Positive definite nonparametric regression using an evolutionary algorithm with application to covariance function estimation

Myeongjong Kang
Texas A&M University
College Station, TX, USA
myeongjong@tamu.edu

ABSTRACT

We propose a novel nonparametric regression framework subject to the positive definiteness constraint. It offers a highly modular approach for estimating covariance functions of stationary processes. Our method can impose positive definiteness, as well as isotropy and monotonicity, on the estimators, and its hyperparameters can be decided using cross validation. We define our estimators by taking integral transforms of kernel-based distribution surrogates. We then use the iterated density estimation evolutionary algorithm, a variant of estimation of distribution algorithms, to fit the estimators. We also extend our method to estimate covariance functions for point-referenced data. Compared to alternative approaches, our method provides more reliable estimates for long-range dependence. Several numerical studies are performed to demonstrate the efficacy and performance of our method. Also, we illustrate our method using precipitation data from the Spatial Interpolation Comparison 97 project.

CCS CONCEPTS

• **Mathematics of computing** → **Nonparametric statistics**; **Regression analysis**; • **Computing methodologies** → **Genetic algorithms**.

KEYWORDS

covariance estimation; shape constrained regression; variogram analysis; kernel smoothing; probabilistic model-building genetic algorithm

1 INTRODUCTION

There have been numerous studies on nonparametric regression analysis under shape constraints, including monotonicity and convexity [26, 27]. Nonparametric regression techniques are typically

considered as solutions to overcome an inherent weakness of parametric regression models. The failure of the parametric approaches arises when the models are so restricted that the relationship between explanatory and response variables cannot be described properly. Since the nonparametric approaches only assume that the models belong to a certain infinite dimensional collection of functions, they are more flexible in estimating the target function [15, 29]. However, a number of regression problems require estimators to have certain features or shapes. For instance, to estimate covariance functions or variograms using regression techniques, positive definiteness has been considered as a crucial constraint since the covariance estimates are ultimately used for kriging or prediction [11, 24]. Also, isotropy and monotonicity are typically regarded as reasonable assumptions in practice. To tackle this issue, the most widely used approach is to use pre-selected parametric covariance functions such as Matérn or squared exponential kernels [22], since positive definiteness has not received abundant attention in nonparametric estimation literature, compared to other shape constraints.

There have been several studies on positive definite nonparametric regression approaches. After [1] emphasized the importance of using positive definite functions for estimating covariance functions, [9] proposed several criteria that allow to check whether a function is positive definite. Although they focused on parametric covariance function estimation, they provided necessary and sufficient conditions of "permissible" covariance function models. Notably, [42] proposed a nonparametric method to estimate isotropic covariance functions using non-negative regression techniques, which is referred to as the Shapiro-Botha method in this paper. Also, [28] suggested a different nonparametric method using kernel regression techniques, which is here referred to as the Hall-Fisher-Hoffmann method. Since both methods provided promising ideas, there have been several related studies: For instance, the Shapiro-Botha method was evaluated against typical parametric covariance models by [8]. [21] and [20] proposed nonparametric variogram estimators based on popular nonparametric regression techniques. Note that they used the Shapiro-Botha method to guarantee positive definiteness of the nonparametric estimators. In addition, [19] discussed the asymptotic properties of a Nadaraya-Watson-type variogram estimator. [24] introduced a procedure for selecting a parametric variogram model based on derivatives of the Shapiro-Botha estimator. A numerical experiment for comparing the estimators was provided in [36]. [18] proposed a bias-corrected version of the Shapiro-Botha estimator.

Both the Shapiro-Botha and Hall-Fisher-Hoffmann methods have advantages and disadvantages. The Shapiro-Botha method is computationally inexpensive and can be easily implemented. Also, one

can impose additional assumptions, such as smoothness or monotonicity. The Hall-Fisher-Hoffmann method has nice large-sample properties. However, one disadvantage of using the Shapiro-Botha method is that there is no clear rule for node selection. Both [42] and [8] considered several ways to select the nodes, for instance, 200 equidistant nodes. But, according to the simulation results in [42] and [8], the use of equidistant nodes causes spurious oscillations of the estimator. To avoid this issue, [23] claimed that the nodes should be chosen as zeros of Bessel functions. The node selection, however, remains a challenging task for using the Shapiro-Botha estimator. On the other hand, the Hall-Fisher-Hoffmann method requires deciding on a proper positive rendering operator for coercing the Fourier transform of kernel-type covariance function estimators into non-negative functions. Arguably, the positive rendering operator selection is also a difficult task, since it is unclear whether positive rendering affects the finite sample properties of the estimator. Furthermore, the Hall-Fisher-Hoffmann method only considers one-dimensional cases and is therefore not applicable to spatial data analysis. Note that several other approaches also consider only one-dimensional cases [e.g. 5]. To the best of our understanding, there is no widely used nonparametric approach for covariance function estimation.

In this paper, we propose a new nonparametric regression framework subject to the positive definiteness constraint, which can be useful for covariance function estimation. It provides a highly modular approach for estimating covariance functions of stationary processes. We introduce three different estimators: Positive definite, isotropic positive definite, and monotone isotropic positive definite estimators. Like the other two popular methods, our approach is based on Bochner's theorem [3] which implies that a continuous positive definite function can be formulated as the Fourier transform of a finite non-negative Borel measure. But, unlike the other two methods, our method allows to decide hyperparameters (e.g., bandwidth of kernel and size of pseudo datasets), for example, using cross validation. To obtain the advantages, the estimators are defined by taking integral transforms of kernel-based distribution surrogates. We suggest using the iterated density estimation evolutionary algorithm which is a variant of estimation of distribution algorithms [EDAs; 34]. Our numerical experiments show that the proposed method can be more accurate and stable than the Shapiro-Botha method, which implies that the estimators can be useful for kriging. The code needed to reproduce our results can be found at https://github.com/myeongjong/NPcov.

The rest of this paper is organized as follows. Our method is presented in Section 2. Our three different estimators are presented in Section 2.1. We provide several closed-form examples of the estimators in Section 2.2. We propose to obtain the estimators using the algorithm presented in Section 2.3. We discuss how to estimate covariance functions using our method in Section 2.4. We evaluate our estimators on synthetic data in Section 3. In specific, we consider a typical regression problem in Section 3.1 and covariance function estimation in Section 3.2. We also illustrate our method with real data from the Spatial Interpolation Comparison (SIC) 97 project. Section 5 concludes and discusses future work. Supplementary material contains proofs, additional numerical results, and empirical evidence on the convergence of the iterated density estimation evolutionary algorithm we use in this paper.

2 METHOD

In this section, we propose a positive definite nonparametric regression framework using kernel smoothing techniques [43, 44]. Our estimators are based on Bochner's theorem [3] and an inference algorithm is obtained by adopting the iterated density estimation evolutionary algorithm [IDEA; 4].

