The Annals of Statistics
2012, Vol. 40, No. 5, 2601–2633
DOI: 10.1214/12-AOS1043

NONPARAMETRIC REGRESSION FOR LOCALLY STATIONARY TIME SERIES[1]

BY MICHAEL VOGT

University of Cambridge

In this paper, we study nonparametric models allowing for locally stationary regressors and a regression function that changes smoothly over time. These models are a natural extension of time series models with time-varying coefficients. We introduce a kernel-based method to estimate the time-varying regression function and provide asymptotic theory for our estimates. Moreover, we show that the main conditions of the theory are satisfied for a large class of nonlinear autoregressive processes with a time-varying regression function. Finally, we examine structured models where the regression function splits up into time-varying additive components. As will be seen, estimation in these models does not suffer from the curse of dimensionality.

1. Introduction. Classical time series analysis is based on the assumption of stationarity. However, many time series exhibit a nonstationary behavior. Examples come from fields as diverse as finance, sound analysis and neuroscience.

One way to model nonstationary behavior is provided by the theory of locally stationary processes introduced by Dahlhaus; cf. [5, 6] and [7]. Intuitively speaking, a process is locally stationary if over short periods of time (i.e., locally in time) it behaves in an approximately stationary way. So far, locally stationary models have been mainly considered within a parametric context. Usually, parametric models are analyzed in which the coefficients are allowed to change smoothly over time.

There is a considerable amount of papers that deal with time series models with time-varying coefficients. Dahlhaus et al. [8], for example, study wavelet estimation in autoregressive models with time-dependent parameters. Dahlhaus and Subba Rao [9] analyze a class of ARCH models with time-varying coefficients. They propose a kernel-based quasi-maximum likelihood method to estimate the parameter functions; a kernel-based normalized-least-squares method is suggested by Fryzlewicz et al. [10]. Hafner and Linton [12] provide estimation theory for a multivariate GARCH model with a time-varying unconditional variance. Finally, a diffusion process with a time-dependent drift and diffusion function is investigated in Koo and Linton [14].

Received October 2011; revised July 2012.
[1]Supported by the DFG-SNF research group FOR916.
MSC2010 subject classifications. Primary 62G08, 62M10; secondary 62G20.
Key words and phrases. Local stationarity, nonparametric regression, smooth backfitting.

In this paper, we introduce a nonparametric framework which can be regarded as a natural extension of time series models with time-varying coefficients. In its most general form, the model is given by

$$Y_{t,T} = m\left(\frac{t}{T}, X_{t,T}\right) + \varepsilon_{t,T} \qquad \text{for } t = 1, \dots, T \tag{1}$$

with $\mathbb{E}[\varepsilon_{t,T}|X_{t,T}] = 0$, where $Y_{t,T}$ and $X_{t,T}$ are random variables of dimension 1 and d, respectively. The model variables are assumed to be locally stationary and the regression function as a whole is allowed to change smoothly over time. As usual in the literature on locally stationary processes, the function m does not depend on real time t but rather on rescaled time $\frac{t}{T}$. This goes along with the model variables forming a triangular array instead of a sequence. Throughout the Introduction, we stick to an intuitive concept of local stationarity. A technically rigorous definition is given in Section 2.

There is a wide range of interesting nonlinear time series models that fit into the general framework (1). An important example is the nonparametric autoregressive model

$$Y_{t,T} = m\left(\frac{t}{T}, Y_{t-1,T}, \dots, Y_{t-d,T}\right) + \varepsilon_{t,T} \qquad \text{for } t = 1, \dots, T \tag{2}$$

with $\mathbb{E}[\varepsilon_{t,T}|Y_{t-1,T}, \dots, Y_{t-d,T}] = 0$, which is analyzed in Section 3. As will be seen there, the process defined in (2) is locally stationary and strongly mixing under suitable conditions on the function m and the error terms $\varepsilon_{t,T}$. Note that independently of the present work, Kristensen [16] has developed results on local stationarity of the process given in (2) under a set of assumptions similar to ours.

In Section 4, we develop estimation theory for the nonparametric regression function in the general framework (1). As described there, the regression function is estimated by nonparametric kernel methods. We provide a complete asymptotic theory for our estimates. In particular, we derive uniform convergence rates and an asymptotic normality result. To do so, we split up the estimates into a variance part and a bias part. In order to control the variance part, we generalize results on uniform convergence rates for kernel estimates as provided, for example, in Bosq [3], Masry [18] and Hansen [13]. The locally stationary behavior of the model variables also changes the asymptotic analysis of the bias part. In particular, it produces an additional bias term which can be regarded as measuring the deviation from stationarity.

Even though model (1) is theoretically interesting, it has an important drawback. Estimating the time-varying regression function in (1) suffers from an even more severe curse of dimensionality problem than in the standard, strictly stationary setting with a time-invariant regression function. The reason is that in model (1), we fit a fully nonparametric function $m(u, \cdot)$ locally around *each* rescaled time point u. Compared to the standard case, this means that we additionally smooth in time direction and thus increase the dimensionality of the estimation problem by one. This makes the procedure even more data consuming than in the standard setting and thus infeasible in many applications.

In order to countervail this severe curse of dimensionality, we impose some structural constraints on the regression function in (1). In particular, we consider additive models of the form

$$Y_{t,T} = \sum_{j=1}^{d} m_j\left(\frac{t}{T}, X_{t,T}^j\right) + \varepsilon_{t,T} \qquad \text{for } t = 1, \dots, T \tag{3}$$

with $X_{t,T} = (X_{t,T}^1, \dots, X_{t,T}^d)$ and $\mathbb{E}[\varepsilon_{t,T}|X_{t,T}] = 0$. In Section 5, we will show that the component functions of this model can be estimated with two-dimensional nonparametric convergence rates, no matter how large the dimension d. In order to do so, we extend the smooth backfitting approach of Mammen et al. [17] to our setting.

2. Local stationarity. Heuristically speaking, a process $\{X_{t,T} : t = 1, \dots, T\}_{T=1}^{\infty}$ is locally stationary if it behaves approximately stationary locally in time. This intuitive concept can be turned into a rigorous definition in different ways. One way is to require that locally around each rescaled time point u, the process $\{X_{t,T}\}$ can be approximated by a stationary process $\{X_t(u) : t \in \mathbb{Z}\}$ in a stochastic sense; cf., for example, Dahlhaus and Subba Rao [9]. This idea also underlies the following definition.

DEFINITION 2.1. The process $\{X_{t,T}\}$ is locally stationary if for each rescaled time point $u \in [0, 1]$ there exists an associated process $\{X_t(u)\}$ with the following two properties:

(i) $\{X_t(u)\}$ is strictly stationary with density $f_{X_t(u)}$;
(ii) it holds that

$$\|X_{t,T} - X_t(u)\| \le \left(\left|\frac{t}{T} - u\right| + \frac{1}{T}\right) U_{t,T}(u) \qquad \text{a.s.},$$

where $\{U_{t,T}(u)\}$ is a process of positive variables satisfying $\mathbb{E}[(U_{t,T}(u))^{\rho}] < C$ for some $\rho > 0$ and $C < \infty$ independent of u, t, and T. $\| \cdot \|$ denotes an arbitrary norm on $\mathbb{R}^d$.

Since the ρth moments of the variables $U_{t,T}(u)$ are uniformly bounded, it holds that $U_{t,T}(u) = O_p(1)$. As a consequence of the above definition, we thus have

$$\|X_{t,T} - X_t(u)\| = O_p\left(\left|\frac{t}{T} - u\right| + \frac{1}{T}\right).$$

The constant ρ can be regarded as a measure of how well $X_{t,T}$ is approximated by $X_t(u)$: the larger ρ can be chosen, the less mass is contained in the tails of the distribution of $U_{t,T}(u)$. Thus, if ρ is large, then the bound $(|\frac{t}{T} - u| + \frac{1}{T})U_{t,T}(u)$ will take rather moderate values for most of the time. In this sense, the bound and thus the approximation of $X_{t,T}$ by $X_t(u)$ is getting better for larger ρ.

3. Locally stationary nonlinear AR models. In this section, we examine a large class of nonlinear autoregressive processes with a time-varying regression function that fit into the general framework (1). We show that these processes are locally stationary and strongly mixing under suitable conditions on the model components. To shorten notation, we repeatedly make use of the following abbreviation: for any array of variables $\{Z_{t,T}\}$, we let $Z_{t,T}^{t-k} := (Z_{t-k,T}, \dots, Z_{t,T})$ for $k > 0$.

3.1. *The time-varying nonlinear AR* (*tvNAR*) *process*. We call an array $\{Y_{t,T} : t \in \mathbb{Z}\}_{T=1}^{\infty}$ a time-varying nonlinear autoregressive (tvNAR) process if $Y_{t,T}$ evolves according to the equation

$$Y_{t,T} = m\left(\frac{t}{T}, Y_{t-1,T}^{t-d}\right) + \sigma\left(\frac{t}{T}, Y_{t-1,T}^{t-d}\right)\varepsilon_t. \tag{4}$$

A tvNAR process is thus an autoregressive process of form (2) with errors $\varepsilon_{t,T} = \sigma(\frac{t}{T}, Y_{t-1,T}^{t-d})\varepsilon_t$. In the above definition, $m(u, y)$ and $\sigma(u, y)$ are smooth functions of rescaled time u and $y \in \mathbb{R}^d$. We stipulate that for $u \le 0$, $m(u, y) = m(0, y)$ and $\sigma(u, y) = \sigma(0, y)$. Analogously, we set $m(u, y) = m(1, y)$ and $\sigma(u, y) = \sigma(1, y)$ for $u \ge 1$. Furthermore, the variables ε_t are assumed to be i.i.d. with mean zero. For each $u \in \mathbb{R}$, we additionally define the associated process $\{Y_t(u) : t \in \mathbb{Z}\}$ by

$$Y_t(u) = m\big(u, Y_{t-1}^{t-d}(u)\big) + \sigma\big(u, Y_{t-1}^{t-d}(u)\big)\varepsilon_t, \tag{5}$$

where the rescaled time argument of the functions m and σ is fixed at u.

As stipulated above, the functions m and σ in (4) do not change over time for $t \le 0$. Put differently, $Y_{t,T} = m(0, Y_{t-1,T}^{t-d}) + \sigma(0, Y_{t-1,T}^{t-d})\varepsilon_t$ for all $t \le 0$. We can thus assume that $Y_{t,T} = Y_t(0)$ for $t \le 0$. Consequently, if there exists a process $\{Y_t(0)\}$ that satisfies the system of equations (5) for $u = 0$, then this immediately implies the existence of a tvNAR process $\{Y_{t,T}\}$ satisfying (4). As will turn out, under appropriate conditions there exists a strictly stationary solution $\{Y_t(u)\}$ to (5) for each $u \in \mathbb{R}$, in particular for $u = 0$. We can thus take for granted that the tvNAR process $\{Y_{t,T}\}$ defined by (4) exists.

Before we turn to the analysis of the tvNAR process, we compare it to the framework of Zhou and Wu [24] and Zhou [23]. Their model is given by the equation $Z_{t,T} = G(\frac{t}{T}, \psi_t)$, where $\psi_t = (\dots, \varepsilon_{t-1}, \varepsilon_t)$ with i.i.d. variables ε_t and G is a measurable function. In their theory, the variables $Z_t(u) = G(u, \psi_t)$ play the role of a stationary approximation at $u \in [0, 1]$. Under suitable assumptions, we can iterate equation (5) to obtain that $Y_t(u) = F(u, \psi_t)$ for some measurable function F. Note, however, that $Y_{t,T} \neq F(\frac{t}{T}, \psi_t)$ in general. This is due to the fact that when iterating (5), we use the same functions $m(u, \cdot)$ and $\sigma(u, \cdot)$ in each step. In contrast to this, different functions show up in each step when iterating the tvNAR variables $Y_{t,T}$. Thus, the relation between the tvNAR process $\{Y_{t,T}\}$ and the approximations $\{Y_t(u)\}$ is in general different from that between the processes $\{Z_{t,T}\}$ and $\{Z_t(u)\}$ in the setting of Zhou and Wu.

3.2. *Assumptions*. We now list some conditions which are sufficient to ensure that the tvNAR process is locally stationary and strongly mixing. To start with, the function m is supposed to satisfy the following conditions:

(M1) m is absolutely bounded by some constant $C_m < \infty$.

(M2) m is Lipschitz continuous with respect to rescaled time u, that is, there exists a constant $L < \infty$ such that $|m(u, y) - m(u', y)| \leq L|u - u'|$ for all $y \in \mathbb{R}^d$.

(M3) m is continuously differentiable with respect to y. The partial derivatives $\partial_j m(u, y) := \frac{\partial}{\partial y_j} m(u, y)$ have the property that for some $K_1 < \infty$,

$$\sup_{u \in \mathbb{R}, \|y\|_\infty > K_1} \left|\partial_j m(u, y)\right| \leq \delta < 1.$$

An exact formula for the bound δ is given in (31) in Appendix A.

The function σ is required to fulfill analogous assumptions.

(Σ1) σ is bounded by some constant $C_\sigma < \infty$ from above and by some constant $c_\sigma > 0$ from below, that is, $0 < c_\sigma \leq \sigma(u, y) \leq C_\sigma < \infty$ for all u and y.

(Σ2) σ is Lipschitz continuous with respect to rescaled time u.

(Σ3) σ is continuously differentiable with respect to y. The partial derivatives $\partial_j \sigma(u, y) := \frac{\partial}{\partial y_j} \sigma(u, y)$ have the property that for some $K_1 < \infty$, $|\partial_j \sigma(u, y)| \leq \delta < 1$ for all $u \in \mathbb{R}$ and $\|y\|_\infty > K_1$.

