Original Article

Journal of Time Series Analysis

First version received June 2008

Published online in Wiley InterScience: 2009

(www.interscience.wiley.com) DOI: 10.1111/j.1467-9892.2009.00639.x

Postmodel selection estimators of variance function for nonlinear autoregression

Piotr Borkowski[a] and Jan Mielniczuk[b,*,†]

We consider a problem of estimating a conditional variance function of an autoregressive process. A finite collection of parametric models for conditional density is studied when both regression and variance are modelled by parametric functions. The proposed estimators are defined as the maximum likelihood estimators in the models chosen by penalized selection criteria. Consistency properties of the resulting estimator of the variance when the conditional density belongs to one of the parametric models are studied as well as its behaviour under misspecification. The autoregressive process does not need to be stationary but only existence of a stationary distribution and ergodicity is required. Analogous results for the pseudolikelihood method are also discussed. A simulation study shows promising behaviour of the proposed estimator in the case of heavy-tailed errors in comparison with local linear smoothers.

Keywords: Heteroscedastic autoregression; variance function estimation; maximum likelihood and pseudolikelihood method; postmodel selection estimators; Kullback–Leibler distance; heavy-tailed data.

1. INTRODUCTION

We focus here on the following real-valued time series $(X_t)_{t\in\mathbb{N}}$ satisfying

$$X_{t+1} = m(X_t) + \sigma(X_t)\varepsilon_{t+1}, \quad t = 0, 1 \ldots, \tag{1}$$

where $m(\cdot)$ and $\sigma(\cdot)$ are some real functions, $(\varepsilon_t)_{t\in\mathbb{N}}$ is an i.i.d sequence such that $\mathbb{E}(\varepsilon_t) = 0, \mathrm{var}(\varepsilon_t) = 1$ and ε_{t+1} is independent from the history $\mathcal{F}_t$ of the process up to the time $t, \mathcal{F}_t = \sigma(X_0 \ldots, X_t)$. Random variable X_0 has an arbitrary initial distribution. We note here that intensively studied autoregressive process with errors being Autoregressive Conditionally Heteroscedastic of order 1 [ARCH(1)] is a special case of (1) for which $\sigma^2(x) = c_0 + b_1x^2$, $c_0 \geq 0$, $b_1 \geq 0$. Also, qualitative threshold ARCH [QTARCH(1)] model

$$X_t = \sum_{j=1}^{J} \alpha_j I\{X_{t-1} \in A_j\} + \sum_{j=1}^{J} \beta_j I\{X_{t-1} \in A_j\}\varepsilon_t, \tag{2}$$

introduced by Gourieroux and Montfort (1992) is also a special case of (1). It follows from (1) that $\mathbb{E}(X_{t+1} \mid X_t) = m(X_t)$, i.e. $m(\cdot)$ is the regression of X_{t+1} given X_t, and

$$\mathrm{var}(X_{t+1} \mid X_t) = \mathbb{E}((X_{t+1} - \mathbb{E}(X_{t+1} \mid X_t))^2 \mid X_t) = \mathbb{E}((X_{t+1} - m(X_t))^2 \mid X_t) = \sigma^2(X_t),$$

provided marginal distribution of X_t has a finite second moment. Thus, $\sigma^2(\cdot)$ coincides with the autocovariance of the process, i.e. the conditional variance function of X_{t+1} given X_t. When $\sigma^2(\cdot)$ is not constant (X_t) obeying (1) is called the heteroscedastic autoregression model.

In this article, we discuss estimation of the conditional variance function $\sigma^2(\cdot)$. This is often of independent interest from estimation of the regression, especially when one would like to assess the heteroscedasticity of considered dependence structure or evaluate volatility or risk. Frequently, the conditional variance is used to evaluate some related characteristics of conditional distribution, as e.g. in Value at Risk (VaR) estimation. Some preliminary estimates of the variance are also needed to construct a variance-stabilizing transformation or weighted regression estimators.

Moreover, let us note that the Euler approximation with step Δ to the Itô time invariant diffusion process $\mathrm{d}X_t = \mu(X_t)\ \mathrm{d}t + \sigma(X_t)\ \mathrm{d}W_t$, where W_t is the standard Wiener process, satisfies (1). For small Δ properties of such approximating process resemble those of the diffusion process (Fan, 2005). Moreover, discretization error is frequently small in comparison with estimation error as argued, e.g for the Cox–Ingersoll–Ross processes by Phillips and Yu (2005, p. 340) and in such cases (1) is a good approximation of the Itô process. Let us note that stock prices were classically modelled by a geometric Brownian motion for which the conditional heteroscedastic standard deviation of the increment is proportional to the value of the process. Thus, reliable estimation of $\sigma(\cdot)$ in (1) is important in

[a]*Institute of Computer Science, Polish Academy of Sciences*
[b]*Warsaw University of Technology*
**Correspondence to: Jan Mielniczuk, Institute of Computer Science, Polish Academy of Sciences, Ordona 21, 01-237 Warsaw, Poland.*
[†]*E-mail: miel@ipipan.waw.pl*

the context of volatility estimation in financial mathematics, especially when a proposed method works satisfactorily for the case of heavy-tailed errors. We refer to Sørensen (2004) for a survey of parametric approaches to discretely observed diffusion processes. We assume throughout that the distribution of ε_t is absolutely continuous with respect to the Lebesgue measure λ on $\mathbb{R}$ with a density $f_\varepsilon(\cdot)$ and $\sigma(x) > 0$ for all $x \in \mathbb{R}$. Observe that for such a process a conditional density of X_i given $X_{i-1} = x_1$ exists and is equal

$$f(x_2 \mid x_1) = \frac{1}{\sigma(x_1)} f_\varepsilon\left(\frac{x_2 - m(x_1)}{\sigma(x_1)}\right). \tag{3}$$

We consider a parametric family $\mathcal{F}_{kl}$ of conditional densities which we use to model the unknown density $f(x_2|x_1)$. Namely, $\mathcal{F}_{kl} = \{f_\theta(x_2 \mid x_1)\}_{\theta \in \Theta_{kl}}$, where $\theta = (\beta', \eta')' \in \Theta_{kl} \subset \mathbb{R}^{k+l}$ and $f_\theta(x_2 \mid x_1)$ is defined as in (3) with $m(\cdot) = m_\theta(\cdot)$ and $\sigma(\cdot) = \sigma_\theta(\cdot)$, where

$$\begin{aligned} m_\theta(x) &= \sum_{i=1}^{k} \beta_i \phi_i(x), \\ \sigma_\theta(x) &= \exp\left\{\sum_{i=1}^{l} \eta_i \psi_i(x)\right\}, \end{aligned} \tag{4}$$

and $(\phi_i(x))_{i=1}^k$ and $(\psi_i(x))_{i=1}^l$ are two sets of linearly independent functions in the sense that their non-trivial linear combination is zero on a set of Lebesgue measure 0 only. The choice of $(\phi_i(x))_{i=1}^k$ and $(\psi_i(x))_{i=1}^l$ may depend on the model. Note, however, that if one considers models $\mathcal{M}_{11}, \mathcal{M}_{21}, \ldots, \mathcal{M}_{kl}, \ldots, \mathcal{M}_{KL}$ of conditional densities given in (4) where the pertaining functions are chosen from fixed sequences $(\phi_i(x))_{i=1}^K$ and $(\psi_i(x))_{i=1}^L$, the problem we consider here reduces to variable selection for variance estimation problem. The parameterization (eqn 4) was considered in Ledwina and Mielniczuk (2007), where the aim was to estimate the variance function using model selection in a random design regression model. In the following, we will assume that Θ_{kl} is a compact subset of $\mathbb{R}^{k+l}$. We will consider a case when f belongs to some members of a given family of parametric models having the aforesaid form as well as the case when none of them is correct. More specifically, given a finite family of models consisting of conditional densities, one of them is chosen using penalized log-likelihood method and then the regression and the variance estimators are defined as maximum likelihood (ML) estimators in this model. Such estimators, following Leeb and Pötscher (2008), will be called the postmodel selection estimators (PMS) in the article. Note that with such approach the regression and the variance functions are simultaneously estimated in contrast to most of the non-parametric procedures when the regression is estimated first and then the variance based on the squared residuals from the regression fit.

The article is structured as follows. In Section 2, we discuss imposed assumptions and prove some auxiliary results including consistency and asymptotic normality of ML estimators for parameters of a non-stationary autoregressive time series. In Section 3, we state and prove the main results of the article including consistency of PMS variance estimator when the true conditional distribution belongs to one of the models on the finite list. Moreover, we discuss behaviour of the considered selection rule under mis-specification as well as analogous results for the pseudolikelihood method. This is useful to assess mis-specification bias of PMS methods. In the last section, we discuss the results of the simulation study in which the PMS estimator of the variance was compared with two-stage estimators using local linear smoothing. In our simulation study functions $\phi_i(\cdot)$ and $\psi_j(\cdot)$ are defined as piecewise Legendre polynomials on an appropriately defined partition of a range of a data set. For an approach based on local linear smoothing we refer, e.g. to Fan and Yao (1998). In a parallel approach Yu and Jones (2004) use the local ML approach instead of the local linear method. We refer to Borkowski and Mielniczuk (2008) for analysis of medium size sample performance of such estimators.

2. ASSUMPTIONS AND AUXILIARY RESULTS

In the following π will denote a stationary marginal distribution pertaining to (X_t) and $(\tilde{X}_1, \tilde{X}_2)$ will stand for a bivariate random vector having the stationary bivariate distribution, i.e. a vector such that $\tilde{X}_1$ is distributed according to π and conditional density of $\tilde{X}_2$ given $\tilde{X}_1 = x_1$ is defined in (3). Without loss of generality we assume that $(\tilde{X}_1, \tilde{X}_2)$ is defined on the same probability space $(\Omega, \mathcal{A}, P)$ where Markov chain (X_t) is defined. Throughout, θ' denotes a transposition of a column vector θ. The following conditions will be considered for the results. All vectors are considered as column vectors. We will *not* assume, what is common in statistical inference for Markov chains, that (X_t) is stationary, see e.g. McKeague and Zhang (1994). We require only that it has a stationary distribution π and is ergodic (see condition C1 and Remark 2).

(C1) Process $(X_t)_{t\in\mathbb{N}}$ is an ergodic Markov chain in the sense of Markov chains theory, i.e. $\|P(X_t \in \cdot \mid X_0 = x_0)\ ) \pi(\cdot)\|_{tv} \to 0$, for all x_0, where $\|\cdot\|_{tv}$ denotes a total variation norm for probability measures $\|P_1 - P_2\|_{tv} = \sup_{A\in\mathcal{B}(R)} |P_1(A) - P_2(A)|$.
(C2) A density $\tilde{f}(\cdot)$ of a distribution π of $\tilde{X}_1$ with respect to the Lebesgue measure exists and $\int s^2 \tilde{f}(s)\, ds < \infty$.
(C3) $\log f_\theta(x_2|x_1)$ is two times continuously differentiable in θ.
(C4) For any $\theta \in \Theta_{kl}, I(\theta) = \mathbb{E}_P(ZZ') < \infty$, where $Z = \frac{\partial}{\partial\theta} \log f_\theta(\tilde{X}_2|\tilde{X}_1)$.
(C5) $\sup_{\theta\in\Theta_{kl}} |\log f_\theta(x_2 \mid x_1)| < g(x_1,x_2)$, where $g(x_1,x_2)$ is $P_{\tilde{X}_1,\tilde{X}_2}$-integrable.
(C6) $\sup_{\theta\in\Theta_{kl}} |\frac{\partial^2}{\partial\theta\partial\theta'} \log f_\theta(x_2 \mid x_1)| \le h(x_1, x_2)$, where $h(x_1,x_2)$ is $P_{\tilde{X}_1,\tilde{X}_2}$-integrable.
(C7) Θ_{kl} is compact.
(C8) A density f_ε is a continuous function satisfying the following condition. If $\sigma_1^{-1} f_\varepsilon((y - m_1)/\sigma_1) = \sigma_2^{-1} f_\varepsilon((y - m_2)/\sigma_2)$ for y having non-zero Lebesgue measure λ, then $m_1 = m_2$ and $\sigma_1 = \sigma_2$.

