N° d'ordre :

REPUBLIQUE ALGERIENNE DEMOCRATIQUE & POPULAIRE

MINISTERE DE L'ENSEIGNEMENT SUPERIEUR & DE LA RECHERCHE SCIENTIFIQUE

UNIVERSITE DJILLALI LIABES
FACULTE DES SCIENCES EXACTES
SIDI BEL ABBÈS

THESE DE DOCTORAT

Présentée par: *Imane METMOUS*

Domaine : Mathématique informatique
Filière : Mathématique
Intitulé de la formation : Probabilité et Statistiques Appliquées

Intitulée

Estimation non paramétrique de quelques modèles statistique pour des données de type surrogate

Soutenue le ../../2022.
Devant le jury composé de :

Président :
Mr BENAISSA Samir — *Professeur à l'université de S.B.A.*

Examinateurs :
Mr GUENDOUZI Toufik — *Professeur à l'université de Saida.*
Mr BENCHIKH Tawfik — *Professeur à l'université de S.B.A.*

Directeur de thèse :
Mr ATTOUCH Mohammed Kadi — *Professeur à l'université de S.B.A.*

Co-Directeur de thèse :
Mr RIGHI Ali — *M.C.B . à l'université de S.B.A.*

Année universitaire : *2021/2022*

Abstract

In this thesis, we are interested in the non-parametric estimation of the distribution function of an incomplete scalar response variable, but with surrogate data, from a functional random variable. We started to construct an estimator of the regression operator by replacing the missing responses with surrogate data. We then established asymptotic properties of the constructed estimator, in terms of convergence in probability, almost complete and root mean square. Then, we applied the results obtained on simulated data.

After with the same method we study the asymptotic properties of the functional parameters in nonparametric statistics for incomplete data of the expectile regression function. More precisely, we are interested in the expectile regression for which we construct an estimator and we study the asymptotic behavior in the functional data model.

We first study the asymptotic properties of a nonparametric estimator of the expectile regression given a functional explanatory variable, when the response is scalar, in the i.i.d. case. We establish the almost complete uniform convergence and the asymptotic normality of these estimators. A simulation study and a real data application are performed to illustrate how this leads to better predictive performance than estimates obtained with quantile regression or classical regression.

Finally, a comparative study on simulated data and real data was carried out in order to highlight the quality of the estimation offered by the expectile regression in comparison with the classical regression models, i.e. the natural regression and the conditional quantile regression. Thus, an R program was developed to confirm the theoretical result obtained.

Key words : Functional Data Analysis (FDA); Conditional distribution function;

Nonparametrickernel estimation ; Surrogate dat ;Expectile regression ; ; Functional data analysis ; almost complete convergence ; asymptotic normality

Résumé

Dans cette thèse nous intéressant à l'estimation non paramétrique de la fonction de répartition conditionnel d'une variable de réponse scalaire incomplète, mais ayant des valeurs de substitution (surrogate data), à partir d'une variable aléatoire fonctionnelle. Nous avons commencé à construire un estimateur de l'opérateur de régression en remplaçant les réponses manquantes par les données de substitution. Nous avons établi ensuite des propriétés asymptotiques de l'estimateur construit, en termes de convergence en probabilité,presque complète et en moyenne quadratique. Puis, nous avons appliquées résultats obtenus sur des données simulées
Aprés avec la même méthode nous avons étudier les propriétés asymptotiques des paramètre fonctionnels en statististique non paramétrique pour des données incomplètes dela fonction de régression par expectile. Plus précisement, nous nous intéressons à la regression expectile pour laquelle nous construisons un estimateur et nous étudions le comportement asymptotique dans le modèle de données fonctionnelles.

Nous avons d'abord étudié les propriétés asymptotiques d'un estimateur non paramétrique de la régression expectile étant donné une variable explicative fonctionnelle, lorsque la réponse est scalaire, dans le cas i.i.d. Nous établissons la convergence uniforme presque complet et la normalité asymptotique de ces estimateurs. Une étude de simulation et une application de données réelles sont réalisées pour illustrer comment cela permet d'obtenir des performances prédictives supérieures à celles obtenues avec des estimations par la régression quantile ou la régression classique.

Enfin, une étude comparative sur des données simulées et des données réelles ont été traiter afin de mettre en évidence la qualité de l'estimation qu'offre la régression par expectile en comparaison avec les mdoèles classiques de la régression à svoir la régression

naturelle et la régression par quantile conditionnel. Ainsi, un programme R a été élaborer afin de de confirmer le résultat théorique obtenu.

Mots- clés : RFunctional Data Analysis (FDA) ; Conditional distribution function ; Nonparametrickernel estimation ; Surrogate dat ;Expectile regression ; ; Functional data analysis ; almost complete convergence ; asymptotic normality.

List of works

Articles published in international journals

1. I.METMOUS, M.K.ATTOUCH, B.MECHAB and T.MEROUAN. Nonparametric Estimation of the Conditional Distribution Function For Surrogate Data by the Regression Model. Applications and Applied Mathematics :An International Journal (AAM), 1932-9466, DOI : https://digitalcommons.pvamu.edu/aam/vol16/iss1/4.

List of communications

1. I.METMOUS. Asymptotic properties of robust nonparametric regression for a functional regressor.Journée Académique Mathématiques Appliquées (JAMA'19) 10 Avril 2019
2. I.METMOUS. Asymptotic study ofrobust nonparametric regression with functional ragrassor. RAMA11'2019. Sidi Bel Abbès, 21-24 Nouvembre 2019.
3. I.METMOUS. The robust nonparametric regression for functional models. JDFSE'2019. Sidi Bel Abbès, 14-15 Décembre 2019.

TABLE DES MATIÈRES

CHAPITRE 1

INTRODUCTION

1.1 L'estimation non paramétriques

Since the beginning of human civilization, the statistic has been used to obtain information about population numbers, warfare, agricultural production ...(etc), and that's back to the nature of the human being who is always trie to understand and explain the phenomena and anticipate their results, this is why many researchers are interested in statistics, of which the main purpose is to provide an analysis, or a description, of a past phenomenon, and to predict a future phenomenon of a similar nature. historically the first result appeared at the beginning of the 17th century (GALILEE and Jérôme CARDAN), after that, many results were published and several models are found to express the relationship between the variables. nowadays, technological advances high-dimensional techniques are more powerful, and the amount of information collected is increasing, that's why the functional statistics became a field of topical research, at once diversified by its fundamental aspects and by the various fields it covers.

Historically, the statistical analysis of functional data goes back to the 1960s when several studies, in different scientific disciplines, were interested in data in the form of curves. The first works can be attributed to meteorologists and chemists we citeObukhov(1960), Holmström(1963) in climatology, Deville(1974) in Econometrics, and Kirkpatrick and Heckman (1989) in genetics). Furthermore, the amount of work on parametric regression models is very large. In fact, in functional statistics, the first innovative results of the statistical analysis of functional data are due to the contribution of Ramsay and Silverman (1997). This monograph allowed statisticians to have a global vision of the treatment of functional data in terms of regression techniques, statistical discrimination, and factor analysis.More generally, in 2002, the same authors delivered a second book

focusing on the applied aspects of this branch of statistics. We can also mention the works of Bosq (2000) concerning dependent linear functional models such as autoregression models, then in 2005, he focused on prediction in high dimensional models. Thanks to the contribution of Ferraty and Vieu which can be considered decisive in the nonparametric functional framework. These authors have studied the asymptotic properties of several non-parametric models such as the regression operator, the regression function, the conditional distribution function, the conditional density and its derivative, the conditional mode and the conditional quantiles) when the explanatory variable takes these values in an infinie dimension space and when the response variable is scalar. These authors established the almost sure convergence of these models with the kernel method. Then, Ferraty et al. (2007) have studies quadratic mean convergence, also asymptotic normality of the regression operator. Also, Laksaci (2007) gave the quadratic error of the conditional density. Ezzahrioui and Ould-Said (2008a, 2008b) showed normality asymptotic mode and conditional quantile in i.i.d and dependent cases, Dabo-Niang and Laksaci (2007) included normality convergence results for conditional mode kernel estimators and conditional quantiles (if i.i.d.). Ferraty et al. (2008) study the almost complet convergence of conditional hasard functions.

The main idea of statistic is to study of the relation between two random variables is a very important problem. One of the most frequently encountered models in non-parametric statistics is the conditional distribution function. The importance of this function comes from the fact that most of the statistical quantities used in practice to understand the link between an explanatory variable X and a response variable Y .On the other hand, both the regression function and the conditional mode have the disadvantage of giving only a predictive value without informing us about the probability that the realization is close to it (without de the distribution of the conditional variable).The conditional distribution function, on the other hand, gives us the probability that the variable of interest belongs to a given interval, while the conditional quantiles allow the construction of prediction intervals, and consequently, they both provide very useful additional information on the behavior of the variable of interest.From a historical point of view The nonparametric estimation of the conditional distribution function when the explanatory variable has values in a space of finite dimension was introduced by Ferraty et al(2006). These authors have constructed a double kernel estimator for the conditional distribution function and have specified the speed of convergence of this estimator when the observations are independent and identically distributed.

On the other hand, several authors have studied the estimation of the conditional distribution function as a preliminary study of the estimation of conditional quantiles. Laksaci et al. (2008) proposed a method for estimating conditional quantiles and were he establish

almost-complete convergence and asymptotic normality of their estimator when observations are functional i.i.d.. In the statistical analysis of functional data, the estimation of the conditional density and its derivatives was introduced by Ferraty et vieu (2006). These authors obtained almost complete convergence in the case where the observations are independent and identically distributed(i.i.d). This work can be considered as the starting point of abundant literature that has developed on the estimation of the conditional density and its derivatives, notably in order to use it to estimate the conditional mode. Indeed, by considering -mixing observations, Ferraty et al(2005).established the almost complete convergence of a kernel estimator of the conditional mode defined by the random variable maximizing the conditional density, in 2010, Laksaci et al. answered the question of the choice of the smoothing parameter in the estimation of the conditional density when the explanatory variable is functional.

1.2 Nonparametric estimation of expectile regression

Recently, the risk measures such as the Expectile and the quantile have become the subject of renewed attention in actuarial and financial risk management. The Value-at-Risk (VaR) and the Expected Shortfall are the most commonly used tools in financial risk analysis (ES). These statistical models, on the other hand, have some flaws, such as non-subadditivity and insensitivity to the severity of extreme losses, see, for instance, Bellini, Negri, and Pyatkova (2019) for the definitions of these properties. We can define the expectile and quantile as solutions to minimization. Unlike quantiles, expectiles are determined by extreme expectations rather than by extreme probabilities and define a consistent risk measure. Thanks to Newey and Powell (1987) who is introduce of the expectile, that has received less attention than the quantile. The expectile is a generalization of the mean, whereas the quantile is a generalization of the median. Furthermore, the quantile is based on the absolute loss function, whereas the expectile is based on the quadratic loss function. The expectile model becomes increasingly popular in the financial literature, we refer, among many others, to Pratesi, Ranalli, and Salvati (2009), Waltrup,Sobotka, Kneib, and Kauermann (2015), Bellini and Bernardino (2017), Farooq and Stein wart (2019), since it is the unique elicitable coherent risk measures, we may refer to Bellini et al. (2019) and the references therein. For the use of the expectile regression in the heteroscedasticity analysis, one can refer to Gu and Zou (2016) and Zhao, Chen, and Zhang (2018). For more motivation on the use of the expectile model, we refer to the recent paper by Daouia, Girard, and Stupfler (2018, 2020). For an overview of the use of expectile curves in regression analysis, we refer to Kneib (2013) and the extensive discussions to that paper, in particular by Eilers (2013) for an appraisal of expectiles and

by Koenker (2013) for a critical viewpoint. Quantiles and expectiles, which contain information about the full distribution for a random variable, are extensions of median and mean,respectively. Expectiles are excellent alternatives to quantiles in different aspects for relevant applications. Motivating advantages are that expectiles are more alert than quantiles to the magnitude of infrequent catastrophic losses, and they depend on both the tail realizations of the predictor and their probability, whereas quantiles depend only on the frequency of tail realizations, see Kuan, Yeh, and Hsu (2009). This high sensitivity of expectiles to tail behaviour allows for more prudent and reactive risk management. Notice that the quantiles are not always satisfactory and can be criticised for being somewhat difficult to compute as the corresponding loss function is not continuously differentiable. The key advantage of the expectile over the quantile is its efficiency and computing experience, although it has not a direct interpretation as the quantile in terms of the relative frequency, see Daouia, Gijbels, and Stupfler (2019). Although they present differences in their construction, both quantiles and expectiles share similar properties. The main reason for this, as shown in Jones (1994), is the fact that expectiles are precisely quantiles but for a transformation of the original distribution. Abdous and Rémillard (1995) established an important feature that the quantiles and expectiles of the same distribution coincide under the hypothesis of weighted symmetry and pointed out that inference on expectiles is much easier than inference on quantiles.

