Nonparametric prediction of nonstationary spatio-temporal processes[1]

Jan Johannes[2] Suhasini Subba Rao[3]

May 31, 2006

Abstract

In spatial statistics often the response variable at a given location and time is observed together with some covariates which are known to influence the response. In several applications the relationship between the response and covariates may be unknown, and to prevent misspecification of the model, a nonparametric approach could be appropriate. In this paper prediction and forecasting of the response variable, for spatially nonstationary, spatio-temporal processes, within a nonparametric framework is developed. The linear prediction of the response, which involves estimation of the covariance structure, and also the more general optimal predictor are investigated. The asymptotic sampling properties of the predictors are studied. It is shown that in order to avoid the curse of dimensionality the covariance estimator should be defined in terms of the dependence structure of the spatio-temporal process. Furthermore the rate of convergence of the prediction estimators depend on the temporal dependence of the covariates and the mixing rates of the spatio-temporal process.

The model defined and the estimation methodology has many possible applications. We consider a specific application and illustrate our methodology by modelling house prices in the Stockport area, United Kingdom, using the deprivation index and district as the covariates.

Keywords and phrases Covariates, mixing, multivariate time series, nonparametric regression, kernel estimation, spatio-temporal processes.

[1]This work was partially supported by the DFG (DA 187/12-3).

[2]Universität Heidelberg, Institut für Angewandte Mathematik, Im Neuenheimer Feld, 294, D-69120 Heidelberg, Germany, johannes@statlab.uni-heidelberg.de

[3]University of Bristol, Department of Mathematics, University Walk, Bristol, BS8 1TW, U.K. S.SubbaRao@bristol.ac.uk

1 Introduction

In spatial statistics often the response variable $Y_t(w,x)$ is observed at time t, together with the location $w \in \mathbb{R}^2$ and some covariates $x = (x_1, \ldots, x_q) \in \mathbb{R}^q$ which are known to influence the response variable. A standard method for modelling the response is to use a multiple regression model with spatio-temporal errors (cf. Cressie (1993), Guttorp, Meiring, and Sampson (1994), Lesch, Strauss, and Rhoades (1995), Cressie and Huang (1999)). However, there are several real situations where the influence of the covariates x may also depend on the location w. To model the location dependent influence of the covariates, one often uses the model

$$Y_t(w,x) = \sum_{j=1}^{q} x_j f_j(w) + \xi_t(w), \tag{1.1}$$

where $\{f_j(\cdot)\}$ are nonrandom functions and $\{\xi_t(\cdot)\}$ is a stationary spatio-temporal process (cf. Yakowitz and Szidaravosky (1985), Luo and Wahba (1998)). An alternative approach advocated in Gelfand, Kim, Sirmans, and Banerjee (2003) is to treat the coefficients of x as if they were random, stationary spatio-temporal processes and write

$$Y_t(w,x) = \sum_{j=1}^{q} x_j \xi_t^{(j)}(w) + \xi_t^{(0)}(w) + V_{t,w,x}, \tag{1.2}$$

where $\{V_{t,w,x}\}$ are independent, identically distributed (iid) random variables and $\{\xi_t^{(j)}(\cdot)\}$ are stationary spatio-temporal processes with separable covariances from a known family of distributions (eg. the Matèrn family, see Matèrn (1986)). We observe that if we were to treat the covariates x as a function of w, $x(w)$,then $Y_t(w, x(w))$ is a spatially nonstationary process. Gelfand, Kim, Sirmans, and Banerjee (2003) use model (1.2) to model house prices in Baton Rouge, LA, U.S.A, where w is the location and x are a range of house characteristics such as size of living area etc.

In certain applications the nature of the relationship between the covariates and response variable may be unknown, as well as the family of distributions the spatio-temporal process belongs to. In such situations, to prevent misspecification of the model, a nonparametric approach may be appropriate. The approach that we adopt is to redefine the location vector $u = (w,x) \in \Omega \subseteq \mathbb{R}^d$, and suppose $\{\Phi_t(u); u \in \Omega, t \in \mathbb{Z}\}$ is a spatio-temporal process, where for each $t \in \mathbb{Z}$, $\{\Phi_t(u); u \in \Omega\}$ is a nonstationary spatial process on the region Ω and $\{\Phi_t(\cdot)\}_t := \{\Phi_t(\cdot); t \in \mathbb{Z}\}$ is a stationary infinite dimensional process (which implies that for every fixed $u \in \Omega$, $\{\Phi_t(u)\}_t$ is a stationary time series). The model we consider in this paper is

$$Y_t(u) = \Phi_t(u) + V_{t,u}, \tag{1.3}$$

where $\{V_{t,u}\}$ are iid random variables. The model (1.3) includes as special cases both the partially linear models (1.1) and (1.2). Furthermore it allows additional flexibility, for

example, we do not require $\{\Phi_t(\cdot)\}$ to have a known parametric form and there may be situations where there is no realistic reason to distinguish between covariates and locations. From now on, unless stated otherwise, we shall refer to $u = (u^{(1)}, \ldots, u^{(d)}) \in \Omega$ as the covariate vector.

Here our object is to predict $\Phi_t(u_0)$, given the observations $\{Y_{t-s}(u_i)\}_{i=1}^{m-1}$ at the time lag $s \in \mathbb{Z}$, for any $u_0 \in \Omega$. In this paper we will consider a nonparametric approach for prediction and forecasting for spatially nonstationary, spatio-temporal processes. In particular we will consider the best linear predictor and also the more general optimal predictor (under the mean squared criterion), $\psi : \mathbb{R}^{d(m+1)-1} \to \mathbb{R}$ defined for all $t \in \mathbb{Z}$ by

$$\psi(\underline{y}, \underline{u}, u_0) = \mathbb{E}(\Phi_t(u_0)|\{Y_{t-s}(u_i) = y_i\}_{i=1}^{m-1}), \tag{1.4}$$

where $\underline{u} = (u_1, \ldots, u_{m-1}) \in \Omega^{m-1}$ and $\underline{y} = (y_1, \ldots, y_{m-1}) \in \mathbb{R}^{m-1}$. We observe that since $\{\Phi_t(\cdot)\}_t$ is a stationary process, $\psi(\cdot)$ does not depend on t. We mention that even though the above predictor function is defined for only one lag s it is easy to generalise the methods and results in this paper to estimate the prediction function $\psi(\underline{y}, \underline{u}, u_0) = \mathbb{E}(\Phi_t(u_0)|\{Y_{t-\tau_i}(u_i) = y_i\}_{i=1}^{m-1})$ for several time lags $\tau_1, \ldots, \tau_m$. We use only one lag to reduce notation.

The process $\{Y_t(\cdot)\}_t$, will only be observed on a finite set of covariate values, which we denote as the random variables $\{U_{t,i}\}$ that take values in Ω. We will suppose that for i fixed $\{U_{t,i}\}_t$ is a time series and we observe $\{(Y_t(U_{t,i}), U_{t,i}); t = 1, \ldots, T, i = 1, \ldots, N\}$, where

$$Y_t(U_{t,i}) = \Phi_t(U_{t,i}) + V_{t,i}, \qquad t = 1, \ldots T, i = 1, \ldots, N, \tag{1.5}$$

and $\{V_{t,i}\}$ are iid random variables with $\mathbb{E}(V_{t,i}) = 0$. Though we shall assume that the error terms $\{V_{t,i}\}_{t,i}$, $\{U_{t,i}\}_{t,i}$ and the process $\{\Phi_t(\cdot)\}_t$ are independent of each other. To illustrate our methods, in Section 6 we model the selling price of houses, sold over a period time in Stockport, UK, and use the district of the house and deprivation index as covariates (though other factors such as age and size of properties could also be used as covariates, when available). In this application the values of the covariate can be different at each time and can be assumed to be random. Furthermore the deprivation index in any given district, is evaluated every two years and can be viewed is a dependent time series.

The crucial point that allows us to use the observations $\{Y_t(U_{t,i})\}$, defined in (1.5), to estimate $\psi(\cdot)$ is that at a given point $(\{y_i, u_i\}_{i=1}^{m-1}, u_0)$ and for any collection $\{i_0, \ldots, i_{m-1}\} \subset \{1, \ldots, N\}$ and for all $t = s+1, \ldots, T$

$$\psi(\underline{y}, \underline{u}, u_0) = \mathbb{E}(Y_t(U_{t,i_0})|\{Y_{t-s}(U_{t-s,i_l}) = y_l, U_{t-s,i_l} = u_l\}_{l=1}^{m-1}, U_{t,i_0} = u_0). \tag{1.6}$$

This formulation motivates us to estimate ψ in a nonparametric way.

In Section 2 we will define the model and the mixing assumptions that will be used in the paper. We work under the relatively weak assumption that for any fixed u and i, $\{\Phi_t(u)\}_t$ and $\{U_{t,i}\}$ are mixing, which we use to show that the composite random process $\{\Phi_t(U_{t,i})\}_t$ is also mixing.

In Section 3 we consider nonparametric estimation in the context of multivariate time series. These results motivate the prediction estimators and unify the theory in the subsequent sections. But we believe they are also of wider interest, and can be applied to other problems.

In Section 4 we consider linear prediction. It is clear there exists a function $a(\cdot)' = (a_1(\cdot), \ldots, a_{m-1}(\cdot))$, where $a : \mathbb{R}^{md} \to \mathbb{R}^{m-1}$, such that

$$Y_t(u_0) = \sum_{j=1}^{m-1} a_j(\underline{u}, u_0) Y_{t-s}(u_j) + \sigma(\underline{u}, u_0)\varepsilon_{t,\underline{u}}, \tag{1.7}$$

where $\underline{u} = (u_1, \ldots, u_{m-1})$, $\varepsilon_{t,\underline{u}}$ is uncorrelated with $\{Y_{t-s}(u_j)\}_{j=1}^{m-1}$, $\mathbb{E}(\varepsilon_{t,\underline{u}}) = 0$ and $\mathrm{var}(\varepsilon_{t,\underline{u}}) = 1$. This yields the linear predictor

$$\psi_L(\{y_{t-s,i}, u_i\}_{i=1}^{m-1}, u_0) = a(\underline{u}, u_0)'\underline{y}$$

where $\underline{y}' = (y_{t-s,1}, \ldots, y_{t-s,m-1})$. Now for Gaussian $Y_t(u)$, we have $\psi_L(\cdot) \equiv \psi(\cdot)$, however if $Y_t(u)$ is non-Gaussian, then $\psi_L(\cdot)$ is the best linear predictor. Since the coefficients $a(\underline{u}, u_0)$ can be obtained from the covariance function $c_s(u_i, u_j) = \mathrm{cov}(Y_t(u_i), Y_{t-s}(u_j))$, we estimate the function $a(\cdot)$ using estimators of the covariance function $c_s(\cdot)$. In particular we consider nonparametric methods for estimating the covariance function, which we use to estimate the function $\psi_L(\cdot)$. We mention that nonparametric estimators of spatial covariances for locations which are fixed over time have been considered previously (cf. Sampson and Guttorp (1992)). In contrast, we consider here estimators of the covariance function for covariates whose values can change over time. In fact, we show that changing covariate values lead to estimators of the covariance function which are consistent as T grows (even for fixed N). Though the rate of convergence depends on several factors; the temporal dependence of the covariates (which we discuss later) and the dependence structure of the spatio-temporal process $\Phi_t(\cdot)$. Addressing the latter point; without any additional assumptions on the process the rate of convergence declines as the dimension d of covariates grows. On the other hand under the much stronger assumption that $\Phi_t(\cdot)$ is a spatially isotropic process the rate of convergence is independent of dimension. Nevertheless, we show that a compromise between the generality of nonstationary and the more restrictive isotropy property is possible. That is, under additional dependence constraints (for example the representations in (1.1) and (1.2)), it is possible to define an estimator of the covariance function of a spatially nonstationary process, whose rate of convergence is independent of dimension. It is worth mentioning that all the results in Section 4 include the case where the number of predictors in $\psi_L(\cdot)$ can be $m = N$ (even when we derive results for $N \to \infty$).

An advantage of the best linear estimator is that its rate of convergence does not depend on the number of covariates m used to define the function ψ_L. However when there is departure from Gaussianity the linear predictor can be far from the optimal predictor. In Section 5 we propose a direct nonparametric estimator of the prediction function ψ (defined

in (1.4)). In Section 5.2 we derive the asymptotic sampling properties of the estimator of ψ.

It is worth noting that due to the spatio-temporal nature of the problem the estimators considered in this paper yield different sampling properties to those often obtained in nonparametric statistics. More precisely, the rate of convergence depends heavily on the temporal dependence (characterised by the mixing rate) of the covariates $\{U_{t,i}\}_t$ and the number of spatial points at a given time. If there is only weak temporal dependence in the covariates then for any N, as $T \to \infty$, the estimator is consistent in probability. Whereas, if the covariates were the same at each time, the estimator is consistent only if $N \to \infty$ and $T \to \infty$. On the other hand if the temporal dependence of the covariates is very slow then for any N as $T \to \infty$ the estimator is consistent but the rate of convergence is slow. However by also allowing $N \to \infty$ at a sufficiently fast rate the usual rate of convergence (as in the weak dependence case) is achieved.

In Section 6 we illustrate our methods by modelling the selling price of properties in the Stockport area, U.K. We use as covariates the district number and the deprivation index. The data we use are the selling prices of several types of accomodation; detached houses, semi-detached houses, town houses and apartments. We show that that the proposed nonparametric linear predictor, predicts well the house price at unobserved locations for several types of houses. Moreover the linear dependence between house prices in areas of large deprivations seems to be much larger than the linear dependence between house prices in areas of low deprivation. Interestingly, this trend is observed over most housing types (with the exception of apartments).

All proofs can be found in the Appendix.

2 The model and observations

In this section we describe the model, observations and state the assumptions and notations we will use. We shall assume throughout that all the necessary densities exist.

We use the following assumptions in the paper.

Assumption 2.1. *$\{\Phi_t(\cdot)\}_t$ is a stationary process, which implies that for each $t \in \mathbb{Z}$, $\Phi_t(\cdot)$ is a nonstationary spatial process on the region Ω and for each $u \in \Omega$, $\{\Phi_t(u)\}_t$ is a stationary time series.*

Suppose we observe $\{(Y_{t,i}(U_{t,i}), U_{t,i}); i = 1, \ldots, N, t = 1, \ldots, T\}$, where $Y_{t,i}(U_{t,i})$ satisfies (1.5) (to minimise notation we let $Y_{t,i} \equiv Y_t(U_{t,i})$). We use the following assumptions.

Assumption 2.2.

(i) For fixed $m \in \mathbb{N}$ and arbitrary indices $(i_1, \ldots, i_{2m}) \in \{1, \ldots, N\}^{2m}$, the vector time series of covariates $\{(U_{t,i_1}, \ldots, U_{t,i_{2m}})\}_t$ is stationary.

(ii) The observation errors $\{V_{t,i}\}_{t,i}$ are iid random variables.

(iii) *The covariates* $\{U_{t,i}\}_{t,i}$, *the observation errors* $\{V_{t,i}\}_{t,i}$ *and the process* $\{\Phi_t(\cdot)\}_t$ *are independent.*

In order to obtain the sampling properties of the estimators defined in the sequel we need to show that the random process $\{\Phi_t(U_{t,i})\}_i$ is 2-mixing. This requires the following assumptions.

Suppose $\underline{u} = (u_1, \ldots, u_m) \in \Omega^m$, and let $\underline{\Phi}_t^{(s)}(\underline{u}) = (\Phi_t(u_1), \Phi_{t-s}(u_2), \ldots, \Phi_{t-s}(u_m))$. Define $\underline{U}_t^{(s,\underline{i})} = (U_{t,i_1}, \{U_{t-s,i_j}\}_{j=2}^m)$ for the distinct indices $\underline{i} = (i_1, \ldots, i_m) \in \{1, \ldots, N\}^m$ and denote by $P_{t,\tau}^{(s,\underline{i})}$ the joint distribution of $(\underline{U}_t^{(s,\underline{i})}, \underline{U}_\tau^{(s,\underline{i})})$.

ASSUMPTION 2.3.

(i) *The vector processes* $\{\underline{\Phi}_t^{(s)}(\underline{u})\}_t$ *are 2-mixing, where for all* $\underline{u}, \underline{v} \in \Omega^m$

$$\sup_{A \in \sigma(\underline{\Phi}_t^{(s)}(\underline{u})), \sigma(\underline{\Phi}_\tau^{(s)}(\underline{v}))} |P(A \cap B) - P(A)P(B)| \leq C_{2m}(\underline{u}, \underline{v})|t - \tau|^{-\beta}$$

and additionally $\int C_{2m}(\underline{u}, \underline{v}) P_{t,\tau}^{(s,\underline{i})}(d\underline{u}, d\underline{v}) < C_{2m} < \infty$, *where the constant* C_{2m} *does not depend on* $t, \tau \in \mathbb{Z}$ *and the distinct indices* $\underline{i} \in \mathbb{N}^m$.

(ii) *Suppose for all distinct indices* $\underline{i}, \underline{j} \in \mathbb{N}^m$ *the vector time series* $\{\underline{U}_t^{(s,\underline{i})}\}_t$ *and* $\{\underline{U}_t^{(s,\underline{j})}\}_t$ *are 2-mixing, that is*

$$\sup_{\substack{A \in \sigma(\underline{U}_t^{(s,\underline{i})}) \\ B \in \sigma(\underline{U}_\tau^{(s,\underline{j})})}} |P(A \cap B) - P(A)P(B)| \leq C \begin{cases} |t - \tau|^{-\alpha}; & \text{if } \underline{i} = \underline{j}, \\ |t - \tau|^{-\gamma}; & \text{if } (\underline{i}, \underline{j}) \text{ distinct indices in } \mathbb{N}^{2m}, \end{cases}$$

where the constant C *does not depend on* t, τ, $\underline{i}$ *and* $\underline{j}$.