(C9) $|s|f_{\varepsilon}(s) \to 0$ when $|s| \to \infty$.

The following comments on assumptions are in order. Observe that when (X_t) is stationary, i.e. $X_1 \sim \pi$ then C1 implies ergodicity in the sense of ergodic theory. Indeed, it follows that a chain satisfying C1 is indecomposable, i.e. there do not exist two disjoint non-empty Borel sets A_1, A_2 such that $P(A_i|x) = 1$ for $x \in A_i$ for $i = 1, 2$ and the assertion follows from Thm 7.16 in Breiman (1992). It follows from Chen and Tsay (1993) that if $|m(x)/x|$ is uniformly bounded by a constant less than 1 and density f_{ε} is positive on the whole line then C1 holds and (X_t) is actually geometrically ergodic. C5 is a crucial condition needed to ensure consistency of conditional ML estimator (cf. Proposition 5) when the considered conditional density belongs to a parametric family. It is frequently used in i.i.d. case [cf., e.g. assumption A3 in White (1982)]. Consider two additional assumptions:

(A1) Functions $\{\psi_i\}_{i=1}^k$ are bounded on $\mathbb{R}$ and $\mathbb{E}\phi_j^2(\tilde{X}_1) < \infty$, $j = 1, \ldots, l$.
(A2) $f_{\varepsilon} \sim N(0, 1)$.

When f_{ε} is the standard normal density, $2 \log f_{\theta}(x_2 \,|\, x_1) = -\log \sigma_{\theta}^2(x_1) - (x_2 - m_{\theta}(x_1))^2/\sigma_{\theta}^2(x_1) - \log 2\pi$. Moreover, in view of C7 and A1 we have $\inf_{\theta \in \Theta_{kl}, x \in \mathbb{R}} \sigma_{\theta}(x) > 0$ and using $\mathbb{E}_P \tilde{X}_1^2 < \infty$ it is easy to see that C5 is satisfied. By the same token and standard calculations it can be checked that C3, C4 and C6 hold in such a case. Conditions C8 and C9 also hold; actually equality in C8 for three points suffice in the case of the normal distribution. Thus, when A2 holds we additionally need only to assume C1, C2 and C7, A1. In the following, we will assume C1–C9 keeping in mind that for normal innovations they can be significantly simplified. It is also easily seen from C3 and C5 by the Lebesgue bounded convergence theorem that $\mathbb{E}_P\{\log p_{\theta}(\tilde{X}_2 \,|\, \tilde{X}_1)\}$ is a continuous function of θ.

In an important paper, Sin and White (1996) consider model selection problem in a very general setting when the underlying data-generating process is allowed to be non-stationary and a criterion function is only assumed to be a sum of contributions pertaining to consecutive observations. The results announced in this article are similar to our Theorems 1(ii) and 2(i) that follow; for a discussion, see Remark 4. However, some of the assumptions imposed by Sin and White are overly restrictive such as their assumption of identifiable uniqueness [(vii) of their Ass A, p. 210] and a strong assumption of almost sure behaviour of the second derivative of the criterion function [Ass (iv) of Propn 4.1, p. 212]. Moreover, all the general assumptions on the Uniform Law of Large Numbers and the Central Limit Theorem (CLT) which have to be satisfied for properly normalized criterion functions and its derivatives are proved here under assumptions tailored to the autoregressive case. Our main assumptions listed concern a uniform majorization (in θ) of the log density and its second derivative by an integrable function and ergodicity of an underlying Markov sequence.

In our approach we use parametric modelling of the *conditional* density of X_i given the previous observation. This seems to be much more convenient for autoregressive processes than, e.g. parametric modelling of *bivariate* density of (X_{i-1}, X_i), as the marginal density involved in the latter depends also, in a very complicated way, on parameters of the model through the regression and variance. Since our main objective is estimation of the conditional variance function focusing on the conditional density seems more suited to our purposes. Let us note that modelling of the conditional density of multivariate random vector in the case of independent data has been considered by Vuong (1989).

First we prove several propositions on identifiability of parameters in a given model and properties of conditional ML estimators. Properties of ML-type estimators for a stationary ergodic processes are discussed, e.g. in Tjøstheim (1986). However, we impose milder conditions on the conditional density, e.g. existence of the third derivative $f_{\theta}(x_2|x_1)$ is not required. Moreover, we do not assume stationarity of (X_i).

PROPOSITION 1. *Assume C8. Parameter $\theta \in \Theta_{kl}$ is identifiable in the following sense: if $f_{\theta_0}(x_2|x_1) = f_{\theta_1}(x_2|x_1)$ for a set of (x_1,x_2) of non-zero Lebesgue measure on $\mathbb{R}^2$, then $\theta_0 = \theta_1$.*

PROOF. Let $C = \{(x_1,x_2): f_{\theta_0}(x_2|x_1) = f_{\theta_1}(x_2|x_1)\}$ and C_{x_1} be x_1 section of C. As $\lambda_2(C) = \int \lambda(C_{x_1})\lambda(\mathrm{d}x_1)$ with λ_2 denoting the Lebesgue measure on $\mathbb{R}^2$, it follows that there exists a Borel set $A \subset \mathbb{R}$ such that $\lambda(A) > 0$ and for $x_1 \in A$ $\lambda(C_{x_1}) > 0$. Thus, condition C8 implies that for such x_1 $m_{\theta_0}(x_1) = m_{\theta_1}(x_1)$ and $\sigma_{\theta_0}(x_1) = \sigma_{\theta_1}(x_1)$. The first equality translates to $\sum_{i=1}^k (\beta_{i,0} - \beta_{i,1})\phi_i(x_1) = 0$ for a set of x_1 of positive Lebesgue measure. This contradicts linear independence of $\{\phi_i\}_{i=1}^k$. The same reasoning holds for variance functions. □

PROPOSITION 2. *Assume that C2 and C8 hold and there exists $\theta_0 \in \Theta_{kl}$ such that $f(x_2|x_1) = f_{\theta_0}(x_2|x_1)$ $P_{\tilde{X}_1,\tilde{X}_2}$-a.e. Then, θ_0 is uniquely determined.*

PROOF. Observe that C2 implies that $P_{\tilde{X}_1,\tilde{X}_2}$ is absolutely continuous with respect to λ_2 and thus existence of different θ_0 and θ_1 satisfying the assumptions would imply $f_{\theta_0}(x_2 \mid x_1) = f_{\theta_1}(x_2 \mid x_1)$ for a set of (x_1,x_2) of non-zero Lebesgue measure, which contradicts Proposition 1. □

PROPOSITION 3. *Assume that the conditions of Proposition 2 hold. Then, $L(\theta) = \mathbb{E}_P \log f_{\theta}(\tilde{X}_2 \,|\, \tilde{X}_1)$ attains its unique maximum at θ_0.*

PROOF. Let $\theta_1 \neq \theta_0$ and C be defined as in the proof of Proposition 1. It follows from Proposition 1 that $\lambda_2(C) = 0$ and thus $P_{\tilde{X}_1,\tilde{X}_2}(C) = 0$. Let $B = \mathbb{R}^2 \setminus C$. Then reasoning as before we have that there exists A such that $P_{\tilde{X}_1}(A) = 1$ such that for $x_1 \in A$ $f_{\theta_0}(x_2 \mid x_1) \neq f_{\theta_1}(x_2 \mid x_1)$ for $x_2 \in B_{x_1}$ such that $P_{\tilde{X}_2 \mid \tilde{X}_1 = x_1}(B_{x_1}) = 1$. Thus, $\lambda(B_{x_1}) > 0$ and the information inequality implies that for $x_1 \in A$

$$\int f_{\theta_0}(x_2|x_1)\log f_{\theta_1}(x_2|x_1)\,\lambda(\mathrm{d}x_2) < \int f_{\theta_0}(x_2|x_1)\log f_{\theta_0}(x_2|x_1)\lambda(\mathrm{d}x_2).$$

Integrating with respect to $P_{\tilde{X}_1} = \pi$ we obtain the conclusion as $P_{\tilde{X}_1}(A) > 0$. □

Assume now that a sample path $X_1, \ldots, X_n$ from the process (1) is observable and consider a conditional density $f(x_1, x_2, \ldots, x_n|x_1)$ of $X_1, \ldots, X_n$ given $X_1 = x_1$. Observe that

$$\begin{aligned} f(x_1, x_2, \ldots, x_n|x_1) &= \prod_{i=1}^{n-1} f(x_1, \ldots, x_{n-i+1}|x_1, \ldots, x_{n-i}) \\ &= \prod_{i=1}^{n-1} f(x_{n-i+1}|x_1, \ldots, x_{n-i}) = \prod_{i=1}^{n-1} f(x_{n-i+1}|x_{n-i}) = \prod_{i=2}^{n} f(x_i|x_{i-1}), \end{aligned}$$

where the penultimate equality follows from the Markov property of (X_i). Observe that no stationarity argument is needed for the aforesaid equalities. Thus when $f(x_2|x_1)$ is modelled by elements of $\mathcal{F}_{kl}$, the following objective function can be considered, being the logarithm of the conditional density $f_\theta(x_1, x_2, \ldots, x_n|x_1)$ of $(X_1, \ldots, X_n)$ given X_1:

$$\mathcal{L}_n(\theta) = \log f_\theta(x_1, x_2, \ldots, x_n|x_1) = \sum_{i=1}^{n-1} \log f_\theta(x_{i+1}|x_i) \tag{5}$$

and we define a (conditional) ML estimator

$$\hat{\theta}_{\mathrm{ML}} = \mathrm{argmax}_{\theta\in\Theta}\mathcal{L}_n(\theta).$$

Observe that as for a considered parameterization in view of continuity of the density f_ε, $f_\theta(x_2|x_1)$ is a continuous function of θ for fixed x_2 and x_1, $\hat{\theta}_{\mathrm{ML}}$ exists as Θ_{kl} is compact. In the case when there are several points for which the maximum is attained, we choose any of them as a ML estimator. First we consider properties of a score $S_n(\theta_0) = \frac{\partial}{\partial\theta}\mathcal{L}_n(\theta)_{|\theta=\theta_0}$, where S_n is treated as a column vector.

