Journal of Machine Learning Research 14 (2013) 2361-2376 Submitted 3/13; Revised 6/13; Published 8/13

On the Mutual Nearest Neighbors Estimate in Regression

Arnaud Guyader ARNAUD.GUYADER@UHB.FR
Université Rennes 2
IRMAR & INRIA Rennes
Campus de Villejean
Place du Recteur Henri Le Moal, CS 24307
35043 Rennes Cedex, France

Nick Hengartner NICKH@LANL.GOV
Information Sciences Group
Los Alamos National Laboratory
TA-3, Bldg 508, Room 133, Mail Stop B-256
Los Alamos, NM 86545, USA

Editor: Ulrike von Luxburg

Abstract

Motivated by promising experimental results, this paper investigates the theoretical properties of a recently proposed nonparametric estimator, called the Mutual Nearest Neighbors rule, which estimates the regression function $m(\mathbf{x}) = \mathbb{E}[Y|\mathbf{X} = \mathbf{x}]$ as follows: first identify the k nearest neighbors of $\mathbf{x}$ in the sample $\mathcal{D}_n$, then keep only those for which $\mathbf{x}$ is itself one of the k nearest neighbors, and finally take the average over the corresponding response variables. We prove that this estimator is consistent and that its rate of convergence is optimal. Since the estimate with the optimal rate of convergence depends on the unknown distribution of the observations, we also present adaptation results by data-splitting.

Keywords: nonparametric estimation, nearest neighbor methods, mathematical statistics

1. Introduction

Let $\mathcal{D}_n = \{(\mathbf{X}_1, Y_1), \dots, (\mathbf{X}_n, Y_n)\}$ be a sample of independent and identically distributed (i.i.d.) copies of an $\mathbb{R}^d \times \mathbb{R}$-valued random pair $(\mathbf{X}, Y)$ satisfying $\mathbb{E}Y^2 < \infty$. For fixed $\mathbf{x}$ in $\mathbb{R}^d$, our goal is to estimate the regression function $m(\mathbf{x}) = \mathbb{E}[Y|\mathbf{X} = \mathbf{x}]$ using the data $\mathcal{D}_n$. A regression function estimate $m_n(\mathbf{x})$ is said to be weakly consistent if the mean integrated squared error $\mathbb{E}[m_n(\mathbf{X}) - m(\mathbf{X})]^2$ tends to 0 as the sample size n goes to infinity, and is said to be universally weakly consistent if this property holds for all distributions of $(\mathbf{X}, Y)$ with $\mathbb{E}Y^2 < \infty$.

Equip the space $\mathbb{R}^d$ with the standard Euclidean metric. Then, for $\mathbf{x}$ in $\mathbb{R}^d$, the k Nearest Neighbors (kNN) estimate for the regression function m is defined by

$$m_n^{k\mathrm{NN}}(\mathbf{x}) = \frac{1}{k} \sum_{i=1}^{k} Y_{(i,n)}(\mathbf{x}),$$

where $(\mathbf{X}_{(1,n)}(\mathbf{x}), Y_{(1,n)}(\mathbf{x})), \dots, (\mathbf{X}_{(n,n)}(\mathbf{x}), Y_{(n,n)}(\mathbf{x}))$ denotes a reordering of the data according to the increasing values of $d_i = d_i(\mathbf{x}) = \|\mathbf{X}_i - \mathbf{x}\|$ (ties are broken in favor of smallest indices). This procedure is one of the oldest approaches to regression analysis, dating back to Fix and Hodges

(1951, 1952), and is among the most popular nonparametric methods. We refer the reader to Devroye et al. (1996) for results and details in the classification context, and to Györfi et al. (2002) for the regression framework considered in the present paper. Accordingly, we adhere as much as possible to their notations.

Let us denote $\mathcal{N}_k(\mathbf{x})$ the set of the k nearest neighbors of $\mathbf{x}$ in $\mathcal{D}_n$, $\mathcal{N}_k'(\mathbf{X}_i)$ the set of the k nearest neighbors of $\mathbf{X}_i$ in $(\mathcal{D}_n \setminus \{\mathbf{X}_i\}) \cup \{\mathbf{x}\}$, and

$$\mathcal{M}_k(\mathbf{x}) = \left\{\mathbf{X}_i \in \mathcal{N}_k(\mathbf{x}) :\ \mathbf{x} \in \mathcal{N}_k'(\mathbf{X}_i)\right\},$$

the set of the Mutual Nearest Neighbors (MNN) of $\mathbf{x}$. Denoting $M_k(\mathbf{x}) = |\mathcal{M}_k(\mathbf{x})|$ the number of mutual nearest neighbors of $\mathbf{x}$, $M_k(\mathbf{x})$ is a random variable taking values between 0 and k. The mutual nearest neighbors regression estimate is then defined as follows

$$m_n(\mathbf{x}) = \frac{1}{M_k(\mathbf{x})} \sum_{i:X_i \in \mathcal{M}_k(\mathbf{x})} Y_i,$$

with the convention that $0/0 = 0$. Two remarks are in order. First, contrarily to the k-NN estimate, the MNN estimate is symmetric. This means that, when averaging over the neighbors of $\mathbf{x}$ in the sample $\mathcal{D}_n$, we only consider the points for which $\mathbf{x}$ is itself one of the k nearest neighbors.

The second remark is that, compared to the standard kNN rule, there might be an additional computational cost for applying the MNN procedure. Specifically, we might consider two different situations. In the first one, it is possible to precompute and sort the distances between all couples of points $(\mathbf{X}_i, \mathbf{X}_j)$ in the sample $\mathcal{D}_n$. Since the cost of computing the distance between a pair of d-dimensional vectors is $O(d)$, and that there are $n(n-1)/2$ such pairs in $\mathcal{D}_n$, and considering that the (quick)sorting of a vector of size n is $O(n \log n)$, the cost of this precomputation is $O((d + \log n)n^2)$. In this case, after computing and sorting the pairwise distances, the computational burden of MNN and kNN are of the same order. Indeed, for a new point $\mathbf{x}$, computing the distances to the $\mathbf{X}_i$'s and finding the k nearest neighbors has a cost in $O((d + \log k)n)$. For the mutual nearest neighbors, for each of these k nearest neighbors, one has also to see if $\mathbf{x}$ is one of its k nearest neighbors, hence an additive cost in $O(k)$. In the second situation, the cost of precomputation is prohibitively expensive, typically due to large sample size n and high dimension d of the covariates. In this case, the algorithmic cost for the kNN rule is of course the same as before, that is in $O((d + \log k)n)$, while the cost for the MNN rule is $O((k+1)(d + \log k)n) = O(k(d + \log k)n)$.

