Econometric Theory, **39**, 2023, 1123–1153.
doi:10.1017/S0266466622000500

ADAPTATION FOR NONPARAMETRIC ESTIMATORS OF LOCALLY STATIONARY PROCESSES

Rainer Dahlhaus
Heidelberg University

Stefan Richter
Heidelberg University

Two adaptive bandwidth selection methods for minimizing the mean squared error of nonparametric estimators in locally stationary processes are proposed. We investigate a cross-validation approach and a method based on contrast minimization and derive asymptotic properties of both methods. The results are applicable for different statistics under a general setting of local stationarity including nonlinear processes. At the same time, we deepen the general framework for local stationarity based on stationary approximations. For example, a general Bernstein inequality is derived for such processes. The properties of the bandwidth selection methods are also investigated in several simulation studies.

1. INTRODUCTION

In this paper, we develop data adaptive bandwidth selection rules for nonstationary processes under a novel paradigm of local stationarity recently introduced in Dahlhaus, Richter, and Wu (2019). Time series sampled at high frequency or long time series exhibit often nonstationarity instead of stationarity, and correspondingly the use of models with time-varying parameters or of locally stationary processes in general has increased a lot. A prominent example from financial econometrics is the use of GARCH models for modeling conditional heteroskedasticity. While in the beginning ordinary GARCH models have been regarded as sufficient to model conditional heteroskedasticity of the volatility, insight has grown that, for example, modeling of the daily pattern can be improved by a time-varying GARCH model (cf. Dahlhaus and Subba Rao, 2006; Amado and Teräsvirta, 2013; Amado and Teräsvirta, 2017). Analogously, time-varying models for the trading intensity have been used such as locally stationary Hawkes models (Roueff, von Sachs, and Sansonnet, 2016). Motivated by the study of financial returns, Koo and Linton (2012) have investigated locally stationary diffusion processes with a time-varying drift and a volatility coefficient varying over time.

We are very grateful to two referees whose comments helped to improve the paper significantly. Address correspondence to Rainer Dahlhaus, Institut für Angewandte Mathematik, Heidelberg University, Im Neuenheimer Feld 205, Heidelberg, Germany; e-mail: dahlhaus@statlab.uni-heidelberg.de.

The main focus of this paper is on deriving methods for adaptive bandwidth selection of general nonparametric estimators. These include estimators of the time-varying covariance function, the autocorrelation function, the time-varying characteristic function, and, more general, estimators that are functionals of moment estimators. Different to nonparametric regression, there exist only very few theoretical results on adaptivity for locally stationary processes. Mallat, Papanicolaou, and Zhang (1998) discussed adaptive covariance estimation for a general class of locally stationary processes. More recently, Niedzwiecki, Ciolek, and Kajikawa (2017) used a cross-validation approach to estimate covariances and the spectrum for locally stationary linear processes. In Beran (2009), explicit expansions of the mean squared error of parameter estimators of locally stationary long-memory linear processes were derived. Other results are constructed for specific models (in particular, the tvAR model) and are partly dependent on further tuning parameters: in Dahlhaus and Giraitis (1998), explicit expansions of the mean squared error of parameter estimators in tvAR models were provided. Giraud, Roueff, and Sanchez-Perez (2015) discussed online-adaptive forecasting of tvAR processes, and Arkoun (2011) and Arkoun and Pergamenchtchikov (2016) proposed methods for sequential and minimax-optimal bandwidth selection for tvAR processes of order 1. In Zhou and Wu (2009), rules of thumb for bandwidth selection were proposed for local linear quantile estimators of general locally stationary time series models.

In Richter and Dahlhaus (2019), adaptive estimation was developed for time-varying parameter curves by means of local M-estimators, i.e., in a locally parametric setting. In this paper, the task is adaptive estimation in a setting which also locally is a nonparametric one. Technically, the difference is that we no longer assume that the observed time series comes from a specific model like tvGARCH or tvAR (and use this knowledge to build the estimator). Instead, we are interested in general functionals of the time series such as covariances, correlations, or characteristic functions.

Beyond that, there is another benefit this paper: in the course of the derivations, we also deepen the general framework for local stationarity of Dahlhaus et al. (2019) based on stationary approximations—the Bernstein inequality at the end of Section 4 being an example. General frameworks for locally stationary processes had been introduced by Dahlhaus (1997) for time-varying linear processes, and by Wu (2005) and Wu and Zhou (2011) for time-varying Bernoulli shifts in combination with the functional dependence measure, and furthermore by Priestley (1965, 1988), the nonasymptotic approach for processes with evolutionary spectra. Former approaches which use the original idea behind local stationarity, namely that at each point in time the observed nonstationary process can be approximated by a stationary process, were considered in the context of time-varying ARCH processes in Dahlhaus and Subba Rao (2006), and investigated further in the context of random coefficient models in Subba Rao (2006). The use of such approximations as a general model was recommended by Vogt (2012), who investigated nonparametric regression for locally stationary time series, and

https://doi.org/10.1017/S0266466622000500 Published online by Cambridge University Press

by Koo and Linton (2012), who investigated semiparametric estimation for locally stationary models.

In Section 2, we introduce the framework of local stationarity based on stationary approximations. We mention some results from Dahlhaus et al. (2019) and extend these results in which we prove an invariance property of the results also for nonlinear transformations based on infinitely many lags. In Section 3, we introduce a global bandwidth selection method based on cross validation and prove asymptotic optimality of this method with respect to a squared-error-type distance measure. We also discuss its behavior in practice via simulations. In Section 4, we introduce a local bandwidth selection procedure by using a contrast minimization approach in the spirit of Lepski, Mammen, and Spokoiny (1997). We prove that the resulting nonparametric estimator attains the optimal rate for the mean squared error up to a log factor. We compare the obtained method with a global optimal selection routine and show the superiority of the method in examples. The section also contains a Bernstein inequality which is of interest beyond the present paper. Section 5 contains some concluding remarks.

Throughout this paper, we use the following notation. For vectors $x,y \in \mathbb{R}^d$ and positive semidefinite matrices $A \in \mathbb{R}^{d\times d}$, $|x|_2 := (\sum_{j=1}^{d}|x_j|^2)^{1/2}$ denotes the euclidean norm, x' the transpose, $x'y$ the euclidean scalar product, and $|x|_A := (x'Ax)^{1/2}$ the weighted euclidean norm.

The Supplementary Material for this article contains several technical results including the proofs of the main theorems and a more general result for the setting in Section 4.

2. MOMENT ESTIMATORS AND OPTIMAL BANDWIDTH SELECTORS

2.1. The Model

We assume that we observe n realizations of a process $X_{t,n}$ at time points $t = 1,\dots,n$. The process is considered to be locally stationary in the following sense (cf. Dahlhaus et al., 2019).

Assumption 2.1. Let $q \geq 1$. There exists some $D > 0$, and, for each $u \in [0,1]$, there exists a strictly stationary process $(\tilde{X}_t(u))_{t\in\mathbb{Z}}$ such that, for all $t = 1,\dots,n$ and $u,u' \in [0,1]$, $\|\tilde{X}_0(u)\|_q \leq D$, $\|X_{t,n}\|_q \leq D$, and

$$\|X_{t,n} - \tilde{X}_t(t/n)\|_q \leq Dn^{-1}, \qquad \|\tilde{X}_0(u) - \tilde{X}_0(u')\|_q \leq D|u-u'|.$$

Here, we use $\|Z\|_q := \mathbb{E}[|Z|^q]^{1/q}$ for random variables Z.

The conditions mean that $X_{t,n}$ can be approximated locally, for $|u - \frac{t}{n}| \ll 1$, by a stationary process $\tilde{X}_t(u)$. The continuity condition stated on $u \mapsto \tilde{X}_t(u)$ implies that the stationary approximations vary smoothly over time. This motivates the interpretation of locally stationary processes as processes which change their (approximate) stationary properties smoothly over time. The main properties of $X_{t,n}$ are therefore encoded in the stationary approximations, and it is therefore of

https://doi.org/10.1017/S0266466622000500 Published online by Cambridge University Press

interest to analyze terms of the form $\mathbb{E}g(\tilde{X}_t(u), \tilde{X}_{t-1}(u), \ldots)$ with some function g which are a natural approximation of $\mathbb{E}g(X_{t,n}, X_{t-1,n}, \ldots)$.

More detailed, define $Y_{t,n} := (X_{t,n}, X_{t-1,n}, \ldots, X_{1,n}, 0, 0, \ldots)$ and $\tilde{Y}_t(u) := (\tilde{X}_s(u) : s \leq t)$. Our goal is to estimate functionals of the form

$$u \mapsto G(u) := \mathbb{E}g(\tilde{Y}_t(u)),$$

where $g : \mathbb{R}^{\mathbb{N}} \to \mathbb{R}^d$ is some measurable function (see Definition 2.3) and, in a second step, also compositions

$$F(G(u)),$$

where $F : \mathbb{R}^d \to \mathbb{R}^{\tilde{d}}$ is some known continuous function. Important examples are:

- time-varying covariances, $c(u,k) := \mathbb{E}\tilde{X}_0(u)\tilde{X}_k(u)$ with $g(x_0, x_1, \ldots, x_k) := x_0 x_k$;
- time-varying characteristic functions, $\phi(t,u) := \mathbb{E}e^{it\tilde{X}_0(u)}$, with $g_t(x) = e^{itx}$;

and, for compositions:

- time-varying autocorrelation functions $\gamma(u,k) := \frac{c(u,k)}{c(u,0)} = \frac{\mathbb{E}\tilde{X}_0(u)\tilde{X}_k(u)}{\mathbb{E}[\tilde{X}_0(u)^2]}$;
- moment estimators of time-varying parameters depending on several moments such as time-varying Yule–Walker estimators, but also time-varying moment estimators for nonlinear models such as tvGARCH models (cf. Section 3.5 for two examples).

2.2. Estimators

A standard estimator for $G(u)$ is given by a localized moment estimator,

$$\hat{G}_h(u) := \frac{1}{n}\sum_{t=1}^{n} K_h(t/n - u) \cdot g(Y_{t,n}), \tag{1}$$

where $h \in (0, \infty)$ is some bandwidth, and $K_h(\cdot) := \frac{1}{h}K\big(\frac{\cdot}{h}\big)$ where K is a kernel function belonging to the class $\mathcal{K}$ defined below.

DEFINITION 2.2. *A function K is in the set $\mathcal{K}$ if K is symmetric, nonnegative, Lipschitz continuous, has support $[-\frac{1}{2}, \frac{1}{2}]$ and $\int K(x)\,\mathrm{d}x = 1$. We set $|K|_\infty := \sup_{x \in [-\frac{1}{2}, \frac{1}{2}]} |K(x)|$, $\mu_K := \int K(x)x^2\,\mathrm{d}x$, and $\sigma_K^2 = \int K(x)^2\,\mathrm{d}x$.*

With respect to $F(G(u))$, we will analyze the plug-in estimator

$$F(\hat{G}_h(u)).$$

In the following, we present a general theory about how to obtain asymptotic results for such estimators with a focus on adaptation, i.e., on choosing the bandwidth h. Let $w : [0,1] \to [0, \infty)$ be some weight function with compact support $\subset (0,1)$. This function is introduced to avoid the discussion of boundary issues. Our aim in this paper is to define:

https://doi.org/10.1017/S0266466622000500 Published online by Cambridge University Press

- a "global" selector $\hat{h}$ such that the integrated squared error

  $$d_{ISE}(h) := \int_0^1 |F(\hat{G}_h(u)) - F(G(u))|_2^2 \, w(u) \, \mathrm{d}u \tag{2}$$

  is minimized;
- and a "local" selector $\hat{h}(u)$ such that the squared error

  $$d_{SE}(h,u) := |F(\hat{G}_h(u)) - F(G(u))|_2^2 \tag{3}$$

  is asymptotically minimized for fixed $u \in [0,1]$.

We first verify that $F(\hat{G}_h(u))$ is a consistent estimator of $F(G(u))$. In order to derive this result, we assume that g belongs to the following class $\mathcal{H}(M,\chi,C)$. For some sequence of nonnegative real-valued numbers $\chi = (\chi_i)_{i\in\mathbb{N}}$ and some sequence of complex-valued numbers $x = (x_i)_{i\in\mathbb{N}}$, we set $|x|_\chi := \sum_{i\in\mathbb{N}} \chi_i |x_i|$.

Definition 2.3. *We say that $g : \mathbb{R}^{\mathbb{N}} \to \mathbb{R}$ belongs to the class $\mathcal{H}(M,\chi,C)$ if there exists some $M \geq 1$, some constant $C > 0, \varepsilon > 0$, and some sequence of nonnegative real numbers $\chi = (\chi)_{i\in\mathbb{N}}$ with $\chi_i = O(i^{-2-\varepsilon})$ such that*

$$\sup_{x\neq y} \frac{|g(x)-g(y)|}{|x-y|_\chi \cdot (1+|x|_\chi^{M-1}+|y|_\chi^{M-1})} \leq C.$$

A function $g : \mathbb{R}^{\mathbb{N}} \to \mathbb{R}^d$ ($d \in \mathbb{N}$) is in $\mathcal{H}(M,\chi,C)$ if each component belongs to $\mathcal{H}(M,\chi,C)$.

The condition on g is a Lipschitz-type condition with polynomially growing constants (i.e., the Lipschitz condition is relaxed for larger x and y).

By Hölder's inequality, it is easy to see that the following "invariance principle" of local stationarity holds.

Proposition 2.4. *If $(X_{t,n})_{t=1,\dots,n}$ is locally stationary in the sense of Assumption 2.1 with some $q > 0$, and $g \in \mathcal{H}(M,\chi,C)$, then the same holds for $g(X_{t,n})$ with $q' := \frac{q}{M}$.*

Based on this result, we can use Theorem 2.7 in Dahlhaus et al. (2019) to obtain the following theorem.