In the following three propositions we assume that the model $\mathcal{F}_{kl}$ is correctly specified and C2 and C8 hold implying that there exists unique $\theta_0 \in \Theta_{kl}$ such that $f(x_2|x_1) = f_{\theta_0}(x_2|x_1)$ $P_{\tilde{X}_1,\tilde{X}_2}$-a.e.

Part (i) of Proposition 4 asserts that S_n is a martingale. This is a well-known property for a general stochastic process under conditions allowing to move differentiation of the joint density under the integral defining it [cf., e.g. Eqn (1.2.5) in Taniguchi and Kakizawa (2000)]. For our special case of the autoregressive process we give a direct proof under a milder condition.

PROPOSITION 4. *Assume C4 and C9. (i) $S_n(\theta_0)$ is a martingale with respect to $\mathcal{G}_n = \sigma(X_1, X_2, \ldots, X_n)$. (ii) $n^{-1/2}S_n(\theta_0) \xrightarrow{\mathcal{D}} \mathrm{N}(0, I(\theta_0))$, where $I(\theta)$ is defined in C4.*

PROOF. Observe that

$$\begin{aligned} \frac{\partial}{\partial\beta_k}\log f_\theta(X_i\,|\,X_{i-1})_{|\theta=\theta_0} &= \frac{\partial}{\partial\beta_k}\left\{-\log\sigma_\theta(X_{i-1}) + \log f_\varepsilon\left(\frac{X_i - m_\theta(X_{i-1})}{\sigma_\theta(X_{i-1})}\right)\right\}_{|\theta=\theta_0} \\ &= \frac{f'_\varepsilon}{f_\varepsilon}(\varepsilon_i)\left(\frac{-\phi_k(X_{i-1})}{\sigma(X_{i-1})}\right) =: Z_i \end{aligned}$$

and

$$\mathbb{E}_{P_{\theta_0}}(Z_i\,|\,\mathcal{G}_{i-1}) = \left(\frac{-\phi_k(X_{i-1})}{\sigma(X_{i-1})}\right)\int\frac{f'_\varepsilon(s)}{f_\varepsilon(s)}f_\varepsilon(s)\,\mathrm{d}s = 0, \tag{6}$$

provided $f(s) \to 0$ for $|s| \to \infty$. In the same way we check that $\mathbb{E}_{P_{\theta_0}}(\frac{\partial}{\partial\eta_l}\log f_\theta(X_i\,|\,X_{i-1})_{|\theta=\theta_0}|\mathcal{G}_{i-1}) = 0$ when $|s|f_\varepsilon(s) \to 0$ for $|s| \to \infty$. Thus, $S_n(\theta_0)$ is a sum of martingale differences and the conclusion follows by an application of a martingale CLT in conjunction with Cramér–Wald approach after proving that for $\mathbf{a} \in \mathbb{R}^{k+l}$

$$\frac{1}{n}\sum_{i=2}^{n}\mathbb{E}_{P_{\theta_0}}\left(\left[\mathbf{a}'\frac{\partial}{\partial\theta}\log f_\theta(X_i\,|\,X_{i-1})_{|\theta=\theta_0}\right]^2|\mathcal{G}_{i-1}\right) \xrightarrow{\mathbf{P}} \mathbf{a}'I(\theta_0)\mathbf{a} \tag{7}$$

and for any $\varepsilon > 0$

$$J_n^\varepsilon(n) \xrightarrow{\mathcal{P}} 0, \tag{8}$$

where, with the derivatives calculated at $\theta = \theta_0$,

$$J_n^{\varepsilon}(b) = \frac{1}{n}\sum_{i=2}^{n} \mathbb{E}_{P_{\theta_0}}\left(\left[\mathbf{a}'\frac{\partial}{\partial\theta}\log f_\theta(X_i \mid X_{i-1})\right]^2 \mathbf{I}\left\{\left|\mathbf{a}'\frac{\partial}{\partial\theta}\log f_\theta(X_i \mid X_{i-1})\right| > (\varepsilon b)^{1/2}\right\}\right)\Big|\mathcal{G}_{i-1}\right). \tag{9}$$

We check (8). Fix $\eta > 0$. Observe that for $n \geq b$

$$P(J_n^{\varepsilon}(n) > 2\eta) \leq P(J_n^{\varepsilon}(b) > 2\eta) \leq P(|J_n^{\varepsilon}(b) - J^{\varepsilon}(b)| > \eta) + P(J^{\varepsilon}(b) > \eta), \tag{10}$$

where

$$J^{\varepsilon}(b) = \mathbb{E}_{P_{\theta_0}}\left(\left[\mathbf{a}'\frac{\partial}{\partial\theta}\log f_\theta(\tilde{X}_2 \mid \tilde{X}_1)\right]^2 \mathbf{I}\left\{\left|\mathbf{a}'\frac{\partial}{\partial\theta}\log f_\theta(\tilde{X}_2 \mid \tilde{X}_1)\right| > (\varepsilon b)^{1/2}\right\}\right).$$

It follows from C4 that the second term is 0 for sufficiently large b. Moreover, as ergodicity implies that (X_t) is Harris recurrent [Propn 6.3 in Nummelin (1983)] and the Strong Law of Large Numbers (SLLN) holds for its functionals (Meyn and Tweedie, 1993, Thm 17.0.1), we have that the first term tends to 0. From this (8) readily follows. Convergence of the conditional variance in (7) follows from SLLN as. □

Remark 1. Asymptotic normality of $n^{-1/2}S_n(\theta_0)$ also follows from a CLT for Markov chains (cf., Ibragimov and Linnik, 1971, Thm 19.1.2) but under more stringent assumptions that (X_t) is geometrically ergodic and existence of the moment of order $2 + \delta$ of $\frac{\partial}{\partial\theta}\log f_\theta(\tilde{X}_i \mid \tilde{X}_{i-1})_{|\theta=\theta_0}$.

Proposition 5. *Assume C1, C5 and C7. Then* $\hat{\theta}_{\mathrm{ML}} \to \theta_0$ *a.e.*

Proof. The following generalization (Lemma 1) of Jennrich's (1969) result will be used.

Lemma 1. *If* (X_i) *is an ergodic Markov chain and* $g(s,t,\theta)$ *a measurable function on* $\mathbb{R}^2 \times \Theta_{kl}$ *such that it is continuous in* θ *and* $\mathbb{E}_{P_{\theta_0}} \sup_{\theta\in\Theta_{kl}} |g(\tilde{X}_1, \tilde{X}_2, \theta)| < \infty$ *then*

$$\sup_{\theta\in\Theta_{kl}} |n^{-1}\sum_{i=2}^{n} gX_{i-1}, (X_i, \theta) - \mathbb{E}_{P_{\theta_0}} g(\tilde{X}_1, \tilde{X}_2, \theta)| \to 0 \quad \text{a.e.} \tag{11}$$

Proof of lemma 1. By the same token as in Lemma 2 one can prove that if (X_i) is an ergodic Markov chain the same is true for the chain $[(X_{i-1},X_i)]$. Thus arguing as before, we have that Thm 17.0.1 in Meyn and Tweedie (1993) implies that SLLN holds for its functionals. In particular, for $h(x_1,x_2) = \sup_{\theta\in\Theta_{kl}}|g(x_{i-1},x_i,\theta)|$ we have that $n^{-1}\sum_{i=1}^{n} h(X_{i-1}, X_i) \to \mathbb{E}_P h(\tilde{X}_1, \tilde{X}_2)$ a.e. Thus the conclusion follows from a slight adaptation of the original proof of Thm 1 in Jennrich (1969). □

Proof of proposition 5. Observe that in view of C5 and C7 it follows from (11) for $g(x_1,x_2,\theta) = \log f_\theta(x_2|x_1)$ that

$$\lim_{n\to\infty} n^{-1}\mathcal{L}_n(\hat{\theta}_{\mathrm{ML}}) = \lim_{n\to\infty} \sup_{\theta\in\Theta_{kl}} n^{-1}\mathcal{L}_n(\theta) \to \sup_{\theta\in\Theta_{kl}} \mathbb{E}_{P_{\theta_0}} \log f_\theta(\tilde{X}_2 \mid \tilde{X}_1) \tag{12}$$

almost everywhere. This implies the conclusion. Indeed, if $\hat{\theta}_n$ does not converge to θ_0 on a set A of a positive measure, then for any $\omega \in A$ $\hat{\theta}_{n_k} \to \theta_1 \neq \theta_0$ for a certain subsequence $n_k(\omega)$. For this subsequence using (11) again together with continuity of $L(\theta)$ and Proposition 3 we have that

$$n_k^{-1}\mathcal{L}_{n_k}(\hat{\theta}_{n_k}) = \sup_{\theta\in\Theta_{kl}} n_k^{-1}\mathcal{L}_{n_k}(\theta) \to \mathbb{E}_{P_{\theta_0}} \log f_{\theta_1}(\tilde{X}_2 \mid \tilde{X}_1) < \mathbb{E}_{P_{\theta_0}} \log f_{\theta_0}(\tilde{X}_2 \mid \tilde{X}_1), \tag{13}$$

which contradicts (12). □

Proposition 6. *Assume that C1–C9 hold and* $\theta_0 \in \mathrm{Int}\Theta_{kl}$. *Then* $\sqrt{n}(\hat{\theta}_{\mathrm{ML}} - \theta_0) \xrightarrow{\mathcal{D}} N(0, A(\theta_0)^{-1}I(\theta_0)A(\theta_0)^{-1})$, *where* $A(\theta_0) = \mathbb{E}_{P_{\theta_0}}(\frac{\partial^2}{\partial\theta\partial\theta'}\log f_\theta(\tilde{X}_2 \mid \tilde{X}_1)_{|\theta=\theta_0})$.

