République Algérienne Démocratique et Populaire

Ministère de l'Enseignement Supérieur et de la Recherche Scientifique

N°Attribué par la bibliothèque

| | | | | | | | | | |

Année Univ.: 2015/2016

* *

About The Nonparametric Estimation with K-Nearest Neighbors for Functional Data

* *

Mémoire présenté en vue de l'obtention du diplôme de

Master Académique

Université Dr Tahar Moulay-Saïda

Discipline: MATHEMATIQUES

Spécialité: Analyse Stochastique; Statistique des processus et Applications

Par

Zahira Arabi[1]

Sous la direction de

Dr. Fethi Madani

Soutenu le 30 Mai 2016 devant le jury composé de

Mlle. F. Benziadi	**Université Dr Tahar Moulay - Saïda**	**Présidente**
Dr. F. Madani	**Université Dr Tahar Moulay - Saïda**	**Rapporteur**
Dr. S. Rahmani	**Université Dr Tahar Moulay - Saïda**	**Examinatrice**
Dr. R. Rouane	**Université Dr Tahar Moulay - Saïda**	**Examinatrice**

[1] e-mail:zahira_arabi@yahoo.fr

Thanks

Thanks first to God who give me the will to complete this work.

To my parents

Apart from that, there are some more people whom I owe a debt of gratitude.

In the first place,

I warmly thank to my rapporteur ***Mr. Fethi Madani****:*

For the confidence that he gave me,

For also the advices and her essential participation in this work.....Thank you.

Thanks to all my teachers, specially: Mme. F.Mokhtari, Dr. A.Kandouci, Mme. F.Benziadi for the help in my study.

I wish to extend my sincere thanks to: Mme. R.Rouane and Mlle. S.Rahmani for agreeing to judge and criticize this work Assuming roles of examiner,

And Mlle. F.Benziadi to chair this jury.

I must thank various people here who have helped me,....Thanks Saliha

My familly, all my sisters and friends who I love them:

Fouzia, Fatima, Lamia, Fatiha, Amina...

Z.Arabi

Contents

List of Figures

Summary

In this work, we are interested in the functional nonparametric estimation using the k nearest neighbors method (k-NN) for a scalar response variable given a random variable taking values in a semi metric space.

In the first part, we will explain how this method work (with their algorithm) by giving some concepts that help us to better understand the basic idea of the k-NN.

Then, and using these concepts, we will give the asymptotic properties for real and vector data.

In the second part of this study, and with the result of Ferraty and Vieu (2006) [15], we will demonstrate the functional case; also for real response.

In the last chapter, and to include classification and regression problems, we will give more fields for the application of this method taking simulation examples to compare between kNN and another parametric and nonparametric methods like the kernel and linear regression.

Introduction

The functional statistic branch is a topical research fields, and diversified by its fundamental aspects and by the different areas that overlap. Parametric statistics in which it is assumed that the distribution follows a model described by a finite number of parameters, and nonparametric statistics; which is based on the idea of not making assumptions about the distribution.

In fact, the nonparametric statistical know a great expand with many authors and in different fields. The proof of this success is the number of scientific publications data in this subject. The first travaux in this domain date back to the 50s , by Roe and Tucker (1958) who studied the maximum Likelihood estimate for the multinomial distribution.

Note that the most frequently encountered by nonparametric statistical model is the regression model that describes the relationship between two or more random variables. This model was already considered by many authors. The properties and results of the estimator of the regression function in the case of independent and identically distributed variables (i.i.d) by Nadaraya (1964). Watson (1964) established the uniform convergence of this estimator, the almost sure uniform convergence is achieved by Devroye (1978) [14] and the asymptotic normality of the same estimator was established by Roussas (1989) [37].

In the same year Gyorfie obtained the asymptotic results for the estimator of the regression function on $\alpha-$mixing process, Vieu (1991) gave the exact asymptotic terms of the square error of the kernel estimator of the regression function. Moreover, Ferraty and Vieu (2000) gave the first results in the functional case. These results have been developed in 2002 by Ferraty and Laksaci et al. treating the fore-casting problem about the continuous time processes. Several authors have generalized these results; Masri (2005) studied the asymptotic normality of the estimator of regression function independence condition, the convergence in quadratic

mean estimator of the same estimator was shown by Laksaci (2007).

Note also that many authors have dealt with other cases variables in the estimating of the regression function. Include Ould-Said (2012) who studied the uniform and asymptotic normality with censored variables. Under the ergodic conditions and Louani Laid (2010-2011) studied the estimator of functional regression function.

0.1 *What are nonparametric statistics for functional data?*

There are different ways for defining what is a nonparametric statistical model in finite dimensional context, and the border between nonparametric and parametric models may sometimes appear to be unclear. Here, we decided to start from the following definition of nonparametric model in finite dimensional context. First, we must give a definition for: *functional variable and functional datasets.*

Functional data:

A random variable $\mathcal{X}$ is called functional variable (f.v) if it takes values in an infinite dimensional space (or functional space). An observation χ of $\mathcal{X}$ is called a functional data.

Note that, when $\mathcal{X}$ (resp.χ) denote a random curve (resp.its observation), we have the following identification: $\mathcal{X} = \{\mathcal{X}(t); t \in T\}$ (resp.$\chi = \{\chi(t); t \in T\}$).
Now, let us define the *Functional datasets.*

Functional datasets:

A functional datasets χ_1,...,χ_n is the observation of n functional variables $\mathcal{X}_1$,...,$\mathcal{X}_n$ identically distributed as $\mathcal{X}$.
Here; we decided to start from the following definition of nonparametric model in finite dimensional context.

Nonparametric model:

Let $\boldsymbol{X}$ be a random vector valued in $\mathbb{R}^p$ and let ϕ be a function defined on $\mathbb{R}^p$ and depending on the distribution of $\boldsymbol{X}$. A model for the estimation of ϕ consists in introducing some constraint of the form: $\phi \in \mathcal{C}$ The model is called a parametric model for the estimation of ϕ if $\mathcal{C}$ is indexed by a finite number of elements of $\mathbb{R}$. Otherwise, the model is called a nonparametric model.

Our decision for choosing this definition was motivated by the fact that it makes definitively clear the border between parametric and nonparametric models,and also because this definition can be easily extended to the functional framework.
Now we can give the definition of *functional nonparametric model.*

Functional nonparametric model:

Let $\mathcal{Z}$ be a random variable valued in some infinite dimensional space F and let ϕ be a mapping defined on F and depending on the distribution of $\mathcal{Z}$. A model for the estimation of ϕ consists in introducing some constraint of the form

$$\phi \in \mathcal{C}.$$

The model is called a functional parametric model for the estimation of ϕ if $\mathcal{C}$ is indexed by a finite number of elements of F.Otherwise,the model is called a functional nonparametric model.

The appellation ***Functional Nonparametric Statistics*** covers all statistical backgrounds involving a nonparametric functional model. In the terminology Functional Nonparametric Statistics,the adjective nonparametric refers to the form of the set of constraints whereas the word functional is linked with the nature of the data.
In other words, nonparametric aspects come from the infinite dimensional feature of the object to be estimated and functional designation is due to the infinite dimensional feature of the data. That is the reason why we may identify this framework to a double infinite dimensional context. Indeed, ϕ can be viewed as a non-linear operator and one could use the terminology model for functional estimation.
To illustrate our purpose concerning these modelling aspects, we focus on the regression models

$$Y = r(X) + error$$

Where Y is a real random variable by considering various situations: linear (parametric) or nonparametric regression models with curves.

0.2 *What is a semi-metric space?*

One of the most popular in $\mathbb{R}^p$ is the usual euclidean norm $\| \, . \, \|$ which is based on the sum of squares of the components of any vector. More precisely , let $x =^t (x_1, ..., x_p)$ be a vector of $\mathbb{R}^p$; then, the classical euclidean norm is defined by

$$\| \, x \, \|^2 = \sum_{j=1}^{p} (x_j)^2 =^t xx.$$

Of course, we can deduce a family of norms based on the euclidean norm by using different definite positive matrix $\boldsymbol{M}$, in the following way

$$\| \, x \, \|^2_{\boldsymbol{M}} =^t x\boldsymbol{M}x.$$

The choice of the norm comes to the same as the choice of $\boldsymbol{M}$.

Semi-norm.

$\| \, . \, \|$ is a semi-norm on some space F as soon as:
1) $\forall (\lambda, x) \in \mathbb{R} \times F, \| \, \lambda x \, \| = |\lambda| \, \| \, x \, \|$;
2) $\forall (x, y) \in F \times F, \| \, x + y \, \| \leq \| \, x \, \| + \| \, y \, \|$.
Note that in fact, a semi-norm $\| \, . \, \|$ is a norm except that $\| \, x \, \| = 0 \Rightarrow x = 0$
Similarly, a semi-metric d can be defined to be a metric but such that

$$d(x, y) = 0 \quad \not\Rightarrow x = y.$$

Semi-metric.

d is a semi-metric on some space F as soon as:
1) $\forall x \in F, d(x, x) = 0$,
2) $\forall (x, y, z) \in F \times F \times F, \; d(x, y) \leq d(x, z) + d(z, y)$.

Semi-metric space.

Let $(\mathcal{X}_i, Y_i)_{i=1,...,n}$ be n independent pairs identically distributed as $(\mathcal{X}, Y)$ and valued in $E \times \mathbb{R}$, (E, d) is a semi-metric space; it mean that $\mathcal{X}$ is a functional random variable and d is a semi metric.

Hölder's inequality.

$$\mathbb{E}[|XY|] \leq \mathbb{E}^{1/p}[|X|^p]\mathbb{E}^{1/q}[|Y|^q], \qquad \frac{1}{p}+\frac{1}{q}=1$$

Schwarz's inequality.

$$\mathbb{E}[|XY|] \leq \mathbb{E}^{1/2}[|X|^2]\mathbb{E}^{1/2}[|Y|^2], \qquad p=q=1$$

If X and Y have second moments, then XY must have a first moment.

Markov's inequality (Theorem).

Let X:S $\to \mathbb{R}$ *be a non-negative random variable. Then, for any* $a > 0$,

$$\mathbb{P}(X \geq a) \leq \frac{\mathbb{E}(X)}{a}$$

Chernoff bound.

Let $X_1, X_2, ..., X_n$ be independent poisson trials with $\mathbb{P}[X_i = 1] = p_i$. Then if X is the sum of the X_i and if μ is $\mathbb{E}[X]$ for any $\delta \in (0, 1]$:

$$\mathbb{P}[X < (1-\delta)\mu] < \left(\frac{e^{-\delta}}{(1-\delta)^{(1-\delta)}} \right)^{\mu}$$

this bound is quite good, but can be clumsy to compute. We can simplify it to a weaker bound which is:

$$\mathbb{P}[X < (1-\delta)\mu] < \exp(-\mu\delta^2/2)$$

the simplifier bound makes it clear that the probability decrease exponentially with distance δ from the mean.

Berry-Esseen inequality(Theorem).

There is a constant C,such that if: $X_1, X_2, ..., X_n$ are i.i.d *r.v*, admitting moments of order 1 to 3, and $\mathbb{E}[X_i] = 0$, $Var(X_i) = \mathbb{E}[X_i^2] = \sigma^2 > 0$ and $\mathbb{E}[|X_i|^3] = \rho < +\infty$, and if there is $Y_n = \frac{X_1+X_2+...+X_n}{n}$ the sample mean of these variables; F_n the

distribution function of $\frac{Y_n\sqrt{n}}{\sigma}$ and ϕ function distribution of the normal distribution, then for all X and n:

$$|F_n(x) - \phi(x)| \leq \frac{C\rho}{\sigma^3\sqrt{n}}$$

Donsker's Theorem (Donsker,1952)

$\mathbf{Z}_n \Rightarrow \mathbf{Z} \equiv \mathbf{U}(F)$ in $D(\mathbb{R}, \| \, . \, \|_\infty)$ where $\mathbf{U}$ is a standard Brownian bridge process on [0,1]. Thus $\mathbf{U}$ is a zero-mean gaussian process with covariance function:

$$\mathbb{E}(\mathbf{U}(s)\mathbf{U}(t)) = s \wedge t - st, \qquad s, t \in [0, 1]$$

This means that we have: $\mathbb{E}g(\mathbf{Z}_n) \to \mathbb{E}g(\mathbf{Z})$ for any bounded, continuous function $g : D(\mathbb{R}, \| \, . \, \|_\infty) \to \mathbb{R}$.

Theorem (Van der Vaart and Wellner)

Suppose that $\mathcal{F}_1, ..., \mathcal{F}_k$ are Donsker classes with $\| \, \mathbb{P} \, \|_{\mathcal{F}_i} \leq \infty$ for each i. Suppose that $\varphi : \mathbb{R}^k \to \mathbb{R}$ satisfies:

$$|\varphi(f(x)) - \varphi(g(x))|^2 \leq \sum_{l=1}^{k} (f_l(x) - g_l(x))^2.$$

for every $f, g \in \mathcal{F}_1 \times ... \times \mathcal{F}_k$ and x. Then the class $\varphi(\mathcal{F}_1, ..., \mathcal{F}_k)$ is Donsker provided that $\varphi(f_1, ..., f_k)$ is square integrable for at least one $(f_1, ..., f_k)$

Bernstein's inequality (Theorem)

Let $X_1, ..., X_n$ be independent Bernoulli random variables taking values +1 , -1 with probability 1/2, then for every positive ε,

$$\mathbb{P}\left(\left| \, \frac{1}{n} \sum_{i=1}^{n} X_i \, \right| > \varepsilon \right) \leq 2 \exp\left(-\frac{n\varepsilon^2}{2(1+\frac{\varepsilon}{3})} \right)$$

Stone's theorem (Stone 1977).

Consider the following five conditions:

(i) *There is a constant C such that, for every Borel measurable function $g : \mathbb{R}^d \longrightarrow \mathbb{R}$ with $\mathbb{E}|g(\boldsymbol{X})| < \infty$*

$$\mathbb{E}[\sum_{i=1}^{n} |W_{ni}(\mathbf{X})||g(\mathbf{X}_i)|] \leq C\mathbb{E}|g(\mathbf{X})| \quad \textit{for all} \quad n \geq 1$$

(ii) *There is a constant $D \geq 1$ such that*

$$\mathbb{P}\{\sum_{i=1}^{n} |W_{ni}(\mathbf{X})| \leq D\} = 1 \quad \textit{for all} \quad n \geq 1$$

(iii) *For all $a > 1$,*

$$\sum_{i=1}^{n} |W_{ni}(\mathbf{X})| \mathbb{1}_{[\|\mathbf{X}_i - \mathbf{X}\| > a]} \to 0 \quad \textit{in probability}$$

(iv) *One has*

$$\sum_{i=1}^{n} W_{ni}(\mathbf{X}) \to 1 \quad \textit{in probability}$$

(v) *One has*

$$\max_{1 \leq i \leq n} |W_{ni}(\mathbf{X})| \to 0 \quad \textit{in probability}$$

If (i)-(v) are satisfied for any distribution of $\mathbf{X}$, then the corresponding regression function estimate r_n is universally L^p-consistent ($p \geq 1$), that is

$$\mathbb{E}|r_n(\mathbf{X}) - r(\mathbf{X})|^p \to 0$$

For all distributions of $(\mathbf{X}, Y)$ with $\mathbb{E}|Y|^p < \infty, p \geq 1$

Suppose, conversely that r_n is universally L^p-consistent. Then (iv) and (v) hold for any distribution of $\mathbf{X}$. Moreover, if the weights are nonnegative for all $n \geq 1$, then (iii) is satisfied. Finally, if the weights are nonnegative for all $n \geq 1$ and (ii) holds,then (i) holds as well.

Jensen's inequality (Theorem).

Let $\mathbf{X}$ *be a real valued random variable such that* $\mathbb{E}|\mathbf{X}| < \infty$*, and let* $g : \mathbb{R} \to \mathbb{R}$ *be convex function such that* $\mathbb{E}|g(\mathbf{X})| < \infty$*. Then*

$$g(\mathbb{E}\mathbf{X}) \quad \leq \quad \mathbb{E}(g(\mathbf{X}))$$

0.3 *Kernel estimator*

We will give the estimate of the density p. Let $X_1, ..., X_n$ be independent identically distributed (i.i.d) random variables that have a probability density p with respect to the Lebesgue measure on $\mathbb{R}$.
Here, the empirical distribution function

$$F_n(x) = \frac{1}{n}\sum_{i=1}^{n} I(X_i \leq x)$$

Where I(.) denotes the indicator function. By the strong law of large numbers, we have

$$F_n(x) \longrightarrow F(x), \forall x \in \mathbb{R} \quad \text{almost surely as} \quad n \longrightarrow \infty.$$

There fore, $F_n(x)$ is a consistent estimator of F(x) for every $x \in \mathbb{R}$. Now, the question is: *How can we estimate the density p?*
for sufficiently small $h > 0$, we can write an approximation

$$p(x) \approx \frac{F(x+h) - F(x-h)}{2h}$$

Replacing F by the estimate F_n, we define

$$\hat{p}_n^R(x) = \frac{F_n(x+h) - F_n(x-h)}{2h}$$

The function $\hat{p}_n^R$ is an estimator of p called the Rosenblatt estimator. We can rewrite it in the form

$$\hat{p}_n^R(x) = \frac{1}{2nh}\sum_{i=1}^{n} I(x-h < X_i \leq x+h) = \frac{1}{nh}\sum_{i=1}^{n} K_0\left(\tfrac{X_i - x}{h}\right)$$

Where $K_0(u) = \frac{1}{2}I(-1 < u \leq 1)$. A simple generalization of the Rosenblatt estimator is given by

$$\hat{p}_n = \frac{1}{nh}\sum_{i=1}^{n} K\left(\tfrac{X_i - x}{h}\right)$$

Where
1) $K : \mathbb{R} \longrightarrow \mathbb{R}$ is an integrable function.
2)$\int K(u)du = 1$.
The function K is called a kernel and the parameter h is called a bandwidth of the estimator $\hat{p}_n$. the function $x \longmapsto \hat{p}_n(x)$ is called the kernel density estimator or the Parzen-Rosenblatt estimator.
In the asymptotic framework, as $n \longrightarrow \infty$, we will consider h_n is the bandwidth h that depends on n, and we will suppose that the sequence $(h_n)_{n\geq 1}$ tends to 0 as $n \longrightarrow \infty$. The notation h without index n will also be used for brevity whenever this causes no ambiguity.
We have here, some classical examples of kernels:

name of kernel	formula
Rosenblatt(rectangular)	$K(u) = \frac{1}{2}I(\|u\| \leq 1)$
triangular	$K(u) = (1 - \|u\|)I(\|u\| \leq 1)$
parabolic(Epanechnikov)	$K(u) = \frac{3}{4}(1 - u^2)I(\|u\| \leq 1)$
biweight	$K(u) = \frac{15}{16}(1 - u^2)^2 I(\|u\| \leq 1)$
Gaussian	$K(u) = \frac{1}{\sqrt{2\pi}} \exp(-u^2/2)$
Silverman	$K(u) = \frac{1}{2} \exp(-u/\sqrt{2}) \sin(\|u\|/\sqrt{2} + \pi/4)$
Cosine	$K(u) = \frac{\pi}{4} \cos(u\pi/2)\mathbb{1}_{[-1,1]}(u)$

Note that if the kernel K takes only nonnegative values and if $X_1, ..., X_n$ are fixed, then the function $x \longmapsto \hat{p}_n(x)$ is a probability density.
The Parzen-Rosenblatt estimator can be generalized to the multidimensional case. for example if we have a kernel density estimator in two dimensions as follows. Suppose that we observe n pairs of random variables (X_1, Y_1),...,(X_n, Y_n) such that (X_i, Y_i) are i.i.d. with a density $p(x, y)$ in $\mathbb{R}^2$.
A kernel estimator of $p(x, y)$ is then given by the formula

$$\hat{p}_n(x, y) = \frac{1}{nh^2} \sum_{i=1}^{n} K\left(\frac{X_i - x}{h} \right) K\left(\frac{Y_i - y}{h} \right)$$

Where: $K : \mathbb{R} \longrightarrow \mathbb{R}$ is a kernel defined as above and $h > 0$ is a bandwidth.

Chapter 1

k-Nearest Neighbors: State of the Art

1.1 *Literature about the k nearest neighbors method*

The method of the kernel is the most used and known in economic nonparametric approaches. However, this method with its window fixed h, sometimes leads to over-smoothing or sub-smoothing. The first when you can have for a given points x many points in the interval $[x-h; x+h]$; and the second when you can get for a point x given, few points in the same gap in certain beaches of data media.

The method k-nearest neighbor (k-NN) tries to find a solution to this problem. Instead of using a fixed bandwidth h and leave the number of points of this interval $[x-h; x+h]$ varies depending to x, the method of kNN rather the number of points of the interval and lets the wooden windows be varied.

In resent years, a wide range of literature in the field of the estimation by the method of the k-nearest neighbors (k-NN) is provided by the literature reviews because of many advantages it offers. The first advantage of this method comes from the nature of the smoothing parameter h which is a positive real number. However, in our method, the latter and by a real random variable is replaced (H_n) for the functional

explanatory variable (X).

Several other aspects of this method; in the functional part, it respects the local structure of the data, which is essential in infinite dimension. It is commonly used in practice and like in Ferraty and Vieu (2006) [15] and proves easy to operate because the user has only one parameter controlling (the number of k nearest neighbor), this parameter takes its values in a finite set. In addition, this method allows to build a neighborhood in every way adapted to the data.

Our aim is to provide the first theoretical rational and practical to the current use of the k-NN method in functional nonparametric estimation.

Proposed for the first time by Loftsgaarden and Quesenberry in 1965 for estimating densities. Then Mack and Rosenblatt(1979) [28] make a detailed study of the use of this method also for estimating densities, that was formulated and applied for the first time by Fix and Hodges (1989) in the part of classification problems.

The bibliographic of the estimate by the k-NN method existed for Royall (1966) and Stone (1977) [39] which is started by estimating the regression function in the multivariate case. It was used then by Nielson (1967) (Cover and Hart,1967) [13] for pattern recognition. Other authors has also the interest of studying estimator of the regression function, find for example Collomb (1980) [12] which showed the different types of convergence (probability and almost complete), Mack (1981) [27] studied the L^2- convergence and the asymptotic distribution, the uniform convergence is given by Devroye (1978) [14] and Devroye (1981).

Liu and Lu (1997) study the use of k-NN method for the semi-parametric regression, and Li and Racine (2004) [23] study the use of this method for nonparametric regression. In the case of functional data, Ferraty and Vieu (2004) began with an introduction on the estimation of k-NN, Burba et al.(2008) [10] obtained the almost complete convergence of the estimator of the regression function with independent and identically distributed data.

Finally Attouch and Benchikh (2012) [04] established the asymptotic normality of the regression function.

One of the important issues in this use is the choice of the number of neighbors k to with hold. This question was dealt with by Ouyang et al.(2006) [34], they proposed

several methods of choosing k; and when k is specified for each variable, the method of kNN enables to remove automatically the non-significant variables (Li and Gong 2008) [24], and besides the time and memory limitation, Gongde Guo selects the value of k using model based approach. The model proposed automatically selects the value of k.

T.M.Cover and P.E.Hart (1967) [13] purpose kNN in which nearest neighbor is calculated on the basis of value of k, that specifies how many nearest neighbors are to be considered to define class of a sample data point. T.Bailey and A.K.Jain(1978) [05] improve kNN which is based on weights. The training points are assigned weights according to their distances from sample data point. But still, the computational complexity and memory requirements remain the main concern always. To overcome memory limitation, size of data set is reduced.

For this, the repeated patterns, which do not add extra information, are eliminated from training samples (K.Chidananda and G.Krishna(1979) [11]) and (E. Alpaydin(1997) [02]). To further in prove, the data points which do not affect the result are also eliminated from training data set (Geoffrey W.Gates [17]).

Similarly, many improvements are proposed to improve speed of classical kNN using concept of ranking; see for example: S.C.Bagui, S.Bagui,K.Pal (2003) . Y.Zeng, Y.Yang, L.Zhou(2009) [42] give the false neighbor information, clustering by: H. Parvin, H.Alizadeh and B.Minaei at 2008 [35]. The NN training data set can be structured using various techniques to improve over memory limitation of kNN. The kNN implementation can be done using ball tree(T.Liu, A.W.Moore, A.Gray (2006)) and (S.N.Omohundro (1989)), k-d tree (R.F Sproull [38]), nearest feature line (S.Z Li, K.L.Chan (2000)), tunable metric (Y.Zhou, C.Zhang (2004) [43]), principal axis search tree (Y.C.Liaw, M.L.Leou [25]) and orthogonal search tree (J.Mcname [30]). However, the k-NN method presents a major technical difficulty: the selection of the nearest neighbors gives a random bandwidth. The second problem, linked with the functional nature of the data, is that we do not suppose the existence of a density; because, in infinite-dimensional spaces, we do not have any standard measure like the Lebesgue measure in the multivariate case.

