremely difficult. Breidt et al. (1998) use the Whittle estimator (see Sect. 5.5.2) applied to the logarithm of the squares instead.

Much easier is the application of semiparametric methods to stochastic volatility models. We consider for simplicity the LMSV model (6.3). Applying the log-

transformation to X_t^2, we obtain a new model

$$Y_t = \mu + 2\sum_{k=1}^{\infty} a_k \varepsilon_{t-k} + Z_t,$$

where $Z_t = \log \xi_t^2 - E(\log \xi_t^2)$, $\mu = E(\log \xi_t^2)$. The semiparametric estimators (in the Fourier or wavelet domain) can be applied directly to the sequence Y_t. We note that Y_t has the form of a long-memory sequence plus i.i.d. noise Z_t. Hence, we are exactly in the situation of the additive noise model considered in Example 5.13. Specifically, if we assume that the spectral density $f_{\tilde{X}}$ of the linear process $\tilde{X}_t := \sum_{k=1}^{\infty} a_k \varepsilon_{t-k}$ has the form $f_{\tilde{X}}(\lambda) = \lambda^{-2d} f_*(\lambda)$, then Y_t has the spectral density

$$\begin{aligned} f_Y(\lambda) &= f_{\tilde{X}}(\lambda) + \sigma_Z^2/(2\pi) = \lambda^{-2d} f_*(\lambda) + \sigma_Z^2/(2\pi) \\ &\approx \lambda^{-2d} f_*(0) + \sigma_Z^2/(2\pi) = \lambda^{-2d} f_*(0)\big(1 + O\big(\lambda^{2d}\big)\big), \end{aligned}$$

where $\sigma_Z^2 = \mathrm{var}(Z_1)$. According to the results in Sect. 5.8, the optimal mean squared error of a semiparametric estimator is then of order

$$m = O\big(n^{-\frac{4d}{4d+1}}\big), \qquad MSE(\hat{d}) = O\big(n^{-\frac{4d}{4d+1}}\big),$$

cf. Deo and Hurvich (2001), Hurvich and Soulier (2002) for log-periodogram regression (GPH), and Arteche (2004) for the local estimator. Hurvich et al. (2005b) show that a modified version of these estimators can outperform the GPH approach.

Furthermore, the techniques considered in Sect. 5.6.4 can be applied to the situation of additive noise as well. Consequently, we obtain the following asymptotic normality of the local Whittle estimator (see Hurvich et al. 2005b; Dalla et al. 2006). The result mimics Theorem 5.5. We have to adapt the bandwidth condition (LW3) there to the present context.

Theorem 6.3 *Consider the LMSV model given in* (6.3). *If*

$$m^{-1} + m^{2d+1} n^{-2d} \to 0, \quad \text{(LW3–SV)}$$

then $m^{1/2}(\hat{d}_{\mathrm{LW}} - d) \to N(0, 1/4)$.

Proof The proof follows similar lines as in the case of a linear process without the additive noise (see Sect. 5.6.4). The main step is asymptotic normality of a weighted sum of periodogram ordinates. Let us recall some notation: $\lambda_j = 2\pi j/n$, $j = 1, \ldots, m$, are Fourier frequencies, $b_j = -2\log \lambda_j$, $I_{n,Y}(\cdot)$ is the periodogram associated with the sequence $Y_1, \ldots, Y_n$. We re-write the decomposition (5.67) in the present context to obtain

$$\sum_{j=1}^{m} b_{j,m}\left[\frac{I_{n,Y}(\lambda_j)}{f_Y(\lambda_j)} - 1\right] + \sum_{j=1}^{m} b_{j,m}\left[\frac{I_{n,Y}(\lambda_j)}{g_Y(\lambda_j)} - \frac{I_{n,Y}(\lambda_j)}{f_Y(\lambda_j)}\right], \tag{6.4}$$

where $b_{j,m} = (b_j - \bar{b})/\sqrt{m}$ and $g_Y(\lambda) = |\lambda|^{-2d} f_*(\lambda)$. We deal with the first part only to illustrate the influence of the additive noise.

Let us decompose the difference between the normalized periodogram of Y_t and $\tilde{X}_t$:

$$\begin{aligned}\frac{I_{n,Y}(\lambda_j)}{f_Y(\lambda_j)} - \frac{I_{n,\tilde{X}}(\lambda_j)}{f_{\tilde{X}}(\lambda_j)} &= \frac{I_{n,\tilde{X}}(\lambda_j)}{f_Y(\lambda_j)} - \frac{I_{n,\tilde{X}}(\lambda_j)}{f_{\tilde{X}}(\lambda_j)} + \frac{I_{n,Z}(\lambda_j)}{f_Y(\lambda_j)} \\ &= \frac{f_{\tilde{X}}(\lambda_j) - f_Y(\lambda_j)}{f_Y(\lambda_j)} \frac{I_{n,\tilde{X}}(\lambda)}{f_{\tilde{X}}(\lambda)} + \frac{I_{n,Z}(\lambda_j)}{f_Y(\lambda_j)} \\ &= \frac{\sigma_Z^2/2\pi}{f_Y(\lambda_j)} \frac{I_{n,\tilde{X}}(\lambda)}{f_{\tilde{X}}(\lambda)} + \frac{I_{n,Z}(\lambda_j)}{f_Y(\lambda_j)}.\end{aligned}$$

We start with the term $I_{n,Z}(\lambda_j)/f_Y(\lambda_j)$. Since the random variables Z_t are i.i.d., the expected value of the normalized periodogram is one (cf. (4.139)). Thus

$$E\left(\frac{I_{n,Z}(\lambda_j)}{f_Y(\lambda_j)}\right) = E\left(\frac{I_{n,Z}(\lambda_j)}{f_Z(\lambda_j)}\right)\frac{f_Z(\lambda_j)}{f_Y(\lambda_j)} \sim \frac{\sigma_Z^2}{2\pi}|\lambda_j|^{2d} f_*^{-1}(\lambda_j) \le C(j/n)^{2d}.$$

Furthermore, we recall that $E[I_{n,\tilde{X}}(\lambda_j)/f_{\tilde{X}}(\lambda_j)]$ is uniformly bounded (in j) and that $f_Y(\lambda_j) = O((j/n)^{-2d})$. Thus we conclude

$$E\left|\frac{I_{n,Y}(\lambda_j)}{f_Y(\lambda_j)} - \frac{I_{n,\tilde{X}}(\lambda_j)}{f_{\tilde{X}}(\lambda_k)}\right| \le C(j/n)^{2d}.$$

Hence,

$$E\left|\sum_{j=1}^{m} b_{j,m}\left\{\frac{I_{n,Y}(\lambda_j)}{f_Y(\lambda_j)} - \frac{I_{n,\tilde{X}}(\lambda_j)}{f_{\tilde{X}}(\lambda_j)}\right\}\right| \le \sum_{j=1}^{m} |b_{j,m}|(j/n)^{2d}.$$

The bound is $\max_{1\le j\le m}|b_{j,m}|\sum_{j=1}^{m}(j/n)^{2d} = o(1)n^{-2d}m^{2d+1}$ which converges to 0 if (LW3–SV) holds. Consequently, the asymptotic behaviour of

$$\sum_{j=1}^{m} b_{j,m}\left[\frac{I_{n,Y}(\lambda_j)}{f_Y(\lambda_j)} - 1\right]$$

is the same as that of

$$\sum_{j=1}^{m} b_{j,m}\left[\frac{I_{n,\tilde{X}}(\lambda_j)}{f_{\tilde{X}}(\lambda_j)} - 1\right].$$

The latter was studied in Sect. 5.6.4. □

The result of Theorem 6.3 can be extended to the case of stochastic volatility models with leverage, i.e. when $\rho_{Z,\varepsilon} = E[Z_t\varepsilon_t] \neq 0$. In this case, the spectral den-

sity of $Y_t = \log X_t^2$ behaves like

$$f_Y(\lambda) \sim f_{\tilde{X}}(\lambda) + \frac{\sigma_Z^2}{2\pi} + \mathrm{Re}\big((1-e^{i\lambda})^{-d}\big)\frac{2\rho_{Z,\varepsilon}\sigma_Z^2\sqrt{f_*(0)}}{\sqrt{2\pi}}.$$

6.1.3 Tail Index Estimation

Consider the stochastic volatility models $X_t = \xi_t\sigma_t$ given in (6.1), where ξ_t are i.i.d. random variables with

$$P(\xi_t > x) \sim A\frac{1+\beta}{2}x^{-\alpha}, \qquad P(\xi_t < -x) \sim A\frac{1-\beta}{2}x^{-\alpha},$$

as $x \to \infty$, and $\alpha > 0$ is the tail index. Furthermore, it is assumed that the sequence σ_t is independent of ξ_t.