Proof. Observe that in view of the previous proposition $\hat{\theta}_{\mathrm{ML}}$ also belongs to the interior of Θ and whence $\frac{\partial}{\partial\theta}\mathcal{L}_n(\hat{\theta}_{\mathrm{ML}}) = 0$. Thus

$$0 = \sqrt{n}\frac{\partial}{\partial\theta}\mathcal{L}_n(\hat{\theta}_{\mathrm{ML}}) = \sqrt{n}\frac{\partial}{\partial\theta}\mathcal{L}_n(\theta_0) + \sqrt{n}\frac{\partial^2}{\partial\theta\partial\theta'}\mathcal{L}_n(\theta^*)(\hat{\theta}_{\mathrm{ML}} - \theta_0), \tag{14}$$

where θ^* is in between θ_0 and $\hat{\theta}_{\mathrm{ML}}$. From conditions C3, C6 and Lemma 1 it follows that

$$\frac{1}{n}\frac{\partial^2}{\partial\theta\partial\theta'}\mathcal{L}_n(\theta^*) = \frac{1}{n}\frac{\partial^2}{\partial\theta\partial\theta'}\mathcal{L}_n(\theta_0) + o_P(1). \tag{15}$$

Moreover, in view of SLLN (cf. proof of Lemma 1)

$$-\frac{1}{n}\frac{\partial^2}{\partial\theta\partial\theta'}\mathcal{L}_n(\theta_0) \to -\mathbb{E}_{P_{\theta_0}}\frac{\partial^2}{\partial\theta\partial\theta'}\log f_\theta(\tilde{X}_2\,|\,\tilde{X}_1)_{|\theta=\theta_0} = -A(\theta_0) \tag{16}$$

almost everywhere. Observe that as $-A(\theta_0) = -\frac{\partial^2}{\partial\theta\partial\theta'}L(\theta)_{|\theta=\theta_0}$ in view of C6, it is positive definite as θ_0 is the unique maximizer of $L(\theta_0)$. Thus, it follows that $-n^{-1}\frac{\partial^2}{\partial\theta\partial\theta}\mathcal{L}_n(\theta^*)$ is positive definite with probability tending to 1. Therefore

$$\sqrt{n}(\hat{\theta}_{\mathrm{ML}} - \theta_0) = \left(-\frac{1}{n}\frac{\partial^2}{\partial\theta\partial\theta'}\mathcal{L}_n(\theta^*)\right)^{-1}\frac{1}{\sqrt{n}}\frac{\partial}{\partial\theta}\mathcal{L}_n(\theta_0) \tag{17}$$

and the conclusion follows from the last equality together with Proposition 4 and (15). □

Remark 2. If (X_t) is a stationary and ergodic sequence (in the sense of ergodic theory), then Propositions 3–5 hold true as the ergodic theorem can be used instead of the SLLN for Markov chains. Finding conditions on m and σ such that there exists stationary and ergodic process satisfying (1) is far from trivial, however, and that is the main reason we employ the presented approach. For some conditions sufficient for stationarity and ergodicity of nonlinear autoregressive process see Wu and Shao's (2004) extension of Diaconis and Freedman (1999) in which it is proved that, given these conditions, the stationary process $(X_t)_{t\in Z}$ which satisfies (1) can be represented as $J(\ldots,\varepsilon_{t-1},\varepsilon_t)$ for some measurable function J, and thus is ergodic.

Remark 3. Under additional regularity conditions ensuing $\int\frac{\partial^2}{\partial\theta\partial\theta}f_\theta(\tilde{x}_2\,|\,\tilde{x}_1)\,\mathrm{d}\tilde{x}_2 = 0$ for almost all $\tilde{x}_1$ with respect to $P_{\tilde{X}_1}$ the asymptotic covariance reduces to $I(\theta_0)^{-1}$.

3. MAIN RESULTS

We consider now the main problem of our article, namely a model selection procedure for the variance estimation when several competing parametric models are considered. Model selection will be decided according to values of the penalized conditional likelihood function $\mathcal{S}$ defined next. We focus here on models given by (3) and (4) with varying k and l and possibly different sets of orthonormal functions. As sets of orthonormal functions may vary from model to model, the models under study do not need to be nested. Examples of such approach will be presented in the simulation part. Two main results that follow correspond to cases when one of the models is correctly specified or all of them are mis-specified. They are stated for the case when one of the two models is chosen, the case of selecting a model from a finite list follows from this.

3.1. Case of correctly specified model

We consider first the case when the conditional distribution $f(x_2\,|\,x_1)$ belongs to one (but not necessarily the only one) of the parametric families $\{\mathcal{F}_{k,l}\}$, $\mathcal{F}_{k_0,l_0}$, say, i.e. there exists $\theta_0\in\Theta_{k_0,l_0}$ such that $f(x_2\,|\,x_1) = f_{\theta_0}(x_2\,|\,x_1)$ $P_{\tilde{X}_1,\tilde{X}_2}$-a.e. When a second model $\mathcal{F}_{k_1,l_1}$ is considered as an alternative model, we choose a model having a larger value of penalized conditional log-likelihood $\mathcal{S}_i, i = 0, 1$

$$\mathcal{S}_0 = \sup_{\theta\in\Theta_{k_0,l_0}}\mathcal{L}_n^{k_0,l_0}(\theta) - (k_0+l_0)c_n \quad \mathcal{S}_1 = \sup_{\theta\in\Theta_{k_1,l_1}}\mathcal{L}_n^{k_1,l_1}(\theta) - (k_1+l_1)c_n,$$

where c_n is some penalty depending on n. In particular, for $c_n = 1$ one obtains the Akaike information criterion (AIC) and for $c_n = 1/2$ log $(n-1)$ Schwarz's rule (Bayesian information criterion or BIC). Term $\log(n-1)$ instead of $\log n$ is motivated by $n-1$ terms appearing in the likelihood function. Choosing the model specified by the larger value of $\mathcal{S}_i$ corresponds to a choice of a more parsimonious model among the two models. We prove the following result, which parallels the result for i.i.d. data (X_i,Y_i) generated from a random design regression model with heteroscedastic normal errors proved in Ledwina and Mielniczuk (2007). We stress that $f_\theta(x_2|x_1)$ is a general notation for parametric conditional densities in *any* of the considered parametric families and which family is considered depends on the context. When two families are under consideration, respective members are denoted by $f_{\theta_0}(x_2\,|\,x_1)$ and $f_{\theta_1}(x_2|x_1)$.

Theorem 1. *Assume that conditions C1–C9 hold for Θ_{k_0,l_0} and Θ_{k_1,l_1} and $f(x_2\,|\,x_1)=f_{\theta_0}(x_2\,|\,x_1)$ $P_{\tilde{X}_1,\tilde{X}_2}$-a.e. for some $\theta_0\in Int(\Theta_{k_0,l_0})$. Then $P(\mathcal{S}_0 > \mathcal{S}_1)\to 1$ provided one of the following conditions hold: (i) $f(x_2\,|\,x_1) = f_{\theta_1}(x_2\,|\,x_1)$ $P_{\tilde{X}_1,\tilde{X}_2}$-a.e. for some $\theta_1\in Int(\Theta_{k_1,l_1})$, $k_1+l_1 > k_0+l_0$ and $c_n\to\infty$ (ii) $f(x_2\,|\,x_1)\in\mathcal{F}_{k_0,l_0}\setminus\mathcal{F}_{k_1,l_1}$ and $c_n = o(n)$.*

PROOF. Consider case (i). Expanding $\mathcal{L}_n^{k_0,l_0}(\theta_0) - \mathcal{L}_n^{k_0,l_0}(\hat{\theta}_{\mathrm{ML}}^0)$ around ML estimator $\hat{\theta}_{\mathrm{ML}}^0$ in $\mathcal{F}_{k_0,l_0}$ we get

$$\mathcal{L}_n^{k_0,l_0}(\theta_0) - \mathcal{L}_n^{k_0,l_0}(\hat{\theta}_{\mathrm{ML}}^0) = \frac{1}{2}(\theta_0 - \hat{\theta}_{\mathrm{ML}}^0)' \frac{\partial^2}{\partial\theta\partial\theta'} \mathcal{L}_n^{k_0,l_0}(\theta^*)(\theta_0 - \hat{\theta}_{\mathrm{ML}}^0) =$$
$$= \frac{1}{2}(\theta_0 - \hat{\theta}_{\mathrm{ML}}^0)' \frac{\partial^2}{\partial\theta\partial\theta'} \mathcal{L}_n^{k_0,l_0}(\theta_0)(\theta_0 - \hat{\theta}_{\mathrm{ML}}^0) + o_p(1),$$

where the last equality follows from Proposition 5 and C6. This implies $\mathcal{Q}_0 := \mathcal{L}_n^{k_0,l_0}(\theta_0) - \mathcal{L}_n^{k_0,l_0}(\hat{\theta}_{\mathrm{ML}}^0) = \mathcal{O}_P(1)$ and analogously $\mathcal{Q}_1 := \mathcal{L}_n^{k_1,l_1}(\theta_1) - \mathcal{L}_n^{k_1,l_1}(\hat{\theta}_{\mathrm{ML}}^1) = \mathcal{O}_P(1)$. Thus, since $\mathcal{L}_n^{k_1,l_1}(\theta_1) = \mathcal{L}_n^{k_0,l_0}(\theta_0)$ we have $\mathcal{L}_n^{k_1,l_1}(\hat{\theta}_{\mathrm{ML}}^1) - \mathcal{L}_n^{k_0,l_0}(\hat{\theta}_{\mathrm{ML}}^0) = \mathcal{Q}_0 - \mathcal{Q}_1 = \mathcal{O}_P(1)$ and

$$P(\mathcal{S}_0 > \mathcal{S}_1) = P(\mathcal{L}_n^{k_1,l_1}(\hat{\theta}_{\mathrm{ML}}^1) - \mathcal{L}_n^{k_0,l_0}(\hat{\theta}_{\mathrm{ML}}^0) < (k_1 + l_1 - k_0 - l_0)c_n) \to 1.$$

For case (ii) the proof follows the proof of Thm 1 in Ledwina and Mielniczuk (2007), after noting that $n^{-1}\sum_{i=2}^n \log f_\theta(X_i \mid X_{i-1}) \to \mathbb{E}_P \log f_\theta(\tilde{X}_2 \mid \tilde{X}_1)$ a.e. uniformly on Θ_{k_1,l_1} and Θ_{k_0,l_0} in view of the proof of Proposition 5 and that $\sup_{\theta\in\Theta_{k_i,l_i}} \mathbb{E}_P \log f_\theta(\tilde{X}_2 \mid \tilde{X}_1)$ for $i = 0,1$ is attained owing to C3, C5 and C7. □

It follows from Theorem 1 that when one chooses a model from a list containing a finite number of models satisfying imposed assumptions, a correctly specified model described by the smallest number of parameters will be chosen with probability tending to 1 by an appropriately defined penalized selection rule. More specifically, consider a finite collection of models $\{\mathcal{F}_{kl}\}_{k\in\mathcal{U},l\in\mathcal{V}}$, where $\mathcal{U}$ and $\mathcal{V}$ are some subsets of integers, such that each of them satisfy assumptions C1–C9. Let $\hat{\theta}_{kl}$ be ML estimator of θ in $\mathcal{F}_{kl}$ and define

$$(\hat{k},\hat{l}) = \mathrm{argmax}_{k\in\mathcal{U},l\in\mathcal{V}}\{\mathcal{L}_n^{kl}(\hat{\theta}_{kl}) - c_n(k+l)\}.$$

In the case of multiple maxima, arbitrary one of them is chosen. Moreover, set

$$(k^*,l^*) = \mathrm{argmin}_{k\in\mathcal{U},l\in\mathcal{V}}\{k+l : f_{\theta_0}(x_2 \mid x_1) \in \mathcal{F}_{kl}\}.$$

Then we have 4 Corollary 1.

COROLLARY 1. *Assume that (k^*,l^*) is uniquely defined, $c_n = o(n)$ and $c_n \to \infty$. Then*

$$\lim_{n\to\infty} P_{\theta_0}((\hat{k},\hat{l}) = (k^*,l^*)) = 1.$$

Estimator $\hat{\sigma}^2_{\hat{k}\hat{l}}$ in the family $\mathcal{F}_{\hat{k}\hat{l}}$ obtained by plugging in ML estimators of η into formula (4) is called PMS estimator of variance and will be denoted by $\hat{\sigma}^2$. We refer to Leeb and Pötscher (2008) for an overview of methods of construction and properties of PMS estimators. It easily follows from Corollary 1 and consistency of ML estimators that the property of Corollary 2 holds.