Finally, the error terms are required to have the following properties.

(E1) The variables ε_t are i.i.d. with $\mathbb{E}[\varepsilon_t] = 0$ and $\mathbb{E}|\varepsilon_t|^{1+\eta} < \infty$ for some $\eta > 0$. Moreover, they have an everywhere positive and continuous density f_ε.

(E2) The density f_ε is bounded and Lipschitz, that is, there exists a constant $L < \infty$ such that $|f_\varepsilon(z) - f_\varepsilon(z')| \leq L|z - z'|$ for all $z, z' \in \mathbb{R}$.

To show that the tvNAR process is strongly mixing, we additionally need the following condition on the density of the error terms:

(E3) Let d_0, d_1 be any constants with $0 \leq d_0 \leq D_0 < \infty$ and $|d_1| \leq D_1 < \infty$. The density f_ε fulfills the condition

$$\int_{\mathbb{R}} \left|f_\varepsilon([1 + d_0]z + d_1) - f_\varepsilon(z)\right| dz \leq C_{D_0,D_1}(d_0 + |d_1|)$$

with $C_{D_0,D_1} < \infty$ only depending on the bounds D_0 and D_1.

We shortly give some remarks on the above conditions:

(i) Our set of assumptions can be regarded as a strengthening of the assumptions needed to show geometric ergodicity of nonlinear AR processes of the form $Y_t = m(Y_{t-1}^{t-d}) + \sigma(Y_{t-1}^{t-d})\varepsilon_t$. The main assumption in this context requires the functions m and σ not to grow too fast outside a large bounded set. More precisely, it requires them to be dominated by linear functions with sufficiently small slopes; cf.

Tjøstheim [21], Bhattacharya and Lee [2], An and Huang [1] or Chen and Chen [4], among others. (M3) and (Σ3) are very close in spirit to this kind of assumption. They restrict the growth of m and σ by requiring the derivatives of these functions to be small outside a large bounded set.

(ii) If we replace (M3) and (Σ3) with the stronger assumption that the partial derivatives $|\partial_j m(u, y)|$ and $|\partial_j \sigma(u, y)|$ are globally bounded by some sufficiently small number $\delta < 1$, then some straightforward modifications allow us to dispense with the boundedness assumptions (M1) and (Σ1) in the local stationarity and mixing proofs.

(iii) Condition (M3) implies that the derivatives $\partial_j m(u, y)$ are absolutely bounded. Hence, there exists a constant $\Delta < \infty$ such that $|\partial_j m(u, y)| \leq \Delta$ for all $u \in \mathbb{R}$ and $y \in \mathbb{R}^d$. Similarly, (Σ3) implies that the derivatives $\partial_j \sigma(u, y)$ are absolutely bounded by some constant $\Delta < \infty$.

(iv) As already noted, (E3) is only needed to prove that the tvNAR process is strongly mixing. It is, for example, fulfilled for the class of bounded densities f_ε whose first derivative f_ε' is bounded, satisfies $\int |z f_\varepsilon'(z)|\, dz < \infty$ and declines monotonically to zero for values $|z| > C$ for some constant $C > 0$; see also Section 3 in Fryzlewicz and Subba Rao [11] who work with assumptions closely related to (E3).

3.3. *Properties of the tvNAR process.* We now show that the tvNAR process is locally stationary and strongly mixing under the assumptions listed above. In addition, we will see that the auxiliary processes $\{Y_t(u)\}$ have densities that vary smoothly over rescaled time u. As will turn out, these three properties are central for the estimation theory developed in Sections 4 and 5.

The first theorem summarizes some properties of the tvNAR process and of the auxiliary processes $\{Y_t(u)\}$ that are needed to prove the main results.

THEOREM 3.1. *Let (M1)–(M3), (Σ1)–(Σ3) and (E1) be fulfilled. Then*:

(i) *for each $u \in \mathbb{R}$, the process $\{Y_t(u), t \in \mathbb{Z}\}$ has a strictly stationary solution with ε_t independent of $Y_{t-k}(u)$ for $k > 0$;*

(ii) *the variables $Y_{t-1}^{t-d}(u)$ have a density $f_{Y_{t-1}^{t-d}(u)}$ w.r.t. Lebesgue measure*;

(iii) *the variables $Y_{t-1,T}^{t-d}$ have densities $f_{Y_{t-1,T}^{t-d}}$ w.r.t. Lebesgue measure.*

The next result states that $\{Y_{t,T}\}$ can be locally approximated by $\{Y_t(u)\}$. Together with Theorem 3.1, it shows that the tvNAR process $\{Y_{t,T}\}$ is locally stationary in the sense of Definition 2.1.

THEOREM 3.2. *Let (M1)–(M3), (Σ1)–(Σ3) and (E1) be fulfilled. Then*

$$|Y_{t,T} - Y_t(u)| \leq \left(\left|\frac{t}{T} - u\right| + \frac{1}{T}\right) U_{t,T}(u) \qquad a.s., \tag{6}$$

where the variables $U_{t,T}(u)$ have the property that $\mathbb{E}[(U_{t,T}(u))^\rho] < C$ for some $\rho > 0$ and $C < \infty$ independent of u, t and T.

To get an idea of the proof of Theorem 3.2, consider the model $Y_{t,T} = m(\frac{t}{T}, Y_{t-1,T}) + \varepsilon_t$ for a moment. Our arguments are based on a backward expansion of the difference $Y_{t,T} - Y_t(u)$. Exploiting the smoothness conditions of (M2) and (M3) together with the boundedness of m, we obtain that

$$\left|Y_{t,T} - Y_t(u)\right| \le C \sum_{r=0}^{n-1} \prod_{k=1}^{r} |\partial m(u, \xi_{t-k})| \left(\left| \frac{t}{T} - u \right| + \frac{r}{T} \right) + C \prod_{k=1}^{n} |\partial m(u, \xi_{t-k})|,$$

where $\partial m(u, y)$ is the derivative of $m(u, y)$ with respect to y and ξ_{t-k} is an intermediate point between $Y_{t-k,T}$ and $Y_{t-k}(u)$. To prove (6), we have to show that the product $\prod_{k=1}^n |\partial m(u, \xi_{t-k})|$ is contracting in some stochastic sense as n tends to infinity. The heuristic idea behind the proof is the following: using conditions (M1) and (E1), we can show that at least a certain fraction of the terms $\xi_{t-1}, \dots, \xi_{t-n}$ take a value in the region $\{y : |y| > K_1\}$ as n grows large. Since the derivative $|\partial m|$ is small in this region according to (M3), this ensures that at least a certain fraction of the elements in the product $\prod_{k=1}^n |\partial m(u, \xi_{t-k})|$ are small in value. This prevents the product from exploding and makes it contract to zero as n goes to infinity.

Next, we come to a result which shows that the densities of the approximating variables $Y_{t-1}^{t-d}(u)$ change smoothly over time.

THEOREM 3.3. *Let $f(u, y) := f_{Y_{t-1}^{t-d}(u)}(y)$ be the density of $Y_{t-1}^{t-d}(u)$ at $y \in \mathbb{R}^d$. If (M1)–(M3), (Σ1)–(Σ3) and (E1), (E2) are fulfilled, then*

$$\left|f(u, y) - f(v, y)\right| \le C_y |u - v|^p$$

with some constant $0 < p < 1$ and $C_y < \infty$ continuously depending on y.

We finally characterize the mixing behavior of the tvNAR process. To do so, we first give a quick reminder of the definitions of an α- and β-mixing array. Let $(\Omega, \mathcal{A}, \mathbb{P})$ be a probability space, and let $\mathcal{B}$ and $\mathcal{C}$ be subfields of $\mathcal{A}$. Define

$$\alpha(\mathcal{B}, \mathcal{C}) = \sup_{B \in \mathcal{B}, C \in \mathcal{C}} \left|\mathbb{P}(B \cap C) - \mathbb{P}(B)\mathbb{P}(C)\right|,$$

$$\beta(\mathcal{B}, \mathcal{C}) = \mathbb{E} \sup_{C \in \mathcal{C}} \left|\mathbb{P}(C) - \mathbb{P}(C|\mathcal{B})\right|.$$

Moreover, for an array $\{Z_{t,T} : 1 \le t \le T\}$, define the coefficients

$$\alpha(k) = \sup_{t,T : 1 \le t \le T-k} \alpha\big(\sigma(Z_{s,T}, 1 \le s \le t), \sigma(Z_{s,T}, t+k \le s \le T)\big), \tag{7}$$

$$\beta(k) = \sup_{t,T : 1 \le t \le T-k} \beta\big(\sigma(Z_{s,T}, 1 \le s \le t), \sigma(Z_{s,T}, t+k \le s \le T)\big), \tag{8}$$

where $\sigma(Z)$ is the σ-field generated by Z. The array $\{Z_{t,T}\}$ is said to be α-mixing (or strongly mixing) if $\alpha(k) \to 0$ as $k \to \infty$. Similarly, it is called β-mixing if $\beta(k) \to 0$. Note that β-mixing implies α-mixing. The final result of this section shows that the tvNAR process is β-mixing with coefficients that converge exponentially fast to zero.

THEOREM 3.4. *If* (M1)–(M3), (Σ1)–(Σ3) *and* (E1)–(E3) *are fulfilled, then the tvNAR process* $\{Y_{t,T}\}$ *is geometrically* β*-mixing, that is, there exist positive constants* $\gamma < 1$ *and* $C < \infty$ *such that* $\beta(k) \le C\gamma^k$.

The strategy of the proof is as follows: the (conditional) probabilities that show up in the definition of the β-coefficient in (8) can be written in terms of the functions m, σ and the error density f_ε. To do so, we derive recursive expressions of the model variables $Y_{t,T}$ and of certain conditional densities of $Y_{t,T}$. Rewriting the β-coefficient with the help of these expressions allows us to derive an appropriate bound for it. The overall strategy is thus similar to that of Fryzlewicz and Subba Rao [11] who also derive bounds of mixing coefficients in terms of conditional densities. The specific steps of the proof, however, are quite different. The details together with the proofs of the other theorems can be found in Appendix A.

4. Kernel estimation. In this section, we consider kernel estimation in the general model (1),

$$Y_{t,T} = m\left(\frac{t}{T}, X_{t,T}\right) + \varepsilon_{t,T} \qquad \text{for } t = 1, \dots, T$$

with $\mathbb{E}[\varepsilon_{t,T}|X_{t,T}] = 0$. Note that $m(\frac{t}{T}, \cdot)$ is the conditional mean function in model (1) at the time point t. The function m is thus identified almost surely on the grid of points $\frac{t}{T}$ for $t = 1, \dots, T$. These points form a dense subset of the unit interval as the sample size grows to infinity. As a consequence, m is identified almost surely at all rescaled time points $u \in [0, 1]$ if it is continuous in time direction (which we will assume in what follows).

4.1. *Estimation procedure.* We restrict attention to Nadaraya–Watson (NW) estimation. It is straightforward to extend the theory to local linear (or more generally local polynomial) estimation. The NW estimator of model (1) is given by

$$\hat{m}(u, x) = \frac{\sum_{t=1}^T K_h(u - t/T) \prod_{j=1}^d K_h(x^j - X_{t,T}^j) Y_{t,T}}{\sum_{t=1}^T K_h(u - t/T) \prod_{j=1}^d K_h(x^j - X_{t,T}^j)}. \tag{9}$$

Here and in what follows, we write $X_{t,T} = (X_{t,T}^1, \dots, X_{t,T}^d)$ and $x = (x^1, \dots, x^d)$ for any vector $x \in \mathbb{R}^d$, that is, we use subscripts to indicate the time point of observation and superscripts to denote the components of the vector. K denotes a one-dimensional kernel function and we use the notation $K_h(v) = K(\frac{v}{h})$. For convenience, we work with a product kernel and assume that the bandwidth h is the same in each direction. Our results can, however, be easily modified to allow for nonproduct kernels and different bandwidths.

The estimate defined in (9) differs from the NW estimator in the standard strictly stationary setting in that there is an additional kernel in time direction. We thus do not only smooth in the direction of the covariates $X_{t,T}$ but also in the time direction. This takes into account that the regression function is varying over time. In what follows, we derive the asymptotic properties of our NW estimate. The proofs are given in Appendix B.

4.2. *Assumptions*. The following three conditions are central to our results:

(C1) The process $\{X_{t,T}\}$ is locally stationary in the sense of Definition 2.1. Thus, for each time point $u \in [0, 1]$, there exists a strictly stationary process $\{X_t(u)\}$ having the property that $\|X_{t,T} - X_t(u)\| \le (|\frac{t}{T} - u| + \frac{1}{T})U_{t,T}(u)$ a.s. with $\mathbb{E}[(U_{t,T}(u))^\rho] \le C$ for some $\rho > 0$.

(C2) The densities $f(u, x) := f_{X_t(u)}(x)$ of the variables $X_t(u)$ are smooth in u. In particular, $f(u, x)$ is differentiable w.r.t. u for each $x \in \mathbb{R}^d$, and the derivative $\partial_0 f(u, x) := \frac{\partial}{\partial u} f(u, x)$ is continuous.

(C3) The array $\{X_{t,T}, \varepsilon_{t,T}\}$ is α-mixing.

As seen in Section 3, these three conditions are essentially fulfilled for the tvNAR process: (C1) and (C3) follow immediately from Theorems 3.2 and 3.4. Moreover, Theorem 3.3 shows that the tvNAR process satisfies a weakened version of (C2) which requires the densities $f_{X_t(u)}$ to be continuous rather than differentiable in time direction. Note that we could do with this weakened version of (C2), however at the cost of getting slower convergence rates for the bias part of the NW estimate.

