Electronic Journal of Statistics
Vol. 5 (2011) 31–40
ISSN: 1935-7524
DOI: 10.1214/11-EJS595

Convergence of functional k-nearest neighbor regression estimate with functional responses

Heng Lian*

Division of Mathematical Sciences
School of Physical and Mathematical Sciences
Nanyang Technological University
Singapore 637371
e-mail: henglian@ntu.edu.sg

Abstract: Let $(X_1, Y_1), \ldots, (X_n, Y_n)$ be independent and identically distributed random elements taking values in $\mathcal{F} \times \mathcal{H}$, where $\mathcal{F}$ is a semi-metric space and $\mathcal{H}$ is a separable Hilbert space. We investigate the rates of strong (almost sure) convergence of the k-nearest neighbor estimate. We give two convergence results assuming a finite moment condition and exponential tail condition on the noises respectively, with the latter requiring less stringent conditions on k for convergence.

AMS 2000 subject classifications: Primary 62G08; secondary 62G20.
Keywords and phrases: Functional response models, martingale difference sequence, nearest neighbor estimate, rates of convergence.

Received November 2010.

1. Introduction

Let $(\mathcal{F}, d(.,.))$ be a semi-metric space, $(\mathcal{H}, \langle\cdot\rangle)$ a separable Hilbert space, and let $(X, Y), (X_1, Y_1), (X_2, Y_2), \ldots, (X_n, Y_n)$ be independent identically distributed $\mathcal{F} \times \mathcal{H}$-valued random pairs. In regression analysis, usually an estimate of the function $m(x) = E(Y|X = x)$ is being sought using n pairs of data points.

In the literature, two related classes of nonparametric estimates have been proposed. The first one is the Nadaraya-Watson estimate or kernel estimate [20, 16], with the well-known drawback that it ignores the local denseness/sparseness of the data and uses a fixed bandwidth parameter on the entire predictor space. The k-nearest neighbor (k-NN) method addresses this problem by using adaptive neighborhood size based on the distance of a point from its neighbors [5, 4, 13].

In the classical setting, the observation pairs reside in the Euclidean spaces. In particular, $\mathcal{F} = R^d$ and $\mathcal{H} = R$ is the most common and most studied case in the statistical literature. With the increasing interest at the present moment in many fields of statistics in which the observations are curves, such as speech recordings, weather data, commodity prices, functional regression analysis as an extension of classical setting has risen to the center stage of statistical research. Two major

*Research supported by Singapore Ministry of Education Tier 1 Grant 36/09

approaches exist for functional data analysis. The parametric modeling approach was masterfully documented in the monograph [19], and the nonparametric approach was proposed in the pioneering work [9] and also popularized by the book [11]. Another nonparametric approach is based on the reproducing kernel Hilbert spaces framework [18, 15].

For some applications, the dependent variable takes values in a more general space than finite-dimensional Euclidean spaces. For example, one might predict annual precipitation using temperature measurements [19], or predict future hourly electricity consumption based on past history [1]. In this note we investigate the convergence rates of functional k-NN estimate when the regression output takes values in a general separable Hilbert space $\mathcal{H}$. Although it is conceptually straightforward to apply k-NN method in this context, the demonstration of its asymptotic properties poses technical difficulties due to the functional responses.

This work can be regarded as an extension of [3] where k-NN method in functional regression with scalar responses is studied. For functional responses, the theoretical investigation involves extra complications. Besides, we use a slightly more general setup (in terms of weights v_{ni} defined in the next section) and also emphasize the role of the assumption on errors.

During the final stage of preparation for this manuscript, the author learned that Dr. Frederic Ferraty and his collaborators have recently obtained corresponding results with functional responses, although in the context of Nadaraya-Watson kernel regression. On the one hand, they used the stronger assumption on the noise (similarly to our Assumption 4 below) while we also obtained rates under finite moment assumption (as in our Assumption 3). On the other hand, they studied inferences using bootstrap and we did not investigate the inference problems here.

