cemmap

centre for microdata methods and practice

Nonparametric regression for locally stationary time series

Michael Vogt

The Institute for Fiscal Studies
Department of Economics, UCL

cemmap working paper CWP22/12

An ESRC Research Centre

Institute for Fiscal Studies, 7 Ridgmount Street, London WC1E 7AE **Tel:** 44 (0)20 7291 4800 **Email:** info@cemmap.ac.uk **www.cemmap.ac.uk**

Nonparametric Regression For Locally Stationary Time Series

Michael Vogt*
University of Cambridge

July 27, 2012

Abstract

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. We complement the technical analysis of the paper by an application to financial data.

Key words: local stationarity, nonparametric regression, smooth backfitting.
AMS 2010 subject classifications: 62G08, 62G20, 62M10, 62P20.

1 Introduction

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

One way to model nonstationary behaviour is provided by the theory of locally stationary processes introduced by Dahlhaus (cf. [6], [7], and [8]). Intuitively speaking, a process is locally stationary if over short periods of time (i.e. locally in time) it behaves approximately stationary. 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.

*I would like to thank Enno Mammen, Oliver Linton and Suhasini Subba Rao for numerous helpful suggestions and comments. Support by the DFG-SNF research group FOR916 is gratefully acknowledged.

There is a considerable amount of papers that deal with time series models with time-varying coefficients. Dahlhaus et al. [9], for example, study wavelet estimation in autoregressive models with time-dependent parameters. Dahlhaus & Subba Rao [10] 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. [11]. Linton & Hafner [19] 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 & Linton [15].

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

$$Y_{t,T} = m\Big(\frac{t}{T}, X_{t,T}\Big) + \varepsilon_{t,T} \quad \text{for } t = 1, \ldots, T \tag{1}$$

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

$$Y_{t,T} = m\Big(\frac{t}{T}, Y_{t-1,T}, \ldots, Y_{t-d,T}\Big) + \varepsilon_{t,T} \quad \text{for } t = 1, \ldots, T \tag{2}$$

with $\mathbb{E}[\varepsilon_{t,T}|Y_{t-1,T}, \ldots, Y_{t-d,T}] = 0$, 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 to the present work, Kristensen [17] 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 [4], Masry [22], and Hansen [13]. The locally stationary behaviour 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

$$Y_{t,T} = \sum_{j=1}^{d} m_j\Big(\frac{t}{T}, X_{t,T}^j\Big) + \varepsilon_{t,T} \quad \text{for } t = 1, \ldots, T \tag{3}$$

with $X_{t,T} = (X_{t,T}^1, \ldots, X_{t,T}^d)$ 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. [20] to our setting.
To show the practical usefulness of our theory, we apply an additive volatility model with time-varying component functions to a sample of financial data in Section 6. The analysis makes visible how the component functions estimated at time points before and during the recent financial crisis differ from each other.

2 Local Stationarity

Heuristically speaking, a process $\{X_{t,T} : t = 1, \ldots, T\}_{T=1}^{\infty}$ 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 & Subba Rao [10]). 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

$$\big\|X_{t,T} - X_t(u)\big\| \le \Big(\Big|\frac{t}{T} - u\Big| + \frac{1}{T}\Big) U_{t,T}(u) \quad a.s.,$$

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 ρ-th 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

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

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 $Z_{t,T}^{t-k} := (Z_{t-k,T}, \ldots, Z_{t,T})$ 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

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

A tvNAR process is thus an autoregressive process of the 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 $u \leq 0$, $m(u,y) = m(0,y)$ and $\sigma(u,y) = \sigma(0,y)$. Analogously, we set $m(u,y) = m(1,y)$ and $\sigma(u,y) = \sigma(1,y)$ for $u \geq 1$. 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 $t \leq 0$. Put differently, $Y_{t,T} = m\big(0, Y_{t-1,T}^{t-d}\big) + \sigma\big(0, Y_{t-1,T}^{t-d}\big)\varepsilon_t$ for all $t \leq 0$. We can thus assume that $Y_{t,T} = Y_t(0)$ for $t \leq 0$. 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 some related nonstationary models. Zhou & Wu [29] and Zhou [30] consider the framework $Z_{t,T} = G(\frac{t}{T}, \psi_t)$, where $\psi_t = (\ldots, \varepsilon_{t-1}, \varepsilon_t)$ 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 & Wu.

Another related model is analyzed in Karlsen & Tjøstheim [14]. They are concerned with estimating the conditional mean function $m(y) = \mathbb{E}[Y_t | Y_{t-1} = y]$, when the Markov process $\{Y_t\}$ is nonstationary. In particular, they consider the situation that $\{Y_t\}$ belongs to the class of null recurrent Markov chains. The concept of null recurrence is quite distinct from that of local stationarity. The latter restricts the behaviour of the process in the time domain: it requires the process to change its stochastic behaviour not too fast over time. Null recurrence in contrast restricts the behaviour in the state domain: it requires the time series to recur to each point in its range almost surely, i.e. the process is not allowed to wander off for good. This difference in concepts is reflected in the fact that the analysis of null recurrent time series requires techniques substantially different from those for the analysis of locally stationary processes.

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, i.e. 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 $K_1 < \infty$,

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

An exact formula for the bound δ is given in (35) 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, i.e. $0 < c_\sigma \leq \sigma(u, y) \leq C_\sigma < \infty$ 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, i.e. 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

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

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 [26], Bhattacharya & Lee [3], An & Huang [2] or Chen & Chen [5] 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 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 $|\partial_j m(u, y)| \le \Delta$ 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 & Subba Rao [12] 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*

$$|Y_{t,T} - Y_t(u)| \le \Big(\Big| \frac{t}{T} - u \Big| + \frac{1}{T} \Big) U_{t,T}(u) \quad a.s., \tag{6}$$

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

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

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 the 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*

$$|f(u, y) - f(v, y)| \leq C_y |u - v|^p$$

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

We finally characterize the mixing behaviour 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

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

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

$$\alpha(k) = \sup_{t,T: 1 \leq t \leq T-k} \alpha\big(\sigma(Z_{s,T}, 1 \leq s \leq t), \sigma(Z_{s,T}, t+k \leq s \leq T)\big) \tag{7}$$
$$\beta(k) = \sup_{t,T: 1 \leq t \leq T-k} \beta\big(\sigma(Z_{s,T}, 1 \leq s \leq t), \sigma(Z_{s,T}, t+k \leq s \leq T)\big), \tag{8}$$

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, i.e. 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 & Subba Rao [12] 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.

3.4 The Additive tvNAR Process

An interesting special case of the tvNAR process arises, when the functions m and σ split up into additive components. In this case, the process is defined as

$$Y_{t,T} = \sum_{j=1}^{d} m_j\Big(\frac{t}{T}, Y_{t-j,T}\Big) + \Big(\sum_{j=1}^{d} \sigma_j\Big(\frac{t}{T}, Y_{t-j,T}\Big)\Big)^{1/2} \varepsilon_t. \tag{9}$$

In this setting, we can replace the conditions (M1)–(M3) and (Σ1)–(Σ3) on the functions m and σ by analogous conditions on the additive component functions. Most importantly, (M3) (and analogously (Σ3)) can be replaced by

($\mathrm{M_{add}}$3) $m_1, \dots, m_d$ are continuously differentiable with respect to y. The partial derivatives $\partial m_j(u, y_j) := \frac{\partial}{\partial y_j} m_j(u, y_j)$ are such that for some $K_1 < \infty$,

$$\sup_{u \in \mathbb{R}, |y_j| > K_1} |\partial m_j(u, y_j)| \le \delta_{\mathrm{add}} < 1.$$

Here, δ_{add} is given by a similar expression as δ in (M3).

Inspecting the proofs of Theorems 3.1–3.4, it is straightforward to see that the theorems still hold true under these modified conditions.

4 Kernel Estimation

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

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

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 $t = 1, \ldots, T$. 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

$$\hat{m}(u,x) = \frac{\sum_{t=1}^{T} K_h\big(u - \frac{t}{T}\big) \prod_{j=1}^{d} K_h\big(x^j - X_{t,T}^j\big) Y_{t,T}}{\sum_{t=1}^{T} K_h\big(u - \frac{t}{T}\big) \prod_{j=1}^{d} K_h\big(x^j - X_{t,T}^j\big)}. \tag{10}$$

Here and in what follows, we write $X_{t,T} = (X_{t,T}^1, \ldots, X_{t,T}^d)$ and $x = (x^1, \ldots, x^d)$ for any vector $x \in \mathbb{R}^d$, i.e. 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 non-product kernels and different bandwidths.

The estimate defined in (10) 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 $\partial^2_{ij} m(u,x)$ for $i,j = 0,\dots,d$.

(C6) The kernel K is symmetric about zero, bounded and has compact support, i.e. $K(v) = 0$ for all $|v| > C_1$ with some $C_1 < \infty$. Furthermore, K is Lipschitz, i.e. $|K(v) - K(v')| \le L|v - v'|$ 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, i.e. 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 (10), we examine kernel averages of the general form

$$\hat{\psi}(u,x) = \frac{1}{Th^{d+1}} \sum_{t=1}^{T} K_h\Big(u - \frac{t}{T}\Big) \prod_{j=1}^{d} K_h(x^j - X^j_{t,T}) W_{t,T} \tag{11}$$

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 (10) can be written as functions of averages of the above form. The asymptotic behaviour 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 (11):

(K1) It holds that $\mathbb{E}|W_{t,T}|^s \le C$ 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 $\alpha(k) \le Ak^{-\beta}$ 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 $\sup_{t,T} \sup_{x \in S} \mathbb{E}\big[|W_{t,T}|^s \big| X_{t,T} = x\big] f_{X_{t,T}}(x) \leq C$. Moreover, there exists a natural number $l^* < \infty$ such that for all $l \geq l^*$, $\sup_{t,T} \sup_{x,x' \in S} \mathbb{E}\big[|W_{t,T}||W_{t+l,T}| \big| X_{t,T} = x, X_{t+l,T} = x'\big] f_{X_{t,T},X_{t+l,T}}(x, x') \leq C$.

The next theorem generalizes uniform convergence results of Hansen [13] for the strictly stationary case to our setting. See Kristensen [16] for related results.

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

$$\beta > \frac{2 + s(1 + (d+1))}{s - 2} \tag{12}$$

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

$$\frac{\phi_T \log T}{T^\theta h^{d+1}} = o(1) \tag{13}$$

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

$$\theta = \frac{\beta(1 - \frac{2}{s}) - \frac{2}{s} - 1 - (d+1)}{\beta + 3 - (d+1)}. \tag{14}$$

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

$$\sup_{u \in [0,1], x \in S} \big|\hat{\psi}(u, x) - \mathbb{E}\hat{\psi}(u, x)\big| = O_p\Big(\sqrt{\frac{\log T}{T h^{d+1}}}\Big). \tag{15}$$

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 (12) and (14), 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, i.e. if $\beta = \infty$. The restriction (13) on the bandwidth is thus a strengthening of the usual condition that $T h^{d+1} \to \infty$.

4.4 Uniform Convergence Rates for NW Estimates

The next theorem characterizes the uniform convergence behaviour 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 (12) and suppose that $\inf_{u \in [0,1], x \in S} f(u, x) > 0$. Moreover, assume that the bandwidth h satisfies*

$$\frac{\phi_T \log T}{T^\theta h^{d+1}} = o(1) \quad \textit{and} \quad \frac{1}{T^r h^{d+r}} = o(1) \tag{16}$$

with θ given in (14), $\phi_T = \log\log T$, $r = \min\{\rho, 1\}$ and ρ introduced in (C1). Defining $I_h = [C_1 h, 1 - C_1 h]$, it then holds that

$$\sup_{u \in I_h, x \in S} \big|\hat{m}(u,x) - m(u,x)\big| = O_p\Big(\sqrt{\frac{\log T}{Th^{d+1}}} + \frac{1}{T^r h^d} + h^2\Big). \tag{17}$$

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*

$$\sqrt{Th^{d+1}}\big(\hat{m}(u,x) - m(u,x)\big) \xrightarrow{d} N(B_{u,x}, V_{u,x}), \tag{18}$$

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

$$\mathbb{E}[Y_{t,T}|X_{t,T} = x] = m_0\Big(\frac{t}{T}\Big) + \sum_{j=1}^d m_j\Big(\frac{t}{T}, x^j\Big). \tag{19}$$

To identify the component functions of model (19) 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. [20] to our setting. To do so, we first introduce the auxiliary estimates

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

$\hat{p}(u,x)$ 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,

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

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

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

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 backfitting estimates.

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

$$\int \Big(\hat{m}(u,w) - g_0 - \sum_{j=1}^{d} g_j(w^j)\Big)^2 \hat{p}(u,w)dw, \tag{20}$$

where the minimization runs over all additive functions $g(x) = g_0 + g_1(x^1) + \ldots + g_d(x^d)$ whose components are normalized to satisfy $\int g_j(w^j)\hat{p}_j(u,w^j)dw^j = 0$ for $j = 1,\ldots,d$. 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 (20), 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 (20) 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 (20) is characterized by the system of integral equations

$$\tilde{m}_j(u,x^j) = \hat{m}_j(u,x^j) - \sum_{k\neq j}\int \tilde{m}_k(u,x^k)\frac{\hat{p}_{j,k}(u,x^j,x^k)}{\hat{p}_j(u,x^j)}dx^k - \tilde{m}_0(u) \tag{21}$$

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

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

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

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

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 (21). This is ensured by the following assumption.

(Add1) (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 condition that $\beta > \max\{4,\frac{2+3s}{s-2}\}$ 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 $\theta = \min\{\frac{\beta-4}{\beta}, \frac{\beta(1-\frac{2}{s})-\frac{2}{s}-3}{\beta+1}\}$, and (iii) $(T^r h)^{-1} = o(h^2)$ and $T^{-\frac{r}{r+1}} = o(h^2)$ with $r = \min\{\rho, 1\}$ and ρ given in (C1).

