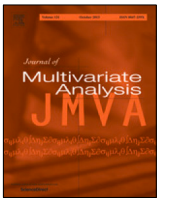
Journal of Multivariate Analysis 183 (2021) 104735

Contents lists available at ScienceDirect

Journal of Multivariate Analysis

journal homepage: www.elsevier.com/locate/jmva



Semiparametric method and theory for continuously indexed spatio-temporal processes

Jialuo Liu [a], Tingjin Chu [b,*], Jun Zhu [c], Haonan Wang [a]

[a] Department of Statistics, Colorado State University, United States of America
[b] School of Mathematics and Statistics, University of Melbourne, Australia
[c] Department of Statistics, University of Wisconsin-Madison, United States of America

ARTICLE INFO

Article history:
Received 19 February 2020
Received in revised form 4 February 2021
Accepted 5 February 2021
Available online 13 February 2021

AMS 2010 subject classifications:
primary 62H11
secondary 62M30

Keywords:
Bimodal kernel
Random fields
Spatial statistics
Spatio-temporal statistics

ABSTRACT

Spatio-temporal processes with a continuous index in space and time are useful for modeling spatio-temporal data in many scientific disciplines such as environmental and health sciences. However, approaches that enable simultaneous estimation of the mean and covariance functions for such spatio-temporal processes are limited. Here, we propose a flexible spatio-temporal model with partially linear regression in the mean function and local stationarity in the covariance function. We develop a profile likelihood method for estimation and an effective bandwidth selection in the presence of spatio-temporally correlated errors. Specifically, we employ a family of bimodal kernels to alleviate bias, which may be of independent interest for semiparametric spatial statistics. The theoretical properties of our profile likelihood estimation, including consistency and asymptotic normality, are established. A simulation study is conducted and suggests sound empirical properties, while a health hazard data example further illustrates the methodology.

1. Introduction

In this paper, we develop new semiparametric methodology and theory for spatio-temporal processes where both space and time are continuously indexed, which often arise in many scientific disciplines [see, e.g., 34]. An illustrative data set comprises measurements of a health hazard taken in an indoor environment by both static sensors at fixed sampling locations and roving sensors at varying sampling locations over time [31]. The spatio-temporal sampling design is non-standard due to data irregularity and sparsity in both space and time, calling for development of novel methodology and theory.

In spatial statistics, geostatistical data with continuous spatial index and lattice data with discrete spatial index often require different modeling techniques. For example, to account for spatial dependence, a Matérn covariance function is typically used for geostatistical data, while a spatial weight matrix is used for lattice data [8]. For spatio-temporal datasets, the distinction between continuous and discrete index applies to both spatial and temporal dimensions. To analyze datasets with continuous spatial index and discrete temporal index, time series methods for temporal data are often combined with geostatistical methods for spatial data. For example, Stroud et al. [41] developed a state space model where spatial variability is captured by a locally weighted mixture of linear regressions while the regression coefficients are allowed to vary with time. Sun et al. [42] proposed a profile likelihood based estimation procedure for a semiparametric

* Corresponding author.
E-mail address: tingjin.chu@unimelb.edu.au (T. Chu).

https://doi.org/10.1016/j.jmva.2021.104735
0047-259X/

spatial dynamic model with a nonlinear spatial trend. Al-Sulami et al. [1] considered a nonlinear spatio-temporal model to investigate the relationship between housing price index (HPI) and consumer price index (CPI) for individual states in the USA. The aforementioned spatial time series methods can capture nonlinearity and nonstationarity in space and/or time, assuming that data are observed at regular and discrete time points. In contrast, for spatio-temporal data with continuous temporal index, approaches that enable simultaneous estimation of the mean and covariance functions are limited. While the existing methods focus primarily on linear regression models [see, e.g., 11], we will develop semiparametric methods and theory for continuously indexed spatio-temporal processes.

The underlying spatio-temporal process can be decomposed into a mean trend and a spatio-temporal error process. Partially linear models offer a flexible way to model the mean trend of spatio-temporal data. For independent data, partially linear models have been extensively studied [see, e.g., 14,22,26,27,37]. For spatio-temporal data, Gao et al. [18] proposed an estimation procedure based on marginal integration for geostatistical partially linear models, and Lu et al. [29] developed spatio-temporal varying coefficient models, which can be applied to spatio-temporal partially linear models. Theoretical property is established for both works under the spatio-temporal mixing conditions. Since both works focus on estimating the mean trend of spatio-temporal data, spatio-temporal error is treated as independent in estimation. In practice, there is considerable interest in spatio-temporal covariance functions, which characterize the spatio-temporal dependence of underlying processes. Furthermore, to interpolate unsampled spatial locations and time points (spatio-temporal kriging), spatio-temporal covariance functions are a key building block. Thus, there is clearly a need for statistical methods to estimate spatio-temporal covariance functions and here, we aim to develop new methodology which allows simultaneous estimation of the mean and covariance functions.

Various types of spatio-temporal covariance functions have been developed [see, e.g., 7,9,16,19,35,39]. However, the dependence structure in spatio-temporal data poses challenges for establishing the asymptotic properties. In spatial statistics, there are three commonly used asymptotic frameworks, namely, increasing-domain asymptotics, fixed-domain asymptotics and mixed-domain asymptotics. For increasing-domain asymptotics, the spatial domain expands as the number of observations increases [see, e.g., 6,10,32]. For fixed-domain asymptotics, the spatial domain is fixed and the sampling locations get denser [see, e.g., 28,38,45,46]. A mixed-domain asymptotic framework allows both spatial domain and sampling density to increase [see, e.g., 4,21,23,30]. For spatio-temporal processes, Bandyopadhyay et al. [3] considered an increasing temporal domain and a mixed spatial domain for a Fourier analysis. Chu et al. [5] proposed a spatio-temporal expanding distance (STED) asymptotic framework in a fixed spatio-temporal domain, which extends the aforementioned asymptotic frameworks for spatial domain to spatio-temporal domain for exploring the asymptotic properties of statistical inference for spatio-temporal processes. The STED framework also paves the way for studying the local behavior of a spatio-temporal process, especially the second-order properties. Here, we will consider a locally stationary spatio-temporal covariance function, introduced by Chu et al. [5], to study the slowly-varying second-order nonstationarity under the STED asymptotic framework.

In essence, the mean trend of spatio-temporal data is modeled through partially linear models, and spatio-temporal dependence is accounted by locally stationary spatio-temporal covariance functions. The resulting model provides a flexible way to analyze continuously indexed spatio-temporal datasets. For estimation, the main challenge is to incorporate spatio-temporal covariance functions, which we overcome by profiling the spatio-temporal likelihood function. In addition, the theoretical property of proposed method is investigated under STED framework, and both consistency and asymptotic normality are established. Furthermore, a proper bandwidth selection is critical for estimation. For *iid* data, various methods have been studied, notably cross validation [see, e.g., 13], but are known to not perform well for non-*iid* data [see, e.g., 2,12,24,33]. Here, we show that cross-validation is asymptotically biased in the presence of spatio-temporal correlated errors for most commonly used kernels. We also propose a cross-validation based method with bimodal kernels to alleviate this bias in bandwidth selection.

The remainder of the paper is organized as follows. Section 2 introduces the spatio-temporal model and the profile likelihood method. The asymptotic properties of the profile likelihood estimation are established in Section 3 under suitable regularity conditions. In Section 4, we discuss the choice of kernel functions and develop a procedure for bandwidth selection. Numerical examples including a simulation study and the health hazard data example are given in Sections 5 and 6, respectively. Appendix A contains the technical details including proofs, while additional simulation results are given as Supplementary Materials.

2. Model and estimation

2.1. Spatio-temporal semiparametric model

We consider the following spatio-temporal process for the response variable $y(\cdot,\cdot)$,

$$y(\boldsymbol{s},t)=\boldsymbol{x}(\boldsymbol{s},t)^{\top}\boldsymbol{\beta}+f(t)+\varepsilon(\boldsymbol{s},t),\quad \boldsymbol{s}\in\mathcal{R},\ t\in\mathcal{T}, \tag{1}$$

where the location $\boldsymbol{s}$ is in the unit hypercube $\mathcal{R}=[0,1]^d$ for $d\geq 1$ and the rescaled time t takes values in $\mathcal{T}=[0,1]$. Here, $\boldsymbol{x}(\boldsymbol{s},t)=(x_1(\boldsymbol{s},t),\ldots,x_p(\boldsymbol{s},t))^{\top}$ is a $p\times 1$ vector of covariates at spatial location $\boldsymbol{s}$ and time point t, $\boldsymbol{\beta}=(\beta_1,\ldots,\beta_p)^{\top}$ is a $p\times 1$ vector of regression coefficients, and $f(t)$ denotes a fixed nonparametric temporal function. In the special case of $\boldsymbol{\beta}=\mathbf{0}$, (1) has a fully nonparametric mean function. Furthermore, the zero-mean spatio-temporal Gaussian random

process $\varepsilon(\boldsymbol{s},t)$ accounts for the local variations unexplained by the mean function (i.e., trend) $\boldsymbol{x}(\boldsymbol{s},t)^\top\boldsymbol{\beta}+f(t)$. In practice, $\varepsilon(\boldsymbol{s},t)=\varepsilon_1(\boldsymbol{s},t)+\varepsilon_2(\boldsymbol{s},t)$, where $\varepsilon_1(\boldsymbol{s},t)$ is a Gaussian spatio-temporal error process, and $\varepsilon_2(\boldsymbol{s},t)$ are Gaussian *iid* errors with mean 0 and equal variance, representing the nugget effect and independent of $\varepsilon_1(\boldsymbol{s},t)$. Denote $\gamma((\boldsymbol{s},t),(\boldsymbol{s}',t');\boldsymbol{\theta})$ as the covariance function of $\varepsilon(\boldsymbol{s},t)$, where $(\boldsymbol{s},t),(\boldsymbol{s}',t')\in\mathcal{R}\times\mathcal{T}$ and $\boldsymbol{\theta}$ is a $q\times 1$ vector of covariance function parameters. For $\gamma((\boldsymbol{s},t),(\boldsymbol{s}',t');\boldsymbol{\theta})$, various types of covariance functions are proposed [see, e.g., 7,9,16,19,35,39]. In this work, we focus on locally stationary covariance function, introduced by Chu et al. [5], and more details are provided in Section 5.

We consider N samples observed at $(\boldsymbol{s}_1,t_1),\ldots,(\boldsymbol{s}_N,t_N)$. Define the $N\times 1$ vector of the response variable as $\boldsymbol{y}=(y(\boldsymbol{s}_1,t_1),\ldots,y(\boldsymbol{s}_N,t_N))^\top$, the $N\times p$ design matrix as $\boldsymbol{X}=[x_j(\boldsymbol{s}_i,t_i)]_{i=1,j=1}^{N,p}$, and the $N\times 1$ vector of the errors as $\boldsymbol{\varepsilon}=(\varepsilon(\boldsymbol{s}_1,t_1),\ldots,\varepsilon(\boldsymbol{s}_N,t_N))^\top$. Let $\boldsymbol{f}=(f(t_1),\ldots,f(t_N))^\top$ denote the temporal function at the N sampling points. We have

$$\boldsymbol{y}=\boldsymbol{X}\boldsymbol{\beta}+\boldsymbol{f}+\boldsymbol{\varepsilon}. \tag{2}$$

The $N\times N$ covariance matrix of $\boldsymbol{\varepsilon}$ is expressed as $\boldsymbol{\Gamma}(\boldsymbol{\theta})=[\gamma((\boldsymbol{s}_i,t_i),(\boldsymbol{s}_j,t_j);\boldsymbol{\theta})]_{i,j=1}^{N}$. Further, let $\boldsymbol{\eta}=(\boldsymbol{\beta}^\top,\boldsymbol{\theta}^\top)^\top$ denote a $(p+q)\times 1$ vector of parameters comprising both the regression coefficients $\boldsymbol{\beta}$ and the covariance function parameters $\boldsymbol{\theta}$.

2.2. Profile likelihood estimation

Since the likelihood principle cannot be easily adopted for semiparametric models like (1), here we develop a profile likelihood method for model estimation. For a given $\boldsymbol{\beta}$, let $y_i^*=y(\boldsymbol{s}_i,t_i)-\boldsymbol{x}(\boldsymbol{s}_i,t_i)^\top\boldsymbol{\beta}$ denote a partially detrended spatio-temporal process for the response variable and let $\boldsymbol{y}^*=(y_1^*,\ldots,y_N^*)^\top$ denote an $N\times 1$ vector of partially detrended spatio-temporal response variables. We obtain an estimate of $\boldsymbol{f}$ by local polynomial regression; that is, we minimize the following criterion, with respect to $\boldsymbol{b}_t=(b_{0,t},b_{1,t})^\top$,

$$\sum_{i=1}^{N}\left\{y_i^*-b_{0,t}-b_{1,t}(t_i-t)\right\}^2K_h(t_i-t), \tag{3}$$

where $K_h=K(\cdot/h)/h$ is a kernel function $K(\cdot)$ with a bandwidth h.

With $\boldsymbol{K}_t=\mathrm{diag}\{K_h(t_1-t),\ldots,K_h(t_N-t)\}$, $\boldsymbol{D}_t=(\mathbf{1}_N,\boldsymbol{d}_{1t})$, $\boldsymbol{d}_{1t}=((t_1-t)/h,\ldots,(t_N-t)/h)^\top$, and $\mathbf{1}_N$ is an $N\times 1$ vector of 1's, it follows from (1) that, $(\widehat{b}_{0,t},h\widehat{b}_{1,t})^\top=\boldsymbol{\omega}(t)\boldsymbol{y}^*$, where $\boldsymbol{\omega}(t)=(\boldsymbol{D}_t^\top\boldsymbol{K}_t\boldsymbol{D}_t)^{-1}\boldsymbol{D}_t^\top\boldsymbol{K}_t$. The resulting estimate of $\boldsymbol{f}$ is $\widetilde{\boldsymbol{f}}=\boldsymbol{S}\boldsymbol{y}^*=\boldsymbol{S}(\boldsymbol{y}-\boldsymbol{X}\boldsymbol{\beta})$, where the smoother matrix is

$$\boldsymbol{S}=\left(\boldsymbol{\omega}_1(t_1)^\top,\ldots,\boldsymbol{\omega}_1(t_N)^\top\right)^\top, \tag{4}$$

and $\boldsymbol{\omega}_1(t)=(1,0)\boldsymbol{\omega}(t)$. Plugging $\widetilde{\boldsymbol{f}}$ into (2), we have the following approximation

$$(\boldsymbol{I}-\boldsymbol{S})\boldsymbol{y}\approx(\boldsymbol{I}-\boldsymbol{S})\boldsymbol{X}\boldsymbol{\beta}+\boldsymbol{\varepsilon}. \tag{5}$$

If $\boldsymbol{\varepsilon}$ is a sequence of independent and identically distributed random variables, the profile method is used to obtain estimates of $\boldsymbol{\beta}$ as

$$\bar{\boldsymbol{\beta}}=\{\boldsymbol{X}^\top(\boldsymbol{I}-\boldsymbol{S})^\top(\boldsymbol{I}-\boldsymbol{S})\boldsymbol{X}\}^{-1}\boldsymbol{X}^\top(\boldsymbol{I}-\boldsymbol{S})^\top(\boldsymbol{I}-\boldsymbol{S})\boldsymbol{y}$$

and $\bar{\boldsymbol{f}}=\boldsymbol{S}(\boldsymbol{y}-\boldsymbol{X}\bar{\boldsymbol{\beta}})$. However, the above method does not account for the spatio-temporal dependence of $\varepsilon(\boldsymbol{s},t)$, and therefore, the spatio-temporal parameter $\boldsymbol{\theta}$ cannot be estimated. In order to estimate both the mean trend and the spatio-temporal parameter $\boldsymbol{\theta}$, we propose to maximize the approximated profile log-likelihood function based on (5),

$$\ell(\boldsymbol{\beta},\boldsymbol{\theta})=-(N/2)\log(2\pi)-(1/2)\log\{\det\boldsymbol{\Gamma}(\boldsymbol{\theta})\}-(1/2)(\boldsymbol{y}-\boldsymbol{X}\boldsymbol{\beta})^\top(\boldsymbol{I}-\boldsymbol{S})^\top\boldsymbol{\Gamma}(\boldsymbol{\theta})^{-1}(\boldsymbol{I}-\boldsymbol{S})(\boldsymbol{y}-\boldsymbol{X}\boldsymbol{\beta}). \tag{6}$$

The estimate of $(\boldsymbol{\beta}^\top,\boldsymbol{\theta}^\top)^\top$ is the maximizer of (6), and is denoted as $(\widehat{\boldsymbol{\beta}}^\top,\widehat{\boldsymbol{\theta}}^\top)^\top$. Consequently, the estimate of $\boldsymbol{f}$ can be expressed as

$$\widehat{\boldsymbol{f}}=\boldsymbol{S}(\boldsymbol{y}-\boldsymbol{X}\widehat{\boldsymbol{\beta}}).$$

In addition, let $\boldsymbol{f}'=(f'(t_1),\ldots,f'(t_N))^\top$ denote an $N\times 1$ vector of the first-order derivatives of the temporal function $f(t)$ evaluated at the sampling time points $\{t_1,\ldots,t_N\}$. Minimizing (3) yields an estimate of $\boldsymbol{f}'$

$$\widehat{\boldsymbol{f}'}=\boldsymbol{L}(\boldsymbol{y}-\boldsymbol{X}\widehat{\boldsymbol{\beta}}),$$

where $\boldsymbol{L}=\left(h^{-1}\boldsymbol{\omega}_2(t_1)^\top,\ldots,h^{-1}\boldsymbol{\omega}_2(t_N)^\top\right)^\top$ and $\boldsymbol{\omega}_2(t)=(0,1)\boldsymbol{\omega}(t)$.

In general, we write the estimate of $\boldsymbol{F}(t)=\left(f(t),hf'(t)\right)^\top$ as $\widehat{\boldsymbol{F}}(t)=\boldsymbol{\omega}(t)(\boldsymbol{y}-\boldsymbol{X}\widehat{\boldsymbol{\beta}})$. In the case of spatio-temporal independence (i.e., $\boldsymbol{\Gamma}=\sigma^2\boldsymbol{I}$), the estimates of $\boldsymbol{\beta}$ and σ^2 in (6) can be expressed in closed form [see, e.g., 14]. In the case of a nonparametric mean function (i.e., $\boldsymbol{\beta}=\mathbf{0}$), (6) can still be maximized to obtain the estimates of $\boldsymbol{\theta}$ and $\boldsymbol{f}$, while the estimate of $\boldsymbol{f}'$ can be obtained by $\widehat{\boldsymbol{f}'}=\boldsymbol{L}\boldsymbol{y}$. The estimation procedure above depends on the choice of bandwidth, which will be discussed in Section 4.