2.1 Model and estimation

Suppose that data $(\mathbf{x}_1, y_1), \cdots, (\mathbf{x}_n, y_n)$ are observed, where $\mathbf{x}_i \in \mathbb{R}^d, d \geq 1$, and $y_i \in \mathbb{R}$ are the ith explanatory and response variables, respectively. Also, assume that the response can be modeled with the unknown regression function f, which implies that

$$y_i = f(\mathbf{x}_i) + \epsilon_i, \quad i = 1, 2, \cdots, n,$$

where ϵ_i is the ith observation error. Our goal here is to estimate the regression function f under reasonable assumptions. We consider three different types of the regression function: (i) In the general case, we assume that f is continuous and positive definite. Also, for the sake of simplicity, we assume that $f(\mathbf{0}) = 1$. (ii) In the isotropic case, it is additionally assumed in the general case that f is isotropic. (iii) In the monotone case, it is additionally assumed in the isotropic case that f is monotone.

For the general case, Bochner's theorem [3] says that f can be represented in the form

$$f(\mathbf{x}) = \int_{\mathbb{R}^d} e^{-2\pi i \mathbf{x} \cdot \mathbf{u}} \, dF(\mathbf{u}), \tag{1}$$

where F is a distribution function on $\mathbb{R}^d$. It is clear that an estimator of f can be obtained by imposing a restriction on F. In this paper, we define a kernel-based surrogate of F by

$$\tilde{F}_m(\mathbf{u}) = \frac{1}{m} \sum_{i=1}^{m} \int_{-\infty}^{\mathbf{u}} \mathbf{K_H}(\tilde{\mathbf{u}} - \mathbf{v}_i) d\tilde{\mathbf{u}},$$

where $m \geq 1$, $\mathbf{v}_1, \cdots, \mathbf{v}_m \in \mathbb{R}^d$, $\mathbf{K}$ is a d-variate kernel function satisfying $\int \mathbf{K}(\mathbf{t}) d\mathbf{t} = 1$, $\mathbf{H}$ is a positive definite $d \times d$ matrix called the bandwidth matrix, and $\mathbf{K_H}(\mathbf{t}) = |\mathbf{H}|^{-\frac{1}{2}} \mathbf{K}(\mathbf{H}^{-\frac{1}{2}}\mathbf{t})$. See [44] for details about the formulation. We should mention that that although $\mathbf{v}_1, \cdots, \mathbf{v}_m$ are typically observed as "data", this is not the case; they are vectors specifying an estimator $\tilde{F}_m$. We call $\mathbf{v}_1, \cdots, \mathbf{v}_m$ pseudo data. Also, we call m the size of pseudo data. Therefore, an estimator of f can be modeled by replacing F in Equation (1) by $\tilde{F}_m$ as follows:

$$\hat{f}_m(\mathbf{x}) = \int_{\mathbb{R}^d} e^{-2\pi i \mathbf{x} \cdot \mathbf{u}} \, d\tilde{F}_m(\mathbf{u}) = \frac{1}{m} \sum_{i=1}^{m} \cos(2\pi \mathbf{x} \cdot \mathbf{v}_i) \mathcal{F}(\mathbf{K_H})(\mathbf{x}),$$

where $\mathcal{F}(\mathbf{K_H})$ is the Fourier transform of $\mathbf{K_H}$. This estimator is referred to as the general positive definite regression estimator or simply general estimator. Note that one can use any popular kernel functions including uniform, Epanechnikov and Gaussian kernels.

Although the general estimator $\hat{f}_m$ is formulated with the commonly used functions and operators, it can be computationally expensive since it is necessary to decide hyperparameters m and $\mathbf{H}$ and optimize the pseudo data $\mathbf{v}_1, \cdots, \mathbf{v}_m$ as well. This can be achieved through the algorithm we propose in Section 2.3, but the computational issue can be avoided by making an additional assumption on the shape of the regression function; in this paper, we consider isotropy and monotonicity. A main reason we consider

isotropic functions is that isotropy is one of the most frequently assumed properties for modeling covariance functions. Furthermore, in practice, it is typically assumed that covariance functions are not only isotropic but also monotonically decreasing.

In the isotropic case that there is a function $g : [0, \infty) \to \mathbb{R}$ such that $f(\mathbf{x}) = g(|\mathbf{x}|)$ for all $\mathbf{x} \in \mathbb{R}^d$, our goal is to estimate the one dimensional function g. Note that g is continuous, positive definite and isotropic. Also, $g(0) = 1$. To obtain an estimator, we use Schoenberg's theorem [40] which can be viewed as a modified version of Bochner's theorem for isotropic functions [7, 16]. By Schoenberg's theorem, g can be represented as

$$g(r) = \int_0^\infty \Omega_d(ru)\mathrm{d}G(u), \tag{2}$$

where G is a distribution function on $\mathbb{R}_0^+ = [0, \infty)$,

$$\Omega_d(r) = \begin{cases} \cos(r) & \text{for } d = 1 \\ \Gamma\left(\frac{d}{2}\right)\left(\frac{2}{r}\right)^{(d-2)/2} \mathbf{J}_{(d-2)/2}(r) & \text{for } d \geq 2 \end{cases},$$

Γ is the gamma function and $\mathbf{J}_\nu$ is the Bessel function of the first kind of order $\nu \in \mathbb{R}$. Similar to the general case, we can replace G using kernel smoothing. There is extensive literature on kernel density estimation with a bounded support [see 32, 47, for details]. Here we use the reflection method [10, 41, 43] due to the fact that one can easily formulate the model and use popular kernel functions with the reflection method. We define a surrogate of G by

$$\tilde{G}_m(u) = \frac{1}{m} \sum_{i=1}^m \int_0^u [K_h(\tilde{u} - v_i) + K_h(\tilde{u} + v_i)]\mathrm{d}\tilde{u}, \tag{3}$$

where $m \geq 1$, $v_1, \cdots, v_m \in \mathbb{R}_0^+$, K is a univariate kernel function satisfying $\int K(u)\mathrm{d}u = 1$, $h > 0$ is the bandwidth, and $K_h(t) = \frac{1}{h}K(\frac{t}{h})$. Therefore, an estimator of g can be obtained by replacing G in Equation (2) by $\tilde{G}_m$ as follows:

$$\hat{g}_m(r) = \frac{1}{m} \sum_{i=1}^m \left[\int_0^\infty \Omega_d(ru)K_h(u - v_i)\mathrm{d}u \right.$$
$$\left. + \int_0^\infty \Omega_d(ru)K_h(u + v_i)\mathrm{d}u \right]. \tag{4}$$

We call this estimator the isotropic positive definite regression estimator or simply isotropic estimator. Note that it is required to decide both m and h and also optimize pseudo data $v_1, \cdots, v_m$.

For the monotone case, we derive the estimator based on the approach presented by [8] in the following manner: From the fact that continuous positive definite isotropic functions on $\mathbb{R}^s$ for all $s \geq 1$ are monotone, [8] proposed to use $\exp(-r^2u^2)$ instead of $\Omega_d(ru)$ in Equation (2). One can easily see that adopting this idea to Equation (4) guarantees monotonicity of the estimator. We propose to estimate the monotone regression function by using

$$\hat{q}_m(r) = \frac{1}{m} \sum_{i=1}^m \left[\int_0^\infty e^{-r^2u^2}K_h(u - v_i)\,\mathrm{d}u \right.$$
$$\left. + \int_0^\infty e^{-r^2u^2}K_h(u + v_i)\,\mathrm{d}u \right].$$

This model is referred to as the monotone isotropic positive definite regression estimator or simply monotone estimator. Note that the monotone estimator requires to decide m and h and also optimize pseudo data $v_1, \ldots, v_m$, similar to the isotropic estimator.