Theorem 2.5 (Consistency). *Suppose $X_{t,n}$ satisfies Assumption 2.1 with $q \geq M$. Let $g \in \mathcal{H}(M,\chi,C)$, let F be continuous, and let $nh \to \infty$, $h \to 0$. Then we have*

$$F(\hat{G}_h(u)) \xrightarrow{p} F(G(u)).$$

2.3. The mean squared error and the optimal bandwidth

We now derive the theoretically optimal bandwidth (depending on the unknown G) based on a second-order Taylor expansion of the bias. For the understanding of

https://doi.org/10.1017/S0266466622000500 Published online by Cambridge University Press

our approach, some principle remarks are essential at this point: we treat in this paper the ISE and the SE as defined above and not the mean squared errors MISE and MSE. Despite this, we use, both in our theoretical results and the simulations, the "optimal" bandwidths h_{opt} and $h_{opt}(u)$ obtained by minimizing some approximation of the MISE/MSE (see (7) for $\hat{G}_h(u)$). This will be justified in particular by Corollary 3.3.

However, for compositions $F(\hat{G}_h(u))$, the situation is more tricky since the MISE/MSE may not exist, especially if F has singularities (for example, for correlation functions, the denominator may cause nonexistence of the MISE/MSE). Here, we circumvent the problem by a linearization of F: by using a first-order Taylor expansion of F, we have

$$|F(\hat{G}_h(u)) - F(G(u))|_2^2 = |\hat{G}_h(u) - G(u)|^2_{A_F(G(u))} + O(|\hat{G}_h(u) - G(u)|_2^3), \tag{4}$$

where

$$A_F(G(u)) := \partial_x F(x) \cdot \partial_x F(x)'_{|x=G(u)} \tag{5}$$

is a positive semidefinite matrix. We then minimize the MISE/MSE of the first term which always exists. With this approach, we avoid stronger assumptions on F. Again, the result of Corollary 3.3 is a justification of this procedure.

For a second-order Taylor expansion of the bias, we introduce the following assumption.

Assumption 2.6. $u \mapsto G(u)$ is twice continuously differentiable.

To guarantee the existence of the variance of $\hat{G}_h(u)$, we have to impose conditions on the dependence of $X_{t,n}$. In the setting of Assumption 2.1, it is sufficient to state these assumptions pointwise on the stationary approximations $\tilde{X}_t(u)$. An elegant way to formulate mixing assumptions is via the functional dependence measure by Wu (2005). Suppose that ζ_t, $t \in \mathbb{Z}$ is a sequence of i.i.d. random variables. We put $\mathcal{F}_t := (\zeta_t, \zeta_{t-1}, \ldots)$, $t \geq 0$. Let ζ_t^*, $t \in \mathbb{Z}$ be an independent copy of ζ_t, $t \in \mathbb{Z}$, and put $\mathcal{F}_t^* := (\zeta_t, \zeta_{t-1}, \ldots, \zeta_1, \zeta_0^*, \zeta_{-1}, \zeta_{-2}, \ldots)$, $t \geq 0$.

Assumption 2.7. Let $q > 0$. Assume that, for each $u \in [0,1]$, $\tilde{X}_t(u) = J(\mathcal{F}_t, u)$ with some measurable function J. Suppose that

$$\delta_q^{\tilde{X}(u)}(k) := \|\tilde{X}_k(u) - \tilde{X}_k^*(u)\|_q, \quad k \geq 0, \tag{6}$$

fulfills $\sum_{k=0}^{\infty} \sup_{u \in [0,1]} \delta_q^{\tilde{X}(u)}(k) < \infty$.

With these assumptions, we can now derive the expansions of the squared error.

THEOREM 2.8. *Let $g \in \mathcal{H}(M, \chi, C)$ with some $M \geq 1$. Suppose that Assumption 2.1 is fulfilled for some $q \geq 2M$. Let Assumption 2.6 hold, and let Assumption 2.7 hold with the same q. Let $K \in \mathcal{K}$. Define the so-called long-run* *variance of* $g\big(\tilde{Y}_t(u)\big)$ *as*

$$\Sigma_g(u) = \sum_{k\in\mathbb{Z}} \mathrm{Cov}\big(g(\tilde{Y}_0(u)), g(\tilde{Y}_k(u))\big).$$

Then it holds uniformly in $u \in [\frac{h}{2}, 1-\frac{h}{2}]$ *that*

$$\mathbb{E}|\hat{G}_h(u) - G(u)|_A^2 = \frac{\sigma_K^2}{nh}\cdot \mathrm{tr}\big(A\Sigma_g(u)\big) + \frac{h^4}{4}\mu_K^2 \cdot |\partial_u^2 G(u)|_A^2 + o\big((nh)^{-1} + h^4\big). \qquad \mathbf{(7)}$$

Let F *be continuously differentiable. Then, it holds for all* $u \in (0,1)$ *that, for* $h \to 0$, $nh \to \infty$,

$$\big|F\big(\hat{G}_h(u)\big) - F\big(G(u)\big)\big|_2^2 = \big|\hat{G}_h(u) - G(u)\big|^2_{A_F(G(u))} + o_p\big(h^4 + (nh)^{-1}\big). \qquad \mathbf{(8)}$$

Here, the second result (8) links the squared error $|F(\hat{G}_h(u)) - F(G(u))|_2^2$ to a weighted squared error of $\hat{G}_h(u)$ as announced above. As a consequence of this result, we define

$$d_{approxMSE}(h,u) := \frac{\sigma_K^2}{nh}\,\mathrm{tr}\big(A_F(G(u))\,\Sigma_g(u)\big) + \frac{h^4}{4}\mu_K^2\,|\partial_u^2 G(u)|^2_{A_F(G(u))}$$

and

$$d_{approxMISE}(h) := \int_0^1 d_{approxMSE}(h,u) w(u) du,$$

leading (if $|\partial_u^2 G(u)|^2_{A_F(G(u))} > 0$) easily to the optimal bandwidths

$$h_{opt}(u) := \mathrm{argmin}_{h\in H_n}\, d_{approxMSE}(h,u) = \left(\frac{4\sigma_K^2 \mathrm{tr}(A_F(G(u))\,\Sigma_g(u))}{\mu_K^2 |\partial_u^2 G(u)|^2_{A_F(G(u))}}\right)^{1/5} n^{-1/5} \qquad \mathbf{(9)}$$

and

$$h_{opt} := \mathrm{argmin}_{h\in H_n}\, d_{approxMISE}(h) = \left(\frac{4\sigma_K^2 \int_0^1 \mathrm{tr}(A_F(G(u))\,\Sigma_g(u)) w(u) du}{\mu_K^2 \int_0^1 |\partial_u^2 G(u)|^2_{A_F(G(u))} w(u) du}\right)^{1/5} n^{-1/5}. \qquad \mathbf{(10)}$$

The quantities $h_{opt}(u)$ and h_{opt} are not directly available in practice due to the unknown $G(u)$ and $\Sigma_g(u)$. The use of h_{opt} is justified by Corollary 3.3.

3. GLOBAL ADAPTIVE BANDWIDTH SELECTION: CROSS VALIDATION

We first discuss cross validation for $G(u)$, and afterward for the more technical case $F(G(u))$.

https://doi.org/10.1017/S0266466622000500 Published online by Cambridge University Press

3.1. Cross validation for *G*

The reasoning for our cross-validation approach is at the beginning very similar to the reasoning for classical cross validation (cf. Härdle and Marron, 1985): we attempt to estimate the goodness-of-fit criterion $d_{ISE}(h) = \int_0^1 |G(u) - \hat{G}_h(u)|_2^2 w(u)du$ from (2) by

$$d_{prelimISE,G}^{(n)}(h) := \frac{1}{n}\sum_{t=1}^{n} \left|g(Y_{t,n}) - \hat{G}_h(t/n)\right|_2^2 w(t/n).$$

If $g(Y_{t,n})$ and $\hat{G}_h(t/n)$ were independent (they are not!), we would obtain $\mathbb{E}d_{prelimISE,G}^{(n)}(h) \approx \mathbb{E}d_{ISE}(h) + \frac{1}{n}\sum_{t=1}^{n}\mathbb{E}|g(Y_{t,n}) - G(t/n)|_2^2$, with the latter term being constant in h. As a consequence, we could minimize $d_{prelimISE,G}^{(n)}(h)$ with respect to h for adaptive bandwidth selection.

To overcome the missing independence between $g(Y_{t,n})$ and $\hat{G}_h(t/n)$, one uses in i.i.d. situations a "leave-one-out estimator" instead of $\hat{G}_h(t/n)$ by omitting $Y_{t,n}$.

In the present situation of a dependent sequence $\big(g(Y_{t,n})\big)_t$, this is not sufficient. We have to leave out a larger portion of data to achieve at least an approximate independence between $g(Y_{t,n})$ and the new $\hat{G}_h(t/n)$. These thoughts lead to the following final definitions for our cross-validation procedure: let

$$d_{ISE,G}^{(n)}(h) := \frac{1}{n}\sum_{t=1}^{n} \left|g(Y_{t,n}) - \hat{G}_h^{-}(t/n)\right|_2^2 w(t/n)$$

with

$$\hat{G}_h^{-}(u) := \left(\frac{1}{n}\sum_{t=1}^{n} K_h^{(n)}(t/n-u)\right)^{-1} \cdot \frac{1}{n}\sum_{t=1}^{n} K_h^{(n)}(t/n-u)\cdot g(Y_{t,n}), \tag{11}$$

where, for $\alpha \in (0,1)$ and $\varepsilon > 0$,

$$K^{(n)}(x) := K^{(n),\alpha}(x) := \begin{cases} K(x), & |x| \geq n^{-\alpha}, \\ 0, & |x| \leq (1-\varepsilon)n^{-\alpha}, \\ \frac{n^{\alpha}K(n^{-\alpha})}{\varepsilon}(x-(1-\varepsilon)n^{-\alpha}), & (1-\varepsilon)n^{-\alpha} < x < n^{-\alpha}, \\ \frac{-n^{\alpha}K(-n^{-\alpha})}{\varepsilon}(x+(1-\varepsilon)n^{-\alpha}), & -n^{-\alpha} < x < -(1-\varepsilon)n^{-\alpha}. \end{cases} \tag{12}$$

In this kernel, we use in the case $(1-\varepsilon)n^{-\alpha} < x < n^{-\alpha}$ a linear interpolation between the endpoints $x = (1-\varepsilon)n^{-\alpha}$ and $x = n^{-\alpha}$ such that $K^{(n)}$ is continuous (similarly for the last case). We can interpret $K^{(n)}$ as a Lipschitz continuous approximation of $K(x)\mathbb{1}_{\{|x|\geq n^{-\alpha}\}}$ with Lipschitz constant being of order $O(n)$. Note that $K^{(n)}$ depends on α which is suppressed in the notation for simplicity.

$\hat{G}_h^{-}(u)$ is still a weighted mean of the observations $g(Y_{t,n})$, but observations with $|\frac{t}{n} - u| \ll 1$ are excluded. Note that $\hat{G}_h^{-}(t/n)$ and $g(Y_{t,n})$ are now approximately independent if the sequence $g(Y_{t,n})$, $t = 1,\ldots,n$, fulfills appropriate mixing

https://doi.org/10.1017/S0266466622000500 Published online by Cambridge University Press

conditions. $\hat{h}_G$ is now defined via

$$\hat{h}_G := \operatorname{argmin}_{h\in H_n} d^{(n)}_{ISE,G}(h), \tag{13}$$

where $H_n \subset (0,1)$ is a suitable set of bandwidths. The final estimator of $G(u)$ is then given by $\hat{G}_{\hat{h}_G}(u)$.

3.2. Cross validation for composited functionals $F \circ G$

To obtain a bandwidth selection procedure for the estimator $F(\hat{G}_h(u))$, we use the Taylor expansion in (4). We replace $d^{(n)}_{ISE,G}(h)$ by

$$d^{(n)}_{ISE,F}(h) := \frac{1}{n}\sum_{t=1}^{n} \left|g(Y_{t,n}) - \hat{G}_h^-(t/n)\right|^2_{A_F(\hat{G}_{\hat{h}_G}(t/n))} w(t/n), \tag{14}$$

where $\hat{h}_G$ is the cross-validation bandwidth from (15), and set

$$\hat{h} := \hat{h}_F := \operatorname{argmin}_{h\in H_n} d^{(n)}_{ISE,F}(h). \tag{15}$$

The final estimator of $F\big(G(u)\big)$ then is given by $F\big(\hat{G}_{\hat{h}_F}(u)\big)$.

It is important to state that cross validation for G from above is obtained as a special case: we have for $F = id$ the relation $A_F(x) = I_{\tilde{d}\times\tilde{d}}$, and therefore

$$|\cdot|^2_{A_F(\hat{G}_{\hat{h}_G}(t/n))} = |\cdot|_2^2, \qquad d^{(n)}_{ISE,F}(h) = d^{(n)}_{ISE,G}(h), \qquad \hat{h} = \hat{h}_F = \hat{h}_G. \tag{16}$$

To justify the above approach, we mention that a "naive" cross validation for $F(\hat{G}_h(u))$, say, by minimizing $\frac{1}{n}\sum_{t=1}^n |F(g(Y_{t,n})) - F(\hat{G}_h^-(t/n))|_2^2 w(t/n)$ with respect to h, does not work due to possible singularities of $F(\cdot)$, both theoretically and in simulations. The solution to this issue is the above use of the Taylor expansion (4). $d^{(n)}_{ISE,F}(h)$ as defined above contains twice an estimator of G—besides $\hat{G}_h^-$ also $\hat{G}_{\hat{h}_G}$ in the norm $|\cdot|^2_{A_F(G(t/n))}$. Simulations show that $d^{(n)}_{ISE,F}(h)$ is not very sensitive toward the choice of $\tilde{h}$ in $|\cdot|^2_{A_F(\hat{G}_{\tilde{h}}(t/n))}$ (mathematically, any consistent estimator $\hat{G}_{\tilde{h}}$ of G would be sufficient for our results).