3. Asymptotic properties

3.1. Asymptotic framework

The spatio-temporal framework is essential for establishing asymptotic property of parameter estimators. In spatial statistics, there are three asymptotic frameworks, namely, increasing-domain asymptotics, fixed-domain asymptotics, and mixed-domain asymptotics. For the likelihood-based method considered here, we employ a spatio-temporal expanding distance (STED) asymptotic framework in a fixed spatio-temporal domain, which provides a flexible tool for exploring the asymptotic properties of statistical inference for spatio-temporal processes [5]. Let n denote the stage of the asymptotics and let $\{A_n\}$ and $\{B_n\}$ be two sequences of increasing positive numbers. The (A_n, B_n)-rate STED asymptotic framework in a fixed spatio-temporal domain is defined as follows. For all n, there exist positive constants c_1, c_2 and c_3, independent of n, such that

(A.1) $\delta_n/\min_{1\le j\le N_n}\delta_{j,n}\le c_1$,
(A.2) $\zeta_n/\min_{1\le j\le N_n}\zeta_{j,n}\le c_2$,
(A.3) $\delta_n^d A_n^d\zeta_n B_n\ge c_3$,

where $\delta_{j,n}=\min\{\|\boldsymbol{s}_i-\boldsymbol{s}_j\|:1\le i\le N_n,\boldsymbol{s}_i\ne\boldsymbol{s}_j\}$, $\delta_n=\max_{1\le j\le N_n}\delta_{j,n}$, $\zeta_{j,n}=\min\{|t_i-t_j|:1\le i\le N_n,t_i\ne t_j\}$ and $\zeta_n=\max_{1\le j\le N_n}\zeta_{j,n}$. We assume that the error process $\varepsilon(\boldsymbol{s},t)$ is locally stationary in the sense that a covariance function $\gamma_n\left((\boldsymbol{s},t),(\boldsymbol{s}',t')\right)$ is said to be locally stationary if there exists a sequence of functions $g_n(\cdot,\cdot,\boldsymbol{s},t)$ such that,

$$\left|\gamma_n\left((\boldsymbol{s},t),(\boldsymbol{s}',t')\right)-g_n\left(\boldsymbol{s}'-\boldsymbol{s},t'-t,\boldsymbol{s},t\right)\right|=\mathcal{O}\left(\|\boldsymbol{s}'-\boldsymbol{s}\|+|t'-t|+\rho_n\right),$$

uniformly for all $(\boldsymbol{s},t),(\boldsymbol{s}',t')\in\mathcal{R}\times\mathcal{T}$, where $\{\rho_n\}$ is a sequence of positive numbers, which does not depend on the location, time or the parameter $\boldsymbol{\theta}$. Furthermore, $\rho_n\to 0$ as $n\to\infty$. In addition, there exists a function g such that, as $n\to\infty$,

$$\lim_{n\to\infty}\left|g_n\left(\boldsymbol{s}'-\boldsymbol{s},t'-t,\boldsymbol{s},t\right)-g(\boldsymbol{u}_1,u_2,\boldsymbol{s},t)\right|\to 0$$

uniformly for all $(\boldsymbol{s},t),(\boldsymbol{s}',t')\in\mathcal{R}\times\mathcal{T}$, where $\boldsymbol{u}_1=A_n(\boldsymbol{s}'-\boldsymbol{s})$ and $u_2=B_n(t'-t)$.

We use a one-dimensional (1D) toy example to illustrate the structure of the locally stationary covariance function. For locations $s\in\mathcal{R}=[0,1]$, we construct a locally stationary covariance function by taking the product of a positive function $D(s)$ and a stationary covariance function such that $\gamma(s,s')=D(s)D(s')\exp(-d/r)$, where r is the range parameter and $d=|s-s'|$ is the distance between s and s'. Fig. A of the supplementary material demonstrates four covariance functions, one stationary covariance function where $D_1(s)=1$ and three locally stationary covariance functions.

3.2. Asymptotic properties

For *iid* data, the maximum profile likelihood estimate $\widehat{\boldsymbol{\beta}}$ is consistent and asymptotically normal [14]. For the spatio-temporal semiparametric model (1), the asymptotic properties of the maximum profile likelihood estimates ${}^n\widehat{\boldsymbol{\beta}}$ and ${}^n\widehat{\boldsymbol{\theta}}$, which maximize (6), will be established as follows.

Theorem 1. *Under (C.1)–(C.13) in the Appendix A, there exists, with probability tending to one, a local maximizer* ${}^n\widehat{\boldsymbol{\eta}}=({}^n\widehat{\boldsymbol{\beta}}^\top,{}^n\widehat{\boldsymbol{\theta}}^\top)^\top$ *of* $\ell(\boldsymbol{\eta})$ *such that* $\|{}^n\widehat{\boldsymbol{\eta}}-\boldsymbol{\eta}_0\|=\mathcal{O}_p(N_n^{-1/2})$. *Moreover, the local maximizer* ${}^n\widehat{\boldsymbol{\eta}}$ *is asymptotic normal; that is, as* $n\to\infty$,

$$N_n^{1/2}({}^n\widehat{\boldsymbol{\beta}}-\boldsymbol{\beta}_0)\xrightarrow{D}\mathcal{N}(\mathbf{0},\boldsymbol{\Pi}^{-1})\quad\text{and}\quad N_n^{1/2}({}^n\widehat{\boldsymbol{\theta}}-\boldsymbol{\theta}_0)\xrightarrow{D}\mathcal{N}(\mathbf{0},\mathcal{I}_0(\boldsymbol{\theta}_0)^{-1}).$$

Theorem 1 establishes that the estimate ${}^n\widehat{\boldsymbol{\eta}}$ is root-N_n consistent. However, the asymptotic variance of ${}^n\widehat{\boldsymbol{\beta}}$ does not converge to the information matrix (13). As will be seen in Appendix A, if $\|\boldsymbol{\Gamma}^{-1}\|_\infty=\mathcal{O}(1)$, then $\boldsymbol{X}^\top\boldsymbol{\Gamma}^{-1}\boldsymbol{X}\succeq\boldsymbol{\Phi}^\top\boldsymbol{\Gamma}^{-1}\boldsymbol{\Phi}$, where $\boldsymbol{A}\succeq\boldsymbol{B}$ if $\boldsymbol{A}-\boldsymbol{B}$ is positive semi-definite. That is, the asymptotic variance of $\widehat{\boldsymbol{\beta}}$ in partially linear models is greater than those in simple linear regression models. Following a series of lemmas in Appendix A, the proof of Theorem 1 is given in Appendix A.

A by-product of the proof for Theorem 1, given in Appendix A, shows that $\mathcal{I}(\boldsymbol{\beta}_0)=\lim_{n\to\infty}N_n^{-1}\boldsymbol{X}^\top(\boldsymbol{I}-\boldsymbol{S})^\top\boldsymbol{\Gamma}^{-1}(\boldsymbol{I}-\boldsymbol{S})\boldsymbol{X}$. Thus, we use $N_n^{-1}\boldsymbol{X}^\top(\boldsymbol{I}-\boldsymbol{S})^\top\boldsymbol{\Gamma}^{-1}(\boldsymbol{I}-\boldsymbol{S})\boldsymbol{X}$ as a finite sample approximation of $\boldsymbol{\Pi}$, the asymptotic variance of ${}^n\widehat{\boldsymbol{\beta}}$. In contrast, for ${}^n\widehat{\boldsymbol{\theta}}$, it can be shown that the asymptotic variance is the same as that for the case when the temporal function $f(\cdot)$ is assumed to be known.

Further, recall that $\widehat{\boldsymbol{F}}(t)=\boldsymbol{\omega}(t)(\boldsymbol{y}-\boldsymbol{X}\widehat{\boldsymbol{\beta}})$ is the estimate of $\boldsymbol{F}(t)=\left(f(t),hf'(t)\right)^\top$, where $\boldsymbol{\omega}(t)=(\boldsymbol{D}_t^\top\boldsymbol{K}_t\boldsymbol{D}_t)^{-1}\boldsymbol{D}_t^\top\boldsymbol{K}_t$. The following Theorem 2 establishes the asymptotic normality of $\widehat{\boldsymbol{F}}(t)$. The proof of Theorem 2 is given in Appendix A.

Theorem 2. *Suppose* $f^{(3)}(t)$ *is bounded. Under (C.1)–(C.13) in the Appendix A, we have, as* $n\to\infty$,

$$(N_nh)^{1/2}\left\{\widehat{\boldsymbol{F}}(t)-\boldsymbol{F}(t)-(1/2)h^2\begin{pmatrix}\mu_2f''(t)\\0\end{pmatrix}+o(h^2)\right\}\xrightarrow{D}\mathcal{N}\left(\mathbf{0},\begin{pmatrix}1&0\\0&\mu_2^{-1}\end{pmatrix}\boldsymbol{\Delta}_t\begin{pmatrix}1&0\\0&\mu_2^{-1}\end{pmatrix}\right),$$

for $t\in(0,1)$, *where* $\mu_k=\int_{-\infty}^{\infty}x^kK(x)dx$.

Similar to Theorems 1 and 2, we can show that, when $\boldsymbol{\beta}$ is known, the asymptotic properties of ${}^{n}\widehat{\boldsymbol{\theta}}$ and $\widehat{\boldsymbol{F}}(t)$ remain the same as in Theorem 2. This may be expected, since $\widehat{\boldsymbol{\beta}}$ is root-N_n consistent. Further, $\boldsymbol{\Delta}_t$ is generally unknown in practice. As suggested by (C.9), $\boldsymbol{\Delta}_t$ can be approximated by $N_n^{-1}hq(t)^{-2}\boldsymbol{k}_t^{\top}\boldsymbol{\Gamma}\boldsymbol{k}_t$, where $\boldsymbol{k}_t = \{K_h(t_i-t)\{(t_i-t)/h\}^{j-1}\}_{i,j=1}^{N_n,2}$ is an $N_n \times 2$ matrix. Since $q(t)$ represents the density of the sampling time points, we may use a kernel density method to estimate $q(t)$. An alternative is to estimate $q(t)$ by $v_{0,t}/N_n$, where $v_{0,t} = \sum_{i=1}^{N_n} K_h(t_i - t)$. Lemma 1 in Appendix A shows that such an approximation is reasonable. In the remainder of this paper, we will refer to the former approximation as a kernel density approximation and the latter a plug-in approximation.

4. Selection of kernel and bandwidth

4.1. Theoretically optimal bandwidth

The selection of bandwidth is crucial in kernel smoothing and thus we derive a theoretically optimal bandwidth. By the results in Theorem 2, the asymptotic mean squared error (AMSE) of $\widehat{f}(t)$ is

$$\mathrm{AMSE}(t) = (1/4)h^4\mu_2^2 f''(t)^2 + (N_n h)^{-1}(1,0)\boldsymbol{\Delta}_t(1,0)^{\top},$$

and the asymptotic weighted mean integrated squared error is

$$\mathrm{AMISE}(h) = \int_0^1 \mathrm{AMSE}(t)q(t)dt = (1/4)h^4\mu_2^2\int_0^1 f''(t)^2 q(t)dt + (N_n h)^{-1}\int_0^1 (1,0)\boldsymbol{\Delta}_t(1,0)^{\top}q(t)dt.$$

Viewing the density function $q(t)$ as a weight function, we obtain an asymptotically optimal bandwidth as

$$h_{\mathrm{opt}} = N_n^{-1/5}\mu_2^{-2/5}\left\{\frac{\int_0^1 (1,0)\boldsymbol{\Delta}_t(1,0)^{\top}q(t)dt}{\int_0^1 f''(t)^2 q(t)dt}\right\}^{1/5}, \tag{7}$$

where the convergence rate is $N_n^{2/5}$ and is the nonparametric optimal rate [40].

The asymptotically optimal bandwidth h_{opt} above depends on several unknown quantities: $\boldsymbol{\Delta}_t$ in the asymptotic variance of $\widehat{\boldsymbol{F}}(t)$, the density of sampling time points $q(t)$, and the second-order derivative of the temporal function $f''(t)$; thus, it is not straightforward to estimate h_{opt}. When $\boldsymbol{\Gamma} = \sigma^2\boldsymbol{I}$ (i.e., the process assumes spatio-temporal independence), $\boldsymbol{\Delta}_t$ can be expressed as $\sigma^2 q(t)^{-1}\mathrm{diag}\{\int_{-\infty}^{\infty}K(u)^2du, \int_{-\infty}^{\infty}u^2K(u)^2du\}$. A rule of thumb for bandwidth selection in this case is available [see, e.g., 13]. The idea is to plug in the estimates of σ^2 and $f''(t)$ to obtain an approximation of h_{opt}. Specifically, after a pilot global polynomial regression of degree 4 is fitted, σ^2 is estimated by the standardized residual sum of squares, and the estimate of $f''(t)$ is obtained by differentiating the resulting global fit. However, for a spatio-temporally correlated error process, the covariance matrix $\boldsymbol{\Gamma}$ needs to be estimated, and this rule of thumb is not directly applicable. Hence, a more practical bandwidth selection procedure is needed.

4.2. Practical bandwidth selection

Under model (1), we have

$$y^*(\boldsymbol{s},t) = f(t) + \varepsilon_1(\boldsymbol{s},t) + \varepsilon_2(\boldsymbol{s},t), \quad \boldsymbol{s}\in\mathcal{R},\ t\in\mathcal{T}. \tag{8}$$

We use a leave-one-out cross-validation criterion [CV; 44]. A straightforward calculation reveals that $\mathrm{CV}(h) = N_n^{-1}\sum_{i=1}^{N_n}\left\{\frac{y_i^* - \widehat{f}(t_i)}{1-S_{ii}}\right\}^2$, where S_{ii} is the (i,i)th element of the smoother matrix $\boldsymbol{S}$. However, as will be seen in the following theorem, cross-validation is asymptotically biased in the presence of correlated errors for most commonly used kernels with $K(0) \neq 0$.

Theorem 3. *Under Assumptions (C.1)–(C.13) in the Appendix A, if there exists a sequence $C_n > 0$ such that $C_n h^{-1} \to 0$ and $1/(B_n\zeta_n)\int_{B_nC_n}^{\infty}\gamma_1(u)du \to 0$, as $n\to\infty$, then we have*

$$\mathrm{E}\{\mathrm{CV}(h)\} = N_n^{-1}\sum_{i=1}^{N_n}\mathrm{E}\{f(t_i) - \widehat{f}^{(-i)}(t_i)\}^2 + \overline{\sigma^2} - K(0)\left\{(2/N_n)\sum_{i=1}^{N_n}\sum_{\substack{j\neq i\\ |t_j-t_i|<C_n}}\frac{\mathrm{Cov}(\varepsilon_i,\varepsilon_j)}{b(t_i)-K(0)}\right\} + o\left(1/(N_n h)\right),$$

where $\widehat{f}^{(-i)}(t_i)$ is the leave-one-out estimator with the ith observation deleted for estimation, $\overline{\sigma^2} = N_n^{-1}\sum_{i=1}^{N_n}\mathrm{Var}(Y_i)$, $\mathrm{CV}(h) = N_n^{-1}\sum_{i=1}^{N_n}\left\{y_i^* - \widehat{f}^{(-i)}(t_i)\right\}^2$ and $b(t_i) = N_n q(t_i)h(\mu_{0,t_i}\mu_{2,t_i} - \mu_{1,t_i}^2)\mu_{2,t_i}^{-1}$.

The proof of Theorem 3 is given in Appendix A. Theorem 3 provides a theoretical basis for the choice of kernel functions. In practice, we propose the following procedure for the selection of bandwidth h.

(i) For a predetermined bandwidth h_0 and a kernel function K_{h_0}, obtain the estimated regression coefficients $\widetilde{\boldsymbol{\beta}}$ by the profile likelihood method.

(ii) For given a kernel function, find the bandwidth h_{opt} that minimizes the cross-validation criterion

$$\text{CV}(h) = N_n^{-1} \sum_{i=1}^{N_n} \left(\frac{\widetilde{y}_i^* - \widetilde{f}_i^*}{1 - S_{ii}} \right)^2, \tag{9}$$

where $\widetilde{y}_i^* = y_i - \boldsymbol{x}(\boldsymbol{s}_i, t_i)^\top \widetilde{\boldsymbol{\beta}}$, and $\widetilde{f}_i^*$ is the profile likelihood estimate of (8).

(iii) Use h_{opt} and the kernel function from Step (ii) to obtain the desired estimates of both the regression coefficients $\boldsymbol{\beta}$ and the covariance function parameters $\boldsymbol{\theta}$.

As to be illustrated in a simulation study, the estimate of $\boldsymbol{\beta}$ is not very sensitive to the choices of bandwidth and kernel function in Step (i). Thus, we suggest to use a pilot bandwidth to yield an underestimate of f and consequently an estimate of $\boldsymbol{\beta}$. In Steps (ii) and (iii), we use a bimodal kernel $K_2(u) = 2\pi^{-1/2} u^2 \exp(-u^2)$; see Fig. B of the supplementary material [12]. Unlike the more commonly used kernels (e.g., Gaussian or Epanechnikov kernel), the bimodal kernel satisfies $K_2(0) = 0$, which can mitigate the influence of the spatio-temporal correlation.

A popular alternative to the cross-validation criterion (9) is the generalized cross-validation [GCV;20] criterion, in which S_{ii} is replaced by $N_n^{-1} \operatorname{tr}(\boldsymbol{S})$. For dependent data, Francisco-Fernandez and Opsomer [15] proposed a bias-corrected generalized cross-validation criterion (GCV_c), replacing S_{ii} by $N_n^{-1} \operatorname{tr}(\boldsymbol{S}\boldsymbol{R}(\boldsymbol{\theta}))$; that is,

$$\text{GCV}_c(h) = \frac{\sum_{i=1}^{N_n} (\widetilde{y}_i^* - \widetilde{f}_i^*)^2}{N_n \{1 - N_n^{-1} \operatorname{tr}(\boldsymbol{S}\boldsymbol{R}(\boldsymbol{\theta}))\}^2}, \tag{10}$$

where $\boldsymbol{R}(\boldsymbol{\theta})$ is a correlation matrix. In practice, a pilot estimate of the covariance parameter vector is required; however, the choice of such an estimate is not obvious, and would impact the overall estimation performance. To ensure the performance of parameter estimation in covariance function, for each candidate bandwidth h, we compute the corresponding estimate of $\boldsymbol{\theta}$ and obtain an estimated GCV_c criterion, denoted by GCV_{ce}. As further demonstrated in the simulation study, the results based on the cross-validation and GCV_{ce} are similar, although GCV_{ce} is computationally more expensive.

5. Simulation study

5.1. Simulation set-up

We sample N_s locations uniformly from the spatial domain $[0, 1]^2$, where $N_s \in \{20, 40, 60\}$. At each sampling location, we randomly sample 4% from the grid of time points $(i - 1/2)/1000$, $i \in \{1, \ldots, 1000\}$. The selected locations and time are labeled as $\{(\boldsymbol{s}_1, t_1), \ldots, (\boldsymbol{s}_{N_n}, t_{N_n})\}$. For $N_s \in \{20, 40, 60\}$, the sample sizes are $N_n \in \{806, 1644, 2449\}$, respectively. The space–time coordinates will remain fixed across iterations once generated.

For the regression mean function and the semiparametric mean function, the vector of regression coefficients is $\boldsymbol{\beta} = (4, 3, 2, 1)^\top$. The covariates are drawn (once) from a multivariate normal distribution with zero mean, unit variance, and a cross-covariate correlation of 0.5. Each covariate is standardized to have zero sample mean and unit sample variance. Further, the nonparametric temporal function in the semiparametric mean function is $f(t) = 2\{1 - \cos(2\pi t)\}$.

We then draw a realization from the mean zero Gaussian error process $\varepsilon(\boldsymbol{s}, t)$ using three different covariance functions. The first covariance function is an exponential spatio-temporal covariance function

$$\text{Cov}\{\varepsilon(\boldsymbol{s}_i, t_i), \varepsilon(\boldsymbol{s}_j, t_j)\} = \begin{cases} \sigma^2(1 - c) \exp\{-\varrho_{1,n} \|\boldsymbol{s}_i - \boldsymbol{s}_j\|/c_s - \varrho_{2,n} |t_i - t_j|/c_t\}, & \text{if } i \neq j, \\ \sigma^2, & \text{if } i = j. \end{cases}$$

Here, σ^2 is the variance of $\varepsilon(\boldsymbol{s}, t)$, $c \in [0, 1]$ is the proportion of random noise such that $c\sigma^2$ is the nugget effect, and c_s and c_t are the positive spatial and temporal range parameters, respectively. We take $\sigma^2 = 9.0$, $c = 0.2$, $c_s = 1$ and $c_t = 1$. This covariance function is stationary and separable and we denote it as `COV-1`.