1.2 *The link between the nearest neighbor and kernel methods*

First, we will start by given the idea of the k-nearest neighbor method that is based on the definition of the probability density,

$$f(x) = \lim_{h \longrightarrow 0} (2h)^{-1} \mathbb{P}(x - h < X < x + h)$$

Then, noting that we expect $k = n(2h)f(x)$ observations falling in a box of width $2h$ and centered at the point of interest x.
Recall that the naive density estimator is based on using a fixed bandwidth h, calculating the number $\hat{k}$ of observations such that :$\hat{k} \in [x - h; x + h]$;
and we have

$$\hat{f}_n(x) = \frac{\hat{k}}{2nh} \qquad (1)$$

In contrast, the nearest neighbor method is based on a fixed number of points k that determines the width of a box in a search.
Thus, we calculate the euclidean distance $\hat{h}$ from the point of interest x to the distant k-th observation and define the k-th nearest neighbor density estimate by

$$\tilde{f}_n(x) = \frac{k}{2n\hat{h}} \qquad (2)$$

Note that for x less than then smallest data point $X_{(1)}$ we have

$$\hat{h}(x) = X_{(k)} - x$$

($X_{(k)}$: the k-th ordered observation), namely, that density is inversely proportional to the size of the box needed to contain a fixed number k of observations. The drawback of the nearest neighbor method is that the derivative of a nearest neighbor estimate is discontinuous. as a result, the estimate can give a wrong impression. Also, this estimate is not integrable due to its heavy tails.
The idea of nearest neighbor method can be used also in a kernel estimator where the bandwidth is chosen to be $\hat{h}$. Such a kernel estimator is called a k-th neighbor kernel estimate

$$\tilde{f}_n(x) = (n\hat{h}(x))^{-1} \sum_{l=1}^{n} K((x - X_l)/\hat{h}(x)) \qquad (3)$$

i.e: (3) is a kernel estimate with a data-driven bandwidth. However, this not an entirely data-driven method, because a choice of k should be made. Note that this

generalized estimate becomes the ordinary *k-th* nearest neighbor estimate when the kernel function is rectangular.
Now, to define a nearest neighbor estimate in s-dimensional space, let $d_k(X)$ be the euclidean distance from X to the *k-th* nearest data point; and let $V_k(x)$ be the volume of the s-dimensional sphere of radius $d_k(X)$. Thus

$$V_k(X) = c_s[d_k(X)]^d$$

where c_s is the volume of the s-dimensional sphere with unit radius , that is, $c_1 = 2, c_2 = \pi, c_3 = 4\pi/3$,...ect Then, the nearest neighbor method is defined by

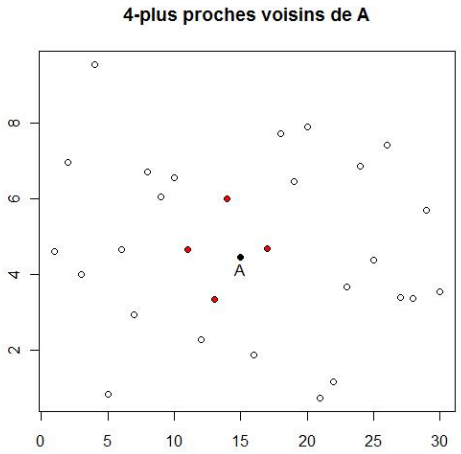
$$\tilde{f}_n(X) = \frac{k}{nV_k(X)} \qquad (4)$$

Note that if we set the kernel function $K(X) = 1/c_k$ within the sphere of unit radius and $K(X) = 0$. otherwise, then the nearest neighbor method is identical to a kernel smoothing. This connection between the kernel and nearest neighbor method demonstrates that a study of the nearest neighbor can be based on the theory of kernel estimation.

1.3 *k nearest neighbor to the real and the vector cases*

Definition of k-nearest neighbors

k nearest neighbor is a supervised learning algorithm where the result of new instance query is classified based on majority of k nearest neighbor category.



The point x is a k nearest neighbors of the point y if and only if

$$card\{z, \mathrm{d}(z, y) \geq \mathrm{d}(x, y)\} \geq n - \mathrm{k}$$

The purpose of this algorithm is to classify a new object based on attributes and training samples. The classifiers do not use any model to fit and only based on memory. Given a query point, we find k number of objects or(training points) closest to the query point. The classification is using majority vote among the classification of the k objects. any ties can be broken at random. k nearest neighbor algorithm used neighborhood classification as the prediction value of the new query instance.

This sense of ordering on many different objects helps us place then in time and space and to make sense of the world. It is what allows us to build clusters/neighbors-both in databases on computers as well as in our daily lives.

This definition of nearness that seems to be ubiquitous also allows us to make predictions, so The nearest neighbor prediction algorithm as:

Objects that are « near » to each other will have similar prediction values as well. Thus if you know the prediction value of one of the objects you can predict it for it's nearest neighbors.

k Nearest neighbor Algorithm

Here is step by step on how to compute k- nearest neighbors (k-NN) algorithm:

1. determine parameter k=number of nearest neighbors.
2. Calculate the distance between the query-instance and all the training samples.
3. Determine nearest neighbors based on the k-*th* minimum distance.
4. Gather the category Y of the nearest neighbors.
5. Use simple majority of the category of nearest neighbors as the prediction value of the query instance or predict the mean for numeric prediction.

A real example of nearest neighbors method

If you look at the people in your neighborhood (those people are in fact geographically near to you). You may notice that, in general, you all have somewhat similar incomes. Thus if your neighbor has an income greater than $ 150.000 chances are good that you too have a high income. Certainly the chances that you have a high income are greater when all of your neighbors have incomes over $ 150.000 than if all of your neighbors have incomes of $ 25.000 . Within your neighborhood there may still be a wide variety of incomes possible among even your «closest» neighbors but if you had to predict some one's income based on only knowing their neighbors you're best chance of being right would be to predict the incomes of the neighbors who live closest to the unknown person.

The nearest neighbor prediction algorithm works in very much the same way except that «nearness» in a data base may consist of a variety of factors not just where the person lives. It may, for instance, be far more important to know which school someone attended and what degree they attained when predicting income. The better definition of «near» might in fact be other people that you graduated from college with rather than the people that you live next to.

Nearest Neighbor techniques are among the easiest to use; and understand because they work in a way similar to the way that people think-by detecting closely matching examples. They also perform quite well in terms of automation, as many of the algorithms are robust with respect to dirty data and missing data.

Univariate case

In this case x is a real number, we denote by : $R_x = R_n(x)$ the euclidean distance between the point x and *k-th* nearest neighbor of x amongst the x_i ; this is the smallest ball of center x contains k points of $\{1,...,n\}$.
The density estimator f at the point x by the kNN method is given by the following formula

$$\hat{f}(x) = \frac{1}{nR_x} \sum_{i=1}^{n} \left(\frac{1}{2} \right) \mathbb{1} \left(\frac{|x-X_i|}{R_x} \leq 1 \right) = \frac{k}{2nR_x}$$

where

$$\mathbb{1}(x) = \begin{cases} 1, & \frac{|x-X_i|}{R_x} \leq 1, \forall i \\ \\ 0, & \text{otherwise} \end{cases}$$

A consistent estimator is obtained for $f(x)$ where $k = k(n)$ is chosen such $k \longrightarrow \infty$ and $\frac{k}{n} \longrightarrow 0$ when $n \longrightarrow \infty$. $\frac{k}{n}$ that plays a similar role in the approach by kernel. Indeed, the condition $k \longrightarrow \infty$ and $\frac{k}{n} \longrightarrow 0$ is $nk \longrightarrow \infty$ and $h \longrightarrow 0$.

Multivariate case

In this case x is a q-vector ($x \in \mathbb{R}^q$). The estimator of f at x by the k-NN approach is as follows

$$\hat{f}(x) = \frac{1}{nR_x^q} \sum_{i=1}^{n} \frac{1}{c_0} \mathbb{1} \left(\frac{\|x - X_i\|}{R_x} \leq 1 \right) = \frac{k}{c_0 n R_x^q} \tag{5}$$

where

$$c_0 = \frac{\pi^{q/2}}{\Gamma(\frac{q+2}{2})}$$

is the volume of the unit ball in $\mathbb{R}^q$, and $\Gamma(.)$ is the function defined by

$$\Gamma(\alpha) = \int_0^\infty t^{\alpha-1} e^{-t} dt.$$

($\Gamma(\alpha + 1) = \alpha\Gamma(\alpha), \Gamma(1/2) = \sqrt{\pi}$, and $\Gamma(1) = 1$).
The general form of equation (5) is given by

$$\hat{f}(x) = \frac{1}{nR_x^q} \sum_{i=1}^{n} w \left(\frac{\|x - X_i\|}{R_x} \right) \tag{6}$$

where $w(.)$ is a function of bounded weight, symmetrical, non-negative integral such that

$$\int_{\mathbb{R}^q} w(v) dv = 1 \tag{7}$$

considering the weight function:

$$w(v) = \begin{cases} \frac{1}{c_0}, & \| v \| \leq 1. \\ 0, & \| v \| > 1 \end{cases}$$

1.3.1 *Some asymptotic results for density estimator*

Let $(X_1, ..., X_n)$ be $i.i.d$ observations with common distribution μ on $\mathbb{R}^d$, equipped with the standard euclidean norm $\| \ . \ \|$. The empirical measure μ_n based on $(X_1, ..., X_n)$ is defined, for any Borel set $A \subset \mathbb{R}^d$ by

$$\mu_n(A) = \frac{1}{n} \sum_{i=1}^{n} \mathbb{1}_{[\mathbf{x}_i \in A]}.$$

Moreover, given a sequence of positive integer $\{k_n\}$ such that $1 \leq k_n \leq n$. for $m_n = k_n/n$ the function d_{μ_n,m_n} takes the simple form

$$d^2_{\mu_n,m_n}(\mathbf{x}) = \frac{1}{k_n}\sum_{j=1}^{k_n} \parallel X_{(j)}(\mathbf{x}) - \mathbf{x} \parallel^2$$

is a weighted sum of the squares of the distance from $\mathbf{x}$ to its first k_n nearest neighbors.
where $X_j(\mathbf{x})$ is the *j-th* nearest neighbor to $\mathbf{x}$ among $X_1, ..., X_n$ and ties are broken arbitrarily.
Thus

$$\parallel X_{(1)}(\mathbf{x}) - \mathbf{x} \parallel \leq ... \leq \parallel X_{(n)}(\mathbf{x}) - \mathbf{x} \parallel .$$

Our goal is to establish some pointwise asymptotic properties of the estimate f_n. To this aim, we note once and for all that for any $\rho > 0$. All quantities of the form $\int_{[0,1]} t^\rho \nu(dt)$ are finite and positive. Moreover, for $\rho \geq 1$ as $k_n \longrightarrow \infty$,

$$\frac{1}{k_n^\rho}\sum_{j=1}^{k_n} p_{nj} j^\rho = \int_{[0,1]} t^\rho \nu(dt) \left(1 + O\left(\frac{1}{k_n} \right) \right).$$

The symbol λ stands for the Lebesgue measure on $\mathbb{R}^d$. We start by establishing the weak pointwise consistency of f_n.

Theorem 1.1

if $k_n \to \infty$ and $k_n/n \to 0$; then generalized k-nearest neighbor estimate f_n is weakly consistent at λ- almost all $\mathbf{x}$; that is $f_n(\mathbf{x}) \to f(\mathbf{x})$ in probability at λ-almost all $\mathbf{x}$ as $n \to \infty$.

Our next result states the mean square consistency of the generalized k-nearest neighbor estimate.

Theorem 1.2

We have, at λ-almost all $\mathbf{x}$, $\mathbb{E}([f_n^2(\mathbf{x})]) < \infty$. whenever $k_n \geq 5$. Furthermore, if $k_n \to \infty$ and $k_n/n \to 0$, then, for such $\mathbf{x}$, $\mathbb{E}[f_n(\mathbf{x}) - f(\mathbf{x})]^2 \to 0$ as $n \to \infty$.

The asymptotic normality of the original Loftsgaarden and Quesenberry k-NN estimate has been established by Moore and Yackel. These authors proved that for f

sufficiently smooth in a neighborhood of $\mathbf{x}$; $f(\mathbf{x}) > 0, k_n \to \infty$ and $k_n/n^{2/(d+2)} \to 0$ as $n \to \infty$,then

$$\sqrt{k_n}\frac{f_n(\mathbf{x}) - f(\mathbf{x})}{f(\mathbf{x})} \to^{\mathcal{D}} N.$$

Where N is a standard normal random variable, $\Gamma(.)$ be the gamma function; and $[\partial^2 f(\mathbf{x})/\partial \mathbf{x}^2]$ is the Hessian matrix of f at $\mathbf{x}$ which is given by

$$\left[\frac{\partial^2 f(\mathbf{x})}{\partial \mathbf{x}^2} \right]_{i,j} = \frac{\partial^2 f(\mathbf{x})}{\partial x_i \partial x_j}$$

Notation.

$tr(A)$ stand for the trace of the square matrix A.
$\{\zeta_n\}$ is a sequences of random variables and $\{u_n\}$ is a deterministic sequence. Where

$$\zeta_n = o(u_n) \Rightarrow \zeta_n/u_n \to 0 \quad \textit{in probability} \quad \textit{as} \quad n \to \infty$$

and

$$\zeta_n = O(u_n) \Rightarrow \zeta_n/u_n \quad \textit{is bounded in probability} \quad \textit{as} \quad n \to \infty$$

Theorem 1.3

Let $\mathbf{x} \in \mathbb{R}^d$ *and assume that* f *has derivatives of second order at* $\mathbf{x}$; *with* $f(\mathbf{x}) > 0$. *Let*:

$$v^2 = \frac{\int_0^1 (1-\phi(t))^2 dt}{[\int_{[0,1]} t\nu(dt)]^2} \quad \text{and} \quad b = \frac{\int_{[0,1]} t^{1+2/d}\nu(dt)}{\int_{[0,1]} t\nu(dt)}$$

$$\textit{with} \qquad \phi(t) = \int_{[0,1]} \nu(du), \quad t \in [0,1]$$

$$\textit{Let also} \qquad c(\mathbf{x}) = \frac{1}{2(d+2)\pi}\Gamma^{2/d}\left(\tfrac{d+2}{2} \right) tr\left[\tfrac{\partial^2 f(\mathbf{x})}{\partial \mathbf{x}^2} \right]$$

Then; if N *denotes a standard normal random variable, and if* $k_n \to \infty$ *and* $\frac{k_n}{n} \to 0$

$$f_n(\mathbf{x}) - f(\mathbf{x}) =^{\mathcal{D}} \frac{f(\mathbf{x})v}{\sqrt{k_n}}N + \frac{c(\mathbf{x})b}{f^{2/d}(\mathbf{x})}\left(\frac{k_n}{n} \right)^{2/d} + o\left(\frac{1}{\sqrt{k_n}} + \left(\frac{k_n}{n} \right)^{2/d} \right)$$

1.3.2 *Proof of Theorems.*

Throughout this section, we let $\mathcal{B}(\mathbf{x}, r)$ be the closed ball in $\mathbb{R}^d$ of radius r centered at $\mathbf{x}$ and denote by μ the probability measure associated with the density f. The collection of all $\mathbf{x}$ with $\mu(\mathcal{B}(\mathbf{x}, \varepsilon)) > 0$ for all $\varepsilon > 0$ is called the support of μ. We denote it by supp μ and note that it may alternatively be defined as the smallest closed subset of $\mathbb{R}^d$ of μ-measure 1.

Two basic Lemmas.

We will make repeated use of the following two lemmas:

Lemma 1.1.

Let $U_1, ..., U_n$ be i.i.d uniform $[0, 1]$ random variables with order statistics $U_{(1)} \leq ... \leq U_{(n)}$. Then

$$(U_{(1)}, ..., U_{(n)}) =^{\mathcal{D}} \left(\frac{\sum_{j=1}^{1} E_j}{n+1}, \cdots, \frac{\sum_{j=1}^{n} E_j}{n+1} \right) (1 + \zeta_n)$$

Where $E_1, ..., E_n$ is a sequence of *i.i.d* standard exponential random variables and $\zeta_n = O_{\mathbb{P}}(n^{-1/2})$ as $n \to \infty$.
Furthermore, for all positive integers

$$\sup_{n \geq 2r} [n^{r/2} \mathbb{E}|\zeta_n|^r] < \infty.$$

Proof.

It is well known that if $E_1, ..., E_{n+1}$ is a sequence of *i.i.d* standard exponential random variables (see, e.g.,Devroye (1986, Chapter 5)), then

$$(U_{(1)}, ..., U_{(n)}) =^{\mathcal{D}} \left(\frac{\sum_{j=1}^{1} E_j}{\sum_{j=1}^{n+1} E_j}, \cdots, \frac{\sum_{j=1}^{n} E_j}{\sum_{j=1}^{n+1} E_j} \right)$$

Let G_{n+1} be the gamma (n+1) random variable $\sum_{j=1}^{n+1} E_j$. Then by the central limit theorem

$$\sqrt{n} \left(\frac{G_{n+1}}{n+1} - 1 \right) \to^{\mathcal{D}} N$$

Where N is a standard normal random variable. Thus, by an application of the delta method, we obtain

$$\sqrt{n}\left(\frac{n+1}{G_{n+1}} - 1\right) \to^{\mathcal{D}} N$$

and the first part of the lemma follows by setting

$$\zeta_n = \frac{n+1}{G_{n+1}} - 1$$

To prove the second statement observe that by the Cauchy-Schwarz inequality

$$\mathbb{E}\left|\frac{n+1}{G_{n+1}} - 1\right|^r \leq \sqrt{\mathbb{E}|G_{n+1} - (n+1)|^{2r}} \times \sqrt{\mathbb{E}G_{n+1}^{-2r}}$$

The first term in the above product is $O(n^{r/2})$ (see e.g, Willink(2003)) whereas the second one is infinite for $n+1 \leq 2r$ and $O(1/n^r)$ otherwise.
It follows that

$$\sup_{n \geq 2r}\left[\, n^{r/2}\mathbb{E}\left|\, \tfrac{n+1}{G_{n+1}} - 1 \,\right|^r \,\right] < \infty$$

Lemma 1.2.

Let $E_1, E_2, ..$ be a sequence of i.i.d standard exponential random variables and let $\{k_n\}$ be a sequence of positive integers. For $j = 1, ..., k_n$; let

$$p_{nj} = \int_{]\frac{j-1}{k_n}, \frac{j}{k_n}]} \nu(dt)$$

where ν is a given probability measure on $[0, 1]$ with no atom at 0. Fix $\rho \geq 1$; then, if $k_n \to \infty$

$$\frac{\sum_{j=1}^{k_n} p_{nj}(E_1 + ... + E_j)^\rho}{\sum_{j=1}^{k_n} p_{nj} j^\rho} = 1 + \zeta_n$$

where $\zeta_n = O_{\mathbb{P}}(k_n^{-1/2})$ and, for all positive integers r

$$\sup_{n \geq 1}[k_n^{r/2}\mathbb{E}|\zeta_n|^r] < \infty$$

In addition, letting: $\phi(t) = \int_{[0,t]} \nu(du), \quad t \in [0, 1]$ and $\sigma^2 = \int_0^1 (1 - \phi(t))^2 dt$. Then, on an appropriate probability space, there exists a standard normal random variable N such that

$$\frac{1}{k_n}\sum_{j=1}^{k_n} p_{nj}(E_1 + ... + E_j) = \int_{[0,1]} tr(dt) + \frac{\sigma}{\sqrt{k_n}}N + \zeta_n'$$

where $\zeta_n' = o_{\mathbb{P}}(k_n^{-1/2})$ and, for all positive integers r

$$\sup_{n\geq 1}[k_n^{r/2}\mathbb{E}|\zeta_n'|^r] < \infty.$$

Proof.

Denote by $\lceil . \rceil$ the ceiling function and observe that, since ν has no atom at 0, $\sum_{j=1}^{k_n} p_{nj}(E_1 + ... + E_j)^\rho = \int_{[0,1]}(E_1 + ... + E_{\lceil tk_n \rceil})^\rho \nu(dt) = \int_{[0,1]}(\lceil tk_n \rceil^\rho \nu(dt)$, where we set:

$$S_{\lceil tk_n \rceil} = \sum_{j=1}^{\lceil tk_n \rceil}(E_j - 1)$$

Note that $S_{\lceil tk_n \rceil}$ is a sum of $i.i.d$ zero mean random variables. There for,

$$\sum_{j=1}^{k_n} p_{nj}(E_1 + ... + E_j)^\rho = \int_{[0,1]} \lceil tk_n \rceil^\rho \nu(dt) + \int_{[0,1]} \left[\left(1 + \frac{S_{\lceil tk_n \rceil}}{\lceil tk_n \rceil} \right)^\rho - 1 \right] \lceil tk_n \rceil^\rho \nu(dt)$$

By an application of Donsker's and continuous mapping theorems (see,e.g.Vander Vaart and Wellner [41]), as $k_n \to \infty$

$$\int_{[0,1]} \left[\left(1 + \frac{S_{\lceil tk_n \rceil}}{\lceil tk_n \rceil} \right)^\rho - 1 \right] \lceil tk_n \rceil^\rho \nu(dt) = \int_{[0,1]} \rho \frac{S_{\lceil tk_n \rceil}}{\lceil tk_n \rceil} \lceil tk_n \rceil^\rho \nu(dt) + k_n^\rho \zeta_{n1}$$

where $\zeta_{n1} = O_{\mathbb{P}}(k_n^{-1})$ and, for all positive integers r; $\sup_{n\geq 1}[k_n^r \mathbb{E}|\zeta_{n1}|^r] < \infty$. Similarly:

$$\int_{[0,1]} \rho \frac{S_{\lceil tk_n \rceil}}{\lceil tk_n \rceil} \lceil tk_n \rceil^\rho \nu(dt) = k_n^\rho \zeta_{n2}$$

where $\zeta_{n2} = O_{\mathbb{P}}(k_n^{-1/2})$ and, for all positive integers r, $\sup_{n\geq 1}[k_n^{r/2}\mathbb{E}|\zeta_{n2}|^r] < \infty$. Consequently,

$$\frac{1}{k_n^\rho} \sum_{j=1}^{k_n} p_{nj}(E_1, ..., E_j)^\rho = \int_{[0,1]} t^\rho \nu(dt) + \zeta_n$$

where $\zeta_n = O_{\mathbb{P}}(k_n^{-1/2})$ and, for all positive integers r, $\sum_{n\geq 1}[k_n^{r/2}\mathbb{E}|\zeta_n|^r] < \infty$. The conclusion of the first assertion follows by observing that, for $\rho \geq 1$,

$$\frac{1}{k_n^\rho} \sum_{j=1}^{k_n} p_{nj} j^\rho = \int_{[0,1]} t^\rho \nu(dt) \left(1 + O\left(\frac{1}{k_n} \right) \right)$$

The proof of the second assertion requires a bit more care we already know that

$$\frac{1}{k_n}\sum_{j=1}^{k_n} p_{nj}(E_1+...+E_j) \;=\; \int_{[0,1]} tr(dt) + \frac{1}{k_n}\int_{[0,1]} S_{\lceil tk_n\rceil}\nu(dt) + \zeta_{n3} \tag{1.1}$$

where $\zeta_{n3} = O(k_n^{-1})$. With respect to the second term on the right-hand side of (1.1) , we have

$$\frac{1}{k_n}\int_{[0,1]} S_{\lceil tk_n\rceil}\nu(dt) = \frac{1}{k_n}\sum_{j=1}^{k_n}[(E_j-1)\int_{]\frac{j-1}{k_n},1]}\nu(dt)]$$

Clearly, letting

$$\sigma_{nj} = \int_{]\frac{j-1}{k_n},1]}\nu(dt), \qquad j=1,...,k_n$$

and

$$\phi(t) = \int_{[0,t]}\nu(du), \quad t\in[0,1]$$

We may write

$$\sum_{j=1}^{k_n}\sigma_{nj}^2 = \sum_{j=1}^{k_n}\Big(\; 1-\phi\Big(\;\frac{j-1}{k_n}\;\Big)\;\Big)^2$$

as a consequence, setting

$$\sigma^2 = \int_0^1 (1-\phi(t))^2 dt,$$

and using the fact that

$$0\le(1-\phi(t))^2\le 1$$

is a monotone nonincreasing function, a Riemannian argument shows that

$$\frac{1}{k_n}\sum_{j=1}^{k_n}\sigma_{nj}^2 \in \Big[\;\sigma^2, \sigma^2+\tfrac{1}{k_n}\;\Big] \tag{1.2}$$

There for, we obtain via the Komlós, Major and Turnády strong approximation result (see Komlós, Major, and Tusnády and Mason) that, on the same probability space, there exists a sequence E_1, E_2,...of *i.i.d* standard exponential random variables and a sequence N_1, N_2,...of standard normal random variables such that, for positive constants C_1 and λ_1 and for all $x\ge 0$,

$$\mathbb{P}\left(\;\sqrt{k_n}\left|\;\frac{1}{\sqrt{\sum_{j=1}^{k_n}\sigma_{nj}^2}}\sum_{j=1}^{k_n}\sigma_{nj}(E_j-1) - N_{k_n}\;\right| > x\;\right) \le C_1 e^{-\lambda_1 x}$$

Using (1.2), we deduce that, for positive constants λ_2, λ_3 and all n large enough

$$\mathbb{P}\left(\sqrt{k_n} \left| \frac{1}{\sqrt{k_n}} \sum_{j=1}^{k_n} \sigma_{nj}(E_j - 1) - \sigma N_{k_n} \right| > x \right) \leq C_1 e^{-\lambda_2 x}$$

$$+\mathbb{P}(|N_{k_n}| > \lambda_3 \sqrt{k_n} x)$$

Thus, writing

$$\zeta_{n4} = \frac{1}{k_n} \sum_{j=1}^{k_n} \sigma_{nj}(E_j - 1) - \frac{\sigma}{\sqrt{k_n}} N_{k_n}$$

We see that

$$\frac{1}{k_n} \int_{[0,1]} S_{\lceil t k_n \rceil} \nu(dt) = \frac{\sigma}{\sqrt{k_n}} N_{k_n} + \zeta_{n4}$$

where

$$\zeta_{n4} = o_{\mathbb{P}}(k_n^{-1/2}) \quad \text{and} \quad \sup_{n \geq 1} [k_n^{r/2} \mathbb{E}|\zeta_{n4}|^r] < \infty$$

for all positive integers r. Plugging this identity into (1.1) leads to the desired result.

Proof of Theorem 1.1.