3.3. Properties of the bandwidth selection

We now derive the properties of the adaptive bandwidth selection procedure. The results are formulated for functionals $F\circ G$, but due to (16), cross validation for G is included for $F = id$. In particular, we then have $|\cdot|^2_{A_F(G(u))} = |\cdot|_2^2$ and $\hat{h} = \hat{h}_F = \hat{h}_G$.

THEOREM 3.1. *Let $g \in \mathcal{H}(M,\chi,C)$, where $\chi_i = O(i^{-\kappa})$ with some $\kappa > 3$. Suppose that Assumptions 2.1 and 2.7 hold for all $q > 0$ with $\sup_{u\in[0,1]} \delta_q^{\tilde{X}(u)}(k) = O(k^{-\kappa})$. Let $K \in \mathcal{K}$ and $K^{(n)}$ as in (12). Assume that the support of w is $\subset [\gamma, 1-\gamma]$ with* *some $\gamma > 0$. For arbitrary small $\eta > 0$, let*

$$H_n = [n^{-1+\alpha+\eta}, n^{\min\{2\alpha-1,0\}-\eta}] \tag{17}$$

with α as in (12). Suppose that F is twice continuously differentiable, and

$$\int_0^1 \operatorname{tr}\big(A_F(G(u))\,\Sigma_g(u)\big)\, w(u)\, du > 0, \qquad \int_0^1 \operatorname{tr}\big(\Sigma_g(u)\big)\, w(u)\, du > 0,$$

$$\inf_{n\in\mathbb{N}} \frac{\int_0^1 |\mathbb{E}\hat{G}_h(u) - G(u)|_2^2\, w(u)\, du}{\int_0^1 |\mathbb{E}\hat{G}_h(u) - G(u)|_{A_F(G(u))}^2\, w(u)\, du} > 0. \tag{18}$$

Then, almost surely,

$$\lim_{n\to\infty} \frac{d_{ISE}(\hat{h})}{\inf_{h\in H_n} d_{ISE}(h)} = 1.$$

Theorem 3.1 states that $\hat{h}$, chosen by the cross-validation procedure (13) or (15), is asymptotically optimal in the sense that $\hat{h}$ (or the estimator $F(\hat{G}_{\hat{h}}(u))$, respectively) attains the minimal distance to $F(G(u))$ measured with d_{ISE} over all possible bandwidths $h \in H_n$. Note that we do not impose Assumption 2.6, that is, the result is true even if no explicit bias expansion of order $O(h^2)$ exists. We now give a discussion on the assumptions.

Remark 3.2. • The parameter α which appears in (17) is the same α which is used in the construction of $\hat{G}_h^-$ through the kernel $K^{(n)}$ in (12). We comment on this connection in Section 3.4.

- The conditions stated in (18) guarantee that the bandwidth selection problem is well posed in the sense we now describe. The first condition in (18) is the leading variance contribution of the MSE of $F(\hat{G}_h(u))$ (cf. Theorem 2.8). We request that this variance does not vanish. The second condition means that the variance of $\hat{G}_h(u)$ does not vanish. The third condition in (18) means that $F(\hat{G}_h(u))$ does not have a bias with faster convergence rate than $\hat{G}_h(u)$. Basically, the last two conditions ensure that the estimation of $F(G(u))$ with $F(\hat{G}_h(u))$ is not easier than the estimation of $G(u)$ with $\hat{G}_h(u)$. These last two technical conditions are needed in our proofs but may be omitted in a much more detailed analysis.
- The decay condition on χ and the functional dependence measure $\delta_q^{\tilde{X}(u)}(k)$, namely $\chi_k, \delta_q^{\tilde{X}(u)} = O(k^{-\kappa})$ with $\kappa > 3$, as well as the moment condition ($\mathbb{E}[|\tilde{X}_0(u)|^q] < \infty$ for all $q > 0$) are used to treat the remainder terms in the proof and to apply a chaining device. Prominent recursively defined time series models like tvARMA or tvGARCH processes fulfill $\mathbb{E}[|\tilde{X}_0(u)|^q] < \infty$, for all $q > 0$, if the corresponding i.i.d. innovation process ζ_t satisfies $\mathbb{E}[|\zeta_0|^q] < \infty$, for all $q > 0$, and additional restrictions on the parameter space hold (cf. Dahlhaus et al. (2019) for general recursively defined locally stationary processes or Dahlhaus and Subba Rao (2006) and Francq and Zakoïan (2004) for ARCH and GARCH processes).

- We comment in detail on the parameters ε (from (12)), η, α, and the choice of H_n in Section 3.5.

  We now prove that $\hat{h}$ behaves asymptotically as h_{opt} from (10).

COROLLARY 3.3. *Let the assumptions of Theorem 3.1 hold. Additionally, suppose that Assumption 2.6 holds and $h_{opt} \in H_n$ for n large enough. Then, almost surely,*

$$\frac{d_{ISE}(\hat{h})}{d_{ISE}(h_{opt})} \to 1 \quad \textit{and} \quad \frac{\hat{h}}{h_{opt}} \to 1.$$

Remark 3.4 (Extension to functional estimation). Suppose that instead of $G(u)$, we are interested in estimating a function $\theta \mapsto G_\theta(u)$ with $G_\theta(u) := \mathbb{E} g_\theta(\tilde{Y}_t(u))$, where $g_\theta : \mathbb{R}^{\mathbb{N}} \to \mathbb{R}^d$ and $\theta \in \Theta \subset \mathbb{R}^{\tilde{d}}$. A prominent example is the characteristic function of $\tilde{X}_0(u)$ (cf. Jentsch et al., 2020) where

$$G_\theta(u) = \mathbb{E} e^{i\theta \tilde{X}_0(u)}.$$

We can still use the estimator $\hat{G}_{\theta,h}(u)$ with $\hat{G}_{\theta,h}(u) := \frac{1}{n}\sum_{t=1}^{n} K_h(t/n-u)\cdot g_\theta(Y_{t,n})$ and $\hat{G}_{\theta,h}^{-}$ as in (11). However, we are now interested in minimizing the integrated squared error

$$d_{ISE}^{func}(h) := \int_\Theta \int_0^1 |\hat{G}_{\theta,h}(u) - G_\theta(u)|_2^2 \, w(u) \, \mathrm{d}u \, \mathrm{d}\theta,$$

with an additional integration over θ. Of course, one has to assume $\int_\Theta 1 \, \mathrm{d}\theta > 0$. The cross-validation procedure from (15) can be modified to cover such cases by using an integrated form:

$$d_{ISE}^{func,(n)}(h) := \int_\Theta \frac{1}{n} \sum_{t=1}^{n} \left| g_\theta(Y_{t,n}) - \hat{G}_{\theta,h}^{-}(t/n) \right|_2^2 \mathrm{d}\theta, \quad \hat{h}^{func} := \operatorname{argmin}_{h \in H_n} d_{ISE}^{func,(n)}.$$

If the conditions of Theorem 3.1 hold uniformly in θ, one may then derive the result

$$\lim_{n\to\infty} \frac{d_{ISE}^{func}(\hat{h}^{func})}{\inf_{h\in H_n} d_{ISE}^{func}(h)} = 1 \qquad \text{almost surely}$$

by a straightforward generalization of the proof.

3.4. Discussion on the parameters and an algorithm

Contrary to our announcement at the beginning of the paper, the cross-validation procedure seems to depend on some regularity parameters. We now discuss the influence of these parameters. We will demonstrate that, essentially, only one parameter, namely α with the restriction $h \geq n^{-\alpha}$ for $h \in H_n$ in (17) and the size of the zero set of $K^{(n)}$ being $2n^{-\alpha}$ in (12), remains from a practical view. We will point out why such a condition cannot be avoided and how α can be selected via "eye inspection."

In detail, the estimator $\hat{G}_h^-(u)$ from (11), which is incorporated in $d_{ISE,F}^{(n)}(h)$, depends on the kernel $K^{(n)}$. The kernel depends on the two parameters $\alpha, \varepsilon > 0$, where α controls the zero set of $K^{(n)}$ via

$$|x| \le (1-\varepsilon)n^{-\alpha} \quad \Leftrightarrow \quad K^{(n)}(x) = 0, \tag{19}$$

and ε controls the Lipschitz constant of $K^{(n)}$. Furthermore, the optimization which leads to $\hat{h}$ has to take place over the set

$$H_n = [n^{-1+\alpha+\eta}, n^{\min\{2\alpha-1,0\}-\eta}]$$

with some $\eta > 0$. We now comment on the three parameters $\varepsilon, \alpha, \eta > 0$.

- Choice of ε: In our theoretical proofs of Theorem 3.1, we need that $K^{(n)}$ is Lipschitz-continuous and thus $\varepsilon > 0$. However, using much more involved empirical process results would yield the same theoretical results also for noncontinuous $K^{(n)}$. Thus, it is also theoretically founded to choose $\varepsilon = 0$. In practical simulations, we did not notice any drawback when choosing $\varepsilon = 0$.
- Choice of η: The proof of Theorem 3.1 requires that there exists some $\eta > 0$ such that $(nh)^{-1} = O(n^{-\eta})$ as well as $h = O(n^{-\eta})$ for any $h \in H_n$. We have seen in several simulations that the upper bound in H_n is typically not necessary since $d_{ISE,F}^{(n)}(h)$ does not behave erratically for large $h \in H_n$. However, the lower bound in H_n is important. From Theorem 3.1, we know that the lower bound of H_n has to have the form $n^{-1+\alpha+\eta}$. Thus, the choice of $\alpha+\eta$ can be replaced by the choice of a single parameter $\alpha > 0$ (being the former $\alpha+\eta$). It follows that η is not a tuning parameter and can be ignored in the procedure.

We therefore only have to discuss the parameter α which simultaneously controls the lower bound $n^{-\alpha}$ of H_n and the zero set of $K^{(n)}$ (cf. (19)). Unfortunately, the theoretical conditions of Theorem 3.1 only ask for $\alpha > 0$ arbitrarily small and do not give any hint how to choose it. However, the theoretical result states that for n large enough, $h \mapsto d_{ISE,F}^{(n)}(h)$ has a distinct local minimum. In Figure 1, we have depicted the quantity $d_{ISE,F}^{(n)}(h)$ based on the variance $g(z) = z^2$ for different $h \in [0.01, 1]$ and $\alpha = 0.23, 0.33, 0,48$, leading to $n^{-\alpha} = 0.24, 0.13, 0.05$, respectively, of two realizations of the tvAR(1) process

$$X_{t,n} = a(t/n)X_{t-1,n} + \varepsilon_t, \qquad a(u) := 0.9\sin(2\pi u), \qquad \varepsilon_t \sim N(0,1), \quad n = 500.$$

We first explain the (irrelevant) global minimum at $h \approx 0$ in Figure 1: as $h \to 0$, also the number of points left out (caused by the zero set in $K_h^{(n)}$) tends to zero, and (in case of $g(z) = z^2$ where neighboring values of $g(Y_{t,n})$ are positively correlated) $\hat{G}_h^-(t/n) \approx g(Y_{t,n})$ and $d_{ISE}^{(n)}(h) \approx 0$. Thus, we have to look for a (local) minimum above (say) $h \approx 0.1$.

It can be seen that for smaller α (larger $n^{-\alpha}$), distinct local minima exist. However, for α chosen too small, too many observations are omitted by $K^{(n)}$ and

https://doi.org/10.1017/S0266466622000500 Published online by Cambridge University Press

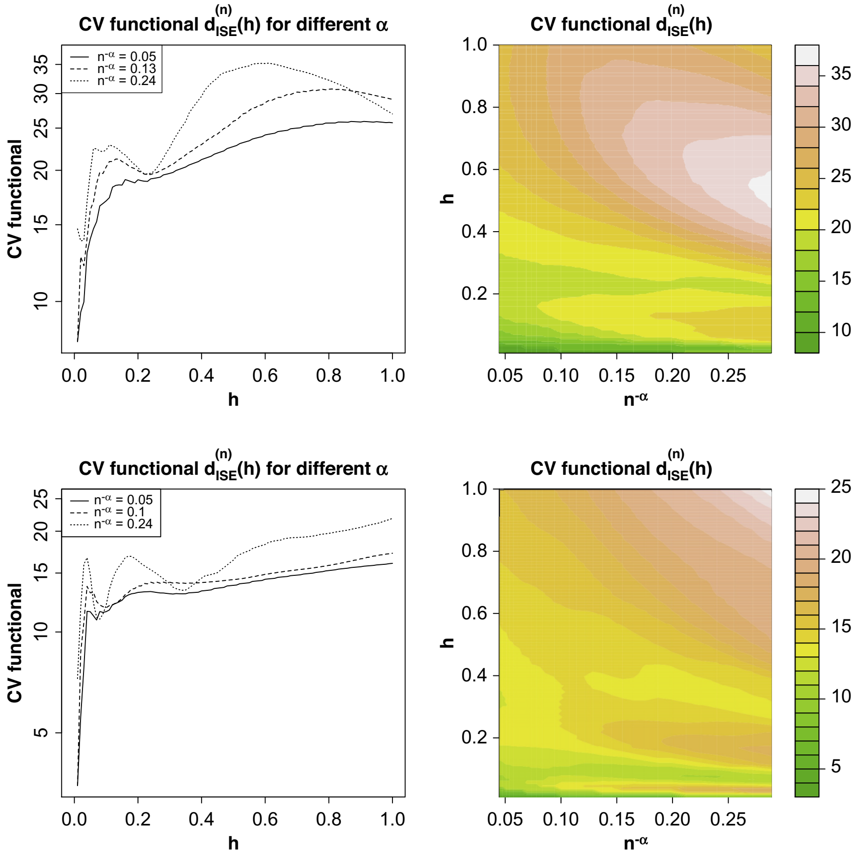

FIGURE 1. Cross-validation functions $d^{(n)}_{ISE,F}(h)$ for different α for two realizations. Top: first realization. Bottom: second realization. Left: logarithmic plot of $h \mapsto d^{(n)}_{ISE,F}(h)$ for three selected α. Right: contour plot of $(n^{-\alpha}, h) \mapsto d^{(n)}_{ISE,F}(h)$.

therefore local minima of $h \mapsto d^{(n)}_{ISE,F}(h)$ may be poor approximations of local minima of $d_{ISE}(h)$. For large α, $h \mapsto d^{(n)}_{ISE,F}(h)$ the local minimum disappears and we only have a global minimum at $h \approx 0$. This happens because the limit behavior of Theorem 3.1 is not yet achieved. Therefore, it is counterproductive to choose $\alpha \in (0,1)$ as large as possible. A heuristic description of the selection procedure therefore reads as follows: choose $\alpha \in (0,1)$ as large as possible *such that* $h \mapsto d^{(n)}_{ISE,F}(h)$ still has a pronounced local minimum which is separated from $h \approx 0$.