In addition to the above three assumptions, we impose the following regularity conditions:

(C4) $f(u, x)$ is partially differentiable w.r.t. x for each $u \in [0, 1]$. The derivatives $\partial_j f(u, x) := \frac{\partial}{\partial x^j} f(u, x)$ are continuous for $j = 1, \dots, d$.

(C5) $m(u, x)$ is twice continuously partially differentiable with first derivatives $\partial_j m(u, x)$ and second derivatives $\partial_{ij}^2 m(u, x)$ for $i, j = 0, \dots, d$.

(C6) The kernel K is symmetric about zero, bounded and has compact support, that is, $K(v) = 0$ for all $|v| > C_1$ with some $C_1 < \infty$. Furthermore, K is Lipschitz, that is, $|K(v) - K(v')| \le L|v - v'|$ for some $L < \infty$ and all $v, v' \in \mathbb{R}$.

Finally, note that throughout the paper the bandwidth h is assumed to converge to zero at least at polynomial rate, that is, there exists a small $\xi > 0$ such that $h \le CT^{-\xi}$ for some constant $C > 0$.

4.3. *Uniform convergence rates for kernel averages.* As a first step in the analysis of the NW estimate (9), we examine kernel averages of the general form

$$\hat{\psi}(u,x) = \frac{1}{Th^{d+1}} \sum_{t=1}^{T} K_h\left(u - \frac{t}{T}\right) \prod_{j=1}^{d} K_h(x^j - X_{t,T}^j) W_{t,T} \tag{10}$$

with $\{W_{t,T}\}$ being an array of one-dimensional random variables. A wide range of kernel-based estimators, including the NW estimator defined in (9), can be written as functions of averages of the above form. The asymptotic behavior of such averages is thus of wider interest. For this reason, we investigate the properties of these averages for a general array of variables $\{W_{t,T}\}$. Later on we will employ the results with $W_{t,T} = 1$ and $W_{t,T} = \varepsilon_{t,T}$.

We now derive the uniform convergence rate of $\hat{\psi}(u,x) - \mathbb{E}\hat{\psi}(u,x)$. To do so, we make the following assumptions on the components in (10):

(K1) It holds that $\mathbb{E}|W_{t,T}|^s \leq C$ for some $s > 2$ and $C < \infty$.

(K2) The array $\{X_{t,T}, W_{t,T}\}$ is α-mixing. The mixing coefficients α have the property that $\alpha(k) \leq Ak^{-\beta}$ for some $A < \infty$ and $\beta > \frac{2s-2}{s-2}$.

(K3) Let $f_{X_{t,T}}$ and $f_{X_{t,T},X_{t+l,T}}$ be the densities of $X_{t,T}$ and $(X_{t,T}, X_{t+l,T})$, respectively. For any compact set $S \subseteq \mathbb{R}^d$, there exists a constant $C = C(S)$ such that $\sup_{t,T} \sup_{x \in S} f_{X_{t,T}}(x) \leq C$ and $\sup_{t,T} \sup_{x \in S} \mathbb{E}[|W_{t,T}|^s | X_{t,T} = x] f_{X_{t,T}}(x) \leq C$. Moreover, there exists a natural number $l^* < \infty$ such that for all $l \geq l^*$, $\sup_{t,T} \sup_{x,x' \in S} \mathbb{E}[|W_{t,T}||W_{t+l,T}||X_{t,T} = x, X_{t+l,T} = x'] f_{X_{t,T},X_{t+l,T}}(x,x') \leq C$.

The next theorem generalizes uniform convergence results of Hansen [13] for the strictly stationary case to our setting. See Kristensen [15] for related results.

THEOREM 4.1. *Assume that (K1)–(K3) are satisfied with*

$$\beta > \frac{2 + s(1 + (d+1))}{s-2} \tag{11}$$

and that the kernel K fulfills (C6). In addition, let the bandwidth satisfy

$$\frac{\phi_T \log T}{T^{\theta} h^{d+1}} = o(1) \tag{12}$$

with ϕ_T slowly diverging to infinity (e.g., $\phi_T = \log\log T$) and

$$\theta = \frac{\beta(1 - 2/s) - 2/s - 1 - (d+1)}{\beta + 3 - (d+1)}. \tag{13}$$

Finally, let S be a compact subset of $\mathbb{R}^d$. Then it holds that

$$\sup_{u \in [0,1], x \in S} \left|\hat{\psi}(u,x) - \mathbb{E}\hat{\psi}(u,x)\right| = O_p\left(\sqrt{\frac{\log T}{Th^{d+1}}}\right). \tag{14}$$

The convergence rate in the above theorem is identical to the rate obtained for a $(d+1)$-dimensional nonparametric estimation problem in the standard strictly stationary setting. This reflects the fact that additionally smoothing in time direction, we essentially have a $(d+1)$-dimensional problem in our case. Moreover, note that with (11) and (13), we can compute that $\theta \in (0, 1-\frac{2}{s}]$. In particular, $\theta = 1-\frac{2}{s}$ if the mixing coefficients decay exponentially fast to zero, that is, if $\beta = \infty$. Restriction (12) on the bandwidth is thus a strengthening of the usual condition that $Th^{d+1} \to \infty$.

4.4. *Uniform convergence rates for NW estimates.* The next theorem characterizes the uniform convergence behavior of our NW estimate.

THEOREM 4.2. *Assume that* (C1)–(C6) *hold and that* (K1)–(K3) *are fulfilled both for* $W_{t,T} = 1$ *and* $W_{t,T} = \varepsilon_{t,T}$. *Let* β *satisfy* (11) *and suppose that* $\inf_{u\in[0,1],x\in S} f(u,x) > 0$. *Moreover, assume that the bandwidth* h *satisfies*

$$\frac{\phi_T \log T}{T^{\theta}h^{d+1}} = o(1) \quad \text{and} \quad \frac{1}{T^r h^{d+r}} = o(1) \tag{15}$$

with θ *given in* (13), $\phi_T = \log\log T$, $r = \min\{\rho, 1\}$ *and* ρ *introduced in* (C1). *Defining* $I_h = [C_1 h, 1 - C_1 h]$, *it then holds that*

$$\sup_{u\in I_h, x\in S} \left|\hat{m}(u,x) - m(u,x)\right| = O_p\left(\sqrt{\frac{\log T}{Th^{d+1}}} + \frac{1}{T^r h^d} + h^2\right). \tag{16}$$

To derive the above result, we decompose the difference $\hat{m}(u,x) - m(u,x)$ into a stochastic part and a bias part. Using Theorem 4.1, the stochastic part can be shown to be of the order $O_p(\sqrt{\log T/Th^{d+1}})$. The bias term splits up into two parts, a standard component of the order $O(h^2)$ and a nonstandard component of the order $O(T^{-r}h^{-d})$. The latter component results from replacing the variables $X_{t,T}$ by $X_t(\frac{t}{T})$ in the bias term. It thus captures how far these variables are from their stationary approximations $X_t(\frac{t}{T})$. Put differently, it measures the deviation from stationarity. As will be seen in Appendix B, handling this nonstationarity bias requires techniques substantially different from those needed to treat the bias term in a strictly stationary setting.

Note that the additional nonstationarity bias converges faster to zero for larger $r = \min\{\rho, 1\}$. This makes perfect sense if we recall from Section 2 that r measures how well $X_{t,T}$ is locally approximated by $X_t(\frac{t}{T})$: the larger r, the smaller the deviation of $X_{t,T}$ from its stationary approximation and thus the smaller the additional nonstationarity bias.

4.5. *Asymptotic normality.* We conclude the asymptotic analysis of our NW estimate with a result on asymptotic normality.

THEOREM 4.3. *Assume that* (C1)–(C6) *hold and that* (K1)–(K3) *are fulfilled both for* $W_{t,T} = 1$ *and* $W_{t,T} = \varepsilon_{t,T}$. *Let* $\beta \geq 4$ *and* $T^r h^{d+2} \to \infty$

with $r = \min\{\rho, 1\}$. *Moreover, suppose that* $f(u,x) > 0$ *and that* $\sigma^2(\frac{t}{T}, x) := \mathbb{E}[\varepsilon_{t,T}^2 | X_{t,T} = x]$ *is continuous. Finally, let* $r > \frac{d+2}{d+5}$ *to ensure that the bandwidth* h *can be chosen to satisfy* $Th^{d+5} \to c_h$ *for a constant* c_h. *Then*

$$\sqrt{Th^{d+1}}\big(\hat{m}(u,x) - m(u,x)\big) \xrightarrow{d} N(B_{u,x}, V_{u,x}), \tag{17}$$

where $B_{u,x} = \sqrt{c_h}\frac{\kappa_2}{2}\sum_{i=0}^d [2\,\partial_i m(u,x)\,\partial_i f(u,x) + \partial_{i,i}^2 m(u,x) f(u,x)]/f(u,x)$ *and* $V_{u,x} = \kappa_0^{d+1}\sigma^2(u,x)/f(u,x)$ *with* $\kappa_0 = \int K^2(\varphi)\,d\varphi$ *and* $\kappa_2 = \int \varphi^2 K(\varphi)\,d\varphi$.

The above theorem parallels the asymptotic normality result for the standard strictly stationary setting. In particular, the bias and variance expressions $B_{u,x}$ and $V_{u,x}$ are very similar to those from the standard case. By requiring that $T^r h^{d+2} \to \infty$, we make sure that the additional nonstationarity bias is asymptotically negligible.

5. Locally stationary additive models. We now put some structural constraints on the regression function m in model (1). In particular, we assume that for all rescaled time points $u \in [0,1]$ and all points x in a compact subset of $\mathbb{R}^d$, say $[0,1]^d$, the regression function can be split up into additive components according to $m(u,x) = m_0(u) + \sum_{j=1}^d m_j(u, x^j)$. This means that for $x \in [0,1]^d$, we have the additive regression model

$$\mathbb{E}[Y_{t,T} | X_{t,T} = x] = m_0\left(\frac{t}{T}\right) + \sum_{j=1}^d m_j\left(\frac{t}{T}, x^j\right). \tag{18}$$

To identify the component functions of model (18) within the unit cube $[0,1]^d$, we impose the condition that $\int m_j(u,x^j) p_j(u,x^j)\,dx^j = 0$ for all $j = 1, \dots, d$ and all rescaled time points $u \in [0,1]$. Here, the functions $p_j(u,x^j) = \int p(u,x)\,dx^{-j}$ are the marginals of the density

$$p(u,x) = \frac{I(x \in [0,1]^d) f(u,x)}{\mathbb{P}(X_0(u) \in [0,1]^d)},$$

where as before $f(u,\cdot)$ is the density of the strictly stationary process $\{X_t(u)\}$. Note that this normalization of the component functions varies over time in the sense that for each rescaled time point u, we integrate with respect to a different density.

To estimate the functions $m_0, \dots, m_d$, we adapt the smooth backfitting technique of Mammen et al. [17] to our setting. To do so, we first introduce the auxiliary estimates

$$\hat{p}(u,x) = \frac{1}{T_{[0,1]^d}} \sum_{t=1}^T I\big(X_{t,T} \in [0,1]^d\big) K_h\left(u, \frac{t}{T}\right) \prod_{j=1}^d K_h\big(x^j, X_{t,T}^j\big),$$

$$\hat{m}(u,x) = \frac{1}{T_{[0,1]^d}} \sum_{t=1}^T I\big(X_{t,T} \in [0,1]^d\big) K_h\left(u, \frac{t}{T}\right) \prod_{j=1}^d K_h\big(x^j, X_{t,T}^j\big) Y_{t,T} / \hat{p}(u,x).$$

$\hat{p}(u,x)$ is a kernel estimate of the density $p(u,x)$, and $\hat{m}(u,x)$ is a $(d+1)$-dimensional NW smoother that estimates $m(u,x)$ for $x \in [0,1]^d$. In the above definitions,

$$T_{[0,1]^d} = \sum_{t=1}^{T} K_h\left(u, \frac{t}{T}\right) I\left(X_{t,T} \in [0,1]^d\right)$$

is the number of observations in the unit cube $[0,1]^d$, where only time points close to u are taken into account, and

$$K_h(v,w) = I\big(v,w \in [0,1]\big) \frac{K_h(v-w)}{\int_0^1 K_h(s-w)\,ds}$$

is a modified kernel weight. This weight has the property that $\int_0^1 K_h(v,w)\,dv = 1$ for all $w \in [0,1]$, which is needed to derive the asymptotic properties of the back-fitting estimates.

Given the smoothers $\hat{p}$ and $\hat{m}$, we define the smooth backfitting estimates $\tilde{m}_0(u), \tilde{m}_1(u,\cdot), \dots, \tilde{m}_d(u,\cdot)$ of the functions $m_0(u), m_1(u,\cdot), \dots, m_d(u,\cdot)$ at the time point $u \in [0,1]$ as the minimizers of the criterion

$$\int \left(\hat{m}(u,w) - g_0 - \sum_{j=1}^{d} g_j(w^j)\right)^2 \hat{p}(u,w)\,dw, \tag{19}$$

where the minimization runs over all additive functions $g(x) = g_0 + g_1(x^1) + \cdots + g_d(x^d)$ whose components are normalized to satisfy $\int g_j(w^j)\hat{p}_j(u,w^j)\,dw^j = 0$ for $j = 1,\dots,d$. Here, $\hat{p}_j(u,x^j) = \int \hat{p}(u,x)\,dx^{-j}$ is the marginal of the kernel density $\hat{p}(u,\cdot)$ at the point x^j.