One of the most important problems when dealing with heavy tails is to estimate the tail index α. A standard (though not quite unproblematic; see, e.g. Resnick 1997) method is Hill's estimator. Setting $\gamma = \alpha^{-1}$, the Hill estimator of γ is defined by

$$\hat{\gamma}_n = \frac{1}{k}\sum_{j=1}^{k}\log\left(\frac{X_{n-j+1:n}}{X_{n-k:n}}\right) = \int_0^\infty \frac{\hat{T}_n(s)}{1+s}\,ds,$$

where

$$\hat{T}_n(s) = \frac{1}{k}\sum_{j=1}^{n} 1\big\{X_j > X_{n-k:n}(1+s)\big\}, \qquad T(s) = (1+s)^{-\alpha},$$

and $X_{k:n}$ are the order statistics of the sample $X_1, \ldots, X_n$. Since $\gamma = \int_0^\infty (1+s)^{-1}T(s)\,ds$, we have

$$\hat{\gamma}_n - \gamma = \int_0^\infty \frac{\hat{e}_n^*(s)}{1+s}\,ds,$$

where $\hat{e}_n^*(s)$ is the tail empirical process defined in Sect. 4.8.5:

$$\hat{e}_n^*(s) = \hat{T}_n(s) - T(s) \quad \big(s \in [0,\infty)\big).$$

Thus we can apply Theorem 4.37 to obtain the asymptotic distribution of the Hill estimator. Heuristically,

$$\sqrt{k_n}(\hat{\gamma}_n - \gamma) \to_d \int_0^\infty \frac{\tilde{B}(T(s))}{1+s}\,ds$$

where $\tilde{B}(u) = B(u) - uB(1)$ $(u \in [0,1])$ is a Brownian bridge. This integral is a centred normal random variable with variance γ^2.

Table 6.1 Simulated average values of and standard deviations of the Hill estimator $\hat{\gamma}$ (where $\gamma = 1/\alpha$) for an LMSV model with standard deviation $\beta = 0.2$ and sample size $n = 1000$

$\gamma = 1/\alpha$	d =0	0.2	0.4	0.45
0.667	mean = 0.6631	0.6670	0.6717	0.6659
	Std. dev. = 0.0664	0.0682	0.0648	0.0648
0.5	0.5001	0.5010	0.4988	0.5010
	0.0506	0.0500	0.0515	0.0503
0.25	0.2513	0.2518	0.2511	0.2530
	0.0249	0.0251	0.0251	0.0246
0.167	0.1711	0.1718	0.1791	0.1833
	0.0174	0.0170	0.0174	0.0188
0.1	0.1208	0.1226	0.1379	0.1452
	0.0114	0.0111	0.0140	0.0165

Corollary 6.1 *Under the assumptions of Theorem* 4.37, $\sqrt{k}(\hat{\gamma}_n - \gamma)$ *converges weakly to the centred Gaussian distribution with variance* γ^2.

This result allows us to construct confidence intervals for γ, with a user-chosen number k of extreme observations. The result is, in fact, the best possible rate of convergence for the Hill estimator for i.i.d. data (see Drees 1998). The surprising result is that it is possible to achieve the i.i.d. rate in spite of long memory. A detailed proof is given in Kulik and Soulier (2011). Further results can be found in Jach et al. (2012). Also, Corollary 6.1 can be extended to stochastic volatility models with leverage, i.e. when the sequences σ_t and ξ_t are not mutually independent, see Luo (2011).

Example 6.1 We simulate an LMSV model $X_t = \xi_t \exp(\beta\zeta_t)$, $(t = 1, \ldots, n = 1000)$ with $\beta > 0$, ξ_t independent Pareto random variables with tail parameter α and ζ_t a long memory FARIMA$(0, d, 0)$ sequence with standard normal innovations and dependence parameter $d \in [0, 1/2)$. We assume that $\{\zeta_t, t = 1, \ldots, n\}$ and $\{\xi_t, t = 1, \ldots, n\}$ are mutually independent. Table 6.1 shows that dependence of ζ_t does not influence tail index estimation, unless α is very large. Note, however, that from a practical point of view, large values of α are not interesting (if $\alpha > 4$, then the squares X_t^2 have a finite variance). Note also that further simulations (not reported here) illustrate that, if the variability coefficient β is large, then dependence may start to play a role for finite samples, although this influence disappears asymptotically, as indicated in Corollary 6.1. We refer to Luo (2011) for further details.

6.2 Statistical Inference for LARCH Processes

In this section, we consider LARCH processes. As in Sect. 6.1, we start with location estimation showing again that the asymptotic distribution of the sample mean

as well as for M-estimators is not affected by long memory, as long as function ψ that defines the M-estimator is antisymmetric and the noise variables ξ_t are symmetrically distributed (Beran 2006).

As for parameter estimation, it is reported in Giraitis et al. (2000b) that the graphical methods (KPSS, V/S, R/S) perform well for LARCH processes. Giraitis et al. (2003) claim further that for LARCH(∞) models V/S is superior to R/S and *KPSS*. There is no existing theory for semiparametric estimators for LARCH processes. Teyssière and Abry (2006) and Jach and Kokoszka (2008) study the numerical performance of wavelet estimators. Giraitis and Robinson (2001) argue that for ARCH(∞)-type models (including LARCH), the Whittle approach is less motivated than the maximum likelihood procedure that yields explicit results. However, it turns out that the issue is actually more complex. This and other detailed theoretical results on MLE type estimation, including asymptotic normality, can be found in Beran and Schützner (2009), and will be discussed below.

6.2.1 Location Estimation

Consider a time series $Y_t = \mu + X_t$ with residuals X_t generated by a long-memory LARCH process

$$X_t = \sigma_t \varepsilon_t, \tag{6.5}$$

$$\sigma_t = a + \sum_{j=1}^{\infty} b_j X_{t-j}. \tag{6.6}$$

Here ε_t are i.i.d. random variables with $E(\varepsilon_t) = 0$ and $E(\varepsilon_t^2) = 1$, and the coefficients are such that $a \neq 0$, $b_j \sim cj^{d-1}$ (as $j \to \infty$) for some $0 < d < \frac{1}{2}$ and $\sum b_j^2 < 1$. Since $cov(X_t, X_{t+k}) = 0$ ($k \neq 0$), the variance of the sample mean is not affected by the dependence in volatility, i.e. $\text{var}(\bar{X}) = \sigma_X^2/n$ where $\sigma_X^2 = \text{var}(X_t)$ (note that $\sigma_X^2 = \sigma_Y^2 = \text{var}(Y_t)$). Beran (2006) defines sufficient moment conditions under which a functional limit theorem holds for partial sums, namely

$$n^{-\frac{1}{2}} \sigma_X^{-1} S_n(u) = n^{-\frac{1}{2}} \sigma_Z^{-1} \sum_{t=1}^{[nu]} X_t \underset{D[0,1]}{\to} B(u)$$

where convergence is in the space $D[0,1]$ of càdlàg functions equipped with the Skorokhod metric and $B(u)$ denotes standard Brownian motion. More generally, for functions ψ satisfying some moment conditions, we can write

$$\begin{aligned} E\big[\psi(X_{t+k})\psi(X_t)\big] &= E\big\{E\big[\psi(X_{t+k})\psi(X_i) \mid \mathscr{F}_{t+k-1}\big]\big\} \\ &= E\big\{\psi(X_t)E\big[\psi(\varepsilon_{t+k}\sigma_{t+k}) \mid \mathscr{F}_{t+k-1}\big]\big\} \end{aligned}$$

