The Annals of Statistics
2019, Vol. 47, No. 4, 2145–2173
https://doi.org/10.1214/18-AOS1743

CROSS VALIDATION FOR LOCALLY STATIONARY PROCESSES[1]

BY STEFAN RICHTER AND RAINER DAHLHAUS

Heidelberg University

We propose an adaptive bandwidth selector via cross validation for local M-estimators in locally stationary processes. We prove asymptotic optimality of the procedure under mild conditions on the underlying parameter curves. The results are applicable to a wide range of locally stationary processes such linear and nonlinear processes. A simulation study shows that the method works fairly well also in misspecified situations.

1. Introduction. Inference for locally stationary time series models is strongly connected to the estimation of parameter curves which determine the degree of nonstationarity. The estimation of these curves was discussed for several specific models such as tvARMA processes [5], the tvARCH and tvGARCH processes [4, 7, 8] and time-varying random coefficient models [16]. Of interest is also a time-varying TAR process which was considered in [18].

Local estimators such as kernel estimators require the selection of a bandwidth. Unlike nonparametric regression, there exist only very few theoretical results about adaptivity for locally stationary processes. We mention [14] who discussed adaptive covariance estimation for a general class of locally stationary processes. Other results are constructed for specific models and are partly dependent on further tuning parameters: Giraud, Roueff and Sanchez-Perez [9] discussed online-adaptive forecasting of tvAR processes and [1, 2] proposed methods for sequential and minimax-optimal bandwidth selection for tvAR processes of order 1.

In this paper, we treat the problem for arbitrary locally stationary time series models determined by a time varying parameter curve. We focus on local M-estimators and use the functional dependence measure introduced in [17] to formulate mixing conditions. We propose an adaptive bandwidth selection procedure inspired by cross validation in the i.i.d. regression model which does not need any tuning parameters. We discuss the theoretic behavior by proving asymptotic optimality of the selector (similar to [12] where nonparametric regression has been treated). We also prove convergence toward the deterministic asymptotic optimal bandwidth.

Received May 2017; revised February 2018.

[1]Supported by Deutsche Forschungsgemeinschaft through the Research Training Group RTG 1653.

MSC2010 subject classifications. Primary 62M10; secondary 62G20.

Key words and phrases. Locally stationary processes, cross validation, adaptive bandwidth selection, asymptotic optimality.

The technical core of the paper is martingale theory applied in particular to the score function of the objective function and several bounds for moments of quadratic and cubic forms of locally stationary processes which are needed to provide convergence of expansions of the estimation error with suitable rates.

In Section 2, we introduce the locally stationary time series model and formalize the separation of the process into a parametric stationary process and unknown parameter curves. We define local M-estimators and the cross validation procedure. We introduce a Kullback–Leibler-type distance measure which can be seen as an analogue to the averaged squared error in nonparametric regression.

In Section 3, we prove asymptotic optimality of the cross-validation procedure with respect to the Kullback–Leibler-type distance measure and convergence of the cross-validation bandwidth toward the deterministic asymptotic optimal bandwidth. Furthermore, we derive the limit distribution of the bandwidth chosen by cross validation. The assumptions are stated in terms of a parametric stationary time series model which is connected to the locally stationary process. This allows for easy verification since most of the conditions are standard in M-estimation theory and were already shown for specific stationary models.

In Section 4, we discuss some processes where the main results are applicable. The performance of the method for different models such as tvAR, tvARCH and tvMA is studied in simulations.

In Section 5, a short conclusion is drawn. Many lemmata used in the proofs are deferred without further reference to the Supplementary Material [15].

2. A cross-validation method for locally stationary processes.

2.1. *The model.* In this paper, we discuss adaptive estimation of a multidimensional parameter curve $\theta_0 : [0,1] \to \Theta \subset \mathbb{R}^p$, that is, we restrict to locally stationary processes $X_{t,n}$, $t = 1, \dots, n$ parameterized by curves. As usual, we are working in the infill asymptotic framework with rescaled time $t/n \in [0,1]$, where n denotes the number of observations.

Following the original idea of locally stationary processes, for fixed $u \in [0,1]$, $X_{t,n}$ should locally (i.e., for $|u - \frac{t}{n}| \ll 1$) behave like a stationary process $\hat{X}_t(u)$. In this paper, we assume that the time dependence of the approximation $\hat{X}_t(u)$ is solely described by θ_0, that is, $\hat{X}_t(u) = \tilde{X}_t(\theta_0(u))$, where $\tilde{X}_t(\theta)$, $\theta \in \Theta$ is some family of parametric stationary processes. We formulate the assumptions in terms of $\tilde{X}_t(\theta)$ instead of $\hat{X}_t(u)$ leading to a clear separation between the properties of the model class and the smoothness assumptions on θ_0. We formalize this by the following.

ASSUMPTION 2.1 (Locally stationary time series model). Let $q \geq 1$ and $\|W\|_q := (\mathbb{E}|W|^q)^{1/q}$. Let $X_{t,n}$, $t = 1, \dots, n$ be a triangular array of observations.

Suppose that for each $\theta \in \Theta$, there exists a stationary process $\tilde{X}_t(\theta)$, $t \in \mathbb{Z}$ such that for all $q \geq 1$, uniformly in $\theta, \theta' \in \Theta$,

$$(1) \quad \|\tilde{X}_t(\theta) - \tilde{X}_t(\theta')\|_q \leq C_A|\theta - \theta'|_1, \qquad \sum_{t=1}^{n} \left\| X_{t,n} - \tilde{X}_t\left(\theta_0\left(\frac{t}{n}\right)\right)\right\|_q \leq C_B,$$

with some $C_A = C_A(q), C_B = C_B(q) \geq 0$, and

$$D_q := \max\left\{\sup_{\theta\in\Theta}\|\tilde{X}_0(\theta)\|_q, \sup_{n\in\mathbb{N}}\sup_{t=1,\ldots,n}\|X_{t,n}\|_q\right\} < \infty.$$

REMARK 2.2. (i) We conjecture that the assumption on the existence of all moments of $X_{t,n}$ and $\tilde{X}_t(\theta)$ can be dropped but the calculations would be very tedious without much additional insight. The number of moments needed for the proofs increases if the Hoelder exponent of the unknown parameter curve decreases.

(ii) In many models, the second condition in (1) basically means that the unknown parameter curve θ_0 has bounded variation; see also Assumption 3.3.

We first give some examples which are covered by our results. These include in particular several classical parametric time series models where the constant parameters have been replaced by time-dependent parameter curves. Let ε_t, $t \in \mathbb{Z}$ be an i.i.d. sequence with mean zero.

EXAMPLE 2.3. (i) The tvARMA(r, s) process: Given parameter curves $a_i, b_j, \sigma : [0, 1] \to \mathbb{R}$ $(i = 0, \ldots, r,\ j = 0, \ldots, s)$ with $a_0(\cdot), b_0(\cdot) = 1$,

$$\sum_{i=0}^{r} a_i\left(\frac{t}{n}\right)X_{t-i,n} = \sum_{j=0}^{s} b_j\left(\frac{t}{n}\right)\sigma\left(\frac{t-j}{n}\right)\varepsilon_{t-j}.$$

(ii) The tvARCH(r) process (cf. [7]): Given parameter curves $a_i : [0, 1] \to \mathbb{R}$ $(i = 0, \ldots, r)$,

$$X_{t,n} = \left(a_0\left(\frac{t}{n}\right) + a_1\left(\frac{t}{n}\right)X_{t-1,n}^2 + \cdots + a_r\left(\frac{t}{n}\right)X_{t-r,n}^2\right)^{1/2}\varepsilon_t.$$

(iii) The tvTAR(1) process (cf. [18]): Given parameter curves $a_1, a_2 : [0, 1] \to \mathbb{R}$, define

$$X_{t,n} = a_1\left(\frac{t}{n}\right)X_{t-1,n}^{+} + a_2\left(\frac{t}{n}\right)X_{t-1,n}^{-} + \varepsilon_t,$$

where $x^+ := \max\{x, 0\}$ and $x^- := \max\{-x, 0\}$.

As an estimator of $\theta_0(\cdot)$ we consider local likelihood (or local M-) estimators weighted by kernels, that is,

$$\hat{\theta}_h(u) := \operatorname*{argmin}_{\theta\in\Theta} L_{n,h}(u,\theta). \tag{2}$$

where

$$L_{n,h}(u,\theta) := \frac{1}{n}\sum_{t=1}^{n} K_h\left(\frac{t}{n}-u\right)\ell_{t,n}(\theta) \tag{3}$$

and $\ell_{t,n}(\theta) := \ell(X_{t,n}, Y^c_{t-1,n}, \theta)$ with $Y^c_{t-1,n} := (X_{t-1,n}, \ldots, X_{1,n}, 0, 0, \ldots)$ consisting of the observed past, where ℓ is a given objective function [localized in $L_{n,h}(u,\theta)$ by the kernel K]. $K : \mathbb{R} \to \mathbb{R}$ is nonnegative with $\int K = 1$, and $h \in (0,\infty)$ is the bandwidth. For shortening the notation, we used $K_h(\cdot) := \frac{1}{h}K(\frac{\cdot}{h})$. In practice, ℓ is often chosen to be the negative logarithm of the infinite past likelihood of $X_{t,n}$ given $Y_{t-1,n} := (X_{s,n} : s \le t-1)$,

$$\ell(x,y,\theta) = -\log p_\theta(X_{t,n} = x | Y_{t-1,n} = y), \tag{4}$$

assuming that $\theta_0(\cdot) = \theta \in \Theta$. In this paper, we allow for general objective functions ℓ which have to obey some smoothness conditions (see Assumption 3.3).

2.2. *Distance measures.* Define $\tilde{Y}_t(\theta) := (\tilde{X}_s(\theta) : s \le t)$. In the following, we will use ∇ to denote the derivative with respect to $\theta \in \Theta$, and x' denotes the transpose of a vector or matrix x. As global distance measures, we use the (infeasible) averaged and the integrated squared error (ASE/ISE) weighted by the Fisher information

$$I(\theta) := \mathbb{E}\big[\nabla\ell\big(\tilde{Y}_0(a),\theta\big)\cdot\nabla\ell\big(\tilde{Y}_0(a),\theta\big)'\big]|_{a=\theta}. \tag{5}$$

and the possibly misspecified Fisher information $V(\theta) := \mathbb{E}\nabla^2\ell(\tilde{Y}_0(a),\theta)|_{a=\theta}$ of the corresponding stationary approximation. In addition, the weight function $w(\cdot) := \mathbb{1}_{[\gamma,1-\gamma]}(\cdot)$ with some $\gamma > 0$ is needed to exclude boundary effects. Since the proof is the same for other weights $w(\cdot)$, we allow in Assumption 3.4 for more general weights.

More precisely, we set (with $|x|_A^2 := x'Ax$ for $x \in \mathbb{R}^p$ and $A \in \mathbb{R}^{p\times p}$),

$$d_A(\hat{\theta}_h,\theta_0) := \frac{1}{n}\sum_{t=1}^{n}\left|\hat{\theta}_h\left(\frac{t}{n}\right)-\theta_0\left(\frac{t}{n}\right)\right|^2_{V(\theta_0(t/n))} w\left(\frac{t}{n}\right) \tag{6}$$

and

$$d_I(\hat{\theta}_h,\theta_0) := \int_0^1 |\hat{\theta}_h(u)-\theta_0(u)|^2_{V(\theta_0(u))} w(u)\,\mathrm{d}u. \tag{7}$$

It can be shown that for $w \equiv 1$, $2d_A$ and $2d_I$ are approximations of the global Kullback–Leibler divergence between models with parameter curves $\hat{\theta}_h(\cdot)$ and

$\theta_0(\cdot)$ which can be seen as follows: If ℓ is the correct likelihood (4) and we assume that all observations including the negative indices are available, we obtain with a Taylor expansion

$$
\begin{aligned}
&\frac{1}{n}\sum_{t=1}^{n}\mathbb{E}_{\theta_0}\log\biggl(\frac{d\mathbb{P}^{X_{t,n}|Y_{t-1,n},\theta_0}}{d\mathbb{P}^{X_{t,n}|Y_{t-1,n},\theta_1}}\biggr)\\
&\quad\approx\frac{1}{n}\sum_{t=1}^{n}\mathbb{E}_{\theta_0}\biggl[\ell\biggl(X_{t,n},Y_{t-1,n},\theta_1\Bigl(\frac{t}{n}\Bigr)\biggr)-\ell\biggl(X_{t,n},Y_{t-1,n},\theta_0\Bigl(\frac{t}{n}\Bigr)\biggr)\biggr]\\
&\quad\approx\frac{1}{n}\sum_{t=1}^{n}\mathbb{E}_{\theta_0}\nabla\ell\biggl(X_{t,n},Y_{t-1,n},\theta_0\Bigl(\frac{t}{n}\Bigr)\biggr)\cdot\biggl(\theta_1\Bigl(\frac{t}{n}\Bigr)-\theta_0\Bigl(\frac{t}{n}\Bigr)\biggr)\\
&\qquad+\frac{1}{2n}\sum_{t=1}^{n}\mathbb{E}_{\theta_0}\biggl|\theta_1\Bigl(\frac{t}{n}\Bigr)-\theta_0\Bigl(\frac{t}{n}\Bigr)\biggr|^2_{\nabla^2\ell(X_{t,n},Y_{t-1,n},\theta_0(\frac{t}{n}))}.
\end{aligned}
\tag{8}
$$

The first approximation holds since the evolution from $Y_{t-1,n}$ to $X_{t,n}$ is mainly affected by θ_0 through $\theta_0(\frac{t}{n})$; see (1). Since ℓ is the correct likelihood, it holds that $\mathbb{E}_{\theta_0}\nabla\ell(X_{t,n},Y_{t-1,n},\theta_0(\frac{t}{n}))\approx 0$ which shows that only the second summand in (8) remains, which is approximately $d_A(\theta_1,\theta_0)$, and thus justifies the definition of d_A.