REMARK 2.1 (The location dependent AR process). An example of a process which has dependence over time and satisfies the assumptions above, is the location dependent spatio-temporal AR process considered in Subba Rao (2005), where $\Phi_t(u)$ satisfies the representation

$$\Phi_t(u) = \sum_{j=1}^p a_j(u)\Phi_{t-j}(u) + \sigma_t(u)\xi_t(u) \tag{2.1}$$

and $\{\xi_t(u); u \in \Omega\}_t$ is a stationary spatial process which is independent over time. Suppose for all $u \in \Omega$, the absolute value of the roots of the characteristic polynomial $\lambda^p - \sum_{j=1}^p a_j(u)\lambda^{p-j}$, are less than δ, where $\delta < 1$. It is straightforward to show that this model satisfies Assumption 2.1. Furthermore, if the innovations $\{\xi_t(\cdot)\}_t$ were a sequence of independent stationary Gaussian spatial process, then we can show that Assumption 2.3(i) is satisfied, where

$$\sup_{A \in \sigma(\underline{\Phi}_{t,s}(\underline{u})), \sigma(\underline{\Phi}_{\tau,s}(\underline{v}))} |P(A \cap B) - P(A)P(B)| \leq C_{2m}\rho^{-|t-\tau|},$$

and $\delta < \rho < 1$. The proof is in the Appendix. □

REMARK 2.2. For reasons that are explained in Section 3, an implication of a condition in the later theorems, is that $\min(\beta, \gamma) > 2$. Loosely speaking this means that the spatio-temporal process is only weakly dependent over time, and if $i \neq j$, then there is very little dependence between the covariates $U_{t,i}$ and $U_{\tau,j}$ when the difference $|t - \tau|$ is large. On the other hand, α can take any value. This includes several cases of interest;

(a) The process $\{\underline{U}_t\}_t$, where $\underline{U}_t = (U_{t,1}, \ldots, U_{t,n})$, is a stationary vector autoregressive, process, in which case $\{\underline{U}_t\}_t$ is geometrically, strongly mixing (see Pham and Tran (1985)). This implies it is two-mixing with the rate $C\rho^{|t-\tau|}$, where $0 < \rho < 1$ (α and γ are same).

(b) At the other extreme the covariates are fixed, or change extremely slowly over time. That is for fixed i, $U_{t,i} \approx U_i$ (where $\approx$ denotes close to). Also suppose $\{U_i\}_i$ are independent random variables. These conditions imply $\alpha = 0$, and because of independence between covariates, roughly speaking, $\gamma = \infty$.

□

Define the random vector $\underline{Y}_t^{(s,\underline{i})} = (Y_{t,i_1}, \{Y_{t-s,i_j}\}_{j=2}^m)$. We now show that the composite random processes $\{(\underline{\Phi}_t^{(s)}(\underline{U}_t^{(s,\underline{i})}), \underline{U}_t^{(s,\underline{i})})\}_t$ and $\{(\underline{Y}_t^{(s,\underline{i})}, \underline{U}_t^{(s,\underline{i})})\}_t$ are also 2-mixing.

PROPOSITION 2.1. *Suppose Assumption 2.3 holds, and let* $\underline{W}_t^{(s,\underline{i})} = (\underline{\Phi}_t^{(s)}(\underline{U}_t^{(s,\underline{i})}), \underline{U}_t^{(s,\underline{i})})$ *or* $\underline{W}_t^{(s,\underline{i})} = (\underline{Y}_t^{(s,\underline{i})}, \underline{U}_t^{(s,\underline{i})})$. *Then for all distinct indices* $\underline{i}, \underline{j} \in \mathbb{N}^m$ *the vector time series* $\{\underline{W}_t^{(s,\underline{i})}\}_t$ *and* $\{\underline{W}_t^{(s,\underline{j})}\}_t$ *are* 2*-mixing with*

$$\sup_{\substack{A \in \sigma(\underline{W}_t^{(s,\underline{i})}) \\ B \in \sigma(\underline{W}_\tau^{(s,\underline{j})})}} |P(A \cap B) - P(A)P(B)| \leq C \begin{cases} |t-\tau|^{-\min(\alpha,\beta)}; & \text{if } \underline{i} = \underline{j}, \\ |t-\tau|^{-\min(\gamma,\beta)}; & \text{if } (\underline{i}, \underline{j}) \text{ distinct indices in } \mathbb{N}^{2m}, \end{cases}$$

where the constant C *does not depend on* t, τ, $\underline{i}$ *and* $\underline{j}$.

REMARK 2.3 (2-mixing of the spatio-temporal process). An immediate consequence of Proposition 2.1 is that for all $j \in \mathbb{N}$ the composite stochastic processes $\{\Phi_t(U_{t,j})\}_t$ and $\{Y_{t,j}\}_t$ are also 2-mixing with the rate

$$\sup_{A \in \mathcal{F}_0, B \in \mathcal{F}_t} |P(A \cap B) - P(A)P(B)| \leq Ct^{-\min(\alpha,\beta)}$$

where $\mathcal{F}_0 = \sigma(\Phi_0(U_{0,j}))$ or $\sigma(Y_{0,j})$ and $\mathcal{F}_t = \sigma(\Phi_t(U_{t,j}))$ or $\sigma(Y_{t,j})$. □

3 Nonparametric regression with multivariate time series

In this section we consider nonparametric estimation in the context of multivariate time series. The results in this section unify the theory in the following sections, where we consider estimators for specific prediction problems. Furthermore, we believe the generality of the results, give them wider appeal, for example, in nonparametric estimation for panel

time series with dependent panels. Nevertheless, though the methods developed here and their asymptotic sampling results are used in later sections, this section can be omitted on first reading.

Let us suppose we observe the multivariate time series $\{(X_{t,i}, Z_{t,i}); i = 1, \ldots, N\}_t$, where the $(1+\eta)$ dimensional random vector $(X_{t,i}, Z_{t,i})$ satisfies

$$\mathbb{E}[X_{t,i}|Z_{t,i} = z] = \varphi(z) \quad \forall z \in \mathbb{R}^\eta, t \in \mathbb{Z}, i \in \mathbb{N}, \tag{3.1}$$

and $\varphi : \mathbb{R}^\eta \to \mathbb{R}$ is an unknown function. The object in this section it to define an estimator for $\varphi(\cdot)$ and study its sampling properties. Our approach is sufficiently general for us not to impose a parametric structure on the multivariate time series, however we will assume it satisfies the following dependence structure.

ASSUMPTION 3.1 (Temporal dependence). *For all $i, j \in \mathbb{N}$, the vector time series $\{(X_{t,i}, Z_{t,i}, X_{t,j}, Z_{t,j})\}_t$ is stationary and 2-mixing where*

$$\sup_{A\in\sigma(X_{t,i},Z_{t,i}),B\in\sigma(X_{\tau,j},Z_{\tau,j})} |P(A\cap B) - P(A)P(B)| \leq C \begin{cases} |t-\tau|^{-\mathfrak{r}}; & \text{if } i = j, \\ |t-\tau|^{-\mathfrak{u}}; & \text{otherwise,} \end{cases}$$

and the constant C does not depend on i, j, t or τ.

In Section 4 and 5 we given examples which satisfy model (3.1) and Assumption 3.1.

In the theorem below, we allow $\mathfrak{r}$ to take any value but impose the restriction $\mathfrak{u} > 2$. Roughly speaking this means that the two vector time series $\{(X_{t,i}, Z_{t,i})\}_t$ and $\{(X_{t,j}, Z_{t,j})\}_t$ can be dependent, but $(X_{t,i}, Z_{t,i})$ and $(X_{\tau,j}, Z_{\tau,j})$ will become asymptotically independent over time, as $|t-\tau| \to \infty$. On the other hand, since there are no restrictions on $\mathfrak{r}$, the time series $\{(X_{t,j}, Z_{t,j})\}_t$ can be highly dependent over time, where, as we shall see below, the dependence affects the rate of convergence of the estimator.

Let $f_i(x, z)$ denote the joint density of the random vector $(X_{t,i}, Z_{t,i})$ for $i \in \mathbb{N}$, which due to stationarity does not depend on t (see Assumption 3.1). Moreover, let $f_i(z)$ denote the marginal density of $Z_{t,i}$. Using these densities we can rewrite (3.1) as

$$\varphi(z) = \mathbb{E}[X_{t,i}|Z_{t,i} = z] = \int x \frac{f_i(x,z)}{f_i(z)} dx =: \frac{g_i(z)}{f_i(z)}, \quad \forall z \in \mathbb{R}^\eta, t \in \mathbb{Z}, i \in \mathbb{N}.$$

Furthermore, using the last identity it is easily verified that

$$\varphi(z) = \frac{\frac{1}{n}\sum_{i=1}^n g_i(z)}{\frac{1}{n}\sum_{i=1}^n f_i(z)}, \quad \forall z \in \mathbb{R}^\eta, n \in \mathbb{N} \tag{3.2}$$

which motivates the following estimator of φ. We observe, if the random vectors $(X_{t,i}, Z_{t,i})$ for $i \in \mathbb{N}$ and $t \in \mathbb{Z}$ are identically distributed with joint density $f(x, z)$, then we obtain equation (3.2) with $g(z) = \int x f(x, z) dx$ and marginal density $f(z)$ of $Z_{t,i}$.

We now use a nonparametric kernel approach to estimate $\varphi(\cdot)$ from the observations $\{(X_{t,i}, Z_{t,i}); t = 1, \ldots, T; i = 1, \ldots, N\}$. Motivated by (3.2) we consider $\hat{\varphi}(z)$ as an estimator

of $\varphi(z)$ where

$$\hat{\varphi}(z) = \frac{\frac{1}{N}\sum_{i=1}^{N} \hat{g}_i(z)}{\frac{1}{N}\sum_{i=1}^{N} \hat{f}_i(z)}, \tag{3.3}$$

using for each $i = 1, \ldots, N$

$$\hat{g}_i(z) := \frac{1}{T}\sum_{t=1}^{T} X_{t,i} K_{b_i}(Z_{t,i} - z) \quad \text{and} \quad \hat{f}_i(z) := \frac{1}{T}\sum_{t=1}^{T} K_{b_i}(Z_{t,i} - z) \tag{3.4}$$

as estimators of $g_i(z)$ and $f_i(z)$, respectively, with for $z \in \mathbb{R}^\nu$, $K_b(z) := b^{-\nu}K(z/b)$, $b > 0$ is a bandwidth and K is multiplicative kernel, defined below (see Scott (1992)).

Having defined the estimator, in the rest of this section we study its sampling properties. In order to do this we require the following definitions and assumptions (which will be used throughout the paper).

DEFINITION 3.1. *For all $w = (w_1, \ldots, w_\eta) \in \mathbb{R}^\eta$, K is a multiplicative kernel of order r, i.e. $K(w) = \Pi_{i=1}^{\eta} \ell(w_i)$ where ℓ is a univariate, bounded, even function such that*

$$\int du\, \ell(u) = 1, \quad \int du\, u^i \ell(u) = 0$$

for all $i = 1, \ldots, r-1$ and there exists a constant S_K such that

$$[\int du\, |u|^r \ell(u)]^\eta = S_K.$$

In later sections we will customise the following assumptions to specific situations. It is worth bearing in mind that they are relatively mild and, roughly speaking, require that the densities and the conditional expectation of $X_{t,i}X_{t,j}$ are p-integrable.

ASSUMPTION 3.2. *[Technical assumptions]*

(i) For all $i \in \mathbb{N}$ and $t \in \mathbb{Z}$ we have $\mathbb{E}[|X_{t,i}|^p] < \infty$ for some $p > 2$ and let

$$g_i^{(p)}(z) := \mathbb{E}[X_{t,i}^r | Z_{t,i} = z] \cdot f_i(z).$$

Then the functions $(g_i^{(p)})^{1/p}$ and f_i are bounded by a constant δ_i and we define $q := 1 - 2/p$.

(ii) For each $t, \tau \in \mathbb{Z}$ and $i, j \in \mathbb{N}$ let $f_{i,j}^{(t,\tau)}$ denotes the joint density of $(Z_{t,i}, Z_{\tau,j})$ and let

$$g_{i,j}^{(t,\tau)}(z_1, z_2) := \mathbb{E}[X_{t,i}X_{\tau,j} | Z_{t,i} = z_1, Z_{\tau,j} = z_2] \cdot f_{i,j}^{(t,\tau)}(z_1, z_2).$$

(a) Define[1] $F_{i,j}^{(t,\tau)} := f_{i,j}^{(t,\tau)} - f_i \otimes f_j$. Then for some $p_F > 2$ there exists a constant C_F such that $\sup_{t,\tau,i,j} \|F_{i,j}^{(t,\tau)}\|_{p_F} < C_F$ and we define $q_F = 1 - 2/p_F$.

(b) Define $G_{i,j}^{(t,\tau)} := g_{i,j}^{(t,\tau)} - g_i \otimes g_j$. Then for some $p_G > 2$ there exists a constant C_G such that $\sup_{t,\tau,i,j} \|G_{i,j}^{(t,\tau)}\|_{p_G} < C_G$ and we define $q_G = 1 - 2/p_G$.

[1] We use the notation $f \otimes g(x, y) = f(x)g(y)$ and $\|f\|_p = (\int |f(x)|^p dx)^{1/p}$.

(iii) *The multiplicative kernel K has a finite κ-moment with $\kappa \geq \max(p, 1-1/p_F, 1-1/p_G)$, i.e. $C_K := \|K\|_\kappa < \infty$.*

We note that due to stationarity $F_{i,j}^{(t,\tau)} = F_{i,j}^{(0,t-\tau)}$ and $G_{i,j}^{(t,\tau)} = F_{i,j}^{(0,t-\tau)}$.

The definition below provides the suitable regularity space in order to prove the results in this paper.

DEFINITION 3.2. *For $s, \triangle > 0$, the space $\mathfrak{G}_{s,\triangle}^{\eta}$ is the class of functions $g : \mathbb{R}^\eta \to \mathbb{R}$ satisfying: g is everywhere $(m-1)$-times partially differentiable for $m - 1 < s \leqslant m$; where for some $\rho > 0$ and for all x, the inequality*

$$\sup_{y:|y-x|<\rho} \frac{|g(y) - g(x) - Q(y-x)|}{|y-x|^s} \leqslant \kappa(x),$$

holds true with $Q = 0$ when $m = 1$ and for $m > 1$, Q is an $(m-1)$th-degree homogeneous polynomial in $y - x$, whose coefficients are the partial derivatives of g of orders 1 to $m-1$ evaluated at x; κ is uniformly bounded by $\triangle$.

We now obtain the rate of convergence of $\hat{\varphi}(z)$, as $T \to \infty$.

THEOREM 3.1. *Suppose Assumptions 3.1 and 3.2 are satisfied, where $\mathfrak{r}$ and $\mathfrak{u}$ are the mixing coefficients associated with the vector time series $\{(X_{t,i}, Z_{t,i}, X_{t,j}, Z_{t,j})\}_t$ given in Assumption 3.1, and $q_G, q_F, q \in (0,1)$ and $\delta_i > 0$, $i = 1, \ldots, N$, are defined in Assumption 3.2. Let $\varpi = \min(q_F, q_G)$ and suppose $\mathfrak{u} > 1/\varpi + 1/q$.*

Let $\hat{\varphi}(z)$ be defined as in (3.3), where K is a multivariate kernel of order $r > 0$ (see Definition 3.1). In addition for each $i = 1, \ldots, N$ assume that $\varphi \cdot f_i$, $f_i \in \mathfrak{G}_{s_i,\triangle_i}^{\eta}$ for some $\triangle_i, s_i > 0$ (see Definition 3.2), that f_i is bounded away from zero and let $\rho_i = \min(r, s_i)$. We have for all $z \in \mathbb{R}^\eta$

(i) *if $\mathfrak{r} > 1/\varpi + 1/q$ and $b_i = O(T^{\frac{-1}{2\rho_i+\eta}})$, $i = 1, \ldots, N$, then*

$$|\hat{\varphi}(z) - \varphi(z)| = O_p\Big(\frac{1}{N}\sum_{i=1}^{N} (\delta_i^2)^{\frac{2\rho_i}{2\rho_i+\eta}} \cdot (\triangle_i^2)^{\frac{\eta}{2\rho_i+\eta}} \cdot T^{\frac{-2\rho_i}{2\rho_i+\eta}}\Big); \tag{3.5}$$

(ii) *if $1/q < \mathfrak{r} \leq 1/\varpi + 1/q$ and $b_i = O((N \cdot T)^{\frac{-1}{2\rho_i+\kappa\eta}})$, $i = 1, \ldots, N$, with $\kappa := 1 + \varpi + q - \varpi q \mathfrak{r}$ then*

$$|\hat{\varphi}(z) - \varphi(z)| = O_p\Big(\frac{1}{N}\sum_{i=1}^{N} (\delta_i^2)^{\frac{2\rho_i}{2\rho_i+\kappa\eta}} \cdot (\triangle_i^2)^{\frac{\kappa\eta}{2\rho_i+\kappa\eta}} \cdot (N \cdot T)^{\frac{-2\rho_i}{2\rho_i+\kappa\eta}}\Big); \tag{3.6}$$

(iii) *if $\mathfrak{r} \leq 1/q$ and $b_i = O((N \cdot T^{q\mathfrak{r}})^{\frac{-1}{2\rho_i+(1+q)\eta}})$, $i = 1, \ldots, N$ then*

$$|\hat{\varphi}(z) - \varphi(z)| = O_p\Big(\frac{1}{N}\sum_{i=1}^{N} (\delta_i^2)^{\frac{2\rho_i}{2\rho_i+q\eta+\eta}} \cdot (\triangle_i^2)^{\frac{q\eta+\eta}{2\rho_i+q\eta+\eta}} \cdot (N \cdot T^{q\mathfrak{r}})^{\frac{-2\rho_i}{2\rho_i+q\eta+\eta}}\Big). \tag{3.7}$$

REMARK 3.1. (i) From the proposition above we see if the mixing rate of the observations, $\mathfrak{r}$ were sufficiently large then we obtain the usual rate of convergence often found in nonparametric estimation (we note that a necessary condition is that $\mathfrak{r} > 2$). In other words, if for all i the rates are the same, with $\rho_i = \rho$, and $\mathfrak{r}$ is sufficiently, then $|\hat{\varphi}(z) - \varphi(z)| = O_p(T^{\frac{-2\rho_i}{2\rho+\eta}})$.