The 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 (12) and (14) from Theorem 4.1, we get a pair of expressions for β and θ for each of the kernel estimates occurring in (21). 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. Firstly, 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),*

$$\sup_{u,x^j \in I_h} \left|\tilde{m}_j(u,x^j) - m_j(u,x^j)\right| = O_p\Big(\sqrt{\frac{\log T}{Th^2}} + h^2\Big). \tag{22}$$

Secondly, 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 $u, x^1, \ldots, x^d \in (0,1)$,*

$$\sqrt{T_{[0,1]^d} h^2} \begin{bmatrix} \tilde{m}_1(u,x^1) - m_1(u,x^1) \\ \vdots \\ \tilde{m}_d(u,x^d) - m_d(u,x^d) \end{bmatrix} \xrightarrow{d} N(B_{u,x}, V_{u,x}). \tag{23}$$

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

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

where the minimization runs over all additive functions $b(x) = b_0 + b_1(x^1) + \ldots + b_d(x^d)$ with $\int b_j(x^j)p_j(u,x^j)dx^j = 0$ and the function β is given in Lemma C4 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 C4.

6 Application

To illustrate our estimation theory, we apply it to a sample of NASDAQ Composite index data from 01/01/2000 to 31/12/2011.[1] For each day, our sample contains the return and the so-called high-low range. The latter is defined as the difference between the highest and lowest logarithmic price of a day. The range is a measure of daily volatility and has a long history in finance. It has been employed for example in the studies of Rogers & Satchell [24], Yang & Zhang [28], Alizadeh et al. [1], and Martens & van Dijk [21].

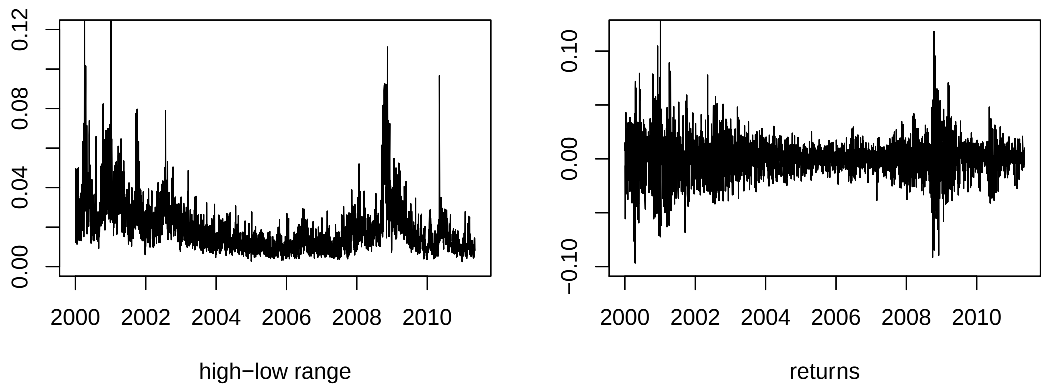

Figure 1: High-low range and returns of the NASDAQ Composite series.

In what follows, $y_{t,T}$ denotes the logarithm of the high-low range and $r_{t,T}$ is the daily return. With this notation, we define the model

$$y_{t,T} = m_0\Big(\frac{t}{T}\Big) + m_1\Big(\frac{t}{T}, y_{t-1,T}\Big) + m_2\Big(\frac{t}{T}, r_{t-1,T}\Big) + \varepsilon_{t,T}, \tag{24}$$

where $\mathbb{E}[\varepsilon_{t,T}|y_{t-1,T}, r_{t-1,T}] = 0$ and the functions m_1 and m_2 are normalized as described at the beginning of Section 5. Model (24) can be regarded as a localized version of the setting studied in Wu & Xiao [27].[2] The function m_1 specifies how today's volatility level depends on yesterday's level. The m_2-component is the news impact curve of the model. It captures how return shocks influence volatility.

We fit model (24) locally around three different time points in our sample, using an Epanechnikov kernel and choosing the bandwidth in time direction to span approximately one year and a half. As a result, we estimate the model for three different time periods, each spanning roughly three years. We include the period from 03/2000 to 03/2003 which corresponds to the aftermath of the technology bubble and the events of 9/11, the period from 11/2007 to 11/2010 which spans a great deal of the recent financial crisis, and an intermediate non-crisis period from 11/2003 to 11/2006.

[1]The data are available on the Yahoo Finance website.

[2]Wu & Xiao consider a model in which the component functions m_1 and m_2 do not depend on time and the first component m_1 is restricted to be linear. Moreover, implied volatility instead of the range is used as a daily volatility measure.

The estimation results are shown in Figure 2. The solid, dashed and dotted lines are the nonparametric fits for the three different periods and the grey shaded areas are 95% pointwise confidence intervals. The estimates are normalized as described in Section 5.

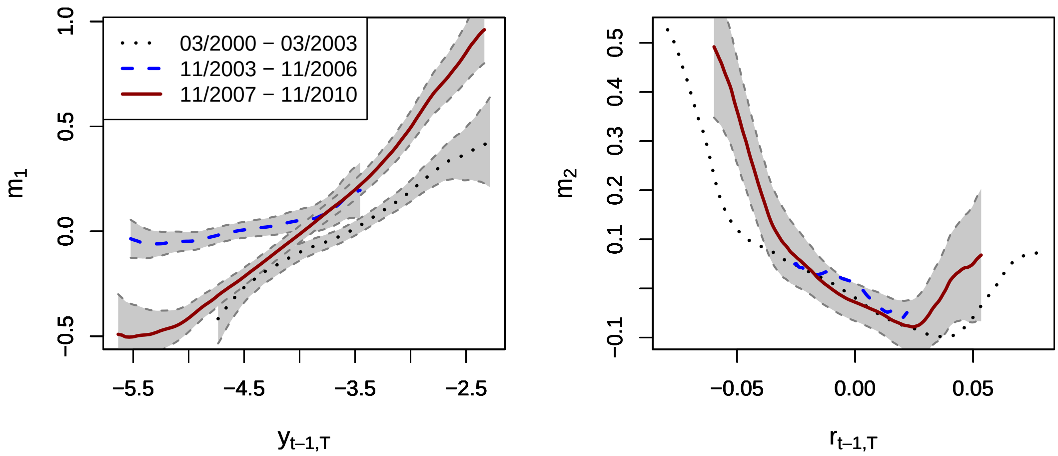

Figure 2: Estimation results for the additive model (24).

We have made several robustness checks. The first one concerns the choice of bandwidths. The bandwidth in time direction is handpicked rather than automatically selected. Given this, the bandwidths with respect to the two covariates are selected via a mean-squared error criterion. To check whether the estimation results are robust against different choices of bandwidth in time direction, we have gradually reduced the bandwidth to span only one year. This has virtually no effect on the fits. Moreover, we have smoothly varied the time points around which the model is estimated. As expected, this results in smooth changes of the nonparametric fits. In particular, shifting the time points only by a couple of months does not have major effects on the fits and preserves their qualitative form.

We now have a closer look at the estimation results in Figure 2. The estimates of m_1 are fairly linear. Interestingly, the fit for the financial crisis period (and presumably also the one for the period from 2000 to 2003) is much steeper than that for the intermediate non-crisis period from 2003 to 2006. This suggests that in more tense economic situations or crisis periods, today's volatility reacts more strongly to changes in yesterday's volatility. Put differently, the market is more sensitive to changes in volatility. The estimates of m_2 suggest that the overall form of the news impact curve is rather robust over time. Moreover, one can clearly see the asymmetric form of the curve which has been reported in numerous other studies before.

In the next step of our empirical analysis, we use the nonparametric fits of (24) as a guideline to set up a parametric model. We choose a specification with a linear m_1-component and a quadratic m_2-component that is flexible enough to allow for asymmetries. The model is given by

$$y_{t,T} = m_{0,\text{par}}\Big(\frac{t}{T}\Big) + m_{1,\text{par}}\Big(\frac{t}{T}, y_{t-1,T}\Big) + m_{2,\text{par}}\Big(\frac{t}{T}, r_{t-1,T}\Big) + \varepsilon_{t,T} \tag{25}$$

with

$$m_{1,\text{par}}\Big(\frac{t}{T}, y_{t-1,T}\Big) = a_1\Big(\frac{t}{T}\Big) y_{t-1,T}$$
$$m_{2,\text{par}}\Big(\frac{t}{T}, r_{t-1,T}\Big) = a_2\Big(\frac{t}{T}\Big) r^2_{t-1,T} I(r_{t-1,T} < 0) + a_3\Big(\frac{t}{T}\Big) r^2_{t-1,T} I(r_{t-1,T} \geq 0),$$

where a_1, a_2 and a_3 are time-varying parameters. We estimate (25) locally around the same time points as the additive model (24) using the same bandwidth in time direction. The estimation is done by minimizing a least-squares criterion localized in time. Rather than reporting the estimates of the time-varying parameters in a table, we plot the fits of $m_{1,\text{par}}$ and $m_{2,\text{par}}$ in Figure 3.

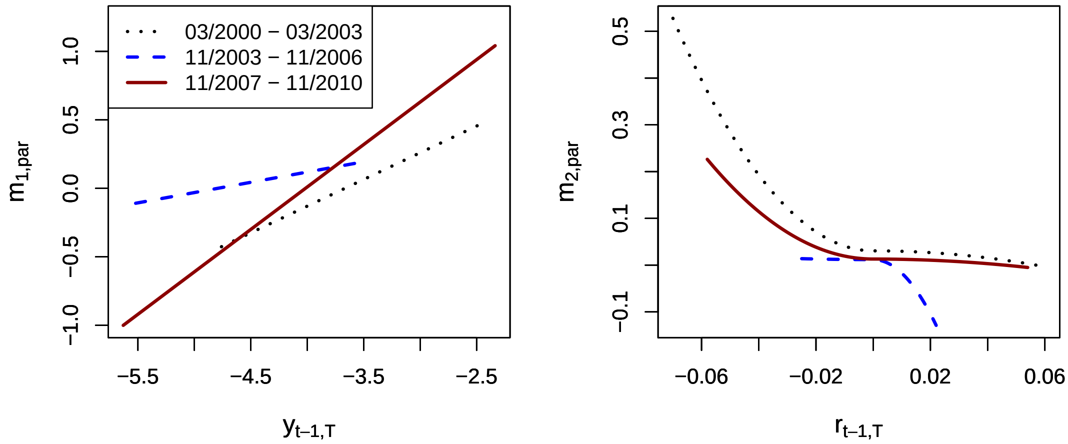

Figure 3: Estimation results for the parametric model (25).

The fits of $m_{1,\text{par}}$ give a very similar picture as their nonparametric counterparts. The estimates of $m_{2,\text{par}}$, however, do not. In particular, they suggest that the news impact curve in the intermediate non-crisis period from 2003 to 2006 substantially differs from the curves in the two crisis periods. Figure 4 makes visible the differences between the parametric and nonparametric fits of the news impact curve.

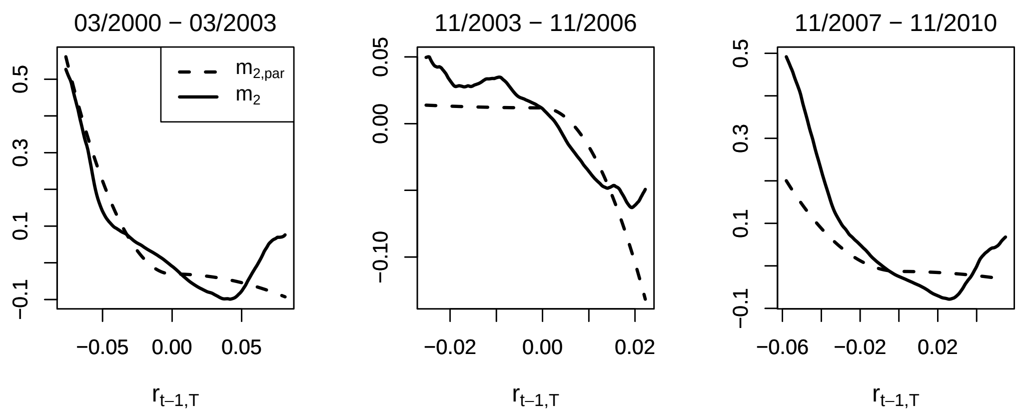

Figure 4: Comparison of the parametric function $m_{2,\text{par}}$ (dashed) and its nonparametric counterpart m_2 (solid).

As can be seen from Figure 4, the parametric estimates roughly capture the overall form of their nonparametric counterparts. However, they are not flexible enough to reproduce all important characteristics. In particular, the parametric estimate for the intermediate non-crisis period strongly exaggerates the slightly concave form of the corresponding nonparametric fit. This gives the impression that the news impact curve in the non-crisis period drastically differs from that in the two crisis periods.

The above considerations make visible an important shortcoming of the parametric analysis: If the parametric model is not flexible enough, then the fits may spuriously generate time-varying effects. Thus, the news impact curve may after all be much more robust over time than suggested by many parametric specifications.