Let $\mathbf{x}$ be a Lebesgue point of f, that is an $\mathbf{x}$ for which

$$\lim_{r \to 0} \frac{\mu(\mathcal{B}(\mathbf{x}, r))}{\lambda(\mathcal{B}(\mathbf{x}, r))} = \lim_{r \to 0} \frac{\int_{\mathcal{B}(\mathbf{x}, r)} f(Y) dY}{\int_{\mathcal{B}(\mathbf{x}, r)} dY} = f(\mathbf{x})$$

as f is a density, we know that $\lambda-$almost all x satisfy the property given above. Assume first that $f(\mathbf{x}) > 0$. Fix $\varepsilon \in (0, 1)$ and find $\delta > 0$ such that

$$\sup_{0 < r \leq \delta} \left| \frac{\int_{\mathcal{B}(\mathbf{x}, r)} f(Y) dY}{\int_{\mathcal{B}(\mathbf{x}, r)} dY} - f(\mathbf{x}) \right| \leq \varepsilon f(\mathbf{x}) \tag{1.3}$$

Let F be the (continuous) univariate distribution function of $W = \| X - \mathbf{x} \|^d$. Note that if $w \leq \delta^d$, then

$$\begin{aligned} F(\mathbf{x}) &= \mathbb{P}(\| X - \mathbf{x} \|^d \leq w) \\ &= \mathbb{P}(X \in \mathcal{B}(\mathbf{x}, w^{1/d})) \\ &= \int_{\mathcal{B}(\mathbf{x}, w^{1/d})} f(Y) dY \in [(1 - \varepsilon) V_d f(\mathbf{x}) w, (1 + \varepsilon) V_d f(\mathbf{x}) w] \end{aligned}$$

Define $W_j = \| X_j - \mathbf{x} \|^d;\ j = 1, ..., n$, and let $W_{(1)} \leq ... \leq W_{(n)}$ be the order statistics, we have in fact the representation $W_{(j)} =^{\mathcal{D}} F^{\text{inv}}(U_{(j)})$ jointly for all j. Thus, provided $U_{(j)} \leq F(\delta^d)$

$$\frac{U_{(j)}}{(1+\varepsilon)V_d f(\mathbf{x})} \leq F^{\text{inv}}(U_{(j)}) \leq \frac{U_{(j)}}{(1-\varepsilon)V_d f(\mathbf{x})} \tag{1.4}$$

There for, on the event $[U_{(k_n)} \leq F(\delta^d)]$ the generalized k-nearest neighbor estimate may be written as follows

$$f_n(\mathbf{x}) =^{\mathcal{D}} \frac{\theta f(\mathbf{x})}{n} \frac{\sum_{j=1}^{k_n} p_{nj} j}{\sum_{j=1}^{k_n} p_{nj} U_{(j)}}$$

Where θ denotes some arbitrary random variable with values in $[1-\varepsilon, 1+\varepsilon]$.
Observe that $F(\delta^d) > 0$ and; as $k_n/n \to 0$, $\mathbb{P}(U_{(k_n)} \leq F(\delta^d)) \to 1$ as $n \to \infty$ (see,e.g.Devroye et al.). Thus, to prove that $f_n(x) \to f(x)$ in probability, it suffices to show that

$$\frac{\sum_{j=1}^{k_n} p_{nj} j}{n \sum_{j=1}^{k_n} p_{nj} U_{(j)}} \to 1 \qquad \textit{in probability}$$

But, by Lemma 1.1, we know that

$$(U_{(1)}, ..., U_{(n)}) =^{\mathcal{D}} \left(\frac{\sum_{j=1}^{1} E_j}{n+1}, ..., \frac{\sum_{j=1}^{n} E_j}{n+1} \right) (1+\zeta_n)$$

Where $E_1, ..., E_n$ are *i.i.d* standard exponential random variables and $\varepsilon \to 0$ in probability. Consequently

$$\frac{\sum_{j=1}^{k_n} p_{nj} j}{n \sum_{j=1}^{k_n} p_{nj} U_{(j)}} =^{\mathcal{D}} \frac{n+1}{n} \times \frac{\sum_{j=1}^{k_n} p_{nj} j}{\sum_{j=1}^{k_n} p_{nj} (E_1 + ... + E_j)} \times \frac{1}{1+\zeta_n}$$

which goes to 1 in probability as $k_n \to \infty$ according to the first statement of Lemma 1.2.
If $f(x) = 0$, two cases are possible. Suppose first that $\mathbf{x}$ belongs to the complement of supp μ. Then, clearly, for some positive constant C and all $n \geq 1$, almost surely,

$$f_n(\mathbf{x}) \leq \frac{C k_n}{n}.$$

But $f(\mathbf{x}) = 0$ and, by using the condition $k_n \to 0$, we deduce that

$$f_n(\mathbf{x}) \to f(\mathbf{x}) \qquad \textit{in probability}$$

as $n \to \infty$.
If x belongs to supp μ, the proof is similar to the case $f(\mathbf{x}) > 0$.
Just fix $\varepsilon \in (0,1)$ and find $\delta > 0$ such that

$$\sup_{0<r\leq\delta} \left| \frac{\int_{\mathcal{B}(\mathbf{x},r)} f(Y)dY}{\int_{\mathcal{B}(\mathbf{x},r)} dY} \right| \leq \varepsilon$$

Proof of Theorem 1.2.

Choose $\mathbf{x}$ a Lebesgue point of f. Assume first that $f(\mathbf{x}) > 0$ and fix ε and δ as in (1.3). Note that

$$f_n^2(\mathbf{x}) \;=\; \frac{1}{n^2 V_d^2} \left(\frac{\sum_{j=1}^{k_n} p_{nj} j}{\sum_{j=1}^{k_n} p_{nj} \|X_{(j)}(\mathbf{x}) - \mathbf{x}\|^d} \right)^2$$

Using $\frac{1}{k_n}\sum_{j=1}^{k_n} p_{nj} j \;\to\; \int_{[0,1]} tr(dt)$ and $\liminf_{n\to\infty} \sum_{\lceil k_n/2 \rceil}^{k_n} p_{nj} \;\geq\; \int_{[1/2,1]} \nu(dt)$, we have, for some positive constant C_1 and all $n \geq 1$

$$\mathbb{E}[f_n^2(\mathbf{x})] \;\leq\; \frac{C_1 k_n^2}{n^2} \mathbb{E}\left[\frac{1}{\|X_{(\lceil k_n/2 \rceil)}(\mathbf{x}) - \mathbf{x}\|^{2d}} \right]$$

If $U_{(1)} \leq \ldots \leq U_{(n)}$ are uniform $[0,1]$ order statistics, we may write, using inequality (1.4)

$$\mathbb{E}\left[\frac{1}{\|X_{(\lceil k_n/2 \rceil)}(\mathbf{x}) - \mathbf{x}\|^{2d}} \right] \;\leq\; C_2 \left(\mathbb{E}\left[\frac{1}{U^2_{(\lceil k_n/2 \rceil)}} \right] + \frac{1}{\delta^{2d}} \right)$$

For some positive constant C_2, it is know that $U_{(\lceil k_n/2 \rceil)}$ is beta distributed with parameters $\lceil k_n/2 \rceil$ and $n + 1 - \lceil k_n/2 \rceil$ (see, e.g,Devroye (1986)). Consequently, for $\lceil k_n/2 \rceil > 2$

$$\mathbb{E}\left[\frac{1}{\|X_{(\lceil k_n/2 \rceil)}(\mathbf{x}) - \mathbf{x}\|^{2d}} \right] \;\leq\; C_3 \left(\frac{n^2}{k_n^2} + \frac{1}{\delta^{2d}} \right)$$

Whence, for $k_n \geq 5$, $\mathbb{E}[f_n^2(\mathbf{x})] \leq C_4$, for some positive constant C_4. Next, if $f(\mathbf{x}) = 0$, two cases are possible. If $\mathbf{x}$ belongs to the complement of supp μ, then, clearly, for some positive constant C_5 and all $n \geq 1$

$$\mathbb{E}[f_n^2(\mathbf{x})] \;\leq\; \frac{C_5 k_n^2}{n^2} \;\leq\; C_5$$

If x belongs to suppμ, the proof is similar to the case $f(\mathbf{x}) > 0$. Just fix $\varepsilon \in (0,1)$ and find $\delta > 0$ such that

$$\sup_{0<r\leq\delta} \left| \frac{\int_{\mathcal{B}(\mathbf{x},r)} f(Y)dY}{\int_{\mathcal{B}(\mathbf{x},r)} dY} \right| \leq \varepsilon.$$

This shows the first part of the theorem. One proves, with similar arguments that there exists a positive constant C_6 such that, for all n large enough, $\mathbb{E}[f_n^3(\mathbf{x})] \leq C_6$. Consequently, for all n large enough, the sequence $\{f_n^2(\mathbf{x})\}$ is uniformly integrable and, since

$$f_n(\mathbf{x}) - f(\mathbf{x}) \to 0 \qquad \textit{in probability} \text{ (by Theorem 1.1)}$$

this implies

$$\mathbb{E}[f_n(\mathbf{x}) - f(\mathbf{x})]^2 \to 0 \quad \text{as} \quad n \to \infty$$

Proof of Theorem 1.3.

fix $\mathbf{x} \in \mathbb{R}^d$ and assume that f has derivatives of second order at $\mathbf{x}$, with $f(\mathbf{x}) > 0$, let $G(u) = \mathbb{P}(\| X - \mathbf{x} \| \leq u) = \int_{\mathcal{B}(\mathbf{x},u)} f(Y)dY$ be the univariate distribution function of $\| X - \mathbf{x} \|$. We may write by a Taylor-Young expansion of f around $\mathbf{x}$,

$$\begin{aligned} G(u) &= V_d f(\mathbf{x}) u^d + \left[\frac{\partial f(\mathbf{x})}{\partial \mathbf{x}} \right]^T \int_{\mathcal{B}(\mathbf{x},u)} (Y - \mathbf{x}) dY \\ &+ \frac{1}{2} \int_{\mathcal{B}(\mathbf{x},u)} (Y - \mathbf{x})^T \left[\frac{\partial^2 f(\mathbf{x})}{\partial \mathbf{x}^2} \right] (Y - \mathbf{x}) dY + o(u^{d+2}) \qquad as \quad u \to 0 \end{aligned} \tag{1.5}$$

where the symbol T denotes transposition and $[\partial f(\mathbf{x})/\partial \mathbf{x}]$ and $[\partial^2 f(\mathbf{x})/\partial \mathbf{x}^2]$ are a vector and a matrix given by

$$\left[\frac{\partial f(\mathbf{x})}{\partial \mathbf{x}} \right] = \left(\frac{\partial f(\mathbf{x})}{\partial \mathbf{x}_1}, \ldots, \frac{\partial f(\mathbf{x})}{\partial \mathbf{x}_d} \right)^T \quad \text{and} \quad \left[\frac{\partial^2 f(\mathbf{x})}{\partial \mathbf{x}^2} \right]_{i,j} = \frac{\partial^2 f(\mathbf{x})}{\partial x_i \partial x_j}.$$

In view of the symmetry of the ball $\mathcal{B}(\mathbf{x}, u)$, the first term in (1.5) is seen to be zero. using the linearity of trace and relations $tr(A\mathbf{Z}\mathbf{Z}^T) = \mathbf{Z}^T A \mathbf{Z}$, $tr(AB) = tr(BA)$ for matrices A, B an vector $\boldsymbol{Z}$, (1.5) becomes

$$G(u) = V_d f(\mathbf{x}) u^d + \frac{1}{2} tr\{[\int_{\mathcal{B}} (y - \mathbf{x})(y - \mathbf{x})^T dy][\partial^2 f(\mathbf{x})/\partial \mathbf{x}^2]\} + o(u^{d+2}).$$

letting $\mathbf{z} = (y - \mathbf{x})/u$ that maps $\mathcal{B}(\mathbf{x}, u)$ to $\mathcal{B}(0, 1)$, and using a hyperspherical coordinate change of variables (see,e.g,Miller[1964,chapter 1]), the integral inside the trace term simplifies to

$$\int_{\mathcal{B}(0,1)} u^2 \mathbf{z}\ \mathbf{z}^T u^d d\mathbf{z} = \left[\ \frac{V_d}{d+2} u^{d+2}\ \right] Id$$

where Id is the $d \times d$ identity matrix, thus, denoting by $\Gamma(.)$ the gamma function an recalling that, for the euclidean norm $V_d = \frac{\pi^{d/2}}{\Gamma(1+d/2)}$, we obtain

$$G_u = V_d f(\mathbf{x}) u^d + c(\mathbf{x}) V_d^{1+d/2} u^{d+2} + o(u^{d+2}) \quad \text{as} \quad u \to 0$$

where

$$c(\mathbf{x}) = \frac{1}{2(d+2)\pi} \Gamma^{2/d} \left(\frac{d+2}{2} \right) tr \left[\frac{\partial^2 f(\mathbf{x})}{\partial \mathbf{x}^2} \right]$$

Consequently

$$G^{\text{inv}}(u) = \frac{1}{V_d^{1/d} f^{1/d}(\mathbf{x})} u^{1/d} - \frac{c(\mathbf{x})}{d V_d^{1/d} f^{1+3/d}(\mathbf{x})} u^{3/d} + o(u^{3/d}) \quad \text{as} \quad u \to 0$$

and

$$[G^{\text{inv}}(u)]^d = \frac{1}{V_d f(\mathbf{x})} u - \frac{c(\mathbf{x})}{V_d f^{2+2/d}(\mathbf{x})} u^{1+2/d} + o(u^{1+2/d}) \quad \text{as} \quad u \to 0$$

Let F be the univariate distribution function of $w = \| X - \mathbf{x} \|^d$. Clearly $F^{\text{inv}}(u) = [G^{\text{inv}}(u)]^d$. Define $W_j = \| X_j - x \|^d$, $j = 1, .., n$ and let $W_{(1)} \leq ... \leq W_{(n)}$ be the order statistics for $W_1, .., W_n$. If $U_{(1)} \leq ... \leq U_{(n)}$ are uniform [0,1] order statistics, using the representation $W_{(j)} =^{\mathcal{D}} F^{\text{inv}}(U_{(j)})$ jointly for all j , we may write

$$f_n(\mathbf{x}) =^{\mathcal{D}} \frac{1}{n} \frac{\sum_{j=1}^{k_n} p_{nj} j}{f^{-1}(\mathbf{x}) \sum_{j=1}^{k_n} p_{nj} U_{(j)} + c'(\mathbf{x}) \sum_{j=1}^{k_n} p_{nj} o(U_{(j)}^{1+2/d})}$$

where

$$c'(\mathbf{x}) = -\frac{c(\mathbf{x})}{f^{2+2/d}(\mathbf{x})}$$

thus

$$f_n^{-1}(\mathbf{x}) =^{\mathcal{D}} n \left(\frac{f^{-1}(x) \sum_{j=1}^{k_n} p_{nj} U_{(j)}}{\sum_{j=1}^{k_n} p_{nj} j} + \frac{c'(\mathbf{x}) \sum_{j=1}^{k_n} p_{nj} U_{(j)}^{1+2/d}}{\sum_{j=1}^{k_n} p_{nj} j} + \frac{\sum_{j=1}^{k_n} p_{nj} o(U_{(j)}^{1+2/d})}{\sum_{j=1}^{k_n} p_{nj} j} \right)$$

Consequently, by lemma 1.1. letting $E_1, ..., E_{n+1}$ be *i.i.d* standard exponential random variables and

$$V_{(j)} = \frac{\sum_{i=1}^{j} E_i}{\sum_{i=1}^{n+1} E_i}.$$

we obtain

$$f_n^{-1}(x) =^{\mathcal{D}} \frac{f^{-1}(x)\sum_{j=1}^{k_n} p_{nj}(E_1 + ... + E_j)}{\sum_{j=1}^{k_n} p_{nj} j}(1+\zeta_{n1})$$

$$+\frac{c'(x)\sum_{j=1}^{k_n} p_{nj}(E_1 + ... + E_j)^{1+2/d}}{n^{2/d}\sum_{j=1}^{k_n} p_{nj} j}(1+\zeta_{n2}) + \frac{n\sum_{j=1}^{k_n} p_{nj} o(V_{(j)}^{1+2/d})}{\sum_{j=1}^{k_n} p_{nj} j}$$

Besides, for $j = 1, 2, \zeta_{nj} = O_{\mathbb{P}}(n^{-1/2})$ and, for all positive integers r,

$$\limsup_{n\to\infty}[n^{r/2}\mathbb{E}|\zeta_{nj}|^r] < \infty.$$

on the one hand, using the second statement of Lemma 1.2. and the identity

$$\frac{1}{k_n}\sum_{j=1}^{k_n} p_{nj} j = \int_{[0,1]} tr(dt)\left(\ 1 + O\left(\ \tfrac{1}{k_n}\ \right)\ \right)$$

as $k_n \to \infty$, we may write, on an appropriate probability space

$$\frac{f^{-1}\sum_{j=1}^{k_n} p_{nj}(E_1 + ... + E_j)}{\sum_{j=1}^{k_n} p_{nj} j} = f^{-1}(\mathbf{x}) + \frac{f^{-1}(\mathbf{x})v}{\sqrt{k_n}}N + \zeta_{n3},$$

where N is a standard normal random variable

$$v^2 = \frac{\int_0^1 (1-\phi(t))^2 dt}{[\int_0^1 tr(dt)]^2}$$

and $\zeta_{n3} = o_{\mathbb{P}}(k_n^{-1/2})$ with, for all positive integers r, $\sup_{n\geq 1}[k_n^{r/2}\mathbb{E}|\zeta_{n3}|^r] < \infty$. Next, recalling that for $\rho \geq 1$;

$$\frac{1}{k_n^{\rho}}\sum_{j=1}^{k_n} p_{nj} j^{\rho} = \int_{[0,1]} t\nu(dt)\left(\ 1 + O\left(\ \frac{1}{k_n}\ \right)\ \right)$$

and applying the first statement of Lemma 1.2, we obtain

$$\frac{c'(\mathbf{x})\sum_{j=1}^{k_n} p_{nj}(E_1 + ... + E_j)^{1+2/d}}{n^{2/d}\sum_{j=1}^{k_n} p_{nj} j} = c'(\mathbf{x})b\left(\ \frac{k_n}{n}\ \right)^{2/d} + \left(\ \frac{k_n}{n}\ \right)^{2/d}\zeta_{n4}$$

where

$$b = \frac{\int_0^1 t^{1+2/d}\nu(dt)}{\int_0^1 t\nu(dt)}$$

and $\zeta_{n4} = o_{\mathbb{P}}(1)$ with, for all positive integers ν , $\sup_{n\geq 1}\mathbb{E}|\zeta_{n4}|^r < \infty$
Similarly

$$\left| \frac{n\sum_{j=1}^{k_n} p_{nj} o(V_{(j)}^{1+2/d})}{\sum_{j=1}^{k_n} p_{nj} j} \right| \leq \frac{(E_1 + ... + E_{k_n})^{1+2/d}}{n^{2/d}\sum_{j=1}^{k_n} p_{nj} j} \times \frac{o(V_{(k_n)}^{1+2/d})}{V_{(k_n)}^{1+2/d}} \times (1 + \zeta_{n5})$$

where $\zeta_{n5} = O_{\mathbb{P}}(n^{-1/2})$ and for all positive integers r

$$\limsup_{n\to\infty}[n^{r/2}\mathbb{E}|\zeta_{n5}|^r] < \infty.$$

thus

$$\left| \frac{n\sum_{j=1}^{k_n} p_{nj} o(V_{(j)}^{1+2/d})}{\sum_{j=1}^{k_n} p_{nj} j} \right| \leq \left(\frac{k_n}{n} \right)^{2/d} \zeta_{n6}$$

where $\zeta_{n6} = o_{\mathbb{P}}(1)$. Moreover, we clearly have, for some $\varepsilon_0 \in (0, 1)$ and all $r > 0$

$$\limsup_{n\to\infty} \mathbb{E}[|\zeta_{n6}|^r \mathbb{1}_{[V_{(k_n)}\leq\varepsilon_0]}] < \infty,$$

Thus, putting all the pieces together, we obtain

$$f_n^{-1}(\mathbf{x}) =^{\mathcal{D}} f^{-1}(\mathbf{x}) + \frac{f^{-1}(\mathbf{x})v}{\sqrt{k_n}} N + c'(\mathbf{x})b\left(\frac{k_n}{n} \right)^{2/d} + \zeta_{n7} + \left(\frac{k_n}{n} \right)^{2/d} \zeta_{n8}$$

where

$$\zeta_{n7} = o_{\mathbb{P}}(k_n^{-1/2})$$

and

$$\zeta_{n8} = o_{\mathbb{P}}(1).$$

Besides, for all positive integers r,

$$\limsup_{n\to\infty}[k_n^{r/2}\mathbb{E}|\zeta_{n7}|] < \infty$$

and

$$\limsup_{n\to\infty} \mathbb{E}[|\zeta_{n8}|^r \mathbb{1}_{[V_{(k_n)}\leq\varepsilon_0]}] < \infty$$

we see in particular that, for all positive integers r and all n large enough, the sequence $\{k_n^{r/2}\zeta_{n7}^r\}$ is uniformly integrable and, consequently, that

$$\mathbb{E}|\zeta_{n7}|^r = o(k_n^{-r/2})$$

(see, e.g,Billing sley[[07],chapter 5]). Likewise

$$\mathbb{E}[|\zeta_{n8}|^r \mathbb{1}_{[V_{(k_n)} \leq \varepsilon_0]}] = o(1)$$

It follows that

$$f_n^{-1}(\mathbf{x}) =^{\mathcal{D}} f^{-1}(\mathbf{x}) + \frac{f^{-1}(\mathbf{x})v}{\sqrt{k_n}}N + c'(\mathbf{x})b\left(\frac{k_n}{n}\right)^{2/d} + \zeta_{n9}$$

where $\zeta_{n9} = o_{\mathbb{P}}(k_n^{-1/2} + (k_n/n)^{2/d})$ and

$$\mathbb{E}[|\zeta_{n9}|^r \mathbb{1}_{[V_{(k_n)} \leq \varepsilon_0]}] = o\left(\frac{1}{k_n^{r/2}} + \left(\frac{k_n}{n}\right)^{2r/d}\right) \tag{1.6}$$

as $k_n \to \infty$ and $k_n/n \to 0$ note that, by definition $f_n^{-1}(\mathbf{x})$ is almost surely finite and positive. Therefore,setting

$$T_n(\mathbf{x}) = \frac{v}{\sqrt{k_n}}N + f(\mathbf{x})c'(\mathbf{x})b\left(\frac{k_n}{n}\right)^{2/d} + f(\mathbf{x})\zeta_{n9}$$

and using the identity $\frac{1}{1+t} = 1 - t + \frac{t^2}{1+t}$ valid for $t \neq -1$, we finally get

$$f_n(\mathbf{x}) =^{\mathcal{D}} f(\mathbf{x}) - \frac{f(\mathbf{x})v}{\sqrt{k_n}}N + \frac{c(\mathbf{x})b}{f^{2/d}(\mathbf{x})}\left(\frac{k_n}{n}\right)^{2/d} + \zeta_{n10} + \frac{f(\mathbf{x})T_n^2(\mathbf{x})}{1 + T_n(\mathbf{x})}$$

where $\zeta_{n10} = o_{\mathbb{P}}(k_n^{-1/2} + (k_n/n)^{2/d})$ and

$$\mathbb{E}[\zeta_{n10}^2 \mathbb{1}_{[V_{(k_n)} \leq \varepsilon_0]}] = o\left(\frac{1}{k_n} + \left(\frac{k_n}{n}\right)^{4/d}\right)$$

Clearly

$$\frac{T_n^2(\mathbf{x})}{1 + T_n(\mathbf{x})} = o_{\mathbb{P}}\left(\frac{1}{\sqrt{k_n}} + \left(\frac{k_n}{n}\right)^{2/d}\right)$$

Next, observing that

$$\mathbb{E}\left[\frac{1}{1+T_n(\mathbf{x})}\right]^4 = f^{-4}(\mathbf{x})\mathbb{E}[f_n^4(\mathbf{x})]$$

it follows from an immediate adaptation of the proof of Theorem 1.2. that

$$\limsup_{n\to\infty} \mathbb{E}\left[\frac{1}{1+T_n(\mathbf{x})} \right]^4 < \infty$$

Thus, using the Cauchy-Schwarz inequality and (1.6), we see that

$$\mathbb{E}\left[\left(\frac{T_n^2(\mathbf{x})}{1+T_n(\mathbf{x})} \right)^2 \mathbb{1}_{[V_{(k_n)}\leq \varepsilon_0]} \right] = o\left(\frac{1}{k_n} + \left(\frac{k_n}{n} \right)^{4/d} \right)$$

In conclusion:

$$f_n(\mathbf{x}) =^{\mathcal{D}} f(\mathbf{x}) - \frac{f(\mathbf{x})v}{\sqrt{k_n}} N + \frac{c(\mathbf{x})b}{f^{2/d}(\mathbf{x})} \left(\frac{k_n}{n} \right)^{2/d} + \zeta_n$$

where $\zeta_n = o_{\mathbb{P}}(k_n^{-1/2} + (k_n/n)^{2/d})$, as desired. In addition

$$\mathbb{E}[\zeta_n^2 \mathbb{1}_{[V_{(k_n)}\leq \varepsilon_0]}] = o\left(\frac{1}{k_n} + \left(\frac{k_n}{n} \right)^{4/d} \right)$$

1.3.3 *Some asymptotic results for regression estimator*

The data in our model can be rewritten as

$$Y_i \quad = \quad r(\mathbf{X}_i) + \varepsilon_i, \qquad\qquad 1 \leq i \leq n$$

where $\varepsilon_i = Y_i - r(\mathbf{X}_i)$ satisfies $\mathbb{E}[\varepsilon_i/\mathbf{X}_i] = 0$.
The nearest neighbor estimate is

$$r_n(\mathbf{x}) = \sum_{i=1}^{n} w_{ni} Y_{(i)}(\mathbf{x})$$

where $(w_{ni}, ..., w_{nn})$ is a given (deterministic) weight vector summing to one.

In this section , we study the local rate of convergence of $r_n(\mathbf{x})$ and to simplify the notation, we will study only the weak convergence properties of $r_n(0) - r(0)$. We let the conditional variance of Y be:

$$\sigma^2(\mathbf{x}) = \mathbb{E}[|Y - r(\mathbf{X})|^2/\mathbf{X} = \mathbf{x}]$$

and assume the following:

i) There exists a sequence of positive integers $\{k\} = \{k_n\}$ with $k \to \infty, \quad k/n \to 0$, and a positive constant c such that

$$|w_{ni}| \leq \begin{cases} \frac{c}{k}, & \text{for } 1 \leq i \leq k \\ 0, & \text{otherwise} \end{cases}$$

and $\sum_{i=1}^{n} w_{ni} = 1$

ii) The random variable $\mathbf{X}$ has a density f on $\mathbb{R}^d$ that is twice continuously differentiable in a neighborhood of 0. Also $f(0) > 0$

iii) The regression function r is twice continuously differentiable in a neighborhood of 0.

iv) One has $\| Y \|_\infty \leq 1$. This condition can be weakened to either $\| Y - r(\mathbf{X}) \|_\infty \leq 1$ or even $:\sup_{x \in \mathbb{R}^d} \mathbb{E}[|Y - r(\mathbf{X})|^3 / \mathbf{X} = \mathbf{x}] < \infty$.

v) The function σ is continuous in a neighborhood of 0 and $\sigma^2(0) > 0$.

Theorem 1.4

Assume that conditions (i), (iv) and (v) are satisfied. Then:

$$\frac{V_n}{\sigma(0)\sqrt{\sum_{i=1}^{n} w_{ni}^2}} \to N$$

Where N is a standard normal random variable.