(ii) There is a continuity between the rates under the three different conditions. In other words as $\mathfrak{r} \to \varpi^{-1} + q^{-1}$ from the left then $T^{-\frac{\rho}{2\rho_i+\kappa\eta}} \to T^{-\frac{\rho}{2\rho_i+\eta}}$ and $\mathfrak{r} \to q^{-1}$ from the left then $T^{-\frac{\rho}{2\rho_i+\eta+q\eta}\cdot\mathfrak{r}q} \to T^{-\frac{\rho}{2\rho_i+\eta\kappa}}$. Roughly speaking this means the conclusions of $(ii) \to (i)$ and $(iii) \to (ii)$ as $\mathfrak{r} \to \varpi^{-1} + q^{-1}$ and $\mathfrak{r} \to q^{-1}$ respectively.

(iii) In order to reduce the number of different cases we have imposed the restriction that $\mathfrak{u} > 1/\varpi + 1/q$ (which basically means asymptotic independence of $(X_{t,i}, Z_{t,i})$ and $(X_{t,j}, Z_{t,j})$). If we were to relax this assumption and allow $\mathfrak{u} \leq 1/\varpi + 1/q$, this would give rise to 6 more cases. The most notable is when $\mathfrak{u}$ is also small and the conditions of Theorem 3.1(iii) hold. This case there is so much dependence within the time series $\{(X_{t,i}, Z_{t,i})\}_t$ and between the different time series $\{(X_{t,i}, Z_{t,i})\}_t$ and $\{(X_{t,j}, Z_{t,j})\}_t$ that $\hat{\varphi}(z)$ converges extremely slowly to the true parameter (even if $N \to \infty$, which is the case we consider below).

□

We now show that it is possible for the estimator $\widehat{\varphi}(\cdot)$ to obtain the rate given in Theorem 3.1(i) even in the case that the observations are only slowly 2-mixing. This is achieved by allowing the number of time series $N \to \infty$.

COROLLARY 3.2. *Suppose the assumptions of Theorem 3.1 are satisfied, and the bandwidth parameters are such that $b_i = O(T^{-1/(2\rho_i+\eta)})$, $i = 1, \ldots, N$,. Let $\rho = \min\{\rho_i; i = 1, \ldots, N\}$). Furthermore, in the case Theorem 3.1(ii), where $q^{-1} < \mathfrak{r} \leq \varpi^{-1} + q^{-1}$, let $N = O(T^{\frac{\eta(\kappa-1)}{2\rho+\eta}})$ while, in the case Theorem 3.1 (iii) where $\mathfrak{r} < q^{-1}$, let $N = O(T^{\frac{q\eta}{2\rho+\eta}+1-\mathfrak{r}q})$.*
Then we have $|\hat{\varphi}(z) - \varphi(z)| = O_p\Big(\frac{1}{N}\sum_{i=1}^{N}(\delta_i^2)^{\frac{2\rho_i}{2\rho_i+\eta}} \cdot (\triangle_i^2)^{\frac{\eta}{2\rho_i+\eta}} \cdot T^{\frac{-2\rho_i}{2\rho_i+\eta}}\Big)$, for all $z \in \mathbb{R}^\eta$.

Roughly speaking, if all the rates are the same with $\rho = \rho_i$, then from the above corollary we have $|\hat{\varphi}(z) - \varphi(z)| = O_p(T^{-2\rho/(2\rho+\eta)}))$, if N grows at a sufficient rate. Hence if we allow the number of time series N to grow also, then we achieve the usual nonparametric rate discussed in Remark 3.1(i).

REMARK 3.2 (Mean square error). In Theorem 3.1 and Corollary 3.2 we obtain the probabilitistic rate of convergence of the estimator $\widehat{\varphi} = \hat{g}/\hat{f}$ defined in (3.3). In order to obtain a similar rate in terms of mean square error (MSE) we need stronger assumptions, see for example Bosq (1998), Theorem 3.1. However we now show that by introducing a regularisation term in the estimator $\widehat{\varphi}$ under the same conditions given in Theorem 3.1 we can derive a MSE which is uniform in z. In the appendix we derive the MSE for the numerator $\hat{g}(\cdot)$ and the denominator $\hat{f}(\cdot)$ (see Lemma A.2). The difficulty in the estimation of the MSE of

$\widehat{\varphi}$ comes from the denominator $\hat{f}$, i.e., the expectation of $\hat{f}^{-1}$ does not, in general, exist. In order to cirumvent this difficulty we can introduce an additional regularization parameter $h > 0$ such that the denominator is bounded away from zero. For example, consider

$$\widehat{\varphi}^{(h)}(z) = \{h + \hat{f}(z)\}^{-1}\hat{g}(z) \tag{3.8}$$

$$\widehat{\varphi}^{(h)}(z) = \hat{f}(z)^{-1}\hat{g}(z)\mathbb{I}\{\hat{f}(z) > h\}. \tag{3.9}$$

We mention that regularizers have been used in several problems. For example, in the context of partially linear models an adaptation of (3.8) is used in Florens, Johannes, and Van Bellegem (2005) whereas Robinson (1988) considered a version of (3.9). □

In the following section we apply these methods to prediction for spatio-temporal processes.

4 Linear prediction

4.1 Covariance estimation

In this section we consider the linear prediction of $\Phi_t(u_0)$ given $Y_{t-s}(\underline{u}) := \{Y_{t-s}(u_i)\}_{i=1}^{m-1}$ with $\underline{u} = (u_1, \ldots, u_{m-1}) \in \Omega^{m-1}$ satisfying model (1.3), which of course is optimal if $(Y_t(u_0), Y_{t-s}(\underline{u}))$ were distributed according to a multivariate Gaussian.

We define for each $s \in \mathbb{Z}$ the covariance function $c_s : \mathbb{R}^{2d} \to \mathbb{R}$ with

$$c_s(u, v) := \mathrm{cov}(Y_t(u), Y_{t-s}(v)).$$

The best linear predictor of $\Phi_t(u_0)$ given $Y_{t-s}(\underline{u}) = \underline{y}$ with $\underline{y} \in \mathbb{R}^{m-1}$ is

$$\psi_L(\underline{y}, \underline{u}, u_0) = r(\underline{u}, u_0)' R(\underline{u})^{-1}\underline{y} \tag{4.1}$$

where $r(\underline{u}, u_0)' = (c_s(u_0, u_1), \ldots, c_s(u_0, u_{m-1}))$ and

$$R(\underline{u}) = \begin{pmatrix} c_0(u_1, u_1) & c_0(u_1, u_2) & \ldots & c_0(u_1, u_{m-1}) \\ c_0(u_1, u_2) & c_0(u_2, u_2) & \ldots & c_0(u_2, u_{m-1}) \\ \vdots & \vdots & \ddots & \ldots \\ c_0(u_{m-1}, u_1) & c_0(u_{m-1}, u_2) & \ldots & c_0(u_{m-1}, u_{m-1}) \end{pmatrix}.$$

Since the parameter $r(\cdot)$ and the matrix $R(\cdot)$ are functions of the covariance, the object of this section is to develop methods for estimating $c_s(\cdot)$, which in turn can be used to estimate $r(\cdot)$ and $R(\cdot)$. The predictor $\psi_L(\cdot)$ can include all observations at a given time, that is $m = N$ (though it is natural to use only those which are near to the unobserved point). For brevity we shall assume that the spatial mean is zero, it is straightforward to extend the results to spatial processes with non-zero mean.

As will become clear below, the method we use to estimate $c_s(\cdot)$ should depend on its covariance structure. Since we are using nonparametric methods to estimate the covariance,

the rate of convergence of the estimator will be affected by the dimension d of the covariates. But this can be remedied if there is a known function $H : \mathbb{R}^{2d} \to \mathbb{R}^{\nu}$ such that there exists a function $c_{H,s} : \mathbb{R}^{\nu} \to \mathbb{R}$ (in general unknown) which satisfies

$$c_{H,s}(H(u,v)) = \mathrm{cov}(\Phi_t(u), \Phi_{t-s}(v)), \quad \forall u, v \in \Omega, t, s \in \mathbb{Z}. \tag{4.2}$$

By using this information we can reduce the dimension from $2d$ to ν, thereby avoiding the curse of dimensionality. Let σ^2 denote the variance of the observation error and define $v_H(\cdot) := c_{H,0}(\cdot) + \sigma^2$. Since the observation errors are independent of the process $\{\Phi_t(\cdot)\}_t$ we have the following characterisation of the covariance function

$$c_s(u,v) = \begin{cases} v_H(H(u,u)) & , \text{ when } s = 0 \text{ and } u = v; \\ c_{H,s}(H(u,v)) & , \text{ otherwise.} \end{cases} \tag{4.3}$$

Example 4.1 (Dimension reduction through a suitable $H(\cdot)$).

(i) In the case that $\{\Phi_t(\cdot)\}$ is spatially nonstationary with no additional assumptions then $H : \mathbb{R}^{2d} \to \mathbb{R}^{2d}$ with $H(u,v) := (u,v)$.

(ii) Often it is reasonable to suppose that the process $\{\Phi_t(\cdot)\}$ is both temporally and spatially stationary. In which case $H : \mathbb{R}^{2d} \to \mathbb{R}^{d}$ with $H(u,v) := (u - v)$.

(iii) In spatial statistics it is common to assume isotropy of the covariance function. In this case $H : \mathbb{R}^{2d} \to \mathbb{R}$ with $H(u,v) := \|u - v\|$, where $\|\cdot\|$ is the Euclidean norm.

(iv) Consider the model in (1.2), where $\{\xi_t^{(i)}(\cdot)\}$ are iid random functions with an isotropic covariance. It is straightforward to show that

$$\mathrm{cov}(\Phi_t(w,x), \Phi_{t-s}(\tilde{w},\tilde{x})) = (1 + x'\tilde{x}) \cdot \gamma_s(\|w - \tilde{w}\|),$$

where $\gamma_s(\|w - \tilde{w}\|) = \mathrm{cov}(\xi_0(w), \xi_s(\tilde{w}))$. Let $u = (w,x)$, $v = (\tilde{w},\tilde{x})$, then in this case $H : \mathbb{R}^{2(p+1)} \to \mathbb{R}^2$ with

$$H(u,v) := ((1 + x'\tilde{x}), \|w - \tilde{w}\|).$$

□

Based on the characterisation (4.3), we use the observations$\{(Y_{t,i}, U_{t,i}), i = 1, \ldots, N; t = 1, \ldots, T\}$ to construct an estimator of the function $c_s(u,v)$ for $s \in \mathbb{Z}$ and $u, v \in \Omega$. Therefore define for $t \in \mathbb{Z}$ and $i, j \in \mathbb{N}$ the random variables $X_{t,i,j}^{(s)} = Y_{t,i} \cdot Y_{t-s,j}$ and $Z_{t,i,j}^{(s)} = H(U_{t,i}, U_{t-s,j})$. Under Assumption 2.2, it is easily verified that for all $z \in \mathbb{R}^{\nu}, t, s \in \mathbb{Z}, i, j \in \mathbb{N}$ with $i \neq j$

$$c_{H,s}(z) = \mathbb{E}(Y_{t,i} \cdot Y_{t-s,j} | H(U_{t,i}, U_{t-s,j}) = z) = \mathbb{E}(X_{t,i,j}^{(s)} | Z_{t,i,j}^{(s)} = z),$$

and

$$v_H(z) = \mathbb{E}(Y_{t,i}^2 | H(U_{t,i}, U_{t,i}) = z) = \mathbb{E}(X_{t,i,i}^{(0)} | Z_{t,i,i}^{(0)} = z).$$

By using, for fixed i, j $\{(U_{t,i}, U_{t-s,j})\}_t$ is a stationary vector process, it follows that $\{(X_{t,i,j}^{(s)}, Z_{t,i,j}^{(s)})\}_t$ are identically distributed random vectors. Comparing this with the general multivariate case considered in Section 3, we see that $\{(X_{t,i,j}^{(s)}, Z_{t,i,j}^{(s)})\}_t$ can be treated as a particular example. This motivates us to let $f_{i,j}^{(s)}(z)$ denote the marginal density of $Z_{t,i,j}^{(s)}$ and define $g_{i,j}^{(s)}(z) := \mathbb{E}(X_{t,i,j}^{(s)} | Z_{t,i,j}^{(s)} = z) \cdot f_{i,j}^{(s)}(z)$. Then, for all $z \in \mathbb{R}^\nu$, we have

$$c_{H,s}(z) = \frac{\frac{1}{N/2}\sum_{i=1}^{N/2} g_{2i-1,2i}^{(s)}(z)}{\frac{1}{N/2}\sum_{i=1}^{N/2} f_{2i-1,2i}^{(s)}(z)} \quad \text{and} \quad v_H(z) = \frac{\frac{1}{N}\sum_{i=1}^{N} g_{i,i}^{(0)}(z)}{\frac{1}{N}\sum_{i=1}^{N} f_{i,i}^{(0)}(z)}. \tag{4.4}$$

For all $i, j \in \mathbb{N}$ and $s \in \mathbb{Z}$ we estimate the functions $g_{i,j}^{(s)}(\cdot)$ and $f_{i,j}^{(s)}(\cdot)$ using the kernel estimators

$$\widehat{g_{i,j}^{(s)}}(\cdot) := \frac{1}{T-s}\sum_{t=1+s}^{T} X_{t,i,j}^{(s)} K_{b_{i,j}^{(s)}}(Z_{t,i,j}^{(s)} - \cdot) \quad \text{and} \quad \widehat{f_{i,j}^{(s)}}(\cdot) := \frac{1}{T-s}\sum_{t=1+s}^{T} K_{b_{i,j}^{(s)}}(Z_{t,i,j}^{(s)} - \cdot),$$

where K denotes a multiplicative kernel (see Definition 3.1) and $b_{i,j}^{(s)} > 0$ a given bandwidth. Replacing in (4.4) the functions $g_{i,j}^{(s)}(\cdot)$ and $f_{i,j}^{(s)}(\cdot)$ by their estimators we obtain

$$\begin{aligned} \widehat{c_{H,s}}(\cdot) &:= \frac{\sum_{i=1}^{N/2}\sum_{t=1+s}^{T} X_{t,2i-1,2i}^{(s)} K_{b_{2i-1,2i}^{(s)}}(Z_{t,2i-1,2i}^{(s)} - \cdot)}{\sum_{i=1}^{N/2}\sum_{t=1+s}^{T} K_{b_{2i-1,2i}^{(s)}}(Z_{t,2i-1,2i}^{(s)} - \cdot)} \\ \text{and} \quad \widehat{v_H}(\cdot) &:= \frac{\sum_{i=1}^{N}\sum_{t=1+s}^{T} X_{t,i,j}^{(s)} K_{b_{i,i}^{(s)}}(Z_{t,i,i}^{(s)} - \cdot)}{\sum_{i=1}^{N}\sum_{t=1+s}^{T} K_{b_{i,i}^{(s)}}(Z_{t,i,i}^{(s)} - \cdot)}, \end{aligned} \tag{4.5}$$

as estimators of $c_{H,s}(\cdot)$ and $v_H(\cdot)$, respectively.

It is worth mentioning that in practice the densities $\{f_{i,j}(\cdot)\}_{i,j}$ maybe the same, hence we would use a universal bandwidth b. Moreover the bandwidth can be selected using cross-validation methods (cf. Hart (1994)).

We are now equiped to define an estimator of the matrix $R(\underline{u})$, that is

$$\widehat{R}(\underline{u}) := \begin{pmatrix} \widehat{v_H}(H(u_1, u_1)) & \dots & \widehat{c_{H,0}}(H(u_1, u_{m-1})) \\ \vdots & \ddots & \vdots \\ \widehat{c_{H,0}}(H(u_{m-1}, u_1)) & \dots & \widehat{v_{H,0}}(H(u_{m-1}, u_{m-1})) \end{pmatrix},$$

while we use $\widehat{r}(\underline{u}, u_0)' = (\widehat{c_{H,s}}(H(u_0, u_1)), \dots, \widehat{c_{H,s}}(H(u_0, u_{m-1})))$ as an estimator of $r(\underline{u}, u_0)$. Altogether we obtain the estimator $\widehat{\psi_L}$ of the linear predictor ψ_L, where

$$\widehat{\psi_L}(\underline{y}, \underline{u}, u_0) := \widehat{r}(\underline{u}, u_0)' \widehat{R}(\underline{u})^{-1} \underline{y}. \tag{4.6}$$

4.2 Sampling properties

In this section we will study the sampling properties of the estimators $\widehat{c_{H,s}}(\cdot)$ and $\widehat{v_H}(\cdot)$, which are derived using the results given in Section 3.

Under Assumptions 2.1, 2.2 and 2.3 (with $m = 2$), we see that Assumption 3.1 holds with $X_{t,i} := X^{(s)}_{t,2i-1,2i}$ and $Z_{t,i} := X^{(s)}_{t,2i-1,2i}$ or $X_{t,i} := X^{(0)}_{t,i,i}$ and $Z_{t,i} := X^{(0)}_{t,i,i}$, and by appealing to Proposition 2.1 we have

$$\sup_{\substack{A\in\sigma(X_{t,i},Z_{t,i})\\ B\in\sigma(X_{\tau,j},Z_{\tau,j})}} |P(A\cap B)-P(A)P(B)| \le C\begin{cases} |t-\tau|^{-\min(\alpha,\beta)}; & \text{if } i=j,\\ |t-\tau|^{-\min(\gamma,\beta)}; & \text{otherwise.}\end{cases}$$

Therefore, if also Assumptions 3.2 is satisfied, the following theorem on the rate of convergence of the covariance and variance estimators is a direct consequence of Theorem 3.1.