Zhang (1994) introduced the nonparametric estimation of the expectile regression and proved the consistency and the asymptotic normality in the finite-dimensional framework. An increasing interest has been given to regression models in which the response variable is real-valued and the explanatory variable takes the form of smooth functions that vary randomly between repeated observations or measurements. Despite this importance, the expectile regression is, in comparison to both competitive regressions (conditional mean and quantile), relatively unexplored and still in full development and our aim is to fill this gap. The first results in this direction are obtained by Mohammedi, Bouzebda, and Laksaci (2021). They investigated the nonparametric expectile regression in the case of a functional predictor and a scalar response and obtained the asymptotic properties of the kernel expectile regression estimator in the i.i.d. setting. The main purpose of the present work is to extend the previous works to serially dependent observations to cope with the case when the return or the loss are linked to the functional time series valued exogenous covariates which could prevent from the problem of the ‘curse of infinite dimensionality’.Indeed, as discussed in Geenens (2011), the nonparametric functional time series analysis repose on the use of semi-metric as proximity measures in infinite-dimensional functional space, which is often presented as a technical tool for dimension reduction purposes. This is one of the principal motivations behind the development of the nonpara-

metric approach in functional statistics.More motivations on the importance of this topic of nonparametric functional statistics can be found in the precursor monograph (Ferraty and Vieu 2006) or in the comprehensive survey paper (Ling and Vieu 2018). Generally, it is well known that the statistical study of functional dependent data originates from the monograph (Bosq 2000) and the literature has flourished since this cited work. In particular, the first study on Nonparametric Functional Time Series Analysis (NFTSA) was investigated by Ferraty, Goia, and Vieu (2002) where the almost complete consistency of the kernel estimator of the regression operator is established. Masry (2005) considered the regression estimation using a kernel-type estimator in a functional setting with mixing samples and established the asymptotic normality. Laib and Louani (2010) stated the asymptotic normality of the regression operator when the functional time series data has an ergodic structure. Based on the ergodicity assumption, Ling and Liu (2017) studied the large-sample properties of the kernel estimator for a nonparametric regression operator when the response variable is randomly censored. Recently, Ling, Wang, and Vieu (2020) have investigated the properties of kernel estimators in a functional regression model when both response and covariate are functional. The authors proved pointwise and uniform almost complete convergence of the examined estimator. Ling, Meng, and Vieu (2019) investigated the k-nearest neighbours(kNN) estimator of a nonparametric regression model for functional time series data and established the uniform almost complete convergence rate of the kNN estimator. Chowd-hury and Chaudhuri (2020) considered a nonparametric regression setup with a particular focus on a parameter of interest associated with the conditional distribution. The authors derived the optimum convergence rate for the kernel estimate of the parameter in this setup. Other NFTSA investigations using alternative nonparametric models can be found in Ezzahrioui and Ould Saïd (2010) and Bouzebda, Chaouch, and Laïb (2016) (for the conditional mode), Quintela-del Río (2008) (for the hazard function), Chen, Ling, Ling, and Liu (2019) (for the M-regression), and Ling, Cheng, Vieu, and Ding (2021) (for the single index model). The functional time series analyses by quantile regression are active and relevant fields of investigation which are very close to the expectile regression model. For example, Ezzahrioui and Ould-Saïd (2008) have established the asymptotic normality for a kernel estimator of the conditional quantile obtained by inverting the double kernel estimator of the cumulative distribution function. Using the L 1 approach, Laksaci, Lemdani, and Ould Saïd (2011) have constructed an alternative estimator of the quantile regression andestablished its asymptotic properties under the strong mixing assumption. We cite Bellini and Bernardino (2017) for the nonparametric estimation of the conditional quantile for the ergodic functional data. Other forms of quantile regressions have been proposed in functional statistics literature. We cite, for instance, Ding, Lu, Zhang, and Zhang (2018) (for semi-functional partial li-

near model), Sang and Cao (2020) (for single index model), and Zhang, Lian, Guodong, and Zhu (2021) (for additive model). For recent advances and trends in FDA, we refer to some survey papers and journal special issues, such as Cuevas (2014), Goia and Vieu (2016), Ling and Vieu (2018), and Aneiros, Cao, Fraiman, Genest, and Vieu (2019).

In this paper, we investigate an alternative way based on the least asymmetrically weighted squares estimation, borrowed from the econometrics literature, that is one of the basic tools in statistical applications. This method often involves Newey and Powell (1987) concept of expectiles, the least-squares analogue of the traditional quantiles. They were so named because they resemble the quantiles of a random variable, but unlike them, they are based on a quadratic loss function, as the case of the expectation.
The advantages of expectile regression are the same as those of least square regression and conditional quantile regression :

1. Expectile regression is a very useful and strong method for exploring the relationship between random variables.
2. Expectation regression estimator depends on the form of the whole distribution, Therefore, the expectations regression estimator contains additional information on the magnitude of the tail distribution.
3. The expectile has a risk measure sensitive to the extreme values which is beneficial is some situations such as survival analysis, insurance, economics or finance.
4. The expectile regression is the only consistent and elicitable risk model, that's why is used as an alternative estimators for both known risk measures such as Conditional Value at Risk (CVaR) or the Conditional Expected Shortfall (CES).

1.3 Incomplete data

Missing data is one of the common traps in statistical issues, the feature of these models is the existence of incomplete observations, for which the variable of interest is not completely observed for all sample data,this data effect on the rigor and strong biases in the analysis models which impact on the performance of estimators. Historically, after the paper of Yates (1933) who formulates the idea of substituting least-square estimates for the missing values, the analysis of the missing data was expanded. Along with this idea of imputing missing values by least-square predictions, Cochran (1968) uses it to reduce bias in observational studies, and Afifi and Elashoff (1969) provide a large sample theoretical analysis. After that many studies interset on the regression models with missing data.
In what follows we present the three categories of the incomplete data :

1. MCAR (missing completely at random) : Missing data completely at random means that the missing value is not linked by observed and unobserved observations but it can be associated with observed covariates(so if the probability of absence is a constant for all observations), in other words, the absence does not relate to the variables observed in the analytical model.
2. MAR (Missing at random) :Missing observations in the data are independent of the missing variables themselves, but possibly dependent on other observed variables. The The data is missing at random if the probability of an observation being incomplete depends only on the other observed values.
3. MNAR (Missing not at random) :the missing value can depend on both the observed values and the missing values of the variable itself, as well as other variables in the analytical model. The data is missing not at random (MNAR), if the probability of absence depends on the variable in question.

there are several methods to treat the incomplete data

1.3.1 Surrogate data

There are several methods to treat the incomplete data, for example remove rows with missing values, mean-based imputations, regression-based imputations, k-nearest-neighbor (kNN) method, hot-deck and cold-deck imputation, and maximization methods likelihood...
In this thesis, we will investigate the surrogate data. In various fields, we can't observe all the the response variable, we can use substitution data to replaces this missing data.Surrogate or analogous data can refer to data used to supplement the available data from which a mathematical model is constructed. According to this definition, it can be generated (i.e., built from synthetic data) or transformed from another source. Surrogate data can be used for statistical forecasting. Data from similar series can be aggregated to improve forecast accuracy.The use of surrogate data can allow a model to account for trends that are not apparent in historical data.

1.3.2 Definitions and estimators

1.3.3 The conditional cumulative distribution function

Let $(X_i, Y_i)_{i=1,\dots,N}$ be a random variables independent and identically distributed as (X, Y), where $X \in \mathcal{F}$, Y take values in $\mathbb{R}$ and $(\mathcal{F}, d)$ is a semi metric space with a metric $d(.,.)$. The conditional cumulative distribution function of Y given $X = x$, denoted by

$F^x(.)$ is defined

$$F^x(.) = \mathbb{P}(Y \leq y | X = x), \quad \forall y \in \mathbb{R},$$

and by the regression model, we have

$$\mathbb{E}\left[H\left(\frac{y - Y_i}{h_H}\right) \Big| X_i = x \right] \xrightarrow{h_H \to 0} F^x(y),$$

where $H(.)$ is a cumulative distribution function and h_H is a sequence of positive real numbers tending to 0 when n go to infinity.

The estimator of conditional distribution function by the kernel method defined by

$$\widehat{F}_C^x(y) = \frac{\sum_{i=1}^N H\left(\frac{y-Y_i}{h_H}\right) K\left(\frac{d(x,X_i)}{h_K}\right)}{\sum_{i=1}^N K\left(\frac{d(x,X_i)}{h_K}\right)}, \quad \forall y \in \mathbb{R}, \ \forall x \in \mathcal{F},$$

where $K(.)$ is a kernel function and h_K is a bandwidth sequence tend toward 0.

The estimator of conditional distribution function with surrogate data defined by

$$\widehat{F}^x(y) = \frac{\sum_{i \in V} H\left(\frac{y-Y_i}{h_H}\right) K\left(\frac{d(x,X_i)}{h_K}\right) + \sum_{j \in \bar{V}} u(X_j, \widetilde{Y}_j) K\left(\frac{d(x,X_j)}{h_K}\right)}{\sum_{i=1}^N K\left(\frac{d(x,X_i)}{h_K}\right)},$$

where

$$u(X_j, \widetilde{Y}_j) = \mathbb{E}\left[H\left(\frac{y - Y_j}{h_H}\right) \Big| X_j, \widetilde{Y}_j \right]$$

and the function $u(.,.)$ is unknown.

So, we estimate this function by validation data set :

$$\widehat{u}(X_j, \widetilde{Y}_j) = \frac{\sum_{i \in V} H\left(\frac{y-Y_i}{h_H}\right) W\left(\frac{d(X_j,X_i)}{a_n}, \frac{\widetilde{Y}_j - \widetilde{Y}_i}{a_n}\right)}{\sum_{i \in V} W\left(\frac{d(X_j,X_i)}{a_n}, \frac{\widetilde{Y}_j - \widetilde{Y}_i}{a_n}\right)}$$

and $W(.,.)$ is the two-dimensional kernel function in $\mathcal{F} \times \mathbb{R}$ and a_n is a sequence of real number which tend to zero when n tend to infinity.

