Estimating Nonlinear Additive Models with Nonstationarities and Correlated Errors[1]

Michael Vogt[2,*]
University of Bonn

Christopher Walsh[3]
University of Vienna

February 10, 2016

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 quasi-maximum likelihood methods. Our theory establishes convergence rates as well as 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.

Key words: semiparametric, nonstationary, smooth backfitting, correlated errors.
AMS 2010 subject classifications: 62G08, 62G20, 62F12, 62P12.

1 Introduction

In many time series applications, the data at hand exhibit seasonal fluctuations as well as a trending behaviour. A common way to incorporate these features is to assume that the data generating 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 semi- and nonparametric methods in order to estimate them.

[1] We would like to thank Enno Mammen, Oliver Linton and Kyusang Yu for numerous helpful discussions and comments.

[2] Corresponding author. Address: Friedrich-Wilhelms-Universität Bonn, Institut für Finanzmarktökonomie & Statistik, Adenauerallee 24-42, 53113 Bonn, Germany. Email: michael.vogt@uni-bonn.de.

[3] Address: University of Vienna, Department of Statistics and Operations Research, Oskar-Morgenstern-Platz 1, 1090 Vienna, Austria. Email: christopher_walsh@univie.ac.at.

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

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

with $\mathbb{E}[\varepsilon_t | X_t] = 0$. Here, m_θ is a periodic function with a known period θ and m_0 is a deterministic time trend. The stochastic component consists of the residual ε_t and of the term $m(X_t)$ which 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 modelled 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 $Y_{t,T} = m_0(\frac{t}{T}) + \varepsilon_t$ has been analyzed for example in Truong [19], Altman [2], Hall & van Keilegom [5], and Shao [18] who provide a variety of methods to estimate the nonparametric trend function m_0 and the AR parameters of the error term. Interestingly, they establish an oracle type result for the estimation of the AR parameters. In particular, they show that the limiting distribution of the parameter estimators is unaffected by the need to estimate the nonparametric function m_0. A 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 & Stone [20], Schick [17] and Lin et al. [11] show 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 estimation of the parametric and nonparametric components in the general model (1). We allow X_t and ε_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

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

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 et al. [12], 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 Subsections 3.1 and 3.2, we describe how to incorporate this type of nonstationarity into the smooth backfitting procedure. Given our estimates $\tilde{m}_\theta$ 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 quasi-maximum likelihood based method, the details of which are given in Subsection 3.3.

Section 4 contains our results on the asymptotic properties of our estimators. In Subsections 4.2 and 4.3, we provide the convergence rates of the nonparametric estimators $\tilde{m}_\theta$ and $\tilde{m}_0, \dots, \tilde{m}_d$ as well as their Gaussian limit distribution. The asymptotic behaviour of the parameter estimators of the AR(p) error process is studied in Subsection 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 the 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 in the analysis of the special cases previously discussed in the literature.

It will also be seen that the oracle type result concerning the estimation of the error parameters does not hold without imposing some conditions on the the dependence structure between the covariates X_t and the errors ε_t. In general, the asymptotic distribution of our parameter estimators differs from that of the oracle estimators constructed under the assumption that the additive component functions are known. Thus, the additional uncertainty which stems from estimating the component functions does have an impact on the asymptotic distribution of our parameter estimators. However, the limiting distribution will coincide with that of the oracle estimators if $\mathbb{E}[\varepsilon_t | X_{t+k}] = 0$ for all $k = -p, \dots, p$, which 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 appyling 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 et al. [10] used a parametric regression model setup with AR errors to analyse 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 variables $\{Y_{t,T}, X_t\}$ for $t = 1, \ldots, T$, where $Y_{t,T}$ is real-valued and $X_t = (X_t^1, \ldots, X_t^d)$ is a strictly stationary $\mathbb{R}^d$-valued random vector. As already noted in the introduction, the data are assumed to follow the process

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

with $\mathbb{E}[\varepsilon_t | X_t] = 0$, where m_θ is a periodic component with some known integer-valued period θ, m_0 is a deterministic trend, and m_j are nonparametric functions of the regressors X_t^j for $j = 1, \ldots, d$. Moreover, $\{\varepsilon_t\}$ is 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 $\phi^* = (\phi_1^*, \ldots, \phi_p^*)$ is the vector of parameters and the residuals η_t are assumed to be a martingale difference.

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, i.e.

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

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 which without loss of generality is taken to be $[0, 1]$ for each $j = 1, \ldots, d$. Throughout the paper, the symbol x_0 is used to denote a point in rescaled time. Moreover, we write $x = (x_0, x_{-0})$ with $x_{-0} = (x_1, \ldots, x_d)$.

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 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 estimates 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, smaller than or equal to x. Our estimate of the periodic component m_θ is defined as

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

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 estimate has a very simple structure. It is the empirical mean of observations that are separated by a multiple of θ periods. Later on, we will show that $\tilde{m}_\theta$ 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 estimate $\tilde{m}_\theta$ can be asymptotically neglected.

3.2 Estimation of $m_0, \dots, m_d$

We next introduce the estimates 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 estimate $\tilde{m}_\theta$. 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

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

In order to estimate the functions $m_0, \dots, m_d$ in (6), we extend the smooth backfitting approach of Mammen et al. [12]. The asymptotic properties of this approach are well understood in a strictly stationary setup. Our setting, however, involves a deterministic time-trend component which makes the model dynamics nonstationary. In what follows, we describe how to extend the smooth backfitting procedure to allow for the nonstationarities present in our setting.

To do so, we first introduce the auxiliary estimates

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

$\hat{q}(x)$ is a kernel estimate of the density $q(x) := I(x_0 \in [0,1]) p(x_{-0})$ with p being the joint density of the regressors $X_t = (X_t^1, \ldots, X_t^d)$. Moreover, $\hat{m}(x)$ is a $(d+1)$-dimensional Nadaraya-Watson estimate of the regression function $m(x) = m_0(x_0) + \ldots + m_d(x_d)$. 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 $K_h(v) = \frac{1}{h} K(\frac{v}{h})$ 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 estimates.

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

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

where the minimization runs over all additive functions $g(x) = g_0(x_0) + \cdots + g_d(x_d)$ whose components satisfy $\int_0^1 g_j(x_j) \hat{p}_j(x_j) dx_j = 0$ for $j = 0, \ldots, d$. Here, $\hat{p}_j$ is a kernel estimator of p_j for $j = 0, \ldots, d$, 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 estimate $\tilde{m} = \tilde{m}_0 + \ldots + \tilde{m}_d$ 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 which 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 behaviour 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

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

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 estimates in (8) in terms of the approximations $\tilde{Z}_{t,T} = Y_{t,T} - \tilde{m}_\theta(t)$. In particular, we let

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

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

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

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 estimate 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,

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

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

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

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 $\hat{p}_0(x_0) = \frac{1}{T}\sum_{t=1}^T K_h(x_0, \frac{t}{T})$. This creates a behaviour of the estimate $\hat{p}_0$ in the boundary region of the support $[0,1]$ analogous to that of $\hat{p}_j$ at the boundary.[4]

[4] 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 $x_0 \in [2C_1 h, 1 - 2C_1 h]$, 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). Note however that in general, the system of equations (8) cannot be solved analytically. Nevertheless, the solution can be approximated by an iterative projection algorithm which converges for arbitrary starting values; see Mammen et al. [12], who establish the asymptotic properties of this algorithm under very general high order conditions. Our technical arguments will show that these high order conditions are satisfied in our framework.