where $\mathscr{F}_t$ denotes the σ-algebra generated by ε_j $(j \le t)$. In particular, if the distribution of ε_t is symmetric and ψ is antisymmetric, i.e. $\psi(-x) = -\psi(x)$, then $E[\psi(\varepsilon_{t+k}\sigma_{t+k}) \mid \mathscr{F}_{t+k-1}] = 0$ so that $\psi(X_t)$ $(t \in \mathbb{Z})$ is a martingale difference and $\mathrm{var}(\sum \psi(X_t)) = O(n)$. This has direct implications for M-estimators of the location parameter μ defined as solutions of $\sum_{i=1}^{n} \psi(Y_t - \hat{\mu}) = 0$. It can be shown that under regularity conditions, the asymptotic distribution of $\hat{\mu}$ is the same as for $S_{n;\psi} = E^{-1}[\psi'(X_1)]S_n(1)$. Thus, we have

$$\sqrt{n}(\hat{\mu} - \mu) \underset{d}{\to} N\big(0, \sigma_\psi^2\big)$$

where $\sigma_\psi^2 = E[\psi^2(X_1)]E^{-2}[\psi'(X_1)]$, cf. Sect. 5.2.3. In other words, the asymptotic distribution of M-estimators of location is undisturbed by LARCH type (long-range) dependence in volatility, and is, in fact, the same as if observations were i.i.d. For detailed conditions on ψ and ε_t, see Beran (2006). In conclusion, approximate $(1-\alpha)$-confidence intervals for μ may be given by

$$\hat{\mu} \pm z_{1-\alpha/2}\sigma_\psi n^{-\frac{1}{2}} \tag{6.7}$$

where $z_{1-\alpha/2}$ is the standard normal $(1-\alpha/2)$-quantile.

A completely different result is obtained, however, if ψ is not antisymmetric or if ε_t are not symmetrically distributed such that $E[X_1\psi(X_1)] \neq 0$. In this case, $\hat{\mu}$ has a slower rate of convergence and limit theorems derived in Berkes and Horváth (2003) apply; see also Sect. 4.2.8. From the applied point of view, this means that it is important to check symmetry of the innovation process.

6.2.2 Estimation of Dependence Parameters

6.2.2.1 Basic Definitions and Problems

Consider a parametric long-memory LARCH process $(X_t, \sigma_t)_{t\in\mathbb{Z}}$ as in (6.5) and (6.6), where ε_t are i.i.d. continuous random variables with density function p_ε, $E(\varepsilon_t) = 0$ and $E(\varepsilon_t^2) = 1$, $a \neq 0$, $b_j \sim cj^{d-1}$ (as $j \to \infty$) for some $0 < d < \frac{1}{2}$, $\sum b_j^2 < 1$ and $b_j = b_j(\theta)$ with $\theta = (d, a, c, \ldots)$ denoting a finite dimensional parameter vector. The true parameter value will be denoted by θ^0. In the following, we summarize results from Beran and Schützner (2009). For simplicity of notation, we will consider the case with exact equality $b_j = cj^{d-1}$ $(j \ge 1)$ which implies that $\theta = (d, a, c)^T$.

Since σ_t is given explicitly as a function of past observations X_s $(s \le t-1)$, a plausible approach to estimating θ is to use the conditional likelihood function of $\varepsilon_t(\theta) = X_t/\sigma_t(\theta)$. If $\sigma_t(\theta)$ can be calculated exactly and θ is equal to the true parameter θ^0, then $\varepsilon_t(\theta)$ $(t \in \mathbb{Z})$ coincides with the innovations ε_t. Since ε_t $(t \in \mathbb{Z})$

are i.i.d. with density p_ε, the log-likelihood function can be written as

$$L_n(\theta) = \sum_{t=1}^{n} \log p_\varepsilon\big(\varepsilon_t(\theta)\big).$$

If differentiation with respect to θ is possible, then the maximum likelihood estimator of θ^0 can be defined as a solution of

$$\dot{L}_n(\hat{\theta}) := \dot{L}_n(\theta)|_{\theta=\hat{\theta}} = 0$$

(where "$\cdot$" denotes differentiation with respect to θ). In particular, if p_ε is a normal density function with mean zero, then

$$-\frac{2}{n} L_n(\theta) = \frac{1}{n} \sum_{t=1}^{n} \frac{X_t^2}{\sigma_t^2(\theta)} + \log \sigma_t^2(\theta) + \log 2\pi \tag{6.8}$$

and

$$\begin{aligned} -\frac{2}{n} \dot{L}_n(\theta) &= \frac{\partial}{\partial \theta} \left[\sum_{t=1}^{n} \varepsilon_t^2(\theta) + \log \sigma_t^2(\theta) \right] \\ &= 2 \sum_{t=1}^{n} \dot{\varepsilon}_t(\theta) \varepsilon_t(\theta) + \frac{\dot{\sigma}_t^2(\theta)}{\sigma_t^2(\theta)}. \end{aligned}$$

If the innovations ε_t are not normally distributed, then this function can still be used to define an estimator $\hat{\theta}$, but the solution no longer coincides with the MLE and is therefore often called a pseudo- or quasi-maximum likelihood estimator (PMLE or QMLE).

If all quantities in the last equation are well defined, then the asymptotic distribution of $\hat{\theta}$ can be derived quite easily because $\dot{\varepsilon}_t(\theta^0)\varepsilon_t(\theta^0)$ is a martingale difference. However, in contrast to short-memory volatility models (Lee and Hansen 1994; Lumsdaine 1996; Berkes et al. 2003; Robinson and Zaffaroni 2006; Francq and Zakoian 2008; Truquet 2008), for LARCH processes with slowly decaying coefficients $b_j \sim cj^{d-1}$ $(0 < d < \frac{1}{2})$ several complications arise. First of all, it is not obvious whether $\sigma_t(\theta)$ is an ergodic process (see, e.g. Walters 2000; Krengel 1985; Petersen 1989). Moreover, for $\theta \neq \theta^0$, it is not even clear whether $\varepsilon_t(\theta) = \sum_{j=1}^{\infty} b_j(\theta) X_{t-j}$ is finite with probability one. (Note that for $\theta = \theta^0$ this problem disappears because $\varepsilon_t(\theta^0)$ is almost surely equal to the random variable ε_t.) The reason is that $\sum b_j(\theta) = \infty$ implies $\sum |b_j X_{t-j}| = \infty$ almost surely unless $P(\varepsilon_t = 0) = 1$. Similarly, it is not clear whether and in which sense the derivative of $\varepsilon_t(\theta)$ with respect to θ exists (this problem occurs even for $\theta = \theta^0$), and whether the derivative is equal to $\sum \dot{b}_j(\theta) X_{t-j}$. An additional technical property that has to be established when studying the asymptotic distribution of $\hat{\theta}$ is the measurability of infima involving $\sigma_t(\theta)$ on the (uncountable) set Θ.

Apart from these questions, there is also the problem that $\sigma_t^2(\theta)$ may become arbitrarily small. In particular, for $\theta \neq \theta^0$, $E[L_n(\theta)]$ may be infinite or not defined. In fact, Francq and Zakoian 2008 (also see Truquet 2008) showed that, because of this reason, even in the case of short memory with a finite number of nonzero coefficients b_j the estimator based on (6.8) is not consistent.

Finally, for long-memory LARCH models, the issue that has to be addressed is that $\sigma_t(\theta)$ depends on the entire past X_s ($s \leq t-1$), whereas the only available observations are $X_1, \ldots, X_n$. This means that σ_t cannot be calculated exactly. Because of the slow decay of b_j, finite approximations may not be very good.