We now comment on the realization whose $d^{(n)}_{ISE,F}$ is depicted in the top row of Figure 1. In a practical situation, $n^{-\alpha}$ can be chosen with this information by "eye inspection" of Figure 1, say, from the left plot by looking for a clear local

https://doi.org/10.1017/S0266466622000500 Published online by Cambridge University Press

minimum separated from $h \approx 0$, leading to (say) $n^{-\alpha} = 0.13$ with minimum 0.241; or from the right plot, leading to (say) $n^{-\alpha} = 0.20$ with minimum 0.230. Both choices are adequate in our opinion. For the second realization, which is depicted in the bottom row of Figure 1, $n^{-\alpha} = 0.24$ leads to two local minima at $h \approx 0.11$ and $h \approx 0.36$. The right plot in the bottom row suggests that decreasing $n^{-\alpha}$ to $n^{-\alpha} = 0.10$ (i.e., increasing α) makes the local minimum at $h \approx 0.36$ to disappear which is confirmed in the left plot for $n^{-\alpha} = 0.10$ (dashed line). This leads to the choice $n^{-\alpha} = 0.10$ with minimum $h \approx 0.105$.

However, in larger simulations (say, with $N = 1{,}000$ replications as below), α must either be fixed beforehand or be chosen automatically separately in each simulation step. The problem then arises how to convert the "eye inspection" into a stable objective algorithm. Heuristically, one could term the problem as the search for the smallest local(!) minimum (smallest in terms of location on the h-axis). However, the curve for $n^{-\alpha} = 0.24$ in the top-left plot of Figure 1 shows the difficulty with this definition: below the minimum of interest, it has two small local minima, whose locations in terms of h are smaller.

Based on practical experience with the simulations, we suggest the following algorithm, which makes use of the existence of distinct local minima of $h \mapsto d_{ISE,F}^{(n)}(h)$ for small α, well separated from $h \approx 0$. We mark the dependence on α of the above functional for the moment by $d_{ISE,F,\alpha}^{(n)}(h)$. The algorithm then searches in step l for $h_l = \operatorname{argmin} d_{ISE,F,\alpha_l}^{(n)}(h)$ with $h \in [h_{l-1} - \Delta, h_{l-1} + \Delta]$, where α_l is increasing, and $h_0 = 0.5$. For large α, h_l will gradually drift to zero—but the idea is that the majority of h_l will be either close to the right minimum or close to zero, thus creating two clusters of minima.

Remark 3.5 (Heuristic CV selection algorithm). Choose $H_n \subset (0,1)$, $\varepsilon = 0$, and $\eta = 0$. Let $\{\alpha_1, \dots, \alpha_L\} \subset [-\frac{\log(0.5)}{\log(n)}, 1]$ be a grid of L equidistant numbers between $-\frac{\log(0.5)}{\log(n)}$ and 1. Let $h_0 = \frac{1}{2}$ and $\Delta := \frac{n^{-1/5}}{L}$. Repeat for $l = 1, \dots, L$:

- Find $h_l = \operatorname{argmin}_{h \in [h_{l-1}-\Delta, h_{l-1}+\Delta] \cap H_n} d_{ISE,F,\alpha_l}^{(n)}(h)$.

Cluster the elements of $(h_l)_{l=1,\dots,L}$ into two clusters with a 2-means algorithm, and choose the larger center of the two clusters as $\hat{h}$.

Since K has support $[-0.5, 0.5]$, α has to be larger than $-\frac{\log(0.5)}{\log(n)}$ so that $K^{(n)}$ is not constant 0.

3.5. Simulations

Since our estimators are model-free, we expect good behavior of the selection procedure for a wide range of locally stationary processes. Here, we inspect tvAR, tvMA, and tvARCH processes:

$$X_{t,n}^{(1)} = a(t/n) \cdot X_{t-1,n}^{(1)} + \sigma(t/n)\zeta_t,$$
$$a(u) = 0.9\sin(2\pi u), \quad \sigma(u) = 0.9 + 0.5\cos(2\pi u),$$

https://doi.org/10.1017/S0266466622000500 Published online by Cambridge University Press

$$X_{t,n}^{(2)} = \zeta_t + b_1(t/n)\zeta_{t-1} + b_2(t/n)\zeta_{t-2},$$
$$b_1(u) = 0.9\sin(2\pi u), \quad b_2(u) = 0.7\cos(2\pi u),$$
$$X_{t,n}^{(3)} = \big(a_1(t/n) + a_2(t/n)(X_{t-1,n}^{(3)})^2\big)^{1/2}\zeta_t,$$
$$a_1(u) = 0.5 + 0.4\sin(2\pi u), \quad a_2(u) = 0.3 + 0.25\sin(2\pi u),$$

where ζ_t are i.i.d. $N(0,1)$. We will estimate the following quantities:

(a) For $X_{t,n}^{(1)}$:
- $c(u,1) = \mathbb{E}\tilde{X}_2(u)\tilde{X}_1(u) \left[= \frac{a(u)}{1-a(u)^2}\right]$,
- $(c(u,1), c(u,0)) := (\mathbb{E}\tilde{X}_2(u)\tilde{X}_1(u), \mathbb{E}\tilde{X}_1(u)^2) \left[= \left(\frac{a(u)}{1-a(u)^2}, \frac{1}{1-a(u)^2}\right)\right]$,
- $\frac{c(u,1)}{c(u,0)} = \frac{\mathbb{E}\tilde{X}_2(u)\tilde{X}_1(u)}{\mathbb{E}\tilde{X}_1(u)^2} \left[= a(u)\right]$.

(b) For $X_{t,n}^{(2)}$:
- $c(u,2) = \mathbb{E}\tilde{X}_3(u)\tilde{X}_1(u) \left[= b_2(u)\right]$,
- $\frac{c(u,2)}{1+c(u,1)} := \frac{\mathbb{E}\tilde{X}_3(u)\tilde{X}_1(u)}{1+\mathbb{E}\tilde{X}_2(u)\tilde{X}_1(u)} \left[= b_1(u)\right]$.

(c) For $X_{t,n}^{(3)}$:
- $c(u,0) = \mathbb{E}\tilde{X}_1(u)^2 \left[= \frac{a_1(u)}{1-a_2(u)}\right]$,
- $\frac{\mathrm{Cov}(\tilde{X}_2(u)^2, \tilde{X}_1(u)^2)}{\mathrm{Cov}(\tilde{X}_1(u)^2, \tilde{X}_1(u)^2)} \left[= a_2(u)\right]$.

(note that the expressions in the [ ]-brackets are not used in the simulations). In all simulations, we use $w(\cdot) = \mathbb{1}_{[0.05,0.95]}(\cdot)$, $H_n = [0.01, 0.6]$, and the Epanechnikov kernel $K(x) = \frac{3}{2}(1-(2x)^2)\mathbb{1}_{[-\frac{1}{2},\frac{1}{2}]}(x)$. We use a time series length of $n = 500$ and the algorithm from Remark 3.5 to select α and therefore also $\hat{h}$. We judge the behavior of this selector by reporting the empirical quantiles of $d_{ISE}(\hat{h})$ over the $N = 1{,}000$ replications.

For comparison, we also report the quantiles of the two *optimal* selectors $d_{ISE}(h_{opt})$ with h_{opt} from (10) and $\min_{h\in H_n} d_{ISE}(h)$ which serve as a benchmark (cf. Tables 1–3). We usually have

$$d_{ISE}(\hat{h}) \geq d_{ISE}(h_{opt}) \geq \min_{h\in H_n} d_{ISE}(h),$$

where the first inequality comes from the fact that h_{opt} is deterministic and is not subject to an additional estimation procedure as it is the case for $\hat{h}$.

Thus, the goal of the simulation study is to find out how close the cross-validation values $\hat{h}$ and $d_{ISE}(\hat{h})$ are to the optimal (but unknown) selectors.

An inspection of the simulation results in Tables 1–3 shows that the median of $d_{ISE}(\hat{h})$ over all $N = 1{,}000$ replications has the same order of magnitude as the median of $d_{ISE}(h_{opt})$ and $\min_{h\in H_n} d_{ISE}(h)$ with a slightly larger variation. When comparing the median of $d_{ISE}(\hat{h})$ and $\min_{h\in H_n} d_{ISE}(h)$, the selection procedure works best for the estimation of $c(u,2)$ and $\frac{c(u,2)}{1+c(u,1)}$ in the tvMA example $X_{t,n}^{(2)}$ (cf. Table 1). In Figure 2, we have considered estimation of $c(u,2)$ and depicted

https://doi.org/10.1017/S0266466622000500 Published online by Cambridge University Press

Table 1. Empirical quantiles of the distances $d_{ISE}(\hat{h})$ of the CV selector, $d_{ISE}(h_{opt})$ of the MSE-optimal deterministic bandwidth, and $\min_{h\in H_n} d_{ISE}(h)$ for the tvMA model $X_{t,n}^{(2)}$ with $n = 500$ based on $N = 1,000$ replications.

Quantity	Selector	Empirical quantiles						
		0.05	0.10	0.25	0.5	0.75	0.9	0.95
$c(u,2)$	$d_{ISE}(\hat{h})$	0.020	0.027	0.04	0.06	0.088	0.117	0.144
	$d_{ISE}(h_{opt})$	0.016	0.023	0.035	0.053	0.079	0.103	0.124
	$\min_{h\in H_n} d_{ISE}(h)$	0.012	0.019	0.03	0.048	0.07	0.096	0.113
$\frac{c(u,2)}{1+c(u,1)}$	$d_{ISE}(\hat{h})$	0.059	0.067	0.083	0.104	0.128	0.157	0.181
	$d_{ISE}(h_{opt})$	0.057	0.064	0.079	0.100	0.120	0.144	0.159
	$\min_{h\in H_n} d_{ISE}(h)$	0.054	0.062	0.076	0.095	0.115	0.135	0.147

Table 2. Empirical quantiles of the distances $d_{ISE}(\hat{h})$ of the CV selector, $d_{ISE}(h_{opt})$ of the MSE-optimal deterministic bandwidth, and $\min_{h\in H_n} d_{ISE}(h)$ for the tvAR model $X_{t,n}^{(1)}$ with $n = 500$ based on $N = 1,000$ replications.

Quantity	Selector	Empirical quantiles						
		0.05	0.10	0.25	0.5	0.75	0.9	0.95
$c(u,1)$	$d_{ISE}(\hat{h})$	0.376	0.475	0.717	0.999	1.409	1.911	2.982
	$d_{ISE}(h_{opt})$	0.365	0.448	0.621	0.898	1.263	1.716	2.372
	$\min_{h\in H_n} d_{ISE}(h)$	0.283	0.368	0.550	0.842	1.204	1.630	2.098
$(c(u,1),c(u,0))$	$d_{ISE}(\hat{h})$	0.876	1.032	1.482	2.134	2.970	3.946	5.269
	$d_{ISE}(h_{opt})$	0.719	0.873	1.264	1.796	2.525	3.493	4.651
	$\min_{h\in H_n} d_{ISE}(h)$	0.563	0.745	1.128	1.686	2.411	3.295	4.119
$\frac{c(u,1)}{c(u,0)}$	$d_{ISE}(\hat{h})$	0.006	0.007	0.010	0.014	0.019	0.025	0.030
	$d_{ISE}(h_{opt})$	0.005	0.006	0.008	0.011	0.015	0.019	0.022
	$\min_{h\in H_n} d_{ISE}(h)$	0.005	0.006	0.008	0.010	0.014	0.018	0.021

an histogram of $\hat{h}$ and $\operatorname{argmin}_{h\in H_n} d_{ISE}(h)$ over the $N = 1,000$ replications. One can see that the values of $\hat{h}$ fluctuate nicely around h_{opt} (red line). The reason is that the tvMA process is m-dependent with $m = 3$; therefore, it is only necessary to guarantee $n^{-\alpha} \geq 6$ to eliminate all dependencies which occur in the cross-validation functional for the estimation of $c(u,2)$ or $c(u,1)$. We therefore expect $h \mapsto d_{ISE,F}^{(n)}(h)$ to have a distinct local minimum for nearly all $\alpha \in (0,1)$.

For the tvAR(1) process $X_{t,n}^{(1)}$ and estimation of $c(u,1)$ and $(c(u,1),c(u,0))$, we observe rather large values for all three quantities $d_{ISE}(\hat{h})$, $d_{ISE}(h_{opt})$, and $\min_{h\in H_n} d_{ISE}(h)$. The reason is twofold: first, even the theoretical quantities attain values of around 10 for some $u \in [0,1]$, and thus a larger error is comprehensible. Second, both quantities $c(u,1)$ and $(c(u,1),c(u,0))$ are relatively hard to estimate

TABLE 3. Empirical quantiles of the distances $d_{ISE}(\hat{h})$ of the CV selector, $d_{ISE}(h_{opt})$ of the MSE-optimal deterministic bandwidth, and $\min_{h\in H_n} d_{ISE}(h)$ for the tvARCH model $X_{t,n}^{(3)}$ with $n = 500$ based on $N = 1{,}000$ replications.

