Strong consistency of a kernel-based rule for spatially dependent data

Ahmad Younso
Department of Mathematical Statistics, Faculty of Sciences, Damascus University, Syrian Arab Republic, and
Ziad Kanaya and Nour Azhari
Department of Mathematics, Faculty of Sciences, Tishreen University, Syrian Arab Republic

Received 1 September 2018
Revised 27 October 2019
Accepted 28 October 2019

Abstract

We consider the kernel-based classifier proposed by Younso (2017). This nonparametric classifier allows for the classification of missing spatially dependent data. The weak consistency of the classifier has been studied by Younso (2017). The purpose of this paper is to establish strong consistency of this classifier under mild conditions. The classifier is discussed in a multi-class case. The results are illustrated with simulation studies and real applications.

Keywords Bayes rule, Kernel rule, Random field, Bandwidth, Strong consistency
Paper type Original Article

1. Introduction

In many applications one needs to classify spatial data that have been collected incompletely. The classification of incomplete-data problem, in which certain features are missing from particular feature vectors, exists in a wide range of fields, including image labeling, computer vision and others. For example, in the remote sensing technology, because of the internal malfunction of satellite sensors and poor atmospheric conditions such as thick cloud, the acquired remote sensing images often suffer from missing information at certain pixels and one wants to classify these pixels using the information in the nearest identified pixels. Many existing classification algorithms assume either certain parametric distributions for the data or certain forms of separating curves or surfaces. These parametric classifiers are suboptimal

The author would like to thank the anonymous referees whose valuable comments led to an improved version of the paper.

The publisher wishes to inform readers that the article "Strong consistency of a kernel-based rule for spatially dependent data" was originally published by the previous publisher of the *Arab Journal of Mathematical Sciences* and the pagination of this article has been subsequently changed. There has been no change to the content of the article. This change was necessary for the journal to transition from the previous publisher to the new one. The publisher sincerely apologises for any inconvenience caused. To access and cite this article, please use "Younso, A., Kanaya, Z., Azhari, N. (2019), "Strong consistency of a kernel-based rule for spatially dependent data", *Arab Journal of Mathematical Sciences*, Vol. 26 No. 1/2, pp. 211-225. The original publication date for this paper was 13/11/2019.

Arab Journal of Mathematical Sciences
Vol. 26 No. 1/2, 2020
pp. 211-225
Emerald Publishing Limited
e-ISSN: 2588-9214
p-ISSN: 1319-5166
DOI 10.1016/j.ajmsc.2019.10.004

and of limited use in practical applications where little information about the underlying distributions is available a priori. In comparison, nonparametric classifiers are usually more flexible in accommodating different data structures, and are hence more desirable. [21] has proposed a nonparametric approach allowing to include contextual features for classifying missing spatial data and has investigated the consistency of the classifier under mild conditions. In nonparametric spatial estimation, the existing works concern mainly the estimation of a probability density and regression functions, see the key references: [2–4,15] and [14]. More recently, [5] has proposed a kernel spatial density estimator allowing for the analysis of spatial clustering. In this work, we establish strong consistency of the classifier proposed by [21] and then, we check its performance with simulation studies and applications. We consider a strictly stationary random field $\{(X_{\mathbf{i}}, Y_{\mathbf{i}})\}_{\mathbf{i}\in\mathbb{Z}^N}$ defined on some probability space $(\Omega, \mathcal{F}, \mathbb{P})$ and taking values in $\mathbb{R}^d \times \{0, \ldots, M\}$, for some integer $M \geq 1$. In the problem of classification, for each $\mathbf{i} \in \mathbb{Z}^N$, $X_{\mathbf{i}}$ is a vector of features and $Y_{\mathbf{i}}$ is the label (class) of $X_{\mathbf{i}}$. A point $\mathbf{i} = (i_1, \ldots, i_N) \in \mathbb{Z}^N$ will be referred to as a site. For $\mathbf{n} = (n_1, \ldots, n_N) \in (\mathbb{N}^*)^N$, we define the rectangular region $\mathcal{I}_{\mathbf{n}}$ by $\mathcal{I}_{\mathbf{n}} = \{\mathbf{i} \in \mathbb{Z}^N : 1 \leq i_k \leq n_k, \forall k = 1, \ldots, N\}$. We will write $\mathbf{n} \to \infty$ if $\min_{k=1,\ldots,N} n_k \to \infty$. Define $\widehat{\mathbf{n}} = n_1 \times \cdots \times n_N = \text{card}(\mathcal{I}_{\mathbf{n}})$ and assume that the random field is observed on a subset $\mathcal{S}_{\mathbf{n}} \subset \mathcal{I}_{\mathbf{n}}$ with $\mathcal{I}_{\mathbf{n}} - \mathcal{S}_{\mathbf{n}}$ is a bounded set for $\widehat{\mathbf{n}}$ large enough. When processing a particular site, its features are not used at all, but only the features of its neighbors will be considered. In other words, we wish to predict the label $Y_{\mathbf{j}}$ of a new site $\mathbf{j}$ based only on observations in a vicinity, say $\nu_{\mathbf{j}} \subset \mathcal{S}_{\mathbf{n}}$, where the set $\nu_{\mathbf{j}}$ is not containing $\mathbf{j}$. Let $\nu_{\mathbf{j}} = \mathbf{j} + \nu$, where $\nu \subset \mathbb{Z}^N$ is a fixed bounded set of sites not containing $\mathbf{0}$ with $\text{card}(\nu) = l$ (l is also the cardinal of each $\nu_{\mathbf{j}}$). We assume that $X_{(\mathbf{j})} = \{X_{\mathbf{i}} : \mathbf{i} \in \nu_{\mathbf{j}}\}$ is a random vector taking values in $\mathbb{R}^{\tilde{d}}$ with $\tilde{d} = ld$, and that the components of $X_{(\mathbf{j})}$ are ordered according to an arbitrary order on indices, for example the lexicographic order. The pair $(X_{(\mathbf{j})}, Y_{\mathbf{j}})$ may be completely described by μ, the probability measure for $X_{(\mathbf{j})}$, and $\eta(x)$, the regression of $Y_{\mathbf{j}}$ on $X_{(\mathbf{j})} = x$. Assume that for each $\mathbf{i} \in \mathbb{Z}^N$, $(X_{(\mathbf{i})}, Y_{\mathbf{i}})$ has the same distribution as the pair $(X_{(\mathbf{1})}, Y_{\mathbf{1}})$. We will create a classifier $g : \mathbb{R}^{\tilde{d}} \to \{0, \ldots, M\}$ mapping $X_{(\mathbf{j})}$ into the predicted label of $X_{\mathbf{j}}$. The error rate, or risk, of a rule g is $L(g) = \mathbb{P}\{g(X_{(\mathbf{j})}) \neq Y_{\mathbf{j}}\}$. This is minimized by the rule

$$g^*(x) = \arg\max_{0\leq k\leq M} \mathbb{P}(Y_{\mathbf{j}} = k | X_{(\mathbf{j})} = x), \tag{1.1}$$

whose error rate $L^* = L(g^*)$ is called the Bayes-optimal risk and $g^*(x)$ is called the Bayes rule. Clearly, $g^*(x)$ predicts the label $Y_{\mathbf{j}}$ of the site $\mathbf{j}$ using only x, the value of $X_{(\mathbf{j})}$, while the features vector $X_{\mathbf{j}}$ does not affect the classification procedure at all. This means that $g^*(x)$ well work event if $X_{\mathbf{j}}$ is completely missing. Unfortunately, we cannot use (1.1) directly because it depends on the distribution of $(X_{(\mathbf{j})}, Y_{\mathbf{j}})$ which is generally unknown. So, we take $\mathcal{J}_{\mathbf{n}} = \{\mathbf{i} \in \mathcal{S}_{\mathbf{n}} : \nu_{\mathbf{i}} \subset \mathcal{S}_{\mathbf{n}}\}$ and we use the training data $D_{\mathbf{n}} = \{(X_{\mathbf{i}}, Y_{\mathbf{i}}) : \mathbf{i} \in \mathcal{J}_{\mathbf{n}}\}$ to construct a classifier $g_{\mathbf{n}}(x)$. We consider the classifier $g_{\mathbf{n}}(x)$ obtained by extending the classifier of [21] to the multi-class case as follows:

$$g_{\mathbf{n}}(x) = \arg\max_{0\leq k\leq M} \sum_{\mathbf{i}\in\mathcal{J}_{\mathbf{n}}} \mathbb{1}_{\{Y_{\mathbf{i}}=k\}} K\left(\frac{x - X_{(\mathbf{i})}}{b_{\mathbf{n}}}\right). \tag{1.2}$$

