



Estimating nonlinear additive models with nonstationarities and correlated errors

Michael Vogt¹  | Christopher Walsh² 

¹Department of Economics and Hausdorff Center for Mathematics, University of Bonn, Bonn, Germany

²Department of Statistics and Operations Research, University of Vienna, Vienna, Austria

Correspondence

Christopher Walsh, Department of Statistics and Operations Research, University of Vienna, Oskar-Morgenstern-Platz 1, 1090 Vienna, Austria.

Email:

christopher.paul.walsh@univie.ac.at

Abstract

In this paper, we study a nonparametric additive regression model suitable for a wide range of time series applications. Our model includes a periodic component, a deterministic time trend, various component functions of stochastic explanatory variables, and an $AR(p)$ error process that accounts for serial correlation in the regression error. We propose an estimation procedure for the nonparametric component functions and the parameters of the error process based on smooth backfitting and quasimaximum likelihood methods. Our theory establishes convergence rates and the asymptotic normality of our estimators. Moreover, we are able to derive an oracle-type result for the estimators of the AR parameters: Under fairly mild conditions, the limiting distribution of our parameter estimators is the same as when the nonparametric component functions are known. Finally, we illustrate our estimation procedure by applying it to a sample of climate and ozone data collected on the Antarctic Peninsula.

KEYWORDS

correlated errors, nonstationary, semiparametric, smooth backfitting

1 | INTRODUCTION

In many time series applications, the data at hand exhibit seasonal fluctuations and a trending behavior. A common way to incorporate these features is to assume that the data generating

The copyright line for this article was changed on 16 August 2018 after original online publication.

This is an open access article under the terms of the Creative Commons Attribution License, which permits use, distribution and reproduction in any medium, provided the original work is properly cited.

© 2018 The Authors Scandinavian Journal of Statistics published by John Wiley & Sons Ltd on behalf of The Board of the Foundation of the Scandinavian Journal of Statistics

process can be written as the sum of a seasonal part, a deterministic time trend, and a stationary stochastic process. In general, the structure of these three components is largely unknown. This necessitates the development of flexible semiparametric and nonparametric methods in order to estimate them.

Let $\{Y_{t,T} : t = 1, \dots, T\}$ be the time series under investigation. A general semiparametric framework that decomposes $Y_{t,T}$ into a seasonal, a trend, and a stationary stochastic component is given by the regression model

$$Y_{t,T} = m_\theta(t) + m_0\left(\frac{t}{T}\right) + m(X_t) + \varepsilon_t \quad \text{for } t = 1, \dots, T \quad (1)$$

with $\mathbb{E}[\varepsilon_t | X_t] = 0$. Here, m_θ is a periodic function with a known integer period θ and m_0 is a deterministic time trend. The stochastic component consists of the residual ε_t and of the term $m(X_t)$ that captures the influence of the d -dimensional stationary covariate vector $X_t = (X_t^1, \dots, X_t^d)$. We do not impose any parametric restrictions on the component functions m_θ , m_0 , and m . Moreover, we allow for correlation in the error terms ε_t , which are modeled as a stationary AR(p) process. Note that, as usual in nonparametric regression, the time argument of the trend function m_0 is rescaled to the unit interval.

Two special cases of model (1) have been considered in the literature. The fixed design setting $Y_{t,T} = m_0\left(\frac{t}{T}\right) + \varepsilon_t$ was analyzed, for example, in the works of Truong (1991), Altman (1993), Hall and van Keilegom (2003), and Shao and Yang (2011), who provided a variety of methods to estimate the nonparametric trend function m_0 and the AR parameters of the error term. Interestingly, they established an oracle-type result for the estimation of the AR parameters. In particular, they showed that the limiting distribution of the parameter estimators is unaffected by the need to estimate the nonparametric function m_0 . The second special case of model (1) is the setting $Y_t = m(X_t) + \varepsilon_t$. The problem of estimating the AR parameters in this setup has been studied under the restriction that $\{X_t\}$ is independent of the error process $\{\varepsilon_t\}$. Truong and Stone (1994), Schick (1994), and Lin, Pourahmadi, and Schick (1999) showed that, under this restriction, an oracle-type result for the parameter estimators holds analogous to that in the fixed design setting.

In this paper, we study the estimation of the parametric and nonparametric components in the general model (1). We allow $\{X_t\}$ and $\{\varepsilon_t\}$ to be dependent, thus dispensing with the very restrictive assumption that the covariate process is independent of the errors. In order to circumvent the well-known curse of dimensionality, we assume the function m to be additive with component functions m_j for $j = 1, \dots, d$, thus yielding

$$Y_{t,T} = m_\theta(t) + m_0\left(\frac{t}{T}\right) + \sum_{j=1}^d m_j(X_t^j) + \varepsilon_t \quad \text{for } t = 1, \dots, T. \quad (2)$$

A full description of model (2) together with a discussion of its components is given in Section 2.

Our estimation procedure is introduced in Section 3. The nonparametric components m_θ and m_0, \dots, m_d are estimated by extending the smooth backfitting approach of Mammen, Linton, and Nielsen (1999), who derived its asymptotic properties in a strictly stationary setup. Due to the inclusion of the periodic and the deterministic trend components, our model dynamics are no longer stationary. In Sections 3.1 and 3.2, we describe how to incorporate this type of nonstationarity into the smooth backfitting procedure. Given our estimators \tilde{m}_θ and $\tilde{m}_0, \dots, \tilde{m}_d$ of the functions m_θ and m_0, \dots, m_d , we can construct approximate expressions $\tilde{\varepsilon}_t$ of ε_t . Using these, the parameters of the AR(p) error process are estimated via a quasimaximum likelihood-based method, the details of which are given in Section 3.3.

Section 4 contains our results on the asymptotic properties of our estimators. In Sections 4.2 and 4.3, we provide the convergence rates of the nonparametric estimators \tilde{m}_θ and $\tilde{m}_0, \dots, \tilde{m}_d$, as well as their Gaussian limit distribution. The asymptotic behavior of the parameter estimators of the AR(p) error process is studied in Section 4.4. There, we show that the parameter estimators are asymptotically normal. Deriving the limit distribution of the parameter estimators is by far the most difficult part of the theory developed in this paper. To do so, we need to establish a higher-order stochastic expansion of the first derivative of the likelihood function. This requires substantially different and much more intricate techniques than those in the analysis of the special cases previously discussed in the literature.

As we will see, the asymptotic distribution of our parameter estimators in general differs from that of the oracle estimators constructed under the assumption that the additive component functions are known. Thus, the additional uncertainty that stems from estimating the component functions does have an impact on the asymptotic distribution of our parameter estimators in general. However, an oracle-type result can be established under additional conditions on the dependence structure between the covariates X_t and the errors ε_t . In particular, the limit distribution of our parameter estimators coincides with that of the oracle estimators if we additionally assume that $\mathbb{E}[\varepsilon_t | X_{t+k}] = 0$ for all $k = -p, \dots, p$. This assumption is evidently much weaker than imposing independence between $\{X_t\}$ and $\{\varepsilon_t\}$ as in the simpler settings discussed above. Our theory thus generalizes the previously found oracle-type results.

We illustrate our estimation procedure by applying it to monthly minimum temperature and ozone data from the Faraday/Vernadsky research station on the Antarctic Peninsula in Section 5. The nice thing about this application is that Hughes, Subba Rao, and Subba Rao (2007) used a parametric regression model setup with AR errors to analyze the same data. Hence, our analysis can be regarded as a semiparametric extension to their study and we can get an impression of what can be gained by using our more flexible specification in this setting.

2 | MODEL

Before we introduce our estimation procedure, we take a closer look at model (2) and comment on some of its features. We observe a sample of data $\{(Y_{t,T}, X_t) : t = 1, \dots, T\}$, where $Y_{t,T}$ are real-valued random variables and $X_t = (X_t^1, \dots, X_t^d)$ are \mathbb{R}^d -valued random vectors that form a strictly stationary process. As already noted in the introduction, the data are assumed to satisfy the model equation

$$Y_{t,T} = m_\theta(t) + m_0\left(\frac{t}{T}\right) + \sum_{j=1}^d m_j(X_t^j) + \varepsilon_t \quad \text{for } t = 1, \dots, T \quad (3)$$

with $\mathbb{E}[\varepsilon_t | X_t] = 0$, where m_θ is a periodic component with some integer-valued period θ , m_0 is a deterministic trend, and m_j are nonparametric functions of the regressors X_t^j for $j = 1, \dots, d$. For simplicity, we assume that the period θ is known. Methods to estimate θ can, for example, be found in the work of Vogt and Linton (2014). By including the periodic component m_θ and the deterministic trend m_0 , the dynamics of $Y_{t,T}$ depend on time and are thus nonstationary. The errors $\{\varepsilon_t\}$ follow a stationary AR(p) process of the form

$$\varepsilon_t = \sum_{i=1}^p \phi_i^* \varepsilon_{t-i} + \eta_t \quad \text{for all } t \in \mathbb{Z},$$

where $\phi^* = (\phi_1^*, \dots, \phi_p^*)$ is the vector of parameters, the AR order p is assumed to be known, and the residuals η_t form a martingale difference with respect to $\mathcal{F}_t = \{X_t, X_{t-1}, \dots, \varepsilon_{t-1}, \varepsilon_{t-2}, \dots\}$.

The additive functions in model (3) are only identified up to an additive constant. To identify them, we assume that the constant is absorbed into the periodic component and the remaining components have zero mean, that is,

$$\int_0^1 m_0(x_0) dx_0 = 0 \quad \text{and} \quad \int m_j(x_j) p_j(x_j) dx_j = 0 \quad \text{for } j = 1, \dots, d, \quad (4)$$

where p_j is the marginal density of X_t^j . The covariates X_t^j are assumed to take values in a bounded interval that, without loss of generality, is taken to be $[0, 1]$ for each $j = 1, \dots, d$. Throughout this paper, the symbol x_0 is used to denote a point in rescaled time. Moreover, we write $x = (x_0, x_{-0})$ with $x_{-0} = (x_1, \dots, x_d)$.

To be able to do reasonable asymptotics, we let the trend function m_0 in model (3) depend on rescaled time $\frac{t}{T}$ rather than on real time t . If we defined m_0 in terms of real time, we would not get additional information on the structure of m_0 locally around a fixed time point t as the sample size increases. Within the framework of rescaled time, in contrast, the function m_0 is observed on a finer and finer grid of rescaled time points on the unit interval as T grows. Thus, we obtain more and more information on the local structure of m_0 around each point in rescaled time. This is the reason why we can make reasonable asymptotic considerations within this framework.

In contrast to m_0 , we let the periodic component m_θ in model (3) be a function of real time t . This allows us to exploit its periodic character when doing asymptotics: Assume that we want to estimate m_θ at a time point $t_\theta \in \{1, \dots, \theta\}$. As m_θ is periodic, it has the same value at $t_\theta, t_\theta + \theta, t_\theta + 2\theta, t_\theta + 3\theta$, and so on. Hence, if m_θ depends on real time t , the number of time points in our sample at which m_θ has the value $m_\theta(t_\theta)$ increases as the sample size grows. This gives us more and more information about the value $m_\theta(t_\theta)$ and thus allows us to do asymptotics.

3 | ESTIMATION PROCEDURE

We now describe how the various components of model (3) are estimated. Our procedure consists of three steps. In the first step, the periodic model component m_θ is estimated. The estimation of the nonparametric functions m_0, \dots, m_d is addressed in the second step. Finally, we use the estimators of the additive component functions to construct estimators of the AR parameters.

3.1 | Estimation of m_θ

For any time point $t = 1, \dots, T$, let $t_\theta = t - \lfloor \frac{t-1}{\theta} \rfloor \theta$ with $\lfloor x \rfloor$ denoting the largest integer that is smaller than or equal to x . Our estimator of the periodic component m_θ is defined as

$$\tilde{m}_\theta(t) = \frac{1}{K_{t_\theta, T}} \sum_{k=1}^{K_{t_\theta, T}} Y_{t_\theta + (k-1)\theta, T} \quad \text{for } t = 1, \dots, T, \quad (5)$$

where $K_{t_\theta, T} = 1 + \lfloor \frac{T-t_\theta}{\theta} \rfloor$ is the number of observations that satisfy $t = t_\theta + k\theta$ for some $k \in \mathbb{N}$. The estimator has a very simple structure. It is the empirical mean of observations that are separated by a multiple of θ time points. Later on, we will show that \tilde{m}_θ is asymptotically normal. Note that this result is robust to the presence of the deterministic trend function m_0 . In particular, we will see that the effect of the unknown time trend m_0 on the estimator \tilde{m}_θ can be asymptotically neglected.

3.2 | Estimation of m_0, \dots, m_d

We next introduce the estimators of the functions m_0, \dots, m_d . For the time being, let us assume that the periodic component m_θ is known. Later on, m_θ will be replaced by its estimator \tilde{m}_θ . Given that m_θ is known, $Z_{t,T} = Y_{t,T} - m_\theta(t)$ is observable. This allows us to rewrite model (3) as

$$Z_{t,T} = m_0\left(\frac{t}{T}\right) + \sum_{j=1}^d m_j(X_t^j) + \varepsilon_t. \quad (6)$$

In order to estimate the functions m_0, \dots, m_d in (6), we extend the smooth backfitting approach of Mammen et al. (1999), who derived the asymptotic properties of this approach in a strictly stationary setup. Model (6) involves a deterministic time-trend component, which makes the model dynamics nonstationary. In what follows, we describe how to modify the smooth backfitting procedure to allow for this type of nonstationarity.

To do so, we first introduce the auxiliary estimators

$$\begin{aligned} \hat{q}(x) &= \frac{1}{T} \sum_{t=1}^T K_h\left(x_0, \frac{t}{T}\right) \prod_{k=1}^d K_h(x_k, X_t^k) \\ \hat{m}(x) &= \frac{1}{T} \sum_{t=1}^T K_h\left(x_0, \frac{t}{T}\right) \prod_{k=1}^d K_h(x_k, X_t^k) Z_{t,T} / \hat{q}(x). \end{aligned}$$

$\hat{q}(x)$ is a kernel estimator of the density $q(x) := I(x_0 \in [0, 1])p(x_{-0})$ with p being the joint density of the regressors $X_t = (X_t^1, \dots, X_t^d)$. Moreover, $\hat{m}(x)$ is a $(d + 1)$ -dimensional Nadaraya–Watson estimator of the regression function $m(x) = m_0(x_0) + \dots + m_d(x_d)$. In these definitions,

$$K_h(v, w) = \frac{K_h(v - w)}{\int_0^1 K_h(s - w) ds}$$

is a modified kernel weight, where h denotes the bandwidth, $K_h(v) = \frac{1}{h} K\left(\frac{v}{h}\right)$, and the kernel function $K(\cdot)$ integrates to one. These weights have the property that $\int_0^1 K_h(v, w) dv = 1$ for all w , which is needed to derive the asymptotic results for the backfitting estimators.

Given the smoothers \hat{q} and \hat{m} , we define the smooth backfitting estimators $\tilde{m}_0, \dots, \tilde{m}_d$ as the minimizers of the criterion

$$\int_{[0,1]^{d+1}} (\hat{m}(x) - g_0(x_0) - \dots - g_d(x_d))^2 \hat{q}(x) dx, \quad (7)$$

where the minimization runs over all additive functions $g(x) = g_0(x_0) + \dots + g_d(x_d)$ whose components satisfy $\int_0^1 g_j(x_j) \hat{p}_j(x_j) dx_j = 0$ for $j = 0, \dots, d$. Here, \hat{p}_j is a kernel estimator of p_j for $j = 0, \dots, d$, where we define $p_0(x_0) = I(x_0 \in [0, 1])$. Explicit expressions for these estimators are given below in (9) and (12).

According to the definition in (7), the backfitting estimator $\tilde{m} = \tilde{m}_0 + \dots + \tilde{m}_d$ is an L^2 -projection of the $(d + 1)$ -dimensional Nadaraya–Watson smoother \hat{m} onto the space of additive functions with respect to the density \hat{q} . Rescaled time is treated as an additional component in this projection. In particular, note that \hat{q} estimates the product of a uniform density over $[0, 1]$ and the density p of the regressors X_t . This shows that rescaled time is treated in a similar way to an additional stochastic regressor that is uniformly distributed over $[0, 1]$ and independent of the variables X_t . The heuristic idea behind this is the following. Firstly, as the variables X_t are strictly stationary, their distribution is time-invariant. In this sense, their stochastic behavior is

independent of rescaled time $\frac{t}{T}$. Thus, rescaled time behaves similarly to an additional stochastic variable that is independent of X_t . Secondly, as the points $\frac{t}{T}$ are evenly spaced over the unit interval, a variable with a uniform distribution closely replicates the pattern of rescaled time.

By differentiation, we can show that the solution to the projection problem (7) is characterized by the system of integral equations

$$\tilde{m}_j(x_j) = \hat{m}_j(x_j) - \sum_{k \neq j} \int_0^1 \tilde{m}_k(x_k) \frac{\hat{p}_{k,j}(x_k, x_j)}{\hat{p}_j(x_j)} dx_k - \tilde{m}_c \quad (8)$$

with $\int_0^1 \tilde{m}_j(x_j) \hat{p}_j(x_j) dx_j = 0$ for $j = 0, \dots, d$. As we do not observe the variables $Z_{t,T} = Y_{t,T} - m_\theta(t)$, we define the kernel estimators in (8) in terms of the approximations $\tilde{Z}_{t,T} = Y_{t,T} - \tilde{m}_\theta(t)$. In particular, we let

$$\hat{p}_j(x_j) = \frac{1}{T} \sum_{t=1}^T K_h(x_j, X_t^j) \quad (9)$$

$$\hat{p}_{j,k}(x_j, x_k) = \frac{1}{T} \sum_{t=1}^T K_h(x_j, X_t^j) K_h(x_k, X_t^k) \quad (10)$$

$$\hat{m}_j(x_j) = \frac{1}{T} \sum_{t=1}^T K_h(x_j, X_t^j) \tilde{Z}_{t,T} / \hat{p}_j(x_j) \quad (11)$$

for $j, k = 1, \dots, d$ with $j \neq k$, where \hat{p}_j is the one-dimensional kernel density estimator of the marginal density p_j of X_t^j , $\hat{p}_{j,k}$ is the two-dimensional kernel density estimator of the joint density $p_{j,k}$ of (X_t^j, X_t^k) , and \hat{m}_j is a one-dimensional Nadaraya–Watson smoother. Moreover,

$$\hat{p}_0(x_0) = \frac{1}{T} \sum_{t=1}^T K_h\left(x_0, \frac{t}{T}\right) \quad (12)$$

$$\hat{p}_{0,k}(x_0, x_k) = \frac{1}{T} \sum_{t=1}^T K_h\left(x_0, \frac{t}{T}\right) K_h(x_k, X_t^k) \quad (13)$$

$$\hat{m}_0(x_0) = \frac{1}{T} \sum_{t=1}^T K_h\left(x_0, \frac{t}{T}\right) \tilde{Z}_{t,T} / \hat{p}_0(x_0) \quad (14)$$

for $k = 1, \dots, d$ and $\tilde{m}_c = \frac{1}{T} \sum_{t=1}^T \tilde{Z}_{t,T}$. Note that it would be more natural to define $\hat{p}_0(x_0) = I(x_0 \in [0, 1])$, as we already know the “true density” of rescaled time. However, for technical reasons, we set $\hat{p}_0(x_0) = \frac{1}{T} \sum_{t=1}^T K_h\left(x_0, \frac{t}{T}\right)$. This creates a behavior of the estimator \hat{p}_0 in the boundary region of the support $[0, 1]$ analogous to that of \hat{p}_j at the boundary. Alternatively, we could define $\hat{p}_0(x_0) = \int_0^1 K_h(x_0, v) dv$. (Note that $\int_0^1 K_h(x_0, v) dv = 1$ for $x_0 \in [2C_1h, 1 - 2C_1h]$, where $[-C_1, C_1]$ is the support of the kernel function K .) Moreover, we could set $\hat{p}_{0,k}(x_0, x_k) = \hat{p}_0(x_0) \hat{p}_k(x_k)$, thereby exploiting the “independence” of rescaled time and the other regressors.