2.2 Closed-form examples

We here obtain closed-form expressions of the estimators for several scenarios. According to our numerical experiments, although the isotropic and monotone estimators require evaluation of indefinite integrals, one can compute them numerically and also efficiently. But, it is also true that the estimators can be explicitly formulated and computed quickly by using proper kernel functions. We here consider three popular kernel functions: The uniform kernel is popular due to its simplicity; the Epanechnikov kernel is also popular mainly due to its theoretical properties; and the Gaussian kernel is widely used for its numerical stability and differentiability. We henceforth restrict ourselves to the cases where regression functions are isotropic. Furthermore, we assume that the dimension of the input domain $d = 2$. Note that the definitions of the kernels and proofs for the following examples are in Supplementary material.

Uniform kernel. Assume $r > 0$. Let K be the uniform kernel. Then the isotropic estimator can be represented as

$$\hat{g}_m(r) = \frac{1}{2mhr} \sum_{j=1}^m \left[\Lambda_1(r(v_j + h)) - \Lambda_1(r(v_j - h)) \right],$$

where $\mathcal{H}_\nu$ is the Sturve function of order ν and Λ_1 is defined as

$$\Lambda_1(r) = r\mathbf{J}_0(r) + \frac{\pi r}{2} \left[\mathbf{J}_1(r)\mathcal{H}_0(r) - \mathbf{J}_0(r)\mathcal{H}_1(r) \right].$$

Also, the monotone estimator can be represented as

$$\hat{q}_m(r) = \frac{\sqrt{\pi}}{2mhr} \sum_{j=1}^m \left[\Phi\left(\sqrt{2}r(v_j + h)\right) - \Phi\left(\sqrt{2}r(v_j - h)\right) \right],$$

where Φ is the distribution function of standard normal distribution.

Epanechnikov kernel. Assume $r > 0$. Let K be the Epanechnikov kernel. Then the isotropic estimator can be represented as

$$\hat{g}_m(r) = \frac{3}{4mhr} \sum_{j=1}^m \left[\left(1 - \frac{v_j^2}{h^2}\right)\left(\Lambda_1(r(v_j + h)) - \Lambda_1(r(v_j - h))\right) \right.$$
$$- \left(1 - \frac{v_j^2}{h^2}\right)\left(\mathbf{J}_1(r(v_j + h)) - \mathbf{J}_1(r(v_j - h))\right)$$
$$\left. + \left(\frac{1}{r^2h^2}\right)\left(\Lambda_0(r(v_j + h)) - \Lambda_0(r(v_j - h))\right) \right],$$

where

$$\Lambda_0(r) = \frac{\pi r}{2} \left[\mathbf{J}_1(r)\mathcal{H}_0(r) - \mathbf{J}_0(r)\mathcal{H}_1(r) \right].$$

Also, the monotone estimator can be represented as

$$\hat{q}_m(r) = \frac{3}{4mhr} \sum_{j=1}^m \left[\sqrt{\pi}\left(1 - \frac{v_j^2}{h^2} - \frac{1}{2r^2h^2}\right)\left(\Phi(r(v_j+h)) - \Phi(r(v_j-h))\right) \right.$$
$$\left. + \left(\frac{1}{r^2h^2}\right)\left((v_j + h)e^{-r^2(v_j-h)^2} - (v_j - h)e^{-r^2(v_j+h)^2}\right) \right].$$

Gaussian kernel. Assume $r > 0$. Let K be the Gaussian kernel. Then the isotropic estimator can be represented as

$$\hat{g}_m(r) = \frac{1}{m} e^{-\frac{1}{4}h^2 r^2} \sum_{j=1}^{m} \mathbf{J}_0^{(2)} \left(\frac{\mathrm{i}}{4} h^2 r^2, r v_j, \mathrm{i}^{-1} \right)$$

$$= \frac{1}{m\pi} e^{-\frac{1}{4}h^2 r^2} \sum_{j=1}^{m} \int_0^\pi \exp\left(\frac{1}{4} h^2 r^2 \cos(2\phi) \right) \cos(r v_j \sin(\phi)) \mathrm{d}\phi,$$

where $\mathrm{i} = \sqrt{-1}$ and $\mathbf{J}_0^{(2)}$ is the $\left[\begin{smallmatrix} (2) \\ 0 \end{smallmatrix} \right]$ order generalized Bessel function [12]. Also, the monotone estimator can be represented as

$$\hat{q}_m(r) = \frac{1}{mh} \sum_{j=1}^{m} \frac{1}{\sqrt{2A(r,h)}} e^{C_j(r,h)},$$

where $A(r,h) = r^2 + \frac{1}{2h^2}$ and $C_j(r,h) = \frac{v_j^2}{4h^4 A(r,h)} - \frac{v_j^2}{2h^2}$ for $j = 1, 2, \cdots, m$. Our numerical experiments show that the estimators with the Gaussian kernel can be obtained more efficiently and differences in estimation performance between the kernels are negligible, as illustrated in Figure 1. Therefore, from now on we only consider the Gaussian kernel.

2.3 Algorithm

To search for proper pseudo data, we suggest to use the iterated density estimation evolutionary algorithm [IDEA; 4], a variant of estimation of distribution algorithms [EDA; 34], mainly because optimizing pseudo data is a non-convex problem. EDAs are probabilistic model fitting genetic algorithms, and their convergence is discussed by [46] and [38]. The concepts are already introduced to statisticians by [6]. Here we focus on introducing an algorithm for obtaining isotropic and monotone estimators.

An unified framework for IDEA is specified in Algorithm 1. To use the framework, one needs to first decide not only kernel-specifying hyperparameters (bandwidth of kernel $h > 0$ and size of pseudo data $m \geq 0$) but also algorithm-specifying hyperparameters (size of collection of pseudo datasets $l \geq 1$ and acceptance rate $\tau > 0$). Fortunately, it appears that l/m and τ do not significantly affect the estimation performance. In numerical experiments, we assume $l = 10m$ and $\tau = 0.1$. We use cross validation to decide kernel-specifying hyperparameters. Note that we provide numerical evidence of convergence of the IDEA in Supplementary material.

The IDEA framework requires specifying six steps: Initialization, evaluation, selection, estimation, replacement and termination. The initialization step is to generate l pseudo datasets randomly from the initialized surrogate $\tilde{G}_m$ of the distribution G. Our numerical studies show that the exponential distribution with rate parameter 1 works properly in any case we consider in this paper.

The evaluation step is to compute an objective function value for each pseudo dataset. Arguably, there are a variety of objective functions we can use for the algorithm. For instance, if the observation errors are independent and homoskedastic, we suggest to use the mean of squared errors (MSE): In specific, the objective function L of ith pseudo-dataset $\Theta_i^{(k)}$ of kth step is defined as

$$L(\Theta_i^{(k)}) = \frac{1}{n} \sum_{j=1}^{n} \left(y_j - \hat{g}_m \left(|\mathbf{x}_j|; \Theta_i^{(k)} \right) \right)^2$$

where $\hat{g}_m(|\mathbf{x}_j|; \Theta_i^{(k)})$ is defined by Equation (4),

$$\Theta_i^{(k)} = \{ v_{1,i}^{(k)}, \cdots, v_{m,i}^{(k)} \}$$

and $v_{j,i}^{(k)}$ is the jth element of ith pseudo dataset of kth step for all $1 \leq j \leq m$, $1 \leq i \leq l$. But, the errors are clearly correlated and heteroscedastic for covariance function estimation scenarios. We will discuss this issue in Section 2.4.