Next, we consider a generalized spatio-temporal Matérn covariance function [5]:

$$\gamma_n((\boldsymbol{s}, t), (\boldsymbol{s}', t'); \boldsymbol{\theta}) = \begin{cases} \frac{D(\boldsymbol{s},t) D(\boldsymbol{s}',t') \sigma^2 \theta_3^{d/2} 2^{1-\nu}}{(\theta_1^2 u_2^2 + 1)^\nu (\theta_1^2 u_2^2 + \theta_3)^{d/2} \Gamma(\nu)} m(\boldsymbol{u}_1, u_2)^\nu K_\nu \{m(\boldsymbol{u}_1, u_2)\}, & \text{if } \|\boldsymbol{u}_1\| > 0, \\ \frac{D(\boldsymbol{s},t) D(\boldsymbol{s}',t') \sigma^2 \theta_3^{d/2}}{(\theta_1^2 u_2^2 + 1)^\nu (\theta_1^2 u_2^2 + \theta_3)^{d/2}}, & \text{if } \|\boldsymbol{u}_1\| = 0, |u_2| > 0, \\ D(\boldsymbol{s}, t)^2 \sigma^2 + c\sigma^2, & \text{if } \|\boldsymbol{u}_1\| = 0, |u_2| = 0, \end{cases} \tag{11}$$

where $\boldsymbol{u}_1 = \rho_{1,n}(\boldsymbol{s}' - \boldsymbol{s})$ and $u_2 = \rho_{2,n}(t' - t)$. In this covariance function, $m(\boldsymbol{u}_1, u_2) = \theta_2 \|\boldsymbol{u}_1\| \{(\theta_1^2 u_2^2 + 1)/(\theta_1^2 u_2^2 + \theta_3)\}^{1/2}$ and $K_\nu(\cdot)$ is the modified Bessel function of the second kind of order ν. Here, θ_1 and θ_2 are nonnegative range parameters of time and space respectively, $\theta_3 > 0$ is a separability parameter. The point-wise variance of $\varepsilon(\boldsymbol{s}, t)$ is $D(\boldsymbol{s}, t) D(\boldsymbol{s}', t') \sigma^2 + c\sigma^2$,

where $c\sigma^2$ accounts for the nugget effect. The parameter ν in $K_\nu(\cdot)$ controls the smoothness of the covariance. If we let $\nu = 1/2$ and $\theta_3 = 1$, then (11) reduces to

$$\mathrm{Cov}\{\varepsilon(\boldsymbol{s}_i, t_i), \varepsilon(\boldsymbol{s}_j, t_j)\} = \begin{cases} D(\boldsymbol{s}_i, t_i)D(\boldsymbol{s}_j, t_j)\frac{\sigma^2}{\{a^2|\varrho_{2,n}(t_i-t_j)|^2+1\}^{3/2}} \exp\{-b\varrho_{1,n}\|\boldsymbol{s}_i - \boldsymbol{s}_j\|\}, & \text{if } i \neq j; \\ D(\boldsymbol{s}_i, t_i)D(\boldsymbol{s}_j, t_j)\sigma^2 + c\sigma^2, & \text{if } i = j. \end{cases} \tag{12}$$

Similarly, σ^2 is the variance of $\varepsilon(\boldsymbol{s}, t)$, $c \in [0, 1]$ is the proportion of random noise such that $c\sigma^2$ is the nugget effect, and a and b are the positive temporal and spatial range parameters, respectively. Here, $D(\boldsymbol{s}_i, t_i) = dt_i + 1$ varies by time, resulting in a nonstationarity covariance function. We set $\sigma^2 = 9, c = 0.2, a = 1, b = 1$ and $d = 1$. This covariance function is still separable in space and time, which we refer to as `COV-2`.

The third covariance function we considered is a slight modification of `COV-2`, with $D(\boldsymbol{s}_i, t_i) = dt_i + es_{1i} + fs_{2i} + 1$. We set $\sigma^2 = 9.0, c = 0.2, a = 1, b = 1, d = 0.5, e = 0.5$ and $f = 0.5$. This covariance function is nonstationary, nonseparable and asymmetric, referred as `COV-3`.

For each combination of the sample size and the covariance function, we generate 400 simulation replicates. We also consider a special case of the semiparametric mean function where the temporal function f is assumed to be zero. As will be demonstrated later, this case will serve as a benchmark in the comparison of the estimation for $\boldsymbol{\beta}$ and $\boldsymbol{\theta}$.

For each simulated data set, a predetermined bandwidth $h_0 = 0.05$ is used to obtain an initial estimate of $\boldsymbol{\beta}$. The estimate of the optimal bandwidth $\widehat{h}$ is then determined by minimizing the cross-validation criterion (9) over a predetermined grid of bandwidth values. Given the estimated optimal bandwidth, the profile likelihood estimates $\widehat{\boldsymbol{\beta}}$, $\widehat{\boldsymbol{\theta}}$ and $\widehat{f}(\cdot)$ are obtained. We further consider two variants of the GCV for determining the bandwidth in (9): GCV_c and GCV_{ce} as described in Section 4.

The profile likelihood method (PLE) results are compared with two alternative methods, namely, ALT_1 and ALT_2. In ALT_1, the parameter estimates and the estimate of the temporal function are obtained by the profile likelihood method ignoring the spatio-temporal dependence. In ALT_2, the regression coefficients $\boldsymbol{\beta}$ and the covariance parameters $\boldsymbol{\theta}$ are estimated by the classical maximum likelihood method assuming the temporal trend $f(\cdot)$ is known. That is, ALT_2 is essentially the maximum likelihood method under the model with the regression mean function.

To assess the performance of estimation by the different methods under the different bandwidth selection criteria, we compute the means and the standard deviations (SD) of $\widehat{\boldsymbol{\beta}}$ and $\widehat{\boldsymbol{\theta}}$ from the 400 simulated data sets. We also compute the estimated standard errors of the parameters for each simulated data set based on the information matrix in Theorem 1 and report the mean estimated standard errors (SDm). For ALT_1, we use $\boldsymbol{\Gamma}(\widehat{\boldsymbol{\theta}}) = \widehat{\sigma}^2\boldsymbol{I}$ to calculate `SDm`. In addition, for the estimated temporal function $\widehat{f}$, we calculate the average squared error (ASE) for each simulated data set, defined as

$$\mathrm{ASE} = N_{\mathrm{grid}}^{-1} \sum_{i=1}^{N_{\mathrm{grid}}} \{f(t_{i,\mathrm{grid}}) - \widehat{f}(t_{i,\mathrm{grid}})\}^2,$$

where $t_{i,\mathrm{grid}} = (i - 1/2)/N_{\mathrm{grid}}$ for $i \in \{1, \ldots, N_{\mathrm{grid}}\}$ and $N_{\mathrm{grid}} = 1000$.

Finally, we generate an additional 10% new sampling locations and, at each new sampling location, new sampling time points are generated as in the simulation set-up. At these new sampling locations and time points, new observations denoted as $y_{i,\mathrm{new}}$ are generated and let $\widetilde{y}_{i,\mathrm{new}}$ denote the predicted value at the ith new sampling location and time, where $i \in \{1, \ldots, N_{\mathrm{new}}\}$, and N_{new} is the total number of new sampling locations and time points. We use the mean squared prediction error (MSPE) to evaluate the performance of the various methods as

$$\mathrm{MSPE} = N_{\mathrm{new}}^{-1} \sum_{i=1}^{N_{\mathrm{new}}} (y_{i,\mathrm{new}} - \widetilde{y}_{i,\mathrm{new}})^2,$$

The results are provided in Tables 1–3, the last two rows of which give the average values of ASE and MSPE.

5.2. Simulation results

As shown in Table 1 for the first scenario of the spatio-temporal covariance function (`COV-1`), the bandwidths chosen by the three selection criteria, CV, GCV_c and GCV_{ce}, are similar for the profile likelihood method. For parameter estimation, both the accuracy and the precision increase as the sample size increases. The empirical standard deviations are well approximated by the standard errors, supporting the information-based asymptotic variance in Theorem 1. Further, under different bandwidth selection criteria, similar ASE and MSPE values are obtained, which may not be surprising due to the similar choices of bandwidths and hence similar estimates.

For the estimation of the regression coefficients, our method PLE and the two alternative methods ALT_1 and ALT_2 have comparable estimation bias, which suggests that the accuracy of $\widehat{\boldsymbol{\beta}}$ is not sensitive to the assumption of covariance structure. However, the simulation standard deviations from ALT_1 are larger than those from PLE and ALT_2, indicating noticeable gain of statistical efficiency in the parameter estimation by accounting for spatio-temporal dependence. In addition, ALT_1 has much larger MSPE and thus poorer prediction than PLE and ALT_2. For estimating the temporal function, the ASEs for ALT_1 and PLE are similar; both decrease as the sample size increases.

Table 1
Sample mean, sample standard deviation (SD), averaged information matrix based standard deviation (SDm) of regression and covariance parameters, averaged squared error (ASE) and mean-squared prediction error (MSPE) for three sample sizes with $N_s = 20, 40, 60$ using bimodal kernel for COV−1 and for three bandwidth selection criteria, CV, GCV_c, GCV_{ce} under profile likelihood estimation (PLE). Comparison is made with two alternatives ALT_1 and ALT_2.

Term	Truth	$N_s = 20$					$N_s = 40$					$N_s = 60$				
Method	–	PLE			ALT_1	ALT_2	PLE			ALT_1	ALT_2	PLE			ALT_1	ALT_2
Criteria	–	CV	GCV_c	GCV_{ce}	CV	–	CV	GCV_c	GCV_{ce}	CV	–	CV	GCV_c	GCV_{ce}	CV	–
h	–	0.079	0.081	0.082	0.079	–	0.072	0.073	0.073	0.072	–	0.068	0.069	0.069	0.068	–
β_1	4.0	3.985	3.985	3.985	3.991	3.985	4.002	4.002	4.002	3.993	4.002	4.002	4.002	4.002	4.003	4.002
SD		0.104	0.104	0.104	0.120	0.104	0.080	0.080	0.080	0.104	0.080	0.067	0.067	0.067	0.082	0.067
SDm		0.112	0.112	0.112	0.132	0.112	0.080	0.080	0.080	0.095	0.080	0.066	0.066	0.066	0.078	0.066
β_2	3.0	3.017	3.017	3.017	3.017	3.018	3.009	3.009	3.009	3.004	3.010	2.996	2.996	2.996	2.989	2.996
SD		0.123	0.123	0.123	0.140	0.122	0.075	0.075	0.075	0.095	0.075	0.065	0.065	0.065	0.079	0.065
SDm		0.116	0.116	0.116	0.136	0.117	0.078	0.078	0.078	0.094	0.078	0.065	0.065	0.065	0.077	0.065
β_3	2.0	2.006	2.006	2.006	2.003	2.005	1.986	1.986	1.986	1.982	1.986	1.997	1.997	1.997	2.001	1.997
SD		0.109	0.109	0.109	0.127	0.109	0.074	0.074	0.074	0.087	0.074	0.067	0.067	0.067	0.082	0.067
SDm		0.114	0.114	0.114	0.132	0.114	0.081	0.081	0.081	0.096	0.081	0.064	0.064	0.064	0.077	0.064
β_4	1.0	0.996	0.996	0.995	0.988	0.996	1.002	1.002	1.002	1.010	1.002	1.000	1.000	1.000	0.999	1.000
SD		0.110	0.110	0.110	0.131	0.110	0.077	0.077	0.077	0.097	0.077	0.065	0.065	0.065	0.082	0.065
SDm		0.115	0.115	0.115	0.134	0.115	0.079	0.079	0.079	0.094	0.079	0.065	0.065	0.065	0.077	0.065
σ^2	9.0	9.111	9.112	9.112	9.101	8.944	9.120	9.120	9.120	9.113	8.997	9.055	9.055	9.055	9.054	8.963
SD		0.569	0.569	0.569	0.589	0.525	0.387	0.387	0.387	0.394	0.372	0.325	0.324	0.324	0.334	0.317
SDm		0.546	0.547	0.547	–	0.528	0.395	0.395	0.395	–	0.384	0.319	0.319	0.319	–	0.313
c	0.2	0.209	0.209	0.209	–	0.202	0.201	0.201	0.201	–	0.196	0.197	0.197	0.197	–	0.193
SD		0.077	0.077	0.077	–	0.078	0.047	0.047	0.048	–	0.047	0.043	0.043	0.043	–	0.043
SDm		0.074	0.074	0.074	–	0.077	0.047	0.047	0.047	–	0.048	0.040	0.040	0.040	–	0.041
c_s	1.0	1.090	1.090	1.091	–	1.025	1.052	1.052	1.052	–	1.007	1.027	1.027	1.027	–	0.994
SD		0.224	0.224	0.224	–	0.198	0.134	0.134	0.134	–	0.124	0.103	0.103	0.103	–	0.098
SDm		0.213	0.213	0.213	–	0.200	0.127	0.127	0.127	–	0.121	0.099	0.099	0.099	–	0.095
c_t	1.0	1.087	1.088	1.088	–	1.029	1.047	1.048	1.048	–	1.008	1.016	1.016	1.016	–	0.987
SD		0.243	0.244	0.245	–	0.224	0.142	0.142	0.142	–	0.134	0.112	0.112	0.112	–	0.108
SDm		0.213	0.213	0.213	–	0.204	0.139	0.139	0.139	–	0.134	0.112	0.112	0.112	–	0.109
ASE	–	0.266	0.265	0.266	0.265	–	0.183	0.183	0.183	0.182	–	0.149	0.149	0.150	0.149	–
MSPE	–	6.501	6.501	6.501	9.245	6.484	7.403	7.403	7.403	9.169	7.386	6.934	6.933	6.933	9.090	6.918

When the sample size is smaller, PLE has less accuracy and precision in the estimation than ALT_2. In particular, both the standard deviations and the standard errors of the estimates from PLE are considerably larger than those of ALT_2, for all the covariance parameters except the nugget proportion c. When the sample size is larger, PLE and ALT_2 have similar estimation results. In particular, the standard deviations and the standard errors of the estimates from PLE are similar to ALT_2, supporting that the asymptotic variance of $\widehat{\boldsymbol{\theta}}$ under the semiparametric mean function is the same as the regression mean function, as shown in Theorem 1. Moreover, ALT_2 has slightly better prediction than PLE due to possible bias in the estimation of $f(\cdot)$ in PLE.

Tables 2 and 3 show results for the second and the third scenario of the spatio-temporal covariance function, COV−2 and COV−3, respectively. Similar conclusions can be drawn. Particularly, in the presence of non-separability and nonstationarity in the spatio-temporal covariance function, the finite-sample performance of the estimation for the semiparametric mean function is sound and supports the asymptotic results. The bandwidths selected by the three criteria, CV, GCV_c and GCV_{ce}, are very similar and so are the resulting estimates. Unlike COV−1 and COV−2, the prediction under COV−3 changes greatly for different sample sizes, which may be attributed to the nonstationarity in space with very different variances at different new spatial locations where the observations are predicted. Tables D–E in Section 3 of the Supplementary Material show that the regression coefficient estimates are robust against the choice of the kernel and the initial bandwidth. For a predetermined bandwidth, it can be seen that, different kernel functions in Step (i) of the bandwidth selection procedure yield very similar results. Moreover, those results are similar to the benchmark case when $\widetilde{\boldsymbol{\beta}} = \boldsymbol{\beta}$.

As demonstrated in Theorem 3, bimodal kernels can effectively alleviate the influence of correlated errors on bandwidth selection. To see this, we compare the results from a bimodal kernel with those from a Gaussian kernel. The estimation results under the Gaussian kernel for the first scenario of spatio-temporal covariance function (COV−1) are given in Table F of the Supplementary Material. Unlike the bimodal kernel, the bandwidths selected by the three criteria, CV, GCV_c and GCV_{ce}, can be quite different. In particular, CV selects much smaller bandwidth than GCV_c and GCV_{ce}, supporting the fact that cross-validation does not handle correlation well for most commonly-used kernels with $K(0) \neq 0$. The regression coefficient estimates are similar for all three bandwidth selection criteria, which suggests that the estimation of $\boldsymbol{\beta}$ is not sensitive to the choice of bandwidth. However, the estimates of covariance parameters are greatly affected by the bias in the bandwidth selection, with CV having the largest bias in parameter estimation and the largest ASE in the estimation of the temporal function, among the three criteria. As an alternative of CV in the presence of spatio-temporal correlation, GCV_c produces a much larger bandwidth, although the resulting estimates are not as accurate as those from the bimodal kernel function.

Table 2
Sample mean, sample standard deviation (SD), averaged information matrix based standard deviation (SDm) of regression and covariance parameters, averaged squared error (ASE), and mean-squared prediction error (MSPE) for three sample sizes with $N_s = 20, 40, 60$ using bimodal kernel for COV−2 and for three bandwidth selection criteria, CV, GCV_c, GCV_{ce} under profile likelihood estimation (PLE). Comparison is made with two alternatives ALT_1 and ALT_2.

Term	Truth	$N_s = 20$					$N_s = 40$					$N_s = 60$				
Method	–	PLE			ALT_1	ALT_2	PLE			ALT_1	ALT_2	PLE			ALT_1	ALT_2
Criteria	–	CV	GCV_c	GCV_{ce}	CV	–	CV	GCV_c	GCV_{ce}	CV	–	CV	GCV_c	GCV_{ce}	CV	–
h	–	0.100	0.103	0.104	0.100	–	0.089	0.091	0.091	0.089	–	0.083	0.086	0.086	0.083	–
β_1	4.0	3.969	3.969	3.969	3.987	3.969	4.002	4.002	4.002	3.987	4.003	4.004	4.004	4.004	4.005	4.004
SD		0.136	0.137	0.137	0.197	0.137	0.102	0.102	0.102	0.169	0.102	0.089	0.089	0.089	0.129	0.089
SDm		0.146	0.146	0.146	0.209	0.146	0.104	0.104	0.104	0.151	0.104	0.086	0.086	0.086	0.123	0.086
β_2	3.0	3.020	3.020	3.020	3.027	3.023	3.010	3.010	3.010	3.008	3.012	2.994	2.994	2.994	2.983	2.994
SD		0.165	0.165	0.165	0.218	0.164	0.097	0.097	0.097	0.155	0.097	0.083	0.083	0.083	0.128	0.082
SDm		0.153	0.153	0.153	0.215	0.153	0.101	0.101	0.101	0.150	0.101	0.082	0.082	0.082	0.122	0.082
β_3	2.0	2.011	2.011	2.011	1.998	2.008	1.984	1.984	1.984	1.970	1.984	1.993	1.993	1.993	2.000	1.993
SD		0.146	0.146	0.146	0.201	0.146	0.100	0.100	0.100	0.136	0.100	0.088	0.088	0.088	0.130	0.088
SDm		0.152	0.152	0.152	0.209	0.152	0.106	0.106	0.106	0.154	0.106	0.082	0.082	0.082	0.121	0.082
β_4	1.0	1.003	1.003	1.003	0.983	1.004	1.002	1.002	1.002	1.021	1.002	1.002	1.002	1.002	1.000	1.002
SD		0.144	0.144	0.144	0.208	0.144	0.101	0.101	0.101	0.161	0.102	0.085	0.085	0.085	0.130	0.085
SDm		0.148	0.149	0.149	0.213	0.148	0.102	0.102	0.102	0.150	0.102	0.084	0.084	0.084	0.122	0.084
σ^2	9.0	9.075	9.073	9.071	22.751	8.934	9.198	9.199	9.199	23.316	9.094	9.163	9.159	9.159	22.736	9.085
SD		1.610	1.609	1.608	1.658	1.543	1.131	1.133	1.133	1.217	1.114	0.871	0.869	0.869	0.995	0.864
SDm		1.569	1.569	1.569	–	1.544	1.114	1.114	1.114	–	1.100	0.887	0.887	0.887	–	0.879
c	0.2	0.229	0.229	0.230	–	0.227	0.200	0.200	0.200	–	0.199	0.195	0.195	0.195	–	0.194
SD		0.121	0.121	0.121	–	0.120	0.060	0.060	0.060	–	0.061	0.051	0.051	0.051	–	0.051
SDm		0.102	0.102	0.102	–	0.103	0.061	0.061	0.061	–	0.062	0.049	0.049	0.049	–	0.050
a	1.0	0.980	0.980	0.980	–	0.996	0.991	0.991	0.991	–	1.002	1.002	1.002	1.002	–	1.010
SD		0.117	0.117	0.117	–	0.118	0.070	0.070	0.070	–	0.071	0.058	0.058	0.058	–	0.058
SDm		0.101	0.101	0.101	–	0.104	0.069	0.069	0.069	–	0.071	0.058	0.058	0.058	–	0.059
b	1.0	0.973	0.973	0.973	–	1.001	0.984	0.984	0.984	–	1.006	0.992	0.992	0.992	–	1.009
SD		0.139	0.139	0.139	–	0.140	0.089	0.089	0.089	–	0.089	0.070	0.070	0.070	–	0.070
SDm		0.131	0.131	0.131	–	0.135	0.086	0.086	0.086	–	0.087	0.069	0.069	0.069	–	0.070
d	1.0	1.020	1.020	1.020	–	1.021	0.999	0.999	0.999	–	1.000	0.991	0.992	0.992	–	0.992
SD		0.246	0.247	0.247	–	0.242	0.167	0.167	0.167	–	0.167	0.127	0.127	0.127	–	0.127
SDm		0.238	0.238	0.238	–	0.237	0.162	0.162	0.162	–	0.161	0.129	0.129	0.129	–	0.129
ASE	–	0.663	0.659	0.657	0.655	–	0.451	0.451	0.451	0.450	–	0.365	0.366	0.367	0.366	–
MSPE	–	13.454	13.456	13.456	24.114	13.429	15.646	15.645	15.646	21.682	15.609	15.629	15.629	15.629	24.123	15.598

Finally, in Tables 1–3, the standard deviations and the standard errors of $\widehat{\boldsymbol{\beta}}$ from PLE are similar to those from ALT_2. This is as expected, since the design matrix in our setting does not vary by time. In Section 1.2 of the Supplementary Material, we investigate a design matrix that varies over time. From Table G in the Supplementary Material, it can be seen that the standard deviations of $\widehat{\boldsymbol{\beta}}$ from PLE are larger than those from ALT_2, which indicates a loss of statistical efficiency in the estimation of $\boldsymbol{\beta}$ when the unknown temporal function is estimated. This finding is consistent with the standard error formula in Theorem 1. In addition, we consider a nonseparable but stationary covariance function. The simulation results are provided in Section 1.3 of the Supplementary Material, and similar lessons can be learned.