where $\mathbb{1}_A$ denotes the indicator of the set A, the kernel $\boldsymbol{K} : \mathbb{R}^{\tilde{d}} \to \mathbb{R}_+$ is a density function on $\mathbb{R}^{\tilde{d}}$, and $b_{\mathbf{n}}$ is a sequence of bandwidths tending to zero as $\mathbf{n}$ tends to infinity. In one hand, the sum in (1.2) is taken over $\mathcal{J}_{\mathbf{n}}$ instead of $\mathcal{S}_{\mathbf{n}}$ just to ensure that $X_{(\mathbf{i})}$ always exists and that the sums make sense. On the other hand, for each new site $\mathbf{j} \notin \mathcal{S}_{\mathbf{n}}$, the classifier $g_{\mathbf{n}}(x)$ predicts the missing label $Y_{\mathbf{j}}$ independently of its features vector $X_{\mathbf{j}}$ which does not belong neither to the training sample $D_{\mathbf{n}}$ nor to the components set of $X_{(\mathbf{j})}$. Consequently, $g_{\mathbf{n}}(x)$ may classify $\mathbf{j}$ even if its own features vector $X_{\mathbf{j}}$ is completely missing and that makes our method exhibit

good performance in comparison with the classical spatial Markovian model. [6] proposes a nonparametric approach to extend the result of [2] to the non-Markovian case by using two kernels in the estimator in order to control both the distance between observations and that between spatial locations without using a specific vicinity for the non-observed site. This latter approach may be developed to classify spatial data but it does not work when one wants to classify sites with missing or incomplete features. Let $L_{\mathbf{n}} = \mathbb{P}\{g_{\mathbf{n}}(X_{(\mathbf{j})}) \neq Y_{\mathbf{j}} | D_{\mathbf{n}}\}$ be the error probability of $g_{\mathbf{n}}(x)$. Generally, we cannot hope to design a classifier that achieve the Bayes error probability L^* but it is possible that the limit behavior of $L_{\mathbf{n}}$ compares favorably to L^*. This idea is encapsulated in the notion of consistency.

Definition 1.1. The classifier $g_{\mathbf{n}}(x)$ is called weakly consistent if

$$\mathbb{E}L_{\mathbf{n}} \to L^* \text{as } \mathbf{n} \to \infty$$

and strongly consistent if

$$L_{\mathbf{n}} \to L^* \text{ as } \mathbf{n} \to \infty \text{ with probability one.}$$

The classifier is called universally (weakly or strongly) consistent if it is (weakly or strongly) consistent for all distribution of $(X_{\mathbf{1}}, Y_{\mathbf{1}})$.

Remark 1.1. Since $L_{\mathbf{n}}$ is bounded, the weak consistency of $L_{\mathbf{n}}$ is equivalent to the convergence of $L_{\mathbf{n}}$ towards L^* in probability which means that strong consistency implies the weak consistency.
In this paper, we investigate the strong consistency of $g_{\mathbf{n}}(x)$ under some mild mixing conditions.

2. Notation and general hypotheses

Let $(\Omega, \mathcal{F}, \mathbb{P})$ be a probability space and let $\mathcal{A}$ and $\mathcal{B}$ be two sub σ-fields of $\mathcal{F}$. The α-mixing coefficient between $\mathcal{A}$ and $\mathcal{B}$ is defined by

$$\alpha = \alpha(\mathcal{A}, \mathcal{B}) = \sup_{A \in \mathcal{A},\, B \in \mathcal{B}} |\mathbb{P}(A \cap B) - \mathbb{P}(A)\mathbb{P}(B)|$$

and the β-mixing coefficient is defined by

$$\beta = \beta(\mathcal{A}, \mathcal{B}) = \mathbb{E}\left\{\sup_{A \in \mathcal{A}} |\mathbb{P}(A|\mathcal{B}) - \mathbb{P}(A)|\right\}.$$

Let $(Z_{\mathbf{i}})_{\mathbf{i} \in \mathbb{Z}^N}$ be a random field on $(\Omega, \mathcal{F}, \mathbb{P})$ and taking values in some space $(\Omega', \mathcal{F}')$.

Definition 2.1. The random field $(Z_{\mathbf{i}})_{\mathbf{i} \in \mathbb{Z}^N}$ is called strongly mixing if there exists $\chi : \mathbb{R} \to \mathbb{R}^+$ with $\chi(t) \searrow 0$ as $t \to \infty$, and for any $E, E' \subset \mathbb{Z}^N$ with finite cardinals,

$$\alpha(\mathcal{B}(E), \mathcal{B}(E')) \leq \chi(\text{dist}(E, E')),$$

where $\text{dist}(E, E')$ denotes the Euclidean distance between E and E'.

The α-mixing condition is one of the most popular mixing conditions. This condition is satisfied by many spatial models. Examples can be found in [17,19] and [11].

Definition 2.2. The random field $(Z_{\mathbf{i}})_{\mathbf{i} \in \mathbb{Z}^N}$ is called β-mixing if there exists $\varphi : \mathbb{R} \to \mathbb{R}^+$ with $\varphi(t) \searrow 0$ as $t \to \infty$, and for any $E, E' \subset \mathbb{Z}^N$ with finite cardinals,

$$\beta(\mathcal{B}(E), \mathcal{B}(E')) \leq \varphi(\text{dist}(E, E')).$$

Linear processes or more generally Markov chains may be β-mixing (see [9]). Similar mixing coefficient is used by [2] to establish some asymptotic properties of the kernel regression

estimator in the spatial case. The two mixing coefficients α and β are related by the inequality $2\alpha \leq \beta$ (see [18]). It means that any β-mixing random field is a strongly mixing one.

Now, we need some regularity assumptions.

Assumption 1. K is a regular kernel, that is, there exist $\delta > 0$ and $c > 0$ such that $c\mathbb{1}_{B(0,\delta)} \leq K(x)$ for all $x \in \mathbb{R}^{\tilde{d}}$ and $\int_{\mathbb{R}^{\tilde{d}}} \text{Sup}_{u \in v+B(0,\delta)} K(u)dv < \infty$, where $B(x,\delta)$ is the closed ball of radius $\delta > 0$ and center at x.

Assumption 2. For each $\mathbf{i}$, $X_{(\mathbf{i})}$ has a density f with respect to Lebesgue measure and for each $\mathbf{i} \neq \mathbf{j}$ with $\nu_{\mathbf{i}} \cap \nu_{\mathbf{j}} = \phi$, $(X_{(\mathbf{i})}, X_{(\mathbf{j})})$ has a density $f_{\mathbf{i},\mathbf{j}}$ such that $\sup_{u,v \in \mathbb{R}^{\tilde{d}}} |f_{\mathbf{i},\mathbf{j}}(u,v) - f(u)f(v)| \leq C$, for some $C > 0$.

Assumption 3. The random field $\{(X_{\mathbf{i}}, Y_{\mathbf{i}})\}_{\mathbf{i} \in \mathbb{Z}^N}$ is β-mixing and there exists $\theta > 0$ such that $\varphi(t) = O(t^{-\theta})$ for all $t \in \mathbb{R}_+^*$.

Assumption 1 is used by [8] and [7] in the i.i.d. case. It may be satisfied if $K(x) = \xi(\|x\|)$ where ξ is a non-negative and decreasing function on $[0, +\infty]$ and $\|.\|$ is the Euclidean norm. Hence, the Gaussian kernel is regular. Assumption 2, used by [21] to prove the weak consistency, is similar to that used by [3]. It is satisfied for example if f and $f_{\mathbf{i},\mathbf{j}}$ are uniformly bounded. Assumption 3 means that the random field is arithmetically β-mixing which implies that it is also strongly mixing with $\alpha(\mathcal{B}(E), \mathcal{B}(E')) \leq \varphi(\text{dist}(E, E'))$ since $2\alpha \leq \beta$.

3. Preliminary lemmas

This section is a collection of technical lemmas which will be used to prove the strong consistency result stated in Theorem 4.1. Let $\|.\|_r$ denote the L_r-norm for any real $r \geq 1$. The following lemma is a direct consequence of the covariance inequality of Ibragimov [12] and the inequality $2\alpha \leq \beta$.

Lemma 3.1. *If r, s and t are strictly positive reals such that $r^{-1} + s^{-1} + t^{-1} = 1$ and Z_1 and Z_2 are two $\mathbb{R}$-valued random variables such that $\|Z_1\|_s < \infty$ and $\|Z_2\|_t < \infty$, then*

$$|\text{cov}(Z_1, Z_2)| \leq 2\{\beta(\sigma(Z_1), \sigma(Z_2))\}^{1/r} \|Z_1\|_s \|Z_2\|_t,$$

where $\sigma(Z_i)$ is the σ-field generated by Z_i for $i = 1, 2$.