In our theoretical analysis, we work with the smooth backfitting estimators characterized as the solution to the system of integral equations (8). In general, however, the system of equations (8) cannot be solved analytically. Nevertheless, the solution can be approximated by a backfitting algorithm that converges for arbitrary starting values. The algorithm can be

summarized as follows. Let $\tilde{m}_j^{[0]}$ be starting values for $j = 0, \dots, d$. In the r th iteration step, one cycles through all the components $j = 0, \dots, d$ and computes

$$\tilde{m}_j^{[r]}(x_j) = \hat{m}_j(x_j) - \sum_{k < j} \int_0^1 \tilde{m}_k^{[r]}(x_k) \frac{\hat{p}_{k,j}(x_k, x_j)}{\hat{p}_j(x_j)} dx_k - \sum_{k > j} \int_0^1 \tilde{m}_k^{[r-1]}(x_k) \frac{\hat{p}_{k,j}(x_k, x_j)}{\hat{p}_j(x_j)} dx_k - \tilde{m}_c$$

for each j . In the work of Mammen et al. (1999), the asymptotic properties of this algorithm are established under very general conditions that are implied by our assumptions in Section 4.1. Under these general conditions, it can be shown that, with probability tending to 1,

$$\int_0^1 \left[\tilde{m}_j^{[r]}(x_j) - \tilde{m}_j(x_j) \right]^2 p_j(x_j) dx_j \leq c\gamma^{2r} \left(1 + \sum_{j=0}^d \int_0^1 \left\{ \tilde{m}_j^{[0]}(x_j) \right\}^2 p_j(x_j) dx_j \right)$$

with some constants $0 < \gamma < 1$ and $c > 0$. Hence, $\tilde{m}_j^{[r]}$ converges to \tilde{m}_j at the order $O(\gamma^{2r})$ in an L_2 -sense. In practice, the backfitting algorithm is iterated until some convergence criterion is satisfied. In the empirical part of this paper, we work with the following stopping rule. The algorithm terminates either after a maximum of 50 iterations or if, for all directions $j = 0, \dots, d$,

$$\frac{\int \left[\tilde{m}_j^{[r]}(x_j) - \tilde{m}_j^{[r-1]}(x_j) \right]^2 dx_j}{\int \left[\tilde{m}_j^{[r-1]}(x_j) \right]^2 dx_j + \delta} \leq \delta,$$

where δ is a small number that is set to 10^{-5} in our code.

Our smooth backfitting estimators are based on Nadaraya–Watson pilot smoothers. Alternatively, local linear smoothers could be used. Throughout this paper, we restrict attention to Nadaraya–Watson-based smooth backfitting because the derivations and the notation would become even more involved in the local linear case. From a theoretical point of view, local linear smoothers have the advantage that they are design adaptive and, thus, do not suffer from boundary problems. As shown in the work of Mammen et al. (1999) in the strictly stationary case, this advantage carries over to the local linear based smooth backfitting estimators; see theorem 4' in the work of Mammen et al. (1999) and the discussion thereafter. In applications where forecasts are performed, a particular interest lies in the value of the trend function m_0 at the rescaled time point $u = 1$. Hence, a good boundary behavior is particularly important for the trend component m_0 . To reduce the boundary bias in time direction, one could use a local linear pilot smoother in time direction while working with Nadaraya–Watson or local linear smoothers in the other directions.

3.3 | Estimation of the AR parameters

To motivate the third step in our estimation procedure, we shall initially consider an infeasible estimator of the AR parameters. Suppose that the functions $m_\theta, m_0, \dots, m_d$ were known. In this situation, the AR(p) error process $\{\varepsilon_t\}$ would be observable because $\varepsilon_t = Y_{t,T} - m_\theta(t) - m_0\left(\frac{t}{T}\right) - \sum_{j=1}^d m_j(X_t^j)$. The parameters $\phi^* = (\phi_1^*, \dots, \phi_p^*)$ of the error process could thus be estimated by standard maximum likelihood methods. In particular, we could use a conditional maximum likelihood estimator of the form

$$\hat{\phi} = \arg \max_{\phi \in \Phi} l_T(\phi), \quad (15)$$

where Φ is a compact parameter space and l_T is the conditional log-likelihood given by

$$l_T(\phi) = - \sum_{t=p+1}^T (\varepsilon_t - \varepsilon_t(\phi))^2 \quad (16)$$

with $\varepsilon_t(\phi) = \sum_{i=1}^p \phi_i \varepsilon_{t-i}$. Note that $\hat{\phi}$ has a closed-form solution that is identical to the usual least squares estimator. We will, however, not work with this closed-form solution in what follows. Instead, we will formulate our proofs in terms of the likelihood function. This makes it easier to apply our arguments to other error structures such as ARCH processes. We give some comments on how to extend our approach in this direction in Section 6.

As the functions $m_\theta, m_0, \dots, m_d$ are not observed, we cannot use the standard approach from above directly. However, given the estimators $\tilde{m}_\theta, \tilde{m}_0, \dots, \tilde{m}_d$ from the previous estimation steps, we can replace the error terms ε_t by the approximations

$$\tilde{\varepsilon}_t = Y_{t,T} - \tilde{m}_\theta(t) - \tilde{m}_0\left(\frac{t}{T}\right) - \sum_{j=1}^d \tilde{m}_j(X_t^j) \quad (17)$$

in the maximum likelihood estimation. The log-likelihood then becomes

$$\tilde{l}_T(\phi) = - \sum_{t=p+1}^T (\tilde{\varepsilon}_t - \tilde{\varepsilon}_t(\phi))^2 \quad (18)$$

with $\tilde{\varepsilon}_t(\phi) = \sum_{i=1}^p \phi_i \tilde{\varepsilon}_{t-i}$. Our estimator $\tilde{\phi}$ of the true parameter values ϕ^* is now defined as

$$\tilde{\phi} = \arg \max_{\phi \in \Phi} \tilde{l}_T(\phi). \quad (19)$$

4 | ASYMPTOTICS

In this section, we analyze the asymptotic properties of our estimators. The first subsection lists the assumptions required for our analysis. The following subsections describe the main asymptotic results, with each subsection dealing with a separate step of our estimation procedure.

4.1 | Assumptions

To derive the asymptotic properties of the estimators $\tilde{m}_\theta, \tilde{m}_0, \dots, \tilde{m}_d$, the following assumptions are needed:

- (H1) *The process $\{(X_t, \varepsilon_t) : t = 1, \dots, T\}$ is strictly stationary and strongly mixing with mixing coefficients α satisfying $\alpha(k) \leq a^k$ for some $0 < a < 1$.*
- (H2) *The variables X_t have compact support, which w.l.o.g. equals $[0, 1]^d$. The density p of X_t and the densities $p_{(0,l)}$ of (X_t, X_{t+l}) , $l = 1, 2, \dots$, are uniformly bounded. Furthermore, p is bounded away from zero on $[0, 1]^d$.*
- (H3) *The functions m_0 and m_j ($j = 1, \dots, d$) are twice continuously differentiable with Lipschitz-continuous second derivatives. The first partial derivatives of p exist and are continuous.*
- (H4) *The kernel K is bounded, symmetric about zero and has compact support $[-C_1, C_1]$. Moreover, it fulfills the Lipschitz condition that there exists a positive constant L with $|K(u) - K(v)| \leq L|u - v|$.*
- (H5) *There exist a real constant C and a natural number l^* such that $\mathbb{E}[|\varepsilon_t|^\rho | X_t] \leq C$ for some $\rho > \frac{8}{3}$ and $\mathbb{E}[|\varepsilon_t \varepsilon_{t+l}| | X_t, X_{t+l}] \leq C$ for all $l \geq l^*$.*

(H6) The bandwidth h satisfies either of the following:

- (a) $T^{\frac{1}{5}}h \rightarrow c_b$ for some constant $c_b > 0$.
- (b) $T^\lambda h \rightarrow c_b$ for some constant $c_b > 0$ and some $\lambda \in \left(\frac{1}{4}, \frac{1}{3}\right)$.

The above assumptions are very similar to the standard conditions for smooth backfitting estimators to be found, for example, in the works of Mammen et al. (1999), Mammen and Park (2006), or Yu, Mammen, and Park (2011). Note that we do not necessarily require exponentially decaying mixing rates, as assumed in (H1). These could alternatively be replaced by sufficiently high polynomial rates. We nevertheless make the stronger assumption (H1) to keep the notation and structure of the proofs as clear as possible. In (H6), we impose two alternative conditions on the bandwidth h . In (H6a), h is assumed to be of the order $T^{-1/5}$, which is optimal for estimating the additive component functions m_0, \dots, m_d . In (H6b), we assume h to be of the order $T^{-\lambda}$ with some $\lambda \in \left(\frac{1}{4}, \frac{1}{3}\right)$ and, thus, undersmooth the backfitting estimators of m_0, \dots, m_d . Undersmoothing is needed to estimate the AR coefficients in the error term. It makes sure that the bias parts of the backfitting estimators can be asymptotically neglected and do not influence the limit distribution of the AR parameter estimators. Assumption (H6b) parallels standard undersmoothing conditions imposed in the context of semiparametric estimation problems.

In order to show that the estimators of the AR parameters are consistent and asymptotically normal, we additionally require the following assumptions:

- (H7) The parameter space Φ is a compact subset of $\{\phi \in \mathbb{R}^p \mid 1 - \phi_1 z - \dots - \phi_p z^p \neq 0 \text{ for all complex } z \text{ with } |z| \leq 1 \text{ and } \phi_p \neq 0\}$. The true parameter vector $\phi^* = (\phi_1^*, \dots, \phi_p^*)$ is an interior point of Φ .
- (H8) $\mathbb{E}[\varepsilon_t^{4+\delta}] < \infty$, for some $\delta > 0$.
- (H9) There exist a real constant C and a natural number l^* such that $\mathbb{E}[|\varepsilon_t| \mid X_{t+k}] \leq C$ and $\mathbb{E}[|\varepsilon_t \varepsilon_{t+l}| \mid X_{t+k}, X_{t+l}] \leq C$ for all l with $|l| \geq l^*$ and $k = -p, \dots, p$.

The compactness assumption in (H7) is required for the proof of consistency. (H8) and (H9) are technical assumptions needed to show asymptotic normality.

4.2 | Asymptotics for \tilde{m}_θ

We start by considering the asymptotic behavior of the estimator \tilde{m}_θ . The next theorem shows that $\sqrt{\frac{T}{\theta}}(\tilde{m}_\theta(t) - m_\theta(t))$ is asymptotically normal for each fixed time point t .

Theorem 1. Let (H1) and (H3) be fulfilled and assume that $\mathbb{E}|\varepsilon_t|^\rho < \infty$ for some $\rho > 2$. Then,

$$\sqrt{\frac{T}{\theta}}(\tilde{m}_\theta(t) - m_\theta(t)) \xrightarrow{d} N(0, V_\theta)$$

for any $t = 1, \dots, T$, where $V_\theta = \sum_{k=-\infty}^{\infty} \text{Cov}(W_0, W_{k\theta})$ with $W_t = Y_{t,T} - m_\theta(t) - m_0\left(\frac{t}{T}\right) = \sum_{j=1}^d m_j(X_t^j) + \varepsilon_t$.

As \tilde{m}_θ and m_θ are periodic and the period θ is a fixed constant that does not depend on the sample size T , Theorem 1 immediately implies that

$$\sup_{t=1, \dots, T} |\tilde{m}_\theta(t) - m_\theta(t)| = \sup_{t=1, \dots, \theta} |\tilde{m}_\theta(t) - m_\theta(t)| = O_p\left(\sqrt{\frac{\theta}{T}}\right) = O_p\left(\frac{1}{\sqrt{T}}\right).$$

The proof of Theorem 1 is straightforward: We have

$$\begin{aligned}\tilde{m}_\theta(t) - m_\theta(t) &= \frac{1}{K_{t_\theta, T}} \sum_{k=1}^{K_{t_\theta, T}} m_0 \left(\frac{t_\theta + (k-1)\theta}{T} \right) + \frac{1}{K_{t_\theta, T}} \sum_{k=1}^{K_{t_\theta, T}} W_{t_\theta + (k-1)\theta} \\ &=: (A) + (B).\end{aligned}$$

The term (A) approximates the integral $\int_0^1 m_0(u)du$. It is easily seen that the convergence rate is $O\left(\frac{\theta}{T}\right)$. As $\int_0^1 m_0(u)du = 0$ by the normalization in (4), we obtain that (A) is of the order $O\left(\frac{\theta}{T}\right)$ and can thus be asymptotically neglected. Noting that $\{W_t\}$ is mixing by (H1) and has mean zero by our normalization, we can now apply a central limit theorem for mixing variables to the term (B) to get the normality result of Theorem 1.

The statement of Theorem 1 is derived under the assumption that the period θ is a fixed constant that does not depend on the sample size T . Heuristically speaking, we thus assume that θ is small compared with T . To better take into account the case where θ is not so small in comparison with T , we could allow θ to grow with T , that is, $\theta = \theta_T$. Because $\tilde{m}_\theta(t)$ is nothing else than an empirical average of the observations $Y_{t_\theta, T}, Y_{t_\theta + \theta, T}, Y_{t_\theta + 2\theta, T}, \dots$, the effective sample size for estimating $m_\theta(t)$ is T/θ . This implies that, for each fixed t , $\tilde{m}_\theta(t)$ converges to $m_\theta(t)$ at the rate $O_p(\sqrt{\theta/T})$. The faster $\theta = \theta_T$ grows with T , the slower this rate becomes. As regards to our theory, we can allow $\theta = \theta_T$ to grow with the sample size T as long as it does not diverge too quickly. In particular, our main theorems (Theorems 2, 3, and 4) and Corollary 1 remain to hold true as long as θ_T does not grow too fast. In the sequel, we stick to the assumption that θ is a fixed constant in order to keep the technical arguments as clear as possible.

4.3 | Asymptotics for $\tilde{m}_0, \dots, \tilde{m}_d$

The main result of this subsection characterizes the limiting behavior of the smooth backfitting estimators $\tilde{m}_0, \dots, \tilde{m}_d$. It shows that the estimators converge uniformly to the true component functions at the one-dimensional nonparametric rate no matter how large the dimension d of the regressors. Moreover, it characterizes the asymptotic distribution of the estimators.

Theorem 2. Suppose that conditions (H1)–(H5) hold.

- (a) Assume that the bandwidth h satisfies (H6a) or (H6b). Then, for $I_h = [2C_1h, 1 - 2C_1h]$ and $I_h^c = [0, 2C_1h) \cup (1 - 2C_1h, 1]$,

$$\sup_{x_j \in I_h} |\tilde{m}_j(x_j) - m_j(x_j)| = O_p \left(\sqrt{\frac{\log T}{Th}} \right) \quad (20)$$

$$\sup_{x_j \in I_h^c} |\tilde{m}_j(x_j) - m_j(x_j)| = O_p(h) \quad (21)$$

for all $j = 0, \dots, d$.

- (b) Assume that the bandwidth h satisfies (H6a). Then, for any $x_0, \dots, x_d \in (0, 1)$,

$$T^{\frac{2}{5}} \begin{bmatrix} \tilde{m}_0(x_0) - m_0(x_0) \\ \vdots \\ \tilde{m}_d(x_d) - m_d(x_d) \end{bmatrix} \xrightarrow{d} N(B(x), V(x))$$

with the bias term $B(x) = [c_b^2(\beta_0(x_0) - \gamma_0), \dots, c_b^2(\beta_d(x_d) - \gamma_d)]'$ and the covariance matrix $V(x) = \text{diag}(v_0(x_0), \dots, v_d(x_d))$. Here, $v_0(x_0) = c_b^{-1}c_K \sum_{l=-\infty}^{\infty} \gamma_\epsilon(l)$ and $v_j(x_j) = c_b^{-1}c_K \sigma_j^2(x_j)/p_j(x_j)$ for $j = 1, \dots, d$ with $\gamma_\epsilon(l) = \text{Cov}(\epsilon_t, \epsilon_{t+l})$, $\sigma_j^2(x_j) = \text{Var}(\epsilon_t | X_t^j = x_j)$

and the constants $c_b = \lim_{T \rightarrow \infty} T^{1/5} h$ and $c_K = \int K^2(u) du$. Furthermore, the functions β_j are the components of the $L^2(q)$ -projection of the function β defined in Lemma 3 in Appendix A onto the space of additive functions satisfying $\int \beta_j(x_j) p_j(x_j) dx_j = 0$. Finally, the constants γ_j can be characterized by the equation $\int_0^1 \alpha_{T,j}(x_j) \hat{p}_j(x_j) dx_j = h^2 \gamma_j + o_p(h^2)$ for $j = 0, \dots, d$, with $\alpha_{T,j}$ also given in Lemma 3 in Appendix A.

As described in Section 3.2, rescaled time $\frac{t}{T}$ behaves similarly to an additional uniformly distributed regressor that is independent of the other regressors. This consideration allows us to derive the above result by extending the proving strategy of Mammen et al. (1999). The details are given in Appendix B.

4.4 | Asymptotics for the AR parameter estimators

We finally establish the asymptotic properties of our estimator $\tilde{\phi}$ of the AR parameters ϕ^* . The technical details can be found in Appendix C. The first theorem shows that $\tilde{\phi}$ is consistent.

Theorem 3. *Suppose that the bandwidth h satisfies (H6a) or (H6b). In addition, let assumptions (H1)–(H5) and (H7) be fulfilled. Then, $\tilde{\phi}$ is a consistent estimator of ϕ^* , that is, $\tilde{\phi} \xrightarrow{P} \phi^*$.*

The central result of our theory specifies the limiting distribution of $\tilde{\phi}$.