2. Estimation and rates of convergence

Consider the simple additive noise model $Y = m(X) + \epsilon$ where ϵ takes values in $\mathcal{H}$, has mean zero (in the sense of Bochner integral, see [14]), and is independent of covariate X. Given n copies of independent observations $\mathcal{D}_n = \{(X_1, Y_1), \ldots, (X_n, Y_n)\}$, the k-NN estimate at any $x \in \mathcal{F}$ is defined by

$$\hat{m}(x) = \sum_{i=1}^{n} v_{ni} Y_i \,, \tag{2.1}$$

where $(v_{n1}, \ldots, v_{nn})$ is a (possibly random) probability vector. Note we consider estimation and convergence at a fixed x and thus we sometimes omit explicitly stating the fixed covariate. For example, a nearest neighbor always refers to the nearest neighbor of a fixed x. Two specific examples of v_{ni} follow.

Example 1. Take $v_{ni} = a_{nj}$ if X_i is the $j-$th nearest neighbor, with $a_{n1} \geq a_{n2} \geq \cdots \geq a_{nn}$ a deterministic probability vector, thus putting more weights in (2.1) for data closer to x. Setting $a_{nj} = 1/k$ if $j \leq k$ and 0 otherwise gives us

back the simple k-NN estimate. We should note that even in this simplest case, v_{ni} depends not just on X_i since all $X_j, j \leq n$ together determine the identities of x's nearest neighbors, which leads to some complications in theoretical analysis.

Example 2. Take $v_{ni} = K(d(X_i, x)/H)/\sum_j K(d(X_j, x)/H)$ where K is a kernel function and H is the distance of the $k-$th nearest neighbor. Mathematically,

$$H = \min\{h \in R : \sum_{i=1}^{n} I\{X_i \in B(x, h)\} \geq k\}, \tag{2.2}$$

where $B(x, h) = \{x' \in \mathcal{F} : d(x', x) \leq h\}$ and $I\{.\}$ denotes the indicator function. For simplicity we only consider the case where the kernel function K is compactly supported and nonincreasing on $[0, 1]$.

Naturally we need the following assumption on the regression function to obtain meaningful rates of convergence.

Assumption 1. *m is bounded and Lipschitz continuous at x, that is, $||m(x)|| \leq B, \forall x \in \mathcal{F}$ and $||m(x) - m(x')|| \leq Md(x, x')^{\alpha}$. The Lipschitz condition only needs to be satisfied locally on an open neighborhood of the fixed x.*

In the following theoretical investigations, we directly take $v_{n1} \geq v_{n2} \geq \cdots \geq v_{nn}$. For our two examples above, this amounts to assuming that the n data pairs have already been ordered according to the distance of X_i to x so that X_1 is the nearest neighbor of x, for example (ties are broken by comparing indices in the original sequence). We assume such reordering has been performed throughout. We need the following conditions on v_{ni}.

Assumption 2. *Suppose $\sum_{i=k+1}^{n} v_{ni} = O(b_n)$ and denote $\|v\|_s = (\sum_{i=1}^{n} v_{ni}^s)^{1/s}$, we assume $b_n \to 0, \|v\|_2 \to 0$, where the asymptotic orders are in the sense of almost sure convergence. We also require that $k/n \to 0$ and $k/\log n \to \infty$.*

Some moment conditions are necessary on the norm of the noise also.

Assumption 3. *$E||\epsilon||^r < \infty$ for some $r > 2$.*

As an alternative, we can instead impose a stronger exponential tail condition.

Assumption 4. *$P(||\epsilon|| > a) \leq exp\{-Ca^p\}$ with $C > 0$ and $p > 0$, for any $a > 0$.*

Although Assumption 4 is much stronger than finite moment condition in Assumption 3, it is satisfied by many Gaussian processes, whose norm typically exhibits sub-Gaussian tails (see for example the Appendix of [23]) and thus satisfies this assumption with $p = 2$.