For any sub σ-fields $\mathcal{A}$ and $\mathcal{B}$ of $\mathcal{F}$, we denote by $\mathcal{A} \vee \mathcal{B}$ the σ-field generated by $\mathcal{A} \cup \mathcal{B}$. The following coupling lemma of Berbee [1] will be needed to establish the asymptotic results.

Lemma 3.2. *Let Z be a random variable on $(\Omega, \mathcal{F}, \mathbb{P})$ with values in some Polish space Ω' and $\mathcal{M}$ a sub σ-field of $\mathcal{F}$. Assume that there exists a random variable U uniformly distributed over $[0,1]$, independent of $\sigma(Z) \vee \mathcal{M}$. Then, there exists a random variable $\tilde{Z}$ measurable with respect to $\sigma(U) \vee \sigma(Z) \vee \mathcal{M}$, distributed as Z and independent of $\mathcal{M}$, such that*

$$\mathbb{P}(Z \neq \tilde{Z}) = \beta(\mathcal{M}, \sigma(Z)).$$

Remark 3.1. We recall that a Polish space Ω' is a topological space which is separable and completely metrizable (see [13]) and that most of the familiar objects of study in analysis involve Polish spaces. For example, $\mathbb{R}^d$ for each integer $d \geq 1$, is Polish with the usual topology and $\{0, 1, \ldots, n\}$, for all $n \in \mathbb{N}$, is Polish with discrete topology. We also recall that a countable product of Polish spaces is Polish.

The following covering lemma can be found in [8].

Lemma 3.3. *Let K be a regular kernel on $\mathbb{R}^{\tilde{d}}$ and $b_{\mathbf{n}}$ be a sequence of bandwidths. Denote $K_{\mathbf{n}}(x) = b_{\mathbf{n}}^{-d} K(x/b_{\mathbf{n}})$. Then, for any probability measure μ,*

$$\sup_{u\in\mathbb{R}^d}\int_{\mathbb{R}^d}\frac{K_{\mathbf{n}}(x-u)}{\mathbb{E}K_{\mathbf{n}}(x-X_{(1)})}\mu(dx)<\rho,$$

for some $\rho>0$ *dependent only on* K.

The proof of the following lemma is in [4] (see also [21]).

Lemma 3.4. *Let* $\zeta=-N-\epsilon+(1-\gamma)Na^{-1}$ *for some* $0<a<1/2$, *with* γ *and* ϵ *being small positive numbers such that* $a^{-1}-(N+\epsilon)(1-\gamma)^{-1}N^{-1}>1$. *If Assumption 3 holds for some* $\theta>2N$, *then for any* $\delta>0$,

$$\sum_{\|\mathbf{i}\|\geq\delta}\|\mathbf{i}\|^{\zeta}\{\varphi(\|\mathbf{i}\|)\}^{1-\gamma}<\infty.$$

The proof of the following lemma follows from the reverse triangle inequality.

Lemma 3.5. *For each* $\mathbf{i},\mathbf{j}\in\mathcal{J}_{\mathbf{n}}$, $dist(\nu_{\mathbf{i}},\nu_{\mathbf{j}})\geq\max\{\|\mathbf{i}-\mathbf{j}\|-\tilde{r},0\}$, *where* $\tilde{r}=\max\{\|\mathbf{i}-\mathbf{j}\|,\mathbf{i},\mathbf{j}\in\nu\}$ *is the diameter of* $\nu\subset\mathbb{Z}^N$.

4. Main result

The weak consistency of the classifier (1.2) has been established by [21]. In this section we study the strong consistency of (1.2). The following theorem states the strong consistency under mild conditions.

Theorem 4.1. *Assume that Assumptions 1–3 hold for some* $\theta>2N$. *If* $\widehat{\mathbf{n}}b_{\mathbf{n}}^{\tilde{d}}\to\infty$ *as* $\mathbf{n}\to\infty$, *then*

$$L_{\mathbf{n}}\to L^{*}\text{ as }\mathbf{n}\to\infty\text{ with probability one.}$$

Remark 4.1. Note that the assumption on the bandwidth, using by [21] to prove the weak consistency, is similar to the classical assumption used by [7] and [8] in the independent case. In addition, the condition on $b_{\mathbf{n}}$ is minimal compared to that used by [4] and [3] since they have studied the rate of uniform convergence for the estimators. However, the restrictive constraints on the bandwidth in [4] and [3] are related to θ and one has to let $\theta\to\infty$ in order to attain the classical assumption.

5. Simulation study including comparison with the classical kernel rule

Our aim in this section is to look at how the classifier (1.2) behaves on simulated samples by comparing it with the classical kernel rule. We use the R statistical programming environment to run a simulation study for $N=2$. Let $\{(X_{(i,j)},Y_{(i,j)})\}$ be the field of interest and suppose that the simulated data are observed on the area $\mathcal{I}_{(n,n)}=\{(i,j)\in\mathbb{Z}^2:1\leq i,j\leq n\}$. Let

$$\begin{aligned}\mathcal{J}_{(n,n)}=\mathcal{I}_{(n,n)}\backslash\{&\{\nu_{(i,j)}\cup\{(i,j)\},(i,j)\in\mathcal{M}\}\\&\cup\{(1,j),(k,1),(n,l),(m,n):1\leq j,k,l,m\leq n\}\},\end{aligned}$$

where $\mathcal{M}=\{(2k,2l),1\leq k,l\leq 10\}$ is the set of non-observed sites which need to be classified. In this particular case, the vicinity of any missing site (i,j) may be taken as in Figure 1.

It is important to note that the vicinity $\nu_{(i,j)}$ may be designed depending on the location of the missing site (see some typical examples in Figure 2) and that samples with larger size give more freedom to design vicinities.

Figure 2 shows some examples of vicinities that can be used when the missing sites are not completely surrounded by already labeled sites (located at the edges of $\mathcal{S}_{\mathbf{n}}$ for example).

We suppose that the simulated fields have the covariance function

$$C(\mathbf{u}) = 4\|\mathbf{u}\|^{-4.5} \text{ for each } \mathbf{u} \in \mathbb{R}^{*2}.$$

We use the classifier (1.2) with $K(x) = \prod_{i=1}^{8} K_i(x_i)$ for $x = (x_1, \ldots, x_8) \in \mathbb{R}^8$ where $K_i(x_i)$ is the standard Gaussian density (Gaussian kernel). We suppose that $\{X_{(i,j)}, 1 \leq i,j \leq n\}$ are observations of a Gaussian mixture model:

$$\pi_0 \mathcal{N}(\mu_0, \sigma_0^2) + \pi_1 \mathcal{N}(\mu_1, \sigma_1^2) + \pi_2 \mathcal{N}(\mu_2, \sigma_2^2),$$

with $\mu_0 < \mu_1 < \mu_2$ and $\pi_1 + \pi_2 + \pi_3 = 1$. In order to illustrate the fact that our method works for multi-class, the data set $\{X_{(i,j)}, 1 \leq i,j \leq n\}$ is partitioned in three clusters as follows:

$$\text{class } (Y_{(i,j)} = 0) : X_{(i,j)} < (\mu_0 + \mu_1)/2$$

$$\text{class } (Y_{(i,j)} = 1) : (\mu_0 + \mu_1)/2 \leq X_{(i,j)} \leq (\mu_1 + \mu_2)/2$$

$$\text{class } (Y_{(i,j)} = 2) : X_{(i,j)} > (\mu_1 + \mu_2)/2.$$

For each $n = 50, 75, 100$, we generate 100 samples on the region $\mathcal{I}_{(n,n)}$ with $\mu_0 = 5$, $\mu_1 = 15$, $\mu_2 = 25$, $\pi_0 = \pi_1 = \pi_2 = 1/3$ and $\sigma_0^2 = \sigma_1^2 = \sigma_2^2 = 4$. In each replication, we use the classifier (1.2), constructed on the basis of the training data observed on $\mathcal{J}_{(n,n)}$, to re-predict the labels of sites in the test set $\mathcal{M}$. Figure 3 displays one replication for $n = 50$.

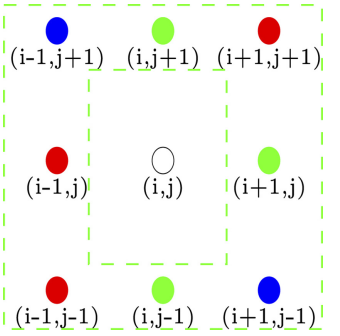

Figure 1.
The missing site (i,j) and its vicinity $\nu_{(i,j)}$ with boundary in green dashed lines.