Theorem 4. *Suppose that the bandwidth h satisfies (H6b) and let assumptions (H1)–(H5) together with (H7)–(H9) be fulfilled. Then, it holds that*

$$\sqrt{T}(\tilde{\phi} - \phi^*) \xrightarrow{d} N(0, V^*)$$

with

$$V^* = \Gamma_p^{-1}(W + \Omega)\Gamma_p^{-1}.$$

Here, Γ_p is the autocovariance matrix of the AR(p) process $\{\varepsilon_t\}$, that is, $\Gamma_p = (\gamma(i - j))_{i,j=1,\dots,p}$ with $\gamma(i - j) = \mathbb{E}[\varepsilon_0 \varepsilon_{i-j}]$. Moreover, $W = (\mathbb{E}[\eta_0^2 \varepsilon_{-i} \varepsilon_{-j}])_{i,j=1,\dots,p}$ and the matrix Ω is defined in Equation (C15) in Appendix C.

Consider for a moment the case in which the functions m_θ and m_0, \dots, m_d are known. In this case, we can use the “oracle” estimator $\hat{\phi}$ defined in (15) to estimate the AR parameters ϕ^* . Standard theory tells us that $\hat{\phi}$ is asymptotically normal with asymptotic variance $\Gamma_p^{-1} W \Gamma_p^{-1}$. Theorem 4 thus shows that, in general, the limiting distribution of our estimator $\tilde{\phi}$ differs from that of the oracle estimator. Note that this difference does not merely result from the fact that the functions m_0, \dots, m_d are nonparametric. Even if they are parametric, $\tilde{\phi}$ will in general have a different distribution than the oracle estimator $\hat{\phi}$.

Even though different in general, the asymptotic distributions of $\tilde{\phi}$ and $\hat{\phi}$ are the same in a wide range of cases. This oracle-type result is stated in the following corollary.

Corollary 1. *Suppose that all the assumptions of Theorem 4 are fulfilled and that $\mathbb{E}[\varepsilon_t | X_{t+k}] = 0$ for all $k = -p, \dots, p$. Then,*

$$\sqrt{T}(\tilde{\phi} - \phi^*) \xrightarrow{d} N(0, \Gamma_p^{-1} W \Gamma_p^{-1}).$$

Corollary 1 follows directly from the proof of Theorem 4: Inspecting the functions defined in Lemma 4 in Appendix C and realizing that they are constantly zero under the assumptions of the corollary, the matrix Ω is immediately seen to be equal to zero as well. The corollary shows that the oracle result holds under fairly mild conditions on the dependence structure between

X_t and ε_t , in particular, under much weaker conditions than independence of the processes $\{X_t\}$ and $\{\varepsilon_t\}$. To give an example where the conditions of the corollary are satisfied but where the processes $\{X_t\}$ and $\{\varepsilon_t\}$ are not independent, consider the following. Let the AR residuals be given by $\varepsilon_t = \sum_{i=1}^p \phi_i^* \varepsilon_{t-i} + \eta_t$ with $\eta_t = \sigma(X_t) \xi_t$, where σ is a continuous volatility function and $\{\xi_t\}$ is a process of zero-mean i.i.d. variables that is independent of $\{X_t\}$. A simple argument shows that $\mathbb{E}[\varepsilon_t | \{X_t\}] = 0$ in this case, that is, the assumptions of the corollary are satisfied. Moreover, it is easily seen that the processes $\{X_t\}$ and $\{\varepsilon_t\}$ are not independent given that the function σ is nonconstant.

Note that our theory re-establishes the oracle result derived in the simpler setup without stochastic covariates, that is, in the model

$$Y_{t,T} = m_\theta(t) + m_0\left(\frac{t}{T}\right) + \varepsilon_t \quad \text{for } t = 1, \dots, T \quad (22)$$

with $\mathbb{E}[\varepsilon_t] = 0$. In this case, the periodic component can be estimated as described in Section 3.1. Moreover, we can use a Nadaraya–Watson smoother of the form (14) to approximate the trend component m_0 . A vastly simplified version of the proof for Theorem 4 shows that the limiting distribution of the AR parameter estimators is identical to that of the oracle estimators in this setting. In particular, the stochastic higher-order expansion derived in Lemma 4 is not required any more. The arguments of the much simpler Lemma 5 in Appendix C are sufficient to derive the result. To understand the main technical reasons why the argument simplifies so substantially, we refer the reader to the remarks given after the proof of Lemma 5.

The normality results of Theorem 4 and Corollary 1 enable us to calculate confidence bands for the AR parameter estimators and to conduct inference based on these. To do so, we need a consistent estimator of the asymptotic variance of $\tilde{\phi}$. Whereas such an estimator is easily obtained under the conditions of Corollary 1, it is not at all trivial to derive a consistent estimator of V^* in Theorem 4. This is due to the very complicated structure of the matrix Ω , which involves functions obtained from a higher-order expansion of the stochastic part of the backfitting estimators (see Theorem 5 in Appendix B). To circumvent these difficulties, one may try to set up a bootstrap approach to estimate confidence bands and to do testing. The normality result of Theorem 4 could be used as a starting point to derive consistency results for such a bootstrap procedure. However, this is beyond the scope of this paper and a substantial project in itself.

5 | APPLICATION

In this section, we apply our estimation procedure to a set of monthly temperature and ozone data from the Faraday/Vernadsky research station on the Antarctic Peninsula. The data can be found online as supporting information. Alternatively, it is available on request from the British Antarctic Survey, Cambridge. A strong warming trend has been identified on the peninsula over the past 50 years. In particular, the monthly mean temperatures at Faraday station have considerably increased during this time, as pointed out, for example, by Turner, King, Lachlan-Cope, and Jones (2002) and Turner et al. (2005).

5.1 | Modeling approach

We closely follow the analysis of Hughes et al. (2007) as our model can be seen as a semiparametric extension to their approach. According to Hughes et al. (2007), the rise of the mean monthly

temperature is mostly due to an increase in the minimum monthly temperature. They argued that, to understand and quantify the warming on the peninsula, an appropriate statistical model of the minimum temperature is called for. Following their lead, we focus on modeling the minimum temperature and consider stratospheric ozone as a potential explanatory variable.

The data used in our analysis are plotted in Figure 1. The upper panel contains the monthly minimum near-surface temperatures at Faraday station from September 1957 to December 2004, whereas the lower panel shows the monthly level of stratospheric ozone concentration measured in Dobson units over the same period. For more information on the data, consult the work of Hughes et al. (2007), where a detailed description of them can be found.

Hughes et al. (2007) proposed a parametric model with a linear trend and a parametrically specified periodic component with a period of 12 months to fit the temperature and ozone data. Their baseline model is given by the equation

$$Y_t = a_0 + a_1 \sin\left(\frac{2\pi}{12}t\right) + a_2 \cos\left(\frac{2\pi}{12}t\right) + a_3t + \varepsilon_t, \quad (23)$$

where Y_t denotes the minimum monthly temperature and $a = (a_0, \dots, a_3)$ is a vector of parameters. In addition, they considered the extended model

$$Y_t = a_0 + a_1 \sin\left(\frac{2\pi}{12}t\right) + a_2 \cos\left(\frac{2\pi}{12}t\right) + a_3t + a_4X_{t-1} + \varepsilon_t, \quad (24)$$

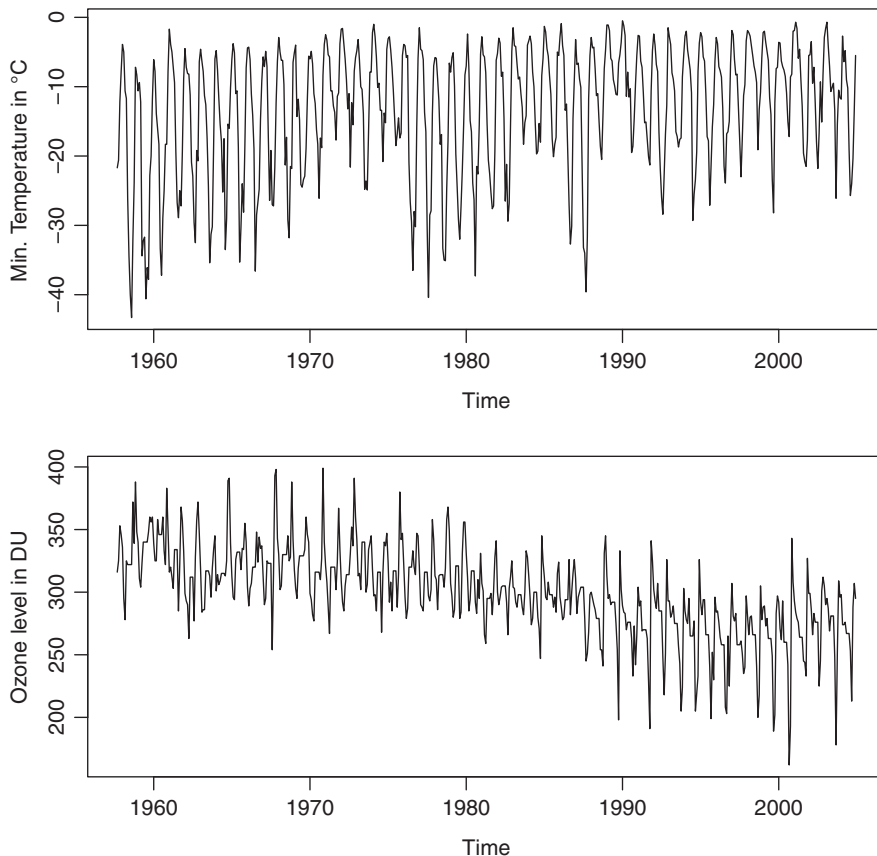


FIGURE 1 The upper panel shows the monthly minimum near-surface temperatures (in degree Celsius) and the lower one shows the monthly stratospheric ozone concentrations (in Dobson units) at Faraday station

where the covariate X_{t-1} , denoting the lagged detrended and deseasonalized ozone concentration, enters linearly. In their analysis, they found a strong linear upward trend in the minimum monthly temperature. Moreover, they observed considerable autocorrelation in the residuals ε_t and proposed an AR process to model them. Using an order selection criterion, they found an AR(1) model to be most suitable, which fits nicely with the preference for AR(1) errors when using discrete time series to model climate data, as mentioned in Mudelsee (2010).

We now introduce a framework that can be regarded as a semiparametric extension to the parametric models (23) and (24). Our baseline model is given by

$$Y_{t,T} = m_\theta(t) + m_0\left(\frac{t}{T}\right) + \varepsilon_t \quad \text{for } t = 1, \dots, T, \quad (25)$$

where $Y_{t,T}$ are minimum monthly temperatures, m_θ is a seasonal component, and m_0 is a non-parametric time trend. We additionally consider an extended version of (25) having the form

$$Y_{t,T} = m_\theta(t) + m_0\left(\frac{t}{T}\right) + m_1(X_{t-1}) + \varepsilon_t \quad \text{for } t = 1, \dots, T, \quad (26)$$

where, as before, the variables X_{t-1} denote lagged monthly stratospheric ozone concentration levels that have been detrended and deseasonalized. The additive functions m_θ , m_0 , and m_1 in the above two models are normalized as described in (4). Following the work of Hughes et al. (2007), we assume the variables ε_t to have an AR(1) structure and allow for the minimum monthly temperature to have a 12-month cycle by setting $\theta = 12$.

Before giving our estimates, we provide the preferred fits of the models (23) and (24) given in the work of Hughes et al. (2007) in order to compare our estimates to theirs. Their models are fitted using observations up until and including December 2003. For the model (23), their preferred fit is

$$Y_t = 6.25 \sin\left(\frac{2\pi}{12}t\right) + 6.95 \cos\left(\frac{2\pi}{12}t\right) + 0.0105t + \varepsilon_t \quad (27)$$

with $\varepsilon_t = 0.566\varepsilon_{t-1} + \eta_t$ and η_t distributed as conv GEV($-0.109, -5.71, 3.65$), where “conv GEV” stands for converse generalized extreme value. A conv GEV(γ, μ, σ) random variable Z has the distribution function $P(Z \leq z) = 1 - \exp\{[1 + \frac{\gamma}{\sigma}(\mu - z)]^{-\frac{1}{\gamma}}\}$. Their preferred fit for the model in (24) is

$$Y_t = 6.61 \sin\left(\frac{2\pi}{12}t\right) + 7.22 \cos\left(\frac{2\pi}{12}t\right) + 0.0091t - 0.0267X_{t-1} + \varepsilon_t \quad (28)$$

with $\varepsilon_t = 0.562\varepsilon_{t-1} + \eta_t$ and η_t a conv GEV($-0.0969, -5.67, 3.59$) random variable.

5.2 | Implementation

We estimate the additive component functions together with the AR parameter in our models (25) and (26) by the three-step procedure outlined in Section 3. Note that, in the simpler model (25), the trend function m_0 is estimated by a Nadaraya–Watson smoother in the second step of the procedure. To maintain comparability with the work of Hughes et al. (2007), we also use the observed data up until December 2003 for the estimation. To compute the estimators of the additive functions m_0 and m_1 and of the AR parameter, we employ an Epanechnikov kernel and choose the bandwidths by the following simple plug-in rule:

$$h_j^* = T^{-1/5} \left(\frac{c_V \int \check{v}_j^2(x_j) dx_j}{4 \int \check{\beta}_j^2(x_j) \check{p}_j(x_j) dx_j} \right)^{1/5}, \quad (29)$$

where

$$\check{\beta}_j(x_j) = c_B \left(\check{b}_j(x_j) - \int \check{b}_j(x_j) \check{p}_j(x_j) dx_j \right)$$

with

$$\check{b}_j(x_j) = \check{m}'_j(x_j) \frac{\check{p}'_j(x_j)}{\check{p}_j(x_j)} + \frac{1}{2} \check{m}''_j(x_j)$$

for $j \in \{0, 1\}$. Here, \check{v}_j^2 , \check{p}_j , \check{p}'_j , \check{m}'_j , and \check{m}''_j are initial estimators that are computed with a pilot bandwidth h . Specifically, $\check{p}_j(x_j) = T^{-1} \sum_{t=1}^T K_h(X_t^j - x_j)$ is a kernel estimator of the marginal density p_j and $\check{p}'_j(x_j)$ is its derivative. Moreover, $\check{m}'_j(x_j)$ and $\check{m}''_j(x_j)$ are estimators of the first and second derivatives $m'_j(x_j)$ and $m''_j(x_j)$, respectively. They are obtained from a local quadratic fit of \check{m}_j , as described on p. 1271 of the work of Mammen and Park (2005), where \check{m}_j is an initial backfitting estimator of the component function m_j . Furthermore, \check{v}_0^2 is an estimator of the long-run error variance v_0^2 , which has the form $v_0^2 = \mathbb{E}[\eta_t^2]/(1 - \phi^*)^2$ given the AR(1) error structure $\varepsilon_t = \phi^* \varepsilon_{t-1} + \eta_t$. To obtain \check{v}_0^2 , we fit an AR(1) model to the initial residuals $\check{\varepsilon}_t = Y_t - \check{m}_0(t) - \check{m}_1(X_{t-1})$ and compute estimates of ϕ^* and $\mathbb{E}[\eta_t^2]$. The term \check{v}_1^2 is an estimator of the conditional error variance $\mathbb{E}[\varepsilon_t^2 | X_{t-1} = \cdot]$. For simplicity, we suppose that the errors ε_t are homoskedastic, implying that $\mathbb{E}[\varepsilon_t^2 | X_{t-1} = \cdot]$ is equal to the unconditional short-run error variance $\mathbb{E}[\varepsilon_t^2]$. This allows us to work with the simple estimator $\check{v}_1^2 = T^{-1} \sum_{t=1}^T \check{\varepsilon}_t^2$. All initial estimators are computed using an Epanechnikov kernel and the starting bandwidth $h = 0.1$. The constants c_V and c_B used in (29) are given by $c_V = \int K^2(u) du$ and $c_B = \int u^2 K(u) du$. For the Epanechnikov kernel, they amount to $c_V = 3/5$ and $c_B = 1/5$.

The plug-in bandwidth h_j^* can be regarded as an estimator of the optimal bandwidth which minimizes the (asymptotic) integrated mean squared error (MSE) of the backfitting estimator \check{m}_j . To compute it, we essentially have to estimate the asymptotic variance and bias expressions appearing in the normality result of Theorem 2. To do so, we make use of the following: (a) natural estimators of the terms γ_j appearing in the asymptotic bias are given by 0 and (b) the expression $\check{\beta}_j(x_j)$ is an estimator of the bias component $\beta_j(x_j)$. In general, the terms $\beta_j(x_j)$ have a very complicated form. Only if $d = 1$ (as in our application) or if the $d > 1$ random regressors X_t^1, \dots, X_t^d are independent from each other, do we obtain $\beta_j(x_j) = c_B(b_j(x_j) - \int b_j(x_j)p_j(x_j)dx_j)$ with the simple formula $b_j(x_j) = m'_j(x_j)p'_j(x_j)/p_j(x_j) + m''_j(x_j)/2$. Hence, under the assumption of independent regressors, our plug-in rule can be easily extended to the case $d > 1$. In practice, this extended plug-in rule can be expected to work as long as the regressors are only moderately dependent, that is, as long as the independence assumption is not too far from the truth. Otherwise, more sophisticated methods are needed to approximate the terms $\beta_j(x_j)$.

The proposed plug-in procedure is of course only a heuristic rule. We do not provide any theory for automated bandwidth selection as this would be an entire project in itself. Indeed, to the best of our knowledge, bandwidth selection for smooth backfitting estimators in the dependent data case is still an open problem. Some theory for the i.i.d. case is provided in the work of Mammen and Park (2005). However, their results have not been extended to the dependent data case so far. Note that we have not used a cross-validation procedure to select the bandwidths for the following reason. As shown, for example, in the works of Altman (1990), Hart (1991), Herrmann, Gasser, and Kneip (1992), and Hart (1994), standard cross-validation tends to perform poorly when used to estimate the optimal bandwidth in the simple fixed design setting $Y_{t,T} = m(\frac{t}{T}) + \varepsilon_t$ with correlated errors. In our setting, analogous difficulties are to be expected when selecting the bandwidth in time direction.

5.3 | Estimation results

The estimate of the periodic component m_θ is given by the circles in Figure 2. The vertical dashed lines illustrate the estimated 95% confidence intervals. Using the dotted line, we have superimposed the estimated periodic function from the parametric model (28). Two differences between our periodic component estimate and the parametric estimate given in (28) become apparent immediately. Firstly, our estimate achieves its minimum in the southern hemisphere winter month of August, whereas the minimum is in July/August when the parametric model is used. Secondly, in contrast to its parametric counterpart, our estimate is not symmetric: The fall in the minimum temperature from January to August is more gradual than the increase from August until January. Interestingly, the median monthly minimum temperature also follows this pattern, as can be seen in the boxplot of the monthly minimum temperatures provided in figure 1(b) of the work of Hughes et al. (2007).