Our convergence results below are stated in terms of the critical quantity $\phi(h) := P(\{x' : x' \in B(x, h)\})$ which is called the small ball probability. Its importance has been demonstrated in [10, 11, 8] for functional kernel regression. The quantity $\phi(h)$ is closely related to the $\epsilon-$covering number of the Banach space $\mathcal{F}$, which is defined as the smallest number of open balls of radius ϵ that cover the set $\mathcal{F}$. A set with finite $\epsilon-$covering number for all $\epsilon > 0$ is called totally

bounded. For our purpose, since we are interested in the regression function at a fixed $x \in \mathcal{F}$, the global property of total boundedness is not necessary. However, if we assume a uniform small ball probability over $\mathcal{F}$, that is $c\psi(h) \leq P(B(x,h)) \leq C\psi(h)$ for some positive increasing function ψ independent of x, then it automatically implies total boundedness of $\mathcal{F}$. In fact, suppose $D(\epsilon)$ is the maximal number of points $x_i \in \mathcal{F}$ with $d(x_i, x_j) \geq \epsilon$ (the so-called $\epsilon-$packing number), we have $1 = P(\mathcal{F}) \geq D(\epsilon) \cdot c\psi(\epsilon/2)$ using the fact that each ball of radius $\epsilon/2$ around a point x_i has probability at least $c\psi(\epsilon/2)$, whence $D(\epsilon)$ is finite and $\mathcal{F}$ is totally bounded by the well known relationship between the packing number and the covering number (see for example [24]).

The main results for k-NN estimates satisfying the above assumptions are the following.

Theorem 1. *If Assumptions 1, 2 and 3 hold and* $\sum_{n=1}^{\infty}(\log n)^{(r-2)/2}(\|v\|_r/\|v\|_2)^r < \infty$, *then* $||\hat{m}(x) - m(x)|| = O(b_n + [\phi^{-1}(2k/n)]^{\alpha} + (\log n)^{1/2}\|v\|_2)$ *almost surely, where* $\phi^{-1}(x) := \inf\{h : \phi(h) \geq x\}$.

Alternatively, assuming exponential tail decay, we have

Theorem 2. *If Assumptions 1, 2 and 4 hold, then* $||\hat{m}(x) - m(x)|| = O(b_n + [\phi^{-1}(2k/n)]^{\alpha} + (\log n)^{1+1/p}\|v\|_2)$ *almost surely.*

Comparing the two related results above, we see that in Theorem 1, when the weaker assumption 3 is used, we require an extra condition on the weight vector v_{ni}. From the discussion after the corollary below, this condition actually imposes some strong constraints on k in some simple examples.

The theorems above are stated for general weight vector $v_{ni}, 1 \leq i \leq n$. When specialized to some commonly used weight vector, we have the following corollary.

Corollary 1. *For the simple k-NN estimates* ($v_{ni} = 1/k$ *for* $i \leq k$ *and* 0 *otherwise), the theorems above hold with* $b_n = 0$ *and* $\|v\|_2 = O(1/\sqrt{k})$. *The same applies to Example 2 (with a kernel compactly supported and bounded away from zero on* $[0,1]$*) presented previously.*

Remark 1. In the above Corollary we only aim for the simplest results while more complicated kernel functions can be dealt with using lengthier arguments and additional assumption on the small ball probability. In these two simple examples we have $v_{ni} \sim 1/k$ for $i \leq k$ and 0 otherwise, and thus the condition $\sum_{n=1}^{\infty}(\log n)^{(r-2)/2}(\|v\|_r/\|v\|_2)^r < \infty$ reduces to $\sum_{n=1}^{\infty}(\log n)^{(r-2)/2}k^{-(r/2-1)} < \infty$ (note k is a function of n). We see this condition generally requires that k increases polynomially in n, with the requirement less stringent for bigger r.