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 model parameters. Suppose that the functions $m_\theta, m_0, \ldots, m_d$ were known. In this situation, the AR(p) error process ε_t would be observable, since

$$\varepsilon_t = Y_{t,T} - m_\theta(t) - m_0\Big(\frac{t}{T}\Big) - \sum_{j=1}^{d} m_j(X_t^j). \tag{15}$$

The parameters $\phi^* := (\phi_1^*, \ldots, \phi_p^*)$ 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

$$\hat{\phi} = \arg\max_{\phi \in \Phi} l_T(\phi), \tag{16}$$

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

$$l_T(\phi) = -\sum_{t=p+1}^{T} \big(\varepsilon_t - \varepsilon_t(\phi)\big)^2 \tag{17}$$

with $\varepsilon_t(\phi) = \sum_{i=1}^{p} \phi_i \varepsilon_{t-i}$. Note that $\hat{\phi}$ has a closed form solution which is identical to the usual least squares estimate. 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 $m_\theta, m_0, \ldots, m_d$ are not observed, we cannot use the standard approach from above directly. However, given the estimates $\tilde{m}_\theta, \tilde{m}_0, \ldots, \tilde{m}_d$ from the previous estimation steps, we can replace ε_t by the estimates

$$\tilde{\varepsilon}_t = Y_{t,T} - \tilde{m}_\theta(t) - \tilde{m}_0\Big(\frac{t}{T}\Big) - \sum_{j=1}^{d} \tilde{m}_j(X_t^j) \tag{18}$$

and use these as approximations to ε_t in the maximum likelihood estimation. The log-likelihood then becomes

$$\tilde{l}_T(\phi) = -\sum_{t=p+1}^{T} \left(\tilde{\varepsilon}_t - \tilde{\varepsilon}_t(\phi)\right)^2 \tag{19}$$

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

$$\tilde{\phi} = \arg\max_{\phi \in \Phi} \tilde{l}_T(\phi). \tag{20}$$

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.

(C1) *The process $\{X_t, \varepsilon_t\}$ is strictly stationary and strongly mixing with mixing coefficients α satisfying $\alpha(k) \leq a^k$ for some $0 < a < 1$.*

(C2) *The variables X_t have compact support, say $[0,1]^d$. 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 $[0,1]^d$.*

(C3) *The functions m_0 and m_j ($j = 1, \dots, d$) are twice differentiable with Lipschitz continuous second derivatives. The first partial derivatives of p exist and are continuous.*

(C4) *The kernel K is bounded, symmetric about zero and has compact support ($[-C_1, C_1]$, say). Moreover, it fulfills the Lipschitz condition that there exists a positive constant L with $|K(u) - K(v)| \leq L|u - v|$.*

(C5) *There exists 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 $l \geq l^*$.*

(C6) *The bandwidth h satisfies either of the following:*

(a) *$T^{\frac{1}{5}} h \to c_h$ for some constant $c_h > 0$.*

(b) $T^{\frac{1}{4}+\delta}h \to c_h$ *for some constant* $c_h > 0$ *and some small* $\delta > 0$.

Note that the above assumptions are very similar to the standard smoothing conditions for smooth backfitting estimators to be found e.g. in Mammen et al. [12], Mammen & Park [14] or Yu et al. [23]. It should also be mentioned that we do not necessarily require exponentially decaying mixing rates as assumed in (C1). These could alternatively be replaced by sufficiently high polynomial rates. We nevertheless make the stronger assumption (C1) to keep the notation and structure of the proofs as clear as possible.
In order to show that the estimators of the AR parameters are consistent and asymptotically normal, we additionally require the following assumptions.

(C7) *The parameter space* Φ *is a compact subset of* $\{\phi \in \mathbb{R}^p \mid \phi(z) = 1 - \phi_1 z - \ldots - \phi_p z^p \neq 0$ *for all complex* z *with* $|z| \leq 1$ *and* $\phi_p \neq 0\}$. *The true parameter vector* $\phi^* = (\phi_1^*, \ldots, \phi_p^*)$ *is an interior point of* Φ.

(C8) $\mathbb{E}[\varepsilon_t^{4+\delta}] < \infty$, *for some* $\delta > 0$.

(C9) *There exists a real constant* C *and a natural number* l^* *such that* $\mathbb{E}[|\varepsilon_t||X_{t+k}] \leq C$ *and* $\mathbb{E}[|\varepsilon_t \varepsilon_{t+l}||X_{t+k}, X_{t+l}] \leq C$ *for all* l *with* $|l| \geq l^*$ *and* $k = -p, \ldots, p$.

The compactness assumption in (C7) is required for the proof of consistency. (C8) and (C9) are technical assumptions needed to show asymptotic normality.

4.2 Asymptotics for $\tilde{m}_\theta$

We start by considering the asymptotic behaviour of the estimate $\tilde{m}_\theta$. The next theorem shows that it is asymptotically normal.

Theorem 4.1. *Assume that* $\mathbb{E}|\varepsilon_t|^\rho < \infty$ *for some* $\rho > 2$ *and let (C1) be fulfilled. Then*

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

for all $t = 1, \ldots, T$, *where*

$$V_\theta = \theta \sum_{k=-\infty}^{\infty} \mathrm{Cov}(W_0, W_{k\theta})$$

with $W_t = Y_{t,T} - m_\theta(t) - m_0(\frac{t}{T}) = \sum_{j=1}^d m_j(X_t^j) + \varepsilon_t$.

As $\tilde{m}_\theta$ and m_θ are periodic, this trivially implies that

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

The proof of Theorem 4.1 is straightforward: We have

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

The term (A) approximates the integral $\int_0^1 m_0(u)du$. It is easily seen that the convergence rate is $O(\frac{1}{T})$. As $\int_0^1 m_0(u)du = 0$ by the normalization in (4), we obtain that (A) is of the order $O(\frac{1}{T})$ and can thus be asymptotically neglected. Noting that $\{W_t\}$ is mixing by (C1) 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 4.1.

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

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

Theorem 4.2. *Suppose that conditions $(C1)$ – $(C5)$ hold.*

(a) Assume that the bandwidth h satisfies (C6a) or (C6b). Then, for $I_h = [2C_1h, 1-2C_1h]$ and $I_h^c = [0, 2C_1h) \cup (1-2C_1h, 1]$,

$$\sup_{x_j \in I_h} \big|\tilde{m}_j(x_j) - m_j(x_j)\big| = O_p\Big(\sqrt{\frac{\log T}{Th}}\Big) \tag{21}$$

$$\sup_{x_j \in I_h^c} \big|\tilde{m}_j(x_j) - m_j(x_j)\big| = O_p(h) \tag{22}$$

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

(b) Assume that the bandwidth h satisfies (C6a). Then, for any $x_0, \ldots, x_d \in (0,1)$,

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

with the bias term $B(x) = [c_h^2(\beta_0(x_0) - \gamma_0), \ldots, c_h^2(\beta_d(x_d) - \gamma_d)]'$ and the covariance matrix $V(x) = \mathrm{diag}(v_0(x_0), \ldots, v_d(x_d))$. Here, $v_0(x_0) = c_h^{-1} c_K \sum_{l=-\infty}^{\infty} \gamma_\varepsilon(l)$ and $v_j(x_j) = c_h^{-1} c_K \sigma_j^2(x_j)/p_j(x_j)$ for $j = 1, \ldots, d$ with $\gamma_\varepsilon(l) = \mathrm{Cov}(\varepsilon_t, \varepsilon_{t+l})$, $\sigma_j^2(x_j) = \mathrm{Var}(\varepsilon_t | X_t^j = x_j)$ and the constants $c_h = \lim_{T\to\infty} T^{1/5}h$ 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 A3 of Appendix A onto the space of additive functions. 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 A3 of Appendix A.

As described in Subsection 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. [12]. The details are given in Appendix B.

4.4 Asymptotics for the AR Parameter Estimates

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 4.3. *Suppose that the bandwidth h satisfies (C6a) or (C6b). In addition, let assumptions (C1) – (C5) and (C7) be fulfilled. Then $\tilde{\phi}$ is a consistent estimator of ϕ^*, i.e. $\tilde{\phi} \xrightarrow{P} \phi^*$.*

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

Theorem 4.4. *Suppose that the bandwidth h satifies (C6b) and let assumptions (C1) – (C5) together with (C7) – (C9) 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 $\{\varepsilon_t\}$, i.e. $\Gamma_p = (\gamma(i - j))_{i,j=1,\dots,p}$ 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 (61) of 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 (16) 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 4.4 thus shows that in general the limiting distribution of our estimator $\tilde{\phi}$ differs from that of the oracle estimator. There is however a wide range of cases where $\tilde{\phi}$ has the same asymptotic distribution as $\hat{\phi}$. This oracle type result is stated in the following corollary.