Figure 3 shows the smooth backfitting estimates of the additive functions m_0 and m_1 in model (26), along with their corresponding estimated 95% pointwise confidence bands. To compute them, we have used the bandwidths $(h_0^*, h_1^*) = (0.15, 0.14)$, which were chosen as described in Section 5.2. The dashed lines in Figure 3 are fits from the parametric model (28). As the Nadaraya–Watson estimate of m_0 in the simpler model (25) is very similar to the estimate in (26), we do not plot it separately. From the shape of \tilde{m}_0 together with the rather tight 95% confidence band in the left-hand panel of Figure 3, there seems to be a strongly nonlinear upward moving trend in the minimum monthly temperature. Not only is the linear parametric trend in (28) not capable of capturing the nonlinear pattern, we can also see that it underestimates the overall trend increase in the monthly minimum temperature over the entire estimation period. However, toward the end of the sample, the fits are very similar. The estimate \tilde{m}_1 in the right-hand panel of Figure 3 suggests that the lagged ozone concentration level has a negative effect on the minimum monthly temperature. Although the effect appears to be nonlinear again, the deviation from linearity does not seem to be as severe as for \tilde{m}_0 .

From the third step of our estimation procedure, we obtain estimated AR parameters of 0.56 and 0.57 for the models (25) and (26), respectively. These are essentially identical to the

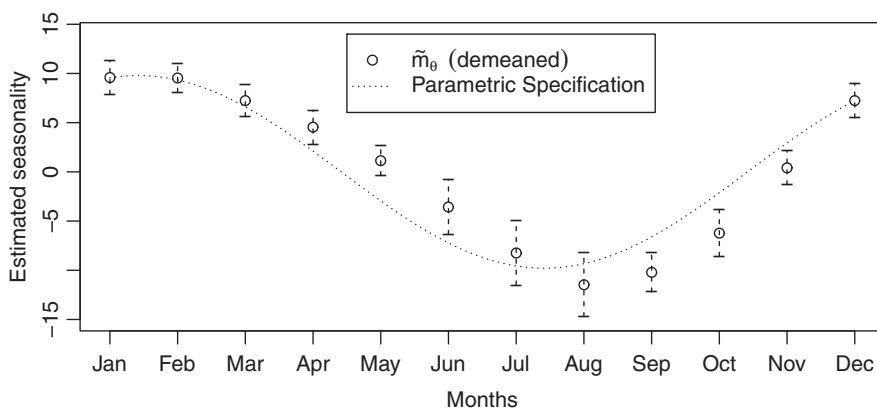


FIGURE 2 The circles represent the demeaned estimate of the seasonal component m_θ of models (25) and (26) along with (dashed vertical lines) the estimated 95% pointwise confidence intervals. The dotted line represents the function $6.61 \sin\left(\frac{2\pi}{12}t\right) + 7.22 \cos\left(\frac{2\pi}{12}t\right)$, which is the estimate of the seasonal component from the fitted parametric model in (28) obtained by Hughes et al. (2007)

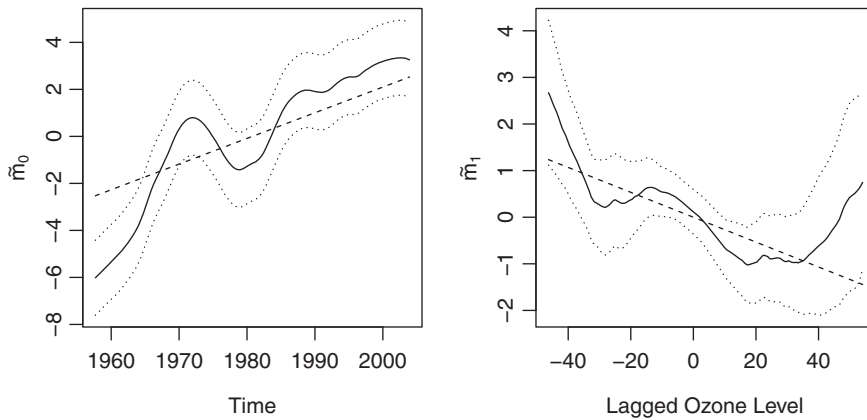


FIGURE 3 Estimation results for model (26). The solid lines are the smooth backfitting estimates \tilde{m}_0 and \tilde{m}_1 , and the dotted lines are pointwise 95% confidence bands. The dashed lines are the (centered) estimates from the fitted parametric model in (28) obtained by Hughes et al. (2007)

estimates obtained by Hughes et al. (2007) in the parametric models (27) and (28). As discussed in Section 4.4, it is straightforward to calculate confidence intervals for the parameter estimate in the simple model (25), whereas this is extremely involved in the extended model (26) if we are not willing to make the assumptions of Corollary 1. Here, we shall be content with giving the 95% confidence band in the simple model (25), which is $[0.50, 0.64]$. Comparing this to the corresponding estimated band of $[0.51, 0.62]$ for the simple parametric model (27), we see that the parameter uncertainty is fairly similar, although the estimated 95% confidence band for the parametric model (27) is slightly narrower and asymmetric due to the assumed conv GEV innovations. To summarize, it seems like the residual process displays significant positive persistence, which, as pointed out by Mudelsee (2010), is a common phenomenon for climate data.

Finally, we compare our semiparametric models with those of Hughes et al. (2007) in terms of forecasting ability. To do so, we repeat the forecasting exercise from the work of Hughes et al. (2007), that is, we compute the one-step-ahead forecasts of the minimum monthly temperatures for the twelve months from January to December 2004. The one-step-ahead forecast for time point $t_0 + 1$ is obtained by estimating the model using observations at $t = 1, \dots, t_0$ and constantly extrapolating the estimated trend function \tilde{m}_0 into the future. As it is unclear what the theoretically optimal smoothing parameter in the forecasting context would be, we have computed the forecasts for bandwidths h on a grid spanning a wide range of values from 0.05 to 1 in each direction.

We first compare the forecasting performance of our simple model (25), the corresponding parametric model (23), and the intermediate model

$$Y_t = m_\theta(t) + bt + \varepsilon_t, \quad (30)$$

where the seasonality is modeled as in our approach but the trend is linear as in the models of Hughes et al. (2007). Note that in (30), the model constant is absorbed into the seasonal component m_θ . For all three models, we take the errors to follow an AR(1) process. For comparison purposes, all three models are estimated using minimum temperature data from September 1957 onwards. The parameters $a = (a_0, \dots, a_3)$ in model (23) are estimated by least squares. For the intermediate model (30), the seasonal component is estimated as in our procedure and the parameter b is obtained from a least squares fit. Finally, our simple model (25) is estimated using a

local constant smoother with a bandwidth varying over the grid mentioned above. In all the above models, an estimator of the AR(1) parameter in the error term is obtained from a least squares fit to the estimated residuals.

Table 1 reports the estimated MSE of the forecasts for the models considered. A graphical illustration is provided in Figure 4. As one can see, our model (25) performs better in terms of MSE than its two competitors (23) and (30) for all bandwidths h considered. The largest reduction in MSE occurs when moving from model (23) to (30), that is, when using our estimate of the periodic component instead of the one considered in (23). The MSE can be seen to further reduce when moving from the intermediate model (30) to our model (25), that is, when replacing the linear trend by a nonparametric trend. The additional gain, however, is much smaller. These results are quite intuitive: The estimated periodic component exhibits a strong variation within a one-year cycle, whereas the estimated trend varies comparatively little over short time periods. Hence, for short-term forecasts, the variation in the trend is much less important than the variation in the seasonal component. Moreover, the parametric and nonparametric fits of the trend function are very close toward the end of the sample. This suggests that, for the data sample at hand, much more can be gained in terms of forecasting by improving the estimate of the seasonal component than the trend estimate.

In addition to the analysis above, we have performed the same forecasting exercise for our model (26), which includes ozone as an additional predictor. We denote the bandwidth for estimating m_j by h_j and consider bandwidths h_j on a grid ranging from 0.05 to 1 for $j = 0, 1$. For all combinations of bandwidths (h_0, h_1) considered, the MSE lies between 11.55 and 9.32.

TABLE 1 Mean squared error (MSE) of forecasts for models (23), (30), and (25) for various bandwidths h

Model (23)	Model (30)	Model (25) with bandwidth:									
		$h = 0.1$	$h = 0.2$	$h = 0.3$	$h = 0.4$	$h = 0.5$	$h = 0.6$	$h = 0.7$	$h = 0.8$	$h = 0.9$	$h = 1$
13.11	10.46	9.75	9.81	9.54	9.48	9.42	9.34	9.31	9.34	9.38	9.45

Note. In the work of Hughes et al. (2007), the best model of type (23) is estimated via a maximum likelihood procedure and an extreme value distribution assumption on a sample starting in January 1951 with a reported MSE of 11.09.

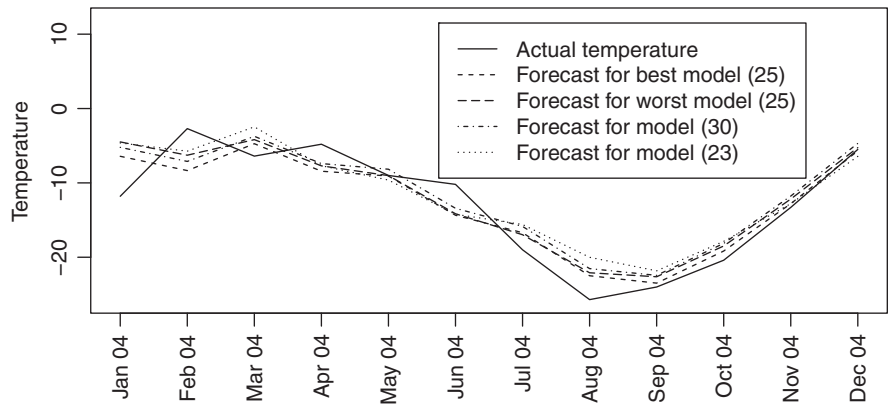


FIGURE 4 Forecasting results for the period from January to December 2004. The solid line shows the actual minimum temperatures in 2004, and the dashed and dotted lines give the one-step-ahead forecasts based on the models (23), (30), and (25). “Forecast for best model (25)” refers to the forecast of our model (25) using the “best” possible bandwidth, that is, the bandwidth that produces the smallest MSE. “Forecast for worst model (25)” analogously refers to the forecast using the “worst” bandwidth

The MSE value reported in the work of Hughes et al. (2007) for the corresponding parametric model (28) with an ozone component amounts to 10.14. The MSE produced by our semiparametric model (26) lies below the reference value 10.14 for all bandwidth combinations with $h_1 > 0.1$. Hence, we can improve on the forecasting performance of the parametric model (28) as long as we do not strongly undersmooth in the direction of ozone. The best MSE for model (26) is obtained for the bandwidth $h_1 = 1$, that is, when ozone is essentially smoothed out. Setting $h_1 = 1$ and varying the bandwidth h_0 in time direction over the grid from 0.05 to 1, the MSE varies from 9.78 to 9.32. Comparing these MSE values to those for the simpler model (25) with a trend component only, there appears to be no further gain from including ozone as a predictor. Thus, it seems that lagged ozone does not help in terms of forecasting, in contrast to the results in the work of Hughes et al. (2007). Finally, one may try to improve the forecast performance of our model (26) by using a local linear pilot smoother in time direction, as discussed at the end of Section 3.2. However, for the data sample at hand, there is no gain from doing so, the best prediction MSE over the grid of bandwidths considered being 9.63.

6 | EXTENSIONS

Our theory and methods can be extended in various directions. In what follows, we discuss some of them.

6.1 | Alternative error structures

As already mentioned in Section 3.3, our theoretical arguments can be generalized to work for other error structures. An important example is the case in which we suspect the residuals to be heteroskedastic and model them via an ARCH(p) process. Going along the lines of the proofs for Theorems 3 and 4, the ARCH parameter estimators can be shown to be consistent and asymptotically normal. The only difference to the AR case is that the conditional likelihood has a more complicated form, making it more tedious to derive the expansion of the first derivative of the likelihood function in the normality proof.

Our proving strategy may also be applied to ARMA(p, q) and GARCH(p, q) residuals. This is most easily seen for a causal and invertible ARMA(1, 1) process $\{\varepsilon_t\}$, which satisfies the equation

$$\varepsilon_t - \phi^* \varepsilon_{t-1} = \eta_t + \theta^* \eta_{t-1}$$

for some white-noise residuals η_t . In this case, the conditional likelihood can be written as

$$l_T(\phi, \theta) = - \sum_{t=2}^T (\varepsilon_t - \varepsilon_t(\phi, \theta))^2 \quad \text{with} \quad \varepsilon_t(\phi, \theta) = \sum_{k=1}^{t-1} (-\theta)^{k-1} (\phi + \theta) \varepsilon_{t-k},$$

which has a very similar structure to the likelihood function of the AR(p) case. The only notable difference is that the sum over k in the definition of $\varepsilon_t(\phi, \theta)$ now has $t - 1$ elements rather than only a fixed number p . As the elements of the sum are weighted by the coefficients $(-\theta)^{k-1}(\phi + \theta)$ that decay exponentially fast to zero, this does however not cause any major problems in the proofs. In particular, we can truncate the sum at $\min\{t - 1, C \log T\}$ for a sufficiently large C , the remainder being asymptotically negligible. After this truncation, the arguments of the AR(p) case apply more or less unchanged.

In the general ARMA(p, q) setup, the structure of the likelihood function becomes much more complicated. It is thus convenient to base the estimation of the parameters on a criterion function

that is a bit simpler to handle. In particular, consider a causal and invertible ARMA(p, q) process $\{\varepsilon_t\}$ of the form

$$\varepsilon_t - \sum_{i=1}^p \phi_i^* \varepsilon_{t-i} = \eta_t + \sum_{j=1}^q \theta_j^* \eta_{t-j}$$

and write $\phi^* = (\phi_1^*, \dots, \phi_p^*)$ and $\theta^* = (\theta_1^*, \dots, \theta_q^*)$. As $1 + \sum_{j=1}^q \theta_j^* z^j \neq 0$ for all complex $|z| \leq 1$, there exist coefficients $\rho_k^* = \rho_k(\theta^*)$ with

$$\left(1 + \sum_{j=1}^q \theta_j^* z^j\right)^{-1} = \sum_{k=0}^{\infty} \rho_k^* z^k$$

for all $|z| \leq 1$. Using this, we obtain that

$$\sum_{k=0}^{\infty} \rho_k^* \left(\varepsilon_{t-k} - \sum_{i=1}^p \phi_i^* \varepsilon_{t-k-i} \right) = \eta_t.$$

Truncating the infinite sum on the left-hand side, we now define the expressions

$$\eta_t(\phi, \theta) = \sum_{k=0}^{t-p-1} \rho_k(\theta) \left(\varepsilon_{t-k} - \sum_{i=1}^p \phi_i \varepsilon_{t-k-i} \right)$$

and estimate the ARMA coefficients ϕ^* and θ^* by minimizing the least squares criterion

$$l_T(\phi, \theta) = \sum_{t=p+1}^T \eta_t(\phi, \theta)^2.$$

This criterion function again has a very similar structure to that of the AR(p) setup. In particular, setting $\rho_0(\theta) = 1$ and $\rho_k(\theta) = 0$ for $k > 0$ yields the conditional likelihood of the AR(p) case. As the coefficients $\rho_k(\theta)$ (as well as their derivatives with respect to θ) decay exponentially fast to zero, a truncation argument as in the ARMA(1,1) case allows us to adapt the proving strategy of Theorems 3 and 4 to the setup at hand.

6.2 | A locally stationary version of our model

Our model decomposes the time series observations $Y_{t,T}$ into the seasonal component $m_\theta(t)$, the time trend component $m_0(\frac{t}{T})$, and the stationary stochastic component $\sum_{j=1}^d m_j(X_{t,T}^j) + \varepsilon_t$. Whether the stationarity of the stochastic component is a reasonable assumption of course depends on the application context. An interesting extension of our framework is to replace the stationary by a locally stationary stochastic component. This results in the model equation

$$Y_{t,T} = m_\theta(t) + m_0\left(\frac{t}{T}\right) + \sum_{j=1}^d m_j\left(\frac{t}{T}, X_{t,T}^j\right) + \varepsilon_{t,T}, \quad (31)$$

where m_θ and m_0 are defined as before and m_j are time-varying nonparametric component functions. In this very general model framework, the covariate process $\{X_{t,T} : t = 1, \dots, T\}$ may be allowed to be locally stationary (rather than strictly stationary) and $\{\varepsilon_{t,T} : t = 1, \dots, T\}$ may be a time-varying AR process of the form

$$\varepsilon_{t,T} = \sum_{i=1}^p \phi_i^* \left(\frac{t}{T}\right) \varepsilon_{t-i,T} + \eta_{t,T} \quad \text{with} \quad \eta_{t,T} = \sigma\left(\frac{t}{T}, X_{t,T}\right) \xi_t, \quad (32)$$

where σ is a time-varying volatility function and ξ_t are i.i.d. innovations. The locally stationary model (31) (without a periodic component m_θ) was studied by Vogt (2012). There, theory for smooth backfitting estimators of the time-varying functions m_j was developed. The results from the work of Vogt (2012) could be used as a starting point to generalize our theoretical results to models (31) and (32). This is however far from trivial. To do so, we would have to derive a uniform stochastic expansion of the smooth backfitting estimators in the time-varying regression model (31), which parallels the expansion from Theorem 5 in Appendix B. (Such an expansion has not been provided in the work of Vogt (2012).) With such an expansion at hand, one could attempt to extend the theoretical results of this paper and to derive the asymptotic distribution of the estimated AR parameters $\tilde{\phi}_l(u)$ at a given rescaled time point u .

6.3 | More efficient estimation by prewhitening techniques

We finally discuss a prewhitening strategy to improve the efficiency of our estimators. For ease of notation, consider the simplified model

$$Y_{t,T} = m_\theta(t) + m_0\left(\frac{t}{T}\right) + m_1(X_t) + \varepsilon_t, \quad (33)$$

where X_t is real valued and $\{\varepsilon_t\}$ is an AR(1) process of the form $\varepsilon_t = \phi^* \varepsilon_{t-1} + \eta_t$ with white-noise residuals η_t . Taking first-order differences in (33) and using the AR(1) equation of the error terms yields

$$\begin{aligned} Y_{t,T} - \phi^* Y_{t-1,T} &= m_\theta(t) - \phi^* m_\theta(t-1) + m_0\left(\frac{t}{T}\right) - \phi^* m_0\left(\frac{t-1}{T}\right) \\ &\quad + m_1(X_t) - \phi^* m_1(X_{t-1}) + \eta_t. \end{aligned} \quad (34)$$

Hence, if we knew the AR parameter ϕ^* , we could prewhiten the model by forming the first-order differences $Y_{t,T} - \phi^* Y_{t-1,T}$. As a result, we would obtain a model with uncorrelated error terms η_t , thus getting rid of the AR structure in the errors. Such a prewhitening strategy has been studied, for example, in the work of Xiao, Linton, Carroll, and Mammen (2003) in the context of a nonparametric regression model with AR errors.