Quantity	Selector	Empirical quantiles						
		0.05	0.10	0.25	0.5	0.75	0.9	0.95
$c(u,0)$	$d_{ISE}(\hat{h})$	0.025	0.034	0.059	0.092	0.136	0.205	0.344
	$d_{ISE}(h_{opt})$	0.027	0.034	0.053	0.083	0.123	0.187	0.307
	$\min_{h\in H_n} d_{ISE}(h)$	0.014	0.021	0.038	0.07	0.113	0.177	0.273
$\frac{\mathrm{Cov}(\tilde{X}_2(u)^2,\tilde{X}_1(u)^2)}{\mathrm{Cov}(\tilde{X}_1(u)^2,\tilde{X}_1(u)^2)}$	$d_{ISE}(\hat{h})$	0.012	0.015	0.022	0.034	0.048	0.064	0.075
	$d_{ISE}(h_{opt})$	0.011	0.014	0.021	0.031	0.046	0.060	0.070
	$\min_{h\in H_n} d_{ISE}(h)$	0.010	0.012	0.018	0.027	0.041	0.052	0.062

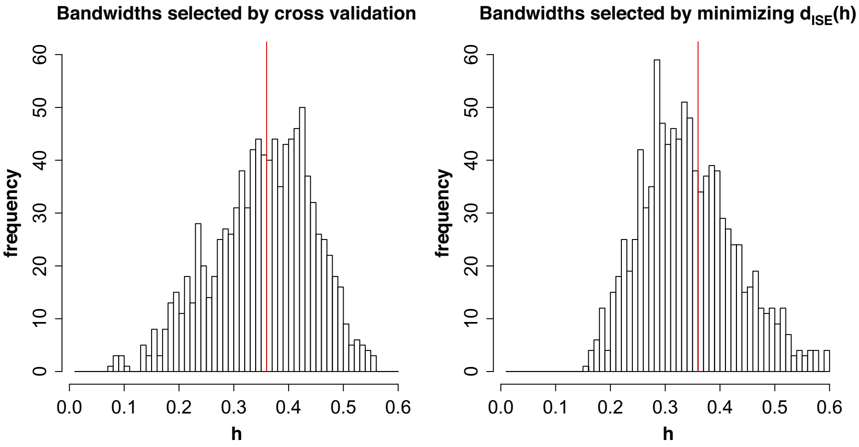

FIGURE 2. tvMA model $X_{t,n}^{(2)}$, estimation of $c(u,2)$. Left: histogram of the cross-validation-selected bandwidths over $N = 1{,}000$ replications. Right: histogram of $\mathrm{argmin}_{h\in H_n} d_{ISE}(h)$ over $N = 1{,}000$ replications. The red vertical line marks h_{opt} in both histograms.

in the sense that $\hat{G}_h(u)$ is a poor estimator for $G(u)$ *for all* bandwidths $h \in H_n$. A large variation of $\hat{G}_h(u)$ clearly leads to a larger variation of $\hat{h}$ which explains the slightly larger medians of $d_{ISE}(\hat{h})$. In Figure 3, we have depicted histograms of $\hat{h}$ and $\mathrm{argmin}_{h\in H_n} d_{ISE}(h)$ for the $N = 1{,}000$ replications. In comparison to $c(u,1)$ and $(c(u,1),c(u,0))$, estimation of $\frac{c(u,1)}{c(u,0)}$ works quite well. The reason is that the composited moment estimator $F(\hat{G}_h(u))$ of $\frac{c(u,1)}{c(u,0)} = a(u)$ can be written as a maximum likelihood estimator; thus, the difference $F(\hat{G}_h(u)) - F(G(u))$ enjoys a martingale difference property. In this case, the cross-validation functional $d_{ISE}^{(n)}$ works nearly as well as in the i.i.d. case (cf. Richter and Dahlhaus, 2019).

https://doi.org/10.1017/S0266466622000500 Published online by Cambridge University Press

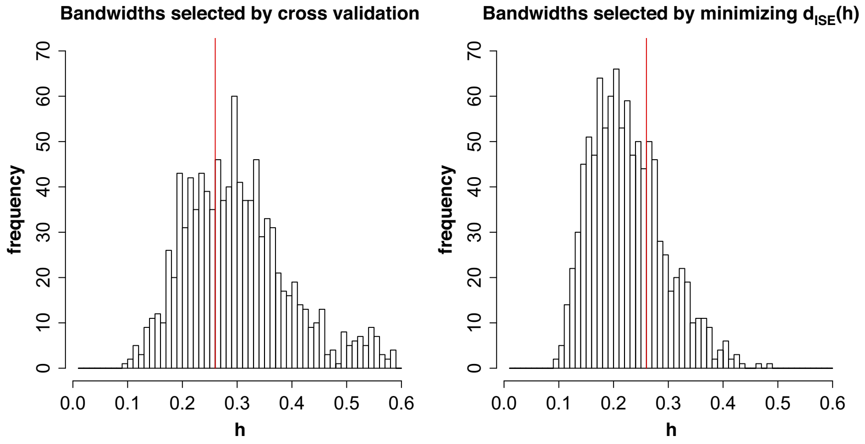

Figure 3. tvAR model $X_{t,n}^{(1)}$ and estimation of $(c(u,0),c(u,1))$. Left: histogram of the cross-validation-selected bandwidths over $N = 1{,}000$ replications. Right: histogram of $\operatorname{argmin}_{h \in H_n} d_{ISE}(h)$ over $N = 1{,}000$ replications. The red vertical line marks h_{opt} in both histograms.

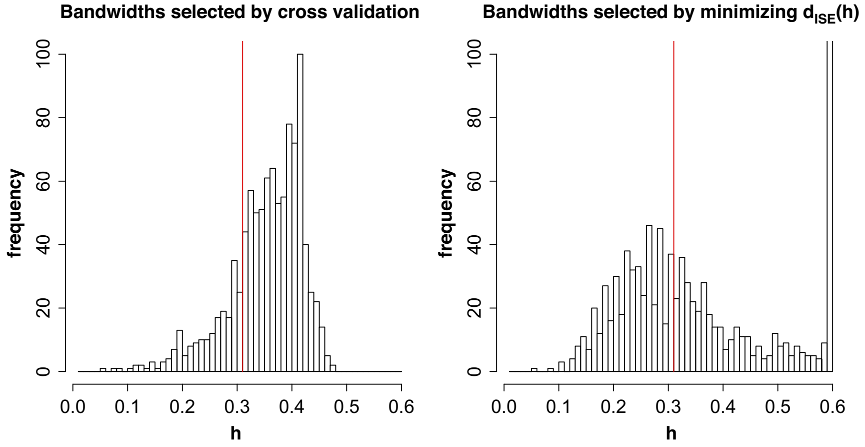

Figure 4. tvARCH model $X_{t,n}^{(3)}$ and estimation of $\frac{\mathrm{Cov}(\tilde{X}_2(u)^2,\tilde{X}_1(u)^2)}{\mathrm{Cov}(\tilde{X}_1(u)^2,\tilde{X}_1(u)^2)}$. Left: histogram of the cross-validation-selected bandwidths over $N = 1{,}000$ replications. Right: histogram of $\operatorname{argmin}_{h \in H_n} d_{ISE}(h)$ over $N = 1{,}000$ replications. The red vertical line marks h_{opt} in both histograms.

In Table 3, we see that the selection procedure $\hat{h}$ also works quite well for the tvARCH process $X_{t,n}^{(3)}$. Estimation in tvARCH processes (in particular parameter estimation) is typically hard due to the more complicated dependence structure. In Figure 4, we have considered estimation of $\frac{\mathrm{Cov}(\tilde{X}_2(u)^2,\tilde{X}_1(u)^2)}{\mathrm{Cov}(\tilde{X}_1(u)^2,\tilde{X}_1(u)^2)}$ and depicted histograms of $\hat{h}$ and $\operatorname{argmin}_{h \in H_n} d_{ISE}(h)$ for the $N = 1{,}000$ replications. Here, we

https://doi.org/10.1017/S0266466622000500 Published online by Cambridge University Press

see that for several realizations, $\operatorname{argmin}_{h\in H_n} d_{ISE}(h)$ is chosen as the largest possible bandwidth, which indicates that $\hat{G}_h(u)$ is a poor estimator of $G(u)$ (at least for this relatively small sample size $n = 500$). However, $\hat{h}$ tries to mimic this behavior of choosing larger bandwidths, which leads to comparable values of $d_{ISE}(\hat{h})$ and $\min_{h\in H_n} d_{ISE}(h)$.

Overall, we see that $\hat{h}$ behaves quite satisfactorily in all examples. If $G(u)$ or $F(G(u))$ is hard to estimate, that is, $\hat{G}_h(u) - G(u)$ or $F(\hat{G}_h(u)) - F(G(u))$ are large for all $h \in H_n$, then $\hat{h}$ suffers from a larger deviation.

In practice, when only a single time series has to be analyzed, we recommend choosing α by eye inspection of one of the plots in Figure 1, and then $\hat{h}$ by the local minimum of the corresponding curve $d^{(n)}_{ISE,\alpha}(h)$. One just has to choose α large, but small enough such that the local minimum of the $d^{(n)}_{ISE,\alpha}(h)$ is well shaped and clearly distinct from $h \approx 0$. Note that the selected $\hat{h}$ then is quite insensitive to the choice of α. The heuristic algorithm from Remark 3.5 may support this choice.

4. LOCAL MODEL SELECTION: A CONTRAST MINIMIZATION APPROACH

The approach presented in the following allows to choose h locally for each $u \in (0,1)$ in the estimator $\hat{G}_h(u)$. This enables the procedure to take into account local smoothness changes of the function $G(u)$. The algorithm is based on a contrast minimization approach which was introduced by Lepski et al. (1997) for a different model. In the following, we use a slightly modified estimator for $G(u)$, namely

$$\hat{G}_h(u) := \frac{\sum_{t=1}^n K_h(t/n-u)g(Y_{t,n})}{\sum_{t=1}^n K_h(t/n-u)},$$

which corrects for the deviation of the kernel sum from its integral $\int K(x)\,\mathrm{d}x = 1$. The kernel K is now assumed to be from the class $\mathcal{K}'$ given by the following definition. Here, $\partial_1 K$ denotes the almost surely existing derivative of the Lipschitz continuous function K.

DEFINITION 4.1. *A function K is in the set $\mathcal{K}'$ if K is symmetric, nonnegative, Lipschitz continuous, satisfies $\{x \in \mathbb{R} : K(x) > 0\} = (-\frac{1}{2}, \frac{1}{2})$, and fulfills $\int K(x)\,dx = 1$, $\sigma_K^2 := \int K(x)^2 dx > 0$, $\mu_K = \int K(x)x^2 dx > 0$, $\int \{y \cdot \partial_1 K(y) + K(y)\}^2 dy > 0$.*

In this section, we restrict ourselves to estimation of $G(u)$ instead of $F(G(u))$.

The approach presented in the following can be interpreted as an iterative algorithm. It compares $\hat{G}_h(u) - \hat{G}_{h'}(u)$, $h' \le h$ with the variance of $\hat{G}_h(u)$ for several bandwidths h from a grid. It then selects the bandwidth h where the former terms are roughly of the same size. The algorithm therefore needs an estimator of the variance of $\hat{G}_h(u)$. From Theorem 2.8, we know that

$$\operatorname{Var}\big(\hat{G}_h(u)\big) = \frac{\sigma_K^2}{nh}\operatorname{tr}\big(\Sigma_g(u)\big) + o\Big(\frac{1}{nh}\Big) \quad \text{where} \quad \Sigma_g(u) = \sum_{k\in\mathbb{Z}} \operatorname{Cov}(g(\tilde{Y}_0(u)), g(\tilde{Y}_k(u))$$

https://doi.org/10.1017/S0266466622000500 Published online by Cambridge University Press

is the long-run variance of $g(\tilde{Y}_0(u))$. Let $\hat{\Sigma}_n(u)$ be an estimator of $\Sigma_g(u)$. We discuss a suitable choice of $\hat{\Sigma}_n(u)$ below. The local bandwidth selection procedure is defined as follows. Let $\lambda(h) := \max\{1, \sqrt{\log(1/h)}\}$ and

$$\hat{v}^2(h,u) := \frac{\sigma_K^2}{nh}\mathrm{tr}(\hat{\Sigma}_n(u)).$$

Let $H_n \subset (0,1]$ be a geometrically decaying grid of bandwidths given by

$$H_n = \{a^k : k \in \mathbb{N}_0\} \cap [\underline{h}, \overline{h}], \tag{20}$$

where $a \in (0,1)$ and $\underline{h}, \overline{h} > 0$ are specified below. For some $C^\# > 0$, define

$$\hat{h}(u) :\in \sup\{h \in H_n : |\hat{G}_h^\circ(u) - \hat{G}_{h'}^\circ(u)|_2 \leq C^\# \cdot \hat{v}(h',u)\lambda(h') \text{ for all } h' < h, h' \in H_n\}. \tag{21}$$

Remark 4.2. $\hat{v}^2(h',u)$ is a proxy for the variance of $\hat{G}_h^\circ(u) - \hat{G}_{h'}^\circ(u)$. It is multiplied by a log factor $\lambda(h')$ in (21) to account for local random deviations. As described in Lepski et al. (1997), the bandwidth (21) can be seen as the largest bandwidth where $\hat{G}_h^\circ(u)$ does not deviate significantly from $G(u)$.

We now present the assumptions on the process $\left(\tilde{X}_t(u)\right)_{t\in\mathbb{Z}}$ under which the theoretical statements for the local bandwidth selection procedure $\hat{h}(u)$ holds.

Assumption 4.3 (Moment and dependence assumptions). Let $\alpha \geq 0$. Let Assumption 2.1 hold for all $q \geq 1$. Define $N_\alpha(q) := \Gamma(\alpha q + 2)$, where $\Gamma(x) = \int_0^\infty e^{-t}t^{x-1}dt$ denotes the Gamma function. Assume that, for all $u \in [0,1]$:

(i) $\sup_{u\in[0,1]} \|\tilde{X}_0(u)\|_q \leq D \cdot N_\alpha(q)$,
(ii) $\sup_{u\in[0,1]} \delta_q^{\tilde{X}(u)}(k) \leq D \cdot N_\alpha(q) \cdot \rho^k$.