Corollary 4.1. *Suppose that all the assumptions of Theorem 4.4 are fulfilled and that* $\mathbb{E}[\varepsilon_t|X_{t+k}] = 0$ *for all* $k = -p, \ldots, p$*. Then*

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

Corollary 4.1 follows directly from the proof of Theorem 4.4: Inspecting the functions defined in Lemma C1 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, i.e. 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 non-constant.

Note that our theory also reestablishes the oracle result derived in the simpler setup without stochastic covariates, i.e. in the model

$$Y_{t,T} = m_\theta(t) + m_0\Big(\frac{t}{T}\Big) + \varepsilon_t \quad \text{for } t = 1, \ldots, T \tag{23}$$

with $\mathbb{E}[\varepsilon_t] = 0$. In this case, the periodic component can be estimated as described in Subsection 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 4.4 shows that the limiting distribution of the AR parameter estimates is identical to that of the oracle estimates in this setting. In particular, the stochastic higher-order expansion derived in Lemma C1 is not required any more. The arguments of the much simpler Lemma C2 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 C2 in Appendix C.

The normality results of Theorem 4.4 and Corollary 4.1 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 4.1, it is not at all trivial to derive a consistent estimator of V^* in Theorem 4.4. 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 estimates (see Theorem B1 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 4.4 could be used as a starting point to derive consistency results for such a bootstrap procedure. However, this is beyond the scope of the present 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.[5] A strong warming trend has been identified on the whole peninsula during the past 50 years. In particular, the monthly mean temperatures at Faraday station have considerably increased over this time (cf. Turner et al. [21], [22]).

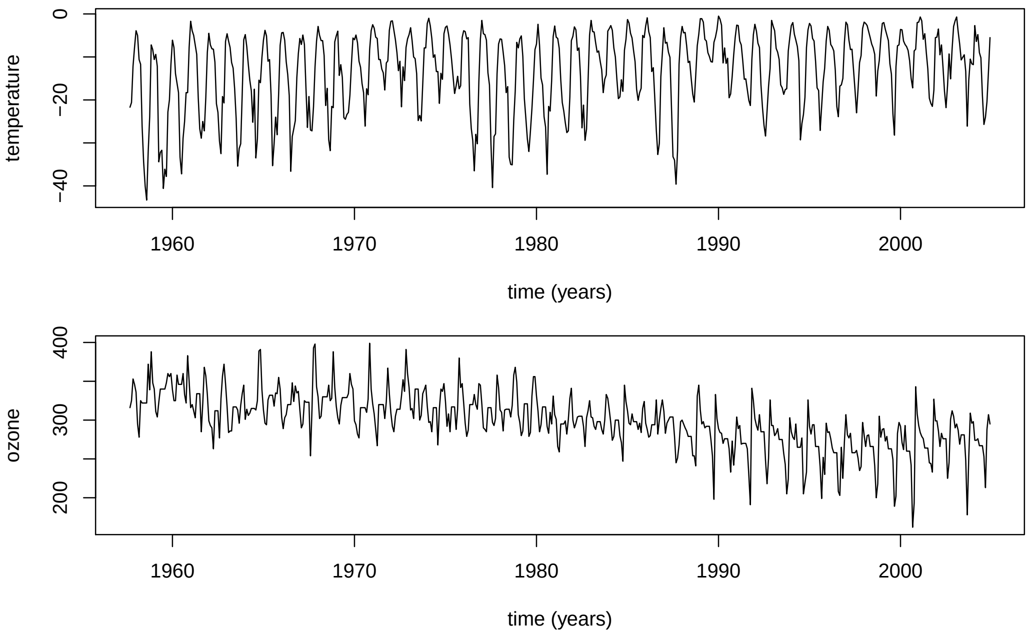

Figure 1: The upper panel shows the monthly minimum near-surface temperatures (in °C), the lower one the monthly stratospheric ozone concentrations (in Dobson units) at Faraday station.

We will closely follow the analysis of Hughes et al. [10] as our model can be seen as a semiparametric extension to their approach. According to Hughes et al. [10], the rise of the mean monthly temperature is mostly due to an increase in the minimum monthly temperature. They argue that to understand and quantify the warming on

[5] The data can be downloaded from the webpage of Suhasini Subba Rao `http://www.stat.tamu.edu/~suhasini/data.html`. Alternatively, it is available on request from the British Antarctic Survey, Cambridge.

the peninsula an appropriate statistical model of the minimum temperature is called for. Following their lead we will focus on modelling the minimum temperature and consider stratospheric ozone as a potential explanatory variable.

The data used in our analysis is plotted in Figure 1. The upper panel contains the monthly minimum near-surface temperatures at Faraday station from September 1957 to December 2004, whilst the lower shows the monthly level of stratospheric ozone concentration measured in Dobson units over the same period. For more information on the data consult Hughes et al. [10], where a detailed description of them can be found.

Hughes et al. [10] propose 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

$$Y_t = a_0 + a_1 \sin\left(\frac{2\pi}{12}t\right) + a_2 \cos\left(\frac{2\pi}{12}t\right) + a_3 t + \varepsilon_t, \tag{24}$$

where Y_t denotes the minimum monthly temperature and $a = (a_1, \dots, a_3)$ is a vector of parameters. In addition, they consider the extended model

$$Y_t = a_0 + a_1 \sin\left(\frac{2\pi}{12}t\right) + a_2 \cos\left(\frac{2\pi}{12}t\right) + a_3 t + a_4 X_{t-1} + \varepsilon_t, \tag{25}$$

where the covariate X_{t-1}, denoting the lagged detrended and deseasonalized ozone concentration, enters linearly. In their analysis, they find a strong linear upward trend in the minimum monthly temperature. Moreover, they observe considerable autocorrelation in the residuals ε_t and propose an AR process to model them. Using an order selection criterion, they find an AR(1) model to be most suitable, which also fits nicely with the preference for AR(1) errors when using discrete time series to model climate data as mentioned in Mudelsee [16].

We now introduce a framework that can be regarded as a semiparametric extension to the parametric models (24) and (25). Our baseline model is given by

$$Y_{t,T} = m_\theta(t) + m_0\Big(\frac{t}{T}\Big) + \varepsilon_t \quad \text{for } t = 1, \dots, T, \tag{26}$$

where $Y_{t,T}$ are minimum monthly temperatures, m_θ is a seasonal component and m_0 is a nonparametric time trend. We additionally consider an extended version of (26) having the form

$$Y_{t,T} = m_\theta(t) + m_0\Big(\frac{t}{T}\Big) + m_1(X_{t-1}) + \varepsilon_t \quad \text{for } t = 1, \dots, T, \tag{27}$$

where as before, the variables X_{t-1} denote lagged monthly stratospheric ozone concentration levels that have been detrended and deseasonalized as in Hughes et al. [10]. The additive functions m_θ, m_0, and m_1 in the above two models are normalized as described in (4). Following Hughes et al. [10], 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 will provide the preferred fits of the models (24) and (25) given in Hughes et al. [10] in order to compare our estimates to theirs. Their models are fitted using observations up until and including December 2003. For the model (24) their preferred fit is

$$Y_t = 6.25 \sin\left(\frac{2\pi}{12}t\right) + 6.95 \cos\left(\frac{2\pi}{12}t\right) + 0.0105t + \varepsilon_t, \tag{28}$$

with $\varepsilon_t = 0.566\varepsilon_{t-1} + \eta_t$ and η_t distributed as a conv GEV(-0.109,-5.71,3.65). [6] Their preferred fit for the model in (25) is

$$Y_t = 6.61 \sin\left(\frac{2\pi}{12}t\right) + 7.22 \cos\left(\frac{2\pi}{12}t\right) + 0.0091t - 0.0267X_{t-1} + \varepsilon_t, \tag{29}$$

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

We now turn to the estimation of our models (26) and (27). To maintain comparability to Hughes et al. [10] we will also estimate our models using the observed data up until December 2003. Using our three step procedure outlined in Section 3, we can estimate the additive component functions of (26) and (27) together with the AR parameter of the error term.

