A k-NEAREST NEIGHBOR APPROACH FOR FUNCTIONAL REGRESSION

Thomas LALOË

Institut de Mathématiques et de Modélisation de Montpellier
UMR CNRS 5149, Equipe de Probabilités et Statistiques
Université Montpellier II, Cc 051
Place Eugène Bataillon, 34095 Montpellier Cedex 5, France
tlaloe@math.univ-montp2.fr

Abstract

Let (X, Y) be a random pair taking values in $\mathcal{H} \times \mathbb{R}$, where $\mathcal{H}$ is an infinite dimensional separable Hilbert space. We establish weak consistency of a nearest neighbor-type estimator of the regression function of Y on X based on independent observations of the pair (X, Y). As a general strategy, we propose to reduce the infinite dimension of $\mathcal{H}$ by considering only the first d coefficients of an expansion of X in an orthonormal system of $\mathcal{H}$, and then to perform k-nearest neighbor regression in $\mathbb{R}^d$. Both the dimension and the number of neighbors are automatically selected from the observations using a simple data-dependent splitting device.

1 Introduction

Regression is the problem of predicting a variable from some observation. An observation is usually supposed to be a collection of numerical measurements represented by a d-dimensional vector. However, in many real-life problems, input data items are in the form of random functions (speech recordings, spectra, images) rather than standard vectors, and this casts the regression

problem into the general class of functional data analysis. Even though in practice such observations are observed at discrete sampling points, the challenge in this context is to infer the data structure by exploiting the infinite-dimensional nature of the observations. The last few years have witnessed important developments in both the theory and practice of functional data analysis, and many traditional data analysis tools have been adapted to handle functional inputs. The book of Ramsay and Silverman [6] provides a comprehensive introduction to the area. For an updated list of references, we refer the reader to Cérou and Guyader [3], Rossi and Villa [7], and Tuleau [8].

In the present paper, we consider the functional regression setting, where the goal is to predict a scalar response Y from some infinite-dimensional observations X. More precisely, we will denote by (X, Y) a random pair taking values in $\mathcal{Z} = \mathcal{H} \times \mathbb{R}$, where $\mathcal{H}$ is an infinite dimensional separable Hilbert space. Throughout the document, we will denote by ρ the (unknown) distribution of (X, Y), and by ρ_X the marginal distribution of X. Based on n independent copies $(X_1, Y_1), \ldots, (X_n, Y_n)$ of (X, Y), we introduce an estimator f_n of the regression function $f_\rho(x) = \mathbb{E}[Y|X = x]$ which is designed as follows: First, we reduce the dimension of $\mathcal{H}$ by considering the first d coefficients of an expansion of each observation in a orthonormal system of $\mathcal{H}$; Second, we perform k-nearest neighbor regression (see Györfi, Kohler, Krzyzak, and Walk [5])in $\mathbb{R}^d$. We select simultaneously the dimension d and the number of neighbors k by a data-splitting device. Our main result states weak consistency of the resulting estimator, thereby extending the general strategy introduced by Biau, Bunea, and Wegkamp [2] in the context of classification (ie, when Y takes its values in a finite set).

The paper is organised as follows. We start in Section 2.1 by introducing some notation. Then, in Section 2.2, we present the construction of the estimator, and state its weak consistency. Proof are collected in Section 3.

2 Consistent functional regression

2.1 Notation

We let the symbols $\langle .|. \rangle$ and $\|.\|$ denote the inner product and the associated norm on $\mathcal{H}$, respectively, and we let $(\phi_j)_{j\geq 1}$ be a complete orthonormal system of $\mathcal{H}$ (Akhiezer and Glazman [1]). For each observation X_i, we set $X_{ij} = \langle X_i|\phi_j\rangle$. We know that

$$X_i = \sum_{j=1}^{\infty} X_{ij}\phi_j,$$

where the consistency holds in the L^2 sense.
Introduce $\mathcal{H}^{(d)}$, the finite-dimensional vector space spanned by the functions $\{\phi_1, \phi_2, \ldots, \phi_d\}$, and let, for each X_i,