To use our approach, it is necessary that this property is still maintained even if ℓ is not the correct likelihood, see Assumption 3.3(2). This is fulfilled for many time series models; cf. Section 4.

To give a feeling of the arising quantities, we discuss them for the simple example of the tvAR(1) process.

EXAMPLE 2.4 (tvAR(1) process). Let $X_{t,n}=\theta_0(\frac{t}{n})X_{t-1,n}+\varepsilon_t$ with i.i.d. ε_t, where $\mathbb{E}\varepsilon_t=0$, $\mathbb{E}\varepsilon_t^2=1$ and $\theta_0:[0,1]\to(-1,1)$. With $\ell(x,y,\theta):=\frac{1}{2}(x-\theta y_1)^2+\text{const}$ chosen as the negative log Gaussian likelihood, we obtain

$$
\hat{\theta}_h(u)=\frac{\hat{c}_{1,h}(u)}{\hat{c}_{0,h}(u)},\qquad \hat{c}_{j,h}(u):=\frac{1}{n}\sum_{t=1}^{n}K_h\Bigl(\frac{t}{n}-u\Bigr)X_{t-j,n}X_{t-1,n},
$$

and $V(\theta)=(1-\theta^2)^{-1}$ which leads to

$$
d_A(\hat{\theta}_h,\theta_0)=\frac{1}{n}\sum_{t=1}^{n}\biggl(1-\theta_0\Bigl(\frac{t}{n}\Bigr)^2\biggr)^{-1}\cdot\biggl(\hat{\theta}_h\Bigl(\frac{t}{n}\Bigr)-\theta_0\Bigl(\frac{t}{n}\Bigr)\biggr)^2 w\Bigl(\frac{t}{n}\Bigr).
$$

Unlike the direct ASE $\frac{1}{n}\sum_{t=1}^{n}(\hat{\theta}_h(\frac{t}{n})-\theta_0(\frac{t}{n}))^2$, the Kullback–Leibler-type distance $d_A(\hat{\theta}_h,\theta_0)$ takes care of the fact that θ_0 has to be well estimated if it attains values near 1 to guarantee that the model described by $\hat{\theta}_h$ is near to the model described by θ_0.

In Theorem 3.8 below, we will prove that under suitable conditions, $d_A(\hat{\theta}_h, \theta_0)$ can be approximated uniformly in h by a deterministic distance measure $d^*_{M,2}(h)$, which has a unique minimizer $h_0 = h_{0,n} \sim n^{-1/5}$. h_0 can be seen as the (deterministic) optimal bandwidth.

2.3. *The cross-validation method.* We now choose the bandwidth h by a generalized cross-validation method. The main idea is to approximate the infeasible distance measure $d_A(\hat{\theta}_h, \theta_0)$ by an estimator $\mathrm{CV}(h)$. Motivated by (8), we replace θ_1 therein with an estimator $\hat{\theta}_{h,-t}$ of θ_0 which guarantees unbiasedness. We define a "quasi-leave-one-out" local likelihood

$$L_{n,h,-t}(u,\theta) := \frac{1}{n} \sum_{s=1, s\neq t}^{n} K_h\left(\frac{s}{n} - u\right) \ell_{s,n}(\theta) \tag{9}$$

and a "quasi-leave-one-out" estimator of θ_0 by

$$\hat{\theta}_{h,-t}(u) := \operatorname*{argmin}_{\theta\in\Theta} L_{n,h,-t}(u,\theta). \tag{10}$$

Here, "leave-one-out" does not mean that we ignore the tth observation of the process $(X_{s,n})_{s=1,\dots,n}$, but that we ignore the term which is contributed by the likelihood $\ell_{t,n}$ at time step t. In case of a Gaussian likelihood, this can be interpreted as leaving out the tth projection error. Because of that, we refer to the estimator as a quasi-leave-one-out method.

We then choose $\hat{h}$ via minimizing the cross-validation functional

$$\mathrm{CV}(h) := \frac{1}{n} \sum_{t=1}^{n} \ell_{t,n}\left(\hat{\theta}_{h,-t}\left(\frac{t}{n}\right)\right) w\left(\frac{t}{n}\right). \tag{11}$$

Note that there may not exist a unique minimizer $\hat{h}$ of $\mathrm{CV}(h)$ due to its piecewise constancy. For the mathematical considerations, we therefore choose some $\hat{h}$ such that

$$\mathrm{CV}(\hat{h}) - \inf_{h\in H_n} \mathrm{CV}(h) \leq \frac{1}{n}, \tag{12}$$

where H_n is a suitable subinterval of $(0, 1)$, see Assumption 3.4, which covers all relevant values of h.

Let us specify the corresponding estimators in the tvAR(1) from Example 2.4 above.

EXAMPLE 2.5 (tvAR(1) process ctd). We have

$$\hat{\theta}_{h,-t}(u) = \frac{\hat{c}_{1,h,-t}(u)}{\hat{c}_{0,h,-t}(u)}, \qquad \hat{c}_{j,h,-t}(u) := \frac{1}{n} \sum_{s=1, s\neq t}^{n} K_h\left(\frac{s}{n} - u\right) X_{s-j,n} X_{s-1,n},$$

and, ignoring the changes for the first summand $t = 1$,

$$
\begin{aligned}
\mathrm{CV}(h) &= \frac{1}{2n}\sum_{t=1}^{n}\Big(X_{t,n} - \hat{\theta}_{h,-t}\Big(\frac{t}{n}\Big)X_{t-1,n}\Big)^2 w\Big(\frac{t}{n}\Big) \\
&= -\frac{1}{n}\sum_{t=1}^{n}\sum_{t=1}^{n}\varepsilon_t X_{t-1,n}\Big(\hat{\theta}_{h,-t}\Big(\frac{t}{n}\Big) - \theta_0\Big(\frac{t}{n}\Big)\Big)w\Big(\frac{t}{n}\Big) \\
&\quad + \frac{1}{2n}\sum_{t=1}^{n}X_{t-1,n}^2\Big(\hat{\theta}_{h,-t}\Big(\frac{t}{n}\Big) - \theta_0\Big(\frac{t}{n}\Big)\Big)^2 w\Big(\frac{t}{n}\Big) \\
&\quad + \frac{1}{n}\sum_{t=1}^{n}\varepsilon_t^2 w\Big(\frac{t}{n}\Big).
\end{aligned} \tag{13}
$$

While the first equation in (13) shows how to use $\mathrm{CV}(h)$ in practice, the second equation gives a glance how $\mathrm{CV}(h)$ is used to approximate $d_A(\hat{\theta}_h, \theta_0)$ in this situation: the second summand $\frac{1}{n}\sum_{t=1}^{n} X_{t-1,n}^2(\hat{\theta}_{h,-t}(\frac{t}{n}) - \theta_0(\frac{t}{n}))^2 w(\frac{t}{n})$ is a direct approximation of $d_A(\hat{\theta}_h, \theta_0)$.

3. Main results. In this chapter, we present our main results concerning the bandwidth $\hat{h}$ chosen by cross validation. Our results are twofold. By assuming that θ_0 is only Hoelder continuous and of bounded variation, we prove in Theorem 3.6 that $\hat{h}$ is asymptotically optimal with respect to d_A, that is,

$$
\lim_{n\to\infty}\frac{d_A(\hat{\theta}_{\hat{h}}, \theta_0)}{\inf_{h\in H_n} d_A(\hat{\theta}_h, \theta_0)} = 1 \qquad \text{a.s.}
$$

This result especially holds for nonsymmetric one-sided kernels which is of special interest in prediction. Recall that $d_A(\hat{\theta}_h, \theta_0)$ can be interpreted as a Kullback–Leibler-type distance between the two time series models associated to $\hat{\theta}_h$ and θ_0. Thus, the cross-validation procedure yields an estimator $\hat{\theta}_{\hat{h}}$ of θ_0 such that the distributions of the associated time series coincide best.

In the special situation that K is a symmetric kernel and θ_0 is twice continuously differentiable, we show in Theorem 3.9 that $\hat{h}$ is consistent in the sense that $\hat{h}/h_0 \to 1$ a.s., where h_0 is the deterministic optimal bandwidth defined in (22). Furthermore, we derive the asymptotic distribution and the convergence rate of $\hat{h}$, more precisely we show that $n^{3/10}(\hat{h} - h_0)$ is asymptotically normal in Theorem 3.10.

3.1. *Assumptions for asymptotic optimality of $\hat{h}$.* We split the assumptions into three parts. Assumption 3.1 asks the stationary approximation $\tilde{X}_t(\theta)$ to fulfill some mixing conditions stated with the dependence measure introduced in [17], which is necessary to prove asymptotic results. Assumption 3.3 states conditions on the

objective function ℓ, ensuring the application of typical maximum likelihood techniques. Assumption 3.4 collects some requirements on the kernel K and the weight function w which are usually satisfied in practice and are only dependent on the choice of the user.

It is important to note that all our assumptions are stated in terms of the stationary approximation $\tilde{X}_t(\theta)$ which are therefore easily verifiable due to known results on stationary time series. In Section 4, it is shown that a large class of time series models such as tvARMA or tvARCH models fulfill these assumptions.

Mixing conditions: We use the functional dependence measure introduced in [17]. Let ε_t, $t \in \mathbb{Z}$ be a sequence of i.i.d. random variables. For $t \geq 0$, let $\mathcal{F}_t := (\varepsilon_t, \varepsilon_{t-1}, \ldots)$ be the shift process and $\mathcal{F}_t^* := (\varepsilon_t, \ldots, \varepsilon_1, \varepsilon_0^*, \varepsilon_{-1}, \ldots)$, where ε_0^* is a random variable which has the same distribution as ε_0 and is independent of all ε_t, $t \in \mathbb{Z}$. For a stationary process $W_t = H(\mathcal{F}_t) \in L^q$ with deterministic $H : \mathbb{R}^\infty \to \mathbb{R}$ define $W_t^* := H_t(\mathcal{F}_t^*)$ and the functional dependence measure

$$\delta_q^W(k) := \|W_t - W_t^*\|_q. \tag{14}$$

ASSUMPTION 3.1 (Dependence assumption). Suppose that for each $\theta \in \Theta$, there exists a representation $\tilde{X}_t(\theta) = H(\theta, \mathcal{F}_t)$ with some measurable $H(\theta, \cdot)$ and $\delta_q(k) := \sup_{\theta\in\Theta} \delta_q^{\tilde{X}(\theta)}(k) = O(k^{-(3+\eta)})$ for some $\eta > 0$.

Conditions on ℓ: To state smoothness conditions on the objective function ℓ in a concise way, we introduce the class of Lipschitz-continuous functions from $\mathbb{R}^\infty$ to $\mathbb{R}$ where we allow the Lipschitz constant to depend on the location at most polynomially.

DEFINITION 3.2 (The class $\mathcal{L}(M, \chi, C)$). We say that a function $g : \mathbb{R}^\infty \times \Theta \to \mathbb{R}$ is in the class $\mathcal{L}(M, \chi, C)$ if $C = (C_1, C_2)$, $M \geq 1$, $\chi = (\chi_i)_{i=1,2,3,\ldots} \in \mathbb{R}_{\geq 0}^\infty$ and for all $z \in \mathbb{R}^\infty$, $\theta \in \Theta$:

$$\begin{aligned} \sup_{z\neq z'} \frac{|g(z,\theta) - g(z',\theta)|}{|z - z'|_{\chi,1}(1 + |z|_{\chi,1}^{M-1} + |z'|_{\chi,1}^{M-1})} &\leq C_1, \\ \sup_{\theta\neq\theta'} \frac{|g(z,\theta) - g(z,\theta')|}{|\theta - \theta'|_1(1 + |z|_{\chi,1}^{M})} &\leq C_2, \end{aligned} \tag{15}$$

where $|z|_{\chi,1} := \sum_{i=1}^\infty \chi_i \cdot |z_i|$ and $\sum_{i=1}^\infty \chi_i < \infty$.