6.2.2.2 Ergodicity

Let us start with the fundamental question of ergodicity. Ergodicity of the process $\sigma_t(\theta)$ ($t \in \mathbb{Z}$) follows once the existence of a measurable function $f : \mathbb{R}^\infty \to \mathbb{R}$ is established for which $\sigma_t(\theta) = f(\varepsilon_{t-1}, \varepsilon_{t-2}, \ldots)$ almost surely (Stout 1974, Theorem 3.5.8). In view of the definition

$$\sigma_t(\theta) = a + a \sum_{k=1}^{\infty} \sum_{j_1,\ldots,j_k=1}^{\infty} b_{j_1}(\theta) \cdots b_{j_k}(\theta) \varepsilon_{t-j_1} \cdots \varepsilon_{t-j_1-\cdots-j_k}, \tag{6.9}$$

the natural choice of f is

$$f = a + a \sum_{k=1}^{\infty} f_k$$

with

$$f_k(x_1, x_2, \ldots) = \sum_{1 \leq m \leq k} \sum_{\substack{j_1,\ldots,j_m=1 \\ j_1+\cdots+j_m=k}}^{\infty} b_{j_1} \cdots b_{j_m} x_{j_1} \cdots x_{j_1+\cdots+j_m}.$$

Almost sure convergence of $\sum f_k$ follows from the fact that, for each fixed t,

$$M_t(k) = f_k(\varepsilon_{t-1}, \varepsilon_{t-2}, \ldots) \quad (k \in \mathbb{N})$$

is a martingale difference with respect to the sequence of σ-algebras $\mathcal{F}_k = \sigma(M_t(l), l \leq k)$. Measurability of f follows, for instance, from Corollary 2.1.3 in Straumann (2004).

6.2.2.3 Summability, Continuity and Differentiability

Next consider the existence of $\sigma_t(\theta)$ ($\theta \in \Theta$) and its derivatives. If the coefficients b_j were absolutely summable, then answering these questions would be straightforward because $\sum |b_j| < \infty$ implies absolute summability of the right-hand side of (6.9) which, in turn, implies that $\sigma_t(\theta)$ inherits the differentiability properties

of $b_j(\theta)$. For nonsummable coefficients, these arguments do not apply. The solution proposed in Beran and Schützner (2009) is to consider $\sigma_t(\theta)$ (for fixed t) as a stochastic process with index $\theta \in \Theta$. To carry over the properties of $b_j(\theta)$ to $\sigma_t(\theta)$, the process $\sigma_t(\theta)$ ($\theta \in \Theta$) is assumed to be separable. More specifically, the technical condition can be written down as follows:

(S) For every $t \in \mathbb{Z}$, $(\sigma_t(\theta))_{\theta\in\Theta}$ is a separable stochastic process on Θ, i.e. for every open set $A \subset \Theta$ and closed interval B, the sets $\{\omega : \sigma_t(\theta) \in B, \forall\theta \in A\}$ and $\{\omega : \sigma_t(\theta) \in B, \forall\theta \in A \cap \mathbb{Q}^3\}$ differ only on a set $N \subset N_0$ where $P(N_0) = 0$.

Note that the original process $(\sigma_t(\theta))_{\theta\in\Theta}$ can always be replaced by a separable version (see Theorem 2.4 in Doob 1953). Before establishing differentiability of $\sigma_t(\theta)$, we recall two different definitions of derivatives that are particularly useful for stochastic processes.

Definition 6.1 A stochastic process $\xi(x)$ ($x \in [a, b]$) is uniformly mean squared differentiable (u.m.s.-differentiable), if there exists a process $\zeta(x) =: \xi'(x)$ ($x \in [a, b]$) such that

$$E\left[\left(\frac{\xi(x+h)-\xi(x)}{h} - \zeta(x)\right)^2\right] \underset{h\to 0}{\to} 0$$

uniformly in $x \in (a, b)$. The process $\xi'(x)$ is also called the L^2-derivative of $\xi(x)$.

Definition 6.2 Let $\Psi(a, b)$ be the set of (test) functions ψ that are infinitely continuously differentiable on (a, b) and such that the closure $\bar{K}_\psi$ of the support $K_\psi = \{x : \psi(x) \neq 0\}$ is a compact subset of (a, b). A function $g \in L^2(a, b)$ is called a generalized (or distributional) derivative of a function $f \in L^2(a, b)$, if

$$\int_a^b g(x)\psi(x)\,dx = -\int_a^b f(x)\psi'(x)\,dx$$

for all $\psi \in D(a, b)$.

Note that generalized derivatives extend differentiation to functions that are not differentiable in the usual sense (or more generally, to generalized functions). For an elementary introduction to generalized derivatives, see, e.g. Lighthill (1958). For a more detailed account of the theory and further references, see, e.g. Gelfand and Shilov (1966–1968), Kanwal (2004), Strichartz (1994), Vladimirov (2002), Zemanian (2010).

Example 6.2 Let $H(x) = 1\{x \geq 0\}$ be the Heaviside function defined on $(-\infty, \infty)$. For $\psi \in \Psi(-\infty, \infty)$, we then have

$$-\int_{-\infty}^{\infty} H(x)\psi'(x)\,dx = -\big[\psi(\infty) - \psi(0)\big] = \psi(0).$$

Thus, the generalized derivative H' is equal to the Dirac delta function δ defined by $\int \delta(x)\psi(x)\,dx = \psi(0)$ (for all $\psi \in \Psi(-\infty, \infty)$).

The following result is derived in Beran and Schützner (2009).

Theorem 6.4 *Suppose that there are constants $d_u < \frac{1}{2}$ and $0 < C < 1$ such that $b_j = cj^{d-1}$ with $d \in [0, d_u]$, $c \in [0, c_u(d)]$ and $c_u(d) = C/\sqrt{\sum_{j=1}^{\infty} j^{2d-2}}$. Assume furthermore that (S) holds. Then $\sigma_t(\theta)$ is almost surely infinitely many times differentiable in θ in the generalized sense, and the kth generalized partial derivative w.r.t. θ is given by*

$$\frac{\partial^k}{\partial\theta_{j_1}\cdots\partial\theta_{j_k}}\sigma_t(\theta) = \sum_{j=1}^{\infty} \frac{\partial^k}{\partial\theta_{j_1}\cdots\partial\theta_{j_k}} b_j(\theta) X_{t-j},$$

i.e. we can write $\dot{\sigma}_t := \partial/\partial\theta\sigma_t = \sum \dot{b}_j X_{t-j}$.

This theorem follows by applying the following results.

Lemma 6.1 *Let $\xi(x)$ $(x \in [a, b])$ be a separable and u.m.s.-differentiable process with the L^2-derivative $\xi'(x)$. Then $\xi'(x)$ is also a generalized derivative of $\xi(x)$.*

Lemma 6.2 (Kolmogorov) *Let $\xi(x)$ $(x \in [a, b])$ be such that $E[\xi(x)] = 0$, $E[\xi^2(x)] < \infty$ and*

$$E\big[\big|\xi(x_1) - \xi(x_2)\big|^{\alpha}\big] \le \text{const}|x_1 - x_2|^{1+\beta}$$

for some $\alpha, \beta > 0$. Then there exists a version of $\xi(x)$ with almost surely continuous paths.

Lemma 6.3 *Let $\xi(x)$ $(x \in [a, b])$ be a separable process, m times u.m.s.-differentiable with the L^2-derivatives $\xi^{(k)}$ $(k \le m)$ and such that the paths of $\xi^{(k)}$ $(k \le m)$ are almost surely continuous. Then $\xi(x)$ is also $(m-1)$-times continuously differentiable in the generalized sense.*

Note that the last lemma is essentially an application of Sobolev's famous embedding theorem (see, e.g. Adams and Fournier 2003). Using these lemmas, the theorem can be proved in three steps. First of all, it is obvious that the only problem with respect to differentiability occurs for d. The lemmas were therefore formulated for the case of a one-dimensional index x only. The other parameter components can be fixed, and we can write $\sigma_t = \sigma_t(d)$ and $\dot{b}_j = \frac{\partial}{\partial d} b_j$.