In our context, we could proceed as follows. To start with, we rewrite (34) as

$$Z_{t,T}^* = m_0\left(\frac{t}{T}\right) - \phi^* m_0\left(\frac{t-1}{T}\right) + m_1(X_t) - \phi^* m_1(X_{t-1}) + \eta_t$$

with $Z_{t,T}^* = \{Y_{t,T} - \phi^* Y_{t-1,T}\} - \{m_\theta(t) - \phi^* m_\theta(t-1)\}$. Because $|m_0(\frac{t}{T}) - m_0(\frac{t-1}{T})| = O(\frac{1}{T})$ under our smoothness conditions, we obtain that

$$\begin{aligned} Z_{t,T}^* &\approx (1 - \phi^*) m_0\left(\frac{t}{T}\right) + m_1(X_t) - \phi^* m_1(X_{t-1}) + \eta_t \\ &=: g_0\left(\frac{t}{T}\right) + g_1(X_t) + g_2(X_{t-1}) + \eta_t, \end{aligned}$$

where $g_0(u) = (1 - \phi^*) m_0(u)$, $g_1(x) = m_1(x)$ and $g_2(x) = -\phi^* m_1(x)$. Moreover, as the variables $Z_{t,T}^*$ are not observed, we replace them by the estimated versions $\tilde{Z}_{t,T} = (Y_{t,T} - \tilde{\phi} Y_{t-1,T}) - (\tilde{m}_\theta(t) - \tilde{\phi} \tilde{m}_\theta(t-1))$, where $\tilde{\phi}$ and \tilde{m}_θ are our estimators from Section 3. This yields the approximate model equation

$$\tilde{Z}_{t,T} \approx g_0\left(\frac{t}{T}\right) + g_1(X_t) + g_2(X_{t-1}) + \eta_t.$$

We now estimate the functions g_0 , g_1 , and g_2 by the smooth backfitting algorithm established in Section 3.2 and denote the resulting estimators by \hat{g}_0 , \hat{g}_1 , and \hat{g}_2 . Finally, we define updated

estimators of m_0 and m_1 by $\hat{m}_0(u) = \hat{g}_0(u)/(1 - \tilde{\phi})$ and $\hat{m}_1(x) = \hat{g}_1(x)$. (More generally, we could also define $\hat{m}_1(x) = \alpha \hat{g}_1(x) - (1 - \alpha) \hat{g}_2(x)/\tilde{\phi}$ for some $\alpha \in [0, 1]$.)

We conjecture that, in terms of asymptotic variance, the updated estimators \hat{m}_0 and \hat{m}_1 should in general be more efficient than the estimators \tilde{m}_0 and \tilde{m}_1 from Section 3. In particular, only the variance of the white-noise errors η_t (rather than that of the AR errors ε_t) should show up in the asymptotic variance of \hat{m}_0 and \hat{m}_1 .

ACKNOWLEDGEMENTS

We would like to thank Enno Mammen, Oliver Linton, and Kyusang Yu for numerous helpful discussions and comments. Moreover, we are grateful for the constructive comments of an associate editor and two anonymous referees.

ORCID

Michael Vogt  <http://orcid.org/0000-0002-5885-4303>

Christopher Walsh  <http://orcid.org/0000-0003-3586-7010>

REFERENCES

- Altman, N. S. (1990). Kernel smoothing of data with correlated errors. *Journal of the American Statistical Association*, 85, 749–759.
- Altman, N. S. (1993). Estimating error correlation in nonparametric regression. *Statistics & Probability Letters*, 18, 213–218.
- Bosq, D. (1998). *Nonparametric statistics for stochastic processes: Estimation and prediction* (2nd ed.). New York, NY: Springer.
- Hall, P., & van Keilegom, I. (2003). Using difference-based methods for inference in nonparametric regression with time series errors. *Journal of the Royal Statistical Society, Series B: Statistical Methodology*, 65, 443–456.
- Hansen, B. E. (2008). Uniform convergence rates for kernel estimation with dependent data. *Econometric Theory*, 24, 726–748.
- Hart, J. D. (1991). Kernel regression estimation with time series errors. *Journal of the Royal Statistical Society, Series B: Statistical Methodology*, 53, 173–187.
- Hart, J. D. (1994). Automated kernel smoothing of dependent data by using time series cross-validation. *Journal of the Royal Statistical Society, Series B: Statistical Methodology*, 56, 529–542.
- Herrmann, E., Gasser, T., & Kneip, A. (1992). Choice of bandwidth for kernel regression when residuals are correlated. *Biometrika*, 79, 783–795.
- Hughes, G. L., Subba Rao, S., & Subba Rao, T. (2007). Statistical analysis and time-series models for minimum/maximum temperatures in the Antarctic Peninsula. *Proceedings of the Royal Society A*, 463, 241–259.
- Liebscher, E. (1996). Central limit theorems for sums of α -mixing random variables. *Stochastics and Stochastic Reports*, 59, 241–258.
- Lin, T. C., Pourahmadi, M., & Schick, A. (1999). Regression models with time series errors. *Journal of Time Series Analysis*, 20, 425–433.
- Mammen, E., Linton, O., & Nielsen, J. (1999). The existence and asymptotic properties of a backfitting projection algorithm under weak conditions. *The Annals of Statistics*, 27, 1443–1490.
- Mammen, E., & Park, B. U. (2005). Bandwidth selection for smooth backfitting in additive models. *The Annals of Statistics*, 33, 1260–1294.
- Mammen, E., & Park, B. U. (2006). A simple smooth backfitting method for additive models. *The Annals of Statistics*, 34, 2252–2271.
- Masry, E. (1996). Multivariate local polynomial regression for time series: Uniform strong consistency and rates. *Journal of Time Series Analysis*, 17, 571–599.
- Mudelsee, M. (2010). *Climate time series analysis: Classical statistical and bootstrap methods*. New York, NY: Springer.

- Schick, A. (1994). Estimation of the autocorrelation coefficient in the presence of a regression trend. *Statistics & Probability Letters*, 21, 371–380.
- Shao, Q., & Yang, L. J. (2011). Autoregressive coefficient estimation in nonparametric analysis. *Journal of Time Series Analysis*, 32, 587–597.
- Truong, Y. K. (1991). Nonparametric curve estimation with time series errors. *Journal of Statistical Planning and Inference*, 28, 167–183.
- Truong, Y. K., & Stone, C. (1994). Semiparametric time series regression. *Journal of Time Series Analysis*, 15, 405–428.
- Turner, J., Colwell, S. R., Marshall, G. J., Lachlan-Cope, T. A., Carleton, A. M., Jones, P. D., ... , Iagovkina, S. (2005). Antarctic climate change during the past 50 years. *International Journal of Climatology*, 25, 279–294.
- Turner, J., King, J. C., Lachlan-Cope, T. A., & Jones, P. D. (2002). Recent temperature trends in the Antarctic. *Nature*, 418, 291–292.
- Vogt, M. (2012). Nonparametric regression for locally stationary time series. *The Annals of Statistics*, 40, 2601–2633.
- Vogt, M., & Linton, O. (2014). Nonparametric estimation of a periodic sequence in the presence of a smooth trend. *Biometrika*, 101, 121–140.
- Xiao, Z., Linton, O., Carroll, R. J., & Mammen, E. (2003). More efficient local polynomial estimation in nonparametric regression with autocorrelated errors. *Journal of the American Statistical Association*, 98, 980–992.
- Yu, K., Mammen, E., & Park, B. U. (2011). Semi-parametric regression: Efficiency gains from modeling the nonparametric part. *Bernoulli*, 17, 736–748.

SUPPORTING INFORMATION

Additional supporting information may be found online in the Supporting Information section at the end of the article.

How to cite this article: Vogt M, Walsh C. Estimating nonlinear additive models with nonstationarities and correlated errors. *Scand J Statist.* 2019;46:160–199. <https://doi.org/10.1111/sjos.12342>

APPENDIX A

AUXILIARY RESULTS

Before deriving our main results, we state and sketch the proofs of some auxiliary lemmas. The first lemma concerns the uniform convergence of the kernel density estimators \hat{p}_j and $\hat{p}_{j,k}$.

Lemma 1. Suppose that (H1)–(H5) hold and that the bandwidth h satisfies (H6a) or (H6b). Then,

$$\sup_{x_j \in I_h} |\hat{p}_j(x_j) - p_j(x_j)| = O_p \left(\sqrt{\frac{\log T}{Th}} \right) + o(h) \quad (\text{A1})$$

$$\sup_{0 \leq x_j \leq 1} |\hat{p}_j(x_j) - \kappa_0(x_j)p_j(x_j)| = O_p \left(\sqrt{\frac{\log T}{Th}} \right) + O(h) \quad (\text{A2})$$

$$\sup_{x_j, x_k \in I_h} |\hat{p}_{j,k}(x_j, x_k) - p_{j,k}(x_j, x_k)| = O_p \left(\sqrt{\frac{\log T}{Th^2}} \right) + o(h) \quad (\text{A3})$$

$$\sup_{0 \leq x_j, x_k \leq 1} |\hat{p}_{j,k}(x_j, x_k) - \kappa_0(x_j)\kappa_0(x_k)p_{j,k}(x_j, x_k)| = O_p \left(\sqrt{\frac{\log T}{Th^2}} \right) + O(h) \quad (\text{A4})$$

for $j, k = 0, \dots, d$ with $j \neq k$, where $p_0(x_0) = I(x_0 \in (0, 1])$, $\kappa_0(v) = \int_0^1 K_h(v, w)dw$ and $I_h = [2C_1h, 1 - 2C_1h]$.

We next consider the convergence behavior of the one-dimensional Nadaraya–Watson smoothers \hat{m}_j defined in (11) and (14). To do so, we decompose $\hat{m}_j = \hat{m}_j^A + \hat{m}_j^B$ into a stochastic part \hat{m}_j^A and a bias part \hat{m}_j^B , which are defined as

$$\hat{m}_j^A(x_j) = \frac{1}{T} \sum_{t=1}^T K_h(x_j, X_t^j) \varepsilon_t / \hat{p}_j(x_j) \quad (\text{A5})$$

$$\hat{m}_j^B(x_j) = \frac{1}{T} \sum_{t=1}^T K_h(x_j, X_t^j) \left[(m_\theta(t) - \tilde{m}_\theta(t)) + m_0\left(\frac{t}{T}\right) + \sum_{k=1}^d m_k(X_t^k) \right] / \hat{p}_j(x_j) \quad (\text{A6})$$

for $j = 0, \dots, d$, where we set $X_t^0 = \frac{t}{T}$ to shorten the notation. For the stochastic part \hat{m}_j^A , we have the following.

Lemma 2. Under (H1)–(H5) together with (H6a) or (H6b),

$$\sup_{x_j \in [0,1]} |\hat{m}_j^A(x_j)| = O_p\left(\sqrt{\frac{\log T}{Th}}\right) \quad (\text{A7})$$

for all $j = 0, \dots, d$.

Lemmas 1 and 2 can be proven by small modifications of standard uniform convergence results for kernel estimators, as given in the works of Bosq (1998), Masry (1996), or Hansen (2008). For the bias part \hat{m}_j^B , we have the following expansion.

Lemma 3. Under (H1)–(H5) together with (H6a) or (H6b),

$$\sup_{x_j \in I_h} |\hat{m}_j^B(x_j) - \hat{\mu}_{T,0} - \hat{\mu}_{T,j}(x_j)| = o_p(h^2) \quad (\text{A8})$$

$$\sup_{x_j \in I_h^c} |\hat{m}_j^B(x_j) - \hat{\mu}_{T,0} - \hat{\mu}_{T,j}(x_j)| = O_p(h^2) \quad (\text{A9})$$

for all $j = 0, \dots, d$, where

$$\begin{aligned} \hat{\mu}_{T,0} &= -\frac{1}{T} \sum_{t=1}^T \left(\sum_{j=1}^d m_j(X_t^j) + \varepsilon_t \right) \\ \hat{\mu}_{T,j}(x_j) &= \alpha_{T,0} + \alpha_{T,j}(x_j) + \sum_{k \neq j} \int_0^1 \alpha_{T,k}(x_k) \frac{\hat{p}_{j,k}(x_j, x_k)}{\hat{p}_j(x_j)} dx_k + h^2 \int \beta(x) \frac{q(x)}{p_j(x_j)} dx_{-j}. \end{aligned}$$

Here, $\alpha_{T,0} = 0$ and

$$\begin{aligned} \alpha_{T,k}(x_k) &= m_k(x_k) + m'_k(x_k) \frac{h\kappa_1(x_k)}{\kappa_0(x_k)} \\ \beta(x) &= \sum_{k=0}^d \int u^2 K(u) du \left(\frac{\partial \log q(x)}{\partial x_k} m'_k(x_k) + \frac{1}{2} m''_k(x_k) \right) \end{aligned}$$

with $\kappa_0(x_k) = \int_0^1 K_h(x_k, w)dw$ and $\kappa_1(x_k) = \int_0^1 K_h(x_k, w) \left(\frac{w-x_k}{h}\right)dw$.

Lemma 3 can be proven by going along the lines of the arguments for theorem 4 in the work of Mammen et al. (1999). To see that

$$\hat{\mu}_{T,0} = -\frac{1}{T} \sum_{t=1}^T \left(\sum_{j=1}^d m_j(X_t^j) + \varepsilon_t \right), \quad (\text{A10})$$

note that

$$\begin{aligned} \hat{m}_j^B(x_j) &= \frac{1}{T} \sum_{t=1}^T K_h(x_j, X_t^j) (m_\theta(t) - \tilde{m}_\theta(t)) / \hat{p}_j(x_j) \\ &\quad + \frac{1}{T} \sum_{t=1}^T K_h(x_j, X_t^j) \left[m_0\left(\frac{t}{T}\right) + \sum_{k=1}^d m_k(X_t^k) \right] / \hat{p}_j(x_j) \end{aligned}$$

for $j = 0, \dots, d$ with $X_t^0 = \frac{t}{T}$. Moreover,

$$\begin{aligned} &\frac{1}{T} \sum_{t=1}^T K_h(x_j, X_t^j) (m_\theta(t) - \tilde{m}_\theta(t)) / \hat{p}_j(x_j) \\ &= \sum_{t_\theta=1}^{\theta} (m_\theta(t_\theta) - \tilde{m}_\theta(t_\theta)) \frac{1}{T} \sum_{k=1}^{K_{t_\theta,T}} K_h(x_j, X_{t_\theta+(k-1)\theta}^j) / \hat{p}_j(x_j) \\ &= \frac{1}{\theta} \sum_{t_\theta=1}^{\theta} (m_\theta(t_\theta) - \tilde{m}_\theta(t_\theta)) \underbrace{\frac{1}{K_{t_\theta,T}} \sum_{k=1}^{K_{t_\theta,T}} K_h(x_j, X_{t_\theta+(k-1)\theta}^j)}_{\xrightarrow{P} \kappa_0(x_j) p_j(x_j) \text{ uniformly in } x_j} / \hat{p}_j(x_j) + o_p(h^2) \\ &= \frac{1}{\theta} \sum_{t_\theta=1}^{\theta} (m_\theta(t_\theta) - \tilde{m}_\theta(t_\theta)) + o_p(h^2) \end{aligned}$$

uniformly in x_j and

$$\begin{aligned} \frac{1}{\theta} \sum_{t_\theta=1}^{\theta} (m_\theta(t_\theta) - \tilde{m}_\theta(t_\theta)) &= -\frac{1}{\theta} \sum_{t_\theta=1}^{\theta} \frac{1}{K_{t_\theta,T}} \sum_{k=1}^{K_{t_\theta,T}} \left(m_0\left(\frac{t_\theta + (k-1)\theta}{T}\right) + \sum_{j=1}^d m_j(X_{t_\theta+(k-1)\theta}^j) + \varepsilon_{t_\theta+(k-1)\theta} \right) \\ &= -\frac{1}{\theta} \sum_{t_\theta=1}^{\theta} \frac{1}{K_{t_\theta,T}} \sum_{k=1}^{K_{t_\theta,T}} \left(\sum_{j=1}^d m_j(X_{t_\theta+(k-1)\theta}^j) + \varepsilon_{t_\theta+(k-1)\theta} \right) + o_p(h^2) \\ &= -\frac{1}{T} \sum_{t=1}^T \left(\sum_{j=1}^d m_j(X_t^j) + \varepsilon_t \right) + o_p(h^2). \end{aligned}$$

Combining the above calculations with the arguments from the proof of theorem 4 in the work of Mammen et al. (1999) yields formula (A10) for $\hat{\mu}_{T,0}$.

APPENDIX B

PROOF OF THEOREM 2

In this appendix, we prove Theorem 2, which describes the asymptotic behavior of our smooth backfitting estimators. For the proof, we split up the estimators into a stochastic and a bias part. In Theorem 5, we provide a uniform expansion of the stochastic part. This result is an extension

of a related expansion given in the work of Mammen and Park (2005) in the context of bandwidth selection in additive models. The bias part is treated in Theorem 6. The proof of both theorems requires the uniform convergence results summarized in Appendix A for the kernel smoothers that enter the backfitting procedure as pilot estimators. Note that Theorems 5 and 6 are needed not only for the second estimation step but also for the derivation of the asymptotics of the AR estimators in the third step. Throughout the appendices, we use the symbol C to denote a finite real constant, which may take a different value on each occurrence.

We now turn to the proof of Theorem 2. To start with, we decompose the backfitting estimators \tilde{m}_j into a stochastic part \tilde{m}_j^A and a bias part \tilde{m}_j^B according to

$$\tilde{m}_j(x_j) = \tilde{m}_j^A(x_j) + \tilde{m}_j^B(x_j).$$

The two components are defined by

$$\tilde{m}_j^S(x_j) = \hat{m}_j^S(x_j) - \sum_{k \neq j} \int_0^1 \tilde{m}_k^S(x_k) \frac{\hat{p}_{k,j}(x_k, x_j)}{\hat{p}_j(x_j)} dx_k - \tilde{m}_c^S \quad (\text{B1})$$

for $S = A, B$, where \hat{m}_k^A and \hat{m}_k^B denote the stochastic and the bias part of the Nadaraya–Watson pilot estimators defined in (A5) and (A6). Moreover, $\tilde{m}_c^A = \frac{1}{T} \sum_{t=1}^T \epsilon_t$ and $\tilde{m}_c^B = \frac{1}{T} \sum_{t=1}^T \{m_\theta(t) - \tilde{m}_\theta(t) + m_0(\frac{t}{T}) + \sum_{k=1}^d m_k(X_t^k)\}$. We now analyze the convergence behavior of \tilde{m}_j^A and \tilde{m}_j^B separately.

We first provide a higher-order expansion of the stochastic part \tilde{m}_j^A . The following result extends theorem 6.1 in the work of Mammen and Park (2005) (in particular, their equation (6.3)) to our setting.

Theorem 5. Suppose that assumptions (H1)–(H5) apply and that the bandwidth h satisfies (H6a) or (H6b). Then,

$$\sup_{x_j \in [0,1]} \left| \tilde{m}_j^A(x_j) - \hat{m}_j^A(x_j) - \frac{1}{T} \sum_{t=1}^T r_{j,t}(x_j) \epsilon_t \right| = o_p \left(\frac{1}{\sqrt{T}} \right),$$

where $r_{j,t}(\cdot) := r_j(\frac{t}{T}, X_t, \cdot)$ are absolutely uniformly bounded functions with

$$|r_{j,t}(x'_j) - r_{j,t}(x_j)| \leq C |x'_j - x_j|$$

for a constant $C > 0$.