According to (19), the backfitting estimate $\tilde{m}(u,\cdot) = \tilde{m}_0(u) + \sum_{j=1}^{d} \tilde{m}_j(u,\cdot)$ is an L_2-projection of the full-dimensional NW estimate $\hat{m}(u,\cdot)$ onto the subspace of additive functions, where the projection is done with respect to the density estimate $\hat{p}(u,\cdot)$. Note that (19) is a d-dimensional projection problem. In particular, rescaled time does not enter as an additional dimension. The projection is rather done separately for each time point $u \in [0,1]$. We thus fit a smooth backfitting estimate to the data separately around each point in time u.

By differentiation, we can show that the minimizer of (19) is characterized by the system of integral equations

$$\tilde{m}_j(u,x^j) = \hat{m}_j(u,x^j) - \sum_{k \neq j} \int \tilde{m}_k(u,x^k) \frac{\hat{p}_{j,k}(u,x^j,x^k)}{\hat{p}_j(u,x^j)}\,dx^k - \tilde{m}_0(u) \tag{20}$$

together with $\int \tilde{m}_j(u,w^j)\hat{p}_j(u,w^j)\,dw^j = 0$ for $j = 1,\dots,d$. Here, $\hat{p}_j$ and $\hat{p}_{j,k}$ are kernel density estimates, and $\hat{m}_j$ is a NW smoother defined as

$$\hat{p}_j(u,x^j) = \frac{1}{T_{[0,1]^d}} \sum_{t=1}^{T} I\left(X_{t,T} \in [0,1]^d\right) K_h\left(u, \frac{t}{T}\right) K_h\left(x^j, X_{t,T}^j\right),$$

$$\hat{p}_{j,k}(u, x^j, x^k) = \frac{1}{T_{[0,1]^d}} \sum_{t=1}^{T} I(X_{t,T} \in [0,1]^d) K_h\left(u, \frac{t}{T}\right) \times K_h(x^j, X^j_{t,T}) K_h(x^k, X^k_{t,T}),$$

$$\hat{m}_j(u, x^j) = \frac{1}{T_{[0,1]^d}} \sum_{t=1}^{T} I(X_{t,T} \in [0,1]^d) K_h\left(u, \frac{t}{T}\right) \times K_h(x^j, X^j_{t,T}) Y_{t,T} / \hat{p}_j(u, x^j).$$

Moreover, the estimate $\tilde{m}_0(u)$ of the model constant at time point u is given by $\tilde{m}_0(u) = T^{-1}_{[0,1]^d} \sum_{t=1}^{T} I(X_{t,T} \in [0,1]^d) K_h(u, \frac{t}{T}) Y_{t,T}$.

We next summarize the assumptions needed to derive the asymptotic properties of the smooth backfitting estimates. First of all, the conditions of Section 4 must be satisfied for the kernel estimates that show up in the system of integral equations (20). This is ensured by the following assumption.

(Add1) Conditions (C1)–(C6) are fulfilled together with (K1)–(K3) for $W_{t,T} = 1$ and $W_{t,T} = \varepsilon_{t,T}$. The parameter β satisfies the inequality $\beta > \max\{4, \frac{2+3s}{s-2}\}$ and $\inf_{u \in [0,1], x \in [0,1]^d} f(u, x) > 0$.

In addition to (Add1), we need some restrictions on the admissible bandwidth. For convenience, we stipulate somewhat stronger conditions than in Section 4 to get rid of the additional nonstationarity bias from the very beginning.

(Add2) The bandwidth h is such that (i) $Th^5 \to \infty$, (ii) $\frac{\phi_T \log T}{T^\theta h^2} = o(1)$ with $\phi_T = \log\log T$ and $\theta = \min\{\frac{\beta-4}{\beta}, \frac{\beta(1-2/s)-2/s-3}{\beta+1}\}$ and (iii) $(T^r h)^{-1} = o(h^2)$ and $T^{-r/(r+1)} = o(h^2)$ with $r = \min\{\rho, 1\}$ and ρ given in (C1).

Condition (ii) is already known from Section 4. As will be seen in Appendix C, (iii) ensures that the additional nonstationarity bias is of smaller order than $O(h^2)$ and can thus be asymptotically neglected. The expressions for β and θ in (Add1) and (Add2) are calculated as follows: using the formulas (11) and (13) from Theorem 4.1, we get a pair of expressions for β and θ for each of the kernel estimates occurring in (20). Combining these expressions yields the formulas in (Add1) and (Add2).

Under the above assumptions, we can establish the following results, the proofs of which are given in Appendix C. First, the backfitting estimates uniformly converge to the true component functions at the two-dimensional rates no matter how large the dimension d of the full regression function.

THEOREM 5.1. *Let $I_h = [2C_1 h, 1 - 2C_1 h]$. Then under (Add1) and (Add2),*

$$\sup_{u, x^j \in I_h} |\tilde{m}_j(u, x^j) - m_j(u, x^j)| = O_p\left(\sqrt{\frac{\log T}{Th^2}} + h^2\right). \tag{21}$$

Second, the estimates are asymptotically normal if rescaled appropriately.

THEOREM 5.2. *Suppose that* (Add1) *and* (Add2) *hold. In addition, let* $\theta > \frac{1}{3}$ *and* $r > \frac{1}{2}$ *to ensure that the bandwidth* h *can be chosen to satisfy* $T_{[0,1]^d} h^6 \to c_h$ *for a constant* c_h. *Then for any* $u, x^1, \ldots, x^d \in (0,1)$,

$$\sqrt{T_{[0,1]^d} h^2} \begin{bmatrix} \tilde{m}_1(u, x^1) - m_1(u, x^1) \\ \vdots \\ \tilde{m}_d(u, x^d) - m_d(u, x^d) \end{bmatrix} \xrightarrow{d} N(B_{u,x}, V_{u,x}). \tag{22}$$

Here, $V_{u,x}$ *is a diagonal matrix whose diagonal entries are given by the expressions* $v_j(u, x^j) = \kappa_0^2 \sigma_j^2(u, x^j)/p_j(u, x^j)$ *with* $\kappa_0 = \int K^2(\varphi)\,d\varphi$. *Moreover, the bias term has the form* $B_{u,x} = \sqrt{c_h}[\beta_1(u, x^1) - \gamma_1(u), \ldots, \beta_d(u, x^d) - \gamma_d(u)]^T$. *The functions* $\beta_j(u, \cdot)$ *in this expression are defined as the minimizers of the problem*

$$\int [\beta(u, x) - b_0 - b_1(x^1) - \cdots - b_d(x^d)]^2 p(u, x)\,dx,$$

where the minimization runs over all additive functions $b(x) = b_0 + b_1(x^1) + \cdots + b_d(x^d)$ *with* $\int b_j(x^j) p_j(u, x^j)\,dx^j = 0$, *and the function* β *is given in Lemma* C.4 *of Appendix* C. *Moreover, the terms* γ_j *can be characterized by the equation* $\int \alpha_{T,j}(u, x^j)\hat{p}_j(u, x^j)\,dx^j = h^2 \gamma_j(u) + o_p(h^2)$, *where the functions* $\alpha_{T,j}$ *are again defined in Lemma* C.4.

6. Concluding remarks. In this paper, we have studied nonparametric models with a time-varying regression function and locally stationary covariates. We have developed a complete asymptotic theory for kernel estimates in these models. In addition, we have shown that the main assumptions of the theory are satisfied for a large class of nonlinear autoregressive processes with a time-varying regression function.

Our analysis can be extended in several directions. An important issue is bandwidth selection in our framework. As shown in Theorem 4.3, the asymptotic bias and variance expressions of our NW estimate are very similar in structure to those from a standard stationary random design. We thus conjecture that the techniques to choose the bandwidth in such a design can be adapted to our setting. In particular, using the formulas for the asymptotic bias and variance from Theorem 4.3, it should be possible to select the bandwidth via plug-in methods.

Another issue concerns forecasting. The convergence results of Theorems 4.2 and 5.1 are only valid for rescaled time lying in a subset $[Ch, 1 - Ch]$ of the unit interval. For forecasting purposes, it would be important to provide convergence rates also in the boundary region $(1 - Ch, 1]$. This can be achieved by using boundary-corrected kernels. Another possibility is to work with one-sided kernels. In both cases, we have to ensure that the kernels have compact support and are Lipschitz to get the theory to work.

APPENDIX A

In this Appendix, we prove the results on the tvNAR process from Section 3. To shorten notation, we frequently make use of the abbreviations $\underline{Y}_{t,T} = Y_{t,T}^{t-d+1}$, $\underline{Y}_t(u) = Y_t^{t-d+1}(u)$ and $\underline{\varepsilon}_t = \varepsilon_t^{t-d+1}$. Moreover, throughout the Appendices, the symbol C denotes a universal real constant which may take a different value on each occurrence.

Preliminaries. Before we come to the proofs of the theorems, we state some useful facts needed for the arguments later on.

LINEARIZATION OF m AND σ. Consider the function m. The mean value theorem allows us to write

$$m(v, \underline{Y}_{t-1}(v)) - m(u, \underline{Y}_{t-1}(u)) = \Delta_{t,0}^m + \sum_{j=1}^{d} \Delta_{t,j}^m (Y_{t-j}(v) - Y_{t-j}(u)), \tag{23}$$

where we have used the shorthands $\Delta_{t,0}^m = m(v, \underline{Y}_{t-1}(v)) - m(u, \underline{Y}_{t-1}(v))$ and $\Delta_{t,j}^m = \Delta_j^m(u, \underline{Y}_{t-1}(u), \underline{Y}_{t-1}(v))$ for $j = 1, \dots, d$ with the functions $\Delta_j^m(u, y, y') = \int_0^1 \partial_j m(u, y + s(y' - y))\, ds$.

The terms $\Delta_{t,j}^m$ have the property that

$$|\Delta_{t,j}^m| \le \Delta_t := \Delta I(\|\underline{\varepsilon}_{t-1}\|_\infty \le K_2) + \delta I(\|\underline{\varepsilon}_{t-1}\|_\infty > K_2) \tag{24}$$

for $j = 1, \dots, d$ with $K_2 = (K_1 + C_m)/c_\sigma$ and $\Delta \ge \sup_{u,y} |\partial_j m(u, y)|$. This is a straightforward consequence of the boundedness assumptions on m and σ. See the supplement [22] for details.

Repeating the above considerations for the function σ, we obtain analogous terms $\Delta_{t,j}^\sigma$ that are again bounded by Δ_t for $j = 1, \dots, d$.

RECURSIVE FORMULAS FOR $Y_{t,T}$. For the proof of Theorem 3.4, we rewrite $Y_{t,T}$ in a recursive fashion: letting $y_{t-k_1}^{t-k_2}$ and $e_{t-k_1}^{t-k_2}$ be values of $Y_{t-k_1}^{t-k_2}$ and $\varepsilon_{t-k_1}^{t-k_2}$, respectively, we recursively define the functions $m_{t,T}^{(i)}$ by $m_{t,T}^{(0)}(y_{t-1}^{t-d}) = m(\frac{t}{T}, y_{t-1}^{t-d})$ and for $i \ge 1$ by

$$\begin{aligned} &m_{t,T}^{(i)}(e_{t-1}^{t-i}, y_{t-i-1}^{t-i-d}) \\ &\quad = m_{t,T}^{(i-1)}(e_{t-1}^{t-i+1}, m_{t-i,T}^{(0)}(y_{t-i-1}^{t-i-d}) + \sigma_{t-i,T}^{(0)}(y_{t-i-1}^{t-i-d}) e_{t-i}, y_{t-i-1}^{t-i-d+1}). \end{aligned}$$

Using analogous recursions for the function σ, we can additionally define functions $\sigma_{t,T}^{(i)}$ for $i \ge 0$. With this notation at hand, $Y_{t,T}$ can represented as

$$Y_{t,T} = m_{t,T}^{(i)}(\varepsilon_{t-1}^{t-i}, Y_{t-i-1,T}^{t-i-d}) + \sigma_{t,T}^{(i)}(\varepsilon_{t-1}^{t-i}, Y_{t-i-1,T}^{t-i-d}) \varepsilon_t.$$

Moreover, for $i \geq d$ we can write

$$\begin{aligned} & m_{t,T}^{(i)}(e_{t-1}^{t-i}, y_{t-i-1}^{t-i-d}) \\ & \quad = m\bigg(\frac{t}{T}, m_{t-1,T}^{(i-1)}(e_{t-2}^{t-i}, y_{t-i-1}^{t-i-d}) + \sigma_{t-1,T}^{(i-1)}(e_{t-2}^{t-i}, y_{t-i-1}^{t-i-d})e_{t-1}, \dots, \\ & \qquad\qquad m_{t-d,T}^{(i-d)}(e_{t-d-1}^{t-i}, y_{t-i-1}^{t-i-d}) + \sigma_{t-d,T}^{(i-d)}(e_{t-d-1}^{t-i}, y_{t-i-1}^{t-i-d})e_{t-d}\bigg). \end{aligned}$$

The term $\sigma_{t,T}^{(i)}(e_{t-1}^{t-i}, y_{t-i-1}^{t-i-d})$ can be reformulated in the same way.