Remark 2. In [11], the authors distinguished two types of processes: the fractal type processes and the exponential type processes. The former is characterized by $\phi(h) \sim h^{\tau}$, for some $\tau > 0$ and the latter characterized by $\phi(h) \sim exp\{-(1/h^{\tau_1})\log(1/h^{\tau_2})\}, \tau_1 > 0, \tau_2 \geq 0$. The fractal type processes are similar to finite dimensional problems in many aspects, while for infinite dimensional case such as when the covariate curves belong to some smoothness class, exponential type processes are more typical. For example, the simple Gaussian

process, Brownian motion, is of exponential type. The paper [22] provides other more complicated Gaussian processes all of which are of exponential type. From the rates obtained in the Corollary, it is easy to see that for exponential type processes the convergence rates are logarithmic in the sample size, much slower than the classical finite-dimensional cases. Note that as discussed above, under Assumption 3, we require that k increases polynomially in n, which seems to make it similar to the finite dimensional case. However, this impression is misleading. For example, when $\phi(h) \sim exp\{-1/h^\tau\}$ as in typical functional contexts, we have $\phi^{-1}(2k/n) \sim \{1/\log(n/(2k))\}^{1/\tau}$, the convergence rate is logarithmic in n whether k increases polynomially or logarithmically in n.

3. Proofs

In the proofs, different appearances of C denote possibly different positive constants, even within the same expression. We start off by showing a relatively simple result on the distance from x to its $k-$th nearest neighbor.

Lemma 1. *Suppose $k/n \to 0$ and $k/\log n \to \infty$. Let H be the distance from x to its $k-$th nearest neighbor as defined in (2.2), then $P(H \geq \phi^{-1}(2k/n), i.o.) \to 0$, where i.o. means "infinitely often", and $\phi^{-1}(x) := \inf\{h : \phi(h) \geq x\}$.*

Proof. First we note that ϕ is right-continuous and non-decreasing and thus $h = \phi^{-1}(x)$ implies $\phi(h) \geq x$. Denote $a = \phi^{-1}(2k/n)$, $p = \phi(a)$ and thus $np \geq 2k$. We have

$$\begin{aligned}
& P(H \geq \phi^{-1}(2k/n)) \\
= \; & P(\sum_i I\{X_i \in B(x,a)\} \leq k) \\
= \; & P(\sum_i I\{X_i \in B(x,a)\} - np \leq k - np) \\
\leq \; & P(|\sum_i I\{X_i \in B(x,a)\} - np| \geq np/2) \\
\leq \; & 2exp\{-\frac{1}{2}(np/2)^2/[np(1-p) + (np/6)]\} \\
\leq \; & 2exp\{-Cnp\}\,,
\end{aligned}$$

where we applied the Bernstein's inequality for Bernoulli random variables (see for example the Appendix in [17]). Then $P(H \geq \phi^{-1}(2k/n), i.o.) \to 0$ can be shown using Borel-Cantelli lemma noting that $k/\log n \to \infty$. □

Proof of Theorem 1. We use the following decomposition into the bias term and the variance term.

$$||\hat{m}(x) - m(x)|| \leq ||\sum_i v_{ni}(m(X_i) - m(x))|| + ||\sum_i v_{ni}\epsilon_i||. \qquad (3.1)$$

The bias term is easier to deal with. In fact,

$$\begin{aligned} ||\sum_i v_{ni}(m(X_i)-m(x))|| &\leq 2B\sum_{i=k+1}^{n} v_{ni} + ||\sum_{i=1}^{k} v_{ni}(m(X_i)-m(x))|| \\ &= O(b_n + [\phi^{-1}(\frac{2k}{n})]^{\alpha}), \end{aligned}$$

by Assumption 1 and Lemma 1.