6. Data example

To illustrate our methodology, we consider a data set collected by static sensors at fixed sampling locations in time and roving sensors traversing the spatial domain in time in an engine facility for evaluating the intensity level of noise as an occupational hazard [25,31]. We focus on the observations between 10:29:00 am and 11:24:00 am when all the sensors are operating. As shown in Fig. 1, there are 56 observations, one per minute, for each of the 17 static sensors. For the two roving sensors, there are a total of 179 observations, observed at irregular time points. Therefore, the total sample size is $N_n = 1131$.

We consider the semiparametric mean function (12) with the generalized spatio-temporal Matérn error covariance function (12). More specifically, for $\boldsymbol{s} = (s_1, s_2)$, we have

$$y(\boldsymbol{s}, t) = \beta_1 s_1 + \beta_2 s_2 + f(t) + \varepsilon(\boldsymbol{s}, t),$$

where the regression is on the coordinates of the spatial location $\boldsymbol{s}$, the temporal function f is nonparametric, and the zero-mean error process $\varepsilon(\boldsymbol{s}, t)$ has the spatio-temporal covariance function (12). We fit three spatio-temporal covariance functions: $D_1(\boldsymbol{s}, t) = 1$ for stationarity, $D_2(\boldsymbol{s}, t) = 1 + dt$ and $D_3(\boldsymbol{s}, t) = 1 + dt + e(t - \kappa)_+$ for nonstationary. In the latter two nonstationary cases, for any fixed time point t_0, $\varepsilon(\boldsymbol{s}, t_0)$ is spatially stationary. For $D_3(\boldsymbol{s}, t)$, κ is chosen around 11:02:00 am, which is expected to capture the temporal change due to an engine shutdown.

Table 3
Sample mean, sample standard deviation (SD), averaged information matrix based standard deviation (SDm) of regression and covariance parameters, averaged squared error (ASE), and mean-squared prediction error (MSPE) for three sample sizes with $N_s = 20, 40, 60$ using bimodal kernel for COV−3 and for three bandwidth selection criteria, CV, GCV_c, GCV_{ce} under profile likelihood estimation (PLE). Comparison is made with two alternatives ALT_1 and ALT_2.

Term	Truth	$N_s = 20$					$N_s = 40$					$N_s = 60$				
Method	–	PLE			ALT_1	ALT_2	PLE			ALT_1	ALT_2	PLE			ALT_1	ALT_2
Criteria	–	CV	GCV_c	GCV_{ce}	CV	–	CV	GCV_c	GCV_{ce}	CV	–	CV	GCV_c	GCV_{ce}	CV	–
h	–	0.106	0.108	0.109	0.106	–	0.095	0.098	0.099	0.095	–	0.087	0.090	0.090	0.087	–
β_1	4.0	3.964	3.963	3.963	3.984	3.963	4.003	4.003	4.003	3.984	4.003	4.005	4.005	4.005	4.004	4.005
SD		0.154	0.154	0.154	0.220	0.154	0.114	0.114	0.114	0.193	0.114	0.098	0.098	0.098	0.148	0.098
SDm		0.165	0.165	0.165	0.238	0.165	0.116	0.116	0.116	0.173	0.116	0.096	0.096	0.096	0.139	0.096
β_2	3.0	3.025	3.025	3.025	3.029	3.027	3.014	3.014	3.014	3.011	3.015	2.992	2.992	2.992	2.983	2.992
SD		0.184	0.184	0.184	0.248	0.183	0.107	0.107	0.107	0.181	0.107	0.092	0.092	0.092	0.144	0.092
SDm		0.171	0.171	0.171	0.246	0.172	0.112	0.112	0.112	0.171	0.112	0.092	0.092	0.092	0.138	0.092
β_3	2.0	2.014	2.014	2.014	1.994	2.011	1.981	1.981	1.981	1.968	1.981	1.991	1.991	1.991	2.002	1.991
SD		0.162	0.162	0.162	0.230	0.161	0.111	0.111	0.111	0.158	0.111	0.097	0.097	0.097	0.148	0.097
SDm		0.172	0.172	0.172	0.238	0.172	0.117	0.117	0.117	0.176	0.117	0.092	0.092	0.092	0.137	0.092
β_4	1.0	1.002	1.002	1.002	0.981	1.003	1.002	1.002	1.002	1.023	1.002	1.004	1.004	1.004	0.999	1.004
SD		0.163	0.163	0.163	0.238	0.163	0.114	0.114	0.114	0.177	0.114	0.096	0.096	0.096	0.145	0.095
SDm		0.168	0.168	0.168	0.242	0.168	0.112	0.112	0.112	0.172	0.112	0.093	0.093	0.093	0.137	0.093
σ^2	9.0	9.322	9.328	9.330	29.520	9.066	9.415	9.418	9.419	30.491	9.275	9.366	9.363	9.363	29.027	9.262
SD		2.418	2.417	2.418	2.115	2.374	1.757	1.758	1.756	1.552	1.728	1.409	1.408	1.408	1.228	1.393
SDm		2.397	2.398	2.398	–	2.343	1.725	1.725	1.725	–	1.700	1.431	1.431	1.431	–	1.416
c	0.2	0.238	0.238	0.238	–	0.240	0.202	0.202	0.202	–	0.201	0.193	0.193	0.193	–	0.192
SD		0.151	0.151	0.151	–	0.155	0.077	0.077	0.077	–	0.077	0.062	0.062	0.062	–	0.062
SDm		0.134	0.134	0.134	–	0.138	0.075	0.075	0.075	–	0.076	0.061	0.061	0.061	–	0.061
a	1.0	0.982	0.981	0.981	–	0.996	0.990	0.989	0.989	–	1.000	1.002	1.002	1.002	–	1.009
SD		0.115	0.115	0.115	–	0.116	0.067	0.067	0.067	–	0.068	0.055	0.055	0.055	–	0.056
SDm		0.098	0.098	0.098	–	0.101	0.066	0.066	0.066	–	0.067	0.056	0.056	0.056	–	0.056
b	1.0	0.977	0.977	0.977	–	1.001	0.986	0.986	0.985	–	1.005	0.993	0.993	0.993	–	1.008
SD		0.135	0.135	0.135	–	0.135	0.087	0.087	0.087	–	0.087	0.067	0.067	0.067	–	0.067
SDm		0.127	0.127	0.127	–	0.130	0.083	0.083	0.083	–	0.084	0.067	0.067	0.067	–	0.068
d	0.5	0.509	0.509	0.508	–	0.513	0.498	0.498	0.498	–	0.500	0.490	0.490	0.491	–	0.491
SD		0.222	0.222	0.222	–	0.221	0.158	0.158	0.158	–	0.158	0.117	0.117	0.117	–	0.116
SDm		0.222	0.222	0.222	–	0.224	0.150	0.150	0.150	–	0.150	0.120	0.120	0.120	–	0.120
e	0.5	0.502	0.501	0.501	–	0.515	0.497	0.497	0.497	–	0.500	0.483	0.483	0.483	–	0.486
SD		0.223	0.223	0.223	–	0.228	0.140	0.141	0.140	–	0.142	0.119	0.119	0.119	–	0.120
SDm		0.217	0.217	0.217	–	0.220	0.145	0.145	0.145	–	0.146	0.118	0.118	0.118	–	0.118
f	0.5	0.514	0.514	0.514	–	0.524	0.487	0.487	0.487	–	0.491	0.493	0.493	0.493	–	0.496
SD		0.193	0.193	0.193	–	0.198	0.151	0.151	0.151	–	0.153	0.116	0.116	0.116	–	0.117
SDm		0.197	0.197	0.197	–	0.200	0.151	0.151	0.151	–	0.151	0.115	0.115	0.115	–	0.115
ASE	–	0.825	0.828	0.829	0.823	–	0.577	0.577	0.578	0.576	–	0.450	0.447	0.447	0.450	–
MSPE	–	13.107	13.103	13.103	23.673	13.062	20.457	20.455	20.455	28.599	20.412	22.368	22.369	22.369	34.789	22.336

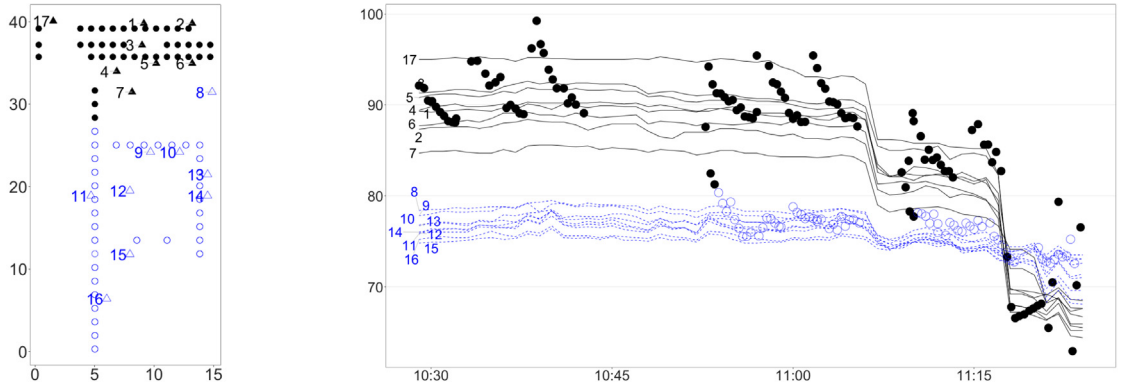


Fig. 1. Left panel: Locations of static and roving sensors (▲: static sensors in group 1, △: static sensors in group 2, •: roving sensors closer to static sensors in group 1, and ○: roving sensors closer to static sensors in group 2). Right panel: noise intensity over time at all static and roving sensors. Here, time series for static sensors in Group 1 are shown in solid line, and those from Group 2 static sensors are shown in dashed line. In addition, measurements of roving sensors recorded near Group 1 sensors are shown in dark solid circles, otherwise, they are shown as open circles.

We apply our method to analyze this data set and summarize the parameter estimates of β_1, β_2 and $\boldsymbol{\theta}$ in Table 4, whereas the estimated temporal function $\widehat{f}(t)$ and the pointwise 95% confidence intervals are plotted in Fig. 2. We

Table 4
Selected bandwidths using bimodal kernel and corresponding parameter estimates for four covariance structures: $D_1(\boldsymbol{s}, t) = 1$, $D_2(\boldsymbol{s}, t) = dt + 1$ and $D_3(\boldsymbol{s}, t) = dt + e(t - \kappa)_+ + 1$. Standard errors are computed based on information matrices from Theorem 1 and given in parentheses.

	$D_1(\boldsymbol{s}, t)$	$D_2(\boldsymbol{s}, t)$	$D_3(\boldsymbol{s}, t)$	$D_3(\boldsymbol{s}, t)$ (penalized)
h	0.0193	0.0193	0.0193	0.0193
Regression parameters				
β_1	−0.3922 (0.0820)	−0.4492 (0.0652)	−0.4872 (0.0608)	−0.4600 (0.0636)
β_2	0.3015 (0.0565)	0.4048 (0.0440)	0.4142 (0.0410)	0.3954 (0.0425)
Covariance parameters				
σ^2	50.8840 (8.7442)	8.8096 (1.7254)	19.0058 (3.6086)	14.7628 (2.7797)
c	0.0007 (0.0001)	0.0020 (0.0005)	0.0007 (0.0002)	0.0009 (0.0002)
c_s	0.1662 (0.0040)	0.1677 (0.0037)	0.1647 (0.0035)	0.1723 (0.0038)
c_t	0.0152 (0.0025)	0.0215 (0.0033)	0.0201 (0.0031)	0.0237 (0.0036)
d	–	1.9218 (0.2647)	−0.3070 (0.1298)	0.1982 (0.1723)
e	–	–	6.5719 (0.4883)	3.0394 (0.4177)

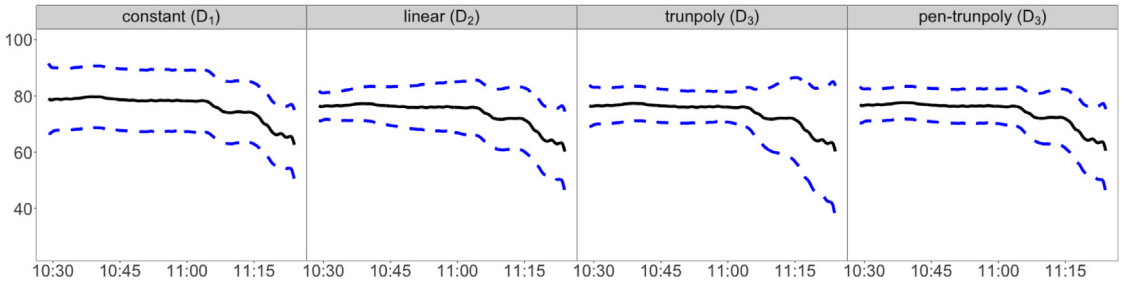


Fig. 2. Estimated temporal function $\widehat{f}(t)$ (solid curve) and 95% pointwise confidence intervals (dash curves) by maximizing the profile-likelihood (6) with four covariance structures: constant $D_1(\boldsymbol{s}, t) = 1$; linear $D_2(\boldsymbol{s}, t) = dt + 1$; truncated polynomial $D_3(\boldsymbol{s}, t) = dt + e(t - \kappa)_+ + 1$; and maximizing a penalized profile-likelihood by adding a penalty term to (6) with D_3.

approximate the pointwise standard deviation of $\widehat{f}(t)$ by Theorem 2. The temporal function estimates $\widehat{f}(t)$ under the three models D_1, D_2, and D_3 are quite similar; however, the pointwise confidence interval based on D_1 is much wider than those based on D_2. For D_3, the pointwise confidence interval is much narrower than D_1 and D_2 when t is small, however it is unusually large when t is large.

This finding is also reflected in Table 4, the estimate of the coefficient of $(t - \kappa)_+$ in D_3 (e) is unusually large. This seems like a common phenomenon in spline smoothing with truncated polynomial basis functions. To circumvent this potential issue, we consider a penalized approach [36]. That is, when maximizing the profile likelihood function (6), we consider adding an additional penalty term $-\lambda|e|$, where λ is a tuning parameter. In practice, we choose λ over a grid of λ values by minimizing the rotated residual sum of squares, for details, see Section 2.2 of the Supplementary Material. In our data analysis, $\hat{\lambda} = 20$, and the resulting parameter estimates are given in the last column of Table 4. The resulting estimate of e is much smaller, the other estimates of e are close to each other. The estimated standard deviation at each time point are plotted in Fig. C of the Supplementary Material. We notice that D_3 from the penalized approach has the smallest area under the curve. As a consequence, the 95% confidence interval of the temporal function $\widehat{f}(t)$ of D_3 from the penalized approach is the narrowest compared to D_1, D_2 and D_3, as presented in the last panel of Fig. 2.

Finally, we consider an interpolation of the noise intensity in space and time by kriging based on D_3 with penalty. Fig. D in the Supplementary Material presents a dynamic evolution of the noise intensity maps over time and suggests a possible noise source in the upper-left corner with high noise intensity. There is also a sharp decrease of the noise intensity at 11:10:00 am when the engine was turned off even though all the sensors remain active, as well as a horizontal separation around $y = 30$ before 11:10:00 am, reflecting the wall that separates the facility [31].

CRediT authorship contribution statement

Jialuo Liu: Methodology, Software, Formal analysis, Writing - original draft. **Tingjin Chu:** Methodology, Software, Validation, Writing - review & editing. **Jun Zhu:** Conceptualization, Methodology, Writing - review & editing. **Haonan Wang:** Conceptualization, Methodology, Validation, Writing - original draft, Writing - review & editing, Project administration.

Acknowledgments

The authors would like to thank the Editor, Associate Editor for their constructive comments. The research of Haonan Wang was partially supported by National Science Foundation (NSF), United States of America grants DMS-1737795,

DMS-1923142 and CNS-1932413. This work is in part supported by the U.S. Geological Survey under a Grant/Cooperative Agreement. The views and conclusions contained in this document are those of the authors and should not be interpreted as representing the opinions or policies of the U.S. Geological Survey. Mention of trade names or commercial products does not constitute their endorsement by the U.S. Geological Survey.