The first step of the proof is to show that $\sum \dot{b}_j X_{t-j}$ is indeed the L^2-derivative of σ_t in the u.m.s.-sense. This can be done directly by showing that

$$E\left[\left(\frac{\sigma_t(d+h) - \sigma_t(d)}{h} - \sum \dot{b}_j X_{t-j}\right)^2\right] \le \text{const} \cdot h^2$$

and similar inequalities for higher derivatives. In a second step, one shows in a similar way that the condition in Lemma 6.2 holds. Since $\sigma_t(d)$ $(d \in [d, d_u])$ is assumed to be separable, almost sure continuity of the paths of $\dot{\sigma}_t(d)$ $(d \in [d, d_u])$ and

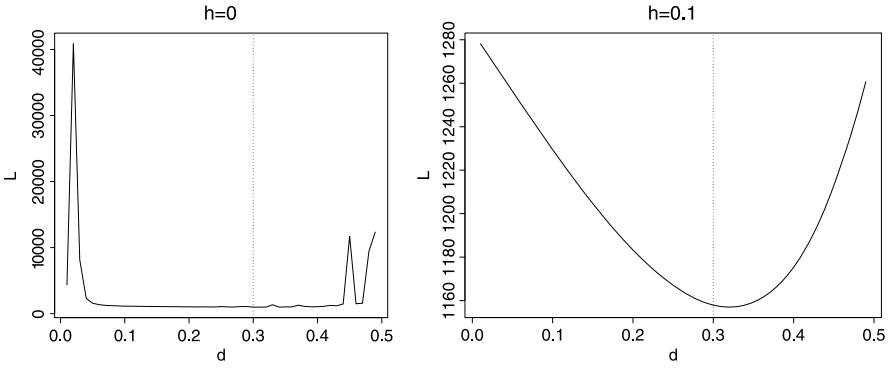

Fig. 6.1 Log-likelihood function $L_{n,h}(d)$ as a function of d for a simulated LARCH process with $b_j = cj^{d-1}$ and $d = 0.3$. The *left panel* shows $L_{n,h}$ for $h = 0$ whereas on the *right* $h = 0.1$ was used

higher order L^2-derivatives follows from Lemma 6.2. Finally, Lemma 6.3 implies that these are also derivatives in the generalized sense and the generalized derivatives $\sigma_t^{(k)}(d) = \frac{\partial^k}{\partial d^k}\sigma_t(d)$ $(k \le m-1)$ are almost surely continuous.

In a similar but slightly more involved manner, it can be shown that, under assumption (S), one can find bounds for $E(\sup_{\theta\in\Theta} |\sigma_t(\theta)|^m)$ $(m \ge 1)$ in terms of $\sup_{\theta\in\Theta} E(|\sigma_t(\theta)|^m)$ and $\sup_{\theta\in\Theta} E(|\dot{\sigma}_t(\theta)|^m)$. This is very useful for proving consistency (see below).

6.2.2.4 A Modified Log-likelihood Function

As mentioned above, a QMLE based on L_n in (6.8) is not consistent even in the case of short memory. The reason is that σ_t can be arbitrarily close to zero. Beran and Schützner (2009) therefore suggest a modified (quasi-) log-likelihood function. Multiplied by -1 it is given by

$$L_{n,h}(\theta) = n^{-1}\sum_{t=1}^{n}\left(\frac{X_t^2}{\sigma_t^2(\theta)+h} + \log\bigl[\sigma_t^2(\theta)+h\bigr]\right) \tag{6.10}$$

for some $h > 0$. Computationally, the effect of the correction is a regularization in the sense that the function $L_{n,h}$ becomes smoother, with clearly identifiable local minima. This is illustrated in Fig. 6.1 where $L_{n,h}$ is plotted against d (for fixed a and c) for $h = 0$ (left) and $h = 0.1$ (right), respectively. The correct value of $d = 0.3$ is indicated by a dotted vertical line. Obviously, for $h = 0$, the function is not suitable for minimization whereas the minimum for $h = 0.1$ is clearly visible and close to the true value.

The function $L_{n,h}$ can also be interpreted as a robust version of L_n in the following sense. Suppose that ε_t are Gaussian and instead of X_t we observe a perturbed process $Y_t = X_t + \zeta_t$ where ζ_t are i.i.d. $N(0,h)$-distributed, and independent of X_t.

Then $\operatorname{var}(Y_t \mid X_s, s \le t-1) = \sigma_t^2 + h$ so that the conditional log-likelihood function of $Y_1, \ldots, Y_n$ is given by

$$L_{n,Y}(\theta) = n^{-1} \sum_{t=1}^{n} \left[\frac{(X_t + \zeta_t)^2}{\sigma_t^2(\theta) + h} + \log\bigl(\sigma_t^2(\theta) + h\bigr) \right].$$

Integrating out ζ_t, we obtain $E_\zeta[L_{n,Y}(\theta)] = L_{n,h}(\theta)$.

6.2.2.5 Consistency

Let $\hat{\theta}_{n,h}$ be defined by minimizing $L_{n,h}$ with respect to θ and denote by θ^0 the true value of θ. Sufficient conditions for almost sure consistency of $\hat{\theta}_{n,h}$ are: (a) $\theta^0 \in \Theta^0$ (with θ^0 denoting the true parameter and Θ^0 the interior of Θ) and Θ is compact; (b) $L_{n,h}(\theta)$ is continuous and $\sup_\theta |L_{n,h}(\theta) - L_h(\theta)|$ converges a.s. to zero where

$$L_h(\theta) = E\bigl[L_{n,h}(\theta)\bigr]$$

and (c) $L_h(\theta)$ has a unique minimum at $\theta = \theta^0$.

Continuity of $L_{n,h}(\theta)$ follows from continuity of $\sigma_t^2(\theta)$ discussed in the previous section. Pointwise a.s. convergence of $|L_{n,h}(\theta) - L_h(\theta)|$ follows from ergodicity of $\sigma_t(\theta)$ (for each $\theta \in \Theta$) and

$$\sup_{\theta\in\Theta} E\left[\left| \frac{X_t^2 + h}{\sigma_t^2(\theta) + h} + \log\bigl(\sigma_t^2(\theta) + h\bigr) \right| \right] \le \text{const} \cdot \left\{ E\bigl[X_t^2\bigr] + h + \sup_{\theta\in\Theta} E\bigl[\sigma_t^2(\theta)\bigr] \right\}.$$

Since Θ is assumed to be compact, $\sup_{\theta\in\Theta} E[\sigma_t^2(\theta)] < \infty$ can be shown and thus Birkhoff's ergodic theorem implies $|L_{n,h}(\theta) - L_h(\theta)| \to 0$ almost surely. The convergence of $\sup_\theta |L_{n,h}(\theta) - L_h(\theta)|$ follows from equicontinuity of $L_{n,h}(\theta)$ which requires slightly more involved arguments (see Beran and Schützner 2009) involving certain moment conditions on ε_t.

The proof of (c) follows from

Lemma 6.4 *If ε_t are continuous random variables with density function p_ε, then*

$$P\bigl(\sigma_t(\theta) = 0\bigr) = 0 \quad (\textit{for all } t \textit{ and } \theta),$$
$$P\bigl(\sigma_t^2(\theta) = \sigma_t^2\bigl(\theta^0\bigr)\bigr) = 1 \quad \Longrightarrow \quad \theta = \theta^0$$

and

$$\theta \neq \theta^0 \quad \Longrightarrow \quad L_h(\theta) > L_h\bigl(\theta^0\bigr).$$

Proof Defining the set $N_t = \{\omega : \sigma_t(\theta) = 0\}$, the first equation means that $P(N_t) = 0$. To prove this, consider $\omega \in N_t \cap N_{t-1}^c$, i.e. we look at a realization

of the process such that $\sigma_t(\theta) = 0$ but $\sigma_{t-1}(\theta) \neq 0$. Then

$$0 = \sigma_t(\theta) = a + b_1(\theta)\overbrace{\varepsilon_{t-1}\sigma_{t-1}(\theta)}^{X_{t-1}} + \sum_{j=2}^{\infty} b_j(\theta) X_{t-j}$$

so that

$$\varepsilon_{t-1} = -\frac{1}{b_1(\theta)\sigma_{t-1}(\theta)}\left(a + \sum_{j=2}^{\infty} b_j(\theta) X_{t-j}\right).$$

However, the right-hand side involves only ε_s $(s \leq t-2)$ which is independent of the left-hand side ε_{t-1}. Therefore, since the ε_t's are assumed to be continuous variables, this equality can only occur with probability zero. In other words, $P(N_t \cap N_{t-1}^c) = 0$. The same arguments lead to $P(N_{t,k}) = 0$ where

$$N_{t,k} = \bigcap_{i=0}^{k-1} N_{t-i} \cap N_{t-k}^c \quad (k \geq 1).$$

Since $N_t = \bigcup_{k=1}^{\infty} N_{t,k}$, we obtain $P(N_t) = P(\sigma_t(\theta) = 0) = 0$.