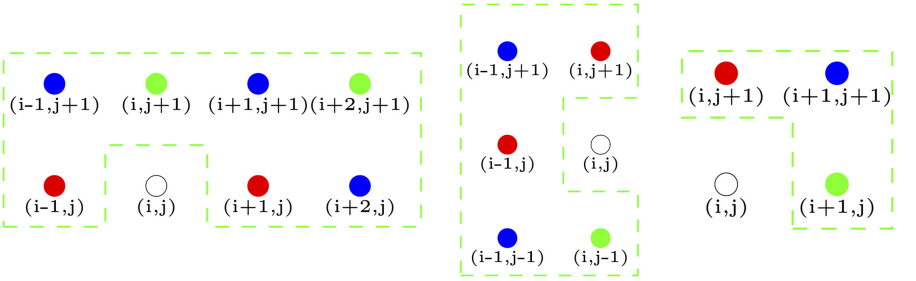

Figure 2.
Three typical vicinities corresponding to three missing sites (i,j) in different locations.

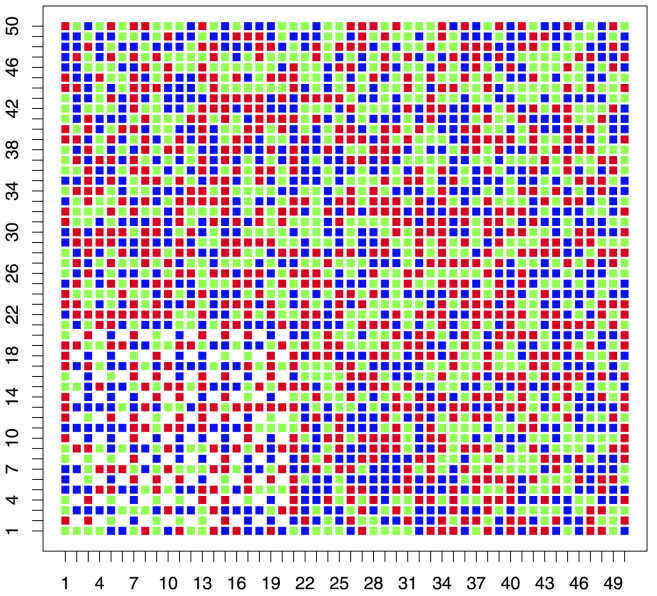

Figure 3. The training sites are colored in red (0), green (1) or blue (2) and the sites to classify are blank.

The optimal bandwidth $\widehat{b}_{opt}$ is obtained by minimizing the cross-validation criterion on a training sample and the misclassification error rate ($E\ R$) is evaluated based on the associated test sample. The average error rate ($A\ E\ R$) is obtained by averaging the error rates associated with the corresponding 100 test samples.

Table 1 shows that the estimated optimal bandwidth and the average error rate decrease when the training sample size increases. This means that the practical results in the simulation study are in line with the theoretical results. Now, let us compare the average error rate ($A\ E\ R$) resulting from application of the proposed classifier with that resulting from application of the classical kernel rule.

5.1 Comparison with the classical kernel rule

The classical kernel rule is given, for any unlabeled site $\mathbf{j}$ with $X_{\mathbf{j}} = x$, by

$$\tilde{g}_{\mathbf{n}}(x) = \arg\max_{0\le k\le M} \sum_{\mathbf{i}\in\mathcal{I}_{\mathbf{n}}} \mathbb{1}_{\{Y_{\mathbf{i}}=k\}} \tilde{K}\left(\frac{x - X_{\mathbf{i}}}{h_{\mathbf{n}}}\right).$$

where $\tilde{K} : \mathbb{R}^d \to \mathbb{R}_+$ is a kernel on $\mathbb{R}^d$ (the Gaussian kernel is considered here), and $h_{\mathbf{n}}$ is a sequence of bandwidths. In order for the classical kernel classifier to be usable in our case, we have to adjust it slightly by taking the sum over $\mathcal{I}_{\mathbf{n}} - \mathcal{M}$ instead of $\mathcal{I}_{\mathbf{n}}$, *i.e.*, for each $\mathbf{j} \in \mathcal{M}$ with $X_{\mathbf{j}} = x$,

Table 1. Estimated optimal bandwidths and average error rates corresponding to the classifier (1.2) with samples of different sizes.

n	50	75	100
$\widehat{b}_{opt}$	2.04	1.93	1.77
AER	28.1%	21.2%	14.8%

$$\tilde{g}_{\mathbf{n}}(x) = \arg \max_{0 \le k \le M} \sum_{\mathbf{i} \in \mathcal{I}_{\mathbf{n}} - \mathcal{M}} \mathbb{1}_{\{Y_{\mathbf{i}} = k\}} \tilde{K}\left(\frac{x - X_{\mathbf{i}}}{h_{\mathbf{n}}}\right).$$

From the theoretical point of view, this is justified by the fact that $\tilde{g}_{\mathbf{n}}$ has the same asymptotic behavior on $\mathcal{I}_{\mathbf{n}}$ as on $\mathcal{I}_{\mathbf{n}} - \mathcal{M}$ since $\mathcal{M}$ is bounded. In this classical kernel method, we consider knowing the features vector $X_{\mathbf{j}}$ of each element $\mathbf{j}$ of $\mathcal{M}$ and we use x, the value of $X_{\mathbf{j}}$, to predict its class while we needed only observations in nearby sites to predict the label of $\mathbf{j}$ by the classifier (1.2). We apply the classical kernel classifier to re-classify the elements of $\mathcal{M}$ using the same training samples generated above and taking into account all the replications for each size $n = 50, 75, 100$. Similar to what we have done in application of (1.2), the optimal bandwidth $\widehat{h}_{opt}$ is chosen by minimizing the cross-validation criterion on a training sample and the misclassification error rate (ER) is evaluated based on the associated test sample. Table 2 reports the average error rate (AER), obtained by averaging the error rates associated with the corresponding 100 test samples.

By comparing Tables 1 and 2, we observe that the corresponding error values in the two tables begin to be close as n increases. This supports the possibility of using the classifier (1.2) as an alternative to the classical kernel classifier when we have to classify sites with missing features.

6. Application to a real data

A digital image is nothing than data numbers indicating variation of *red*, *green* and *blue* (*RGB*) at a particular location on a grid of pixels. An *RGB* color value is specified with: $rgb(red, green, blue)$. Each parameter (red, $green$, $blue$) defines the intensity of the color as an integer between 0 and 255. For example, $rgb(0, 0, 255)$ is rendered as blue, because the blue parameter is set to its highest value 255 and the others are set to 0. One can divide *RGB* color values by 255 in order to provide values in the interval $[0, 1]$. Let us have an image of Eiffel tower with 100 missing pixels as in Figure 3.

We use the R package *jpeg* to convert a *jpg* image into 3-d array of numbers. The package *jpeg* offers the *read JPEG*() function which can read raster graphics (consisting of "pixel matrices") in *jpg* format into R. It returns either a single matrix with gray values in $[0, 1]$ or 3-d array with the *RGB* values in $[0, 1]$, say E. In our example of Figure 3, the dimensions of E are $306 \times 165 \times 3$. Thus, the elements of $E[, j]$ represent the intensities of the color j, for $j =$ "*red*", "*green*" or "*blue*", at all pixels of the grid $\mathcal{I}_{(306,165)}$. For example, the matrix $E[55 : 60, 1 : 6, 1]$ displays the intensities of *red* in each pixel of the region:

$$\{(i, j), 55 \le i \le 60, 1 \le j \le 6\}.$$

Le $X_{(i,j)} = (X^{(1)}_{(i,j)}, X^{(2)}_{(i,j)}, X^{(3)}_{(i,j)})$ where $X^{(k)}_{(i,j)}$ is the intensity of the color k at the pixel (i, j). Since our purpose is to classify new sites with completely missing features, we set an arbitrary threshold of 0.4 and we define labels as follow:

$$Y_{(i,j)} = \begin{cases} 1, & \text{if } \min_{1 \le k \le 3} X^{(k)}_{(i,j)} > 0.4 \\ 0, & \text{otherwise.} \end{cases}$$

Table 2. Estimated optimal bandwidths and average error rates corresponding to the classical kernel classifier with samples of different sizes.

n	25	50	80
$\widehat{h}_{opt}$	1.85	1.72	1.69
AER	23.4%	18.7%	13.2%

The set of 100 missing pixels is taken as a test set, say $\mathcal{M}$. We use the classifier (1.2) (see (1.7) for the binary version) to classify each element of $\mathcal{M}$ based on its eight-neighbors. The optimal bandwidth is evaluated by minimizing the cross-validation criterion on the known sites where we get $\widehat{b}_{opt} \approx 0.72$. The misclassification error rate ($E\,R$) is evaluated on $\mathcal{M}$ where we obtain $E\,R = 0.04$ which indicates that there are only four misclassified cases out of 100 classified cases (see Figure 4).

Now let us use the support vector machine (S V M) classifier to re-classify the elements of $\mathcal{M}$. In this case we should suppose that the RGB value is known for each element of $\mathcal{M}$. For implementing support vector machine in R programming language, we use the package $e1071$. According to this classifier, we get a misclassification error of $E\,R = 0.11$ and this permits to conclude that our kernel classifier in this example proceeds well compared to the (SVM) procedure.