Appendix A. Technical details

We use $\boldsymbol{\beta}_0$ to denote the vector of true regression coefficients and $\boldsymbol{\theta}_0$ to denote the vector of true covariance parameters. We denote the log-likelihood of $(\boldsymbol{\beta}, \boldsymbol{\theta})$ in (1), when $f(t)$ is known, as

$$\ell_0(\boldsymbol{\beta}, \boldsymbol{\theta}) = -(N_n/2)\log(2\pi) - (1/2)\log\{\det\boldsymbol{\Gamma}(\boldsymbol{\theta})\} - (1/2)(\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta} - \boldsymbol{f})^\top \boldsymbol{\Gamma}(\boldsymbol{\theta})^{-1}(\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta} - \boldsymbol{f}).$$

Let $\ell_0'(\boldsymbol{\beta}) = \partial\ell_0(\boldsymbol{\beta}, \boldsymbol{\theta})/\partial\boldsymbol{\beta}$ and $\ell_0'(\boldsymbol{\theta}) = \partial\ell_0(\boldsymbol{\beta}, \boldsymbol{\theta})/\partial\boldsymbol{\theta}$ denote the first-order partial derivatives of $\ell_0(\boldsymbol{\beta}, \boldsymbol{\theta})$ with respect to $\boldsymbol{\beta}$ and $\boldsymbol{\theta}$, respectively. For ease of notation, we suppress $\boldsymbol{\theta}$ in matrices relying on $\boldsymbol{\theta}$. For example, we write $\boldsymbol{\Gamma} = \boldsymbol{\Gamma}(\boldsymbol{\theta})$. Then, we have $\ell_0'(\boldsymbol{\beta}) = \boldsymbol{X}^\top\boldsymbol{\Gamma}^{-1}(\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta} - \boldsymbol{f})$ and the kth element of $\ell_0'(\boldsymbol{\theta})$ is $-(1/2)\mathrm{tr}(\boldsymbol{\Gamma}^{-1}\boldsymbol{\Gamma}_k) - (1/2)(\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta} - \boldsymbol{f})^\top\boldsymbol{\Gamma}^k(\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta} - \boldsymbol{f})$, where $\boldsymbol{\Gamma}_k = \partial\boldsymbol{\Gamma}/\partial\theta_k$ and $\boldsymbol{\Gamma}^k = \partial\boldsymbol{\Gamma}^{-1}/\partial\theta_k = -\boldsymbol{\Gamma}^{-1}\boldsymbol{\Gamma}_k\boldsymbol{\Gamma}^{-1}$ for $k \in \{1, \ldots, q\}$.

Further, let $\ell_0''(\boldsymbol{\beta}, \boldsymbol{\beta}) = \partial^2\ell_0(\boldsymbol{\beta}, \boldsymbol{\theta})/\partial\boldsymbol{\beta}^2$, $\ell_0''(\boldsymbol{\theta}, \boldsymbol{\theta}) = \partial^2\ell_0(\boldsymbol{\beta}, \boldsymbol{\theta})/\partial\boldsymbol{\theta}^2$ and $\ell_0''(\boldsymbol{\beta}, \boldsymbol{\theta}) = \partial^2\ell_0(\boldsymbol{\beta}, \boldsymbol{\theta})/\partial\boldsymbol{\beta}\partial\boldsymbol{\theta}$ denote the second-order partial derivatives with respect to $\boldsymbol{\beta}$ and $\boldsymbol{\theta}$. Let $\mathcal{J}_n(\boldsymbol{\beta}) = E\{-\ell_0''(\boldsymbol{\beta}, \boldsymbol{\beta})\}$ and $\mathcal{J}_n(\boldsymbol{\theta}) = E\{-\ell_0''(\boldsymbol{\theta}, \boldsymbol{\theta})\}$ denote the information matrices of $\boldsymbol{\beta}$ and $\boldsymbol{\theta}$, respectively. In particular, $\ell_0''(\boldsymbol{\beta}, \boldsymbol{\beta}) = -\boldsymbol{X}^\top\boldsymbol{\Gamma}^{-1}\boldsymbol{X}$, the kth column of $\ell_0''(\boldsymbol{\beta}, \boldsymbol{\theta})$ is $\boldsymbol{X}^\top\boldsymbol{\Gamma}^k(\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta} - \boldsymbol{f})$, and the (k, k')th entry of $\ell_0''(\boldsymbol{\theta}, \boldsymbol{\theta})$ is $-(1/2)\{\mathrm{tr}(\boldsymbol{\Gamma}^{-1}\boldsymbol{\Gamma}_{kk'} + \boldsymbol{\Gamma}^k\boldsymbol{\Gamma}_{k'}) + (\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta} - \boldsymbol{f})^\top\boldsymbol{\Gamma}^{kk'}(\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta} - \boldsymbol{f})\}$, where $\boldsymbol{\Gamma}_{kk'} = \partial^2\boldsymbol{\Gamma}/\partial\theta_k\partial\theta_{k'}$ and $\boldsymbol{\Gamma}^{kk'} = \partial^2\boldsymbol{\Gamma}^{-1}/\partial\theta_k\partial\theta_{k'} = \boldsymbol{\Gamma}^{-1}(\boldsymbol{\Gamma}_k\boldsymbol{\Gamma}^{-1}\boldsymbol{\Gamma}_{k'} + \boldsymbol{\Gamma}_{k'}\boldsymbol{\Gamma}^{-1}\boldsymbol{\Gamma}_k - \boldsymbol{\Gamma}_{kk'})\boldsymbol{\Gamma}^{-1}$ for $k, k' \in \{1, \ldots, q\}$. It can be shown that $E\{\ell_0''(\boldsymbol{\beta}, \boldsymbol{\theta})\} = \boldsymbol{0}$, so the information matrix of $\boldsymbol{\eta}$ is $\mathcal{J}_n(\boldsymbol{\eta}) = \mathrm{diag}\{\mathcal{J}_n(\boldsymbol{\beta}), \mathcal{J}_n(\boldsymbol{\theta})\}$, where

$$\mathcal{J}_n(\boldsymbol{\beta}) = E\{-\ell_0''(\boldsymbol{\beta}, \boldsymbol{\beta})\} = \boldsymbol{X}^\top\boldsymbol{\Gamma}^{-1}\boldsymbol{X} \tag{13}$$

and the (k, k')th entry of $\mathcal{J}_n(\boldsymbol{\theta}) = E\{-\ell_0''(\boldsymbol{\theta}, \boldsymbol{\theta})\}$ is $t_{kk'}/2$ with $t_{kk'} = \mathrm{tr}(\boldsymbol{\Gamma}^{-1}\boldsymbol{\Gamma}_k\boldsymbol{\Gamma}^{-1}\boldsymbol{\Gamma}_{k'}) = \mathrm{tr}(\boldsymbol{\Gamma}\boldsymbol{\Gamma}^k\boldsymbol{\Gamma}\boldsymbol{\Gamma}^{k'})$.

For a matrix $\boldsymbol{A} = [a_{ii'}]_{i,i'=1}^{N_n}$, we let $\mu_i(\boldsymbol{A})$ denote its ith largest eigenvalue, let $\|\boldsymbol{A}\|_2 = \mu_1(\boldsymbol{A})$ denote its spectral norm, let $\|\boldsymbol{A}\|_F = \left(\sum_{i=1}^{N_n}\sum_{i'=1}^{N_n} a_{ii'}^2\right)^{1/2}$ denote its Frobenius norm, let $\|\boldsymbol{A}\|_{\max} = \max_{i,i'}|a_{ii'}|$ denote its max norm, and let $\|\boldsymbol{A}\|_\infty = \max_{1\le i\le N_n}\sum_{i'=1}^{N_n}|a_{ii'}|$ denote the maximum absolute column sum of the matrix. Finally, let $\xrightarrow{P}$ denote convergence in probability and $\xrightarrow{D}$ denote convergence in distribution, as $n \to \infty$.

The theoretical properties of the methods developed in Section 2 are established under the following additional regularity conditions.

(C.1) There exists a nondecreasing function $Q(t)$ with $Q(0) = 0$ and $Q(1) = 1$ such that (i) $\sup_{t\in[0,1]}|Q_{N_n}(t) - Q(t)| = \mathcal{O}(\zeta_n)$, where $Q_{N_n}(t) = N_n^{-1}\sum_{i=1}^{N_n} I(t_i \le t)$; (ii) its first-order derivative function $q(t)$ is bounded away from zero and infinity and has continuous second partial derivatives.

(C.2) For $j \in \{1, \ldots, p\}$, there exists a function $g_j(\cdot)$ on $\mathcal{T}$ with a bounded second derivative satisfying

$$x_j(\boldsymbol{s}_i, t_i) = g_j(t_i) + \phi_{ij}, \quad i \in \{1, \ldots, N_n\},$$

where $\{\phi_{ij}\}$ is a sequence of real numbers such that

$$\lim_{n\to\infty} N_n^{-1}\boldsymbol{\Phi}^\top\boldsymbol{\Gamma}^{-1}\boldsymbol{\Phi} = \boldsymbol{\Pi},$$

where $\boldsymbol{\phi}_i = (\phi_{i1}, \ldots, \phi_{iN_n})^\top$, $\boldsymbol{\Phi} = (\boldsymbol{\phi}_1, \ldots, \boldsymbol{\phi}_p)$, and $\boldsymbol{\Pi}$ is a positive definite matrix. In addition, for $j \in \{1, \ldots, p\}$, $\limsup_{n\to\infty}(1/a_n)\max_{1\le k\le N_n}\left|\sum_{m=1}^{k}\phi_{i_m j}\right| < \infty$ for all permutations $(i_1, \ldots, i_{N_n})$ of $(1, \ldots, N_n)$, where $a_n = N_n^{1/2}\log N_n$.

(C.3) The temporal function $f(t)$ is twice differentiable with a bounded second-order derivative on $\mathcal{T}$.

(C.4) The kernel $K(\cdot)$ is a symmetric, nonnegative, and bounded function with a compact support in $\mathbb{R}$ and with a bounded first-order derivative.

(C.5) The bandwidth h satisfies $h \to 0$, $N_n h^4 \to \infty$, $N_n h^8 \to 0$ and $\zeta_n h^{-1} \to 0$ as $n \to \infty$.

(C.6) For $k \in \{1, \ldots, q\}$, $\|\boldsymbol{\Gamma}_k\|_F^{-2} \le D_k N_n^{-1/2-\iota}$ for some $\iota > 0$ and $D_k > 0$.

(C.7) It holds that $\|\boldsymbol{\Gamma}^{-1}\|_2 < C^* < \infty$ for some constant C^*.

(C.8) $\lim_{n\to\infty} N_n^{-1}\mathcal{J}_n(\boldsymbol{\theta}) \longrightarrow \mathcal{I}_0(\boldsymbol{\theta})$, where $\mathcal{I}_0(\boldsymbol{\theta})$ is non-singular.

(C.9) Given $t \in (0, 1)$, there exists a 2×2 matrix $\boldsymbol{\Delta}_t$, such that $(N_n^{-1}h)\boldsymbol{k}_t^\top\boldsymbol{\Gamma}\boldsymbol{k}_t \longrightarrow q(t)^2\boldsymbol{\Delta}_t$, where $\boldsymbol{k}_t = \{K_h(t_i - t)\{(t_i - t)/h\}^{j-1}\}_{i,j=1}^{N_n,2}$ is an $N_n \times 2$ matrix.

(C.10) Define $g(\boldsymbol{s}, t) = g(\boldsymbol{0}, 0, \boldsymbol{s}, t)$. Assume $g(\boldsymbol{s}, t)$ satisfies $|g(\boldsymbol{s}, t) - g(\boldsymbol{s}', t')| \le C_1\|\boldsymbol{s} - \boldsymbol{s}'\| + C_2|t - t'|$ for all $(\boldsymbol{s}, t), (\boldsymbol{s}', t') \in \mathcal{R} \times \mathcal{T}$, where C_1, C_2 are positive constants.

(C.11) There exist two positive nonincreasing functions γ_0 and γ_1 such that $|\gamma_n((\boldsymbol{s}, t), (\boldsymbol{s} + \boldsymbol{u}_1/A_n, t + u_2/B_n))| \le \gamma_0(\|\boldsymbol{u}_1\|)\gamma_1(|u_2|)$ for all n and $\|\boldsymbol{u}_1\|, |u_2| \in [0, \infty)$ such that $(\boldsymbol{s}, t), (\boldsymbol{s} + \boldsymbol{u}_1/A_n, t + u_2/B_n) \in \mathcal{R} \times \mathcal{T}$. In addition, $\int_0^\infty u^{d-1}\gamma_0(u)du < \infty$ and $\int_0^\infty \gamma_1(u)du < \infty$.

(C.12) The covariance function $\gamma_n(\cdot,\cdot;\boldsymbol{\theta})$ is bounded and is twice continuously differentiable with respect to $\boldsymbol{\theta}$ in an open set.

(C.13) There exist two positive nonincreasing functions γ_2 and γ_3 such that

$$\max\{|\gamma_{n,k}((\boldsymbol{s},t),(\boldsymbol{s}+\boldsymbol{u}_1/A_n,t+u_2/B_n))|,|\gamma_{n,kk'}((\boldsymbol{s},t),(\boldsymbol{s}+\boldsymbol{u}_1/A_n,t+u_2/B_n))|\}\le\gamma_2(\|\boldsymbol{u}_1\|)\gamma_3(|u_2|)$$

for all n and $\|\boldsymbol{u}_1\|,|u_2|\in[0,\infty)$ with $(\boldsymbol{s},t),(\boldsymbol{s}+\boldsymbol{u}_1/A_n,t+u_2/B_n)\in\mathcal{R}\times\mathcal{T}$ and $1\le k,k'\le q$. Further, $\int_0^\infty u^{d-1}\gamma_2(u)du<\infty$ and $\int_0^\infty\gamma_3(u)du<\infty$.

In the following proofs, we suppress n in ${}^n t_{kk'}$, ${}^n a_{kk'}$, ${}^n\boldsymbol{\Gamma}$, ${}^n\boldsymbol{\Gamma}_k$, ${}^n\boldsymbol{\Gamma}_{kk'}$, $\boldsymbol{I}_n$, $\boldsymbol{A}_n$, ${}^n\widehat{\boldsymbol{\eta}}$, ${}^n\widehat{\boldsymbol{\beta}}$ and ${}^n\widehat{\boldsymbol{\theta}}$ for ease of notation.

Remarks. (C.1) is a condition on fixed time points for the spatio-temporal sampling design. (C.2) is a mild assumption about the relationship between the fixed design points and $\{t_i\}$ in the partially linear model, which is similar to Assumption 2.2 (i) in Gao and Liang [17]. (C.3)–(C.5) are common assumptions in kernel smoothing. (C.3) ensures the smoothness of the temporal function [27,43]. (C.4) is a standard assumption for kernel functions and can be relaxed further such that $K(t)$ satisfies a Lipschitz condition $\left|K(t)-K(t')\right|\le c\left|t-t'\right|$ for any $t,t'\in\mathbb{R}$ and some $c>0$. In addition, (C.5) is a condition for the rate of bandwidth with respect to N_n and ζ_n. (C.6) assures that the first-order partial derivatives of the covariance matrix have a higher order than root-N_n. (C.7) imposes a lower bound on the smallest eigenvalue of the covariance matrix. (C.8) guarantees that the growth of the information matrix is at the rate of the total sample size [6]. Moreover, (C.9) is an assumption for the fixed sampling design under the spatio-temporal dependence. Finally, (C.10)–(C.13) are regularity conditions for locally stationary processes. In (C.11), the covariance function of locally stationary processes is bounded by a product of two functions, whose integrals are finite.

A Remark on Assumption (C.2) Here, we will show that if $\|{}^n\boldsymbol{\Gamma}^{-1}\|_\infty=\mathcal{O}(1)$, we have $\boldsymbol{X}^\top\boldsymbol{\Gamma}^{-1}\boldsymbol{X}\succeq\boldsymbol{\Phi}^\top\boldsymbol{\Gamma}^{-1}\boldsymbol{\Phi}$, where $\boldsymbol{A}\succeq\boldsymbol{B}$ if $\boldsymbol{A}-\boldsymbol{B}$ is positive semi-definite. To see this, we write

$$\boldsymbol{X}^\top\boldsymbol{\Gamma}^{-1}\boldsymbol{X}=\boldsymbol{G}^\top\boldsymbol{\Gamma}^{-1}\boldsymbol{G}+\boldsymbol{G}^\top\boldsymbol{\Gamma}^{-1}\boldsymbol{\Phi}+\boldsymbol{\Phi}^\top\boldsymbol{\Gamma}^{-1}\boldsymbol{G}+\boldsymbol{\Phi}^\top\boldsymbol{\Gamma}^{-1}\boldsymbol{\Phi}.$$

Since $\|{}^n\boldsymbol{\Gamma}^{-1}\|_\infty=\mathcal{O}(1)$, $\boldsymbol{G}^\top\boldsymbol{\Gamma}^{-1}$ is uniformly bounded elementwise. Together with (C.2), $\boldsymbol{G}^\top\boldsymbol{\Gamma}^{-1}\boldsymbol{\Phi}=\mathcal{O}(N_n^{1/2}\log N_n)$. Recall that $\lim_{n\to\infty}N_n^{-1}\boldsymbol{\Phi}^\top\boldsymbol{\Gamma}^{-1}\boldsymbol{\Phi}=\boldsymbol{\Pi}$. Thus, $\boldsymbol{G}^\top\boldsymbol{\Gamma}^{-1}\boldsymbol{\Phi}$ is dominated by $\boldsymbol{\Phi}^\top\boldsymbol{\Gamma}^{-1}\boldsymbol{\Phi}$, and thus,

$$\boldsymbol{X}^\top\boldsymbol{\Gamma}^{-1}\boldsymbol{X}\succeq\boldsymbol{\Phi}^\top\boldsymbol{\Gamma}^{-1}\boldsymbol{\Phi},$$

in which the equality holds if $g(\cdot)=0$. This result indicates that the asymptotic variances of $\widehat{\boldsymbol{\beta}}$ in the partially linear model are greater than those in the simple linear regression model.

In the following Lemmas 1–6, we generalize some classical results for random sampling designs [14] to fixed sampling design, which will be used in the proofs of Theorems 1–3.

Lemma 1. *Under Assumptions (C.1), (C.4) and (C.5), for $k\ge0$,*

$$\sup_{t\in[0,1]}\left|v_{k,t}-N_n\mu_{k,t}q(t)\right|=\mathcal{O}(N_nh+N_n\zeta_nh^{-1}),$$

where $v_{k,t}=h^{-k}\sum_{i=1}^{N_n}(t_i-t)^kK_h(t_i-t)$, $K_h(t)=(1/h)K(t/h)$,

$$\mu_{k,t}=\begin{cases}\int_{-t/h}^{\infty}x^kK(x)dx, & \text{if } t<Mh,\\ \int_{-\infty}^{\infty}x^kK(x)dx:=\mu_k, & \text{if } Mh\le t\le1-Mh,\\ \int_{-\infty}^{(1-t)/h}x^kK(x)dx, & \text{if } t>1-Mh,\end{cases}$$

and $[-M,M]$ is the compact support of $K(\cdot)$.