Remark 4.4. Assumption 4.3(i) asks for a quantification of the growth of the moments $\|\tilde{X}_0(u)\|_q^q$ in q. It can be easily seen that the condition $\sup_{u\in[0,1]} \|\tilde{X}_0(u)\|_q \leq D \cdot N_\alpha(q)$ follows if $\tilde{X}_0(u)$ has a Lebesgue density which is bounded from above by $K\exp(-x^{1/\alpha})$. Assumption 4.3(ii) asks the process to have a geometrically decaying dependence measure $\delta_q^{\tilde{X}(u)}(k)$. This is, for instance, fulfilled for tvARMA and tvGARCH processes (cf. Wu, 2011).

We now formulate assumptions on the pre-estimator $\hat{\Sigma}_n(u)$ of $\Sigma_g(u)$.

Assumption 4.5 (Assumptions on $\hat{\Sigma}_n(u)$). There exists some constant $c_\Sigma > 0$ such that, for each $u \in [0,1]$:

- $\hat{\Sigma}_n(u)$ is consistent in the following sense: for every $\varepsilon > 0$, $\mathbb{P}(|\hat{\Sigma}_n(u) - \Sigma_g(u)|_\infty > \varepsilon) = O(n^{-2})$.
- $\| \, |\hat{\Sigma}_n(u)|_\infty \, \|_2 \leq c_\Sigma$.

We now have the following near-optimality result for $\hat{h}(u)$.

https://doi.org/10.1017/S0266466622000500 Published online by Cambridge University Press

THEOREM 4.6. *Let $g \in \mathcal{H}(M,\chi,C)$ be such that, for all $j \geq 0$, $|\chi_j| \leq \kappa\rho^j$ with some $\rho \in (0,1)$, $\kappa > 0$. Suppose that $K \in \mathcal{K}'$.*

Fix $u \in (0,1)$. Suppose that Assumptions 2.6, 4.3 (with the same ρ as above), and 4.5 are fulfilled. Suppose that $|\partial_u^2 G(u)|_2 > 0$ and $\Sigma_g(v)$ is positive definite for all $v \in [0,1]$.

Then there exist constants $c_H, C_H, c' > 0$ independent of n and some universal constant $c > 0$ such that $\hat{h}(u)$ defined in (21) has the following property: if $[\underline{h},\overline{h}] = [c_H \log(n)^{5+2\alpha M} n^{-1}, C_H]$ and $C^{\#} \geq 64$, then

$$\mathbb{E}|\hat{G}_{\hat{h}(u)}(u) - G(u)|_2^2$$
$$\leq c' \log(n)^2 n^{-1} + \frac{c}{1-a} \cdot \min_{h \in H_n} \left\{ \frac{h^4}{4} \mu_K^2 |\partial_u^2 G(u)|_2^2 + \sigma_K^2 \mathrm{tr}(\Sigma(u)) \frac{\log(n)}{nh} \right\}. \qquad \textbf{(22)}$$

Remark 4.7.
- The theorem states that the mean squared error $\mathbb{E}|\hat{G}_{\hat{h}(u)}(u) - G(u)|_2^2$ with the selector $\hat{h}(u)$ is bounded by two terms: the first term is of the smaller rate $\log(n)^2 n^{-1}$ and negligible compared to the second term. The second term is a universal constant times the *minimal* mean squared error of $\mathbb{E}|\hat{G}_h(u) - G(u)|_2^2$ (cf. (7) in Theorem 2.8) up to an additional log-factor $\log(n)$ in the variance part.
- The set $H_n = \{a^k : k \in \mathbb{N}\} \cap [\underline{h},\overline{h}]$ contains all theoretically interesting bandwidths. Especially, $h_{opt}(u)$ from (9), which has rate $n^{-1/5}$, is included for n large enough in the sense that there exists $h \in H_n$ such that $ah \leq h_{opt}(u) \leq h$.
- The positivity assumptions on $|\partial_u^2 G(u)|_2$ and $\Sigma_g(u)$ are needed to ensure that both leading terms of the mean squared error decomposition exist.
- The proof can be adopted in such a way that $C^{\#} \geq 64$ can be replaced by $C^{\#} \geq 1+\varepsilon$ with $\varepsilon > 0$ arbitrarily small.
- The geometric decay in Assumption 4.3 on the functional dependence measure is used to derive a Bernstein inequality (see Theorem 4.8) which is connected to the $\log(n)$ term in the variance in (22). The existence of all moments in the prescribed way is necessary to guarantee the result with $\underline{h} = c_H \log(n)^{5+2\alpha M} n^{-1}$. In principle, both assumptions may be relaxed. This then leads to larger $\underline{h}$ (for instance, additional polynomial factors in n) and a larger additional factor in the variance in (22). As pointed out in Remark 3.2, these assumptions are fulfilled for several recursively defined time series models like tvARMA and tvGARCH under appropriate conditions on the underlying parameter space and the distribution of the innovations.
- Assumption 4.5 basically asks that $\hat{\Sigma}_n(u)$ is consistent for $\Sigma_g(u)$. The main message here is that $\hat{\Sigma}_n(u)$ has not to be "optimal" in any sense, but it is sufficient if $\hat{\Sigma}_n(u)$ is of the same order as $\Sigma_g(u)$. We present reasonable candidates in Section 3.5.

Assumption 4.3 is used to prove a Bernstein-type inequality which is a key ingredient to obtain optimality results for contrast minimization methods. The geometric decay of the functional dependence measure is needed to apply a

https://doi.org/10.1017/S0266466622000500 Published online by Cambridge University Press

Bernstein-type result from Doukhan and Neumann (2007), while the moment conditions are necessary to allow for a large set of bandwidths H_n by establishing a simple exponential inequality. The Bernstein inequality is formulated in terms of the process

$$\tilde{G}_h(u) := \frac{\sum_{t=1}^{n} K_h(t/n-u)\cdot g(\tilde{Y}_t(t/n))}{\sum_{t=1}^{n} K_h(t/n-u)},$$

which is a theoretical approximation of $\hat{G}_h(u)$.

THEOREM 4.8 (Bernstein inequality for $\tilde{G}_h(u)$). *Fix* $u \in [0,1]$. *Assume that* $g \in \mathcal{H}(M,\chi,C)$ *and Assumption 4.3 holds. Then there exist some constants* $c_1, c_2, c_H, C_H > 0$ *independent of* n *such that, for all* $h \in [c_H n^{-1}, C_H]$ *and all* $j \in \{1,\ldots,d\}$,

$$\mathbb{P}\Big((nh)\big|\tilde{G}_h(u)_j - \mathbb{E}\tilde{G}_h(u)_j\big| > \gamma\Big)$$
$$\leq 2\exp\Big(-\frac{\gamma^2}{32(nh)^2 v_j^2(h,u) + c_1\log(n)^{\alpha M/3}\gamma^{5/3}}\Big) + c_2\frac{n^{-2}}{\gamma^2},$$

where $v_j^2(h,u) := \frac{\sigma_K^2}{nh}\Sigma(u)_{jj}$.

Remark 4.9. The above Bernstein inequality states that the deviation $\tilde{G}_h(u)_j - \mathbb{E}\tilde{G}_h(u)_j$ can be bounded by the deviation of a Gaussian distribution with variance $v_j^2(h,u) = 64\frac{\sigma_K^2}{nh}\Sigma(u)_{jj}$ for a certain regime of the threshold γ where

$$32(nh)^2 v_j^2(h,u) \geq c_1\log(n)^{\alpha M/3}\gamma^{5/3}.$$

"Optimal" Bernstein inequalities for independent variables are formulated with γ instead of $\gamma^{5/3}$. Here, we obtain $\gamma^{5/3}$ due to dependence by using a result from Doukhan and Neumann (2007). It turns out that this result is just adequate for the above model selection result. c_1, c_2 are complicated functions of the constants given in the assumptions. In the proof, one has to approximate $g(\tilde{Y}_t(t/n))$ by $g(\tilde{Y}_t(u))$ in the variance term leading to the condition $h \in [c_H n^{-1}, C_H]$. This condition basically asks $\tilde{G}_h(u)$ to contain at least a certain finite number of summands and is further needed to neglect bias-type terms which arise in the variance decomposition. The additional summand $c_2\frac{n^{-2}}{\gamma^2}$ should be regarded as a remainder term which arises by excluding rare events $\{g(\tilde{Y}_t(t/n)) > \log(n)\}$.

4.1. Discussion on the tuning parameters

Theorem 4.6 states a nonasymptotic result which provides near optimality of $\hat{h}(u)$ under conditions on H_n, $\hat{\Sigma}_n$, and $C^{\#}$. However, the constants $c' > 0$, $c > 0$ arising in (22) may be relatively large, which undermines the nice theoretical statement

https://doi.org/10.1017/S0266466622000500 Published online by Cambridge University Press

for moderate sample sizes n. For practical purposes, a more detailed discussion on the influence of the parameters on the procedure is necessary.

The selector $\hat{h}(u)$ depends on several parameters and pre-estimators, namely:

- the choice of $\hat{\Sigma}_n$,
- the choice of $C^{\#}$,
- the set H_n, especially $a \in (0,1)$ and $\underline{h}, \overline{h}$.

4.1.1. *Choice of* $\hat{\Sigma}_n(u)$. As mentioned above, the requirements on the quality of $\hat{\Sigma}_n(u)$ from Assumption 4.5 are not too strong; however, we need that $\hat{\Sigma}_n(u)$ has the same order of magnitude as $\Sigma_g(u)$. A possible choice for $\hat{\Sigma}_n(u)$ is given by

$$\hat{\Sigma}_n(u) := \sum_{k=-r_n}^{r_n} \hat{c}_\eta^g(u,k), \tag{23}$$

where

$$\hat{c}_\eta^g(u,k) := \frac{\frac{1}{n}\sum_{t=1}^n K_\eta(t/n-u)\cdot\{g(Y_{t,n})-\hat{G}_\eta(u)\}\{g(Y_{t-k,n})-\hat{G}_\eta(u)\}'}{\frac{1}{n}\sum_{t=1}^n K_\eta(t/n-u)}$$

with some $r_n \in \mathbb{N}$, $\eta > 0$. The following result holds.

LEMMA 4.10. *Let the assumptions of Theorem 4.6 hold. Let $s > 0$. Then there exist constants $c_r, C_r, c_\eta, C_\eta > 0$ independent of n such that for any $r_n \in [c_r \log(n), C_r \log(n)]$ and $\eta \in [c_\eta n^{-1+s}, C_\eta n^{-s}]$, $\hat{\Sigma}_n$ satisfies Assumption 4.5.*

Note that, here, we allow for a large area of $\eta \in [c_\eta n^{-1+s}, C_\eta n^{-s}]$, that is, no optimal choice of η is needed. In practice, one may use a rather ad hoc choice via

$$r_n \approx \log(n), \qquad \eta \approx n^{-1/5} \tag{24}$$

($n^{-1/5}$ is motivated by the MSE-optimal bandwidth for twice continuously differentiable objectives) to obtain stable results.

4.1.2. *Choice of* H_n. H_n depends on $\underline{h}$, $\overline{h}$, and a. As seen in Theorem 4.6, there are restrictions on $\underline{h}$, $\overline{h}$ of the form

$$\underline{h} \geq c_H \log(n)^{5+2\alpha M}, \qquad \overline{h} \leq C_H$$

with some (potentially large) constant $c_H > 0$ and some (potentially small) constant $C_H > 0$. Since $\hat{h}(u)$ is a supremum and therefore formulated as a bottom-up procedure and therefore already for smaller $h \in H_n$ the condition in $\hat{h}(u)$ is violated, in practice an upper bound on H_n via $\overline{h}$ is *not necessary*.

Clearly, a choice $a \approx 1$ is preferable in practice to obtain a fine grid H_n of possible bandwidths, especially for small samples sizes n. In Theorem 4.6 and (22), it is seen that the choice of a has a direct influence of the quality of $\hat{h}(u)$ due

https://doi.org/10.1017/S0266466622000500 Published online by Cambridge University Press

to the factor $\frac{1}{1-a}$ on the right-hand side. Generally speaking, $a \in (0,1)$ should be chosen roughly in such a way that

$$\frac{1}{1-a} \le 4, \qquad \text{that is,} \qquad a \ge \frac{3}{4},$$

which restricts the impact of this factor in (22).

Lastly, the choice of $\underline{h}$ in practice is strongly connected to the choice of $C^{\#}$, which is discussed in the next section.

4.2. Choice of $\underline{h}$ from H_n and choice of $C^{\#}$

The selector $\hat{h}(u)$ is quite sensitive to the choice of $C^{\#}$. In this way, the whole method should be interpreted as a procedure which is able to reduce the choice of several tuning parameters (bandwidth choices for all $u \in (0,1)$) to one tuning parameter $C^{\#}$. Theorem 4.6 states that $C^{\#} \ge 64$ should provide good results at least for large n, while in principle this condition can be reduced to $C^{\#} \ge 1$ (cf. the remark after the theorem). We now investigate the choice of $C^{\#}$ connected to the choice of $\underline{h}$ for estimation of $G(u) = c(u,1) = \mathbb{E}\tilde{X}_0(u)\tilde{X}_1(u)$ in the tvAR(1) process

$$X_{t,n} = a(\frac{t}{n})X_{t-1,n} + \varepsilon_t, \qquad a(u) = 0.8\sin(2\pi u), \qquad \varepsilon_t \sim N(0,1), \qquad n = 1{,}000. \tag{25}$$

In Figure 5, we depict the behavior of $\hat{h}(u)$ and the corresponding estimator $\hat{G}_{\hat{h}(u)}(u)$ based on the estimator $\hat{\Sigma}_n$ defined in (23) with $\eta = 0.3$, $r_n = 5$, and $a = \frac{3}{4}$. We have chosen $C^{\#} \in \{0.5, 1, 2\}$ and $H_n = \{a^k : k \in \{0,\dots,10\}\}$ and $H_n = \{a^k : k \in \{0,\dots,20\}\}$. We see that for $C^{\#} = 1$ and $H_n = \{a^k : k \in \{0,\dots,10\}\}$, the estimator works quite nicely and detects changes of the smoothness behavior of the underlying curve $G(u)$ by using smaller or larger bandwidths. If the set $H_n = \{a^k : k \in \{0,\dots,20\}\}$ includes much smaller bandwidths, one can observe several "overshoots," that is, the smallest possible bandwidth is selected at certain $u \in (0,1)$, whereas for values $v \approx u$ near u, $\hat{h}(v)$ is much larger. This is due to the fact that for too small bandwidths (and thus, too few observations inside the sum $\hat{G}_h(u)$), the perturbation theory based on the Bernstein inequality does no longer hold. If $C^{\#} = 2$, one can see that the estimator acts quite conservatively and chooses only large bandwidths. If $C^{\#} = 0.5$, the estimator already tends to select too small bandwidths.