$$X_i^{(d)} = \sum_{j=1}^{d} X_{ij}\phi_j.$$

Finally, denote by f_ρ and $f_{\rho,d}$ the regression functions in $\mathcal{H}$ and $\mathcal{H}^{(d)}$, respectively, and by σ_ρ^2 and $\sigma_{\rho,d}^2$ their respective L^2 errors. More precisely, we have $f_\rho(x) = \mathbb{E}[Y|X = x]$, $\sigma_\rho^2 = \int_{\mathcal{Z}}(y - f_\rho(x))^2 d\rho(x, y)$, and the same in $\mathcal{H}^{(d)} \times \mathbb{R}$ for $f_{\rho,d}$ and $\sigma_{\rho,d}^2$. Throughout the document, we suppose that $Y < \infty$ $a.s.$, and all the integrals are to be understood over ρ or ρ_X.

2.2 k-nearest neighbor in $\mathcal{H}^{(d)}$

Let us first formally define our k-nearest neighbor type estimator. To this aim, we consider the sequence $(X_1^{(d)}, Y_1), \ldots, (X_n^{(d)}, Y_n)$ where the observations have been projected onto $\mathcal{H}^{(d)}$. For x in $\mathcal{H}^{(d)}$, we reorder the data:

$$\left(X_{(1)}^{(d)}(x), Y_{(1)}(x)\right), \ldots, \left(X_{(n)}^{(d)}(x), Y_{(n)}(x)\right),$$

according to the increasing Euclidean distances $\|X_i^{(d)} - x\|$ of the $X_i^{(d)}$ to x. In other words, $X_{(i)}^{(d)}(x)$ is the i-th nearest neighbor of x amongst $X_j^{(d)}$. If $\|X_i^{(d)} - x\| = \|X_j^{(d)} - x\|$, $X_i^{(d)}$ is declared closer to x if $i < j$. The k-nearest

neighbor estimator of f_ρ is then defined (Györfi, Kohler, Krzyzak, and Walk [5]) as

$$f_{n,k,d}(x) = \frac{1}{k}\sum_{i=1}^{k} Y_{(i)}(x). \tag{1}$$

To select simultaneously the dimension d and the number of neighbors k, we suggest the following data-splitting device. First we split the data into a training set $\{(X_i, Y_i), i \in \mathcal{I}_\ell\}$ of length ℓ, and a validation set $\{(X_j, Y_j), j \in \mathcal{J}_m\}$ of length m, with $m+\ell = n$ (ℓ and m possibly function of n). For each $d \geq 1$, $1 \leq k \leq \ell$, we construct a k-nearest neighbor estimator based on the training set. Second we use the validation set to select $\hat{d}$ and $\hat{k}$ as follows:

$$(\hat{d}, \hat{k}) \in \underset{d\geq 1, 1\leq k\leq \ell}{\arg\min} \left[\frac{1}{m}\sum_{j\in\mathcal{J}_m}\left(Y_j - f_{\ell,k,d}(X_j^{(d)})\right)^2 + \frac{\lambda_d}{\sqrt{m}}\right]. \tag{2}$$

Here, the term $\lambda_d/\sqrt{m}$ is a given penalty term which tends to infinity with d to prevent overfitting.
This method, which is computationnaly simple, leads to the estimator

$$\hat{f}_n(x) := f_{\ell,\hat{k},\hat{d}}(x^{(\hat{d})}), \tag{3}$$

which has an error

$$\mathcal{E}(\hat{f}_n) = \int_{\mathcal{Z}}\left(y - \hat{f}_n(x)\right)^2 d\rho(x,y) = \int_{\mathcal{H}}\left(\hat{f}_n(x) - f_\rho(x)\right)^2 d\rho_X(x) \ + \sigma_\rho^2 \, .$$

The estimator f_n satisfies the following oracle inequality:

Proposition 2.1 *Let M be a positive constant such that $(Y - f_{\ell,k,d}(x))^2 \leq M$ a.s., and suppose that*

$$\Delta := \sum_{d=1}^{\infty} e^{-2(\lambda_d/M)^2} < \infty. \tag{4}$$

Then there exists a constant $c > 0$, only depending on Δ and M, such that, for every integer $\ell > 1/\Delta$ and $m = n - \ell$,

$$\mathbb{E}\int_{\mathcal{H}}\left(\hat{f}_n(x)-f_\rho(x)\right)^2$$

$$\leq \inf_{d\geq 1}\left[(\sigma^2_{\rho,d}-\sigma^2_\rho)+\inf_{1\leq k\leq \ell}\left(\mathbb{E}\int_{\mathcal{H}^{(d)}}\left(f_{\ell,k,d}(x)-f_{\rho,d}(x)\right)^2\right)+\frac{\lambda_d}{\sqrt{m}}\right]+c\sqrt{\frac{\ln \ell}{m}}\,. \tag{5}$$

