On the excess of average squared error for data-driven bandwidths in nonparametric trend estimation

Sur l'excès de la moyenne quadratique des erreurs associées à des fenêtres adaptatives dans l'estimation non-paramétrique de la tendance

Karim Benhenni[1], Didier A. Girard[2], and Sana Louhichi[1,*]

[1]Laboratoire Jean Kuntzmann, Université Grenoble Alpes, 700 Avenue Centrale, 38401 Saint-Martin-d'Hères, France
karim.benhenni@univ-grenoble-alpes.fr, sana.louhichi@univ-grenoble-alpes.fr
[2]CNRS, Laboratoire Jean Kuntzmann
didier.girard@univ-grenoble-alpes.fr

ABSTRACT. We consider the problem of the optimal selection of the smoothing parameter h in kernel estimation of a trend in nonparametric regression models with strictly stationary errors. We suppose that the errors are stochastic volatility sequences. Three types of volatility sequences are studied: the log-normal volatility, the Gamma volatility and the log-linear volatility with Bernoulli innovations. We take the weighted average squared error (ASE) as the global measure of performance of the trend estimation using h and we study two classical criteria for selecting h from the data, namely the adjusted generalized cross validation and Mallows-type criteria. We establish the asymptotic distribution of the gap between the ASE evaluated at one of these selectors and the smallest possible ASE. A Monte-Carlo simulation for a log-normal stochastic volatility model illustrates that this asymptotic approximation can be accurate even for small sample sizes.

2010 Mathematics Subject Classification: 62G08. 62G20. 60G10.
KEYWORDS. Nonparametric trend estimation, Kernel nonparametric models, Smoothing parameter selection, Average squared error, Excess of average squared error, Mean average squared error, Mallows criterion, Cross validation, Generalized cross validation, SV models.

1. Introduction

Nonparametric trend estimation is a very popular field of research in Statistics and is used in different domains of application. There are several nonparametric estimates of the trend in time series models or the mean function of stochastic processes. Many of these estimates are constructed from a kernel function that depends on a smoothing parameter h known, in nonparametric statistical literature, as a bandwidth. The choice of this parameter is crucial since it has an important impact on the performance of the kernel estimates. Some criteria are available for choosing this parameter but there are mostly based on models with independent errors; the plug-in method, see for instance Ruppert, Sheather and Wand (1995) and Fan, Gijbels, Hu and Huang (1996), the cross-validation (CV) and the generalized cross-validation (GCV), see for instance Rice (1984), Härdle, Hall and Marron (1988) and Girard (1998) among others. However, in the case of dependent errors, there are very limited available results in the literature concerning the selection methods of the smoothing parameter, see for instance the review by Opsomer, Wang and Yang (2001). Hall, Lahiri and Polzehl (1995) develop bootstrap and cross-validation methods to select the smoothing parameter under short and long range dependance.

In Benhenni, Girard and Louhichi (2021), we considered dependent strictly stationary martingale difference errors with an application to ARCH(1) errors. Based on four criteria defining, via minimization,

[*] Corresponding author.

two optimal and two data-driven smoothing parameters (or "two selectors"), we showed that the minimizers of these four criteria are "first-order equivalent" in probability (using now a classical terminology as specified in equation (9) below). Moreover, we gave a normal asymptotic behavior of the difference between, in particular, the minimizer of the average squared error (ASE, in short, defined by equation (3) below) and that of the Mallows-type criterion. In this paper, we are mainly interested in studying the excess of the ASE for such selectors. Let us recall that the excess of the ASE ("ASE-excess" in short) associated with any particular selector, is defined as the increase of ASE when one compares it with its lowest possible value when varying h for the available data. Early studies of the ASE-excess, Hall and Marron (1987), and Mammen (1990), put an emphasis on the fact that the asymptotic behavior of this excess is interpretable, more than the asymptotic behavior of the difference between selectors.

In this paper we provide some extensions to the theoretical results of Benhenni, Girard and Louhichi (2021) (abbreviated as BGL henceforth) in two directions. Firstly, we show, under the same assumptions as in Theorem 3.1 of BGL, that the asymptotic behavior of the ASE-excess which has been established by Härdle, Hall and Marron (1988) for independent and identically distributed (i.i.d., in short) errors, still holds. Secondly, we show that the required assumptions for the sequence of errors are fulfilled by three common time series volatility models. Volatility has been one of the most active areas of research in time series econometrics and economic forecasting. It may be modeled as an unobserved component following some latent stochastic process, such as autoregression. The resulting models are called stochastic volatility (SV) models and have been the focus of considerable attention, see for instance Taylor (1994), Ghysels, Harvey and Renault (1996), Shephard (1996) and Billo and Sartore (2005). Here, we concentrate on three stochastic volatility models (SV): the log-normal volatility, the Gamma volatility and the log-linear volatility with Bernoulli innovations. The first two SV models satisfy a strong mixing condition with a decreasing power bound whereas the Bernoulli SV model does not satisfy any mixing condition.

In addition, an extensive Monte-Carlo simulation study for the log-normal stochastic volatility model with various model parameters is provided in section 4 and demonstrates that the asymptotic behavior established here can be quite realistic for a moderate number of observations, n, over a large range of parameters.

The paper is organized as follows. Section 2 introduces the nonparametric model, defines the different criteria for the selection of the smoothing parameter h, explains the main assumptions and recalls previous results that are required for the present study. Section 3 states the main results of this paper. Theorem 3.1 studies the asymptotic distribution of the ASE-excess for optimal bandwidths. Proposition 3.1 is an application of Theorem 3.1 to the SV models. Corollaries 3.1, 3.2 and 3.3 study, respectively, the log-normal SV, the Gamma SV and the log-linear SV with Bernoulli innovations. Section 4 is the Monte Carlo simulation study mentioned above. Finally, Section 5 is devoted to the proofs.

2. Model, selection criteria and useful tools

Let $(\epsilon_i)_{i\geq 0}$ be a strictly stationary sequence of centered random variables with finite second moment. Let $\sigma^2 = \text{Var}(\epsilon_1)$ and R be the correlation matrix of the vector $(\epsilon_1, \cdots, \epsilon_n)^t$ (where v^t denotes the transpose of the vector v). Consider the following regression model, defined for $i = 1, \cdots, n$, by

$$Y_i = r(x_i) + \epsilon_i, \quad x_i = \frac{i}{n}, \tag{1}$$

where r is an unknown regression function with second order continuous derivative and the x_i's are equally spaced fixed design. We are interested, here, in the Priestley-Chao estimator of r defined, for $x \in \mathbb{R}$, by

$$\hat{r}(x) = \sum_{i=1}^{n} l_i(x) Y_i, \text{ with } l_i(x) = \frac{1}{nh} K\left(\frac{x - x_i}{h}\right),$$

where K is a compactly supported, even kernel, with class $\mathcal{C}^2([-1, 1])$ and h is a positive bandwidth smaller than 0.5. The above curve estimator can also be written in the following matrix form:

$$\hat{r} = LY, \tag{2}$$

with

$$\hat{r} = (\hat{r}(x_1), \cdots, \hat{r}(x_n))^t, \;\; Y = (Y_1, \cdots, Y_n)^t$$

and $L = (l_j(x_i))_{1 \leq i,j \leq n}$ is known as the smoothing matrix or the hat matrix. Since the estimator $\hat{r}$ depends on a smoothing parameter h, some criteria are needed for choosing h. One first criterion is the ASE, $T_n(h)$, defined by

$$T_n(h) = \frac{1}{n} \sum_{i=1}^{n} u(x_i)(\hat{r}(x_i) - r(x_i))^2, \tag{3}$$

where $u = u_\epsilon$ (for a fixed positive ϵ less than 0.5) is a known positive function of class $\mathcal{C}^1$ and $[\epsilon, 1 - \epsilon]$-compactly supported. This function is introduced in order to eliminate the boundary effects of the compactly supported kernel K. We know from Lemma 2.1, in BGL, that if $\sum_{k=1}^{\infty} k |\mathrm{Cov}(\epsilon_0, \epsilon_k)| < \infty$ then the mean weighted average squared error, $\mathbb{E}(T_n(h))$, is very close to the quantity $D_n(h)$, defined by

$$\begin{aligned} D_n(h) &= \frac{h^4}{4} \int_0^1 u(x) r''^2(x) dx \left(\int_{-1}^{1} t^2 K(t) dt \right)^2 \\ &+ \frac{1}{nh} (\int_0^1 u(x) dx) \int_{-1}^{1} K^2(y) dy \left(\sigma^2 + 2 \sum_{k=1}^{\infty} \mathrm{Cov}(\epsilon_0, \epsilon_k) \right). \end{aligned} \tag{4}$$

Let $h_n^* \in \mathrm{argmin}_{h>0} D_n(h)$. We suppose in the sequel that $\int_0^1 u(x) r''^2(x) dx \neq 0$. Then clearly,

$$h_n^* = c n^{-1/5}, \text{ with } c = \left(\frac{(\int_0^1 u(x)dx) \int_{-1}^{1} K^2(y) dy \left(\sigma^2 + 2 \sum_{k=1}^{\infty} \mathrm{Cov}(\epsilon_0, \epsilon_k)\right)}{\int_0^1 u(x) r''^2(x) dx (\int_{-1}^{1} t^2 K(t) dt)^2} \right)^{1/5}. \tag{5}$$

Let H_n be a neighborhood of h_n^*, i.e, $H_n = [a n^{-1/5}, b n^{-1/5}]$ for some fixed $a < c < b$. Define also,

$$\hat{h}_n \in \mathrm{argmin}_{h \in H_n} T_n(h).$$

Both those two criteria ($D_n(h)$ and $T_n(h)$) depend, in particular, on the unknown function r, so they cannot, of course, directly provide a data-driven selection of bandwidth. A well-known criterion, which is an unbiased estimator (up to an additive constant) of the mean weighted average squared error $\mathbb{E}(T_n(h))$, is the following variant of Mallows' criterion introduced for other purposes than ours and adapted to dependent errors with known covariance matrix $\sigma^2 R$:

$$\mathrm{CL}(h) = n^{-1} \| U^{1/2}(I - L)Y \|^2 + 2\sigma^2 n^{-1} tr(URL), \tag{6}$$

where I is the identical matrix, U is the diagonal matrix defined by $diag(u(x_1), \cdots, u(x_n))$, $Y = (Y_1, \cdots, Y_n)^t$, $L = (l_j(x_i))_{1 \leq i,j \leq n}$ as defined by (2), and $tr(M)$ denotes the trace of a square matrix M. We consider, according to our purpose, $\hat{h}_M$ to be the minimizer of this dependent version of the Mallows-type criterion:

$$\hat{h}_M \in \mathrm{argmin}_{h \in H_n} \mathrm{CL}(h).$$

Finally, let $G_{\mathrm{X}}(h)$ be the classical GCV-type criterion and $\hat{h}_G$ be its minimizer:

$$\hat{h}_G \in \mathrm{argmin}_{h \in H_n} G_{\mathrm{X}}(h) \;\; \text{with} \;\; G_{\mathrm{X}}(h) = n^{-1} \| U^{1/2}(I - L)Y \|^2 \times \Xi_{\mathrm{X}} \left(\frac{tr(UL)}{tr(U)} \right), \tag{7}$$

where Ξ_{X} satisfies, for small values of $|t|$,

$$\Xi_{\mathrm{X}}(t) = 1 + 2t + O(t^2) \text{ with second derivative } \Xi''_{\mathrm{X}} \text{ bounded on a neighborhood of } 0. \tag{8}$$

This GCV-type criterion G_{X} is well adapted to strictly stationary uncorrelated dependent errors.