Similar observations can be made for other processes (see Section 3.5). In general, it is a good start to choose $C^{\#} = 1$. A direct connection between an optimal choice of $C^{\#}$ and the sample size n as well as the properties of the underlying process is not obvious to us. One may choose slightly smaller values $C^{\#} < 1$ after a first inspection of the estimator, for instance, $C^{\#} \in [0.5, 1]$. Here, one also has to adapt the lower bound $\underline{h}$ of H_n to stabilize $\hat{h}(u)$. More precisely, one has to select $\underline{h}$ large enough such that no overshoots arise. We summarize this procedure in the following heuristic algorithm.

https://doi.org/10.1017/S0266466622000500 Published online by Cambridge University Press

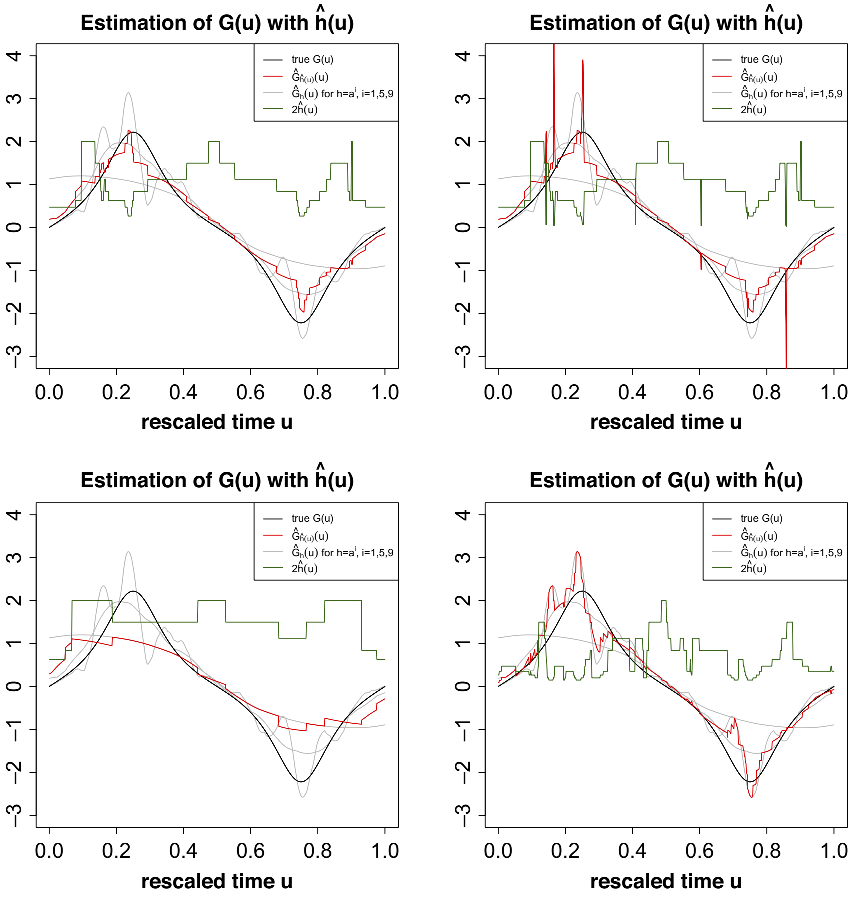

FIGURE 5. Behavior of $\hat{h}(u)$ for estimation of $G(u) = \mathbb{E}\tilde{X}_0(u)\tilde{X}_1(u)$ for one observation of the process given in (25). Top left: $C^{\#} = 1$ and $H_n = \{(\frac{3}{4})^k : k \in \{0,\ldots,10\}\}$; top right: $C^{\#} = 1$, $H_n = \{(\frac{3}{4})^k : k \in \{0,\ldots,20\}\}$; bottom left: $C^{\#} = 2$, $H_n = \{(\frac{3}{4})^k : k \in \{0,\ldots,10\}\}$; bottom right: $C^{\#} = 0.5$, $H_n = \{(\frac{3}{4})^k : k \in \{0,\ldots,10\}\}$.

Remark 4.11 (Heuristic algorithm for local bandwidth selection). We summarize our finding in the following heuristic algorithm, which applies to a large range of time series.

Let $r_n = \lceil \log(n) \rceil$, $\eta = n^{-1/5}$, and $a = \frac{3}{4}$. Put $C^{\#} = 1$, $H_{n,Z} = \{a^k : k \in \{0,\ldots,N\}\}$ for $Z \in \mathbb{N}$ with $Z_0 = \lceil -\frac{\log(n)}{\log(a)} \rceil$. Iterate for $Z = Z_0, Z_0 - 1, \ldots$:

- For $u \in (0,1)$, calculate $\hat{h}(u)$ from (21) with $H_{n,Z}$.
- Check if $\hat{G}_{\hat{h}(u)}(u)$ or $\hat{h}(u)$, respectively, have "overshoots." If not, stop the procedure.

The selector $\hat{h}(u)$ based on $H_{n,Z}$ is the final selector.

https://doi.org/10.1017/S0266466622000500 Published online by Cambridge University Press

4.3. Simulations

Since our estimators are model-free, we expect good behavior of the selection procedure for a wide range of locally stationary processes. Here, we inspect tvAR, tvMA, and tvARCH processes:

$$\begin{aligned}
X_{t,n}^{(1)} &= a(t/n)\cdot X_{t-1,n}^{(1)}+\zeta_t, \qquad a(u)=0.8-1.6\cdot \mathbb{1}_{[0,0.5]}(u),\\
X_{t,n}^{(2)} &= \zeta_t+b_1(t/n)\zeta_{t-1}+b_2(t/n)\zeta_{t-2},\\
&\qquad\qquad b_1(u)=0.9\sin(2\pi u), \quad b_2(u)=3\sin(2\pi u),\\
X_{t,n}^{(3)} &= \big(a_1(t/n)+0.5(X_{t-1,n}^{(3)})^2\big)^{1/2}\zeta_t, \qquad a_1(u)=0.5+0.4\cos(2\pi u),
\end{aligned}$$

where ζ_t are i.i.d. $N(0,1)$. We will estimate the following quantities:

(a) for $X_{t,n}^{(1)}$: $c(u,1)=\mathbb{E}\tilde{X}_2(u)\tilde{X}_1(u)\Big[=\frac{a(u)}{1-a(u)^2}\Big]$,

(b) for $X_{t,n}^{(2)}$: $c(u,2)=\mathbb{E}\tilde{X}_3(u)\tilde{X}_1(u)\Big[=b_2(u)\Big]$,

(c) for $X_{t,n}^{(3)}$: $c(u,0)=\mathbb{E}\tilde{X}_1(u)^2\Big[=\frac{a_1(u)}{1-a_2(u)}\Big]$.

In all simulations, we use the parameters from the algorithm in Remark 4.11 and the Epanechnikov kernel $K(x)=\frac{3}{2}(1-(2x)^2)\mathbb{1}_{[-\frac{1}{2},\frac{1}{2}]}(x)$. We use a time series of length $n=1,000$ and restrict ourselves to

$$H_{n,Z}=\{a^k : k\in\{0,\ldots,Z\}\}, \qquad Z=10. \tag{26}$$

We do $N=1,000$ replications and investigate the behavior of certain local bandwidth selectors $u\mapsto h(u)$ via $d_{SE}(h,u)$ from (3) (with $F=id$). For a bandwidth curve $h(\cdot)$, we therefore obtain with $w(u)=\mathbb{1}_{[0.05,0.95]}(u)$ the integrated squared error

$$\tilde{d}_{ISE}(h(\cdot))=\int_0^1 d_{SE}(h(u),u)w(u)du=\int_0^1 \mathbb{E}|\hat{G}_{h(u)}(u)-G(u)|_2^2 w(u)du,$$

which differs from $d_{ISE}(h)$ given in (2) through the possibility to insert different bandwidths $h(u)$ for every $u\in[0,1]$. We compare $\tilde{d}_{ISE}(\hat{h}(\cdot))$ with

$$\min_{h\in H_n}\tilde{d}_{ISE}(h)=\min_{h\in H_n}d_{ISE}(h), \tag{27}$$

which reflects the minimal value of $\tilde{d}_{ISE}(h(\cdot))$ a *global* bandwidth selector could achieve (that is, only one bandwidth $h(\cdot)=h\in H_n$ is chosen for all $u\in[0,1]$) and

$$\min_{h:[0,1]\to H_n}\tilde{d}_{ISE}(h(\cdot)), \tag{28}$$

which reflects the minimal value of $\tilde{d}_{ISE}(h(\cdot))$ a *local* bandwidth selector could achieve (that is, for every $u\in[0,1]$ a bandwidth $h(u)\in H_n$ is chosen). The comparison with $\min_{h\in H_n}\tilde{d}_{ISE}(h)$ and $\min_{h:[0,1]\to H_n}\tilde{d}_{ISE}(h)$ allows to judge how far $\hat{h}(u)$ is away from a global and a local optimal selection procedure. Furthermore, we can analyze if at least to some extent, $\hat{h}(u)$ outperforms a global optimal selection procedure. Note that both distances (27) and (28) are achieved with

https://doi.org/10.1017/S0266466622000500 Published online by Cambridge University Press

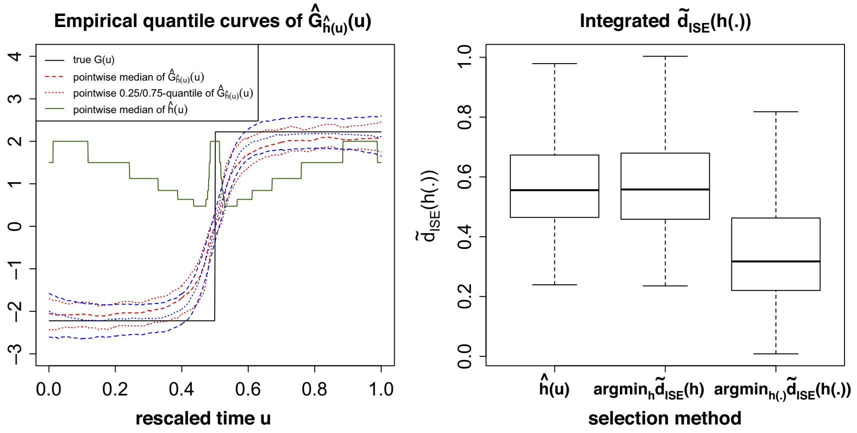

FIGURE 6. Behavior of $\hat{h}(u)$ for estimation of $G(u) = \mathbb{E}\tilde{X}_0(u)\tilde{X}_1(u)$ based on the tvAR process $X_{t,n}^{(1)}$. Left: empirical pointwise quantile curves of $\hat{G}_{\hat{h}(u)}(u)$ (red) and $\hat{G}_{h^*}(u)$ (blue) together with the median of $\hat{h}(u)$ (green) over $N = 1{,}000$ replications. Right: boxplots of the achieved distances $\tilde{d}_{ISE}(h)$ for $\hat{h}(u)$, a global and a local optimal selector.

knowing the true function G and therefore do not correspond to bandwidth selectors which are available in practice. In general, we expect

$$\tilde{d}_{ISE}(\hat{h}(\cdot)), \min_{h\in H_n} \tilde{d}_{ISE}(h) \geq \min_{h:[0,1]\to H_n} \tilde{d}_{ISE}(h(\cdot)),$$

but nothing can be said about the relation between

$$\tilde{d}_{ISE}(\hat{h}(\cdot)) \quad \text{and} \quad \min_{h\in H_n} \tilde{d}_{ISE}(h).$$

If the left distance is smaller than the right distance, this clearly shows that $\hat{h}(\cdot)$ outperforms any global bandwidth selection procedure. However, even if the distances are of comparable size, this is a remarkable result since the local estimation method $\hat{h}(u)$ has no access to $G(u)$ and suffers from the additional estimation error in the selection procedure. To analyze this in more detail, let

$$h^* \in \text{argmin}_{h\in H_n} d_{ISE}(h) = \text{argmin}_{h\in H_n} \tilde{d}_{ISE}(h)$$

be the bandwidth of a global optimal selector.