Proof. For any $t\in[0,1]$,

$$\begin{aligned}\left|v_{k,t}-N_n\mu_{k,t}q(t)\right|\le&\left|N_nh^{-k}\int_0^1(z-t)^kK_h(z-t)\,d(Q_{N_n}-Q)(z)\right|\\&+\left|N_nh^{-k}\int_0^1(z-t)^kK_h(z-t)\,dQ(z)-N_n\mu_{k,t}q(t)\right|\equiv(I_{1,1})+(I_{1,2}),\end{aligned}$$

where

$$\begin{aligned}(I_{1,1})&=N_nh^{-k}\left|\int_0^1(z-t)^kK_h(z-t)\,d(Q_{N_n}-Q)(z)\right|\\&=N_nh^{-k}\left|(z-t)^kK_h(z-t)(Q_{N_n}-Q)(z)|_0^1-\int_0^1(Q_{N_n}-Q)(z)\left[(z-t)^kK_h(z-t)\right]'dz\right.\end{aligned}$$

$$= N_n h^{-2} \left| \int_0^1 (Q_{N_n} - Q)(z) k \left(\frac{z-t}{h}\right)^{k-1} K\left(\frac{z-t}{h}\right) dz + \int_0^1 (Q_{N_n} - Q)(z) \left(\frac{z-t}{h}\right)^{k} K'\left(\frac{z-t}{h}\right) dz \right|$$
$$\leq N_n h^{-1} \sup_{z\in[0,1]} |(Q_{N_n} - Q)(z)| \left(\int_{-M}^{M} \left| k u^{k-1} K(u) \right| du + \int_{-M}^{M} \left| u^k K'(u) \right| du \right).$$

The second equality uses the fact that $(Q_{N_n} - Q)(1) = (Q_{N_n} - Q)(0) = 0$. By (C.1), $\int_{-M}^{M} \left| k u^{k-1} K(u) \right| du + \int_{-M}^{M} \left| u^k K'(u) \right| du = \mathcal{O}(1)$. Together with (C.4), $(I_{1,1}) = \mathcal{O}(N_n \zeta_n h^{-1})$. For $(I_{1,2})$,

$$(I_{1,2}) = \sup_{t\in[0,1]} \left| N_n h^{-k} \int_0^1 (z-t)^k K_h(z-t)\, dQ(z) - N_n q(t) \mu_{k,t} \right|$$
$$= \sup_{t\in[0,1]} \left| N_n \int_{-\frac{t}{h}}^{\frac{1-t}{h}} u^k K(u) \left(q(t) + q'(t) uh + \frac{q''(\tilde{t}) u^2 h^2}{2} \right) du - N_n q(t) \mu_{k,t} \right|$$
$$\leq N_n h \left\{ \sup_{t\in[0,1]} |q'(t)| \int |u^{k+1} K(u)| du + (h/2) \sup_{t\in[0,1]} |q''(t)| \int |u^{k+2} K(u)| du \right\} = \mathcal{O}(N_n h),$$

where $\tilde{t} \in [t, t+uh]$. Thus, Lemma 1 holds. □

Lemma 2. *Under Assumptions (C.1) and (C.3)–(C.5),* $\sup_{t\in[0,1]} |\boldsymbol{\omega}_1(t)\boldsymbol{f} - f(t)| = \mathcal{O}(h^2)$.

Proof. First, straightforward calculation yields $\boldsymbol{D}_t^\top \boldsymbol{K}_t \boldsymbol{D}_t = \begin{pmatrix} v_{0,t} & v_{1,t} \\ v_{1,t} & v_{2,t} \end{pmatrix}$. By Lemma 1, uniformly on $[0, 1]$, we have $v_{0,t} = N_n q(t) \mu_{0,t} + \mathcal{O}(N_n h + N_n \zeta_n h^{-1})$, $v_{1,t} = N_n q(t) \mu_{1,t} + \mathcal{O}(N_n h + N_n \zeta_n h^{-1})$ and $v_{2,t} = N_n q(t) \mu_{2,t} + \mathcal{O}(N_n h + N_n \zeta_n h^{-1})$. In addition, notice that

$$(1,0)(\boldsymbol{D}_t^\top \boldsymbol{K}_t \boldsymbol{D}_t)^{-1} = \left(\frac{v_{2,t}}{v_{0,t} v_{2,t} - v_{1,t}^2}, \frac{-v_{1,t}}{v_{0,t} v_{2,t} - v_{1,t}^2} \right).$$

Thus,

$$\frac{v_{2,t}}{v_{0,t} v_{2,t} - v_{1,t}^2} = N_n^{-1} (q(t))^{-1} \frac{\mu_{2,t}}{\mu_{0,t} \mu_{2,t} - \mu_{1,t}^2} + \mathcal{O}(N_n^{-1} h + N_n^{-1} \zeta_n h^{-1}),$$
$$\frac{-v_{1,t}}{v_{0,t} v_{2,t} - v_{1,t}^2} = N_n^{-1} (q(t))^{-1} \frac{\mu_{1,t}}{\mu_{0,t} \mu_{2,t} - \mu_{1,t}^2} + \mathcal{O}(N_n^{-1} h + N_n^{-1} \zeta_n h^{-1})$$

uniformly on [0, 1].

Recall that $\boldsymbol{\omega}_1(t)\boldsymbol{f} - f(t) = (1,0)(\boldsymbol{D}_t^\top \boldsymbol{K}_t \boldsymbol{D}_t)^{-1} \boldsymbol{D}_t^\top \boldsymbol{K}_t \boldsymbol{f} - f(t)$. A Taylor's expansion yields $f(t_i) = f(t) + f'(t)(t_i - t) + 1/2 f''(\xi_i)(t_i - t)^2$, where ξ_i is between t and t_i. Thus,

$$\boldsymbol{\omega}_1(t)\boldsymbol{f} - f(t) = (1,0)(\boldsymbol{D}_t^\top \boldsymbol{K}_t \boldsymbol{D}_t)^{-1} \boldsymbol{D}_t^\top \boldsymbol{K}_t \boldsymbol{D}_t (f(t), hf'(t))^\top + (1/2)(1,0)(\boldsymbol{D}_t^\top \boldsymbol{K}_t \boldsymbol{D}_t)^{-1} \boldsymbol{D}_t^\top \boldsymbol{K}_t \boldsymbol{d}_\xi - f(t)$$
$$= (1/2)(1,0)(\boldsymbol{D}_t^\top \boldsymbol{K}_t \boldsymbol{D}_t)^{-1} \boldsymbol{D}_t^\top \boldsymbol{K}_t \boldsymbol{d}_\xi,$$

where $\boldsymbol{d}_\xi = (f''(\xi_1)(t_1 - t)^2, \ldots, f''(\xi_{N_n})(t_{N_n} - t)^2)^\top$. In addition,

$$\sup_{t\in[0,1]} \left| (1,0)(\boldsymbol{D}_t^\top \boldsymbol{K}_t \boldsymbol{D}_t)^{-1} \boldsymbol{D}_t^\top \boldsymbol{K}_t \boldsymbol{d}_\xi \right| \leq \max_{x\in[0,1]} |f''(x)| \sup_{t\in[0,1]} \left(\left| \frac{v_{2,t}^2}{v_{0,t} v_{2,t} - v_{1,t}^2} \right| + \left| \frac{v_{1,t} v_{3,t}}{v_{0,t} v_{2,t} - v_{1,t}^2} \right| \right) h^2 = \mathcal{O}(h^2).$$

Thus, Lemma 2 holds. □

Lemma 3. *Suppose that Assumptions (C.1) and (C.3)–(C.5) hold. For any random vector* $\boldsymbol{\varepsilon}$ *of zero mean,*

$$\sup_{t\in[0,1]} |\boldsymbol{\omega}_1(t)\boldsymbol{\varepsilon}| = \mathcal{O}_p \left\{ \left(\frac{\log N_n}{N_n h} \right)^{1/2} \right\}.$$

Proof. For a random vector $\boldsymbol{\varepsilon} = (\varepsilon_1, \ldots, \varepsilon_{N_n})^\top$, we have $\boldsymbol{\omega}_1(t)\boldsymbol{\varepsilon} = (1,0)(\boldsymbol{D}_t^\top \boldsymbol{K}_t \boldsymbol{D}_t)^{-1} \boldsymbol{D}_t^\top \boldsymbol{K}_t \boldsymbol{\varepsilon} = (I_{3,1}) - (I_{3,2})$, where $(I_{3,1}) = \frac{v_{2,t}}{v_{0,t} v_{2,t} - v_{1,t}^2} \sum_{i=1}^{N_n} K_h(t_i - t) \varepsilon_i$ and $(I_{3,2}) = \frac{v_{1,t}}{v_{0,t} v_{2,t} - v_{1,t}^2} \sum_{i=1}^{N_n} K_h(t_i - t)(t_i - t) h^{-1} \varepsilon_i$.

Note that $(I_{3,1}) = \frac{v_{0,t}v_{2,t}}{v_{0,t}v_{2,t}-v_{1,t}^2}\frac{\sum_{i=1}^{N_n}K_h(t_i-t)\varepsilon_i}{\sum_{i=1}^{N_n}K_h(t_i-t)}$, by Lemma 5,

$$\sup_{t\in[0,1]}\left|\frac{\sum_{i=1}^{N_n}K_h(t_i-t)\varepsilon_i}{\sum_{i=1}^{N_n}K_h(t_i-t)}\right| = \mathcal{O}_p\left\{\left(\frac{\log N_n}{N_nh}\right)^{1/2}\right\}, \quad \sup_{t\in[0,1]}\left|\frac{\sum_{i=1}^{N_n}(t_i-t)h^{-1}K_h(t_i-t)\varepsilon_i}{\sum_{i=1}^{N_n}K_h(t_i-t)}\right| = \mathcal{O}_p\left\{\left(\frac{\log N_n}{N_nh}\right)^{1/2}\right\}.$$

Following similar arguments in Lemma 2, $\sup_{t\in[0,1]}|(I_{3,1})| = \mathcal{O}_p\left\{\left(\frac{\log N_n}{N_nh}\right)^{1/2}\right\}$ and $\sup_{t\in[0,1]}|(I_{3,2})| = \mathcal{O}_p\left\{\left(\frac{\log N_n}{N_nh}\right)^{1/2}\right\}$. □

Lemma 4. *Under Assumptions (C.1) and (C.3)–(C.5),*

$$\sup_{t\in[0,1]}\left|\widetilde{f}(t)-f(t)\right| = \mathcal{O}_p\left\{h^2+\left(\frac{\log N_n}{N_nh}\right)^{1/2}\right\},$$

where $\widetilde{f}(t) = \boldsymbol{\omega}_1(t)\boldsymbol{y}^* = \boldsymbol{\omega}_1(t)(\boldsymbol{y}-\boldsymbol{X}\boldsymbol{\beta})$.

Proof. First, $\widetilde{f}(t)-f(t) = \boldsymbol{\omega}_1(t)\left\{\boldsymbol{f}+\boldsymbol{\varepsilon}-f(t)\mathbf{1}_{N_n}\right\}$ since $\boldsymbol{\omega}_1(t)\mathbf{1}_{N_n}-1=0$. Next, we have $\sup_{t\in[0,1]}\left|\widetilde{f}(t)-f(t)\right| \le \sup_{t\in[0,1]}|\boldsymbol{\omega}_1(t)(\boldsymbol{f}-f(t)\mathbf{1}_{N_n})| + \sup_{t\in[0,1]}|\boldsymbol{\omega}_1(t)\boldsymbol{\varepsilon}|$. The desired result follows from Lemmas 2 and 3. □

Lemma 5. *Suppose Assumptions (C.1), (C.4) and (C.5) hold. For any random vector* $\boldsymbol{\varepsilon}$ *of zero mean,*

$$\sup_{t\in[0,1]}\left|\frac{\sum_{i=1}^{N_n}(t_i-t)^jh^{-j}K_h(t_i-t)\varepsilon_i}{\sum_{i=1}^{N_n}K_h(t_i-t)}\right| = \mathcal{O}_p\left\{\left(\frac{\log N_n}{N_nh}\right)^{1/2}\right\}, \quad j\in\{0,1\}.$$

Proof. Let I_k be the interval centered at c_k with the length $\iota_{N_n} = \{\log N_n/(N_nh)\}^{1/2}h^{3+j}$. There exist $r_{N_n} = \lfloor\iota_{N_n}^{-1}\rfloor+1$ intervals satisfying $[0,1]\subset\bigcup_{k=1}^{r_{N_n}}I_k$. First, $\sup_{t\in[0,1]}\left|\frac{\sum_{i=1}^{N_n}(t_i-t)^jh^{-j}K_h(t_i-t)\varepsilon_i}{\sum_{i=1}^{N_n}K_h(t_i-t)}\right| \le (I_{6,1})+(I_{6,2})$, where

$$(I_{6,1}) = \max_{1\le k\le r_{N_n}}\sup_{t\in I_k}\left|\frac{\sum_{i=1}^{N_n}(t_i-t)^jK_h(t_i-t)\varepsilon_i}{h^j\sum_{i=1}^{N_n}K_h(t_i-t)} - \frac{\sum_{i=1}^{N_n}(t_i-c_k)^jK_h(t_i-c_k)\varepsilon_i}{h^j\sum_{i=1}^{N_n}K_h(t_i-c_k)}\right|,$$
$$(I_{6,2}) = \max_{1\le k\le r_{N_n}}\left|\frac{\sum_{i=1}^{N_n}(t_i-c_k)^jh^{-j}K_h(t_i-c_k)\varepsilon_i}{\sum_{i=1}^{N_n}K_h(t_i-c_k)}\right|.$$

For $(I_{6,1})$,

$$\begin{aligned}(I_{6,1}) &\le \max_{1\le k\le r_{N_n}}\sup_{t\in I_k}\left|\frac{1}{v_{0,t}}\left[\sum_{i=1}^{N_n}\bar{K}(t,t_i,c_k)(t_i-t)^jh^{-j}\varepsilon_i\right]\right| \\ &\quad+\max_{1\le k\le r_{N_n}}\sup_{t\in I_k}\left|\frac{1}{v_{0,t}}\left[\sum_{i=1}^{N_n}\{(t_i-t)^j-(t_i-c_k)^j\}h^{-j}K_h(t_i-c_k)\varepsilon_i\right]\right| \\ &\quad+\max_{1\le k\le r_{N_n}}\sup_{t\in I_k}\left|\frac{1}{v_{0,t}v_{0,c_k}}\sum_{i=1}^{N_n}\bar{K}(t,t_i,c_k)\sum_{i=1}^{N_n}(t_i-c_k)^jh^{-j}K_h(t_i-c_k)\varepsilon_i\right| \\ &= (I_{6,1A})+(I_{6,1B})+(I_{6,1C}),\end{aligned}$$

where $\bar{K}(t,t_i,c_k) = K_h(t_i-t)-K_h(t_i-c_k)$.

By Lemma 1, it can be shown that $\sup_{t\in[0,1]}\left|v_{0,t}^{-1}\right| = \mathcal{O}(N_n^{-1})$. In addition, by (C.4), for any $t\in I_k$, $|\bar{K}(t,t_i,c_k)| \le h^{-1}\max_{x\in\mathbb{R}}|K'(x)|\left|\frac{t_i-t}{h}-\frac{t_i-c_k}{h}\right| = \mathcal{O}(h^{-2}\iota_{N_n})$. Therefore,

$$\begin{aligned}(I_{6,1A}) &= \max_{1\le k\le r_{N_n}}\sup_{t\in I_k}\left|\frac{1}{v_{0,t}}\left[\sum_{i=1}^{N_n}\bar{K}(t,t_i,c_k)(t_i-t)^jh^{-j}\varepsilon_i\right]\right| \\ &\le \max_{1\le k\le r_{N_n}}\sup_{t\in I_k}|\bar{K}(t,t_i,c_k)|\max_{1\le k\le r_{N_n}}\sup_{t\in I_k}\left\{\frac{1}{v_{0,t}}\sum_{i=1}^{N_n}\left|(t_i-t)^jh^{-j}\varepsilon_i\right|\right\} = \mathcal{O}(N_n^{-1}\iota_{N_n}h^{-2-j})\sum_{i=1}^{N_n}|\varepsilon_i|.\end{aligned}$$

We further note that

$$(I_{6,1B}) = \max_{1\le k\le r_{N_n}} \sup_{t\in I_k} \left| \frac{1}{v_{0,t}} \left[\sum_{i=1}^{N_n} \{(t_i - t)^j - (t_i - c_k)^j\} h^{-j} K_h(t_i - c_k)\varepsilon_i \right] \right|$$
$$= \begin{cases} 0, & \text{if } j = 0, \\ \mathcal{O}(N_n^{-1}\iota_{N_n} h^{-1-j}) \sum_{i=1}^{N_n} |\varepsilon_i|, & \text{if } j = 1. \end{cases}$$

Moreover,

$$(I_{6,1C}) = \max_{1\le k\le r_{N_n}} \sup_{t\in I_k} \left| \frac{1}{v_{0,t} v_{0,c_k}} \sum_{i=1}^{N_n} \bar{K}(t, t_i, c_k) \sum_{i=1}^{N_n} (t_i - c_k)^j h^{-j} K_h(t_i - c_k)\varepsilon_i \right|$$
$$\le \max_{1\le k\le r_{N_n}} \sup_{t\in I_k} \left\{ \frac{1}{v_{0,t} v_{0,c_k}} \sum_{i=1}^{N_n} |\bar{K}(t, t_i, c_k)| \right\} \max_{1\le k\le r_{N_n}} \left\{ \sum_{i=1}^{N_n} |(t_i - c_k)^j h^{-j} \varepsilon_i| K_h(t_i - c_k) \right\}$$
$$= \mathcal{O}(N_n^{-2})\mathcal{O}(N_n \iota_{N_n} h^{-2})\mathcal{O}(h^{-1-j}) \sum_{i=1}^{N_n} |\varepsilon_i| = \mathcal{O}(N_n^{-1}\iota_{N_n} h^{-3-j}) \sum_{i=1}^{N_n} |\varepsilon_i|.$$

Since $\sum_{i=1}^{N_n} |\varepsilon_i| = \mathcal{O}_p(N_n)$, $(I_{6,1}) = \mathcal{O}_p\left\{ \left(\frac{\log N_n}{N_n h} \right)^{1/2} \right\}$.

For $(I_{6,2})$, let $\boldsymbol{e} = \boldsymbol{\Gamma}^{-1/2}\boldsymbol{\varepsilon}$ be a sequence of *iid* $N(0, 1)$, and we have $\sum_{i=1}^{N_n} (t_i - t)^j K_h(t_i - t)\varepsilon_i = h^j \boldsymbol{k}_{t,j+1}^\top \boldsymbol{\Gamma}^{1/2}\boldsymbol{e}$, where $\boldsymbol{k}_{t,j+1}$ is the $(j + 1)$th column of $\boldsymbol{k}_t$, $j \in \{0, 1\}$. For any $\lambda > 0$ and $t \in [0, 1]$, by Bernstein inequality,

$$P\left(\left| \sum_{i=1}^{N_n} (t_i - t)^j K_h(t_i - t)\varepsilon_i \right| > 2\lambda v_{0,t} \left(\frac{\log N_n}{N_n h} \right)^{\frac{1}{2}} \right) < \exp\left\{ \frac{-\lambda^2 v_{0,t}^2 \frac{\log N_n}{N_n h}}{h^{2j} \boldsymbol{k}_{t,j+1}^\top \boldsymbol{\Gamma} \boldsymbol{k}_{t,j+1}} \right\}.$$

In addition, we have $h^{2j}\boldsymbol{k}_{t,j+1}^\top \boldsymbol{\Gamma}\boldsymbol{k}_{t,j+1} \le \|\boldsymbol{\Gamma}\|_2 \sum_{i=1}^{N_n} (t_i - t)^{2j} K_h(t_i - t)^2 \le \|\boldsymbol{\Gamma}\|_2 \sum_{i=1}^{N_n} K_h(t_i - t)^2$. By similar arguments as in Lemma 1, we can show that, $\sup_{t\in[0,1]} \left| \sum_{i=1}^{N_n} K_h(t_i - t)^2 \right| = \mathcal{O}(N_n h^{-1})$. From Lemma 1, we also have $\inf_{t\in[0,1]} v_{0,t}^2 = \mathcal{O}(N_n^2)$, and therefore, by choosing a large enough λ,

$$\sup_{t\in[0,1]} P\left(\left| \sum_{i=1}^{N_n} (t_i - t)^j K_h(t_i - t)\varepsilon_i \right| > \lambda v_{0,t} \left(\frac{\log N_n}{N_n h} \right)^{1/2} \right) = \mathcal{O}(N_n^{-2}).$$

Since

$$P\left((I_{6,2}) > \lambda \left(\frac{\log N_n}{N_n h} \right)^{1/2} \right) \le \sum_{k=1}^{r_{N_n}} P\left(\left| \frac{\sum_{i=1}^{N_n} (t_i - c_k)^j K_h(t_i - c_k)\varepsilon_i}{v_{0,c_k}} \right| > \lambda \left(\frac{\log N_n}{N_n h} \right)^{1/2} \right)$$
$$= \mathcal{O}(r_{N_n} N_n^{-2}) = o(1),$$

we have $(I_{6,2}) = \mathcal{O}_p\left\{ \left(\frac{\log N_n}{N_n h} \right)^{1/2} \right\}$. Thus, we have the result. □

Lemma 6. *Suppose that Assumptions (C.1)–(C.5) hold, we have* $\sup_{t\in[0,1]} |\boldsymbol{\omega}(t)\boldsymbol{X}| = \mathcal{O}(1)$.