Now we deal with the variance term. Let $S_n = \sum_{i=1}^n v_{ni}\epsilon_i$ and the following arguments are conditional on $\{X_1,\ldots,X_n\}$ (in effect treating v_{ni} as nonrandom weights). Following the idea of Section 6.3 in [14], we write $||S_n|| - E||S_n|| = ||\sum_{i=1}^n v_{ni}\epsilon_i|| - E||\sum_{i=1}^n v_{ni}\epsilon_i|| = \sum_{i=1}^n d_i$ (where we remind the reader that the expectation is conditional on $\{X_1,\ldots,X_n\}$), with $d_i = E[||S_n||\,|\mathcal{G}_i] - E[||S_n||\,|\mathcal{G}_{i-1}]$ where $\mathcal{G}_i$ is the $\sigma-$algebra generated by $\epsilon_1,\ldots,\epsilon_i$ ($\mathcal{G}_0$ is the trivial $\sigma-$algebra). It is easy to see that $\{d_i\}$ is a *real-valued* martingale difference sequence which enables us to use relevant exponential type inequalities below. Citing Lemma 6.16 in [14], we know

$$|d_i| \leq ||\epsilon_i||v_{ni} + v_{ni}E||\epsilon_i|| \leq ||\epsilon_i||v_{ni} + Cv_{ni} \tag{3.2}$$

and

$$E(d_i^2|\mathcal{G}_{i-1}) \leq v_{ni}^2 E||\epsilon_i||^2. \tag{3.3}$$

We bound the variance term in four steps.

Step 1: We show $E||S_n|| = O(\|v\|_2)$.

$$\begin{aligned} & E||S_n|| \\ = \; & E||\sum_{i=1}^{n} v_{ni}\epsilon_i|| \\ \leq \; & \sqrt{E\langle\sum_{i=1}^{n} v_{ni}\epsilon_i, \sum_{i=1}^{n} v_{ni}\epsilon_i\rangle} \\ = \; & O(\sqrt{\sum_i v_{ni}^2}) \\ = \; & O(\|v\|_2). \end{aligned}$$

Step 2: Let $d_i' = d_i I\{|d_i| \leq L)$ for some $L > 0$ to be specified later. We have $P(\sum_{i=1}^n (d_i' - E(d_i'|\mathcal{G}_{i-1})) > a) \leq exp\{-Ca^2/(aL + (\sum_i v_{ni}^2))\}, \forall a > 0$.

Using (3.3), $E[(d_i' - E(d_i'|\mathcal{G}_{i-1}))^2|\mathcal{G}_{i-1}] \leq E(d_i'^2|\mathcal{G}_{i-1}) \leq E(d_i^2|\mathcal{G}_{i-1}) = O(v_{ni}^2)$ and together with $|d_i' - E(d_i'|\mathcal{G}_{i-1})| \leq 2L$, we get $E(|d_i' - E(d_i'|\mathcal{G}_{i-1})|^m|\mathcal{G}_{i-1}) \leq C(2L)^{m-2}v_{ni}^2$. Since $d_i' - E(d_i'|\mathcal{G}_{i-1}), i \leq n$ is a martingale difference sequence, using Lemma 8.9 in [21] (Berstein's inequality for martingales), we obtain the desired bound.

Step 3: Let $d_i'' = d_i - d_i' = d_i I\{|d_i| > L\}$. We have $P(\sum_i |d_i'' - E(d_i''|\mathcal{G}_{i-1})| > a) \leq C(\sum_i v_{ni}^r)L^{1-r}/a$.

Using Hölder's inequality and Markov's inequality, we have

$$\begin{aligned} & E(|d_i'' - E(d_i''|\mathcal{G}_{i-1})|) \\ \leq\ & 2E(|d_i''|) \\ =\ & 2E(|d_i| I\{|d_i| > L\}) \\ \leq\ & 2\{E(|d_i|^r)\}^{1/r} P(|d_i| > L)^{1-1/r} \\ \leq\ & 2\{E(|d_i|^r)\}^{1/r} \{\frac{E(|d_i|^r)}{L^r}\}^{1-1/r} \\ =\ & 2E(|d_i|^r) L^{1-r} \\ \leq\ & C v_{ni}^r L^{1-r}\,, \end{aligned}$$

and note that in the last line above we used the bound (3.2). Thus we have $P(\sum_i |d_i'' - E(d_i''|\mathcal{G}_{i-1})| > a) \leq E[\sum_i |d_i'' - E(d_i''|\mathcal{G}_{i-1})|]/a \leq C(\sum_i v_{ni}^r) L^{1-r}/a$.