Proof. As Mammen and Park (2005) have worked in an i.i.d. setting, we cannot apply theorem 6.1 of their work directly. In what follows, we outline the arguments needed to extend their proof to our framework. For an additive function $g(x) = g_0(x_0) + \dots + g_d(x_d)$, let

$$\hat{g}_j g(x) = g_0(x_0) + \dots + g_{j-1}(x_{j-1}) + g_j^*(x_j) + g_{j+1}(x_{j+1}) + \dots + g_d(x_d)$$

with

$$g_j^*(x_j) = - \sum_{k \neq j} \int_0^1 g_k(x_k) \frac{\hat{p}_{j,k}(x_j, x_k)}{\hat{p}_j(x_j)} dx_k + \sum_{k=0}^d \int_0^1 g_k(x_k) \hat{p}_k(x_k) dx_k.$$

Using the uniform convergence results from Appendix A and exploiting our model assumptions, we can show that lemma 3 in the work of Mammen et al. (1999) applies in our case. For $\tilde{m}^A(x) = \tilde{m}_0^A(x_0) + \dots + \tilde{m}_d^A(x_d)$, we therefore have the expansion

$$\tilde{m}^A(x) = \sum_{r=0}^{\infty} \hat{S}^r \hat{\tau}(x),$$

where $\hat{S} = \hat{\psi}_d \cdots \hat{\psi}_0$ and $\hat{\tau}(x) = \hat{\psi}_d \cdots \hat{\psi}_1 [\hat{m}_0^A(x_0) - \hat{m}_{c,0}^A] + \cdots + \hat{\psi}_d [\hat{m}_{d-1}^A(x_{d-1}) - \hat{m}_{c,d-1}^A] + [\hat{m}_d^A(x_d) - \hat{m}_{c,d}^A]$ with $\hat{m}_{c,j}^A = \int_0^1 \hat{m}_j^A(x_j) \hat{p}_j(x_j) dx_j$. Now decompose $\tilde{m}^A(x)$ according to

$$\tilde{m}^A(x) = \hat{m}^A(x) - \hat{m}_c^A + \sum_{r=0}^{\infty} \hat{S}^r (\hat{\tau}(x) - (\hat{m}^A(x) - \hat{m}_c^A)) + \sum_{r=1}^{\infty} \hat{S}^r (\hat{m}^A(x) - \hat{m}_c^A)$$

with $\hat{m}^A(x) = \hat{m}_0^A(x_0) + \cdots + \hat{m}_d^A(x_d)$ and $\hat{m}_c^A = \hat{m}_{c,0}^A + \cdots + \hat{m}_{c,d}^A$. We show that there exist absolutely bounded functions $a_t(x)$ with $|a_t(x) - a_t(y)| \leq C\|x - y\|$ for a constant C s.t.

$$\sum_{r=1}^{\infty} \hat{S}^r (\hat{m}^A(x) - \hat{m}_c^A) = \frac{1}{T} \sum_{t=1}^T a_t(x) \varepsilon_t + o_p \left(\frac{1}{\sqrt{T}} \right) \quad (\text{B2})$$

uniformly in x . A similar claim holds for the term $\sum_{r=0}^{\infty} \hat{S}^r (\hat{\tau}(x) - (\hat{m}^A(x) - \hat{m}_c^A))$. As $\hat{m}_c^A = (d+1) \frac{1}{T} \sum_{t=1}^T \varepsilon_t$, this implies the statement of Theorem 5.

The idea behind the proof of (B2) is as follows. From the definition of the operators $\hat{\psi}_j$, it can be seen that

$$\hat{S} (\hat{m}^A(x) - \hat{m}_c^A) = \sum_{j=0}^{d-1} \hat{\psi}_d \cdots \hat{\psi}_{j+1} \left(\sum_{k=j+1}^d S_{j,k}(x_j) \right) \quad (\text{B3})$$

with

$$S_{j,k}(x_j) = - \int_0^1 \frac{\hat{p}_{j,k}(x_j, x_k)}{\hat{p}_j(x_j)} (\hat{m}_k^A(x_k) - \hat{m}_{c,k}^A) dx_k.$$

In what follows, we show that the terms $S_{j,k}(x_j)$ have the representation

$$S_{j,k}(x_j) = - \frac{1}{T} \sum_{t=1}^T \left(\frac{p_{j,k}(x_j, X_t^k)}{p_j(x_j) p_k(X_t^k)} - 1 \right) \varepsilon_t + o_p \left(\frac{1}{\sqrt{T}} \right) \quad (\text{B4})$$

uniformly in x_j . Thus, they essentially have the desired form $\frac{1}{T} \sum_t w_{t,k}(x_j) \varepsilon_t$ with some weights $w_{t,k}$. This allows us to infer that

$$\hat{S} (\hat{m}^A(x) - \hat{m}_c^A) = \frac{1}{T} \sum_{t=1}^T b_t(x) \varepsilon_t + o_p \left(\frac{1}{\sqrt{T}} \right) \quad (\text{B5})$$

uniformly in x with some absolutely bounded functions b_t satisfying $|b_t(x) - b_t(y)| \leq C\|x - y\|$ for some $C > 0$. Moreover, using the uniform convergence results from Appendix A, it can be shown that

$$\sum_{r=0}^{\infty} \hat{S}^r (\hat{m}^A(x) - \hat{m}_c^A) = \sum_{r=0}^{\infty} S^{r-1} \hat{S} (\hat{m}^A(x) - \hat{m}_c^A) + o_p \left(\frac{1}{\sqrt{T}} \right) \quad (\text{B6})$$

uniformly in x , where S is defined analogously to \hat{S} with the density estimators replaced by the true densities. Combining (B5) and (B6) completes the proof of (B2).

To show (B4), we exploit the mixing behavior of the variables X_t . Plugging the definition of \hat{m}_k^A into the term $S_{j,k}$, we can write

$$S_{j,k}(x_j) = - \frac{1}{T} \sum_{t=1}^T \left(\int_0^1 \frac{\hat{p}_{j,k}(x_j, x_k)}{\hat{p}_j(x_j) \hat{p}_k(x_k)} K_h(x_k, X_t^k) dx_k - 1 \right) \varepsilon_t.$$

Then, applying the uniform convergence results from Appendix A, we can replace the density estimators in the above expression by the true densities. This yields

$$\begin{aligned} S_{j,k}(x_j) &= -\frac{1}{T} \sum_{t=1}^T \left(\int_0^1 \frac{p_{j,k}(x_j, x_k)}{p_j(x_j)p_k(x_k)} K_h(x_k, X_t^k) dx_k - 1 \right) \varepsilon_t + o_p \left(\frac{1}{\sqrt{T}} \right) \\ &=: S_{j,k}^*(x_j) + o_p \left(\frac{1}{\sqrt{T}} \right) \end{aligned}$$

uniformly for $x_j \in [0, 1]$. In the final step, we show that

$$S_{j,k}^*(x_j) = -\frac{1}{T} \sum_{t=1}^T \left(\frac{p_{j,k}(x_j, X_t^k)}{p_j(x_j)p_k(X_t^k)} - 1 \right) \varepsilon_t + o_p \left(\frac{1}{\sqrt{T}} \right)$$

again uniformly in x_j . This is done by applying a covering argument together with an exponential inequality for mixing variables. The employed techniques are similar to those used to establish the results of Appendix A. \square

We now turn to the bias part \tilde{m}^B .

Theorem 6. Suppose that (H1)–(H5) hold. If the bandwidth h satisfies (H6a), then

$$\sup_{x_j \in I_h} |\tilde{m}_j^B(x_j) - m_j(x_j)| = O_p(h^2) \quad (\text{B7})$$

$$\sup_{x_j \in I_h^c} |\tilde{m}_j^B(x_j) - m_j(x_j)| = O_p(h) \quad (\text{B8})$$

for $j = 0, \dots, d$. If the bandwidth satisfies (H6b), we have

$$\sup_{x_j \in I_h} \left| \tilde{m}_j^B(x_j) + \frac{1}{T} \sum_{t=1}^T m_j(X_t^j) - m_j(x_j) \right| = O_p(h^2) \quad (\text{B9})$$

$$\sup_{x_j \in I_h^c} \left| \tilde{m}_j^B(x_j) + \frac{1}{T} \sum_{t=1}^T m_j(X_t^j) - m_j(x_j) \right| = O_p(h) \quad (\text{B10})$$

for $j = 0, \dots, d$.

Proof. The result follows from theorem 3 in the work of Mammen et al. (1999). Note that (A6) is not needed for the proof of theorem 3, as opposed to the statement in the work of Mammen et al. (1999). Thus, to make sure that theorem 3 applies in our case, we have to show that the high-order conditions (A1)–(A5), (A8), and (A9) from the work of Mammen et al. (1999) are fulfilled in our setting. This can be achieved by using the results from Appendix A, in particular, by using the expansion of \tilde{m}_j^B given in Lemma 3, and by following the arguments for the proof of theorem 4 in the work of Mammen et al. (1999). To see that (B7) and (B8) have to be replaced by (B9) and (B10) in the undersmoothing case with $h = O(T^{-(\frac{1}{4}+\delta)})$, note that

$$\int_0^1 \alpha_{T,j}(x_j) \hat{p}_j(x_j) dx_j = \frac{1}{T} \sum_{t=1}^T m_j(X_t^j) + O_p(h^2)$$

with $\frac{1}{T} \sum_{t=1}^T m_j(X_t^j) = O_p(\frac{1}{\sqrt{T}})$, where $\alpha_{T,j}(x_j)$ is defined in Lemma 3. Using this in the proof of theorem 3 in the work of Mammen et al. (1999) instead of $\int_0^1 \alpha_{T,j}(x_j) \hat{p}_j(x_j) dx_j = \gamma_{T,j} + o_p(h^2)$ with $\gamma_{T,j} = O(h^2)$ gives (B9) and (B10). \square

By combining Theorems 5 and 6, it is now straightforward to complete the proof of Theorem 2.

APPENDIX C

PROOF OF THEOREMS 3 AND 4

This appendix contains the proofs of Theorems 3 and 4, which show consistency and asymptotic normality of the AR estimators. By far the most difficult part is the proof of asymptotic normality. After giving some auxiliary results and proving consistency, we run through the main steps of the normality proof postponing the major technical difficulties to a series of lemmas. The main challenge of the proof is to derive a stochastic expansion of $\frac{1}{\sqrt{T}} \frac{\partial \tilde{l}_T(\phi^*)}{\partial \phi}$. This expansion is given in Lemmas 4–7. Note that, as in Appendix B, C denotes a finite real constant that may take a different value on each occurrence.

Auxiliary results

Before we come to the proofs, we list some simple facts that are frequently used throughout this appendix. For ease of notation, we work with the likelihood functions

$$l_T(\phi) = - \sum_{t=1}^T (\varepsilon_t - \varepsilon_t(\phi))^2$$

$$\tilde{l}_T(\phi) = - \sum_{t=1}^T (\tilde{\varepsilon}_t - \tilde{\varepsilon}_t(\phi))^2,$$

where $\varepsilon_t(\phi) = \sum_{i=1}^p \phi_i \varepsilon_{t-i}$ and $\tilde{\varepsilon}_t(\phi) = \sum_{i=1}^p \phi_i \tilde{\varepsilon}_{t-i}$. These differ from the functions defined in (16) and (18) only in that the sum over t starts at the time point $t = 1$ rather than at $t = p + 1$. Trivially, the error resulting from this modification can be neglected in the proofs.

To bound the distance between l_T and \tilde{l}_T , the following facts are useful. From the convergence results on the estimators $\tilde{m}_\theta, \tilde{m}_0, \dots, \tilde{m}_d$, it is easily seen that

$$\max_{t=1, \dots, T} |\varepsilon_t - \tilde{\varepsilon}_t| = O_p(h). \quad (C1)$$

Using (C1), we can immediately infer that

$$\max_{t=1, \dots, T} \sup_{\phi \in \Phi} |\varepsilon_t(\phi) - \tilde{\varepsilon}_t(\phi)| = O_p(h). \quad (C2)$$

Moreover, noting that $\frac{\partial \varepsilon_t(\phi)}{\partial \phi_i} = \varepsilon_{t-i}$ and, analogously, $\frac{\partial \tilde{\varepsilon}_t(\phi)}{\partial \phi_i} = \tilde{\varepsilon}_{t-i}$, we get

$$\max_{t=1, \dots, T} \sup_{\phi \in \Phi} \left| \frac{\partial \varepsilon_t(\phi)}{\partial \phi_i} - \frac{\partial \tilde{\varepsilon}_t(\phi)}{\partial \phi_i} \right| = O_p(h). \quad (C3)$$

Proof of Theorem 3

Let $l_T(\phi)$ and $\tilde{l}_T(\phi)$ be the likelihood functions introduced in the previous subsection. We show that

$$\sup_{\phi \in \Phi} \left| \frac{1}{T} \tilde{l}_T(\phi) - \frac{1}{T} l_T(\phi) \right| = o_p(1). \quad (C4)$$

This, together with standard arguments, yields consistency of $\tilde{\phi}$. In order to prove (C4), we decompose $\frac{1}{T}\tilde{l}_T(\phi) - \frac{1}{T}l_T(\phi)$ into

$$\begin{aligned} \frac{1}{T}\tilde{l}_T(\phi) - \frac{1}{T}l_T(\phi) &= \frac{1}{T} \sum_{t=1}^T (\varepsilon_t^2 - \tilde{\varepsilon}_t^2) + \frac{2}{T} \sum_{t=1}^T (\tilde{\varepsilon}_t - \varepsilon_t) \tilde{\varepsilon}_t(\phi) \\ &\quad + \frac{2}{T} \sum_{t=1}^T \varepsilon_t (\tilde{\varepsilon}_t(\phi) - \varepsilon_t(\phi)) + \frac{1}{T} \sum_{t=1}^T (\varepsilon_t^2(\phi) - \tilde{\varepsilon}_t^2(\phi)). \end{aligned}$$

Using (C1)–(C3), it is straightforward to show that the four terms on the right-hand side of the above equation are all $o_p(1)$ uniformly in ϕ . This shows (C4). \square

Proof of Theorem 4

By the usual Taylor expansion argument, we obtain

$$0 = \frac{1}{T} \frac{\partial \tilde{l}_T(\tilde{\phi})}{\partial \phi} = \frac{1}{T} \frac{\partial \tilde{l}_T(\phi^*)}{\partial \phi} + \frac{1}{T} \tilde{H}_T(\tilde{\phi}, \phi^*)(\tilde{\phi} - \phi^*),$$

where $\tilde{H}_T(\tilde{\phi}, \phi^*)$ is the $p \times p$ matrix whose i th row is given by

$$\frac{\partial^2 \tilde{l}_T(\bar{\phi}^{[i]})}{\partial \phi_i \partial \phi^T}$$

for some intermediate point $\bar{\phi}^{[i]}$ between ϕ^* and $\tilde{\phi}$. Rearranging and premultiplying by \sqrt{T} yields

$$\sqrt{T}(\tilde{\phi} - \phi^*) = - \left(\frac{1}{T} \tilde{H}_T(\tilde{\phi}, \phi^*) \right)^{-1} \frac{1}{\sqrt{T}} \frac{\partial \tilde{l}_T(\phi^*)}{\partial \phi}.$$

In what follows, we show that

$$\frac{1}{T} \tilde{H}_T(\tilde{\phi}, \phi^*) \xrightarrow{P} H \tag{C5}$$

$$\frac{1}{\sqrt{T}} \frac{\partial \tilde{l}_T(\phi^*)}{\partial \phi} \xrightarrow{d} N(0, \Psi) \tag{C6}$$

with $\Psi = 4W + 4\Omega$ and $H = -2\Gamma_p$, where Γ_p is the autocovariance matrix of the AR process $\{\varepsilon_t\}$, $W = (\mathbb{E}[\eta_0^2 \varepsilon_{-i} \varepsilon_{-j}])_{i,j=1,\dots,p}$ and Ω is given in (C15). This completes the proof.

Proof of (C5). By straightforward calculations, it can be seen that

$$\sup_{\phi \in \Phi} \left\| \frac{1}{T} \frac{\partial^2 \tilde{l}_T(\phi)}{\partial \phi \partial \phi^T} - \frac{1}{T} \frac{\partial^2 l_T(\phi)}{\partial \phi \partial \phi^T} \right\| = o_p(1).$$

Defining the $p \times p$ matrix $H_T(\tilde{\phi}, \phi^*)$ analogously to $\tilde{H}_T(\tilde{\phi}, \phi^*)$ with \tilde{l}_T replaced by l_T , it is further easy to show that $\frac{1}{T} H_T(\tilde{\phi}, \phi^*) \xrightarrow{P} H$, yielding (C5). \square

Proof of (C6). We write

$$\frac{1}{\sqrt{T}} \frac{\partial \tilde{l}_T(\phi^*)}{\partial \phi_i} = \frac{1}{\sqrt{T}} \frac{\partial l_T(\phi^*)}{\partial \phi_i} + \left(\frac{1}{\sqrt{T}} \frac{\partial \tilde{l}_T(\phi^*)}{\partial \phi_i} - \frac{1}{\sqrt{T}} \frac{\partial l_T(\phi^*)}{\partial \phi_i} \right).$$

Introducing the notation $\phi_0^* = -1$, we obtain that

$$\begin{aligned} \frac{1}{\sqrt{T}} \frac{\partial \tilde{l}_T(\phi^*)}{\partial \phi_i} - \frac{1}{\sqrt{T}} \frac{\partial l_T(\phi^*)}{\partial \phi_i} &= \sum_{k=0}^p 2\phi_k^* \left(\frac{1}{\sqrt{T}} \sum_{t=1}^T (\varepsilon_{t-k} - \tilde{\varepsilon}_{t-k}) \varepsilon_{t-i} \right) \\ &\quad + \sum_{k=0}^p 2\phi_k^* \left(\frac{1}{\sqrt{T}} \sum_{t=1}^T (\varepsilon_{t-i} - \tilde{\varepsilon}_{t-i}) \tilde{\varepsilon}_{t-k} \right) \\ &= \sum_{k=0}^p 2\phi_k^* \left(\frac{1}{\sqrt{T}} \sum_{t=1}^T (\varepsilon_{t-k} - \tilde{\varepsilon}_{t-k}) \varepsilon_{t-i} \right) \\ &\quad + \sum_{k=0}^p 2\phi_k^* \left(\frac{1}{\sqrt{T}} \sum_{t=1}^T (\varepsilon_{t-i} - \tilde{\varepsilon}_{t-i}) \varepsilon_{t-k} \right) + o_p(1), \end{aligned} \quad (C7)$$

where the last equality follows from the fact that $(\varepsilon_{t-i} - \tilde{\varepsilon}_{t-i})(\tilde{\varepsilon}_{t-k} - \varepsilon_{t-k}) = O_p(h^2) = o_p(\sqrt{T})$ uniformly in t, k , and i by (C1). In what follows, we derive a stochastic expansion of the terms

$$Q_T = Q_T^{[k,i]} := \frac{1}{\sqrt{T}} \sum_{t=1}^T (\varepsilon_{t-k} - \tilde{\varepsilon}_{t-k}) \varepsilon_{t-i}.$$

By symmetry, this also gives us an expansion for $Q_T^{[i,k]}$ and thus, by (C7), also for the difference $\frac{1}{\sqrt{T}} \frac{\partial \tilde{l}_T(\phi^*)}{\partial \phi_i} - \frac{1}{\sqrt{T}} \frac{\partial l_T(\phi^*)}{\partial \phi_i}$.