The term of *mutual nearest neighbors* seems to date back to Chidananda Gowda and Krishna (1978, 1979) in the context of clustering. In the past few years, it has raised an increasing interest in image analysis for object retrieval (see for example Jégou et al. (2010) and Qin et al. (2011)) as well as for classification purposes (see Liu et al., 2010). Interestingly, the latter reports that experimental results show that, on standard data sets, the MNN estimates have better performances than standard nearest neighbors estimates as well as other widely used classification rules.

Without claiming that MNN estimates always outperform standard nearest neighbors estimates, a heuristic explanation for this better behavior in some situations is related to the existence of hubs in high dimensional data. Specifically, a hub is a point which appears in many more kNN lists than the others, making it very influential in kNN estimates. As explained in Radovanović et al. (2010), hubness is an aspect of the curse of dimensionality as increasing the dimensionality results in the emergence of hubs under widely applicable conditions. These authors have also conducted several simulations to show how the existence of "bad" hubs negatively affects the kNN classifier (see Section 7.1.2 in Radovanović et al. 2010). In our context, the existence of hubs might not affect the performance of MNN estimates and one could even consider the MNN rule as a variant of the kNN rule which allows to automatically reduce the role of these hubs.

However, to the best of our knowledge, little if nothing is known about the theoretical properties of the mutual nearest neighbors estimator. Our goal in this paper is to investigate its statistical properties, focusing our attention on the regression viewpoint. In Section 2, we present strong and weak consistency results. In Section 3, we go one step further and show that the rate of convergence of this estimate is, in fact, optimal when $d \geq 2$. Since the parameter $k = k_n$ of the estimate with the optimal rate of convergence depends on the unknown distribution of $(\mathbf{X}, Y)$, especially on the smoothness of the regression function, we also present adaptive (i.e., data-dependent) choices for k_n that preserve the minimax optimality of the estimate.

2. Consistency

To prove the consistency of the MNN estimator, we write

$$m_n(\mathbf{x}) = \sum_{i=1}^{n} W_i(\mathbf{x}, \mathbf{X}_1, \ldots, \mathbf{X}_n) Y_i = \sum_{i=1}^{n} W_i Y_i,$$

where the weights W_i are non negative random variables defined by

$$W_i = \begin{cases} \frac{1}{M_k(\mathbf{x})} & \text{if } M_k(\mathbf{x}) > 0 \text{ and } \mathbf{X}_i \in \mathcal{M}_k(\mathbf{x}), \\ 0 & \text{otherwise.} \end{cases}$$

This representation brings the MNN estimator into the general framework of weighted nearest neighbors, as studied for example in Stone (1977). But, contrarily to the standard kNN estimator for which the weights are deterministically linked to the order statistics $\mathbf{X}_{(1,n)}(\mathbf{x}), \ldots, \mathbf{X}_{(n,n)}(\mathbf{x})$, notice that this is not the case in our situation.

Nonetheless, in order to control the random weights W_i, we will exploit the following observation: for all $\mathbf{X}_i$ in $\mathcal{N}'_k(\mathbf{x})$, we have the following assertion

$$\|\mathbf{X}_i - \mathbf{x}\| < \frac{d_{(k+1)}}{2} = \frac{\|\mathbf{X}_{(k+1,n)}(\mathbf{x}) - \mathbf{x}\|}{2} \quad \Rightarrow \quad \mathbf{X}_i \in \mathcal{M}_k(\mathbf{x}). \tag{1}$$

Indeed, if not, there would exist k points $\tilde{\mathbf{X}}_1, \ldots, \tilde{\mathbf{X}}_k$, different from $\mathbf{X}_i$, and such that for all $j = 1, \ldots, k$,

$$\|\tilde{\mathbf{X}}_j - \mathbf{X}_i\| < \|\mathbf{X}_i - \mathbf{x}\| < \frac{d_{(k+1)}}{2}.$$

By the triangle inequality,

$$\|\tilde{\mathbf{X}}_j - \mathbf{x}\| < \|\mathbf{X}_i - \mathbf{x}\| + \|\tilde{\mathbf{X}}_j - \mathbf{X}_i\| < d_{(k+1)},$$

which implies that there are at least $(k+1)$ points in the open ball $S_{\mathbf{x}, d_{(k+1)}}$ centered at $\mathbf{x}$ of radius $d_{(k+1)}$, which contradicts the definition of $d_{(k+1)}$.

Accordingly, let us define the random variable B as the number of nearest neighbors $\mathbf{X}_i$'s which belong to $S_{\mathbf{x}, d_{(k+1)}/2}$. Given $d_{(k+1)}$ and defining p_k as

$$p_k = \frac{\mu(S_{\mathbf{x}, d_{(k+1)}/2})}{\mu(S_{\mathbf{x}, d_{(k+1)}})},$$

where μ stands for the law of $\mathbf{X}$, we will justify in the proof of Theorem 1 that B has a binomial distribution with parameters k and p_k. As a consequence, Assertion (1) reads as an inequality between random variables

$$M_k(\mathbf{x}) \geq B.$$

This latter remark is of crucial importance for showing the following consistency results as well as for establishing the rates of convergence of Section 3. We begin with a strong consistency result.

Theorem 1 *Suppose that the distribution μ of $\mathbf{X}$ is absolutely continuous on $\mathbb{R}^d$, that Y is bounded and that the regression function m is μ almost everywhere continuous. If $k \to \infty$, $k/n \to 0$, and $k/\log n \to \infty$, then m_n is strongly consistent, that is*

$$m_n(\mathbf{X}) - m(\mathbf{X}) \to 0,$$

with probability one.

The proof of Theorem 1 reveals that local convergence in probability holds without the assumption that $k/\log n \to \infty$. Indeed, for μ almost every $\mathbf{x}$ and for every $\varepsilon > 0$,

$$\mathbb{P}(|m_n(\mathbf{x}) - m(\mathbf{x})| > \varepsilon) \to 0,$$

when n goes to infinity, provided that $k \to \infty$ and $k/n \to 0$. Since Y is bounded, the weak (i.e., L_2) consistency of Theorem 2 below is just a straightforward consequence of the dominated convergence theorem.

Notice that a standard way to prove the weak consistency of weighted nearest neighbors rules is to check the five conditions of Stone's universal consistency theorem (see Stone, 1977, Theorem 1). As is often the case, one of them is in fact particularly hard to verify in our situation, namely that there exists $C \geq 1$ such that for any nonnegative Borel function f on $\mathbb{R}^d$,

$$\mathbb{E}\left[\sum_{i=1}^{n} W_i f(\mathbf{X}_i)\right] \leq C\, \mathbb{E}\left[f(\mathbf{X})\right].$$

The additional constraints in Theorem 2 are sufficient and are in fact the same as for the layered nearest neighbor estimate studied in Biau and Devroye (2010), as well as for the affine invariant nearest neighbor estimate investigated in Biau et al. (2012).