7 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

$$m\big(v, \underline{Y}_{t-1}(v)\big) - m\big(u, \underline{Y}_{t-1}(u)\big) = \Delta^m_{t,0} + \sum_{j=1}^{d} \Delta^m_{t,j} \big(Y_{t-j}(v) - Y_{t-j}(u)\big), \tag{26}$$

where we have used the shorthands $\Delta^m_{t,0} = m(v, \underline{Y}_{t-1}(v)) - m(u, \underline{Y}_{t-1}(v))$ and $\Delta^m_{t,j} = \Delta^m_j(u, \underline{Y}_{t-1}(u), \underline{Y}_{t-1}(v))$ for $j = 1, \ldots, d$ with the functions $\Delta^m_j(u, y, y') = \int_0^1 \partial_j m(u, y + s(y' - y))ds$.
The terms $\Delta^m_{t,j}$ have the property that

$$\big|\Delta^m_{t,j}\big| \le \Delta_t := \Delta I(\|\underline{\varepsilon}_{t-1}\|_\infty \le K_2) + \delta I(\|\underline{\varepsilon}_{t-1}\|_\infty > K_2) \tag{27}$$

for $j = 1, \ldots, d$ with $K_2 = (K_1 + C_m)/c_\sigma$ and $\Delta \ge \sup_{u,y} |\partial_j m(u, y)|$. This can be seen as follows: Using the shorthands $m_{u,k} = m(u, \underline{Y}_{t-k-1}(u))$ and $\sigma_{u,k} = \sigma(u, \underline{Y}_{t-k-1}(u))$, we obtain

$$\begin{aligned}
&\|\underline{Y}_{t-1}(u) + s(\underline{Y}_{t-1}(v) - \underline{Y}_{t-1}(u))\|_\infty \\
&= \max_{k=1,\ldots,d} |Y_{t-k}(u) + s(Y_{t-k}(v) - Y_{t-k}(u))| \\
&= \max_{k=1,\ldots,d} |m_{u,k} + s(m_{v,k} - m_{u,k}) + \varepsilon_{t-k}(\sigma_{u,k} + s(\sigma_{v,k} - \sigma_{u,k}))| \\
&\ge c_\sigma \|\underline{\varepsilon}_{t-1}\|_\infty - C_m,
\end{aligned} \tag{28}$$

since $|m_{u,k} + s(m_{v,k} - m_{u,k})| \le C_m$ and $|\sigma_{u,k} + s(\sigma_{v,k} - \sigma_{u,k})| \ge c_\sigma > 0$. Now assume that $\|\underline{\varepsilon}_{t-1}\|_\infty > K_2$. In this case, (28) implies that $\|\underline{Y}_{t-1}(u) + s(\underline{Y}_{t-1}(v) - \underline{Y}_{t-1}(u))\|_\infty > K_1$ for all $s \in [0, 1]$. Hence, the region over which the integral in $\Delta^m_{t,j}$ runs completely lies outside the area $[-K_1, K_1]^d$. Therefore, the integrand $\partial_j m$ is always smaller than δ in absolute value, which immediately implies that $|\Delta^m_{t,j}| \le \delta$. Next let $\|\underline{\varepsilon}_{t-1}\|_\infty \le K_2$. As $\sup_{u,y} |\partial_j m(u, y)| \le \Delta < \infty$, the term $|\Delta^m_{t,j}|$ is always bounded by Δ, in particular for $\|\underline{\varepsilon}_{t-1}\|_\infty \le K_2$.
Repeating the above considerations for the function σ, we obtain analogous terms $\Delta^\sigma_{t,j}$ that are again bounded by Δ_t for $j = 1, \ldots, d$.

Recursive formulas for $Y_{t,T}$

For the proof of Theorem 3.4, we rewrite $Y_{t,T}$ in a recursive fashion: Letting $y^{t-k_2}_{t-k_1}$ and $e^{t-k_2}_{t-k_1}$ be values of $Y^{t-k_2}_{t-k_1}$ and $\varepsilon^{t-k_2}_{t-k_1}$, respectively, we recursively define the functions $m^{(i)}_{t,T}$

by $m^{(0)}_{t,T}(y^{t-d}_{t-1}) = m(\frac{t}{T}, y^{t-d}_{t-1})$ and for $i \geq 1$ by

$$m^{(i)}_{t,T}\big(e^{t-i}_{t-1}, y^{t-i-d}_{t-i-1}\big) = m^{(i-1)}_{t,T}\big(e^{t-i+1}_{t-1}, m^{(0)}_{t-i,T}(y^{t-i-d}_{t-i-1}) + \sigma^{(0)}_{t-i,T}(y^{t-i-d}_{t-i-1})e_{t-i}, y^{t-i-d+1}_{t-i-1}\big).$$

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

$$Y_{t,T} = m^{(i)}_{t,T}\big(\varepsilon^{t-i}_{t-1}, Y^{t-i-d}_{t-i-1,T}\big) + \sigma^{(i)}_{t,T}\big(\varepsilon^{t-i}_{t-1}, Y^{t-i-d}_{t-i-1,T}\big)\varepsilon_t.$$

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

$$m^{(i)}_{t,T}\big(e^{t-i}_{t-1}, y^{t-i-d}_{t-i-1}\big) = m\Big(\frac{t}{T}, m^{(i-1)}_{t-1,T}(e^{t-i}_{t-2}, y^{t-i-d}_{t-i-1}) + \sigma^{(i-1)}_{t-1,T}(e^{t-i}_{t-2}, y^{t-i-d}_{t-i-1})e_{t-1}, \ldots \\ \ldots, m^{(i-d)}_{t-d,T}(e^{t-i}_{t-d-1}, y^{t-i-d}_{t-i-1}) + \sigma^{(i-d)}_{t-d,T}(e^{t-i}_{t-d-1}, y^{t-i-d}_{t-i-1})e_{t-d}\Big).$$

The term $\sigma^{(i)}_{t,T}(e^{t-i}_{t-1}, y^{t-i-d}_{t-i-1})$ 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 $1 \leq r \leq d$,

$$f_{Y_{t,T}|Y^{t-r+1}_{t-1,T}, \varepsilon^{-s}_{t-r}, Y^{-s-d}_{-s-1,T}}(y_t|y^{t-r+1}_{t-1}, e^{-s}_{t-r}, z) = \frac{1}{\sigma_{t,T}} f_\varepsilon\Big(\frac{y_t - m_{t,T}}{\sigma_{t,T}}\Big). \tag{29}$$

Here, y_t, y^{t-r+1}_{t-1}, e^{-s}_{t-r}, and z are values of $Y_{t,T}$, $Y^{t-r+1}_{t-1,T}$, ε^{-s}_{t-r}, and $Y^{-s-d}_{-s-1,T}$, respectively. Moreover,

$$m_{t,T} = m\Big(\frac{t}{T}, y^{t-r+1}_{t-1}, m^{(t-r+s)}_{t-r,T}(e^{-s}_{t-r-1}, z) + \sigma^{(t-r+s)}_{t-r,T}(e^{-s}_{t-r-1}, z)e_{t-r}, \ldots \\ \ldots, m^{(t-d+s)}_{t-d,T}(e^{-s}_{t-d-1}, z) + \sigma^{(t-d+s)}_{t-d,T}(e^{-s}_{t-d-1}, z)e_{t-d}\Big)$$

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

Proof of Theorem 3.1

(i) follows by standard arguments to be found for example in Chen & Chen [5]. (ii) immediately follows with the help of (29). For (iii), recall that $Y^{t-d}_{t-1,T} = Y^{t-d}_{t-1}(0)$ for $t \leq 1$. This allows us to write the density of $Y^{t-d}_{t-1,T}$ as

$$f_{Y^{t-d}_{t-1,T}}(y) = \int f_{Y^{t-d}_{t-1,T}|\varepsilon^1_{t-d-1}, Y^{-d+1}_0(0)}(y|e,z) \prod_{i=1}^{t-d-1} f_\varepsilon(e_i) f_{Y^{-d+1}_0(0)}(z)\, dedz,$$

where $e = e^1_{t-d-1}$ and the conditional density $f_{Y^{t-d}_{t-1,T}|\varepsilon^1_{t-d-1}, Y^{-d+1}_0(0)}$ can be expressed in terms of the error density f_ε with the help of (29). $\square$

Proof of Theorem 3.2

We apply the triangle inequality to get

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

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

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

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,

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

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 (26). The terms $\Delta_{t,j}^\sigma$ for $j = 0, \dots, d$ are defined analogously. In matrix notation, we obtain

$$\underline{Y}_t\Big(\frac{t}{T}\Big) - \underline{Y}_t(u) = A_t\Big(\underline{Y}_{t-1}\Big(\frac{t}{T}\Big) - \underline{Y}_{t-1}(u)\Big) + \underline{\xi}_t \tag{30}$$

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

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

Iterating (30) n times yields

$$\Big\|\underline{Y}_t\Big(\frac{t}{T}\Big) - \underline{Y}_t(u)\Big\| \le \|\underline{\xi}_t\| + \Big\|\sum_{r=0}^{n-1} \prod_{k=0}^{r} A_{t-k} \underline{\xi}_{t-r-1}\Big\| + \Big\|\prod_{k=0}^{n} A_{t-k}\Big(\underline{Y}_{t-n-1}\Big(\frac{t}{T}\Big) - \underline{Y}_{t-n-1}(u)\Big)\Big\|.$$

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

$$B_t = (1 + |\varepsilon_t|)B(\Delta_t) \tag{31}$$

with $\Delta_t = \Delta I(\|\underline{\varepsilon}_{t-1}\|_\infty \leq K_2) + \delta I(\|\underline{\varepsilon}_{t-1}\|_\infty > K_2)$. As shown in the preliminaries section of the appendix, $|\Delta^m_{t,j} + \Delta^\sigma_{t,j}\varepsilon_t| \leq \Delta_t(1 + |\varepsilon_t|)$ for $j = 1, \ldots, d$. Therefore, the entries of the matrix B_t are all weakly larger in absolute value than those of A_t. This implies that $\| \prod_{k=0}^n A_{t-k} z\| \leq \| \prod_{k=0}^n B_{t-k} |z|\|$ with $z = (|z_1|, \ldots, |z_d|)$. Using this together with the boundedness of m and σ and the fact that $|\Delta^m_{t,0} + \Delta^\sigma_{t,0}\varepsilon_t| \leq C|\frac{t}{T} - u|(1 + |\varepsilon_t|)$, we finally arrive at

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

with

$$\begin{aligned} V_{t,n} &= C(1 + |\varepsilon_t|) + C \sum_{r=0}^{n-1} \big(1 + |\varepsilon_{t-r-1}|\big) \Big\| \prod_{k=0}^{r} B_{t-k} \Big\| \\ R_{t,n} &= C(1 + \|\underline{\varepsilon}_{t-n-1}\|) \Big\| \prod_{k=0}^{n} B_{t-k} \Big\|. \end{aligned}$$

Bounding $V_{t,n}$ and $R_{t,n}$. The convergence behaviour 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 behaviour of the latter is described by the following lemma.

Lemma A1. *If δ is sufficiently small, in particular if it satisfies (35), then there exists a constant $\rho > 0$ such that for some $\gamma < 1$,*

$$\mathbb{E}\Big[\Big\| \prod_{k=0}^{n} B_{t-k} \Big\|^\rho\Big] \leq C\gamma^n. \tag{32}$$

The proof of Lemma A1 is postponed until the arguments for Theorem 3.2 are completed. The following statement is a direct consequence of Lemma A1.

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

In addition, it holds that

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

This can be seen as follows. First note that

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

Using the monotone convergence theorem and Loève's inequality with $\rho < 1$, we obtain $\mathbb{E}[V_t^\rho] \leq C\mathbb{E}(1+|\varepsilon_t|)^\rho + \sum_{r=0}^\infty \mathbb{E}[R_{t,r}^\rho]$. As the right-hand side of the previous inequality is finite by (R), we arrive at (V).
(R) and (V) imply that $|Y_t(\frac{t}{T}) - Y_t(u)| \leq |\frac{t}{T} - u|V_t$ a.s. with variables V_t whose ρ-th 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. $\square$

Proof of Lemma A1. We want to show that the ρ-th 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. [23]) or they heavily draw on the independence of the random matrices involved (see e.g. Proposition 2.1 in Subba Rao [25]).
We now describe our proving strategy in detail. To start with, we replace the spectral norm $\|\cdot\|$ in (32) 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 $\| \prod_{k=0}^n B_{t-k}\| \leq C\mathcal{B}_n$ 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 $J_k = I(\min_{l=1,\ldots,d} |\varepsilon_{t-k-l}| \leq K_2)$ and a constant $0 < \kappa < 1$ to be specified later on. Lemma A1 is a direct consequence of the following two facts:

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

(ii) $\mathbb{E}[\mathcal{B}_{n,2}] \leq C\gamma^n$ 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

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

It is easy to see that $\mathbb{E}[\mathcal{B}_{n,1}^{2\rho}] \leq C^{\rho n}$ 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

$$\mathbb{P}\big(\mathcal{B}_{n,1} > \phi_n\big) \leq C\gamma^n \quad \text{for some } \gamma < 1. \tag{33}$$

For the proof of (33), we write

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

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

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

for some $\gamma < 1$. This yields (33) and thus completes the proof of (i).
Let us now turn to the proof of (ii). We have that

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

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 $\|\underline{\varepsilon}_{t-k-1}\|_\infty \leq K_2$, 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 $I_n = 0$, i.e. 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

$$\mathbb{E}\big[\mathcal{B}_{n,2}\big] \leq \mathbb{E}\Big[\prod_{k=0}^{n}(1 + |\varepsilon_{t-k}|)\Big]\big\|B(\Delta)\big\|_1^{\kappa n}\big\|B(\delta)^d\big\|_1^{\frac{(1-\kappa)n}{d}}. \tag{34}$$

By direct calculations, we can verify that $\|B(\delta)^d\|_1 \leq C_d\delta$ 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, $\|B(\Delta)\|_1 \leq (\Delta + 1)$. Plugging this into (34) yields

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

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

$$\delta < \Big[(1 + \mathbb{E}|\varepsilon_0|)^{\frac{d}{1-\kappa}}(\Delta + 1)^{\frac{\kappa d}{1-\kappa}}C_d\Big]^{-1}. \tag{35}$$

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

Proof of Theorem 3.3

Throughout the proof, we use the following notation: y_j and z_j are values of the variables $Y_{t-j}(u)$ and $Y_{t-d-j}(u)$ for $j = 1, \dots, d$. Moreover, we write $y = (y_1, \dots, y_d)$ together with $z = (z_1, \dots, z_d)$ and define

F_u : distribution function of $Y_{t-d-1}^{t-2d}(u)$
$F_{u,v}$: joint distribution function of $Y_{t-d-1}^{t-2d}(u)$ and $Y_{t-d-1}^{t-2d}(v)$
$f_u(y)$: density of $Y_{t-1}^{t-d}(u)$ at y
$f_u(y|z)$: density of $Y_{t-1}^{t-d}(u)$ at y conditional on $Y_{t-d-1}^{t-2d}(u) = z$.