THEOREM 4.1. *Suppose Assumptions 2.1, 2.2 and 2.3 (with $m = 2$) holds, where α and γ are the mixing coefficients of the covariates $\{U_{t,i}\}_t$ over time and space, respectively, β is the mixing coefficient of the process $\{\Phi_t(u)\}_t$ (u is arbitrary but fixed). Let the multivariate vector time series $\{X^{(s)}_{t,2i-1,2i}, Z^{(s)}_{t,2i-1,2i}\}_{t,i}$ and $\{X^{(0)}_{t,i,i}, Z^{(0)}_{t,i,i}\}_{t,i}$ satisfy the Assumption 3.2 with common constants $q_G, q_F, q \in (0,1)$ and individual constants $\delta_{i,s}$ and $\delta_i > 0$, respectively. Let $\varpi = \min(q_F, q_G)$ and suppose $\min(\gamma,\beta) > 1/\varpi + 1/q$.*

Let $\widehat{c_{H,s}}(z)$ and $\widehat{v_H}(z)$ be defined as in (4.5), where K is a multivariate kernel of order $r > 0$ (Definition 3.1). In addition for each $i, j \in \mathbb{N}$, $i \neq j$ assume $c_{H,s}(z)\cdot f^{(s)}_{i,j}$, $f^{(s)}_{i,j} \in \mathfrak{G}^{\eta}_{l_{i,s},\triangle_{i,s}}$ with $l_{i,s}, \triangle_{i,s} > 0$ (see Definition 3.2), while $v_H(z)\cdot f^{(0)}_{i,i}$, $f^{(0)}_{i,i} \in \mathfrak{G}^{\eta}_{l_i,\triangle_i}$ for some $l_i, \triangle_i > 0$, and suppose that the functions $f^{(s)}_{i,j}$ and $f^{(0)}_{i,i}$ are bounded away from zero. Let $\rho_{i,s} = \min(r, l_{i,s})$ and $\rho_i = \min(r, l_i)$ for $i \in \mathbb{N}$. Then for all $z \in \mathbb{R}^\nu$

(i) if $\alpha > 1/\varpi + 1/q$, $b^{(s)}_{2i-1,2i} = O(T^{-1/(2\rho_{i,s}+\nu)})$ and $b^{(0)}_{i,i} = O(T^{-1/(2\rho_i+\nu)})$, we have

$$|\widehat{c_{H,s}}(z) - c_{H,s}(z)| = O_p\Big(\frac{1}{N/2}\sum_{i=1}^{N/2}(\delta^2_{i,s})^{\frac{2\rho_{i,s}}{2\rho_{i,s}+\nu}}\cdot(\triangle^2_{i,s})^{\frac{\nu}{2\rho_{i,s}+\nu}}\cdot T^{\frac{-2\rho_{i,s}}{2\rho_{i,s}+\nu}}\Big); \tag{4.7}$$

$$|\widehat{v_H}(z) - v_H(z)| = O_p\Big(\frac{1}{N}\sum_{i=1}^{N}(\delta^2_{i})^{\frac{2\rho_{i}}{2\rho_{i}+\nu}}\cdot(\triangle^2_{i})^{\frac{\nu}{2\rho_{i}+\nu}}\cdot T^{\frac{-2\rho_{i}}{2\rho_{i}+\nu}}\Big) \tag{4.8}$$

(ii) if $1/q < \alpha \le 1/\delta + 1/q$, $b^{(s)}_{2i-1,2i} = O((N\cdot T)^{-1/(2\rho_{i,s}+\kappa\nu)})$ and $b^{(0)}_{i,i} = O((N\cdot T)^{-1/(2\rho_i+\kappa\nu)})$ with $\kappa := 1 + \varpi + q - \varpi\cdot q\cdot\alpha$ we have

$$|\widehat{c_{H,s}}(z) - c_{H,s}(z)| = O_p\Big(\frac{1}{N/2}\sum_{i=1}^{N/2}(\delta^2_{i,s})^{\frac{2\rho_{i,s}}{2\rho_{i,s}+\kappa\nu}}\cdot(\triangle^2_{i,s})^{\frac{\kappa\nu}{2\rho_{i,s}+\kappa\nu}}\cdot (N\cdot T)^{\frac{-2\rho_{i,s}}{2\rho_{i,s}+\kappa\nu}}\Big); \tag{4.9}$$

$$|\widehat{v_H}(z) - v_H(z)| = O_p\Big(\frac{1}{N}\sum_{i=1}^{N}(\delta^2_{i})^{\frac{2\rho_{i}}{2\rho_{i}+\kappa\nu}}\cdot(\triangle^2_{i})^{\frac{\kappa\nu}{2\rho_{i}+\kappa\nu}}\cdot (N\cdot T)^{\frac{-2\rho_{i}}{2\rho_{i}+\kappa\nu}}\Big) \tag{4.10}$$

(iii) if $\alpha \leq 1/q$, $b^{(s)}_{2i-1,2i} = O((N \cdot T^{\alpha q})^{\frac{-1}{2\rho_{i,s}+q\nu+\nu}})$ *and* $b^{(0)}_{i,i} = O((N \cdot T^{\alpha q})^{\frac{-1}{2\rho_i+q\nu+\nu}})$ *we obtain*

$$|\widehat{c_{H,s}}(z) - c_{H,s}(z)| = O_p\Big(\frac{1}{N/2}\sum_{i=1}^{N/2}(\delta_{i,s}^2)^{\frac{2\rho_{i,s}}{2\rho_{i,s}+q\nu+\nu}} \cdot (\triangle_{i,s}^2)^{\frac{\kappa\nu}{2\rho_{i,s}+q\nu+\nu}} \cdot (N \cdot T^{\alpha q})^{\frac{-2\rho_{i,s}}{2\rho_{i,s}+q\nu+\nu}}\Big); \tag{4.11}$$

$$|\widehat{v_H}(z) - v_H(z)| = O_p\Big(\frac{1}{N}\sum_{i=1}^{N}(\delta_i^2)^{\frac{2\rho_i}{2\rho_i+q\nu+\nu}} \cdot (\triangle_i^2)^{\frac{(1+q)\nu}{2\rho_i+q\nu+\nu}} \cdot (N \cdot T^{\alpha q})^{\frac{-2\rho_i}{2\rho_i+q\nu+\nu}}\Big). \tag{4.12}$$

In the theorem above we see that purpose of the summands is to accomodate the different smoothness classes of $f^{(s)}_{i,j}$. If the densities were to belong to the same class, then the summands are avoided. Moreover, since $f^{(s)}_{i,j}$ is the density of $H(U_{t,i}, U_{t-s,j})$, the rate of convergence depends both on the smoothness of the covariance function $c_{s,H}(\cdot)$, as one would expect, as well as the smoothness of the covariate densites.

Remark 4.1 (Local stationarity). We note that, since we are using 'local' smoothing to estimate the covariance function, the assumption $c_{H,s}(z), v_H(z) \in \mathfrak{G}^{\nu}_{l,\triangle}$ (in Theorem 4.1) imposes a 'local stationarity' condition on the spatially nonstationary spatio-temporal process $\Phi_t(\cdot)$. For instance, consider two observations from the spatio-temporal process; $\Phi_t(u_1)$ and $\Phi_t(u_2)$ whose covariates u_1 and u_2 are 'close'. Since $\mathfrak{G}^{\nu}_{l,\triangle}$ characterises a class of 'smooth' functions, we have for all $u \in \Omega$, that $c_s(u_1, u) \approx c_s(u_2, u)$, this implies that they have a similar covariance structure, leading to a process which at least in the second order sense can be described as 'locally' stationary. □

Remark 4.2 (MSE, uniform convergence, asymptotic normality). It is worth mentioning that a small adaption of the estimator $\hat{c}_{H,s}(\cdot)$ or $\hat{v}_H(\cdot)$ yields an estimator whose mean squared error can easily be evaluated. We refer to Remark 3.2 in Section 3 for the details.

It is possible to show that the estimators are uniformly convergent almost surely, over an increasing sequence of compact sets, under much stronger conditions on the processes $\{\Phi_t(\cdot)\}_t$ and $\{U_{t,i}\}_t$. To summarise, if the 2-mixing rates in Assumption 2.3 were replaced by strong mixing rates and the rate of convergence is strengthed to a geometric mixing rate, f, $c_{H,s} \in \mathfrak{G}^{\eta}_{2,\triangle}$ and there exists an a such that $\mathbb{E}(\exp(aY_{t,j})) < \infty$, then $\widehat{c_{H,s}}(z)$ and $\widehat{v_H}(z)$ converge uniformly in a compact set to $c_{H,s}(z)$ and $v_H(z)$, respectively. Under a similar set of assumptions asymptotic normality of $\widehat{c_{H,s}}(z)$ and $\widehat{v_H}(z)$ can also be shown. For details we refer to Bosq (1998), Theorems 3.3 and 3.4. □

In the previous theorem we have obtained a rate of convergence for the covariance estimators $\widehat{c_{H,s}}$ and $\widehat{v_H}$. An interesting aspect in the proof of this result, is that the rates are independent of of z. This observation immediately yields the following result on the rate of convergence of $\widehat{\psi_L}$.

Corollary 4.2. *Suppose the assumptions of Theorem 4.1 are satisfied. Let* $\widehat{\psi_L}$ *be defined as in (4.6), with* $m \leq N$. *Assume, that* $\sup_i((\delta_{i,0}, \delta_{i,s}, \delta_i) < \infty$, $\sup_i(\triangle_{i,0}, \triangle_{i,s}, \triangle_i) < \infty$

and $\rho := \inf_i\{\rho_{i,0}, \rho_{i,s}, \rho_i\} > 0$. *Then for all* $\underline{y} \in \mathbb{R}^{m-1}$, $(\underline{u}, u_0) \in \Omega^m$ *we have*

(i) If $\alpha > 1/\varpi + 1/q$ *and* $b_{i,j}^{(s)} = O(T^{-1/(2\rho+\nu)})$, *then*

$$|\widehat{\psi_L}(\underline{y}, \underline{u}, u_0) - \psi_L(\underline{y}, \underline{u}, u_0)| = O_p\Big(T^{-2\rho/(2\rho+\nu)}\Big).$$

(ii) If $1/q < \alpha \leq 1/\varpi + 1/q$ *and* $b_{i,j}^{(s)} = O((N \cdot T)^{-1/(2\rho+\kappa\nu)})$, *with* $\kappa := 1 + \varpi + q - \varpi \cdot q \cdot \alpha$ *then*

$$|\widehat{\psi_L}(\underline{y}, \underline{u}, u_0) - \psi_L(\underline{y}, \underline{u}, u_0)| = O_p\Big((N \cdot T)^{-2\rho/(2\rho+\kappa\nu)}\Big).$$

(iii) If $\alpha \leq 1/q$ *and* $b_{i,j}^{(s)} = O((N \cdot T^{\alpha q})^{-1/(2\rho+(1+q)\nu)})$, *then*

$$|\widehat{\psi_L}(\underline{y}, \underline{u}, u_0) - \psi_L(\underline{y}, \underline{u}, u_0)| = O_p\Big((N \cdot T^{\alpha q})^{-2\rho/(2\rho+(1+q)\nu)}\Big).$$

It is worth drawing to attention that, the rate in (4.13) does not depend on m (since ν is independent of m). Therefore it includes the case where $\psi_L(\cdot)$ is defined with $m = N$ predictors.

In spatial statistics the asymptotic results often use the assumptions that the number of 'locations' $N \to \infty$ (see, for example, Guan, Sherman, and Calvin (2004) and Mukherjee and Lahiri (2004)). Similarly, we now consider the case that the number of covariates at each time point $N \to \infty$ as well as $T \to \infty$. We shall show that in this case it is possible to obtain the rate $T^{-\frac{\rho}{2\rho+\eta}}$ for the estimator $\widehat{\psi}(\cdot)$ even in the case that the covariates are not mixing much.

COROLLARY 4.3. *Suppose the assumptions of Theorem 4.1 and Corollary 4.2 are satisfied. In addition, if the assumptions Corollary 4.2(ii) are satisfied, with* $q^{-1} < \alpha \leq \varpi^{-1} + q^{-1}$, *let* $N = O(T^{\frac{\nu(\kappa-1)}{2\rho+\nu}})$ *while, if Corollary 4.2(iii) holds with* $\alpha < q^{-1}$, *let* $N = O(T^{\frac{q\nu}{2\rho+\nu}+1-\alpha q})$. *Then*

$$\Big|\widehat{\psi}_L(\underline{y}, \underline{u}, u_0) - \psi_L(\underline{y}, \underline{u}, u_0)\Big| = O_p(T^{-\frac{\rho}{2\rho+\nu}}), \quad \text{for all } \underline{y} \in \mathbb{R}^{m-1}, (\underline{u}, u_0) \in \mathbb{R}^{md}. \quad (4.13)$$

REMARK 4.3 (Different rates). From the theorem and corollary above we notice that the rate of convergence depends on two factors:

(i) The dimension of the image of the function $H : \mathbb{R}^{2d} \to \mathbb{R}^{\nu}$, which is based on the dependence structure of the spatial temporal process. For example, in the case of nonstationary, with no additional assumptions, the rate is slowest, but if the process were isotropic the estimator can be modified to yield an estimator which is independent of dimension. What is more, even for nonstationary processes with some structure (see Example 4.1(iv)) we can still obtain an estimator which is independent of dimension.

(ii) The dependence structure of the covariates. We see that for N kept fixed, as the dependence of the covariates grow the rate of convergence of $\widehat{\psi}_L$ becomes slower. As an illustration we consider the examples in Remark 2.2:

(a) The covariates satisfy a vector AR process, hence the temporal dependence of the covariates is weak and α is large. This means for N kept fixed (we can even have $N = 2$), the usual rate of convergence in nonparametric estimation is incurred.

(b) Here there is little or no temporal mixing of the covariates. Thus for N kept fixed the estimator $\widehat{\psi}_L$ is inconsistent. However because there is independence between the covariates if $N \to \infty$ and $T \to \infty$ at a sufficient rate the usual nonparametric rate can be obtained.

□

REMARK 4.4 (Dimension reduction for unknown $H(\cdot)$). There arises applications where we know that there exists a function $H : \mathbb{R}^{2d} \to \mathbb{R}^{\nu}$, with $\nu < 2d$ such that $c_{H,s}(H(u,v)) = \mathrm{cov}(\Phi_t(u)\Phi_{t-s}(v))$ but the actual function $H(\cdot)$ may be unknown. Examples include

(i) A generalisation of (1.2) where $H : \mathbb{R}^{2d} \to \mathbb{R}$, defined by

$$H(w,x,\tilde{w},\tilde{x}) = \{\sum_{k=1}^{q} f_k(x^{(k)}\tilde{x}^{(k)})\}\|w-\tilde{w}\|, \text{ with } x = (x^{(1)},\ldots,x^{(d)}), \tilde{x} = (\tilde{x}^{(1)},\ldots,\tilde{x}^{(d)}),$$

and the functions $f_k : \mathbb{R} \to \mathbb{R}$ are unknown.

(ii) Anistropic processes are a generalisation of an isotropic process (c.f. Cressie and Huang (1999)). Here $H : \mathbb{R}^{2d} \to \mathbb{R}$, is defined by

$$H(u,v) = \sqrt{v'Au} = \{\sum_{i=1}^{d}\sum_{j=1}^{d} A_{ij}v^{(i)}u^{(j)}\}^{1/2}, \text{ with } u = (u^{(1)},\ldots,u^{(d)}), v = (v^{(1)},\ldots,v^{(d)}),$$

and A is an unknown positive definite matrix.

In order to reduce the affect of dimension in the estimation of $c_s(\cdot)$ it is of interest to obtain estimators of the function $H(\cdot)$. There are several ways to approach this problem, however the obvious is to use the representation

$$Y_{t,i}Y_{t-s,j} = c_{H,s}(H(U_{t,i},U_{t-s,j})) + \varepsilon_t^{(i,j)}, \tag{4.14}$$

where $\varepsilon_t^{(i,j)} = Y_{t,i}Y_{t-s,j} - c_{H,s}(H(U_{t,i},U_{t-s,j}))$ and $\mathbb{E}(\varepsilon_t^{(i,j)}|H(U_{t,i},U_{t-s,j}))) = 0$. We observe that (4.14) resembles the classical setup where we observe a nonparametric function which has been corrupted by additive noise. For this reason it is possible to draw on dimension reduction methods, developed in nonparametric statistics, to estimate $c_{H,s}(\cdot)$ (though the noise stucture $\varepsilon_t^{(i,j)}$ is more complicated than usual). For example, suppose, $c_{H,s}(H(u,v)) = \mathrm{cov}(\Phi_t(u),\Phi_{t-s}(v))$ where $H : \mathbb{R}^{2d} \to \mathbb{R}$ with

$$H(u,v) = \sum_{k=1}^{d} f_k(u^{(k)} - v^{(k)}), \quad \text{with } u = (u^{(1)},\ldots,u^{(d)}), v = (v^{(1)},\ldots,v^{(d)}),$$

but $f_k(\cdot)$ is unknown. We now consider an iterative scheme to estimate $c_{H,s}(\cdot)$ and $H(\cdot)$, when $f_k(\cdot)$ is a linear functional; $f_k(x) = \alpha_k \ell_k(x)$, where α_k is unknown, but the function $\ell_k(\cdot)$ is known. This can be done by using methods developed in Lu, Tjostheim, and Yao (2005) and Fan, Yao, and Cai (2003), who consider the model

$$Z_t = h_0(\mathbf{a}'\mathbf{X}_t) + \mathbf{X}_t' \mathbf{g}_0(\mathbf{a}'\mathbf{X}_t) + \varepsilon_t, \tag{4.15}$$

where $\{\mathbf{X}_t\}_t$ is a predictor vector, $\mathbb{E}(\varepsilon_t|\mathbf{X}_t) = 0$, α is a vector of unknown parameters and $h_0(\cdot)$ and $\mathbf{g}_0(\cdot)$ are unknown functions. Comparing (4.15) with

$$Y_{t,i}Y_{t-s,j} = c_{H,s}(\sum_{k=1}^{d} \alpha_k \cdot \ell_k(U_{t,i}^{(k)} - U_{t-s,j}^{(k)})) + \varepsilon_t^{(i,j)}, \tag{4.16}$$

we see that (4.16) can be written in terms of (4.15), with $Z_t = Y_{t,i}Y_{t-s,j}$, $\mathbf{X}_t = (\ell_1(U_{t,i}^{(1)} - U_{t-s,j}^{(1)}), \ldots, \ell_d(U_{t,i}^{(d)} - U_{t-s,j}^{(d)}))$, $h_0(\cdot) = c_{H,s}(\cdot)$, $\mathbf{g}_0(\cdot) \equiv 0$, and $\mathbf{a}' = (\alpha_1, \ldots, \alpha_d)$. Given this representation, we can estimate the parameter vector $\mathbf{a}$ and the function $c_{H,s}(\cdot)$, using the iterative local least squares scheme advocated in Fan, Yao, and Cai (2003) and Lu, Tjostheim, and Yao (2005). A small difference is that for each time t we have multiple observations $\{(Y_{t,i}Y_{t,j}, U_{t,i}, U_{t,j}); 1 \leq i \leq j \leq N\}$, however it is straightforward to modify the estimation scheme to the current problem. The main advantage of using this scheme is, if it can be shown that the conditions stated in Lu, Tjostheim, and Yao (2005) are satisfied then the curse of dimensionality is avoided. That is, the estimator of the parameter vector $\mathbf{a}$ is $\sqrt{T}$-convergent and the rate of convergence of the nonparametric estimator of $c_{H,s}(\cdot)$ does not depend on the dimension d.