Theorem 2 *Suppose that the distribution μ of $\mathbf{X}$ is absolutely continuous on $\mathbb{R}^d$, that Y is bounded and that the regression function m is μ almost everywhere continuous. If $k \to \infty$ and $k/n \to 0$, then m_n is weakly consistent, that is*

$$\mathbb{E}[(m_n(\mathbf{X}) - m(\mathbf{X}))^2] \to 0.$$

We may lighten the assumption that $\mathbf{X}$ has a density. Indeed, an inspection of the proof of Theorem 1 indicates that consistency holds as long as for μ almost every $\mathbf{x}$,

$$\liminf_{h \to 0} \frac{\mu(S_{\mathbf{x},h/2})}{\mu(S_{\mathbf{x},h})} > 0.$$

Interestingly, this condition is linked to the notion of *doubling measure* in geometric measure theory. We refer the interested reader to the monographs of Ambrosio and Tilli (2004), Heinonen (2001), and to the paper of Ambrosio et al. (2004).

Recall that the support $\mathcal{S}(\mu)$ is defined as the collection of all $\mathbf{x}$ with $\mu(S_{\mathbf{x},h}) > 0$ for all $h > 0$. In our context, a probability measure μ is said to be doubling on its support $\mathcal{S}(\mu)$ equipped with the Euclidean norm if there exists a constant $c > 0$ such that, for every $\mathbf{x}$ in $\mathcal{S}(\mu)$,

$$\frac{\mu(S_{\mathbf{x},h/2})}{\mu(S_{\mathbf{x},h})} > c, \tag{2}$$

and μ is said to be asymptotically doubling if, for every $\mathbf{x}$ in $\mathcal{S}(\mu)$,

$$\liminf_{h\to 0} \frac{\mu(S_{\mathbf{x},h/2})}{\mu(S_{\mathbf{x},h})} > 0.$$

Thus, we can relax the condition of Theorems 1 and 2 to only requiring that the probability measure μ is asymptotically doubling almost surely. To see that this condition is weaker than requiring a density, note that if μ admits the density f, then a consequence of Lebesgue's differentiation Theorem (see for example Theorem A.10 in Devroye et al. (1996)) is that for μ-almost every $\mathbf{x}$ in $\mathcal{S}(\mu)$,

$$\frac{\mu(S_{\mathbf{x},h/2})}{\mu(S_{\mathbf{x},h})} = \frac{\int_{S_{\mathbf{x},h/2}} f(\mathbf{u})\mathrm{d}\mathbf{u}}{\int_{S_{\mathbf{x},h}} f(\mathbf{u})\mathrm{d}\mathbf{u}} \to \frac{1}{2^d},$$

when h tends to 0. Hence μ is asymptotically doubling almost surely.

It is also readily seen that any discrete probability measure is asymptotically doubling almost surely. Singular continuous probability measures can also be asymptotically doubling as is seen on the following example. Consider the uniform distribution on the standard Cantor ternary set $\mathcal{C}$. Recall that the uniform probability measure μ on $\mathcal{C}$ is the weak limit of the uniform probability measures μ_N on the sets $\mathcal{C}_N$ defined for every integer N as the union of 2^N disjoint intervals with common length 3^{-N}. It is easy to see that for every integer N and for every $\mathbf{x}$ in $\mathcal{C}_N$,

$$\frac{1}{2} \le \frac{\mu_N(S_{\mathbf{x},h/2})}{\mu_N(S_{\mathbf{x},h})} \le 1.$$

Hence, for every $\mathbf{x}$ in $\mathcal{C}$,

$$\liminf_{h\to 0} \frac{\mu(S_{\mathbf{x},h/2})}{\mu(S_{\mathbf{x},h})} \ge \frac{1}{2},$$

and μ is asymptotically doubling almost surely.

Next, we give an example of a singular continuous distribution that is non-asymptotically doubling with probability one. Given a sequence (U_N) of independent Bernoulli distributed random variables with respective parameters $N/(N+1)$, which means that for all $N \ge 1$, $\mathbb{P}(U_N = 1) = N/(N+1)$, define the random variable

$$\mathbf{X} = \sum_{N=1}^{\infty} \frac{2U_N}{3^N}.$$

Note that $\mathbf{X}$ takes values in the standard Cantor ternary set $\mathcal{C}$, but that the law μ of $\mathbf{X}$ is not the uniform law on it: obviously, in the triadic expansion of $\mathbf{X}$, the 2's are much more likely than the

0's. Nevertheless, a direct application of Borel-Cantelli Lemma ensures that μ almost surely, there is an infinite number of 0's in the triadic expansion of $\mathbf{X}$. For such an $\mathbf{x}=\sum_{N=1}^{\infty} 2u_N/3^N$, consider the infinite set of indices

$$I_{\mathbf{x}} = \{N \geq 1 : \ u_N = 0\},$$

and denote μ_N the restriction of μ to the set C_N defined as above, that is, the union of 2^N disjoint intervals with common length 3^{-N}. Then, by construction, for each N in $I_{\mathbf{x}}$, there exists an $h = h_N(\mathbf{x}) \in [1/3^N, 2/3^N]$ such that

$$\frac{\mu_N(S_{\mathbf{x},h/2})}{\mu_N(S_{\mathbf{x},h})} = \frac{1}{N} \ \Rightarrow \ \liminf_{h\to 0} \frac{\mu(S_{\mathbf{x},h/2})}{\mu(S_{\mathbf{x},h})} = 0.$$

Consequently, μ is almost surely not asymptotically doubling. However, even on this pathological probability space, it is not obvious that we can define a regression function m and a distribution for Y such that the mutual nearest neighbors rule would fail to be consistent.

To conclude this section, let us finally notice that, in the context of adaptation to local intrinsic dimension of kNN regression, similar ideas related to the doubling property also appear in a recent paper by Kpotufe (2011).

3. Rates of Convergence

In this section, we are interested in rate of convergence results for the class $\mathcal{F}$ of $(1, C, \rho, \sigma^2)$-smooth distributions $(\mathbf{X}, Y)$ such that $\mathbf{X}$ has compact support with diameter 2ρ, the regression function m is Lipschitz with constant C and, for all $\mathbf{x} \in \mathbb{R}^d$, $\mathbb{V}[Y\,|\,\mathbf{X}=\mathbf{x}] \leq \sigma^2 < \infty$ (the symbol $\mathbb{V}$ denotes variance).

It is known (see, for example, Ibragimov and Khasminskii 1980, 1981, 1982, Stone 1980, 1982, or Györfi et al. 2002) that for the class $\mathcal{F}$, the optimal minimax rate of convergence is $n^{-2/(d+2)}$. In particular, one has that

$$\liminf_{n\to\infty} \inf_{\hat{m}_n} \sup_{(\mathbf{X},Y)\in\mathcal{F}} \frac{\mathbb{E}[\hat{m}_n(\mathbf{X}) - m(\mathbf{X})]^2}{\left((\rho C)^d \sigma^2\right)^{\frac{2}{d+2}} n^{-\frac{2}{d+2}}} \geq \Delta,$$

for some positive constant Δ independent of C, ρ and σ^2. Here the infimum is taken over all estimates $\hat{m}_n$, that is, over all measurable functions of the data.