The selection step is to select $\lfloor \tau l \rfloor$ best pseudo datasets generated from the previous step with respect to the values of the objective function L. This action of k-step can be described as follows:

· Rearrange l pseudo datasets $\Theta_1^{(k)}, \cdots, \Theta_l^{(k)}$ in ascending order of objective function values $L(\Theta_1^{(k)}), \cdots, L(\Theta_l^{(k)})$ and denote the ordered pseudo datasets by $\Theta_{(1)}^{(k)}, \cdots, \Theta_{(l)}^{(k)}$.

· Choose $\lfloor \tau l \rfloor$ pseudo datasets $\Theta_{(1)}^{(k)}, \cdots, \Theta_{(\lfloor \tau l \rfloor)}^{(k)}$ of which objective function values are smaller than those of other pseudo datasets $\Theta_{(\lfloor \tau l \rfloor + 1)}^{(k)}, \cdots, \Theta_{(l)}^{(k)}$.

The estimation step is to estimate G with the $\lfloor \tau l \rfloor$ selected pseudo datasets. In the kth step, one can obtain the k-step distribution surrogate $\tilde{G}_m^{(k)}$ using Equation (3) with the merged pseudo dataset $\Theta^{(k)}$ defined by $\Theta^{(k)} = \Theta_{(1)}^{(k)} \cup \cdots \cup \Theta_{(\lfloor \tau l \rfloor)}^{(k)}$.

Algorithm 1: Iterated density estimation evolutionary algorithm (IDEA). Note that objective function and pseudo-dataset are called OBF and PSD, respectively.

1: **procedure** IDEA(h, m, l, τ)
2:   Generate a collection of l PDs ▷*Initialization*
3:   **while** the termination rule is true **do**
4:     Evaluate OBF values of PSDs ▷*Evaluation*
5:     Select $\lfloor \tau l \rfloor$ best PSDs ▷*Selection*
6:     Estimate G with the selected PSDs ▷*Estimation*
7:     Check the termination rule ▷*Termination*
8:     **if** the termination rule is false **then**
9:       Generate $l - \lfloor \tau l \rfloor$ PSDs ▷*Replacement*
10:    **end if**
11:   **end while**
12:   **return** a collection of $\lfloor \tau l \rfloor$ selected PSDs
13: **end procedure**

The replacement step is to replace $l - \lfloor \tau l \rfloor$ worst pseudo datasets contained in the collection of pseudo datasets with new pseudo datasets. This can be accomplished using the simulation approach proposed by [43]. Replacement of kth step can be described as follows:

· Denote the merged pseudo dataset $\Theta^{(k)}$ by $\{\tilde{v}_1, \cdots, \tilde{v}_{m\lfloor \tau l \rfloor}\}$.
· Choose $I_1, \cdots, I_{m(l - \lfloor \tau l \rfloor)}$ uniformly with replacement from $\{1, \cdots, m\lfloor \tau l \rfloor\}$.
· Draw random samples $e_1, \cdots, e_{m(l - \lfloor \tau l \rfloor)}$ from the kernel K.
· Set $\tilde{v}_i' = |\tilde{v}_{I_i} + e_i|$ for all $i = 1, 2, \cdots, m(l - \lfloor \tau l \rfloor)$.
· Divide $\{\tilde{v}_1', \cdots, \tilde{v}_{m(l - \lfloor \tau l \rfloor)}'\}$ into $l - \lfloor \tau l \rfloor$ pseudo datasets $\Theta_{\lfloor \tau l \rfloor + 1}^{(k+1)}, \cdots, \Theta_l^{(k+1)}$ uniformly.
· Set $\Theta_1^{(k+1)} = \Theta_{(1)}^{(k)}, \cdots, \Theta_{\lfloor \tau l \rfloor}^{(k+1)} = \Theta_{(\lfloor \tau l \rfloor)}^{(k)}$.
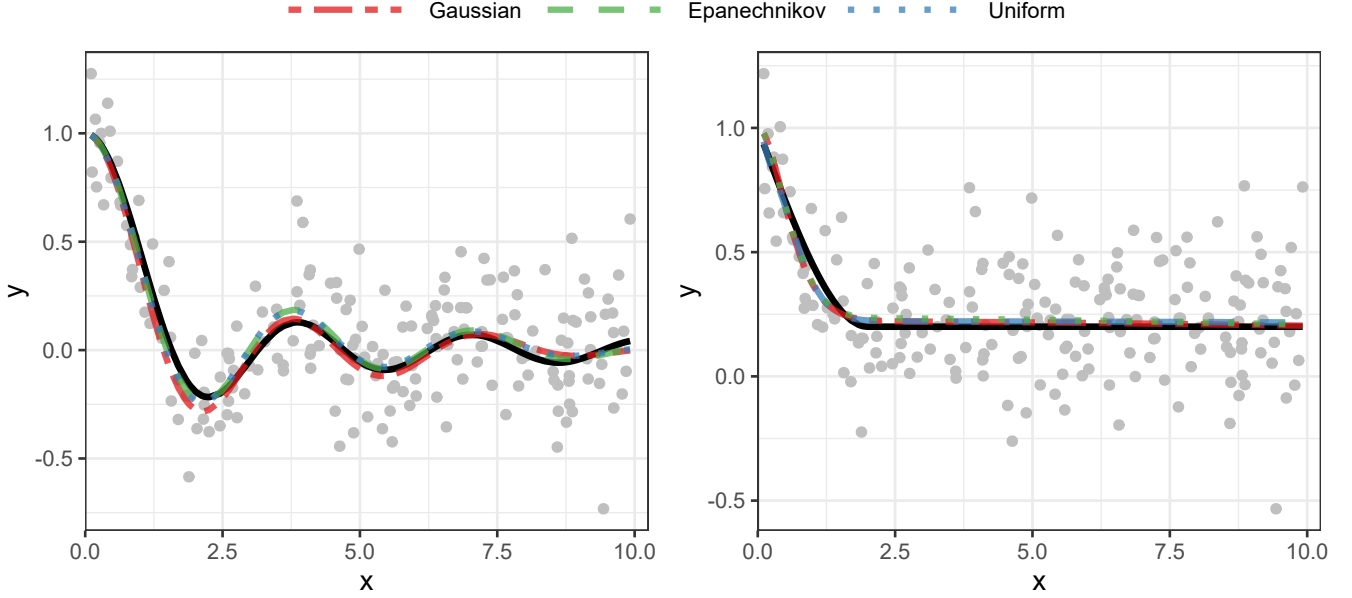
Figure 1: Comparison of three popular kernels for the isotropic (left) and monotone (right) estimators. We consider Gaussian, Epanechnikov and Uniform kernels. The regression equations (5) and (6) are considered on the left and right panels, respectively. The black lines are the true regression equations, and gray points are noisy observations. The estimators are fitted using the iterated density estimation evolutionary algorithm. Bandwidth h and pseudo-data size m are selected using 5-fold cross validation.