A useful generalisation is to develop dimension reduction techniques where $f_k(\cdot)$ is an unknown non-linear functional (possibly by adapting the dimension reduction methods considered in Horowitz and Mammen (2004) and Fan, Härdle, and Mammen (1998)). This problem will be considered in the future. □

5 Estimation of the optimal predictior

5.1 The estimator

If the observations depart from Gaussianity the linear predictor can be far from the optimal predictor. In this section we propose a method to estimate directly the function $\psi(\underline{y}, \underline{u}, u_0) = \mathbb{E}(\Phi_t(u_0)|\{Y_{t-s}(u_i)\}_{i=1}^{m-1} = \underline{y})$ with $\underline{y} \in \mathbb{R}^{m-1}$ and $\underline{u} = (u_1, \ldots, u_{m-1}) \in \Omega^{m-1}$.

Recall that $\underline{Y}_{t-s}^{(0,\underline{i})} = \{Y_{t-s,i_j}\}_{j=1}^{m-1}$ and $\underline{U}_{t-s}^{(0,\underline{i})} = \{U_{t-s,i_j}\}_{j=1}^{m-1}$ for $\underline{i} = (i_1, \ldots, i_{m-1}) \in \mathbb{N}^{m-1}$. Under Assumption 2.2 it is straightforward to show that for all distinct indices $(i_0, \underline{i}) \in \mathbb{N}^m$

$$\psi(\underline{y}, \underline{u}, u_0) = \mathbb{E}(Y_{t,i_0}|\underline{Y}_{t-s}^{(0,\underline{i})} = \underline{y}, \underline{U}_{t-s}^{(0,\underline{i})} = \underline{u}, U_{t,i_0} = u_0), \quad \forall\, (u_0, \underline{u}) \in \Omega^m,\ \underline{y} \in \mathbb{R}^{m-1}.$$

Define for $i = 1, \ldots, N/m$ the η-dimensional random vector $Z_{t,i}^{(s)} := (\underline{Y}_{t-s}^{(0,\underline{i}_i)}, \underline{U}_{t-s}^{(0,\underline{i}_i)}, U_{t,im})$ with $\underline{i}_i = ((i-1)m + 1, \ldots, im - 1)$ and $\eta = (d+1)m - 1$. By using that for all $i \in$

$\mathbb{N}$, $\{(U_{t,im}, \underline{U}_{t-s}^{(0,i_i)})\}_t$ is a stationary vector time series, it follows that $\{Y_{t,im}, Z_{t,i}^{(s)}\}_t$ are identically distributed random variables. As in the previous section we see that $\{Y_{t,im}, Z_{t,i}^{(s)}\}_t$ is an example of the multivariate time series considered in Section 3. Therefore to estimate $\psi(\cdot)$, first let $f_i^{(s)}(\cdot)$ denote the marginal density of $Z_{t,i}^{(s)}$ and define $g_i^{(s)}(z) := \mathbb{E}[Y_{t,im}|Z_{t,i}^{(s)} = z] \cdot f_i^{(s)}(z)$, then we have for all $z \in \mathbb{R}^\eta$

$$\psi(z) = \frac{\frac{1}{N/m}\sum_{i=1}^{N/m} g_i^{(s)}(z)}{\frac{1}{N/m}\sum_{i=1}^{N/m} f_i^{(s)}(z)}. \tag{5.1}$$

For all $i \in \mathbb{N}$ and $s \in \mathbb{Z}$ we estimate the functions $g_i^{(s)}(\cdot)$ and $f_i^{(s)}(\cdot)$ with

$$\widehat{g_i^{(s)}}(\cdot) := \frac{1}{T-s}\sum_{t=1+s}^{T} Y_{t,im} K_{b_i^{(s)}}(Z_{t,i}^{(s)} - \cdot) \quad \text{and} \quad \widehat{f_i^{(s)}}(\cdot) := \frac{1}{T-s}\sum_{t=1+s}^{T} K_{b_i^{(s)}}(Z_{t,i}^{(s)} - \cdot),$$

where K denotes a multiplicative kernel (see Definition 3.1) and $b_i^{(s)} > 0$ a given bandwidth. Replacing in (5.1) the functions $g_i^{(s)}(\cdot)$ and $f_i^{(s)}(\cdot)$ by their estimators we obtain

$$\widehat{\psi}(\cdot) := \frac{\sum_{i=1}^{N/m}\sum_{t=1+s}^{T} Y_{t,im} K_{b_i^{(s)}}(Z_{t,i}^{(s)} - \cdot)}{\sum_{i=1}^{N/m}\sum_{t=1+s}^{T} K_{b_i^{(s)}}(Z_{t,i}^{(s)} - \cdot)}, \tag{5.2}$$

as estimator of the optimal predictor ψ.

REMARK 5.1. We observe that $\psi(\cdot)$ is a function on $(d+1)m-1$ variables, hence the quality of the estimator depends on the dimension. It is possible, as in Section 4, to incorporate isotropy into the model, however meaningful methods for reducing the dimension of ψ, unlike the covariance estimator, are not so obvious, which we now demonstrate.

Recall the linear predictor $\psi_L(\underline{y}, \underline{u}, u_0) = a'(\underline{u}, u_0)\underline{y}$ with $a(\underline{u}, u_0) := R(\underline{u}, u_0)^{-1} r(\underline{u}, u_0)$ defined in (4.1). We observe that without any additional assumptions on the structure of the process $\{\Phi_t(\cdot)\}_t$, $a(\cdot)$ is a function of $d \cdot m$ variables. Now suppose the covariances are isotropic, and let us define the function $H : \mathbb{R}^{dm} \supseteq \Omega^m \to \mathbb{R}^{m(m-1)/2}$ with $H(\underline{u}, u_0) := (\|u_0 - u_1\|, \ldots, \|u_{m-1} - u_{m-2}\|)$ for all $\underline{u} = (u_1, \ldots, u_{m-1}) \in \Omega^{m-1}$. Then there exists a function $a_H : \mathbb{R}^{m(m-1)/2} \to \mathbb{R}^{m-1}$ such that $a_H(H(\underline{u}, u_0)) = a(\underline{u}, u_0)$ for all $(\underline{u}, u_0) \in \Omega^m$. Therefore isotropy implies that there exists a function $\psi_{L,H} : \mathbb{R}^{(m+2)(m-1)/2} \to \mathbb{R}$ such that

$$\psi_{L,H}(\underline{y}, H(\underline{z})) = a_H'(\underline{z})\underline{y} = \psi_L(\underline{y}, \underline{z}), \quad \forall\, \underline{y} \in \mathbb{R}^{m-1}, \underline{z} \in \mathbb{R}^{(d+1)m-1}.$$

Now consider the optimal predictor $\psi : \mathbb{R}^{(d+1)m-1}$. Motivated by above we see that one way to introduce the notion of isotropy into the optimal predictor is to define a function $\psi_H : \mathbb{R}^{(m+2)(m-1)/2} \to \mathbb{R}$ such that $\psi(\underline{y}, \underline{u}, u_0) = \psi_H(\underline{y}, H(\underline{u}, u_0))$ (where $H(\cdot)$ is defined as above). Thus we can estimate $\psi_H(\cdot)$, using similar techniques to those discussed in Sections 4 and 5. Comparing the two dimensions $(((d+1)m-1)$ and $(m+2)(m-1)/2)$, we see the benefit of estimating $\psi_H(\cdot)$ rather than $\psi(\cdot)$ arises when d is large and m is small.

However when m is also large, there is no advantage in estimating $\psi_H(\cdot)$ over $\psi(\cdot)$. This example illustrates that the benefits of using a known dependence structure to estimate $\psi(\cdot)$ are not as apparent as covariance estimation. Though it is of interest to investigate whether the dimension of $\psi(\cdot)$ can be reduced in a realistic way, in order to avoid the curse of dimensionality.

Nevertheless it is clear that if $Y_t(\cdot)$ departs significantly from stationarity estimating the optimal predictor rather than the best linear predict is preferable, even if the estimator converges to the parameter of interest at a slower rate. □

5.2 Consistency and rates of convergence

In this section we study the asymptotic sampling properties of the estimators $\widehat{\psi}(\cdot)$, derived using the results in Section 3.

Under Assumptions 2.1, 2.2 and 2.3 (with $n = 2$), Assumption 3.1 holds with $X_{t,i} = Y_{t,im}$ and $Z_{t,i} = Z_{t,i}^{(s)}$ and by appealing to Proposition 2.1 we have

$$\sup_{\substack{A\in\sigma(X_{t,i},Z_{t,i})\\ B\in\sigma(X_{\tau,j},Z_{\tau,j})}} |P(A\cap B) - P(A)P(B)| \leq C \begin{cases} |t-\tau|^{-\min(\alpha,\beta)}; & \text{if } i=j,\\ |t-\tau|^{-\min(\gamma,\beta)}; & \text{otherwise.}\end{cases}$$

Therefore, as in the previous section under the additional technical assumptions the theorem below follows immediately from Theorem 3.1.

THEOREM 5.1. *Suppose Assumptions 2.1, 2.2 and 2.3 holds, where α and γ are the mixing coefficients of the covariates $\{U_{t,i}\}_t$ over time and space, respectively, β is the mixing coefficient of the process $\{\Phi_t(u)\}_t$ (u is arbitrary but fixed). Let the multivariate vector time series $\{Y_{t,im}, Z_{t,i}^{(s)}\}_{t,i}$ satisfies Assumption 3.2 with constants $q_G, q_F, q \in (0,1)$ and $\delta_i > 0$. Let $\varpi = \min(q_F, q_G)$ and suppose $\min(\gamma,\beta) > 1/\varpi + 1/q$.*

Let $\widehat{\psi}(z)$ be defined as in (5.2), where K is a multivariate kernel of order $r > 0$ (Definition 3.1). In addition for each $i \in \mathbb{N}$ assume $\psi \cdot f_i^{(s)}$ and $f_i^{(s)}$ belong to $\mathfrak{G}^{\eta}_{l_i,\triangle_i}$ for $l_i, \triangle_i > 0$ (Definition 3.2), and that the function $f_i^{(s)}$ is bounded away from zero. Let $\rho_i = \min(r, l_i)$ for $i \in \mathbb{N}$. Then for all $z \in \mathbb{R}^{\eta}$ (with $\eta = (d+1)m - 1$)

(i) if $\alpha > 1/\varpi + 1/q$ and $b_i^{(s)} = O(T^{-1/(2\rho_i+\nu)})$, $i = 1,\ldots,N/m$, we have

$$|\widehat{\psi}(z) - \psi(z)| = O_p\Big(\frac{1}{N/m}\sum_{i=1}^{N/m} (\delta_i^2)^{\frac{2\rho_i}{2\rho_i+\eta}} \cdot (\triangle_i^2)^{\frac{\eta}{2\rho_i+\eta}} \cdot T^{\frac{-2\rho_i}{2\rho_i+\eta}}\Big) \tag{5.3}$$

(ii) if $1/q < \alpha \leq 1/\varpi + 1/q$ and $b_i^{(s)} = O((N\cdot T)^{-1/(2\rho_i+\kappa\eta)})$, $i = 1,\ldots,N/m$, with $\kappa := 1 + \varpi + q - \varpi\cdot q\cdot\alpha$ we have

$$|\widehat{\psi}(z) - \psi(z)| = O_p\Big(\frac{1}{N/m}\sum_{i=1}^{N/m} (\delta_i^2)^{\frac{2\rho_i}{2\rho_i+\kappa\eta}} \cdot (\triangle_i^2)^{\frac{\kappa\eta}{2\rho_i+\kappa\eta}} \cdot (N\cdot T)^{\frac{-2\rho_i}{2\rho_i+\kappa\eta}}\Big) \tag{5.4}$$

(iii) if $\alpha \leq 1/q$ *and* $b_i^{(s)} = O((N \cdot T^{\alpha q})^{-1/(2\rho_i+(1+q)\eta)})$, $i = 1, \ldots, N/m$ *we obtain*

$$|\widehat{\psi}(z) - \psi)| = O_p\Big(\frac{1}{N/m} \sum_{i=1}^{N/m} (\delta_i^2)^{\frac{2\rho_i}{2\rho_i+q\eta+\eta}} \cdot (\triangle_i^2)^{\frac{(1+q)\eta}{2\rho_i+q\eta+\eta}} \cdot (N \cdot T^{\alpha q})^{\frac{-2\rho_i}{2\rho_i+q\eta+\eta}}\Big) \tag{5.5}$$

REMARK 5.2. It is worth noting that Remarks 3.2, 4.2 and 4.3(ii), as well as Corollary 4.2 (with $\nu = \eta$) are also true for the optimal predictor $\widehat{\psi}(z)$. □

6 An Example

In this section we illustrate the methods proposed in Section 4 on real data. The data we consider are the prices of houses sold (in pounds, Stirling), in Stockport, Greater Manchester, United Kingdom, during the period January 2002 to December 2005. The Stockport region is about 100 square miles, and is divided into 79 districts. We identify the 79 districts by their postcode (districts SK1 2 - SK23 9, which we label as $1.2 - 23.9$, and note that districts which are geographically close have similar postcodes). We focus on modelling the log selling price of detached, semi-detached and town houses and individual apartments, denoted as D, S, T and A. The data was obtained from the National Land register, `http://www.landreg.gov.uk/propertyprice/interactive/`, where for each district the average selling prices of D, S, T and A evaluated over each quarter of the year (3 months) is given. We consider each quarter as one time unit. Since we are considering the data from 2002-2005, for any one district and property type, there are a maximum of 16 observations over time. However, for most districts, there are far fewer than 16 observations, since houses are not necessarily sold every quarter.

Despite the Stockport region being relatively small, there is quite a large economic disparity in deprivation over the region, and this is measured by the deprivation index. Usually this effects the property prices. The deprivation index, for each district, is evaluated every two years, and can be obtained from the Office of National Statistics, `http://www.neighbourhood.statistics.gov.uk/dissemination/`. Nationally, the deprivation level ranges from 0 to 32000, where 0 indicates the highest level of deprivation and 32000 the lowest. In Stockport the deprivation level ranges from 3737 (in district one) to 31205 (in district nine). The deprivation index in any given district is evaluated using a mixture of health and economic indicators, such as the unemployment rate, the state of health and the average income, in that area. All the data used in the analysis is available on the authors' websites.

In our analysis, the response variable is the log average selling price of property I at time t (where $I \in \{D, S, T, A\}$) denoted as $Y_t^{(I)}(w, x)$, and we use the district, w and the log deprivation value u as the covariates, and let $u = (w, x)$. We assume $Y_t^{(I)}(w, x)$ satisfies

(1.3). Our object here is to predict $Y_t^{(I)}(w_0, x_0)$ given the house prices at $(m-1)$ different covariate values at time t, $\{Y_t^{(I)}(w_i, x_i)\}_{i=1}^{m-1}$. From (4.1) the best linear predictor is

$$\tilde{\psi}_t(\{y_i, (w_i, x_i)\}, (w_0, x_0)) = \mathbb{E}[Y_t(w_0, x_0)] + \sum_{j=1}^{m-1} a_j(\{(w_i, x_i)\})\{(y_j - \mathbb{E}[Y_t(w_j, x_j)])\}, \quad (6.1)$$

where $\{a_j(\{w_i, x_i\})\}_j$ is a function of the unknown covariances and variances $c^{(I)}(0, (w, x), (v, y)) = \mathrm{cov}(Y_t^{(I)}(w, x), Y_t^{(I)}(v, y))$ and $v^{(I)}(w, x) = \mathrm{var}(Y_t^{(I)}(w, x))$ (see Section 4). By estimating the covariances nonparametrically we obtain an estimator of the linear predictor. We model the covariances using, for $I \in \{D, S, T, A\}$, the two covariance models

$$\begin{aligned} \text{Model 1} \quad & \mathrm{cov}(Y_t^{(I)}(w, x), Y_t^{(I)}(v, y)) = c_1^{(I)}(0, x, y, w, v) \\ \text{Model 2} \quad & \mathrm{cov}(Y_t^{(I)}(w, x), Y_t^{(I)}(v, y)) = c_2^{(I)}(0, x, y, |w - v|), \end{aligned}$$

(clearly Model 1 includes as a special case Model 2). We see that Model 1, assumes that the house prices are 'spatially' nonstationary, whereas Model 2 assumes that the house prices are spatially isotropic (the dependence between two locations depends only their distance). We will estimate the two covariances and use them to obtain the predictions.

Define the residuals $\xi_t(w, x) = Y_t^{(I)}(w, x) - \mathbb{E}(Y_t^{(I)}(w, x))$. Using the data from January 2002 - September 2005 we estimate the mean $f^{(I)}(w, x) := \mathbb{E}[Y_t^{(I)}(w, x)]$ nonparametrically, which we denote as $\hat{f}^{(I)}(w, x)$. We use $\hat{\xi}_t^{(I)}(w, x) = Y_t^{(I)}(w, x) - \hat{f}^{(I)}(w, x)$ as an estimator of the residuals $\xi_t^{(I)}(w, x)$. Using the estimated residuals $\{\hat{\xi}_t^{(I)}(w, x)\}$ from the period January 2002 - September 2005 and the methods described in Section 4, we estimate the covariances $c_1^{(I)}(\cdot)$ and $c_2^{(I)}(\cdot)$, by smoothing over districts and the log deprivation indices.