In addition, we let $f_{u,j} = f_u(y_j|y_{j+1},\dots,y_d,z_1,\dots,z_j)$ denote the conditional density of $Y_{t-j}(u)$ given $Y_{t-j-1}^{t-j-d}(u)$. Note that $f_u(y|z) = \prod_{j=1}^d f_{u,j}$ and that the conditional densities $f_{u,j}$ can be expressed in terms of the error density as $f_{u,j} = \frac{1}{\sigma_{u,j}} f_\varepsilon(\frac{y_j - m_{u,j}}{\sigma_{u,j}})$, where $m_{u,j} = m(u,(y_{j+1},\dots,y_d,z_1,\dots,z_j))$ and $\sigma_{u,j} = \sigma(u,(y_{j+1},\dots,y_d,z_1,\dots,z_j))$.
With this notation at hand, we can now analyze the term $|f_u(y) - f_v(y)|$. Letting $z' = (z'_1,\dots,z'_d)$ be some value taken by the random vector $Y_{t-d-1}^{t-2d}(v)$, we can apply a telescoping argument to obtain

$$\begin{aligned}\big|f_u(y) - f_v(y)\big| &= \Big|\int_{\mathbb{R}^{2d}} \big[f_u(y|z) - f_v(y|z')\big] dF_{u,v}(z,z')\Big| \\ &\le \sum_{k=1}^d \int_{\mathbb{R}^{2d}} \big|f_{u,k} - f_{v,k}\big| dF_{u,v}(z,z') =: \sum_{k=1}^d Q_{u,v}^{[k]}(y).\end{aligned}$$

Using the boundedness of m, σ, and f_ε yields

$$\begin{aligned}Q_{u,v}^{[k]}(y) &= \int_{\mathbb{R}^{2d}} \Big|\frac{1}{\sigma_{u,k}} f_\varepsilon\Big(\frac{y_k - m_{u,k}}{\sigma_{u,k}}\Big) - \frac{1}{\sigma_{v,k}} f_\varepsilon\Big(\frac{y_k - m_{v,k}}{\sigma_{v,k}}\Big)\Big| dF_{u,v}(z,z') \\ &\le C \int_{\mathbb{R}^{2d}} \Big|f_\varepsilon\Big(\frac{y_k - m_{u,k}}{\sigma_{u,k}}\Big) - f_\varepsilon\Big(\frac{y_k - m_{v,k}}{\sigma_{v,k}}\Big)\Big| dF_{u,v}(z,z') \\ &\quad + C \int_{\mathbb{R}^{2d}} \big|\sigma_{u,k} - \sigma_{v,k}\big| dF_{u,v}(z,z') \\ &=: Q_{u,v}^{[k,1]}(y) + Q_{u,v}^{[k,2]}(y).\end{aligned}$$

Exploiting the Lipschitz continuity of f_ε together with the smoothness conditions on m and σ, we further obtain

$$\begin{aligned}Q_{u,v}^{[k,1]}(y) &\le C(1+|y_k|) \int \big(|u-v| + |z_1 - z'_1| + \ldots + |z_k - z'_k|\big) dF_{u,v}(z,z') \\ &= C(1+|y_k|)\Big(|u-v| + \sum_{j=1}^k \mathbb{E}\big|Y_{t-d-j}(u) - Y_{t-d-j}(v)\big|\Big)\end{aligned}$$

together with an analogous expression for $Q_{u,v}^{[k,2]}(y)$. As an intermediate result, we have thus shown that

$$\big|f_u(y) - f_v(y)\big| \le C(1+\|y\|_1)\Big(|u-v| + \mathbb{E}\big|Y_t(u) - Y_t(v)\big|\Big), \tag{36}$$

where $C < \infty$ is some sufficiently large constant and $\|\cdot\|_1$ denotes the usual l_1-norm for $\mathbb{R}^d$-valued vectors.
In the remainder of the proof, we derive a bound for the expression $\mathbb{E}|Y_t(u) - Y_t(v)|$. By the proof of Theorem 3.2, it holds that $|Y_t(u) - Y_t(v)| \le |u-v| U_t$ with random variables U_t having the property that $\mathbb{E}[U_t^\rho] \le C$ for some $\rho > 0$. Letting q be a constant with $0 < q < \rho$, we arrive at

$$\begin{aligned}\mathbb{E}\big|Y_t(u) - Y_t(v)\big| &= \mathbb{E}\Big[\big|Y_t(u) - Y_t(v)\big| I\Big(U_t \le \frac{C}{|u-v|^q}\Big)\Big] \\ &\quad + \mathbb{E}\Big[\big|Y_t(u) - Y_t(v)\big| I\Big(U_t > \frac{C}{|u-v|^q}\Big)\Big] \\ &=: E_1(u,v) + E_2(u,v).\end{aligned}$$

It is straightforward to show that $E_1(u,v) \leq C|u-v|^{1-q}$ and $E_2(u,v) \leq C|u-v|^r$ for some $r > 0$. This completes the proof. $\square$

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

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

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

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

for $j = 0, \ldots, d-1$, where we suppress the dependence on the arguments y_{t+j-1}^t and e in the notation. Finally, note that by (29), the above conditional density can be expressed in terms of the error density f_ε as

$$f_j(y_{t+j}|z) = \frac{1}{\sigma_{t,T,j}(z)} f_\varepsilon\Big(\frac{y_{t+j} - m_{t,T,j}(z)}{\sigma_{t,T,j}(z)}\Big) \tag{37}$$

with

$$\begin{aligned} m_{t,T,j}(z) = m\Big(&\frac{t+j}{T}, y_{t+j-1}^t, m_{t-1,T}^{(k-2)}(e_{t-2}^{t-k+1}, z) + \sigma_{t-1,T}^{(k-2)}(e_{t-2}^{t-k+1}, z)e_{t-1}, \ldots \\ &\ldots, m_{t+j-d,T}^{(k-j+d-1)}(e_{t+j-d-1}^{t-k+1}, z) + \sigma_{t+j-d,T}^{(k-j+d-1)}(e_{t+j-d-1}^{t-k+1}, z)e_{t+j-d}\Big) \end{aligned}$$

and $\sigma_{t,T,j}(z)$ defined analogously. The functions $m_{t-1,T}^{(k-2)}$, $\sigma_{t-1,T}^{(k-2)}, \ldots$ were introduced in the preliminaries section of the appendix.
With this notation at hand, we can write

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

and likewise $\mathbb{P}(S) = \int I(y \in S) \prod_{j=0}^{d-1} f_j(y_{t+j}|z) \prod_{l=1}^{k-1} f_\varepsilon(e_{t-l}) f_{\underline{Y}_{t-k,T}}(z) dedzdy$. Using the shorthand $\underline{Y} = \underline{Y}_{t-k,T}$, we thus arrive at

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

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

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

where the last inequality exploits the fact that $\int \prod_{j=i+1}^{d-1} f_j(y_{t+j}|z) dy_{t+d-1} \dots dy_{t+i+1}$ is a conditional probability and thus almost surely bounded by one. Using the formula (37) together with (E3), it is straightforward to see that

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

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

$$\begin{aligned}
\big| m_{t,T,i}(z) - &m_{t,T,i}(\underline{Y}) \big| + \big| \sigma_{t,T,i}(z) - \sigma_{t,T,i}(\underline{Y}) \big| \\
&\le C \sum_{j=1}^{d-i} \Big\| \prod_{m=0}^{n} B_{t-j-m} \Big\| \big(1 + \| e_{t-j-n-1}^{t-j-n-d} \| \big), \qquad (38)
\end{aligned}$$

where $\|\cdot\|$ denotes the Euclidean norm for vectors and the spectral norm for matrices. The matrix B_t was introduced in (31). 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 (38) into the bound for $(**)$ and insert this into the bound for $(*)$ to arrive at

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

As a consequence,

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

Using the arguments from Lemma A1, 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

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

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

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 B1. *Suppose the kernel K satisfies (C6) and let $I_h = [C_1 h, 1 - C_1 h]$. Then for $k = 0, 1, 2$,*

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

Lemma B2. *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 $S \subset \mathbb{R}^d$,*

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

Proof of Theorem 4.1

The proof extends Theorem 2 of Hansen [13]. Define $B = \{(u,x) \in \mathbb{R}^{d+1} : u \in [0,1], x \in S\}$ and $\tau_T = \rho_T T^{\frac{1}{s}}$ with ρ_T slowly diverging to infinity as $T \to \infty$. To simplify the calculations in later parts of the proof, we choose

$$\rho_T = (\log T)^{\frac{1}{1+\beta}} \phi_T^{(1+\frac{\beta-d}{2})\frac{1}{1+\beta}}$$

with $\phi_T = \log\log T$. With this notation at hand, we write

$$\hat{\psi}(u,x) - \mathbb{E}[\hat{\psi}(u,x)] = \big(\hat{\psi}_1(u,x) - \mathbb{E}[\hat{\psi}_1(u,x)]\big) + \big(\hat{\psi}_2(u,x) - \mathbb{E}[\hat{\psi}_2(u,x)]\big), \qquad (39)$$

where

$$\hat{\psi}_1(u,x) = \frac{1}{Th^{d+1}} \sum_{t=1}^{T} K_h\Big(u - \frac{t}{T}\Big) \prod_{j=1}^{d} K_h(x^j - X_{t,T}^j) W_{t,T} I(|W_{t,T}| \le \tau_T)$$

$$\hat{\psi}_2(u,x) = \frac{1}{Th^{d+1}} \sum_{t=1}^{T} K_h\Big(u - \frac{t}{T}\Big) \prod_{j=1}^{d} K_h(x^j - X_{t,T}^j) W_{t,T} I(|W_{t,T}| > \tau_T).$$

In what follows, we compute the uniform convergence rates of the two terms on the right-hand side of (39). We proceed in several steps.

Step 1. First consider the term $\hat{\psi}_2(u,x) - \mathbb{E}[\hat{\psi}_2(u,x)]$. Defining $a_T = \sqrt{\log T / Th^{d+1}}$, it holds that

$$\mathbb{P}\Big(\sup_{(u,x)\in B} |\hat{\psi}_2(u,x)| > C a_T\Big) \le P\big(|W_{t,T}| > \tau_T \text{ for some } 1 \le t \le T\big)$$
$$\le \tau_T^{-s} \sum_{t=1}^{T} \mathbb{E}|W_{t,T}|^s \le C T \tau_T^{-s} = \rho_T^{-s} \to 0.$$

Using (K3), we additionally obtain that

$$\begin{aligned}
\mathbb{E}\big|\hat{\psi}_2(u,x)\big| &\le \frac{1}{Th^{d+1}} \sum_{t=1}^{T} K_h\Big(u - \frac{t}{T}\Big) \int_{\mathbb{R}^d} \prod_{j=1}^{d} K_h(x^j - w^j) \\
&\qquad \times \mathbb{E}\big[|W_{t,T}| I(|W_{t,T}| > \tau_T) \big| X_{t,T} = w\big] f_{X_{t,T}}(w) dw \\
&= \frac{1}{Th} \sum_{t=1}^{T} K_h\Big(u - \frac{t}{T}\Big) \int_{\mathbb{R}^d} \prod_{j=1}^{d} K(\varphi^j) \\
&\qquad \times \mathbb{E}\big[|W_{t,T}| I(|W_{t,T}| > \tau_T) \big| X_{t,T} = x - h\varphi\big] f_{X_{t,T}}(x - h\varphi) d\varphi \\
&\le \frac{1}{Th} \sum_{t=1}^{T} K_h\Big(u - \frac{t}{T}\Big) \frac{1}{\tau_T^{s-1}} \int_{\mathbb{R}^d} \prod_{j=1}^{d} K(\varphi^j) \\
&\qquad \times \mathbb{E}\big[|W_{t,T}|^s I(|W_{t,T}| > \tau_T) \big| X_{t,T} = x - h\varphi\big] f_{X_{t,T}}(x - h\varphi) d\varphi \\
&\le \frac{C}{\tau_T^{s-1}} \underbrace{\frac{1}{Th} \sum_{t=1}^{T} K_h\Big(u - \frac{t}{T}\Big)}_{\le C \text{ uniformly in u}} \le \frac{C}{\tau_T^{s-1}} = C \rho_T^{-(s-1)} T^{-\frac{s-1}{s}} \le C a_T.
\end{aligned}$$

As a result,

$$\sup_{(u,x)\in B} \big|\hat{\psi}_2(u,x) - \mathbb{E}\hat{\psi}_2(u,x)\big| = O_p(a_T).$$

Note that Hansen [13] uses the more slowly diverging truncation sequence $\tau_T = a_T^{-1/(s-1)}$. He shows that with this choice of τ_T, $|\hat{\psi}_2(u,x) - \mathbb{E}\hat{\psi}_2(u,x)| = O_p(a_T)$. It is however not clear at all whether $\sup_{u,x} |\hat{\psi}_2(u,x) - \mathbb{E}\hat{\psi}_2(u,x)| = O_p(a_T)$ in his case, which is needed for the proof. To be on the safe side, we work with the sequence $\tau_T = \rho_T T^{1/s}$.