Proof. Using similar arguments as in Lemma 2, $v_{k,t}/(v_{0,t}v_{2,t} - v_{1,t}^2) = \mathcal{O}(N_n^{-1})$, for $k \in \{0, 1, 2\}$. The ith element of the first row of $\boldsymbol{\omega}(t)$ is $v_{2,t}/(v_{0,t}v_{2,t} - v_{1,t}^2)K_h(t_i - t) - v_{1,t}/(v_{0,t}v_{2,t} - v_{1,t}^2)K_h(t_i - t)(t_i - t)/h = \mathcal{O}(N_n^{-1}h^{-1})$. Similarly, the ith element of the second row of $\boldsymbol{\omega}(t)$ is $\mathcal{O}(N_n^{-1}h^{-1})$. Thus, $\boldsymbol{\omega}(t)\boldsymbol{\phi}_j = \mathcal{O}(N_n^{-1/2}h^{-1}\log N_n)$. Using similar arguments in Lemma 2, we obtain $(0, 1)\boldsymbol{\omega}(t)\boldsymbol{g}_j - hg_j'(t) = (0, 1/2)(\boldsymbol{D}_t^\top \boldsymbol{K}_t \boldsymbol{D}_t)^{-1}\boldsymbol{D}_t^\top \boldsymbol{K}_t \boldsymbol{d}_{\xi,j} = \mathcal{O}(h^2)$, where ξ_i is between t and t_i and $\boldsymbol{d}_{\xi,j} = (g''(\xi_1)(t_1 - t)^2, \ldots, g''(\xi_{N_n})(t_{N_n} - t)^2)^\top$. Thus, $\boldsymbol{\omega}(t)\boldsymbol{X}_j = \boldsymbol{\omega}(t)(\boldsymbol{\phi}_j + \boldsymbol{g}_j) = \mathcal{O}(1)$. □

Proof of Theorem 1. By Mardia and Marshall [32], the convergence property of $\ell_0'(\boldsymbol{\beta})$, $\ell_0'(\boldsymbol{\theta})$, $\ell_0''(\boldsymbol{\beta}, \boldsymbol{\beta})$, $\ell_0''(\boldsymbol{\beta}, \boldsymbol{\theta})$ and $\ell_0''(\boldsymbol{\theta}, \boldsymbol{\theta})$ can be established. By (C.6)–(C.7), together with proof of Theorem 1 in Chu et al. [5], we have

$$N_n^{-1/2}\ell_0'(\boldsymbol{\theta}) \xrightarrow{D} N(\boldsymbol{0}, \mathcal{I}_0(\boldsymbol{\theta})), N_n^{-1}\ell_0''(\boldsymbol{\theta}, \boldsymbol{\theta}) \xrightarrow{p} -\mathcal{I}_0(\boldsymbol{\theta}).$$

Under (C.1)–(C.9), we first show the following results

$$N_n^{-1/2}\ell'(\boldsymbol{\beta}) \xrightarrow{D} N(\boldsymbol{0}, \boldsymbol{\Pi}), N_n^{-1/2}\ell'(\boldsymbol{\theta}) \xrightarrow{D} N(\boldsymbol{0}, \mathcal{I}_0(\boldsymbol{\theta})), N_n^{-1}\ell''(\boldsymbol{\beta}, \boldsymbol{\beta}) \xrightarrow{p} -\boldsymbol{\Pi}, \tag{14a}$$
$$N_n^{-1}\ell''(\boldsymbol{\theta}, \boldsymbol{\theta}) \xrightarrow{p} -\mathcal{I}_0(\boldsymbol{\theta}), N_n^{-1}\ell''(\boldsymbol{\beta}, \boldsymbol{\theta}) \xrightarrow{p} \boldsymbol{0}. \tag{14b}$$

First, straightforward calculation yields

$$\begin{aligned}\ell'(\boldsymbol{\beta}) =& \boldsymbol{X}^\top(\boldsymbol{I}-\boldsymbol{S})^\top\boldsymbol{\Gamma}^{-1}(\boldsymbol{I}-\boldsymbol{S})(\boldsymbol{Y}-\boldsymbol{X}\boldsymbol{\beta})\\ =& \boldsymbol{X}^\top(\boldsymbol{I}-\boldsymbol{S})^\top\boldsymbol{\Gamma}^{-1}(\boldsymbol{I}-\boldsymbol{S})\boldsymbol{\varepsilon}+\boldsymbol{X}^\top(\boldsymbol{I}-\boldsymbol{S})^\top\boldsymbol{\Gamma}^{-1}(\boldsymbol{I}-\boldsymbol{S})\boldsymbol{f}\equiv(I_1)+(I_2).\end{aligned}$$

By Assumption (C.2),

$$\begin{aligned}(I_1) =& N_n^{-1/2}\{\boldsymbol{\phi}_j+\boldsymbol{g}_j\}^\top(\boldsymbol{I}-\boldsymbol{S})^\top\boldsymbol{\Gamma}^{-1}(\boldsymbol{I}-\boldsymbol{S})\boldsymbol{\varepsilon}\\ =& N_n^{-1/2}\boldsymbol{\phi}_j^\top\boldsymbol{\Gamma}^{-1}\boldsymbol{\varepsilon}+N_n^{-1/2}\boldsymbol{\phi}_j^\top\boldsymbol{\Gamma}^{-1}\boldsymbol{S}\boldsymbol{\varepsilon}+N_n^{-1/2}\boldsymbol{g}_j^\top(\boldsymbol{I}-\boldsymbol{S})^\top\boldsymbol{\Gamma}^{-1}\boldsymbol{S}\boldsymbol{\varepsilon}+\\ & N_n^{-1/2}\boldsymbol{\phi}_j^\top\boldsymbol{S}^\top\boldsymbol{\Gamma}^{-1}\boldsymbol{\varepsilon}+N_n^{-1/2}\boldsymbol{g}_j^\top(\boldsymbol{I}-\boldsymbol{S})^\top\boldsymbol{\Gamma}^{-1}\boldsymbol{\varepsilon}+N_n^{-1/2}\boldsymbol{\phi}_j^\top\boldsymbol{S}^\top\boldsymbol{\Gamma}^{-1}\boldsymbol{S}\boldsymbol{\varepsilon}\\ \equiv&(I_{11})+(I_{12})+(I_{13})+(I_{14})+(I_{15})+(I_{16}).\end{aligned}$$

For (I_{11}), it can be shown that $N_n^{-1/2}\boldsymbol{\phi}_j^\top\boldsymbol{\Gamma}^{-1}\boldsymbol{\varepsilon}\xrightarrow{D}N(\boldsymbol{0},N_n^{-1}\boldsymbol{\phi}_j^\top\boldsymbol{\Gamma}^{-1}\boldsymbol{\phi}_j)$. In addition, we have

$$\begin{aligned}&N_n^{-1}\,\mathrm{E}(\boldsymbol{\phi}_j^\top\boldsymbol{\Gamma}^{-1}\boldsymbol{S}\boldsymbol{\varepsilon})^2\le N_n^{-1}\|\boldsymbol{\Gamma}^{-1}\|_2^2\,\mathrm{E}\,|\boldsymbol{\phi}_j^\top\boldsymbol{S}\boldsymbol{\varepsilon}|^2=\mathcal{O}((\log N_n)^3N_n^{-1}h^{-1}),\\ &N_n^{-1}\,\mathrm{E}(\boldsymbol{g}_j^\top(\boldsymbol{I}-\boldsymbol{S})^\top\boldsymbol{\Gamma}^{-1}\boldsymbol{S}\boldsymbol{\varepsilon})^2\le N_n^{-1}\|\boldsymbol{g}_j^\top(\boldsymbol{I}-\boldsymbol{S})^\top\boldsymbol{\Gamma}^{-1}\|^2\,\mathrm{E}\,\|\boldsymbol{S}\boldsymbol{\varepsilon}\|^2=\mathcal{O}(h^3\log N_n),\\ &N_n^{-1}\,\mathrm{E}(\boldsymbol{\phi}_j^\top\boldsymbol{S}^\top\boldsymbol{\Gamma}^{-1}\boldsymbol{\varepsilon})^2\le N_n^{-1}\|\boldsymbol{\phi}_j^\top\boldsymbol{S}^\top\|^2\|\boldsymbol{\Gamma}^{-1}\|_2^2=\mathcal{O}(N_n^{-1}h^{-2}(\log N_n)^2),\\ &N_n^{-1}\,\mathrm{E}(\boldsymbol{g}_j^\top(\boldsymbol{I}-\boldsymbol{S})^\top\boldsymbol{\Gamma}^{-1}\boldsymbol{\varepsilon})^2\le N_n^{-1}\|\boldsymbol{g}_j^\top(\boldsymbol{I}-\boldsymbol{S})^\top\|^2\|\boldsymbol{\Gamma}^{-1}\|_2^2=\mathcal{O}(h^4).\end{aligned}$$

By Lemma 1 and Assumption (C.2), $\|\boldsymbol{S}\boldsymbol{\phi}_j\|=\mathcal{O}(N_n^{1/2}N_n^{-1}h^{-1}N_n^{1/2}\log N_n)=\mathcal{O}(h^{-1}\log N_n)$. Thus, for (I_{16}),

$$N_n^{-1}\,\mathrm{E}(\boldsymbol{\phi}_j^\top\boldsymbol{S}^\top\boldsymbol{\Gamma}^{-1}\boldsymbol{S}\boldsymbol{\varepsilon})^2\le N_n^{-1}\|\boldsymbol{\phi}_j^\top\boldsymbol{S}^\top\|^2\|\boldsymbol{\Gamma}^{-1}\|_2^2\,\mathrm{E}\,\|\boldsymbol{S}\boldsymbol{\varepsilon}\|^2=\mathcal{O}(N_n^{-1}h^{-3}(\log N_n)^3).$$

Similarly, for (I_2), we obtain

$$\begin{aligned}&N_n^{-1/2}\boldsymbol{g}_j^\top(\boldsymbol{I}-\boldsymbol{S})^\top\boldsymbol{\Gamma}^{-1}(\boldsymbol{I}-\boldsymbol{S})\boldsymbol{f}=\mathcal{O}(N_n^{-1/2}N_n^{1/2}h^2N_n^{1/2}h^2)=\mathcal{O}(N_n^{1/2}h^4),\\ &N_n^{-1/2}\boldsymbol{\phi}_j^\top\boldsymbol{\Gamma}^{-1}(\boldsymbol{I}-\boldsymbol{S})\boldsymbol{f}=\mathcal{O}(N_n^{-1/2}N_n^{1/2}\log N_nh^2)=\mathcal{O}(h^2\log N_n),\\ &N_n^{-1/2}\boldsymbol{\phi}_j^\top\boldsymbol{S}^\top\boldsymbol{\Gamma}^{-1}(\boldsymbol{I}-\boldsymbol{S})\boldsymbol{f}=\mathcal{O}(N_n^{-1/2}h^{-1}\log N_nN_n^{1/2}h^2)=\mathcal{O}(h\log N_n).\end{aligned}$$

Thus, $N_n^{-1/2}\ell'(\boldsymbol{\beta})\xrightarrow{D}N(\boldsymbol{0},\boldsymbol{\Pi})$. The kth column of $-\ell''(\boldsymbol{\beta},\boldsymbol{\theta})$ is $\boldsymbol{X}^\top(\boldsymbol{I}-\boldsymbol{S})^\top\boldsymbol{\Gamma}^k(\boldsymbol{I}-\boldsymbol{S})(\boldsymbol{y}-\boldsymbol{X}\boldsymbol{\beta})=\boldsymbol{X}^\top(\boldsymbol{I}-\boldsymbol{S})^\top\boldsymbol{\Gamma}^k(\boldsymbol{I}-\boldsymbol{S})(\boldsymbol{f}+\boldsymbol{\varepsilon})$. The same argument can be used to show $N_n^{-1}\ell''(\boldsymbol{\beta},\boldsymbol{\theta})\xrightarrow{p}\boldsymbol{0}$.

Similarly, we can show

$$N_n^{-1}\ell''(\boldsymbol{\beta})=-N_n^{-1}\boldsymbol{X}^\top(\boldsymbol{I}-\boldsymbol{S})^\top\boldsymbol{\Gamma}^{-1}(\boldsymbol{I}-\boldsymbol{S})\boldsymbol{X}\xrightarrow{p}-N_n^{-1}\boldsymbol{\Phi}^\top\boldsymbol{\Gamma}^{-1}\boldsymbol{\Phi}=-\boldsymbol{\Pi}.$$

Therefore, $N_n^{-1}\ell''(\boldsymbol{\beta},\boldsymbol{\beta})\xrightarrow{p}-\boldsymbol{\Pi}$.

By Lemmas 2–3,

$$\|(\boldsymbol{I}-\boldsymbol{S})\boldsymbol{f}\|^2=N_n\mathcal{O}(h^4)=\mathcal{O}(N_nh^4),\ \|\boldsymbol{S}\boldsymbol{\varepsilon}\|^2=N_n\mathcal{O}_p\left(\frac{\log N_n}{N_nh}\right)=\mathcal{O}_p\left(\frac{\log N_n}{h}\right).$$

Since the kth element of $-2\ell'(\boldsymbol{\theta})$ is $\mathrm{tr}(\boldsymbol{\Gamma}^{-1}\boldsymbol{\Gamma}_k)+(\boldsymbol{f}+\boldsymbol{\varepsilon})^\top(\boldsymbol{I}-\boldsymbol{S})^\top\boldsymbol{\Gamma}^k(\boldsymbol{I}-\boldsymbol{S})(\boldsymbol{f}+\boldsymbol{\varepsilon})$, we can further show that

$$\begin{aligned}&N_n^{-1/2}\boldsymbol{f}^\top(\boldsymbol{I}-\boldsymbol{S})^\top\boldsymbol{\Gamma}^k(\boldsymbol{I}-\boldsymbol{S})\boldsymbol{f}\le N_n^{-1/2}\|\boldsymbol{\Gamma}^k\|_2\|(\boldsymbol{I}-\boldsymbol{S})\boldsymbol{f}\|^2=\mathcal{O}(N_n^{1/2}h^4),\\ &N_n^{-1/2}\boldsymbol{f}^\top(\boldsymbol{I}-\boldsymbol{S})^\top\boldsymbol{\Gamma}^k\boldsymbol{\varepsilon}\le N_n^{-1/2}\mathcal{O}(h^2)\boldsymbol{1}_{N_n}^\top\boldsymbol{\varepsilon}\xrightarrow{p}0,\\ &N_n^{-1/2}|\boldsymbol{f}^\top(\boldsymbol{I}-\boldsymbol{S})^\top\boldsymbol{\Gamma}^k\boldsymbol{S}\boldsymbol{\varepsilon}|\le N_n^{-1/2}\mathcal{O}_p\left(N_n^{1/2}h^3\log N_n\right),\\ &\mathrm{E}\left|N_n^{-1/2}\boldsymbol{\varepsilon}^\top\boldsymbol{S}^\top\boldsymbol{\Gamma}^k\boldsymbol{S}\boldsymbol{\varepsilon}\right|\le N_n^{-1/2}\|\boldsymbol{\Gamma}^k\|_2\,\mathrm{E}\,\|\boldsymbol{S}\boldsymbol{\varepsilon}\|^2=\mathcal{O}_p\left(\frac{\log N_n}{N_n^{1/2}h}\right).\end{aligned}$$

By Lemma 3, $N_n^{-1/2}\boldsymbol{\varepsilon}^\top\boldsymbol{S}^\top\boldsymbol{\Gamma}^k\boldsymbol{\varepsilon}=N_n^{-1/2}\boldsymbol{1}_{N_n}^\top\boldsymbol{\varepsilon}o_p(1)\xrightarrow{p}0$. Therefore, $N_n^{-1/2}\boldsymbol{\varepsilon}^\top(\boldsymbol{I}-\boldsymbol{S})^\top\boldsymbol{\Gamma}^k(\boldsymbol{I}-\boldsymbol{S})\boldsymbol{\varepsilon}\xrightarrow{D}N_n^{-1/2}\boldsymbol{\varepsilon}^\top\boldsymbol{\Gamma}^k\boldsymbol{\varepsilon}$. Thus, $N_n^{-1/2}\ell'(\boldsymbol{\theta})\xrightarrow{D}N(\boldsymbol{0},\mathcal{I}_0(\boldsymbol{\theta}))$, and similar argument can be applied to show that $N_n^{-1}\ell''(\boldsymbol{\theta},\boldsymbol{\theta})\xrightarrow{p}-\mathcal{I}_0(\boldsymbol{\theta})$.

Next, we show the consistency and asymptotic normality of parameter estimates. To establish $\|\widehat{\boldsymbol{\eta}}-\boldsymbol{\eta}_0\|=\mathcal{O}_p(N_n^{-1/2})$, it suffices to show that, for a given constant $\epsilon>0$, there is a constant C such that, for a sufficiently large n,

$$P\left\{\sup_{\|\boldsymbol{u}\|=C}\ell(\boldsymbol{\eta}_0+N_n^{-1/2}\boldsymbol{u})<\ell(\boldsymbol{\eta}_0)\right\}\ge1-\epsilon,\tag{15}$$

where $\boldsymbol{u}\in\mathbb{R}^{p+q}$. By Taylor's expansion, we obtain

$$\ell(\boldsymbol{\eta}_0+N_n^{-1/2}\boldsymbol{u})-\ell(\boldsymbol{\eta}_0)=N_n^{-1/2}\ell'(\boldsymbol{\eta}_0)^\top\boldsymbol{u}-(1/2)N_n^{-1}\boldsymbol{u}^\top\ell''(\boldsymbol{\eta}_0)\boldsymbol{u}\{1+o_p(1)\}.\tag{16}$$

By (14a)–(14b), we have $N_n^{-1/2}\ell'(\boldsymbol{\eta}_0) = \mathcal{O}_p(1)$ and $N_n^{-1}\ell''(\boldsymbol{\eta}_0) = \mathcal{O}_p(1)$. Therefore, for a sufficiently large C, the second term dominates the first term in (16), and therefore, (15) holds.

To further establish the asymptotic normality of $\widehat{\boldsymbol{\eta}}$, it can be shown that $\widehat{\boldsymbol{\eta}} = (\widehat{\boldsymbol{\beta}}^\top, \widehat{\boldsymbol{\theta}}^\top)^\top$ satisfies

$$\mathbf{0} = \ell'(\boldsymbol{\eta})\big|_{\boldsymbol{\eta}=\widehat{\boldsymbol{\eta}}} = \ell'(\boldsymbol{\eta}_0) + \{\ell''(\boldsymbol{\eta}_0) + o_p(1)\}(\widehat{\boldsymbol{\eta}} - \boldsymbol{\eta}_0).$$

Together with (14a)–(14b), Theorem 1 holds. □

Proof of Theorem 2. Write $\widehat{\boldsymbol{F}}(t) - \boldsymbol{F}(t) = \boldsymbol{\omega}(t)\left\{\boldsymbol{f} + \boldsymbol{\varepsilon} - \boldsymbol{X}(\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta})\right\} - \boldsymbol{F}(t) = (II_1) + (II_2) - (II_3)$, where $(II_1) = \boldsymbol{\omega}(t)\boldsymbol{f} - \boldsymbol{F}(t)$, $(II_2) = \boldsymbol{\omega}(t)\boldsymbol{\varepsilon}$ and $(II_3) = \boldsymbol{\omega}(t)\boldsymbol{X}(\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta})$.