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 dashed line we have superimposed the estimated periodic function from the parametric model (29). Two differences between our periodic component estimate and the parametric estimate given in (29) become apparent immediately. Firstly, our periodic component gives the lowest estimated monthly effect in the southern hemisphere winter month of August, whereas the lowest estimated monthly effect is in July and August, when using the parametric model. Secondly, in contrast to the parametric component 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 Hughes et al. [10].

[6] The conv GEV stands for converse Generalized Extreme Value. A conv GEV(γ, μ, σ random variable Z has a distribution function

$$P(Z \leq z) = 1 - \exp\left\{\left[1 + \frac{\gamma}{\sigma}(\mu - z)\right]^{-\frac{1}{\gamma}}\right\}.$$

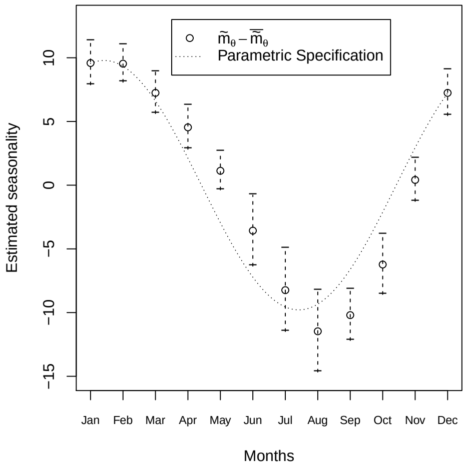

Figure 2: The circles represent the demeaned estimates of the seasonal component m_θ of models (26) and (27) along with the estimated 95% pointwise confidence intervals (dashed vertical lines). The dotted line is $6.61 \sin\left(\frac{2\pi}{12}t\right) + 7.22 \cos\left(\frac{2\pi}{12}t\right)$, the estimate of the seasonal component from the fitted parametric model in (29) obtained by Hughes et al. [10].

In Figure 3 the smooth backfitting estimates of the additive functions m_0 and m_1 in model (27) are given by the solid lines along with their corresponding estimated 95% pointwise confidence bands given by the dotted lines. The dashed lines are fits from the parametric model (29). As the Nadaraya-Watson estimate of m_0 in the simpler model (26) is very similar to the estimate in (27), we do not plot it separately. For the estimation of the functions m_0 and m_1, we have used an Epanechnikov kernel and bandwidths selected by a simple plug-in rule. To check the robustness of our results, we have additionally repeated our analysis for a wide range of different bandwidths. As the results are very similar, we only report the findings for the bandwidths chosen by the plug-in rule.

From the shape of $\tilde{m}_0$ together with the rather tight 95% confidence bands 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 (29) not capable of capturing the nonlinear pattern, we can also see that it overestimates the overall trend increase in the monthly minimum temperature over the entire estimation period. 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 $\tilde{m}_0$.

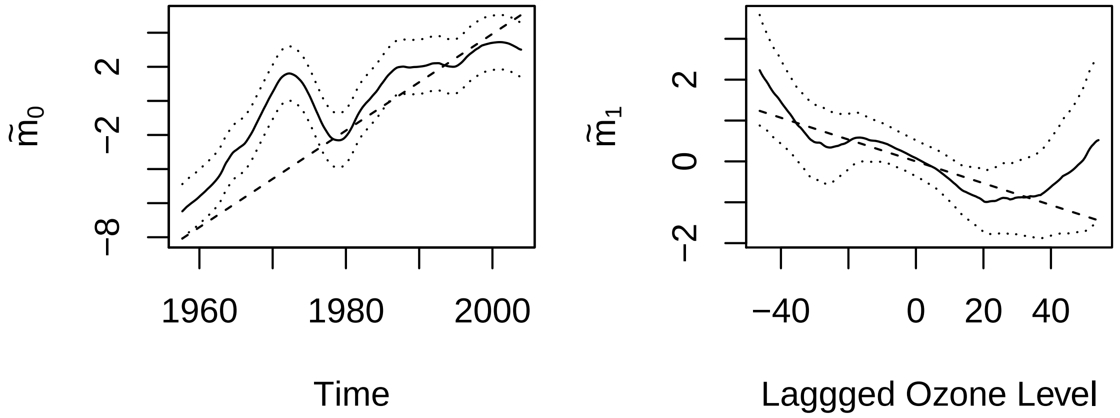

Figure 3: Estimation results for model (27). The solid lines are the smooth backfitting estimates $\tilde{m}_0$ and $\tilde{m}_1$, the dotted lines are pointwise 95% confidence bands. The dashed lines are the estimates from the fitted parametric model in (29) obtained by Hughes et al. [10].

From the third step of our estimation procedure, we obtain estimated AR parameters of 0.57 and 0.58 for the models (26) and (27) respectively. These are essentially identical to the estimates obtained by Hughes et al. [10] in the parametric models (28) and (29). As discussed in Subsection 4.4, it is straightforward to calculate confidence intervals for the parameter estimate in the simple model (26), whereas this is extremely involved in the extended model (27) if we are not willing to make the assumptions of Corollary 4.1. Here, we shall be content with giving the 95% confidence band in the simple model (26), which is $[0.49, 0.67]$. Comparing this to the corresponding estimated band of $[0.51, 0.62]$ for the simple parametric model (28), we see that the parameter uncertainty is fairly similar although the estimated 95% confidence band for the parametric model (28) is slightly narrower than the one for our simple model (26) and asymmetric due to the assumed converse GEV innovations. To summarize, it seems like the residual process displays significant positive persistence which is a common phenomenon for climate data (see Mudelsee [16]).

Finally, we compare the forecasting abilities of our models versus those of Hughes et al. [10] by repeating their forecasting exercise, i.e. 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 $t = 1, \ldots, t_0$ and constantly extrapolating the estimated trend function $\tilde{m}_0$ into the future. The resulting forecasts for our model and the actual minimum monthly temperature of the forecasting period are given in Figure 4. The estimated mean squared error (MSE) of the forecasts based on model (27) is 10.27, whereas for the simple model (26) it amounts to 9.70. The prediction MSE for (28) and (29) are reported as 11.09 and 10.14. This suggests that for this forecasting exercise at least our simple model (26) is best at forecasting. Contrary to the finding in Hughes et al. [10] we do not seem to gain in terms of forecasting performance from including lagged stratospheric ozone as an additional covariate.

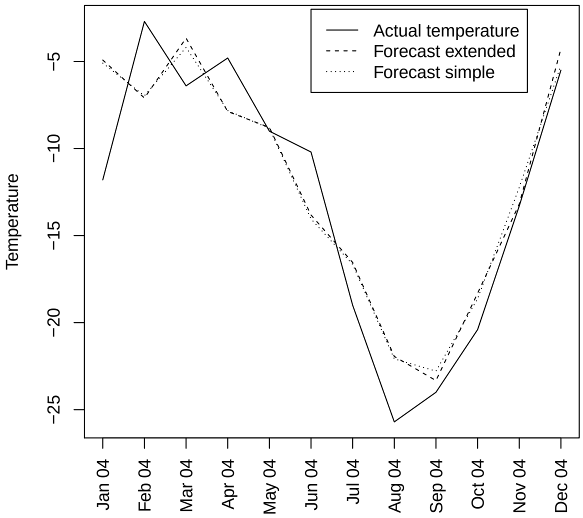

Figure 4: Forecasting results for the period from January 2004 to December 2004. The solid line shows the actual minimum temperatures in 2004, the dashed line gives the one-step ahead forecasts based on the extended model (27), and the dotted line depicts the corresponding forecasts based on the simple model (26).

6 Concluding Remarks

Our theory can be extended to allow for other error structures as mentioned in Subsection 3.3. 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 4.3 and 4.4, 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

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

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

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

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

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

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

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

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

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

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 4.3 and 4.4 to the setup at hand.