We now state the necessary conditions on ℓ.

ASSUMPTION 3.3. Suppose that ℓ is three times differentiable with respect to θ, and:

(1) $\Theta \subset \mathbb{R}^d$ is compact. For all $u \in [0,1]$, $\theta_0(u)$ lies in the interior of Θ and θ_0 is Hoelder continuous with exponent $\beta > 0$ and has componentwise bounded variation B_{θ_0}.

(2) $\theta_0(u)$ is the unique minimizer of $L(u,\theta) := \mathbb{E}\ell(\tilde{Y}_0(\theta_0(u)),\theta)$.

(3) the minimal eigenvalue of $V(\theta) := \mathbb{E}[\nabla^2 \ell(\tilde{Y}_0(\theta'),\theta)|_{\theta'=\theta}]$ is bounded from below by some constant λ_0 uniformly in $\theta \in \Theta$.

(4) $\nabla\ell(\tilde{Y}_0(\theta'),\theta)|_{\theta'=\theta}$ is an uncorrelated sequence.

(5) each component of $g \in \{\ell, \nabla\ell, \nabla^2, \nabla^3\ell\}$ lies in $\mathcal{L}(M,\chi,C)$ for some $\chi = (\chi_j)_{j=1,2,\ldots}$, where $\chi_j = O(j^{-(3+\eta)})$ for some $\eta > 0$.

The conditions are discussed more detailed in Remark 3.5.

Conditions on K, H_n, w: Finally, let us formalize the conditions on the set of bandwidths H_n, the localizing kernel K appearing in the estimation procedure and the weight function w which arises in the cross-validation functional and the distance measures.

ASSUMPTION 3.4. Suppose that:

(1) For $n \in \mathbb{N}$, $H_n = [\underline{h}, \overline{h}]$, where $\underline{h} = \underline{h}_n \geq c_0 n^{\delta-1}$, $\overline{h} = \overline{h}_n \leq c_1 n^{-\delta}$ for some constants $c_0, c_1 > 0$, $\delta \in (0,1)$.

(2) The kernel $K : \mathbb{R} \to \mathbb{R}$ has compact support $\subset [-\frac{1}{2}, \frac{1}{2}]$, fulfills $\int K(x)\,dx = 1$ and is Lipschitz continuous with Lipschitz constant L_K.

(3) The weight function $w : [0,1] \to \mathbb{R}_{\geq 0}$ is bounded by $|w|_\infty$, has bounded variation B_w and compact support $\subset [\gamma, 1-\gamma]$ with some $\gamma > 0$.

REMARK 3.5 (Discussion of the assumptions). 1. Note that Assumptions 3.3(1), 3.3(2), 3.3(3) are standard conditions on the objective function ℓ which ensure the validity of basic results (such as Taylor expansions) from maximum likelihood theory. Condition 3.3(2) also implies that $\mathbb{E}\nabla\ell(\tilde{Y}_0(\theta_0(u)),\theta_0(u)) = 0$ which is important to support the interpretation of $d_A(\hat{\theta}_h,\theta_0)$ as a Kullback–Leibler-type distance measure in (8). Furthermore, it ensures the approximation of $d_A(\hat{\theta}_h,\theta_0)$ by $\mathrm{CV}(h)$; cf. Section 3.6.

2. Assumption 3.3(4) is crucial to prove that $\mathrm{CV}(h)$ is an unbiased estimator of $d_A(\hat{\theta}_h,\theta_0)$ which leads to the necessary rate of convergence. In many time series models which are based on i.i.d. innovations, $\nabla\ell(\tilde{Y}_0(\theta'),\theta)|_{\theta'=\theta}$ is a martingale difference which is even stronger; cf. Section 4. The Lipschitz assumptions 3.3(5) are used in three ways: They allow to replace $X_{t,n}$ by its stationary approximations, they guarantee uniform convergence results in θ which are needed in maximum likelihood theory and they are used to transfer the mixing conditions of $\tilde{X}_t(\theta)$ to functions of $\tilde{X}_t(\theta)$ such as $\ell(\tilde{Y}_t(\theta'),\theta)$. Let us emphasize that we have to ask ℓ and its derivatives to decay with a certain rate χ in y to deal with the truncated past which is used in our approach (3).

3. In principle, $\delta \in (0,1)$ and $c_0 > 0$ in Assumption 3.4 can be chosen very small and $c_1 > 0$ arbitrarily large, which ensures that all relevant bandwidths are covered by H_n. In most practical applications, one can even choose $H_n = [0, \infty)$ without having any drawbacks. A standard choice for the weight function is $w(\cdot) = \mathbb{1}_{[\gamma, 1-\gamma]}(\cdot)$ with some $\gamma > 0$.

3.2. *Asymptotic optimality of $\hat{h}$.* Let us emphasize that the following result asks θ_0 to be only Hoelder continuous and of bounded variation. The kernel is allowed to be one-sided which may be of interest in prediction settings.

THEOREM 3.6 (Asymptotic optimality of cross validation). *Under Assumptions 2.1, 3.1, 3.3 and 3.4, the bandwidth $\hat{h}$ chosen by cross validation is asymptotically optimal in the sense that*

$$\lim_{n\to\infty} \frac{d(\hat{\theta}_{\hat{h}}, \theta_0)}{\inf_{h\in H_n} d(\hat{\theta}_h, \theta_0)} = 1 \qquad a.s.,$$

where d is d_A or d_I from (6) and (7).

3.3. *Assumptions for twice continuously differentiable θ_0.* To ensure that usual second-order bias decompositions hold for d_A, we state natural specifications of the smoothness properties of ℓ and the underlying process $\tilde{X}_t(\theta)$.

ASSUMPTION 3.7 (Bias expansion conditions). Suppose that:

(1) K is symmetric and θ_0 is twice continuously differentiable.

(2) For all $\theta \in \Theta$, $z \in \mathbb{R}^\infty$, $z \mapsto \nabla \ell(z, \theta)$ is twice partially differentiable and $\partial_{z_i}\partial_{z_j}\nabla\ell(\cdot,\theta) \in \mathcal{L}(\max\{M-2,1\}, \chi, \tilde{\psi}_1(i)\tilde{\psi}_2(j))$ for all $i, j \geq 1$ with absolutely summable sequences $\tilde{\psi}_1, \tilde{\psi}_2$.

(3) $\theta \mapsto \tilde{X}_t(\theta)$ is twice continuously differentiable almost surely. It holds that $\|\sup_{\theta\in\Theta} |\nabla \tilde{X}_0(\theta)|_1\|_M$ and $\|\sup_{\theta\in\Theta} |\nabla^2 \tilde{X}_0(\theta)|_1\|_M$ are finite, and

$$\sup_{\theta\neq\theta'} \frac{\||\nabla^2 \tilde{X}_0(\theta) - \nabla^2 \tilde{X}_0(\theta')|_1\|_M}{|\theta - \theta'|_1} < \infty.$$

3.4. *Results on convergence rates of $\hat{h}$.* We know from standard asymptotics that

$$\begin{aligned} \hat{\theta}_h(u) - \theta_0(u) &\approx -\nabla^2 L_{n,h}\big(u, \bar{\theta}(u)\big)^{-1} \nabla L_{n,h}\big(u, \theta_0(u)\big) \\ &\approx -V\big(\theta_0(u)\big)^{-1} \nabla L_{n,h}\big(u, \theta_0(u)\big), \end{aligned} \tag{16}$$

which motivates the following approximation of $d_I(\hat{\theta}_h, \theta_0)$:

$$d_I^*(h) := \int_0^1 \big|\nabla L_{n,h}\big(u, \theta_0(u)\big)\big|^2_{V(\theta_0(u))^{-1}} w(u)\,\mathrm{d}u. \tag{17}$$

While $d_I(\hat{\theta}_h, \theta_0)$ contains the implicitly defined $\hat{\theta}_h$, the quantity $d_I^*(h)$ can be stated explicitly which allows the explicit calculation of its expectation. We now set (with "M" for mean)

$$d_M^*(h) := \mathbb{E} d_I^*(h),$$

which can be seen as an approximation of the weighted mean integrated squared error $\mathbb{E} d_I(\hat{\theta}_h, \theta_0)$ of $\hat{\theta}_h$. If additionally to Assumptions 2.1, 3.1, 3.3, 3.4, we suppose Assumption 3.7, Proposition 1.1 implies the usual bias-variance decomposition for d_M^*:

$$d_M^*(h) = \frac{\mu_K V_0}{nh} + \frac{h^4}{4} d_K^2 B_0 + o\big((nh)^{-1}\big) + o\big(h^4\big) \tag{18}$$

uniformly in $h \in H_n$, where $\mu_K := \int K(x)^2 \,\mathrm{d}x$, $d_K := \int K(x) x^2 \,\mathrm{d}x$ and

$$V_0 := \int_0^1 \operatorname{tr}\big\{V\big(\theta_0(u)\big)^{-1} I\big(\theta_0(u)\big)\big\} w(u) \,\mathrm{d}u > 0, \tag{19}$$

$$B_0 := \int_0^1 \big|\mathbb{E}\big[\partial_u^2 \nabla \ell\big(\tilde{Y}_t\big(\theta_0(u)\big), \theta\big)|_{\theta=\theta_0(u)}\big]\big|^2_{V(\theta_0(u))^{-1}} w(u) \,\mathrm{d}u \geq 0, \tag{20}$$

leading to the definition of the deterministic bias-variance decomposition $d_{M,2}^*(h)$ without any smaller-order terms and the resulting asymptotically optimal bandwidth in the following two theorems.

THEOREM 3.8 (Approximation of distance measures). *Let Assumptions 2.1, 3.1, 3.3, 3.4 and 3.7 hold. Define*

$$d_{M,2}^*(h) := \frac{\mu_K V_0}{nh} + \frac{h^4}{4} d_K^2 B_0. \tag{21}$$

If the bias B_0 is not degenerated, that is, $B_0 > 0$, then it holds that

$$\sup_{h \in H_n} \left| \frac{d(\hat{\theta}_h, \theta_0) - d_{M,2}^*(h)}{d_{M,2}^*(h)} \right| \to 0 \qquad a.s.,$$

where d is d_A or d_I from (6) and (7).

THEOREM 3.9 (Consistency of the cross-validation bandwidth). *Let Assumptions 2.1, 3.1, 3.3, 3.4 and 3.7 hold and assume that $B_0 > 0$. Then the bandwidth $\hat{h}$ chosen by cross validation fulfils*

$$\frac{\hat{h}}{h_0} \to 1 \qquad a.s.,$$

where

$$h_0 = \left(\frac{V_0 \mu_K}{B_0 d_K^2}\right)^{1/5} n^{-1/5} \tag{22}$$

is the unique minimizer of $d_{M,2}^*(h)$ from (21).

Note that Theorem 3.9 does not give any information about the convergence rate of $\hat{h}$ toward h_0. Under some additional regularity assumptions on the kernel K, it is possible to obtain the exact asymptotic behavior.

THEOREM 3.10 (Asymptotic normality of the cross-validation bandwidth). *Let Assumptions 2.1, 3.1, 3.3, 3.4 and 3.7 hold. Additionally, assume that $B_0 > 0$, the second derivative of θ_0 is Lipschitz continuous and that K is continuously differentiable with Lipschitz continuous derivative K'. Put $\hat{K}(x) := -K'(x)x$ and $\tilde{K}(x) = K - \hat{K}$. Then it holds with $C_0 := nh_0^5$ that*

$$n^{3/10}(\hat{h} - h_0) \xrightarrow{d} N\biggl(0, \frac{8}{25} \cdot \frac{\int f_{\mathrm{var}}(u)\,\mathrm{d}u}{V_0^2} \cdot \frac{\int (\tilde{K} - K * \tilde{K})^2}{\mu_K^2} \cdot C_0^{3/5}\biggr), \tag{23}$$

where $*$ denotes convolution, V_0 is defined in (19), μ_K is defined below (18), the matrices I, V are defined in (5) and

$$f_{\mathrm{var}}(u) := w(u)^2 \operatorname{tr}\bigl\{V\bigl(\theta_0(u)\bigr)^{-1} I\bigl(\theta_0(u)\bigr) V\bigl(\theta_0(u)\bigr)^{-1} I\bigl(\theta_0(u)\bigr)\bigr\}.$$

Since $h_0 \sim n^{-1/5}$ in the above situation, the relative proportion $\frac{\hat{h}-h_0}{h_0}$ has a convergence rate of order $n^{1/10}$ which is common for standard cross-validation selectors (see [11]). Note especially that our model covers the i.i.d. regression case which was discussed in [11]. The additional Lipschitz assumption on the second derivative of θ_0 is necessary to quantify the residual terms of $d_M^*(h)$ in (18) more detailed.