The bandwidth h_n^* has the advantage of being explicitly derived, but its drawback is that it cannot be implemented in practice since it depends on unknown quantities such as σ^2 and r''. The bandwidth $\hat{h}_G$ has no explicit expression but on the other hand it can be implemented in practice since it is obtained by minimizing an observable criterion. Because $\hat{h}_n$ is recognized as the ideal, albeit unobservable bandwidth, then a comparison of $\hat{h}_G$ (or $\hat{h}_M$) with $\hat{h}_n$ has been an initial subject of interest for many authors.

Since we are interested in this study to move out of the basic framework where the errors $(\epsilon_i)_{1 \leq i \leq n}$ form a sequence of i.i.d. random variables, a dependency condition should be assumed.

We consider here the rather general condition of “martingale difference sequences” (MDS, see Conditions (A) below) since, on the one hand, there are many models satisfying this notion (see the next section) and since, on the other hand, studying this framework of dependency makes it possible to study the most general case of strictly stationary random variables as is done by Peligrad, Utev and Wu (2007). Note also, that MDS are uncorrelated dependent random variables, so that the above GCV-type criterion, G_{X}, is well adapted. We also need, for technical reasons, a higher moment condition on the distribution of ϵ_1. All these assumptions are summarized in the following Conditions (A).

Conditions (A). Assume that the errors $(\epsilon_i)_{i \geq 0}$ form a strictly stationary MDS with respect to some natural filtration $(\mathcal{F}_i)_{i \geq 1}$, i.e, for any $i > 0$, ϵ_i is $\mathcal{F}_i$-measurable and $\mathbb{E}(\epsilon_i | \mathcal{F}_{i-1}) = 0$ almost surely (a.s., in short). Suppose also that $\mathbb{E}(\epsilon_1^{2p}) < \infty$ for some $p > 8$.

Under Conditions (A), the bandwidths $h_n^*, \hat{h}_n, \hat{h}_M$ and $\hat{h}_G$ are first-order equivalent in probability (and then both the CL criterion and the GCV-type criterion enjoy the "asymptotic optimality" property), this means that

$$\frac{\hat{h}_n}{h_n^*}, \frac{\hat{h}_M}{h_n^*}, \frac{\hat{h}_G}{h_n^*} \tag{9}$$

all converge in probability to 1 as n tends to infinity (see Proposition 3.1 in BGL).

Theorem 3.1 in BGL gives the rate at which $\hat{h}_M - \hat{h}_n$ and $\hat{h}_G - \hat{h}_n$ converge in distribution to a centered normal law, and it needs an additional dependence condition, Condition (B) below,

Condition (B). There exists a positive decreasing function Φ defined on $\mathbb{R}^+$ satisfying

$$\sum_{s=1}^{\infty} s^4 \Phi(s) < \infty,$$

and for any positive integer $q \leq 6$, $1 \leq i_1 \leq \cdots \leq i_k < i_{k+1} \leq \cdots \leq i_q \leq n$ such that $i_{k+1} - i_k \geq \max_{1 \leq l \leq q-1}(i_{l+1} - i_l)$,

$$|\mathrm{Cov}(\epsilon_{i_1} \cdots \epsilon_{i_k}, \epsilon_{i_{k+1}} \cdots \epsilon_{i_q})| \leq \Phi(i_{k+1} - i_k), \tag{10}$$

where $\epsilon_{i_1} \cdots \epsilon_{i_k}$ denotes the product $\prod_{\ell=1}^{k} \epsilon_{i_\ell}$ (and likewise for $\epsilon_{i_{k+1}} \cdots \epsilon_{i_q}$).

Condition (B) is known in the literature (see Doukhan and Louhichi (1999)). It allows, in particular, to control the higher order moments of partial sums of dependent random sequences.

Now, if both Conditions (A) and (B) are satisfied, then it holds that, by Theorem 3.1 in BGL,

$$n^{3/10}(\hat{h}_M - \hat{h}_n) \text{ and } n^{3/10}(\hat{h}_G - \hat{h}_n) \tag{11}$$

both converge in distribution to a centered normal law with variance Σ^2 given by

$$\Sigma^2 = \frac{4\sigma^{6/5}}{5^2 A^{8/5} B^{2/5}} (\int t^2 K(t) dt)^2 \int_0^1 u^2(x) r''^2(x) dx$$
$$+ \frac{8\sigma^{6/5}}{5^2 A^{3/5} B^{7/5}} \int_0^1 u^2(x) dx \int (K - G)^2(u) du, \tag{12}$$

where $\sigma^2 = \mathbb{E}(\epsilon_1^2)$, G is the function defined for any $x \in \mathbb{R}$ by $G(x) = -xK'(x)$ and

$$A = \int_0^1 u(x) r''^2(x) dx \left(\int t^2 K(t) dt \right)^2, \quad B = \int_0^1 u(x) dx \int K^2(t) dt.$$

3. Main results with applications to stochastic volatility models

We have thus studied sufficient conditions on the sequence of errors under which these four bandwidths ($\hat{h}_M$, $\hat{h}_n$, $\hat{h}_G$ and h_n^*) are asymptotically equivalent and their differences have asymptotic normal distributions. Now, let us recall that selecting the bandwidth is only a means to an end: minimising the error in the estimation of r. Since the measure of error that we consider is the ASE, as it is often the case in supervised learning, it is now important to compare these selectors via their corresponding ASEs rather than comparing them directly with each other. Our purpose is thus to study the asymptotic behaviors of the gaps, called the ASE-excesses, $T_n(\hat{h}_M) - T_n(\hat{h}_n)$ and $T_n(\hat{h}_G) - T_n(\hat{h}_n)$. That is, we are interested in the study of the deviations of the error $T_n(\hat{h}_M)$ (respectively $T_n(\hat{h}_G)$) from the minimal possible value of the ASE, which is $T_n(\hat{h}_n)$.

The following theorem gives, under Conditions (A) and (B), the rate at which these gaps tend to 0, extending the results stated by Härdle, Hall and Marron (1988) for i.i.d. errors.

Theorem 3.1. *Suppose that Conditions (A) and (B) are satisfied. Then both*

$$n(T_n(\hat{h}_M) - T_n(\hat{h}_n)) \text{ and } n(T_n(\hat{h}_G) - T_n(\hat{h}_n))$$

converge in distribution to a $C\mathcal{X}_2(1)$ law, where $\mathcal{X}_2(1)$ is the chi-square distribution with one degree of freedom and C is the positive constant given by,

$$C = \frac{2\sigma^2}{5A}\left(\left(\int t^2K(t)dt\right)^2 \int_0^1 u^2(x)r''^2(x)dx + \frac{2A}{B}\int_0^1 u^2(x)dx \int (K-G)^2(t)dt\right),$$

where $\sigma^2 = \mathbb{E}(\epsilon_1^2)$, G, A and B are defined as in (12).

Remark. As expected, the constant C in this asymptotic distribution is the same as in the i.i.d. case established by Härdle, Hall and Marron (1988). If we choose the weight function u to be proportional to u^2, then it is easy to check that C gets so simplified that it no longer depends on the function r. In this case and as an immediate consequence of Theorem 3.1, it is possible to construct asymptotic prediction intervals for the excess error using the quantiles of the chi-square distribution. See Section 4 for an example of such a simplified value of C, since, in this simulation study, u is trivially proportional to u^2.

We give, in the sequel, examples of strictly stationary MDS of errors satisfying the requirements of Theorem 3.1 (and also the asymptotic optimality (9) and the asymptotic normality (11)-(12). Both results (9) and (11)-(12) are proved in BGL but the class of examples below has not been considered there. More precisely, BGL only studies standard ARCH(1) sequences. Let us recall that an ARCH(1) sequence satisfies the Conditions (A) and (B) provided its so-called "persistence" parameter stays lower than $2025027^{-1/8} \approx 0.162796$). The examples we consider here belong to the class of stochastic volatility processes. Recall that a Stochastic Volatility process $(\epsilon_i)_{i\in\mathbb{N}}$, SV in short, is defined as

$$\epsilon_i = \sigma_i Z_i, \quad i \in \mathbb{N}, \tag{13}$$

where the volatility sequence $(\sigma_i)_{i\in\mathbb{N}}$ is a strictly stationary sequence of positive random variables which is independent of the i.i.d. centered noise sequence $(Z_i)_{i\in\mathbb{N}}$. We refer, for instance, to Davis and Mikosh (2009) for the main properties of SV models.

The following proposition gives conditions under which the requirements of Theorem 3.1 are satisfied for SV error processes.

Proposition 3.1. *Let $(\epsilon_i)_{i\in\mathbb{N}}$ be as defined in (13). Suppose that, there exists a positive decreasing function $\tilde{\Phi}$ defined on $\mathbb{R}^+$ satisfying*

$$\sum_{s=1}^{\infty} s^4\tilde{\Phi}(s) < \infty,$$

and for any positive integer $q \le 6$, $1 \le i_1 \le \cdots \le i_k < i_{k+1} \le \cdots \le i_q \le n$ such that $i_{k+1} - i_k \ge \max_{1\le l\le q-1}(i_{l+1} - i_l)$,

$$|\mathrm{Cov}(\sigma_{i_1}\cdots\sigma_{i_k}, \sigma_{i_{k+1}}\cdots\sigma_{i_q})| \le \tilde{\Phi}(i_{k+1} - i_k), \text{ and moreover } |\mathbb{E}(Z_{i_1}\cdots Z_{i_q})| < \infty. \tag{14}$$

If $\mathbb{E}(\epsilon_1^{2p}) < \infty$ for some $p > 8$ then the asymptotic optimality (9), the asymptotic normality (11)-(12) and the conclusions of Theorem 3.1 hold.