Theorem 1.5 (Pointwise rate of convergence)

Assume that conditions (i),(v) are satisfied. Then the corresponding nearest neighbor regression function estimate r_n satisfies:

$$r_n(0) - r(0) =^{\mathcal{D}} \sigma(0) \sqrt{\sum_{i=1}^{n} w_{ni}^2} (N + o_{\mathbb{P}}(1))$$

$$+\beta \left(\frac{k}{n} \right)^{2/d} \left(\sum_{i=1}^{k} w_{ni} \left(\frac{i}{k} \right)^{2/d} \right) (1 + o_{\mathbb{P}}(1)) + o_{\mathbb{P}} \left(\left(\frac{k}{n} \right)^{2/d} \right)$$

Where N *is a standard normal random variable and*

$$\beta =^{def} \frac{f(0)tr(r''(0)) + 2r'(0)^T f'(0)}{2dV_d^{2/d} f^{1+2/d}(0)}$$

for the standard k nearest neighbor estimate, one has:

$$w_{ni} = \begin{cases} \frac{1}{k}, & \text{for } 1 \leq i \leq k \\ \\ 0, & \text{for } k < i \leq n \end{cases}$$

where $\{k\} = \{k_n\}$ is a sequence of integers such that $1 \leq k \leq n$. In this case, $\sum_{i=1}^n w_{ni}^2 = \frac{1}{k}$ and $\sum_{i=1}^n w_{ni} \left(\frac{i}{k} \right)^{2/d} = \frac{d}{d+2}(1 + o(1))$

Theorem 1.6(L^2 rates of convergence)

Let $r_n(\mathbf{x}) = \sum_{i=1}^n w_{ni} Y_{(i)}(\mathbf{x})$ *be the nearest neighbor regression function estimate, where* $(w_{n1}, ..., w_{nn})$ *is a probability weight vector. Assume that* $\mathbf{X}$ *takes values in* $[0, 1]^d$. *Assume,in addition, that for all* $\mathbf{x}$ *and* $\mathbf{x}' \in \mathbb{R}^d$,

$$|r(\mathbf{x}) - r(\mathbf{x}')| \leq L \parallel \mathbf{x} - \mathbf{x}' \parallel$$

and

$$\sup_{x \in \mathbb{R}^d} \sigma^2(\mathbf{x}) \leq \sigma^2$$

for some positive constants L *and* σ^2. *Then*

(i) *For* $d = 1$

$$\mathbb{E}|r_n(\mathbf{X}) - r(\mathbf{X})|^2 \leq \sigma^2 \sum_{i=1}^n w_{ni}^2 + 8L^2 \sum_{i=1}^n w_{ni} \frac{i}{n}$$

(ii) *For* $d \geq 2$

$$\mathbb{E}|r_n(\mathbf{X}) - r(\mathbf{X})|^2 \leq \sigma^2 \sum_{i=1}^n w_{ni}^2 + c_d' L^2 \sum_{i=1}^n w_{ni} \left(\frac{i}{n} \right)^{2/d}$$

where

$$c_d' = \frac{2^{3+\frac{2}{d}}(1 + \sqrt{d})^2}{V_d^{2/d}}$$

for the standard k nearest neighbor estimate, we have the following corollary:

Corollary 1.1

Let r_n be the k nearest neighbor regression function estimate, then under the condition of theorem 1.6

(i) *for $d = 1$, there exists a sequence $\{k\} = \{k_n\}$, for some positive universal constant Λ_1*

(ii) *for $d \leq 2$, there exists a sequence $\{k\} = \{k_n\}$ with $k \sim \left(\frac{\sigma^2}{L^2} \right)^{\frac{d}{d+2}} n^{\frac{2}{d+2}}$ such that*

$$\mathbb{E}|r_n(\mathbf{X}) - r(\mathbf{X})|^2 \leq \Lambda_d \left(\frac{\sigma^2 L^2}{n} \right)^{\frac{2}{d+2}}$$

for some positive universal constant Λ_d.

1.3.4 *Proof of Theorems*

Proof of Theorem 1.4

It is useful to recall the Berry-Essen inequality
For sums of *i.r.v* $W_1, ..., W_n$ such that $\mathbb{E}W_i = 0$, $\sum_{i=1}^n \mathbb{E}W_i^2 > 0$, and $\mathbb{E}|W_i|^3 < \infty$:

$$\sup_{t \in \mathbb{R}^d} \left| \mathbb{P} \left\{ \frac{\sum_{i=1}^n W_i}{\sqrt{\sum_{i=1}^n \mathbb{E}W_i^2}} \leq t \right\} - \mathbb{P}\{N \leq t\} \right| \leq \frac{\gamma \sum_{i=1}^n \mathbb{E}|W_i|^3}{(\sum_{i=1}^n \mathbb{E}W_i^2)^{3/2}} \quad (1.7)$$

for some universal constant $\gamma > 0$.
We apply this inequality with the formal replacement $W_i = w_{ni}(Y_{(i)} - m(Z_{(i)})$, conditional on $Z_1, ..., Z_n$. Since, conditional on $Z_1, ..., Z_n$;

$$\mathbb{E}W_i^2 = w_{ni}^2 \tau^2(Z_{(i)}) \quad \text{and} \quad \mathbb{E}|W_i|^3 \leq \frac{8c}{k} w_{ni}^2$$

The bound in (1.7) becomes

$$\begin{aligned} \frac{8c\gamma \sum_{i=1}^n w_{ni}^2}{k(\sum_{i=1}^n w_{ni}^2 \tau^2(Z_{(i)}))^{3/2}} &\leq \frac{8c\gamma}{k\sqrt{\sum_{i=1}^n w_{ni}^2} \times \min^{3/2}(\tau^2(Z_{(1)}), ..., \tau^2(Z_{(k)}))} \\ &\leq \frac{8c\gamma}{k^{1/2} \min^{3/2}(\tau^2(Z_{(1)}), ..., \tau^2(Z_{(k)}))} \end{aligned}$$

(since $\sum_{i=1}^{n} w_{ni}^2 \geq \frac{1}{k}$; by the Cauchy-Schwarz inequality), observe that:

$$\frac{\sum_{i=1}^{n} w_{ni}(Y_{(i)} - m(Z_{(i)}))}{\tau(0)\sqrt{\sum_{i=1}^{n} w_{ni}^2}} = \frac{\sum_{i=1}^{n} w_{ni}(Y_{(i)} - m(Z_{(i)}))}{\sqrt{\sum_{i=1}^{n} w_{ni}^2 \tau^2(Z_{(i)})}} \times \frac{\sqrt{\sum_{i=1}^{n} w_{ni}^2 \tau^2(Z_{(i)})}}{\tau(0)\sqrt{\sum_{i=1}^{n} w_{ni}^2}}$$

$$=^{def} \mathbf{I} \times \mathbf{II}$$

Now: $\mathbf{II} \to 1$ in probability as noted earlier.
For $\mathbf{I}$, we have:

$$\sup_{t \in \mathbb{R}} |\mathbb{P}\{\mathbf{I} \leq t / Z_1, ..., Z_n\} - \mathbb{P}\{N \leq t\}| = \frac{O(1/\sqrt{k})}{\min^{3/2}(\tau^2(Z_{(1)}), ..., \tau^2(Z_{(k)})}$$

Hence

$$\sup_{t \in \mathbb{R}} \Big| \ \mathbb{P}\{\mathbf{I} \leq t\} - \mathbb{P}\{N \leq t\} \ \Big| = \sup_{t \in \mathbb{R}} |\mathbb{E}\mathbb{P}\{\mathbf{I} \leq t / Z_1, ..., Z_n\} - \mathbb{P}\{N \leq t\}|$$

$$\leq \frac{O(1/\sqrt{k})}{\tau^3(0)} + \mathbb{P}\{\min(\tau^2(Z_{(1)}, .., Z_{(k)})) < \frac{\tau^2(0)}{2}\}$$

The latter probability tends to zero since $\tau(0) > 0, \tau$ is continuous at 0, and $Z_{(k)} \to 0$ in probability. Thus, $\mathbf{I} \to^{\mathcal{D}} N$; so that $\mathbf{I} \times \mathbf{II} \to^{\mathcal{D}} N$.

Proof of Theorem 1.5

We will applies the following result

Proposition 1.1

Assume that f and r are twice continuously differentiable in a neighborhood of 0,and $f(0) > 0$. Then, as $z \downarrow 0$

$$m(z) = r(0) + \alpha z^2 + o(z^2)$$

where

$$\alpha = \frac{f(0) tr(r''(0)) + 2r'(0)^T f'(0)}{2df(0)}$$

Lemma 1.3

For $\mathbf{x} \in \mathbb{R}^d$; set $\rho_{\mathbf{x}} = \inf\{\| Y - \mathbf{x} \|: Y \in supp(\mu)\}$. If $k/n \to 0$, then

$$\| \mathbf{X}_{(k)}(\mathbf{x}) - \mathbf{x} \| \to \rho_{\mathbf{x}} \quad \text{almost surely}$$

In particular, if $\mathbf{x} \in supp(\mu)$ and $k/n \longrightarrow 0$; then

$$\| \mathbf{X}_{(k)}(\mathbf{x}) - \mathbf{x} \| \longrightarrow 0 \quad \text{almost surely.}$$

Lemma 1.4

If $k \to \infty$, then

$$\frac{U_{(k)}}{k/n} \quad \to \quad 1 \qquad \textit{in probability.}$$

Proposition 1.2

Assume that condition (i) is satisfied. Then

$$W_n = \left(\tfrac{k}{n} \right)^{2/d} \left(\sum_{i=1}^{k} w_{ni} \left(\tfrac{i}{k} \right)^{2/d} \right) (1 + o_{\mathbb{P}}(1)) + o_{\mathbb{P}} \left(\left(\tfrac{k}{n} \right)^{2/d} \right)$$

Proof.

By proposition 1.1, where α is defined, we have

$$\begin{aligned} B_n &= \sum_{i=1}^{n} w_{ni}(m(Z_{(i)}) - m(0)) \\ &= \alpha \sum_{i=1}^{n} w_{ni} Z_{(i)}^2 + \sum_{i=1}^{n} w_{ni} \varphi(Z_{(i)}) \\ &= \mathbf{I} + \mathbf{II} \end{aligned}$$

where $\varphi(z) = O(z^2)$ as $z \downarrow (0)$. Clearly

$$\begin{aligned} |\mathbf{II}| &\leq \sum_{i=1}^{n} |w_{ni}| \sup_{0<z\leq Z_{(k)}} |\varphi(z)| \\ &\leq \sum_{i=1}^{n} |w_{ni}| Z_{(k)}^2 \sup_{0<z\leq Z_{(k)}} \left| \tfrac{\varphi(z)}{z^2} \right| \\ &\leq cZ_{(k)}^2 \sup_{0<z\leq Z_{(k)}} \left| \tfrac{\varphi(z)}{z^2} \right| \\ &= o_{\mathbb{P}}(Z_{(k)}^2) \end{aligned}$$

since $Z_{(k)} \to 0$ in probability (by Lemma 1.3 and the fact that 0 belongs to the support of $\mathbf{X}$ (see condition (ii))).

Next, recall the decomposition

$$Z_{(i)} =^{\mathcal{D}} \left(\tfrac{U_{(i)}}{V_d f(0)} \right)^{1/d} + \psi(U_{(i)})$$

where $\psi(u) = O(u^{1/d})$ as $u \downarrow 0$ and

⋆ $(U_{(1)}, ..., U_{(n)}) =^{\mathcal{D}} \left(\tfrac{G_1}{G_{n+1}}, ..., \tfrac{G_n}{G_{n+1}} \right)$ where

$$G_i = \sum_{j=1}^{i} E_j; 1 \leq i \leq n+1$$

and $E_1, ..., E_{n+1}$ are independent standard exponential random variables.

⋆ $(Z_{(1)}, ..., Z_{(n)}) =^{\mathcal{D}} (G^{-1}(U_{(1)}), ..., G^{-1}(U_{(n)}))$ and

$$G^{-1}(u) = \inf\{t \geq 0, G(t) \geq u\}, u \in [0, 1]$$

⋆ Since: $G^{-1}(u) = \left(\tfrac{u'}{V_d f(0)} \right)^{1/d} + \psi(u)$; it will be convenient to replace $Z_{(i)}$ by

$$Z_{(i)} =^{\mathcal{D}} \left(\tfrac{U_{(j)}}{V_d f(0)} \right)^{1/d} + \psi(U_{(i)})$$

Thus

$$\mathbf{I} = \beta \sum_{i=1}^{n} w_{ni} U_{(i)}^{2/d} + 2\alpha \sum_{i=1}^{n} w_{ni} \left(\tfrac{U_{(i)}}{V_d f(0)} \right)^{1/d} \psi(U_{(i)}) + \sum_{i=1}^{n} w_{ni} \psi^2(U_{(i)})$$

Where $\beta = \frac{\alpha}{V_d^{2/d} f^{2/d}(0)}$. Using the fact that $U_{(k)} \longrightarrow 0$ in probability and $|w_{ni}| \leq c/k$ for $1 \leq i \leq k$, it is easy to see that

$$\mathbf{I} = \beta \sum_{i=1}^{n} w_{ni} U_{(i)}^{2/d} + O_{\mathbb{P}}(U_{(k)}^{2/d}) = \beta \sum_{i=1}^{n} w_{ni} U_{(i)}^{2/d} + O_{\mathbb{P}}\left(\left(\frac{k}{n} \right)^{2/d} \right)$$

By the well-known fact(Lemma 1.4) that $U_{(k)} = O_{\mathbb{P}}(k/n)$
Combining this result with proposition 1.2 proves the theorem.

Proof of Theorem 1.6

Lemma 1.5

Let $\mathbf{X}$ *takes values in* $[0,1]^d$. *Then,for* $d \geq 2$;

$$\mathbb{E} \parallel \mathbf{X}_{(k)}(\mathbf{X}) - \mathbf{X} \parallel^2 \leq c'_d \left(\frac{k}{n} \right)^{2/d}$$

where

$$c'_d = \frac{2^{3+\frac{2}{d}}(1+\sqrt{d})^2}{V_d^{2/d}}$$

for $d = 1$ *we have*

$$E \parallel \mathbf{X}_{(k)}(\mathbf{X}) - \mathbf{X} \parallel^2 \leq \frac{8k}{n}$$

The proof of this theorem relies on lemma 1.5, which bounds the expected square distance between $\mathbf{X}$ and its *i-th* nearest neighbor. Letting $\tilde{r}_n(\mathbf{x}) = \sum_{i=1}^{n} w_{ni} r(\mathbf{X}_{(i)}(\mathbf{x}))$, we start with the variance/bias decomposition

$$\mathbb{E}|r_n(\mathbf{X}) - r(\mathbf{X})|^2 = \mathbb{E}|r_n(\mathbf{X}) - \tilde{r}_n(\mathbf{X})|^2 + \mathbb{E}|\tilde{r}_n(\mathbf{X}) - r(\mathbf{X})|^2$$

to bound the first term, note that

$$\mathbb{E}\left| \; r_n(\mathbf{X}) - \tilde{r}_n(\mathbf{X}) \; \right|^2 = \mathbb{E}\left| \; \sum_{i=1}^{n} W_{ni}(\mathbf{X})(Y_i - r(\mathbf{X}_i)) \; \right|^2$$

where $W_{ni}(\mathbf{X}) = w_{n \sum_i}$ and $(\sum_1, ..., \sum_n)$ is a permutation of (1,...,n) such that $\mathbf{X}_i$ is the $\sum_i - th$ nearest neighbor estimate

$$\begin{aligned} \mathbb{E}\left| \; \sum_{i=1}^{n} W_{ni}(\mathbf{X})(Y_i - r(\mathbf{X}_i)) \; \right|^2 &= \mathbb{E}\left[\; \sum_{i=1}^{n} W_{ni}^2(\mathbf{X}) |Y_i - r(\mathbf{X}_i)|^2 \; \right] \\ &= \mathbb{E}\left[\; \sum_{i=1}^{n} W_{ni}(\mathbf{X}) \sigma^2(\mathbf{X}_i) \; \right] \end{aligned}$$

So that

$$\mathbb{E}|r_n(\mathbf{X}) - \tilde{r}_n(\mathbf{X})|^2 \leq \sigma^2 \sum_{i=1}^{n} w_{ni}^2$$

Finally

$$\begin{aligned}
\mathbb{E}|\tilde{r}_n(\mathbf{X}) - r(\mathbf{X})|^2 &= \mathbb{E}\left| \textstyle\sum_{i=1}^n w_{ni}(r(\mathbf{X}_{(i)}(\mathbf{X}) - r(\mathbf{X})) \right|^2 \\
&\leq \mathbb{E}\left[\left(\textstyle\sum_{i=1}^n w_{ni}|r(\mathbf{X}_i(\mathbf{X})) - r(\mathbf{X})| \right)^2 \right] \\
&\leq L^2\mathbb{E}\left[\left(\textstyle\sum_{i=1}^n w_{ni} \parallel \mathbf{X}_{(i)}(\mathbf{X}) - \mathbf{X} \parallel \right)^2 \right] \\
&\leq L^2 \left(\textstyle\sum_{i=1}^n w_{ni} E \parallel \mathbf{X}_{(i)}(\mathbf{X}) - \mathbf{X} \parallel^2 \right)
\end{aligned}$$

(by Jensen's inequality). The conclusion follows by applying lemma 1.5.

1.4 *Cross-validation with k nearest neighbors estimation*

This section present for the kNN locally constant estimator three k parameter selection methods from $\hat{g}(x)$. We always considered the nonparametric regression:

$$Y_i = g(X_i) + u_i$$

with $\mathbb{E}(u/X) = 0$, $var(u/X) = \sigma^2(X)$

The three methods below to select the value of k. These methods have been studied by Li(1987).

1. C_L or C_p of Mallows (Mallows (1973)):

This method consists of selecting the $\hat{k}$ that minimizes the objective function below

$$C_L = n^{-1} \sum_{i=1}^{n} (y_i - \hat{g}(x_i))^2 + 2\sigma^2 tr(M_n(k))/n$$

where σ^2 is the variance of u_i. In practice σ^2 is estimated by

$$\hat{\sigma} = n^{-1} \sum_{i=1}^{n} \hat{u}_i^2; \quad u_i = y_i - \hat{g}(x).$$

2. Generalised Cross-Validation by Craven & Wahba (1979):

This method consists to select the $\hat{k}$ that minimizes the objective function below

$$GCV_k \frac{\sum_{i=1}^{n}(y_i - \hat{g}(x_i))^2}{(1 - n^{-1}tr(M_n(k)))^2}$$

3. Cross-Validation ("Leave one out" (Stone 1974)):

This method is to select the $\hat{k}$ that minimizes the objective function below

$$CV_k = \sum_{i=1}^{n}(y_i - \hat{g}^{(-i)}(x_i))^2$$

where

$$\hat{g}^{(-i)}(x_i) = \sum_{j \neq i} y_i W_{ij} / \sum_{j \neq i} W_{ij} \left(\; W_{ij} = w\left(\; \tfrac{x_i - x_j}{R_i} \; \right) \; \right)$$

is the "Leave one out" kNN estimator of $g(x_i)$.
The methods C_L and GCV_k are less costly in terms of computing time unlike the CV_k method.
Li (1987) showed that the three approaches are asymptotically equivalent and provide an optimal smoothing in the sense

$$\frac{\int[\hat{g}_{\hat{k}}(x) - g(x)]^2 dF(x)}{\int[\hat{g}(x) - g(x)]^2 dF(x)} \longrightarrow^{p} 1$$

where $\hat{g}_{\hat{k}}(x)$ is the kNN estimator using one of the above approaches to select k.
Li and Ouyang (2004) showed that for all values of $k \in \Lambda = [n^{\epsilon}, n^{1-\epsilon}]$, $\epsilon \in (0, 1/2)$. CV_k can be put in the form below

$$CV_{kc} = \phi_1 (k/n)^{4/q} + \phi_2 k^{-1} + o((k/n)^{4/q} + k^{-1})$$

where:

$$\phi_1 = c_0^{-4/q} k_2^2 \int ([\frac{1}{2} f(x) tr])$$

$$\phi_2 = c_0 k \int \sigma^2(x) M(x) f(x) dx.$$

with:

$$c_0 = \pi^{q/2} / \Gamma((q+2)/2).$$

In the second part, we redo the same analysis but considering the local linear kNN estimator. Recall that $\delta(x) = (g(x), \nabla g(x)').\delta(x)$ is a vector $(q+1) \times 1$ where the first element is $g(x)$ and the other elements are the partial derivatives of $g(x)$. the optimal number of neighbors is selected by minimizing the objective function of Cross-validation below

$$CV_{kL} = n^{-1} \sum_{i=1}^{n} (y_i - \hat{g}^{(-i,L)}(x_i))^2 M(x_i)$$

with $M(.)$ a weight function.
Li and Ouyang(2004) showed that, for all $k \in \nabla$

$$CV_{kL}(k) = \phi_{1,L}(\frac{k}{n})^{4/q} + \phi_2 k^{-1} + o_{\mathbb{P}}((\frac{k}{n})^{4/q} + k^{-1})$$

where ϕ_2 is defined in the same manner as previously and

$$\phi_{1,L} = c_0^{-4/q} k_2^2 \int (\frac{1}{2} f(x) tr([\nabla^2 g(x)]))^2 \frac{M(x)}{f(x)^{(q+4)/4}} dx$$

1.5 *Automatic selection of k the number of nearest*

for choosing the tuning parameter k it remains to introduce a loss function Loss. Among the kNN estimators, we retain the loss function that allowing us to build a local version of our kNN estimator.

Loss function

Loess was introduced by Cleveland(1988), and is a multivariate version of Lowess Cleveland (1979), which is another version of LPR. Loess is described by

$$\hat{f}(x) \quad = \quad \sum_{i=1}^{n} a_i(x) Y_i,$$

where $a(x) = I_1^T \hat{\beta}_x$ and $I_1^T = (1, 0, ..., 0)$, where the polynomial degree is one $(d = 1)$ or two $(d = 2)$. For the bandwidth selection and weight calculation, loess is similar to kNN . Its weights are calculated with: $K_b(u) = \frac{1}{b} K(\frac{D(u)}{b})$, where $u = (x_i - x)$, and D(.) is u's L_2- norm in the predictor space and b is the euclidean distance between the input vector x and its k^{th} nearest neighbor. The weight function chosen

by Cleveland and Delvin(1988) was the Tricube kernel, however it is not mandatory. The main goal is to compute the quantity

$$p_g^{LCV}(x) = \frac{\sum_{\{i:y_i=g\}} K(d(x_i,x)/h_{LCV}(x_{i_0}))}{\sum_{i=1}^n K(d(x_i,x)/h_{LCV}(x_{i_0}))}$$

where

$$i_0 = \arg\min_{i=1,\ldots,n} d(x,x_i) \quad \text{and} \quad h_{LCV}(x_{i_0})$$

is the bandwidth corresponding to the optimal number of neighbors at x_{i_0} obtained by the following Cross-Validation procedure

$$k_{LCV}(x_{i_0}) = \arg\min_k LCV(k,i_0),$$

where

$$LCV(k,i_0) = \sum_{g=1}^{G} (1_{[y_{i_0}=g]} - p_{g,k}^{(-i_0)}(x_{i_0}))^2;$$

and

$$p_{g,k}^{(-i_0)}(x_{i_0}) = \frac{\sum_{\{i:y_i=g,i\neq i_0\}}^n K(d(x_i,x_{i_0})/h_{k(x_{i_0})})}{\sum_{i=1,i\neq i_0}^n K(d(x_i,x_{i_0})/h_{k(x_{i_0})})}$$

The main feature of such an estimator concerns the local behavior of the bandwidth. More precisely, the optimal number of neighbors depends on the functional point at which the kNN estimator is evaluated. This is the reason why we use the term local selection. Note that many other loss functions can be built as in the prediction setting. Now, the estimation procedure is entirely determinated as soon as a semi-matric d(.,.) and a kernel function K(.) are fixed.
In order to give an idea of the performance of the procedure, we include the computation of the misclassification rate for the learning sample $(x_i,y_i)_{i=1,..,n}$ (i.e: the sample of curves for which the class numbers are observed): for

$$i \in \{1,2,..,n\} : y_i^{LCV} \longleftarrow \arg\max_{g\in\{1,..,G\}} p_g^{LCV}(x_i)$$

end do

$$Misclas \longleftarrow \frac{1}{n}\sum_{i=1}^n \mathbb{1}_{[y_i\neq y_i^{LCV}]}.$$

Chapter 2

The k nearest neighbor method for functional data

2.1 *Introduction*

In contrast to the first chapter, here our aim is to studies the kNN method in the functional case with the independent identically distributed data. A regression analysis is a statistical technique for estimating the value of a variable as a function of independent variables. They are widely applied in science and engineering, they are used in problems like function estimation, financial forecasting, and time series prediction.

In many practical situations, one is faced with functional type phenomena. It is now possible to take into account their functional nature thanks to technological improvements permitted to collect data discretized on thinner grids. The statistical problems involved in the modelization of functional random variables have received an increasing interest in recent literature, we only refer to the good overviews in parametric models given by Bosq(2000) [08], Ramsay& Silverman [36].

The literature of the kNN method for estimation of regression function date bakes to

Royall (1966) [14] & Stone (1977) [39] for the multivariate case. For the functional data studies, the kNN kernel estimate was first introduced in the monograph of Ferraty & Vieu (2006) [15], Burba et al.(2009) [09] obtained the rate of almost complete convergence of the regression function using the kNN method for independent data.

2.2 *Models and estimators*

Let $(\mathcal{X}_i, Y_i)_{i=1,..,n}$ be n pairs independent and identically distributed as $(\mathcal{X}, Y)$ and valued in $E \times \mathbb{R}$. (E,d) is a semi-metric space, E is not necessarily of finite dimension and we do not suppose the existence of a density for the functional random variable *(f.r.v)* X. The general frame is the functional nonparametric regression:

$$Y = r(\mathcal{X}) + \varepsilon \quad \text{with} \quad \mathbb{E}(\varepsilon/\mathcal{X}) = 0.$$

then, the object we want to estimate is the non-linear operator $r(.) = \mathbb{E}[Y/\mathcal{X} = .]$

for a fixed $\chi \in E$, the k NN kernel estimator can be written as

$$\hat{r}_{kNN}(\chi) = \sum_{i=1}^{n} Y_i w_{i,n}(\chi)$$

where

$$w_{i,n}(\chi) = \frac{K(H_{n,k}(\chi)^{-1} d(\chi, \mathcal{X}_i))}{\sum_{i=1}^{n} K(H_{n,k}(\chi)^{-1} d(\chi, \mathcal{X}_i))}$$

where K is an asymmetrical kernel and $H_{n,k}(\chi)$ is defined as follows

$$H_{n,k}(\chi) \;=\; \min \left\{ \; h \in \mathbb{R}^+ / \textstyle\sum_{i=1}^{n} \mathbb{1}_{\mathcal{B}(\chi,h)}(\mathcal{X}_i) = k \; \right\} \tag{2.1}$$

It is clear that $H_{n,k}(\chi)$ is a positive random variable *(r.v)* which depends on $(\mathcal{X}_1, ..., \mathcal{X}_n)$. The random feature of the kNN bandwidth represents both its main quality and also its major disadvantage.