In Figure 6, we have depicted empirical quantile curves of $\hat{G}_{\hat{h}(u)}(u)$ and $\hat{G}_{h^*}(u)$ as well as the median of $\hat{h}(u)$ (scaled with the factor 2) when applied to model (a). It can be seen that, on average, $\hat{h}(u)$ adapts quite nicely to the smoothness of $G(u)$. While for values $u \in (0,1)$ near the boundaries $0, 1$, $\hat{h}(u)$ is chosen large to use the constancy of $G(u)$ to reduce the variance, it is getting smaller toward the step at $u = 0.5$. The "anomaly" of a large bandwidth choice for $u = 0.5$ comes from the fact that $G(0.5) = 0 = \int G(u)du$ can be estimated best when averaging over all

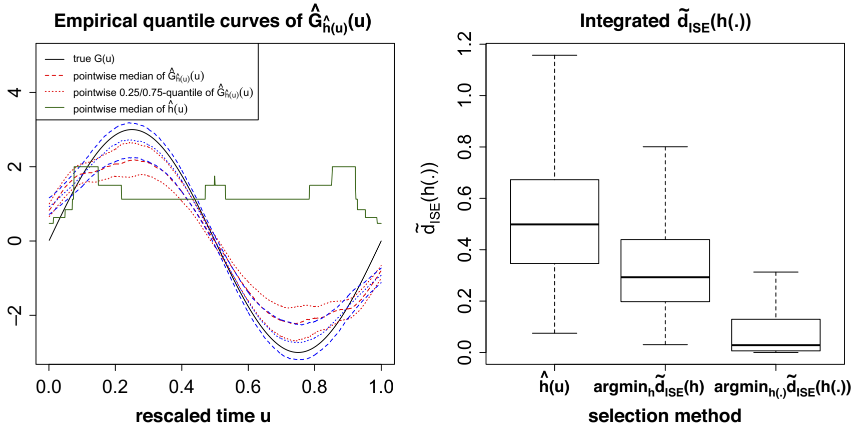

Figure 7. Behavior of $\hat{h}(u)$ for estimation of $G(u) = \mathbb{E}\tilde{X}_0(u)\tilde{X}_2(u)$ based on the tvMA process $X_{t,n}^{(2)}$. Left: empirical pointwise quantile curves of $\hat{G}_{\hat{h}(u)}(u)$ (red) and $\hat{G}_{h^*}(u)$ (blue) together with the median of $\hat{h}(u)$ (green) over $N = 1{,}000$ replications. Right: boxplots of the achieved distances $\tilde{d}_{ISE}(h)$ for $\hat{h}(u)$, a global and a local optimal selector.

observations. Compared with the global optimal selector $\hat{G}_{h^*}(u)$, we observe that $\hat{G}_{\hat{h}(u)}(u)$ has a smaller variance for boundary points $u \in (0,1)$ near $0,1$. From the boxplot, we see that $\tilde{d}_{ISE}(\hat{h}(\cdot))$ and $\min_{h\in H_n} \tilde{d}_{ISE}(h)$ have comparable size, but are clearly larger than $\min_{h:[0,1]\to H_n} \tilde{d}_{ISE}(h(\cdot))$.

In Figure 7, the simulations results for model (b) are depicted. Here, the objective $G(u)$ is much smoother than in model (a). It can be seen that $\hat{h}(u)$ adapts to the smoothness of $G(u)$, but in general is too conservative, leading to too large bandwidths. Here, a smaller $C^{\#}$ may have led to better results. Since $G(u)$ is smooth over the whole interval $u \in [0,1]$, the estimator based on the global optimal selector $\hat{G}_{h^*}(u)$ outperforms $\hat{G}_{\hat{h}(u)}(u)$. However, from the boxplot, we see that $\tilde{d}_{ISE}(\hat{h}(\cdot))$ and $\min_{h\in H_n} \tilde{d}_{ISE}(h)$ still have comparable size.

In Figure 8, the simulations results for model (c) are depicted. Here, the objective $G(u)$ is smooth. Again, it can be seen that $\hat{h}(u)$ adapts to the smoothness of $G(u)$ and is too conservative. However, it has a better performance than the estimator based on the global optimal selector $\hat{G}_{h^*}(u)$ for $u \approx 0.5$ where a smaller bandwidth is necessary to capture the hill of $G(u)$. From the boxplot, we see that $\tilde{d}_{ISE}(\hat{h}(\cdot))$ and $\min_{h\in H_n} \tilde{d}_{ISE}(h)$ still have comparable size, whereas $\min_{h:[0,1]\to H_n} \tilde{d}_{ISE}(h(\cdot))$ is much smaller.

Let us summarize the following points. First of all, comparison with $\min_{h:[0,1]\to H_n} \tilde{d}_{ISE}(h(\cdot))$ is “unfair” since for any $u \in (0,1)$ there may exist a good $h(u) \in H_n$ to minimize $|\hat{G}_{h(u)}(u) - G(u)|$ even if $\hat{G}_h(u) - G(u)$ does not behave like the nonasymptotic theory derived. Therefore, it should not be overestimated that $\min_{h:[0,1]\to H_n} \tilde{d}_{ISE}(h(\cdot))$ is much smaller in all examples. Second, the choice

https://doi.org/10.1017/S0266466622000500 Published online by Cambridge University Press

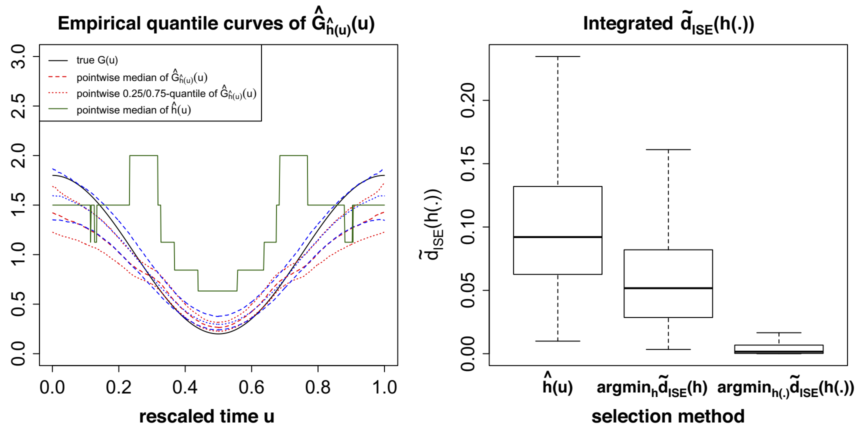

Figure 8. Behavior of $\hat{h}(u)$ for estimation of $G(u) = \mathbb{E}\tilde{X}_0(u)^2$ based on the tvARCH process $X_{t,n}^{(3)}$. Left: empirical pointwise quantile curves of $\hat{G}_{\hat{h}(u)}(u)$ (red) and $\hat{G}_{h^*}(u)$ (blue) together with the median of $\hat{h}(u)$ (green) over $N = 1{,}000$ replications. Right: boxplots of the achieved distances $\tilde{d}_{ISE}(h)$ for $\hat{h}(u)$, a global and a local optimal selector.

$C^\# = 1$ and a large lower bound $\underline{h}$ on H_n as given in (26) produce stable estimators $\hat{h}(u)$ in general. However, $\hat{h}(u)$ may select too large bandwidths to compete with a global bandwidth selector. Therefore, one typically has to consider also smaller choices of $C^\#$. Even if $G_{\hat{h}(u)}(u)$ itself may not be useful, $\hat{h}(u)$ clearly captures some information on the smoothness behavior of $G(u)$ and therefore can serve as a pre-estimator for other selection procedures.

5. CONCLUSION

In this paper, we have developed two methods for adaptive bandwidth selection for nonparametric moment estimators of locally stationary processes of some curve $G(u)$. We have derived theoretical results for their optimality with respect to mean-squared-error-type distance measures and found with simulations that they work well for different time series models.

The first method is based on a cross-validation approach and allows for global bandwidth selection. A critical issue is to deal with the dependency of the observed time series whose influence is controlled by some parameter α. This parameter can be selected by eye inspection from a plot of the empirical integrated squared error $d_{ISE,\alpha}^{(n)}(h)$ as a function of h for different α. The selected bandwidth $\hat{h}$ then is quite insensitive to the choice of α. The selection of α may be supported by the heuristic algorithm from Remark 3.5.

The second method is for local bandwidth selection, i.e., for each time point a different bandwidth is chosen. This allows for taking into account local smoothness properties of the unknown curve. The method needs an estimator $\hat{\Sigma}_n(u)$ of the

https://doi.org/10.1017/S0266466622000500 Published online by Cambridge University Press

asymptotic long-run variance and is also dependent on a tuning parameter $C^{\#}$. We have theoretically justified that the quality of $\hat{\Sigma}_n(u)$ does not influence the bandwidth selection procedure very much, while the choice of $C^{\#}$ and the lower bound of the set of bandwidths H_n is important for meaningful results. With theoretical justifications, we have derived a heuristic algorithm to obtain a stable procedure.

From an abstract view, we have presented two new methods for global and local bandwidth selection, but at the expense of introducing with α and $C^{\#}$ two new regularity parameters. The important point is, however, that in the case of cross validation, the new tuning parameter α is much less influential on the quality of the final estimate than the bandwidth, and, in the case of the local bandwidth, the selection of several bandwidths ($h(u)$ for all $u \in (0,1)$) is replaced by the selection of the single tuning parameter $C^{\#}$. Additionally, some heuristic methods are presented which allow for choosing both parameters reasonably.

One may try to improve the theory for the presented methods by allowing for more general structures of $G(u)$ and its estimators or investigating $G(u)$ with moments of two-sided functions. These problems are left to further research.

SUPPLEMENTARY MATERIAL

Dahlhaus, R. and Richter, S. (2022). Supplement to “Adaptation for nonparametric estimators of locally stationary processes,” Econometric Theory Supplementary Material. To view, please visit: https://doi.org/10.1017/S0266466622000500.

REFERENCES

Amado, C. & T. Teräsvirta (2013) Modelling volatility by variance decomposition. *Journal of Econometrics* 175(2), 142–153.

Amado, C. & T. Teräsvirta (2017) Specification and testing of multiplicative time-varying GARCH models with applications. *Econometric Reviews* 36(4), 421–446.

Arkoun, O. (2011) Sequential adaptive estimators in nonparametric autoregressive models. *Sequential Analysis* 30(2), 229–247.

Arkoun, O. & S. Pergamenchtchikov (2016) Sequential robust estimation for nonparametric autoregressive models. *Sequential Analysis* 35(4), 489–515.

Beran, J. (2009) On parameter estimation for locally stationary long-memory processes. *Journal of Statistical Planning and Inference* 139(3), 900–915.

Dahlhaus, R. (1997) Fitting time series models to nonstationary processes. *Annals of Statistics* 25(1), 1–37.

Dahlhaus, R. & L. Giraitis (1998) On the optimal segment length for parameter estimates for locally stationary time series. *Journal of Time Series Analysis* 19(6), 629–655.

Dahlhaus, R., S. Richter, & W.B. Wu (2019) Towards a general theory for nonlinear locally stationary processes. *Bernoulli* 25(2), 1013–1044.

Dahlhaus, R. & S. Subba Rao (2006) Statistical inference for time-varying ARCH processes. *Annals of Statistics* 34(3), 1075–1114.

Doukhan, P. & M.H. Neumann (2007) Probability and moment inequalities for sums of weakly dependent random variables, with applications. *Stochastic Processes and their Applications* 117(7), 878–903.

https://doi.org/10.1017/S0266466622000500 Published online by Cambridge University Press

Francq, C. & J.-M. Zakoïan (2004) Maximum likelihood estimation of pure GARCH and ARMA-GARCH processes. *Bernoulli* 10(4), 605–637.

Giraud, C., F. Roueff, & A. Sanchez-Perez (2015) Aggregation of predictors for nonstationary sub-linear processes and online adaptive forecasting of time varying autoregressive processes. *Annals of Statistics* 43(6), 2412–2450.

Härdle, W. & J.S. Marron (1985) Optimal bandwidth selection in nonparametric regression function estimation. *Annals of Statistics* 13(4), 1465–1481.

Jentsch, C., A. Leucht, M. Meyer, & C. Beering (2020) Empirical characteristic functions-based estimation and distance correlation for locally stationary processes. *Journal of Time Series Analysis* 41(1), 110–133.

Koo, B. & O. Linton (2012) Estimation of semiparametric locally stationary diffusion models. *Journal of Econometrics* 170(1), 210–233.

Lepski, O.V., E. Mammen, & V.G. Spokoiny (1997) Optimal spatial adaptation to inhomogeneous smoothness: An approach based on kernel estimates with variable bandwidth selectors. *Annals of Statistics* 25(3), 929–947.

Mallat, S., G. Papanicolaou, & Z. Zhang (1998) Adaptive covariance estimation of locally stationary processes. *Annals of Statistics* 26(1), 1–47.

Niedzwiecki, M., M. Ciolek, & Y. Kajikawa (2017) On adaptive covariance and spectrum estimation of locally stationary multivariate processes. *Automatica* 82, 1–12.

Priestley, M.B. (1965) Evolutionary spectra and non-stationary processes. *Journal of the Royal Statistical Society, Series B* 27, 204–237 (with discussion).

Priestley, M.B. (1988) *Nonlinear and Nonstationary Time Series Analysis*. Academic Press, Inc. [Harcourt Brace Jovanovich, Publishers].

Richter, S. & R. Dahlhaus (2019) Cross validation for locally stationary processes. *Annals of Statistics* 47(4), 2145–2173.

Roueff, F., R. von Sachs, & L. Sansonnet (2016) Locally stationary Hawkes processes. *Stochastic Processes and their Applications* 126(6), 1710–1743.

Subba Rao, S. (2006) On some nonstationary, nonlinear random processes and their stationary approximations. *Advances in Applied Probability* 38(4), 1155–1172.

Vogt, M. (2012) Nonparametric regression for locally stationary time series. *Annals of Statistics* 40(5), 2601–2633.

Wu, W.B. (2005) Nonlinear system theory: Another look at dependence. *Proceedings of the National Academy of Sciences USA* 102(40), 14150–14154.

Wu, W.B. (2011) Asymptotic theory for stationary processes. *Statistics and Its Interface* 4(2), 207–226.

Wu, W.B. & Z. Zhou (2011) Gaussian approximations for non-stationary multiple time series. *Statistica Sinica* 21(3), 1397–1413.

Zhou, Z. & W.B. Wu (2009) Local linear quantile estimation for nonstationary time series. *Annals of Statistics* 37(5B), 2696–2729.

https://doi.org/10.1017/S0266466622000500 Published online by Cambridge University Press