For each house type I we randomly select 6 different locations $\{(w_i^{(I)}, x_i^{(I)}); i = 1, \ldots, 6\}$ (hence $m = 7$) and use the corresponding house prices $\{Y_t(w_i^{(I)}, x_i^{(I)}); i = 1, \ldots, 6\}$ to predict the selling price at other $11 - 19$ (depending on the availability of data) randomly chosen locations $\{(w_{0,j}^{(I)}, x_{0,j}^{(I)}) : j = 1, \ldots, n_I\}$. Depending on which covariance $c_1^{(I)}(\cdot)$ or $c_2^{(I)}(\cdot)$ is used to define the predictor $\tilde{\psi}(\cdot)$ in (6.1) we denote the predictor for each location as $\{\hat{Y}_{1,t}(w_{0,j}^{(I)}, x_{0,j}^{(I)}) : j = 1, \ldots, n_I\}$ or $\{\hat{Y}_{2,t}(w_{0,j}^{(I)}, x_{0,j}^{(I)}) : j = 1, \ldots, n_I\}$, respectively. We calculate the mean squared error for $k \in \{1, 2\}$ and $I \in \{D, S, T, A\}$ as

$$\sigma_{k,I}^2 = \frac{1}{n_I} \sum_{j=1}^{n_I} (\hat{Y}_{k,t}(w_{0,j}^{(I)}, x_{0,j}^{(I)}) - Y_t(w_{0,j}^{(I)}, x_{0,j}^{(I)}))^2,$$

which we compare with the mean squared error obtained using the average value $\hat{f}^{(I)}$ as the predictor at the point $(w_{0,j}^{(I)}, x_{0,j}^{(I)})$:

$$s_I^2 = \frac{1}{n_I} \sum_{j=1}^{n_I} (\hat{f}^{(I)}(w_{0,j}^{(I)}, x_{0,j}^{x}(I)) - Y_t(w_{0,j}^{(I)}, x_{0,j}^{(I)}))^2.$$

The results are summarised in Table 1. To see how the deprivation index may influence

	Detached	Semi-detacted	Town House	Apartment
Model 1: $\sigma^2_{1,I}$	0.1177	0.3239	0.03055	0.12269
Model 2: $\sigma^2_{2,I}$	0.1085	0.3329	0.03043	0.11247
Model Mean: s^2_I	0.1709	0.1222	0.12696	0.18503

Table 1: Mean squared errors of house selling price

the dependence between house prices, we integrate over the district values in the covariance function, and estimate the function $c_3^{(I)}(x, y)$, where

$$c_3^{(I)}(x,y) = \mathbb{E}\left\{\xi_t^I(W,X)\xi_t^I(V,Y)|X=x, Y=y\right\} = \mathbb{E}\left\{\text{cov}(Y_t^{(I)}(W,X), Y_t^{(I)}(V,Y))|X=x, Y=y\right\},$$

which is function of the deprivation indices only. We plot these functional covariance estimates in Figure 1.

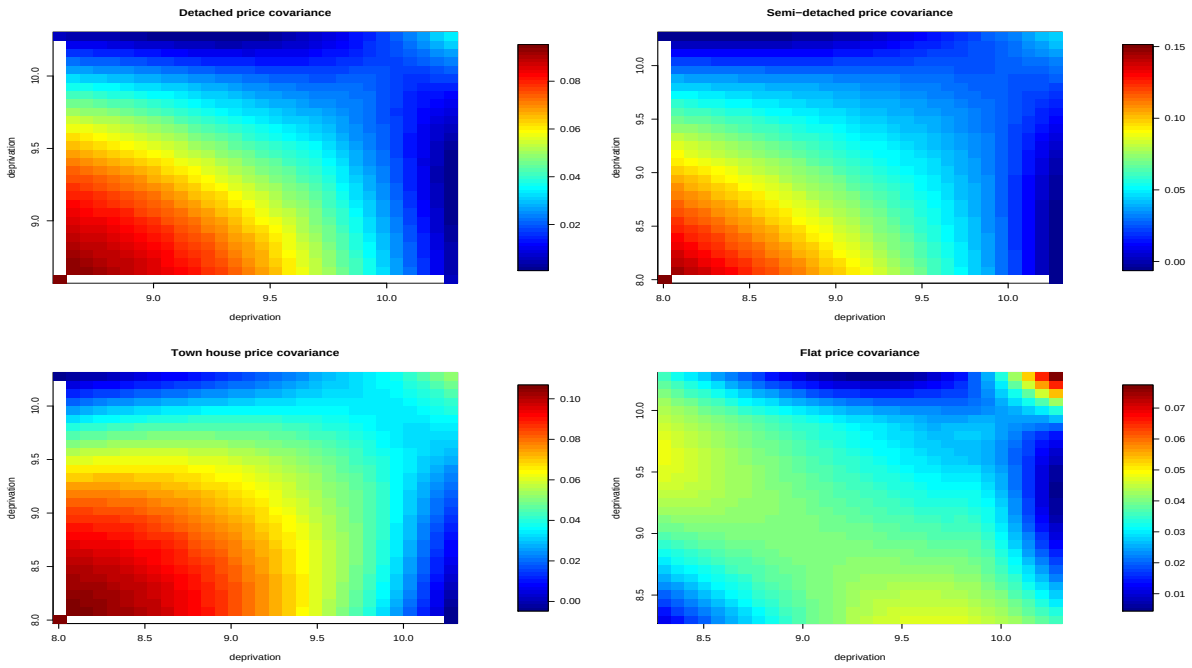

Figure 1: The top left is $c_3^{(D)}(\cdot,\cdot)$, top right is $c_3^{(S)}(\cdot,\cdot)$, the bottom left is $c_3^{(T)}(\cdot,\cdot)$ and the bottom right is $c_3^{(A)}(\cdot,\cdot)$.

From Table 1 we see that in general predicting the house price using house prices at other locations is better than using the mean estimator for that location. The exception is the predictor of the Semi-detached houses prices, where the mean estimator performed better. The covariance deprivation plots in Figure 1 appears to add weight to this. We see that for house types D, S and T there is greater linear dependence between houses in

areas with high deprivation (low deprivation index) than similar houses in low deprivation areas (large deprivation index). Indicating that for areas where there is less deprivation and more prosperity, the relationship between the houses may actually be nonlinear (meaning that $\{\Phi_t(\cdot)\}$ is non-Gaussian). The dependence also declines as the difference between the deprivation indices increases. Overall there appears to be higher dependence for both Detached and Town houses compared to Semi-detached houses. This may explain why the linear predictor for Semi-detached houses does not perform so well. The covariance trend seems to be different for the apartment covariance plot. This could be attributed to the fact that ownership of apartments is new in the UK, and in areas of high deprivation most apartments are rented.

Furthermore, the reduced Model 2 seems to adequately predict the house prices and nothing is gained by using the more general Model 1 in the prediction. In fact the mean squared error is slightly worse for the more general model, which is probably due to the curse of dimensionality in the estimation.

7 Concluding remarks

In this paper we have considered a class of spatio-temporal processes and studied spatial and temporal prediction using these models. We have proposed a nonparametric method to estimate the prediction function and we have shown that the convergence rates of the estimators depend on the temporal dependence of the observed covariates. We have shown that the estimators defined here belong to a quite general nonparametric setup in multivariate time series. And we have defined the estimator and sampling properties within this framework.

We believe that other estimation methods other than the Nadaraya-Watson type estimator can be used, for example, by using local polynomials, the methods of sieves or recursive estimation methods. Such methods may yield estimators which are faster to implement.

An advantage in defining the model (1.3) and (1.5) is that it includes many types of models, and prevents misspecification, which may occur using a parametric model.

The novel approach considered in this paper, motivates several possible avenues of future research, which we now briefly outline.

It is of interest to investigate the relevance of the model (1.3) in the prediction of financial assets given the price of other assets, where $Y_t(u)$ is the observed price of an asset described by the covariates u. However as the discussion below suggests this may require relaxing some of the assumptions on the model, and thus a different estimation approach. Suppose our object is to estimate $\psi(y,u,v) := \mathbb{E}(\Phi_t(v)|Y_{t-1}(u) = y)$ from a given set of observations. It is often the case in asset price modelling (see Ekeland, Heckman, and Nesheim (2002) for a labour market example) that the only available observations are $\{Y_t(U_{t,i}), t = 1,\ldots,T, i = 1,\ldots,N\}$ (see (1.5)) where the covariates $U_{t,i}$ are endogenous, i.e., the covariate $U_{t,i}$ and the observation error $V_{t,i}$ are correlated or more generally $\mathbb{E}(V_{t,i}|U_{t,i}) \neq 0$. In such a situation

we have $\psi(y, u_B, u_A) \neq \mathbb{E}(Y_t(U_{t,j})|Y_{t-1}(U_{t-1,i}) = y, U_{t-1,i} = u_B, U_{t,j} = u_A)$, in other words (1.6) does not hold true. Therefore $\hat{\psi}_{T,N}$ is an inappropiate estimator of the prediction function ψ. In this case a nonparametric instrumental variables approach (cf. Florens, Johannes, and Van Bellegem (2005)) may be required to obtain a suitable estimator.

We note that the assumption of temporal stationarity of the infinite dimension process $\{\Phi_t(\cdot)\}_t$ can be relaxed, to include locally stationary processes, where asymptotic results similar to those discussed in Dahlhaus and Subba Rao (2006) can be derived. From a practical point of view, this relaxation would include the case that the mean of $\phi_t(u)$ has a slowly changing time dependent mean, which may be of interest.

Acknowledgements

The authors would like to thank Professors Peter Green and Rainer Dahlhaus for making several useful suggestions and Gregory Berkolaiko, who helped us obtain the data.

A Appendix: Proofs

In this section we prove the results stated in the sections above.

A.1 Proof of Proposition 2.1

We prove Proposition 2.1 for the case $\underline{W}_t^{s,\underline{i}} = (\underline{\Phi}_t^{(s)}(\underline{U}_t^{(s,\underline{i})}), \underline{U}_t^{(s,\underline{i})})$. The case where $\underline{W}_t^{(s,\underline{i})} = (\underline{Y}_t^{(s,\underline{i})}, \underline{U}_t^{(s,\underline{i})})$ follows immediately, since the errors $V_{t,i}$ are iid and independent of $U_{t,i}$ and $\Phi_t(\cdot)$. Let $P_{\underline{W}}^{s,\underline{i}}$ denote the distribution of $\underline{W}_t^{s,\underline{i}}$, $P_{\underline{W}_t^{s,\underline{i}},\underline{W}_\tau^{s,\underline{j}}}$ denote the joint distribution of $\underline{W}_t^{s,\underline{i}}$ and $\underline{W}_\tau^{s,\underline{j}}$. In addition, we define the distribution of $\underline{U}_t^{s,\underline{i}}$ as $P_{\underline{U}_t^{s,\underline{j}}}$ and the joint distribution of $\underline{U}_t^{s,\underline{i}}$ and $\underline{U}_\tau^{s,\underline{j}}$ as $P_{\underline{U}_t^{s,\underline{i}},\underline{U}_\tau^{s,\underline{j}}}$. Suppose $\underline{u}$ and $\underline{v}$ are fixed and let $P_{\underline{\Phi}_{t,s}(\underline{u})}$ denote the distribution of $\underline{\Phi}_{t,s}(\underline{u})$ and $P_{\underline{\Phi}_{t,s}(\underline{u}),\Phi_{\tau,s}(\underline{v})}$ denote the joint distribution of $\underline{\Phi}_{t,s}(\underline{u})$ and $\underline{\Phi}_{\tau,s}(\underline{v})$. Conditioning on $\underline{U}_t^{s,\underline{i}}$ and $\underline{U}_\tau^{s,\underline{j}}$ we have

$$P_{\underline{W}_t^{s,\underline{i}},\underline{W}_\tau^{s,\underline{j}}} - P_{\underline{W}^{s,\underline{i}}} \otimes P_{\underline{W}^{s,\underline{i}}} = H_{\underline{\Phi}_{t,s}(\underline{u}),\underline{\Phi}_{\tau,s}(\underline{v})}\left(P_{\underline{U}_t^{s,\underline{i}},\underline{U}_\tau^{s,\underline{j}}}\right) + H_{\underline{U}_t^{s,\underline{i}},\underline{U}_\tau^{s,\underline{j}}}\left(P_{\underline{\Phi}_{t,s}(\underline{u})} \otimes P_{\underline{\Phi}_{\tau,s}(\underline{v})}\right), \quad \text{(A.1)}$$

where

$$\begin{aligned} H_{\underline{\Phi}_{t,s}(\underline{u}),\underline{\Phi}_{\tau,s}(\underline{v})} &= P_{\underline{\Phi}_{t,s}(\underline{u}),\underline{\Phi}_{\tau,s}(\underline{v})} - P_{\underline{\Phi}_{t,s}(\underline{u})} \otimes P_{\underline{\Phi}_{\tau,s}(\underline{v})} \\ \text{and } H_{\underline{U}_t^{s,\underline{i}},\underline{U}_\tau^{s,\underline{j}}} &= P_{\underline{U}_t^{s,\underline{i}},\underline{U}_\tau^{s,\underline{j}}} - P_{\underline{U}_t^{s,\underline{i}}} \otimes P_{\underline{U}_\tau^{s,\underline{j}}}. \end{aligned}$$

By using (A.1) we have

$$\sup_{\substack{A\in\sigma(\underline{W}_t^{s,\underline{i}}) \\ B\in\sigma(\underline{W}_\tau^{s,\underline{j}})}} |\mathbb{P}(A\cap B) - \mathbb{P}(A)\mathbb{P}(B)| \leq I + II \quad \text{(A.2)}$$

where

$$I = \sup_{\substack{A\in\sigma(\underline{W}_t^{s,\underline{i}})\\ B\in\sigma(\underline{W}_\tau^{s,\underline{j}})}} \Big| \int I_A(\underline{x},\underline{u}) I_B(\underline{y},\underline{v}) H_{\underline{\Phi}_{t,s}(\underline{u}),\underline{\Phi}_{\tau,s}(\underline{v})}(d\underline{x},d\underline{y}) P_{\underline{U}_t^{s,\underline{i}},\underline{U}_\tau^{s,\underline{j}}}(d\underline{u},d\underline{v})\Big|$$

$$II = \sup_{\substack{A\in\sigma(\underline{W}_t^{s,\underline{i}})\\ B\in\sigma(\underline{W}_\tau^{s,\underline{j}})}} \Big| \int I_A(\underline{x},\underline{u}) I_B(\underline{y},\underline{v}) P_{\underline{\Phi}_{t,s}(\underline{u})}(d\underline{x}) P_{\underline{\Phi}_{t,s}(\underline{v})}(d\underline{y}) H_{\underline{U}_t^{s,\underline{i}},\underline{U}_\tau^{s,\underline{j}}}(d\underline{u},d\underline{v})\Big|.$$

Under the assumption that $\{\underline{\Phi}_{t,s}(\underline{u})\}$ and $\{\underline{\Phi}_{t,s}(\underline{v})\}$ are 2-mixing (see Assumption 2.3 (i)), we appeal to Hall and Heyde (1980), Theorem A.5, to obtain

$$I \leq 4|t-\tau|^{-\beta} \cdot \left| \int C_{2m}(\underline{u},\underline{v}) P_{\underline{U}_t^{s,\underline{i}},\underline{U}_\tau^{s,\underline{j}}}(d\underline{u},d\underline{v}) \right| \leq 4C_{2m}|t-\tau|^{-\beta}. \tag{A.3}$$

Using similar arguments and Assumption 2.3(ii) it can be shown that

$$II \leq C \begin{cases} |t-\tau|^{-\alpha}; & \text{if } \underline{i}=\underline{j}, \\ |t-\tau|^{-\gamma}; & \text{if } (\underline{i},\underline{j}) \text{ distinct indices in } \mathbb{N}^{2n}, \end{cases}$$

Substituting this and (A.3) into (A.2) we obtain the result.

A.2 Proofs: Nonparametric regression with multivariate time series

LEMMA A.1. *Suppose the Assumptions 3.1 and 3.2 are satisfied, where* $\mathfrak{r}$ *and* $\mathfrak{u}$ *are the mixing coefficients of the vector time series* $\{(X_{t,i}, Z_{t,i}, X_{t,j}, Z_{t,j})\}_t$, *and where* $q_G, q_F, q \in (0,1)$ *and* $\delta_i > 0$, $i = 1,\ldots,N$ *are defined in Assumption 3.2.*

(i) If $1 \leq t,\tau \leq T$ *and* $1 \leq i < j \leq N$, *then*

$$|cov\big\{X_{t,i}K_{b_i}(Z_{t,i}-z), X_{\tau,j}K_{b_j}(Z_{\tau,j}-z)\big\}| \leq C \cdot \min\Big((b_ib_j)^{-\frac{\eta}{2}(1-q_G)}; \delta_i\delta_j \cdot (b_ib_j)^{-\frac{\eta}{2}(q+1)}|t-\tau|^{-q\mathfrak{u}}\Big); \tag{A.4}$$

$$|cov\big\{K_{b_i}(Z_{t,i}-z), K_{b_j}(Z_{\tau,j}-z)\big\}| \leq C \cdot \min\Big((b_ib_j)^{-\frac{\eta}{2}(1-q_F)}; \delta_i\delta_j \cdot (b_ib_j)^{-\frac{\eta}{2}(q+1)}|t-\tau|^{-q\mathfrak{u}}\Big), \tag{A.5}$$

where the constant C *does not depend on* i, j, t *or* τ.

(ii) If $1 \leq t,\tau \leq T$ *and* $1 \leq i \leq N$, *then*

$$|cov\{X_{t,i}K_{b_i}(Z_{t,i}-z), X_{\tau,i}K_{b_i}(Z_{\tau,i}-z)\}| \leq C \cdot \min\Big(b_i^{-\eta(1-q_G)}; \delta_i^2 \cdot b_i^{-\eta(q+1)}|t-\tau|^{-q\mathfrak{r}}\Big); \tag{A.6}$$

$$|cov\{K_{b_i}(Z_{t,i}-z), K_{b_i}(Z_{\tau,i}-z)\}| \leq C \cdot \min\Big(b_i^{-\eta(1-q_F)}; \delta_i^2 \cdot b_i^{-\eta(q+1)}|t-\tau|^{-q\mathfrak{r}}\Big), \tag{A.7}$$

where the constant C *does not depend on* i, t *or* τ.

PROOF. We only give the details for the proof of (A.4) and (A.6). The proofs of the other results are very similar and we omit the details.