1.3.4 The conditional expectile

the conditional expectile of order p has been introduced by Newey and Powell (1987) as the minimizer of an asymmetric quadratic loss

$$\xi_p(x) = \underset{t \in \mathbb{R}}{\operatorname{argmin}}\, \delta_p(x, t),$$

where

$$\delta_p(x,t) = \mathbb{E}[p(Y-t)^2\mathbf{1}_{(Y-t)>0}{}'X = x] + \mathbb{E}[(1-p)(Y-t)^2\mathbf{1}_{(Y-t)\leqslant 0}{}'X = x],$$

where $\mathbf{1}_Z$ is the indicator function of set Z, we can show that ξ_p is the solution of

$$\frac{p}{1-p} = \frac{A_1(x,t)}{A_2(x,t)}$$

where

$$\begin{cases} A_1(x,t) = -\mathbb{E}[(Y-t)\mathbf{1}_{(Y-t)\leqslant 0}{}'X = x], \\ A_2(x,t) = \mathbb{E}(Y-t)\mathbf{1}_{(Y-t)>0}{}'X = x]. \end{cases} \tag{1.1}$$

We use the fact that the function $A(x,t) := \frac{A_1(x,t)}{A_2(x,t)}$ is an increasing function so we can express the conditional expectileξ_p of order p as follows :

$$\xi_p = \inf\{t \in \mathbb{R} : A(x,t) \geqslant \frac{p}{1-p}\}.$$

To build an estimator of the conditional expectile of order p when there are missing data in the response variable, let N and n ($n < N$) the respetive sizes of the sample set and the validation set, we assume that the observations are independent and identically distributed, V is the index set of individuals in the sampled validation set and $\bar{V} = \{1,2,\ldots,N\} - V$. Since

$$\begin{cases} \mathbb{E}[\mathbb{E}\{(Y_j-t)\mathbf{1}_{(Y_j-t)\leqslant 0}{}'X_j, \widetilde{Y}_j\}] = \mathbb{E}[(Y_j-t)\mathbf{1}_{(Y_j-t)\leqslant 0}{}'X_j = x] = A_1(x,t) \\ \mathbb{E}[\mathbb{E}\{(Y_j-t)\mathbf{1}_{(Y_j-t)>0}{}'X_j, \widetilde{Y}_j\}] = \mathbb{E}[(Y_j-t)\mathbf{1}_{(Y_j-t)>0}{}'X_j = x] = A_2(x,t) \end{cases}$$

where $\widetilde{Y}$ is a surrogate variable of Y.
we can estimate ξ_p by

$$\widehat{\xi}_p(x) = \inf\{t \in \mathbb{R} : \widehat{A}(x,t) \geqslant \frac{p}{1-p}\}$$

we define

$$\widehat{A}(x,t) := \frac{\widehat{A}_1(x,t)}{\widehat{A}_2(x,t)}$$

with

$$\begin{cases} \widehat{A}_1(x,t) = \frac{\sum_{i\in V} K(h_K^{-1}d(x,X_i))(Y_i-t)\mathbf{1}_{(Y_i-t)\leqslant 0} + \sum_{j\in\bar{V}} u_1(X_j,Y_j)K(h_K^{-1}d(x,X_i))}{\sum_{i=1}^N K(h_K^{-1}d(x,X_i))} \\ \widehat{A}_2(x,t) = \frac{\sum_{i\in V} K(h_K^{-1}d(x,X_i))(Y_i-t)\mathbf{1}_{(Y_i-t)>0} + \sum_{j\in\bar{V}} u_2(X_j,Y_j)K(h_K^{-1}d(x,X_i))}{\sum_{i=1}^N K(h_K^{-1}d(x,X_i))} \end{cases}$$

where

$$\begin{cases} u_1(X_j,\widetilde{Y}_j) = \mathbb{E}\{(Y_j-t)\mathbf{1}_{(Y_j-t)\leqslant 0}{}',X_j,\widetilde{Y}_j\} \\ u_2(X_j,\widetilde{Y}_j) = \mathbb{E}\{(Y_j-t)\mathbf{1}_{(Y_j-t)>0}{}',X_j,\widetilde{Y}_j\} \end{cases}$$

for $j \in \bar{V}$

Recall that the functions $u_1(.,.)$ and $u_2(.,.)$ are unknown,so we estimate those functions by validation data set :

$$\begin{cases} \widehat{u}_1(X_j,\widetilde{Y}_j) = \frac{\sum_{i\in V}(Y_i-t)\mathbf{1}_{(Y_i-t)\leqslant 0}W(\frac{d(X_j,X_i)}{a_n},\frac{\widetilde{Y}_j-\widetilde{Y}_i}{a_n})}{\sum_{i\in V}W(\frac{d(X_j,X_i)}{a_n},\frac{\widetilde{Y}_j-\widetilde{Y}_i}{a_n})} \\ \widehat{u}_2(X_j,\widetilde{Y}_j) = \frac{\sum_{i\in V}(Y_i-t)\mathbf{1}_{(Y_i-t)>0}W(\frac{d(X_j,X_i)}{a_n},\frac{\widetilde{Y}_j-\widetilde{Y}_i}{a_n})}{\sum_{i\in V}W(\frac{d(X_j,X_i)}{a_n},\frac{\widetilde{Y}_j-\widetilde{Y}_i}{a_n})}, \qquad \text{for} \quad j \in \bar{V} \end{cases}$$

$W(.,.)$ is a kernel function which is define on $\mathbb{R}^2$ and a_n is a sequence of real number which tend to zero when n tend to infinity.

1.3.5 Results

Uniform almost complete consistency

Theorem 1.1. *Under assumptions, we obtain*

$$\sup_{x\in S_{\mathcal{F}}}\sup_{y\in S_{\mathbb{R}}}|\widehat{F}^x(y)-F^x(y)| = O(h_K^{A_1}+h_H^{A_2}+a_n^{A_1})+O_{a.co.}\left(\sqrt{\frac{\log d_n}{n\phi(a_n)}}\right)+O_{a.co.}\left(\sqrt{\frac{\log N}{N\phi(h_K)}}\right).$$

The quadratic error of the conditional distribution function for surrogate data

Theorem 1.2. *Under assumptions* $(A1)-(A4)$*, we obtain*

$$\mathbb{E}\left(\widehat{F}^x(y)-F^x(y)\right)^2 = o\left(\frac{\alpha}{N\phi(h_K)^{3/2}\phi(a_n)^{3/2}}\right)+o\left(\frac{1-\alpha}{N\phi(h_k)\phi(h_H)}\right)+o\left(h_H^{\beta}h_K^{\beta}\right)+o\left(h_Kh_H\right)$$

The asymptotic normality of the conditional distribution function for surrogate data

Theorem 1.3. *Under assumptions, we obtain*

$$\sqrt{Nh_N}\left(\widehat{F}^x(y)-F^x(y)\right)^2 \xrightarrow{L} \mathcal{N}\left(0,\sigma^2(x)\right)$$

where

$$\sigma_p^2(x) = \frac{\alpha_2(x)\lambda_p(\theta(p;x);x)}{\alpha_1^2(x)}\Gamma_p^2(\theta(p;x);x) \qquad \left(\text{with } \alpha_j(x) = K^j(1) - \int_0^1 (K^j)'(s)\beta_x(s)ds \text{ for } j = 1,2\right)$$

and

$$\lambda_p(\theta(p;x);x) = \left(\frac{p}{1-p}\right)^2 R_+^x(\theta(p;x)) + R_-^x(\theta(p;x))$$

where

$$R_+(\theta(p;x);x) = \mathbb{E}\left[(Y_1 - \theta(p;x))^2 \mathbf{1}_{(Y_1>\theta(p;x)}|X = x\right]$$
$$R_-(\theta(p;x);x) = \mathbb{E}\left[(Y_1 - \theta(p;x))^2 \mathbf{1}_{(Y_1\leq\theta(p;x)}|X = x\right]$$

and

$$\Lambda_p(\theta(p;x);x) = A_1'(\theta(p;x);x) - \left(\frac{p}{1-p}\right) A_2'(\theta(p;x);x).$$

The asymptotic properteis of the conditional expectile regression estimator for surrogate data

Theorem 1.4. *Under the assumptions, and if in addition*

$$\frac{\partial A(x,\xi_p(x))}{\partial t} > 0$$

then

$$\sup_{x\in\mathcal{F}}|\widehat{\xi_p}(x) - \xi_p(x)| = O(h^{k_{l_K}}) + O(a^{\min k_{l_n}})$$

$$+O_{a.co.}\left(\sqrt{\frac{\log n}{n\phi(a_n)}}\right) + O_{a.co.}\left(\sqrt{\frac{\log N}{N\phi(h_K)}}\right).$$

The asymptotic Normality of the conditional expectile regression estimator for surrogate data

Theorem 1.5. *Under the Hypotheses we have when* $n \to \infty$

$$\left(\frac{n\phi_x(h_n)}{\sigma_p^2}(x)\right)^{1/2}\left(\hat{\xi}_p(x) - \xi_p(x)\right) \xrightarrow{D} \mathcal{N}(0,1)$$

where

$$\sigma_p^2(x) = \frac{\alpha_2(x)\lambda_p(\theta(p;x);x)}{\alpha_1^2(x)}\Gamma_p^2(\theta(p;x);x) \qquad \left(\text{with } \alpha_j(x) = K^j(1) - \int_0^1 (K^j)'(s)\beta_x(s)ds \text{ for } j = 1,2\right)$$

and

$$\lambda_p(\theta(p;x);x) = \left(\frac{p}{1-p}\right)^2 R_+^x(\theta(p;x)) + R_-^x(\theta(p;x))$$

where

$$R_+(\theta(p;x);x) = \mathbb{E}\left[(Y_1 - \theta(p;x))^2 \mathbf{1}_{(Y_1 > \theta(p;x)}|X = x\right]$$
$$R_-(\theta(p;x);x) = \mathbb{E}\left[(Y_1 - \theta(p;x))^2 \mathbf{1}_{(Y_1 \leq \theta(p;x)}|X = x\right]$$

and

$$\Lambda_p(\theta(p;x);x) = A_1'(\theta(p;x);x) - \left(\frac{p}{1-p}\right) A_2'(\theta(p;x);x).$$

CHAPITRE 2

ASYMPTOTIC STUDY OF THE CONDITIONAL DISTRIBUTION FUNCTION FOR SURROGATE DATA BY THE REGRESSION MODEL

2.1 Introduction

Conditional distribution function (CFD) estimation is an essential field in nonparametric statistical analysis ; this technique helps us understand the relationship between a response variable and covariates set.

One of the branches of modern statistics is Functional Data Analysis (FDA) ; this has become possible thanks to the computing techniques' progress, both in terms of memory and storage capacities, which allows us to consider increasingly voluminous data, regarded as an observation of curve or surface. The reader can consult the books of Ramsay and Silverman (1997), Ramsay and Silverman (2002), Bosq (2000) and Ferraty and Vieu (2006), which offer a good introduction both for the theoretical or applied aspect with various applications, including economics, sociology, and biology. It should be noted that extensions of probability theory to random variables taking values in normed spaces (e.g., Banach and Hilbert spaces), including extensions of some classical theorems, are handy tools in the literature.

Note first that the study of the conditional distribution function of real data was obtained in the early 1960s by Roussas (1968) who studied the kernel estimator's asymptotic properties conditional distribution function where it showed convergence in probability. In the case of functional data, many researchers have been interested in the study of this function. For example, we cite, Ferraty et al. (2006) who estimate the conditional distribution characteristics in nonparametric functional models. In the same framework, Ferraty

et al. (2005) use the conditional distribution function to obtain a nonparametric estimator of the conditional quantile when the data is weakly dependent.

It should be noted that most of the results involved in the nonparametric literature (and not only on the conditional distribution) only deal with completely observed samples. While in many practical works, including, for example, sample survey, reliability, or pharmaceutical tracing where data is often observed incompletely, and parts of the responses are missing randomly (MAR).

The most popular method to involve missing data is the imputation method that fills or retrieves the missing data in the response variable Y. In this context, we can cite various works that used this technique : We can cite, Yates (1933) for the linear regression model. The kernel estimation of the mean functions is considered in Cheng (1994), the nearest neighbor imputation for the data survey is addressed in Chen and Shao (2000), the robust regression model with missing data is considered in Pérez-González et al. (2009), the asymptotic properties of the regression operator estimator when the regressor is functional and completely observed, and that missing data at random in the scalar response variable are investigated in Ferraty et al. (2013), in the case of dependent data, the reader may refer to Ling et al. (2015). In this work, we investigate the unavailability of response data because sometimes it is default or very expensive to measure some response observations ; the main idea is to recover (or fill) this missing data by a surrogate validation data set. In this context, we cite Duncan and Hill (1985), Wittes et al. (1989), Carroll and Wand (1991) and Pepe (1992). The principle of this method is to incorporate both surrogate data and the corresponding observations of the covariate X.