The following corollaries give more explicit examples of SV models satisfying the assumptions of Proposition 3.1, so that Theorem 3.1 holds true in these cases.

3.1. Log-normal volatility sequences

The log-normal SV models are due to Taylor (1986). For these models, the volatility sequence $(\sigma_i)_{i\in\mathbb{N}}$ is an exponential weight of a Gaussian moving average. They are a basic alternative to ARCH-type processes, since unlike ARCH-type models, their variances always remain positive without the need of further conditions.

Corollary 3.1 below proves in particular that for log-normal SV models the volatility sequence $(\sigma_i)_{i\in\mathbb{N}}$ is a strictly stationary strong mixing sequence in the sense of Rosenblatt (1956). Recall that $(\sigma_i)_{i\in\mathbb{N}}$ is a strongly mixing sequence if its strong mixing coefficient $(\alpha_s)_{s\geq 0}$ defined by

$$\alpha_s = \sup_{k\in\mathbb{N}} \alpha(\sigma(\sigma_i,\ i\leq k), \sigma(\sigma_i,\ i\geq k+s)),$$

tends to 0 as s tends to infinity, where for two sigma-fields $\mathcal{A}$ and $\mathcal{B}$,

$$\alpha(\mathcal{A},\mathcal{B}) = \sup_{A\in\mathcal{A}, B\in\mathcal{B}} |\mathrm{Cov}(\mathbb{I}_A, \mathbb{I}_B)|.$$

Corollary 3.1. *Suppose that the volatility sequence $(\sigma_i)_{i\in\mathbb{N}}$ is defined for $i \in \mathbb{N}$, by $\sigma_i = \beta \exp\left(\sum_{j=0}^{\infty} \gamma^j \eta_{i-j}\right)$ with $|\gamma| < 1$, $\beta > 0$ and $(\eta_i)_{i\in\mathbb{Z}}$ is an i.i.d. centered sequence distributed as a Gaussian law with finite variance. Suppose also that Z_1 follows a standard Gaussian law. Then the process $(\epsilon_i)_{i\in\mathbb{N}}$, as defined in (13), is a strictly stationary MDS, with finite all integer moments, strongly mixing with $\alpha_s = O(|\gamma|^{\frac{2}{3}s})$, and the asymptotic optimality (9), the asymptotic normality (11)-(12) together with the conclusions of Theorem 3.1 hold.*

We illustrate the conclusions of Corollary 3.1 in Section 4.

3.2. Gamma volatility sequences

We consider here the Gamma stochastic model as defined by (13), with $(Z_i)_{i\in\mathbb{N}}$ being a sequence of i.i.d. standard normal random variables and for $i \in \mathbb{N}$, $\sigma_i = \sqrt{h_i}$ where $(h_i)_{i\in\mathbb{N}}$ is a positive time-homogeneous strictly stationary Markov chain. We suppose that the marginal distribution of $(h_i)_{i\in\mathbb{N}}$ is a Gamma $\Gamma(p,\lambda)$ distribution, i.e., denoting by π the invariant measure of this Markov chain,

$$\pi(dx) = f(x)dx,\ \ f(x) = \frac{\lambda^p}{\Gamma(p)} x^{p-1} e^{-\lambda x} \mathbb{I}_{x\geq 0},\ p,\lambda > 0,$$

where $\Gamma(p) = \int_0^\infty u^{p-1} e^{-u} du$. Suppose that this Markov chain is geometrically ergodic. This means that there exists a positive constant c and a Borel positive function $a(\cdot)$ such that the following inequality holds for any π-almost everywhere $x \in \mathbb{R}$: for any $n \in \mathbb{N}\setminus\{0\}$, and Borel set B

$$|P^n(x,B) - \pi(B)| \leq a(x)e^{-cn}, \tag{15}$$

recall that the transition probability P is defined, for suitable set A and x, by

$$P(x, A) = \mathbb{P}\left(h_1 \in A | h_0 = x\right)$$

and for $n \in \mathbb{N} \setminus \{0\}$

$$P^n(x, A) = \mathbb{P}\left(h_n \in A | h_0 = x\right).$$

In this case, it is well-known that the Markov chain $(h_i)_{i\in\mathbb{N}}$ is β-mixing with geometrically decaying mixing coefficients $(\beta_n)_{n\geq 1}$ (cf. for instance Theorem 3.7 in Bradley (2005) and the references therein). Recall that, for a sequence $(X_n)_{n\in\mathbb{N}}$, the β-mixing coefficients $(\beta_n)_{n\geq 1}$ are defined by (see for instance Doukhan (1994) (Sec 1.1))

$$\beta_n = \sup_{m\in\mathbb{N}} \mathbb{E}\left(\sup_{B\in\sigma(X_i,\, i\geq m+n)} |\mathbb{P}(B|\sigma(X_0, \cdots, X_m)) - \mathbb{P}(B)| \right).$$

The Gamma AR(1) process, stated in the example below, is a Markov chain satisfying all the above assumptions.

Example. Let h_0 be distributed as $\Gamma(p, \lambda)$ distribution. Define, for $\rho \in]0, 1[$, h_n recursively by,

$$h_n = \rho h_{n-1} + \xi_n,$$

where $(\xi_n)_n$ is an i.i.d. sequence of random variables with characteristic function $\mathbb{E}\left(e^{-it\xi_1}\right) = \left(\frac{\lambda - it}{\lambda - it\rho}\right)^{-p}$. The process $(h_n)_{n\geq 1}$ is then a strictly stationary Markov chain with Gamma $\Gamma(p, \lambda)$ univariate marginal distribution (see for instance Gaver and Lewis (1980)). This Markov chain is also geometrically ergodic in the sense of (15) (see for instance Kesten (1974)).

The following corollary gives conditions under which the Gamma stochastic volatility models satisfy the requirements of Proposition 3.1 and Theorem 3.1.

Corollary 3.2. *Suppose that the volatility sequence $(\sigma_i)_{i\in\mathbb{N}}$ is defined for $i \in \mathbb{N}$, by $\sigma_i = \sqrt{h_i}$ where $(h_i)_{i\in\mathbb{N}}$ is a positive time-homogeneous strictly stationary and geometrically ergodic Markov chain with marginal Gamma $\Gamma(p, \lambda)$ distribution. Let $(Z_i)_{i\in\mathbb{N}}$ be a sequence of i.i.d. standard normal random variables. Then the process $(\epsilon_i)_{i\in\mathbb{N}}$, as defined by (13), is a strictly stationary MDS, β-mixing with $\beta_n = O(e^{-\rho n})$, for some $\rho > 0$, and with finite all integer moments. The asymptotic optimality (9), the asymptotic normality (11)-(12) and the conclusions of Theorem 3.1 hold.*

3.3. Log-linear volatility sequences with Bernoulli innovations

The following corollary studies another class of SV models, $(\epsilon_i)_{i\in\mathbb{N}}$ as introduced in (13), for which the conclusions of Theorem 3.1 still hold. In these models, we suppose that $(\log(\sigma_i))_{i\in\mathbb{N}}$ is a linear process with Bernoulli innovation having coefficients $(2^{-k})_{k\in\mathbb{N}}$. Unlike the log-normal SV, in this case the volatility sequence $(\sigma_i)_{i\in\mathbb{N}}$ is not strongly mixing (see Bradley (1986)) but it is associated in the sense of Esary, Proschan and Walkup (1967). Recall that a sequence $(\sigma_i)_{i\in\mathbb{N}}$ is said to be associated if for any non-decreasing and bounded functions f and g,

$$\mathrm{Cov}(f(\sigma_1, \cdots, \sigma_n), g(\sigma_1, \cdots, \sigma_n)) \geq 0. \tag{16}$$

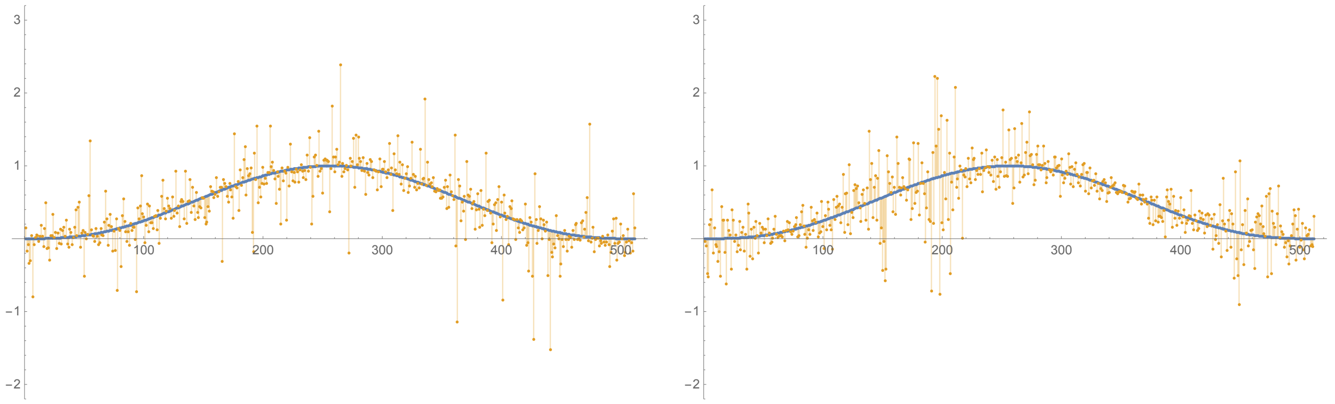

Figure 1. $n = 2^9$. Each of these 2 panels displays one data set Y and the "smooth" deterministic trend $r(\cdot)$ when the noise is a log-normal SV sequence. In the 2 panels $\tau = 0.75$, and they only differ by $\gamma = 0.01$ (left) and $\gamma = 0.98$ (right); see (17) for the definition of these two parameters.