Step 2. We now turn to the analysis of $\hat{\psi}_1(u,x) - \mathbb{E}[\hat{\psi}_1(u,x)]$. Cover the region B with $N \le C h^{-(d+1)} a_T^{-(d+1)}$ balls $B_n = \{(u,x) \in \mathbb{R}^{d+1} : \|(u,x) - (u_n,x_n)\|_\infty \le a_T h\}$ and use (u_n, x_n) to denote the midpoint of B_n. In addition, let $K^*(v) = C \prod_{j=0}^d I(|v^j| \le 2C_1)$ for $v \in \mathbb{R}^{d+1}$ and note that for $(u,x) \in B_n$ and T sufficiently large,

$$\Big| K_h\Big(u - \frac{t}{T}\Big) \prod_{j=1}^d K_h(x^j - X_{t,T}^j) - K_h\Big(u_n - \frac{t}{T}\Big) \prod_{j=1}^d K_h(x_n^j - X_{t,T}^j) \Big| \le a_T K_h^*\Big(u_n - \frac{t}{T}, x_n - X_{t,T}\Big)$$

with $K_h^*(v) = K^*(\frac{v}{h})$. Defining

$$\tilde{\psi}_1(u,x) = \frac{1}{T h^{d+1}} \sum_{t=1}^T K_h^*\Big(u - \frac{t}{T}, x - X_{t,T}\Big) |W_{t,T}| I(|W_{t,T}| \le \tau_T)$$

and noticing that $\mathbb{E}|\tilde{\psi}_1(u,x)| \le M < \infty$ for some sufficiently large M, we obtain

$$\begin{aligned}
&\sup_{(u,x)\in B_n} \big|\hat{\psi}_1(u,x) - \mathbb{E}\hat{\psi}_1(u,x)\big| \\
&\quad \le \big|\hat{\psi}_1(u_n,x_n) - \mathbb{E}\hat{\psi}_1(u_n,x_n)\big| + a_T\big(\big|\tilde{\psi}_1(u_n,x_n)\big| + \mathbb{E}\big|\tilde{\psi}_1(u_n,x_n)\big|\big) \\
&\quad \le \big|\hat{\psi}_1(u_n,x_n) - \mathbb{E}\hat{\psi}_1(u_n,x_n)\big| + \big|\tilde{\psi}_1(u_n,x_n) - \mathbb{E}\tilde{\psi}_1(u_n,x_n)\big| + 2M a_T.
\end{aligned}$$

As a consequence,

$$\begin{aligned}
&\mathbb{P}\Big(\sup_{(u,x)\in B} \big|\hat{\psi}_1(u,x) - \mathbb{E}\hat{\psi}_1(u,x)\big| > 4M a_T\Big) \\
&\quad \le N \max_{1\le n\le N} \mathbb{P}\Big(\sup_{(u,x)\in B_n} \big|\hat{\psi}_1(u,x) - \mathbb{E}\hat{\psi}_1(u,x)\big| > 4M a_T\Big) \le \hat{Q}_T + \tilde{Q}_T
\end{aligned}$$

with

$$\begin{aligned}
\hat{Q}_T &= N \max_{1\le n\le N} \mathbb{P}\Big(\big|\hat{\psi}_1(u_n,x_n) - \mathbb{E}\hat{\psi}_1(u_n,x_n)\big| > M a_T\Big) \\
\tilde{Q}_T &= N \max_{1\le n\le N} \mathbb{P}\Big(\big|\tilde{\psi}_1(u_n,x_n) - \mathbb{E}\tilde{\psi}_1(u_n,x_n)\big| > M a_T\Big).
\end{aligned}$$

As $\hat{Q}_T$ and $\tilde{Q}_T$ can be analyzed in the same way, we restrict attention to $\hat{Q}_T$ in what follows. To bound $\hat{Q}_T$, we write

$$\mathbb{P}\Big(\big|\hat{\psi}_1(u,x) - \mathbb{E}\hat{\psi}_1(u,x)\big| > M a_T\Big) = \mathbb{P}\Big(\Big|\sum_{t=1}^T Z_{t,T}(u,x)\Big| > M a_T T h^{d+1}\Big) \tag{40}$$

with

$$\begin{aligned}
Z_{t,T}(u,x) = K_h\Big(u - \frac{t}{T}\Big)\Big\{ &\prod_{j=1}^d K_h(x^j - X_{t,T}^j) W_{t,T} I(|W_{t,T}| \le \tau_T) \\
&- \mathbb{E}\Big[\prod_{j=1}^d K_h(x^j - X_{t,T}^j) W_{t,T} I(|W_{t,T}| \le \tau_T)\Big]\Big\}.
\end{aligned}$$

Note that for each fixed (u,x), the array $\{Z_{t,T}(u,x)\}$ is α-mixing with mixing coefficients α_T^Z satisfying $\alpha_T^Z(k) \le \alpha(k)$. This allows us to apply an exponential inequality for mixing arrays to the right-hand side of (40) in the next step.

Step 3. We now bound $\hat{Q}_T$ with the help of an exponential inequality from Liebscher (see Theorem 2.1 in [18]).

Lemma. *Let $Z_{t,T}$ be a zero-mean triangular array such that $|Z_{t,T}| \le b_T$ with strong mixing coefficients $\alpha(k)$. Then for any $\varepsilon > 0$ and $S_T \le T$ with $\varepsilon > 4S_T b_T$,*

$$\mathbb{P}\Big(\Big|\sum_{t=1}^{T} Z_{t,T}\Big| > \varepsilon\Big) \le 4\exp\Big(-\frac{\varepsilon^2}{64\sigma^2_{S_T,T}\frac{T}{S_T} + \frac{8}{3}\varepsilon b_T S_T}\Big) + 4\frac{T}{S_T}\alpha(S_T),$$

where $\sigma^2_{S_T,T} = \sup_{0\le j\le T-1}\mathbb{E}[(\sum_{t=j+1}^{\min\{j+S_T,T\}} Z_{t,T})^2]$.

We apply this exponential inequality with $\varepsilon = Ma_T T h^{d+1}$, $b_T = C\tau_T$ for some sufficiently large C, and $S_T = a_T^{-1}\tau_T^{-1}$. Moreover, a straightforward extension of Theorem 1 in Hansen [13] shows that $\sigma^2_{S_T,T} \le \Theta S_T h^{d+1}$ with a constant Θ independent of (u,x). It is easy to see that with these choices, the conditions of the above lemma are fulfilled. For any fixed (u,x) and T sufficiently large, we now get

$$\begin{aligned}
&\mathbb{P}\Big(\Big|\sum_{t=1}^{T} Z_{t,T}(u,x)\Big| > Ma_T T h^{d+1}\Big)\\
&\le 4\exp\Big(-\frac{\varepsilon^2}{64\Theta S_T h^{d+1}\frac{T}{S_T} + \frac{8}{3}\varepsilon S_T b_T}\Big) + 4\frac{T}{S_T}\alpha(S_T)\\
&\le 4\exp\Big(-\frac{M^2\log T}{64\Theta + \frac{8}{3}CM}\Big) + 4\frac{T}{S_T}AS_T^{-\beta}\\
&\le 4\exp\Big(-\frac{M\log T}{64 + 3C}\Big) + 4ATS_T^{-1-\beta}\\
&= 4T^{-\frac{M}{64+3C}} + 4ATS_T^{-1-\beta},
\end{aligned}$$

where we choose $M > \Theta$ to get the last inequality. Since $N \le Ch^{-(d+1)}a_T^{-(d+1)}$, it follows that

$$\hat{Q}_T \le O(R_{1T}) + O(R_{2T})$$

with

$$\begin{aligned}
R_{1T} &= h^{-(d+1)}a_T^{-(d+1)}T^{-\frac{M}{64+3C}}\\
R_{2T} &= h^{-(d+1)}a_T^{-(d+1)}TS_T^{-1-\beta}.
\end{aligned}$$

Choosing M sufficiently large, we obtain that $R_{1T} \le T^{-\eta}$ for some small $\eta > 0$. As $\frac{\phi_T \log T}{T^\theta h^{d+1}} = o(1)$ by assumption, we further get that

$$\begin{aligned}
R_{2T} &= h^{-(d+1)}a_T^{-(d+1)}T(a_T\tau_T)^{1+\beta}\\
&= \Big(\frac{\phi_T\log T}{h^{d+1}}\Big)^{1+\frac{\beta-d}{2}}T^{1-\frac{\beta-d}{2}+\frac{1+\beta}{s}}\\
&= o\big(T^{\theta(1+\frac{\beta-d}{2})+1-\frac{\beta-d}{2}+\frac{1+\beta}{s}}\big).
\end{aligned}$$

By our assumptions on θ and β, it holds that $R_{2T} = o(1)$. This shows the result. $\square$

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

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

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

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

We first derive some intermediate results for the above expressions:

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

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

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

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

(iii) It holds that

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

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

(iv) We have that

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

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

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

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

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

with

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

and

$$
\begin{aligned}
q_1(u,x) &= \mathbb{E}\Big[\prod_{j=1}^d \bar{K}_h(x^j - X^j_{t,T})\Big\{\prod_{j=1}^d K_h(x^j - X^j_{t,T}) \\
&\qquad - \prod_{j=1}^d K_h\big(x^j - X^j_t(\tfrac{t}{T})\big)\Big\}\Big\{m\Big(\frac{t}{T}, X_{t,T}\Big) - m(u,x)\Big\}\Big] \\
q_2(u,x) &= \mathbb{E}\Big[\prod_{j=1}^d \bar{K}_h(x^j - X^j_{t,T}) \prod_{j=1}^d K_h\big(x^j - X^j_t(\tfrac{t}{T})\big) \\
&\qquad \times \Big\{m\Big(\frac{t}{T}, X_{t,T}\Big) - m\Big(\frac{t}{T}, X_t(\tfrac{t}{T})\Big)\Big\}\Big] \\
q_3(u,x) &= \mathbb{E}\Big[\Big\{\prod_{j=1}^d \bar{K}_h(x^j - X^j_{t,T}) - \prod_{j=1}^d \bar{K}_h\big(x^j - X^j_t(\tfrac{t}{T})\big)\Big\} \\
&\qquad \times \prod_{j=1}^d K_h\big(x^j - X^j_t(\tfrac{t}{T})\big)\Big\{m\Big(\frac{t}{T}, X_t(\tfrac{t}{T})\Big) - m(u,x)\Big\} \\
q_4(u,x) &= \mathbb{E}\Big[\prod_{j=1}^d K_h\big(x^j - X^j_t(\tfrac{t}{T})\big)\Big\{m\Big(\frac{t}{T}, X_t(\tfrac{t}{T})\Big) - m(u,x)\Big\}\Big].
\end{aligned}
$$

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

$X_t^k(\frac{t}{T}))|^r$ for $r = \min\{\rho, 1\}$. Hence,

$$\begin{aligned} &\Big|\prod_{j=1}^d K_h(x^j - X_{t,T}^j) - \prod_{j=1}^d K_h\big(x^j - X_t^j(\tfrac{t}{T})\big)\Big| \\ &\qquad \le C\sum_{k=1}^d \big|K_h(x^k - X_{t,T}^k) - K_h\big(x^k - X_t^k(\tfrac{t}{T})\big)\big|^r. \end{aligned} \tag{41}$$

Using (41), we obtain

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

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 r-th moment, we can infer that

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

uniformly in u and x.

We next turn to $Q_2(u,x)$. Note that the expression in the expectation of $q_2(u,x)$ is non-zero only if $X_{t,T} \in [x^j - 2C_1 h, x^j + 2C_1 h]_{j=1}^d$ and $X_t(\frac{t}{T}) \in [x^j - C_1 h, x^j + C_1 h]_{j=1}^d$. As m is continuous, this implies that $|m(\frac{t}{T}, X_{t,T}) - m(\frac{t}{T}, X_t(\frac{t}{T}))| \le C$ for some constant $C < \infty$, whenever the expression in the expectation is non-zero. This allows us to use the bound

$$\Big|m\Big(\frac{t}{T}, X_{t,T}\Big) - m\Big(\frac{t}{T}, X_t(\tfrac{t}{T})\Big)\Big| \le C\Big|m\Big(\frac{t}{T}, X_{t,T}\Big) - m\Big(\frac{t}{T}, X_t(\tfrac{t}{T})\Big)\Big|^r$$

with $r = \min\{\rho, 1\}$ and some constant $C < \infty$. We thus arrive at

$$\begin{aligned} |q_2(u,x)| &\le C\mathbb{E}\Big[\prod_{j=1}^d \bar{K}_h(x^j - X_{t,T}^j)\prod_{j=1}^d K_h\big(x^j - X_t^j(\tfrac{t}{T})\big) \\ &\qquad \times \Big|m\Big(\frac{t}{T}, X_{t,T}\Big) - m\Big(\frac{t}{T}, X_t(\tfrac{t}{T})\Big)\Big|^r\Big] \\ &\le C\mathbb{E}\Big[\Big(\sum_{j=1}^d |X_{t,T}^j - X_t^j(\tfrac{t}{T})|\Big)^r\Big] \\ &\le C\mathbb{E}\Big[\Big(\frac{1}{T}U_{t,T}(\tfrac{t}{T})\Big)^r\Big] \le \frac{C}{T^r} \end{aligned}$$

uniformly in u and x. As a result, $\sup_{u,x} |Q_2(u,x)| \leq \frac{C}{T^r h^d}$.
Using analogous arguments as for $Q_1(u,x)$, we can further show that $\sup_{u,x} |Q_3(u,x)| \leq \frac{C}{T^r h^{d-1+r}}$. Finally, applying Lemmas B1 and B2 and exploiting the smoothness conditions on m and f, we obtain that

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

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

Proof of Theorem 4.3

With $\hat{g}^V(u,x)$ and $\hat{g}^B(u,x)$ as in the proof of Theorem 4.2, we let

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

and use the shorthands

$$\begin{aligned} B(u,x) &= \sqrt{Th^{d+1}} \big(\hat{g}^B(u,x) - m(u,x) \hat{f}(u,x) \big) \\ V(u,x) &= \sqrt{Th^{d+1}} \hat{g}^V(u,x). \end{aligned}$$

In what follows, we refer to $B(u,x)$ as the bias part and to $V(u,x)$ as the stochastic part.
The bias part converges in probability to the term $B_{u,x}$ defined in the statement of Theorem 4.3. This follows from (iii) in the proof of Theorem 4.2 and the fact that $B(u,x) - \mathbb{E}[B(u,x)] = o_p(1)$. In order to prove the latter, it suffices to show that $\text{Var}(B(u,x)) = o(1)$, which can be achieved by slightly varying the arguments of Theorem 1 in Hansen [13].
The stochastic part is asymptotically normal. In particular,

$$V(u,x) \stackrel{d}{\longrightarrow} N(0, \kappa_0^{d+1} \sigma^2(u,x) f(u,x)) \tag{42}$$

with $\kappa_0 = \int K^2(\varphi) d\varphi$. The proof proceeds by the usual blocking argument. Decomposing $V(u,x)$ alternately into big blocks and small blocks, we can neglect the small blocks and exploit the mixing conditions to replace the big blocks by independent random variables. This allows us to apply a Lindeberg theorem to get the result. We omit the details, as the proof is very similar to that for the standard strictly stationary setting. We however shortly comment on how to calculate the variance of $V(u,x)$. First, by the same steps as in Theorem 1 of Hansen [13],

$$\begin{aligned} \text{Var}(V(u,x)) &= \text{Var}\Big(\frac{1}{\sqrt{Th^{d+1}}} \sum_{t=1}^{T} K_h\Big(u - \frac{t}{T}\Big) \prod_{j=1}^{d} K_h(x^j - X_{t,T}^j) \varepsilon_{t,T} \Big) \\ &= \frac{1}{Th^{d+1}} \sum_{t=1}^{T} K_h^2\Big(u - \frac{t}{T}\Big) \mathbb{E}\Big[\prod_{j=1}^{d} K_h^2(x^j - X_{t,T}^j) \varepsilon_{t,T}^2 \Big] + o(1). \end{aligned}$$