FORMULAS FOR CONDITIONAL DENSITIES. Throughout the Appendix, the symbol $f_{V|W}$ is used to denote the density of V conditional on W. If the residuals ε_t have a density f_ε, then it can be shown that for $1 \leq r \leq d$,

$$f_{Y_{t,T}|Y_{t-1,T}^{t-r+1}, \varepsilon_{t-r}^{-s}, Y_{-s-1,T}^{-s-d}}(y_t | y_{t-1}^{t-r+1}, e_{t-r}^{-s}, z) = \frac{1}{\sigma_{t,T}} f_\varepsilon\bigg(\frac{y_t - m_{t,T}}{\sigma_{t,T}}\bigg). \tag{25}$$

Here, y_t, y_{t-1}^{t-r+1}, e_{t-r}^{-s} and z are values of $Y_{t,T}$, $Y_{t-1,T}^{t-r+1}$, ε_{t-r}^{-s} and $Y_{-s-1,T}^{-s-d}$, respectively. Moreover,

$$\begin{aligned} m_{t,T} = m\bigg(&\frac{t}{T}, y_{t-1}^{t-r+1}, m_{t-r,T}^{(t-r+s)}(e_{t-r-1}^{-s}, z) + \sigma_{t-r,T}^{(t-r+s)}(e_{t-r-1}^{-s}, z)e_{t-r}, \dots, \\ & \qquad m_{t-d,T}^{(t-d+s)}(e_{t-d-1}^{-s}, z) + \sigma_{t-d,T}^{(t-d+s)}(e_{t-d-1}^{-s}, z)e_{t-d}\bigg), \end{aligned}$$

and $\sigma_{t,T}$ is defined analogously.

Proof of Theorem 3.1. Property (i) follows by standard arguments to be found, for example, in Chen and Chen [4]. Property (ii) immediately follows with the help of (25). Recalling that $Y_{t-1,T}^{t-d} = Y_{t-1}^{t-d}(0)$ for $t \leq 1$, (iii) can again be shown by using (25).

Proof of Theorem 3.2. We apply the triangle inequality to get

$$|Y_{t,T} - Y_t(u)| \leq \bigg|Y_{t,T} - Y_t\bigg(\frac{t}{T}\bigg)\bigg| + \bigg|Y_t\bigg(\frac{t}{T}\bigg) - Y_t(u)\bigg|$$

and bound the terms $|Y_{t,T} - Y_t(\frac{t}{T})|$ and $|Y_t(\frac{t}{T}) - Y_t(u)|$ separately. In what follows, we restrict attention to the term $|Y_t(\frac{t}{T}) - Y_t(u)|$, the arguments for $|Y_{t,T} - Y_t(\frac{t}{T})|$ being analogous.

NOTATION. Throughout the proof, the symbol $\|z\|$ denotes the Euclidean norm for vectors $z \in \mathbb{R}^d$, and $\|A\|$ is the spectral norm for $d \times d$ matrices $A = (a_{ik})_{i,k=1,\dots,d}$. In addition, $\|A\|_1 = \max_{k=1,\dots,d} \sum_{j=1}^d |a_{jk}|$. Furthermore, for

$z \in \mathbb{R}$, we define the family of matrices

$$B(z) = \begin{pmatrix} z & \cdots & z & z \\ 1 & & 0 & 0 \\ & \ddots & & \vdots \\ 0 & & 1 & 0 \end{pmatrix}.$$

Finally, as already noted at the beginning of the Appendix, we make use of the shorthands $\underline{Y}_{t,T} = Y_{t,T}^{t-d+1}$, $\underline{Y}_t(u) = Y_t^{t-d+1}(u)$ and $\underline{\varepsilon}_t = \varepsilon_t^{t-d+1}$.

BACKWARD ITERATION. By the smoothness conditions on m and σ,

$$Y_t\left(\frac{t}{T}\right) - Y_t(u) = \left(\Delta_{t,0}^m + \Delta_{t,0}^\sigma \varepsilon_t\right) + \sum_{j=1}^d \left(\Delta_{t,j}^m + \Delta_{t,j}^\sigma \varepsilon_t\right)\left(Y_{t-j}\left(\frac{t}{T}\right) - Y_{t-j}(u)\right)$$

with $\Delta_{t,0}^m = m(\frac{t}{T}, \underline{Y}_{t-1}(\frac{t}{T})) - m(u, \underline{Y}_{t-1}(\frac{t}{T}))$ and $\Delta_{t,j}^m = \Delta_j^m(u, \underline{Y}_{t-1}(u), \underline{Y}_{t-1}(\frac{t}{T}))$ for $j = 1, \dots, d$ as introduced in (23). The terms $\Delta_{t,j}^\sigma$ for $j = 0, \dots, d$ are defined analogously. In matrix notation, we obtain

$$\underline{Y}_t\left(\frac{t}{T}\right) - \underline{Y}_t(u) = A_t\left(\underline{Y}_{t-1}\left(\frac{t}{T}\right) - \underline{Y}_{t-1}(u)\right) + \underline{\xi}_t \tag{26}$$

with $\underline{\xi}_t = (\Delta_{t,0}^m + \Delta_{t,0}^\sigma \varepsilon_t, 0, \dots, 0)^T$ and

$$A_t = \begin{pmatrix} \Delta_{t,1}^m + \Delta_{t,1}^\sigma \varepsilon_t & \cdots & \Delta_{t,d-1}^m + \Delta_{t,d-1}^\sigma \varepsilon_t & \Delta_{t,d}^m + \Delta_{t,d}^\sigma \varepsilon_t \\ 1 & & 0 & 0 \\ & \ddots & & \vdots \\ 0 & & 1 & 0 \end{pmatrix}.$$

Iterating (26) n times yields

$$\begin{aligned} \left\| \underline{Y}_t\left(\frac{t}{T}\right) - \underline{Y}_t(u) \right\| &\le \|\underline{\xi}_t\| + \left\| \sum_{r=0}^{n-1} \prod_{k=0}^{r} A_{t-k} \underline{\xi}_{t-r-1} \right\| \\ &\quad + \left\| \prod_{k=0}^{n} A_{t-k}\left(\underline{Y}_{t-n-1}\left(\frac{t}{T}\right) - \underline{Y}_{t-n-1}(u)\right) \right\|. \end{aligned}$$

Note that the rescaled time argument $\frac{t}{T}$ plays the same role as the argument u and thus remains fixed when iterating backward. Next define matrices B_t by

$$B_t = \left(1 + |\varepsilon_t|\right) B(\Delta_t) \tag{27}$$

with $\Delta_t = \Delta I(\|\underline{\varepsilon}_{t-1}\|_\infty \le K_2) + \delta I(\|\underline{\varepsilon}_{t-1}\|_\infty > K_2)$. As shown in the preliminaries section of the Appendix, $|\Delta_{t,j}^m + \Delta_{t,j}^\sigma \varepsilon_t| \le \Delta_t(1 + |\varepsilon_t|)$ for $j = 1, \dots, d$. Therefore, the entries of the matrix B_t are all weakly larger in absolute value than those of A_t. This implies that $\| \prod_{k=0}^n A_{t-k} z\| \le \| \prod_{k=0}^n B_{t-k} |z| \|$ with $z =$

$(|z_1|, \dots, |z_d|)$. Using this together with the boundedness of m and σ and the fact that $|\Delta^m_{t,0} + \Delta^\sigma_{t,0}\varepsilon_t| \le C|\frac{t}{T} - u|(1 + |\varepsilon_t|)$, we finally arrive at

$$\left\| \underline{Y}_t\left(\frac{t}{T}\right) - \underline{Y}_t(u) \right\| \le \left| \frac{t}{T} - u \right| V_{t,n} + R_{t,n}$$

with

$$V_{t,n} = C\bigl(1 + |\varepsilon_t|\bigr) + C \sum_{r=0}^{n-1} \bigl(1 + |\varepsilon_{t-r-1}|\bigr) \left\| \prod_{k=0}^{r} B_{t-k} \right\|,$$

$$R_{t,n} = C\bigl(1 + \|\underline{\varepsilon}_{t-n-1}\|\bigr) \left\| \prod_{k=0}^{n} B_{t-k} \right\|.$$

BOUNDING $V_{t,n}$ AND $R_{t,n}$. The convergence behavior of $V_{t,n}$ and $R_{t,n}$ for $n \to \infty$ mainly depends on the properties of the product $\|\prod_{k=0}^n B_{t-k}\|$. The behavior of the latter is described by the following lemma.

LEMMA A.1. *If δ is sufficiently small, in particular, if it satisfies* (31), *then there exists a constant $\rho > 0$ such that for some $\gamma < 1$,*

$$\mathbb{E}\left[\left\| \prod_{k=0}^{n} B_{t-k} \right\|^\rho \right] \le C\gamma^n. \tag{28}$$

The proof of Lemma A.1 is postponed until the arguments for Theorem 3.2 are completed. The following statement is a direct consequence of Lemma A.1.

(R) There exists a constant $\rho > 0$ such that $\mathbb{E}[R^\rho_{t,n}] \le C\gamma^n$ for some $\gamma < 1$. In particular, $R_{t,n} \xrightarrow{\text{a.s.}} 0$ as $n \to \infty$.

In addition, it holds that:

(V) $V_{t,n} \le V_t$, where the variables V_t have the property that $\mathbb{E}[V_t^\rho] \le C$ for a positive constant $\rho < 1$ and all t.

This can be seen as follows. First note that

$$V_{t,n} \le C\bigl(1 + |\varepsilon_t|\bigr) + \sum_{r=0}^{n-1} R_{t,r} \le V_t := C\bigl(1 + |\varepsilon_t|\bigr) + \sum_{r=0}^{\infty} R_{t,r}.$$

Using the monotone convergence theorem and Loève's inequality with $\rho < 1$, we obtain $\mathbb{E}[V_t^\rho] \le C\mathbb{E}(1 + |\varepsilon_t|)^\rho + \sum_{r=0}^\infty \mathbb{E}[R^\rho_{t,r}]$. As the right-hand side of the previous inequality is finite by (R), we arrive at (V).

(R) and (V) imply that $|Y_t(\frac{t}{T}) - Y_t(u)| \le |\frac{t}{T} - u|V_t$ a.s. with variables V_t whose ρth moment is uniformly bounded by some finite constant C. An analogous result can be derived for $|Y_{t,T} - Y_t(\frac{t}{T})|$. This completes the proof.

PROOF OF LEMMA A.1. We want to show that the ρth moment of the product $\|\prod_{k=0}^{n} B_{t-k}\|$ converges exponentially fast to zero as $n \to \infty$. This is a highly nontrivial problem, and as far as we can see, it cannot be solved by simply adapting techniques from related papers on models with time-varying coefficients. The problem is that the techniques used therein are either tailored to products of deterministic matrices (see, e.g., Proposition 13 in Moulines et al. [19]) or they heavily draw on the independence of the random matrices involved (see, e.g., Proposition 2.1 in Subba Rao [20]).

We now describe our proving strategy in detail. To start with, we replace the spectral norm $\|\cdot\|$ in (28) by the norm $\|\cdot\|_1$ which is much easier to handle. As these two norms are equivalent, there exists a finite constant C such that $\|\prod_{k=0}^{n} B_{t-k}\| \le C\mathcal{B}_n$ with $\mathcal{B}_n = \|\prod_{k=0}^{n} B_{t-k}\|_1$. Next, we split up the term $\mathcal{B}_n$ into two parts,

$$\mathcal{B}_n = I_n \mathcal{B}_n + (1 - I_n)\mathcal{B}_n =: \mathcal{B}_{n,1} + \mathcal{B}_{n,2},$$

where $I_n = I(\sum_{k=0}^{n} J_k > \kappa n)$ with $J_k = I(\min_{l=1,\dots,d} |\varepsilon_{t-k-l}| \le K_2)$ and a constant $0 < \kappa < 1$ to be specified later on. Lemma A.1 is a direct consequence of the following two facts:

(i) There exists a constant $\rho > 0$ such that $\mathbb{E}[\mathcal{B}_{n,1}^{\rho}] \le C\gamma^n$ for some $\gamma < 1$.
(ii) $\mathbb{E}[\mathcal{B}_{n,2}] \le C\gamma^n$ for some $\gamma < 1$.

We start with the proof of (i). Letting $\phi_n = \lambda^n$ with some positive constant $\lambda < 1$, we can write

$$\begin{aligned}\mathbb{E}[\mathcal{B}_{n,1}^{\rho}] &= \mathbb{E}[I(\mathcal{B}_{n,1} > \phi_n)\mathcal{B}_{n,1}^{\rho}] + \mathbb{E}[I(\mathcal{B}_{n,1} \le \phi_n)\mathcal{B}_{n,1}^{\rho}] \\ &\le \left(\mathbb{E}[\mathcal{B}_{n,1}^{2\rho}]\mathbb{P}(\mathcal{B}_{n,1} > \phi_n)\right)^{1/2} + \phi_n^{\rho}.\end{aligned}$$

It is easy to see that $\mathbb{E}[\mathcal{B}_{n,1}^{2\rho}] \le C^{\rho n}$ for a sufficiently large constant C, where C^{ρ} can be made arbitrarily close to one by choosing $\rho > 0$ small enough. To show (i), it thus suffices to verify that

$$\mathbb{P}(\mathcal{B}_{n,1} > \phi_n) \le C\gamma^n \qquad \text{for some } \gamma < 1. \tag{29}$$

For the proof of (29), we write

$$\mathbb{P}(\mathcal{B}_{n,1} > \phi_n) \le \mathbb{P}(I_n > 0) = \mathbb{P}\left(\sum_{k=0}^{n} (J_k - \mathbb{E}[J_k]) > \kappa_0 n\right)$$

with $\kappa_0 := \kappa - \mathbb{E}[J_k]$. As the variables ε_t have an everywhere positive density by assumption, the expectation $\mathbb{E}[J_k]$ is strictly smaller than one. We can thus choose $0 < \kappa < 1$ slightly larger than $\mathbb{E}[J_k]$ to get that $0 < \kappa_0 < 1$. As the variables $J_k - \mathbb{E}[J_k]$ for $k = 0, \dots, n$ are $2d$-dependent, a simple blocking argument together with Hoeffding's inequality shows that

$$\mathbb{P}\left(\sum_{k=0}^{n} (J_k - \mathbb{E}[J_k]) > \kappa_0 n\right) \le C\gamma^n$$

for some $\gamma < 1$. This yields (29) and thus completes the proof of (i).