Appendix A - Auxiliary Results

Before considering the proof of our main results, we will state and sketch the proofs of some auxiliary results. These will be needed at several parts of the main proofs later on, in particular for Theorem 4.2. The first auxiliary result concerns the uniform convergence of the kernel density estimators $\hat{p}_j$ and $\hat{p}_{j,k}$. The extensions from the i.i.d. setting to dependent data are given for example in Bosq [3], Masry [15] or Hansen [6]. Using the notation $p_0(x_0) = I(x_0 \in (0,1])$, we have the following result.

Lemma A1. *Suppose that (C1) – (C5) hold and that the bandwidth h satisfies (C6a) or (C6b). Then*

$$\sup_{x_j \in I_h} \big|\hat{p}_j(x_j) - p_j(x_j)\big| = O_p\Big(\sqrt{\frac{\log T}{Th}}\Big) + o(h) \tag{30}$$

$$\sup_{0 \le x_j \le 1} \big|\hat{p}_j(x_j) - \kappa_0(x_j)p_j(x_j)\big| = O_p\Big(\sqrt{\frac{\log T}{Th}}\Big) + O(h) \tag{31}$$

$$\sup_{x_j, x_k \in I_h} \big|\hat{p}_{j,k}(x_j,x_k) - p_{j,k}(x_j,x_k)\big| = O_p\Big(\sqrt{\frac{\log T}{Th^2}}\Big) + o(h) \tag{32}$$

$$\sup_{0 \le x_j, x_k \le 1} \big|\hat{p}_{j,k}(x_j,x_k) - \kappa_0(x_j)\kappa_0(x_k)p_{j,k}(x_j,x_k)\big| = O_p\Big(\sqrt{\frac{\log T}{Th^2}}\Big) + O(h) \tag{33}$$

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

We next consider the convergence behaviour of the one-dimensional Nadaraya-Watson smoothers $\hat{m}_j$ defined in (11) and (14). For the stochastic part $\hat{m}_j^A$, we have

Lemma A2. *Under (C1) – (C5) together with (C6a) or (C6b),*

$$\sup_{x_j\in[0,1]} \big|\hat{m}_j^A(x_j)\big| = O_p\Big(\sqrt{\frac{\log T}{Th}}\Big) \tag{34}$$

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

Proof of Lemma A2. Starting with the definition of $\hat{m}_j^A(x_j)$ using the uniform convergence results for the kernel density estimator, the Borel-Cantelli lemma can be used to show that

$$\sup_{x_j\in[0,1]} \big|\hat{m}_j^A(x_j)\big| = \sup_{x_j\in[0,1]} \big|\frac{1}{T}\sum_{t=1}^{T} K_h(x_j, X_t^j)\varepsilon_t I(\varepsilon_t \le \tau_T)\big| + o_p\Big(\sqrt{\frac{\log T}{Th}}\Big)$$

for a suitably chosen truncation sequence τ_T. The result is then established using a covering argument, an exponential inequality and the mixing conditions.

□

For the bias part $\hat{m}_j^B$, we have the following expansion:

Lemma A3. *Under (C1) – (C5) together with (C6a) or (C6b),*

$$\sup_{x_j\in I_h} \big|\hat{m}_j^B(x_j) - \hat{\mu}_{T,0} - \hat{\mu}_{T,j}(x_j)\big| = o_p(h^2) \tag{35}$$

$$\sup_{x_j\in I_h^c} \big|\hat{m}_j^B(x_j) - \hat{\mu}_{T,0} - \hat{\mu}_{T,j}(x_j)\big| = O_p(h^2) \tag{36}$$

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

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

$$\hat{\mu}_{T,j}(x_j) = \alpha_{T,0} + \alpha_{T,j}(x_j) + \sum_{k\neq j}\int_0^1 \alpha_{T,k}(x_k)\frac{\hat{p}_{j,k}(x_j,x_k)}{\hat{p}_j(x_j)}dx_k + h^2\int \beta(x)\frac{q(x)}{p_j(x_j)}dx_{-j}.$$

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

$$\alpha_{T,k}(x_k) = m_k(x_k) + m_k'(x_k)\frac{h\kappa_1(x_k)}{\kappa_0(x_k)}$$

$$\beta(x) = \sum_{k=0}^{d}\int u^2K(u)du\Big(\frac{\partial \log q(x)}{\partial x_k}m_k'(x_k) + \frac{1}{2}m_k''(x_k)\Big)$$

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

Lemma A3 can be proven by going along the lines of the arguments for Theorem 4 in Mammen et al. [12]. To see that

$$\hat{\mu}_{T,0} = -\frac{1}{T}\sum_{t=1}^{T}\Big(\sum_{j=1}^{d} m_j(X_t^j) + \varepsilon_t\Big), \tag{37}$$

note that

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

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

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

uniformly in x_j and

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

Combining the above calculations with the arguments from the proof of Theorem 4 in [12] yields formula (37) for $\hat{\mu}_{T,0}$.

Appendix B - Proof of Theorem 4.2

In this appendix, we prove Theorem 4.2, which describes the asymptotic behaviour of our smooth backfitting estimates. For the proof, we split up the estimates into a "stochastic" part and a "bias" part. In Theorem B1, we provide a uniform expansion of the stochastic part. This result is an extension of a related expansion given in Mammen & Park [13] in the context of bandwidth selection in additive models. The bias part is treated in Theorem B2. 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 estimates. Note that the two theorems B1 and B2 are not only needed for the second estimation step but also for the derivation of the asymptotics of the AR estimates in the third step. Throughout the appendix, we use the symbol C to denote a finite real constant which may take a different value on each occurrence.

Proof of Theorem 4.2

We decompose the backfitting estimates $\tilde{m}_j$ 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

$$\tilde{m}_j^S(x_j) = \hat{m}_j^S(x_j) - \sum_{k \neq j} \int_0^1 \tilde{m}_k^S(x_k) \frac{\hat{p}_{k,j}(x_k, x_j)}{\hat{p}_j(x_j)} \, dx_k - \tilde{m}_c^S \tag{38}$$

for $S = A$, B. Here, $\hat{m}_k^A$ and $\hat{m}_k^B$ denote the stochastic part and the bias part of the Nadaraya-Watson pilote estimates defined as

$$\hat{m}_j^A(x_j) = \frac{1}{T} \sum_{t=1}^T K_h(x_j, X_t^j) \varepsilon_t / \hat{p}_j(x_j) \tag{39}$$

$$\begin{aligned} \hat{m}_j^B(x_j) = \frac{1}{T} \sum_{t=1}^T K_h(x_j, X_t^j) \big[(m_\theta(t) - \tilde{m}_\theta(t)) \\ + m_0\Big(\frac{t}{T}\Big) + \sum_{k=1}^d m_k(X_t^k) \big] / \hat{p}_j(x_j) \end{aligned} \tag{40}$$

for $j = 0, \ldots, d$, where we set $X_t^0 = \frac{t}{T}$ to shorten the notation. Furthermore, $\tilde{m}_c^A = \frac{1}{T} \sum_{t=1}^T \varepsilon_t$ 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 analyse the convergence behaviour 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 Mammen & Park [13] (in particular their equation (6.3)) to our setting.

Theorem B1. *Suppose that assumptions (C1) – (C5) apply and that the bandwidth h satisfies (C6a) or (C6b). Then*

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

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 & Park [13] work in an i.i.d. setting, we cannot apply their Theorem 6.1 directly. In what follows, we outline the arguments needed to extend their proof to our framework. For an additive function $g(x) = g_0(x_0) + \ldots + g_d(x_d)$, let

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

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 Mammen et al. [12] applies in our case. For $\tilde{m}^A(x) = \tilde{m}_0^A(x_0) + \ldots + \tilde{m}_d^A(x_d)$, 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 $\hat{\tau}(x) = \hat{\psi}_d \cdots \hat{\psi}_1[\hat{m}_0^A(x_0) - \hat{m}_{c,0}^A] + \ldots + \hat{\psi}_d[\hat{m}_{d-1}^A(x_{d-1}) - \hat{m}_{c,d-1}^A] + [\hat{m}_d^A(x_d) - \hat{m}_{c,d}^A]$ 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 $\hat{m}^A(x) = \hat{m}_0^A(x_0) + ... + \hat{m}_d^A(x_d)$ and $\hat{m}_c^A = \hat{m}_{c,0}^A + \ldots + \hat{m}_{c,d}^A$. 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.

$$\sum_{r=1}^{\infty} \hat{S}^r(\hat{m}^A(x) - \hat{m}_c^A) = \frac{1}{T}\sum_{t=1}^{T} a_t(x)\varepsilon_t + o_p\Big(\frac{1}{\sqrt{T}}\Big) \tag{41}$$

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 result.