This paper aims to study the conditional models (conditional distribution function and the conditional quantile) for missing response by the kernel method ; we explore in this work, the aspect of missing data in the response variable. First, we consider the estimator of the conditional distribution for complete data, then by using the validation data set (see, Ibrahim et al. (2020) and Wang (2006), we build our new estimator with surrogate data and we obtain some asymptotic results for the conditional distribution and the quantiles. In the end, we realized a simulation study to improve the efficacity of our estimator.

The rest of the paper is organized as follows. We present our model in Section 2 ; the required notations and assumptions are introduced in Section 3, the main results of strong uniform consistency (with rate) and the quantile estimation as a direct consequence of our asymptotic result obtained from CFD estimation are formulated in section 4. For the numerical results, a simulation study that shows the performance of the proposed estimator is presented in Section 5.

2.2 Model and estimator of the conditional distribution function

2.2.1 Estimation of the *cond-cdf* with complete data

Let $(X_i, Y_i)_{i=1,...,N}$ be a random variables independent and identically distributed as (X, Y), where $X \in \mathcal{F}$, Y take values in $\mathbb{R}$ and $(\mathcal{F}, d)$ is a semi metric space with a metric $d(.,.)$. The conditional cumulative distribution function of Y given $X = x$, denoted by $F^x(.)$ is defined

$$F^x(.) = \mathbb{P}(Y \leq y | X = x), \quad \forall y \in \mathbb{R},$$

and by the regression model, we have

$$\mathbb{E}\left[H\left(\frac{y - Y_i}{h_H}\right) \Big| X_i = x \right] \xrightarrow{h_H \to 0} F^x(y),$$

where $H(.)$ is a cumulative distribution function and h_H is a sequence of positive real numbers tending to 0 when n go to infinity.

The estimator of conditional distribution function by the kernel method defined by

$$\widehat{F}_C^x(y) = \frac{\sum_{i=1}^N H\left(\frac{y-Y_i}{h_H}\right) K\left(\frac{d(x,X_i)}{h_K}\right)}{\sum_{i=1}^N K\left(\frac{d(x,X_i)}{h_K}\right)}, \quad \forall y \in \mathbb{R}, \ \forall x \in \mathcal{F}, \tag{2.1}$$

where $K(.)$ is a kernel function and h_K is a bandwidth sequence tend toward 0.

2.2.2 Estimation of the *cond-cdf* with surrogate data

We have the sample set of the size N and the validation set of size n, where the observations are independent and identically distributed. Here, Y is not accessible(available), so we replaced it by a surrogate variable $\widetilde{Y}$.

Let V the index set of the sampled validation set and $\bar{V} = \{1, ..., N\} \setminus V$. Note that, for the surrogate data we have

$$\mathbb{E}\left[H\left(\frac{y - Y_j}{h_H}\right) \Big| X_j, \widetilde{Y}_j \right] \xrightarrow{h_H \to 0} F^x(y)$$

and

$$\mathbb{E}\left[\mathbb{E}\left[H\left(\frac{y - Y_j}{h_H}\right) \Big| X_j, \widetilde{Y}_j \right] \Big| X_j = x \right] = \mathbb{E}\left[H\left(\frac{y - Y_j}{h_H}\right) \Big| X_j = x \right], \tag{2.2}$$

then, the distribution function can be estimated by

$$\widehat{F}^x(y) = \frac{\sum_{i\in V} H\left(\frac{y-Y_i}{h_H}\right) K\left(\frac{d(x,X_i)}{h_K}\right) + \sum_{j\in\bar{V}} u(X_j,\widetilde{Y}_j) K\left(\frac{d(x,X_j)}{h_K}\right)}{\sum_{i=1}^{N} K\left(\frac{d(x,X_i)}{h_K}\right)},$$

where

$$u(X_j,\widetilde{Y}_j) = \mathbb{E}\left[H\left(\frac{y-Y_j}{h_H}\right)\Big|X_j,\widetilde{Y}_j\right]$$

and the function $u(.,.)$ is unknown.

So, we estimate this function by validation data set :

$$\widehat{u}(X_j,\widetilde{Y}_j) = \frac{\sum_{i\in V} H\left(\frac{y-Y_i}{h_H}\right) W\left(\frac{d(X_j,X_i)}{a_n}, \frac{\widetilde{Y}_j-\widetilde{Y}_i}{a_n}\right)}{\sum_{i\in V} W\left(\frac{d(X_j,X_i)}{a_n}, \frac{\widetilde{Y}_j-\widetilde{Y}_i}{a_n}\right)}$$

and $W(.,.)$ is the two-dimensional kernel function in $\mathcal{F}\times\mathbb{R}$ and a_n is a sequence of real number which tend to zero when n tend to infinity.

2.3 Assumptions

Let $S_{\mathcal{F}}$ be some subset of $\mathcal{F}$ such that $S_{\mathcal{F}} \subset \bigcup_{k=1}^{d_n} B(x_k, r_n)$, where $x_k \in \mathcal{F}$, and (d_n) is a sequence of integers which satisfies the assumption (A5).

Let us introduce $B(x,h_K)$ a ball of the center x and radius h_K defined as $B(x,h_K) = \{x_1 \in \mathcal{F} : d(x_1,x) \le h_K\}$. Furthermore, we have x a fixed point in $\mathcal{F}$, and $S_{\mathbb{R}}$ a fixed compact subset of $\mathbb{R}$.

Our assumptions are gathered below for easy references.

(A1) $\forall h_K > 0,\ \mathbb{P}(X \in B(x,h_K)) =: \phi(h_K) > 0.$

(A2) The operators $F^x(.)$ and $u(.,.)$ are Lipschitzian, such that, $\forall (y_1,y_2) \in S_{\mathbb{R}}^2,\ \forall (x_1,x_2) \in S_{\mathcal{F}}^2$ and $C, A_1, A_2 > 0$

(a) $|F^{x_1}(y_1) - F^{x_2}(y_2)| \le C\left(d(x_1,x_2)^{A_1} + |y_1-y_2|^{A_2}\right).$

(b) $|u(x_1,y_1) - u(x_2,y_2)| \le C\left(d(x_1,x_2)^{A_1} + |y_1-y_2|^{A_2}\right).$

(A3) The distribution function $H(.)$ satisfy

$$\begin{cases} \forall\,(y_1,y_2)\in\mathbb{R}^2, |H(y_1) - H(y_2)| \le C|y_1-y_2|, \\ \int |t|^{A_2} H'(t)\,\mathrm{d}t < \infty. \end{cases}$$

(A4) The bandwidths h_K and a_n satisfy

$$\lim_{N\to\infty} h_K = \lim_{n\to\infty} a_n = 0 \quad \text{and} \quad \lim_{N\to\infty} N\phi(h_K) = +\infty$$

and

$$\lim_{N\to\infty} \frac{\log N}{N\phi(h_K)} = 0 \quad \text{and} \quad \lim_{n\to\infty} \frac{\log n}{n\phi(a_n)} = 0.$$

(A5) For some $\beta > 0$,

$$\lim_{N\to\infty} h_H = 0 \quad \text{with} \lim_{N\to\infty} N^{\beta} h_H = \infty,$$

and for $r_n = O\left(\frac{\log N}{N}\right)$ the sequence d_n satisfy

$$\frac{\log^2 N}{N\phi(a_n)} \leq d_n \leq \frac{N\phi(a_n)}{\log N} \quad \text{and} \quad \sum_{n=1}^{\infty} n^{\beta} \exp\{(1-\eta)\log d_n\} < \infty \text{ where } \beta > 0 \text{ and } \eta > 1. \tag{2.3}$$

(A6) The kernel $K(.)$ is a continuous function from $\mathbb{R}$ into $\mathbb{R}^+$ such that $\int K = 1$, and there exist some positive constants C and C' such that

$$C\mathbf{1}_{(0,1)} \leq K \leq C'\mathbf{1}_{(0,1)} \tag{2.4}$$

where $\mathbf{1}_A$ denotes the indicator function on the set A.
We assume the two-dimensional kernel $W(x,y) = W_1(x)W_2(y)$ is a continuous function with a compact support satisfies (4.2), however, there exist positive finite real constants C_3 and C_4, such that

$$C_3\phi(a_n) \leq \mathbb{E}\left[W\left(\frac{d(X_j,X_i)}{a_n}, \frac{\widetilde{Y}_j - \widetilde{Y}_i}{a_n}\right)\right] \leq C_4\phi(a_n).$$

Remark 2.3.1. *The concentration assumption* $(A1)$ *depend to the distribution of* X *and has an important role, which is linked with the semi-metric* $d(.,.)$*. Note that the correct choice for* $d(.,.)$ *is through the corresponding function* $\phi(.)$ *a key to the curse of dimensionality. The assumption* $(A2)$ *is linked with the nonparametric structure of the model and it's used it for determine the bias term. The assumptions* $(A3)$ – $(A6)$ *are a technical condition similar to the hypothesis in Ferraty et al. (2006) for obtain our results.*

2.4 Results

2.4.1 Uniform almost complete consistency

The uniform almost complete $(O_{a.co.})$ convergence of $\widehat{F}^x(.)$ is given by the following Theorem and Lemmas.

Theorem 2.1. *Under assumptions* $(A1)$ – $(A6)$*, we obtain*

$$\sup_{x\in S_{\mathcal{F}}}\sup_{y\in S_{\mathbb{R}}}|\widehat{F}^x(y)-F^x(y)| = O(h_K^{A_1}+h_H^{A_2}+a_n^{A_1})+O_{a.co.}\left(\sqrt{\frac{\log d_n}{n\phi(a_n)}}\right)+O_{a.co.}\left(\sqrt{\frac{\log N}{N\phi(h_K)}}\right).$$

Démonstration. Let $\widehat{F}_N^x(y)$ and $\widehat{F}_D^x(y)$, defined by

$$\widehat{F}_N^x(y) = \frac{1}{N}\sum_{i\in V}\frac{H\left(\frac{y-Y_i}{h_H}\right)K\left(\frac{d(x,X_i)}{h_K}\right)}{\mathbb{E}\left[K\left(\frac{d(x,X_i)}{h_K}\right)\right]} + \frac{1}{N}\sum_{j\in\bar{V}}\frac{\widehat{u}(X_j,\widetilde{Y}_j)K\left(\frac{d(x,X_j)}{h_K}\right)}{\mathbb{E}\left[K\left(\frac{d(x,X_j)}{h_K}\right)\right]}$$

and

$$\widehat{F}_D^x = \frac{1}{N}\sum_{i=1}^{N}\frac{K\left(\frac{d(x,X_i)}{h_K}\right)}{\mathbb{E}\left[K\left(\frac{d(x,X_i)}{h_K}\right)\right]}.$$

The proof is based on the following decomposition and the Lemmas 2.4.2, 2.4.3 and 2.4.4 given below.

$$\begin{aligned}\widehat{F}^x(y)-F^x(y) = \frac{1}{\widehat{F}_D^x}\left\{\left(\widehat{F}_N^x(y)-\mathbb{E}[\widehat{F}_N^x(y)]\right)-\left(F^x(y)-\mathbb{E}[\widehat{F}_N^x(y)]\right)\right\}\\ -\frac{F^x(y)}{\widehat{F}_D^x}\left\{\widehat{F}_D^x-1\right\}. \qquad (2.5)\end{aligned}$$

□

Auxiliary results

We put the quantities, for $x\in\mathcal{F}$, $(y,\widetilde{y})\in\mathbb{R}^2$ and $i,j=1,...,N$:

$$K_i := K\left(\frac{d(x,X_i)}{h_K}\right),\ H_i(y) := H\left(\frac{y-Y_i}{h_H}\right), W_{ij} := W\left(\frac{d(X_j,X_i)}{a_n},\frac{\widetilde{Y}_j-\widetilde{Y}_i}{a_n}\right).$$

We note for $j\in\bar{V}$:

$$\widehat{u}(X_j,\widetilde{Y}_j) = \frac{\sum_{i\in V}H_i(y)W_{ij}}{\sum_{i\in V}W_{ij}} := \frac{\widehat{u}_N^x(y)}{\widehat{u}_D^x}.$$

We need the following lemma to establish the uniform almost complete convergence.