Let us now turn to the proof of (ii). We have that

$$\mathcal{B}_{n,2} = (1 - I_n) \prod_{k=0}^{n} \big(1 - |\varepsilon_{t-k}|\big) \left\| \prod_{k=0}^{n} B(\Delta_{t-k}) \right\|_1 .$$

The random matrix $B(\Delta_{t-k})$ in the above expression can only take two forms: if $\|\underline{\varepsilon}_{t-k-1}\|_\infty > K_2$, it equals $B(\delta)$, and if $\|\underline{\varepsilon}_{t-k-1}\|_\infty \le K_2$, it equals $B(\Delta)$. Moreover, if $\min_{l=1,\dots,d} |\varepsilon_{t-k-l}| > K_2$, it holds that $\|\underline{\varepsilon}_{t-k-l}\|_\infty > K_2$ for all $l = 1, \dots, d$ and thus $\prod_{l=0}^{d-1} B(\Delta_{t-k-l}) = B(\delta)^d$. Importantly, the term $\mathcal{B}_{n,2}$ is unequal to zero only if $I_n = 0$, that is, only if $\min_{l=1,\dots,d} |\varepsilon_{t-k-l}| > K_2$ for at least $(1-\kappa)n$ terms. From this, we can infer that

$$\mathbb{E}[\mathcal{B}_{n,2}] \le \mathbb{E}\left[\prod_{k=0}^{n} \big(1 + |\varepsilon_{t-k}|\big)\right] \|B(\Delta)\|_1^{\kappa n} \|B(\delta)^d\|_1^{(1-\kappa)n/d}. \tag{30}$$

By direct calculations, we can verify that $\|B(\delta)^d\|_1 \le C_d \delta$ with the constant $C_d = \sum_{l=0}^{d-1} \sum_{k=0}^{l} \binom{l}{k}$ that only depends on the dimension d. Moreover, $\|B(\Delta)\|_1 \le (\Delta + 1)$. Plugging this into (30) yields

$$\mathbb{E}[\mathcal{B}_{n,2}] \le \big(1 + \mathbb{E}|\varepsilon_0|\big)\big[\big(1 + \mathbb{E}|\varepsilon_0|\big)(\Delta + 1)^{\kappa} (C_d \delta)^{(1-\kappa)/d}\big]^n.$$

Straightforward calculations show that the term in square brackets is strictly smaller than one for

$$\delta < \big[\big(1 + \mathbb{E}|\varepsilon_0|\big)^{d/(1-\kappa)} (\Delta + 1)^{\kappa d/(1-\kappa)} C_d\big]^{-1}. \tag{31}$$

Assuming that δ satisfies the above condition, we thus arrive at (ii). $\square$

Proof of Theorem 3.3. The proof can be found in the supplement [22].

Proof of Theorem 3.4. To start with, note that the process $\{Y_{t,T}\}$ is d-Markovian. This implies that

$$\beta(k) = \sup_{T \in \mathbb{Z}} \sup_{t \in \mathbb{Z}} \beta\big(\sigma(\underline{Y}_{t-k,T}), \sigma(\underline{Y}_{t+d-1,T})\big)$$

with

$$\beta\big(\sigma(\underline{Y}_{t-k,T}), \sigma(\underline{Y}_{t+d-1,T})\big) = \mathbb{E}\Big[\sup_{S \in \sigma(\underline{Y}_{t+d-1,T})} \big|\mathbb{P}(S) - \mathbb{P}\big(S|\sigma(\underline{Y}_{t-k,T})\big)\big|\Big].$$

In the following, we bound the expression $|\mathbb{P}(S) - \mathbb{P}(S|\sigma(\underline{Y}_{t-k,T}))|$ for arbitrary sets $S \in \sigma(\underline{Y}_{t+d-1,T})$. This provides us with a bound for the mixing coefficients $\beta(k)$ of the process $\{Y_{t,T}\}$.

We use the following notation: throughout the proof, we let $y = y_{t+d-1}^{t}$, $e = e_{t-1}^{t-k+1}$ and $z = z_{t-k}^{t-k-d+1}$ be values of $\underline{Y}_{t+d-1,T}$, $\varepsilon_{t-1}^{t-k+1}$ and $\underline{Y}_{t-k,T}$, respectively.

Moreover, we use the shorthand

$$f_j(y_{t+j}|z) = f_{Y_{t+j,T}|Y^t_{t+j-1,T},\varepsilon^{t-k+1}_{t-1},\underline{Y}_{t-k,T}}(y_{t+j}|y^t_{t+j-1}, e, z)$$

for $j = 0, \dots, d-1$, where we suppress the dependence on the arguments y^t_{t+j-1} and e in the notation. Finally, note that by (25), the above conditional density can be expressed in terms of the error density f_ε as

$$f_j(y_{t+j}|z) = \frac{1}{\sigma_{t,T,j}(z)} f_\varepsilon\left(\frac{y_{t+j} - m_{t,T,j}(z)}{\sigma_{t,T,j}(z)}\right) \tag{32}$$

with

$$\begin{aligned} m_{t,T,j}(z) = m\biggl(\frac{t+j}{T}, y^t_{t+j-1}, m^{(k-2)}_{t-1,T}(e^{t-k+1}_{t-2}, z) + \sigma^{(k-2)}_{t-1,T}(e^{t-k+1}_{t-2}, z)e_{t-1}, \dots, \\ m^{(k-j+d-1)}_{t+j-d,T}(e^{t-k+1}_{t+j-d-1}, z) + \sigma^{(k-j+d-1)}_{t+j-d,T}(e^{t-k+1}_{t+j-d-1}, z)e_{t+j-d}\biggr) \end{aligned}$$

and $\sigma_{t,T,j}(z)$ defined analogously. The functions $m^{(k-2)}_{t-1,T}$, $\sigma^{(k-2)}_{t-1,T}, \dots$ were introduced in the preliminaries section of the Appendix.

With this notation at hand, we can write

$$\begin{aligned} &\mathbb{P}(S|\sigma(\underline{Y}_{t-k,T})) \\ &\quad = \mathbb{E}[\mathbb{E}[I(\underline{Y}_{t+d-1,T} \in S)|\varepsilon^{t-k+1}_{t-1}, \underline{Y}_{t-k,T}]|\underline{Y}_{t-k,T}] \\ &\quad = \int I(y \in S) f_{\underline{Y}_{t+d-1,T}|\varepsilon^{t-k+1}_{t-1},\underline{Y}_{t-k,T}}(y|e, \underline{Y}_{t-k,T}) \prod_{l=1}^{k-1} f_\varepsilon(e_{t-l})\,de\,dy \\ &\quad = \int I(y \in S) \prod_{j=0}^{d-1} f_j(y_{t+j}|\underline{Y}_{t-k,T}) \prod_{l=1}^{k-1} f_\varepsilon(e_{t-l})\,de\,dy \end{aligned}$$

and likewise

$$\mathbb{P}(S) = \int I(y \in S) \prod_{j=0}^{d-1} f_j(y_{t+j}|z) \prod_{l=1}^{k-1} f_\varepsilon(e_{t-l}) f_{\underline{Y}_{t-k,T}}(z)\,de\,dz\,dy.$$

Using the shorthand $\underline{Y} = \underline{Y}_{t-k,T}$, we thus arrive at

$$\begin{aligned} &|\mathbb{P}(S) - \mathbb{P}(S|\sigma(\underline{Y}))| \\ &\quad \le \int \underbrace{\left[\int \left|\prod_{j=0}^{d-1} f_j(y_{t+j}|z) - \prod_{j=0}^{d-1} f_j(y_{t+j}|\underline{Y})\right| dy\right]}_{=:(*)} \prod_{l=1}^{k-1} f_\varepsilon(e_{t-l}) f_{\underline{Y}}(z)\,de\,dz. \end{aligned}$$

We next consider $(*)$ more closely. A telescoping argument together with Fubini's theorem yields that

$$
\begin{aligned}
(*) &\le \sum_{i=0}^{d-1} \int \Biggl[\prod_{j=0}^{i-1} f_j(y_{t+j}|\underline{Y}) \bigl| f_i(y_{t+i}|z) - f_i(y_{t+i}|\underline{Y}) \bigr| \prod_{j=i+1}^{d-1} f_j(y_{t+j}|z) \Biggr] dy \\
&= \sum_{i=0}^{d-1} \int \Biggl[\int \Biggl[\int \prod_{j=i+1}^{d-1} f_j(y_{t+j}|z)\, dy_{t+d-1} \cdots dy_{t+i+1} \Biggr] \\
&\qquad\qquad \times \bigl| f_i(y_{t+i}|z) - f_i(y_{t+i}|\underline{Y}) \bigr|\, dy_{t+i} \Biggr] \\
&\qquad \times \prod_{j=0}^{i-1} f_j(y_{t+j}|\underline{Y})\, dy_{t+i-1} \cdots dy_t \\
&\le \sum_{i=0}^{d-1} \int \underbrace{\Biggl[\int \bigl| f_i(y_{t+i}|z) - f_i(y_{t+i}|\underline{Y}) \bigr|\, dy_{t+i} \Biggr]}_{=:(**)} \prod_{j=0}^{i-1} f_j(y_{t+j}|\underline{Y})\, dy_{t+i-1} \cdots dy_t,
\end{aligned}
$$

where the last inequality exploits the fact that

$$
\int \prod_{j=i+1}^{d-1} f_j(y_{t+j}|z)\, dy_{t+d-1} \cdots dy_{t+i+1}
$$

is a conditional probability and thus almost surely bounded by one. Using formula (32) together with (E3), it is straightforward to see that

$$
\begin{aligned}
(**) &= \int \Biggl| \frac{1}{\sigma_{t,T,i}(z)} f_\varepsilon\Biggl(\frac{y_{t+i} - m_{t,T,i}(z)}{\sigma_{t,T,i}(z)} \Biggr) \\
&\qquad - \frac{1}{\sigma_{t,T,i}(\underline{Y})} f_\varepsilon\Biggl(\frac{y_{t+i} - m_{t,T,i}(\underline{Y})}{\sigma_{t,T,i}(\underline{Y})} \Biggr) \Biggr| dy_{t+i} \\
&\le C\bigl(\bigl| m_{t,T,i}(z) - m_{t,T,i}(\underline{Y}) \bigr| + \bigl| \sigma_{t,T,i}(z) - \sigma_{t,T,i}(\underline{Y}) \bigr| \bigr) \\
&\le C(2C_m + 2C_\sigma)\bigl(\bigl| m_{t,T,i}(z) - m_{t,T,i}(\underline{Y}) \bigr| + \bigl| \sigma_{t,T,i}(z) - \sigma_{t,T,i}(\underline{Y}) \bigr| \bigr)^p,
\end{aligned}
$$

where p is some constant with $0 < p < 1$. Iterating backward $n \le k - 2d$ times in the same way as in Theorem 3.2, we can further show that

$$
\begin{aligned}
&\bigl| m_{t,T,i}(z) - m_{t,T,i}(\underline{Y}) \bigr| + \bigl| \sigma_{t,T,i}(z) - \sigma_{t,T,i}(\underline{Y}) \bigr| \\
&\qquad \le C \sum_{j=1}^{d-i} \Biggl\| \prod_{m=0}^{n} B_{t-j-m} \Biggr\| \bigl(1 + \| e_{t-j-n-1}^{t-j-n-d} \| \bigr),
\end{aligned}
\tag{33}
$$

where $\|\cdot\|$ denotes the Euclidean norm for vectors and the spectral norm for matrices. The matrix B_t was introduced in (27). Note that B_t was defined there in terms of the random vector ε_t^{t-d}. Slightly abusing notation, we here use the symbol B_t to denote the matrix with ε_t^{t-d} replaced by the realization e_t^{t-d}. Keeping in mind that the matrix B_t only depends on the residual values e_t^{t-d}, we can plug (33) into the bound for $(**)$ and insert this into the bound for $(*)$ to arrive at

$$(*) \leq C\left(\sum_{j=1}^{d}\left\|\prod_{m=0}^{n} B_{t-j-m}\right\|\left(1+\|e_{t-j-n-1}^{t-j-n-d}\|\right)\right)^{p}.$$

As a consequence,

$$\left|\mathbb{P}(S)-\mathbb{P}\left(S|\sigma(\underline{Y})\right)\right| \leq C\mathbb{E}\left(\sum_{j=1}^{d}\left\|\prod_{m=0}^{n} B_{t-j-m}\right\|\left(1+\|\varepsilon_{t-j-n-1}^{t-j-n-d}\|\right)\right)^{p}.$$

Using the arguments from Lemma A.1, we can show that for $p>0$ sufficiently small, the expectation on the right-hand side is bounded by $C\lambda^n$ for some positive constant $\lambda<1$. Choosing $n=k-2d$, for instance, we thus arrive at

$$\left|\mathbb{P}(S)-\mathbb{P}\left(S|\sigma(\underline{Y}_{t-k,T})\right)\right| \leq C\lambda^{k-(d+1)} \leq C\gamma^{k}$$

for some constant $\gamma<1$. This immediately implies that $\beta(k) \leq C\gamma^k$.

APPENDIX B

In this Appendix, we prove the results of Section 4. Before we turn to the proofs, we state two auxiliary lemmas which are repeatedly used throughout the Appendix. The proofs are straightforward and thus omitted.