Moreover, by similar steps as for (iii) in the proof of Theorem 4.2,

$$\frac{1}{Th^{d+1}} \sum_{t=1}^{T} K_h^2\Big(u - \frac{t}{T}\Big) \mathbb{E}\Big[\prod_{j=1}^{d} K_h^2(x^j - X_{t,T}^j)\varepsilon_{t,T}^2\Big] = \kappa_0^{d+1}\sigma^2(u,x)f(u,x) + o(1)$$

with $\kappa_0 = \int K^2(\varphi)d\varphi$. Hence,

$$\mathrm{Var}\big(V(u,x)\big) = \kappa_0^{d+1}\sigma^2(u,x)f(u,x) + o(1).$$

As $\hat{f}(u,x) - f(u,x) = o_p(1)$ and $\hat{f}(u,x)^{-1} = O_p(1)$, we can now combine the asymptotic normality result (42) with the fact that $B(u,x) = B_{u,x} + o_p(1)$ to complete the proof. □

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 C1. *Define $T_0 = \mathbb{E}[T_{[0,1]^d}]$. Then uniformly for $u \in I_h$,*

$$\frac{T_0}{T} = \mathbb{P}(X_0(u) \in [0,1]^d) + O(T^{-\frac{\rho}{1+\rho}}) + o(h) \tag{43}$$

with ρ defined in assumption (C1) and

$$\frac{T_{[0,1]^d} - T_0}{T_0} = O_p\Big(\sqrt{\frac{\log T}{Th}}\Big). \tag{44}$$

Proof. We first show (43). Let $U_{t,T} := U_{t,T}(\frac{t}{T})$ for short and recall that $\|X_{t,T} - X_t(\frac{t}{T})\| \le \frac{1}{T}U_{t,T}$ almost surely with $\mathbb{E}[U_{t,T}^\rho] \le C$ for some $\rho > 0$. It holds that

$$\mathbb{E}\big[I(X_{t,T} \in [0,1]^d)\big] = \mathbb{E}\big[I\big(X_{t,T} \in [0,1]^d, \|X_{t,T} - X_t(\tfrac{t}{T})\| \le \tfrac{1}{T}U_{t,T}\big)\big]$$
$$\begin{cases} \ge \mathbb{E}\big[I\big(X_t(\frac{t}{T}) \in \big[\frac{C}{T}U_{t,T}, 1 - \frac{C}{T}U_{t,T}\big]^d\big)\big] \\ \le \mathbb{E}\big[I\big(X_t(\frac{t}{T}) \in \big[-\frac{C}{T}U_{t,T}, 1 + \frac{C}{T}U_{t,T}\big]^d\big)\big] \end{cases}$$

for some sufficiently large $C < \infty$. Hence, defining

$$B_L = \frac{1}{T}\sum_{t=1}^{T} K_h\Big(u, \frac{t}{T}\Big)\mathbb{E}\big[I\big(X_t(\tfrac{t}{T}) \in \big[\tfrac{C}{T}U_{t,T}, 1 - \tfrac{C}{T}U_{t,T}\big]^d\big)\big]$$
$$B_U = \frac{1}{T}\sum_{t=1}^{T} K_h\Big(u, \frac{t}{T}\Big)\mathbb{E}\big[I\big(X_t(\tfrac{t}{T}) \in \big[-\tfrac{C}{T}U_{t,T}, 1 + \tfrac{C}{T}U_{t,T}\big]^d\big)\big],$$

we get $B_L \le \frac{T_0}{T} \le B_U$. Now let $0 < q < 1$ and write $B_U = B_{U,1} + B_{U,2}$ with

$$B_{U,1} = \frac{1}{T}\sum_{t=1}^{T} K_h\Big(u, \frac{t}{T}\Big)\mathbb{E}\big[I\big(X_t(\tfrac{t}{T}) \in \big[-\tfrac{C}{T}U_{t,T}, 1 + \tfrac{C}{T}U_{t,T}\big]^d, U_{t,T} \le T^q\big)\big]$$

and $B_{U,2} = B_U - B_{U,1}$. Using Lemma B2, we can show that uniformly for $u \in I_h$,

$$\begin{aligned} B_{U,1} &\le \frac{1}{T}\sum_{t=1}^{T} K_h\Big(u, \frac{t}{T}\Big)\mathbb{E}\big[I\big(X_t(\tfrac{t}{T}) \in \big[-\tfrac{C}{T^{1-q}}, 1 + \tfrac{C}{T^{1-q}}\big]^d\big)\big] \\ &= \int I\big(x \in \big[-\tfrac{C}{T^{1-q}}, 1 + \tfrac{C}{T^{1-q}}\big]^d\big)\frac{1}{T}\sum_{t=1}^{T} K_h\Big(u, \frac{t}{T}\Big) f\Big(\frac{t}{T}, x\Big)dx \\ &= \int I\big(x \in [0,1]^d\big) f(u,x)dx + O\Big(\frac{1}{T^{1-q}}\Big) + o(h). \end{aligned}$$

Moreover, it is easy to see that $B_{U,2} \le CT^{-q\rho}$. Setting $q = (1+\rho)^{-1}$, we thus arrive at

$$B_U \le \mathbb{P}(X_0(u) \in [0,1]^d) + O(T^{-\frac{\rho}{1+\rho}}) + o(h) \tag{45}$$

uniformly in u. By similar arguments, $B_L \ge \mathbb{P}(X_0(u) \in [0,1]^d) + O(T^{-\frac{\rho}{1+\rho}}) + o(h)$. This yields (43). Equation (44) now follows immediately:

$$\frac{T_{[0,1]^d} - T_0}{T_0} = \frac{T}{T_0} \cdot \frac{1}{T}(T_{[0,1]^d} - T_0) = O_p\Big(\sqrt{\frac{\log T}{Th}}\Big)$$

uniformly in u, since $\frac{1}{T}(T_{[0,1]^d} - T_0) = O_p(\sqrt{\log T/Th})$ by Theorem 4.1 and $\frac{T_0}{T} = O_p(1)$ by (43). □

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

Lemma C2. *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*

$$\begin{aligned} \sup_{u,x^j \in I_h} \big|\hat{p}_j(u,x^j) - p_j(u,x^j)\big| &= O_p(v_{T,2}) + O(b_{T,r}) + o(h) \\ \sup_{u \in I_h,\, x^j \in [0,1]} \big|\hat{p}_j(u,x^j) - \kappa_0(x^j)p_j(u,x^j)\big| &= O_p(v_{T,2}) + O(b_{T,r}) + O(h) \\ \sup_{u,x^j,x^k \in I_h} \big|\hat{p}_{j,k}(u,x^j,x^k) - p_{j,k}(u,x^j,x^k)\big| &= O_p(v_{T,3}) + O(b_{T,r}) + o(h) \\ \sup_{\substack{u \in I_h, \\ x^j,x^k \in [0,1]}} \big|\hat{p}_{j,k}(u,x^j,x^k) - \kappa_0(x^j)\kappa_0(x^k)p_{j,k}(u,x^j,x^k)\big| &= O_p(v_{T,3}) + O(b_{T,r}) + O(h). \end{aligned}$$

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

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

By (44) from Lemma C1, 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. □

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

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

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

Lemma C3. *It holds that*

$$\sup_{u,x^j \in [0,1]} \big|\hat{m}_j^A(u,x^j)\big| = O_p\Big(\sqrt{\frac{\log T}{Th^2}}\Big). \tag{46}$$

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 C4. *It holds that*

$$\sup_{u,x^j \in I_h} \big|\hat{m}_j^B(u,x^j) - \hat{\mu}_{T,j}(u,x^j)\big| = o_p(h^2) \tag{47}$$

$$\sup_{u \in I_h,\, x^j \in I_h^c} \big|\hat{m}_j^B(u,x^j) - \hat{\mu}_{T,j}(u,x^j)\big| = O_p(h^2) \tag{48}$$

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

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

Here,

$$\begin{aligned}
\alpha_{T,0}(u) &= m_0(u) + h\kappa_1(u)\partial_u m_0(u) + \frac{h^2}{2}\kappa_2(u)\partial_{uu}^2 m_0(u) \\
\alpha_{T,k}(u,x^k) &= m_k(u,x^k) + h\Big[\kappa_1(u)\partial_u m_k(u,x^k) + \frac{\kappa_0(u)\kappa_1(x^k)}{\kappa_0(x^k)}\partial_{x^k} m_k(u,x^k)\Big]
\end{aligned}$$

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

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. By definition

$$
\begin{aligned}
\hat{m}_j^B(u,x^j) = \; & \frac{1}{T_{[0,1]^d}} \sum_{t=1}^{T} I_{t,T} K_h\Big(u, \frac{t}{T}\Big) K_h(x^j, X_{t,T}^j) m_0\Big(\frac{t}{T}\Big) / \hat{p}_j(u,x^j) \\
& + \frac{1}{T_{[0,1]^d}} \sum_{t=1}^{T} I_{t,T} K_h\Big(u, \frac{t}{T}\Big) K_h(x^j, X_{t,T}^j) m_j\Big(\frac{t}{T}, X_{t,T}^j\Big) / \hat{p}_j(u,x^j) \\
& + \sum_{k \neq j} \frac{1}{T_{[0,1]^d}} \sum_{t=1}^{T} I_{t,T} K_h\Big(u, \frac{t}{T}\Big) K_h(x^j, X_{t,T}^j) m_k\Big(\frac{t}{T}, X_{t,T}^k\Big) / \hat{p}_j(u,x^j) \\
=: \; & \hat{m}_j^{B,0}(u,x^j) + \hat{m}_j^{B,j}(u,x^j) + \sum_{k \neq j} \hat{m}_j^{B,k}(u,x^j),
\end{aligned}
$$

where we have used the shorthand $I_{t,T} = I(X_{t,T} \in [0,1]^d)$. We show that

$$
\begin{aligned}
\hat{m}_j^{B,0}(u,x^j) = \; & m_0(u) + h\kappa_1(u) \partial_u m_0(u) \\
& + h^2 \Big[\kappa_2(u) \partial_u m_0(u) \frac{\partial_u p_j(u,x^j)}{p_j(u,x^j)} + \frac{1}{2} \kappa_2(u) \partial^2_{uu} m_0(u) \Big] \\
& + R_T^0(u,x^j)
\end{aligned} \tag{49}
$$

with $\sup_{u,x^j \in I_h} |R_T^0(u,x^j)| = o_p(h^2)$ and $\sup_{u \in I_h, x^j \in I_h^c} |R_T^0(u,x^j)| = O_p(h^2)$,

$$
\begin{aligned}
\hat{m}_j^{B,j}(u,x^j) = \; & m_j(u,x^j) \\
& + h \Big[\kappa_1(u) \partial_u m_j(u,x^j) + \frac{\kappa_0(u)\kappa_1(x^j)}{\kappa_0(x^j)} \partial_{x^j} m_j(u,x^j) \Big] \\
& + h^2 \Big[\kappa_2(u) \partial_u m_j(u,x^j) \frac{\partial_u p_j(u,x^j)}{p_j(u,x^j)} + \frac{1}{2} \kappa_2(u) \partial^2_{uu} m_j(u,x^j) \\
& \qquad + \frac{\kappa_0(u)\kappa_2(x^j)}{\kappa_0(x^j)} \partial_{x^j} m_j(u,x^j) \frac{\partial_{x^j} p_j(u,x^j)}{p_j(u,x^j)} \\
& \qquad + \frac{1}{2} \frac{\kappa_0(u)\kappa_2(x^j)}{\kappa_0(x^j)} \partial^2_{x^j x^j} m_j(u,x^j) \Big] \\
& + R_T^j(u,x^j),
\end{aligned} \tag{50}
$$

where R_T^j is of the same uniform order as R_T^0, and

$$
\begin{aligned}
\hat{m}_j^{B,k}(u,x^j) = &\int m_k(u,x^k)\frac{\hat{p}_{j,k}(u,x^j,x^k)}{\hat{p}_j(u,x^j)}dx^k \\
&+ h\int\Big[\kappa_1(u)\partial_u m_k(u,x^k) \\
&\qquad + \frac{\kappa_0(u)\kappa_1(x^k)}{\kappa_0(x^k)}\partial_{x^k} m_k(u,x^k)\Big]\frac{\hat{p}_{j,k}(u,x^j,x^k)}{\hat{p}_j(u,x^j)}dx^k \\
&+ h^2\Big[\kappa_2(u)\int\kappa_0(x^k)\partial_u m_k(u,x^k)\frac{\partial_u p_{j,k}(u,x^j,x^k)}{p_j(u,x^j)}dx^k \\
&\quad + \kappa_0(u)\int\kappa_2(x^k)\partial_{x^k} m_k(u,x^k)\frac{\partial_{x^k} p_{j,k}(u,x^j,x^k)}{p_j(u,x^j)}dx^k \\
&\quad + \kappa_2(u)\int\kappa_0(x^k)\frac{1}{2}\partial_{uu}^2 m_k(u,x^k)\frac{p_{j,k}(u,x^j,x^k)}{p_j(u,x^j)}dx^k \\
&\quad + \kappa_0(u)\int\kappa_2(x^k)\frac{1}{2}\partial_{x^k x^k}^2 m_k(u,x^k)\frac{p_{j,k}(u,x^j,x^k)}{p_j(u,x^j)}dx^k\Big] \\
&+ R_T^k(u,x^j),
\end{aligned} \tag{51}
$$

where again R_T^k is of the same uniform order as R_T^0. Combining (49)–(51) completes the proof.