Then, $\Theta_1^{(k+1)}, \ldots, \Theta_l^{(k+1)}$ are the $(k+1)$-step pseudo datasets of Algorithm 1. The termination step is based on the concept of Kullback-Leibler (KL) divergence [33]. Let $\hat{\tilde{\psi}}^{(k)}$ denote the kernel density surrogate corresponding to $\tilde{G}_m^{(k)}$. Then, KL divergence from $\tilde{G}_m^{(k-1)}$ to $\tilde{G}_m^{(k)}$ can be computed by

$$D_{KL}(\tilde{G}_m^{(k)}, \tilde{G}_m^{(k-1)}) = \frac{1}{m\lfloor \tau l \rfloor} \sum_{i=1}^{\lfloor \tau l \rfloor} \sum_{j=1}^{m} \log \left(\frac{\hat{\psi}_m^{(k)}(v_{j,i}^{(k-1)})}{\hat{\psi}_m^{(k-1)}(v_{j,i}^{(k-1)})} \right).$$

We terminate Algorithm 1 when $D_{KL} < 10^{-3}$ has been met five times consecutively for stability. Finally, we obtain the estimator using Equation (4) with the pseudo data obtained by merging a collection of $\lfloor \tau l \rfloor$ selected pseudo-datasets at the last step of Algorithm 1.

2.4 Application to covariance function estimation

In this section, we propose a covariance function estimation procedure using our nonparametric positive definite regression estimators. Consider a stationary process $\{Z(\mathbf{s}) \mid \mathbf{s} \in D\}$, where the domain D is a bounded subset of $\mathbb{R}^d$. The covariance function $C : \mathbb{R}^d \to \mathbb{R}$ can be defined as

$$C(\mathbf{s} - \mathbf{s}') = \text{cov}\left(Z(\mathbf{s}), Z(\mathbf{s}')\right) \quad \text{for all } \mathbf{s}, \mathbf{s}' \in D.$$

Let $Z(\mathbf{s}_1), Z(\mathbf{s}_2), \cdots, Z(\mathbf{s}_w)$ denote a realization of the stationary process Z. Then, one can obtain point estimates of the covariance

function using the method proposed by [35]:

$$\hat{c}_{i,j} = \hat{C}(\mathbf{s}_i - \mathbf{s}_j) = \left(Z(\mathbf{s}_i) - \bar{Z}\right)\left(Z(\mathbf{s}_j) - \bar{Z}\right)$$

where $\bar{Z} = \frac{1}{w} \sum_{i=1}^{w} Z(\mathbf{s}_i)$. According to [28], the set of observations may either include or exclude the diagonal terms corresponding to $i = j$. Here, we use diagonal terms only for initialization. This can be framed as a heteroscedastic regression problem with observations $\left(\mathbf{r}_{i,j}, \hat{c}_{i,j}\right)$ where $\mathbf{r}_{i,j} = \mathbf{s}_i - \mathbf{s}_j$. The presence of heteroscedasticity is induced by the fact that the classical point estimates are highly correlated and variable. To tackle this issue, [31] recommended to take large enough groups of distinct pairs according to their distances and computing the average for each group or tolerance region [11, Section 2.4]. Note that [37] and [39] suggested to use weighted averages to obtain more resistant and efficient estimates, respectively. But in this paper, we simply use the classical point estimates, since our estimators are robust and resistant due to their shape constraints, as illustrated in Figure 3.

Now we propose an approach to inferring the covariance function C with our estimators. Since we assume that $f(\mathbf{0}) = 1$ to develop our method in the regression setting, we here need to estimate the variance parameter $\sigma^2 = var(Z(\mathbf{s}))$ as well. We suggest to estimate C as follows: First, we initialize a variance estimator $\hat{\sigma}^2$ using the sample variance $\frac{1}{w} \sum_{i=1}^{w} \left(Z(\mathbf{s}_i) - \bar{Z}\right)^2$ which is widely considered as a reasonable first estimate of σ^2 according to [2]. Second, we infer the function $C_0(\mathbf{r}) = C(\mathbf{r})/\hat{\sigma}^2$ with one of the estimators proposed in this paper. Here we only use $\left(\mathbf{r}_{i,j}, \hat{c}_{i,j}\right)$ where $i \neq j$. We can update $\hat{\sigma}^2$ using the minimizer of the sum of squared differences between the point estimates and $\hat{C}(\mathbf{r}_{i,j}) = \hat{\sigma}^2 \hat{C}_0(\mathbf{r}_{i,j})$,

$1 \leq i, j \leq w$, where $\hat{C}_0(\mathbf{r})$ is the estimator of $C_0(\mathbf{r})$ obtained in the second step. Then, the final covariance function estimator can be obtained by repeating the above steps.

The computational cost of our approach for covariance function estimation is proportional to the number of the point estimates, so it can be high for large point-referenced data. To reduce the computational cost, one can adopt the tolerance-region strategy mentioned above. See [11] for further details.

3 NUMERICAL EXPERIMENTS

We conducted several simulations to demonstrate the performance of our method. We considered two different scenarios: In Section 3.1, we assumed that response y was observed at $\mathbf{x}$ with additive Gaussian noise. Note that two regression equations motivated by popular covariance functions are considered. In Section 3.2, we used a stationary Gaussian process with mean zero and a wave covariance function to illustrate the efficacy of our method. Additional numerical numerical results are presented in Supplementary material.

3.1 Regression with white noise errors

We considered two different regression equations in this section. To demonstrate performance of our isotropic estimator, we considered the wave function [11] which is defined as follows:

$$g(r) = \begin{cases} 1 & \text{for } r = 0 \\ \frac{\sin(2r)}{2r} & \text{for } r > 0. \end{cases} \tag{5}$$

On the other hand, to demonstrate performance of our monotone estimator, we considered the spherical function [11] which is defined as follows:

$$g(r) = \begin{cases} 1 & \text{for } r = 0 \\ 1 - \frac{1}{20}\left(12r - r^3\right) & \text{for } 0 < r \leq 2 \\ 0.2 & \text{for } 2 < r. \end{cases} \tag{6}$$

The shapes of the functions are illustrated in Figure 2 as solid black lines. The 200 inputs are randomly selected from the interval $[0, 10]$ and the white noises are randomly drawn from the Gaussian distribution with mean of zero and standard deviation of 0.2. As competing methods, we considered local constant (Nadaraya-Watson) and local linear estimators, since these estimators are most widely used for kernel-based regression. We used the R package **np** to obtain them [30]. The bandwidths of the competing estimators are decided using cross validation; we use $(h, m) = (0.16, 4)$ for the isotropic estimator and $(h, m) = (0.01, 6)$ for the monotone estimator. Note that we repeated the simulation 200 times.

Our estimators outperformed the other competing estimators, as illustrated in Figure 2 and presented in Table 1. The plots in Figure 2 show that our estimators provided smooth curves, unlike the competing methods. Furthermore, since our approach allowed to use prior knowledge of shapes of functions including positive definiteness and monotonicity, our estimators provided better predictive performance in terms of root mean squared prediction error (RMSPE). We randomly selected test inputs from the same interval for the result presented in Table 1.