It turns out that, for $d \geq 2$ and a suitable choice of the sequence (k_n), the MNN estimate m_n achieves the optimum rate for the class $\mathcal{F}$, that is

$$\limsup_{n\to\infty} \sup_{(\mathbf{X},Y)\in\mathcal{F}} \frac{\mathbb{E}[m_n(\mathbf{X}) - m(\mathbf{X})]^2}{\left((\rho C)^d \sigma^2\right)^{\frac{2}{d+2}} n^{-\frac{2}{d+2}}} \leq \Lambda,$$

for some positive Λ independent of C, ρ and σ^2.

Before precisely stating this result, we need an additional notation. Let μ be a probability measure on $\mathbb{R}^d$ with compact support $\mathcal{S}(\mu)$ with diameter 2ρ. We will assume that μ is doubling, as defined in (2), and let

$$p = \inf_{(\mathbf{x},h)\in\mathcal{S}(\mu)\times(0,2\rho]} \frac{\mu(S_{\mathbf{x},h/2})}{\mu(S_{\mathbf{x},h})} > 0. \tag{3}$$

It is readily seen that if μ is absolutely continuous with density f, a sufficient condition is that there exist two strictly positive real numbers a and A such that for almost every $\mathbf{x}$ in $\mathcal{S}(\mu)$, we have $a \leq f(\mathbf{x}) \leq A$.

Theorem 3 *Assume that ties occur with probability 0. Suppose that the law μ of $\mathbf{X}$ has a compact support $\mathcal{S}(\mu)$ with diameter 2ρ, and that μ is doubling, with p defined as in (3). Suppose in addition that, for all $\mathbf{x}$ and $\mathbf{x}' \in \mathbb{R}^d$,*

$$\sigma^2(\mathbf{x}) = \mathbb{V}[Y \,|\, \mathbf{X} = \mathbf{x}] \leq \sigma^2,$$

and

$$\left|m(\mathbf{x}) - m(\mathbf{x}')\right| \leq C\|\mathbf{x} - \mathbf{x}'\|,$$

for some positive constants σ^2 and C. Denote by L_m an upper-bound of the continuous mapping m on the compact $\mathcal{S}(\mu)$. Then

(i) *If $d = 1$,*

$$\mathbb{E}\left[m_n(\mathbf{X}) - m(\mathbf{X})\right]^2 \leq \frac{2\sigma^2}{kp} + \frac{16\rho^2 C^2 k}{n} + L_m^2(1-p)^k.$$

(ii) *If $d = 2$,*

$$\mathbb{E}\left[m_n(\mathbf{X}) - m(\mathbf{X})\right]^2 \leq \frac{2\sigma^2}{kp} + \frac{32\rho^2 C^2}{n} + L_m^2(1-p)^k.$$

(iii) *If $d \geq 3$,*

$$\mathbb{E}\left[m_n(\mathbf{X}) - m(\mathbf{X})\right]^2 \leq \frac{2\sigma^2}{kp} + \frac{8\rho^2 C^2 \lfloor n/k \rfloor^{-2/d}}{1 - 2/d} + L_m^2(1-p)^k.$$

By balancing the terms in Theorem 3, we are led to the following corollary:

Corollary 1 *Under the assumptions of Theorem 3,*

(i) *If $d = 1$, there exists a sequence (k_n) with $k_n \propto \sqrt{n}$ such that*

$$\mathbb{E}\left[m_n(\mathbf{X}) - m(\mathbf{X})\right]^2 \leq (\Lambda + o(1))\frac{\rho C \sigma}{\sqrt{n}},$$

for some positive constant Λ independent of ρ, C and σ^2.

(ii) *If $d \geq 2$, there exists a sequence (k_n) with $k_n \propto n^{\frac{2}{d+2}}$ such that*

$$\mathbb{E}\left[m_n(\mathbf{X}) - m(\mathbf{X})\right]^2 \leq (\Lambda + o(1))\left(\frac{(\rho C)^d \sigma^2}{n}\right)^{\frac{2}{d+2}},$$

for some positive constant Λ independent of ρ, C and σ^2.

Two remarks are in order.

1. We note that, for $d \geq 2$ and a suitable choice of k_n, the MNN estimate achieves both the minimax $n^{-2/(d+2)}$ rate and the optimal order of magnitude $((\rho C)^d \sigma^2)^{2/(d+2)}$ in the constant, for the class $\mathcal{F}$ of $(1, C, \rho, \sigma^2)$-smooth distributions $(\mathbf{X}, Y)$ such that $\mathbf{X}$ has compact support with covering radius ρ, the regression function m is Lipschitz with constant C and, for all $\mathbf{x} \in \mathbb{R}^d$, $\mathbb{V}[Y \,|\, \mathbf{X} = \mathbf{x}] \leq \sigma^2$.

2. For $d = 1$, the obtained rate is not optimal. This low-dimensional phenomenon is also known to hold for the traditional kNN regression estimate, which does not achieve the optimal rate in dimension 1 (see Problem 6.1 in Györfi et al. 2002).

In Corollary 1, the parameter k_n of the estimate with the optimal rate of convergence for the class $\mathcal{F}$ depends on the unknown distribution of $(\mathbf{X}, Y)$, especially on the smoothness of the regression function as measured by the Lipschitz constant C. To conclude this section, we present a data-dependent way for choosing the resampling size k_n and show that, for bounded Y, the estimate with parameter chosen in such an adaptive way achieves the optimal rate of convergence.

To this end, we split the sample $\mathcal{D}_n = \{(\mathbf{X}_1, Y_1), \ldots, (\mathbf{X}_n, Y_n)\}$ in two parts, denoted by $\mathcal{D}_n^\ell$ (learning set) and $\mathcal{D}_n^t$ (testing set), of size $\lfloor n/2 \rfloor$ and $n - \lfloor n/2 \rfloor$, respectively. The first half is used to construct the MNN estimate

$$m_{\lfloor n/2 \rfloor}(\mathbf{x}, \mathcal{D}_n^\ell) = m_{k, \lfloor n/2 \rfloor}(\mathbf{x}, \mathcal{D}_n^\ell).$$

The second half is used to choose k by picking $\hat{k}_n \in \mathcal{K} = \{1, \ldots, \lfloor n/2 \rfloor\}$ to minimize the empirical risk

$$\frac{1}{n - \lfloor n/2 \rfloor} \sum_{i=\lfloor n/2 \rfloor + 1}^{n} \left(Y_i - m_{k, \lfloor n/2 \rfloor}(\mathbf{X}_i, \mathcal{D}_n^\ell) \right)^2.$$

