Statistics and Probability Letters 106 (2015) 247–255

Contents lists available at ScienceDirect

Statistics and Probability Letters

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

Estimation of a jump point in random design regression

Michael Kohler[a], Adam Krzyżak[b,*]

[a] Fachbereich Mathematik, Technische Universität Darmstadt, Schlossgartenstr. 7, 64289 Darmstadt, Germany

[b] Department of Computer Science and Software Engineering, Concordia University, 1455 De Maisonneuve Blvd. West, Montreal, Quebec, Canada H3G 1M8

ARTICLE INFO

Article history:
Received 28 April 2015
Received in revised form 5 July 2015
Accepted 6 July 2015
Available online 23 July 2015

MSC:
primary 62G08
secondary 62G20

Keywords:
Jump point
Nonparametric regression
Strong consistency

ABSTRACT

Given an i.i.d. sample of an $\mathbb{R} \times \mathbb{R}$-valued random vector (X, Y), we estimate the location and the size of the maximal jump of the piecewise continuous regression function $m(x) = \mathbf{E}\{Y|X = x\}$. The proposed estimates are shown to converge almost surely to the maximal jump point under weak conditions.

1. Introduction

Let $(X, Y), (X_1, Y_1), (X_2, Y_2) \ldots$ be independent and identically distributed random variables with values in $\mathbb{R} \times \mathbb{R}$. Assume $\mathbf{E}|Y| < \infty$, let $m(x) = \mathbf{E}\{Y|X = x\}$ be the so-called regression function, and let $\mu = \mathbf{P}_X$ be the distribution of the design variable X. Assume that m is uniformly continuous except for finitely many jump points, i.e., assume that there exist $N \in \mathbb{N}, z_1, \ldots, z_N \in \mathbb{R}$ and $L : \mathbb{R}_+ \to \mathbb{R}_+$ such that $L(h) \to 0$ $(h \to 0)$ and for all $x, y \in \mathbb{R},\ x < y$ with the property that $[x, y]$ does not contain any of the $z_1, \ldots, z_N$ we have

$$|m(x) - m(y)| \leq L(|x - y|).$$

In this paper we consider the problem of estimating the location and the size of the maximal jump of m. More precisely, let

$$m^+(x) = \lim_{h\to 0, h>0} m(x + h) \quad \text{and} \quad m^-(x) = \lim_{h\to 0, h>0} m(x - h)$$

be the right-hand and left-hand limits of m. Then

$$\Delta(z) = |m^+(z) - m^-(z)|$$

is the size of the jump of m at z. Let $[a, b]$ be the support of X, which we assume to be a compact interval, and denote by z^* the location of the jump with the maximal size within (a, b), i.e.,

$$\Delta := \Delta(z^*) = \sup_{z\in(a,b)} \Delta(z). \tag{1}$$

[*] Corresponding author. Tel.: +1 514 848 2424x3007; fax: +1 514 848 2830.
E-mail addresses: kohler@mathematik.tu-darmstadt.de (M. Kohler), krzyzak@cs.concordia.ca (A. Krzyżak).

http://dx.doi.org/10.1016/j.spl.2015.07.009
0167-7152/© 2015 Elsevier B.V. All rights reserved.

Given the data

$$\mathcal{D}_n = \{(X_1, Y_1), \ldots, (X_n, Y_n)\}$$

we want to construct estimates

$$\hat{\Delta}_n = \hat{\Delta}_n(\mathcal{D}_n) \quad \text{and} \quad \hat{z}_n = \hat{z}_n(\mathcal{D}_n)$$

such that

$$\hat{\Delta}_n \to \Delta \quad \text{a.s.}$$

and

$$\hat{z}_n \to z^* \quad \text{a.s.}$$

as $n \to \infty$.

Of course, the last convergence will be only possible in case that z^* is unique.

The most popular estimates for nonparametric regression include kernel regression estimate (cf., e.g., Watson, 1964, Stone, 1977 or Devroye and Krzyżak, 1989), partitioning regression estimate (cf., e.g., Beirlant and Györfi, 1998), nearest neighbor regression estimate (cf., e.g., Devroye et al., 1994, or Zhao, 1987), local polynomial kernel estimates (cf., e.g., Stone, 1982), least squares estimates (cf., e.g., Lugosi and Zeger, 1995) or smoothing spline estimates (cf., e.g., Kohler and Krzyżak, 2001). The main theoretical results are summarized in the monograph by Györfi et al. (2002). Modifications of several of these estimates have already been applied to jump point regression in a random design setting in various papers. E.g., rate of convergence results have been derived in Gijbels et al. (1999) and Ma and Yang (2011), a data-driven choice of the bandwidth of kernel based jump point estimators has been investigated in Gijbels and Goderniaux (2004a), and jump points of the derivative of a regression function have been estimated in Gijbels and Goderniaux (2004b). But most papers for jump point regression derive results in the fixed design regression setting, see, e.g., Desmet and Gijbels (2011), Gijbels et al. (2007), Jose and Ismail (2001) or Wu and Chu (1993) and the literature cited therein. Related techniques are also applied in change point estimation in connection with time series, see, e.g., Carlstein (1988), Hariz et al. (2007), Lee (2011) or Rafajłowicz et al. (2010).

In this paper we consider a standard kernel estimate of $\Delta(z)$, and the aim is to derive consistency results for this estimate under much more general conditions than usually considered in the literature, in particular we avoid any assumptions stipulating that the distribution of the design has a density with respect to the Lebesgue–Borel measure. If we want to avoid this assumption, we could use techniques from Devroye (1978a,b) and try to construct estimates of m^+ and m^- which are consistent in the supremum norm whenever the distribution of X has the property that the probability of an interval is always greater than or equal to a constant times the length of the interval. However, in this paper we want to avoid even such an assumption. The key trick which allows us to derive consistency results for the estimates under even weaker conditions is that we use a data-dependent modification of the bandwidth of the kernel estimates: we start with some fixed value depending on the sample size and increase it until the intervals to the left and to the right of the point considered contain enough data points. We show consistency of our method under rather weak conditions: we assume that the regression function is uniformly continuous except for finitely many jump points, that the support of X is a compact interval and that Y satisfies some rather weak integrability condition. We prove that our estimates are strongly consistent in a sense that the estimates of the maximal jump size and of the jump point converge almost surely to the true values provided the sample size goes to infinity.

1.1. Notation

Throughout this paper we use the following notations: μ denotes the distribution of X and $m(x) = \mathbf{E}\{Y|X = x\}$ is the regression function of (X, Y).

Let $D \subseteq \mathbb{R}$ and let $f : \mathbb{R} \to \mathbb{R}$ be a real-valued function defined on $\mathbb{R}$. We write $x = \arg\max_{z\in D} f(z)$ if $\max_{z\in\mathcal{D}} f(z)$ exists and if x satisfies

$$x \in D \quad \text{and} \quad f(x) = \max_{z\in\mathcal{D}} f(z).$$

For $A \subseteq \mathbb{R}$ let I_A be the indicator function of A, i.e.,

$$I_A(x) := \begin{cases} 1 & \text{if } x \in A, \\ 0 & \text{if } x \notin A, \end{cases}$$

for $x \in \mathbb{R}$. Furthermore we define

$$\log_+ z := \begin{cases} \log(z) & \text{if } z \geq 1, \\ 0 & \text{if } z < 1, \end{cases}$$

for $z \in \mathbb{R}_+$.

1.2. Outline

The definition of the estimates are given in Section 2, the main result is formulated in Section 3 and proven in Section 4.