7. Proof of Theorem 4.1

Without loss of generality, we prove the theorem in the binary case where $Y_{\mathbf{j}}$ takes values in $\{0, 1\}$ since no additional argument is required to prove it in the multi-class case. However, the Bayes classifier (1.1) in the binary case is given by

$$g^*(x) = \begin{cases} 0 \text{ if } \mathbb{P}\{Y_{\mathbf{j}} = 0 | X_{(\mathbf{j})} = x\} \geq \mathbb{P}\{Y_{\mathbf{j}} = 1 | X_{(\mathbf{j})} = x\} \\ 1 \text{ otherwise,} \end{cases}$$

and the classifier (1.2) is given by

$$g_n(x) = \begin{cases} 0 \text{ if } \displaystyle\sum_{\mathbf{i} \in \mathcal{J}_{\mathbf{n}}} \mathbb{1}_{\{Y_{\mathbf{i}}=0\}} K\left(\frac{x - X_{(\mathbf{i})}}{b_{\mathbf{n}}}\right) \geq \sum_{\mathbf{i} \in \mathcal{J}_{\mathbf{n}}} \mathbb{1}_{\{Y_{\mathbf{i}}=1\}} K\left(\frac{x - X_{(\mathbf{i})}}{b_{\mathbf{n}}}\right) \\ 1 \text{ otherwise.} \end{cases} \tag{7.1}$$

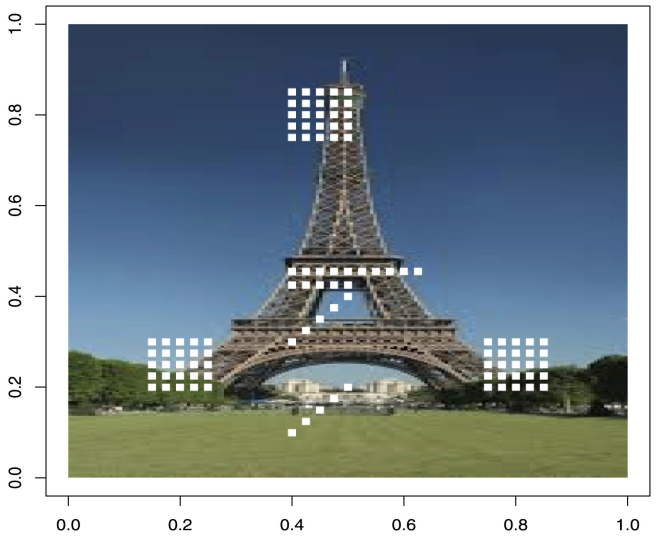

Figure 4. Digital image of Eiffel tower with 100 missing pixels (blank pixels).

Define

$$\eta_{\mathbf{n}}(x) = \frac{\sum_{\mathbf{i}\in\mathcal{J}_{\mathbf{n}}} Y_{\mathbf{i}} K_{\mathbf{n}}(x - X_{(\mathbf{i})})}{\widehat{\mathbf{n}}\mathbb{E}K_{\mathbf{n}}(x - X_{(1)})}.$$

Consequently, the classifier (7.1) can be written as

$$g_{\mathbf{n}}(x) = \begin{cases} 0 \text{ if } \eta_{\mathbf{n}}(x) \leq \dfrac{\sum_{\mathbf{i}\in\mathcal{J}_{\mathbf{n}}} (1 - Y_{\mathbf{i}}) K_{\mathbf{n}}(x - X_{(\mathbf{i})})}{\widehat{\mathbf{n}}\mathbb{E}K_{\mathbf{n}}(x - X_{(1)})} \\ 1 \text{ otherwise.} \end{cases}$$

By Theorem 2.3 in [7], the consistency will be proved if we show that

$$\int_{\mathbb{R}^{\tilde{d}}} |\eta(x) - \eta_{\mathbf{n}}(x)| \mu(dx) \to 0 \text{ as } \mathbf{n} \to \infty \text{ with probability one.} \tag{7.2}$$

But

$$|\eta(x) - \eta_{\mathbf{n}}(x)| \leq |\eta(x) - \mathbb{E}\eta_{\mathbf{n}}(x)| + |\eta_{\mathbf{n}}(x) - \mathbb{E}\eta_{\mathbf{n}}(x)|, \forall x \in \mathbb{R}^{\tilde{d}}.$$

Hence, in order to prove (7.1), it suffices to show that

$$\int_{\mathbb{R}^{\tilde{d}}} |\eta(x) - \mathbb{E}\eta_{\mathbf{n}}(x)| \mu(dx) \to 0 \text{ as } \mathbf{n} \to \infty \tag{7.3}$$

and

$$\int_{\mathbb{R}^{\tilde{d}}} |\eta_{\mathbf{n}}(x) - \mathbb{E}\eta_{\mathbf{n}}(x)| \mu(dx) \to 0 \text{ as } \mathbf{n} \to \infty \text{ with probability one.} \tag{7.4}$$

The proof of (7.3) is the same as in the i.i.d. case (see [7], pp. 156–157). So, it suffices to prove (7.4). To do that, we will employ the blocking technique used in [4]. Let $p = p_{\mathbf{n}} = [\widehat{\mathbf{n}}^{\gamma}]$ for some $1/\theta < \gamma < 1/(2N)$ (where [.] stands for the integer part). Without loss of generality, we suppose that there exists a positive integer q_k such that $n_k = 2pq_k$ for each $k = 1, \ldots, N$. Let

$$\mathbb{J}_q = \{\mathbf{j} = (j_1, \ldots, j_N) \in \mathbb{N}^N : \ 0 \leq j_k \leq q_k - 1, \ \forall k = 1, \ldots, N\}.$$

We define blocks as follow, for each $\mathbf{j} \in \mathbb{J}_q$,

$$\mathbb{S}_{\mathbf{j}}^{(1)} = \{\mathbf{i} \in \mathcal{I}_{\mathbf{n}} : 2j_k p + 1 \leq i_k \leq (2j_k + 1)p, k = 1, \ldots, N\}$$

$$\mathbb{S}_{\mathbf{j}}^{(2)} = \{\mathbf{i} \in \mathcal{I}_{\mathbf{n}} : 2j_k p + 1 \leq i_k \leq (2j_k + 1)p, k = 1, \ldots, N - 1$$

$$\text{and } (2j_N + 1)p + 1 \leq i_N \leq 2(j_N + 1)p\}$$

$$\ldots$$

$$\mathbb{S}_{\mathbf{j}}^{(2^N - 1)} = \{\mathbf{i} \in \mathcal{I}_{\mathbf{n}} : (2j_k + 1)p + 1 \leq i_k \leq 2(j_k + 1)p, k = 1, \ldots, N - 1$$

$$\text{and } 2j_N p + 1 \leq i_N \leq (2j_N + 1)p\}$$

$$\mathbb{S}_{\mathbf{j}}^{(2^N)} = \{\mathbf{i} \in \mathcal{I}_{\mathbf{n}} : (2j_k + 1)p + 1 \leq i_k \leq 2(j_k + 1)p, k = 1, \ldots, N\}.$$

As a consequence, we have $I_{\mathbf{n}} = \bigcup_{k=1}^{2^N} \bigcup_{\mathbf{j}\in\mathbb{J}_q} \mathbb{S}_{\mathbf{j}}^{(k)}$, and for each $k = 1, \ldots, 2^N$, $\text{card}(S_{\mathbf{j}}^{(k)}) = p^N$ and $\text{dist}(\mathbb{S}_{\mathbf{j}}^{(k)}, S_{\mathbf{j}'}^{(k)}) \geq p$ for any $\mathbf{j} \neq \mathbf{j}'$. Let $I_{\mathbf{j}}^{(k)} = \{\mathbf{i} \in \mathbb{S}_{\mathbf{j}}^{(k)} : \nu_{\mathbf{i}} \subset \mathcal{S}_{\mathbf{n}}\}$, for each $k = 1, \ldots, 2^N$

and $\mathbf{j} \in \mathbb{J}_q$. Hence, for a fixed k, we have $\text{dist}(\Gamma_{\mathbf{j}}^{(k)}, \Gamma_{j'}^{\mathbf{k}}) \geq p$ for any $\mathbf{j} \neq \mathbf{j}$, $\text{card}(\Gamma_{\mathbf{j}}^{(k)}) \leq \text{card}(\mathbb{S}_{\mathbf{j}}^{(k)}) = p^N$ and

$$\mathcal{J}_{\mathbf{n}} = \bigcup_{k=1}^{2^N} \bigcup_{\mathbf{j} \in \mathbb{J}_q} \Gamma_{\mathbf{j}}^{(k)}. \tag{7.5}$$