Define the estimate

$$m_n(\mathbf{x}) = m_{\hat{k}_n, \lfloor n/2 \rfloor}(\mathbf{x}, \mathcal{D}_n^\ell),$$

and note that m_n depends on the entire data $\mathcal{D}_n$. If $|Y| \leq L < \infty$ almost surely, a straightforward adaptation of Theorem 7.1 in Györfi et al. (2002) shows that, for any $\delta > 0$,

$$\mathbb{E}[m_n(\mathbf{X}) - m(\mathbf{X})]^2 \leq (1 + \delta) \inf_{k \in \mathcal{K}} \mathbb{E}[m_{k, \lfloor n/2 \rfloor}(\mathbf{X}, \mathcal{D}_n^\ell) - m(\mathbf{X})]^2 + \Xi \frac{\ln n}{n},$$

for some positive constant Ξ depending only on L, d and δ. Immediately from Corollary 1, we can conclude:

Theorem 4 *Suppose that $|Y| \leq L$ almost surely, and let m_n be the MNN estimate with $k \in \mathcal{K} = \{1, \ldots, \lfloor n/2 \rfloor\}$ chosen by data-splitting. Then the condition $(\ln n)^{(d+2)/(2d)} n^{-1/2} \leq \rho C$ together with $d \geq 2$ implies*

$$\mathbb{E}[m_n(\mathbf{X}) - m(\mathbf{X})]^2 \leq (\Lambda + \mathrm{o}(1)) \left(\frac{(\rho C)^d}{n} \right)^{\frac{2}{d+2}},$$

for some positive constant Λ which depends only on L and d.

Thus, the expected error of the estimate obtained via data-splitting is bounded from above up to a constant by the corresponding minimax lower bound for the class $\mathcal{F}$ of regression functions, with the optimal dependence in C and ρ.

4. Proofs

Proofs of the main results are gathered in this section.

4.1 Proof of Theorem 1

Let us fix $\varepsilon > 0$ and $\mathbf{x}$ in $\mathcal{S}(\mu)$ such that m is continuous at $\mathbf{x}$. Setting

$$\tilde{m}_n(\mathbf{x}) = \sum_{i=1}^{n} W_i\, m(\mathbf{X}_i),$$

we have

$$\begin{aligned}\mathbb{P}\left(|m_n(\mathbf{x}) - m(\mathbf{x})| > 2\varepsilon\right) \leq \mathbb{P}&\left(M_k(x) < \frac{k}{2^{d+1}}\right)\\ &+\mathbb{P}\left(|m_n(\mathbf{x}) - \tilde{m}_n(\mathbf{x})| > \varepsilon, M_k(x) \geq \frac{k}{2^{d+1}}\right)\\ &+\mathbb{P}\left(|\tilde{m}_n(\mathbf{x}) - m(\mathbf{x})| > \varepsilon, M_k(x) \geq \frac{k}{2^{d+1}}\right).\end{aligned} \tag{4}$$

First, remark that rearranging the k (ordered) statistics $\mathbf{X}_{(1,n)}, \dots, \mathbf{X}_{(k,n)}$ in the original order of their outcome, one obtains the k (non-ordered) random variables $\mathbf{X}_1^\star, \dots, \mathbf{X}_k^\star$. Let $\tilde{\mathbf{X}}_1, \dots, \tilde{\mathbf{X}}_k$ be i.i.d. random variables, with common law (conditional on $d_{(k+1)}$) the restriction $\tilde{\mu}$ of μ to the open ball $S_{\mathbf{x},d_{(k+1)}}$, then it can be shown (see for example Lemma A.1 in Cérou and Guyader 2006) that

$$\mathcal{L}(\mathbf{X}_1^\star, \dots, \mathbf{X}_k^\star | d_{(k+1)}) = \mathcal{L}(\tilde{\mathbf{X}}_1, \dots, \tilde{\mathbf{X}}_k). \tag{5}$$

Next, given $d_{(k+1)}$, denote

$$p_k = \mathbb{P}\left(\|\tilde{\mathbf{X}} - \mathbf{x}\| < \frac{d_{(k+1)}}{2} \,\middle|\, \tilde{\mathbf{X}} \sim \tilde{\mu}\right) = \frac{\int_{S_{\mathbf{x},d_{(k+1)}/2}} f(\mathbf{u})\mathrm{d}\mathbf{u}}{\int_{S_{\mathbf{x},d_{(k+1)}}} f(\mathbf{u})\mathrm{d}\mathbf{u}},$$

where the denominator is strictly positive since $\mathbf{x}$ belongs to the support of μ. Concerning p_k, recall that Lebesgue's differentiation Theorem ensures that for λ-almost all $\mathbf{x} \in \mathbb{R}^d$,

$$\frac{1}{\lambda(S_{x,\delta})} \int_{S_{x,\delta}} f(\mathbf{u})\mathrm{d}\mathbf{u} \to f(\mathbf{x}),$$

when δ tends to 0 (see for example Theorem A.10 in Devroye et al. 1996). Notice that for μ almost every $\mathbf{x}$ in the support of μ, we have $f(\mathbf{x}) > 0$. Consequently, since $\lambda(S_{\mathbf{x},h}) = V_d h^d$ with V_d the volume of the unit ball of $\mathbb{R}^d$, we have that for μ-almost every $\mathbf{x}$ in $\mathbb{R}^d$,

$$p(\delta) := \frac{\int_{S_{\mathbf{x},\delta/2}} f(\mathbf{u})\mathrm{d}\mathbf{u}}{\int_{S_{\mathbf{x},\delta}} f(\mathbf{u})\mathrm{d}\mathbf{u}} \to \frac{1}{2^d}, \tag{6}$$

when h tends to 0. Hence, let us choose $\delta_0 > 0$ such that

$$\delta \in (0, \delta_0] \;\Rightarrow \left|p(\delta) - \frac{1}{2^d}\right| < \frac{1}{2^{d+2}}.$$

Then we may write

$$\mathbb{P}\left(M_k(x) < \frac{k}{2^{d+1}}\right) \leq \mathbb{P}\left(M_k(x) < \frac{k}{2^{d+1}}, d_{(k+1)} \leq \delta_0\right) + \mathbb{P}(d_{(k+1)} > \delta_0).$$