Lemma 2.4.1. *Under assumptions* $(A1)$ – $(A6)$*, we get*

- $F_1 = \sup_{x\in S_{\mathcal{F}}} |\widehat{u}_D^x - 1]| = O_{a.co.}\left(\sqrt{\frac{\log d_n}{n\phi(a_n)}}\right)$ *and* $\sum_{n=1}^{\infty} \mathbb{P}\left(|\widehat{u}_D^x| \le 1/2\right) < \infty$.
- $F_2 = \sup_{x\in S_{\mathcal{F}}} \sup_{y\in S_{\mathbb{R}}} |\widehat{u}_N^x(y) - \mathbb{E}[\widehat{u}_N^x(y)]| = O_{a.co.}\left(\sqrt{\frac{\log d_n}{n\phi(a_n)}}\right)$.
- $F_3 = \sup_{x\in S_{\mathcal{F}}} \sup_{y\in S_{\mathbb{R}}} |u(x_j, \widetilde{y}_j) - \mathbb{E}[\widehat{u}_N^x(y)]| = O(a_n^{A_1}) + O(h_H^{A_2})$.

Démonstration. 1. As F_1 is a particular case of F_2 (by taking $H(.) \equiv 1$), then it proof will be omitted.

Now, we have

$$\mathbb{P}\left(|\widehat{u}_D^x| \le 1/2\right) \le \mathbb{P}\left(|\widehat{u}_D^x - 1| > 1/2\right),$$

thus, by applying the result above, we get $\sum_{i=1}^{\infty} \mathbb{P}\left(|\widehat{u}_D^x| \le 1/2\right) < \infty$.

2. We conceive the following decomposition, where for all $x \in S_{\mathcal{F}}$, we set $k(x) = \operatorname{argmin}_{k\in\{1,\dots,d_n\}} |x - x_k|$ and we use the compactness of $S_{\mathbb{R}}$, where, we can write $S_{\mathbb{R}} \subset \bigcup_{j=1}^{q_n} S_j, S_j = (l_j - l_n, l_j + l_n)$ and take $y_t = \operatorname{argmin}_{l\in\{l_1,\dots,l_{q_n}\}} |y - l|$, to obtain

$$\begin{aligned}
\underbrace{\sup_{x\in S_{\mathcal{F}}} \sup_{y\in S_{\mathbb{R}}} |\widehat{u}_N^x(y) - \mathbb{E}[\widehat{u}_N^x(y)]|}_{F_2} \le{}& \underbrace{\sup_{x\in S_{\mathcal{F}}} \sup_{y\in S_{\mathbb{R}}} |\widehat{u}_N^x(y) - \widehat{u}_N^{x_{k(x)}}(y)|}_{P_1} \\
&+ \underbrace{\sup_{x\in S_{\mathcal{F}}} \sup_{y\in S_{\mathbb{R}}} |\widehat{u}_N^{x_{k(x)}}(y) - \widehat{u}_N^{x_{k(x)}}(y_t)|}_{P_2} \\
&+ \underbrace{\sup_{x\in S_{\mathcal{F}}} \sup_{y\in S_{\mathbb{R}}} |\widehat{u}_N^{x_{k(x)}}(y_t) - \mathbb{E}[\widehat{u}_N^{x_{k(x)}}(y_t)]|}_{P_3} \\
&+ \underbrace{\sup_{x\in S_{\mathcal{F}}} \sup_{y\in S_{\mathbb{R}}} |\mathbb{E}[\widehat{u}_N^{x_{k(x)}}(y_t)] - \mathbb{E}[\widehat{u}_N^{x_{k(x)}}(y)]|}_{P_4} \\
&+ \underbrace{\sup_{x\in S_{\mathcal{F}}} \sup_{y\in S_{\mathbb{R}}} |\mathbb{E}[\widehat{u}_N^{x_{k(x)}}(y)] - \mathbb{E}[\widehat{u}_N^{x}(y)]|}_{P_5}.
\end{aligned}$$

- For P_1 and P_5, we have from $(A3)$ and the boundness of $W(.,.)$ we can write

$$\begin{aligned}
P_1 &\le \frac{C}{\phi(a_n)} \sup_{x\in S_{\mathcal{F}}} \sup_{y\in S_{\mathbb{R}}} \frac{1}{n} \sum_{i\in V} |W(x, \tilde{y}) - W(x_{k(x)}, \tilde{y})| \\
&\le \frac{C d_n q_n}{a_n \phi(a_n)}
\end{aligned}$$

and analogously, for P_2 we obtain

$$P_2 \leq \frac{Cd_nq_n}{a_n\phi(a_n)}\mathbf{1}_{B(x,a_n)\cup B(x_{k(x)},a_n)}$$

by applying Bernstein's inequality, with

$$Z_i = \frac{\epsilon}{a_n\phi(a_n)}\mathbf{1}_{B(x,a_n)\cup B(x_{k(x)},a_n)},$$

which gives, for n tending to infinity

$$P_1 = O\left(\sqrt{\frac{\log d_n}{n\phi\left(a_n\right)}}\right) \quad \text{and} \quad P_2 = O\left(\sqrt{\frac{\log d_n}{n\phi\left(a_n\right)}}\right).$$

Moreover, using the fact that $P_5 \leq P_1$ and $P_4 \leq P_2$ to get, for n tending to infinity

$$P_5 = O\left(\sqrt{\frac{\log d_n}{n\phi\left(a_n\right)}}\right) \quad \text{and} \quad P_4 = O\left(\sqrt{\frac{\log d_n}{n\phi\left(a_n\right)}}\right).$$

- Now concerning P_3. For all $\eta > 0$, we have

$$\mathbb{P}\left(P_3 > \eta\sqrt{\frac{\log d_n}{n\phi(a_n)}}\right) \leq q_nd_n \max_{x_k\in\{1,\ldots,d_n\}}\max_{y_t\in\{1,\ldots,t_{q_n}\}} \mathbb{P}\left(|\widehat{u}_N^{x_{k(x)}}(y_t) - \mathbb{E}[\widehat{u}_N^{x_{k(x)}}(y_t)]| > \eta\sqrt{\frac{\log d_n}{n\phi(a_n)}}\right)$$

we can use the Bernstein's exponential inequality to Γ_i, where

$$\Gamma_i = \frac{1}{n\phi(a_n)}\left\{W_{i,j}(x_{k(x)},y_t)H_i(y_t) - \mathbb{E}\left[W_{i,j}(x_{k(x)},y_t)H_i(y_t)\right]\right\}, \quad \text{for} j \in \bar{V},$$

and we have $|\Gamma_i| \leq C_4/\phi(a_n),\ \mathbb{E}|\Gamma_i|^2 \leq C/\phi(a_n)$.
However, take $C\eta^2 = 2\beta$ and $q_n = O(l_n^{-1})$, we get

$$q_nd_n\mathbb{P}\left(\sup_{x\in S_{\mathcal{F}}}\sup_{y\in S_{\mathbb{R}}}|\widehat{u}_N^{x_{k(x)}}(y_t) - \mathbb{E}[\widehat{u}_N^{x_{k(x)}}(y_t)]| > \eta\sqrt{\frac{\log d_n}{n\phi(a_n)}}\right) \leq q_nd_n2\exp\{-C\eta^2\ln d_n\}$$

then, by $(A6)$ we get

$$P_3 = O_{\text{a.co.}}\left(\sqrt{\frac{\log d_n}{n\phi(a_n)}}\right). \tag{2.6}$$

3. We have for $j \in \bar{V}$:

$$\begin{aligned} F_3 &:= \mathbb{E}[\widehat{u}_N^x(y)] - u(x,\widetilde{y}) \\ &= \mathbb{E}\left[W_{ij}\left(\mathbb{E}\left(H_1(y)\,|X,\widetilde{Y}\right) - u(x,\widetilde{y})\right)\right] \end{aligned}$$

and we have $\mathbb{E}(H_1(y)|X,\widetilde{Y}) = u(X,\widetilde{Y})$,
then, from $(A2)$, we get

$$|u(X,\widetilde{Y}) - u(x,\widetilde{y})| \leq C(a_n^{A_1} + h_H^{A_2}).$$

Finally, from (F_1), (F_2) and (F_3), we finished the proof of Lemma 2.4.1. □

Lemma 2.4.2. *Under the assumptions* $(A1) - (A6)$*, we obtain*

$$\sup_{x\in S_{\mathcal{F}}}\sup_{y\in S_{\mathbb{R}}}|F^x(y) - \mathbb{E}[\widehat{F}_N^x(y)]| = O(h_K^{A_1}) + O(h_H^{A_2}) + O(a_n^{A_1}) + O_{a.co.}\left(\sqrt{\frac{\log d_n}{n\phi(a_n)}}\right).$$

Démonstration. We have $|V| = n, |\bar{V}| = N - n$

$$\begin{aligned} F^x(y) - \mathbb{E}[\widehat{F}_N^x(y)] &= F^x(y) - \mathbb{E}\left[n\frac{H_1(y)K_1}{\mathbb{E}[K_1]} + (N-n)\frac{\widehat{u}(X_j,\widetilde{Y}_j)K_1}{\mathbb{E}[K_1]}\right] \\ &= F^x(y) - n\mathbb{E}\left[\frac{H_1(y)K_1}{E[K_1]}\right] - (N-n)\mathbb{E}\left[\frac{\widehat{u}(X_j,\widetilde{Y}_j)K_j}{\mathbb{E}[K_1]}\right] := T_1 + T_2. \end{aligned}$$

- Concerning the term T_1 :

$$\begin{aligned} F^x(y) - \mathbb{E}\left[\frac{H_1(y)K_1}{\mathbb{E}[K_1]}\right] &= F^x(y) - \mathbb{E}\left[\mathbb{E}\left[\frac{H_1(y)K_1}{\mathbb{E}[K_1]}\Big|X_1\right]\right] \\ &= F^x(y) - \mathbb{E}\left[H_1(y)|X_1\right]. \end{aligned}$$

We know that

$$\mathbb{E}\left[H_1(y)|X_1\right] = \int_{\mathbb{R}} H'(t)F^{X_1}(y - h_H t)\mathrm{d}t$$

and

$$|\mathbb{E}\left[H_1(y)|X_1\right] - F^x(y)| \leq \int_{\mathbb{R}} H'(t)|F^{X_1}(y - h_H t) - F^x(y)|\mathrm{d}t.$$

So, from $(A2)$, we get

$$|\mathbb{E}\left[H_1(y)|X_1\right] - F^x(y)| \leq C\int_{\mathbb{R}} H'(t)\left(h_K^{A_1} + |t|_2^A h_H^{A_2}\right)\mathrm{d}t,$$

then, $T_1 = O(h_K^{A_1}) + O(h_H^{A_2})$.