Introducing the shorthand $X_t^0 = \frac{t}{T}$, we have

$$\varepsilon_t - \tilde{\varepsilon}_t = (\tilde{m}_\theta(t) - m_\theta(t)) + \sum_{j=0}^d (\tilde{m}_j(X_t^j) - m_j(X_t^j)).$$

From Appendix B, we know that the backfitting estimators $\tilde{m}_j(x_j)$ can be decomposed into a stochastic part $\tilde{m}_j^A(x_j)$ and a bias part $\tilde{m}_j^B(x_j)$. This allows us to rewrite the term Q_T as

$$Q_T = Q_{T,\theta} + \sum_{j=0}^d Q_{T,V,j} + \sum_{j=0}^d Q_{T,B,j} \quad (C8)$$

with

$$\begin{aligned} Q_{T,\theta} &= \frac{1}{\sqrt{T}} \sum_{t=1}^T \varepsilon_{t-i} \left[\tilde{m}_\theta(t-k) - m_\theta(t-k) - \sum_{j=0}^d \frac{1}{T} \sum_{s=1}^T m_j(X_s^j) \right] \\ Q_{T,V,j} &= \frac{1}{\sqrt{T}} \sum_{t=1}^T \varepsilon_{t-i} \tilde{m}_j^A(X_{t-k}^j) \\ Q_{T,B,j} &= \frac{1}{\sqrt{T}} \sum_{t=1}^T \varepsilon_{t-i} \left[\tilde{m}_j^B(X_{t-k}^j) + \frac{1}{T} \sum_{s=1}^T m_j(X_s^j) - m_j(X_{t-k}^j) \right] \end{aligned}$$

for $j = 0, \dots, d$. In Lemmas 6 and 7, we will show that

$$Q_{T,\theta} = o_p(1) \quad (C9)$$

$$Q_{T,B,j} = o_p(1) \quad \text{for } j = 0, \dots, d. \quad (C10)$$

Moreover, Lemmas 4 and 5 establish that

$$Q_{T,V,0} = o_p(1) \quad (C11)$$

$$Q_{T,V,j} = \frac{1}{\sqrt{T}} \sum_{t=1}^T g_j \left(\frac{t}{T}, X_t \right) \varepsilon_t + o_p(1) \quad \text{for } j = 1, \dots, d, \quad (\text{C12})$$

where $g_j = g_j^{[k,i]}$ are deterministic functions whose exact forms are given in the statement of Lemma 4. These functions are easily seen to be absolutely bounded by a constant independent of T . Inserting the above results in (C8), we obtain

$$Q_T = \frac{1}{\sqrt{T}} \sum_{t=1}^T \left[\sum_{j=1}^d g_j \left(\frac{t}{T}, X_t \right) \right] \varepsilon_t + o_p(1).$$

Using this together with (C7) yields

$$\frac{1}{\sqrt{T}} \frac{\partial \tilde{l}_T(\phi^*)}{\partial \phi_i} - \frac{1}{\sqrt{T}} \frac{\partial l_T(\phi^*)}{\partial \phi_i} = \frac{1}{\sqrt{T}} \sum_{t=1}^T h_i \left(\frac{t}{T}, X_t \right) \varepsilon_t + o_p(1) \quad (\text{C13})$$

with the absolutely bounded function

$$h_i \left(\frac{t}{T}, X_t \right) = \sum_{j=1}^d \sum_{k=0}^p 2\phi_k^* \left[g_j^{[k,i]} \left(\frac{t}{T}, X_t \right) + g_j^{[i,k]} \left(\frac{t}{T}, X_t \right) \right], \quad (\text{C14})$$

where we suppress the dependence of h_i on the parameter vector ϕ^* in the notation. As a result,

$$\begin{aligned} \frac{1}{\sqrt{T}} \frac{\partial \tilde{l}_T(\phi^*)}{\partial \phi_i} &= \frac{1}{\sqrt{T}} \frac{\partial l_T(\phi^*)}{\partial \phi_i} + \frac{1}{\sqrt{T}} \sum_{t=1}^T h_i \left(\frac{t}{T}, X_t \right) \varepsilon_t + o_p(1) \\ &= \frac{1}{\sqrt{T}} \sum_{t=1}^T \left[2\eta_t \varepsilon_{t-i} + h_i \left(\frac{t}{T}, X_t \right) \varepsilon_t \right] + o_p(1) \\ &=: \frac{1}{\sqrt{T}} \sum_{t=1}^T U_{t,T} + o_p(1), \end{aligned}$$

that is, the term of interest can be written as a normalized sum of random variables $U_{t,T}$ plus a term that is asymptotically negligible. Using the mixing assumptions in (H1), it is straightforward to see that the variables $\{U_{t,T} : t = 1, \dots, T\}$ form an α -mixing array with mixing coefficients that decay exponentially fast to zero. We can thus apply a central limit theorem for mixing arrays to obtain that

$$\frac{1}{\sqrt{T}} \frac{\partial \tilde{l}_T(\phi^*)}{\partial \phi_i} \xrightarrow{d} N(0, \psi_{ii})$$

with $\psi_{ii} = \lim_{T \rightarrow \infty} \mathbb{E} \left(\frac{1}{\sqrt{T}} \sum_{t=1}^T U_{t,T} \right)^2$. More precisely, we apply theorem 2.1 from the work of Liebscher (1996) to the normalized sum $\frac{1}{\sqrt{T}} \sum_{t=1}^T \frac{U_{t,T}}{\sqrt{\psi_{ii}}}$. To do so, we need to verify conditions (2.1)–(2.3) of this theorem. Conditions (2.2)–(2.3) are easy to see. Moreover, the Lindeberg condition (2.1) is implied by the Lyapunov condition on p. 244 of the work of Liebscher (1996), which is straightforward to verify in our case. With the help of the Cramer–Wold device, we can finally show that

$$\frac{1}{\sqrt{T}} \frac{\partial \tilde{l}_T(\phi^*)}{\partial \phi} \xrightarrow{d} N(0, \Psi)$$

with $\Psi = (\psi_{ij})_{i,j=1,\dots,p}$, where $\Psi = 4W + 4\Omega$ and $\Omega = (\omega_{ij})_{i,j=1,\dots,p}$ with

$$\begin{aligned} \omega_{ij} = & \frac{1}{2} \sum_{l=-\infty}^{\infty} \mathbb{E} \left[\eta_0 \varepsilon_{-i} \varepsilon_l \int_0^1 h_j(u, X_l) du \right] + \frac{1}{2} \sum_{l=-\infty}^{\infty} \mathbb{E} \left[\eta_0 \varepsilon_{-j} \varepsilon_l \int_0^1 h_i(u, X_l) du \right] \\ & + \frac{1}{4} \sum_{l=-\infty}^{\infty} \mathbb{E} \left[\varepsilon_0 \varepsilon_l \int_0^1 h_i(u, X_0) h_j(u, X_l) du \right]. \end{aligned} \quad (\text{C15})$$

□

In order to complete the proof of asymptotic normality, we still need to show that Equations (C9)–(C12) are fulfilled for the terms $Q_{T,\theta}$, $Q_{T,V,j}$, and $Q_{T,B,j}$. We begin with the expansion of the variance components $Q_{T,V,j}$ for $j = 1, \dots, d$, as this is technically the most interesting part.

Lemma 4. *It holds that*

$$Q_{T,V,j} = \frac{1}{\sqrt{T}} \sum_{s=1}^T g_j \left(\frac{s}{T}, X_s \right) \varepsilon_s + o_p(1)$$

for $j = 1, \dots, d$. The functions g_j are given by

$$g_j \left(\frac{s}{T}, X_s \right) = g_j^{\text{NW}}(X_s^j) + g_j^{\text{SBF}} \left(\frac{s}{T}, X_s \right)$$

with

$$\begin{aligned} g_j^{\text{NW}}(X_s^j) &= \mathbb{E}_{-s} \left[\frac{K_h(X_{-k}^j, X_s^j) \varepsilon_{-i}}{\int_0^1 K_h(X_{-k}^j, w) dw p_j(X_{-k}^j)} \right] \\ g_j^{\text{SBF}} \left(\frac{s}{T}, X_s \right) &= \mathbb{E}_{-s} [r_{j,s}(X_{-k}^j) \varepsilon_{-i}], \end{aligned}$$

where $\mathbb{E}_{-s}[\cdot]$ is the expectation with respect to all variables except for those depending on the index s and the functions $r_{j,s}(\cdot) = r_j(\frac{s}{T}, X_s, \cdot)$ are defined in Theorem 5 in Appendix B.

Proof. By Theorem 5, the stochastic part \tilde{m}_j^A of the smooth backfitting estimator \tilde{m}_j has the expansion

$$\tilde{m}_j^A(x_j) = \hat{m}_j^A(x_j) + \frac{1}{T} \sum_{s=1}^T r_{j,s}(x_j) \varepsilon_s + o_p \left(\frac{1}{\sqrt{T}} \right)$$

uniformly in x_j , where \hat{m}_j^A is the stochastic part of the Nadaraya–Watson pilot estimator and $r_{j,s}(\cdot) = r_j(\frac{s}{T}, X_s, \cdot)$ is Lipschitz continuous and absolutely bounded. With this result, we can decompose $Q_{T,V,j}$ as follows:

$$\begin{aligned} Q_{T,V,j} &= \frac{1}{\sqrt{T}} \sum_{t=1}^T \varepsilon_{t-i} \hat{m}_j^A(X_{t-k}^j) + \frac{1}{\sqrt{T}} \sum_{t=1}^T \varepsilon_{t-i} \left[\frac{1}{T} \sum_{s=1}^T r_{j,s}(X_{t-k}^j) \varepsilon_s \right] + o_p(1) \\ &=: Q_{T,V,j}^{\text{NW}} + Q_{T,V,j}^{\text{SBF}} + o_p(1). \end{aligned}$$

In the following, we will give the arguments needed to treat $Q_{T,V,j}^{\text{NW}}$. The line of argument for $Q_{T,V,j}^{\text{SBF}}$ is essentially identical, although some of the steps are easier due to the properties of the $r_{j,s}$ functions.

Plugging the definition (A5) of the estimator $\hat{m}_j^A(x_j)$ into the term $Q_{T,V,j}^{\text{NW}}$, we get

$$Q_{T,V,j}^{\text{NW}} = \frac{1}{\sqrt{T}} \sum_{s=1}^T \left(\frac{1}{T} \sum_{t=1}^T \frac{K_h(X_{t-k}^j, X_s^j)}{\frac{1}{T} \sum_{v=1}^T K_h(X_{t-k}^j, X_v^j)} \varepsilon_{t-i} \right) \varepsilon_s. \quad (\text{C16})$$

In the first step, we show that

$$Q_{T,V,j}^{\text{NW}} = \frac{1}{\sqrt{T}} \sum_{s=1}^T \left(\frac{1}{T} \sum_{t=1}^T K_h(X_{t-k}^j, X_s^j) \mu_t \right) \varepsilon_s + o_p(1), \quad (\text{C17})$$

where $\mu_t := q_j^{-1}(X_{t-k}^j) \varepsilon_{t-i}$ with $q_j(x_j) = \int_0^1 K_h(x_j, w) dw p_j(x_j)$. To do so, decompose $\frac{1}{T} \sum_{v=1}^T K_h(x_j, X_v^j)$ as $\frac{1}{T} \sum_{v=1}^T K_h(x_j, X_v^j) = q_j(x_j) + B_j(x_j) + V_j(x_j)$ with

$$B_j(x_j) = \frac{1}{T} \sum_{v=1}^T \mathbb{E} [K_h(x_j, X_v^j)] - q_j(x_j)$$

$$V_j(x_j) = \frac{1}{T} \sum_{v=1}^T (K_h(x_j, X_v^j) - \mathbb{E} [K_h(x_j, X_v^j)]).$$

Notice that $\sup_{x_j \in [0,1]} |B_j(x_j)| = O_p(h)$ and $\sup_{x_j \in [0,1]} |V_j(x_j)| = O_p(\sqrt{\log T / Th})$. Using a second-order Taylor expansion of $f(z) = (1+z)^{-1}$, we arrive at

$$\begin{aligned} \frac{1}{\frac{1}{T} \sum_{v=1}^T K_h(x_j, X_v^j)} &= \frac{1}{q_j(x_j)} \left(1 + \frac{B_j(x_j) + V_j(x_j)}{q_j(x_j)} \right)^{-1} \\ &= \frac{1}{q_j(x_j)} \left(1 - \frac{B_j(x_j) + V_j(x_j)}{q_j(x_j)} + O_p(h^2) \right) \end{aligned}$$

uniformly in x_j . Plugging this decomposition into (C16), we obtain

$$Q_{T,V,j}^{\text{NW}} = \frac{1}{\sqrt{T}} \sum_{s=1}^T \frac{1}{T} \sum_{t=1}^T \frac{K_h(X_{t-k}^j, X_s^j)}{q_j(X_{t-k}^j)} \varepsilon_{t-i} \varepsilon_s - Q_{T,V,j}^{\text{NW},B} - Q_{T,V,j}^{\text{NW},V} + o_p(1)$$

with

$$Q_{T,V,j}^{\text{NW},B} = \frac{1}{\sqrt{T}} \sum_{s=1}^T \frac{1}{T} \sum_{t=1}^T K_h(X_{t-k}^j, X_s^j) \frac{B_j(X_{t-k}^j)}{q_j^2(X_{t-k}^j)} \varepsilon_{t-i} \varepsilon_s$$

$$Q_{T,V,j}^{\text{NW},V} = \frac{1}{\sqrt{T}} \sum_{s=1}^T \frac{1}{T} \sum_{t=1}^T K_h(X_{t-k}^j, X_s^j) \frac{V_j(X_{t-k}^j)}{q_j^2(X_{t-k}^j)} \varepsilon_{t-i} \varepsilon_s.$$

All that is required to establish (C17) is to show that both $Q_{T,V,j}^{\text{NW},B}$ and $Q_{T,V,j}^{\text{NW},V}$ are $o_p(1)$. As $\sup_{x_j \in I_h} |B_j(x_j)| = O_p(h^2)$ and $\sup_{x_j \in I_h^c} |B_j(x_j)| = O_p(h)$, we can use Markov's inequality

together with (H9) to get that $Q_{T,V,j}^{\text{NW},B} = o_p(1)$. In order to show that $Q_{T,V,j}^{\text{NW},V} = o_p(1)$, let $\mathbb{E}_v[\cdot]$ denote the expectation with respect to the variables indexed by v . Then,

$$\begin{aligned} |Q_{T,V,j}^{\text{NW},V}| &= \left| \frac{1}{\sqrt{T}} \sum_{s=1}^T \frac{1}{T} \sum_{t=1}^T \frac{K_h(X_{t-k}^j, X_s^j)}{q_j^2(X_{t-k}^j)} \varepsilon_{t-i} \right. \\ &\quad \left. \times \left(\frac{1}{T} \sum_{v=1}^T (K_h(X_{t-k}^j, X_v^j) - \mathbb{E}_v[K_h(X_{t-k}^j, X_v^j)]) \right) \varepsilon_s \right| \\ &\leq \frac{1}{\sqrt{T}} \sum_{t=1}^T \frac{|\varepsilon_{t-i}|}{q_j^2(X_{t-k}^j)} \sup_{x_j \in [0,1]} \left| \frac{1}{T} \sum_{s=1}^T K_h(x_j, X_s^j) \varepsilon_s \right| \\ &\quad \times \sup_{x_j \in [0,1]} \left| \frac{1}{T} \sum_{v=1}^T (K_h(x_j, X_v^j) - \mathbb{E}_v[K_h(x_j, X_v^j)]) \right| \\ &= O_p\left(\frac{\log T}{Th}\right) \left(\frac{1}{\sqrt{T}} \sum_{t=1}^T \frac{|\varepsilon_{t-i}|}{q_j^2(X_{t-k}^j)} \right) = O_p\left(\frac{\log T}{Th} \sqrt{T}\right) = o_p(1), \end{aligned}$$

as $\frac{1}{\sqrt{T}} \sum_{t=1}^T |\varepsilon_{t-i}| q_j^{-2}(X_{t-k}^j) = O_p(\sqrt{T})$ by Markov's inequality.

In the next step, we replace the inner sum over t in (C17) by a deterministic function that only depends on X_s^j and show that the resulting error can be asymptotically neglected. Define

$$\psi_{t,s} = K_h(X_{t-k}^j, X_s^j) \mu_t - \mathbb{E}_{-s}[K_h(X_{t-k}^j, X_s^j) \mu_t],$$

where $\mathbb{E}_{-s}[\cdot]$ is the expectation with respect to all variables except for those depending on the index s . With the above notation at hand, we can rewrite (C17) as

$$Q_{T,V,j}^{\text{NW}} = \frac{1}{\sqrt{T}} \sum_{s=1}^T \left(\frac{1}{T} \sum_{t=1}^T \mathbb{E}_{-s}[K_h(X_{t-k}^j, X_s^j) \mu_t] \right) \varepsilon_s + R_{T,V,j}^{\text{NW}} + o_p(1),$$

where

$$R_{T,V,j}^{\text{NW}} = \frac{1}{\sqrt{T}} \sum_{s=1}^T \frac{1}{T} \sum_{t=1}^T \psi_{t,s} \varepsilon_s. \quad (\text{C18})$$

Once we show that $R_{T,V,j}^{\text{NW}} = o_p(1)$, we are left with

$$\begin{aligned} Q_{T,V,j}^{\text{NW}} &= \frac{1}{\sqrt{T}} \sum_{s=1}^T \left(\frac{1}{T} \sum_{t=1}^T \mathbb{E}_{-s}[K_h(X_{t-k}^j, X_s^j) \mu_t] \right) \varepsilon_s + o_p(1) \\ &= \frac{1}{\sqrt{T}} \sum_{s=1}^T \mathbb{E}_{-s}[K_h(X_{-k}^j, X_s^j) \mu_0] \varepsilon_s + o_p(1) \\ &=: \frac{1}{\sqrt{T}} \sum_{s=1}^T g_j^{\text{NW}}(X_s^j) \varepsilon_s + o_p(1) \end{aligned}$$

with $\mu_0 = q_j^{-1}(X_{-k}^j) \varepsilon_{-i}$ and $g_j(X_{-k}^j) = \int_0^1 K_h(X_{-k}^j, w) dw p_j(X_{-k}^j)$.