Step 4: Finally, we demonstrate the bound for the variance term in (3.1).

Using $E(d_i|\mathcal{G}_{i-1}) = E(d_i'|\mathcal{G}_{i-1}) + E(d_i''|\mathcal{G}_{i-1}) = 0$, we have that $d_i = d_i' - E(d_i'|\mathcal{G}_{i-1}) + (d_i'' - E(d_i''|\mathcal{G}_{i-1}))$ and then

$$\begin{aligned} & P(||S_n|| - E||S_n|| > 2a) \\ \leq\ & P(\sum_i (d_i' - E(d_i'|\mathcal{G}_{i-1})) > a) + P(\sum_i (d_i'' - E(d_i''|\mathcal{G}_{i-1})) > a) \\ \leq\ & exp\{-Ca^2/(aL + (\sum_i v_{ni}^2))\} + C(\sum_i v_{ni}^r) L^{1-r}/a\,, \end{aligned}$$

by the previous two steps. Setting $a = C(\log n)^{1/2}\|v\|_2$ for a constant C large enough and $L = \|v\|_2 (\log n)^{-1/2}$, an application of the Borel-Cantelli Lemma leads to $||S_n|| - E||S_n|| = O((\log n)^{1/2}\|v\|_2)$, using the assumption that $\sum_i (\log n)^{(r-2)/2}(\|v\|_r/\|v\|_2)^r < \infty$. Combining this with the result from Step 1, the variance term is thus $||S_n|| = O((\log n)^{1/2}\|v\|_2)$. □

Proof of Theorem 2. The general proof strategy is the same as Theorem 1. In particular, the bias term is bounded in the same way. For the variance term, only Step 3 and Step 4 need to be replaced by the following.

Step 3': We show $P(\sum_i E(d_i'|\mathcal{G}_{i-1}) > a) + P(\text{for some } i, |d_i| > L) = O(n \cdot exp\{-CL^p/v_{n1}^p\})$, if we set $a = C(\log n)^{1+1/p}\|v\|_2$ and $L = C(\log n)^{1/p} v_{n1}$ for C large enough.

Consider the first probability, we have

$$\begin{aligned} E(d_i'|\mathcal{G}_{i-1}) \leq\ & E(|d_i| I\{|d_i| > L|\mathcal{G}_{i-1}) \\ \leq\ & (E|d_i|^r|\mathcal{G}_{i-1}))^{1/r} P(|d_i| > L|\mathcal{G}_{i-1})^{1-1/r} \\ \leq\ & C(E||\epsilon_i||^r v_{ni}^r)^{1/r} exp\{-C(L - Cv_{ni})^p/v_{ni}^p\} \\ \leq\ & Cv_{ni} exp\{-CL^p/v_{ni}^p\} \\ \leq\ & Cexp\{-CL^p/v_{n1}^p\}\,, \end{aligned}$$

using (3.2) and Assumption 4 in the third inequality above. Thus $E(d_i'|\mathcal{G}_{i-1}) \leq a/n$ if we set $a = C(\log n)^{1+1/p}\|v\|_2$ (note that $a \geq \|v\|_2 \geq v_{n1} \geq 1/n$) and $L = C(\log n)^{1/p} v_{n1}$, and then $P(\sum_i E(d_i'|\mathcal{G}_{i-1}) > a) = 0$.