Denoting

$$q_0 = \mathbb{P}(\|\mathbf{X}-\mathbf{x}\| \leq \delta_0) = \int_{S_{\mathbf{x},\delta_0}} f(\mathbf{u})\mathrm{d}\mathbf{u},$$

we have $q_0 > 0$. Following the proof of Lemma 4 in Devroye (1982), denote Z a binomial (n, q_0) random variable. If $k/n \to 0$, then for n large enough, Hoeffding's inequality yields

$$\mathbb{P}(d_{(k+1)} > \delta_0) \leq \mathbb{P}(Z < k+1) \leq \mathbb{P}\left(Z - nq_0 < -\frac{nq_0}{2}\right) \leq e^{-nq_0^2/2},$$

which is summable in n for all $\delta_0 > 0$. Next, observe that

$$\mathbb{P}\left(M_k(x) < \frac{k}{2^{d+1}}, d_{(k+1)} \leq \delta_0\right)$$
$$= \int_0^{\delta_0} \mathbb{P}\left(M_k(x) < \frac{k}{2^{d+1}} \middle| d_{(k+1)} = \delta\right) \mathrm{dP}_{d_{(k+1)}}(\delta).$$

Given δ and defining B as the number of $\mathbf{X}_i$'s among the k nearest neighbors of $\mathbf{x}$ which belong to $S_{\mathbf{x},\delta/2}$, then according to (5), the random variable B has binomial distribution $\mathcal{B}(k, p(\delta))$ and (1) implies $M_k(\mathbf{x}) \geq B$, so that

$$\mathbb{P}\left(M_k(\mathbf{x}) < \frac{k}{2^{d+1}} \middle| d_{(k+1)} = \delta\right) \leq \mathbb{P}\left(B < \frac{k}{2^{d+1}} \middle| p(\delta)\right).$$

In this respect, Hoeffding's inequality and (6) lead to

$$\mathbb{P}\left(B < \frac{k}{2^{d+1}} \middle| p(\delta)\right) \leq \exp\left(-2\left(p(\delta) - \frac{1}{2^{d+1}}\right)^2 k\right) \leq \exp\left(-\frac{k}{2^{2d+3}}\right),$$

which is summable in n provided that $k/\log n \to \infty$.

Let us turn now to second term of (4). This time, we write

$$\mathbb{P}\left(|m_n(\mathbf{x}) - \tilde{m}_n(\mathbf{x})| > \varepsilon, M_k(\mathbf{x}) \geq \frac{k}{2^{d+1}}\right)$$
$$= \mathbb{E}\left[\mathbf{1}_{\{M_k(\mathbf{x}) \geq \frac{k}{2^{d+1}}\}} \mathbb{P}\left(|m_n(\mathbf{x}) - \tilde{m}_n(\mathbf{x})| > \varepsilon | \mathbf{X}_1, \ldots, \mathbf{X}_n\right)\right]$$
$$= \mathbb{E}\left[\mathbf{1}_{\{M_k(\mathbf{x}) \geq \frac{k}{2^{d+1}}\}} \mathbb{P}\left(\left|\sum_{i=1}^n W_i(Y_i - m(\mathbf{X}_i))\right| > \varepsilon \middle| \mathbf{X}_1, \ldots, \mathbf{X}_n\right)\right].$$

Given $\mathbf{X}_1, \ldots, \mathbf{X}_n$, the random variables $Y_1 - m(\mathbf{X}_1), \ldots, Y_n - m(\mathbf{X}_n)$ are independent, centered, and bounded by $2L$. Moreover, the weights $W_1, \ldots, W_n$ are deterministic and, since $M_k(\mathbf{x}) \geq k/2^{d+1}$, bounded by $2^{d+1}/k$. Consequently, Lemma 6 in Devroye (1982) leads to

$$\mathbb{P}\left(|m_n(\mathbf{x}) - \tilde{m}_n(\mathbf{x})| > \varepsilon, M_k(\mathbf{x}) \geq \frac{k}{2^{d+1}}\right) \leq 2\exp\left(-\frac{k\varepsilon^2}{2^{d+3}(2L^2 + L\varepsilon)}\right),$$

which is summable in n for all $\varepsilon > 0$, provided that $k/\log n \to \infty$.

The last term of (4) is easier. First we notice that, since m is assumed continuous at point $\mathbf{x}$, there exists $\delta_1 = \delta_1(\varepsilon)$ such that

$$\|\mathbf{x}' - \mathbf{x}\| \leq \delta_1 \Rightarrow \left|m(\mathbf{x}') - m(\mathbf{x})\right| \leq \varepsilon.$$

The following inequalities are then straightforward

$$\begin{aligned}
&\mathbb{P}\left(|\tilde{m}_n(\mathbf{x})-m(\mathbf{x})|>\varepsilon, M_k(x)\geq\frac{k}{2^{d+1}}\right)\\
&\quad=\mathbb{P}\left(\left|\sum_{i=1}^{n}W_i(m(\mathbf{X}_i)-m(\mathbf{x}))\right|>\varepsilon, M_k(x)\geq\frac{k}{2^{d+1}}\right)\\
&\quad\leq\mathbb{P}\left(\max_{1\leq i\leq k}\left|m(\mathbf{X}_{(i)})-m(\mathbf{x})\right|>\varepsilon\right)\\
&\quad\leq\mathbb{P}\left(\left\|\mathbf{X}_{(k)}-\mathbf{x}\right\|>\delta_1\right),
\end{aligned}$$

and the same reasoning as before yields

$$\mathbb{P}\left(|\tilde{m}_n(\mathbf{x})-m(\mathbf{x})|>\varepsilon, M_k(x)\geq\frac{k}{2^{d+1}}\right)\leq e^{-nq_1^2/2},$$

where

$$q_1=q_1(\varepsilon)=\mathbb{P}(\|\mathbf{X}-\mathbf{x}\|\leq\delta_1)=\int_{S_{\mathbf{x},\delta_1}}f(\mathbf{u})\mathrm{d}\mathbf{u}.$$

Putting all things together, we have proved that for any $\varepsilon>0$, if $k/n\to 0$, then for n large enough we have

$$\begin{aligned}
&\mathbb{P}\left(|m_n(\mathbf{x})-m(\mathbf{x})|>2\varepsilon\right)\\
&\leq 2\exp\left(\frac{-k\varepsilon^2}{2^{d+3}(2L^2+L\varepsilon)}\right)+\exp\left(\frac{-k}{2^{2d+3}}\right)+\exp\left(\frac{-nq_0^2}{2}\right)+\exp\left(\frac{-nq_1^2}{2}\right),
\end{aligned}$$

which is summable in n for all $\varepsilon>0$, provided that $k/\log n\to\infty$. Since this is true for μ almost every $\mathbf{x}$, the strong consistency is established.