Let $\{(X_{(\mathbf{i})}^*, Y_{\mathbf{i}}^*)\}_{\mathbf{i} \in \mathcal{I}_{\mathbf{n}} - \mathcal{J}_{\mathbf{n}}}$ be a set of independent and identically distributed random vectors such that they are independent of $\{(X_{(\mathbf{i})}, Y_{\mathbf{i}})\}_{\mathbf{i} \in \mathcal{J}_{\mathbf{n}}}$ and $(X_{(\mathbf{i})}^*, Y_{\mathbf{i}}^*)$ is identically distributed with $(X_{(\mathbf{1})}, Y_{\mathbf{1}})$. In order to make sense to the blocking technique, we define random vectors as follow: for each $\mathbf{i} \in \mathcal{I}_{\mathbf{n}}$,

$$(\mathcal{X}_{(\mathbf{i})}, \mathcal{Y}_{\mathbf{i}}) = \begin{cases} (X_{(\mathbf{i})}, Y_{\mathbf{i}}) \text{ if } \nu_{\mathbf{i}} \subset \mathcal{S}_{\mathbf{n}} \\ (X_{(\mathbf{i})}^*, Y_{\mathbf{i}}^*) \text{ if } \nu_{\mathbf{i}} \not\subset \mathcal{S}_{\mathbf{n}}. \end{cases}$$

It is clear that $\{(\mathcal{X}_{(\mathbf{i})}, \mathcal{Y}_{\mathbf{i}}),\ \mathbf{i} \in \mathcal{J}_{\mathbf{n}}\} = \{(X_{(\mathbf{i})}, Y_{\mathbf{i}}),\ \mathbf{i} \in \mathcal{J}_{\mathbf{n}}\}$ and $\{(\mathcal{X}_{(\mathbf{i})}, \mathcal{Y}_{\mathbf{i}}),\ \mathbf{i} \in \Gamma_{\mathbf{j}}^{(k)}\} = \{(X_{(\mathbf{i})}, Y_{\mathbf{i}}),\ \mathbf{i} \in \Gamma_{\mathbf{j}}^{(k)}\}$. Now, for a fixed k and each $\mathbf{j} \in \mathbb{J}_q$, let $W_{\mathbf{j}}^{(k)} = \{(\mathcal{X}_{(\mathbf{i})}, \mathcal{Y}_{\mathbf{i}}),\ \mathbf{i} \in \mathbb{S}_{\mathbf{j}}^{(k)}\}$ be a vector whose components are ordered according to a given order on indices. Applying Lemma 3.2 together with the blocks decomposition introduced by [10] (see also [20]) on the family of vectors $\{W_{\mathbf{j}}^{(k)}, \mathbf{j} \in \mathbb{J}_q\}$, we can generate independent copies $\{\tilde{W}_{\mathbf{j}}^{(k)}, \mathbf{j} \in \mathbb{J}_q\}$ such that: they are mutually independent, and for each $\mathbf{j} \in \mathbb{J}_q$, $\tilde{W}_{\mathbf{j}}^{(k)} = \{(\tilde{\mathcal{X}}_{(\mathbf{i})}, \tilde{\mathcal{Y}}_{\mathbf{i}}),\ \mathbf{i} \in \mathbb{S}_{\mathbf{j}}^{(k)}\}$ has the same distribution as $W_{\mathbf{j}}^{(k)} = \{(\tilde{\mathcal{X}}_{(\mathbf{i})}, \tilde{\mathcal{Y}}_{\mathbf{i}}), \mathbf{i} \in S_{\mathbf{j}}^{(k)}\}$. Furthermore, by Lemma 3.5, we have $\mathbb{P}(W_{\mathbf{j}}^{(k)} \neq \tilde{W}_{\mathbf{j}}^{(k)}) \leq \varphi(p - \tilde{r})$ since $p \geq \tilde{r}$ for $\widehat{\mathbf{n}}$ large enough. Thus, the two vectors $(\tilde{\mathcal{X}}_{(\mathbf{i})}, \tilde{\mathcal{Y}}_{(\mathbf{i})})$ and $(\tilde{\mathcal{X}}_{(\mathbf{i}')}, \tilde{\mathcal{Y}}_{\mathbf{i}'})$ are independent for each $\mathbf{i} \in \mathbb{S}_{\mathbf{j}}^{(k)}$ and $\mathbf{i}' \in \mathbb{S}_{\mathbf{j}'}^{(k)}$ with $\mathbf{j} \neq \mathbf{j}'$. Now, for each $\mathbf{i} \in \mathcal{J}_{\mathbf{n}}$, there exists $\mathbf{j} \in \mathbb{J}_q$ such that $\{(\mathcal{X}_{(\mathbf{i})}, \mathcal{Y}_{\mathbf{i}}) \neq (\tilde{\mathcal{X}}_{(\mathbf{i})}, \tilde{\mathcal{Y}}_{\mathbf{i}})\} \subseteq (W_{\mathbf{j}}^{(k)} \neq \tilde{W}_{\mathbf{j}}^{(k)})$. Since $(\tilde{\mathcal{X}}_{(\mathbf{i})}, \tilde{\mathcal{Y}}_{\mathbf{i}}) = (\tilde{\mathcal{X}}_{(\mathbf{i})}, \tilde{\mathcal{Y}}_{\mathbf{i}})$ for each $\mathbf{i} \in \mathcal{J}_{\mathbf{n}}$, denote $(\tilde{\mathcal{X}}_{(\mathbf{i})}, \tilde{\mathcal{Y}}_{\mathbf{i}}) = (\tilde{X}_{(\mathbf{i})}, \tilde{Y}_{\mathbf{i}})$, for each $\mathbf{i} \in \mathcal{J}_{\mathbf{n}}$ (or $\mathbf{i} \in \Gamma_{\mathbf{j}}^{(k)}$). As a consequence

$$\mathbb{P}\left\{(X_{(\mathbf{i})}, Y_{\mathbf{i}}) \neq (\tilde{X}_{(\mathbf{i})}, \tilde{Y}_{\mathbf{i}})\right\} \leq \varphi(p - \tilde{r}), \text{ for each } \mathbf{i} \in \mathcal{J}_{\mathbf{n}}. \tag{7.6}$$

By (7.5), we can write

$$\sum_{\mathbf{i} \in \mathcal{J}_{\mathbf{n}}} \tilde{Y}_{\mathbf{i}} K_{\mathbf{n}}(x - \tilde{X}_{(\mathbf{i})}) = \sum_{k=1}^{2^N} \sum_{\mathbf{j} \in J_q} \sum_{\mathbf{i} \in \Gamma_{\mathbf{j}}^{(k)}} \tilde{Y}_{\mathbf{i}} K_{\mathbf{n}}(x - \tilde{X}_{(\mathbf{i})}).$$

If we denote

$$\tilde{\eta}_{\mathbf{n}}(x) = \frac{\sum_{\mathbf{i} \in \mathcal{J}_{\mathbf{n}}} \tilde{Y}_{\mathbf{i}} K_{\mathbf{n}}(x - \tilde{X}_{(\mathbf{i})})}{\widehat{\mathbf{n}} \mathbb{E} K_{\mathbf{n}}(x - X_{(\mathbf{1})})} \text{ and } \tilde{\eta}_{\mathbf{n},k}(x) = \frac{\sum_{\mathbf{j} \in J_q} \sum_{\mathbf{i} \in \Gamma_{\mathbf{j}}^{(k)}} \tilde{Y}_{\mathbf{i}} K_{\mathbf{n}}(x - \tilde{X}_{(\mathbf{i})})}{\widehat{\mathbf{n}} \mathbb{E} K_n(x - X_{(\mathbf{1})})}, \tag{7.7}$$

then

$$\tilde{\eta}_{\mathbf{n}}(x) = \sum_{k=1}^{2^N} \tilde{\eta}_{\mathbf{n},k}(x). \tag{7.8}$$