Indeed, the fact that $H_{n,k}(\chi)$ is a *r.v* creates technical difficulties in proofs because we can not use the same tools as in the standard kernel method. But the randomness of $H_{n,k}(\chi)$ allows to define a neighborhood adapted to χ and to respect the local structure of the data.

Continuity-type and Lipschitz-type

In order to link the existing literature with this work, and to emphasise differences between the kNN method and the traditional kernel approach, we remind that the functional version of the Nadaraya-Watson kernel type estimator(introduces in Ferraty and Vieu (2006) [15]) of nonparametric functional regression is

$$\hat{r}(\chi) = \frac{\sum_{i=1}^{n} Y_i K(h^{-1} d(\chi, \mathcal{X}_i))}{\sum_{i=1}^{n} K(h^{-1} d(\chi, \mathcal{X}_i))} \tag{2.2}$$

where $\chi \in E$ is fixed, K is an asymmetrical kernel and h is non-random bandwidth.
we will consider two kinds of nonparametric models:

Continuity-type

This model is defined as:

$$r \in \mathcal{C}_E^0 = \{f : E \to \mathbb{R} / \lim_{d(\chi,\chi')\to 0} f(\chi') = f(\chi)\} \tag{2.3}$$

and will issue pointwise consistency results (see Thoerem 2.1 and 2.3 below).

Lipschitz-type

The model assumes the existence of an $\alpha > 0$ and $r \in Lip_{E,\alpha}$. Such that

$$Lip_{E,\alpha} = \{f : E \to \mathbb{R} / \exists C > 0, \forall \chi' \in E, |f(\chi) - f(\chi')| < C d(\chi, \chi')^{\alpha}\} \tag{2.4}$$

and will allow to obtain the rates of convergence (see Theorem 2.2 and Theorem 2.4 below)

2.3 *Asymptotic properties*

First, let us introduce the notation

$$\varphi_{\chi}(\varepsilon) = \mathbb{P}(\mathcal{X} \in \mathcal{B}(\chi, \varepsilon))$$

The concentration function $\varphi_{\chi}(\varepsilon)$ can be interpreted as a small ball probability (when ε is small) and will play a major role in our methodology. From one side,

it will avoid introducing density assumptions on $\mathcal{X}$, wile from the order side it will control rates of convergence of the estimate.

To establish asymptotic properties we need some hypotheses on the distribution of $(\mathcal{X}, Y)$ and on the estimator $\hat{r}_{kNN}$

(H_1) *Concentration of the f.r.v. $\mathcal{X}$.*

$\forall \varepsilon > 0, \varphi_\chi(\varepsilon) > 0$ with $\varphi_\chi(.)$ continuous and strictly increasing on a neighborhood of 0 and $\varphi_\chi(0) = 0$.

(H_2) *Conditional moments of the response r.v.y*

$\forall m \geq 2, \mathbb{E}[|Y|^m / \mathcal{X} = \chi] = \sigma_m(\chi) < \infty$ with $\sigma_m(.)$ continuous in χ

(H_3) *kernel K.*

there exist two constants $0 < C_1 < C_2 < \infty$ such that

$$C_1 \mathbb{1}_{[0,1]} \leq k \leq C_2 \mathbb{1}_{[0,1]}$$

Note that hypothesis (H_3) can be extended, to continuous kernels:

(H_3') The support of K is $[0,1]$, the derivative K' exists on $[0,1]$ and satisfies, for two real numbers

$$-\infty < C_2 < C_1 < 0 \quad \text{and} \quad C_2 \leq K' \leq C_1$$

In this case of (H_3'), we also suppose that

$$\exists C_3 > 0, \exists \varepsilon_0, \forall \varepsilon < \varepsilon_0, \int_0^\varepsilon \varphi_\chi(u) du \quad > \quad C_3 \varepsilon \varphi_\chi(\varepsilon) \tag{2.5}$$

Before studying the kNN estimator, we remind asymptotic properties of $\hat{r}$ defined by Equation (2.2). Ferraty and Vieu first showed the almost complete convergence of this estimator.

2.3.1 Some results of kernel estimator of regression for functional data

Theorem 2.1.

Under the continuity-type model (2.3), suppose (H_1)*-*(H_3) *or* $((H_3')$ *and Equation (2.5)), and suppose also that* $h = h_n$ *is a sequence of positive real numbers such that* $h \to 0$ *and* $logn/n\varphi_\chi(h) \to 0$*; then we have*

$$\hat{r}(\chi) \to^{(aco)} r(\chi)$$

They also established the rate of almost complete convergence.

Theorem 2.2.

Under the Lipschitz-type model (2.4), suppose (H_1)*-*(H_3) *or* $((H_3')$ *and Equation (2.5)), and suppose also that* $h = h_n$ *is a sequence of positive real numbers such that* $h \to 0$ *and* $logn/n\varphi_\chi(h) \to 0$*, then we have*

$$\hat{r}(\chi) - r(\chi) = O(h^\alpha) + O_{aco}\left(\sqrt{\frac{logn}{n\varphi_\chi(h)}}. \right)$$

2.3.2 Asymptotic properties of *k*-NN method estimator of regression function

Remark 2.1.

The rate of convergence of $\hat{r}$ *is divided into two parts.*
The first part comes from the bias of the estimator, The second part comes from the dispersion of $\hat{r}$ *.*

Now let us focus on the kNN method. First, we state the almost complete convergence

of $\hat{r}_{kNN}$ defined by Equation (2.1).

Theorem 2.3.

Equation under the continuity-type model (2.3), suppose $(H_1)-(H_3)$ or $((H_3')$ and equation (2.5)), and suppose also that $k = k_n$ is a sequence of positive real numbers such that $k/n \to 0$ and $logn/k \to 0$ then we have

$$\hat{r}_{kNN}(\chi) \to^{(aco)} r(\chi).$$

then, we establish the rate of almost complete convergence:

Theorem 2.4.

Under the Lipschitz-type model (2.4), suppose (H_1)-(H_3) or $((H_3')$ and equation (2.5)), and suppose also that $k = k_n$ is a sequence of positive real numbers such that $k/n \to 0$ and $logn/k \to 0$, then we have

$$\hat{r}_{kNN}(\chi) - r(\chi) = O\left(\varphi_\chi^{-1}\left(\frac{k}{n}\right)^\alpha\right) + O_{aco}\left(\sqrt{\frac{logn}{k}}\right)$$

2.3.3 *Proof of Theorems.*

Lemma 2.1

Let $(D_n)_{n\in N}$ be a sequence of *r.r.v* and $(u_n)_{n\in N}$ a decreasing positive sequence.

(i) If $l = \lim u_n \neq 0$ and if for all increasing sequence $\beta_n \in]0,1[$, there exist two sequences of r.r.v $(D_n^-(\beta_n))_{n\in N}$ and $(D_n^+(\beta_n))_{n\in N}$ such that

(L_1) $D_n^- \leq D_n^+$; $\quad \forall n \in N$ and $\mathbb{1}_{\{D_n^- \leq D_n \leq D_n^+\}} \longrightarrow^{(aco)} 1$

(L_2) $\sum_{i=1}^n G(D_n^-, A_i)/\sum_{i=1}^n G(D_n^+, A_i) - \beta_n = O_{aco}(u_n)$

(L_3) $c_n(D_n^-) - c = O_{aco}(u_n)$ and $c_n(D_n^+) - c = O_{aco}(u_n)$

Then $\quad c_n(D_n) - c = O_{aco}(u_n)$

(ii) If $l = 0$ and if $(L_1), (L_2)$ and (L_3) are checked for any increasing sequence $\beta_n \in]0,1[$ with limit 1. then the same result holds.

We give now the same kind of results but using the o_{aco}

Lemma 2.2

Let $(D_n)_{n\in N}$ be a sequence of *r.r.v* and $(v_n)_{n\in N}$ a decreasing positive sequence.

(i) If $l' = \lim v_n \neq 0$ and if, for all increasing sequence $\beta_n \in]0,1[$, there exist two sequences of r.r.v $(D_n^-(\beta_n))_{n\in N}$ and $(D_n^+(\beta_n))_{n\in N}$ such that

(L_1) $D_n^- \leq D_n^+$; $\quad \forall n \in N$ and $\mathbb{1}_{\{D_n^- \leq D_n \leq D_n^+\}} \longrightarrow^{(aco)} 1$

(L_2') $\sum_{i=1}^n G(D_n^-, A_i)/\sum_{i=1}^n G(D_n^+, A_i) - \beta_n = o_{aco}(v_n)$

(L_3') $c_n(D_n^-) - c = o_{aco}(v_n)$ and $c_n(D_n^+) - c = o_{aco}(v_n)$

Then , $\quad c_n(D_n) - c = o_{aco}(v_n)$.

(ii) If $l' = 0$ and if L_1 L_2' and L_3' are checked for any increasing sequence $\beta_n \in]0,1[$ with limit 1, then the same result holds.

Now, let us use Chernoff-type exponential inequality for Bernoulli random variables to give the essential technical tool in the verification of (L_1)

Lemma 2.3

Let $X_1, ..., X_n$ be independent *r.v's* in $\{0,1\}$. Note $X = \sum_{i=1}^n$ and $\mu = \mathbb{E}(X)$ then
$\forall \delta > 0$:

$\star$ $\mathbb{P}(X > (1+\delta)\mu) < (e^\delta/(1+\delta)^{1+\delta})^\mu$

$\star$ $\mathbb{P}(X < (1-\delta)\mu) < e^{-\delta^2/2\mu}$

We will give now a quick demonstration for Lemmas and for Theorems

Proof of Lemmas

Proof of Lemma 2.1

For technical reasons, we supposed in this proof that the $r.v$ B_i are positive. The result for any real valued r.v B_i can be deduce by taking $B_i = B_i^+ - B_i^-$ where

$$B_i^+ = \max(B_i, 0) \quad \text{and} \quad B_i^- = -\min(B_i, 0)$$

we prove simultaneously both assertion (i) and (ii).

First, remark that, for all sequence $\beta_n \in]0,1[$, (L_2) and (L_3) give

$$c_n^-(\beta_n) = \frac{\sum_{i=1}^n B_i G(D_n^-(\beta_n), A_i)}{\sum_{i=1}^n G(D_n^+(\beta_n), A_i} = \beta_n c + O_{aco}(u_n) \tag{2.6}$$

and

$$c_n^+(\beta_n) = \frac{\sum_{i=1}^n B_i G(D_n^+(\beta_n), A_i)}{\sum_{i=1}^n G(D_n^-(\beta_n), A_i} = \frac{c}{\beta_n} + O_{aco}(u_n) \tag{2.7}$$

For all $\epsilon > 0$; we note

$$T_n(\epsilon) = \{c - \epsilon u_n \leq c_n(D_n) \leq c + \epsilon u_n\}$$

and for all sequence $\beta_n \in]0,1[$:

$$S_n^-(\epsilon, \beta_n) = \{c - \epsilon u_n \leq c_n^-(\beta_n) \leq c + \epsilon u_n\}$$

$$S_n^+(\epsilon, \beta_n) = \{c - \epsilon u_n \leq c_n^+(\beta_n) \leq c + \epsilon u_n\}$$

$$S_n(\beta_n) = \{c_n^-(\beta_n) \leq c_n(D) \leq c_n^+(\beta_n)\}$$

It is obvious that

$$\forall \epsilon > 0; \forall \beta_n \in]0,1[; \quad S_n^-(\epsilon, \beta_n) \cap S_n^+(\epsilon, \beta_n) \cap S_n(\beta_n) \subset T_n(\epsilon) \tag{2.8}$$

Under (ii), we choose

$$\beta_n = \beta_{n,\epsilon} = 1 - \frac{\epsilon u_n}{3c} \quad ; \quad \forall \epsilon < \epsilon_0 = 1 \tag{2.9}$$

whereas, under (i), we take

$$\beta_n = \beta_{n,\epsilon} \quad = \quad 1 - \frac{\epsilon l}{3c} \quad ; \qquad \forall \epsilon < \epsilon_0 = \frac{3c}{l}. \tag{2.10}$$

By denoting

$$G_n^-(\epsilon) = \left\{ \ \beta_{n,\epsilon} c - \tfrac{\epsilon u_n}{3} \leq c_n^-(\beta_{n,\epsilon}) \leq \beta_{n,\epsilon} c + \tfrac{\epsilon u_n}{3} \ \right\}$$

$$G_n^+(\epsilon) = \left\{ \ \tfrac{c}{\beta_{n,\epsilon}} - \tfrac{\epsilon u_n}{3} \leq c_n^+(\beta_{n,\epsilon}) \leq \tfrac{c}{\beta_{n,\epsilon}} + \tfrac{\epsilon u_n}{3} \ \right\}$$

$$G_n(\epsilon) = \left\{ \ D_n^-(\beta_{n,\epsilon}) \leq D_n \leq D_n^+(\beta_{n,\epsilon}) \ \right\}$$

We see that Equation (2.9) and (2.10) imply that

$$c - \epsilon u_n \leq \beta_{n,\epsilon} c - \epsilon u_n/3$$

$$\beta_{n,\epsilon} c + \epsilon u_n/3 \leq c + \epsilon u_n$$

and

$$c - \epsilon u_n \leq c/\beta_{n,\epsilon} - \epsilon u_n/3$$

$$c/\beta_{n,\epsilon} + \epsilon u_n/3 \leq c + \epsilon u_n.$$

So, we have

$$G_n^-(\epsilon) \subset S_n^-(\epsilon, \beta_{n,\epsilon}) \quad \text{and} \quad G_n^+(\epsilon) \quad \subset \quad S_n^+(\epsilon, \beta_{n,\epsilon}) \tag{2.11}$$

(L_0) implies that $G_n(\epsilon) \subset S_n(\beta_{n,\epsilon})$; so by combining Equation (2.8) and (2.11), we obtain

$$\forall \epsilon \in]0, \epsilon_0[, \quad T_n(\epsilon)^c \subset G_n^-(\epsilon)^c \cup G_n^+(\epsilon)^c \cup G_n(\epsilon)^c.$$

Then

$$\mathbb{P}(|c_n(D_n) - c| > \epsilon u_n) \leq \mathbb{P}\left(\ |c_n^-(\beta_{n,\epsilon}) - \beta_{n,\epsilon} c| > \tfrac{\epsilon u_n}{3} \ \right)$$

$$+\mathbb{P}\left(\ \left| \ c_n^+(\beta_{n,\epsilon}) - \tfrac{c}{\beta_{n,\epsilon}} \ \right| > \tfrac{\epsilon u_n}{3} \ \right) + \mathbb{P}(D_n \notin [D_n^-(\beta_{n,\epsilon}), D_n^+(\beta_{n,\epsilon})])$$

According to Equation (2.6) and (2.7), there exists $0 < \epsilon_1 < \epsilon_0$ such that

$$\sum_{n \in N} \mathbb{P}\left(\ |c_n^-(\beta_{n,\epsilon_1} c| > \tfrac{\epsilon_1 u_n}{3} \ \right) < \infty$$

and

$$\sum_{n \in N} \mathbb{P}\left(\ \left| \ c_n^+(\beta_{n,\epsilon_1}) - \tfrac{c}{\beta_{n,\epsilon_1}} \ \right| > \tfrac{\epsilon_1 u_n}{3} \ \right) < \infty$$

Now, according to $(L_1), \forall \epsilon > 0$

$$\sum_{n\in N} \mathbb{P}(D_n \notin [D_n^-(\beta_{n,\epsilon}), D_n^+(\beta_{n,\epsilon})]) < \infty$$

There, there exists $0 < \epsilon_1 < \epsilon_0$ such that $\sum_{n\in N} \mathbb{P}(|c_n(D_n) - c| > \epsilon_1 u_n) < \infty$.

Proof of Lemma 2.2

For all sequence $\beta_n \in]0, 1[$, (L_2') and (L_3') give

$$c_n^-(\beta_n) = \frac{\sum_{i=1}^n B_i G(D_n^-(\beta_n), A_i)}{\sum_{i=1}^n G(D_n^+(\beta_n), A_i)} = \beta_n c + o_{aco}(v_n) \tag{2.12}$$

and

$$c_n^+(\beta_n) = \frac{\sum_{i=1}^n B_i G(D_n^+(\beta_n), A_i)}{\sum_{i=1}^n G(D_n^-(\beta_n), A_i)} = \frac{c}{\beta_n} + o_{aco}(v_n) \tag{2.13}$$

With the same arguments as in the previous proof, we arrive at $\forall \epsilon > 0$

$$\mathbb{P}(|c_n(D_n - c| > \epsilon v_n) \leq \mathbb{P}\left(|c_n^-(\beta_{n,\epsilon}) - \beta_{n,\epsilon} c| > \tfrac{\epsilon v_n}{3} \right)$$

$$+\mathbb{P}\left(\left| c_n^+(\beta_{n,\epsilon}) - \tfrac{c}{\beta_{n,\epsilon}} \right| > \tfrac{\epsilon v_n}{3} \right) + \mathbb{P}(D_n \notin [D_n^-(\beta_{n,\epsilon}), D_n^+(\beta_{n,\epsilon})])$$

According to equation (2.12) and (2.13), $\forall \epsilon > 0$

$$\sum_{n\in N} \mathbb{P}\left(|c_n^-(\beta_{n,\epsilon}) - \beta_{n,\epsilon} c| > \tfrac{\epsilon v_n}{3} \right) < \infty$$

and

$$\sum_{n\in N} \mathbb{P}\left(|c_n^+(\beta_{n,\epsilon}) - \tfrac{c}{\beta_{n,\epsilon}}| > \tfrac{\epsilon v_n}{3} \right) < \infty$$

Then, $\forall \epsilon > 0, \sum_{n\in N} \mathbb{P}(|c_n(D_n) - c| > \epsilon v_n) < \infty$.

Proof of Lemma 2.3

For $t > 0$, we use Markov inequality to obtain

$$\mathbb{P}(X > (1+\delta)\mu) \leq \frac{\mathbb{E}[e^{tX}]}{e^{(1+\delta)t\mu}} \tag{2.14}$$

Then, the independence of the Bernoulli variables X, gives

$$\mathbb{E}[e^{tX}] = \prod_{i=1}^{n}(1 + \mathbb{P}(X_i = 1)(e^t - 1))$$

Using the fact that $\forall x > 0, 1 + x < e^x$, we arrive at

$$\mathbb{E}[e^{tX}] \quad \leq \quad e^{(e^t-1)\mu} \tag{2.15}$$

The first result comes by combining equation (2.14) and (2.15) and taking $t = \ln(\delta + 1)$.
For the second point, we follow the same way to have

$$\mathbb{P}(X < (1 - \delta)\mu) \leq \frac{e^{(e^{-t}-1)\mu}}{e^{(\delta-1)t\mu}}$$

Which is minimized in $t = \ln(1/1 - \delta)$. Then, we obtain

$$\mathbb{P}(X < (1 - \delta)\mu) < \left(\frac{e^{-\delta}}{(1-\delta)^{(1-\delta)}} \right)^{\mu}$$

Using Taylor expansion, we show that $(1-\delta)^{(1-\delta)} > e^{-\delta}+\delta^2/2$ and the result follows.

Proof of Theorems

Proof of Theorem 2.3

We use Lemma 2.2 (i) with $w_n = 1, c_n(H_{n,k}(\chi)) = \hat{r}_{kNN}(\chi)$ and $c = r(\chi)$. We first remind that, under the same conditions as in Theorem 2.1. Ferraty and Vieu (2006) [15] showed that

$$\frac{1}{n\varphi_\chi(h)} \sum_{i=1}^{n} K(h^{-1}d(\chi, \mathcal{X}_i)) \longrightarrow^{(aco)} 1 \tag{2.16}$$

Let $\beta \in]0, 1[$, we choose D_n^- and D_n^+ such that

$$\varphi_\chi(D_n^-) = \sqrt{\beta}\frac{k}{n}$$

$$\varphi_\chi(D_n^+) = \frac{1}{\sqrt{\beta}}\frac{k}{n}$$

This choice and hypotheses on k allow us to use Theorem 2.1 with $h^{-1} = D_n^- = \varphi_\chi^{-1}(\sqrt{\beta}(k/n))$ and with $h^+ = D_n^+ = \varphi_\chi^{-1}((1/\sqrt{\beta})(k/n))$ to have

$$c_n(D_n^-) \longrightarrow^{(aco)} c.$$

$$c_n(D_n^+) \longrightarrow^{(aco)} c.$$

So that (L_3') is checked. Now, by applying equation (2.16) both with h^- and h^+, we have that

$$\frac{1}{n\varphi_\chi(D_n^-)} \sum_{i=1}^{n} K((D_n^-)^{-1} d(\chi, \mathcal{X}_i)) \to^{(aco)} 1$$

$$\frac{1}{n\varphi_\chi(D_n^+)} \sum_{i=1}^{n} K((D_n^+)^{-1} d(\chi, \mathcal{X}_i)) \to^{(aco)} 1$$

then

$$\frac{\sum_{i=1}^{n} K((D_n^-)^{-1} d(\chi, \mathcal{X}_i))}{\sum_{i=1}^{n} K((D_n^+)^{-1} d(\chi, \mathcal{X}_i))} \to^{(aco)} \beta$$

So (L_2') is checked. Finally, we have to verify (L_1): the first part is obvious and the second one does not deal with rates of convergence. We have to show that, for all $\epsilon > 0$

$$\sum_{n \in N} \mathbb{P}\left(\left| \mathbb{1}_{\{D_n^- \leq H_{n,k}(\chi) \leq D_n^+\}} - 1 \right| > \epsilon \right) < \infty$$

Let $\epsilon > 0$, we have

$$\mathbb{P}\left(\left| \mathbb{1}_{\{D_n^- \leq H_{n,k}(\chi) \leq D_n^+\}} - 1 \right| > \epsilon \right) \leq \mathbb{P}(H_{n,k}(\chi) < D_n^-) + \mathbb{P}(H_{n,k}(\chi) > D_n^+)$$

which can be written as

$$\mathbb{P}(|\mathbb{1}_{\{D_n^- \leq H_{n,k}(\chi) \leq D_n^+\}} - 1| > \epsilon) \leq$$

$$\mathbb{P}\left(\sum_{i=1}^{n} \mathbb{1}_{B(\chi, D_n^-)}(\mathcal{X}_i) > k \right) + \mathbb{P}\left(\sum_{i=1}^{n} \mathbb{1}_{B(\chi, D_n^+)}(\mathcal{X}_i) < k \right)$$

Now, we use Lemma 2.3 to show that

$$\mathbb{P}\left(\sum_{i=1}^{n} \mathbb{1}_{B(\chi, D_n^-)}(\mathcal{X}_i) > k \right) < \left(n^{-\log[\sqrt{\beta}\exp(1-\sqrt{\beta})]} \right)^{-k/\log n}$$

and

$$\mathbb{P}\left(\sum_{i=1}^{n} \mathbb{1}_{B(\chi, D_n^+)}(\mathcal{X}_i) < k \right) < \left(n^{(1-\sqrt{\beta})^2/2\sqrt{\beta}} \right)^{-k/\log n}$$

Then, because $\log n/k \to 0$

$$\sum_{n \in N} \mathbb{P}(|\mathbb{1}_{\{D_n^- \leq H_{n,k}(\chi) \leq D_n^+\}} - 1| > \epsilon) < \infty \qquad \forall \epsilon > 0$$

So, (L_1) is checked and this end the proof of Theorem 2.3.

Proof of Theorem 2.4

The scheme of the proof is likely the same as for Theorem 2.3 before. The main change consists in using Lemma 2.1 (ii) in place of Lemma 2.2 (i). First, we remind that, under the same conditions as Theorem 2.2. Ferraty and Vieu (2006) [15] showed that

$$\frac{1}{n}\sum_{i=1}^{n} K(h^{-1}d(\chi,\mathcal{X}_i)) - \varphi_\chi(h) \quad = \quad O_{aco}\left(\sqrt{\frac{\log n}{n\varphi_\chi(h)}}\right) \tag{2.17}$$

Let $\beta_n \in]0,1[$ be an increasing sequence with limit 1, we choose D_n^- and D_n^+ such that

$$\varphi_\chi(D_n^-) = \sqrt{\beta_n}\frac{k}{n}$$

$$\varphi_\chi(D_n^+) = \frac{1}{\sqrt{\beta_n}}\frac{k}{n}$$

So, we can use Theorem 2.2 with $h^- = D_n^- = \varphi_\chi^{-1}(\sqrt{\beta_n}(k/n))$ and with $h^+ = D_n^+ = \varphi_\chi^{-1}((1/\sqrt{\beta_n})(k/n))$ and, because β_n is bounded by 1, we have

$$c_n(D_n^-) - c = O\left(\varphi_\chi^{-1}\left(\tfrac{k}{n}\right)^\alpha\right) + O_{aco}\left(\sqrt{\tfrac{\log n}{k}}\right)$$

$$c_n(D_n^+) - c = O\left(\varphi_\chi^{-1}\left(\tfrac{k}{n}\right)^\alpha\right) + O_{aco}\left(\sqrt{\tfrac{\log n}{k}}\right)$$

So that (L_3) is checked. Now, by applying Equation (2.17) both with h^+ and h^- and, because β_n is bounded by 1, we have that

$$\frac{1}{n}\sum_{i=1}^{n} K((D_n^-)^{-1}d(\chi,\mathcal{X}_i)) = \sqrt{\beta_n}\frac{k}{n} + O_{aco}\left(\sqrt{\tfrac{\log n}{k}}\right)$$

and

$$\frac{1}{n}\sum_{i=1}^{n} K((D_n^+)^{-1}d(\chi,\mathcal{X}_i)) = \frac{1}{\sqrt{\beta_n}}\frac{k}{n} + O_{aco}\left(\sqrt{\tfrac{\log n}{k}}\right)$$

Then, we have

$$\frac{\sum_{i=1}^{n} K((D_n^-)^{-1}d(\chi,\mathcal{X}_i))}{\sum_{i=1}^{n} K((D_n^+)^{-1}d(\chi,\mathcal{X}_i))} - \beta_n = O_{aco}\left(\sqrt{\tfrac{\log n}{k}}\right)$$

and (L_2) is checked. The verification of (L_1) is the same as in previous proof.

2.3.4 *Other results about the rate of convergence by k nearest neighbors method*

Consider the simple additive noise model $Y = r(\mathcal{X}) + \epsilon$ where ϵ takes values in $\mathcal{H}$, and $\mathbb{E}[\epsilon/\chi] = 0$. Given n copies of independent observations $\mathcal{D}_n = \{(\mathcal{X}_1, Y_1), ..., (\mathcal{X}_n, Y_n)\}$, the kNN estimate at any $x \in \mathcal{F}$ is defined by

$$\hat{r}(\chi) = \sum_{i=1}^{n} w_{ni} Y_i \tag{2.18}$$

where $(w_{n1}, ..., w_{nn})$ 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. We have here an example of w_{ni} follow.