Corollary 3.3. *Suppose that the volatility sequence $(\sigma_i)_{i\in\mathbb{N}}$ is defined for $i \in \mathbb{N}$, by $\sigma_i = \exp\left(\sum_{j=0}^{\infty} 2^{-j}\eta_{i-j}\right)$ where $(\eta_i)_{i\in\mathbb{Z}}$ is an i.i.d. centered sequence distributed as a Bernoulli $\mathcal{B}(1/2)$ distribution. Suppose also that Z_1 follows a standard normal law. Then the volatility sequence $(\sigma_i)_{i\in\mathbb{N}}$ is associated and the asymptotic optimality (9), the asymptotic normality (11)-(12) together with the conclusions of Theorem 3.1 hold.*

4. A Monte-carlo simulation study for a "trend plus a log-normal SV process"

The purpose of this section is to illustrate by simulations the asymptotic optimality (a.o, in short) as stated in (9), and to check whether the asymptotic normality (11)-(12) and the asymptotic scaled $\mathcal{X}_2(1)$ distribution of the ASE-excess, as stated in Theorem 3.1, are realistic (and thus useful) descriptions. Our experiments are similar to those in BGL, the main difference being that the noise sequence is now a common log-normal SV sequence, as analysed in Corollary 3.1, instead of an ARCH(1). Precisely, we consider the smooth function $r(x) = (4x(1-x))^3$ as "deterministic trend", an equispaced design and a noise level σ for which the noise-to-signal ratio is "moderate" (see Figure 1 for $n = 512$) and finally K is taken as the well-known bi-weight kernel. Notice that this trend offers a simple solution to the boundary effect issue: since it is "smoothly periodic", it allows us to consider a circular design, that is, as is explained in Härdle, Hall and Marron (1988), the estimation of r near $i = 1$ is done by setting, for $i \leq 0$, $y_i := y_{n-i}$ and similarly at the other end (such a "periodic padding" is used by many other authors). We are then allowed to take $u(\cdot)$ equals to the characteristic function on $[0, 1]$ as a weight function. Moreover, an important appealing consequence of this periodicity assumption, it that the main computation for any criterion discussed here can be reduced to fast Fourier transforms and this makes very affordable extensive large-scale simulations. See BGL for more details. To guard against possible local minimisers, the numerical minimisation of the ASE ($T_n(h)$) or of the Mallows criterion $\mathrm{CL}(h)$ consists of a global search over a dense enough grid (precisely, of size 401) of equispaced (on log-scale) values of h. We used Mathematica for these simulations; in fact the code is very similar to the one published in the demo Girard (2013) except for the noise generation, of course, since this demo is restricted to i.i.d. errors.

As parameters, in order to define the noise process, in addition to σ and γ (the serial correlation introduced in Corollary 3.1), we introduce

$$\tau := \sqrt{\mathrm{Var}(\eta_1)/(1-\gamma^2)}. \tag{17}$$

Let us remark that, to generate a noise process with variance σ^2, it is easy to check that the parameter β (used in Corollary 3.1) has to be set to $\sigma \exp\left(-\tau^2\right)$. The advantage of using (σ, τ^2, γ) instead of $(\beta, \mathrm{Var}(\eta_1), \gamma)$ is deduced from the fact that τ is the unique shape-parameter for the marginal density of the sequence of conditional variances σ_i^2's. Such a parameterization is common (see e.g. Taylor (1994)).

We consider three values for τ, and three values for the serial-correlation parameter γ, precisely

$$\tau \in \{0.2, 0.4, 0.75\}, \quad \gamma \in \{0.01, 0.9, 0.98\},$$

with a common value $\sigma = 0.32$. Note that the intermediate value 0.4 for τ is representative of values often obtained by fitting such a log-normal SV model to real econometric series; see Taylor (1994) (especially its Section 3.4, where τ is denoted by β) for an interesting review. Any large value of τ (say, greater than 1) implies a very fat tail for the marginal density of the amplitude of the noise $|\epsilon_i|$ which may cause a large instability of the classical kernel curve-estimate (a "robust" version kernel smoothing would be much more appropriate in such case). On the other hand, recall that a value very close to 0 for τ would imply that the density of the conditional variance σ_i^2 is concentrated around 1 and thus the serial correlation would have virtually no impact on the dependence between the ϵ_i's (which is then a "quasi-i.i.d.-normal" sequence). Thus we restrict the present study to the range $[0.2, 0.75]$ for τ.

The a.o. property. As is well-known, a result like (9) generally stems from a uniform relative accuracy result which states that $\mathrm{CL}(h) - n^{-1}\|U^{1/2}(Y-r)\|^2$ uniformly approximates $T_n(h)$ (or its expectation, say $\mathrm{MASE}(h)$) with a small (in probability and in sup norm over the domain of candidate h's) error, "small" being defined relatively to $\mathrm{MASE}(h)$.

We illustrate in Figure 2 that a uniform relative accuracy is well observed for all the considered values of τ and γ. Note that the results for $\gamma = 0.01$ and $\tau \in \{0.2, 0.4, 0.75\}$ have produced plots very similar to the plot for $\gamma = 0.9, \tau = 0.2$ (top-left panel in Figure 2) so they are not included in Figure 2. Only a slight deterioration is observed for $(\tau, \gamma) = (0.75, 0.98)$. Notice that this in contrast with the Monte Carlo study for ARCH(1) noise sequences in BGL where a large sample a.o. was no longer observed when the chosen persistence parameter (denoted, there, by α) was not small enough; indeed, even for $n = 2^{15}$ it was required that $\alpha < 0.75$ (precisely, a reasonable "uniform relative accuracy" was not observed for $\alpha \in \{0.75, 0.9, 0.98\}$; we refer to that paper for a discussion of this).

Asymptotic normal distribution. This Section aims to assess the usefulness of the theoretical asymptotic normal approximation stated in (11)-(12) for reasonable dataset sizes n. We are going to demonstrate that both τ and γ have an impact on the speed of convergence (with respect to n) toward this approximation.

Let us first consider $\tau = 0.2$. By inspecting Figure 3, we clearly see, in the three top panels, that the asymptotic approximation fits rather well already for $n = 2^9$ and for any $\gamma \in \{0.01, 0.9, 0.98\}$. For $n = 2^{15}$ the three bottom panels illustrate that the asymptotic theory provides a very accurate prediction of the finite sample "truth". Notice that, as expected the accuracy for $n = 2^{12}$ (not displayed here) is observed to be intermediary between the one for $n = 2^9$ and that for $n = 2^{15}$, and is thus also quite good.

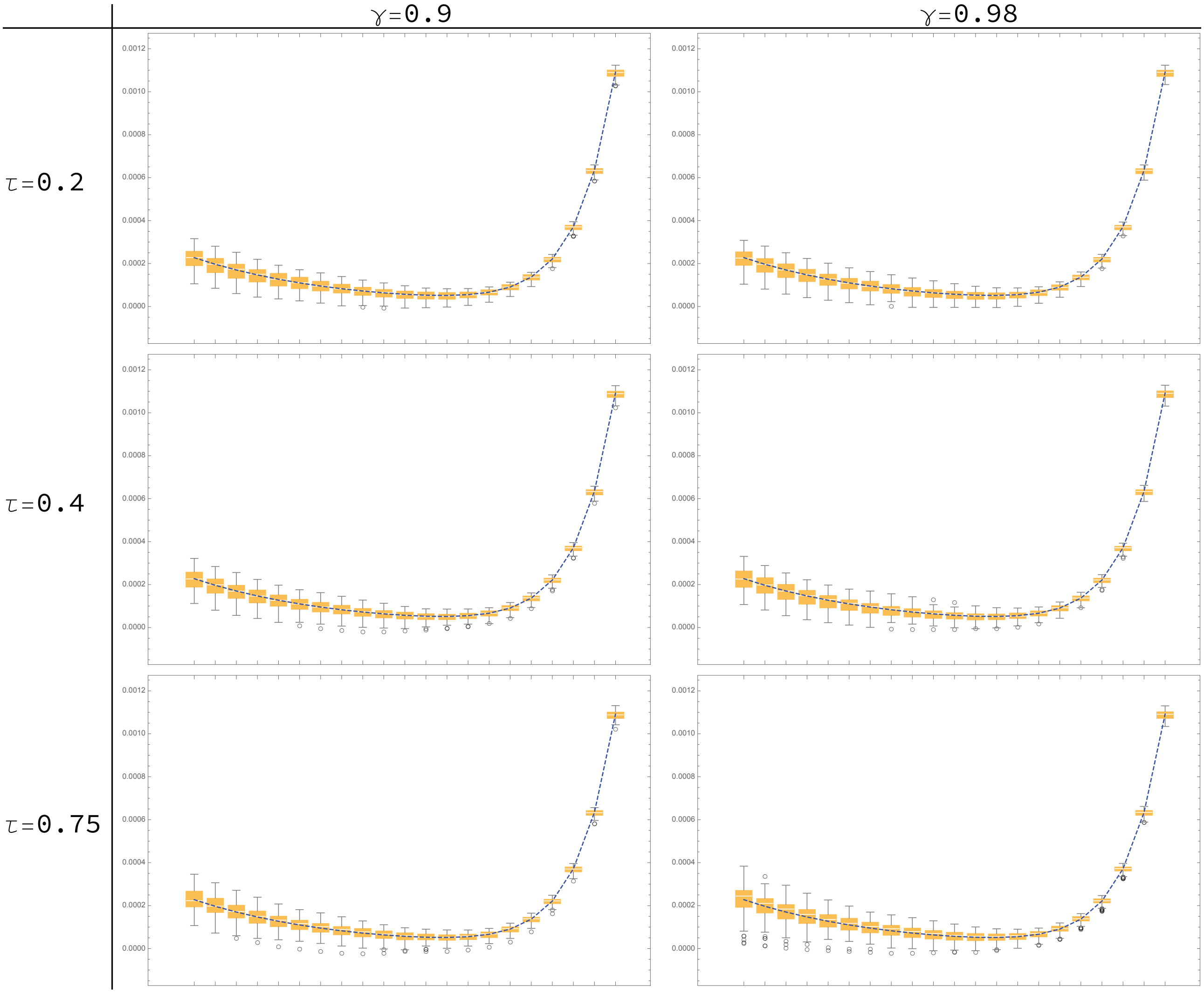

Figure 2. $n = 2^{15}$. These 6 panels only differ by (τ, γ) varying in $\{0.2, 0.4, 0.75\} \times \{0.9, 0.98\}$. In each panel, the dashed blue curve is the "empirical MASE", precisely the average (over the 3000 replicates) of the $T_n(h)$ curves. Each of the 21 boxplots (located at 21 fixed discrete values for h) are built from the first 100 replicates of $\mathrm{CL}(h) - n^{-1}\|U^{1/2}(Y - r)\|^2$.