Analogous arguments can be used to show that $P(\sigma_t^2(\theta) = \sigma_t^2(\theta^0)) = 1$ implies $\theta = \theta^0$. Finally, the last statement in the lemma follows from

$$L_h(\theta) - L_h\big(\theta^0\big) = E\left[\frac{\sigma_t^2(\theta^0) + h}{\sigma_t^2(\theta) + h} - \log\frac{\sigma_t^2(\theta^0) + h}{\sigma_t^2(\theta) + h} - 1\right]$$

and the inequality $u - \log u - 1 > 0$ $(u \neq 1)$. □

6.2.2.6 Asymptotic Normality

By similar arguments as for $L_{n,h}$, one can show that $\sup_\theta \|\dot{L}_{n,h}(\theta) - \dot{L}_h(\theta)\|$ and $\sup_\theta \|\ddot{L}_{n,h}(\theta) - \ddot{L}_h(\theta)\|$ (with the matrix norm $\|A\| = \sqrt{\mathrm{tr}(A^T A)}$) converge to zero almost surely. The asymptotic distribution of $\hat{\theta}_{n,h}$ can therefore be obtained by the Taylor approximation

$$\begin{aligned} 0 = L_{n,h}(\hat{\theta}_{n,h}) &\approx \dot{L}_{n,h}\big(\theta^0\big) + \ddot{L}_{n,h}\big(\theta^0\big)\big(\hat{\theta}_{n,h} - \theta^0\big) \\ &\approx \dot{L}_{n,h}\big(\theta^0\big) + \ddot{L}_h\big(\theta^0\big)\big(\hat{\theta}_{n,h} - \theta^0\big) \end{aligned} \tag{6.11}$$

implying

$$\hat{\theta}_{n,h} - \theta^0 \approx -\big[\ddot{L}_h\big(\theta^0\big)\big]^{-1}\dot{L}_{n,h}\big(\theta^0\big),$$

where $\ddot{L}_h(\theta) = E[\ddot{L}_{n,h}(\theta)]$. Thus, apart from the deterministic matrix $\ddot{L}_h(\theta^0)$, the asymptotic distribution of $\hat{\theta}_{n,h}$ is determined by the asymptotic distribution of

$\dot{L}_{n,h}(\theta^0)$ where

$$\dot{L}_{n,h}(\theta) = n^{-1}\frac{\partial}{\partial\theta}\left\{\sum_{t=1}^{n}\frac{X_t^2}{\sigma_t^2(\theta)+h} + \log\left[\sigma_t^2(\theta)+h\right]\right\}$$
$$= n^{-1}\sum_{t=1}^{n}\dot{\ell}_{t,h}(\theta)$$

with

$$\dot{\ell}_{t,h}(\theta) = 2\left(1 - \frac{X_t^2+h}{\sigma_t^2(\theta)+h}\right)\frac{\sigma_t(\theta)}{\sigma_t^2(\theta)+h}\dot{\sigma}_t(\theta).$$

For $\theta = \theta^0$, $E[\dot{\ell}_{t,h}(\theta^0) \mid \varepsilon_s, s \le t-1] = 0$ so that $\dot{\ell}_{t,h}(\theta^0)$ is a martingale difference. Therefore,

$$\sqrt{n}\dot{L}_{n,h}\left(\theta^0\right) \underset{d}{\to} Z_1$$

where Z_1 is a normal random vector with zero mean and covariance matrix

$$G_h = E\left[\dot{\ell}_{t,h}\left(\theta^0\right)\dot{\ell}_{t,h}^T\left(\theta^0\right)\right]$$
$$= 4E\left\{\frac{\sigma_t^6(\theta^0)[E(\varepsilon_t^4)-1]}{(\sigma_t^2(\theta^0)+h)^4}\dot{\sigma}_t\left(\theta^0\right)\dot{\sigma}_t^T\left(\theta^0\right)\right\}.$$

For the matrix $\ddot{L}_h(\theta^0)$, we have

$$\ddot{L}_h\left(\theta^0\right) = H_h = 4E\left[\frac{\sigma_t^2(\theta^0)}{(\sigma_t^2(\theta^0)+h)^2}\dot{\sigma}_t\left(\theta^0\right)\dot{\sigma}_t^T\left(\theta^0\right)\right].$$

Thus, we obtain (see Beran and Schützner 2009):

Theorem 6.5 *Suppose that H_h is nonsingular. Then, under suitable moment conditions,*

$$\sqrt{n}\left(\hat{\theta}_{n,h} - \theta^0\right) \to Z \underset{d}{\sim} N(0, V_h)$$

with covariance matrix

$$V_h = H_h^{-1}G_hH_h^{-1}.$$

It is interesting to see that in general H_h need not be of full rank. A sufficient condition for nonsingularity of H_h is that ε_t are continuous random variables. The proof essentially follows from $P(\sigma_t = 0) = 0$. To see this, we have to consider the quadratic form

$$u^TH_hu = 4E\left[\frac{\sigma_t^2}{(\sigma_t^2+h)^2}u^T\dot{\sigma}_t\dot{\sigma}_t^Tu\right].$$

Since σ_t is not zero with probability one, the condition $u^T H_h u = 0$ can only be true if $P(\dot{\sigma}_t = 0) > 0$ or if $u = 0$. Considering, for instance, the specific case with $\theta = (a, c, d)^T$ and $b_j = cj^{d-1}$ $(j \geq 1)$, the equation $u^T \dot{\sigma}_t = 0$ can be written as

$$\begin{aligned} 0 &= u_1 \frac{\partial}{\partial a}\sigma_t + u_2 \frac{\partial}{\partial c}\sigma_t + u_3 \frac{\partial}{\partial d}\sigma_t \\ &= u_1 + \sum_{j=2}^{\infty} \bigl(u_2 j^{d-1} + u_3 c \log j \cdot j^{d-1}\bigr) X_{t-j} + u_2 \sigma_{t-1} \varepsilon_{t-1}. \end{aligned}$$

Since $P(\sigma_{t-1} = 0) = 0$, this can be rewritten as

$$-u_2 \varepsilon_{t-1} = \sigma_{t-1}^{-1} \left[u_1 + \sum_{j=2}^{\infty} \bigl(u_2 j^{d-1} + u_3 c \log j \cdot j^{d-1}\bigr) X_{t-j} \right].$$

However, the left-hand side is independent of the right-hand side. Since ε_t (and hence also X_t) has a continuous distribution, equality can only occur with positive probability if all components of u are zero. In other words, H_h is of full rank. Note that in a similar manner G_h can be shown to be positive definite.

It is interesting to look at the asymptotic covariance matrix of $\hat{\theta}_{n,h}$ for small values of h. Letting h tend to zero, we obtain in the limit

$$\lim_{h \to 0} V_h = \bigl[E\bigl(\varepsilon_t^4\bigr) - 1\bigr] H_0^{-1}$$

with

$$H_0 = 4E\left[\frac{\dot{\sigma}_t(\theta^0) \dot{\sigma}_t^T(\theta^0)}{\sigma_t^2(\theta^0)} \right].$$

In particular, if $E[\sigma_t^{-2}(\theta^0)] = \infty$, then the asymptotic variance of $\hat{\theta}_1 = \hat{a}$ is zero. (Note, however, that this does not necessarily follow for the other components $\hat{\theta}_2 = \hat{c}$ and $\hat{\theta}_3 = \hat{d}$.) It is also remarkable that $\hat{\theta}_{n,h}$ has the same rate of convergence, and formally also the same type of asymptotic covariance matrix, as estimators of comparable parameters for GARCH(p, q) and ARCH(∞) processes (cf. Berkes et al. 2003; Robinson and Zaffaroni 2006).