LEMMA B.1. *Suppose the kernel K satisfies (C6) and let $I_h=[C_1h, 1-C_1h]$. Then for $k=0,1,2$,*

$$\sup_{u\in I_h}\left|\frac{1}{Th}\sum_{t=1}^{T}K_h\left(u-\frac{t}{T}\right)\left(\frac{u-t/T}{h}\right)^k-\int_0^1\frac{1}{h}K_h(u-\varphi)\left(\frac{u-\varphi}{h}\right)^k d\varphi\right|$$
$$=O\left(\frac{1}{Th^2}\right).$$

LEMMA B.2. *Suppose K satisfies (C6) and let $g:[0,1]\times\mathbb{R}^d\to\mathbb{R}$, $(u,x)\mapsto g(u,x)$ be continuously differentiable w.r.t. u. Then for any compact set $S\subset\mathbb{R}^d$,*

$$\sup_{u\in I_h, x\in S}\left|\frac{1}{Th}\sum_{t=1}^{T}K_h\left(u-\frac{t}{T}\right)g\left(\frac{t}{T},x\right)-g(u,x)\right|=O\left(\frac{1}{Th^2}\right)+o(h).$$

Proof of Theorem 4.1. To show the result, we use a blocking argument together with an exponential inequality for mixing arrays, thus following the common proving strategy to be found, for example, in Bosq [3], Masry [18] or

Hansen [13]. In particular, we go along the lines of Hansen's proof of Theorem 2 in [13], modifying his arguments to allow for local stationarity in the data. A detailed version of the arguments can be found in the supplement [22].

Proof of Theorem 4.2. We write

$$\hat{m}(u,x) - m(u,x) = \frac{1}{\hat{f}(u,x)}\big(\hat{g}^V(u,x) + \hat{g}^B(u,x) - m(u,x)\hat{f}(u,x)\big)$$

with

$$\hat{f}(u,x) = \frac{1}{Th^{d+1}}\sum_{t=1}^{T} K_h\left(u - \frac{t}{T}\right)\prod_{j=1}^{d} K_h\big(x^j - X_{t,T}^j\big),$$

$$\hat{g}^V(u,x) = \frac{1}{Th^{d+1}}\sum_{t=1}^{T} K_h\left(u - \frac{t}{T}\right)\prod_{j=1}^{d} K_h\big(x^j - X_{t,T}^j\big)\varepsilon_{t,T},$$

$$\hat{g}^B(u,x) = \frac{1}{Th^{d+1}}\sum_{t=1}^{T} K_h\left(u - \frac{t}{T}\right)\prod_{j=1}^{d} K_h\big(x^j - X_{t,T}^j\big)m\left(\frac{t}{T}, X_{t,T}\right).$$

We first derive some intermediate results for the above expressions:

(i) By Theorem 4.1 with $W_{t,T} = \varepsilon_{t,T}$,

$$\sup_{u\in[0,1],x\in S}\big|\hat{g}^V(u,x)\big| = O_p\left(\sqrt{\frac{\log T}{Th^{d+1}}}\right).$$

(ii) Applying the arguments for Theorem 4.1 to $\hat{g}^B(u,x) - m(u,x)\hat{f}(u,x)$ yields

$$\begin{aligned}\sup_{u\in[0,1],x\in S}\big|&\hat{g}^B(u,x) - m(u,x)\hat{f}(u,x)\\ &- \mathbb{E}\big[\hat{g}^B(u,x) - m(u,x)\hat{f}(u,x)\big]\big|\\ = O_p&\left(\sqrt{\frac{\log T}{Th^{d+1}}}\right).\end{aligned}$$

(iii) It holds that

$$\begin{aligned}&\sup_{u\in I_h,x\in S}\big|\mathbb{E}\big[\hat{g}^B(u,x) - m(u,x)\hat{f}(u,x)\big]\big|\\ &\quad = h^2\frac{\kappa_2}{2}\sum_{i=0}^{d}\big(2\,\partial_i m(u,x)\,\partial_i f(u,x) + \partial_{ii}^2 m(u,x)f(u,x)\big)\\ &\qquad + O\left(\frac{1}{T^r h^d}\right) + o\big(h^2\big)\end{aligned}$$

with $r = \min\{\rho, 1\}$. The proof is postponed until the arguments for Theorem 4.2 are completed.

(iv) We have that

$$\sup_{u \in I_h, x \in S} \left|\hat{f}(u,x) - f(u,x)\right| = o_p(1).$$

For the proof, we split up the term $\hat{f}(u,x) - f(u,x)$ into a variance part $\hat{f}(u,x) - \mathbb{E}\hat{f}(u,x)$ and a bias part $\mathbb{E}\hat{f}(u,x) - f(u,x)$. Applying Theorem 4.1 with $W_{t,T} = 1$ yields that the variance part is $o_p(1)$ uniformly in u. The bias part can be analyzed by a simplified version of the arguments used to prove (iii).

Combining the intermediate results (i)–(iii), we arrive at

$$\begin{aligned}
&\sup_{u \in I_h, x \in S} \left|\hat{m}(u,x) - m(u,x)\right| \\
&\quad \le \left(\sup \hat{f}(u,x)^{-1}\right)\left(\sup\left|\hat{g}^V(u,x)\right| + \sup\left|\hat{g}^B(u,x) - m(u,x)\hat{f}(u,x)\right|\right) \\
&\quad = \left(\sup \hat{f}(u,x)^{-1}\right) O_p\left(\sqrt{\frac{\log T}{Th^{d+1}}} + \frac{1}{T^r h^d} + h^2\right)
\end{aligned}$$

with $r = \min\{\rho, 1\}$. Moreover, (iv) and the condition that $\inf_{u \in [0,1], x \in S} f(u,x) > 0$ immediately imply that $\sup \hat{f}(u,x)^{-1} = O_p(1)$. This completes the proof.

PROOF OF (iii). Let $\bar{K} : \mathbb{R} \to \mathbb{R}$ be a Lipschitz continuous function with support $[-qC_1, qC_1]$ for some $q > 1$. Assume that $\bar{K}(x) = 1$ for all $x \in [-C_1, C_1]$ and write $\bar{K}_h(x) = \bar{K}(\frac{x}{h})$. Then

$$\mathbb{E}\left[\hat{g}^B(u,x) - m(u,x)\hat{f}(u,x)\right] = Q_1(u,x) + \cdots + Q_4(u,x)$$

with

$$Q_i(u,x) = \frac{1}{Th^{d+1}} \sum_{t=1}^{T} K_h\left(u - \frac{t}{T}\right) q_i(u,x)$$

and

$$\begin{aligned}
q_1(u,x) = \mathbb{E}\Bigg[\prod_{j=1}^{d} \bar{K}_h\left(x^j - X_{t,T}^j\right) &\Bigg\{\prod_{j=1}^{d} K_h\left(x^j - X_{t,T}^j\right) \\
&\quad - \prod_{j=1}^{d} K_h\left(x^j - X_t^j\left(\frac{t}{T}\right)\right)\Bigg\} \\
&\quad \times \left\{m\left(\frac{t}{T}, X_{t,T}\right) - m(u,x)\right\}\Bigg],
\end{aligned}$$

$$q_2(u,x) = \mathbb{E}\Bigg[\prod_{j=1}^{d} \bar{K}_h(x^j - X_{t,T}^j) \prod_{j=1}^{d} K_h\Big(x^j - X_t^j\Big(\frac{t}{T}\Big)\Big) \times \Big\{m\Big(\frac{t}{T}, X_{t,T}\Big) - m\Big(\frac{t}{T}, X_t\Big(\frac{t}{T}\Big)\Big)\Big\}\Bigg],$$

$$q_3(u,x) = \mathbb{E}\Bigg[\Big\{\prod_{j=1}^{d} \bar{K}_h(x^j - X_{t,T}^j) - \prod_{j=1}^{d} \bar{K}_h\Big(x^j - X_t^j\Big(\frac{t}{T}\Big)\Big)\Big\} \times \prod_{j=1}^{d} K_h\Big(x^j - X_t^j\Big(\frac{t}{T}\Big)\Big)\Big\{m\Big(\frac{t}{T}, X_t\Big(\frac{t}{T}\Big)\Big) - m(u,x)\Big\}\Bigg],$$

$$q_4(u,x) = \mathbb{E}\Bigg[\prod_{j=1}^{d} K_h\Big(x^j - X_t^j\Big(\frac{t}{T}\Big)\Big)\Big\{m\Big(\frac{t}{T}, X_t\Big(\frac{t}{T}\Big)\Big) - m(u,x)\Big\}\Bigg].$$

We first consider $Q_1(u,x)$. As the kernel K is bounded, we can use a telescoping argument to get that $|\prod_{j=1}^{d} K_h(x^j - X_{t,T}^j) - \prod_{j=1}^{d} K_h(x^j - X_t^j(\frac{t}{T}))| \le C\sum_{k=1}^{d} |K_h(x^k - X_{t,T}^k) - K_h(x^k - X_t^k(\frac{t}{T}))|$. Once again exploiting the boundedness of K, we can find a constant $C < \infty$ with $|K_h(x^k - X_{t,T}^k) - K_h(x^k - X_t^k(\frac{t}{T}))| \le C|K_h(x^k - X_{t,T}^k) - K_h(x^k - X_t^k(\frac{t}{T}))|^r$ for $r = \min\{\rho, 1\}$. Hence,

$$\Bigg|\prod_{j=1}^{d} K_h(x^j - X_{t,T}^j) - \prod_{j=1}^{d} K_h\Big(x^j - X_t^j\Big(\frac{t}{T}\Big)\Big)\Bigg| \le C\sum_{k=1}^{d}\Bigg|K_h(x^k - X_{t,T}^k) - K_h\Big(x^k - X_t^k\Big(\frac{t}{T}\Big)\Big)\Bigg|^r. \tag{34}$$

Using (34), we obtain

$$\begin{aligned}|Q_1(u,x)| &\le \frac{C}{Th^{d+1}}\sum_{t=1}^{T} K_h\Big(u - \frac{t}{T}\Big) \\ &\quad \times \mathbb{E}\Bigg[\sum_{k=1}^{d}\Bigg|K_h(x^k - X_{t,T}^k) - K_h\Big(x^k - X_t^k\Big(\frac{t}{T}\Big)\Big)\Bigg|^r \\ &\quad \times \prod_{j=1}^{d} \bar{K}_h(x^j - X_{t,T}^j)\Bigg|m\Big(\frac{t}{T}, X_{t,T}\Big) - m(u,x)\Bigg|\Bigg]\end{aligned}$$

with $r = \min\{\rho, 1\}$. The term $\prod_{j=1}^{d} \bar{K}_h(x^j - X_{t,T}^j)|m(\frac{t}{T}, X_{t,T}) - m(u,x)|$ in the above expression can be bounded by Ch. Since K is Lipschitz, $|X_{t,T}^k - X_t^k(\frac{t}{T})| \le$

$\frac{C}{T}U_{t,T}(\frac{t}{T})$ and the variables $U_{t,T}(\frac{t}{T})$ have finite rth moment, we can infer that

$$\begin{aligned}
&|Q_1(u,x)| \\
&\quad \le \frac{C}{Th^d}\sum_{t=1}^{T} K_h\left(u-\frac{t}{T}\right)\mathbb{E}\left[\sum_{k=1}^{d}\left|K_h(x^k - X_{t,T}^k) - K_h\left(x^k - X_t^k\left(\frac{t}{T}\right)\right)\right|^r\right] \\
&\quad \le \frac{C}{Th^d}\sum_{t=1}^{T} K_h\left(u-\frac{t}{T}\right)\mathbb{E}\left[\sum_{k=1}^{d}\left|\frac{1}{Th}U_{t,T}\left(\frac{t}{T}\right)\right|^r\right] \\
&\quad \le \frac{C}{T^r h^{d-1+r}}
\end{aligned}$$

uniformly in u and x. Using similar arguments, we can further show that $\sup_{u,x}|Q_2(u,x)| \le \frac{C}{T^r h^d}$ and $\sup_{u,x}|Q_3(u,x)| \le \frac{C}{T^r h^{d-1+r}}$. Finally, applying Lemmas B.1 and B.2 and exploiting the smoothness conditions on m and f, we obtain that uniformly in u and x,

$$Q_4(u,x) = h^2\frac{\kappa_2}{2}\sum_{i=0}^{d}\big(2\,\partial_i m(u,x)\,\partial_i f(u,x) + \partial_{ii}^2 m(u,x) f(u,x)\big) + o(h^2).$$

Combining the results on $Q_1(u,x), \dots, Q_4(u,x)$ yields (iii). □

Proof of Theorem 4.3. The result can be shown by using the techniques from Theorem 4.2 together with a blocking argument. More details are given in the supplement [22].

APPENDIX C

In this Appendix, we prove the results concerning the smooth backfitting estimates of Section 5. Throughout the Appendix, conditions (Add1) and (Add2) are assumed to be satisfied.

Auxiliary results. Before we come to the proof of Theorems 5.1 and 5.2, we provide results on uniform convergence rates for the kernel smoothers that are used as pilot estimates in the smooth backfitting procedure. We start with an auxiliary lemma which is needed to derive the various rates.

LEMMA C.1. *Define $T_0 = \mathbb{E}[T_{[0,1]^d}]$. Then uniformly for $u \in I_h$,*

$$\frac{T_0}{T} = \mathbb{P}\big(X_0(u) \in [0,1]^d\big) + O\big(T^{-\rho/(1+\rho)}\big) + o(h) \tag{35}$$

with ρ defined in assumption (C1) and

$$\frac{T_{[0,1]^d} - T_0}{T_0} = O_p\left(\sqrt{\frac{\log T}{Th}}\right). \tag{36}$$

PROOF. The proof can be found in the supplement [22]. $\square$

We now examine the convergence behavior of the pilot estimates of the backfitting procedure. We first consider the density estimates $\hat{p}_j$ and $\hat{p}_{j,k}$.