The idea behind the proof of (41) is as follows: From the definition of the operators $\hat{\psi}_j$, it can be seen that

$$\hat{S}(\hat{m}^A(x) - \hat{m}_c^A) = \sum_{j=0}^{d-1} \hat{\psi}_d \cdots \hat{\psi}_{j+1}\Big(\sum_{k=j+1}^{d} S_{j,k}(x_j)\Big) \tag{42}$$

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

$$S_{j,k}(x_j) = -\frac{1}{T}\sum_{t=1}^T \Big(\frac{p_{j,k}(x_j,X_t^k)}{p_j(x_j)p_k(X_t^k)} - 1\Big)\varepsilon_t + o_p\Big(\frac{1}{\sqrt{T}}\Big) \tag{43}$$

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

$$\hat{S}(\hat{m}^A(x) - \hat{m}_c^A) = \frac{1}{T}\sum_{t=1}^T b_t(x)\varepsilon_t + o_p\Big(\frac{1}{\sqrt{T}}\Big) \tag{44}$$

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

$$\sum_{r=0}^\infty \hat{S}^r(\hat{m}^A(x) - \hat{m}_c^A) = \sum_{r=0}^\infty S^{r-1}\hat{S}(\hat{m}^A(x) - \hat{m}_c^A) + o_p\Big(\frac{1}{\sqrt{T}}\Big) \tag{45}$$

uniformly in x, where S is defined analogously to $\hat{S}$ with the density estimates replaced by the true densities. Combining (44) and (45) completes the proof.

To show (43), we exploit the mixing behaviour of the variables X_t. Plugging the definition of $\hat{m}_k^A$ into the term $S_{j,k}$, we can write

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

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

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

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

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

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. □

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

Theorem B2. *Suppose that (C1) – (C5) hold. If the bandwidth h satisfies (C6a), then*

$$\sup_{x_j \in I_h} |\tilde{m}_j^B(x_j) - m_j(x_j)| = O_p(h^2) \tag{46}$$

$$\sup_{x_j \in I_h^c} |\tilde{m}_j^B(x_j) - m_j(x_j)| = O_p(h) \tag{47}$$

for $j = 0, \ldots, d$. If the bandwidth satisfies (C6b), we have

$$\sup_{x_j \in I_h} \left| \tilde{m}_j^B(x_j) + \frac{1}{T} \sum_{t=1}^{T} m_j(X_t^j) - m_j(x_j) \right| = O_p(h^2) \tag{48}$$

$$\sup_{x_j \in I_h^c} \left| \tilde{m}_j^B(x_j) + \frac{1}{T} \sum_{t=1}^{T} m_j(X_t^j) - m_j(x_j) \right| = O_p(h) \tag{49}$$

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

Proof. The result follows from Theorem 3 in Mammen et al. [12]. To make sure that the latter theorem applies in our case, we have to show that the high-order conditions (A1) – (A5), (A8), and (A9) from [12] are fulfilled in our setting.[7] This can be achieved by using the results from Appendix A, in particular the expansion of $\hat{m}_j^B$ given in Lemma A3, and by following the arguments for the proof of Theorem 4 in [12]. To see that (46) – (47) have to be replaced by (48) – (49) 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 A3. Using this in the proof of Theorem 3 of [12] 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 (48) – (49). □

By combining Theorems B1 and B2, it is now straightforward to complete the proof of Theorem 4.2. □

Appendix C - Proof of Theorems 4.3 and 4.4

This appendix contains the proofs of Theorems 4.3 and 4.4, which show consistency and asymptotic normality of the AR estimates. 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

[7] Note that (A6) is not needed for the proof of Theorem 3 as opposed to the statement in [12].

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 C1 – C4. Note that as in Appendix B, C denotes a finite real constant which 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

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

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 (17) and (19) 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 estimates $\tilde{m}_\theta, \tilde{m}_0, \ldots, \tilde{m}_d$, it is easily seen that

$$\max_{t=1,\ldots,T} |\varepsilon_t - \tilde{\varepsilon}_t| = O_p(h). \tag{R1}$$

Using (R1), we can immediately infer that

$$\max_{t=1,\ldots,T} \sup_{\phi\in\Phi} |\varepsilon_t(\phi) - \tilde{\varepsilon}_t(\phi)| = O_p(h). \tag{R2}$$

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

$$\max_{t=1,\ldots,T} \sup_{\phi\in\Phi} \Big| \frac{\partial \varepsilon_t(\phi)}{\partial \phi_i} - \frac{\partial \tilde{\varepsilon}_t(\phi)}{\partial \phi_i} \Big| = O_p(h). \tag{R3}$$

Proof of Theorem 4.3

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

$$\sup_{\phi\in\Phi} \Big| \frac{1}{T}\tilde{l}_T(\phi) - \frac{1}{T} l_T(\phi) \Big| = o_p(1). \tag{50}$$

This together with standard arguments yields consistency of $\tilde{\phi}$. In order to prove (50), we decompose $\frac{1}{T}\tilde{l}_T(\phi) - \frac{1}{T}l_T(\phi)$ into

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

Using (R1) – (R3), 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 (50). □

Proof of Theorem 4.4

By the usual Taylor expansion argument, we obtain

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

with $\tilde{\mathcal{H}}_T(\tilde{\phi},\phi^*)$ 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

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

In what follows, we show that

$$\frac{1}{T}\tilde{\mathcal{H}}_T(\tilde{\phi},\phi^*) \xrightarrow{P} H \tag{51}$$

$$\frac{1}{\sqrt{T}}\frac{\partial\tilde{l}_T(\phi^*)}{\partial\phi} \xrightarrow{d} N(0,\Psi) \tag{52}$$

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 (61). This completes the proof.

Proof of (51). By straightforward calculations it can be seen that

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

Defininig the $p \times p$ matrix $\mathcal{H}_T(\tilde{\phi},\phi^*)$ analogously to $\tilde{\mathcal{H}}_T(\tilde{\phi},\phi^*)$ with $\tilde{l}_T$ replaced by l_T it is easy to show that $\frac{1}{T}\mathcal{H}_T(\tilde{\phi},\phi^*) \xrightarrow{P} H$, yielding (51). □

Proof of (52). We write

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

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

$$\begin{aligned}\frac{1}{\sqrt{T}}\frac{\partial \tilde{l}_T(\phi^*)}{\partial \phi_i} - \frac{1}{\sqrt{T}}\frac{\partial l_T(\phi^*)}{\partial \phi_i} &= \sum_{k=0}^{p} 2\phi_k^* \Big(\frac{1}{\sqrt{T}}\sum_{t=1}^{T}(\varepsilon_{t-k} - \tilde{\varepsilon}_{t-k})\varepsilon_{t-i}\Big) \\ &\quad + \sum_{k=0}^{p} 2\phi_k^* \Big(\frac{1}{\sqrt{T}}\sum_{t=1}^{T}(\varepsilon_{t-i} - \tilde{\varepsilon}_{t-i})\tilde{\varepsilon}_{t-k}\Big) \\ &= \sum_{k=0}^{p} 2\phi_k^* \Big(\frac{1}{\sqrt{T}}\sum_{t=1}^{T}(\varepsilon_{t-k} - \tilde{\varepsilon}_{t-k})\varepsilon_{t-i}\Big) \\ &\quad + \sum_{k=0}^{p} 2\phi_k^* \Big(\frac{1}{\sqrt{T}}\sum_{t=1}^{T}(\varepsilon_{t-i} - \tilde{\varepsilon}_{t-i})\varepsilon_{t-k}\Big) + o_p(1), \quad (53)\end{aligned}$$