Proof of (A.4) and (A.6). Writing the covariance as an integral, and using the notation in Assumption 3.2 we have

$$\operatorname{cov}\left\{X_{t,i}K_{b_i}(Z_{t,i}-z), X_{\tau,j}K_{b_j}(Z_{\tau,j}-z)\right\} = \int K_{b_i}(u-z)K_{b_j}(v-z)G_{t,\tau}^{(i,j)}(u,v)dudv.$$

Now by using Hölder's inequality with $p_G^{-1}+\bar{p}_G^{-1}=1$ it is clear that

$$|\operatorname{cov}\left\{X_{t,i}K_{b_i}(Z_{t,i}-z), X_{\tau,j}K_{b_j}(Z_{\tau,j}-z)\right\}| \leq (b_i\cdot b_j)^{-\eta/p_G}C_K^2\cdot C_G,$$

because under Assumption 3.2 we have $\|K\|_{\bar{p}_G}<C_K$ and $\|G_{t,\tau}^{(i,j)}\|_{p_G}<C_G$. This gives us the common bound in (A.4) and (A.6). On the other hand, under Assumption 3.2 (i) we have $\mathbb{E}[|X_{t,i}K_{b_i}(Z_{t,i}-z)|^p]<\infty$ for some $p=2/(1-q)>2$. Therefore, using the 2-mixing property of $\{(X_{t,i},Z_{t,i},X_{t,j},Z_{t,j})\}_t$ given in Assumption 3.1 together with Hall and Heyde (1980), Theorem A.6, for $i\neq j$, we obtain

$$\begin{aligned}&|\operatorname{cov}\left\{X_{t,i}K_{b_i}(Z_{t,i}-z), X_{\tau,j}K_{b_j}(Z_{\tau,j}-z)\right\}|\\ &\quad\leq 2q(2C)^q\cdot\{\mathbb{E}[|X_{1,i}K_{b_i}(Z_{1,i}-z)|^p]\cdot\mathbb{E}[|X_{1,j}K_{b_j}(Z_{1,j}-z)|^p]\}^{1/p}\cdot|t-\tau|^{-q\mathfrak{u}},\end{aligned}\tag{A.8}$$

while for $i=j$

$$\begin{aligned}&|\operatorname{cov}\left\{X_{t,i}K_{b_i}(Z_{t,i}-z), X_{\tau,i}K_{b_i}(Z_{\tau,i}-z)\right\}|\\ &\qquad\leq 2q(2C)^q\cdot\{\mathbb{E}[|X_{1,i}K_{b_i}(Z_{1,i}-z)|^p]\}^{2/p}\cdot|t-\tau|^{-q\mathfrak{r}}.\end{aligned}\tag{A.9}$$

Since under the Assumption 3.2 the p-th moment of the multivariate kernel K is bounded by C_K and the function $g_i^{(p)}(\cdot)=\mathbb{E}[|X_{1,i}|^p|Z_{1,i}=\cdot]f_i(\cdot)$ by δ_i^p, we have

$$\mathbb{E}[|X_{1,i}K_{b_i}(Z_{1,i}-z)|^p]^{1/p}\leq C_K\cdot\delta_i\cdot b_i^{-\frac{\eta}{2}(q+1)}.\tag{A.10}$$

Therefore, (A.8) together with (A.10) gives the second bound in (A.4), where (A.9) and (A.10) leads to the second bound in (A.6), which proves the result. □

LEMMA A.2. *Suppose Assumptions 3.1 and 3.2 are satisfied, where* $\mathfrak{r}$ *and* $\mathfrak{u}$ *are the mixing coefficients associated with the vector time series* $\{(X_{t,i},Z_{t,i},X_{t,j},Z_{t,j})\}_t$ *given in Assumption 3.1, and* $q_G,q_F,q\in(0,1)$ *and* $\delta_i>0$, $i=1,\ldots,N$ *are defined in Assumption 3.2. Let* $\varpi=\min(q_F,q_G)$ *and suppose* $\mathfrak{u}>1/\varpi+1/q$.

Consider the nonparametric estimators (3.4) constructed using a multivariate kernel of order $r>0$ *(Definition 3.1). In addition assume for each* $i=1,\ldots,N$, *that the functions* f_i *and* g_i *belong to* $\mathfrak{G}_{s_i,\triangle_i}^{\eta}$ *for* $s_i,\triangle_i>0$ *(Definition 3.2) and let* $\rho_i=\min(r,s_i)$.

(i) If $\mathfrak{r} > 1/\varpi + 1/q$, *then let* $b_i = O(T^{\frac{-1}{2\rho_i+\eta}})$, $i = 1, \ldots, N$ *and we have*

$$\mathbb{E}|\hat{g}(z) - g(z)|^2 = O\Big(\frac{1}{N}\sum_{i=1}^{N}(\delta_i^2)^{\frac{2\rho_i}{2\rho_i+\eta}} \cdot (\triangle_i^2)^{\frac{\eta}{2\rho_i+\eta}} \cdot T^{\frac{-2\rho_i}{2\rho_i+\eta}}\Big), \tag{A.11}$$

$$\mathbb{E}|\hat{f}(z) - f(z)|^2 = O\Big(\frac{1}{N}\sum_{i=1}^{N}(\delta_i^2)^{\frac{2\rho_i}{2\rho_i+\eta}} \cdot (\triangle_i^2)^{\frac{\eta}{2\rho_i+\eta}} \cdot T^{\frac{-2\rho_i}{2\rho_i+\eta}}\Big); \tag{A.12}$$

(ii) If $1/q < \mathfrak{r} \leq 1/\varpi + 1/q$, *then assume* $\kappa := 1 + \varpi + q - \varpi q\mathfrak{r}$ *and* $b_i = O((N \cdot T)^{\frac{-1}{2\rho_i+\kappa\eta}})$, $i = 1, \ldots, N$ *and we obtain*

$$\mathbb{E}|\hat{g}(z) - g(z)|^2 = O\Big(\frac{1}{N}\sum_{i=1}^{N}(\delta_i^2)^{\frac{2\rho_i}{2\rho_i+\kappa\eta}} \cdot (\triangle_i^2)^{\frac{\kappa\eta}{2\rho_i+\kappa\eta}} \cdot (N \cdot T)^{\frac{-2\rho_i}{2\rho_i+\kappa\eta}}\Big), \tag{A.13}$$

$$\mathbb{E}|\hat{f}(z) - f(z)|^2 = O\Big(\frac{1}{N}\sum_{i=1}^{N}(\delta_i^2)^{\frac{2\rho_i}{2\rho_i+\kappa\eta}} \cdot (\triangle_i^2)^{\frac{\kappa\eta}{2\rho_i+\kappa\eta}} \cdot (N \cdot T)^{\frac{-2\rho_i}{2\rho_i+\kappa\eta}}\Big); \tag{A.14}$$

(iii) If $\mathfrak{r} \leq 1/q$, *then given* $b_i = O((N \cdot T^{q\mathfrak{r}})^{\frac{-1}{2\rho_i+(1+q)\eta}})$, $i = 1, \ldots, N$ *we have*

$$\mathbb{E}|\hat{g}(z) - g(z)|^2 = O\Big(\frac{1}{N}\sum_{i=1}^{N}(\delta_i^2)^{\frac{2\rho_i}{2\rho_i+q\eta+\eta}} \cdot (\triangle_i^2)^{\frac{q\eta+\eta}{2\rho_i+q\eta+\eta}} \cdot (N \cdot T^{q\mathfrak{r}})^{\frac{-2\rho_i}{2\rho_i+q\eta+\eta}}\Big), \tag{A.15}$$

$$\mathbb{E}|\hat{f}(z) - f(z)|^2 = O\Big(\frac{1}{N}\sum_{i=1}^{N}(\delta_i^2)^{\frac{2\rho_i}{2\rho_i+q\eta+\eta}} \cdot (\triangle_i^2)^{\frac{q\eta+\eta}{2\rho_i+q\eta+\eta}} \cdot (N \cdot T^{\mathfrak{r}q})^{\frac{-2\rho_i}{2\rho_i+q\eta+\eta}}\Big). \tag{A.16}$$

Proof. We mention that parts of the following proof are motivated by techniques used in Bosq (1998), where nonparametric smoothing was considered for univariate time series. We only give the details for the proofs of the MSE of the estimator $\hat{g}$ in the three different cases (i) -(iii). The proofs of the other results are very similar and we omit the details. Consider the standard variance bias decomposition

$$\mathbb{E}|\hat{g}(z) - g(z)|^2 = \mathrm{var}(\hat{g}(z)) + |\mathbb{E}\hat{g}(z) - g(z)|^2. \tag{A.17}$$

Under the stated assumptions we will derive the following four bounds. The bias is bounded by

$$|\mathbb{E}\hat{g}(z) - g(z)|^2 \leq C \cdot \frac{1}{N}\sum_{i=1}^{N} \triangle_i^2 \cdot b_i^{2\rho_i}. \tag{A.18}$$

For the variance, if $\mathfrak{r} > 1/q_G + 1/q$, then

$$\mathrm{var}(\hat{g}(z)) \leq T^{-1} \cdot C \cdot \frac{1}{N}\sum_{i=1}^{N} \delta_i^2 \cdot b_i^{-\eta}; \tag{A.19}$$

if $1/q < \mathfrak{r} \leq 1/q_G + 1/q$

$$\text{var}(\hat{g}(z)) \leq (N \cdot T)^{-1} \cdot C \cdot \frac{1}{N} \sum_{i=1}^{N} \delta_i^2 \cdot b_i^{-\eta(1+q_g+q-q_g q\mathfrak{r})} + T^{-1} \cdot C \cdot \frac{1}{N} \sum_{i=1}^{N} \delta_i^2 \cdot b_i^{-\eta}; \tag{A.20}$$

while if $\mathfrak{r} \leq 1/q$

$$\text{var}(\hat{g}(z)) \leq (N \cdot T^{q\mathfrak{r}})^{-1} \cdot C \cdot \frac{1}{N} \sum_{i=1}^{N} \delta_i^2 \cdot b_i^{-\eta(q+1)} + T^{-1} \cdot C \cdot \frac{1}{N} \sum_{i=1}^{N} \delta_i^2 \cdot b_i^{-\eta}; \tag{A.21}$$

where the constant C does not depend on N or T. Furthermore, the stated bandwidths b_i, $i = 1, \ldots, N$ ensure the balance between the variance and the bias terms and lead to the bounds given in (A.11), (A.13) and (A.15).

Proof of (A.18). Using iterative conditional expectation we can write

$$\mathbb{E}\hat{g}(z) = \frac{1}{NT} \sum_{i=1}^{N} \sum_{t=1}^{T} \mathbb{E}\Big(\mathbb{E}[X_{t,i}|Z_{t,i}] K_{b_i}(Z_{t,i} - z) \Big) = \frac{1}{N} \sum_{i=1}^{N} \int du \; g_i(u) K_{b_i}(u - z)$$

where $g_i(\cdot) = \mathbb{E}[X_{t,i}|Z_{t,i} = \cdot] f_i(\cdot)$ for all t, i. Since $g = \frac{1}{N}\sum_{i=1}^{N} g_i$ with $g_i \in \mathfrak{G}^{\eta}_{s_i,\triangle_i}$ and K is a multivariate kernel of order r with $\int du |u|^r K(u) \leq S_K$, using a Taylor expansion up to the power $\rho_i = \min(r, s_i)$ leads to $\mathbb{E}\hat{g}(z) = g(z) + \frac{1}{N}\sum_{i=1}^{N} b_i^{\rho_i} R_i$ with reminder $|R_i| \leq \triangle_i S_K < \infty$. Thus applying Jensens inequality we obtain (A.18).

In order to proof (A.19)-(A.21), we consider the expansion

$$\text{var}(\hat{g}(z)) = A_1 + A_2 + A_3 + A_4 \tag{A.22}$$

with

$$\begin{aligned}
A_1 &= \frac{1}{N^2T^2} \sum_{t=1}^{T} \sum_{i=1}^{N} \text{var}\left\{ X_{t,i} K_{b_i}(Z_{t,i} - z) \right\} , \\
A_2 &= \frac{2}{N^2T^2} \sum_{t=1}^{T} \sum_{j>i} \text{cov}\left\{ X_{t,i} K_{b_i}(Z_{t,i} - z), X_{t,j} K_{b_i}(Z_{t,j} - z) \right\} , \\
A_3 &= \frac{4}{N^2T^2} \sum_{t>\tau} \sum_{j>i} \text{cov}\left\{ X_{t,j} K_{b_i}(Z_{t,i} - z), X_{\tau,j} K_{b_j}(Z_{\tau,j} - z) \right\} , \\
A_4 &= \frac{2}{N^2T^2} \sum_{t>\tau} \sum_{i=1}^{N} \text{cov}\left\{ X_{t,i} K_{b_i}(Z_{t,i} - z), X_{\tau,i} K_{b_i}(Z_{\tau,i} - z) \right\} .
\end{aligned}$$

We will show that $|A_1|, |A_2|, |A_3| \leq T^{-1} \cdot C \cdot \frac{1}{N}\sum_{i=1}^{N} \delta_i^2 \cdot b_i^{-\eta}$. Furthermore, if $0 \leq \mathfrak{r} \leq 1/q_G + 1/q$ then these terms are dominated by $|A_4|$. Whereas for $\mathfrak{r} > 1/q_G + 1/q$ all the terms are of the same order. Therefore, the bounds derived for $|A_4|$ will lead to the estimates in (A.19)-(A.21).

First let us consider A_1. Due to the stationarity of the process, we have the bound

$$N \cdot T \cdot A_1 \leq \frac{1}{N}\sum_{i=1}^{N}\mathbb{E}[X_{1,i}^2 K_{b_i}^2(Z_{1,i}-z)] = \frac{1}{N}\sum_{i=1}^{N}\int du\; g_i^{(2)}(u)K_{b_i}^2(u-z)$$
$$\leq \frac{1}{N}\sum_{i=1}^{N}\int du\; \{g_i^{(p)}(u)\}^{2/p}K_{b_i}^2(u-z),$$

where $g_i^{(p)}(\cdot) := \mathbb{E}[|X_{1,i}|^p|Z_{1,i}=\cdot]f_i(\cdot)$ is well defined because $\mathbb{E}(|X_{1,i}|^p) < \infty$. Since under the stated assumptions $\|K\|_2 < C_K$ and the function $(g_i^{(p)})^{1/p}$ is bounded by δ_i this leads to $A_1 \leq (N\cdot T)^{-1}\cdot C_K^2 \cdot \frac{1}{N}\sum_{i=1}^{N}\delta_i^2 \cdot b_i^{-\eta}$.

It is straightforward to show that $T\cdot|A_2|$ is bounded by $\frac{2}{N^2}\sum_{j>i}\mathrm{cov}\left\{X_{1,i}K_{b_i}(Z_{1,i}-z), X_{1,j}K_{b_j}(Z_{1,j}-z)\right\}$. Furthermore by using the Cauchy-Schwarz inequality we have

$$T\cdot|A_2| \leq \frac{2}{N^2}\sum_{j>i}\mathrm{var}(X_{1,i}K_{b_i}(Z_{1,i}-z))^{1/2}\mathrm{var}(X_{1,j}K_{b_j}(Z_{1,j}-z))^{1/2}$$
$$\leq C_K^2\frac{2}{N^2}\sum_{j>i}\delta_i\delta_j\cdot(b_ib_j)^{-\eta/2},$$

where the last line of the above follows by applying the same arguments as those used for for A_1. Therefore usings Jensen's inequality we obtain $|A_2| \leq T^{-1}\cdot C_K^2\cdot\frac{1}{N}\sum_{i=1}^{N}\delta_i^2\cdot b_i^{-\eta}$.

The term $T\cdot|A_3|$ is bound by the sum

$$T\cdot|A_3| \leq \frac{8}{N^2}\sum_{j>i}\sum_{t=1}^{T}|\mathrm{cov}\left\{X_{t,i}K_{b_i}(Z_{t,i}-z), X_{1,j}K_{b_j}(Z_{1,j}-z)\right\}|.$$

To bound the above we partition the inner sum into two parts which we estimate separately using the bounds in (A.4) of Lemma A.1, thus giving us

$$T\cdot|A_3| \leq C_K^2C_{\mathfrak{u}}\frac{8}{N^2}\sum_{j>i}\Big\{\sum_{t=1}^{u_T^{(i,j)}}(b_ib_j)^{-\frac{\eta}{2}(1-q_G)} + \sum_{t=u_T^{(i,j)}+1}^{T}\delta_i\delta_j(b_ib_j)^{-\frac{\eta}{2}(q+1)}t^{-q\mathfrak{u}}\Big\}$$
$$\leq C_K^2C_{\mathfrak{u}}\frac{8}{N^2}\sum_{j>i}\{u_T^{(i,j)}(b_ib_j)^{-\frac{\eta}{2}(1-q_G)} + (u_T^{(i,j)})^{-q\mathfrak{u}+1}\delta_i\delta_j(b_ib_j)^{-\frac{\eta}{2}(q+1)}\}.$$

Thereby using $u_T^{(i,j)} \approx (b_ib_j)^{-\frac{\eta}{2}q_G}\cdot\delta_i\delta_j$ we obtain

$$T\cdot|A_3| \leq C_K^2C_{\mathfrak{u}}\frac{8}{N^2}\sum_{j>i}\{\delta_i\delta_j\cdot(b_ib_j)^{-\eta/2} + (b_ib_j)^{-\frac{\eta}{2}(1+q_G+q-q_Gq\mathfrak{u})}\cdot(\delta_i\delta_j)^{-q\mathfrak{u}+2}\}.$$

Since under the assumptions of the Lemma $\mathfrak{u} > 1/q_G + 1/q$, the second summand is bounded by the first, this together with Jensen's inequality leads to $|A_3| \leq T^{-1}\cdot 16C_K^2C_{\mathfrak{u}}\cdot\frac{1}{N}\sum_{i=1}^{N}\delta_i^2\cdot b_i^{-\eta}$.