REMARK 3.11. It is seen in the examples in Section 4 that if the model is correctly specified but higher moments of ε_0 are not known, it may hold that $I(\theta) = \kappa \cdot I(\theta)$ with some real number $\kappa > 0$, usually only depending on properties of the i.i.d. innovations ε_t. In this case, it holds that $V_0 = p \cdot \kappa \cdot \int w(u)\,\mathrm{d}u$ (p is the dimension of the parameter space) and $\int f_{\mathrm{var}}(u)\,\mathrm{d}u = p \cdot \kappa^2 \cdot \int w(u)^2\,\mathrm{d}u$, leading to simpler forms of V_0 and the asymptotic variance term in (23). Especially in the case that the whole model (including the distribution of ε_0) is correctly specified, it holds that $\kappa = 1$.

REMARK 3.12. Theorem 3.10 can also be used to provide confidence intervals for h_0. Such results may be useful to adjust the cross validation chosen bandwidth. If the simplifications from Remark 3.11 do not hold, one may estimate V_0 and $\int f_{\mathrm{var}}(u)\,\mathrm{d}u$ by $\hat{V}_0$ and $\int \hat{f}_{\mathrm{var}}(u)\,\mathrm{d}u$ which are obtained by replacing $V(\theta_0(u))$, $I(\theta_0(u))$ by $\hat{V}_{n,\hat{h}}(u, \hat{\theta}_{\hat{h}}(u))$, $\hat{I}_{n,\hat{h}}(u, \hat{\theta}_{\hat{h}}(u))$, respectively, where

$$\hat{V}_{n,h}(u,\theta) := \frac{1}{nh} \sum_{t=1}^{n} K_h\biggl(\frac{t}{n} - u\biggr) \nabla^2 \ell\bigl(Y_{t,n}^c, \theta\bigr),$$

$$\hat{I}_{n,h}(u,\theta) := \frac{1}{nh} \sum_{t=1}^{n} K_h\biggl(\frac{t}{n} - u\biggr) \nabla \ell\bigl(Y_{t,n}^c, \theta\bigr) \cdot \nabla \ell\bigl(Y_{t,n}^c, \theta\bigr)'.$$

The asymptotic theory is provided in the Supplementary Material [15]; see Lemma 3.6 (applied to $g = \nabla^2 \ell$ or $g = \nabla \ell \cdot \nabla \ell'$ therein). Then an asymptotic $(1-\alpha)$ confidence interval for h_0 is given by

$$[\hat{h} - \hat{D}, \hat{h} + \hat{D}],$$

where

$$\hat{D} := \frac{q_{1-\frac{\alpha}{2}}\sqrt{8}}{5} \cdot \frac{\sqrt{\int \hat{f}_{\mathrm{var}}(u)\,\mathrm{d}u}}{\hat{V}_0} \cdot \frac{\sqrt{\int (\tilde{K} - K * \tilde{K})^2}}{\mu_K} \cdot \hat{h}^{3/2}$$

and $q_{1-\frac{\alpha}{2}}$ denotes the $(1-\frac{\alpha}{2})$-quantile of the standard normal distribution. Note especially that by the form of the asymptotic variance (23), the bias term B_0 does not have to be estimated separately.

3.5. *Possible generalizations*. In the following, some immediate generalizations of the cross-validation approach (11), (12) are discussed.

REMARK 3.13 (Local linear estimation). Let $\tilde{\Theta} = [-R, R]^p$ with some $R > 0$ large enough. If the parameter curve θ_0 is known to be twice continuously differentiable, instead of (9), (10) and (11) one can also use a local linear approach to estimate h_0 via minimizing

$$\mathrm{CV}^{\mathrm{lin}}(h) := \frac{1}{n}\sum_{t=1}^{n} \ell_{t,n}\Big(\tilde{\theta}_{h,-t}\Big(\frac{t}{n}\Big)\Big) w\Big(\frac{t}{n}\Big),$$

where

$$\big(\tilde{\theta}_{h,-t}(u), \widetilde{\theta'_{h,-t}(u)}\big) := \operatorname*{argmin}_{(\theta,\tilde{\theta})\in\Theta\times\tilde{\Theta}} L^{\mathrm{lin}}_{n,h,-t}(u,\theta,\tilde{\theta})$$

and

$$L^{\mathrm{lin}}_{n,h,-t}(u,\theta,\tilde{\theta}) := \frac{1}{n}\sum_{s=1,s\neq t}^{n} K_h\Big(\frac{s}{n} - u\Big)\ell_{s,n}\Big(\theta + \Big(\frac{s}{n} - u\Big)\tilde{\theta}\Big).$$

We conjecture that similar results as given in Theorems 3.6, 3.8, 3.9 and 3.10 can be shown under the stated assumptions. The main difference is the change of the bias term B_0 to

$$\tilde{B}_0 = \int_0^1 |\theta_0''(u)|^2_{V(\theta_0(u))^{-1}} w(u)\,\mathrm{d}u,$$

due to local linear estimation; cf. [13].

REMARK 3.14 (Computational time). In general, the calculation of $\mathrm{CV}(h)$ proposed in (11) asks to provide n estimators $\hat{\theta}_{h,-t}(\frac{t}{n})$ which are obtained by nonlinear optimizations in (10). If one needs to evaluate $\mathrm{CV}(h)$ for m different values

of h, one has to perform $O(n \cdot m)$ nonlinear optimizations which may be computationally hard.

Note that in some special models like tvAR(r) processes, explicit estimators are available (cf. Remark 4.2). Due to the structure of the estimators, it is even possible to calculate all estimators $\hat{\theta}_{h,-t}(t/n)$, $t = 1, \dots, n$ simultaneously via a convolution approach, which can be used to speed up computation.

For time series with length at most $n = 1000$, the computation of $\hat{h}$ usually only takes seconds. For larger time series, we propose to use a reduced J-fold cross-validation approach as described in Remark 3.15.

REMARK 3.15 (Reduced J-fold cross validation). The typical J-fold cross-validation routine ($J \in \mathbb{N}$) from i.i.d. regression can be adapted in our model: Based on the decomposition $\{1, \dots, n\} = \bigcup_{j=1}^{J} T_{n,j}$, where $T_{n,j} := \{j + J \cdot i : i \in \mathbb{N}_0\} \cap \{1, \dots, n\}$, the J-fold cross-validation functional $\mathrm{CV}^{(J)}$ reads

$$\mathrm{CV}^{(J)}(h) = \frac{1}{J} \sum_{j=1}^{J} \mathrm{CV}^{(J,j)}(h),$$

where the "reduced" J-fold cross-validation functional is based on the validation set $T_{n,j}$,

$$\mathrm{CV}^{(J,j)}(h) = \frac{1}{\#T_{n,j}} \sum_{t \in T_{n,j}} \ell_{t,n}\left(\hat{\theta}_h^{(-j)}\left(\frac{t}{n}\right)\right) w\left(\frac{t}{n}\right),$$

and the corresponding estimators $\hat{\theta}_h^{(-j)}(u) := \operatorname{argmin}_{\theta \in \Theta} L_{n,h}^{(-j)}(u, \theta)$ with

$$L_{n,h}^{(-j)}(u, \theta) := \frac{1}{(n - \#T_{n,j})} \sum_{t \in \{1,\dots,n\} \setminus T_{n,j}} K_h\left(\frac{t}{n} - u\right) \ell_{t,n}(\theta)$$

are based on the training set $\{1, \dots, n\} \setminus T_{n,j}$.

Note that $\mathrm{CV}^{(n)}(h)$ coincides with the original cross-validation routine $\mathrm{CV}(h)$. In view of computational time, $\mathrm{CV}^{(J)}$ has no advantage compared to $\mathrm{CV}(h)$ since still n possibly nonlinear optimizations have to be performed for calculating $\hat{\theta}_h^{(-j)}(t/n)$, $t \in T_{n,j}$, $j = 1, \dots, J$.

We therefore propose to fix some $j_0 \in \{1, \dots, J\}$ and choose $\hat{h}^{(j_0)}$ as a minimizer of only one "reduced" functional $\mathrm{CV}^{(J,j_0)}(h)$. Since then the "effective" training data has only size $n \cdot (1 - \frac{1}{J})$, we expect that $\hat{h}^{(j_0)}$ provides a reasonable estimator of $h_0 \cdot (1 - \frac{1}{J})^{-1/5}$. As long as J is constant in n, we conjecture that our proofs for the properties of $\hat{h}$ and $\mathrm{CV}(h)$ also apply in this situation which means, especially that if Assumptions 2.1,3.1, 3.3, 3.4 and 3.7 are fulfilled,

$$n^{3/10}\left(\hat{h}^{(j_0)} - h_0\left(1 - \frac{1}{J}\right)^{-1/5}\right) \xrightarrow{d} N\left(0, \sigma_{\hat{h}}^2 \cdot (J-1) \cdot \left(1 - \frac{1}{J}\right)^{-3/5}\right),$$

where $\sigma_{\hat{h}}^2$ is the variance of the limit distribution of $\hat{h}$ given in Theorem 3.10, (23).

Using this approach, only $\frac{n}{J}$ nonlinear optimizations for calculating $\hat{\theta}_h^{(-j_0)}(t/n)$, $t \in T_{n,j_0}$ have to be performed, but in turn the cross-validation routine has a higher variance. A typical choice of J is 5 or 10.

3.6. *Proofs.* Here, we present the main ideas of the proofs of the theorems. For the proof of Theorem 3.6, we only discuss the result for $d = d_A$, the proof for d_I is similar. The main idea is to show that $2\,\mathrm{CV}(h)$ approximates $d_A(\hat{\theta}_h, \theta_0)$ uniformly in $h \in H_n$, which then shows that their minima $\hat{h}$ and $\operatorname{argmin}_{h\in H_n} d_A(\hat{\theta}_h, \theta_0)$ converge to each other, giving the result.

Let Assumptions 2.1, 3.1, 3.3 and 3.4 hold. Recall from (18) that $d_M^*(h)$ can be seen as a deterministic MSE of the estimation problem which has a typical bias-variance decomposition and, therefore, describes the squared rate with which θ_0 is estimated by $\hat{\theta}_h$. In the following we show that certain quantities can be approximated by each other with a rate smaller than that given by $d_M^*(h)$. Define

$$d_{A,-}(h) := \frac{1}{n}\sum_{t=1}^{n} \left| \hat{\theta}_{h,-t}\left(\frac{t}{n}\right) - \theta_0\left(\frac{t}{n}\right)\right|^2_{V(\theta_0(t/n))} w\left(\frac{t}{n}\right), \tag{24}$$

which is the same as $d_A(\hat{\theta}_h, \theta_0)$ but with $\hat{\theta}_h$ replaced by the corresponding leave-one-out estimators $\hat{\theta}_{h,-t}$. In a sequence of lemmas in the Supplementary Material [15] (cf. Section 2, Lemmas 2.1, 2.2, 2.3 and 2.4 therein), we show that

$$\sup_{h\in H_n} \left| \frac{d_A(\hat{\theta}_h, \theta_0) - d_{A,-}(h)}{d_M^*(h)} \right| \to 0 \qquad \text{a.s.}, \tag{25}$$

which means that omitting the tth prediction error in $d_A(\hat{\theta}_h, \theta_0)$ is negligible in comparison to the MSE rate $d_M^*(h)$. Furthermore, we show that

$$\sup_{h\in H_n} \left| \frac{d_A(\hat{\theta}_h, \theta_0) - d_M^*(h)}{d_M^*(h)} \right| \to 0 \qquad \text{a.s.}, \tag{26}$$

that is, $d_A(\hat{\theta}_h, \theta_0)$ can be approximated by $d_M^*(h)$ with a rate which is negligible in comparison to $d_M^*(h)$. As a second auxiliary result, we need the following.

LEMMA 3.16. *Let Assumptions 2.1, 3.1, 3.3, 3.4 hold. Then*

$$\sup_{h\in H_n} \left| \frac{2[\mathrm{CV}(h) - \frac{1}{n}\sum_{t=1}^{n} \ell_{t,n}(\theta_0(\frac{t}{n}))w(\frac{t}{n})] - d_{A,-}(h)}{d_M^*(h)} \right| \to 0 \qquad \text{a.s.}, \tag{27}$$

which contains the connection between $\mathrm{CV}(h)$ *and* $d_{A,-}(h)$.