- Concerning the term T_2 :

$$F^x(y) - \mathbb{E}\left[\frac{\widehat{u}(X_j,\widetilde{Y}_j)K_1}{\mathbb{E}[K_1]}\right] = \mathbb{E}\left(u(X_j,\widetilde{Y}_j) - \widehat{u}(X_j,\widetilde{Y}_j)\frac{K_1}{\mathbb{E}[K_1]}\right) + \mathbb{E}\left(F^x(y) - H_1(y)\frac{K_1}{\mathbb{E}[K_1]}\right) + \mathbb{E}\left(\left(H_1(y) - u(X_j,\widetilde{Y}_j)\right)\frac{K_1}{\mathbb{E}[K_1]}\right).$$

Thus,

(a) Firstly, we have

$$\sup_{x\in S_{\mathcal{F}}}\sup_{y\in S_{\mathbb{R}}}\left|\mathbb{E}\left(u(X_j,\widetilde{Y}_j) - \widehat{u}(X_j,\widetilde{Y}_j)\frac{K_1}{\mathbb{E}[K_1]}\right)\right| = O\left(\sup_{x\in S_{\mathcal{F}}}\sup_{y\in S_{\mathbb{R}}}\left|u(x,y) - \widehat{u}(x,y)\right|\right)$$

by the following decomposition for $j \in \bar{V}$:

$$\widehat{u}(X_j,\widetilde{Y}_j) - u(X_j,\widetilde{Y}_j) = -\frac{u}{\widehat{u}_D^x}\left(\widehat{u}_D^x - 1\right) + \frac{1}{\widehat{u}_D^x}\left\{\widehat{u}_N^x(y) - \mathbb{E}[\widehat{u}_N^x(y)] - (u - \mathbb{E}[\widehat{u}_N^x(y)])\right\}$$
$$:= -\frac{u}{\widehat{u}_D^x}T_{2,1} + \frac{1}{\widehat{u}_D^x}\left(T_{2,2} - T_{2,3}\right)$$

then, from (lemma 2.4.1), we get

$$T_{2,1} = T_{2,2} = O_{\text{a.co.}}\left(\sqrt{\frac{\log d_n}{n\phi(a_n)}}\right) \text{ and } T_{2,3} = O(a_n^{A_1}) + O(h_H^{A_2}).$$

(b) Secondly, we have

$$\left|\mathbb{E}\left(F^x(y) - H_1(y)\frac{K_1}{\mathbb{E}[K_1]}\right)\right| = \left|F^x(y) - \mathbb{E}[H_1(y)|X_1]\right|$$

and $\mathbb{E}[H_1(y)|X_1] = \int_{\mathbb{R}} H'(t)F^X(y - h_H t)\mathrm{d}t$,
so, from the hypothesis $(A2)$, we get

$$\left|\mathbb{E}\left(F^x(y) - H_1(y)\frac{K_1}{\mathbb{E}[K_1]}\right)\right| = O(h_K^{A_1}) + O(h_H^{A_2}). \tag{2.7}$$

(c) Thirdly, its clear that after (b), we get

$$\left|\mathbb{E}\left(\left(H_1(y) - u(X_j,\widetilde{Y}_j)\right)\frac{K_1}{\mathbb{E}[K_1]}\right)\right| = 0. \tag{2.8}$$

Finally, from T_1 and T_2 the proof of lemma 2.4.2 is achieved. □

Lemma 2.4.3. *Under the assumptions* $(A1)$ *and* $(A3)$ – $(A6)$*, we obtain*

$$\sup_{x\in S_{\mathcal{F}}}\sup_{y\in S_{\mathbb{R}}}|\widehat{F}_N^x(y)-\mathbb{E}[\widehat{F}_N^x(y)]| = O_{a.co.}\left(\sqrt{\frac{\log N}{N\phi(h_K)}}\right).$$

Démonstration. We keep the same notation used previously, in the definitions of $k(x)$ and y_t. The proof is based on the following decomposition

$$\begin{aligned}\sup_{x\in S_{\mathcal{F}}}\sup_{y\in S_{\mathbb{R}}}|\widehat{F}_N^x(y)-\mathbb{E}[\widehat{F}_N^x(y)]| \le & \sup_{x\in S_{\mathcal{F}}}\sup_{y\in S_{\mathbb{R}}}|\widehat{F}_N^x(y)-\widehat{F}_N^{x_{k(x)}}(y)| + \sup_{x\in S_{\mathcal{F}}}\sup_{y\in S_{\mathbb{R}}}|\widehat{F}_N^{x_{k(x)}}(y)-\widehat{F}_N^{x_{k(x)}}(y_t)| \\ & + \sup_{x\in S_{\mathcal{F}}}\sup_{y\in S_{\mathbb{R}}}|\widehat{F}_N^{x_{k(x)}}(y_t)-\mathbb{E}[\widehat{F}_N^{x_{k(x)}}(y_t)]| \\ & + \sup_{x\in S_{\mathcal{F}}}\sup_{y\in S_{\mathbb{R}}}|\mathbb{E}[\widehat{F}_N^{x_{k(x)}}(y_t)]-\mathbb{E}[\widehat{F}_N^{x_{k(x)}}(y)]| \\ & + \sup_{x\in S_{\mathcal{F}}}\sup_{y\in S_{\mathbb{R}}}|\mathbb{E}[\widehat{F}_N^{x_{k(x)}}(y)]-\widehat{F}_N^{x}(y)| \\ =: & E_1+E_2+E_3+E_4+E_5. \end{aligned} \tag{2.9}$$

- Concerning E_1 and E_5, by following the same lines as for studying the terms P_1 and P_5, we obtain :

$$E_1 = O_{\text{a.co.}}\left(\sqrt{\frac{\log N}{N\phi(h_K)}}\right) \quad \text{and} \quad E_5 = \left(\sqrt{\frac{\log N}{N\phi(h_K)}}\right).$$

- Concerning the term E_2, by using the Lipschitz's condition on the kernel $H(.)$, we can write

$$|\widehat{F}_N^{x_{k(x)}}(y)-\widehat{F}_N^{x_{k(x)}}(y_t)| \le Ch_H^{-1}\underbrace{|y-y_t|}_{l_n}\Big(\underbrace{\frac{1}{N\mathbb{E}[K_1]}\sum_{i\in V}K_i}_{\widehat{F}_D^x}+\underbrace{\sum_{j\in\bar{V}}\frac{K_j}{\mathbb{E}[K_1]}}_{\widehat{F}_D^x}\Big)$$

under $(A6), (A4), (A5)$ and from the almost comply consistency of $\widehat{F}_D$ (Lemma 2.4.4), and take $l_n = N^{-\beta}$, we get

$$E_2 = O_{\text{a.co.}}\left(\sqrt{\frac{\log N}{N\phi(h_K)}}\right) \quad \text{and} \quad E_4 = O\left(\sqrt{\frac{\log N}{N\phi(h_K)}}\right). \tag{2.10}$$

- For E_3, we have

$$\begin{aligned} E_3 &= \sup_{x\in S_{\mathcal{F}}}\sup_{y\in S_{\mathbb{R}}}|\widehat{F}_N^x(z_y)-\mathbb{E}[\widehat{F}_N^x(z_y)]| \\ &\leq \sup_{x\in S_{\mathcal{F}}}\sup_{y\in S_{\mathbb{R}}}\Big|\frac{1}{N}\left(\sum_{i\in V}\frac{H_i(y_t)K_i(x_{k(x)})}{\mathbb{E}[K_1]}-\mathbb{E}\left(\sum_{i\in V}\frac{H_i(y_t)K_i(x_{k(x)})}{\mathbb{E}[K_1]}\right)\right)\Big| \\ &\quad+\sup_{x\in S_{\mathcal{F}}}\sup_{y\in S_{\mathbb{R}}}\Big|\sum_{j\in\bar{V}}\frac{K_j(x_{k(x)})}{\mathbb{E}[K_1]}-\mathbb{E}\left(\sum_{j\in\bar{V}}\frac{K_j(x_{k(x)})}{\mathbb{E}[K_1]}\right)\Big| \\ &=: E_{2,1}+E_{2,2} \end{aligned} \tag{2.11}$$

then, for $E_{2,1}$:

$$\mathbb{P}\left(E_{2,1}>\kappa\sqrt{\frac{\log N}{N\phi(h_K)}}\right)\leq q_n d_n\max_{x\in S_{\mathcal{F}}}\max_{y_t\in S_{\mathbb{R}}}\mathbb{P}\left(\Big|\frac{1}{N}\sum_{i\in V}(\Lambda_i)\Big|>\kappa\sqrt{\frac{\log N}{N\phi(h_K)}}\right)$$

with

$$\Lambda_i=\frac{H_i(y_t)K_i(x_k)}{\mathbb{E}[K_1]}-\mathbb{E}\left(\frac{H_i(y_t)K_i(x_{k(x)})}{\mathbb{E}[K_1]}\right)$$

So, by the Bernstein's exponential inequality for Λ_i, where, $|\Lambda_i|\leq C/\phi(h_K)$ and $\mathbb{E}|\Lambda_i|^2\leq C'/\phi(h_K)$, as usually, we take $q_n=O(l_n^{-1})$, $C\kappa^2=2\beta+1$, such that

$$\begin{aligned} q_n\max_{y_t\in S_{\mathbb{R}}}\mathbb{P}\left(\Big|\frac{1}{N}\sum_{i\in V}\Lambda_i\Big|>\kappa\sqrt{\frac{\log N}{N\phi(h_K)}}\right)&\leq q_n 2\exp\{-C\kappa^2\log N\} \\ &\leq CN^{\beta}N^{-2\beta-1} \end{aligned}$$

so,

$$\mathbb{P}\left(E_{2,1}>\kappa\sqrt{\frac{\log N}{N\phi(h_K)}}\right)\leq CN^{-\beta-1},$$

now, by take $H(y_t)=1$ for $E_{2,1}$, we obtain $E_{2,2}$ in very easy manner.
So,

$$E_3=O_{\text{a.co.}}\left(\sqrt{\frac{\log N}{N\phi(h_K)}}\right). \tag{2.12}$$

Finally, the lemma 2.4.3 is achieved. □

Lemma 2.4.4. *Under the assumptions* $(A1)$ *and* $(A3)-(A6)$*, we obtain*

$$\sup_{x\in S_{\mathcal{F}}}\left|\widehat{F}_D^x-1\right|=O_{a.co.}\left(\sqrt{\frac{\log N}{N\phi(h_K)}}\right).$$

and

$$\sum_{i\in\mathbb{N}} \mathbb{P}(\widehat{F}_D^x < 1/2) < \infty.$$

Démonstration. We have

$$\widehat{F}_D^x - 1 = \frac{1}{N}\sum_{i=1}^{N} \frac{K_i}{\mathbb{E}K_1} - \frac{1}{N}\mathbb{E}(\sum_{i=1}^{N} \frac{K_i}{\mathbb{E}K_1}) \tag{2.13}$$

$$= \frac{1}{N}\sum_{i=1}^{N} \frac{K_i}{\mathbb{E}K_1} - \frac{\mathbb{E}K_i}{\mathbb{E}K_1} \tag{2.14}$$

$$= \frac{1}{N}\sum_{i=1}^{N} \Delta_i \tag{2.15}$$

where $\Delta_i = \dfrac{K_i}{\mathbb{E}K_1} - \mathbb{E}\dfrac{K_i}{\mathbb{E}K_1}$ Under (A6), for $m = 1, 2$, we have

$$0 < \frac{C'}{\phi(h_K)} < \mathbb{E}(K_1^m) < \frac{C}{\phi(h_K)}$$

then

$$|\Delta_i| < \frac{C}{\phi(h_K)} = \theta_1$$

and

$$\mathbb{E}\Delta_i^2 < \frac{C'}{\phi(h_K)} = \theta_2.$$

We apply the Bernstein-type exponential inequality, for all $\varepsilon \in]0, \frac{\theta_1}{\theta_2}[$, we get

$$\mathbb{P}\left(\left|\widehat{F}_D^x - 1\right| > \varepsilon\sqrt{\frac{\log N}{N\phi(h_K)}}\right) \leq 2\exp\left(\frac{-\varepsilon^2 \log N}{4\phi(h_K)\theta_2}\right) \tag{2.16}$$

$$= 2N^{-\varepsilon^2/4\phi(h_K)\theta_2}$$

$$= 2N^{-C\varepsilon^2}.$$

It follows that for ε^2 large enough

$$\sum_{i=1}^{\infty} \mathbb{P}\left(\left|\widehat{F}_D^x - 1\right| > \varepsilon\sqrt{\frac{\log N}{N\phi(h_K)}}\right) < +\infty.$$

For the second part, we have

$$\mathbb{P}\{|\widehat{F}_D^x| \leq 1/2\} \leq \mathbb{P}\{|\widehat{F}_D^x - 1| > 1/2\} \tag{2.17}$$

$$\leq \mathbb{P}\{|\widehat{F}_D^x - \mathbb{E}\widehat{F}_D^x| > 1/2\}$$

we deduce that

$$\sum_{i\in\mathbb{N}} \mathbb{P}\left(\widehat{F}_D^x < 1/2\right) < \infty.$$

□

2.4.2 The consistency of the conditional quantile estimator

In this section we study the asymptotic behaviour of the conditional quantile, obviously we will estimate it by mean of the conditional distribution estimator, we introduce $\hat{q}_\gamma$ the estimator of q_γ defined as

$$\hat{F}^x\left(\hat{q}_\gamma\right) = \gamma$$

where $\gamma \in]0,1[$.