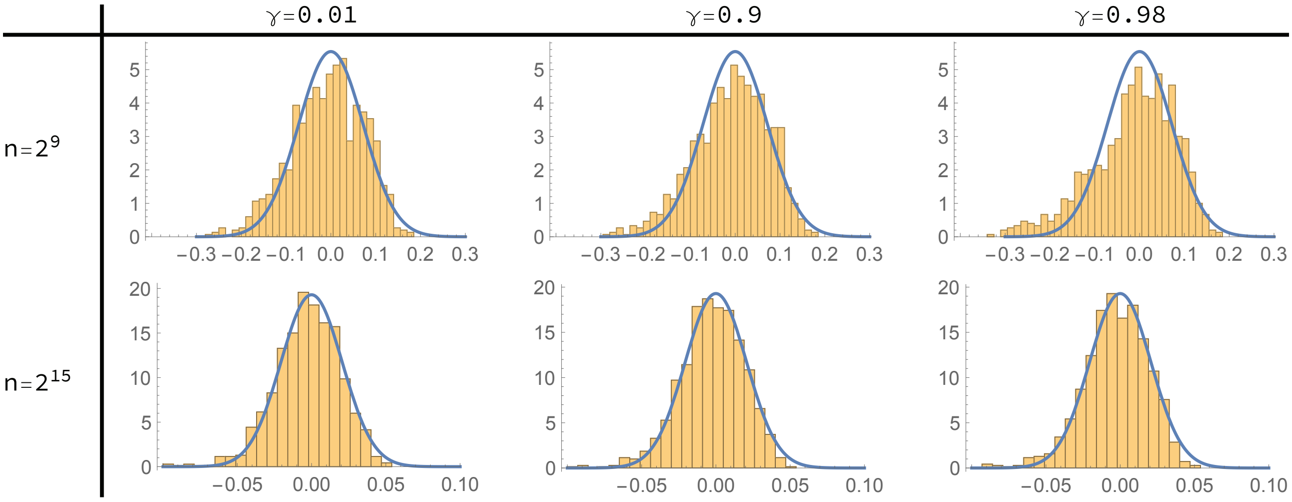

Figure 3. $\tau = 0.2$. . These 6 panels only differ by n ($= 2^9$ in the top row and 2^{15} in the bottom row) and by γ varying in $\{0.01, 0.9, 0.98\}$. In each panel, the displayed normalized histogram is that of the 3000 replicates of $\hat{h}_M - \hat{h}_n$. The superposed blue curve is the normal distribution of $\hat{h}_M - \hat{h}_n$ as predicted by the asymptotic theory.

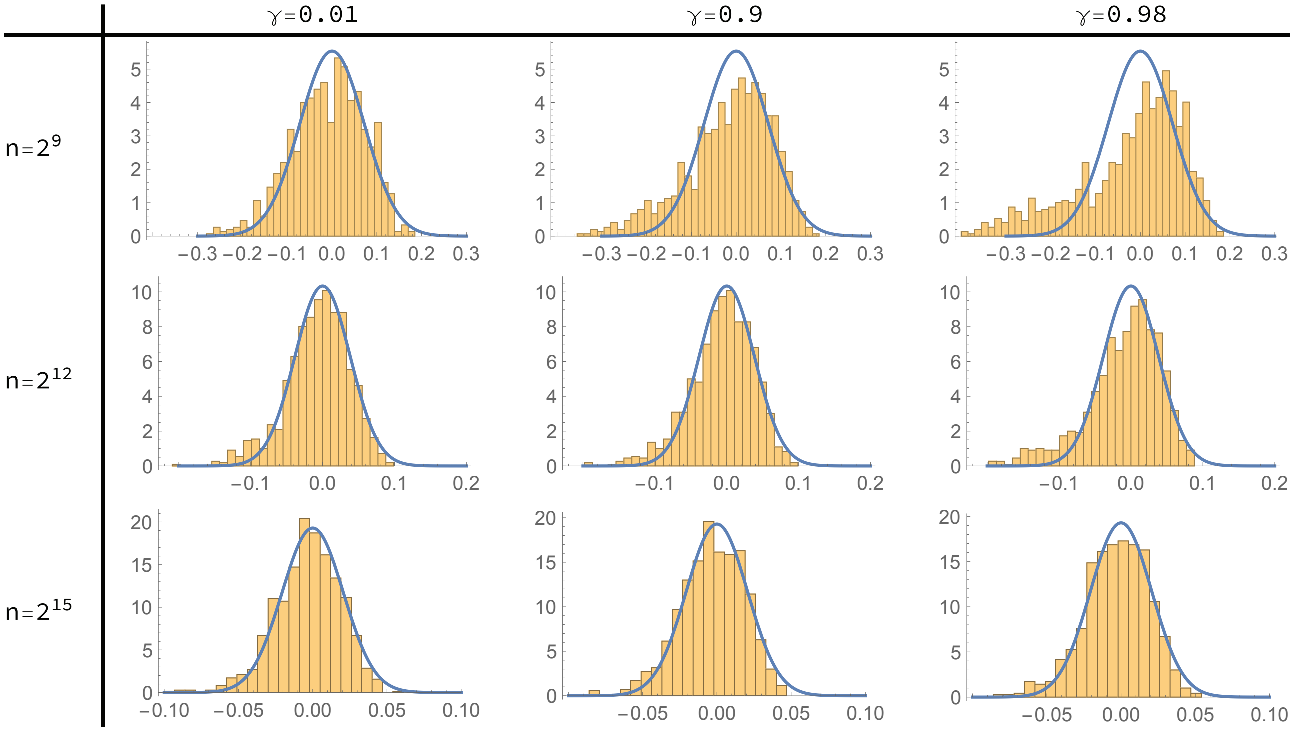

Figure 4. $\tau = 0.4$. These 9 panels only differ by n ($= 2^9$ in the top row, 2^{12} (middle) and 2^{15} in the bottom row) and by γ varying in $\{001, 0.9, 0.98\}$. In each panel, the displayed normalized histogram is that of the 3000 replicates of $\hat{h}_M - \hat{h}_n$. The superposed blue curve is the normal distribution of $\hat{h}_M - \hat{h}_n$ as predicted by the asymptotic theory.

Notice that, again as expected, the range of the abscissae (h-differences) increases by moving from $n = 2^{15}$ (bottom) to $n = 2^9$ (top).

It is good news that the approximation given by (11) and (12) is very useful for n as small as 512.

The simulation results for $\tau = 0.4$ are described in Figure 4. Here we add the three panels corresponding to $n = 2^{12}$. The analog figure for $\tau = 0.75$ is Figure 5. Now one clearly sees that, for $\gamma = 0.01$ (first column in these two 3×3 arrays of histograms) the smallest value of n ($= 2^9$) is always sufficient for the usefulness of the asymptotic normal approximation - although there is a slight deterioration for $\tau = 0.75$ (precisely the histogram in the top-left panel in Figure 5 exhibits a non-negligible proportion of “too large” negative values for $\hat{h}_M - \hat{h}_n$ which almost always are associated with too-small $\hat{h}_M$'s). One observes that the latter deterioration is softened if n is increased to 2^{12} (middle panel of first column of Figure 5). Next, an inspection of the second column (thus $\gamma = 0.9$) of both these two arrays shows that $n = 2^9$ is "just sufficient" only for the smaller $\tau = 0.4$ and provided one accepts a slight inaccuracy of the same type as the one mentioned above. But $n = 2^{12}$ is clearly required for $\tau = 0.75$. Notice that, because of the observed jump in the observed accuracy by going from $(\tau, \gamma) = (0.75, 0.01)$ to $(0.75, 0.9)$, we also repeated the same simulations for the case $(\tau, \gamma) = (0.75, 0.5)$: they produced a histogram rather close to the one for $(0.75, 0.01)$; this demonstrates that it is only for “large” γ (that is, near 0.9 or above) that the asymptotic approximation is not accurate for $n = 2^9$.

Next, the third columns (that is, for $\gamma = 0.98$) shows that $n = 2^{12}$ is required for $\tau = 0.4$, and $n = 2^{15}$ is required for $\tau = 0.75$.

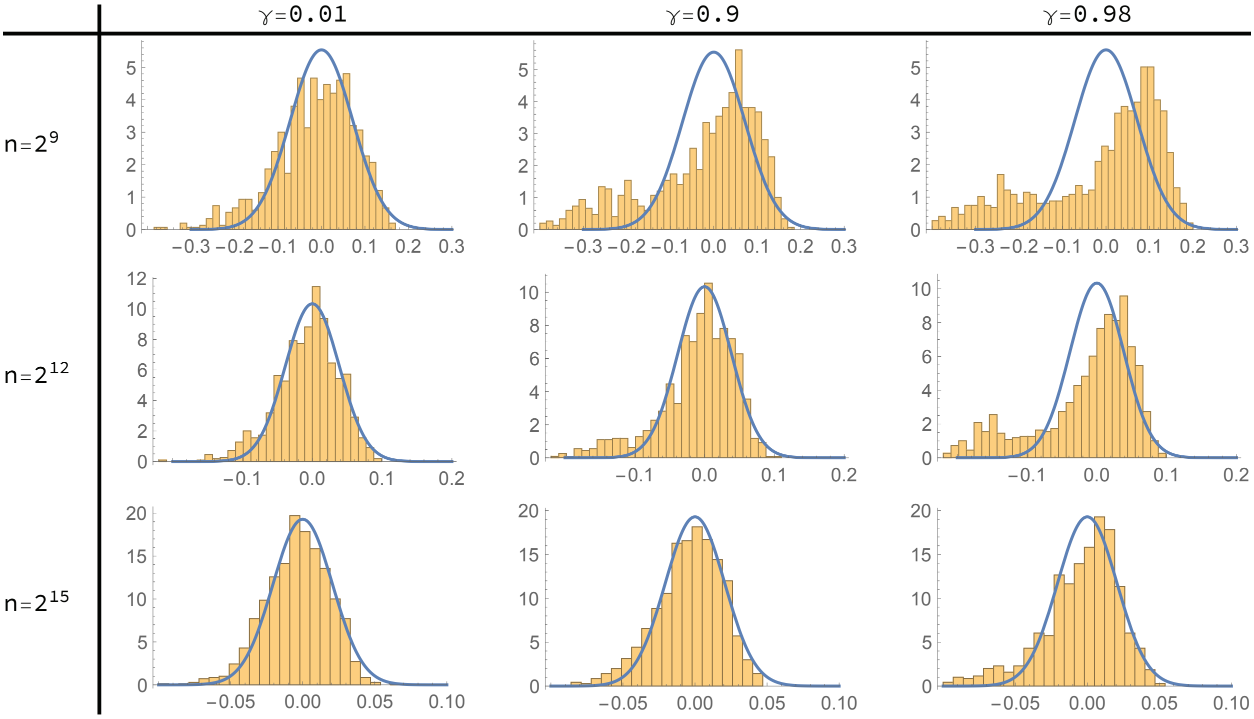

Figure 5. $\tau = 0.75$. $n = 2^9$ (top), $n = 2^{12}$ (middle) and $n = 2^{15}$ (bottom). These 9 panels only differ by n and by γ varying in $\{001, 0.9, 0.98\}$. In each panel, the displayed normalized histogram is that of the 3000 replicates of $\hat{h}_M - \hat{h}_n$. The superposed blue curve is the normal distribution of $\hat{h}_M - \hat{h}_n$ as predicted by the asymptotic theory.