Thus, it remains to prove that $R_{T,V,j}^{\text{NW}} = o_p(1)$. To do so, define

$$P := \mathbb{P} \left(\left| \frac{1}{\sqrt{T}} \sum_{s=1}^T \frac{1}{T} \sum_{t=1}^T \psi_{t,s} \varepsilon_s \right| > \delta \right)$$

for a fixed $\delta > 0$. Then, by Chebyshev's inequality

$$\begin{aligned} P &\leq \frac{1}{T^3 \delta^2} \sum_{s,s'=1}^T \sum_{t,t'=1}^T \mathbb{E} [\psi_{t,s} \epsilon_s \psi_{t',s'} \epsilon_{s'}] \\ &= \frac{1}{T^3 \delta^2} \sum_{(s,s',t,t') \in S} \mathbb{E} [\psi_{t,s} \epsilon_s \psi_{t',s'} \epsilon_{s'}] + \frac{1}{T^3 \delta^2} \sum_{(s,s',t,t') \in S^c} \mathbb{E} [\psi_{t,s} \epsilon_s \psi_{t',s'} \epsilon_{s'}] \\ &=: P_S + P_{S^c}, \end{aligned}$$

where S is the set of tuples (s, s', t, t') with $1 \leq s, s', t, t' \leq T$ such that (at least) one index is separated from the others and S^c is its complement. We say that an index, for instance, t , is separated from the others if $\min\{|t - t'|, |t - s|, |t - s'|\} > C_2 \log T$, that is, if it is further away from the other indices than $C_2 \log T$ for a constant C_2 to be chosen later on. We now analyze P_S and P_{S^c} separately.

- (a) First, consider P_{S^c} . If a tuple (s, s', t, t') is an element of S^c , then no index is separated from the others. Because the index t is not separated, there exists an index, for example, t' , such that $|t - t'| \leq C_2 \log T$. Now, take an index different from t and t' , for instance, s . Then, by the same argument, there exists an index, for example, s' , such that $|s - s'| \leq C_2 \log T$. As a consequence, the number of tuples $(s, s', t, t') \in S^c$ is smaller than $CT^2(\log T)^2$ for some constant C . Using (H8), this suffices to infer that

$$|P_{S^c}| \leq \frac{1}{T^3 \delta^2} \sum_{(s,s',t,t') \in S^c} \frac{C}{h^2} \leq \frac{C}{\delta^2} \frac{(\log T)^2}{Th^2} \rightarrow 0.$$

- (b) The term P_S is more difficult to handle. First, note that S can be written as the union of the disjoint sets

$$\begin{aligned} S_1 &= \{(s, s', t, t') \in S \mid \text{the index } t \text{ is separated}\} \\ S_2 &= \{(s, s', t, t') \in S \mid (s, s', t, t') \notin S_1 \text{ and the index } s \text{ is separated}\} \\ S_3 &= \{(s, s', t, t') \in S \mid (s, s', t, t') \notin S_1 \cup S_2 \text{ and the index } t' \text{ is separated}\} \\ S_4 &= \{(s, s', t, t') \in S \mid (s, s', t, t') \notin S_1 \cup S_2 \cup S_3 \text{ and the index } s' \text{ is separated}\}. \end{aligned}$$

Thus, $P_S = P_{S_1} + P_{S_2} + P_{S_3} + P_{S_4}$ with

$$P_{S_r} = \frac{1}{T^3 \delta^2} \sum_{(s,s',t,t') \in S_r} \mathbb{E} [\psi_{t,s} \epsilon_s \psi_{t',s'} \epsilon_{s'}]$$

for $r = 1, \dots, 4$. In what follows, we show that $P_{S_r} \rightarrow 0$ for $r = 1, \dots, 4$. As the four terms can be treated in exactly the same way, we restrict attention to the analysis of P_{S_1} .

We start by taking a cover $\{I_m\}_{m=1}^{M_T}$ of the compact support $[0, 1]$ of X_{t-k}^j . The elements I_m are intervals of length $1/M_T$ given by $I_m = [\frac{m-1}{M_T}, \frac{m}{M_T})$ for $m = 1, \dots, M_T - 1$ and $I_{M_T} = [1 - \frac{1}{M_T}, 1]$. The midpoint of the interval I_m is denoted by x_m . With this, we can write

$$K_h(X_{t-k}^j, X_s^j) = \sum_{m=1}^{M_T} I(X_{t-k}^j \in I_m) [K_h(x_m, X_s^j) + (K_h(X_{t-k}^j, X_s^j) - K_h(x_m, X_s^j))]. \quad (\text{C19})$$

Using (C19), we can further write

$$\begin{aligned} \psi_{t,s} &= \sum_{m=1}^{M_T} \left\{ I(X_{t-k}^j \in I_m) K_h(x_m, X_s^j) \mu_t - \mathbb{E}_{-s} [I(X_{t-k}^j \in I_m) K_h(x_m, X_s^j) \mu_t] \right\} \\ &\quad + \sum_{m=1}^{M_T} \left\{ I(X_{t-k}^j \in I_m) (K_h(X_{t-k}^j, X_s^j) - K_h(x_m, X_s^j)) \mu_t \right. \\ &\quad \left. - \mathbb{E}_{-s} [I(X_{t-k}^j \in I_m) (K_h(X_{t-k}^j, X_s^j) - K_h(x_m, X_s^j)) \mu_t] \right\} \\ &=: \psi_{t,s}^A + \psi_{t,s}^B \end{aligned}$$

and

$$\begin{aligned} P_{S_1} &= \frac{1}{T^3 \delta^2} \sum_{(s,s',t,t') \in S_1} \mathbb{E} [\psi_{t,s}^A \varepsilon_s \psi_{t',s'}^A \varepsilon_{s'}] + \frac{1}{T^3 \delta^2} \sum_{(s,s',t,t') \in S_1} \mathbb{E} [\psi_{t,s}^B \varepsilon_s \psi_{t',s'}^B \varepsilon_{s'}] \\ &=: P_{S_1}^A + P_{S_1}^B. \end{aligned}$$

We first consider $P_{S_1}^B$. Set $M_T = CT(\log T)h^{-3}$ and exploit the Lipschitz continuity of the kernel K to get that $|K_h(X_{t-k}^j, X_s^j) - K_h(x_m, X_s^j)| \leq \frac{C}{h^2} |X_{t-k}^j - x_m|$. This gives us

$$\begin{aligned} |\psi_{t,s}^B| &\leq \frac{C}{h^2} \sum_{m=1}^{M_T} \left(\underbrace{I(X_{t-k}^j \in I_m) |X_{t-k}^j - x_m|}_{\leq I(X_{t-k}^j \in I_m) M_T^{-1}} |\mu_t| + \mathbb{E} \left[\underbrace{I(X_{t-k}^j \in I_m) |X_{t-k}^j - x_m|}_{\leq I(X_{t-k}^j \in I_m) M_T^{-1}} |\mu_t| \right] \right) \\ &\leq \frac{C}{M_T h^2} (|\mu_t| + \mathbb{E} |\mu_t|). \end{aligned}$$

Plugging this into the expression for $P_{S_1}^B$, we arrive at

$$|P_{S_1}^B| \leq \frac{1}{T^3 \delta^2} \frac{C}{M_T h^2} \sum_{(s,s',t,t') \in S_1} \underbrace{\mathbb{E} [(|\mu_t| + \mathbb{E} |\mu_t|) |\varepsilon_s \psi_{t',s'}^B \varepsilon_{s'}|]}_{\leq Ch^{-1}} \leq \frac{C}{\delta^2 \log T} \rightarrow 0.$$

We next turn to $P_{S_1}^A$. Write

$$P_{S_1}^A = \frac{1}{T^3 \delta^2} \sum_{(s,s',t,t') \in S_1} \left(\sum_{m=1}^{M_T} \gamma_m \right)$$

with

$$\gamma_m = \mathbb{E} \left[\left\{ I(X_{t-k}^j \in I_m) K_h(x_m, X_s^j) \mu_t - \mathbb{E}_{-s} [I(X_{t-k}^j \in I_m) K_h(x_m, X_s^j) \mu_t] \right\} \varepsilon_s \psi_{t',s'}^A \varepsilon_{s'} \right].$$

By Davydov's inequality, it holds that

$$\begin{aligned} \gamma_m &= \text{Cov} (I(X_{t-k}^j \in I_m) \mu_t - \mathbb{E} [I(X_{t-k}^j \in I_m) \mu_t], K_h(x_m, X_s^j) \varepsilon_s \psi_{t',s'}^A \varepsilon_{s'}) \\ &\leq \frac{C}{h^2} (\alpha(C_2 \log T))^{1-\frac{1}{q}-\frac{1}{r}} \leq \frac{C}{h^2} (a C_2 \log T)^{1-\frac{1}{q}-\frac{1}{r}} \leq \frac{C}{h^2} T^{-C_3} \end{aligned}$$

with some $C_3 > 0$, where q and r are chosen slightly larger than $\frac{4}{3}$ and 4, respectively. Note that we can make C_3 arbitrarily large by choosing C_2 large enough. From this, it is easily seen that $P_{S_1}^A \rightarrow 0$.

Combining (a) and (b) yields that $P \rightarrow 0$ for each fixed $\delta > 0$. As a result,

$$R_{T,V,j}^{\text{NW},V} = o_p(1),$$

which completes the proof for the term $Q_{T,V,j}^{\text{NW}}$. As stated at the beginning of the proof, exactly the same arguments can be used to analyze the term $Q_{T,V,j}^{\text{SBF}}$. \square

Lemma 5. *It holds that*

$$Q_{T,V,0} = o_p(1).$$

Proof. As in Lemma 4, we can write

$$\begin{aligned} Q_{T,V,0} &= \frac{1}{\sqrt{T}} \sum_{t=1}^T \varepsilon_{t-i} \hat{m}_0^A \left(\frac{t-k}{T} \right) + \frac{1}{\sqrt{T}} \sum_{t=1}^T \varepsilon_{t-i} \left[\frac{1}{T} \sum_{s=1}^T r_{0,s} \left(\frac{t-k}{T} \right) \varepsilon_s \right] + o_p(1) \\ &=: Q_{T,V,0}^{\text{NW}} + Q_{T,V,0}^{\text{SBF}} + o_p(1). \end{aligned}$$

We again restrict attention to the arguments for $Q_{T,V,0}^{\text{NW}}$, those for $Q_{T,V,0}^{\text{SBF}}$ being essentially the same. Plugging the definition of $\hat{m}_0^A(x_0)$ into the term $Q_{T,V,0}^{\text{NW}}$ yields

$$Q_{T,V,0}^{\text{NW}} = \frac{1}{\sqrt{T}} \sum_{s=1}^T \frac{1}{T} \sum_{t=1}^T w_{t,s} \varepsilon_{t-i} \varepsilon_s$$

with $w_{t,s} = K_h(\frac{t-k}{T}, \frac{s}{T}) / \frac{1}{T} \sum_{v=1}^T K_h(\frac{t-k}{T}, \frac{v}{T})$. Now, let $\{\rho_T\}$ be some sequence that slowly converges to zero, for example, $\rho_T = (\log T)^{-1}$. By Chebyshev's inequality,

$$\mathbb{P} \left(|Q_{T,V,0}^{\text{NW}}| > C \rho_T \right) \leq C \frac{\mathbb{E} \left(Q_{T,V,0}^{\text{NW}} \right)^2}{\rho_T^2}$$

with

$$\mathbb{E} \left(Q_{T,V,j}^{\text{NW}} \right)^2 = \frac{1}{T^3} \sum_{s,s',t,t'=1}^T w_{t,s} w_{t',s'} \mathbb{E} [\varepsilon_{t-i} \varepsilon_s \varepsilon_{t'-i} \varepsilon_{s'}].$$

The moments $\mathbb{E}[\varepsilon_{t-i} \varepsilon_s \varepsilon_{t'-i} \varepsilon_{s'}]$ can be written as covariances if one of the indices s, s', t , or t' is different from the others. Exploiting our mixing assumptions, these covariances can be bounded by Davydov's inequality. With the help of the resulting bounds, it is straightforward to show that $\mathbb{E}(Q_{T,V,j}^{\text{NW}})^2 / \rho_T^2$ goes to zero, which in turn yields that $Q_{T,V,j}^{\text{NW}} = o_p(1)$. \square

Note that the above argument for $Q_{T,V,0}$ is much easier than that for $Q_{T,V,j}$ presented in Lemma 4. The main reason is that the weights $w_{t,s}$ and $w_{t',s'}$ are deterministic, allowing us to separate the expectation $\mathbb{E}[\varepsilon_{t-i} \varepsilon_s \varepsilon_{t'-i} \varepsilon_{s'}]$ from them. In contrast, in Lemma 4, we have the situation that

$$Q_{T,V,j}^{\text{NW}} = \frac{1}{\sqrt{T}} \sum_{s=1}^T \frac{1}{T} \sum_{t=1}^T w_{t,s} \varepsilon_{t-i} \varepsilon_s$$

with $w_{t,s} = K_h(X_{t-k}^j, X_s^j) / \frac{1}{T} \sum_{v=1}^T K_h(X_{t-k}^j, X_v^j)$. In this case,

$$\mathbb{E} \left(Q_{T,V,j}^{\text{NW}} \right)^2 = \frac{1}{T^3} \sum_{s,s',t,t'=1}^T \mathbb{E} [w_{t,s} w_{t',s'} \varepsilon_{t-i} \varepsilon_s \varepsilon_{t'-i} \varepsilon_{s'}]. \quad (\text{C20})$$

If the covariate process $\{X_t\}$ is independent of $\{\varepsilon_t\}$, then $\mathbb{E}[w_{t,s} w_{t',s'} \varepsilon_{t-i} \varepsilon_s \varepsilon_{t'-i} \varepsilon_{s'}] = \mathbb{E}[w_{t,s} w_{t',s'}] \mathbb{E}[\varepsilon_{t-i} \varepsilon_s \varepsilon_{t'-i} \varepsilon_{s'}]$ and similar arguments as those for the term $Q_{T,V,0}^{\text{NW}}$ yield that $Q_{T,V,j}^{\text{NW}} = o_p(1)$. However, if we allow X_t and ε_t to be dependent, then the expectations in (C20) do not split

up into two separate parts any more. Moreover, because the weights $w_{t,s}$ and $w_{t',s'}$ depend on all the X_t^j for $t = 1, \dots, T$, applying covariance inequalities like Davydov's inequality to the expressions $\mathbb{E}[w_{t,s}w_{t',s'}\varepsilon_{t-i}\varepsilon_s\varepsilon_{t'-i}\varepsilon_{s'}]$ is of no use any more. This necessitates the much more subtle arguments of Lemma 4.

We finally turn to the analysis of the terms $Q_{T,\theta}$ and $Q_{T,B,j}$.

Lemma 6. *It holds that*

$$Q_{T,\theta} = o_p(1).$$

Proof. We write

$$\begin{aligned} Q_{T,\theta} &= \frac{1}{\sqrt{T}} \sum_{t=1}^T \varepsilon_{t-i} [\tilde{m}_\theta(t-k) - m_\theta(t-k)] - \frac{1}{\sqrt{T}} \sum_{t=1}^T \varepsilon_{t-i} \left[\sum_{j=0}^d \frac{1}{T} \sum_{s=1}^T m_j(X_s^j) \right] \\ &=: Q_{T,\theta,a} + Q_{T,\theta,b} \end{aligned}$$

and consider the two terms $Q_{T,\theta,a}$ and $Q_{T,\theta,b}$ separately. For $Q_{T,\theta,a}$, we have

$$\begin{aligned} Q_{T,\theta,a} &= \sum_{t_\theta=1}^{\theta} \frac{1}{\sqrt{T}} \sum_{r=1}^{K_{t_\theta,T}} \varepsilon_{t_\theta+(r-1)\theta-i} (\tilde{m}_\theta(t_\theta-k) - m_\theta(t_\theta-k)) \\ &= \sum_{t_\theta=1}^{\theta} \underbrace{(\tilde{m}_\theta(t_\theta-k) - m_\theta(t_\theta-k))}_{=o_p(1)} \underbrace{\left(\frac{1}{\sqrt{T}} \sum_{r=1}^{K_{t_\theta,T}} \varepsilon_{t_\theta+(r-1)\theta-i} \right)}_{=O_p(1)} = o_p(1). \end{aligned}$$

Recalling the normalization of the functions m_j in (4), a similar argument yields that $Q_{T,\theta,b} = o_p(1)$ as well. \square

Lemma 7. *It holds that*

$$Q_{T,B,j} = o_p(1)$$

for $j = 0, \dots, d$.

Proof. We start by considering the case $j \neq 0$: Let $I_h = [2C_1h, 1 - 2C_1h]$ and $I_h^c = [0, 2C_1h] \cup (1 - 2C_1h, 1]$, as defined in Theorem 2. Using the uniform convergence rates from Theorem 6, we get

$$\begin{aligned} |Q_{T,B,j}| &= \left| \frac{1}{\sqrt{T}} \sum_{t=1}^T \varepsilon_{t-i} \left[\tilde{m}_j^B(X_{t-k}^j) + \frac{1}{T} \sum_{s=1}^T m_j(X_s^j) - m_j(X_{t-k}^j) \right] \right| \\ &\leq O_p(h^2) \frac{1}{\sqrt{T}} \sum_{t=1}^T |\varepsilon_{t-i}| I(X_{t-k}^j \in I_h) + O_p(h) \frac{1}{\sqrt{T}} \sum_{t=1}^T |\varepsilon_{t-i}| I(X_{t-k}^j \notin I_h). \end{aligned}$$

By Markov's inequality, the first term on the right-hand side is $O_p(h^2 \sqrt{T}) = o_p(1)$. Recognizing that, by (H9), $\mathbb{E}[|\varepsilon_{t-i}| I(X_{t-k}^j \notin I_h)] \leq Ch$ for a sufficiently large constant C , another appeal to Markov's inequality yields that the second term is $O_p(h^2 \sqrt{T}) = o_p(1)$ as well. This completes the proof for $j \neq 0$.

The proof for $j = 0$ is essentially the same: We have

$$\begin{aligned} |Q_{T,B,0}| &= \left| \frac{1}{\sqrt{T}} \sum_{t=1}^T \varepsilon_{t-i} \left[\tilde{m}_0^B \left(\frac{t-k}{T} \right) + \frac{1}{T} \sum_{s=1}^T m_0 \left(\frac{s}{T} \right) - m_0 \left(\frac{t-k}{T} \right) \right] \right| \\ &\leq O_p(h^2) \frac{1}{\sqrt{T}} \sum_{t=1}^T |\varepsilon_{t-i}| I \left(\frac{t-k}{T} \in I_h \right) + O_p(h) \frac{1}{\sqrt{T}} \sum_{t=1}^T |\varepsilon_{t-i}| I \left(\frac{t-k}{T} \in I_h^c \right) \\ &= O_p(h^2 \sqrt{T}) + O_p(h) \frac{1}{\sqrt{T}} \sum_{t=1}^T |\varepsilon_{t-i}| I \left(\frac{t-k}{T} \in I_h^c \right). \end{aligned}$$

As $\sum_{t=1}^T I \left(\frac{t-k}{T} \in I_h^c \right) \leq CTh$ for a sufficiently large constant C , Markov's inequality yields that the second term on the right-hand side is $O_p(h^2 \sqrt{T}) = o_p(1)$ as well. \square