To achieve our result, we need the following hypotheses.

(A7) $H(.)$ is strictly increasing *cond-cdf*

(A8) The distribution $F^x(.)$ is strictly increasing, continuous and differentiable in neighborhood of q_γ.

Note that (A8) control the flatness of the conditional c.d.f. around the quantile to be estimated.

Corollary 2.4.1. *Under assumptions of the Theorem 4.1 and* $(A8)$*, we obtain*

$$|\widehat{q}_\gamma - q_\gamma| = O(h_K^{A_1} + h_H^{A_2} + a_n^{A_1}) + O_{a.co.}\left(\left(\frac{\log N}{N\phi(h_K)}\right)^{1/2}\right) + O_{a.co.}\left(\left(\frac{\log d_n}{n\phi(a_n)}\right)^{1/2}\right).$$

Démonstration. We present briefly the proof, where, Taylor expansion of $F^x(.)$ drive to the existence of some q^* between $\hat{q}_\gamma$ and q_γ and under the condition $(A8)$ we get :

$$\hat{F}^x\left(\hat{q}_\gamma\right) - \hat{F}^x\left(q_\gamma\right) = \left(\hat{q}_\gamma - q_\gamma\right)\hat{F}^{x^{(1)}}\left(q_\gamma^*\right)$$

$$|\hat{q}_\gamma - q_\gamma| = \frac{1}{\hat{F}^{x^{(1)}}\left(q_\gamma^*\right)}\left[\left|\hat{F}^x\left(\hat{q}_\gamma\right) - \hat{F}^x\left(q_\gamma\right)\right|\right].$$

If we could confirm that

$$\exists\delta > 0, \sum_{n=1}^{\infty} \mathbb{P}\left(\hat{F}^{x^{(1)}}\left(q^*\right) < \delta\right) < \infty,$$

we obtain

$$\begin{aligned}\mathbb{P}\left(|\hat{q}_\gamma - q_\gamma| > \epsilon\right) &\leq \mathbb{P}\left(\left|\hat{F}^x\left(\hat{q}_\gamma\right) - \hat{F}^x\left(q_\gamma\right)\right| > \delta\left(\epsilon\right)\right) \\ &= \mathbb{P}\left(\left|F^x\left(q_\gamma\right) - \hat{F}^x\left(q_\gamma\right)\right| > \delta\left(\epsilon\right)\right) \\ &\leq \mathbb{P}\left(\sup_{x\in S_{\mathcal{F}}}\sup_{y\in S_{\mathbb{R}}}\left|\hat{F}^x\left(y\right) - F^x\left(y\right)\right| > \delta\left(\epsilon\right)\right).\end{aligned}$$

Under assumption (A8), and by comparing the rates of convergence given in Theorem 4.1, we have

$$\sum_n \mathbb{P}\left(|\hat{q}_\gamma - q_\gamma| > \epsilon\right) \leq \sum_n \mathbb{P}\left(\sup_{x\in S_{\mathcal{F}}}\sup_{y\in S_{\mathbb{R}}}\left|\hat{F}^x\left(y\right) - F^x\left(y\right)\right| > \delta\left(\epsilon\right)\right) < \infty.$$

□

2.5 Simulation

In this section, we evaluate the behavior of the proposed estimator by conducting a number of simulation studies. Let $\widehat{F}_V^x(y)$ be the standard Nadaraya-Watson estimator with the true observations in the validation data set. That is,

$$\widehat{F}_V^x(y) = \frac{\sum_{i\in V} H\left(\frac{y-Y_i}{h_H}\right) K\left(\frac{d(x,X_i)}{h_K}\right)}{\sum_{i\in V} K\left(\frac{d(x,X_i)}{h_K}\right)}$$

A simulation was conducted to compare the proposed estimators $\widehat{F}_R^x(y)$ with $\widehat{F}_V^x(y)$ and $\widehat{F}_C^x(y)$, where $\widehat{F}_C^x(y)$ is defined above in equation (2.1). It should be pointed out that $\widehat{F}_C^x(y)$ can serve as a gold standard in the simulation study, even though it is practically unachievable because of the measurement errors.
We generated the response variables Y, such that

$$Y_i = m(X_i) + \varepsilon_i \quad \text{for } i = 1, ..., 250,$$

where the functional regressors X_i are defined (see Figure 2.1), for any $t \in [0, \frac{\pi}{2}]$, by :

$$X_i(t) = 3W_i \sin(2\pi t) + A_i t \quad \text{with } W_i \sim \mathcal{N}(1, 0.5) \text{ and } A_i \sim \mathcal{N}(0, 1),$$

the error ε has the standard normal distribution and it's independent of X, and $m(X_i)$ is given by

$$m(X_i) = \frac{5}{1 + \int_0^{\frac{\pi}{2}} X_i(t)\mathrm{d}t}.$$

A sample of smooth curves $X_i(t)$ are plotted in Figure 2.1.

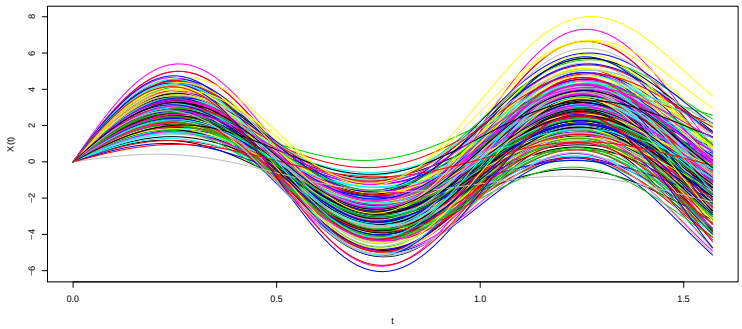

FIGURE 2.1 – Curves (N=250)

Now, let $S_0 = \{1, \ldots, 200\}$ and $S_1 = \{201, \ldots, 250\}$ be two subsets of indices. Then, we choose $\mathcal{L} = \{(X_i, Y_i)\}_{i \in S_0}$ as the learning sample and $\mathcal{T} = \{(X_i, Y_i)\}_{i \in S_1}$ as the test sample. We have from Y_i, for all $i \in S_0$, was generated from

$$\widetilde{Y}_i = \rho Z_i + e_i,$$

where Z_i is the standard score of Y_i and $e_i \sim \mathcal{N}(0, \sqrt{1-\rho^2})$. In such a way that the correlation coefficient between Y_i and $\widetilde{Y}_i$ is approximately equal to ρ which would not be controllable in practice. In the sequel of this simulation study, we take $\rho = 0.35$ or $\rho = 0.75$.

From the learning sample containing $N = 200$ functional data, we randomly choose a set V of n validation data $\{(X_i, Y_i)\}_{i \in V}$ which allows to build the functional kernel estimator $\hat{F}_V^x(y)$ of $m(x)$. The estimator $\hat{F}_R^x(y)$ is then constructed by using the surrogate data $\left\{\left(X_i, \widetilde{Y}_i\right)\right\}_{i \in \bar{V}}$ with the help of the validation data, where $\bar{V} = \{1, \ldots, N\} \backslash V$. It should be pointed out that for $N = n$ (complete observations), we have

$$\hat{F}_V^x(y) = \hat{F}_R^x(y) = \hat{F}_C^x(y).$$

The bandwidths h_H and h_K are selected by a cross-validation method. Because of the smoothness of the curves, we have built the predictors through the semi-metric based on the first derivatives (see, Benhenni et al. (2007)). For the bandwidths a_n, we used the same principal steps in kernels $K(.)$ and $W(.,.)$ are chosen to be the quadratic and the integrate quadratic kernels, these latter are Epanechnikov kernels.

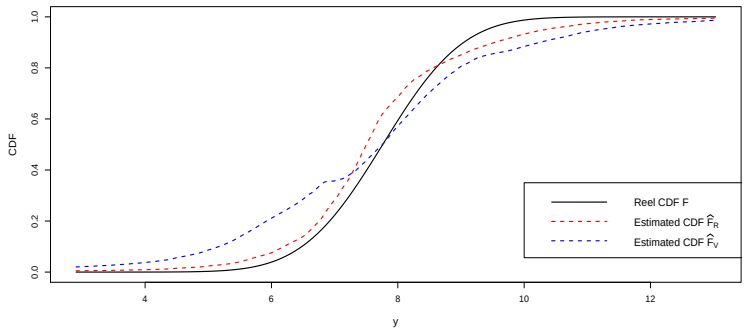

FIGURE 2.2 – CDF comparison

Figure 2.2 represents the curves of the CDF with $F^x(y) = \int_0^y \frac{1}{2\pi} \exp \frac{-(z-m(x))^2}{2} \mathrm{d}z$, where, it is clear that our $\hat{F}_R^x(y)$ is closer to the real curve which represents the complete sample and consequently, $\hat{F}_R(y)$ performs better than $\hat{F}_V^x(y)$.

Hereafter, we will apply our result on the median and obtained results are given in figure 2.3.

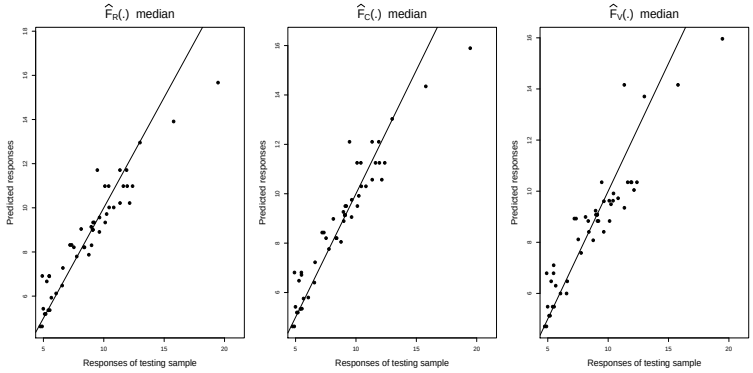

FIGURE 2.3 – Comparative prediction between the median for each : $\hat{F}_R^x(y)$, $\hat{F}_C^x(y)$ and $\hat{F}_V^x(y)$

TABLE 2.1 – MSE result

	$n/N \rightarrow$ ρ $\downarrow$	0.125	0.25
$\hat{F}_V^x(y)$	0.35	0.6543	0.7127
	0.75	0.6729	0.7149
$\hat{F}_R^x(y)$	0.35	0.5503	0.5922
	0.75	0.5692	0.6018
$\hat{F}_C^x(y)$	–	0.5248	0.5248

It can be noticed from Figure 2.3 that the estimator $\hat{F}_R^x(y)$ is better than the estimator $\hat{F}_V^x(y)$. Also, it appears clearly that in this case the performance of both estimates is closely linked to the correlation coefficient and the ration n/N since the values of MSE-error increase substantially with respect to those parameters (see, Table 2.1). In this table we summarize the MSE-error for two values of n/N and ρ, this error increases with respect to those parameters. It is noted that the results are sufficiently good for all sample size, and further results are given for large sample sizes in Figure 2.4.