COROLLARY 2. *Assume that functions $\{\psi_i(\cdot)\}_{i=1}^k$ are bounded. Then under assumptions of Corollary* 1

$$\sup_{x\in\mathbb{R}} |\hat{\sigma}^2(x) - \sigma^2(x)| \xrightarrow{\mathcal{P}} 0.$$

3.2. Mis-specification case

We consider now the case when the conditional density $f(x_2 \mid x_1)$ does not belong to any of the models on the list. Next, we state a result concerning such a case which essentially asserts that when two models are under consideration the one for which the minimal value of $-\mathbb{E}_P \log f_\theta(\tilde{X}_2 \mid \tilde{X}_1)$ over the corresponding parameter set is smaller is selected. Note that minimization of this quantity is equivalent to minimization of averaged Kullback–Leibler distance $\mathbb{E}_{\tilde{X}_1} KL(f(\tilde{X}_2 \mid \tilde{X}_1), f_\theta(\tilde{X}_2 \mid \tilde{X}_1))$ with respect to θ. In the situation when the value is the same for both models, the one having smaller number of parameters is chosen. Define pseudo-true parameter values as:

$$\theta_0^* = \mathrm{argmin}_{\theta\in\Theta_{k_0,l_0}}\{-\mathbb{E}_P \log f_\theta(\tilde{X}_2 \mid \tilde{X}_1)\}, \quad \theta_1^* = \mathrm{argmin}_{\theta\in\Theta_{k_1,l_1}}\{-\mathbb{E}_P \log f_\theta(\tilde{X}_2 \mid \tilde{X}_1)\}.$$

Observe that provided C3, C5 and C7 hold for both Θ_{k_0,l_0} and Θ_{k_1,l_1} then θ_0^* and θ_1^* exist.
In the following, we use:
(C10) θ_0^* and θ_1^* are unique and belong to the interior of the respective parameter set.
Denote by D_0, D_1 the pertaining minimal values. A Markov chain $(X_n)_{n\in\mathbb{N}}$ is called geometrically ergodic if for any x_0

$$\|P(X_n \in \cdot \mid X_0 = x_0) - P(\tilde{X}_1 \in \cdot)\|_{\mathrm{tv}} \le M(x_0)\rho^n, \tag{18}$$

for some $\rho < 1$ and $M(x_0) < \infty$, where $\|\cdot\|_{\mathrm{tv}}$ denotes the total variation norm on $\mathcal{B}(\mathbb{R})$.

We then have Theorem 2.

THEOREM 2. *Assume that C1–C10 hold for both Θ_{k_0,l_0} and Θ_{k_1,l_1}. Then $P(\mathcal{S}_0 > \mathcal{S}_1) \to 1$ provided any of the following conditions hold: (i) $D_0 < D_1$ and $c_n = o(n)$; (ii) $D_0 = D_1$, $f_{\theta_0^*}(x_2 \mid x_1) = f_{\theta_1^*}(x_2 \mid x_1)$ $P_{\tilde{X}_1,\tilde{X}_2}$-a.e., $k_0 + l_0 < k_1 + l_1$, $c_n \to \infty$; (iii) $D_0 = D_1$, $f_{\theta_0^*}(x_2 \mid x_1)$ is not equal $f_{\theta_1^*}(x_2 \mid x_1)$ $P_{\tilde{X}_1,\tilde{X}_2}$-a.e. and $k_0 + l_0 < k_1 + l_1$ when (X_i) is geometrically ergodic, $n^{1/2} = o(c_n)$ and $\mathbb{E}_P|\log_{\theta_i^*}(\tilde{X}_2 \mid \tilde{X}_1)|^{2+\delta} < \infty$ for $i = 0,1$ and some $\delta > 0$.*

It follows from comparison of conditions in Theorem 2 that the most difficult case of choosing the more parsimonious model is when the Kullback–Leibler distances between P and the two models coincide and the densities corresponding to the respective pseudo-true values are different. Such situation may occur for non-nested models. We note that C10 is only necessary for parts (ii) and (iii).

REMARK 4. Sin and White (1996) considered a quantity Δ_n, which in the case of the conditional log-likelihood function equals

$$\Delta_n = n^{-1}\left(\sum_{i=1}^{n}\mathbb{E}\log f_{\theta_{n1}^*}(X_i \mid X_{i-1}) - \sum_{i=1}^{n}\mathbb{E}\log f_{\theta_{n2}^*}(X_i \mid X_{i-1})\right),$$

where $\theta_{ni}^*, i = 1,2$ is the maximizer of the respective sum in the aforegiven expression. They proved in Prop 4.2(i) that if $\liminf\Delta_n > 0$ then an analogue of our Theorem 1(ii) and Theorem 2(i) hold. As (X_i) is not necessarily stationary the assumptions in both cases are different but we conjecture that under appropriate conditions $D_0 < D_1$ implies $\liminf\Delta_n > 0$.

In the proof, Lemma 2 will be used.

LEMMA 2. *If (X_i) is the geometrically ergodic Markov chain then the same is true for (X_{i-1}, X_i).*

PROOF Only checking geometric ergodicity is non-trivial. It suffices to prove that uniformly in $A, B \in \mathcal{B}(R)$

$$|P((X_n, X_{n+1}) \in A \times B \mid (X_0, X_1) = (x_0, x_1)) - P((\tilde{X}_1, \tilde{X}_2 \in A \times B)| \le M(x_0, x_1)\rho^n,$$

where $M(x_0,x_1) < \infty$. Denote by $p^{(n)}(x,y)$ probability density of n-step transition from x to y and $P^{(n)}$ corresponding distribution. Then the expression within the absolute value equals

$$\int_A P^{(1)}(x,B)p^{(n-1)}(x_1,x)\mathrm{d}x - \int_A P^{(1)}(x,B)\tilde{f}(x)\mathrm{d}x,$$

and its absolute value can be bounded by

$$\begin{aligned}\int_A |p^{(n-1)}(x_1,x) - \tilde{f}(x)|\mathrm{d}x &\le \int_R |p^{(n-1)}(x_1,x) - \tilde{f}(x)|\mathrm{d}x \\ &= 2\|P^{(n-1)}(x_1,\cdot) - P(\tilde{X}_1 \in \cdot)\|_{\mathrm{tv}} \le 2M(x_1)\rho^{n-1},\end{aligned} \tag{19}$$

where the equality is the result of Scheffe's theorem (Devroye, 1987, p. 2). Obviously, $M(x_1) < \infty$. From this, the conclusion follows. □

PROOF OF THEOREM 2. Observe first that the obvious analogues of Propositions 5 and 6 hold with θ_0 replaced by the pseudo-true parameter with almost the same proof with the only essential change that conclusion of Proposition 3 used in the proof is replaced by the assumption that $-\mathbb{E}_P \log f_\theta(\tilde{X}_2 \mid \tilde{X}_1)$ has the *unique* minimum over both parameter sets. Now (i) can be proved using the same reasoning as in proof of Thm 2 in Ledwina and Mielniczuk (2007). It is based on the fact that n^{-1} times log-likelihood calculated at the ML estimator on Θ_{k_i,l_i} converges in probability to $\mathbb{E}_P \log f_{\theta_i^*}(\tilde{X}_2 \mid \tilde{X}_1)$ for $i = 1,2$. This follows from uniform in θ a.s. convergence of averaged log-likelihood to its expected value. We omit the details. To prove (ii) and (iii) denote by $\hat{\theta}^0_{\mathrm{ML}}$ and $\hat{\theta}^1_{\mathrm{ML}}$ maximum likelihood estimators of θ in $\mathcal{F}_{k_0,l_0}$ and $\mathcal{F}_{k_1,l_1}$ and observe that

$$\mathcal{L}_n^{k_0,l_0}(\theta_0^*) = \mathcal{L}_n^{k_0,l_0}(\hat{\theta}^0_{\mathrm{ML}}) + \frac{\sqrt{n}}{2}(\hat{\theta}^0_{\mathrm{ML}} - \theta_0^*)'\frac{1}{n}\frac{\partial^2}{\partial\theta\partial\theta'}\mathcal{L}_n^{k_0,l_0}(\tilde{\theta}^*)\sqrt{n}(\hat{\theta}^0_{\mathrm{ML}} - \theta_0^*), \tag{20}$$

where $\tilde{\theta}^*$ is in between θ_0^* and $\hat{\theta}^0_{\mathrm{ML}}$. Reasoning as in Proposition 6 [cf. eqns (15) and (16)] we obtain that $n^{-1}\frac{\partial^2}{\partial\theta\partial\theta'}\mathcal{L}_n^{k_0,l_0}(\tilde{\theta}^*) = \mathbb{E}_P\frac{\partial^2}{\partial\theta\partial\theta'}\log f_\theta(\tilde{X}_2 \mid \tilde{X}_1)_{|\theta=\theta_0^*} + o_P(1)$. Thus in view of an analogue of Proposition 5 mentioned before it follows that

$$\mathcal{L}_n^{k_0,l_0}(\theta_0^*) = \mathcal{L}_n^{k_0,l_0}(\hat{\theta}^0_{\mathrm{ML}}) + \frac{\sqrt{n}}{2}(\hat{\theta}^0_{\mathrm{ML}} - \theta_0^*)'\mathbb{E}_P\frac{\partial^2}{\partial\theta\partial\theta'}\log f_\theta(\tilde{X}_2 \mid \tilde{X}_1)_{|\theta=\theta_0^*}\sqrt{n}(\hat{\theta}^0_{\mathrm{ML}} - \theta_0^*) + o_P(1) \tag{21}$$

and

$$\mathcal{L}_n^{k_1,l_1}(\theta_1^*) = \mathcal{L}_n^{k_1,l_1}(\hat{\theta}_{\mathrm{ML}}^1) + \frac{\sqrt{n}}{2}(\hat{\theta}_{\mathrm{ML}}^1 - \theta_1^*)' \mathbb{E}_P \frac{\partial^2}{\partial\theta\partial\theta'} \log f_\theta(\tilde{X}_2 \,|\, \tilde{X}_1)_{|\theta=\theta_1^*} \sqrt{n}(\hat{\theta}_{\mathrm{ML}}^1 - \theta_1^*) + o_P(1). \tag{22}$$

As in case (ii) we have that $\mathcal{L}_n^{k_0,l_0}(\theta_0^*) = \mathcal{L}_n^{k_1,l_1}(\theta_1^*)$ it follows from (21) and (22) that

$$\mathcal{L}_n^{k_1,l_1}(\hat{\theta}_{\mathrm{ML}}^1) - \mathcal{L}_n^{k_0,l_0}(\hat{\theta}_{\mathrm{ML}}^0) = \mathcal{O}_P(1) \tag{23}$$

and thus as $(k_1 + l_1 - k_0 - l_0)c_n \to \infty$

$$P(\mathcal{S}_0 > \mathcal{S}_1) = P((k_1 + l_1 - k_0 - l_0)c_n > \mathcal{L}_n^{k_1,l_1}(\hat{\theta}_{\mathrm{ML}}^1) - \mathcal{L}_n^{k_0,l_0}(\hat{\theta}_{\mathrm{ML}}^0)) \to 1. \tag{24}$$

In case (iii) we have instead of (23)

$$\mathcal{L}_n^{k_0,l_0}(\hat{\theta}_{\mathrm{ML}}^0) - \mathcal{L}_n^{k_1,l_1}(\hat{\theta}_{\mathrm{ML}}^1) = \mathcal{L}_n^{k_0,l_0}(\theta_0^*) - \mathcal{L}_n^{k_1,l_1}(\theta_1^*) + \mathcal{O}_P(1). \tag{25}$$