We only give the proof of (51), as this is the most complicated term: Recalling that $\int K_h(x^k, X_{t,T}^k)dx^k = 1$, a second-order Taylor expansion of $m_k(\frac{t}{T}, X_{t,T}^k)$ around (u, x^k) yields

$$
\begin{aligned}
\hat{m}_j^{B,k}(u,x^j) = &\int \frac{\hat{p}_{j,k}(u,x^j,x^k)}{\hat{p}_j(u,x^j)} m_k(u,x^k)dx^k \\
&+ \frac{1}{T_{[0,1]^d}}\sum_{t=1}^{T}\big(V_{t,T}^k(u,x^j) + W_{t,T}^k(u,x^j)\big)/\hat{p}_j(u,x^j) + o_p(h^2)
\end{aligned}
$$

uniformly for $u \in I_h$ and $x^j \in [0,1]$ with

$$
\begin{aligned}
V_{t,T}^k(u,x^j) = &\, I(X_{t,T}\in[0,1]^d)K_h\Big(u,\frac{t}{T}\Big)K_h(x^j,X_{t,T}^j)\int K_h(x^k,X_{t,T}^k) \\
&\times\Big[\partial_u m_k(u,x^k)\Big(\frac{t}{T}-u\Big) + \partial_{x^k} m_k(u,x^k)(X_{t,T}^k - x^k)\Big]dx^k \\
W_{t,T}^k(u,x^j) = &\, I(X_{t,T}\in[0,1]^d)K_h\Big(u,\frac{t}{T}\Big)K_h(x^j,X_{t,T}^j)\int K_h(x^k,X_{t,T}^k) \\
&\times\Big[\frac{1}{2}\partial_{uu}^2 m_k(u,x^k)\Big(\frac{t}{T}-u\Big)^2 + \partial_{ux^k}^2 m_k(u,x^k)\Big(\frac{t}{T}-u\Big)(X_{t,T}^k - x^k) \\
&+ \frac{1}{2}\partial_{x^k x^k}^2 m_k(u,x^k)(X_{t,T}^k - x^k)^2\Big]dx^k.
\end{aligned}
$$

We now have a closer look at the expectations of $V_{t,T}^k(u,x^j)$ and $W_{t,T}^k(u,x^j)$. First, note that

$$\begin{aligned}
&\mathbb{E}[V_{t,T}^k(u,x^j)] \\
&= \mathbb{E}\Big[I\big(X_t(\tfrac{t}{T}) \in [0,1]^d\big) K_h\Big(u, \frac{t}{T}\Big) K_h\big(x^j, X_t^j(\tfrac{t}{T})\big) \int K_h\big(x^k, X_t^k(\tfrac{t}{T})\big) \\
&\qquad \times \Big\{ \partial_u m_k(u,x^k) \Big(\frac{t}{T} - u\Big) + \partial_{x^k} m_k(u,x^k) \big(X_t^k(\tfrac{t}{T}) - x^k\big) \Big\} dx^k \Big] \\
&\quad + O\Big(\frac{1}{T^{\frac{r}{r+1}}} + \frac{1}{T^r h}\Big) \qquad (52)
\end{aligned}$$

with $r = \min\{\rho, 1\}$ uniformly for $u \in I_h$ and $x^j \in [0,1]$. This is shown by successively replacing the occurrences of $X_{t,T}$ in $\mathbb{E}[V_{t,T}^k(u,x^j)]$ by $X_t(\frac{t}{T})$. In order to replace the occurrence in the indicator function $I(X_{t,T} \in [0,1]^d)$, similar arguments as in Lemma C1 can be used. For replacing the occurrences in $K_h(x^j, X_{t,T}^j)$ and $K_h(x^k, X_{t,T}^k)$, we exploit the Lipschitz continuity of K and use arguments similar to those for (iii) in the proof of Theorem 4.2. With (52), we can now write

$$\begin{aligned}
\frac{1}{T_{[0,1]^d}} \sum_{t=1}^T \mathbb{E}[V_{t,T}^k(u,x^j)] &= \frac{1}{T_{[0,1]^d}} \sum_{t=1}^T K_h\Big(u, \frac{t}{T}\Big) \int K_h(x^j, w^j) K_h(x^k, w^k) \\
&\qquad \times \Big[\partial_u m_k(u,x^k) \Big(\frac{t}{T} - u\Big) + \partial_{x^k} m_k(u,x^k)(w^k - x^k) \Big] \\
&\qquad \times \Big(\int I(w \in [0,1]^d) f\Big(\frac{t}{T}, w\Big) dw^{-j,k} \Big) dw^j dw^k dx^k \\
&\quad + O\Big(\frac{1}{T^{\frac{r}{r+1}}} + \frac{1}{T^r h}\Big)
\end{aligned}$$

uniformly for $u \in I_h$ and $x^j \in [0,1]$, where $w^{-j,k}$ denotes all but the j-th and k-th component of the vector w. Noting that $O(T^{-\frac{r}{r+1}} + \frac{1}{T^r h}) = o(h^2)$ by (Add2), using a first-order Taylor expansion of $f(\frac{t}{T}, w)$ and recalling the definition of the density p, we can infer that

$$\begin{aligned}
&\frac{1}{T_{[0,1]^d}} \sum_{t=1}^T \mathbb{E}[V_{t,T}^k(u,x^j)] \\
&= \frac{T}{T_{[0,1]^d}} P(X_0(u) \in [0,1]^d) \\
&\quad \times \Big\{ \frac{1}{T} \sum_{t=1}^T K_h\Big(u, \frac{t}{T}\Big) \Big(\frac{t}{T} - u\Big) \int \kappa_0(x^j) \kappa_0(x^k) \partial_u m_k(u,x^k) p_{j,k}(u, x^j, x^k) dx^k \\
&\qquad + \frac{1}{T} \sum_{t=1}^T K_h\Big(u, \frac{t}{T}\Big) \int h \kappa_0(x^j) \kappa_1(x^k) \partial_{x^k} m_k(u,x^k) p_{j,k}(u, x^j, x^k) dx^k \\
&\qquad + \frac{1}{T} \sum_{t=1}^T K_h\Big(u, \frac{t}{T}\Big) \Big(\frac{t}{T} - u\Big)^2 \int \kappa_0(x^j) \kappa_0(x^k) \partial_u m_k(u,x^k) \partial_u p_{j,k}(u, x^j, x^k) dx^k \\
&\qquad + \frac{1}{T} \sum_{t=1}^T K_h\Big(u, \frac{t}{T}\Big) \int h^2 \kappa_0(x^j) \kappa_2(x^k) \partial_{x^k} m_k(u,x^k) \partial_{x^k} p_{j,k}(u, x^j, x^k) dx^k \Big\} \\
&\quad + o_p(h^2)
\end{aligned}$$

uniformly for $u \in I_h$ and $x^j \in [0,1]$. Next note that by Lemma C1,

$$\frac{T}{T_{[0,1]^d}} P(X_0(u) \in [0,1]^d) = 1 + O\Big(\sqrt{\frac{\log T}{Th}}\Big) + O(T^{-\frac{\rho}{1+\rho}}) + o(h)$$

uniformly in u. Using this together with Lemmas B1 and B2 from Appendix B, we get

$$\begin{aligned}
&\frac{1}{T_{[0,1]^d}} \sum_{t=1}^{T} \mathbb{E}[V_{t,T}^k(u,x^j)] \\
&= h\Big[\kappa_1(u)\kappa_0(x^j) \int \kappa_0(x^k)\partial_u m_k(u,x^k) p_{j,k}(u,x^j,x^k) dx^k \\
&\qquad + \kappa_0(u)\kappa_0(x^j) \int \kappa_1(x^k)\partial_{x^k} m_k(u,x^k) p_{j,k}(u,x^j,x^k) dx^k\Big] \\
&\quad + h^2\Big[\kappa_2(u)\kappa_0(x^j) \int \kappa_0(x^k)\partial_u m_k(u,x^k)\partial_u p_{j,k}(u,x^j,x^k) dx^k \\
&\qquad + \kappa_0(u)\kappa_0(x^j) \int \kappa_2(x^k)\partial_{x^k} m_k(u,x^k)\partial_{x^k} p_{j,k}(u,x^j,x^k) dx^k\Big] \\
&\quad + R_T^V(u,x^j) \qquad (53)
\end{aligned}$$

with $\sup_{u,x^j \in I_h} |R_T^V(u,x^j)| = o(h^2)$ and $\sup_{u \in I_h, x^j \in I_h^c} |R_T^V(u,x^j)| = O(h^2)$. Since $\kappa_1(u) = 0$ for all $u \in I_h$ and

$$\int \partial_{x^k} m_k(u,x^k)\Big[\frac{1}{\kappa_0(x^k)} \hat{p}_{j,k}(u,x^j,x^k) - \kappa_0(x^j) p_{j,k}(u,x^j,x^k)\Big] h\kappa_1(x^k) dx^k = O_p(h^2)$$

uniformly for $u \in I_h$ and $x^j \in [0,1]$, we can rewrite (53) as

$$\begin{aligned}
&\frac{1}{T_{[0,1]^d}} \sum_{t=1}^{T} \mathbb{E}[V_{t,T}^k(u,x^j)] \\
&= h\Big[\kappa_1(u) \int \partial_u m_k(u,x^k) \hat{p}_{j,k}(u,x^j,x^k) dx^k \\
&\qquad + \kappa_0(u) \int \frac{\kappa_1(x^k)}{\kappa_0(x^k)} \partial_{x^k} m_k(u,x^k) \hat{p}_{j,k}(u,x^j,x^k) dx^k\Big] \\
&\quad + h^2\Big[\kappa_2(u)\kappa_0(x^j) \int \kappa_0(x^k)\partial_u m_k(u,x^k)\partial_u p_{j,k}(u,x^j,x^k) dx^k \\
&\qquad + \kappa_0(u)\kappa_0(x^j) \int \kappa_2(x^k)\partial_{x^k} m_k(u,x^k)\partial_{x^k} p_{j,k}(u,x^j,x^k) dx^k\Big] \\
&\quad + \tilde{R}_T^V(u,x^j), \qquad (54)
\end{aligned}$$

where $\tilde{R}_T^V$ is of the same uniform order as R_T^V. By analogous arguments as above, we can further show that

$$\begin{aligned}
&\frac{1}{T_{[0,1]^d}} \sum_{t=1}^{T} \mathbb{E}[W_{t,T}^k(u,x^j)] \\
&= \frac{h^2}{2}\Big[\kappa_2(u)\kappa_0(x^j) \int \kappa_0(x^k)\partial_{uu}^2 m_k(u,x^k) p_{j,k}(u,x^j,x^k) dx^k
\end{aligned}$$

$$
\begin{aligned}
&\qquad + \kappa_0(u)\kappa_0(x^j) \int \kappa_2(x^k)\partial^2_{x^k x^k} m_k(u,x^k) p_{j,k}(u,x^j,x^k) dx^k \Big] \\
&+ R_T^W(u,x^j) \qquad (55)
\end{aligned}
$$

with $\sup_{u,x^j \in I_h} |R_T^W(u,x^j)| = o(h^2)$ and $\sup_{u \in I_h, x^j \in I_h^c} |R_T^W(u,x^j)| = O(h^2)$. Finally, applying the same proving strategy as in Theorem 4.1, one can show that

$$
\sup_{u \in I_h,\, x^j \in [0,1]} \Big| \frac{1}{T_{[0,1]^d}} \sum_{t=1}^{T} \big(V_{t,T}^k(u,x^j) - \mathbb{E}[V_{t,T}^k(u,x^j)] \big) \Big| = o_p(h^2)
$$

$$
\sup_{u \in I_h,\, x^j \in [0,1]} \Big| \frac{1}{T_{[0,1]^d}} \sum_{t=1}^{T} \big(W_{t,T}^k(u,x^j) - \mathbb{E}[W_{t,T}^k(u,x^j)] \big) \Big| = o_p(h^2).
$$

Therefore,

$$
\begin{aligned}
\hat{m}_j^{B,k}(u,x^j) = &\int m_k(u,x^k) \frac{\hat{p}_{j,k}(u,x^j,x^k)}{\hat{p}_j(u,x^j)} dx^k \\
&+ \frac{1}{T_{[0,1]^d}} \sum_{t=1}^{T} \big(\mathbb{E}[V_{t,T}^k(u,x^j)] + \mathbb{E}[W_{t,T}^k(u,x^j)] \big) / \hat{p}_j(u,x^j) + o_p(h^2)
\end{aligned}
$$

uniformly for $u \in I_h$ and $x^j \in [0,1]$. Plugging (54) and (55) into the above expression and using the fact that $\hat{p}_j(u,x^j)$ converges uniformly to $\kappa_0(x^j)p_j(u,x^j)$ yields (51). □

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

Lemma C5. *It holds that*

$$
\sup_{u \in I_h} \big| \tilde{m}_0(u) - m_0(u) \big| = O_p\Big(\sqrt{\frac{\log T}{Th}} + h^2 \Big). \qquad (56)
$$

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

To prove Theorems 5.1 and 5.2, it suffices to show that the high-level conditions (A1)–(A6), (A8), and (A9) of Mammen et al. [20] are satisfied. This allows us to apply their Theorems 1–3, which imply the statements of Theorems 5.1 and 5.2. As will be seen, the high-level conditions are satisfied uniformly for $u \in I_h$ rather than only pointwise. Inspecting the proofs of Theorems 1–3 in [20], we can thus infer that the convergence rates in (22) hold uniformly over $u \in I_h$ rather than only pointwise. In what follows, we consider the high-level conditions one after the other.