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 (R1). 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 (53) 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

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

From Appendix B, we know that the backfitting estimates $\tilde{m}_j(x_j)$ 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

$$Q_T = Q_{T,\theta} + \sum_{j=0}^{d} Q_{T,V,j} + \sum_{j=0}^{d} Q_{T,B,j} \quad (54)$$

with

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

for $j = 0, \dots, d$. In Lemmas C3 and C4, we will show that

$$Q_{T,\theta} = o_p(1) \tag{55}$$

$$Q_{T,B,j} = o_p(1) \quad \text{for } j = 0, \dots, d. \tag{56}$$

Moreover, Lemmas C1 and C2 establish that

$$Q_{T,V,0} = o_p(1) \tag{57}$$

$$Q_{T,V,j} = \frac{1}{\sqrt{T}} \sum_{t=1}^{T} g_j\Big(\frac{t}{T}, X_t\Big)\varepsilon_t + o_p(1) \quad \text{for } j = 1, \dots, d, \tag{58}$$

where $g_j = g_j^{[k,i]}$ are deterministic functions whose exact forms are given in the statement of Lemma C1. These functions are easily seen to be absolutely bounded by a constant independent of T. Inserting the above results in (54), we obtain

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

Using this together with (53) now yields

$$\frac{1}{\sqrt{T}} \frac{\partial \tilde{l}_T(\phi^*)}{\partial \phi_i} - \frac{1}{\sqrt{T}} \frac{\partial l_T(\phi^*)}{\partial \phi_i} = \frac{1}{\sqrt{T}} \sum_{t=1}^{T} h_i\Big(\frac{t}{T}, X_t\Big)\varepsilon_t + o_p(1) \tag{59}$$

with the absolutely bounded function

$$h_i\Big(\frac{t}{T}, X_t\Big) = \sum_{j=1}^{d} \sum_{k=0}^{p} 2\phi_k^* \Big[g_j^{[k,i]}\Big(\frac{t}{T}, X_t\Big) + g_j^{[i,k]}\Big(\frac{t}{T}, X_t\Big)\Big], \tag{60}$$

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

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

i.e. the term of interest can be written as a normalized sum of random variables $U_{t,T}$ plus a term which is asymptotically negligible. Using the mixing assumptions in (C1), 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 $\psi_{ii} = \lim_{T\to\infty} \mathbb{E}(\frac{1}{\sqrt{T}} \sum_{t=1}^T U_{t,T})^2$. Using the Cramer-Wold device, it is now easy to 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

$$\begin{aligned} \omega_{ij} = \frac{1}{2} \sum_{l=-\infty}^{\infty} \mathbb{E}\Big[\eta_0 \varepsilon_{-i} \varepsilon_l \int_0^1 h_j(u, X_l) du\Big] + \frac{1}{2} \sum_{l=-\infty}^{\infty} \mathbb{E}\Big[\eta_0 \varepsilon_{-j} \varepsilon_l \int_0^1 h_i(u, X_l) du\Big] \\ + \frac{1}{4} \sum_{l=-\infty}^{\infty} \mathbb{E}\Big[\varepsilon_0 \varepsilon_l \int_0^1 h_i(u, X_0) h_j(u, X_l) du\Big]. \qquad (61) \end{aligned}$$

□

In order to complete the proof of asymptotic normality, we still need to show that equations (55) – (58) 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 the technically most interesting part.

Lemma C1. *It holds that*

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

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

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

with

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

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 B1 of Appendix B.

Proof. By Theorem B1, the stochastic part $\tilde{m}_j^A$ of the smooth backfitting estimate $\tilde{m}_j$ has the expansion

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

uniformly in x_j, where $\hat{m}_j^A$ is the stochastic part of the Nadaraya-Watson pilot estimate 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:

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

In the following, we will give the arguments needed to treat $Q_{T,V,j}^{NW}$. The line of argument for $Q_{T,V,j}^{SBF}$ is essentially identical although some of the steps are easier due to the properties of the $r_{j,s}$ functions.
Plugging the definition (39) of the estimate $\hat{m}_j^A(x_j)$ into the term $Q_{T,V,j}^{NW}$, we get

$$Q_{T,V,j}^{NW} = \frac{1}{\sqrt{T}}\sum_{s=1}^{T}\Big(\frac{1}{T}\sum_{t=1}^{T}\frac{K_h(X_{t-k}^j, X_s^j)}{\frac{1}{T}\sum_{v=1}^{T} K_h(X_{t-k}^j, X_v^j)}\varepsilon_{t-i}\Big)\varepsilon_s. \tag{62}$$

In a first step, we show that

$$Q_{T,V,j}^{NW} = \frac{1}{\sqrt{T}}\sum_{s=1}^{T}\Big(\frac{1}{T}\sum_{t=1}^{T} K_h(X_{t-k}^j, X_s^j)\mu_t\Big)\varepsilon_s + o_p(1), \tag{63}$$

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

$$\begin{aligned} B_j(x_j) &= \frac{1}{T}\sum_{v=1}^{T}\mathbb{E}[K_h(x_j, X_v^j)] - q_j(x_j) \\ V_j(x_j) &= \frac{1}{T}\sum_{v=1}^{T}\big(K_h(x_j, X_v^j) - \mathbb{E}[K_h(x_j, X_v^j)]\big). \end{aligned}$$

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

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

uniformly in x_j. Plugging this decomposition into (62), we obtain

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

with

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

All that is required to establish (63) is to show that both $Q_{T,V,j}^{NW,B}$ and $Q_{T,V,j}^{NW,V}$ 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 (C9) to get that $Q_{T,V,j}^{NW,B} = o_p(1)$. In order to show that $Q_{T,V,j}^{NW,V} = o_p(1)$, let $\mathbb{E}_v[\cdot]$ denote the expectation with respect to the variables indexed by v. Then

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

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 (63) 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 (63) as

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

where

$$R_{T,V,j}^{NW} = \frac{1}{\sqrt{T}} \sum_{s=1}^{T} \frac{1}{T} \sum_{t=1}^{T} \psi_{t,s} \varepsilon_s. \tag{64}$$

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

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

with $\mu_0 = q_j^{-1}(X_{-k}^j)\varepsilon_{-i}$ and $q_j(X_{-k}^j) = \int_0^1 K_h(X_{-k}^j, w) dw \; p_j(X_{-k}^j)$. Thus it remains to prove that $R_{T,V,j}^{NW} = o_p(1)$. To do so, define

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

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

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

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 $\min\{|t - t'|, |t - s|, |t - s'|\} > C_2 \log T$, i.e. 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 analyse 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. Since the index t is not separated, there exists an index, say 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, say 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 (C8), this suffices to infer that

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

(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

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

for $r = 1, \ldots, 4$. In what follows, we show that $P_{S_r} \to 0$ for $r = 1, \ldots, 4$. 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 $m = 1, \ldots, M_T - 1$ 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

$$
\begin{aligned}
K_h(X_{t-k}^j, X_s^j) = \sum_{m=1}^{M_T} & I(X_{t-k}^j \in I_m) \\
& \times \big[K_h(x_m, X_s^j) + (K_h(X_{t-k}^j, X_s^j) - K_h(x_m, X_s^j))\big]. \qquad (65)
\end{aligned}
$$