6.2.2.7 Estimation Given the Finite Past

Since σ_t depends on the complete past X_s $(s \leq t - 1)$, it cannot be calculated exactly. The simplest approximation is obtained by truncating the sum, i.e. setting all unobserved values X_s $(s \leq 0)$ equal to zero. This leads to the approximate estimator

$$\theta_{n,h}^* := \arg\min_{\theta \in \Theta} L_{n,h}^*(\theta),$$

where

$$L^*_{n,h}(\theta) := \frac{1}{n}\sum_{t=1}^{n} \frac{X_t^2 + h}{\bar\sigma_t^2(\theta) + h} + \ln\bigl(\bar\sigma_t^2(\theta) + h\bigr)$$

and

$$\bar\sigma_t(\theta) = a(\theta) + \sum_{j=1}^{t-1} b_j(\theta) X_{t-j}.$$

However, because of the slow decay $b_j \sim cj^{d-1}$, the error $\sigma_t(\theta) - \bar\sigma_t(\theta)$ may be quite large (note that the error is larger for smaller values of t). In fact, we have, as $t \to \infty$,

$$E\bigl[\bigl(\sigma_t(\theta) - \bar\sigma_t(\theta)\bigr)^2\bigr] = \sum_{j=t}^{\infty} b_j^2(c,d) \sim c_1 t^{2d-1}.$$

The question is therefore whether this approximation changes the asymptotic distribution of the estimator. As before, a Taylor expansion yields (cf. (6.11))

$$0 = \dot L^*_n\bigl(\theta^*_{n,h}\bigr) = \dot L^*_{n,h}(\theta_0) + \ddot L^*_{n,h}(\tilde\theta) \cdot \bigl(\theta^*_{n,h} - \theta^0\bigr)$$

so that the asymptotic distribution of $\theta^*_{n,h}$ follows from the asymptotic distribution of $\dot L^*_{n,h}(\theta^0)$. The latter is the same as for $\dot L_{n,h}(\theta^0)$ provided that

$$\Delta_n := \sqrt{n}\bigl(\dot L^*_{n,h}\bigl(\theta^0\bigr) - \dot L_{n,h}\bigl(\theta^0\bigr)\bigr) \xrightarrow{p} 0$$

as $n \to \infty$ which means that

$$\frac{1}{\sqrt{n}}\sum_{t=1}^{n} \frac{\dot{\bar\sigma}_t(\theta)\bar\sigma_t(\theta)(X_t^2 + h)}{\bar\sigma_t^2(\theta) + h}\left(\frac{1}{\bar\sigma_t^2(\theta) + h} - \frac{1}{\sigma_t^2(\theta) + h}\right) \to_p 0.$$

Using the mean value theorem for $(x^2 + h)^{-1}$ and the asymptotic behaviour of $E[(\sigma_t(\theta) - \bar\sigma_t(\theta))^2]$, an upper bound for $E(|\Delta_n|)$ can be given by $E(|\Delta_n|) \le \text{const} \cdot n^d$. Unfortunately, for $d > 0$, this bound does not converge to zero. The errors $E[(\sigma_t(\theta) - \bar\sigma_t(\theta))^2]$ do not decay fast enough (in t) to be negligible when summing over all values of t. As a simple remedy, Beran and Schützner (2009) propose to use only those time points where a sufficient number of past observations is available. Specifically, let $m_n = [n^\beta] - 1$ for some $0 < \beta < 1$ where $[\cdot]$ is denotes the integer part,

$$L_{n,h;\beta}(\theta) := \frac{1}{m_n}\sum_{t=n-m_n}^{n} \frac{X_t^2 + \varepsilon}{\bar\sigma_t^2(\theta) + \varepsilon} + \ln\bigl(\bar\sigma_t^2(\theta) + \varepsilon\bigr)$$

and

$$\theta^{(\beta)}_{n,h} := \arg\min_{\theta\in\Theta} L_{n,h;\beta}(\theta).$$

Then, by similar arguments as before, and under suitable moment conditions,

$$n^{\frac{\beta}{2}}\big(\theta_{n,h}^{(\beta)} - \theta^0\big) \xrightarrow{d} N\big(0, H_h^{-1} G_h H_h^{-1}\big)$$

provided that $0 < \beta < 1 - 2d$. This means that the asymptotic normal distribution is the same as for $\hat{\theta}_{n,h}$; however, the rate of convergence is much slower than $n^{-\frac{1}{2}}$. For the "best" rate of $n^{d-\frac{1}{2}}$, one can at least show $E[|\theta_{n,h}^{(\beta)} - \theta^0|] \sim c_2 n^{-(\frac{1}{2}-d)}$, but it seems more difficult to derive the asymptotic distribution. The problem with a slower rate becomes worse if the long memory becomes stronger because β cannot exceed $1 - 2d$. For instance, for $d = 0.1$ we have $n^{\frac{1}{2}-d} = n^{-0.4}$ whereas for $d = 0.4$ the rate of convergence is $n^{-0.1}$ only. This makes a huge difference even for moderate sample sizes. For instance, for $n = 1000$, $n^{-0.1}/n^{-0.4} \approx 7.9$.

Although the explicit proofs in Beran and Schützner (2009) are written down for the specific case $b_j = cj^{d-1}$ ($\theta = (a, c, d)^T$) the generalization to general weights with $b_j \sim cj^{d-1}$ follows directly. A natural starting point is for instance given by coefficients defined by the fractional differencing operator, i.e. coefficients in the series (in $z \in \mathbb{C}$)

$$\sum_{j=1}^{\infty} b_j z^j = c(d)\big[(1-z)^{-d} - 1\big]$$

where

$$c^2(d) \leq \left[\sum_{j=1}^{\infty} \binom{-d}{j}^2\right]^{-1}$$

(to ensure stationarity, see Sect. 2.1.3.6). This can easily be extended by multiplying the $\sum_{j=1}^{\infty} b_j z^j$ by a function $\psi(z)/\varphi(z)$ corresponding to an ARMA filter and adjusting the constant to satisfy the stationarity condition $\sum b_j^2 < 1$.

6.3 Statistical Inference for ARCH(∞) Processes

In this section, we briefly mention the existing theory for ARCH(∞) models. Location estimation mimics the results for SV and LARCH models; however, there are no available theorems for M-estimators. As for parametric estimation of dependence parameters, we note that the maximum likelihood estimation is much easier than in the LARCH(∞) case (Berkes and Horváth 2004). Furthermore, the MLE seems to be the most suitable approach. The Whittle estimator applied to squared sequences is no longer an approximation of the MLE and is indeed less efficient than the actual MLE (Giraitis and Robinson 2001).

6.3.1 Location Estimation

As in Sect. 6.2.1, we consider a time series $Y_t = \mu + X_t$; however, now the residuals X_t are generated by an ARCH(∞) process

$$X_t = \xi_t \sigma_t, \tag{6.12}$$

$$\sigma_t^2 = b_0 + \sum_{j=1}^{\infty} b_j X_{t-j}^2. \tag{6.13}$$

The random variables ξ_t are such that $E(\xi_t) = 0$ and $\sigma_\xi^2 = E(\xi_t^2) = 1$. Furthermore, $b_0 > 0$, $b_j \geq 0$ and $\sum b_j < 1$ (see Sect. 4.2.7). Then the central limit theorem holds for $S_n = \sum_{t=1}^n X_t$ (see Corollary 4.4) so that

$$\sqrt{n}(\bar{y} - \mu) \underset{d}{\to} N\left(0, \sigma_X^2\right)$$

with

$$\sigma_X^2 = \frac{b_0}{1 - \sum_{j=1}^{\infty} b_j}.$$

Thus, an approximate $(1-\alpha)$-confidence interval for μ can be given by

$$\bar{x} \pm z_{1-\frac{\alpha}{2}} \frac{\sigma_X}{\sqrt{n}}.$$

Since $\operatorname{var}(Y_1) = \operatorname{var}(X_1)$, the parameter σ_X can be estimated based on the observed data $Y_1, \ldots, Y_n$.