The term $\sigma^2_{\rho,d}-\sigma^2_\rho$ may be viewed as the price to be paid for using a finite dimensional approximation of the observations, and it converges to zero by Lemma 3.1 below. The term $\inf_{1\leq k\leq \ell}\left(\mathbb{E}\int_{\mathcal{H}^{(d)}}(f_{\ell,k,d}(x)-f_{\rho,d}(x))^2\right)$ converges also to zero by Lemma 3.2. Since the infimum is taken over all $d\geq 1$, weak convergence of $\hat{f}_n(x)$ to $f_\rho(x)$ is ensured.

Theorem 2.1 *Under the assumption* (4) *and*

$$\lim_{n\to\infty}\ell=\infty,\ \lim_{\ell\to\infty}k=\infty,\ \lim_{\ell\to\infty}\frac{k}{\ell}=0,\ \textit{and}\ \lim_{n\to\infty}\frac{\ln\ell}{m}=0,$$

$\hat{f}_n$ *weakly converges to* f_ρ, *i.e.*

$$\mathbb{E}\int_{\mathcal{H}}\left(\hat{f}_n(x)-f_\rho(x)\right)^2\to 0\ \textit{as}\ n\to\infty.$$

Pratically speaking, as discussed in Biau, Bunea, and Wegkamp [2], choosing the penalty in (2) is not an easy task. Indeed, an abusive penalisation of high dimensions can mask helpful information. For a more involved discussion about the penalty choice, and experimental results, we refer the reader to Tuleau [8], who shows that that adding a penalty term improves the stability of the selected dimension d.

3 Proof

Proof of Proposition 2.1 Let

$$L(k,d)=\mathbb{E}\left[\left(Y-f_{\ell,k,d}\left(X^{(d)}\right)\right)^2\mid (X_i,Y_i),1\leq i\leq n\right],$$

and

$$\hat{L}(k,d) = \frac{1}{m} \sum_{j \in \mathcal{J}_m} \left(Y_j - f_{\ell,k,d}(X_j^{(d)}) \right)^2 .$$

We have to minimize $\hat{L}(k,d) + \lambda_d/m$ in k and d.

Fix $\varepsilon > 0$. For every $d \geq 1$ and every k satisfying $1 \leq k \leq \ell$, we may write

$$\mathbb{P}\left\{ L(\hat{k},\hat{d}) - \hat{L}(k,d) > \frac{\lambda_d}{\sqrt{m}} + \varepsilon \right\} \leq \mathbb{P}\left\{ L(\hat{k},\hat{d}) - \hat{L}(\hat{k},\hat{d}) > \frac{\lambda_{\hat{d}}}{\sqrt{m}} + \varepsilon \right\},$$

since, by definition of $(\hat{k},\hat{d})$,

$$\hat{L}(\hat{k},\hat{d}) + \frac{\lambda_{\hat{d}}}{\sqrt{m}} \leq \hat{L}(k,d) + \frac{\lambda_d}{\sqrt{m}} .$$

Therefore,

$$\begin{aligned}
&\mathbb{P}\left\{ L(\hat{k},\hat{d}) - \hat{L}(k,d) > \lambda_d/\sqrt{m} + \varepsilon \right\} \\
&\quad \leq \sum_{d=1}^{\infty} \sum_{k=1}^{\ell} \mathbb{P}\left\{ L(k,d) - \hat{L}(k,d) > \lambda_d/\sqrt{m} + \varepsilon \right\} \\
&\quad \text{(by the union bound)} \\
&\quad = \sum_{d=1}^{\infty} \sum_{k=1}^{\ell} \mathbb{E}\mathbb{P}\left\{ L(k,d) - \hat{L}(k,d) > \lambda_d/\sqrt{m} + \varepsilon \mid (X_i,Y_i), i \in \mathcal{I}_\ell \right\} \\
&\quad \leq \sum_{d=1}^{\infty} \ell \exp\left\{ -2[(\lambda_d/\sqrt{m}) + \varepsilon]^2 \times (m/M^2) \right\} \\
&\quad \text{(by Hoeffding's inequality)} \\
&\quad \leq \ell e^{-2m\varepsilon^2/M^2} \sum_{d=1}^{\infty} e^{-2(\lambda_d/M)^2} \\
&\quad = \Delta \ell e^{-2m\varepsilon^2/M^2},
\end{aligned}$$