Using (65), we can further write

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

and

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

We first consider $P^B_{S_1}$. Set $M_T = CT(\log T)h^{-3}$ and exploit the Lipschitz continuity of the kernel K to get that $|K_h(X^j_{t-k}, X^j_s) - K_h(x_m, X^j_s)| \le \frac{C}{h^2}|X^j_{t-k} - x_m|$. This gives us

$$\begin{aligned} \big|\psi^B_{t,s}\big| \le \frac{C}{h^2} \sum_{m=1}^{M_T} \Big(& \underbrace{I(X^j_{t-k} \in I_m)|X^j_{t-k} - x_m|}_{\le I(X^j_{t-k} \in I_m) M_T^{-1}} |\mu_t| \\ & + \mathbb{E}\big[\underbrace{I(X^j_{t-k} \in I_m)|X^j_{t-k} - x_m|}_{\le I(X^j_{t-k} \in I_m) M_T^{-1}} |\mu_t|\big] \Big) \le \frac{C}{M_T h^2} \big(|\mu_t| + \mathbb{E}|\mu_t|\big). \end{aligned}$$

Plugging this into the expression for $P^B_{S_1}$, we arrive at

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

We next turn to $P^A_{S_1}$. Write

$$P^A_{S_1} = \frac{1}{T^3\delta^2} \sum_{(s,s',t,t') \in S_1} \Big(\sum_{m=1}^{M_T} \gamma_m \Big)$$

with

$$\begin{aligned} \gamma_m = \mathbb{E}\Big[\big\{ & I(X^j_{t-k} \in I_m) K_h(x_m, X^j_s)\mu_t \\ & - \mathbb{E}_{-s}[I(X^j_{t-k} \in I_m) K_h(x_m, X^j_s)\mu_t] \big\} \varepsilon_s \psi_{t',s'} \varepsilon_{s'} \Big]. \end{aligned}$$

By Davydov's inequality, it holds that

$$\begin{aligned} \gamma_m &= \mathrm{Cov}\Big(I(X^j_{t-k} \in I_m)\mu_t - \mathbb{E}[I(X^j_{t-k} \in I_m)\mu_t], K_h(x_m, X^j_s) \varepsilon_s \psi_{t',s'} \varepsilon_{s'} \Big) \\ &\le \frac{C}{h^2} \big(\alpha(C_2 \log T)\big)^{1 - \frac{1}{q} - \frac{1}{r}} \le \frac{C}{h^2} \big(a^{C_2 \log T}\big)^{1 - \frac{1}{q} - \frac{1}{r}} \le \frac{C}{h^2} T^{-C_3} \end{aligned}$$

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 $P^A_{S_1} \to 0$.

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

$$R^{NW,V}_{T,V,j} = o_p(1),$$

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

Lemma C2. *It holds that*

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

Proof. As in Lemma C1, we can write

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

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

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

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, e.g. $\rho_T = (\log T)^{-1}$. By Chebychev's inequality,

$$\mathbb{P}\big(|Q_{T,V,0}^{NW}| > C\rho_T\big) \le C \frac{\mathbb{E}(Q_{T,V,0}^{NW})^2}{\rho_T^2}$$

with

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

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, 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 $\mathbb{E}(Q_{T,V,j}^{NW})^2/\rho_T^2$ goes to zero, which in turn yields that $Q_{T,V,j}^{NW} = o_p(1)$. □

Note that the above argument for $Q_{T,V,0}$ is much easier than that for $Q_{T,V,j}$ presented in Lemma C1. The main reason is that the weights $w_{t,s}$ and $w_{t',s'}$ are deterministic allowing us to separate the expectations $\mathbb{E}[\varepsilon_{t-i} \varepsilon_s \varepsilon_{t'-i} \varepsilon_{s'}]$ from the weights. In contrast, in Lemma C1 we have the situation that

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

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,

$$\mathbb{E}(Q_{T,V,j}^{NW})^2 = \frac{1}{T^3} \sum_{s,s',t,t'=1}^{T} \mathbb{E}[w_{t,s} w_{t',s'} \varepsilon_{t-i} \varepsilon_s \varepsilon_{t'-i} \varepsilon_{s'}]. \tag{66}$$

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 $Q_{T,V,0}^{\mathrm{NW}}$ yield that $Q^{\mathrm{NW}}_{T,V,j} = o_p(1)$. However, if we allow X_t and ε_t to be dependent, then the expectations in (66) do not split up into two separate parts any more. Moreover, since the weights $w_{t,s}$ and $w_{t',s'}$ depend on all the X^j_t 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 C1 to exploit the covariance structure of the processes $\{X_t\}$ and $\{\varepsilon_t\}$.

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

Lemma C3. *It holds that*

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

Proof. We write

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

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

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

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

Lemma C4. *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 $I_h = [2C_1 h, 1 - 2C_1 h]$ and $I^c_h = [0, 2C_1 h) \cup (1 - 2C_1 h, 1]$ as defined in Theorem 4.2. Using the uniform convergence rates from Theorem B2, we get

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

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 (C9), $\mathbb{E}[|\varepsilon_{t-i}|I(X^j_{t-k} \notin I_h)] \le Ch$ 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

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

As $\sum_{t=1}^T I(\frac{t-k}{T} \in I_h^c) \le CTh$ 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. □

References

[1] Altman, N. S. (1990). Kernel smoothing of data with correlated errors. *Journal of the American Statistical Association* **85** 749-759.

[2] Altman, N. S. (1993). Estimating error correlation in nonparametric regression. *Statistics & Probability Letters* **18** 213-218.

[3] Bosq, D. (1998). *Nonparametric statistics for stochastic processes: estimation and prediction, 2nd ed.* Springer, New York.

[4] Gonçalves, S. & Kilian, L. (2004). Bootstrapping autoregressions with conditional heteroskedasticity of unknown form. *Journal of Econometrics* **123** 89-120.

[5] Hall, P. & van Keilegom, I. (2003). Using difference-based methods for inference in nonparametric regression with time series errors. *Journal of the Statistical Royal Society B* **65** 443-456.

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

[7] Hart, J. D. (1991). Kernel regression estimation with time series errors. *Journal of the Statistical Royal Society B* **53** 173-187.

[8] Hart, J. D. (1994). Automated kernel smoothing of dependent data by using time series cross-validation. *Journal of the Statistical Royal Society B* **56** 529-542.

[9] Herrmann, E., Gasser, T. & Kneip, A. (1992). Choice of bandwidth for kernel regression when residuals are correlated. *Biometrika* **79** 783-795.

[10] 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.

[11] Lin, T. C., Pourahmadi, M. & Schick, A. (1999). Regression models with time series errors. *Journal of Time Series Analysis* **20** 425-433.

[12] 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.

[13] Mammen, E. & Park, B. U. (2005). Bandwidth selection for smooth backfitting in additive models. *The Annals of Statistics* **33** 1260-1294.

[14] Mammen, E. & Park, B. U. (2006). A simple smooth backfitting method for additive models. *The Annals of Statistics* **34** 2252-2271.

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

[16] Mudelsee, M. (2010). Climate Time Series Analysis: Classical Statistical and Bootstrap Methods. *Atmospheric and oceanographic sciences library.* Springer. New York.

[17] Schick, A. (1994). Estimation of the autocorrelation coefficient in the presence of a regression trend. *Statistics & Probability Letters* **21** 371-380.

[18] Shao, Q. & Yang, L. J. (2011). Autoregressive coefficient estimation in nonparametric analysis. *Journal of Time Series Analysis* **32** 587-597.

[19] Truong, Y. K. (1991). Nonparametric curve estimation with time series errors. *Journal of Statistical Planning and Inference* **28** 167-183.

[20] Truong, Y. K. & Stone, C. (1994). Semiparametric time series regression. *Journal of Time Series Analysis* **15** 405-428.

[21] Turner, J., King, J. C., Lachlan-Cope, T. A. & Jones, P. D. (2002). Recent temperature trends in the Antarctic. *Nature* **418** 291-292.

[22] Turner, J., Colwell, S. R., Marshall, G. J., Lachlan-Cope, T. A., Carleton, A. M., Jones, P. D., Lagin, V., Reid, P. A. & Igovkina, S. (2005). Antarctic climate change during the past 50 years. *International Journal of Climatology* **25** 279-294.

[23] Yu, K., Mammen, E. & Park, B. U. (2011). Semi-parametric regression: efficiency gains from modeling the nonparametric part. *Bernoulli* **17** 736-748.