The proof of Lemma 3.16 is based on a Taylor argument similar to (8): By a Taylor expansion, it holds that

$$
\begin{aligned}
&2\left[\mathrm{CV}(h) - \frac{1}{n}\sum_{t=1}^{n} \ell_{t,n}\left(\theta_0\left(\frac{t}{n}\right)\right) w\left(\frac{t}{n}\right)\right] \\
&\quad = \frac{2}{n}\sum_{t=1}^{n} \nabla \ell_{t,n}\left(\theta_0\left(\frac{t}{n}\right)\right)' \left\{\hat{\theta}_{h,-t}\left(\frac{t}{n}\right) - \theta_0\left(\frac{t}{n}\right)\right\} w\left(\frac{t}{n}\right) \\
&\qquad + \frac{1}{n}\sum_{t=1}^{n} \left|\hat{\theta}_{h,-t}\left(\frac{t}{n}\right) - \theta_0\left(\frac{t}{n}\right)\right|^2_{\nabla^2 \ell_{t,n}(\theta_0(t/n))} w\left(\frac{t}{n}\right) \\
&\qquad + \frac{1}{n}\sum_{t=1}^{n} \left|\hat{\theta}_{h,-t}\left(\frac{t}{n}\right) - \theta_0\left(\frac{t}{n}\right)\right|^2_{\nabla^2 \ell_{t,n}(\tilde{\theta}_{h,-t}(t/n)) - \nabla^2 \ell_{t,n}(\theta_0(t/n))} \\
&\qquad \times w\left(\frac{t}{n}\right),
\end{aligned} \tag{28}
$$

where $\tilde{\theta}_{h,-t}(t/n)$ is some intermediate value between $\hat{\theta}_{h,-t}(t/n)$ and $\theta_0(t/n)$. Using (16), the first summand in (28) can be approximated by

$$
\begin{aligned}
2\,\mathrm{CV}^*(h) = -\frac{2}{n}\sum_{t=1}^{n} &\nabla \ell_{t,n}\left(\theta_0\left(\frac{t}{n}\right)\right)' V\left(\theta_0\left(\frac{t}{n}\right)\right)^{-1} \\
&\times \nabla L_{n,h,-t}\left(\frac{t}{n}, \theta_0\left(\frac{t}{n}\right)\right) w\left(\frac{t}{n}\right)
\end{aligned} \tag{29}
$$

which has approximately expectation 0 zero due to Assumption 3.3(4) which mainly justifies $2\,\mathrm{CV}(h)$ as an unbiased estimator of d_A and shows that $\mathrm{CV}^*(h)$ has a smaller rate than $d_M^*(h)$. The second summand in (28) is approximately $d_{A,-}(h)$ due to $\mathbb{E}[\nabla^2 \ell_{t,n}(\theta_0(t/n))] \approx V(\theta_0(t/n))$, and thus eliminated in the difference (27). Finally, the third term can be shown to be of smaller order than $d_M^*(h)$ since it has order $O((\hat{\theta}_{h,-t}(t/n) - \theta_0(t/n))^3)$. Details for the proof of Lemma 3.16 can be found in the Supplementary Material [15], Section 2 therein.

To prove the results (25), (26) and (27), we use as a main tool a general bound for moments on quadratic and cubic forms of functions of locally stationary processes (cf. Proposition 8.1 in the Supplementary Material [15]) which may be of independent interest. Note that for instance (29) can be seen as a quadratic form in the terms $\nabla \ell_{t,n}(\theta_0(t/n))$ and $\nabla \ell_{s,n}(\theta_0(t/n))$ [the last one coming from $\nabla L_{n,h,-t}(t/n, \theta_0(t/n))$].

With the help of these results, we can now prove Theorem 3.6.

PROOF OF THEOREM 3.6. Using the result (25), Lemma 3.16 and (26) [which allows to replace $d_M^*(h)$ in the denumerator], we have

$$
\sup_{h \in H_n} \left| \frac{2[\mathrm{CV}(h) - \frac{1}{n}\sum_{t=1}^{n} \ell_{t,n}(\theta_0(\frac{t}{n}))w(t/n)] - d_A(\hat{\theta}_h, \theta_0)}{d_A(\hat{\theta}_h, \theta_0)} \right| \to 0 \qquad \text{a.s.} \tag{30}
$$

This shows that $2\,\mathrm{CV}(h)$ approximates $d_A(\hat{\theta}_h, \theta_0)$ uniformly in $h \in H_n$ (up to a constant) with a rate smaller than $d_A(\hat{\theta}_h, \theta_0)$. In the following, we show that this implies that the minimizer $\hat{h}$ of $\mathrm{CV}(h)$ (up to a term n^{-1}) converges to the minimizer h' of $d_A(\hat{\theta}_h, \theta_0)$ (up to a term n^{-1}) which then shows the result.

An immediate consequence of (30) is (use $\frac{x_1+x_2}{y_1+y_2} \le \frac{x_1}{y_1} + \frac{x_2}{y_2}$ for positive numbers $x_1, x_2, y_1, y_2 > 0$)

$$\sup_{h,h' \in H_n} \left| \frac{d_A(\hat{\theta}_h, \theta_0) - d_A(\hat{\theta}_{h'}, \theta_0) - 2(\mathrm{CV}(h) - \mathrm{CV}(h'))}{d_A(\hat{\theta}_h, \theta_0) + d_A(\hat{\theta}_{h'}, \theta_0)} \right| \to 0 \qquad \text{a.s.}$$

Choosing $h = \hat{h}$ and h' such that

$$d_A(\hat{\theta}_{h'}, \theta_0) - \inf_{h \in H_n} d_A(\hat{\theta}_h, \theta_0) \le n^{-1}$$

yields

$$\begin{aligned} 0 \leftarrow{}& \frac{d_A(\hat{\theta}_{\hat{h}}, \theta_0) - d_A(\hat{\theta}_{h'}, \theta_0) - (\mathrm{CV}(\hat{h}) - \mathrm{CV}(h'))}{d_A(\hat{\theta}_{\hat{h}}, \theta_0) + d_A(\hat{\theta}_{h'}, \theta_0)} \\ \ge{}& \frac{d_A(\hat{\theta}_{\hat{h}}, \theta_0) - \inf_{h \in H_n} d_A(\hat{\theta}_h, \theta_0) - (\inf_{h \in H_n} \mathrm{CV}(h) - \mathrm{CV}(h'))}{d_A(\hat{\theta}_{\hat{h}}, \theta_0) + \inf_{h \in H_n} d_A(\hat{\theta}_h, \theta_0) + n^{-1}} \\ &+ \frac{2n^{-1}}{d_A(\hat{\theta}_{\hat{h}}, \theta_0) + d_A(\hat{\theta}_{h'}, \theta_0)} \end{aligned}$$

almost surely. By Proposition 1.1, it holds that $d_M^*(h) = \frac{\mu_K V_0}{nh} + B_h + o((nh)^{-1})$ uniformly in $h \in H_n$, where B_h is some nonnegative bias term. Together with (26), we conclude that $\sup_{h \in H_n} \frac{n^{-1}}{d_A(\theta_h, \theta_0)} \to 0$ a.s. Thus,

$$\frac{d_A(\hat{\theta}_{\hat{h}}, \theta_0) - \inf_{h \in H_n} d_A(\hat{\theta}_h, \theta_0)}{d_A(\hat{\theta}_{\hat{h}}, \theta_0) + \inf_{h \in H_n} d_A(\hat{\theta}_h, \theta_0)} \to 0 \qquad \text{a.s.},$$

from which

$$\frac{d_A(\hat{\theta}_{\hat{h}}, \theta_0)}{\inf_{h \in H_n} d_A(\hat{\theta}_h, \theta_0)} \to 1 \qquad \text{a.s.}$$

follows. The same can be done for d_I. □

The work done for the proof of Theorem 3.6 directly allows to prove Theorem 3.8 and 3.9.

PROOF OF THEOREM 3.8. Because of $B_0 > 0$ and (18), we have

$$\sup_{h \in H_n} \left| \frac{d_M^*(h) - d_{M,2}^*(h)}{d_{M,2}^*(h)} \right| \to 0 \qquad \text{a.s.} \tag{31}$$

Application of (26), that is, $\sup_{h\in H_n} |\frac{d_A(\hat{\theta}_h,\theta_0)-d^*_M(h)}{d^*_M(h)}| \to 0$ a.s., completes the proof. □

PROOF OF THEOREM 3.9. We start with (30) from the proof of Theorem 3.6. Using (31) from the proof of Theorem 3.8 and (26), we obtain

$$\sup_{h\in H_n}\left|\frac{\mathrm{CV}(h) - \frac{1}{n}\sum_{t=1}^n \ell_{t,n}(\theta_0(t/n))w(t/n) - d^*_{M,2}(h)}{d^*_{M,2}(h)}\right| \to 0 \qquad \text{a.s.}$$

Using the same methods as in the proof of Theorem 3.6, we have almost surely

$$\frac{d^*_{M,2}(\hat{h})}{d^*_{M,2}(h_0)} = \frac{d^*_{M,2}(\hat{h})}{\inf_{h\in H_n} d^*_{M,2}(h)} \to 1.$$

The structure of $d^*_{M,2}(h)$ implies $\hat{h}/h_0 \to 1$ a.s. □

Finally, we state the proof of Theorem 3.10, the asymptotic normality of $\hat{h}$. Again some lemmas from the Supplementary Material [15], Section 4 are used which provide uniform convergences of arising quadratic or cubic forms of locally stationary processes. The core result for proving asymptotic normality is Lemma 4.8 which is based on a general central limit theorem for quadratic forms of locally stationary processes, Theorem 7.1, which may be of independent interest.

PROOF OF THEOREM 3.10. If K is differentiable, then $h \mapsto \mathrm{CV}(h)$ is differentiable in h and $\hat{h}$ can be chosen as a minimizer. $\hat{h}$ is in the interior of H_n for n large enough due to Theorem 3.9. The proof is based on the following expansion:

$$\begin{aligned} 0 &= 2\partial_h \mathrm{CV}(\hat{h}) = \partial_h d^*_{M,2}(\hat{h}) + \partial_h D(\hat{h}) \\ &= \partial_h^2 d^*_{M,2}(h^*)\cdot(\hat{h}-h_0) + \partial_h D(\hat{h}), \end{aligned}$$

where $D(h) := 2\,\mathrm{CV}(h) - d^*_{M,2}(h)$, h_0 is the unique minimizer of $d^*_{M,2}(h)$ defined in (22), and h^* is some intermediate value between $\hat{h}$ and h_0. Thus

$$\hat{h} - h_0 = -\frac{\partial_h D(\hat{h})}{\partial_h^2 d^*_{M,2}(h^*)}. \tag{32}$$

By Theorem 3.9, we have $\hat{h}/h_0 \to 1$ a.s. and thus $h^*/h_0 \to 1$ a.s. The structure of $\partial_h^2 d^*_{M,2}$ implies that $\frac{\partial_h^2 d^*_{M,2}(h^*)}{\partial_h^2 d^*_{M,2}(h_0)} \to 1$. We conclude that

$$n^{3/10}(\hat{h}-h_0) = \frac{n^{7/10}\partial_h D(\hat{h})}{n^{2/5}\partial_h^2 d^*_{M,2}(h_0)} + o(1) \qquad \text{a.s.},$$

with $n^{2/5}\partial_h^2 d^*_{M,2}(h_0) = 5(\mu_K V_0)^{2/5}(B_0 d_K^2)^{3/5}$. In Lemma 4.3, it is shown that

$$\sup_{h\in\tilde{H}_n} h^{1/2}\left|\frac{\partial_h D(h) - \partial_h \tilde{D}(h)}{d^*_M(h)}\right| \to 0, \tag{33}$$

where $\tilde{D}(h) := \{d^*_I(h) - d^*_M(h)\} + 2\,\mathrm{CV}^*(h)$ and $\mathrm{CV}^*(h)$ is defined in (29) and $\tilde{H}_n = [c_0 n^{-\frac{1}{3}+\delta}, c_1 n^{-\delta}]$.