4.2 Proof of Theorem 3

As previously, setting

$$\tilde{m}_n(\mathbf{x})=\sum_{i=1}^{n}W_i\,m(\mathbf{X}_i),$$

the proof of Theorem 3 will rely on the variance/bias decomposition

$$\mathbb{E}\left[m_n(\mathbf{x})-m(\mathbf{x})\right]^2=\mathbb{E}\left[m_n(\mathbf{x})-\tilde{m}_n(\mathbf{x})\right]^2+\mathbb{E}\left[\tilde{m}_n(\mathbf{x})-m(\mathbf{x})\right]^2. \tag{7}$$

The first term is easily bounded by noting that, for all $\mathbf{x}$ in $\mathbb{R}^d$,

$$\begin{aligned}
\mathbb{E}\left[m_n(\mathbf{x})-\tilde{m}_n(\mathbf{x})\right]^2 &= \mathbb{E}\left[\sum_{i=1}^{n} W_i\left(Y_i-m(\mathbf{X}_i)\right)\right]^2 \\
&= \mathbb{E}\left[\sum_{i=1}^{n} W_i^2\left(Y_i-m(\mathbf{X}_i)\right)^2\right] \\
&= \mathbb{E}\left[\sum_{i=1}^{n} W_i^2 \mathbb{E}\left[\left(Y_i-m(\mathbf{X}_i)\right)^2 \middle| \mathbf{X}_i\right]\right] \\
&= \mathbb{E}\left[\sum_{i=1}^{n} W_i^2 \sigma^2\left(\mathbf{X}_i\right)\right] \\
&\leq \sigma^2\, \mathbb{E}\left[\sum_{i=1}^{n} W_i^2\right].
\end{aligned}$$

With the convention that 0/0=0, notice that by definition of the weights W_i,

$$\mathbb{E}\left[\sum_{i=1}^{n} W_i^2\right] = \mathbb{E}\left[\frac{1}{M_k(\mathbf{X})}\mathbf{1}_{M_k(\mathbf{X})\neq 0}\right].$$

As in the proof of Theorem 1, given $d_{(k+1)}$, denote

$$p_k = \mathbb{P}\left(\|\tilde{\mathbf{X}}-\mathbf{x}\| < \frac{d_{(k+1)}}{2} \middle| \tilde{\mathbf{X}} \sim \tilde{\mu}\right) = \frac{\mu\left(S_{\mathbf{x},d_{(k+1)}/2}\right)}{\mu\left(S_{\mathbf{x},d_{(k+1)}}\right)},$$

and define B as the number of $\mathbf{X}_i$'s among the k nearest neighbors of $\mathbf{x}$ which belong to $S_{\mathbf{x},d_{(k+1)}/2}$. Then, given p_k, the random variable B has binomial distribution $\mathcal{B}(k,p_k)$, and (1) implies

$$M_k(\mathbf{x}) \geq B. \tag{8}$$

In particular,

$$\frac{1}{M_k(\mathbf{x})}\mathbf{1}_{M_k(\mathbf{X})\neq 0} \leq \frac{2}{1+M_k(\mathbf{x})} \leq \frac{2}{1+B},$$

so that

$$\mathbb{E}\left[\sum_{i=1}^{n} W_i^2\right] \leq \mathbb{E}\left[\mathbb{E}\left[\frac{2}{1+B}\middle| p_k\right]\right] = 2\mathbb{E}\left[\frac{1-(1-p_k)^k}{(k+1)p_k}\right].$$

Since $p_k \geq p$, we are led to

$$\mathbb{E}\left[m_n(\mathbf{x})-\tilde{m}_n(\mathbf{x})\right]^2 \leq \frac{2\sigma^2}{kp},$$

and integrating with respect to the distribution of $\mathbf{X}$ yields

$$\mathbb{E}\left[m_n(\mathbf{X})-\tilde{m}_n(\mathbf{X})\right]^2 \leq \frac{2\sigma^2}{kp}.$$

Concerning the bias term in (7), again fix $\mathbf{x}$ in $\mathbb{R}^d$, denote by L_m an upper-bound of the continuous function m on the compact $\mathcal{S}(\mu)$, and write

$$\mathbb{E}\left[\tilde{m}_n(\mathbf{x}) - m(\mathbf{x})\right]^2 \leq \mathbb{E}\left[(\tilde{m}_n(\mathbf{x}) - m(\mathbf{x}))^2 \mathbf{1}_{\{M_k(\mathbf{x})>0\}}\right] + L_m^2\, \mathbb{P}(M_k(\mathbf{x}) = 0).$$

The second term is bounded thanks to (8),

$$\mathbb{P}(M_k(\mathbf{x}) = 0|p_k) \leq \mathbb{P}(B = 0|p_k) = (1-p_k)^k,$$

so that

$$\mathbb{P}(M_k(\mathbf{x}) = 0) \leq \mathbb{E}\left[(1-p_k)^k\right],$$

and since $p_k \geq p$,

$$\mathbb{P}(M_k(\mathbf{x}) = 0) \leq \mathbb{E}\left[(1-p_k)^k\right] \leq (1-p)^k.$$

For the first term, with the convention $0/0 = 0$, one has

$$\begin{aligned}
&\mathbb{E}\left[(\tilde{m}_n(\mathbf{x}) - m(\mathbf{x}))^2 \mathbf{1}_{\{M_k(\mathbf{x})>0\}}\right] \\
&\quad = \mathbb{E}\left[\left(\frac{1}{M_k(\mathbf{x})} \sum_{i:\mathbf{X}_i \in \mathcal{M}_k(\mathbf{x})} (m(\mathbf{X}_i) - m(\mathbf{x}))\right)^2 \mathbf{1}_{\{M_k(\mathbf{x})>0\}}\right] \\
&\quad \leq C^2 \mathbb{E}\left[\left(\frac{1}{M_k(\mathbf{x})} \sum_{i:\mathbf{X}_i \in \mathcal{M}_k(\mathbf{x})} \|\mathbf{X}_i - \mathbf{x}\|\right)^2 \mathbf{1}_{\{M_k(\mathbf{x})>0\}}\right].
\end{aligned}$$

Next we apply Jensen's inequality to get

$$\mathbb{E}\left[(\tilde{m}_n(\mathbf{x}) - m(\mathbf{x}))^2 \mathbf{1}_{\{M_k(\mathbf{x})>0\}}\right] \leq C^2 \mathbb{E}\left[\frac{\mathbf{1}_{\{M_k(\mathbf{x})>0\}}}{M_k(\mathbf{x})} \sum_{i:\mathbf{X}_i \in \mathcal{M}_k(\mathbf{x})} \|\mathbf{X}_i - \mathbf{x}\|^2\right].$$

Since any mutual nearest neighbor of $\mathbf{x}$ belongs to its k nearest neighbors, we deduce

$$\mathbb{E}\left[(\tilde{m}_n(\mathbf{x}) - m(\mathbf{x}))^2 \mathbf{1}_{\{M_k(\mathbf{x})>0\}}\right] \leq C^2 \mathbb{E}\left[\left\|\mathbf{X}_{(k,n)} - \mathbf{x}\right\|^2\right].$$

Therefore, by integrating with respect to the distribution of $\mathbf{X}$, we obtain the following upper-bound for the bias term

$$\mathbb{E}\left[\tilde{m}_n(\mathbf{X}) - m(\mathbf{X})\right]^2 \leq C^2 \mathbb{E}\left[\|\mathbf{X}_{(k,n)} - \mathbf{X}\|^2\right] + L_m^2 (1-p)^k.$$