All these experiments are thus well in agreement with (11)-(12). But for certain settings, which are not un-common in practice (see Taylor (1994)), this normal approximation is accurate only for quite large n (for example, n larger than 2^{12} is required for $(\tau, \gamma) = (0.4, 0.98)$). And this so-required value for n is shown to be an increasing function of both τ and γ.

Asymptotic scaled $\mathcal{X}_2(1)$ distribution of the ASE-excess. Now let us look at the usefulness of the asymptotic approximation stated in Theorem 3.1 for the excess of ASE. The constants in this approximation can be quite simplified in the context of this Monte Carlo study. Indeed, because we have taken $u \equiv 1$, by a simple algebra and using the fact that for the bi-weight kernel we have $\int (K - G)^2 = \int G^2 = \int K^2$ (see the proof of Lemma 3.1 in Girard (2010) for the first equality, and Table 1 there, for the second), we can check that the constant C in Theorem 3.1 has the simplified expression $C = (6/5)\sigma^2$. As a side remark, the fact that the asymptotic approximation is thus not a function of r when $u \equiv 1$, is quite appealing in practice: the distribution of the possible excesses of ASE can then be predicted for n "large enough", requiring only the value of σ (recall that this distribution could also be estimated in much more general contexts via the simulation of randomized choices; cf Girard (2010) for the large-n theoretical justification of this approach in the i.i.d. case).

A natural question for a practitioner is thus the meaning of "large enough" in the above discussion. We summarize in Figure 6 the simulation results for 4 settings: the "worst" setting, precisely $n = 512$ and $(\tau, \gamma) = (0.75, 0.98)$ in panel A, and three other settings, where "worst" means that the goodness of the approximation of the observed histogram by the asymptotic scaled chi-square law of Theorem 3.1 is the worst in the 9 configurations of (τ, γ)-values. This goodness can be of course improved (in agreement with the theoretical result) when n is increased: Panel B, compared to A, demonstrates that $n = 4096$ is sufficient for $(\tau, \gamma) = (0.75, 0.98)$, while $n = 512$ is not. But we also observed two facts: 1) such a goodness is also restored by decreasing γ. Panel C demonstrates that $\gamma = 0.9$ is sufficient even for $n = 512$; and we always observed this behavior for lower γ. 2) the shape parameter τ has the same impact. Indeed Panel D, where, again, $\gamma = 0.98$, demonstrates that for $\tau = 0.4$ the asymptotic scaled $\mathcal{X}_2(1)$ approximation is quite good even for $n = 512$. For the configurations $(\tau, \gamma) \in \{(0.75, 0.01), (0.4, 0.01), (0.4, 0.9), (0.2, 0.01), (0.2, 0.9), (0.2, 0.98)\}$, the results are not displayed, for sake of place, because the figures would be quasi-identical to those in either Panel B, C or D.

By comparing with the experimental results for the normal approximation of the differences of bandwidths, this latter configuration $(\tau, \gamma) = (0.4, 0.98)$ is one of many where the scaled $\mathcal{X}_2(1)$ approximation for the excess of ASE does not require very large n, while it was not the case for the bandwidths' differences.

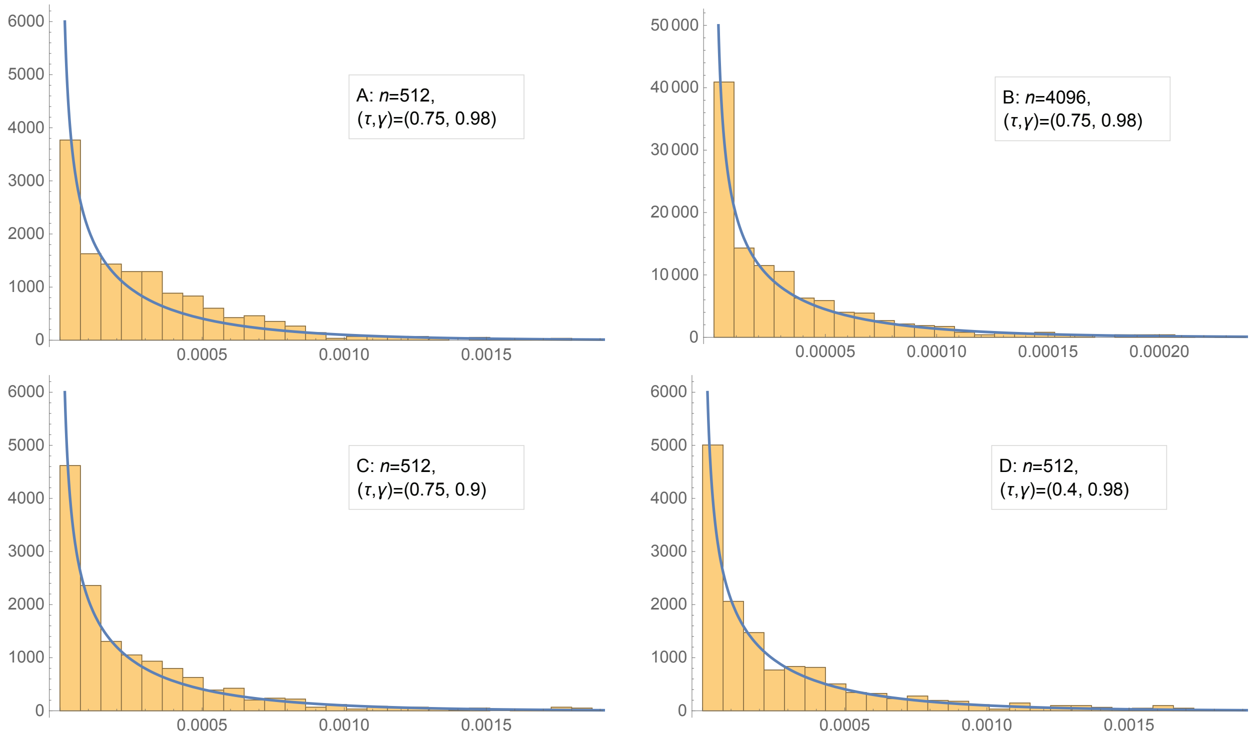

Figure 6. Assessment of the simplified $\frac{6}{5n}\sigma^2\mathcal{X}_2(1)$ approximation of the ASE-excess. In each panel, the displayed normalized histogram is that of the 3000 replicates of $T_n(\hat{h}_M) - T_n(\hat{h}_n)$. The superposed blue curve is the $\frac{6}{5n}\sigma^2\mathcal{X}_2(1)$ density as suggested by the asymptotic theory. (A): $n = 2^9$ and $(\tau, \gamma) = (0.75, 0.98)$. (B): $n = 2^{12}$ and $(\tau, \gamma) = (0.75, 0.98)$. (C): $n = 2^9$ and $(\tau, \gamma) = (0.75, 0.9)$. (D): $n = 2^9$ and $(\tau, \gamma) = (0.4, 0.98)$.

5. Proofs

In all the proofs, we denote by cst a positive constant that may be different from line to line.

5.1. Proof of Theorem 3.1

Let $\hat{h}$ be either $\hat{h}_M$ or $\hat{h}_G$. We know, using $T_n'(\hat{h}_n) = 0$, that there exists h^* between $\hat{h}$ and $\hat{h}_n$ for which,

$$\begin{aligned} n(T_n(\hat{h}) - T_n(\hat{h}_n)) &= n(\hat{h} - \hat{h}_n)T_n'(\hat{h}_n) + \frac{n}{2}(\hat{h} - \hat{h}_n)^2 T_n''(h^*) = \frac{n}{2}(\hat{h} - \hat{h}_n)^2 T_n''(h^*) \\ &= \frac{n}{2}(\hat{h} - \hat{h}_n)^2 \mathbb{E}(T_n''(h_n^*)) + \frac{n}{2}(\hat{h} - \hat{h}_n)^2 \left(T_n''(h^*) - \mathbb{E}(T_n''(h_n^*))\right), \end{aligned} \tag{18}$$

where h_n^* is as defined in (5) and is the minimizer of $D_n(h)$ (defined in (4)). Let us control the two terms of the right hand side of (18).

Control of $\frac{n}{2}(\hat{h}-\hat{h}_n)^2\mathbb{E}(T_n^{''}(h_n^*))$. Clearly,

$$\frac{n}{2}(\hat{h}-\hat{h}_n)^2\mathbb{E}(T_n^{''}(h_n^*)) = \left(\Sigma^{-1}n^{3/10}(\hat{h}-\hat{h}_n)\right)^2\frac{\Sigma^2}{2}n^{2/5}\mathbb{E}(T_n^{''}(h_n^*), \tag{19}$$

where Σ is as defined in (12). We know from (11) and (12) that

$$\Sigma^{-1}n^{3/10}(\hat{h}-\hat{h}_n) \Longrightarrow \mathcal{N}(0,1),$$

in distribution, as n tends to infinity and then

$$\left(\Sigma^{-1}n^{3/10}(\hat{h}-\hat{h}_n)\right)^2 \Longrightarrow \mathcal{X}_2(1), \tag{20}$$

in distribution, as n tends to infinity. We get (using the same arguments as for the proof of Lemma 2.1 in BGL), for any h sufficiently small and for any n sufficiently large,

$$\mathbb{E}(T_n^{''}(h)) = 3h^2\int_0^1 u(x)r^{''2}(x)dx\left(\int_{-1}^1 t^2K(t)dt\right)^2 + \frac{2\sigma^2}{nh^3}(\int_0^1 u(x)dx)\int_{-1}^1 K^2(y)dy$$
$$+o(h^2) + O(\frac{1}{n^2h^4}) + o(\frac{1}{nh^3}).$$

We deduce (recall that $h_n^* = cn^{-1/5}$),

$$n^{2/5}\mathbb{E}(T_n^{''}(h_n^*)) = 3c^2\int_0^1 u(x)r^{''2}(x)dx\left(\int_{-1}^1 t^2K(t)dt\right)^2 + \frac{2\sigma^2}{c^3}(\int_0^1 u(x)dx)\int_{-1}^1 K^2(y)dy + o(1).$$

Hence,

$$\lim_{n\to\infty} n^{2/5}\mathbb{E}(T_n^{''}(h_n^*)) = 5\sigma^{4/5}B^{2/5}A^{3/5},$$

and

$$\lim_{n\to\infty}\frac{\Sigma^2}{2}n^{2/5}\mathbb{E}(T_n^{''}(h_n^*)) = C, \tag{21}$$

with

$$C = \frac{2\sigma^2}{5A}\left(\left(\int t^2K(t)dt\right)^2\int_0^1 u^2(x)r''^2(x)dx + \frac{2A}{B}\int_0^1 u^2(x)dx\int(K-G)^2(t)dt\right).$$

We obtain collecting (19), (20) and (21),

$$\frac{n}{2}(\hat{h}-\hat{h}_n)^2\mathbb{E}(T_n^{''}(h_n^*)) \Longrightarrow C\mathcal{X}_2(1), \tag{22}$$

in distribution, as n tends to infinity.