Observe that in view of Thm 19.1.2 in Ibragimov and Linnik (1971) $n^{-1/2}(\mathcal{L}_n^{k_0,l_0}(\theta_0^*) - \mathcal{L}_n^{k_1,l_1}(\theta_1^*))$ is asymptotically normal as it can be rewritten as $n^{-1/2}\sum_{i=1}^n g(X_{i-1}, X_i)$, where $\mathbb{E}_P g((X_{i-1}, X_i)) = 0$, $\mathbb{E}_P \,|\, g((\tilde{X}_{i-1}, \tilde{X}_i))|^{2+\delta} < \infty$ and (X_{i-1}, X_i) is a geometrically ergodic Markov chain in view of Lemma 2. Thus, $n^{-1/2}(\mathcal{L}_n^{k_0,l_0}(\hat{\theta}_{\mathrm{ML}}^0) - \mathcal{L}_n^{k_1,l_1}(\hat{\theta}_{\mathrm{ML}}^1)) = \mathcal{O}_P(1)$ and

$$P(\mathcal{S}_0 > \mathcal{S}_1) = P(n^{-1/2}(k_1 + l_1 - k_0 - l_0)c_n > n^{-1/2}(\mathcal{L}_n^{k_0,l_0}(\hat{\theta}_{\mathrm{ML}}^0) - \mathcal{L}_n^{k_1,l_1}(\hat{\theta}_{\mathrm{ML}}^1))).$$

As $n^{-1/2}(k_1 + l_1 - k_0 - l_0)c_n \to \infty$ we obtain the conclusion. □

Remark 5. It is easy to generalize Theorem 2(i) to the case of a finite family of models and prove that if there exists the unique model such that the Kullback–Leibler distance from P to this model is the smallest one, then this model is chosen by the penalized ML method with probability tending to 1.

Remark 6. Observe that using the aforegiven methods the following two statements can be proved about asymptotic behaviour of the likelihood ratio (LR) statistic. Let $s_i = k_i + l_i$, $i = 0,1$.

(i) If (X_i) is geometrically ergodic, $f_{\theta_0^*}(x_2 \,|\, x_1)$ is not equal $f_{\theta_1^*}(x_2 \,|\, x_1)$ $P_{\tilde{X}_1,\tilde{X}_2}$-a.e. and $\mathbb{E}_P |\log_{\theta_i^*}(\tilde{X}_2 \,|\, \tilde{X}_1)|^{2+\delta} < \infty$ for $i = 0,1$ then

$$n^{-1/2}\left(\mathcal{L}_n^{k_0,l_0}(\hat{\theta}_{\mathrm{ML}}^0) - \mathcal{L}_n^{k_1,l_1}(\hat{\theta}_{\mathrm{ML}}^1) - n\mathbb{E}_P\left(\log\frac{f_{\theta_0^*}}{f_{\theta_1^*}}(\tilde{X}_2 \,|\, \tilde{X}_1)\right)\right) \xrightarrow{\mathcal{D}} \mathrm{N}(0, \sigma^2),$$

where $\sigma^2 = \mathrm{var}_P(\log\frac{f_{\theta_0^*}}{f_{\theta_1^*}}(\tilde{X}_2 \,|\, \tilde{X}_1)) + 2\sum_{i=2}^\infty \mathbb{E}(\log\frac{f_{\theta_0^*}}{f_{\theta_1^*}}(\tilde{X}_2 \,|\, \tilde{X}_1)\log\frac{f_{\theta_0^*}}{f_{\theta_1^*}}(\tilde{X}_{i+1} \,|\, \tilde{X}_i))$.

(ii) When $f_{\theta_0^*}(x_2 \,|\, x_1) = f_{\theta_1^*}(x_2 \,|\, x_1)$ $P_{\tilde{X}_1,\tilde{X}_2}$-a.e. and the remaining previous conditions are satisfied we have

$$2(\mathcal{L}_n(\hat{\theta}_{\mathrm{ML}}^0) - \mathcal{L}_n(\hat{\theta}_{\mathrm{ML}}^1)) \xrightarrow{\mathcal{D}} M_{s_0+s_1}(\lambda),$$

where $M_{s_0+s_1}(\lambda) = \sum_{i=1}^{s_0+s_1} \lambda_i Z_i^2$, where Z_i are i.i.d. N(0,1) r.vs and $\lambda_1,\ldots,\lambda_{s_0+s_1}$ are eigenvalues of the $(s_0 + s_1) \times (s_0 + s_1)$ matrix W

$$W = \begin{bmatrix} -\tilde{I}(\theta_0^*)A^{-1}(\theta_0^*) & -B(\theta_0^*, \theta_1^*)A^{-1}(\theta_1^*) \\ B(\theta_1^*, \theta_0^*)A^{-1}(\theta_0^*) & \tilde{I}(\theta_1^*)A^{-1}(\theta_1^*) \end{bmatrix},$$

where $A(\theta)$ is defined in Proposition 6 and $\tilde{I}(\theta_i^*), i = 0,1$, equals

$$\mathbb{E}\left(\frac{\partial}{\partial\theta_i}\log f_{\theta_i}(\tilde{X}_2 \,|\, \tilde{X}_1)_{|\theta_i=\theta_i^*} \cdot \left(\frac{\partial}{\partial\theta_i}\log f_{\theta_i}(\tilde{X}_2 \,|\, \tilde{X}_1)_{|\theta_i=\theta_i^*}\right)'\right)$$
$$+ 2\sum_{j=2}^\infty E\left(\frac{\partial}{\partial\theta_i}\log f_{\theta_i}(\tilde{X}_2 \,|\, \tilde{X}_1)_{|\theta_i=\theta_i^*} \cdot \left(\frac{\partial}{\partial\theta_i}\log f_{\theta_i}(\tilde{X}_{j+1} \,|\, \tilde{X}_j)_{|\theta_i=\theta_i^*}\right)'\right)$$

and $B(\theta_0^*, \theta_1^*)$ equals

$$\mathbb{E}\left(\frac{\partial}{\partial\theta_0}\log f_{\theta_0}(\tilde{X}_2 \,|\, \tilde{X}_1)_{|\theta_0=\theta_0^*} \cdot \left(\frac{\partial}{\partial\theta_1}\log f_{\theta_1}(\tilde{X}_2 \,|\, \tilde{X}_1)_{|\theta_1=\theta_1^*}\right)'\right)$$
$$+ 2\sum_{j=2}^\infty \mathbb{E}\left(\frac{\partial}{\partial\theta_0}\log f_{\theta_0}(\tilde{X}_2 \,|\, \tilde{X}_1)_{|\theta_0=\theta_0^*} \cdot \left(\frac{\partial}{\partial\theta_1}\log f_{\theta_1}(\tilde{X}_{j+1} \,|\, \tilde{X}_j)_{|\theta_1=\theta_1^*}\right)'\right)$$

and $f_{\theta_0}(f_{\theta_1})$ denotes the conditional density in $\mathcal{F}_{k_0,l_0}(\mathcal{F}_{k_1,l_1})$ respectively. $B(\theta_1^*, \theta_0^*)$ is defined analogously. Note that we do not necessarily have that $B(\theta_1^*, \theta_0^*) = B(\theta_0^*, \theta_1^*)'$.

The proof of both statements follows the proof of Thm 3.3 in Vuong (1989) with one essential difference: to prove that

$$\sqrt{n}((\hat{\theta}^0_{n,\mathrm{ML}} - \theta^*_0)', (\hat{\theta}^1_{n,\mathrm{ML}} - \theta^*_1)')' \xrightarrow{\mathcal{D}} \mathrm{N}(0, \Sigma),$$

where

$$\Sigma = \begin{bmatrix} A^{-1}(\theta^*_0)\tilde{I}(\theta^*_0)A^{-1}(\theta^*_0) & A^{-1}(\theta^*_0)B(\theta^*_0, \theta^*_1)A^{-1}(\theta^*_1) \\ A^{-1}(\theta^*_1)B(\theta^*_1, \theta^*_0)A^{-1}(\theta^*_0) & A^{-1}(\theta^*_1)\,\tilde{I}(\theta^*_1)A^{-1}(\theta^*_1) \end{bmatrix},$$

we use CLT for geometrically ergodic Markov chains (cf. Ibragimov and Linnik, 1971) and Cramér–Wold approach. Thus, analogously as in the case of i.i.d. data we have that behaviour of LR statistic is different in cases (i) and (ii).

Remark 7. When the Markov chain (X_i) is uniformly geometrically ergodic which means that the constant $M(x_0)$ in (18) does not depend on x_0, then $2 + \delta$ integrability can be weakened to square integrability in all the aforegiven results owing to recent CLT by Bednorz *et al.* (2008).

3.3. Pseudolikelihood method

Consider a different objective function from (5) defined as follows (cf. Gourieroux *et al.*, 1984):

$$\mathcal{N}_n(\theta) = \sum_{i=1}^{n-1} \log\left(\sigma_\theta(X_i)^{-1} f_{\mathrm{N}(0,1)}\left(\frac{X_{i+1} - m_\theta(X_i)}{\sigma_\theta(X_i)}\right)\right) =: \sum_{i=1}^{n-1} l_\theta(X_{i+1} \mid X_i),$$

i.e. we replace the density f_ε by the standard normal density $f_{\mathrm{N}(0,1)}$ irrespective of whether the errors are normally distributed or not. This has an obvious advantage that the density of errors does not need to be known. The estimator $\tilde{\theta} = \mathrm{argmax}_{\theta \in \Theta_{kl}} \mathcal{N}_n(\theta)$ will be called the maximum pseudolikelihood estimator of θ in the model $\mathcal{F}_{kl}$. It turns out that the results of the article are also valid when $\mathcal{N}_n(\theta)$ is used instead of $\mathcal{L}_n(\theta)$ and $\log f_\theta(x_2|x_1)$ is replaced by $l_\theta(x_2|x_1)$ in assumptions C3–C6. The statement follows from the observation that the proofs of the results relied on pivotal Propositions 3 and 6 and their analogues can be proved in the case of l_θ. In particular, $\tilde{S}_n(\theta_0) = \sum_{t=1}^{n-1} \frac{\partial}{\partial\theta} l_\theta(X_{t+1} \mid X_t)_{|\theta=\theta_0}$ is a martingale. Indeed, provided that $P = P_{\theta_0}$ it is easy to see that

$$\mathbb{E}_{P_{\theta_0}}\left(\frac{\partial}{\partial\beta_k} l_\theta(X_t \mid X_{t-1})_{|\theta=\theta_0} | \mathcal{G}_{t-1}\right) = \mathbb{E}_{P_{\theta_0}}\left(\frac{(X_t - m_{\theta_0}(X_{t-1}))}{\sigma^2_{\theta_0}(X_{t-1})}\phi_k(X_{t-1}) | \mathcal{G}_{t-1}\right) = 0$$

and

$$\mathbb{E}_{P_{\theta_0}}\left(\frac{\partial}{\partial\eta_l} l_\theta(X_t \mid X_{t-1})_{|\theta=\theta_0} | \mathcal{G}_{t-1}\right) = \mathbb{E}_{P_{\theta_0}}\left(-\psi_l(X_{t-1}) + \frac{(X_t - m_{\theta_0}(X_{t-1}))^2}{\sigma^2_{\theta_0}(X_{t-1})}\psi_l(X_{t-1}) | \mathcal{G}_{t-1}\right) = 0.$$

Moreover, we have Proposition 7.