Next, let us denote

$$\rho = \inf\left\{r > 0 \ : \ \exists\, \mathbf{x}_0 \in \mathbb{R}^d \text{ such that } \mathcal{S}(\mu) \subset S_{\mathbf{x}_0,r}\right\},$$

and notice that 2ρ is an upper-bound of the diameter of $\mathcal{S}(\mu)$. Then we are in a position to apply Proposition 2.3 in Biau et al. (2010), that is for $d = 1$,

$$\mathbb{E}\left[\|\mathbf{X}_{(k,n)} - \mathbf{X}\|^2\right] \leq \frac{16\rho^2 k}{n},$$

and for $d \geq 3$,

$$\mathbb{E}\left[\|\mathbf{X}_{(k,n)} - \mathbf{X}\|^2\right] \leq \frac{8\rho^2 \lfloor n/k \rfloor^{-\frac{2}{d}}}{1-2/d}.$$

It turns out that, for $d = 2$, the bound given in Biau et al. (2010) is not optimal, since it leads to

$$\mathbb{E}\left[\|\mathbf{X}_{(k,n)} - \mathbf{X}\|^2\right] \leq \frac{8\rho^2 k}{n}\left(1 + \log\frac{n}{k}\right),$$

whereas Theorem 3.2 in Liitiäinen et al. (2010) allows to get rid of the logarithmic term. Namely, the application of their result in our context leads to

$$\mathbb{E}\left[\|\mathbf{X}_{(k,n)} - \mathbf{X}\|^2\right] \leq \frac{32\rho^2 k}{n}.$$

This terminates the proof of Theorem 3.

Acknowledgments

Arnaud Guyader wish to thank Hervé Jégou to have made him aware of the mutual nearest neighbors rule, and Frédéric Cérou for helpful discussions on the second version of the article. The authors are greatly indebted to Luigi Ambrosio and Giovanni Alberti for fruitful discussions on doubling measure theory. They also thank three referees and the Action Editor for their thoughtful critical comments.

References

L. Ambrosio and P. Tilli. *Topics on Analysis in Metric Spaces*. Oxford University Press, Oxford, 2004.

L. Ambrosio, M. Miranda, Jr., and D. Pallara. Special functions of bounded variation in doubling metric measure spaces. In *Calculus of Variations: Topics from the Mathematical Heritage of E. De Giorgi*, volume 14 of *Quad. Mat.*, pages 1–45. Dept. Math., Seconda Univ. Napoli, Caserta, 2004.

G. Biau and L. Devroye. On the layered nearest neighbour estimate, the bagged nearest neighbour estimate and the random forest method in regression and classification. *Journal of Multivariate Analysis*, 101(10):2499–2518, 2010.

G. Biau, F. Cérou, and A. Guyader. On the rate of convergence of the bagged nearest neighbor estimate. *Journal of Machine Learning Research (JMLR)*, 11:687–712, 2010.

G. Biau, L. Devroye, V. Dujmović, and A. Krzyżak. An affine invariant *k*-nearest neighbor regression estimate. *Journal of Multivariate Analysis*, 112:24–34, 2012.

F. Cérou and A. Guyader. Nearest neighbor classification in infinite dimension. *ESAIM: Probability and Statistics*, 10:340–355, 2006.

K. Chidananda Gowda and G. Krishna. Agglomerative clustering using the concept of mutual nearest neighbourhood. *Pattern Recognition*, 10(2):105–112, 1978.

K. Chidananda Gowda and G. Krishna. The condensed nearest neighbor rule using the concept of mutual nearest neighborhood. *IEEE Transactions on Information Theory*, 25(4), 1979.

L. Devroye. Necessary and sufficient conditions for the pointwise convergence of nearest neighbor regression function estimates. *Zeitschrift für Wahrscheinlichkeitstheorie und Verwandte Gebiete*, 61(4):467–481, 1982.

L. Devroye, L. Györfi, and G. Lugosi. *A Probabilistic Theory of Pattern Recognition.* Springer-Verlag, New York, 1996.

E. Fix and J.L. Hodges. Discriminatory analysis, non-parametric discrimination: consistency properties. Technical report, USAF school of aviation and medicine, Randolph Field, 1951.

E. Fix and J.L. Hodges. Discriminatory analysis: Small sample performance. Technical report, USAF school of aviation and medicine, Randolph Field, 1952.

L. Györfi, M. Kohler, A. Krzyżak, and H. Walk. *A Distribution-Free Theory of Nonparametric Regression*. Springer-Verlag, New York, 2002.

J. Heinonen. *Lectures on Analysis on Metric Spaces*. Universitext. Springer-Verlag, New York, 2001.

I.A. Ibragimov and R.Z. Khasminskii. Nonparametric regression estimation. *Doklady Akademii Nauk SSSR*, 252(4):780–784, 1980.

I.A. Ibragimov and R.Z. Khasminskii. *Statistical Estimation: Asymptotic Theory*. Springer-Verlag, New York, 1981.

I.A. Ibragimov and R.Z. Khasminskii. Bounds for the quality of nonparametric estimation of regression. *Akademiya Nauk SSSR. Teoriya Veroyatnosteĭ i ee Primeneniya*, 27(1):81–94, 1982.

H. Jégou, C. Schmid, H. Harzallah, and J. Verbeek. Accurate image search using the contextual dissimilarity measure. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 32(1): 2–11, 2010.

S. Kpotufe. *k*-NN regression adapts to local intrinsic dimension. In *NIPS Proceedings*, pages 729–737, 2011.

E. Liitiäinen, F. Corona, and A. Lendasse. Residual variance estimation using a nearest neighbor statistic. *Journal of Multivariate Analysis*, 101(4):811–823, 2010.

H. Liu, S. Zhang, J. Zhao, X. Zhao, and Y. Mo. A new classification algorithm using mutual nearest neighbors. In *Conference on Grid and Cloud Computing*, pages 52–57, 2010.

D. Qin, S. Gammeter, L. Bossard, T. Quack, and L.J. Van Gool. Hello neighbor: accurate object retrieval with *k*-reciprocal nearest neighbors. In *Computer Vision and Pattern Recognition*, pages 777–784, 2011.

M. Radovanović, A. Nanopoulos, and M. Ivanović. Hubs in space: popular nearest neighbors in high-dimensional data. *Journal of Machine Learning Research (JMLR)*, 11:2487–2531, 2010.

C.J. Stone. Consistent nonparametric regression. *The Annals of Statistics*, 5(4):595–645, 1977.

C.J. Stone. Optimal rates of convergence for nonparametric estimators. *The Annals of Statistics*, 8 (6):1348–1360, 1980.

C.J. Stone. Optimal global rates of convergence for nonparametric regression. *The Annals of Statistics*, 10(4):1040–1053, 1982.