On the Nadaraya-Watson estimator for irregularly spaced spatial data

Mohamed EL MACHKOURI and Lucas REDING
Laboratoire de Mathématiques Raphaël Salem
UMR-6085, CNRS-Université de Rouen Normandie, France
Xiequan FAN
Center for Applied Mathematics
Tianjin University, China
lucas.reding@etu.univ-rouen.fr

1. Introduction

In many situations, practicians want to know the relationship between some predictors and a response. In general, a nonparametric approach is necessary. The Nadaraya-Watson estimator is a well-know tool in nonparametric estimation and the study of its asymptotic properties for dependent spatial data (random field indexed by a lattice) is still investigated nowadays. In this work, we study the asymptotic normality of the Nadaraya-Watson estimator for strongly mixing random fields in the sense of Rosenblatt (1956) and weakly dependent random fields in the sense of Wu (2005). We lay emphasis on the fact that our results require minimal conditions on the bandwidth parameter.

2. Definitions

Let $(X_i)_{i\in\mathbb{Z}^d}$ be a stationary real random field such that the distribution of X_0 is absolutely continuous with respect to Lebesgue measure on $\mathbb{R}$.
Let Λ be a finite subset of $\mathbb{Z}^d$. We consider the following regression model $Y_i = R(X_i, \eta_i)$ for $i \in \Lambda$ where R is an unknown functional and $(\eta_i)_{i\in\Lambda}$ are i.i.d. real random variables with zero-mean and finite variance which are independent of the predictors $(X_i)_{i\in\mathbb{Z}^d}$.

The goal is to approximate the regression function r defined for all $x \in \mathbb{R}$ such that $f(x) > 0$ by $r(x) = \mathbb{E}\left[R(X_0, \eta_0)\,|X_0 = x\right] = \mathbb{E}\left[R(x, \eta_0)\right]$.

So, we introduce **the Nadaraya-Watson kernel regression estimator** defined for all $x \in \mathbb{R}$ and for all $n \in \mathbb{N}$ by

$$r_n(x) = \frac{\sum_{i\in\Lambda_n} Y_i K\left(\frac{x-X_i}{b_n}\right)}{\sum_{i\in\Lambda_n} K\left(\frac{x-X_i}{b_n}\right)},$$

where K is a probability density function (kernel), Λ_n is a finite subset of $\mathbb{Z}^d$ and $b_n \in \mathbb{R}_+^*$ (bandwidth parameter) goes to zero as n goes to infinity.

We consider the following cases:

Strong mixing in the sense of Rosenblatt (1956)

Let $\mathcal{F}$ and $\mathcal{G}$ be two σ-algebra, we define the coefficient

$$\alpha(\mathcal{F}, \mathcal{G}) = \sup\{|\mathbb{P}(A \cap B) - \mathbb{P}(A)\mathbb{P}(B)|, A \in \mathcal{F}, B \in \mathcal{G}\}.$$

Let $p \in \mathbb{N}^* \cup \{+\infty\}$, we set

$$\alpha_{1,p}(n) = \sup\{\alpha(\sigma(X_k), \mathcal{F}_\Gamma), k \in \mathbb{Z}^d, \Gamma \subset \mathbb{Z}^d, |\Gamma| \leq p, \rho(\Gamma, \{k\}) \geq n\},$$

where $\mathcal{F}_\Gamma = \sigma(X_i, i \in \Gamma)$, $|\Gamma|$ is the cardinal of the set Γ and ρ is a metric.

The random field $(X_i)_{i\in\mathbb{Z}^d}$ is said to be strongly mixing (or α-mixing) if $\lim_{n\to+\infty} \alpha_{1,p}(n) = 0$.

Weak mixing in the sense of Wu (2005)

Assume that $X_i = g(\epsilon_{i-s}, s \in \mathbb{Z}^d)$ for all $i \in \mathbb{Z}^d$ where $g : \mathbb{R}^{\mathbb{Z}^d} \to \mathbb{R}$ is a measurable function and $(\epsilon_i)_{i\in\mathbb{Z}^d}$ are i.i.d. random variables.

If $(\epsilon_i')_{i\in\mathbb{Z}^d}$ is an independent copy of $(\epsilon_i)_{i\in\mathbb{Z}^d}$ and

$$X_i^* = g(\epsilon_{i-s}^*, s \in \mathbb{Z}^d)$$

where $\epsilon_j^* = \epsilon_j$ if $j \neq 0$ and $\epsilon_0^* = \epsilon_0'$ then we define the so-called physical dependency measure coefficient $\delta_i = ||X_i - X_i^*||_2$.

The random field $(X_i)_{i\in\mathbb{Z}^d}$ is said to be stable if $\sum_{i\in\mathbb{Z}^d} \delta_i < \infty$.

3. Main result

Let us consider the following assumptions:

(H_1) $b_n \to 0$ and $|\Lambda_n| b_n \to \infty$ such that $|\Lambda_n| b_n^3 \to 0$ when n goes to infinity.

(H_2) There exists $\theta > 0$ such that $\mathbb{E}\left[|Y_0|^{2+\theta}\right] < \infty$ and one of the following conditions $\sum_{n\geq 1} n^{\frac{(2d-1)\theta+6d-2}{2+\theta}} \alpha_{1,\infty}^{\frac{\theta}{2+\theta}}(n) < \infty$ or $\sum_{i\in\mathbb{Z}^d} |i|^{\frac{d(5\theta^2+18\theta+8)}{2\theta(2+\theta)}} \delta_i^{\frac{\theta}{2+\theta}}(n) < \infty$ is satisfied.

The result established in [2] is the following central limit theorem.

Theorem. *Under (H_1), (H_2) and some mild technical assumptions, for all $x \in \mathbb{R}$ such that $f(x) > 0$, we have*

$$\sqrt{|\Lambda_n| b_n}\,(r_n(x) - r(x)) \underset{n\to\infty}{\overset{\mathcal{D}}{\to}} \mathcal{N}(0, \sigma^2(x)),$$

where $\sigma^2(x) = \frac{\mathbb{E}[Y_0^2|X_0=x] - r^2(x)}{f(x)} \int_{\mathbb{R}} K^2(t)dt$.

The above theorem improves a previous result by Biau and Cadre ([1], Theorem 2.2) for strongly mixing random fields in which stronger hypothesis are required on both the bandwidth parameter and on the strong mixing coefficients. On the other hand, to the best of our knowledge, our result is the first central limit theorem for the Nadaraya-Watson estimator built from dependent spatial data under minimal bandwidth conditions.

References

[1] G. Biau and B. Cadre. Nonparametric spatial prediction. *Statistical Inference for Stochastic Processes*, 7(3) :327–349, 2004.

[2] M. El Machkouri, X. Fan, and L. Reding. On the Nadaraya-Watson kernel regression estimator for irregularly spaced spatial data. 2018. Submitted for publication.