Proposition 7. *Under assumptions of Proposition* 3 $\tilde{L}(\theta) = \mathbb{E}_{P_{\theta_0}} l_\theta(\tilde{X}_2 \mid \tilde{X}_1)$ *attains the unique maximum at* θ_0.

Proof Let $f_{N_\theta}(\cdot \mid x_1)$ denote the normal density with the mean $m_\theta(x_1)$ and the variance $\sigma^2_\theta(x_1)$. It is easy to see that $T(f) = \int \log f_{N_\theta}(x_2|x_1) f(x_2)\mathrm{d}x_2$ depends only on the first two moments of the density f. Thus in view of the information inequality as the first two moments of $f_{N_\theta}(\cdot|x_1)$ and $f_\theta(\cdot|x_1)$ coincide, we have

$$\begin{aligned} \int \log f_{N_\theta}(x_2 \mid x_1) f_{\theta_0}(x_2 \mid x_1)\mathrm{d}x_2 = \int \log f_{N_\theta}(x_2 \mid x_1) f_{N_{\theta_0}}(x_2 \mid x_1)\mathrm{d}x_2 \le \\ \int \log f_{N_{\theta_0}}(x_2 \mid x_1) f_{N_{\theta_0}}(x_2 \mid x_1)\mathrm{d}x_2 = \int \log f_{N_{\theta_0}}(x_2 \mid x_1) f_{\theta_0}(x_2 \mid x_1)\mathrm{d}x_2 \end{aligned} \tag{26}$$

and the equality holds only when $m_{\theta_0}(x_1) = m_\theta(x_1)$ and $\sigma^2_{\theta_0}(x_1) = \sigma^2_\theta(x_1)$. Thus integrating this inequality with respect to $\pi(\mathrm{d}x_1)$ in view of C2 and C8 we obtain the conclusion. The analogue of Theorem 1 when the pseudolikelihood is considered instead of the likelihood can be interpreted as follows. Even if the density of errors is mis-specified and erroneously assumed to be the normal density but at least one of the finite collection of models contains conditional density with correctly specified regression and variance functions, under appropriate assumption the smallest model with this property will be identified with probability tending to 1. □

4. SIMULATION STUDY

We first discuss a parameterization of $f(x_2|x_1)$ on a fixed compact interval $[a,b]$. Consider an equipartition of $[a,b]$ into p intervals, where $1 \le p \le P$ and let $\phi_{i,0}(x),\ldots,\phi_{i,s-1}(x)$ denote the Legendre polynomials of order $0,1,\ldots,s-1$ respectively, transformed to the ith

interval of the partition. Moreover, $b_i = (\beta_{i,0},\ldots,\beta_{i,s-1})'$ is a corresponding vector of coefficients. For a given p and s, regression m is modelled with the use of $k = sp$ parameters by

$$m_{sp}(x) = \sum_{i=1}^{p}\sum_{k=0}^{s-1}\beta_{i,k}\phi_{i,k}(x).$$

Thus, the independent functions $\phi_i(\cdot)$ are defined in our study as piecewise Legendre polynomials coinciding with some polynomial $\phi_{i,k}(\cdot)$ with $k \le s-1$ on the ith interval of the partition. The variance function is parameterized in an analogous manner. Namely, for a given order t and $q \le Q$ consider partition of $[a,b]$ into q intervals and $\psi_{j,0}(x),\ldots,\psi_{j,t-1}(x)$ are the Legendre polynomials up to the order $t-1$ transformed to the jth interval of the partition. Denote by $e_j = (\eta_{j,0},\ldots,\eta_{j,t-1})$ a pertaining vector of coefficients belonging to $\mathbb{R}^t$. Then we define a parametric form of a standard deviation based on $l = tq$ parameters

$$\sigma_{tq}(x) = \exp\left\{\sum_{j=1}^{q}\sum_{k=0}^{t-1}\eta_{j,k}\psi_{j,k}(x)\right\}.$$

A parameter $\theta = (b_1',\ldots,b_p',e_1',\ldots,e_q')'$ is assumed to belong to a compact subset $\Theta_{kl} \in \mathbb{R}^{k+l}$. For a fixed (k,l) we consider a parametric model $\mathcal{F}_{kl}$ of conditional densities defined as [cf. (3)]

$$f_\theta(x_2 \mid x_1) = \frac{1}{\sigma_{tq}(x_1)} f_\varepsilon\left(\frac{x_2 - m_{sp}(x_1)}{\sigma_{tq}(x_1)}\right).$$

Observe that for $p = q$, $s = t = 1$ and partition intervals coinciding with A_j in (2) we obtain QTARCH(1) model. A collection $\mathcal{F}_{kl}, k \le K = sP, l \le L = tQ$ is a family of available models from which a model yielding a parsimonious fit to the data is chosen. In the first group of examples ε_t has a standard normal distribution. The conditional log-likelihood for the model $\mathcal{F}_{kl}$ has the form

$$\mathcal{L}_n(X_1,X_2,\ldots,X_n) = -\frac{n-1}{2}\log 2\pi - \sum_{i=1}^{n-1}\log\sigma_{tq}(X_i) - \frac{(X_{i+1} - m_{sp}(X_i))^2}{2\sigma_{tq}^2(X_i)} \tag{27}$$

and the penalized criterion is

$$2\mathcal{L}_n(X_1,X_2,\ldots,X_n) - c_n \times (sp + tq).$$

In the following we take $[a,b] = [X_{1:n},X_{n:n}]$; i.e. we construct models on the range of available data $X_1,X_2,\ldots,X_n$. Note that a data-dependent choice of $[a,b]$ is not covered by our theoretical results. However, as supports of stationary densities are not known and they significantly vary in the considered examples, this approach allows comparing empirical findings from different models objectively. Considered sample was size $n = 500$ and the number of repetitions was $k = 5000$. Samples were generated starting from $x_0 = 0$. Moreover, we took $P = 10$, $Q = 15$, $s = 2$ and $t = 1$, i.e. the regression was modelled by a piecewise linear and the variance by a piecewise constant function. Different pairs (s,t) were considered before choosing $(s,t) = (2,1)$. The pair $(s,t) = (3,1)$ performed similar to $(2,1)$, whereas $(s,t) = (2,2)$ performed worse. As a main selection rule we considered BIC with $c_n = \log(n-1)$. PMS estimator of variance with $(s,t) = (2,1)$ and $c_n = \log(n-1)$ will be called $S1$. We also considered its modification, called $S1m$ which is a polygon joining the values of the constructed histogram $S1$ at midpoints of bins. Moreover, the following estimator $S2$ taking into account optimal models for different penalties was considered. First training sample consisting of 66% of the original sample was considered and for each $c_n = 2,3,\ldots,[\log(n-1)]$ the most parsimonious model for the penalty equal to c_n was chosen based on the training sample. Then the final model is the one for which the pertaining estimator has the largest value of log-likelihood for the remaining part of the observations. The estimator $S2$ is the estimator corresponding to the final chosen model recalculated on the whole sample. However, this estimator performed on average worse than $S1m$, possibly as a result of sample splitting, and it was discarded from further considerations.

For comparison, we considered two two-stage non-parametric estimators of the variance, which use preliminary non-parametric regression estimator to calculate the squared residuals, which in their turn, are used to estimate the variance. The first method, which is called $LL1$ for further reference, uses local linear smoother at both the stages and respective bandwidths are calculated using *dpill* method proposed by Ruppert *et al.* (1995) and implemented in R package *kernsmooth*. The second, $LL2$, differs from the first only in bandwidth choice and is based on the proposal of Fan and Yao (1998). Their open source code is used to calculate $LL2$. We refer to Borkowski and Mielniczuk (2008) for simulation study of properties of two-stage estimators and comparison of their performance in the autoregressive case. In general, $LL2$ works better for pronouncedly variable variances, whereas $LL1$ is superior for slowly varying ones.

4.1. Parametric models

The following regressions and conditional standard deviations have been considered:

(a) $m_1(x) = 0.8x$;
(b) $m_2(x) = 0.8xI\{x > 0\} - 0.3xI\{x \le 0\}$;

(i) $\sigma_1(x) = 0.5$;
(ii) $\sigma_2(x) = 0.4I_{(-\infty,-0.5)} + 0.8I_{[-0.5,0.5)} + 0.6I_{[0.5,\infty)}$;

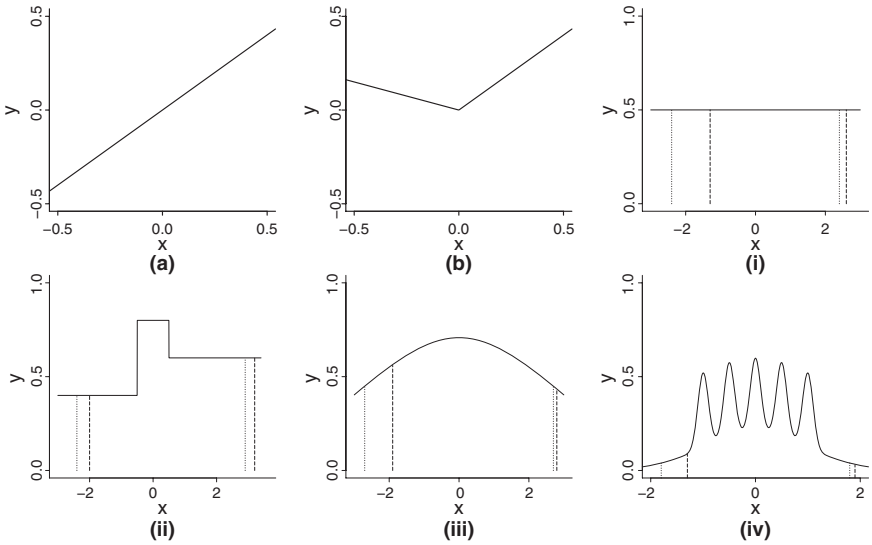

Figure 1. (a, b) Regressions $m_1 - m_2$; (i)–(iv) Standard deviations $\sigma_1 - \sigma_4$

(iii) $\sigma_3(x) = 0.75 \exp(-x^2/8)$;
(iv) $\sigma_4(x) = \frac{1}{2}\mathrm{N}(0,1) + \sum_{l=0}^{4} \frac{1}{10}\mathrm{N}(l/2 - 1, (\frac{1}{10})^2)$, where N(0,1) is the standard Gaussian density.

Plots of considered standard deviations are given in Figure 1. In the last three examples, the maximal value of the variance is 0.64, 0.56 and 0.36 respectively. Moreover, Fig 1 to check performance of the pseudolikelihood method for heavy-tailed data, we considered errors having Laplace distribution with $\lambda = 1/\sqrt{2}$ instead of normal N(0,1) errors, and, to allow for even slower decrease of tails, errors having Student distribution $t(6)$ with six degrees of freedom.