3.2 Covariance function estimation

To assess the efficacy of our estimators, we considered a realization of Gaussian process with mean zero and the wave covariance

Estimator	Wave equation	Spherical equation
Local constant	0.0533	0.0499
Local linear	0.0503	0.0430
Our isotropic	0.0419	-
Our monotone	-	0.0399

Table 1: Average of the root mean squared prediction error (RMSPE) of the estimators over 200 simulated data from the regression scenario. We consider two different regression equations: For the wave equation, we use the isotropic estimator. For the spherical equation, we use the monotone estimator. These estimators are compared with local-constant and local-linear regression estimators.

function: For $\mathbf{x}, \mathbf{x}' \in [0, 10/\sqrt{2}]^2$, we define the wave covariance function as

$$\text{cov}(\mathbf{x}, \mathbf{x}') = \frac{\sin\left(\|\mathbf{x} - \mathbf{x}'\|\right)}{\|\mathbf{x} - \mathbf{x}'\|}.$$

A visual comparison is provided in Figure 3. We randomly selected 200 locations in the domain so that we had 19,900 point estimates of the covariance function. Note that the competing methods considered in the previous section are not directly applicable in this setting. So, we considered the Shapiro-Botha method implemented in the R package **npsp** (0.7-8) as a competing method [17]. The discretization nodes of the Shapiro-Botha method are decided following the guidelines in [25]. We used cross validation to select the bandwidth h and size of pseudo-datasets for our method. In specific, our isotropic estimator used $(h, m) = (0.2, 5)$.

It appears that our method provided better performance in the setting we considered. The Shapiro-Botha estimator provided a poor estimate of the variance (or sill). Further, it did not provide covariance estimates at large distances. This is mainly because the number of data pairs whose distances are large is typically very small, so the state-of-the-art estimators can be extremely unstable. According to [31], the maximum distance of reliability for variogram estimation is one-half the maximum distance between locations. In contrast, our method was reliable and robust even at large distances.

4 SPATIAL INTERPOLATION COMPARISON 97 DATA ANALYSIS

We considered the data set SIC100 from the Spatial Interpolation Comparison (SIC) 97 project [14] which includes 100 daily rainfall observations made in Switzerland on the 8th of May 1986 that were randomly selected from a dataset of 467 observations. A main reason we decided to use only 100 observations instead of all observations is that the project aims to estimate the rainfall at the 367 remaining locations. Furthermore, this is an ideal example to demonstrate the main advantage of our method that it provides positive definite covariance function estimators even when the sample size is very small. See https://wiki.52north.org/bin/view/AI_GEOSTATS/EventsSIC97 for details. The data set can be found in the R package **geoR** [13]. Note that we scaled distances between observations to be from 0 to 8 for numerical stability and retrieved
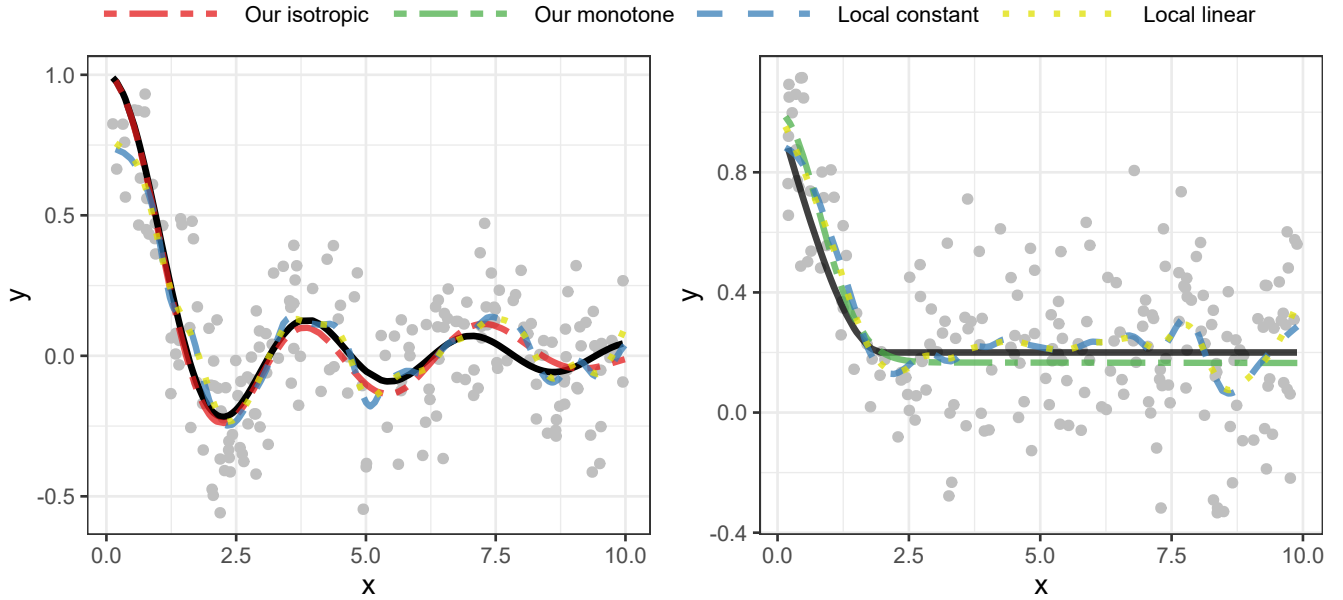
Figure 2: Comparison of our estimators to two popular nonparametric regression estimators: Local-constant and local-linear estimators. We consider the regression equations (5) and (6) and these are presented as solid black lines on the left and right panels, respectively. The gray points are their noisy observations. Our estimators are fitted using the iterated density estimation evolutionary algorithm. Bandwidth h and pseudo-data size m are selected using 5-fold cross validation.

them in Figure 4, and decided the bandwidth h and size of pseudo data m by cross validation with the initial variance estimate.

Figure 4 shows our isotropic and monotone estimators of the underlying covariance function. Also, we again considered the Shapiro-Botha method as a competing method. Although it is not easy to make comparisons directly due to the fact that the true covariance function is unknown, it appears that our isotropic estimator is flexible enough to capture almost any underlying features of the covariance function and our monotone estimator does not suffer from spurious modes, unlike the Shapiro-Botha estimator. If there is good evidence that the covariance function is monotonically decreasing, one can simply use the monotone estimator which only needs a few minutes for hyperparameter selection using cross validation and a few tens of seconds for estimation. Note that a 64-bit workstation with 16 GB RAM and an Intel Core i7-8700K CPU running at 3.70 GHz was used. For the isotropic estimator, the bandwidth h the pseudo-data size m are selected to be 0.18 and 6, respectively, by cross validation. For the monotone estimator, selected h and m are 0.02 and 3, respectively.

5 DISCUSSION

We have introduced a new positive definite nonparametric regression method. To our understanding, this work provides a promising solution to the issue that there was no general framework to obtain positive definite nonparametric covariance function models for stationary processes. Our method allows to use popular kernels, for example, Gaussian and Epanechnikov kernels, and decide hyperparameters using cross validation, like typical nonparametric

regression methods. Furthermore, our method always returns positive definite function estimators and therefore can be useful for kriging. Our numerical experiments suggested that the state-of-the-art method suggested by [42] was outperformed by our estimators. But, due to the presence of heteroscedasticity for the covariance function estimation task, it was not straightforward to make fair comparisons between our method and the Shapiro-Botha method. In specific, since all the point estimates are correlated, the RMSPE cannot be a good measure to compare them.