LEMMA C.2. *Define $v_{T,2} = \sqrt{\log T/Th^2}$, $v_{T,3} = \sqrt{\log T/Th^3}$ and $b_{T,r} = T^{-r}h^{-(d+r)}$ with $r = \min\{\rho, 1\}$. Moreover, let $\kappa_0(w) = \int K_h(w, v)\,dv$. Then*

$$\sup_{u,x^j \in I_h} |\hat{p}_j(u, x^j) - p_j(u, x^j)| = O_p(v_{T,2}) + O(b_{T,r}) + o(h),$$

$$\sup_{u \in I_h, x^j \in [0,1]} |\hat{p}_j(u, x^j) - \kappa_0(x^j) p_j(u, x^j)| = O_p(v_{T,2}) + O(b_{T,r}) + O(h),$$

$$\sup_{u,x^j,x^k \in I_h} |\hat{p}_{j,k}(u, x^j, x^k) - p_{j,k}(u, x^j, x^k)| = O_p(v_{T,3}) + O(b_{T,r}) + o(h),$$

$$\begin{aligned}\sup_{\substack{u \in I_h,\\ x^j, x^k \in [0,1]}} &|\hat{p}_{j,k}(u, x^j, x^k) - \kappa_0(x^j)\kappa_0(x^k) p_{j,k}(u, x^j, x^k)| \\ &= O_p(v_{T,3}) + O(b_{T,r}) + O(h).\end{aligned}$$

PROOF. We only consider the term $\hat{p}_j$, the proof for $\hat{p}_{j,k}$ being analogous. Defining $\check{p}_j(u, x^j) = (T_0)^{-1} \sum_{t=1}^T I(X_{t,T} \in [0,1]^d) K_h(u, \frac{t}{T}) K_h(x^j, X_{t,T}^j)$ with $T_0 = \mathbb{E}[T_{[0,1]^d}]$, we obtain that

$$\begin{aligned}\hat{p}_j(u, x^j) &= \left[1 + \frac{T_{[0,1]^d} - T_0}{T_0}\right]^{-1} \check{p}_j(u, x^j) \\ &= \left[1 - \frac{T_{[0,1]^d} - T_0}{T_0} + O_p\left(\frac{T_{[0,1]^d} - T_0}{T_0}\right)^2\right] \check{p}_j(u, x^j).\end{aligned}$$

By (36) from Lemma C.1, this implies that

$$\hat{p}_j(u, x^j) = \check{p}_j(u, x^j) + O_p(\sqrt{\log T/Th})$$

uniformly for $u \in I_h$ and $x^j \in [0, 1]$. Applying the proving strategy of Theorem 4.2 to $\check{p}_j(u, x^j)$ completes the proof. $\square$

We next examine the Nadaraya–Watson smoother $\hat{m}_j$. To this purpose, we decompose it into a variance part $\hat{m}_j^A$ and a bias part $\hat{m}_j^B$. The decomposition is given by $\hat{m}_j(u, x^j) = \hat{m}_j^A(u, x^j) + \hat{m}_j^B(u, x^j)$ with

$$\hat{m}_j^A(u, x^j) = \frac{1}{T_{[0,1]^d}} \sum_{t=1}^{T} I(X_{t,T} \in [0,1]^d) K_h\left(u, \frac{t}{T}\right) K_h(x^j, X_{t,T}^j) \varepsilon_{t,T} / \hat{p}_j(u, x^j),$$

$$\hat{m}_j^B(u, x^j) = \frac{1}{T_{[0,1]^d}} \sum_{t=1}^{T} I(X_{t,T} \in [0,1]^d) K_h\left(u, \frac{t}{T}\right) K_h(x^j, X_{t,T}^j) \\ \times \left(m_0\left(\frac{t}{T}\right) + \sum_{k=1}^{d} m_k\left(\frac{t}{T}, X_{t,T}^k\right)\right) \Big/ \hat{p}_j(u, x^j).$$

The next two lemmas characterize the asymptotic behavior of $\hat{m}_j^A$ and $\hat{m}_j^B$.

LEMMA C.3. *It holds that*

$$\sup_{u, x^j \in [0,1]} \left|\hat{m}_j^A(u, x^j)\right| = O_p\left(\sqrt{\frac{\log T}{Th^2}}\right). \tag{37}$$

PROOF. Replacing the occurrences of $T_{[0,1]^d}$ in $\hat{m}_j^A$ by $T_0 = \mathbb{E}[T_{[0,1]^d}]$ and then applying Theorem 4.1 gives the result. □

LEMMA C.4. *It holds that*

$$\sup_{u, x^j \in I_h} \left|\hat{m}_j^B(u, x^j) - \hat{\mu}_{T,j}(u, x^j)\right| = o_p(h^2), \tag{38}$$

$$\sup_{u \in I_h, x^j \in I_h^c} \left|\hat{m}_j^B(u, x^j) - \hat{\mu}_{T,j}(u, x^j)\right| = O_p(h^2) \tag{39}$$

with $I_h^c = [0,1] \setminus I_h$ *and*

$$\hat{\mu}_{T,j}(u, x^j) = \alpha_{T,0}(u) + \alpha_{T,j}(u, x^j) \\ + \sum_{k \neq j} \int \alpha_{T,k}(u, x^k) \frac{\hat{p}_{j,k}(u, x^j, x^k)}{\hat{p}_j(u, x^j)} \, dx^k \\ + h^2 \int \beta(u, x) \frac{p(u, x)}{p_j(u, x^j)} \, dx^{-j}.$$

Here,

$$\alpha_{T,0}(u) = m_0(u) + h\kappa_1(u)\, \partial_u m_0(u) + \frac{h^2}{2} \kappa_2(u)\, \partial_{uu}^2 m_0(u),$$

$$\alpha_{T,k}(u, x^k) = m_k(u, x^k) + h\left[\kappa_1(u)\, \partial_u m_k(u, x^k) + \frac{\kappa_0(u)\kappa_1(x^k)}{\kappa_0(x^k)}\, \partial_{x^k} m_k(u, x^k)\right],$$

$$
\begin{aligned}
\beta(u,x) = {} & \kappa_2\, \partial_u m_0(u)\, \partial_u \log p(u,x) \\
& + \sum_{k=1}^{d} \biggl\{ \kappa_2\, \partial_u m_k\bigl(u,x^k\bigr)\, \partial_u \log p(u,x) + \frac{\kappa_2}{2}\, \partial_{uu}^2 m_k\bigl(u,x^k\bigr) \\
& \qquad + \kappa_2\, \partial_{x^k} m_k\bigl(u,x^k\bigr)\, \partial_{x^k} \log p(u,x) + \frac{\kappa_2}{2}\, \partial_{x^k x^k}^2 m_k\bigl(u,x^k\bigr) \biggr\},
\end{aligned}
$$

where the symbol $\partial_z g$ denotes the partial derivative of the function g with respect to z and $\kappa_2 = \int w^2 K(w)\,dw$ as well as $\kappa_l(v) = \int w^l K_h(v,w)\,dw$ for $l = 0,1,2$.

PROOF. As the proof is rather lengthy and involved, we only sketch its idea. A detailed version can be found in the supplement [22]. To provide the stochastic expansion of $\hat{m}_j^B(u,x^j)$ in (38) and (39), we follow the proving strategy of Theorem 4 in Mammen et al. [17]. Adapting this strategy is, however, not completely straightforward. The complication mainly results from the fact that we cannot work with the variables $X_{t,T}$ directly but have to replace them by the approximations $X_t(\frac{t}{T})$. To cope with the resulting difficulties, we exploit (35) and (36) of Lemma C.1 and use arguments similar to those for Theorem 4.2. □

We finally state a result on the convergence behavior of the term $\tilde{m}_0(u)$.

LEMMA C.5. *It holds that*

$$
\sup_{u \in I_h} \bigl| \tilde{m}_0(u) - m_0(u) \bigr| = O_p\biggl(\sqrt{\frac{\log T}{Th}} + h^2 \biggr). \tag{40}
$$

PROOF. The claim can be shown by replacing $T_{[0,1]^d}$ with $T_0 = \mathbb{E}[T_{[0,1]^d}]$ in the expression for $\tilde{m}_0(u)$ and then using arguments from Theorem 4.2. □

Proof of Theorems 5.1 and 5.2. Using the auxiliary results from the previous subsection, it is not difficult to show that the high-level conditions (A1)–(A6), (A8) and (A9) of Mammen et al. [17] are satisfied. We can thus apply their Theorems 1–3, which imply the statements of Theorems 5.1 and 5.2. Note that the high-level conditions are satisfied uniformly for $u \in I_h$ rather than only pointwise. Inspecting the proofs of Theorems 1–3 in [17], this allows us to infer that the convergence rates in (21) hold uniformly over $u \in I_h$ rather than only pointwise. A list of the high-level conditions together with the details of the proof can be found in the supplement [22].

Acknowledgments. I am grateful to Enno Mammen, Oliver Linton and Suhasini Subba Rao for numerous helpful suggestions and comments. Moreover, I would like to thank an Associate Editor and three anonymous referees for their constructive comments which helped a lot to improve the paper.

SUPPLEMENTARY MATERIAL

Additional technical details (DOI: 10.1214/12-AOS1043SUPP; .pdf). The proofs and technical details that are omitted in the Appendices are provided in the supplement that accompanies the paper.

REFERENCES

[1] AN, H. Z. and HUANG, F. C. (1996). The geometrical ergodicity of nonlinear autoregressive models. *Statist. Sinica* **6** 943–956. MR1422412

[2] BHATTACHARYA, R. and LEE, C. (1995). On geometric ergodicity of nonlinear autoregressive models. *Statist. Probab. Lett.* **22** 311–315. MR1333189

[3] BOSQ, D. (1998). *Nonparametric Statistics for Stochastic Processes: Estimation and Prediction*, 2nd ed. *Lecture Notes in Statistics* **110**. Springer, New York. MR1640691

[4] CHEN, M. and CHEN, G. (2000). Geometric ergodicity of nonlinear autoregressive models with changing conditional variances. *Canad. J. Statist.* **28** 605–613. MR1793114

[5] DAHLHAUS, R. (1996). On the Kullback–Leibler information divergence of locally stationary processes. *Stochastic Process. Appl.* **62** 139–168. MR1388767

[6] DAHLHAUS, R. (1996). Asymptotic statistical inference for nonstationary processes with evolutionary spectra. In *Athens Conference on Applied Probability and Time Series Analysis, Vol. II* (1995) (P. M. Robinson and M. Rosenblatt, eds.). *Lecture Notes in Statistics* **115** 145–159. Springer, New York. MR1466743

[7] DAHLHAUS, R. (1997). Fitting time series models to nonstationary processes. *Ann. Statist.* **25** 1–37. MR1429916

[8] DAHLHAUS, R., NEUMANN, M. H. and VON SACHS, R. (1999). Nonlinear wavelet estimation of time-varying autoregressive processes. *Bernoulli* **5** 873–906. MR1715443

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

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

[11] FRYZLEWICZ, P. and SUBBA RAO, S. (2011). Mixing properties of ARCH and time-varying ARCH processes. *Bernoulli* **17** 320–346. MR2797994

[12] HAFNER, C. M. and LINTON, O. (2010). Efficient estimation of a multivariate multiplicative volatility model. *J. Econometrics* **159** 55–73. MR2720844

[13] HANSEN, B. E. (2008). Uniform convergence rates for kernel estimation with dependent data. *Econometric Theory* **24** 726–748. MR2409261

[14] KOO, B. and LINTON, O. (2012). Semiparametric estimation of locally stationary diffusion models. *J. Econometrics* **170** 210–233.

[15] KRISTENSEN, D. (2009). Uniform convergence rates of kernel estimators with heterogeneous dependent data. *Econometric Theory* **25** 1433–1445. MR2540506

[16] KRISTENSEN, D. (2011). Stationary approximations of time-inhomogenous Markov chains with applications. Mimeo.

[17] MAMMEN, E., LINTON, O. and NIELSEN, J. (1999). The existence and asymptotic properties of a backfitting projection algorithm under weak conditions. *Ann. Statist.* **27** 1443–1490. MR1742496

[18] MASRY, E. (1996). Multivariate local polynomial regression for time series: Uniform strong consistency and rates. *J. Time Series Anal.* **17** 571–599. MR1424907

[19] MOULINES, E., PRIOURET, P. and ROUEFF, F. (2005). On recursive estimation for time varying autoregressive processes. *Ann. Statist.* **33** 2610–2654. MR2253097

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

[21] TJØSTHEIM, D. (1990). Non-linear time series and Markov chains. *Adv. in Appl. Probab.* **22** 587–611.
[22] VOGT, M. (2012). Supplement to "Nonparametric regression for locally stationary time series." DOI:10.1214/12-AOS1043SUPP.
[23] ZHOU, Z. (2010). Nonparametric inference of quantile curves for nonstationary time series. *Ann. Statist.* **38** 2187–2217. MR2676887
[24] ZHOU, Z. and WU, W. B. (2009). Local linear quantile estimation for nonstationary time series. *Ann. Statist.* **37** 2696–2729. MR2541444

DEPARTMENT OF ECONOMICS
UNIVERSITY OF CAMBRIDGE
SIDGWICK AVENUE
CAMBRIDGE CB3 9DD
UNITED KINGDOM
E-MAIL: mv346@cam.ac.uk