Since $\frac{\hat{h}}{h_0} \to 1$ a.s., we have that almost surely, $\hat{h} \in \tilde{H}_n$ for n large enough. By (31) we have $\frac{d^*_M(\hat{h})}{d^*_{M,2}(\hat{h})} \to 1$ a.s. By the structure of $d^*_{M,2}$, we obtain $\frac{d^*_{M,2}(\hat{h})}{d^*_{M,2}(h_0)} \to 1$. So inserting $\hat{h}$ in (33) yields

$$n^{7/10}\left|\partial_h D(\hat{h}) - \partial_h \tilde{D}(\hat{h})\right| \to 0 \qquad \text{a.s.},$$

that is,

$$n^{3/10}(\hat{h} - h_0) = \frac{n^{7/10}\partial_h \tilde{D}(\hat{h})}{n^{2/5}\partial_h^2 d^*_{M,2}(h_0)} + o(1) \qquad \text{a.s.} \tag{34}$$

By Lemmas 4.4 and 4.7 we have for each $\gamma > 0$ that

$$\sup_{h\in\tilde{H}_n} n^{-\gamma}\cdot h^{1/2}\frac{|\partial_h \tilde{D}(h)|}{d^*_M(h)} \to 0 \qquad \text{a.s.}$$

Together with (34) we conclude that

$$n^{3/10}(\hat{h} - h_0) = O\big(n^{\gamma}\big) \qquad \text{a.s.} \tag{35}$$

By Lemmas 4.4, 4.5 and 4.6 it holds for each $\tilde{\gamma} > 0$ that

$$\sup_{h,h'\in\tilde{H}_n,\,\frac{|h-h'|}{h}\le n^{-\tilde{\gamma}}} h^{1/2}\frac{|\partial_h \tilde{D}(h) - \partial_h \tilde{D}(h')|}{d^*_M(h)} \to 0 \qquad \text{a.s.}$$

Inserting $h = h_0$, $h' = \hat{h}$ [which is possible due to (35) with $\gamma = \frac{1}{10} - \tilde{\gamma}$, $\tilde{\gamma} \in (0, \frac{1}{10})$], we obtain

$$n^{7/10}\big(\partial_h \tilde{D}(h_0) - \partial_h \tilde{D}(\hat{h})\big) \to 0 \qquad \text{a.s.}$$

Inserting this into (34) yields

$$n^{3/10}(\hat{h} - h_0) = \frac{n^{7/10}\partial_h \tilde{D}(h_0)}{n^{2/5}\partial_h^2 d^*_{M,2}(h_0)} + o(1) \qquad \text{a.s.}$$

Lemma 4.8 in connection with Lemma 4.4 provides a central limit theorem for the joint vector $(2\partial_h \mathrm{CV}^*(h), \partial_h\{d^*_I(h) - d^*_M(h)\})'$, that is,

$$\begin{aligned}&(n^2 h_0^3)^{1/2}\begin{pmatrix} 2\partial_h \mathrm{CV}^*(h) \\ \partial_h\{d^*_I(h) - d^*_M(h)\}\end{pmatrix} \\ &\quad\xrightarrow{d} N\left(0, 8\int f_{\mathrm{var}}(u)\,\mathrm{d}u\cdot\Sigma_K + 4C_0 d_K^2\int f_{\mathrm{bias}}(u)\,\mathrm{d}u\cdot\begin{pmatrix}1 & -1\\ -1 & 1\end{pmatrix}\right),\end{aligned}$$

where $*$ denotes convolution, $C_0 = nh_0^5 = \frac{V_0\mu_K}{B_0 d_K^2}$ and

$$\Sigma_K := \begin{pmatrix} \int (\tilde{K})^2 & -\int \tilde{K}\cdot(K*\tilde{K}) \\ -\int \tilde{K}\cdot(K*\tilde{K}) & \int (K*\tilde{K})^2 \end{pmatrix},$$

$$f_{\text{bias}}(u) := w(u)^2 \operatorname{tr}\{\text{bias}(u)' V(\theta_0(u))^{-1} I(\theta_0(u)) V(\theta_0(u))^{-1} \text{bias}(u)\},$$

$$\text{bias}(u) = \mathbb{E}[\partial_u^2 \nabla \ell(\tilde{Y}_t(\theta_0(u)), \theta)|_{\theta=\theta_0(u)}].$$

Furthermore, $n^{2/5}\partial_h^2 d^*_{M,2}(h_0) = n^{2/5}(\frac{2\mu_K V_0}{nh_0^3} + 3h_0^2 d_K^2 B_0) = 5(B_0 d_K^2)^{3/5}(V_0\mu_K)^{2/5}$ and $n^{7/10}(n^2 h_0^3)^{-1/2} = (\frac{V_0\mu_K}{B_0 d_K^2})^{-3/10}$. We conclude that

$$n^{3/10}(\hat{h} - h_0) \xrightarrow{d} N\left(0, \frac{8\int f_{\text{var}}(u)\,\mathrm{d}u \cdot \int (\tilde{K} - K*\tilde{K})^2}{25(V_0\mu_K)^{7/5}\cdot(B_0 d_K^2)^{3/5}}\right)$$

$$= N\left(0, \frac{8}{25}\frac{\int f_{\text{var}}(u)\,\mathrm{d}u}{V_0^2}\cdot\frac{\int (\tilde{K} - K*\tilde{K})^2}{\mu_K^2}\cdot C_0^{3/5}\right). \qquad \square$$

4. Examples and simulations.

4.1. *Examples*. Assumptions 2.1, 3.1, 3.3 and 3.7 are fulfilled for a large class of locally stationary time series models. Here, we discuss how the conditions transform in the case of some special linear and recursively defined time series. The proofs of this section can be found in the Supplementary Material [15] (Section 5 therein). There one can also find a more general statement about linear time series in Proposition 5.1.

Recall that ε_t, $t \in \mathbb{Z}$ is a sequence of i.i.d. real random variables. We will use a Gaussian likelihood for ℓ defined in (4), but allow for a non-Gaussian distribution of ε_t.

An important special case of locally stationary linear processes is given by tvARMA processes, see also Proposition 2.4. in [5]. Since in this case, the linear filter $A_\theta(\lambda) = \sigma \cdot \frac{\beta(e^{i\lambda})}{\alpha(e^{i\lambda})}$ and the spectral density $f_\theta(\lambda) = \frac{\sigma^2}{2\pi}\cdot|\frac{\beta(e^{i\lambda})}{\alpha(e^{i\lambda})}|^2$ have a simple form, the conditions in Proposition 5.1 are obviously fulfilled. The likelihood (4) takes the form

$$\ell(z,\theta) = \frac{1}{2}\log\left(\frac{2\pi}{\gamma_\theta(0)^2}\right) + \frac{1}{2}\left(\sum_{j=0}^{\infty}\gamma_\theta(j) z_{j+1}\right)^2, \tag{36}$$

where $\gamma_\theta(j) := \frac{1}{2\pi}\int_{-\pi}^{\pi} A_\theta(\lambda)^{-1} e^{-i\lambda j}\,\mathrm{d}\lambda$.

EXAMPLE 4.1 (tvARMA(r, s) process). Assume that $\varepsilon_t, t \in \mathbb{Z}$ are i.i.d. with existing moments of all order. Suppose that $\mathbb{E}\varepsilon_0 = 0$ and $\mathbb{E}\varepsilon_0^2 = 1$. Let Assumption 3.3(1) hold. Assume that $X_{t,n}$ obeys

$$X_{t,n} + \sum_{j=1}^{r} \alpha_j\left(\frac{t}{n}\right) X_{t-j,n}$$
$$= \sigma\left(\frac{t}{n}\right)\varepsilon_t + \sum_{k=1}^{s} \beta_k\left(\frac{t}{n}\right)\sigma\left(\frac{t-k}{n}\right)\varepsilon_{t-k}, \qquad t = 1, \ldots, n,$$

where $\theta_0 = (\alpha_1, \ldots, \alpha_r, \beta_1, \ldots, \beta_s, \sigma)' : [0, 1] \to \mathbb{R}^{r+s+1}$. Define $\beta(z) := 1 + \sum_{k=0}^{s} \beta_k z^k$, $\alpha(z) := 1 + \sum_{k=0}^{r} \alpha_k z^k$, and let Θ be an arbitrary compact subset of

$$\{\theta = (\alpha_1, \ldots, \alpha_r, \beta_1, \ldots, \beta_s, \sigma)' \in \mathbb{R}^{r+s+1} : \sigma > 0,$$
$$\alpha(z), \beta(z) \text{ have no zeros in common and}$$
$$\text{only zeros outside the unit circle}\}.$$

Then Assumptions 2.1, 3.1, 3.3 are fulfilled for ℓ chosen as in (36). If additionally Assumption 3.7(1) is fulfilled, then Assumption 3.7 is fulfilled. It holds that $V(\theta) = \frac{1}{4\pi}\int \nabla \log f_\theta(\lambda) \cdot \nabla \log f_\theta(\lambda)' \, d\lambda$ and $I(\theta) = V(\theta) + \kappa_4(\varepsilon_0) \cdot \frac{\nabla\gamma_\theta(0)\nabla\gamma_\theta(0)'}{\gamma_\theta(0)^2}$, where $\kappa_4(\varepsilon_0)$ is the fourth cumulant of ε_0.

Explicit formulas for the bias (20) are available and can be found in the Supplementary Material [15], Proposition 5.1.

REMARK 4.2 (tvAR(r) processes). In the special case of $tvAR(r)$ processes, closed forms for the estimators based on $\ell(z, \theta) = \frac{1}{2}\log(2\pi\sigma^2) + \frac{1}{2\sigma^2}(z_1 + \sum_{j=1}^{r} \alpha_j z_{j+1})^2$ are available: $\hat{\alpha}_h(u) = -\hat{\Gamma}_h(u)^{-1}\hat{\gamma}_h(u)$ and $\hat{\sigma}_h(u)^2 = \frac{1}{n}\sum_{t=r+1}^{n}(X_{t,n} + \sum_{j=1}^{r} \hat{\alpha}_j(u) X_{t-j,n})^2$, where $Y^\circ_{t-1,n} = (X_{t-1,n}, \ldots, X_{t-r,n})'$ and

$$\hat{\Gamma}_h(u) := \frac{1}{n}\sum_{t=r+1}^{n} K_h\left(\frac{t}{n} - u\right) Y^\circ_{t-1,n}\big(Y^\circ_{t-1,n}\big)',$$

$$\hat{\gamma}_h(u) := \frac{1}{n}\sum_{t=r+1}^{n} K_h\left(\frac{t}{n} - u\right) X_{t,n} Y^\circ_{t-1,n}.$$

We now discuss recursively defined nonlinear time series models with additive innovations ε_t. Let us fix some $r > 0$ and define the vectors of the last r lags $Y^\circ_{t-1,n} = (X_{t-1,n}, \ldots, X_{t-r,n})'$, $\tilde{Y}^\circ_{t-1}(\theta) = (\tilde{X}_{t-1}(\theta), \ldots, \tilde{X}_{t-r}(\theta))'$ as the vector of the r past values of the locally stationary and the stationary time series, respectively. Here, we use the superscript $^\circ$ to clearly separate between the infinite-dimensional vector $Y^c_{t-1,n}$ used in the likelihood (3) and $Y^\circ_{t-1,n}$, the lags used to

create the next observations of the model. Many popular locally stationary models assume that the conditional mean and/or variance is a linear combination of unknown parameter curves and functions of $Y^{\circ}_{t-1,n}$, that is,

$$X_{t,n} = \mu\big(Y^{\circ}_{t-1,n}, \theta_0(t/n)\big) + \sigma\big(Y^{\circ}_{t-1,n}, \theta_0(t/n)\big)\varepsilon_t, \qquad t = 1, \dots, n,$$

with some measurable μ, σ. In this case, the likelihood (4) with $y^{\circ} = (y_1, \dots, y_r)'$ for $y = (y_1, y_2, y_3, \dots)$ takes the form

$$\ell(x, y, \theta) := \frac{1}{2}\log\big(2\pi\sigma\big(y^{\circ}, \theta\big)^2\big) + \frac{1}{2}\bigg(\frac{x - \mu(y^{\circ}, \theta)}{\sigma(y^{\circ}, \theta)}\bigg)^2. \tag{37}$$

We adapt a result from [13] (Example 5.1 therein) which deals with μ, σ^2 having a linear structure in the parameters. The following example covers tvAR-, tvTAR and tvARCH processes.

PROPOSITION 4.3 (Time-varying recursively defined time series models). *Consider the recursion*

$$X_{t,n} = \mu\big(Y^{\circ}_{t-1,n}, \theta_0(t/n)\big) + \sigma\big(Y^{\circ}_{t-1,n}, \theta_0(t/n)\big)\varepsilon_t, \tag{38}$$

where $\theta_0 = (\alpha_1, \dots, \alpha_k, \beta_0, \dots, \beta_l)'$ *and*

$$\mu(y, \theta) := \sum_{i=1}^{k} \alpha_i m_i(y), \qquad \sigma(y, \theta) := \bigg(\sum_{i=0}^{l} \beta_i \nu_i(y)\bigg)^{1/2},$$

with some functions $m = (m_1, \dots, m_k) : \mathbb{R}^r \to \mathbb{R}^k$, $\nu = (\nu_0, \dots, \nu_l) : \mathbb{R}^r \to \mathbb{R}^{l+1}_{\geq 0}$. *Assume that*:

1. ε_i *are i.i.d. with* $\mathbb{E}\varepsilon_i = 0$, $\mathbb{E}\varepsilon_i^2 = 1$ *and* $\mathbb{E}\varepsilon_i^q < \infty$ *for all* $q > 0$.
2. *For all* $\theta \in \Theta$, *the sets*

$$\big\{m_1\big(\tilde{Y}^{\circ}_0(\theta)\big), \dots, m_k\big(\tilde{Y}^{\circ}_0(\theta)\big)\big\}, \qquad \big\{\nu_0\big(\tilde{Y}^{\circ}_0(\theta)\big), \dots, \nu_l\big(\tilde{Y}^{\circ}_0(\theta)\big)\big\}$$

are (*separately*) *linearly independent in* $\mathcal{L}_2$.