Our approach can be easily modified for various purposes. There are several functions that can be characterized by integral transformations. For instance, completely monotone and multiply monotone functions can be expressed as integral transforms of nonnegative Borel measures [45] so that our method can be applied for estimating these functions by using different integrands. One can use different performance measures instead of the MSE: Examples include prediction scores on hold-out observations of a spatial process and the KL divergence values of the termination step of the IDEA. We believe that further investigation is required to extend our approach.
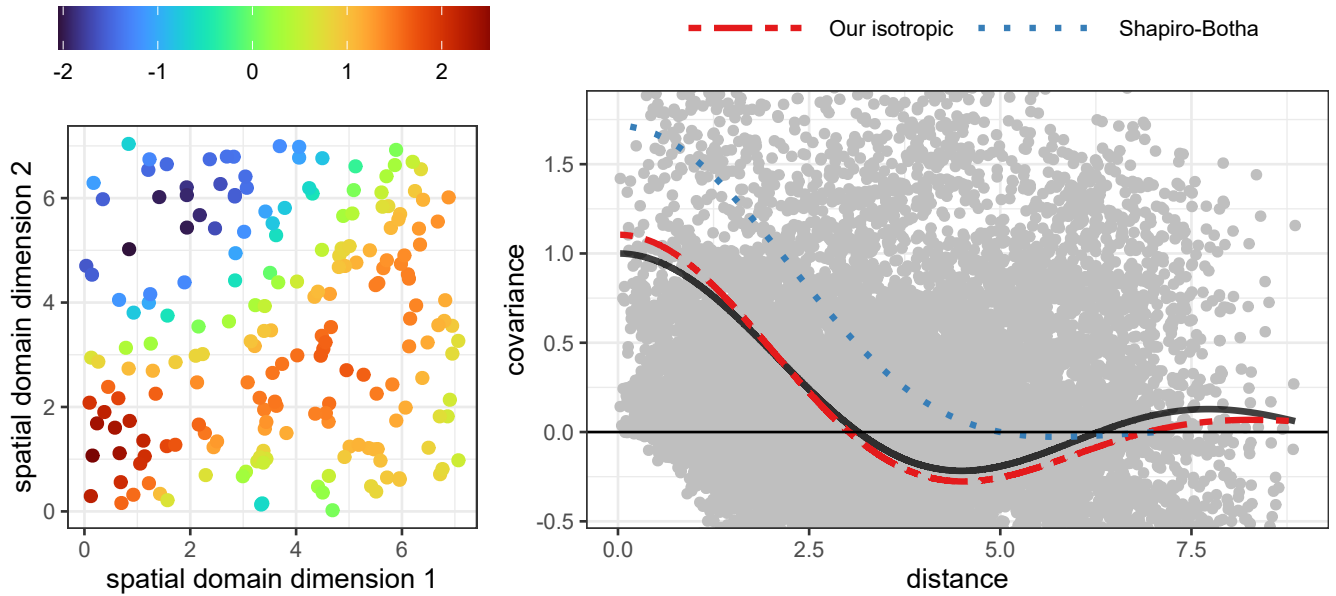
Figure 3: Comparison of our isotropic estimator to the Shapiro-Botha estimator for estimating the wave covariance function of mean-zero Gaussian process. The detailed estimation procedure is described in Section 2.4. A realization of the Gaussian process is depicted in the left panel. The true covariance function is the solid black line on the right panel. The isotropic and Shapiro-Botha estimators are two-dashed red and dotted blue lines, respectively. The gray points are point covariance estimates.
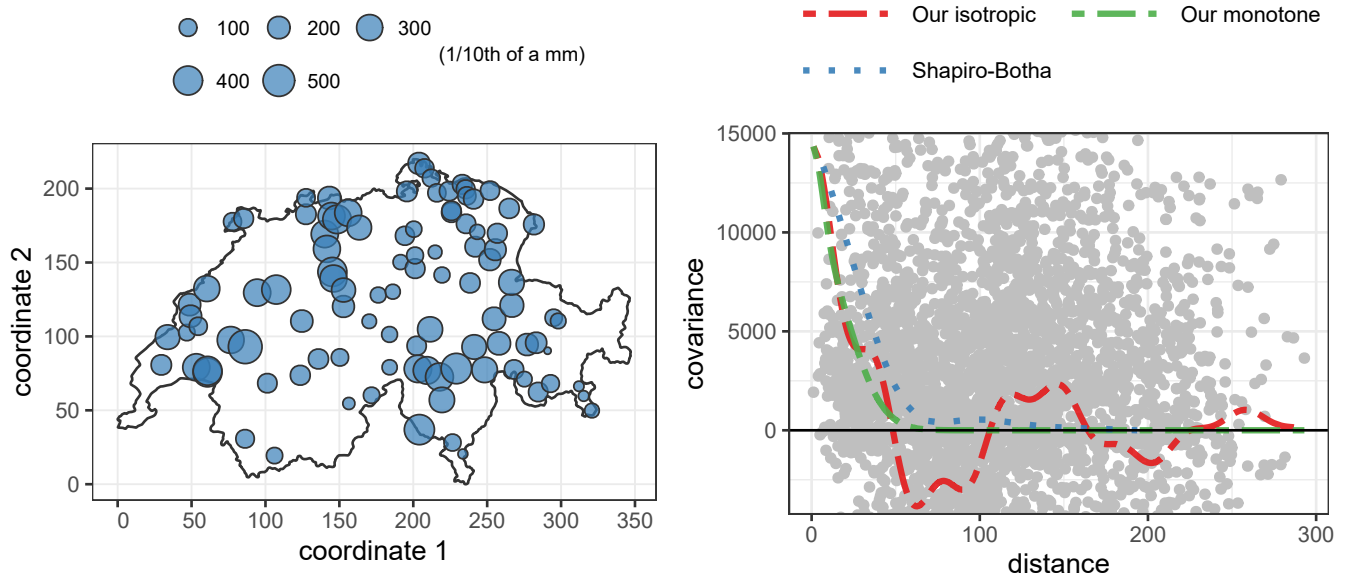

Figure 4: Proportional circle map of SIC100 data from the Spatial Interpolation Comparison (SIC) 97 project (left) and comparison of our isotropic and monotone estimators to the Shapiro-Botha estimator for covariance function estimation with SIC100 data (right). The isotropic and monotone estimators are two-dashed red and green lines, respectively, on the right panel. The Shapiro-Botha estimator is the dotted blue line. The gray points are point covariance estimates.

REFERENCES

[1] Margaret Armstrong and Romain Jabin. 1981. Variogram models must be positive-definite. *Mathematical Geology* 13, 5 (1981), 455–459.

[2] Randal J. Barnes. 1991. The variogram sill and the sample variance. *Mathematical Geology* 23 (1991), 673–678.

[3] Salomon Bochner. 1933. Integration von funktionen, deren werte die elemente eines vektorraumes sind. *Fundamenta Mathematicae* 20, 1 (1933), 262–176.

[4] Peter A. N. Bosman and Dirk Thierens. 1999. An algorithmic framework for density estimation based evolutionary algorithms. *Technical Report, Utrecht University* (1999).

[5] Tony Cai and Ming Yuan. 2010. Nonparametric covariance function estimation for functional and longitudinal data. *Technical Report, University of Pennsylvania and Georgia Institute of Technology* (2010).

[6] Sangit Chatterjee, Matthew Laudato, and Lucy A. Lynch. 1996. Genetic algorithms and their statistical applications: An introduction. *Computational Statistics & Data Analysis* 22, 6 (1996), 633–651.