The term $N{\cdot}T{\cdot}|A_4|$ is bounded by $\frac{4}{NT}\sum_{i=1}^{N}\sum_{t=1}^{T}|\mathrm{cov}\{X_{t,i}K_{b_i}(Z_{t,i}-z),X_{1,i}K_{b_i}(Z_{1,i}-z)\}|$. We now derive bounds for $N\cdot T\cdot|A_4|$ for different mixing rates. If $\mathfrak{r}\le 1/q$ then we estimate the sum using the second bound in (A.6) of Lemma A.1, i.e., $N\cdot T\cdot|A_4|\le T^{-q\mathfrak{r}+1}\cdot 4C_K^2C_{\mathfrak{r}}\cdot\frac{1}{N}\sum_{i=1}^{N}\delta_i^2\cdot b_i^{-\eta(q+1)}$, which is (A.19). On the other hand if $\mathfrak{r}>1/q$ we partition the inner sum into two parts (similar to A_3) and estimate them separately using now the two bounds in (A.6), which leads with $u_T^{(i,i)}\approx b_i^{-\eta q_G}\cdot\delta_i^2$ to

$$N\cdot T\cdot|A_4|\le C_K^2C_{\mathfrak{r}}\frac{4}{N}\sum_{i=1}^{N}\{\delta_i^2\cdot b_i^{-\eta}+\delta_i^{2(-q\mathfrak{r}+2)}\cdot b_i^{-\eta(1+q_G+q-q_Gq\mathfrak{r})}\}.$$

Therefore, if $\mathfrak{r}>1/q+1/q_G$, then the second term of the above is negligible wrt. to the first and we obtain $|A_4|\le(N\cdot T)^{-1}\cdot 8C_K^2C_{\mathfrak{r}}\cdot\frac{1}{N}\sum_{i=1}^{N}\delta_i^2\cdot b_i^{-\eta}$ which proves (A.21). While if $1/q<\mathfrak{r}\le 1/q+1/q_G$, we partition the inner sum using $u_T^{(i,i)}\approx b_i^{-\eta q_G}$ and obtain

$$N\cdot T\cdot|A_4|\le C_K^2C_{\mathfrak{r}}\cdot\frac{4}{N}\sum_{i=1}^{n}\{b_i^{-\eta}+\delta_i^2\cdot b_i^{-\eta(1+q_G+q-q_Gq\mathfrak{r})}\}.$$

The second term of the above is now the leading one and we have $|A_4|\le(N\cdot T)^{-1}\cdot 8C_K^2C_{\mathfrak{r}}\cdot\frac{1}{N}\sum_{i=1}^{n}\delta_i^2\cdot b_i^{-\eta(1+q_G+q-q_Gq\mathfrak{r})}$, which gives (A.20). □

Corollary A.3. *Suppose the assumptions of Lemma A.2 are satisfied. Let the bandwidth parameters are such that* $b_i=O(T^{-1/(2\rho_i+\eta)})$, $i=1,\dots,N$, *and define* $\rho=\min\{\rho_i;i=1,\dots,N\}$). *In addition, in the case Lemma A.2(ii), where* $q^{-1}<\mathfrak{r}\le\varpi^{-1}+q^{-1}$, *assume* $N=O(T^{\frac{\eta(\kappa-1)}{2\rho+\eta}})$; *while, in the case Lemma A.2(iii) where* $\mathfrak{r}<q^{-1}$, *assume* $N=O(T^{\frac{q\eta}{2\rho+\eta}+1-\mathfrak{r}q})$. *Then*

$$\mathbb{E}|\hat{g}(z)-g(z)|^2=O\Big(\frac{1}{N}\sum_{i=1}^{N}(\delta_i^2)^{\frac{2\rho_i}{2\rho_i+\eta}}\cdot(\triangle_i^2)^{\frac{\eta}{2\rho_i+\eta}}\cdot T^{\frac{-2\rho_i}{2\rho_i+\eta}}\Big),\tag{A.23}$$

$$\mathbb{E}|\hat{f}(z)-f(z)|^2=O\Big(\frac{1}{N}\sum_{i=1}^{N}(\delta_i^2)^{\frac{2\rho_i}{2\rho_i+\eta}}\cdot(\triangle_i^2)^{\frac{\eta}{2\rho_i+\eta}}\cdot T^{\frac{-2\rho_i}{2\rho_i+\eta}}\Big).\tag{A.24}$$

Proof. The proof follows in the same spirit as the proof of Lemma A.2. Consider the bounds (A.18)-(A.21) of the bias and the variance term of the estimator $\hat{g}$, where their sum estimates the MSE using its standard decomposition. The bounds are still valid under the assumptions of Corollary A.3. Moreover the conditions on the bandwidth b and the additional assumption on the number of locations N ensures the balance between the variance and the bias term and leads to the result. The proofs of the other results are very similar and we omit the details. □

A.3 Proofs for Sections 4 and 5

PROOF OF THEOREM 3.1. Consider the decomposition

$$\begin{aligned}\hat{\psi}(z) - \psi(z) &= \frac{\hat{g}(z)}{\hat{f}(z)} - \frac{\hat{f}(z)}{\hat{f}(z)}\psi(z) \\ &= \frac{\hat{g}(z) - \hat{f}(z)\psi(z)}{f(z)} + \frac{f(z) - \hat{f}(z)}{\hat{f}(z)} \cdot \frac{\hat{g}(z) - \hat{f}(z)\psi(z)}{f(z)}.\end{aligned}$$

We first show that the second term in the decomposition is negligiable in comparision to the first term. Using this we can apply Lemma A.2, to obtain the result. Now Lemma A.2 gives $\mathbb{E}|f(z) - \hat{f}(z)|^2 = o(1)$ which implies that $|\hat{f}(z)^{-1}|$ is bounded in probability and therefore the second term is of order $o_p(\{\hat{g}(z) - \hat{f}(z)\phi(z)\}/f(z))$. □

PROOF OF COROLLARY 3.2. By using Corollary A.3 and the same proof of Theorem 3.1 we obtain the result. □

PROOF OF THEOREM 4.1. By using Corollary A.3 and the same proof of Theorem 3.1 we obtain the result. □

PROOF OF COROLLARY 4.2. By using Corollary A.3 and the same proof of Theorem 3.1 we obtain the result. □

PROOF OF COROLLARY 4.3. By using Corollary A.3 and the same proof of Theorem 3.1 we obtain the result. □

PROOF OF THEOREM 5.1. By using Corollary A.3 and the same proof of Theorem 3.1 we obtain the result. □

A.4 Mixing properies of the location dependent spatio-temporal AR process

We now show that the location dependent spatio-temporal model defined in (2.1) satisfies the mixing conditions stated in Assumption 2.3(i). Let $\underline{u}_n = (u_1, \ldots, u_n) \in \Omega^n$ and define $\underline{\phi}_t(\underline{u}_n) = (\Phi_t(u_1), \ldots, \Phi_t(u_n))$ and $\underline{\xi}_t(\underline{u}_n) = (\xi_t(u_1), \ldots, \xi_t(u_n))$. We will show that the vector process $\{\underline{\phi}_t(\underline{u}_n)\}_t$ is α-mixing (with a geometric rate that is same for all $\underline{u}_n \in \Omega$), which is stronger than the required 2-mixing assumption.

Suppose the process $\{\Phi_t(u)\}_t$ satisfies (2.1). We shall assume that for all u the absolute values of the roots of the characteristic polynomial associated with the AR process in (2.1) are less than δ, where $0 < \delta < 1$, and $\sup_{u\in\Omega} \sigma(u) < \sigma$ for some $\sigma < \infty$. We observe that $\Phi_t(u)$ has the unique causal solution

$$\Phi_t(u) = \sum_{j=0}^{\infty} c_j(u)\xi_t(u), \tag{A.25}$$

where there exists a $\mathcal{C} < \infty$ and $\delta < \rho < 1$ such that $\sup_u |c_j(u)| < \mathcal{C}\rho^j$. In order to obtain the mixing rate we define the sigma-algebras $\mathcal{F}_t^{\infty}(\underline{u}_n) = \sigma(\underline{\phi}_t(\underline{u}_n), \underline{\phi}_{t+1}(\underline{u}_n), \ldots)$,

$$\mathcal{F}_{-\infty}^{0}(\underline{u}_n) = \sigma(\ldots, \underline{\phi}_{-1}(\underline{u}_n), \underline{\phi}_0(\underline{u}_n)) = \sigma(\ldots, \underline{\xi}_{-1}(\underline{u}_n), \underline{\xi}_0(\underline{u}_n))$$

and $\mathcal{F}_t^{t+p-1}(\underline{u}_n) = \sigma(\underline{\phi}_t(\underline{u}_n), \ldots, \underline{\phi}_{t+p}(\underline{u}_n))$. It is clear that for the location dependent AR process we do not need to use the entire upper tail $\mathcal{F}_t^{\infty}$ to obtain the mixing coefficient, in other words

$$\beta_t(\underline{u}_n) = \sup_{\substack{A\in\mathcal{F}_{-\infty}^{0}(\underline{u}_n) \\ B\in\mathcal{F}_t^{\infty}(\underline{u}_n)}} |P(A\cap B) - P(A)P(B)| = \sup_{\substack{A\in\mathcal{F}_{-\infty}^{0}(\underline{u}_n) \\ B\in\mathcal{F}_t^{t+p-1}(\underline{u}_n)}} |P(A\cap B) - P(A)P(B)|. \quad \text{(A.26)}$$

Let $\Sigma(\underline{u}_n) = \text{var}(\underline{\xi}_t(\underline{u}_n))$, and define the vector $\underline{\tilde{\eta}}_t(\underline{u}_n) = \tilde{\eta}_t(1, \underline{u}_n), \ldots, \tilde{\eta}_t(n, \underline{u}_n)) = \Sigma(\underline{u}_n)^{-1/2}\xi_t(\underline{u}_n)$ (if the inverse does not exist we use the generalised inverse). We define the vector $\underline{\eta}_t(\underline{u}_n) = (\eta_t(1, \underline{u}_n), \ldots, \eta_t(n, \underline{u}_n)) = \text{var}(\underline{\tilde{\eta}}_t(\underline{u}_n))^{-1/2}\underline{\tilde{\eta}}_t(\underline{u}_n)$, and it is clear that the transformed innovations $\underline{\eta}_t(\underline{u}_n) \sim MVN(0, I_{np})$ (where I_{np} is a $np \times np$ identity matrix, but the diagonal can contain zeros if $\Sigma(\underline{u}_n)$ is singular). Then we have $\mathcal{F}_{-\infty}^{0}(\underline{u}_n) = \otimes_{i=1}^{n}\mathcal{F}_{-\infty}^{0}(\eta_i)$, where $\mathcal{F}_{-\infty}^{0}(\eta_i) = \sigma(\ldots, \eta_{-1}(i, \underline{u}_n), \eta_0(i, \underline{u}_n))$. We now make a similar decomposition of the sigma-algebra $\mathcal{F}_t^{t+p-1}(\underline{u}_n)$.

We define the stochastic process $\{Y_t(\eta_i, u_j)\}_t$, which for $1 \leq i, j \leq n$ has the representation

$$Y_t(\eta_i, u_j) = \sum_{r=1}^{p} a_r(u_j)Y_{t-r}(\eta_i, u_j) + \sigma(u_j)\eta_t(i, \underline{u}_n) = \sum_{k=0}^{\infty} c_k(u_j)\eta_{t-k}(i, \underline{u}_n), \quad \text{(A.27)}$$

we note the last term of the above was obtained by using (A.25). Since

$$\Phi_t(u_j) = \sum_{i=1}^{n} \frac{\rho_{i,j}}{\sigma_i} Y_t(\eta_i, u_j),$$

where $\sigma_i^2 = \text{var}(\tilde{\eta}_{t-k}(i, \underline{u}_n))$ $\Phi_t(u_j)$ can linearly be transformed into $Y_t(\eta_i, u_j)$ and we have

$$\mathcal{F}_t^{t+p-1}(\underline{u}_n) = \otimes_{i=1}^{n}\mathcal{F}_t^{t+p-1}(\eta_i, \underline{u}_n),$$

where $\mathcal{F}_t^{t+p-1}(\eta_i, \underline{u}_n) = \sigma(Y_t(\eta_i, u_1), \ldots, Y_t(\eta_i, u_n), \ldots, Y_{t+p-1}(\eta_i, u_1), \ldots, Y_{t+p-1}(\eta_i, u_n))$. Finally we decompose $\mathcal{F}_t^{t+p-1}(\eta_i, \underline{u}_n)$ into independent sigma algebras.

Let $\Delta(\eta_i)$ be a $(np \times np)$-dimensional matrix, where

$$\Delta(\eta_i) = \text{var}\left\{(Y_t(\eta_i, u_1), \ldots, Y_t(\eta_i, u_n), \ldots, Y_{t+p-1}(\eta_i, u_1), \ldots, Y_{t+p-1}(\eta_i, u_n))\right\}$$

and define $\Lambda = \Delta(\eta_i)^{-1/2}$. We now decompose the Gaussian random vector $\underline{W}_t^i$ into independent random variables. For $s = 1, \ldots, np$ and $t \in \mathbb{Z}$ let

$$Z_{t,s}(\eta_i) = \frac{1}{\sqrt{\sum_{r=0}^{n-1}\sum_{j=0}^{p-1}\Lambda_{pr+j,s}^2}} \sum_{r=0}^{n-1}\sum_{j=0}^{p-1} \Lambda_{pr+j,s}Y_{t+j}(\eta_i, u_r). \quad \text{(A.28)}$$

Again by Gaussianity it is clear that $\{Z_{t,s}(\eta_i)\}_{s=1}^{mp}$ are independent random variables. By substituting (A.27) into (A.28) we have that $Z_{t,s}^i$ has the MA(∞) solution

$$Z_{t,s}(\eta_i) = \frac{1}{\sqrt{\sum_{r=0}^{n-1}\sum_{j=0}^{p-1}\Lambda_{pr+j,s}^2}} \sum_{k=0}^{\infty}\sum_{r=0}^{n-1}\sum_{j=0}^{p-1}\Lambda_{pr+j,s}c_k(u_j)\eta_{t+r-k,i} = \sum_{k=0}^{\infty}\alpha_k(\eta_i,\underline{u}_n)\eta_{t+n-k,i}.$$

From the above it is clear that $\{Z_{t,s}(\eta_i)\}_t$ is an MA(∞) process. Furthermore, by using that $\sup_k c_k(u_j) < \mathcal{C}\rho^k$ we have for all i and k that $\alpha_k(\eta_i,\underline{u}_n) < \mathcal{C}np\rho^k$. Using this we appeal to the strong mixing result for MA(∞) processes in Pham and Tran (1985), Theorem 2.1 and obtain

$$\begin{aligned}\beta_t(\underline{u}_n) &\leq \sup_{\substack{A\in\otimes_{i=1}^n\mathcal{F}_{-\infty}^0(\eta_i)\\ B\in\otimes_{i=1}^n\otimes_{s=1}^{np}\sigma(Z_{t,s}(\eta_i))}} |P(A\cap B)-P(A)P(B)| \\ &\leq \sum_{i=1}^{n}\sum_{s=1}^{np}\sup_{\substack{A\in\mathcal{F}_{-\infty}^0(\eta_i)\\ B\in\sigma(Z_{t,s}(\eta_i))}} |P(A\cap B)-P(A)P(B)| \leq C\rho^t\end{aligned}$$

Therefore for every n the vector process $\{\underline{\phi}_t(\underline{u}_n)\}_t$ is α-mixing with a rate independent of $\underline{u}_n$.

References

Bosq, D. (1998): *Nonparametric statistics for stochastic processes*. Springer, New York.

Cressie, N., and H. Huang (1999): "Classes of non-separable, spatio-temporal covariance functions," *Journal of the American Statistical Association*, 94, 1330–1340.

Cressie, N. A. (1993): *Statistics for Spatial Data*. John Wiley and sons, New York.

Dahlhaus, R., and S. Subba Rao (2006): "Statistical inference for time-varying ARCH processes," *Ann. Statist.*, 34.

Ekeland, I., J. Heckman, and L. Nesheim (2002): "Identification and estimation of hedonic models," Cemmap working paper CWP07/02, The Institute for Fiscal Studies, Department of Economics, UCL.

Fan, J., W. Härdle, and E. Mammen (1998): "Direct estimation of low dimension components in additive models," *Ann. Statist.*, 26, 943–971.

Fan, J., Q. Yao, and Z. Cai (2003): "Adaptive varying-coefficient linear models," *Journal of the Royal Statistical Society (B)*, 65, 57–580.

Florens, J., J. Johannes, and S. Van Bellegem (2005): "Instrumental regression in partially linear models.," *Submitted*.

Gelfand, A., H.-J. Kim, C. Sirmans, and S. Banerjee (2003): "Spatial modelling with spatially varying coefficients processes," *Journal of the American Statistical Association*, 98.

Guan, Y., M. Sherman, and J. Calvin (2004): "A nonparametric test for spatial isoptropy using subsampling," *Journal of the American Statistical Association*, 99, 810–821.

Guttorp, P., W. Meiring, and P. D. Sampson (1994): "A space-time analysis of ground level ozone data," *Environmetrics*, 5, 241–254.

Hall, P., and C. Heyde (1980): *Martingale Limit Theory and its Application*. Academic Press, New York.

Hart, J. (1994): "Some automated methods for smoothing time-dependent data," *J. Roy. Statist. Soc.*, 56, 529–542.

Horowitz, J., and E. Mammen (2004): "Nonparametric estimation of an additive model model with link function," *Ann. Statist.*, 32, 2412–2443.

Lesch, S., D. Strauss, and J. Rhoades (1995): "Spatial prediction of soil salinity using electromagnetic induction techniques 1. Statistical prediction models: A comparison of multiple regression and cokriging," *Water Resources Research*, pp. 373–386.

Lu, Z., D. Tjostheim, and Q. Yao (2005): "Adaptive varying-coefficient linear models for stochastic processes: Asymptotic theory," *Submitted*.

Luo, Z., and G. Wahba (1998): "Spatio-temporal analogues of temperature using smooth spline ANOVA," *Journal of Climatology*, 11, 18–28.

Matèrn, B. (1986): *Spatial Variations*, Lecture Notes in Statistics. Springer, New York, 2nd edn.

Mukherjee, K., and S. Lahiri (2004): "Asymptotic distribution of M-estimators in spatial regression under fixed and some stochastic spatial sampling designs," *Annals of the Institute of Statistical Mathematics*, 56, 225–250.

Pham, D. T., and T. T. Tran (1985): "Some mixing properties of time series models," *Stochastic processes and their applications*, 19, 297–303.

Robinson, P. M. (1988): "Root-N-consistent semiparametric regression," *Econometrica*, 56, 931–954.

Sampson, P., and P. Guttorp (1992): "Nonparametric estimation of nonstationary spatial covariance structure," *Journal of the American Statistical Association*, 87.

Scott, D. W. (1992): *Multivariate Density Estimation*. Wiley, New York.

Subba Rao, S. (2005): "Statistical analysis of a spatio-temporal model with location dependent parameters," *Under revision*.

Yakowitz, S., and F. Szidaravosky (1985): "A comparison of Kriging with nonparametric regression methods," *Journal of Multivariate Analysis*, 16, 21–53.