(A1) For all $j \neq k$, it holds that

$$
\int \frac{p_{j,k}^2(u,x^j,x^k)}{p_k(u,x^k)p_j(u,x^j)} dx^j dx^k < \infty
$$

uniformly for $u \in I_h$.

This condition follows immediately from the assumptions on the density $f(u,x)$. These imply that $p_j(u,x^j) \geq c > 0$ and $p_{j,k}(u,x^j,x^k) \leq C < \infty$ for all $u \in [0,1]$ and x^j, $x^k \in [0,1]$ with some appropriately chosen constants c and C.

(A2) For all $j \neq k$, it holds that

$$\int \Big[\frac{\hat{p}_j(u,x^j) - p_j(u,x^j)}{p_j(u,x^j)}\Big]^2 p_j(u,x^j)dx^j = o_p(1)$$

$$\int \Big[\frac{\hat{p}_{j,k}(u,x^j,x^k)}{p_k(u,x^k)p_j(u,x^j)} - \frac{p_{j,k}(u,x^j,x^k)}{p_k(u,x^k)p_j(u,x^j)}\Big]^2 p_k(u,x^k)p_j(u,x^j)dx^j dx^k = o_p(1)$$

$$\int \Big[\frac{\hat{p}_{j,k}(u,x^j,x^k)}{p_k(u,x^k)\hat{p}_j(u,x^j)} - \frac{p_{j,k}(u,x^j,x^k)}{p_k(u,x^k)p_j(u,x^j)}\Big]^2 p_k(u,x^k)p_j(u,x^j)dx^j dx^k = o_p(1)$$

uniformly for $u \in I_h$. Furthermore, for each $u \in I_h$, $\hat{p}_j(u,\cdot)$ and $\hat{p}_{j,k}(u,\cdot)$ vanish outside the support of $p_j(u,\cdot)$ and $p_{j,k}(u,\cdot)$, respectively.

This condition as well as (A4) and (A8) can easily be proven by using the uniform convergence results for the kernel densities derived in Lemma C2.

(A3) There exists a finite constant C such that with probability tending to 1,

$$\int \hat{m}_j^2(u,x^j)p_j(u,x^j)dx^j < \infty$$

uniformly for $u \in I_h$.

Both this condition and (A5) directly follow from Lemmas C3 and C4, which describe the asymptotic behaviour of the variance part $\hat{m}_j^A$ and the bias part $\hat{m}_j^B$ of the Nadaraya-Watson estimate $\hat{m}_j$.

(A4) There exists a finite constant C such that with probability tending to 1,

$$\sup_{x^k \in I_h} \int \frac{\hat{p}_{j,k}^2(u,x^j,x^k)}{\hat{p}_k^2(u,x^k)p_j(u,x^j)}dx^j \leq C$$

for all $j \neq k$ uniformly for $u \in I_h$.

(A5) There exists a finite constant C such that with probability tending to 1,

$$\int \hat{m}_j^A(u,x^j)^2 p_j(u,x^j)dx^j \leq C$$

$$\int \hat{m}_j^B(u,x^j)^2 p_j(u,x^j)dx^j \leq C$$

uniformly for $u \in I_h$.

(A6) For $j \neq k$, it holds that

$$\sup_{x^j \in I_h} \Big| \int \frac{\hat{p}_{j,k}(u,x^j,x^k)}{\hat{p}_j(u,x^j)} \hat{m}_k^A(u,x^k) dx^k \Big| = o_p(h^2)$$
$$\Big\| \int \frac{\hat{p}_{j,k}(u,x^j,x^k)}{\hat{p}_j(u,x^j)} \hat{m}_k^A(u,x^k) dx^k \Big\|_2 = o_p(h^2)$$

uniformly for $u \in I_h$, where $\| \cdot \|_2$ denotes the norm in the space $L_2(p_j(u,\cdot))$.

To prove (A6), it suffices to show that

$$\sup_{u \in I_h,\, x^j \in [0,1]} \Big| \int \frac{\hat{p}_{j,k}(u,x^j,x^k)}{\hat{p}_j(u,x^j)} \hat{m}_k^A(u,x^k) dx^k \Big| = O_p\Big(\sqrt{\frac{\log T}{Th}}\Big). \tag{57}$$

For the proof of (57), we write

$$\begin{aligned} S_{k,j}(u,x^j) &= \int \frac{\hat{p}_{j,k}(u,x^j,x^k)}{\hat{p}_j(u,x^j)} \hat{m}_k^A(u,x^k) dx^k \\ &= \int \frac{\hat{p}_{j,k}(u,x^j,x^k)}{\hat{p}_j(u,x^j)\hat{p}_k(u,x^k)} \hat{\psi}_k(u,x^k) dx^k, \end{aligned}$$

where $\hat{m}_k^A(u,x^k) = \hat{\psi}_k(u,x^k)/\hat{p}_k(u,x^k)$ with

$$\hat{\psi}_k(u,x^k) = \frac{1}{T_{[0,1]^d}} \sum_{t=1}^T I(X_{t,T} \in [0,1]^d) K_h\Big(u, \frac{t}{T}\Big) K_h(x^k, X_{t,T}^k) \varepsilon_{t,T}.$$

In a first step, we replace $S_{k,j}(u,x^j)$ by the term

$$S_{k,j}^*(u,x^j) = \int \frac{p_{j,k}(u,x^j,x^k)}{p_j(u,x^j)p_k(u,x^k)} \hat{\psi}_k(u,x^k) dx^k$$

and show that the resulting error is asymptotically negligible. This is done as follows:

$$\begin{aligned} &\sup_{u \in I_h, x^j \in [0,1]} \big| S_{k,j}(u,x^j) - S_{k,j}^*(u,x^j) \big| \\ &= \sup_{u,x^j} \Big| \int \Big\{ \frac{\hat{p}_{j,k}(u,x^j,x^k)}{\hat{p}_j(u,x^j)\hat{p}_k(u,x^k)} - \frac{\kappa_0(x^j)\kappa_0(x^k)p_{j,k}(u,x^j,x^k)}{\kappa_0(x^j)p_j(u,x^j)\kappa_0(x^k)p_k(u,x^k)} \Big\} \hat{\psi}_k(u,x^k) dx^k \Big| \\ &= O_p\Big(\sqrt{\frac{\log T}{Th^3}} + h\Big) O_p\Big(\sqrt{\frac{\log T}{Th^2}}\Big) = O_p\Big(\frac{\log T}{Th^{5/2}} + \sqrt{\frac{\log T}{T}}\Big), \end{aligned}$$

as $\hat{\psi}_k(u,x^k) = O_p(\sqrt{\log T/Th^2})$ and the term in curly brackets is of the order $O_p(\sqrt{\log T/Th^3} + h)$ uniformly in u, x^j, and x^k. In a second step, we show that

$$\sup_{u \in I_h, x^j \in [0,1]} \big| S_{k,j}^*(u,x^j) \big| = O_p\Big(\sqrt{\frac{\log T}{Th}}\Big).$$

To prove this, we write

$$S_{k,j}^*(u,x^j) = \frac{1}{T_{[0,1]^d}} \sum_{t=1}^T w_{k,j}(u,x^j,X_{t,T}^k) \varepsilon_{t,T} \tag{58}$$

with

$$w_{k,j}(u, x^j, X_{t,T}^k) = I(X_{t,T} \in [0,1]^d) K_h\Big(u, \frac{t}{T}\Big) \times \Big(\int \frac{p_{j,k}(u, x^j, x^k)}{p_j(u, x^j) p_k(u, x^k)} K_h(x^k, X_{t,T}^k) dx^k\Big).$$

Applying the techniques from the proof of Theorem 4.1 to (58) completes the proof of (57), which in turn yields (A6).

(A8) It holds that

$$\sup_{x^j \in I_h} \int \Big| \frac{p_{j,k}(u, x^j, x^k)}{p_j(u, x^j) p_k(u, x^k)} - \frac{\hat{p}_{j,k}(u, x^j, x^k)}{\hat{p}_j(u, x^j) \hat{p}_k(u, x^k)} \Big| p_k(u, x^k) dx^k = o_p(1)$$

uniformly for $u \in I_h$.

(A9) There exist deterministic functions

$$\alpha_{T,0}(u), \alpha_{T,1}(u, x^1), \ldots, \alpha_{T,d}(u, x^d)$$

$$\gamma_{T,1}(u), \ldots, \gamma_{T,d}(u)$$

and a function $\beta(u, x)$ (not depending on T) such that uniformly for $u \in I_h$

$$\int \alpha_{T,j}^2(u, x^j) p_j(u, x^j) dx^j < \infty \tag{59}$$

$$\int \beta^2(u, x) p(u, x) dx < \infty \tag{60}$$

$$\sup_{x^1 \in I_h, \ldots, x^d \in I_h} |\beta(u, x)| < \infty \tag{61}$$

$$\int \alpha_{T,j}(u, x^j) \hat{p}_j(u, x^j) dx^j = \gamma_{T,j}(u) + o_p(h^2) \tag{62}$$

with $\gamma_{T,j}(u) = O(h^2)$ and

$$\sup_{u, x^j \in I_h} \big| \hat{m}_j^B(u, x^j) - \hat{\mu}_{T,0}(u) - \hat{\mu}_{T,j}(u, x^j) \big| = o_p(h^2) \tag{63}$$

$$\sup_{u \in I_h} \int \big| \hat{m}_j^B(u, x^j) - \hat{\mu}_{T,0}(u) - \hat{\mu}_{T,j}(u, x^j) \big|^2 p_j(u, x^j) dx^j = o_p(h^4). \tag{64}$$

Here, $\hat{\mu}_{T,0}(u)$ is some random function and

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

We finally prove (A9). Equations (63) and (64) immediately follow from the uniform expansion of the bias part $\hat{m}_j^B$ proven in Lemma C4. Furthermore, it is trivial to see that (59)–(61) are fulfilled for $\alpha_{T,j}(u, x^j)$ and $\beta(u, x)$ as defined in Lemma C4. Finally, straightforward calculations yield a term $\gamma_{T,j}(u)$ in (62) which is of order h^2 uniformly for $u \in I_h$.

This completes the proof of Theorems 5.1 and 5.2. □

References

[1] Alizadeh, S., Brandt, M.W. & Diebold, F.X. (2002). Range-based estimation of stochastic volatility models. *The Journal of Finance* **57** 1047-1091.

[2] An, H. Z. & Huang, F. C. (1996). The geometrical ergodicity of nonlinear autoregressive models. *Statistica Sinica* **6** 943-956.

[3] Bhattacharya, R. & Lee, C. (1995). On geometric ergodicity of nonlinear autoregressive models. *Statistics and Probability Letters* **22** 311-315.

[4] Bosq, D. (1998). *Nonparametric Statistics for Stochastic Processes: Estimation and Prediction.* 2nd edition, Springer.

[5] Chen, M. & Chen, G. (2000). Geometric ergodicity of nonlinear autoregressive models with changing conditional variances. *The Canadian Journal of Statistics* **28** 605-613.

[6] Dahlhaus, R. (1996). On the Kullback-Leibler information divergence of locally stationary processes. *Stochastic Process. Appl.* **62** 139-168.

[7] Dahlhaus, R. (1996). Asymptotic statistical inference for nonstationary processes with evolutionary spectra. In *Athens Conference on Applied Probability and Time Series Analysis*, (Eds. P. M. Robinson and M. Rosenblatt) **2**. Springer, New York.

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

[9] Dahlhaus, R., Neumann, M. H. & von Sachs, R. (1999). Nonlinear wavelet estimation of time-varying autoregressive processes. *Bernoulli* **5** 873-906.

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

[11] Fryzlewicz, P., Sapatinas, T. & Subba Rao, S. (2008). Normalised least squares estimation in time-varying ARCH models. *Annals of Statistics* **36** 742-786.

[12] Fryzlewicz, P. & Subba Rao, S. (2011). Mixing properties of ARCH and time-varying ARCH processes. *Bernoulli* **17** 320-346.

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

[14] Karlsen, H. A. & Tjøstheim, D. (2001). Nonparametric estimation in null recurrent time series. *Annals of Statistics* **29** 372-416.

[15] Koo, B. & Linton, O. (2010). Semiparametric estimation of locally stationary diffusion models. Forthcoming in *Journal of Econometrics*.

[16] Kristensen, D. (2009). Uniform convergence rates of kernel estimators with heterogeneous, dependent data. *Econometric Theory* **25** 1433-1445.

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

[18] Liebscher, E. (1996). Strong convergence of sums of α-mixing random variables with applications to density estimation. *Stochastic Processes and Their Applications* **65** 69-80.

[19] Linton, O. & Hafner, C. M. (2010). Efficient estimation of a multivariate multiplicative volatility model. *The Journal of Econometrics* **159** 55-73.

[20] Mammen, E., Linton, O. & Nielsen, J. (1999). The existence and asymptotic properties of a backfitting projection algorithm under weak conditions. *Annals of Statistics* **27** 1443-1490.

[21] Martens, M. & van Dijk, D. (2007). Measuring volatility with the realized range. *Journal of Econometrics* **138** 181-207.

[22] Masry, E. (1996). Multivariate local polynomial regression for time series: Uniform strong consistency and rates. *Journal of Time Series Analysis* **17** 571-599.

[23] Moulines, E., Priouret, P. & Roueff, F. (2005). On recursive estimation for time varying autoregressive processes. *Annals of Statistics* **33** 2610-2654.

[24] Rogers, L.C.G. & Satchell, S.E. (1991). Estimating variance from high, low and closing prices. *Annals of Applied Probability* **1** 504-512.

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

[26] Tjøstheim, D. (1990). Non-linear time series and Markov chains. *Advances in Applied Probability* **22** 587-611.

[27] Wu, G. & Xiao, Z. (2002). A generalized partially linear model of asymmetric volatility. *Journal of Empirical Finance* **9** 287-319.

[28] Yang, D. & Zhang, Q. (2000). Drift-independent volatility estimation based on high, low, open, and close prices. *Journal of Business* **73** 477-491.

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

[30] Zhou, Z. (2010). Nonparametric inference of quantile curves for nonstationary time series. *Annals of Statistics* **38** 2187-2217.