[7] Elliott W. Cheney and William A. Light. 2009. *A Course in Approximation Theory*. Vol. 101. American Mathematical Society.

[8] S. Cherry. 1996. An evaluation of a non-parametric method of estimating semi-variograms of isotropic spatial processes. *Journal of Applied Statistics* 23, 4 (1996), 435–449.

[9] George Christakos. 1984. On the problem of permissible covariance and variogram models. *Water Resources Research* 20, 2 (1984), 251–265.

[10] Daren B. H. Cline and Jeffrey D. Hart. 1991. Kernel estimation of densities with discontinuities or discontinuous derivatives. *Statistics: A Journal of Theoretical and Applied Statistics* 22, 1 (1991), 69–84.

[11] Noel Cressie. 2015. *Statistics for Spatial Data*. John Wiley & Sons.

[12] G. Dattoli, C. Chiccoli, S. Lorenzutta, G. Maino, M. Richetta, and A. Torre. 1993. Advances on the theory of generalized Bessel functions and applications to multiphoton processes. *Journal of Scientific Computing* 8, 1 (1993), 69–109.

[13] Peter J. Diggle and Paulo J. Ribeiro. 2007. *Model-based Geostatistics*. Springer, New York.

[14] Grégoire Dubois. 1998. Spatial interpolation comparison 97: Foreword and introduction. *Journal of Geographic Information and Decision Analysis* 2, 2 (1998), 1–10.

[15] Randall L. Eubank. 1999. *Nonparametric Regression and Spline Smoothing*. CRC Press.

[16] Gregory E. Fasshauer. 2007. *Meshfree Approximation Methods with MATLAB*. Vol. 6. World Scientific.

[17] Ruben Fernández-Casal. 2021. npsp: Nonparametric Spatial Statistics. *R package version 0.7-8* (2021).

[18] Rubén Fernández-Casal and Mario Francisco-Fernández. 2014. Nonparametric bias-corrected variogram estimation under non-constant trend. *Stochastic Environmental Research and Risk Assessment* 28 (2014), 1247–1259.

[19] Pilar García-Soidán. 2007. Asymptotic normality of the Nadaraya-Watson semi-variogram estimators. *Test* 16, 3 (2007), 479–503.

[20] Pilar H. Garcìa-Soidán, Manuel Febrero-Bande, and Wenceslao González-Manteiga. 2004. Nonparametric kernel estimation of an isotropic variogram. *Journal of Statistical Planning and Inference* 121, 1 (2004), 65–92.

[21] Pilar H. García-Soidán, Wenceslao González-Manteiga, and Manuel Febrero-Bande. 2003. Local linear regression estimation of the variogram. *Statistics & Probability Letters* 64, 2 (2003), 169–179.

[22] Marc G. Genton. 2001. Classes of kernels for machine learning: A statistics perspective. *Journal of Machine Learning Research* 2, Dec (2001), 299–312.

[23] Marc G. Genton and David J. Gorsich. 2002. Nonparametric variogram and covariogram estimation with Fourier-Bessel matrices. *Computational Statistics & Data Analysis* 41, 1 (2002), 47–57.

[24] David J. Gorsich and Marc G. Genton. 2000. Variogram model selection via nonparametric derivative estimation. *Mathematical Geology* 32, 3 (2000), 249–270.

[25] David J. Gorsich and Marc G. Genton. 2004. On the discretization of nonparametric isotropic covariogram estimators. *Statistics and Computing* 14 (2004), 99–108.

[26] Piet Groeneboom and Geurt Jongbloed. 2014. *Nonparametric Estimation under Shape Constraints*. Vol. 38. Cambridge University Press.

[27] Adityanand Guntuboyina and Bodhisattva Sen. 2018. Nonparametric shape-restricted regression. *Statist. Sci.* 33, 4 (2018), 568–594.

[28] Peter Hall, Nicholas I. Fisher, and Branka Hoffmann. 1994. On the nonparametric estimation of covariance functions. *The Annals of Statistics* 22, 4 (1994), 2115–2134.

[29] Wolfgang Härdle. 1990. *Applied Nonparametric Regression*. Number 19. Cambridge University Press.

[30] Tristen Hayfield and Jeffrey S. Racine. 2008. Nonparametric econometrics: The np package. *Journal of Statistical Software* 27 (2008), 1–32.

[31] André G. Journel and Charles J. Huijbregts. 1978. *Mining Geostatistics*. Academic Press, London.

[32] Rhoana J. Karunamuni and Tom Alberts. 2005. On boundary correction in kernel density estimation. *Statistical Methodology* 2, 3 (2005), 191–212.

[33] Solomon Kullback and Richard A. Leibler. 1951. On information and sufficiency. *The Annals of Mathematical Statistics* 22, 1 (1951), 79–86.

[34] Pedro Larranaga. 2002. A review on estimation of distribution algorithms. In *Estimation of Distribution Algorithms*. Springer, 57–100.

[35] Georges Matheron. 1962. *Traité de Géostatistique Appliquée, Tome I: Mémoires du Bureau de Recherches Géologiques et Minières*. Number 14. Editions Technip, Paris.

[36] Raquel Menezes, Pilar Garcia-Soidán, and Manuel Febrero-Bande. 2005. A comparison of approaches for valid variogram achievement. *Computational Statistics* 20, 4 (2005), 623–642.

[37] Henning Omre. 1984. The variogram and its estimation. *Geostatistics for Natural Resources Characterization: Part 1* (1984), 107–125.

[38] Reza Rastegar and Mohammad R. Meybodi. 2005. A study on the global convergence time complexity of estimation of distribution algorithms. In *Proceedings of the 10th International Conference on Rough Sets, Fuzzy Sets, Data Mining, and Granular Computing-Volume Part I*. Springer, 441–450.

[39] Cavan Reilly and Andrew Gelman. 2007. Weighted classical variogram estimation for data with clustering. *Technometrics* 49, 2 (2007), 184–194.

[40] Isaac J Schoenberg. 1938. Metric spaces and completely monotone functions. *Annals of Mathematics* 39, 4 (1938), 811–841.

[41] Eugene F. Schuster. 1985. Incorporating support constraints into nonparametric estimators of densities. *Communications in Statistics-Theory and methods* 14, 5 (1985), 1123–1136.

[42] Alexander Shapiro and Johan D. Botha. 1991. Variogram fitting with a general class of conditionally nonnegative definite functions. *Computational Statistics & Data Analysis* 11, 1 (1991), 87–96.

[43] Bernard W. Silverman. 1986. *Density Estimation for Statistics and Data Analysis*. Vol. 26. CRC Press.

[44] Matt P. Wand and M. Chris Jones. 1994. *Kernel Smoothing*. CRC Press.

[45] Richard E. Williamson. 1955. *On Multiply Monotone Functions and their Laplace Transforms*. Mathematics Division, Office of Scientific Research, US Air Force.

[46] Qingfu Zhang and Heinz Muhlenbein. 2004. On the convergence of a class of estimation of distribution algorithms. *IEEE Transactions on Evolutionary Computation* 8, 2 (2004), 127–136.

[47] Shunpu Zhang, Rohana J. Karunamuni, and M. Chris Jones. 1999. An improved estimator of the density function at the boundary. *J. Amer. Statist. Assoc.* 94, 448 (1999), 1231–1240.