For the other probability term, again using (3.2) and Assumption 4, we have

$$\begin{aligned} & P(\text{for some } i, |d_i| > L) \\ \leq\ & 1 - P(\forall i, v_{ni}||\epsilon_i|| \leq L - Cv_{ni}) \\ \leq\ & 1 - (1 - exp\{-C(L - Cv_{ni})^p / v_{ni}^p\})^n \\ \leq\ & 1 - (1 - exp\{-CL^p / v_{n1}^p\})^n \\ \leq\ & n \cdot exp\{-CL^p / v_{n1}^p\}, \end{aligned}$$

where in the last line above we used the simple inequality $(1-x)^n \geq 1 - nx$.

Step 4': To demonstrate the bound for the variance term, we use

$$\begin{aligned} & P(||S_n|| - E||S_n|| > 2a) \\ =\ & P(\sum_i d_i > 2a) \\ \leq\ & P(\sum_i (d_i' - E(d_i'|\mathcal{G}_{i-1})) > a) + P(E(d_i'|\mathcal{G}_{i-1}) > a) + P(\text{for some } i, |d_i| > L) \\ \leq\ & exp\{-Ca^2/(aL + \sum_i v_{ni}^2)\} + n \cdot exp\{-CL^p / v_{n1}^p\}, \end{aligned}$$

by the bounds obtained in Step 2 and Step 3'. Finally set $a = C_1 (\log n)^{1+1/p} \|v\|_2$ and $L = C_2 (\log n)^{1/p} v_{n1}$ (choose C_2 large enough to make the second term above summable and then choose C_1 large enough to make the first term summable) and apply the Borel-Cantelli Lemma and then use the result from Step 1 to get $||S_n|| = O((\log n)^{1+1/p} \|v\|_2)$. □

Proof of Corollary 1. For the simple k-NN method this is obvious. For kernel k-NN, it is also obvious that $b_n = 0$ by the definition of H. Since $v_{ni} = K(d(X_i, x)/H) / \sum_j K(d(X_j, x)/H) \leq C / \sum_j K(d(X_j, x)/H)$ and $K(d(X_j, x)/H)$ is bounded away from zero for $j \leq k$ and 0 for $j > k$ by the assumptions made on K, we have $v_{ni} = O(1/k)$ for $i \leq k$ and 0 otherwise. It then follows that $\|v\|_2 = O(1/\sqrt{k})$. □

4. Discussion

We assumed in the paper that $\mathcal{H}$ is a Hilbert space while the covariate space is a much more general semi-metric space. That the response is in a Hilbert space is necessary for applying the results in [14] and thus it seems difficult to consider the response in a semi-metric space. However, it is possible to assume that $\mathcal{H}$ is a Banach space. The proofs go through without change for Banach space except for Step 1 in the proof where we used the inner product. In general Banach space, it is not clear how to deal with $E\|S_n\|$ in Step 1. However, under the additional assumption that $\mathcal{H}$ is a Banach space of type p, then by Proposition 9.11 in [14] or Definition 2.3 in [2], we have $E\|S_n\| = O((E\|S_n\|^p)^{1/p}) = O((\sum_i v_{ni}^p E\|\epsilon_i\|^p)^{1/p}) = O(\|v\|_p)$ and thus an extra $\|v\|_p$ will appear in the convergence rates.

Finally, we mention some possibilities for further studies. For functional regression with scalar responses, uniform convergence was obtained in [12], asymptotic normality was shown in [8, 6] for the independent case and α-mixing case respectively, and [7] studied inferences using bootstrap. We expect these results can be extended to k-NN estimates with functional responses under stronger assumptions.

Acknowledgements

We thank the Editor Professor David Ruppert, Associate Editor and referee for their comments and in particular their pointers to extend the scope of the estimator to more general spaces, which helped us improve the paper significantly.