6.3.2 Estimation of Dependence Parameters

Consider a parametric ARCH(∞) process with $\mu = 0$ and coefficients $b_j = b_j(\theta^0)$ ($j \geq 0$) depending on a finite dimensional parameter vector $\theta^0 = (b_0^0, \vartheta^0)$. As in the LARCH case, quasi maximum likelihood estimation of θ^0 can be obtained by maximizing the Gaussian conditional log-likelihood function

$$-\frac{2}{n} L_n(\theta) = \frac{1}{n} \sum_{t=1}^{n} \frac{X_t^2}{\sigma_t^2(\theta)} + \log \sigma_t^2(\theta) \tag{6.14}$$

where $\sigma_t^2(\theta) = b_0 + \sum_{j=1}^{\infty} b_j X_{t-j}^2$. In contrast to LARCH processes, no problems with respect to summability and differentiability of $\sigma_t^2(\theta)$ occur because the coefficients b_j are absolutely summable. For the same reason, the approximation of σ_t^2 by the truncated sum $b_0 + \sum_{j=1}^{t-1} b_j X_{t-j}^2$ is accurate enough to be negligible asymptotically. Moreover, by definition, σ_t^2 is bounded away from zero by b_0. Asymptotic

normality of $\hat{\theta}_{\mathrm{MLE}} = \arg\max L_n$ is shown in Weiss (1986) for ARCH(p) processes, Lee and Hansen (1994) and Lumsdaine (1996) for the GARCH(1, 1) model and Hall and Yao (2003) for GARCH(p, q) models. Similar results are also given in Berkes et al. (2003), Berkes and Horváth (2004). For more general ARCH(∞) processes, including the case of hyperbolically decaying coefficients b_j, Robinson and Zaffaroni (2006) derived the consistency of $\hat{\theta}_{\mathrm{MLE}}$.

Results on the asymptotic distribution for general ARCH(∞) processes are known for an alternative estimator (Giraitis and Robinson 2001), namely the Whittle estimator $\hat{\theta}_{\mathrm{Whittle}}$ based on the squared observations X_t^2 (see also Bollerslev 1986 and Robinson and Zaffaroni 1997, 1998 for earlier uses of Whittle estimation in volatility models). The idea is to write X_t^2 in the autoregressive form

$$\begin{aligned} X_t^2 &= E\big[X_t^2 \mid \mathscr{F}_{t-1}\big] + X_t^2 - E\big[X_t^2 \mid \mathscr{F}_{t-1}\big] \\ &= \sigma_t^2 + X_t^2 - \sigma_t^2 = b_0 + \sum_{j=1}^{\infty} b_j X_{t-j}^2 + \zeta_t \end{aligned}$$

with $\zeta_t = X_t^2 - \sigma_t^2$ and $\mathscr{F}_t$ the σ-algebra generated by X_s ($s \le t$). The residual process is a martingale difference with variance $\sigma_\zeta^2 = \mathrm{var}(\zeta_t)$. Since the equation can also be written as

$$\begin{aligned} \tilde{X}_t^2 = b_0^{-1} X_t^2 &= 1 + \sum_{j=1}^{\infty} b_0^{-1} b_j X_{t-j}^2 + b_0^{-1}\zeta_t \\ &= 1 + \sum_{j=1}^{\infty} \tilde{b}_j \tilde{X}_{t-j}^2 + \tilde{\zeta}_t, \end{aligned}$$

we may assume without loss of generality that $b_0 = 1$. Under moment assumptions (in particular, fourth-order stationarity of X_t), X_t^2 then has the spectral density

$$f_{X^2}\big(\lambda;\theta^0\big) = \frac{\sigma_\zeta^2}{2\pi} g_{X^2}\big(\lambda;\theta^0\big) = \frac{\sigma_\zeta^2}{2\pi}\left|1 - \sum_{j=1}^{\infty} b_j e^{-j\lambda}\right|^{-2}.$$

The Whittle estimator $\hat{\theta}_{\mathrm{Whittle}}$ of θ^0 based on this spectral density is obtained by minimizing

$$\mathscr{L}_{n,\mathrm{Whittle}}(\theta) = \frac{2}{n}\sum_{j=1}^{[(n-1)/2]} \frac{I_{n,X^2}(\lambda_j)}{g_{X^2}(\lambda_j;\theta)}$$

with respect to θ, where I_{n,X^2} is the periodogram of the sequence X_t^2 evaluated at the Fourier frequencies $\lambda_j = 2\pi j/n$ (cf. (5.42)). It should be noted, however, that, in contrast to L_n, the function $\mathscr{L}_{n,\mathrm{Whittle}}$ is not associated with a likelihood. In particular, for the case of Gaussian innovations ξ_t, L_n essentially corresponds to a (conditional) log-likelihood function whereas this is not the case for $\mathscr{L}_{n,\mathrm{Whittle}}$.

The reason is simply that the process X_t^2 is not Gaussian. This implies that, for Gaussian ξ_t, $\hat{\theta}_{\text{Whittle}}$ is asymptotically less efficient than $\hat{\theta}_{\text{MLE}}$. Specifically, Giraitis and Robinson (2001) derive for general ARCH(∞) processes (and suitable moment conditions) the limit

$$\sqrt{n}\big(\hat{\theta}_{\text{Whittle}} - \theta^0\big) \underset{d}{\to} N\big(0, 2W^{-1} + W^{-1} V W^{-1}\big)$$

where

$$W = \frac{1}{2\pi} \int_{-\pi}^{\pi} \frac{\partial}{\partial\theta} \log g_{X^2}(\lambda) \left[\frac{\partial}{\partial\theta} \log g_{X^2}(\lambda) \right]^T d\lambda,$$

$$V = \frac{2\pi}{\sigma_\xi^2} \int_{-\pi}^{\pi} \frac{\partial}{\partial\theta} \frac{1}{g_{X^2}(\lambda_1)} \left[\frac{\partial}{\partial\theta} \frac{1}{g_{X^2}(\lambda_2)} \right]^T h(\lambda_1, -\lambda_2, \lambda_2)\, d\lambda_1\, d\lambda_2.$$

Here $h(\lambda_1, -\lambda_2, \lambda_2)$ denotes the fourth-order cumulant spectral density of X_t^2 defined by

$$h(\lambda_1, \lambda_2, \lambda_3) = \frac{1}{(2\pi)^3} \sum_{k_1,k_2,k_3=-\infty}^{\infty} \exp\big(-i(k_1\lambda_1 + k_2\lambda_2 + k_3\lambda_3)\big) c_{0,k_1,k_2,k_3}$$

where $c_{0,k_1,k_2,k_3} = cum(X_t^2, X_{t+k_1}^2, X_{t+k_2}^2, X_{t+k_3}^2)$ is the joint cumulant of the variables $Y_1 = X_t^2$, $Y_2 = X_{t+k_1}^2$, $Y_3 = X_{t+k_2}^2$, $Y_4 = X_{t+k_3}^2$. Recall that the cumulants $\kappa_{j_1,\ldots,j_m} = cum(Y_1^{j_1}, Y_2^{j_2}, \ldots)$ of a random vector $Y \in \mathbb{R}^m$ are the coefficients in the series expansion of the cumulant generating function

$$\kappa(u) = \log E\big[\exp\big(iu^T Y\big)\big] = \sum_{j_1,\ldots,j_m=0}^{\infty} \kappa_{j_1,\ldots,j_m} \frac{u_1^{j_1} \cdots u_m^{j_m}}{j_1! \cdots j_m!} i^{j_1+\cdots+j_m}.$$

For other estimators and a nice overview on estimation for ARCH(∞) processes, see, e.g. Giraitis et al. (2006).