All eight pairs $(m_i,\sigma_j), i = 1,2, j = 1,\ldots,4$ have been studied using as a measure of performance of variance estimators empirical ISE (EISE) defined as an average of squared differences between variance estimators and target values over all observations:

$$\mathrm{EISE} = \frac{1}{n-1}\sum_{i=2}^{n}(\tilde{\sigma}^2(X_i) - \sigma^2(X_i))^2,$$

where $\tilde{\sigma}^2(\cdot)$ is any of the considered variance estimators. Sample means of EISE (MEISE) over all simulation, i.e. $\mathrm{MEISE} = k^{-1}\sum_{i=1}^{k}\mathrm{EISE}_i$ together with its standard errors equal to $k^{-1/2}$ times the sample standard deviations of $\{\mathrm{EISE}_i\}_{i=1}^{k}$ were calculated based on $k = 5000$ simulations. All reported values are multiplied by 10^4. We stress that EISE is a measure of *overall* performance of an variance estimator. An L^1 analogue of EISE with its summands replaced by $|\tilde{\sigma}^2(X_i) - \sigma^2(X_i)|$ was also considered and its relative performance for the studied models turned out to be similar.

4.2. Optimization method

Levenberg–Marquardt (cf., e.g. Press *et al.*, 1986) version of the Gauss–Newton algorithm was used to find a maximum of (27) for given p and q. Usual version of the Gauss–Newton algorithm which does not use damping factors frequently achieved maximal allowed number of iterations $N_{\max} = 60$, especially for the most variable variance $\sigma_4^2(\cdot)$. Note that for a given sample the model is chosen among $10 \times 15 = 150$ candidate models. Incorporating damping factors in the algorithm has substantial positive effect on convergence of iteration scheme – the maximal number of models for which $N_{\max}$ was attained was 2 (out of 150) in the case of (m_2,σ_4) autoregression. Also, $N_{\max}$ was attained at least once in at most 30 repetitions (out of 5000). For cases when non-convergence occurs, the modification of the selection procedure was used consisting of rejecting such models and selecting the winner only from among those models, for which iterations approximating ML estimator converged.

4.3. Results

The results given in Tables 1 and 2, pertain respectively to normal errors and errors having Laplace distribution. It is seen that the form of regression function does not have much influence on variance estimation for all methods considered. For normal errors, *S*1 and *S*1*m* perform much better than *LL*1 and *LL*2 in the case of constant standard deviation σ_1 and the most variable σ_4, are worse for σ_3 and they are comparable with local linear smoothers in the case of σ_2. For σ_1 MEISE of the proposed estimators was more than three times smaller in the case of m_1 than for linear smoothers. The property that PMS estimators work very well for the least and the most variable standard deviation considered is in contrast to the behaviour of local linear estimators which work well either for variable standard deviations (*LL*2) or slowly varying ones (*LL*1). Modified estimator *S*1*m* exhibits significantly improved performance in comparison with *S*1 apart from the case of the piecewise constant standard deviation σ_2 which is understandable in view of the nature of modification.

Table 1. Means of integrated squared error of variance estimators and their standard errors; the case of normal distribution

	*S*1	*S*1*m*	LL1	*LL*2
m_1				
σ_1	2.87 (0.07)	2.69 (0.05)	9.33 (0.14)	17.45 (0.13)
σ_2	134.35 (0.75)	166.26 (0.88)	140.78 (0.65)	116.51 (0.53)
σ_3	88.48 (0.57)	60.85 (0.73)	26.1 (0.27)	46.11 (0.4)
σ_4	60.28 (0.28)	61.9 (0.3)	99.74 (0.14)	91.52 (0.32)
m_2				
σ_1	3.03 (0.07)	2.77 (0.06)	10.62 (0.27)	16.39 (0.12)
σ_2	139.04 (0.88)	146.22 (0.88)	102.4 (0.8)	96.28 (0.44)
σ_3	81.75 (0.52)	53.09 (0.58)	29.85 (0.33)	42.62 (0.39)
σ_4	53.9 (0.22)	53.12 (0.21)	101.18 (0.22)	93.05 (0.31)

Table 2. Means of integrated squared error of variance estimators and their standard errors; the case of Laplace distribution

	*S*1	*S*1*m*	*LL*1	*LL*2
m_1				
σ_1	19.11 (0.53)	13.15 (0.38)	30.36 (4.11)	31.74 (0.34)
σ_2	200.69 (1.65)	195 (1.23)	213.99 (8.25)	192.27 (1)
σ_3	138.05 (1.37)	81.85 (0.94)	81.88 (9.01)	62.17 (0.64)
σ_4	93.37 (0.42)	94.69 (0.4)	112.8 (2.04)	104.33 (0.18)
m_2				
σ_1	19.88 (0.55)	13.21 (0.37)	31.82 (1.72)	28.52 (0.32)
σ_2	208.47 (1.66)	182.09 (1.24)	167.63 (5.04)	172.86 (10.03)
σ_3	138.37 (1.51)	83.28 (1.06)	103.09 (9.27)	63.16 (1.16)
σ_4	83.79 (0.46)	84.57 (0.44)	120.9 (2.36)	127.92 (11.26)

Table 3. Means of integrated squared error of variance estimators and their standard errors; the case of Student(6) distribution

	*S*1	*S*1*m*	*LL*1
m_1			
σ_1	17.85 (1.03)	12.92 (0.81)	28.19 (3.03)
σ_2	189.4 (3.94)	193.61 (2.27)	222.32 (9.17)
σ_3	133.11 (1.79)	84.33 (1.29)	102.62 (13.63)
σ_4	87.1 (0.56)	89.49 (0.54)	120.28 (3.95)
m_2			
σ_1	20.43 (2.04)	14.63 (1.48)	39.12 (3.48)
σ_2	204.4 (4.74)	186.54 (2.43)	204.32 (18.06)
σ_3	137.26 (3.07)	88.31 (2.23)	128.6 (15.14)
σ_4	79.52 (0.69)	80.22 (0.54)	131.29 (7.2)

For the errors having Laplace distribution all considered estimators performed worse than in a normal case as a result of relatively larger magnitude of errors and much pronounced sparsity of marginal distribution in the tails. Remarkably, estimator *S*1*m* worked better overall than *S*1 and both of them were superior to the local smoothers for the standard deviations σ_1 and σ_4. For the remaining cases *S*1*m* performed comparably to both local linear estimators except the cases pertaining to σ_3 in which it performed worse than LL2 (in the case of both regressions considered).

For errors having Student $t(6)$ distribution *LL*2 method ceases to be reliable and Table 3 presents results for the remaining estimators only. In this case one of the PMS estimators performs best in all cases. For example, ratio of MEISE (*LL*1)/MEISE (*S*1*m*) is greater than 2 for models pertaining to (m_1,σ_1) and (m_2,σ_1). Moreover, it performed much more stably, e.g. the standard error of *S*1*m* is 10 times smaller than that of *LL*1 for (m_1,σ_3). It seems that this estimator is worth considering especially when, as is typical for financial data, the occurrence of heavy-tailed errors is likely.

The performance of the proposed PMS estimators of variance is affected by the choice of an initial interval on which partitions are built, especially when the variance has a form specified by one of the parametric models as in the case of σ_2 but the pertaining points of discontinuities are far from those determined by the models.

Acknowledgements

Comments from two anonymous referees which helped in improving the original version of the article are appreciated.

REFERENCES

Bednorz, W., Latuszyński, K. and Latala, R. (2008) A regeneration proof of the central limit theorem for uniformly ergodic Markov chains. *Electronic Communications in Probability* **13**, 85–98.
Borkowski, P. and Mielniczuk, J. (2008) Comparison of performance of variance function estimators for autoregressive time series, submitted.

Breiman, L. (1992) *Probability*. Philadelphia: SIAM.
Chen, R. and Tsay, R. (1993) Functional coefficient autoregressive models. *Journal of the American Statistical Association* **88**, 298–308.
Devroye, L. (1987) *A Course in Density Estimation*. Boston: Birkhäuser.
Diaconis, P. and Freedman, D. (1999) Iterated random functions. *SIAM Review* **41**, 45–76.
Fan, J. (2005) A selective overview of nonparametric methods in financial econometrics. *Statistical Science* **20**, 317–37.
Fan, J. and Yao, Q. (1998) Efficient estimation of conditional variance functions in stochastic regression. *Biometrika* **85**, 645–60.
Gourieroux, C. and Monfort, A. (1992) Qualitative threshold ARCH models. *Journal of Econometrics* **52**, 159–99.
Gourieroux, C., Monfort, A. and Trognon, A. (1984) Pseudo maximum likelihood methods: theory. *Econometrica* **52**, 681–700.
Ibragimov, I. A. and Linnik, Y. V. (1971) *Independent and Stationary Sequences of Random Variables*. Groningen: Wolters-Noordhof.
Jennrich, R. I. (1969) Asymptotic properties of non-linear least squares estimators. *The Annals of Mathematical Statistics* **40**, 633–43.
Ledwina, T. and Mielniczuk, J. (2007) Variance function estimation via model selection, submitted.
Leeb, H. and Pötscher, B. M. (2008) Model selection. In: *Handbook of Financial Time Series* (eds T. G. Andersen, R. A. Davis, J. P. Kreiss and T. Mikosch). Berlin: Springer-Verlag.
McKeague, I. and Zhang, M. (1994) Identification of nonlinear time series from first order cumulative characteristic. *The Annals of Statistics* **26**, 1570–613.
Meyn, S. P. and Tweedie, R. L. (1993) *Markov Chains and Stochastic Stability*. London: Springer.
Nummelin, E. (1984) *General Irreducible Markov Chains and Non-Negative Operators*. Cambridge: Cambridge University Press.
Phillips, P. and Yu, J. (2005) Comment: a selective overview of nonparametric methods in financial econometrics. *Statistical Science* **20**, 338–43.
Press, W., Teukolsky, S., Vetterling, W. and Flannery, B. (1986) *Numerical Recipes: The Art of Scientific Computing*. Cambridge: Cambridge University Press.
Ruppert, D., Sheather, S. J. and Wand, M. P. (1995) An effective bandwidth selector for local least squares regression. *Journal of the American Statistical Association* **90**, 1257–70.
Sin, C. Y. and White, H. (1996) Information criteria for selecting possibly misspecified parametric models. *Journal of Econometrics* **71**, 207–25.
Sørensen, H. (2004) Parametric inference for diffusion processes observed at discrete points in time: a survey. *International Statistical Review* **72**, 2337–54.
Taniguchi, M. and Kakizawa, Y. (2000) *Asymptotic Theory of Statistical Inference for Time Series*. New York: Springer.
Tjøstheim, D. (1986) Estimation of non-linear time series models. *Stochastic Processes and their Applications* **21**, 251–73.
Vuong, Q. H. (1989) Likelihood ratio tests for model selection and non-nested hypothesis. *Econometrica* **57**, 307–33.
White, H. (1982) Maximum likelihood estimation of misspecified models. *Econometrica* **50**, 1–25.
Wu, W. B. and Shao, X. (2004) Limit theorems for iterated random functions. *Journal of Applied Probability* **41**, 425–36.
Yu, K. and Jones, M. C. (2004) Likelihood-based local linear estimation of the conditional variance function. *Journal of the American Statistical Association* **99**, 139–44.