3. *There exist* $(\kappa_{ij}) \in \mathbb{R}^{k\times r}_{\geq 0}$, $(\rho_{ij}) \in \mathbb{R}^{(l+1)\times r}_{\geq 0}$ *such that for all* i:

$$\sup_{y\neq y'} \frac{|m_i(y) - m_i(y')|}{|y - y'|_{\kappa_{i\cdot},1}} \leq 1, \qquad \sup_{y\neq y'} \frac{|\sqrt{\nu_i(y)} - \sqrt{\nu_i(y')}|}{|y - y'|_{\rho_{i\cdot},1}} \leq 1. \tag{39}$$

Let $\nu_{\min} > 0$ *be some constant such that for all* $y \in \mathbb{R}^r$, $\nu_0(y) \geq \nu_{\min}$. *With some* $\beta_{\min} > 0$, *choose* $\tilde{\Theta} \subset \mathbb{R}^k \times \mathbb{R}^{l+1}_{\geq \beta_{\min}}$ *such that for all* $q > 0$,

$$\sum_{j=1}^{p}\bigg(\sup_{\theta\in\tilde{\Theta}} \sum_{i=1}^{k} |\alpha_i|\kappa_{ij} + \|\varepsilon_0\|_q \cdot \sup_{\theta\in\tilde{\Theta}} \sum_{i=0}^{l} \sqrt{\beta_i}\rho_{ij}\bigg) < 1. \tag{40}$$

4. *Assumption* 3.3(1) *is valid with some* $\Theta \subset \tilde{\Theta}$.

Then Assumptions 2.1, 3.1 and 3.3 are fulfilled for ℓ chosen to be proportional to the negative log Gaussian conditional likelihood (37) with $M = 3$. In the special case $\sigma(x,\theta)^2 \equiv \beta_0$, one can choose $M = 2$.

With the shortcuts $m = m(\tilde{Y}_0^\circ(\theta))$, $\nu = \nu(\tilde{Y}_0^\circ(\theta))$ it holds that

$$V(\theta) = \begin{pmatrix} \mathbb{E}\dfrac{mm'}{\langle \beta, \nu\rangle} & 0 \\ 0 & \mathbb{E}\dfrac{\nu\nu'}{2\langle \beta, \nu\rangle^2} \end{pmatrix}, \tag{41}$$

$$I(\theta) = \begin{pmatrix} \mathbb{E}\dfrac{mm'}{\langle \beta, \nu\rangle} & \mathbb{E}[\varepsilon_0^3]\cdot \mathbb{E}\dfrac{m\nu'}{2\langle \beta, \nu\rangle^{3/2}} \\ \mathbb{E}[\varepsilon_0^3]\cdot \mathbb{E}\dfrac{\nu m'}{2\langle \beta, \nu\rangle^{3/2}} & \dfrac{\mathbb{E}\varepsilon_0^4 - 1}{4}\cdot \mathbb{E}\dfrac{\nu\nu'}{2\langle \beta, \nu\rangle^2}, \end{pmatrix}. \tag{42}$$

If additionally, Assumption 3.7(1) is fulfilled and m_i, ν_i are twice continuously differentiable such that for all $j_1, j_2 = 1, \dots, r$ and all i,

$$\sup_{y \neq y'} \frac{|\partial_{y_{j_1}} \partial_{y_{j_2}} m_i(y) - \partial_{y_{j_1}} \partial_{y_{j_2}} m_i(y')|}{|y - y'|_1} < \infty, \tag{43}$$

$$\sup_{y \neq y'} \frac{|\partial_{y_{j_1}} \partial_{y_{j_2}} \nu_i(y) - \partial_{y_{j_1}} \partial_{y_{j_2}} \nu_i(y')|}{|y - y'|_1} < \infty, \tag{44}$$

then Assumption 3.7 is fulfilled for ℓ from (37).

REMARK 4.4. 1. If (i) $\mathbb{E}\zeta_0^3 = 0$, or (ii) $\mu(z,\theta) \equiv 0$ or (iii) $\sigma(z,\theta) \equiv \beta_0$ and $\mathbb{E}m(\tilde{Y}_0^\circ(\theta)) = 0$, then

$$I(\theta) = \begin{pmatrix} I_k & 0 \\ 0 & (\mathbb{E}\varepsilon_0^4 - 1) I_{l+1}/2 \end{pmatrix} \cdot V(\theta),$$

where I_d denotes the d-dimensional identity matrix. If additionally $\mathbb{E}\varepsilon_0^4 = 3$ (as it is the case for ε_0 having a standard normal distribution), we have $I(\theta) = V(\theta)$.

2. Note that in many special cases (for instance, tvAR or tvARCH processes) where m_i, ν_i have simple forms and explicit representations of the processes are available, the restrictive conditions on the parameter space (40) can be relaxed by rewriting the recursion (38) as a r-dimensional recursion with only one lag and using matrix arguments.

3. In the tvARCH case [or, more general in cases where $\sigma(z,\theta)$ is dependent on z in a nontrivial way], condition (40) can only be satisfied if there exists $C_\varepsilon > 0$ such that $\|\varepsilon_0\|_q \le C_\varepsilon$ for all $q \ge 1$. By Markov's inequality, this directly implies that ε_0 has to be bounded almost surely, that ist, $|\varepsilon_0| \le C_\varepsilon$ a.s.

4. Explicit formulas for the bias (20) are available in the Supplementary Material [15], Lemma 5.2.

A simulation study. Here, we study the behavior of the presented cross- validation algorithm for different time series models. We assume that ε_t is standard Gaussian distributed, and consider:

(a) tvAR(1) processes $X_{t,n} = \alpha(\frac{t}{n})X_{t-1,n} + \sigma(\frac{t}{n})\varepsilon_t$, with $\alpha(u) = 0.9\sin(2\pi u)$ and $\sigma(u) = 0.3\sin(2\pi u) + 0.5$.

(b) tvMA(1) processes $X_{t,n} = \sigma(\frac{t}{n})\varepsilon_t + \alpha(\frac{t}{n})\sigma(\frac{t-1}{n})\varepsilon_{t-1}$, with $\alpha(u) = 0.9 \times \sin(2\pi u)$ and $\sigma(u) = 0.3\sin(2\pi u) + 0.5$.

(c) tvARCH(1) processes $X_{t,n} = \sqrt{\alpha_1(\frac{t}{n}) + \alpha_2(\frac{t}{n})X_{t-1,n}^2} \cdot \varepsilon_{t-1}$, with $\alpha_1(u) = 0.2\sin(2\pi u) + 0.4$ and $\alpha_2(u) = 0.1\sin(2\pi u) + 0.2$.

(d) tvTAR(1) processes $X_{t,n} = \alpha_1(\frac{t}{n})X_{t-1,n}^+ + \alpha_2(\frac{t}{n})X_{t-1,n}^- + \varepsilon_t$, with $\alpha_1(u) = 0.4\sin(2\pi u)$ and $\alpha_2(u) = 0.5\cos(2\pi u)$ and $y^+ := \max\{y, 0\}$, $y^- := \max\{-y, 0\}$ for real numbers y.

We performed a Monte Carlo study by generating in each case $N = 2000$ realizations of time series with length $n \in \{200, 500\}$. For estimation, we used the weight function $w(\cdot) = \mathbb{1}_{[0.01,0.99]}(\cdot)$ which already excludes most of the boundary effects and the Epanechnikov kernel $K(x) = \frac{3}{2}(1 - (2x)^2)\mathbb{1}_{[-\frac{1}{2},\frac{1}{2}]}(x)$.

We chose $H_n = [0.01, 1.0]$ and calculated the cross-validation bandwidth $\hat{h}$, the ao-bandwidth h_0 from Theorem 3.9 [for models (a)–(c), model (d) does not satisfy the smoothness conditions] and the optimal theoretical bandwidth

$$h^* = \operatorname*{argmin}_{h \in H_n} d_A(\hat{\theta}_h, \theta_0).$$

Note that $\hat{h}, h^*$ depend on the current realization while h_0 is deterministic and fixed. h^* and h_0 depend on the unknown true curve $\theta_0(\cdot)$ and are unavailable in practice. More explicit formulas for the bias term (20) which is necessary to calculate h_0 can be found in the Supplementary Material [15], Section 6.

Figure 1 shows the results $\hat{h}$, h^* for the four models, respectively. The histograms show the chosen cross-validation bandwidths $\hat{h}$, the bandwidth h_0 is marked via a black vertical line and the dashed normal distribution is the theoretical expected limit distribution of $\hat{h}$ given by Theorem 3.10. The boxplots show the achieved values of $d_A(\hat{\theta}_h, \theta_0)$ for the different selectors $h \in \{\hat{h}, h_0, h^*\}$ (labeled as "CV," "Plugin" and "Optimal"). Each box contains 50% while the whiskers contain 90% of the values of $d_A(\hat{\theta}_h, \theta_0)$. It can be seen that the cross-validation procedure works well even for the case of a time series length of only $n = 200$. Compared to the theoretical limit distribution of $\hat{h}$ given by Theorem 3.10, we observe that $\hat{h}$ seems to be biased, tending to be slightly greater than h_0, depending on the variance of the limit distribution. The bias reduces significantly if n increases. For the models (a), (d), we observe that the distances d_A attained by the cross-validation approach are nearly as good as the distances obtained by the optimal selector h^* which is remarkable. For the models (b) and (c), the values of d_A asso-

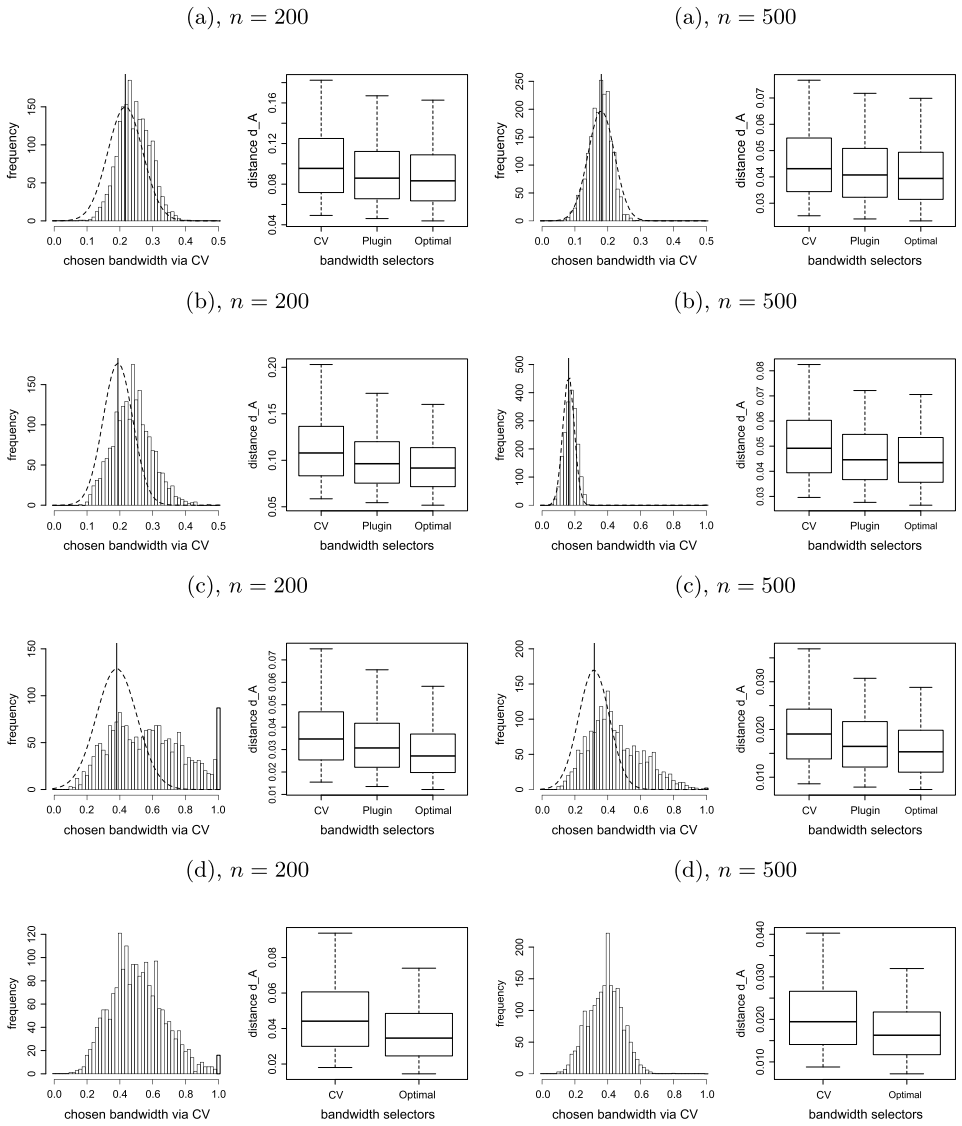

FIG. 1. *Simulation results for the models* (a), (b), (c), (d) *for time series lengths* $n = 200$ (*left*) *and* $n = 500$ (*right*) *and* $N = 2000$ *replications. The left plot shows a histogram of the chosen cross-validation bandwidths* $\hat{h}$, *the vertical line therein represents the asymptotically optimal bandwidth* h_0. *The right box plots show the values of* $d_A(\hat{\theta}_h, \theta_0)$ *achieved for* $h \in \{\hat{h}, h_0, h^*\}$.