where $\Delta = \sum_{d=1}^{\infty} e^{-2(\lambda_d/M)^2}$. Since, for every $d \geq 1$ and k with $1 \leq k \leq \ell$,

$$\mathbb{E}L(\hat{k},\hat{d}) \leq \mathbb{E}\hat{L}(k,d) + \frac{\lambda_d}{\sqrt{m}} + \int_0^{\infty} \mathbb{P}\left\{ L(\hat{k},\hat{d}) - \hat{L}(k,d) > \frac{\lambda_d}{\sqrt{m}} + \varepsilon \right\} d\varepsilon ,$$

we obtain, for every $u > 0$,

$$\mathbb{E}L(\hat{k},\hat{d}) \leq \mathbb{E}\hat{L}(k,d) + \frac{\lambda_d}{\sqrt{m}} + u + \Delta\ell \int_u^{\infty} e^{-2m\varepsilon^2/M^2} d\varepsilon .$$

Note that

$$\begin{aligned}\int_u^\infty e^{-2m\varepsilon^2/M^2}d\varepsilon &\leq \frac{1}{2}\int_u^\infty \left(2+\frac{M^2}{2m\varepsilon}\right)e^{-2m\varepsilon^2/M^2}d\varepsilon \\ &= -\frac{1}{2}\left[\frac{M^2}{2m\varepsilon}e^{-2m\varepsilon^2/M^2}\right]_u^\infty \\ &= \frac{M^2}{4mu}e^{-2mu^2/M^2}.\end{aligned}$$

Whence, choosing $u = M\sqrt{\ln(\Delta\ell)/2m}$, we obtain

$$\mathbb{E}L(\hat{k},\hat{d}) \leq \mathbb{E}\hat{L}(k,d) + \frac{\lambda_d}{\sqrt{m}} + M\sqrt{\frac{\ln(\Delta\ell)}{2m}} + \frac{M}{2\sqrt{2m\ln(\Delta\ell)}}\,.$$

Since k et d are arbitrary,

$$\mathbb{E}L(\hat{k},\hat{d}) \leq \inf_{d\geq 1, 1\leq k\leq \ell} \mathbb{E}\hat{L}(k,d) + \frac{\lambda_d}{\sqrt{m}} + M\sqrt{\frac{\ln(\Delta\ell)}{2m}} + \frac{M}{2\sqrt{2m\ln(\Delta\ell)}}\,.$$

The fact that $\mathbb{E}\hat{L}(k,d) = \mathbb{E}L(k,d)$ for each fixed k,d leads to the inequality (5). □

Proof of Theorem 2.1 will rely on the following lemma.

Lemma 3.1 *We have*

$$\sigma^2_{\rho,d} - \sigma^2_\rho \to 0 \text{ when } d \to \infty\,.$$

Proof of Lemma 3.1 :

$$\begin{aligned}\sigma^2_{\rho,d} - \sigma^2_\rho &= \mathbb{E}\Big[Y - \mathbb{E}\big[Y|X^{(d)}\big]\Big]^2 - \mathbb{E}\Big[Y - \mathbb{E}\big[Y|X\big]\Big]^2 \\ &= \mathbb{E}\Big[Y - \mathbb{E}\big[Y|X\big]\Big]^2 + \mathbb{E}\Big[\mathbb{E}\big[Y|X\big] - \mathbb{E}\big[Y|X^{(d)}\big]\Big]^2 - \mathbb{E}\Big[Y - \mathbb{E}\big[Y|X\big]\Big]^2 \\ &= \mathbb{E}\Big[\mathbb{E}\big[Y|X\big] - \mathbb{E}\big[Y|X^{(d)}\big]\Big]^2.\end{aligned}$$

