STAPRO: 8208 Model 3G pp. 1–12 (col. fig: NIL)

ARTICLE IN PRESS

Statistics and Probability Letters xx (xxxx) xxx–xxx

Contents lists available at ScienceDirect

Statistics and Probability Letters

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

Nonparametric recursive method for kernel-type function estimators for spatial data

Salim Bouzebda [a,*], Yousri Slaoui [b]

[a] *LMAC, Université de Technologie de Compiégne, France*
[b] *Univ. Poitiers, Lab. Math. et Appl., Futuroscope Chasseneuil, France*

ARTICLE INFO

Article history:
Received 5 September 2017
Received in revised form 18 February 2018
Accepted 26 March 2018
Available online xxxx

Keywords:
Bandwidth selection
Regression estimation
Spatial data
Stochastic approximation algorithm

ABSTRACT

In the present paper we propose recursive general kernel-type estimators for spatial data defined by the stochastic approximation algorithm. We obtain the central limit theorem and strong pointwise convergence rate for the nonparametric recursive general kernel-type estimators under some mild conditions. Finally, we investigate the MISE of the proposed estimators and provide the optimal bandwidth.

1. Introduction

Over years ago, Parzen (1962) studied some properties of kernel density estimators introduced by Akaike (1954) and Rosenblatt (1956). Nonparametric density and regression function estimation has been the subject of intense investigation by both statisticians and probabilists for many years and this has led to the development of a large variety of methods. Kernel nonparametric function estimation methods have long attracted a great deal of attention, for good sources of references to research literature in this area along with statistical applications consult Tapia and Thompson (1978), Wertz (1978), Devroye and Györfi (1985), Devroye (1987), Silverman (1986), Nadaraya (1989), Härdle (1990), Scott (1992), Wand and Jones (1995), Eggermont and LaRiccia (2001) and Devroye and Lugosi (2001) and the references therein. There are basically no restrictions on the choice of the kernel $K(\cdot)$ in our setup, apart from satisfying classical conditions. The selection of the bandwidth, however, is more problematic. The choice of the bandwidth is crucial to obtain a good rate of consistency for of the kernel-type estimators. It has a big influence on the size of the bias. One has to find an appropriate bandwidth that produces an estimator which has a good balance between the bias and the variance of the kernel-type estimator, for more discussion refer to Mason (2012). It is worth noticing that the bandwidth selection methods studied in the literature can be divided into three broad classes: the cross-validation techniques, the plug-in ideas and the bootstrap. Recently, some general methods based upon empirical process techniques are developed in order to prove uniform in bandwidth consistency of a class of kernel-type function estimators (density, regression, entropy and copula), we may refer to Einmahl and Mason (2000, 2005), Bouzebda and Elhattab (2009, 2011), Bouzebda (2012) and Bouzebda et al. (2018). Further, plug-in bandwidth selection method for recursive kernel density estimators defined by stochastic approximation method have been done by Slaoui (2014a) and for recursive kernel distribution estimators have been done by Slaoui (2014b).

This work concerns a nonparametric estimation of the recursive general kernel-type estimators for spatial data defined by the stochastic approximation algorithm. To the best of our knowledge, the results presented here, respond to a problem that has not been studied systematically up to the present, which was the basic motivation of the paper.

* Corresponding author.
E-mail addresses: salim.bouzebda@utc.fr (S. Bouzebda), Yousri.Slaoui@math.univ-poitiers.fr (Y. Slaoui).

https://doi.org/10.1016/j.spl.2018.03.017

We start by giving some notation and definitions that are needed for the forthcoming sections. We consider a spatial process $(\mathbf{Z_i} = (\mathbf{X_i}, Y_\mathbf{i}) \in \mathbb{R}^d \times \mathbb{R} : \mathbf{i} \in \mathbb{Z}^N)$ defined over some probability space $(\Omega, \mathcal{F}, \mathbb{P})$ with same distribution as $(\mathbf{X}, Y)$ having unknown density $g_{\mathbf{X},Y}(\cdot)$ on $\mathbb{R}^{d+1}$. The density function of $\mathbf{X}$ on $\mathbb{R}^d$ is $g_\mathbf{X}(\cdot)$. In this paper, we are interested in the following regression model

$$Y_\mathbf{i} = r(\mathbf{X_i}) + \varepsilon_\mathbf{i},$$

where $r(\mathbf{x}) = \mathbb{E}(Y|\mathbf{X} = \mathbf{x})$ whenever it exists, is an unknown function, with real values. The process is observed over the spatial set of sites $\mathcal{I}_n = \{\mathbf{i} = (i_1, \ldots, i_N), 1 \le i_k \le n_k, k = 1, \ldots, N\}$, which is a finite subset of a potentially observable region $\mathcal{S} \subset \mathbb{R}^N$. We denote by $(s_1, \ldots, s_n)$ the localized sites in $\mathcal{S}$ and we denote $\mathbf{n} = (n_1, \ldots, n_N)$; let $\widehat{\mathbf{n}} := n_1 \times \cdots \times n_N$ be the sample size. From now on, we assume for simplicity that $n_1 = n_2 = \cdots = n_N = n$. We let $\Pi_\mathbf{j} = \prod_{\mathbf{i}\in\mathcal{I}_\mathbf{j}} (1 - \gamma_\mathbf{i})$, for $\mathbf{j} \in \{\mathbf{1}, \ldots, \mathbf{n}\}$ and we will study the following process