ciated to $\hat{h}$ have a higher variance. This can be explained by the higher variance of the maximum likelihood estimators $\hat{\theta}_h$ in these models; a theoretical justification can be found in the corresponding limit distribution of $\hat{h}$ given in Theorem 3.10. In all cases, the distances produced by the estimator based on the cross-validation

TABLE 1
Minimizers $(\alpha^{ms}(u), \sigma^{ms}(u))$ *of* $(\alpha^{ms}, \sigma^{ms}) = \theta \mapsto L(u, \theta) = (\sigma^{ms})^{-2} \cdot \mathbb{E}[\tilde{X}_t(u) - \alpha^{ms} \cdot \tilde{X}_{t-1}(u))]^2 + \log[(\sigma^{ms})^2]$ *in the case of model misspecification*

True model	$\boldsymbol{\alpha^{ms}(u)}$	$\boldsymbol{\sigma^{ms}(u)}$
tvMA	$\frac{\alpha(u)}{1+\alpha(u)^2}$	$(\frac{1+\alpha(u)^2+\alpha(u)^4}{1+\alpha(u)^2})^{1/2} \cdot \sigma(u)$
tvARCH	0	$(\frac{\alpha_1(u)}{1-\alpha_2(u)})^{1/2}$

procedure are of course greater in average, but they still look quite satisfying in our opinion.

Note that in case of a more general theory for derivative processes (see [6] for a discussion) it is possible to show similar results as given in Theorems 3.9, 3.10 for the TAR process (d).

Model misspecifications. We observed in simulations that the performance of the cross-validation procedure is robust against the distribution of ε_t, leading to similar results even if ε_t is uniformly, exponentially or Pareto distributed (meaning that the moment conditions from Assumption 2.1 are violated).

Due to the fact that our cross-validation method is a natural generalization of the version for i.i.d. regression it works even well if the underlying model itself is misspecified. In the following, we estimate parameters with a Gaussian likelihood which assumes that the time series model follows a tvAR(1) model $X_{t,n} = \alpha^{ms}(t/n)X_{t-1,n} + \sigma^{ms}(t/n)\varepsilon_t$, but in fact the underlying model is either tvMA (b) or tvARCH (c). The cross-validation method then tries to estimate the minimizer $\theta_0^{ms}(u) = (\alpha^{ms}(u), \sigma^{ms}(u))'$ of $\theta \mapsto L(u, \theta)$, that is, $\alpha^{ms}(u) = \frac{c(1,u)}{c(0,u)}$ and $\sigma^{ms}(u) = (\frac{c(0)^2 - c(1)^2}{c(0)})^{1/2}$ with the covariances $c(k, u) := \mathbb{E}[\tilde{X}_0(\theta_0(u))\tilde{X}_k(\theta_0(u))]$ (see Table 1). To compare the distances, we use $d_A(\hat{\theta}_h(u), \theta_0^{ms}(u))$ with V from the tvAR(1) model. The simulations are performed in the same way as for the correctly specified case above. In Figure 2, it is seen that even in the misspecified case the bandwidth selector $\hat{h}$ produces reasonable estimators which are comparable with the optimal bandwidth choice h^* in the case of tvMA estimators and still satisfying in the tvARCH case [note that a lot of information is lost due to the fact that $\alpha^{ms}(u) \equiv 0$ in this case].

5. Concluding remarks. In this paper we have introduced a data adaptive bandwidth selector via cross validation which is applicable for a large class of locally stationary processes. An important property of the method is the fact that it does not involve any tuning parameters.

In simulations, we have seen that the proposed cross-validation method yields nearly optimal bandwidth choices with respect to an Kullback–Leibler-type dis-

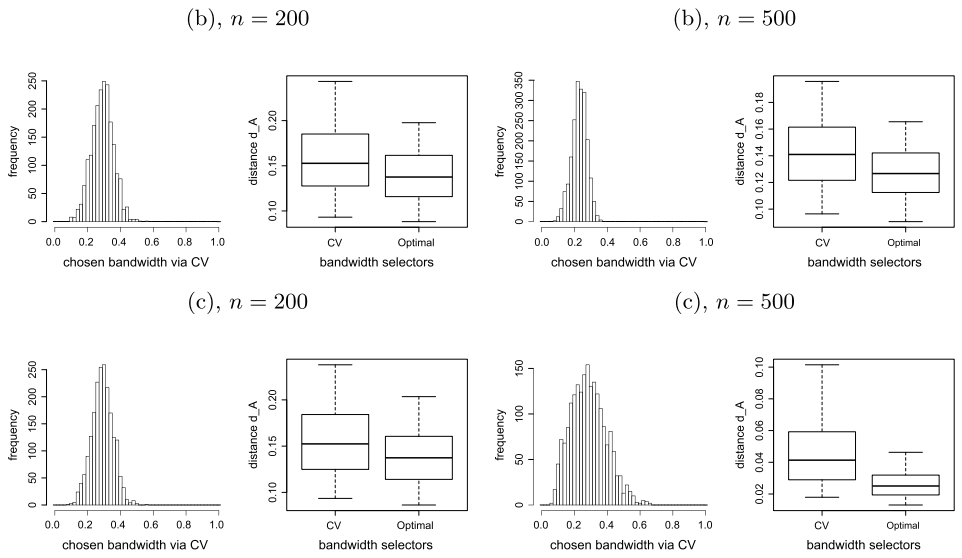

FIG. 2. *Simulation results in the misspecified case*: *The underlying processes are either* (b), (c) *but for the estimation it is assumed that a* tvAR(1) *process is present*.

tance measure in the case of correctly specified models and still leads to satisfying results in the case of model misspecification. It remains an open question if a similar cross-validation procedure can be defined which is asymptotically optimal with respect to a simple quadratic distance measure (i.e., without a weighting matrix) which would then lead to estimates of θ_0 which do not optimize the prediction properties of the associated model but the estimation quality of the parameter curve θ_0 itself.

We worked out the convergence rate of $\hat{h}$ toward the asymptotically optimal bandwidth h_0, which is $\hat{h} = h_0 + O_p(n^{-3/10})$, and showed that $n^{3/10}(\hat{h} - h_0) \xrightarrow{d} N(0, \sigma^2_{\hat{h}})$ with some explicit formula for $\sigma^2_{\hat{h}}$. From this, it could be seen that the convergence rate in practice is strongly dependent on the underlying model; $\sigma^2_{\hat{h}}$ can be large if θ_0 is hard to estimate. This raises the question if there are improved cross-validation methods like [3] (via Fourier transform) or [10] (via presmoothing) proved in the i.i.d. kernel density estimation case that attain the optimal rate of $n^{1/2}$ if further smoothness assumptions on θ_0 are supposed.

We mention that it is not hard to generalize the proposed method and the proofs to multidimensional time series which may be of interest in many practical applications.

An interesting open problem is the adaptive estimation in time series models with several parameter curves coming from different smoothness class, in particular since these curves are not observed separately but via a single time series.

Let us point out the fact that cross-validation procedures in general are not stable if applied locally. Thus it remains an open question to find an local adaptive bandwidth selector.

Acknowledgments. We are grateful to two anonymous referees whose comments lead to a considerable improvement of the organisation of the paper. In particular, showing the asymptotic normality of $\hat{h}$ and providing the strong connection between the Kullback–Leibler divergence and the cross-validation functional was motivated by their remarks.

SUPPLEMENTARY MATERIAL

Supplement: Technical proofs (DOI: 10.1214/18-AOS1743SUPP; .pdf). This material contains some details of the proofs in the paper as well as the proofs of the examples.

REFERENCES

[1] ARKOUN, O. (2011). Sequential adaptive estimators in nonparametric autoregressive models. *Sequential Anal.* **30** 229–247. MR2801140

[2] ARKOUN, O. and PERGAMENCHTCHIKOV, S. (2016). Sequential robust estimation for nonparametric autoregressive models. *Sequential Anal.* **35** 489–515. MR3574333

[3] CHIU, S.-T. (1991). Bandwidth selection for kernel density estimation. *Ann. Statist.* **19** 1883–1905. MR1135154

[4] DAHLHAUS, R. (2012). 13—Locally stationary processes. In *Time Series Analysis*: *Methods and Applications* (T. S. Rao, S. S. Rao and C. R. Rao, eds.). *Handbook of Statistics* **30** 351–413. Elsevier, Amsterdam. DOI:10.1016/B978-0-444-53858-1.00013-2.

[5] DAHLHAUS, R. and POLONIK, W. (2009). Empirical spectral processes for locally stationary time series. *Bernoulli* **15** 1–39. MR2546797

[6] DAHLHAUS, R., RICHTER, S. and WU, W. B. (2018). Towards a general theory for non-linear locally stationary processes. *Bernoulli*. To appear.

[7] DAHLHAUS, R. and SUBBA RAO, S. (2006). Statistical inference for time-varying ARCH processes. *Ann. Statist.* **34** 1075–1114. MR2278352

[8] FRYZLEWICZ, P., SAPATINAS, T. and SUBBA RAO, S. (2008). Normalized least-squares estimation in time-varying ARCH models. *Ann. Statist.* **36** 742–786. MR2396814

[9] GIRAUD, C., ROUEFF, F. and SANCHEZ-PEREZ, A. (2015). Aggregation of predictors for nonstationary sub-linear processes and online adaptive forecasting of time varying autoregressive processes. *Ann. Statist.* **43** 2412–2450. MR3405599

[10] HALL, P., MARRON, J. S. and PARK, B. U. (1992). Smoothed cross-validation. *Probab. Theory Related Fields* **92** 1–20. MR1156447

[11] HÄRDLE, W., HALL, P. and MARRON, J. S. (1988). How far are automatically chosen regression smoothing parameters from their optimum? *J. Amer. Statist. Assoc.* **83** 86–101. MR0941001

[12] HÄRDLE, W. and MARRON, J. S. (1985). Optimal bandwidth selection in nonparametric regression function estimation. *Ann. Statist.* **13** 1465–1481. MR0811503

[13] KARMAKAR, S., RICHTER, S. and WU, W. B. (2018). Bahadur representation and simultaneous inference for time-varying models. Technical report.

[14] MALLAT, S., PAPANICOLAOU, G. and ZHANG, Z. (1998). Adaptive covariance estimation of locally stationary processes. *Ann. Statist.* **26** 1–47. MR1611808

[15] RICHTER, S. and DAHLHAUS, R. (2019). Supplement to "Cross validation for locally stationary processes." DOI:10.1214/18-AOS1743SUPP.

[16] SUBBA RAO, S. (2006). On some nonstationary, nonlinear random processes and their stationary approximations. *Adv. in Appl. Probab.* **38** 1155–1172. MR2285698

[17] WU, W. B. (2005). Nonlinear system theory: Another look at dependence. *Proc. Natl. Acad. Sci. USA* **102** 14150–14154. MR2172215

[18] ZHOU, Z. and WU, W. B. (2009). Local linear quantile estimation for nonstationary time series. *Ann. Statist.* **37** 2696–2729. MR2541444

INSTITUT FÜR ANGEWANDTE MATHEMATIK
UNIVERSITÄT HEIDELBERG
IM NEUENHEIMER FELD 205
69120 HEIDELBERG
GERMANY
E-MAIL: stefan.richter@iwr.uni-heidelberg.de
dahlhaus@statlab.uni-heidelberg.de