Example 2.1.

Take $w_{ni} = K(d(\mathcal{X}_i, \chi)/H)/\sum_j K(d(\mathcal{X}_j, \chi)/H)$ where K is a kernel function and H is the distance of the *k-th* nearest neighbor. Mathematically

$$H = \min\{h \in \mathbb{R} : \sum_{i=1}^{n} I\{\mathcal{X}_i \in B(\chi, h)\} \geq k\} \tag{2.19}$$

where $B(\chi, h) = \{\chi' \in \mathcal{F} : d(\chi, \chi') \leq h\}$ and $I\{.\}$ denotes the indicator function. For simplicity we 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 rate of convergence.

Assumption 2.1. *r is bounded and Lipschitz continuous at χ, that is $\| r(\chi) \| \leq B, \forall \chi \in \mathcal{F}$ and $\| r(\chi) - r(\chi) \| \leq M d(\chi, \chi')^{\alpha}$. The Lipschitz condition only needs to be satisfied locally on an open neighborhood of the fixed χ.*

Assumption 2.2. *Suppose that $\sum_{i=k+1}^{n} w_{ni} = O(b_n)$ and denote $\| w \|_s = (\sum_{i=1}^{n} w_{ni}^8)^{1/8}$; we assume $b_n \to 0, \| w \|_2 \to 0$, where the asymptotic orders are in the sense of al-*

most sure convergence. We also require that $k/n \to 0$ *and* $k/\log n \to \infty$.

Assumption 2.3. $\mathbb{E} \parallel \epsilon \parallel^r < \infty$ *for some* $r > 2$.

Assumption 2.4. $\mathbb{P}(\parallel \epsilon \parallel > a) \leq \exp\{-Ca^p\}$ *with* $C > 0$ *and* $p > 0$*; for any* $a > 0$.

Theorem 2.5.

If Assumption 1,2 and 3 hold and $\sum_{n=1}^{\infty}(\log n)^{(r-2)/2}(\parallel w \parallel_r / \parallel w \parallel_2)^r < \infty$, *then* $\parallel \hat{r}(\chi) - r(\chi) \parallel = O(b_n + [\phi^{-1}(2k/n)]^\alpha + (\log n)^{1/2} \parallel w \parallel_2)$ *almost surely, where*

$$\phi^{-1}(\chi) := \inf\{h : \phi(h) \geq \chi\}$$

.

Alternatively, assuming exponential tail decay, we have

Theorem 2.6.

If Assumptions 1,2 and 4 hold, then $\parallel \hat{r}(\chi) - r(\chi) \parallel = O(b_n + [\phi^{-1}(2k/n)]^\alpha + (\log n)^{1+1/p} \parallel w \parallel_2)$ *almost surely*

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

Corollary 2.1.

For the simple k-NN estimates ($w_{ni} = 1/k$ *for* $i \leq k$ *and 0 otherwise), the theorems above hold with* $b_n = 0$ *and* $\parallel w \parallel_2 = O(1/\sqrt{k})$. *The same applies to Example(with a kernel compactly supported and bounded away from zero on [0,1]) presented previously.*

2.3.5 *Proof of Theorems*

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

Lemma 2.4.

Suppose $k/n \to 0$ and $k/\log n \to \infty$. Let H be the distance from χ to its k-th nearest neighbor as defined in (2.19),then $\mathbb{P}(H \geq \phi^{-1}(2k/n), i.o.) \to 0$, where i.o. mean «infinitely often», and $\phi^{-1}(\chi) := \inf\{h : \phi(h) \geq \chi\}$.

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

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

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

Proof of Theorem 2.5

We use the following decomposition into the bias term and the variance term.

$$\| \hat{r}(\chi) - r(\chi) \| \leq \| \sum_i w_{ni}(r(\mathcal{X}_i) - r(\chi)) \| + \| \sum_i w_{ni}\epsilon_i \| \tag{2.20}$$

The bias term is easier to deal with. In fact

$$\begin{aligned}
\| \sum_i w_{ni}(r(\mathcal{X}_i) - r(\chi)) \| &\leq 2B \sum_{i=k+1}^{n} w_{ni} + \| \sum_{i=1}^{k} w_{ni}(r(\mathcal{X}_i) - r(\chi)) \| \\
&= O(b_n + [\phi^{-1}(\frac{2k}{n})]^{\alpha})
\end{aligned}$$

by assumption 2.1 and Lemma 2.4.

Now we deal with the variance term. Let

$$S_n = \sum_{i=1}^{n} w_{ni}\epsilon_i$$

and the following arguments are conditional on $\{\mathcal{X}_1, ..., \mathcal{X}_n\}$ (in effect treating w_{ni} as non random weights). We will write

$$\| S_n \| - \mathbb{E} \| S_n \| = \| \sum_{i=1}^{n} w_{ni}\epsilon_i \| - \mathbb{E} \| \sum_{i=1}^{n} w_{ni}\epsilon_i \| = \sum_{i=1}^{n} d_i$$

where we remind the reader that the expectation is conditional on $\{\mathcal{X}_1, ..., \mathcal{X}_n\}$, with

$$d_i = \mathbb{E}[\| S_n \| |\mathcal{G}_i] - \mathbb{E}[\| S_n \| |\mathcal{G}_{i-1}]$$

where $\mathcal{G}_i$ is the $\sigma-$algebra generated by $\epsilon_1, ..., \epsilon_i$ ($\mathcal{G}_0$ 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.
we know that

$$|d_i| \leq \| \epsilon \| w_{ni} + w_{ni}\mathbb{E} \| \epsilon_i \| \leq \| \epsilon_i \| w_{ni} + C w_{ni} \tag{2.21}$$

and

$$\mathbb{E}(d_i^2 | \mathcal{G}_{i-1}) \leq w_{ni}^2 \mathbb{E} \| \epsilon \|^2 \tag{2.22}$$

We bound the variance term in four steps
Step 1: We show

$$\begin{aligned}
\mathbb{E} \| S_n \| &= O(\| w \|_2) \\
\mathbb{E} \| S_n \| &= \mathbb{E} \| \sum_{i=1}^{n} w_{ni}\epsilon_i \| \\
&\leq \sqrt{\mathbb{E} < \sum_{i=1}^{n} w_{ni}\epsilon_i, \sum_{i=1}^{n} w_{ni}\epsilon_i >} \\
&= O(\sqrt{\sum_i w_{ni}^2}) \\
&= O(\| w \|_2).
\end{aligned}$$

Step 2: Let $d_i' = d_i I\{|d_i| \leq L\}$ for some $L > 0$ to be specified later. We have

$$\mathbb{P}(\sum_{i=1}^{n}(d_i'|\mathcal{G}_{i-1})) > a) \leq \exp\{-Ca^2/(aL + (\sum_i w_{ni}^2))\}, \forall a > 0$$

Using (2.22)

$$\mathbb{E}[(d_i' - \mathbb{E}(d_i'|\mathcal{G}_{i-1}))^2|\mathcal{G}_{i-1}] \leq \mathbb{E}(d_i'^2|\mathcal{G}_{i-1}) \leq \mathbb{E}(d_i'^2|\mathcal{G}_{i-1}) = O(w_{ni}^2)$$

and together with

$$|d_i' - \mathbb{E}(d_i'|\mathcal{G}_{i-1})| \leq 2L$$

we get

$$\mathbb{E}(|d_i' - \mathbb{E}(d_i'|\mathcal{G}_{i-1})|^r|\mathcal{G}_{i-1}) \leq C(2L)^{r-2}w_{ni}^2.$$

Since

$$d_i' - \mathbb{E}(d_i'|\mathcal{G}_{i-1}); \quad i \leq n$$

is a martingale difference sequence, (using Bernstein'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

$$\mathbb{P}(\sum_i |d_i'' - \mathbb{E}(d_i''|\mathcal{G}_{i-1})| > a) \leq C(\sum_i w_{ni}^m)L^{1-m}/a$$

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

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

and note that in the last line above we used the bound (2.21). Thus we have

$$\begin{aligned}
\mathbb{P}(\sum_i |d_i'' - \mathbb{E}(d_i''|\mathcal{G}_{i-1})| > a) &\leq \mathbb{E}[\sum_i |d_i'' - \mathbb{E}(d_i''|\mathcal{G}_{i-1})|]/a \\
&\leq C(\sum_i w_{ni}^m)L^{1-m}/a
\end{aligned}$$

Step 4: Finally, we demonstrate the bound for the variance term in (2.20). Using

$$\mathbb{E}(d_i|\mathcal{G}_{i-1}) = \mathbb{E}(d_i'|\mathcal{G}_{i-1}) + \mathbb{E}(d_i''|\mathcal{G}_{i-1}) = 0$$

we have that

$$d_i = d_i' - \mathbb{E}(d_i'|\mathcal{G}_{i-1}) + (d_i'' - \mathbb{E}(d_i''|\mathcal{G}_{i-1}))$$

and then

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

By the previous two steps. Setting

$$a = C(\log n)^{1/2} \parallel w \parallel_2$$

for a constant C large enough and

$$L = \parallel w \parallel_2 (\log n)^{-1/2}$$

an application of the Borel-Cantelli Lemma leads to

$$\parallel S_n \parallel - \mathbb{E} \parallel S_n \parallel = O((\log n)^{1/2} \parallel w \parallel_2)$$

using the assumption that

$$\sum_i (\log n)^{(m-2)/2} (\parallel w \parallel_m / \parallel w \parallel_2)^m < \infty$$

Combining this with the result from Step 1, the variance term is thus

$$\parallel S_n \parallel = O((\log n)^{1/2} \parallel w \parallel_2)$$

Proof of Theorem 2.6.

The general proof strategy is the same as Theorem 2.5. 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

$$\mathbb{P}(\sum_i \mathbb{E}(d_i'|\mathcal{G}_{i-1}) > a) + \mathbb{P}(\textit{for some } i, |d_i| > L) = O(n.\exp\{-CL^p/w_{n1}^p\})$$

if we set $a = C(\log n)^{1+1/p} \parallel w \parallel_2$ and $L = C(\log n)^{1/p} w_{n1}$
for C large enough. Consider the first probability, we have

$$\begin{aligned}
\mathbb{E}(d_i'|\mathcal{G}) &\leq \mathbb{E}(|d_i| I\{|d_i| > L|\mathcal{G}_{i-1}\}) \\
&\leq (\mathbb{E}|d_i|^m|\mathcal{G}_{i-1}))^{1-m} P(|d_i| > L|\mathcal{G}_{i-1})^{1-1/m} \\
&\leq C(\mathbb{E} \parallel \epsilon_i \parallel^m w_{ni}^m)^{1/m} \exp\{-C(L - Cw_{ni})^p/w_{ni}^p\} \\
&\leq Cw_{ni} \exp\{-CL^p/w_{ni}^p\} \\
&\leq C \exp\{-CL^p/w_{n1}^p\}
\end{aligned}$$

using (2.22) and assumption 2.4 in the third inequality above.
Thus

$$\mathbb{E}(d_i' \; \mathcal{G}_{i-1}) \leq a/n$$

if we set

$$a = C(\log n)^{1+1/p} \parallel w \parallel_2$$

(note that $a \geq \parallel w \parallel_2 \geq w_{n1} \geq 1/n$)
and

$$L = C(\log n)^{1/p} w_{n1}$$

and then

$$\mathbb{P}(\sum_i \mathbb{E}(d_i'|\mathcal{G}_{i-1}) > a) = 0.$$

For the other probability term, again using (2.22) and assumption 2.4, we have

$$\begin{aligned}
\mathbb{P}(\textit{for some } i, |d_i| > L) &\leq 1 - \mathbb{P}(\forall i, w_{ni} \parallel \epsilon_i \parallel \leq L - Cw_{ni}) \\
&\leq 1 - (1 - \exp\{-C(L - Cw_{ni})^p/w_{ni}^p\})^n \\
&\leq 1 - (1 - \exp\{-CL^p/w_{n1}^p\})^n \\
&\leq n.\exp\{-CL^p/w_{n1}^p\}
\end{aligned}$$

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

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

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

By the bounds obtained in Step 2 and Step 3'. Finally set

$$a = C_1(\log n)^{1+1/p} \parallel w \parallel_2 \qquad \text{and} \qquad L = C_2(\log n)^{1/p} w_{n1}$$

(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

$$\parallel S_n \parallel = O((\log n)^{1+1/p} \parallel w \parallel_2)$$

Proof of Corollary 2.1.

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

$$w_{ni} = K(d(\mathcal{X}_i, \chi)/H) / \sum_j K(d(\mathcal{X}_j, \chi)/H) \leq C / \sum_j K(d(\mathcal{X}_j, \chi)/H)$$

and

$$K(d(\mathcal{X}_j, \chi)/H)$$

is bounded away from zero for $j \leq k$ and 0 for $j > k$ by the assumptions made on K, we have

$$w_{ni} = \begin{cases} O(1/k), & \text{for } i \leq k \\ 0, & \text{otherwise} \end{cases}$$

It then follows that $\parallel w \parallel_2 = O(1/\sqrt{k})$.

Chapter 3

Simulation using the k nearest neighbors method

3.1 *Regression versus classification problems*

In this chapter, we will give a simulation study of qualitative and quantitative responses using the k nearest neighbors method, also we want to compare this method with linear regression method, this one is given by Gareth James, not forget the classification problems.

Some statistical methods such as k nearest neighbors and boosting, can be used in the case of either quantitative or qualitative responses (also known as categorical). Quantitative variables take on numerical values. Examples: person's age, height, or income. In contrast, qualitative variables take on values in one of k different classes, or categories. Examples: person's gender (male or female), a parson's defaults on a debt(Yes or No). We tend to refer to problems with a quantitative responses as regression problems, while those involving a qualitative response are often referred to as classification problems.

Classification problem

In this part, we discuss some of the most important concepts that arise in selecting a statistical learning procedure for a specific data set. We will explain how the concepts presented here can be applied in practice.
In theory we would always like to predict qualitative responses using the Bayes classifier. But for real data, we do not know the conditional distribution of Y given X, and so computing the Bayes classifier is impossible.There for; we use the k nearest neighbors classifier. Given a positive integer k and a test observation x_0, the k-NN classifier first identifies the k points in the training data that are closed to x_0, represented by $\mathcal{N}_0$. It then estimates the conditional probability for class j as the fraction of points in $\mathcal{N}_0$ whose response values equal j:

$$Pr(Y = j/X = x_0) \quad = \quad \frac{1}{k} \sum_{i \in \mathcal{N}_0} I(y_i = j)$$

Finally,k-NN applies Bayes rule and classifies the test observation x_0 to the class with the largest probability.

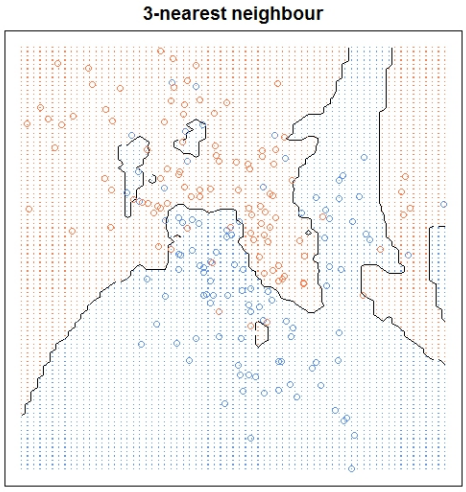


Figure 3.1: 3-nearest neighbors

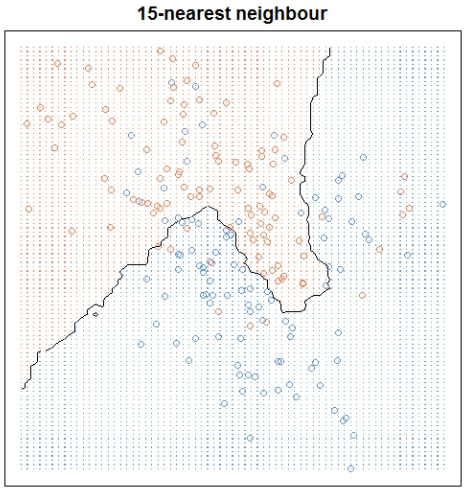


Figure 3.2: 15-nearest neighbors

Figure 3.1: provides an example of k-NN approach. We have plotted a large training data of ElemStatLearn Library of mixture.example set consisting of three blue; and

three orange observations. Our goal is to make a prediction for the points labeled by the black cross. Suppose that we choose $k = 3$.
Then k-NN will first identify the three observations that are closest the cross. This neighborhood is shown as a circle. It consists of two blue points and one orange point,resulting in estimated probabilities of 2/3 for the blue class and 1/3 for the orange class. Hence k-NN will predict that the black cross belongs to the blue class. In the next Figure 3.2, we use the same Library, we have applied the kNN approach with $k = 15$ at all of the possible values for X_1 and X_2, and have drown in the corresponding kNN decision boundary.

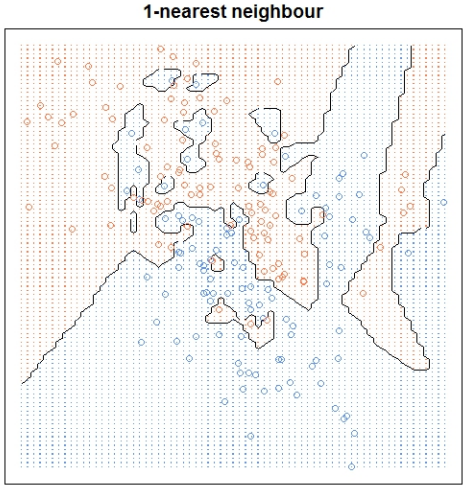


Figure 3.3: 1-nearest neighbors

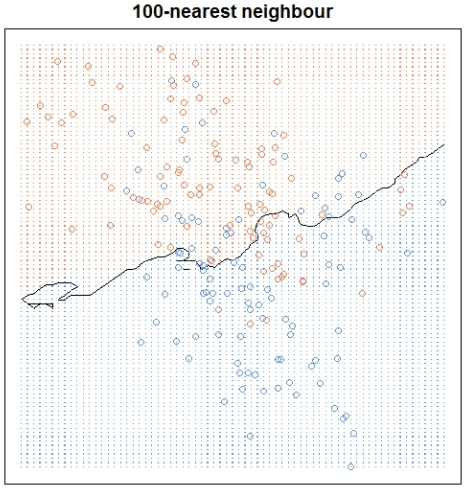


Figure 3.4: 100-nearest neighbors

The choice of k has a drastic effect on the kNN classifier obtained. In the Figure 3.3 and Figure 3.4 , using *k=1 and* k=100. When k=1, the decision boundary is overly flexible and finds patterns in the data. This corresponds to a classifier that has low bias but very high variance. As k grows, the method becomes less flexible and produces a decision boundary that is close to linear.

This corresponds to a low-variance but high-bias classifier. On this simulated data set, neither k=1 nor k=100 give good predictions: they have test error rates of 0.1695 and 0.1925, respectively.

Just as in the regression setting, there is not a strong relationship between the

training error rate and the test error rate. With k=1, the k-NN training error rate is 0, but the test error rate may be quite high. In general, as we use more flexible classification methods, the training error rate will decline but the test error rate may not.

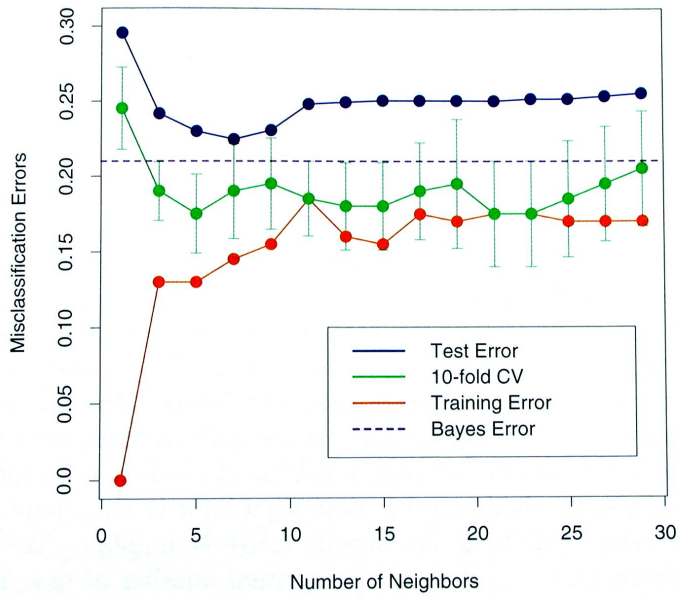


Figure 3.5: Misclassification error rate

In the Figure 3.5, we have plotted the k-NN test and training errors as a function of k the number of neighborhood. using the Misclassification method given in chapter 1; and also we compare between CV (also given in chapter 1); and Bayes error rate. As in the regression setting, the training error rate consistently declines as the flexibility increases. However, the test error exhibits a characteristic U-shape, declining at first (with a minimum at approximately $k = 10$) before increasing again when the method becomes excessively flexible and overfits.

In both the regression and classification settings, choosing the correct level of flexibility is critical to the success of any statistical learning method.

3.2 *Comparison of Linear regression with k nearest neighbors*

3.2.1 *Linear regression*

We want in this section to compare between linear regression as a parametric method and our method k nearest neighbors, first we give linear regression in R. Linear regression is an example of a parametric approach because it assumes a linear functional form for f(X). Parametric methods have several advantages. They are often easy to fit, because one need estimate only a small number of coefficients. But parametric methods do have a disadvantage: by construction, they make strong assumptions about the form of f(X). If the specifier functional form is far from the truth, and prediction accuracy is our goal, then the parametric method will perform poorly.

In contract, nonparametric methods do not explicitly assume a parametric form for f(X), and there by provide an alternative and more flexible approach for performing regression. Here, we consider one of the simplest and best-known nonparametric methods, k-nearest neighbors regression.

The k-NN regression method is closely related to the k-NN classifier discussed in last section. Given a value for k and a prediction points x_0, kNN regression first identifier the k training observations that are closest to x_0, represented by $\mathcal{N}_0$. It then estimates $f(x_0)$ using the average of all the training responses in $\mathcal{N}_0$. In other words

$$\hat{f}(x_0) = \frac{1}{k} \sum_{x_i \in \mathcal{N}_0} y_i$$

In general, the optimal value for k will depend on the bias-variance tradeoff, which a small value for k provides the most flexible fit, which will have low bias but hight variance. This variance is due to the fact that the prediction in a given region is entirely dependent on just one observation.

In contrast, larger values of k provide a smoother and less variable fit; the prediction in a region is an average of several points, and so changing one observation has a smaller effect. However, the smoothing may cause bias by masking some of the

structure in f(X).

In next section, we introduce several approaches for estimating test error rates. These methods can be used to identify the optimal value of k in kNN regression.
In what setting will a parametric approach such as least squares linear regression out perform a nonparametric approach such as kNN regression?
The answer is simple: the parametric approach will out perform the nonparametric approach if the parametric form that has been selected is close to the true form of f.

Simple linear regression in R

Here we load the MASS package, which is very large collection of data sets and

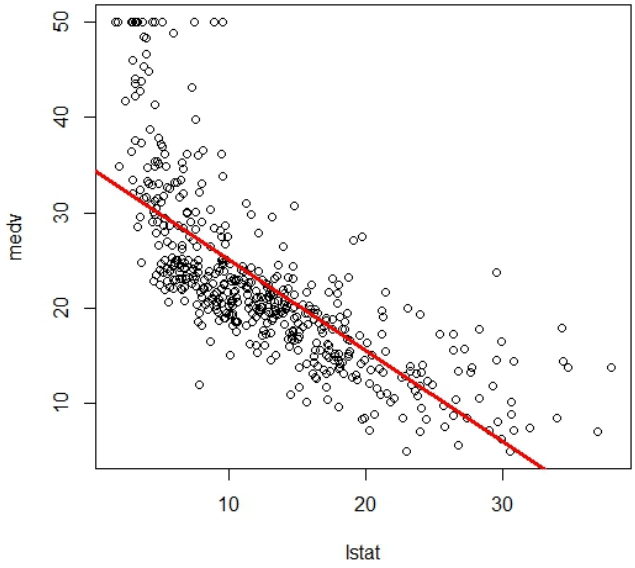


Figure 3.6: Linear regression

function. We also load the ISLR package, which includes the data sets associated.
The MASS library contains the Boston data set, which records medv(median house value) for 506 neighborhoods around Boston. We will seek to predict medv using 13 predictors such as rm (average number of rooms per house), age (average age of houses), and Istat(percent of households with low socioeconomic status).
For instance, the 95% confidence interval associated with a Istat value of 10 is

(24.47,25.63), and the 95% prediction interval is (12.828,37.28). As expected, the confidence and prediction intervals are centered around the same point(a predicted value of 25.05 for medv when Istat equals 10), but the latter are substantially wider. There is some evidence for non-linearity in the relationship between Istat and medv.

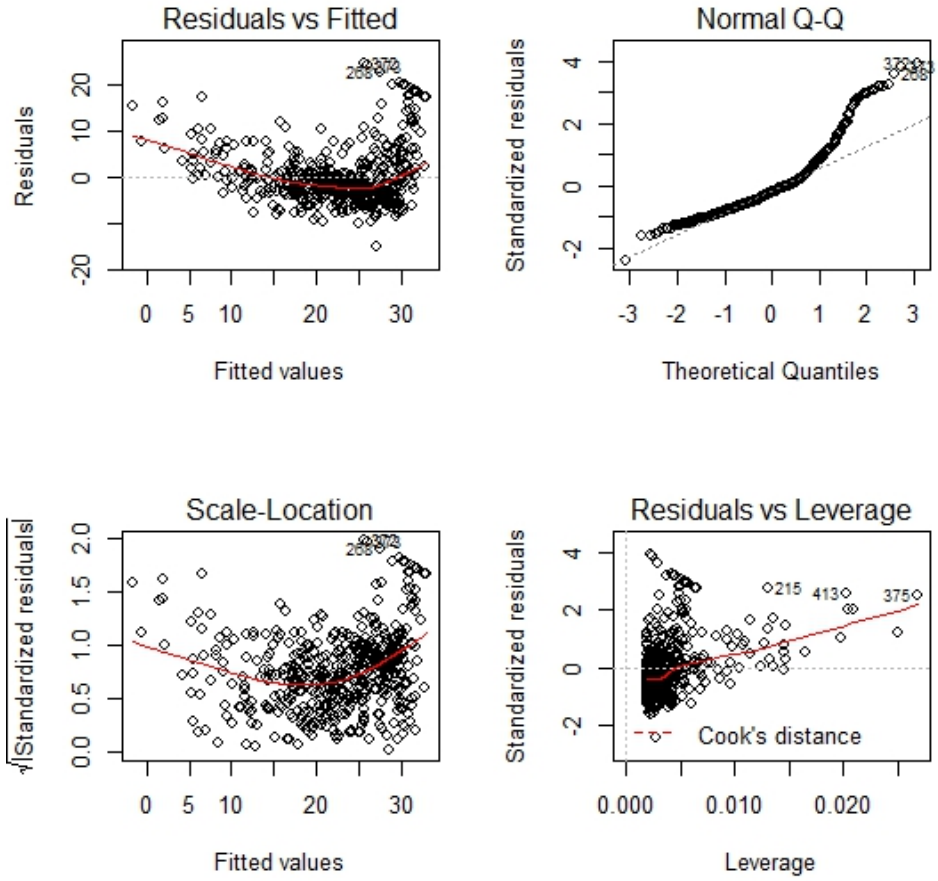


Figure 3.7: Diagnostic plots

Next, in Figure 3.7 we examine some diagnostic plots (several of which were discussed) four diagnostic plots are automatically produced by applying the plot().