For (II_1), a Taylor's expansion yields

$$\begin{aligned} f(t_i) &= f(t) + f'(t)(t_i - t) + 1/2f''(t)(t_i - t)^2 + 1/6f^{(3)}(\xi_i)(t_i - t)^3 \\ &= \left(1, \frac{t_i - t}{h}\right)\boldsymbol{F}(t) + 1/2f''(t)(t_i - t)^2 + 1/6f^{(3)}(\xi_i)(t_i - t)^3, \end{aligned}$$

where ξ_i is between t and t_i. Therefore, $(II_1) = (1/2)\boldsymbol{\omega}(t)\boldsymbol{d}_2 f''(t) + \frac{1}{6}\boldsymbol{\omega}(t)\boldsymbol{d}_{3,\xi}$, where $\boldsymbol{d}_2 = ((t_1 - t)^2, \ldots, (t_{N_n} - t)^2)^\top$ and $\boldsymbol{d}_{3,\xi} = (f^{(3)}(\xi_1)(t_1 - t)^3, \ldots, f^{(3)}(\xi_{N_n})(t_{N_n} - t)^3)^\top$. Given $t \in (0, 1)$, by Lemma 1,

$$\boldsymbol{\omega}(t)\boldsymbol{d}_2 f''(t) = \begin{pmatrix} v_{2,t}^2 - v_{1,t}v_{3,t} \\ v_{0,t}v_{3,t} - v_{1,t}v_{2,t} \end{pmatrix} \frac{h^2 f''(t)}{v_{0,t}v_{2,t} - v_{1,t}^2} = h^2 \begin{pmatrix} \mu_2 f''(t) \\ 0 \end{pmatrix} + o(h^2).$$

Moreover, we have

$$\left|\boldsymbol{\omega}(t)\boldsymbol{d}_{3,\xi}\right| \le \max_{x\in\mathbb{R}} \frac{4|f^{(3)}(x)|h^3}{v_{0,t}v_{2,t} - v_{1,t}^2} \left\{\left|\begin{pmatrix} v_{2,t}v_{3,t} - v_{1,t}v_{4,t} \\ -v_{1,t}v_{3,t} + v_{0,t}v_{4,t} \end{pmatrix}\right|\right\} = \mathcal{O}(h^3).$$

Therefore, $(II_1) = h^2 \begin{pmatrix} \mu_2 f''(t) \\ 0 \end{pmatrix} + o(h^2)$.

For (II_2), let $A(\boldsymbol{\varepsilon}) = \sum_{i=1}^{N_n} K_h(t_i - t)\varepsilon_i$ and $B(\boldsymbol{\varepsilon}) = \sum_{i=1}^{N_n} K_h(t_i - t)\frac{t_i - t}{h}\varepsilon_i$, we have

$$\boldsymbol{\omega}(t)\boldsymbol{\varepsilon} = \frac{1}{v_{0,t}v_{2,t} - v_{1,t}^2} \begin{pmatrix} v_{2,t}A(\boldsymbol{\varepsilon}) - v_{1,t}B(\boldsymbol{\varepsilon}) \\ -v_{1,t}A(\boldsymbol{\varepsilon}) + v_{0,t}B(\boldsymbol{\varepsilon}) \end{pmatrix}.$$

For $t \in (0, 1)$, by Lemma 1, we have $\frac{N_n v_{0,t}}{v_{0,t}v_{2,t} - v_{1,t}^2} \to \mu_2^{-1}q(t)^{-1}$, $\frac{N_n v_{1,t}}{v_{0,t}v_{2,t} - v_{1,t}^2} \to 0$ and $\frac{N_n v_{2,t}}{v_{0,t}v_{2,t} - v_{1,t}^2} \to q(t)^{-1}$. Since $\boldsymbol{\varepsilon}$ is a Gaussian process, by (C.9) and Slusky's Theorem,

$$(N_n h)^{1/2}\boldsymbol{\omega}(t)\boldsymbol{\varepsilon} \xrightarrow{D} N\left(0, \begin{pmatrix} 1 & 0 \\ 0 & \mu_2^{-1} \end{pmatrix} \boldsymbol{\Delta}_t \begin{pmatrix} 1 & 0 \\ 0 & \mu_2^{-1} \end{pmatrix}\right).$$

For (II_3), by Lemma 6, $\boldsymbol{\omega}(t)\boldsymbol{X} = \mathcal{O}(1)$. By Theorem 1, we have $\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta} = \mathcal{O}_p(N_n^{-1/2})$, and therefore, $(II_3) = \mathcal{O}_p(N_n^{-1/2})$. □

Proof of Theorem 3. Let $\boldsymbol{\omega}_1^{(-i)}(t) = (1, 0)\left[\left\{D_t^{(-i)}\right\}^\top K_t^{(-i)} D_t^{(-i)}\right]^{-1}\left\{D_t^{(-i)}\right\}^\top K_t^{(-i)}$, where $D_t^{(-i)}$ is the matrix of D_t with ith row deleted, and $K_t^{(-i)}$ is the matrix K_t with both ith row and column deleted. In addition, we let $\boldsymbol{y}^{*(-i)}$ denote the vector of response variables with the ith entry left out. Straightforward calculation reveals

$$\widehat{f}^{(-i)}(t_i) = \boldsymbol{\omega}_1^{(-i)}(t_i)\boldsymbol{y}^{*(-i)} = \frac{\begin{pmatrix} v_{2,t_i}^{(-i)} & -v_{1,t_i}^{(-i)} \end{pmatrix}}{v_{2,t_i}^{(-i)}v_{0,t_i}^{(-i)} - (v_{1,t_i}^{(-i)})^2}\left\{D_{t_i}^{(-i)}\right\}^\top K_{t_i}^{(-i)}\boldsymbol{y}^{*(-i)} = \sum_{j\neq i} a_j^{(-i)}(t_i)y_j^*,$$

where $a_j^{(-i)}(t_i) = \frac{\begin{pmatrix} v_{2,t_i}^{(-i)} & -v_{1,t_i}^{(-i)} \end{pmatrix}}{v_{2,t_i}^{(-i)}v_{0,t_i}^{(-i)} - (v_{1,t_i}^{(-i)})^2}\left\{D_t^{(-i)}\right\}^\top K_{t_i}^{(-i)}$, $v_{0,t_i}^{(-i)} = \sum_{j\neq i} K_h(t_j - t_i) = v_{0,t_i} - K_h(0)$, $v_{1,t_i}^{(-i)} = \sum_{j\neq i} K_h(t_j - t_i)\frac{t_j - t_i}{h} = v_{1,t_i}$ and $v_{2,t_i}^{(-i)} = \sum_{j\neq i} K_h(t_j - t_i)\left(\frac{t_j - t_i}{h}\right)^2 = v_{2,t_i}$.

By Lemma 1,

$$\begin{aligned} a_j^{(-i)}(t_i) &= \frac{v_{2,t_i}^{(-i)}K_h(t_j - t_i)}{v_{2,t_i}^{(-i)}v_{0,t_i}^{(-i)} - (v_{1,t_i}^{(-i)})^2} - \frac{v_{1,t_i}^{(-i)}K_h(t_j - t_i)(t_j - t_i)/h}{v_{2,t_i}^{(-i)}v_{0,t_i}^{(-i)} - (v_{1,t_i}^{(-i)})^2} = \frac{v_{2,t_i}K_h(t_j - t_i)}{\left\{v_{0,t_i} - K_h(0)\right\}v_{2,t_i} - v_{1,t_i}^2} - \frac{v_{1,t_i}K_h(t_j - t_i)(t_j - t_i)/h}{\left\{v_{0,t_i} - K_h(0)\right\}v_{2,t_i} - v_{1,t_i}^2} \\ &= \frac{K\left(\frac{t_j - t_i}{h}\right) - \frac{\mu_{1,t_i}}{\mu_{2,t_i}}K\left(\frac{t_j - t_i}{h}\right)\left(\frac{t_j - t_i}{h}\right)}{N_n h q(t_i)(\mu_{0,t_i}\mu_{2,t_i} - \mu_{1,t_i}^2)/\mu_{2,t_i} - K(0)} + \mathcal{O}\left(N_n^{-1} + N_n^{-1}\zeta_n h^{-2}\right). \end{aligned}$$

Thus, for the leave-one-out cross-validation (CV) score function,

$$\mathrm{E}\{\mathrm{CV}(h)\} = \mathrm{E}\left[\frac{1}{N_n}\sum_{i=1}^{N_n}\{f(t_i)+\varepsilon_i-\widehat{f}^{(-i)}(t_i)\}^2\right] = \frac{1}{N_n}\sum_{i=1}^{N_n}\mathrm{E}\{f(t_i)-\widehat{f}^{(-i)}(t_i)\}^2+\frac{1}{N_n}\sum_{i=1}^{N_n}\mathrm{Var}(Y_i)-\frac{2}{N_n}\sum_{i=1}^{N_n}\mathrm{Cov}\{\widehat{f}^{(-i)}(t_i),\varepsilon_i\}.$$

Denote $A(h)=\frac{2}{N_n}\sum_{i=1}^{N_n}\mathrm{Cov}\left\{\widehat{f}^{(-i)}(t_i),\varepsilon_i\right\}$, we have

$$A(h)=\frac{2}{N_n}\sum_{i=1}^{N_n}\sum_{j\neq i}\frac{K\left(\frac{t_j-t_i}{h}\right)\left(1-\frac{\mu_{1,t_i}}{\mu_{2,t_i}}\left(\frac{t_j-t_i}{h}\right)\right)}{b(t_i)-K(0)}\mathrm{Cov}(\varepsilon_i,\varepsilon_j)+\mathcal{O}\left(N_n^{-1}\right),$$

since $\|\Gamma\|_\infty=\mathcal{O}(1)$ as shown in the proof of Theorem 1 in Chu et al. [5]. Under the asymptotic framework (A.1), and since $K(\cdot)$ has a bounded first-order derivative at the origin, we obtain

$$\frac{2}{N_n}\sum_{i=1}^{N_n}\sum_{\substack{j\neq i\\ |t_j-t_i|\le C_n}}\frac{K\left(\frac{t_j-t_i}{h}\right)\left(1-\frac{\mu_{1,t_i}}{\mu_{2,t_i}}\left(\frac{t_j-t_i}{h}\right)\right)}{b(t_i)-K(0)}\mathrm{Cov}(\varepsilon_i,\varepsilon_j)=\frac{K(0)}{b(t_i)-K(0)}\left(\frac{2}{N_n}\sum_{i=1}^{N_n}\sum_{\substack{j\neq i\\ B_n|t_j-t_i|<C_n}}\mathrm{Cov}(\varepsilon_i,\varepsilon_j)\right)+o\left(\frac{1}{N_nh}\right).$$

Note that

$$\sum_{\substack{j\neq i\\ |t_j-t_i|>C_n}}\mathrm{Cov}(\varepsilon_i,\varepsilon_j)=\sum_{m'=\lfloor\left(\frac{B_nC_n}{b}\right)\rfloor}\mathcal{O}\left(\frac{b}{B_n\zeta_n}\right)\max_{mb\le|u_2|\le(m+1)b}\gamma_1(|u|)\le\mathcal{O}\left(\frac{b}{B_n\zeta_n}\right)\int_{B_nC_n}^{\infty}\gamma_1(u)du\to 0,$$

so we have

$$\frac{2}{N_n}\sum_{i=1}^{N_n}\sum_{|t_j-t_i|>C_n}\frac{K\left(\frac{t_j-t_i}{h}\right)\left(1-\frac{\mu_{1,t_i}}{\mu_{2,t_i}}\left(\frac{t_j-t_i}{h}\right)\right)}{b(t_i)-K(0)}\mathrm{Cov}(\varepsilon_i,\varepsilon_j)=o\left(\frac{1}{N_nh}\right).$$

Therefore, $A(h)=K(0)\left(\frac{2}{N_n}\sum_{i=1}^{N_n}\sum_{\substack{j\neq i\\ |t_j-t_i|<C_n}}\frac{\mathrm{Cov}(\varepsilon_i,\varepsilon_j)}{b(t_i)-K(0)}\right)+o\left(\frac{1}{N_nh}\right)$. Thus, the desired results are shown. □

Appendix B. Supplementary material

Supplementary material related to this article can be found online at https://doi.org/10.1016/j.jmva.2021.104735.

References

[1] D. Al-Sulami, Z. Jiang, Z. Lu, J. Zhu, On a semiparametric data-driven nonlinear model with penalized spatio-temporal lag interactions, J. Time Series Anal. 40 (2019) 327–342.
[2] N.S. Altman, Kernel smoothing of data with correlated errors, J. Am. Stat. Assoc. 85 (1990) 749–759.
[3] S. Bandyopadhyay, C. Jentsch, S. Subba Rao, A spectral domain test for stationarity of spatio-temporal data, J. Time Series Anal. 38 (2017) 326–351.
[4] S. Bandyopadhyay, S.S. Rao, A test for stationarity for irregularly spaced spatial data, J. R. Stat. Soc. Ser. B Stat. Methodol. 79 (2017) 95–123.
[5] T. Chu, J. Liu, J. Zhu, H. Wang, Spatio-temporal expanding distance asymptotic framework for locally stationary processes, Sankhya A (2020) 1–25.
[6] T. Chu, J. Zhu, H. Wang, Penalized maximum likelihood estimation and variable selection in geostatistics, Ann. Statist. 39 (2011) 2607–2625.
[7] T. Chu, J. Zhu, H. Wang, Semiparametric modeling with nonseparable and nonstationary spatio-temporal covariance functions and its inference, Statist. Sinica 29 (2019) 1233–1252.
[8] N. Cressie, Statistics for Spatial Data, revised edition, Wiley, New York, 1993.
[9] N. Cressie, H.-C. Huang, Classes of nonseparable spatio-temporal stationary covariance functions, J. Amer. Statist. Assoc. 94 (1999) 1330–1340.
[10] N. Cressie, S. Lahiri, The asymptotic distribution of REML estimators, J. Multivariate Anal. 45 (1993) 217–233.
[11] A. Datta, S. Banerjee, A.O. Finley, N.A. Hamm, M. Schaap, Nonseparable dynamic nearest neighbor gaussian process models for large spatio-temporal data with an application to particulate matter analysis, Ann. Appl. Stat. 10 (2016) 1286.
[12] K. De Brabanter, J. De Brabanter, J.A.K. Suykens, B. De Moor, Kernel regression in the presence of correlated errors, J. Mach. Learn. Res. 12 (2011) 1955–1976.
[13] J. Fan, I. Gijbels, Local Polynomial Modelling and its Applications, Chapmann and Hall/CRC, London, 1996.
[14] J. Fan, T. Huang, Profile likelihood inferences on semiparametric varying-coefficient partially linear models, Bernoulli 11 (2005) 1031–1057.
[15] M. Francisco-Fernandez, J.D. Opsomer, Smoothing parameter selection methods for nonparametric regression with spatially correlated errors, Canad. J. Statist. 33 (2005) 279–295.
[16] M. Fuentes, L. Chen, J.M. Davis, A class of nonseparable and nonstationary spatial temporal covariance functions, Environmetrics 19 (2008) 487–507.
[17] J. Gao, H. Liang, Statistical inference in single-index and partially nonlinear models, Ann. Inst. Statist. Math. 49 (1997) 493–517.
[18] J. Gao, Z. Lu, D. Tjostheim, Estimation in semiparametric spatial regression, Ann. Statist. 34 (2006) 1395–1435.
[19] T. Gneiting, Nonseparable, stationary covariance functions for space–time data, J. Amer. Statist. Assoc. 97 (2002) 590–600.
[20] G.H. Golub, M. Heath, G. Wahba, Generalized cross-validation as a method for choosing a good ridge parameter, Technometrics 21 (1979) 215–223.

[21] P. Hall, P. Patil, Properties of nonparametric estimators of autocovariance for stationary random fields, Probab. Theory Related Fields 99 (1994) 399–424.
[22] W. Hä, H. Liang, J.T. Gao, Partially Linear Models, Springer, New York, 2000.
[23] S. Lahiri, Central limit theorems for weighted sums of a spatial process under a class of stochastic and fixed designs, Sankhya A 65 (2003a) 356–388.
[24] S. Lahiri, Resampling Methods for Dependent Data, Springer, New York, 2003b.
[25] K. Lake, J. Zhu, H. Wang, J. Volckens, K.A. Koehler, Effects of data sparsity and spatiotemporal variability on hazard maps of workplace noise, J. Occup. Environ. Hygiene 12 (2015) 256–265.
[26] H. Liang, W. Härdle, R.J. Carroll, Estimation in a semiparametric partially linear errors-in-variables model, Ann. Statist. 27 (1999) 1519–1535.
[27] H. Liang, R. Li, Variable selection for partially linear models with measurement errors, J. Amer. Statist. Assoc. 104 (2009) 234–248.
[28] W.-L. Loh, Fixed-domain asymptotics for a subclass of Matérn-type gaussian random fields, Ann. Statist. 33 (2005) 2344–2394.
[29] Z. Lu, D.J. Steinskog, D. Tjøstheim, Q. Yao, Adaptively varying coefficient spatiotemporal models, J. R. Stat. Soc. Ser. B Stat. Methodol. 71 (2009) 859–880.
[30] Z. Lu, D. Tjøstheim, Nonparametric estimation of probability density functions for irregularly observed spatial data, J. Amer. Statist. Assoc. 109 (2014) 1546–1564.
[31] G. Ludwig, T. Chu, J. Zhu, H. Wang, K. Koehler, Static and roving sensor data fusion for spatio-temporal hazard mapping with application to occupational exposure assessment, Ann. Appl. Stat. 11 (2017) 139–160.
[32] K.V. Mardia, R.J. Marshall, Maximum likelihood estimation of models for residual covariance in spatial regression, Biometrika 71 (1984) 135–146.
[33] J. Opsomer, Y. Wang, Y. Yang, Nonparametric regression with correlated errors, Statist. Sci. 16 (2001) 134–153.
[34] E. Porcu, A. Alegria, R. Furrer, Modeling temporally evolving and spatially globally dependent data, Internat. Statist. Rev. 86 (2018) 344–377.
[35] A. Rodrigues, P.J. Diggle, A class of convolution-based models for spatio-temporal processes with non-separable covariance structure, Scand. J. Stat. 37 (2010) 553–567.
[36] D. Ruppert, M. Wand, R. Carroll, Semiparametric Regression, Cambridge University Press, New York, 2003.
[37] P. Speckman, Kernel smoothing in partial linear models, J. R. Stat. Soc. Ser. B Stat. Methodol. 50 (1988) 413–436.
[38] M.L. Stein, Interpolation of Spatial Data: Some Theory for Kriging, Springer, New York, 1999.
[39] M.L. Stein, Space–time covariance functions, J. Amer. Statist. Assoc. 100 (2005) 310–321.
[40] C.J. Stone, Optimal global rates of convergence for nonparametric regression, Ann. Statist. 10 (1982) 1040–1053.
[41] J.R. Stroud, P. Müller, B. Sansó, Dynamic models for spatiotemporal data, J. R. Stat. Soc. Ser. B Stat. Methodol. 63 (2001) 673–689.
[42] Y. Sun, H. Yan, W. Zhang, Z. Lu, A semiparametric spatial dynamic model, Ann. Statist. 42 (2014) 700–727.
[43] M. Vogt, O. Linton, Nonparametric estimation of a periodic sequence in the presence of a smooth trend, Biometrika 101 (2014) 121–140.
[44] L. Wasserman, All of Nonparametric Statistics, Springer, New York, 2010.
[45] Z. Ying, Maximum likelihood estimation of parameters under a spatial sampling scheme, Ann. Statist. 21 (1993) 1567–1590.
[46] H. Zhang, Inconsistent estimation and asymptotically equivalent interpolations in model-based geostatistics, J. Amer. Statist. Assoc. 99 (2004) 250–261.