Using Markov's inequality and Lemma 3.3 together with (7.7), we have for any $\epsilon > 0$,

$$\begin{aligned}
&\mathbb{P}\left(\left|\int_{\mathbb{R}^{\tilde{d}}}|\eta_{\mathbf{n}}(x)-\mathbb{E}\eta_{\mathbf{n}}(x)|\mu(dx)-\int_{\mathbb{R}^{\tilde{d}}}|\tilde{\eta}_{\mathbf{n}}(x)-\mathbb{E}\tilde{\eta}_{\mathbf{n}}(x)|\mu(dx)\right|>\epsilon\right)\\
&\le \epsilon^{-1}\mathbb{E}\left|\int_{\mathbb{R}^{\tilde{d}}}|\eta_{\mathbf{n}}(x)-\mathbb{E}\eta_{\mathbf{n}}(x)|\mu(dx)-\int_{\mathbb{R}^{\tilde{d}}}|\tilde{\eta}_{\mathbf{n}}(x)-\mathbb{E}\tilde{\eta}_{\mathbf{n}}(x)|\mu(dx)\right|\\
&\le \epsilon^{-1}\mathbb{E}\left(\int_{\mathbb{R}^{\tilde{d}}}|\tilde{\eta}_{\mathbf{n}}(x)-\eta_{\mathbf{n}}(x)|\mu(dx)+\mathbb{E}\int_{\mathbb{R}^{\tilde{d}}}|\tilde{\eta}_{\mathbf{n}}(x)-\eta_{\mathbf{n}}(x)|\mu(dx)\right)\\
&= 2\epsilon^{-1}\mathbb{E}\int_{\mathbb{R}^{\tilde{d}}}|\tilde{\eta}_{\mathbf{n}}(x)-\eta_{\mathbf{n}}(x)|\mu(dx)\\
&= 2\epsilon^{-1}\mathbb{E}\int_{\mathbb{R}^{\tilde{d}}}\left|\frac{\sum_{\mathbf{i}\in\mathcal{J}_{\mathbf{n}}}\tilde{Y}_{\mathbf{i}}K_{\mathbf{n}}(x-\tilde{X}_{(\mathbf{i})})}{\widehat{\mathbf{n}}\mathbb{E}K_{\mathbf{n}}(x-X_{(\mathbf{1})})}-\frac{\sum_{\mathbf{i}\in\mathcal{J}_{\mathbf{n}}}Y_{\mathbf{i}}K_{\mathbf{n}}(x-X_{(\mathbf{i})})}{\widehat{n}\mathbb{E}K_{\mathbf{n}}(x-X_{(\mathbf{1})})}\right|\mu(dx)\\
&\le 4\epsilon^{-1}\sum_{\mathbf{i}\in\mathcal{J}_{\mathbf{n}}}\mathbb{E}\mathbb{1}_{\{(\tilde{X}_{(\mathbf{i})},\tilde{Y}_{\mathbf{i}})\neq(X_{(\mathbf{i})},Y_{\mathbf{i}})\}}\sup_{u\in\mathbb{R}^{\tilde{d}}}\int_{\mathbb{R}^{\tilde{d}}}\frac{K_{\mathbf{n}}(x-u)}{\widehat{\mathbf{n}}\mathbb{E}K_{\mathbf{n}}(x-X_{(\mathbb{1})})}\mu(dx)\\
&\le 4(\epsilon\widehat{\mathbf{n}})^{-1}\rho\sum_{\mathbf{i}\in\mathcal{J}_{\mathbf{n}}}\mathbb{E}\mathbb{1}_{\{(\tilde{X}_{(\mathbf{i})},\tilde{Y}_{\mathbf{i}})\neq(X_{(\mathbf{i})},Y_{\mathbf{i}})\}}\le 4\epsilon^{-1}\rho\varphi(p-\tilde{r}),
\end{aligned}$$

where $\rho > 0$ is the constant defined in Lemma 3.3. Since $\tilde{r}$ is bounded and $p \to \infty$ as $n \to \infty$, so $p - \tilde{r} \ge p/2$ for $\widehat{\mathbf{n}}$ large enough. Therefore, we get

$$\begin{aligned}
&\mathbb{P}\left(\left|\int_{\mathbb{R}^{\tilde{d}}}|\eta_{\mathbf{n}}(x)-\mathbb{E}\eta_{\mathbf{n}}(x)|\mu(dx)-\int_{\mathbb{R}^{\tilde{d}}}|\tilde{\eta}_{\mathbf{n}}(x)-\mathbb{E}\tilde{\eta}_{\mathbf{n}}(x)|\mu(dx)\right|>\epsilon\right)\\
&\le 4\epsilon^{-1}\rho\varphi(p/2)\le C\epsilon^{-1}\rho\widehat{\mathbf{n}}^{-\gamma\theta},
\end{aligned}$$

for some generic positive constant $C > 0$. Since $\gamma\theta > 1$, by Borel–Cantelli lemma, we have

$$\int_{\mathbb{R}^{\tilde{d}}}|\eta_{\mathbf{n}}(x)-\mathbb{E}\eta_{\mathbf{n}}(x)|\mu(dx)-\int_{\mathbb{R}^{\tilde{d}}}|\tilde{\eta}_{\mathbf{n}}(x)-\mathbb{E}\tilde{\eta}_{\mathbf{n}}(x)|\mu(dx)\to 0, \tag{7.9}$$

with probability one. Now, we will show that

$$\int_{\mathbb{R}^{\tilde{d}}}|\tilde{\eta}_{\mathbf{n}}(x)-\mathbb{E}\tilde{\eta}_{\mathbf{n}}(x)|\mu(dx)\to 0 \text{ with probability one.} \tag{7.10}$$

By (7.7) and (7.8), we have

$$\int_{\mathbb{R}^{\tilde{d}}}|\tilde{\eta}_{\mathbf{n}}(x)-\mathbb{E}\tilde{\eta}_{\mathbf{n}}(x)|\mu(dx)\le\sum_{k=1}^{2^N}\int_{\mathbb{R}^{\tilde{d}}}|\tilde{\eta}_{n,k}(x)-\mathbb{E}\tilde{\eta}_{n,k}(x)|\mu(dx). \tag{7.11}$$

Consequently, in order to establish (7.10), it is sufficient to show that

$$\int_{\mathbb{R}^{\tilde{d}}}|\tilde{\eta}_{\mathbf{n},k}(x)-\mathbb{E}\tilde{\eta}_{\mathbf{n},k}(x)|\mu(dx)\to 0 \text{ as } \mathbf{n}\to\infty \text{ with probability one ,} \tag{7.12}$$

for each $1 \leq k \leq 2^N$. Without loss of generality, we show (7.12) for $k = 1$. If the elements of $\mathbb{J}_q$ are enumerated in an arbitrary manner, we can write $\mathbb{J}_q = \{1, \ldots, m\}$ with $m = \mathrm{card}(\mathbb{J}_q) = \prod_{k=1}^{N} q_k$. Denote $\tilde{Z}_j = \{(\tilde{\mathcal{X}}_{(i)}, \tilde{\mathcal{Y}}_i), \mathbf{i} \in \mathbb{S}_j^{(1)}\}$, for each $j = 1, \ldots, m$, where the components of $\tilde{Z}_j$ are ordered according to an arbitrary order on indices. Recall that $(\tilde{\mathcal{X}}_{(\mathbf{i})}, \tilde{\mathcal{Y}}_{\mathbf{i}}) = (\tilde{X}_{(\mathbf{i})}, \tilde{Y}_{\mathbf{i}})$ for $\mathbf{i} \in \Gamma_j^{(1)}$ and suppose that $(\tilde{\mathcal{X}}_{(\mathbf{i})}, \mathcal{Y}_{\mathbf{i}})$ is replaced by $(\mathbf{0}_{\tilde{d}}, 0)$ if $\mathbf{i} \notin \Gamma_j^{(1)}$ where $\mathbf{0}_{\tilde{d}} = (0, \ldots, 0) \in \mathbb{R}^{\tilde{d}}$. Hence, by the blocks decomposition, the random vectors $\tilde{Z}_1, \ldots, \tilde{Z}_m$ are independent. Let $F : ((\mathbb{R}^{\tilde{d}} \times \{0,1\})^{p^N})^m \to \mathbb{R}$ be a real function defined as follows

$$F(\tilde{Z}_1, \ldots, \tilde{Z}_m) = \int_{\mathbb{R}^{\tilde{d}}} \left| \sum_{j=1}^{m} \sum_{\mathbf{i} \in \mathbb{S}_j^{(1)}} \left(\frac{\tilde{\mathcal{Y}}_{\mathbf{i}} K_{\mathbf{n}}(x - \tilde{\mathcal{X}}_{(\mathbf{i})})}{\widehat{\mathbf{n}} \mathbb{E} K_{\mathbf{n}}(x - X_{(1)})} - \frac{\mathbb{E}\tilde{Y}_1 K_{\mathbf{n}}(x - \tilde{X}_{(1)})}{\widehat{\mathbf{n}} \mathbb{E} K_{\mathbf{n}}(x - X_{(1)})} \right) \right| \mu(dx)$$

$$= \int_{\mathbb{R}^{\tilde{d}}} \left| \sum_{j=1}^{m} \sum_{i \in \Gamma_j^{(1)}} \left(\frac{\tilde{Y}_{\mathbf{i}} K_{\mathbf{n}}(x - \tilde{X}_{(\mathbf{i})})}{\widehat{\mathbf{n}} \mathbb{E} K_{\mathbf{n}}(x - X_{(1)})} - \frac{\mathbb{E}\tilde{Y}_1 K_{\mathbf{n}}(x - \tilde{X}_{(\mathbf{1})})}{\widehat{\mathbf{n}} \mathbb{E} K_{\mathbf{n}}(x - X_{(\mathbf{1})})} \right) \right| \mu(dx)$$

$$= \int_{\mathbb{R}^{\tilde{d}}} |\tilde{\eta}_{\mathbf{n},1}(x) - \mathbb{E}\tilde{\eta}_{\mathbf{n},1}(x)| \mu(dx).$$