Alternatively, in Figure 3.8 we can compute the residuals from a linear regression fit using the residuals() function. The function rstudent() will return the studentized residuals, and we can use this function to plot the residuals against the fitted values.

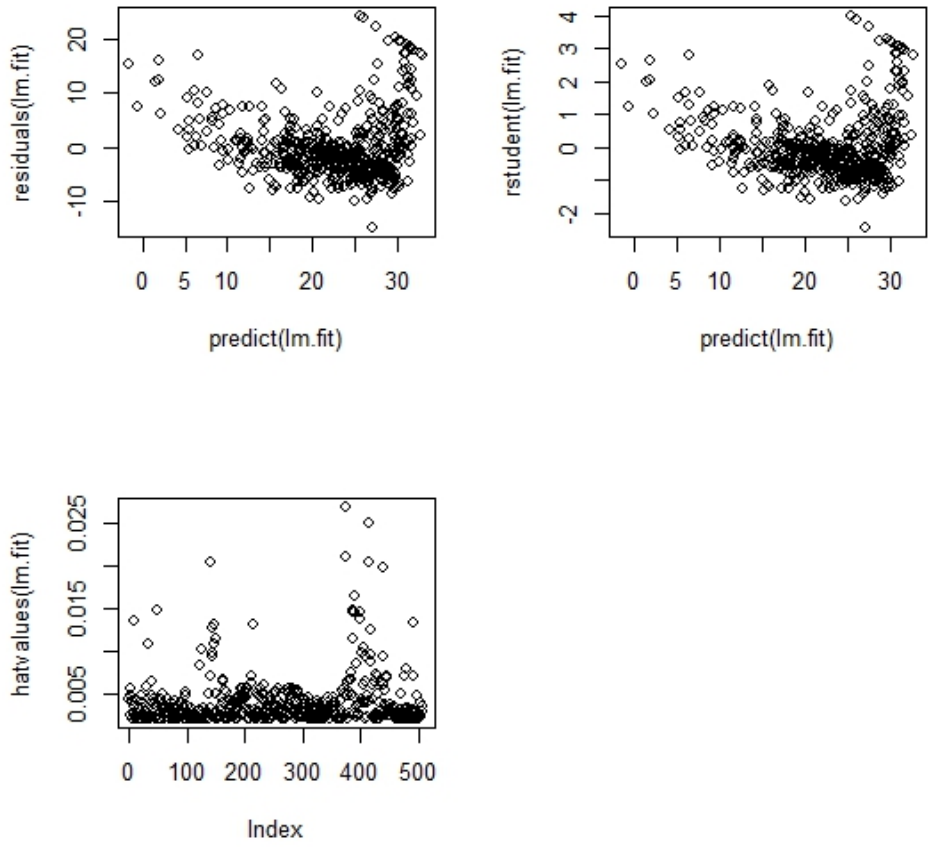


Figure 3.8: Diagnostic plots for residuals

Multiple linear regression

In order to fit a multiple linear regression model using least squares. We again use the Under the lm() function. The syntax lm(y ∼ x1+x2+x3) is used to fit a model with three predictors.
As the last section, the Boston data set would be cumbersome to have to type all of these, in order to perform a regression using all of the predictors. Instead, we can use the following short-hand:

>lm.fit=lm(medv .,data=Boston).

What if we would like to perform a regression using all of the variables but one?

>library(car)
>vif(lm.fit)
For example, in the above regression out put, vif() function is a part of the car package that we must install it in R, age has a high p-value. So we may wish to run a regression excluding this predictor. The following syntax results in regression using all except age
>lm.fit1=lm(medv .,-age,data=Boston)
Alternatively, the update() function can be used
lm.fit1=update(lm.fit, .-age).

3.2.2 *Non-linear transformations of the predictors*

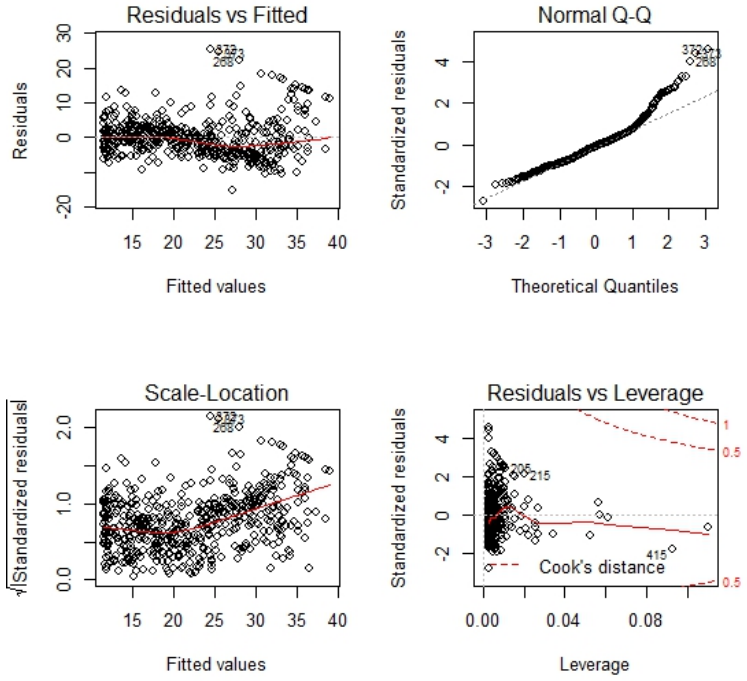


Figure 3.9: Non-linear regression

The lm() function can also accommodate non-linear transformations of the predictors. For the instance, The function I() is neededWe now perform a regression of medv onto *Istat* and $Istat^2$, we use the anova() function to further quantify the extent to which the quadratic fit is superior to the linear fit.
In the Figure 3.9; the near-zero p-value associated with the quadratic term suggests that it leads to an improved model.

Qualitative predictors

In the Figure 3.10 , we will examine the Carseats data, which is part of the ISLR library, and it includes qualitative predictors such as Shelveloc (an indicator of the quality of the shelving location). We will attempt to predict Sales (child car seat sales) in 400 locations based on a number of predictors.

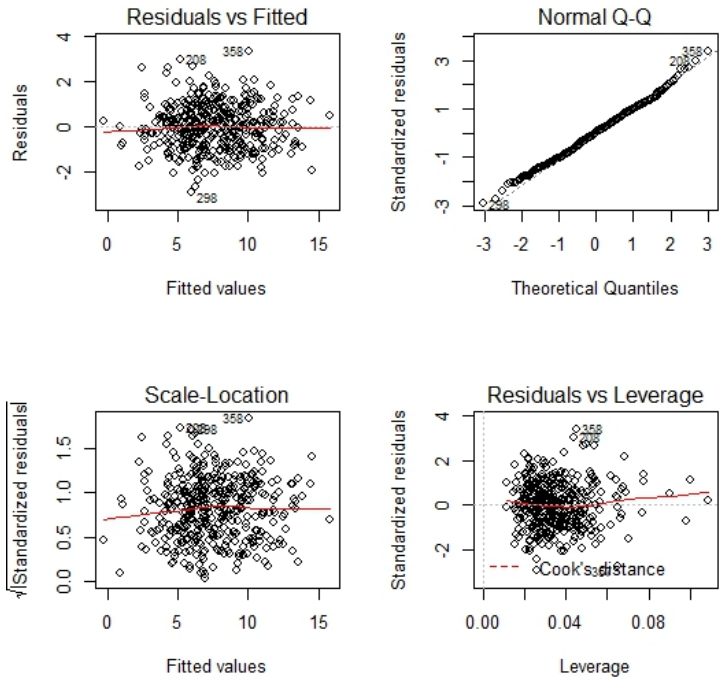


Figure 3.10: Qualitative prediction

3.3 *The regression of the density function by k nearest neighbors*

In this section, we want to give the simulation of the section(1.7.1).
we must first install the nor1mix package and loading her library to estimate the density function. It make a comparison of the true density function of a mixture with the estimate from the function fknn; using different kernel like : Cosine , Silvermen, uniform and Epanechnikov kernel, and with different values of k using the cross-validation method given in chapter one. Also, we will compare between the kernel and the k nearest neighbors methods.
The Figures 3.11,3.12,3.13,3.14,3.15,3.16,3.17 and 3.18 gives this results.

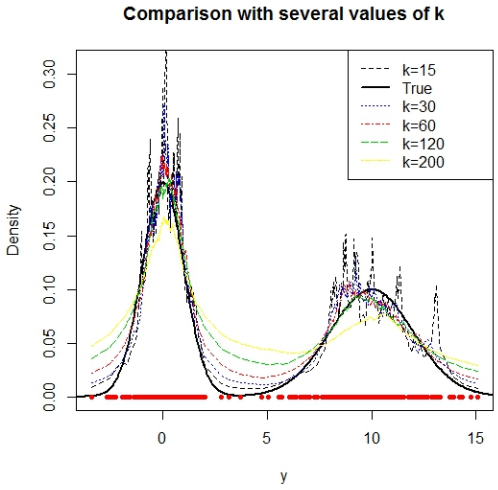


Figure 3.11

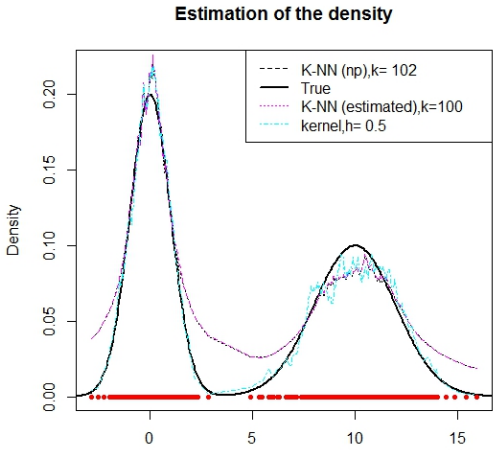


Figure 3.12

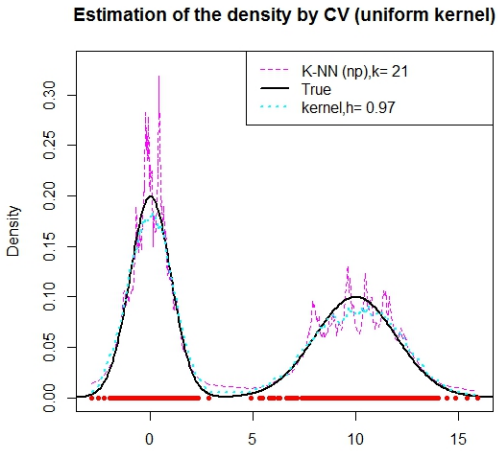


Figure 3.13

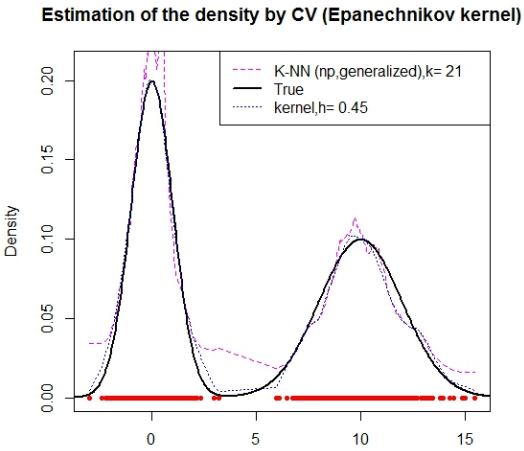


Figure 3.14

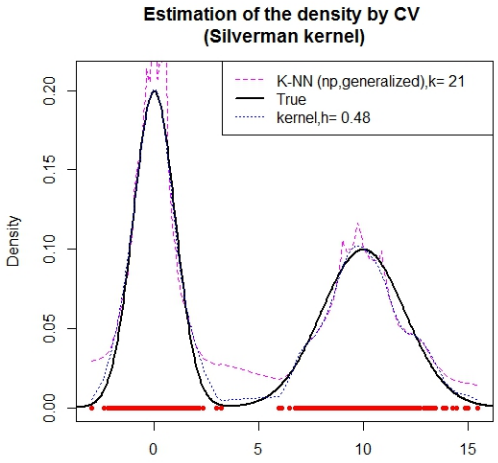


Figure 3.15

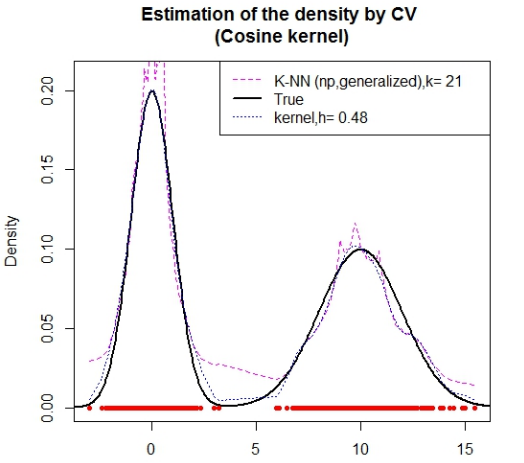


Figure 3.16

Figure 3.19 and Figure 3.20 : The next application concerns a comparison between the kernel and k-NN locally linear estimator by Cross-validation . Using the Prestige database for car library to studying the effect of the salary and the prestige of education. this given in Figure 3.19 and Figure 3.20.
Figure 3.21 : the same results given in the three dimensional space.

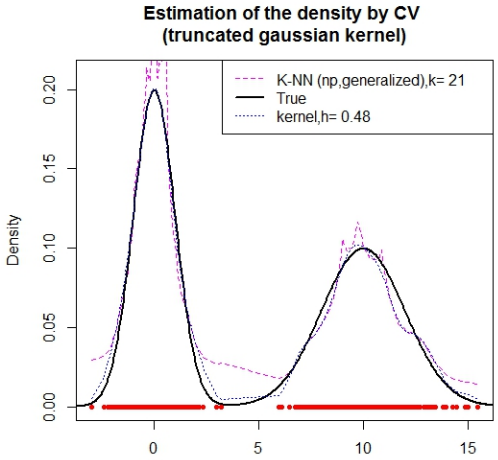


Figure 3.17

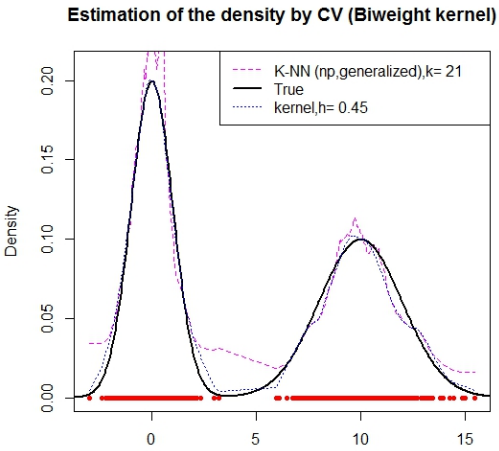


Figure 3.18

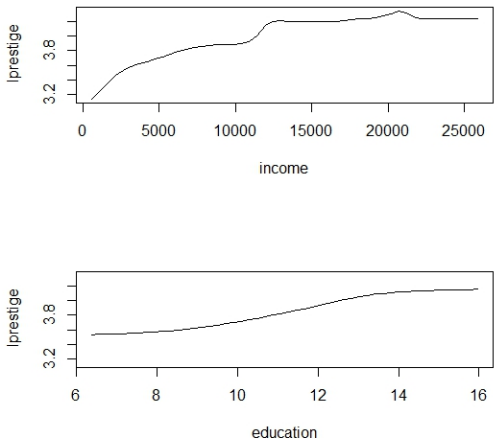


Figure 3.19

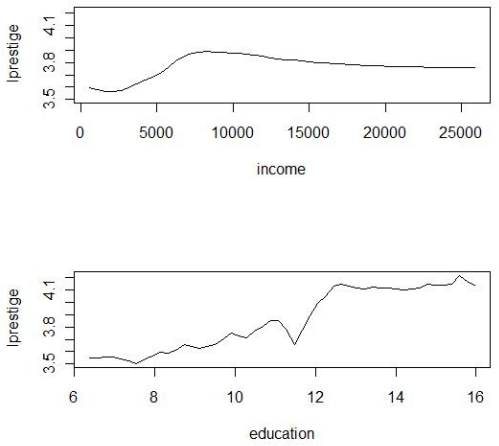


Figure 3.20

3.4 *Classification & Logistic regression*

The linear regression model discussed in the last section assumes that the response variable Y is quantitative. But in many situations; the response variable is instead qualitative. For example, eye color is qualitative, taking on values blue,

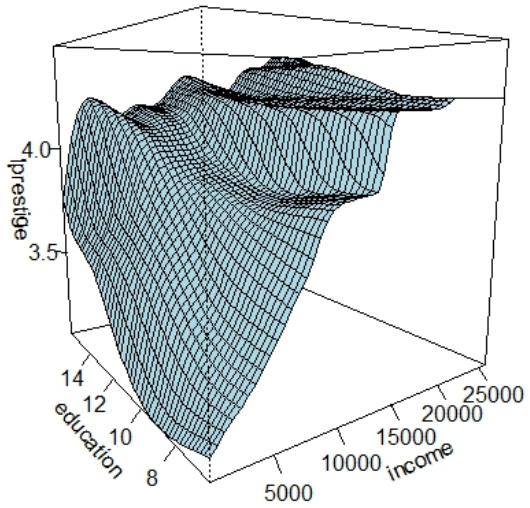


Figure 3.21

brown,..ect. Often qualitative variables are referred to as categorical. We will predict this qualitative responses, a process that is known as classification.Predicting a qualitative response for an observation can be referred to as classifying that observation. On the other hand,often the methods used for classification first predict the probability of each of the categories of a qualitative variable, as the basis for making the classification. In this sense; they also behave like regression methods.

In this section, we discuss one of the most widely-used classifiers: *Logistic regression and k nearest neighbors.*

An overview of classification

Classification problems occur often, perhaps even more so than regression problems. Some examples include:

1) A person arrives at the emergency room with a set of symptoms that could possibly be attributed to one of three medical conditions. Which of the three

conditions does the individual have?

2) An online banking service must be able to determine whether or not a transaction being performed on the site is fraudulent, on the basis of the user's IP address, past transaction history, and so forth.

3) On the basis of DNA sequence data for a number of patients with and without a given disease, a biologist would like to figure out which DNA mutations are deleterious (disease-causing) and which are not.

Now, we will illustrate the concept of classification using the simulated Default data set. e are interested in predicting whether an individual will default on his or her credit card payment, on the basis of annual income and monthly credit card balance.

3.4.1 *Logistic regression*

Consider again the Default data set, where the response default falls into one of two categories, Yes or Not. Rather than modeling this response Y directly, logistic regression models the probability that Y belongs to a particular category. For example:the probability of default given balance can be written as:
Pr(default=Yes/balance)
The values of Pr(default=Yes/balance), which we abbreviate P(balance), will range between 0 and 1. Then for any given value of balance, a prediction can be made for default.
For example,one might predict default=Yes for any individual for whom $p(balance) > 0.5$. Alternatively, if a company wishes to be conservative in predicting individuals who are at risk for default, then they may choose to use a lower threshold, such as $p(balance) > 0.1$.

The Logistic model

How should we model the relationship between $p(X) = Pr(Y = 1/X)$ and X? (For convenience we are using the generic 0/1 coding for the response).
We will talked of using a linear regression model to represent these probabilities:

$$p(X) = \beta_0 + \beta_1 X \tag{3.0}$$

If we use this approach to predict default=Yes using balance, we see the problem with this approach: for balances close to zero we predict a negative probability of

default; if we were to predict for very large balances, we would get values bigger than 1. These predictions are not sensible,since of course the true probability of default, regardless of credit card balance, must fall between 0 and 1.
This problem is not unique to the credit default data. Any time a straight line is fit to a binary response ;that is coded as 0 or 1, in principle; we can always predict $p(X) < 0$ for some values of X and $p(X) > 1$ for others (unless the range of X is limited).
To avoid this problem, we must model $p(X)$ using a function that gives outputs between 0 and 1 for all values of X. Many functions meets this description. In logistic regression , we use the logistic function

$$p(X) = \frac{e^{\beta_0+\beta_1 X}}{1+e^{\beta_0+\beta_1 X}} \tag{3.1}$$

To fit the model (3.1), we use a *maximum likelihood* method.

3.5 *A Comparison of Classification Methods*

In this section we give an example to make a comparison between the nearest neighbors method and the logistic regression model.

How the k nearest neighbors algorithm works in R ?

We will now perform kNN using the knn() function, which is part of the class library. This function works rather differently from the other model-fitting functions that we have encountered thus far. Rather than a two steps approach in which we first fit the model and then we use the model to make predictions, knn() forms predictions using a single command. The function requires four inputs.

1) A matrix containing the predictors associated with the training data, labeled train.X below.

2) A matrix containing the predictors associated with the data for which we wish to make predictions,labeled test.X below.

3) A vector containing the class labels for the training observations labeled train.Direction below.

4) A value for k, the number of nearest neighbors to be used by the classifier.

```
> Library(class)
> train.X=cbind(Lag1,Lag2)[train,]
```

```
> test.X=cbind(Lag1,Lag2)[!train,]
> train.Direction=Direction[train]
```

Now the knn() function can be used to predict the market's movement for the dates in 2005. We set a random seed before we apply knn() because if several observations are tied as nearest neighbors, then R will randomly break the tie. Therefore, a seed must be set in order to ensure reproducibility of results.

for k=1:

```
> set.seed(1)
> knn.pred=knn(train.X,test.X,train.Direction,k=1)
> table(knn.pred,Direction.2005)
```

for k=3:

```
> knn.pred=knn(train.X,test.X,train.Direction,k=3)
> table(knn.pred,Direction.2005)
> mean(knn.pred==Direction.2005)
```

The results have improved slightly. But increasing k further turns out to provides no further improvements. It appears that for this data, QDA provides the best results.

3.5.1 *An application to Caravan Insurance Data*

Finally, we will apply the kNN approach to the Caravan dataset, which is part of the ISLR library, and it includes 85 predictors that measure demographic characteristics for 5.822 individuals. The response variable is Purchase, which indicates whether or not a given individual purchases a caravan insurance policy.
In this dataset, only 6% of people purchased caravan insurance.

```
>dim(Caravan)
>attach(Caravan)
>summary(Purchase)
```

Because the kNN classifier predicts the class of a given test observation by identifying the observations that are nearest to it, the scale of the variables matters. Any variables that are on a large scale will have a much larger effect on the distance between the observations, and hence on the kNN classifier, than variables that are on a small scale. For instance, imagine a data set that contains two variables, salary and age(measured in dollars and years, respectively). As far as kNN is concerned,

a difference of $1,000 in salary is enormous compared to a difference of 50 years in age. Consequently, salary will drive the kNN classification results, and age will have almost no effect. This is contrary to our intuition that a salary difference of $1,000 is quite small compared to an age difference of 50 years. Furthermore, the importance of scale to the kNN classifier leads to another issue: if we measured salary in Japanese yen, or if we measure age in minutes, then we'd get quite different classification results from what we get if these two variables are measured in dollars and years.

A good way to handle this problem is to standardize the data, so that all variables are given a mean of zero and a standard deviation of one. Then all variables will be on a comparable scale. The scale() function does just this. In *standardizing* the data, we exclude column 86, because that is the qualitative Purchase variable.

```
> standardized.X=scale(Caravan[,-86])
```

Now, every column of standardized.X has a standard deviation of one and a mean of zero.

We now split the observations into a test set, containing the first 1,000 observations, and a training set containing the remaining observations. We fit a kNN model on the training data using $k = 1$, and evaluate its performance on the test data.

```
>test=1:1000
>train.X=standardized.X[-test,]
>test.X=standardized.X[test,]
> train.Y=Purchase[-test]
>test.Y=Purchase[test]
>set.seed(1)
>knn.pred=knn(train.X,test.X,train.Y,k=1)
> mean(test.Y!=knn.pred)
> mean(test.Y!="No")
```

The k-NN error rate on the 1,000 test observations is just under 12%. At first glance, this may appear to be fairly good. However, since only 6% of customers purchased insurance, we could get the error rate down to 6% by always predicting No regardless of the values of the predictors.

Suppose that there is some non-trivial cost to trying to sell insurance to a given individual. For instance, perhaps a salesperson must visit each potential customer. If the company tries to sell insurance to a random selection of customers, then the success rate will be only 6%, which many be far too low given the costs involved. Instead, the company would like to try to sell insurance only to customers who are

likely to buy it. So the overall error rate is not of interest. Instead, the fraction of individuals that are correctly predicted to buy insurance is of interest.

It turns out that k-NN which k=1 does far better than random guessing among the customers that are predicted to buy insurance. Among 77 such customers,9, or 11.7%, actually do Purchase insurance. This is double the rate that one would obtain from random guessing.
If we use k=3, the success rate increases to 19%, and with k=5 the rate is 26,7. This is over four times the rate that results from random guessing. It appears that k-NN is finding some real patterns in a difficult data set!

How the Logistic regression algorithm works in R ?