Control of $\frac{n}{2}(\hat{h}-\hat{h}_n)^2\left(T_n^{''}(h^*)-\mathbb{E}(T_n^{''}(h_n^*))\right)$. Clearly,

$$\begin{aligned}&\frac{n}{2}(\hat{h}-\hat{h}_n)^2\left(T_n^{''}(h^*)-\mathbb{E}(T_n^{''}(h_n^*))\right)\\&=\left(n^{3/10}(\hat{h}-\hat{h}_n)\right)^2\frac{n^{2/5}}{2}\left(T_n^{''}(h^*)-\mathbb{E}(T_n^{''}(h_n^*))\right).\end{aligned}\tag{23}$$

We have,

$$\begin{aligned}&n^{2/5}\left|T_n^{''}(h^*)-\mathbb{E}(T_n^{''}(h_n^*))\right| \qquad (24)\\&\le n^{2/5}\left|\frac{T_n^{''}(h^*)}{D_n(h^*)}-1\right|\left|D_n^{''}(h^*)\right|+n^{2/5}\left|D_n^{''}(h^*)-\mathbb{E}(T_n^{''}(h_n^*))\right|,\end{aligned}$$

and

$$n^{2/5}\left|\frac{T_n^{''}(h^*)}{D_n^{''}(h^*)}-1\right|\left|D_n^{''}(h^*)\right|\le\sup_{h\in H_n}\left|\frac{T_n^{''}(h)}{D_n^{''}(h)}-1\right|\left(n^{2/5}\sup_{h\in H_n}\left|D_n^{''}(h)\right|\right).\tag{25}$$

We need the following lemma proved in BGL,

Lemma 5.1. *It holds, under the previous notations,*

$$\begin{aligned}&\limsup_{n\to\infty}\left(n^{2/5}\sup_{h\in H_n}\left|\mathbb{E}(T_n^{''}(h))\right|\right)<\infty,\ \limsup_{n\to\infty}\left(n^{2/5}\sup_{h\in H_n}\left|D_n^{''}(h)\right|\right)<\infty\\&\lim_{n\to\infty}\sup_{h\in H_n}\left|\frac{D_n^{''}(h)}{\mathbb{E}(T_n^{''}(h))}-1\right|=0\\&\lim_{n\to\infty}\sup_{h\in H_n}\left|\frac{T_n^{''}(h)}{D_n^{''}(h)}-1\right|=0,\quad \textit{in probability}.\end{aligned}$$

We deduce, using Lemma 5.1 and (25),

$$\lim_{n\to\infty}n^{2/5}\left|\frac{T_n^{''}(h^*)}{D_n^{''}(h^*)}-1\right|\left|D_n^{''}(h^*)\right|=0,\ \text{ in probability.}\tag{26}$$

Now,

$$\begin{aligned}&n^{2/5}\left|D_n^{''}(h^*)-\mathbb{E}(T_n^{''}(h_n^*))\right|\\&\le n^{2/5}\left|D_n^{''}(h^*)-D_n^{''}(h_n^*)\right|+n^{2/5}\left|\mathbb{E}(T_n^{''}(h_n^*))\right|\left|\frac{D_n^{''}(h_n^*)}{\mathbb{E}(T_n^{''}(h_n^*))}-1\right|\\&\le cst\left|\frac{h^*}{h_n^*}-1\right|+\left(n^{2/5}\sup_{h\in H_n}\left|\mathbb{E}(T_n^{''}(h))\right|\right)\sup_{h\in H_n}\left|\frac{D_n^{''}(h)}{\mathbb{E}(T_n^{''}(h))}-1\right|\\&\le cst\left|\frac{\hat{h}}{h_n^*}-1\right|+cst\left|\frac{\hat{h}_G}{\hat{h}_M}-1\right|+\left(n^{2/5}\sup_{h\in H_n}\left|\mathbb{E}(T_n^{''}(h))\right|\right)\sup_{h\in H_n}\left|\frac{D_n^{''}(h)}{\mathbb{E}(T_n^{''}(h))}-1\right|.\end{aligned}\tag{27}$$

The asymptotic optimality property (9) allows to deduce that $\left|\frac{\hat{h}}{h_n^*}-1\right|+\left|\frac{\hat{h}_G}{\hat{h}_M}-1\right|$ converges, in probability, to 0 as n tends to infinity. Consequently, the bound (27) together with Lemma 5.1, give

$$\lim_{n\to\infty} n^{2/5}\left|D_n^{''}(h^*)-\mathbb{E}(T_n^{''}(h_n^*))\right|=0, \text{ in probability.} \tag{28}$$

We deduce, collecting (24), (26) and (28),

$$\lim_{n\to\infty} n^{2/5}\left|T_n^{''}(h^*)-\mathbb{E}(T_n^{''}(h_n^*))\right|=0, \text{ in probability.} \tag{29}$$

Consequently, we deduce thanks to (23) and (20),

$$\lim_{n\to\infty} \frac{n}{2}(\hat{h}-\hat{h}_n)^2\left(T_n^{''}(h^*)-\mathbb{E}(T_n^{''}(h_n^*))\right)=0, \text{ in probability.} \tag{30}$$

The proof of Theorem 3.1 is complete by collecting (18), (22) and (30). □

5.2. Proof of Proposition 3.1

The sequence $(\epsilon_i)_{i\in\mathbb{N}}$, as defined in (13), is strictly stationary. This property follows from the strictly stationary property of the sequence $(\sigma_i)_{i\in\mathbb{N}}$. Define, now, the sequence of filtration $(\mathcal{F}_i)_{i\geq 0}$ by, for $i\in\mathbb{N}$, $\mathcal{F}_i=\sigma(Z_0,\cdots,Z_i,\sigma_i,\sigma_{i+1})$. Then ϵ_i is $\mathcal{F}_i$-measurable. Since σ_i is $\mathcal{F}_{i-1}$-measurable, Z_i is independent of $\mathcal{F}_{i-1}$ and $\mathbb{E}(Z_i)=0$, we have

$$\mathbb{E}\left(\epsilon_i|\mathcal{F}_{i-1}\right)=\sigma_i\mathbb{E}\left(Z_i|\mathcal{F}_{i-1}\right)=0,$$

almost surely. This sequence $(\epsilon_i)_{i\in\mathbb{N}}$ is then a strictly stationary MDS. We have, by the requirements of Proposition 3.1, that $\mathbb{E}(\epsilon_1^{2p})<\infty$ for some $p>8$. Conditions (A) are then satisfied. Our task now is to check Condition (B), i.e., the inequality in (10). We have, by the definition of SV model, (denoting by P the distribution of Z_1), for $i_1\leq\cdots\leq i_k<i_{k+1}\leq\cdots\leq i_q$,

$$\begin{aligned}
&\mathrm{Cov}(\epsilon_{i_1}\cdots\epsilon_{i_k},\epsilon_{i_{k+1}}\cdots\epsilon_{i_q})\\
&=\mathrm{Cov}(\sigma_{i_1}Z_{i_1}\cdots\sigma_{i_k}Z_{i_k},\sigma_{i_{k+1}}Z_{i_{k+1}}\cdots\sigma_{i_q}Z_{i_q})\\
&=\int\int z_{i_1}\cdots z_{i_q}\mathrm{Cov}(\sigma_{i_1}\cdots\sigma_{i_k},\sigma_{i_{k+1}}\cdots\sigma_{i_q})dP(z_{i_1})\cdots dP(z_{i_q})\\
&=\mathbb{E}(Z_{i_1}\cdots Z_{i_q})\mathrm{Cov}(\sigma_{i_1}\cdots\sigma_{i_k},\sigma_{i_{k+1}}\cdots\sigma_{i_q}).
\end{aligned}$$

Consequently,

$$\begin{aligned}
&\left|\mathrm{Cov}(\epsilon_{i_1}\cdots\epsilon_{i_k},\epsilon_{i_{k+1}}\cdots\epsilon_{i_q})\right|=|\mathbb{E}(Z_{i_1}\cdots Z_{i_q})||\mathrm{Cov}(\sigma_{i_1}\cdots\sigma_{i_k},\sigma_{i_{k+1}}\cdots\sigma_{i_q})|\\
&\leq cst\left|\mathrm{Cov}(\sigma_{i_1}\cdots\sigma_{i_k},\sigma_{i_{k+1}}\cdots\sigma_{i_q})\right|.
\end{aligned} \tag{31}$$

Condition (10) is then satisfied from the last inequality together with Condition (14). Condition (B) holds. The proof of Proposition 3.1 is complete since all the requirements of Theorem 3.1 are satisfied (the asymptotic optimality (9), the asymptotic normality (11)-(12) are already proven in BGL.) □

5.3. Proof of Corollary 3.1

All the higher moments of σ_1 and ϵ_1 are finite, we refer for instance to Cox, Hinkley and Barndorff-Nielsen (1996) (page 22). We have now to check that $(\epsilon_i)_{i\in\mathbb{N}}$ is strongly mixing and that Condition (14) is satisfied. Since the density of η_0 is in $\mathbb{L}^1$, the linear process $\left(\sum_{j=0}^{\infty}\gamma^j\eta_{i-j}\right)_i$ is strongly mixing (see Pham and Tran (1985)) with,

$$\alpha_s \leq K\sum_{j\geq s}\left(\sum_{k\geq j}|\gamma|^k\right)^{2/3} = O(|\gamma|^{\frac{2}{3}s}),$$

for some constant K. Similarly, the sequence $(\sigma_i)_{i\geq 0}$ is still strongly mixing with the same mixing coefficients $(\alpha_s)_s$. We deduce, from (31), that $(\epsilon_i)_{i\geq 0}$ is also strongly mixing with mixing coefficient of order $O(|\gamma|^{\frac{2}{3}s})$. We have, for $1\leq i_1\leq\cdots\leq i_k < i_{k+1}\leq\cdots\leq i_q\leq n$ such that, defining $j = i_{k+1}-i_k$, $j\geq\max_{1\leq l\leq q-1}(i_{l+1}-i_l)$, and for s, l, r strictly positive reals for which $1/s+1/l+1/r=1$,

$$\begin{aligned}&|\mathrm{Cov}(\sigma_{i_1}\cdots\sigma_{i_k},\sigma_{i_{k+1}}\cdots\sigma_{i_q})| \leq 8\alpha_j^{1/r}\|\sigma_{i_1}\cdots\sigma_{i_k}\|_s\|\sigma_{i_{k+1}}\cdots\sigma_{i_q}\|_l \\ &\leq cst\,\alpha_j^{1/r}\leq cst\,|\gamma|^{\frac{2j}{3r}},\end{aligned} \tag{32}$$

(see Davydov (1968)). Condition (14) is then satisfied since $\sum_{j\geq 1}j^4|\gamma|^{\frac{2j}{3r}}<\infty$ (recall that in this model all the moments of σ_1 are finite).