References

[1] Antoch, J., Prchal, L., De Rosa, M. R. and Sarda, P. (2008). Functional linear regression with functional response: Application to prediction of electricity consumption. In *Functional and Operatorial Statistics* (S. DaboNiang and F. Ferraty, eds.) 23-29. MR2478483

[2] Bosq, D. (2000). *Linear processes in function spaces: theory and applications*. Springer Verlag. MR1783138

[3] Burba, F., Ferraty, F. and Vieu, P. (2009). k-Nearest Neighbour method in functional nonparametric regression. *Journal of Nonparametric Statistics* **21** 453–469. MR2571722

[4] Cover, T. M. (1968). Estimation by the nearest neighbor rule. *IEEE Transactions on Information Theory* **14** 50-55.

[5] Cover, T. M. and Hart, P. E. (1967). Nearest neighbor pattern classification. *IEEE Transactions on Information Theory* **13** 21-27.

[6] Delsol, L. (2009). Advances on asymptotic normality in non-parametric functional time series analysis. *Statistics* **43** 13–33. MR2499359

[7] Ferraty, F., Keilegom, I. and Vieu, P. (2010). On the Validity of the Bootstrap in Non-Parametric Functional Regression. *Scandinavian Journal of Statistics* **37** 286–306. MR2682301

[8] Ferraty, F., Mas, A. and Vieu, P. (2007). Nonparametric regression on functional data: Inference and practical aspects. *Australian & New Zealand Journal of Statistics* **49** 267-286. MR2396496

[9] Ferraty, F. and Vieu, P. (2002). The functional nonparametric model and application to spectrometric data. *Computational Statistics* **17** 545-564. MR1952697

[10] Ferraty, F. and Vieu, P. (2004). Nonparametric models for functional data, with application in regression, time-series prediction and curve discrimination. *Journal of nonparametric statistics* **16** 111-125. MR2053065

[11] Ferraty, F. and Vieu, P. (2006). *Nonparametric functional data analysis: theory and practice. Springer series in statistics*. Springer, New York, NY. MR2229687

[12] Ferraty, F., Laksaci, A., Tadj, A. and Vieu, P. (2010). Rate of uniform consistency for nonparametric estimates with functional variables. *Journal of Statistical Planning and Inference* **140** 335–352. MR2558367

[13] Fix, E. and Hodges, J. L. (1989). Discriminatory analysis. nonparametric discrimination: consistency properties. *International Statistical Review* **57** 238-247.

[14] Ledoux, M. and Talagrand, M. (1991). *Probability in Banach spaces: isoperimetry and processes*. Springer-Verlag, Berlin; New York. MR1102015

[15] Lian, H. (2007). Nonlinear functional models for functional responses in reproducing kernel Hilbert spaces. *Canadian Journal of Statistics-Revue Canadienne De Statistique* **35** 597-606. MR2381399

[16] Parzen, E. (1962). On the estimation of a probability density function and mode. *Annals of Mathematical Statistics* **33** 1065-1076. MR0143282

[17] Pollard, D. (1984). *Convergence of stochastic processes*. Springer-Verlag, New York. MR0762984

[18] Preda, C. (2007). Regression models for functional data by reproducing kernel Hilbert spaces methods. *Journal of Statistical Planning and Inference* **137** 829-840. MR2301719

[19] Ramsay, J. O. and Silverman, B. W. (2005). *Functional data analysis*, 2nd ed. *Springer series in statistics*. Springer, New York. MR2168993

[20] Rosenblatt, M. (1956). Remarks on some nonparametric estimates of density function. *Annals of Mathematical Statistics* **27** 832-837. MR0079873

[21] van der Geer, S. A. (2000). *Applications of empirical process theory*. Cambridge University Press, Cambridge.

[22] van der Vaart, A. W. and Van Zanten, J. H. (2008). Rates of contraction of posterior distributions based on Gaussian process priors. *Annals of Statistics* **36** 1435-1463. MR2418663

[23] van der Vaart, A. W. and Wellner, J. A. (1996). *Weak convergence and empirical processes. Springer series in statistics*. Springer, New York. MR1385671

[24] Yang, Y. H. and Barron, A. (1999). Information-theoretic determination of minimax rates of convergence. *Annals of Statistics* **27** 1564-1599. MR1742500