As a comparison, we can also fit a logistic regression model to the data. If we use 0.5 as the predicted probability cut-off for the classifier, then we have a problem: only seven of the test observations are predicted to Purchase insurance. Even worse, we are wrong about all of these! However we are not required to use a cut-off of 0,5. If we instead predict a Purchase exceeds 0,25, we get much better results: we predict that 33 people will Purchase insurance, and we are correct for about 33% of these people. This is over five times better than random guessing!

```
>glm.fit=glm(Purchase .,data=Caravan,family=binomial,subset=-test)
>glm.probs=predict(glm.fit,Caravan[test,],type="response")
>glm.pred=rep("No",1000)
>glm.pred[glm.probs>0.5]="Yes"
>table(glm.pred,test.Y)
>glm.pred=rep("No",1000)
>glm.pred[glm.probs>0.25]="Yes"
>table(glm.pred,test.Y)
```

3.6 *The usefulness of the k-NN method*

In this section, we propose to illustrate the effectiveness of the kNN method. We will show that the more heterogeneous the dataset is, the more the kNN method is able to capture this heterogeneity. We first detail a complete study in a simulated finite sample situation. This example will be voluntarily simple in order to clearly and educationally illustrate our purpose. Then, we present a concrete situation which deals with spectrometric curves.
We start by presenting the simulated datasets and we explain what homogeneity

and heterogeneity are. In the second paragraph we describe our study (regression operator, model, parameters, ect) an then we give various results and comments. Finally, we present the spectrometric dataset and give the results of prediction.

The simulated curves

Now, we will present the dataset we use in our study. To illustrate how the kNN method works, it is important to compare the results when data becomes more heterogeneous. Indeed, we saw previously that the local structure of the data has a major role in infinite-dimensional problems. More precisely, we talked about the importance of the concentration function. so, in the following, the notions of homogeneity and heterogeneity will refer to the concentration of the data. The idea here is to simulate a dataset which presents an homogeneous concentration and then to make it more and more heterogeneous by allowing the concentration of the data to differ noticeably from one location to another one. Now , we explain how to simulate this kind of dataset and we illustrate homogeneity in terms of concentration.
We simulate $n = 300$ pairs $(\mathcal{X}_i, Y_i)_{i=1,...,n}$ such that:

$$\mathcal{X}_i(t) \quad = \quad a_i \cos(2t)$$

Where t takes 100 values in $[0, \pi]$ and where

$$a_i \quad \sim \quad \mathcal{N}(0,1) \qquad \text{for} \quad i = 1, ..., 150$$

$$a_i \quad \sim \quad \mathcal{N}(3, \sigma^2) \qquad \text{for} \quad i = 151, ..., n.$$

We take different values for σ^2. This creates two groups of curves inside the dataset with concentration being different from one group to the other one. We detail the results for the two extreme cases the most homogeneous one when $\sigma^2 = 1$ and the most heterogeneous one when $\sigma^2 = 0.1$.
Note that in both cases we will take 250 curves to construct the testing sample and the other 50 will constitute the learning sample.
As we can see in Figure 3.22 and 3.23, it is not easy to see which dataset is more homogeneous or more heterogeneous and the second more heterogeneous, in terms of concentration of the data. To do this, we study the concentration function and more precisely distances between curves. We estimate the values of the concentration function by

$$\hat{\varphi}_{\mathcal{X}_i}(h) \quad = \quad \frac{1}{n} \sum_{j=1}^{n} \mathbb{1}_{B(\chi_i, h)}(\chi_j)$$

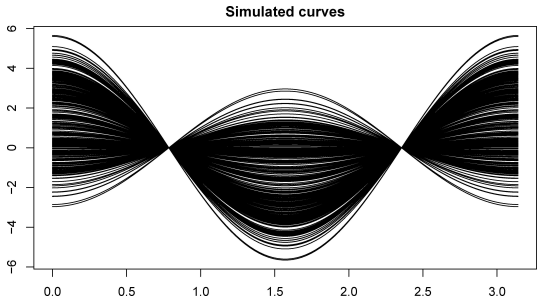


Figure 3.22: represents the most homogeneous case.

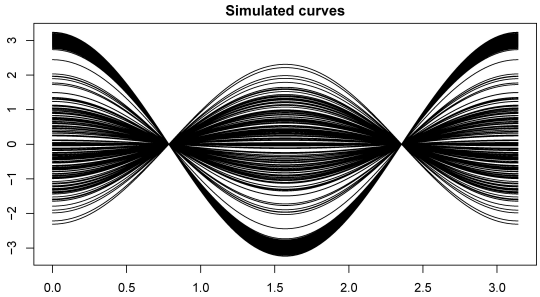


Figure 3.23: represents the most heterogeneous case.

Where $\mathcal{X} = \{\mathcal{X}_i, i = 1, ..., n\}$ is the dataset and h is a fixed bandwidth. We use here the standard L^2 semi-metric:

$$d(\mathcal{X}_i, \mathcal{X}_j) \quad = \quad \sqrt{\int (\mathcal{X}_i(t) - \mathcal{X}_j(t))^2 dt}$$

We represent in Figure 3.24 the values of $\hat{\varphi}_{\mathcal{X}_i}(h)(i = 1, ..., n)$ in our two extreme cases. For each case, the bandwidth h is the one obtained by a cross-validation procedure (see later). Other plots for other values of h looked similar and are therefore not presented here.
Figure 3.24 provides real evidence of the high difference between both datasets, in terms of concentration of the data. For the first datasets (left plot), each of the 300 curves has roughly the same number of neighbors(around 20% of the dataset in each ball of radius h). At the opposite(right plot), the second dataset shows high heterogeneity since both groups of curves have very different numbers of neighbors(20% for one group and 50% for the other one). Note that this kind of concentration plot allows us to have information which was hard to obtain directly from the simple plot of the curves like in Figure 3.22 and 3.23.

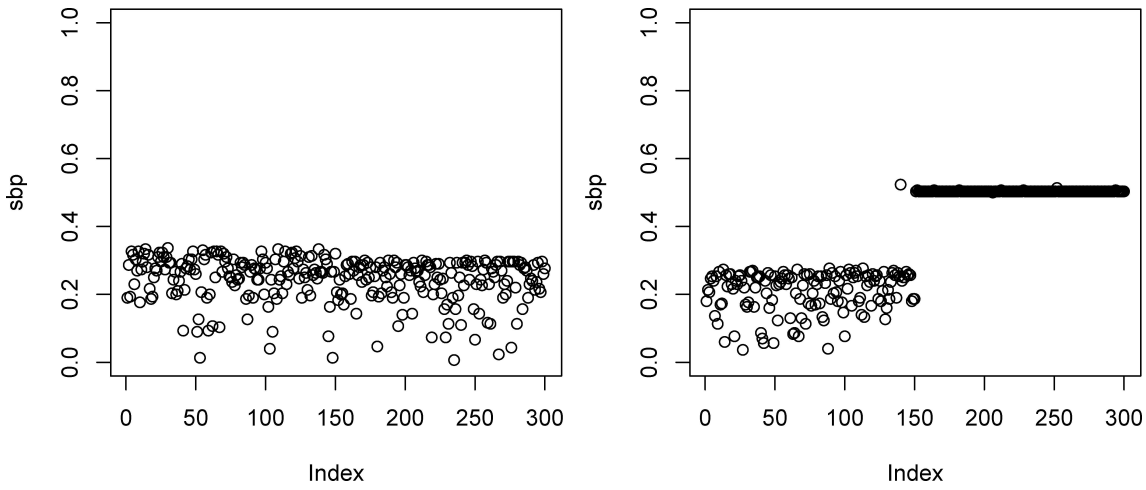


Figure 3.24: Values of $\hat{\varphi}_{\chi_i}(h)$ for the optimal h(left: the most homogeneous case,right: the most heterogeneous case)

3.6.1 *Description of the study*

Our aim is to study the following regression model:

$$Y_i \quad = \quad r(\mathcal{X}_i) + \epsilon_i$$

Where $\epsilon_i \sim \mathcal{N}(0, 0.05)$ and χ_i will be the dataset with $\sigma^2 = 1$ or $\sigma^2 = 0.1$. We will observe the behavior of the two methods(kNN method and kernel method) in these two extreme cases and in order to catch precisely how the kNN method works, we voluntarily choose a quite simple regression operator:

$$r(\mathcal{X}_i) \quad = \quad a_i^2$$

Now, we give a few important informations from a practical point of view. The R procedures used for predictions, we will use two of them to estimate the different regression operators: the first one is based on the kNN method and is called *funopare.knn.gcv*, the second one uses the traditional kernel method and is called *funopare.kernel.cv*. These two R procedures use global smoothing parameters selected by an automatic cross-validation type procedure. The optimal bandwidth h_{opt} of the kernel estimator is defined such that

$$h_{opt} \quad = \quad \arg\min_h CV(h)$$

Where

$$CV(h) = \sum_{i=1}^{n}(Y_i - \hat{r}^{(-i)}(\mathcal{X}_i))^2$$

With

$$\hat{r}^{(-i)}(x) = \frac{\sum_{j=1,j\neq i}^{n} Y_j K(d(\mathcal{X}_j, x)/h)}{\sum_{j=1,j\neq i}^{n} K(d(\mathcal{X}_j, x)/h)}$$

Whereas the optimal number of neighbors k_{opt} is defined by

$$k_{opt} = \arg\min_{k} CV(k)$$

where

$$CV(k) = \sum_{i=1}^{n}(Y_i - \hat{r}_{kNN}^{(-i)}(\mathcal{X}_i))^2$$

with

$$\hat{r}_{kNN}^{(-i)}(x) = \frac{\sum_{j=1,j\neq i}^{n} Y_j K(d(\mathcal{X}_j, x)/h_k(x))}{\sum_{j=1,j\neq i}^{n} K(d(\mathcal{X}_j, x)/h_k(x))}$$

for d, we use the standard L^2 semi-metric and the kernel is

$$K(u) = \frac{3}{4}(1-u^2)\mathbb{1}_{[0,1]}(u)$$

3.7 *Results prediction of mean square error for kernel and k-NN methods*

We present in this section the results of predictions. Note that, in the whole following. MSEP is the mean square error of prediction (i.e, the sum of square errors between predicted values and responses of the testing sample).We start by presenting the results for the two extreme cases ($\sigma^2 = 0.1, \sigma^2 = 1$). Figure 3.25 shows the prediction for the most homogeneous case, while Figure 3.26 concerns the most heterogeneous case.
While for the homogeneous dataset (Figure 3.25) both methods give good results, the kNN one is much more efficient in the heterogeneous situation with a MSEP of

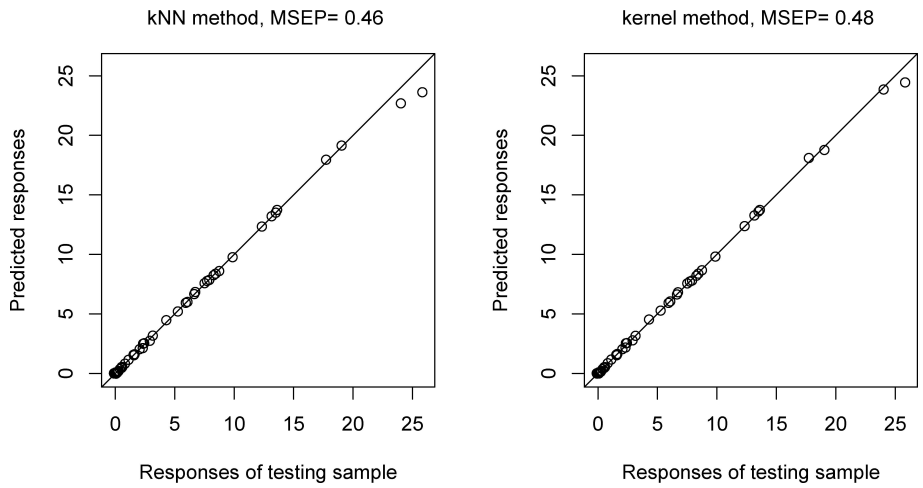


Figure 3.25: kNN method *vs* kernel method in the most homogeneous case ($\sigma^2 = 1$).

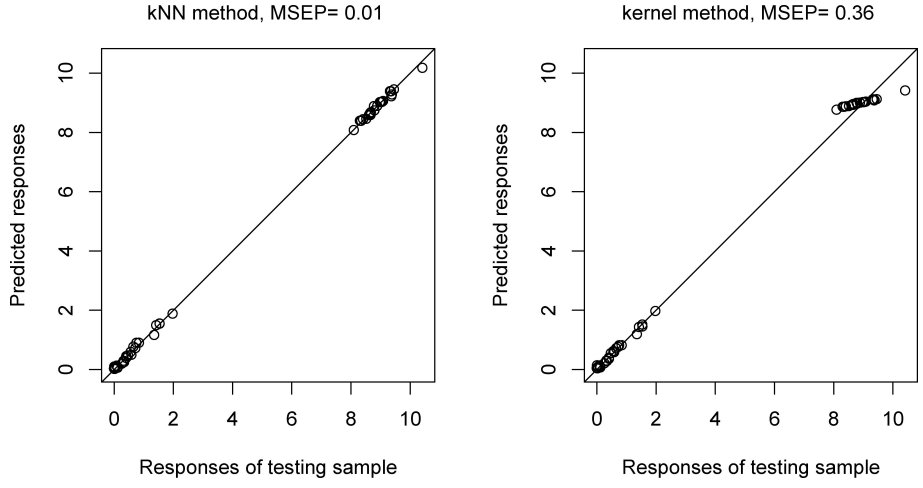


Figure 3.26: kNN method *vs* kernel method in the most heterogeneous case($\sigma^2 = 0.1$).

σ^2	1	0.7	0.5	0.3	0.1
ρ	0.958	0.836	0.253	0.072	0.027

Figure 3.27: k-NN method *vs* kernel method in the most heterogeneous case ($\sigma^2 = 0.1$).

0.01(against 0.36 for the kernel method).

Comparison between kernel and k-NN methods using MSEP

In the previous paragraph, we compared the two extreme cases, the most homogeneous one and the most heterogeneous one. We illustrate in this paragraph what happens when we take different values for σ^2 in order to see the effects of kNN method when the dataset become more and more heterogeneous. We study the behavior of the two methods with $\sigma^2 = 0.1, 0.3, 0.5, 0.7, 1$.
To make things clearer, we consider the following quantity:

$$\rho = \frac{MSEP(kNN)}{MSEP(kernel)}.$$

where MSEP(kNN) is the mean square error of prediction of the kNN method and MSEP(kernel) is the mean square error of prediction of kernel method. When ρ is close to 1, the two methods are equivalent and when ρ approaches to 0, the kNN method is better.
Now, we present in Figure 3.28 the values of ρ according to σ^2.

3.7.1 *Simulated example*

This study allows us to illustrate the real interest of the kNN method on finite sample situation. We see that, in cases where the concentration of the data is homogeneous, the two methods give Equivalent results. In the most heterogeneous case, we see in Figure 3.26 that k-NN method is better. On one hand, the MSEP is much smaller for kNN method and, on the other hand, the k-NN method is well adapted for the second group of curves. In fact, since the bandwidth is fixed in the kernel method, the predictions are precise when the data have an homogeneous concentration, but, in the second group of curves; where concentration is heterogeneous, the fixed bandwidth is not adapted because it does not create a neighborhood adapted to sparse data. This can be clearly seen in Figure 3.26.

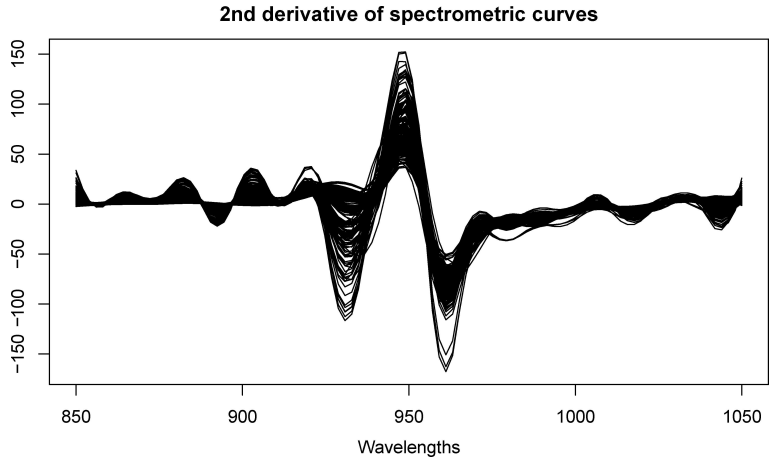


Figure 3.28: The spectrometric dataset.

Now, we can complete this practical section with a concrete example which really illustrates the effectiveness of the kNN method.

3.8 *A real dataset application*

Spectrometry is a modern and useful tool for analysing the chemical composition of any substance.
It provides many datasets which are useful for developing a functional nonparametric methodology. We focus here on a quality control problem in the food industry. The original data concerns a sample of finely chopped meat, Each curve represents the second derivative of the absorbance versus wavelength and the aim is to predict the fact content(Ferraty and Vieu (2006) [15]).These data are presented in Figure 3.28.
As we can see in Figure 3.27, the shape of the second derivative of the spectrometric curves reveals some peaks and valleys. This sample of size 215 was splitted into a learning sample of size 160; and a testing sample of size 55. The parameters k and d are the same as in the simulated example before.

In previous paragraphs, we pointed out the great importance of heterogeneity and homogeneity, so it is important in this concrete case to see what happens. Figure 3.29 displays the concentration function for each spectrometric curve.

We can clearly see in Figure 3.29 some heterogeneity in the structure of the spectrometric curves. Now, we can compare how the kNN method and traditional kernel

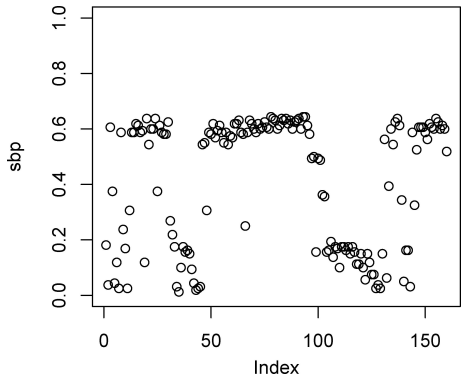


Figure 3.29: Values of $\hat{\varphi}_{\chi_i}(h)$ for the spectrometric dataset.

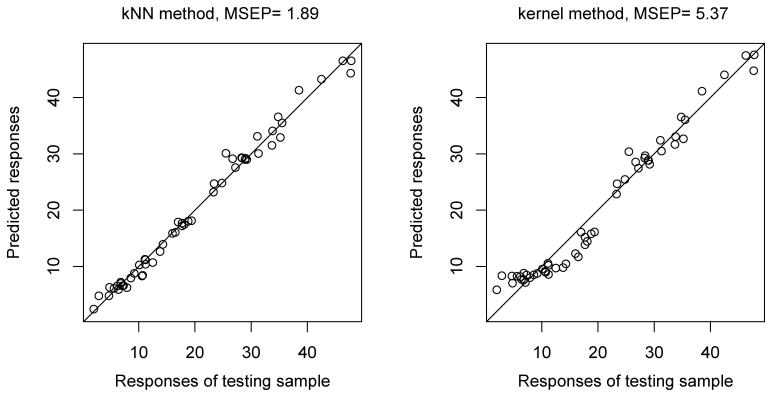


Figure 3.30: k-NN method *vs* kernel method for predicting fat content from spectrometric curves.

method work in this example.
Figure 3.30 shows the predicted values obtained by these two methods on the testing sample.
The results are very clear: in this situation where the data set has a very heterogeneous structure , the kNN method gives better predictions than kernel method.

3.9 *Conclusion*

We present in this work a few convergence results of the k-NN kernel estimator in nonparametric regression for the functional, real and vector data. Also we illustrate How the k-NN algorithm work in R and in our daily life.
For nonparametric estimation, we showed some of many asymptotic property that this method gives; the almost complete pointwise convergence of this estimator and we established its rate of convergence. We remarked that this rate is similar to the rate of convergence of Nadaraya-Watson type kernel estimator (optimal rates are the same). So, from a theoretical point of view, these two methods have the same asymptotical properties and we do not have any loss of effectiveness. This is in concordance with the knowledge in multivariate (unfunctional) nonparametric situations for which methods are known to achieve optimal rates of convergence [40], however, the infinite dimension of the data makes the use of the k-NN method more natural. The real interest of the k-NN method appears on practical examples. The fact that the smoothing parameter k takes its values in a discrete set makes things more simple from an implementation points of view. Moreover, we make example in finance to compare between the k-NN method and the logistic regression, and also we showed in examples that k-NN method takes into account the local structure of the data and gives better predictions when the data are heterogeneously concentrated.

Bibliography

[01] Alexender B, Tsybakov, *Introduction to nonparametric estimation*, Springer, 2009.

[02] Alpaydin.E ,*Voting Over Multiple Condensed Nearest Neighbors*, Artificial Intelligent Review 11:115-132, 1997.

[03] Arnaud Guyader, Nick Hengartner,*On the Mutual nearest neighbors estimate in regression*,France,2013.

[04] Attouch, M.; Benchikh, T.(2012).*Asymptotic distribution of robust k-nearest neighbour estimator for functional nonparametric models.* Mathematic Vesnic, 64,No. 4,pp. 275-285.

[05] Bailey.T, A. K. Jain,*A note on Distance weighted k-nearest neighbor rules*, IEEE Trans. Systems, Man Cybernatics, Vol.8, pp 311-313, 1978.

[06] Belabed Fatima Zohra, Attouch Medkadi,*Estimation non paramétrique fonctionnelle de la fonction de hasard conditionnelle par la méthode des k plus proches voisins (k-NN).*

[07] Billingsley.P, *Probability and Measure*. Third Edition. John Wiley and Sons, New York, 1995.

[08] Bosq, *Linear processes in function spaces: theory and applica- tions.* Springer Verlag. D. (2000).

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

[10] Burba.F, F.Ferraty, P. Vieu,*Convergence de l'estimateur noyau des k plus proches voisins en régression fonctionnelle non-paramétrique.* C. R. Math. Acad. Sci. Paris 346 (2008),5-6,339-342.

[11] Chidananda. K, Krishna.G , *The condensed nearest neighbor rule using the concept of mutual nearest neighbor*, IEEE Trans. Information Theory, Vol IT- 25 pp. 488-490, 1979.

[12] Collomb, *Estimation de la régression par la méthode des k points les plus proches avec noyau: quelques propriétés de convergence ponctuelle, in Statistique non Paramétrique Asymptotique*,ed. by J.-P. Raoult. Lecture Notes in Mathematics, vol. 821 (Springer, Berlin, 1980),pp. 159-175.

[13] Cover.T, Hart.P, *Nearest neighbor pattern classification.* IEEE Transactions on information theory, 13(1) :21-27, Jan 1967.

[14] Devroye. L.P, *The uniform convergence of nearest neighbour regression function estimators and their application in optimization,* IEEE Trans.Inform. Theory, 24.(1978).142-151.

[15] Ferraty.F, Vieu.P, *Nonparametric Functional Data Analysis: Theory and Practice, Springer Series in Statistics*, Springer, New York, 2006.

[16] Florent Burba, Frédéric Ferraty, philippe Vieu, *k Nearest neighbour method in functional nonparametric regression*, Toulouse, May 2009.

[17] Geoffrey W. Gates, *Reduced Nearest Neighbor Rule*, IEEE Trans Information Theory,Vol. 18 No. 3, pp 431-433.

[18] Gérard Biau, Luc Devroye, *Lectures on the nearest neighbor method*, Springer, February 2010.

[19] Gérard Biau, Luc Devroye, *A weighted k nearest neighbor density estimate for Geometric inference*, France, 2010.

[20] Gilles Koumou, *Méthode des k-plus proches voisins dans les approches non paramétrique*, 20 mai 2013.

[21] Heng Lian, *Convergence of functional k-nearest neighbor regression estimate with functional responses*, page 31-39, Vol.5, 2011.

[22] Jussi Klemelä, *multivariate nonparametric regression and visualization with R and applications to finance*, 2014.

[23] Li.Q, Racine.J.S, *Nonparametric Econometrics: Theory and Practice.* Princeton University Press, 2004.

[24] Li. R, GONG. G,*K-nearest-neighbour non-parametric estimation of regression functions in the presence of irrelevant variables.* Econometrics Journal, 11 :396-408, 2008.

[25] Liaw.Y. C, Leou. M. L,*Fast Exact k Nearest Neighbor Search using Orthogonal Search Tree*, Pattern Recognition 43 No. 6, pp 2351-2358.

[26] Luc Devroye, *Necessary and sufficient conditions for the pointwise convergence of nearest neighbor regression function estimates*,Springer, 1982.

[27] Mack. Y. P, *Local properties of k-NN regression estimates.* SIAM J.Algebr Discrete Methods,2. (1981). 311-323.

[28] Mack, M. Rosenblatt, *Multivariate k-nearest neighbor density estimates.* J. Multivar. Anal. 9,1-15 (1979).

[29] Mack, *Local properties of k-NN regression estimates.* SIAM J. Algorithms Discret. Meth. 2,311-323 (1981).

[30] Mcname.J , *Fast Nearest Neighbor Algorithm based on Principal Axis Search Tree*, IEEE Trans on Pattern Analysis and Machine Intelligence, Vol 23, pp 964-976.

[31] Mohammed kadi attouch, Tawfik Benchikh, *Asymptotic distribution of robust k-nearest neighbour estimator for functional nonparametric models*, December 2012.

[32] Moore. D.S, Yackel. J.W, *Consistency properties of nearest neighbor density function estimators.* Ann. Stat. 5, 143-154 (1977b)

[33] Nitin. Bhatia, vandana, *Survey of nearest neighbor Techniques*, India, vol.8.No.2, 2010. page 302.

[34] Ouyang.D , Li. D, Li. Q, *Cross-validation and non-parametric k nearest neighbour estimation.* Econometrics Journal, 9 :448-471, 2006.

[35] Parvin.H , Alizadeh. H, Minaei.B, *Modified k Nearest Neighbor*, Proceedings of the world congress on Engg. and computer science 2008.

[36] Ramsay. J, Silverman.B.W, *Functional Data Analysis, 2nd ed.* Springer, New York, 2005.

[37] Roussas. G, *Nonparametric estimation in mixing sequences of random variables.* J.Statist. Plann, 18(1989)135-149.

[38] Sproull. R. F,*Refinements to Nearest Neighbor Searching*, Technical Report, International Computer Science, ACM (18) 9, pp 507-517.

[39] Stone .C. J, *Consistent nonparametric regression,* Ann.Statist, 5(1977).595-645.

[40] Stone .C. J, *Optimal global rates of convergences for non parametric regression,*Ann. Statist. 10(1982), pp.1040-1053.

[41] Van der. A.W. Vaart, J.A. Wellner, *Weak Convergence and Empirical Processes. With Applications to Statistics.* Springer-Verlag, New York, 1996.

[42] Zeng.Y, Yang. Y, Zhou. L, *Pseudo Nearest Neighbor Rule for Pattern Recognition*, Expert Systems with Applications (36) pp 3587-3595, 2009.

[43] Zhou.Y, Zhang. C, *Tunable Nearest Neighbor Classifier*, DAGM 2004, LNCS 3175.