Therefore the requirements of Proposition 3.1 are satisfied so that the asymptotic optimality (9), the asymptotic normality (11)-(12) and the conclusions of Theorem 3.1 hold. □

5.4. Proof of Corollary 3.2

This sequence $(\epsilon_n)_{n\geq 1}$ has finite moments at any order. In particular, for $r\in\mathbb{N}$,

$$\mathbb{E}(\epsilon_1^{2r}) = (2r-1)(2r-3)\cdots 3\times 1\frac{\Gamma(p+r)}{\Gamma(p)}\lambda^{-r},$$

(see for instance Abraham, Balakrishna and Sivakumar (2007) and the references therein). Now the sequence $(\sigma_i)_i$ is also β-mixing (since, by definition, it's the square root of a β-mixing positive sequence $(h_i)_{i\geq 1}$) so it is also strongly mixing with

$$\alpha_n\leq\beta_n\leq a_1e^{-\rho n},$$

for some positive real numbers a_1 and ρ. This bound together with (32) prove that, for $1\leq i_1\leq\cdots\leq i_k<i_{k+1}\leq\cdots\leq i_q\leq n$ such that, letting $s=i_{k+1}-i_k$, $s\geq\max_{1\leq l\leq q-1}(i_{l+1}-i_l)$,

$$|\mathrm{Cov}(\sigma_{i_1}\cdots\sigma_{i_k},\sigma_{i_{k+1}}\cdots\sigma_{i_q})|\leq cst\;\alpha_s^{1/r}\leq cst\;e^{-\rho\frac{s}{r}}, \tag{33}$$

for some fixe $r>1$. Clearly,

$$\sum_{s\geq 1}s^4e^{-\rho\frac{s}{r}}<\infty. \tag{34}$$

Condition (14) is then satisfied from (33) and (34). As in the proof of Corollary 3.1, we deduce that all the requirements of Proposition 3.1 are satisfied. Therefore the asymptotic optimality (9), the asymptotic normality (11)-(12) and the conclusions of Theorem 3.1 hold. □

5.5. Proof of Corollary 3.3

The sequence $(\sigma_i)_{i\in\mathbb{N}}$ is associated since it is a non-decreasing function of independent random variables (see Esary, Proschan and Walkup (1967)). The random variable σ_i is bounded, $|\sigma_i| \leq e^2$ and, by (see Birkel (1988)),

$$0 \leq \mathrm{Cov}(\sigma_i, \sigma_l) \leq e^2 \sum_{j=0}^{\infty} 2^{-j} \sum_{k=0}^{\infty} 2^{-k} \mathrm{Cov}(\eta_{i-j}, \eta_{l-k}) \leq cst\, 2^{-|i-l|}.$$

For $1 \leq i_1 \leq \cdots \leq i_k < i_{k+1} \leq \cdots \leq i_q \leq n$ such that, letting $s = i_{k+1} - i_k$, $s \geq \max_{1\leq l\leq q-1}(i_{l+1} - i_l)$, we have using Birkel (1988), and the above bound,

$$|\mathrm{Cov}(\sigma_{i_1} \cdots \sigma_{i_k}, \sigma_{i_{k+1}} \cdots \sigma_{i_q})| \leq e^{2(q-2)} \sum_{i=i_1}^{i_k} \sum_{l=i_{k+1}}^{i_q} \mathrm{Cov}(\sigma_i, \sigma_l) \leq cst\, 2^{-s}.$$

Condition (14) is then satisfied, since $\sum_{s=1}^{\infty} s^4 2^{-s} < \infty$. Now, we have since σ_1 is bounded and Z_1 is normally distributed $\mathbb{E}(\epsilon_1^{2p}) < \infty$ for any $p \in \mathbb{N}$.
All the requirements of Proposition 3.1 are satisfied and therefore the asymptotic optimality (9), the asymptotic normality (11)-(12) and the conclusions of Theorem 3.1 hold. □

Acknowledgements

We thank the reviewer for his/her accurate and constructive comments. This led to a significant improvement of the original manuscript. This paper was developed in the framework of Grenoble Alpes Data Institute (ANR-15-IDEX-02).

References

Abraham, B., Balakrishna, N., and Sivakumar, R. (2007). Gamma stochastic volatility models. *Journal of Forecasting*. 25, 153-171.

Benhenni, K. Girard, D. and Louhichi, S. (2021). On bandwidth selection problems in nonparametric trend estimation under martingale difference errors. To appear in *Bernoulli*.

Billo, M. and Sartore D. (2005). Stochastic Volatility Models: A Survey with Applications to Option Pricing and Value at Risk DOI: 10.1002/0470013265.ch8 In book: Applied Quantitative Methods for Trading and Investment.

Birkel, T. (1988). The invariance principle for associated processes. *Stochastic Process. Appl.* 27, 57-71.

Bradley, R. C. (2005). Basic Properties of Strong Mixing Conditions. A Survey and Some Open Questions. *Probab. Surveys*, 2 107-144.

Bradley, R. C. (1986). Basic properties of strong mixing conditions, in: Eberlein E., Taqqu M.S. (Eds.), Dependence in Probability and Statistics, Birkhäuser, Boston, 165-192.

Cox, D.R., Hinkley, D.V. and Barndorff-Nielsen, O.E. (1996). Time Series Models: In econometrics, finance and other fields. Chapman & Hall.

Davis, R. A., Mikosch, T. (2009). Probabilistic properties of stochastic volatility models. In: Andersen, T.G., Davis, R.A., Kreiss, J.-P. and Mikosch, T. (Eds.) Handbook of Financial Time Series. Springer, 255-268.

Davydov, Y. A. (1968). Convergence of distributions generated by stationary stochastic processes. *Theory Probab. Appl.,* 13-4, 691-696.

Doukhan, P. (1994). Mixing: Properties and Examples. Lecture Notes in Statistics. New York. Springer-Verlag.

Doukhan, P. and Louhichi, S. (1999). A new weak dependence condition and applications to moment inequalities. *Stochastic Process. Appl.*, 84-2, 313-342.

Engle, R. F. (1982). Autoregressive Conditional Heteroscedasticity with Estimates of the Variance of United Kingdom Inflation. *Econometrica* **50-4**, 987-1007.

Esary, J., Proschan, F., Walkup, D. (1967). Association of random variables with applications. *Ann. Math. Stat.* 38, 1466-1476.

Fan, J., Gijbels, I., Hu, T.C. and Huang, L.S. (1996). A study of variable bandwidth selection for local polynomial regression. *Statist. Sinica* 6, 113-127.

Gaver, D. P. and Lewis, P. A. W. (1980). First-Order Autoregressive Gamma Sequences and Point Processes. *Adv. Appl. Prob.* 12-3, 727-745.

Ghysels, E., A. Harvey and E. Renault (1996). Stochastic volatility, in: *Statistical Methods in Finance*, Rao C. and Maddala G., eds., North-Holland, Amsterdam.

Girard, D. A. (1998). Asymptotic comparison of (partial) cross-validation, GCV and randomized GCV in non-parametric regression. *Ann. Stat.*, 26, 315-334.

Girard, D. A. (2010). Estimating the accuracy of (local) cross-validation via randomised GCV choices in kernel or smoothing spline regression. *Journal of Nonparametric Statistics*, 22, 41-64.

Girard, D. A. (2013). “Nonparametric Curve Estimation by Kernel Smoothers: Efficiency of Unbiased Risk Estimate and GCV Selector”. *Wolfram Demonstrations Project - A Wolfram Web Resource*, Available at `https://demonstrations.wolfram.com/NonparametricCurveEstimationByKernelSmoothersEfficiencyOfUnb/`

Härdle, W., Hall, P. and Marron, J.S. (1988). How far are automatically chosen regression smoothing parameters from their optimum? *Journal of the American Statistical Association* 83, 86-95.

Hall, P., Lahiri, S. N. and Polzehl, J. (1995). On bandwidth choice in nonparametric regression with both short and long-range dependent errors. *Ann. Statist.* 23-6, 1921-1936.

Hall, P. and Marron, J.S. (1987). Extent to which least-squares cross-validation minimises integrated square error in nonparametric density estimation. *Probability Theory and Related Fields* 74, 567-581.

Kesten, H. (1974). Renewal Theory for Functionals of a Markov Chain with General State Space. *Ann. Probab.* 2-3, 355-386.

Mammen, E. (1990) A short note on optimal bandwidth selection for kernel estimators *Statistics and Probability letters* 9, 23-25

Opsomer, J., Wang, Y., and Yang, Y. (2001). Nonparametric regression with correlated errors. *Statist. Sci.*, 61, 85-109.

Peligrad, M., Utev, S. and Wu, W. B. (2007). A maximal Lp-inequality for stationary sequences and its applications. *Proc. Am. Math. Soc.* **135**, 541-550.

Pham, T. D., Tran, L. T. (1985). Some mixing properties of time series models. *Stoch. Proc. Appl.* 19, 297-303.

Rice, J. (1984). Bandwidth choice for nonparametric regression. *Ann. Statist.* 12, 1215-1230.

Rosenblatt, M. (1956). A central limit theorem and a strong mixing condition. *Proc. Natl. Acad. Sci. USA*, 42, 43-47.

Ruppert, D., Sheather, S. J. and Wand, M. P. (1995). An effective bandwidth selector for local least squares regression. *J. Amer. Statist. Assoc.* 90, 1257-1270.

Shephard, N. (1996). Statistical aspects of ARCH and stochastic volatility, in *Time Series Models with Econometric, Finance and Other Applications*, Cox, D.R., Hinkley, D.V., Barndorff-Nielsen, O.E. (eds.), Chapman & Hall, London, 1-677.

Taylor, S. J. (1986). Modelling Financial Time Series. John Wiley and Sons, Ltd., Chichester.

Taylor, S.J. (1994). Modelling stochastic volatility: a review and comparative study. *Mathematical Finance*, 4, 183-204.