For $\tilde{z}_j \neq \tilde{z}_j'$ where $\tilde{z}_j = \{(\tilde{x}_{(\mathbf{i})}, \tilde{y}_{\mathbf{i}}), \mathbf{i} \in \mathbb{S}_j^{(1)}\}, \tilde{z}_j' = \{(\tilde{x}_{(\mathbf{i})}', \tilde{y}_{\mathbf{i}}'), \mathbf{i} \in \mathbb{S}_j^{(1)}\} \in (\mathbb{R}^{\tilde{d}} \times \{0,1\})^{p^N}$ and $(\tilde{x}_{(\mathbf{i})}, \tilde{y}_{\mathbf{i}}) = (\tilde{x}_{(\mathbf{i})}', \tilde{y}_{\mathbf{i}}') = (\mathbf{0}_{\tilde{d}}, 0)$ for each $\mathbf{i} \notin \Gamma_j^{(1)}$, using Lemma 3.3, we have

$$\left| F(\tilde{Z}_1, \ldots, \tilde{z}_j, \ldots, \tilde{Z}_m) - F\left(\tilde{Z}_1, \ldots, \tilde{z}_j', \ldots, \tilde{Z}_m\right) \right|$$

$$\leq \int_{\mathbb{R}^{\tilde{d}}} \left| \sum_{i \in \Gamma_j^{(1)}} \frac{\tilde{y}_{\mathbf{i}} K_{\mathbf{n}}(x - \tilde{x}_{(\mathbf{i})})}{\widehat{\mathbf{n}} \mathbb{E} K_{\mathbf{n}}(x - X_{(1)})} - \sum_{i \in \Gamma_j^{(1)}} \frac{\tilde{y}_{\mathbf{i}}' K_{\mathbf{n}}(x - \tilde{x}_{(\mathbf{i})}')}{\widehat{\mathbf{n}} \mathbb{E} K_{\mathbf{n}}(x - X_{(1)})} \right| \mu(dx)$$

$$\leq 2p^N \sup_{u \in \mathbb{R}^{\tilde{d}}} \int_{\mathbb{R}^{\tilde{d}}} \frac{K_{\mathbf{n}}(x - u)}{\widehat{\mathbf{n}} \mathbb{E} K_{\mathbf{n}}(x - X_{(1)})} \mu(dx) \leq 2\rho p^N \widehat{\mathbf{n}}^{-1}.$$

Hence, since $\widehat{\mathbf{n}} = 2^N p^N m$ with $m = \prod_{k=1}^{N} q_k$, by McDiarmid's inequality [16], we have for every $\epsilon > 0$,

$$\mathbb{P}(|F(\tilde{Z}_1, \ldots, \tilde{Z}_m) - \mathbb{E}F(\tilde{Z}_1 ..., \tilde{Z}_m)| > \epsilon) \leq 2 \exp\left(-\frac{2^{N-1} \epsilon^2 \widehat{\mathbf{n}}}{\rho^2 p^N} \right).$$

Since $p = [\widehat{\mathbf{n}}^{\gamma}]$ with $\mathbf{1}/\theta < \gamma < \mathbf{1}/(2N)$, then $\widehat{\mathbf{n}}^{1-\gamma^N}/\log(\widehat{\mathbf{n}}) \to \infty$ and Borel–Cantelli lemma yields

$$F(\tilde{Z}_1, \ldots, \tilde{Z}_m) - \mathbb{E}F(\tilde{Z}_1 ..., \tilde{Z}_m) \to 0 \text{ with probability one.}$$

As a consequence

$$\int_{\mathbb{R}^{\tilde{d}}} |\tilde{\eta}_{\mathbf{n},1}(x) - \mathbb{E}\tilde{\eta}_{\mathbf{n},1}(x)| \mu(dx) - \mathbb{E} \int_{\mathbb{R}^{\tilde{d}}} |\tilde{\eta}_{\mathbf{n},1}(x) - \mathbb{E}\tilde{\eta}_{\mathbf{n},1}(x)| \mu(dx) \to 0 \tag{7.13}$$

with probability one. In order to complete the proof of (7.12) for $k = 1$, it remains to show that

$$\mathbb{E}F(\tilde{Z}_1, \ldots, \tilde{Z}_m) = \mathbb{E}\int_{\mathbb{R}^d} \left|\tilde{\eta}_{\mathbf{n},1}(x) - \mathbb{E}\tilde{\eta}_{\mathbf{n},1}(x)\right| \mu(dx) \to 0. \tag{7.14}$$

The proof of (7.14) can be achieved by the same arguments used by ([21], Section 5), in addition to benefiting from Lemmas 3.1, 3.4 and 3.5. Combining (7.9), (7.10), (7.12)–(7.14), we get (7.4). Finally, (7.3) and (7.4) yield (7.2) and the proof is completed. □

References

[1] H.C.P. Berbee, Random Walks with Stationary Increments and Renewal Theory, Math. Cent. Tract., Amsterdam, 1979.

[2] G. Biau, B. Cadre, Nonparametric spatial prediction, Stat. Inference Stoch. Process. 7 (2004) 327–349.

[3] M. Carbon, C. Francq, L.T. Tran, Kernel regression estimation for random fields, J. Statist. Plann. Inference 137 (2007) 778–798.

[4] M. Carbon, L. Tran, B. Wu, Kernel density estimation for random fields, Statist. Probab. Lett. 36 (1997) 115–125.

[5] S. Dabo-Niang, L. Hamdad, C. Ternynck, A kernel spatial density estimation allowing for the analysis of spatial clustering. application to monsoon asia drought atlas data, Stoch. Environ. Res. Risk. Assess. 28 (2014) 2075.

[6] S. Dabo-Niang, C. Ternynck, A.-F. Yao, Nonparametric prediction of spatial multivariate data, J. Nonparametr. Stat. 28 (2016) 428–458.

[7] L. Devroye, L. Györfi, G. Lugosi, A probabilitic Theory of Pattern Recognition, Spriner-Verlag, New York, 1996.

[8] L. Devroye, A. Krzyżak, An equivalence theorem for L1convergence of the kernel regression estimate, J. Statist. Plann. Inference 23 (1989) 71–82.

[9] P. Doukhan, P. Massart, E. Rio, The functional central limit theorem for strongly mixing processes, Ann. Inst. H. Poincaré Probab. Statist. 30 (1) (1994) 63–82.

[10] P. Doukhan, P. Massart, E. Rio, Invariance principles for absolutely regular empirical processes, Ann. Inst. H. Poincaré Probab. Statist. 31 (2) (1995) 393–427.

[11] X. Guyon, Estimation d'un champ par pseudo-vraisemblance conditionnelle: Etude asymptotique et application au cas markovien, in: Proc.6th Franco-Belgian Meeting of Statisticians, 1987.

[12] I.A. Ibragimov, Some limit theorems for for stationary processes, Theory Probab. Appl. 7 (2011) 349–382.

[13] A.S. Kechris, Classical Descriptive Set Theory, Spriner-Verlag, New York, 1995.

[14] M.E. Machkouri, Asymptotic normality of the parzen–rosenblatt density estimator for strongly mixing random fields, J. Statist. Plann. Inference 14 (2011) 73–84.

[15] M.E. Machkouri, R. Stoica, Asymptotic normality of kernel estimates in a regression model for random fields, J. Nonparametr. Stat. 22 (2010) 366–377.

[16] C. McDiarmid, On the method of bounded differences, Surveys in combinatorics 1989, Cambridge University Press, Cambridge, 1989, pp. 148–188.

[17] C.C. Nedearhouser, Convergence of blocks spins defined by a random fields, J. Stat. Phys. 22 (1980) 673–684.

[18] E. Rio, Théorie Asymptotique des Processus Aléatoires Faiblement Dépendants. Mathématiques et Applications, Spriner, Berlin, 2000.

[19] M. Rosenblatt, Stationary Sequences and Random Fields, Birkhäuser, Boston, 1985.

[20] G. Viennet, Inequalities for absolutely sequence. Application to density estimation, Probab. Theory Related Fields 107 (4) (1967) 467–492.

[21] A. Younso, On the consistency of a new kernel rule for spatially dependent data, Statist. Probab. Lett. 131 (2017) 64–71.

Corresponding author
Ahmad Younso can be contacted at: ahyounso@yahoo.fr