Since $\mathbb{E}[Y^2] < \infty$, the sequence $\Big(\mathbb{E}[Y|X^{(d)}]\Big)_{d\geq 1}$ is a L^2 bounded Martingale, therefore we have

$$\mathbb{E}[Y|X^{(d)}] \to \mathbb{E}[Y|X] \text{ in the } L^2 \text{ sense as } d \to \infty. \qquad \square$$

Lemma 3.2 *Assume that $k \to \infty$ and $k/\ell \to 0$ as $\ell \to \infty$. Then, for any fixed d,*

$$\mathbb{E} \int_{\mathcal{H}^{(d)}} \Big(f_{\ell,k,d}(x) - f_{\rho,d}(x) \Big)^2 \to 0 \; as \; \ell \to \infty$$

Proof of Lemma 3.2 See Györfi, Kohler, Krzyzak, and Walk [5], Theorem 6.1, page 88. $\square$

We are now in a position to prove Theorem 2.1.

Proof of Theorem 2.1 Fix $\varepsilon > 0$. By Lemma 3.1, we know there exist d_0 such that $\sigma_{\rho,d} - \sigma_\rho^2 \leq \varepsilon$ for all $d \geq d_0$. Then, by Lemma 3.2, we have

$$\mathbb{E} \int_{\mathcal{H}^{(d_0)}} \Big(f_{\ell,k,d_0}(x) - f_{\rho,d_0}(x) \Big)^2 \to 0 \; as \; \ell \to \infty.$$

Finally by Proposition 2.1 we have :

$$\begin{aligned}
&\mathbb{E} \int_{\mathcal{H}} \Big(\hat{f}_n(x) - f_\rho(x) \Big) \\
&\leq \inf_{d \geq 1} \left[(\sigma_{\rho,d}^2 - \sigma_\rho^2) + \inf_{1 \leq k \leq \ell} \left(\mathbb{E} \int_{\mathcal{H}^{(d)}} \Big(f_{\ell,k,d}(x) - f_{\rho,d}(x) \Big)^2 \right) + \frac{\lambda_d}{\sqrt{m}} \right] + c\sqrt{\frac{\ln \ell}{m}} \\
&\leq (\sigma_{\rho,d_0}^2 - \sigma_\rho^2) + \inf_{1 \leq k \leq \ell} \left(\mathbb{E} \int_{\mathcal{H}^{(d_0)}} \Big(f_{\ell,k,d_0}(x) - f_{\rho,d_0}(x) \Big)^2 \right) + \frac{\lambda_{d_0}}{\sqrt{m}} + c\sqrt{\frac{\ln \ell}{m}} \\
&\leq \varepsilon + o(1), \text{ as n} \to \infty.
\end{aligned}$$

Since ε is arbitrary the convergence is ensured. $\square$

References

[1] N.I. AKHIEZER and I.M. GLAZMAN (1961). *Theory of linear operators in Hilbert space*, Frederick Ungar Publishing Co., New York.

[2] G.BIAU, F. BUNEA, and M.H. WEGKAMP (2005). Functional classification in Hilbert Spaces, *IEEE Transactions on Information Theory*, Vol. 51, pp. 2163-2172.

[3] F. CEROU and A. GUYADER (2005). Nearest neighbor classification in infinite dimension, *Research Report INRIA,* RR 5536.

[4] F. CUCKER and S. SMALE (2001). On the mathematical foundations of learning, *bulletin (New Series) of the american mathematical society*, Vol. 39, pp. 1-49.

[5] L. GYÖRFI, M. KOHLER, A. KRZYZAK, and H. WALK (2002). *A distribution free theory of nonparametric regression*, Springer Verlag, New York.

[6] J.O. RAMSAY and B.W. SILVERMAN (2002). *Functional data analysis*, Springer, New-York.

[7] F. ROSSI and N. VILLA (2006). Support Vector Machine For Functional Data Classification, *Neurocomputing* Vol 69, pp. 730-742.

[8] C. TULEAU (2005). Sélection de variables pour la discrimination en grande dimension, classification de données fonctionnelles, *PhD thesis, University Paris XI.*