$$\widehat{\Psi}_{n,h_n}(\mathbf{x}, f, K) = \Pi_\mathbf{n} \sum_{\mathbf{i}\in\mathcal{I}_n} \Pi_\mathbf{i}^{-1} \gamma_\mathbf{i} h_\mathbf{i}^{-d} \left\{(c_f(\mathbf{x})f(Y_\mathbf{i}) + d_f(\mathbf{x}))K(h_\mathbf{i}^{-1}(\mathbf{x} - \mathbf{X_i}))\right\}, \tag{1.1}$$

where $(\gamma_\mathbf{n})$ is a nonrandom positive sequence tending to zero as $\widehat{\mathbf{n}} \to \infty$, $(h_\mathbf{n})$ is a nonrandom positive sequence tending to zero as $\widehat{\mathbf{n}} \to \infty$, called bandwidth. For convenience, we treat the observations sites as an array that is $\mathcal{I}_n = \left\{s_j, j = 1, \ldots, n\right\}$. By enumerating the sites, we let $\Pi_j = \prod_{i=1}^{j} \left(1 - \gamma_{s_i}\right)$, for $j \in \{1, \ldots, n\}$, one may rewrite $\widehat{\Psi}_{n,h_n}(\mathbf{x}, f, K)$ as

$$\widehat{\Psi}_{n,h_n}(\mathbf{x}, f, K) = \Pi_n \sum_{j=1}^{n} \Pi_j^{-1} \gamma_{s_j} h_{s_j}^{-d} \left\{(c_f(\mathbf{x})f(Y_{s_j}) + d_f(\mathbf{x}))K\left(h_{s_j}^{-1}\left(\mathbf{x} - \mathbf{X}_{s_j}\right)\right)\right\}. \tag{1.2}$$

Noting that, the proposed estimators can be written recursively as follows:

$$\begin{aligned}\widehat{\Psi}_{n,h_n}(\mathbf{x}, f, K) = {} & \left(1 - \gamma_{s_n}\right) \widehat{\Psi}_{n-1,h_{n-1}}(\mathbf{x}, f, K) \\ & + \gamma_{s_n} h_{s_n}^{-d} \left\{(c_f(\mathbf{x})f(Y_{s_n}) + d_f(\mathbf{x}))K\left(h_{s_n}^{-1}\left(\mathbf{x} - \mathbf{X}_{s_n}\right)\right)\right\}.\end{aligned} \tag{1.3}$$

This recursive property is particularly useful when the number of the spatial sites increases on space since $\widehat{\Psi}_{n,h_n}(\mathbf{x}, f, K)$ can be easily updated with each additional observation. In fact, if X_{s_n} is a new observation of the process at a site s_n added to $\mathcal{I}_{n-1}$, the estimators $\widehat{\Psi}_{n,h_n}(\mathbf{x}, f, K)$ can be updated recursively by the relation (1.3). From a practical point of view, this arrangement provides important savings in computational time and storage memory which is a consequence of the fact that the estimate updating is independent of the history of the data. The main drawback of the classical kernel estimator is the use of all data at each step of estimation. From a theoretical point of view, the main advantage of the investigation of such processes is that we can prove almost sure consistency with exact rate for several kernel-type estimators simultaneously. It is worth noting that the quantity $\widehat{\Psi}_{n,h_n}(\mathbf{x}, f, K)$ includes as particular cases: the kernel type density estimator, the Nadaraya Watson estimator and the kernel type estimator of the conditional distribution, we may refer to Einmahl and Mason (2000, 2005) for more details. In this sense, the present paper extends, in non trivial way, some previous results by considering general kernel-type estimators given in (1.2).

The remainder of this paper is organized as follows. In Section 2 we give the assumption and the main results. More precisely, we provide the bias and the asymptotic variance. We establish the asymptotic normality of $\widehat{\Psi}_{n,h_n}(\mathbf{x}, f, K)$ in Theorem 1. Finally we obtain the consistency with exact rate in Theorem 2. We calculate the MISE and provide the optimal bandwidth. Some concluding remarks and possible future developments are mentioned in Section 3. To avoid interrupting the flow of the presentation, all mathematical developments are relegated to Section 4.

2. Assumptions and main results

We define the following class of regularly varying sequences.

Definition 1. Let $\gamma \in \mathbb{R}$ and $\left(v_{s_n}\right)_{n\ge1}$ be a nonrandom positive sequence. We say that $\left(v_{s_n}\right) \in \mathcal{GS}(\gamma)$ if

$$\lim_{n\to+\infty} n \left[1 - \frac{v_{s_{n-1}}}{v_{s_n}}\right] = \gamma. \tag{2.1}$$

Condition (2.1) was introduced by Galambos and Seneta (1973) to define regularly varying sequences (see also Bojanic and Seneta (1973)). Note that the acronym $\mathcal{GS}$ stands for (Galambos and Seneta). Typical sequences in $\mathcal{GS}(\gamma)$ are, for $b \in \mathbb{R}$, $n^\gamma (\log n)^b$, $n^\gamma (\log\log n)^b$, and so on.

In this section, we investigate asymptotic properties of the proposed estimators (1.2). The assumptions to which we shall refer are the following:

(A1) $K : \mathbb{R}^d \to \mathbb{R}$ is a continuous, bounded function satisfying $\int_{\mathbb{R}^d} K(\mathbf{z})\, d\mathbf{z} = 1$, and, for all $j \in \{1, \ldots, d\}$, $\int_{\mathbb{R}^d} z_j K(\mathbf{z})\, d\mathbf{z} = 0$ and $\int_{\mathbb{R}^d} z_j^2 \|K(\mathbf{z})\|\, d\mathbf{z} < \infty$.