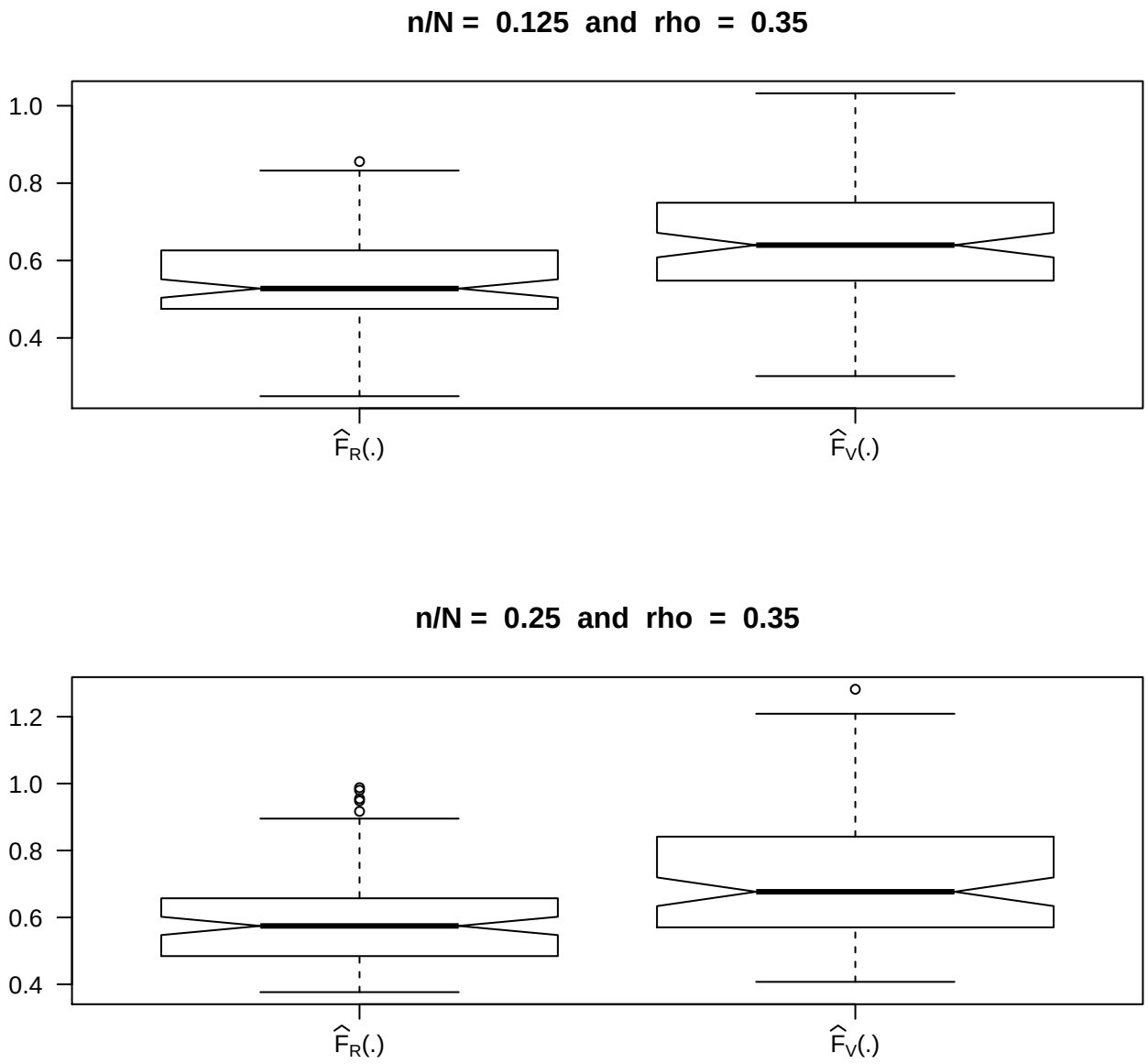

FIGURE 2.4 – A boxplots of the MSE of $\hat{F}_R^x(y)$ and $\hat{F}_V^x(y)$

Figure 2.4 displays the boxplot of MSE. It can be seen from this figure that our estimator $\hat{F}_R^x(y)$ remains more stable than $\hat{F}_V^x(y)$, and we can conclude to good asymptotic performance of $\hat{F}_R^x(y)$.

Conclusion

This paper presents the conditional distribution function's estimator using the kernel method for a surrogated scalar response variable given a functional random one. This estimator is built from the validation data. We obtained the uniform, almost complete convergence of this model using kernel estimate and the conditional quantile estimator under some classical assumptions. To improve the performance of our proposed estimator and the theoretical results, we realized a simulation study. Other research issues are possible, such as extensions to local linear method estimation and the semiparametric linear regression model which can also be studied using this kind of data. Finally, the k nearest neighbor method can be adapted to treat the outliers in the data set as proposed in the literature by Attouch et al.

Acknowledgment :

The authors greatly thank the Editor in chief and the reviewers for the careful reading, constructive comments and relevant remarks which permit us to improve the paper.

CHAPITRE 3

NORMALITY AND SQUARE ERROR OF THE CONDITIONAL DISTRIBUTION FUNCTION FOR SURROGATE

3.1 Introduction

Recently, increasing attention has been paid to cohort studies with a validation sample, where true observations are measured only on a sample from the full study cohort. Studies of this type may arise when collecting complete observa- tions for the entire cohort is difficult, time consuming or expensive. Although in such studies complete true observations can only be obtained on a subset of the cohort, surrogate observations, which are more easily obtained using some relatively simple measuring methods, are available on every study subject. For example, damage to the heart muscle caused by a myocardial infarction can be assessed accurately using arterioscintography. However, this is an invasive and expensive procedure. Instead, the peak cardiac enzyme level in the bloodstream is a more easily obtained variable, and is frequently used as a surrogate measure of heart muscle damage. Expensive and invasive arterioscintography can only be performed for a small subset of the full study cohort, to yield an accurate measure of damage to the heart muscle. The exact measurements obtained by expensive arterioscintography for a small subset of subjects, together with their surrogate observations, are usually treated as validation sample, and the remaining surrogate observations are called primary data. In an example of Rosner et al, related to the nurse health study, long-term dietary saturated fat,X, is an important variable. The primary data set consisted of a cohort of 89538 women, but instead of observing X, a $\tilde{x}$ was observed, namely, a self-administered quantitative food surrogate X and $\tilde{x}$. To understand the relationship between X and $\tilde{x}$ were observed. nurses became part of a validation study, in which, X and $\tilde{x}$.

We address the following question : how to incorporate the information contained in the primary data set and the validation sample into the estimation of the probability density of the true variable X ? Throughout this paper, all the probability densities are assumed to be with respect to Lebesgue measure. There is a vast literature on nonparametric density estimation when all data are observed exactly. For discussions of nonparametric density estimators among others. The problem of nonparametric estimation of a probability density function when the sample observations are measured with error is also studied by many authors. They estimate the F of a random variable X by kernel density and deconvolution based $\tilde{X} = X + \epsilon$, where on independent and identically distributed observations from X is a measurement error with a known distribution.

Generally, however, the relationship between the surrogate variables and the true variables can be rather complicated compared to the additive error model assumed above. That is, the additive error model with error distribution known may not be true in practice. In these

cases, the deconvolution kernel density estimators mentioned before may not be used directly here. One solution is to use the help of validation data to capture the underlying relation between $\tilde{X}$ With the help of validation data, some statisticians developed and X. statistical inference techniques for various statistical models without specifying any error structure and the distribution assumption of the true variable given the surrogate variable. Stefanski and Carroll applied conditional scores and optimal scores to generalized linear measure- ment error model. Carroll and Wand developed a semiparametric approach using the kernel regression technique for logistic measurement error models.applied the method to nonlinear errors-in-variables models. Wang extended it to partial linear errors-in-covariables models and Wang and Rao developed empirical likelihood approach in linear error- in-covariables models. Chen proposed an estimation procedure for the Cox model with incomplete covariate data. Wang and Härdle developedempirical likelihood-based dimension reduction inference for linear error-in-responses models. To the best of our knowledge, the problem of nonparametric estimation of the probability density of X to incorporate information contained in both the validation data and the primary data has not been considered. Intuitively, the information contained in the surrogate variates would be useful to recover part of the efficiency that is lost by incomplete observations. In this paper, we first develop a regression calibration kernel approach to define the estimator of the probability density f of X such that information contained in both the validation sample and the primary data can be incorporated. This approach defines the probability density estimator to be the standard probability density kernel estimator with the terms, where X_j are not available, replaced by the kernel estimators of the regression functions of $\tilde{X}_j$, respectively. Then, we define two weighted estimators based the terms on X on two regression calibration estimators and the standard probability density estimator by weighting, respectively, where the standard probability density estimator is defined with the exact observations in the validation data set. All the estimators are proved to be asymptotically normal. The two weighted estimators are proved to have less asymptotic variances but generally have bigger bias than the regression calibration kernel estimator. Also, we establish the asymptotic representations of the mean square error and mean integrated square error of $F^n(t)$.

3.2 Estimator of the c.d.f. with surrogate data

Let N and n be the respective sizes of the full sample and the validation set,V the index set of the sampled validation set and $\bar{V}$ = $\{1, ..., N\} \setminus V$. We assume that the observations are independent and identically distributed.
Our target, in this article, is to estimate is the conditional distribution function $F^x(y)$

when we have unavailable response variable Y for this kind of data we have :

$$E\left[H\left(\frac{y-Y_j}{h_H}\right)\Big|\widetilde{Y}_j, X_j\right] \xrightarrow{h_H\to 0} F^x(y)$$

as well

$$E\left[E\left[H\left(\frac{y-Y_j}{h_H}\right)\Big|\widetilde{Y}_j, X_j\right]\Big|X_j = x\right] = E\left[H\left(\frac{y-Y_j}{h_H}\right)\Big|X_j = x\right], \tag{3.1}$$

so, by using the kernel method we can defined the distribution function estimator by :

$$\widehat{F}^x(y) = \frac{\sum_{i\in V} H\left(\frac{y-Y_i}{h_H}\right) K\left(\frac{d(x,X_i)}{h_K}\right) + \sum_{j\in\bar{V}} u(X_j,\widetilde{Y}_j) K\left(\frac{d(x,X_j)}{h_K}\right)}{\sum_{i=1}^{N} K\left(\frac{d(x,X_i)}{h_K}\right)}$$

where, $W(.,.)$ is a kernel function in $\mathbb{R}^2$ and a_n is a sequence of real number which tend to zero when n tend to infinty, and

$$u(X_j,\widetilde{Y}_j) = E\left[H\left(\frac{y-Y_j}{h_H}\right)\Big|\widetilde{Y}_j, X_j\right].$$

Noting that the function $u(.,.)$ is unknown. So, by validation data set we have the estimator of this function :

$$\widehat{u}(X_j,\widetilde{Y}_j) = \frac{\sum_{i\in V} H\left(\frac{y-Y_i}{h_H}\right) W\left(\frac{d(X_j,X_i)}{a_n}, \frac{\widetilde{Y}_j-\widetilde{Y}_i}{a_n}\right)}{\sum_{i\in V} W\left(\frac{d(X_j,X_i)}{a_n}, \frac{\widetilde{Y}_j-\widetilde{Y}_i}{a_n}\right)}$$

3.3 Assumptions

All along this paper, when no confusion will be possible, we will denote by C and C' some strictly positive generic constants whose values are allowed to change, to prove our result we need the following assumptions :

(A1) For $h_K > 0, \mathbb{P}(X \in B(x,h_K)) =: \phi_x(h_K) > 0$,
$\phi_x(h) = f(x)\phi(h) + o(\phi(h)), \quad \forall h > 0.$

(A2) The operators $F^x(.)$ and $u(.,.)$ satisfies the holder condition, such that $\forall (y_1,y_2) \in S^2_{\mathbb{R}}$, $\forall (x_1,x_2) \in \mathcal{N}^2_x$ and $C, \beta_1, \beta_2 > 0$

(a) $|F^{x_1}(t_1) - F^{x_2}(t_2)| \le C\left(d(x_1,x_2)^{\beta_1} + |t_1-t_2|^{\beta_2}\right)$

(b) $|u(x_1,t_1) - u(x_2,t_2)| \le C\left(d(x_1,x_2)^{\beta_1} + |t_1-t_2|^{\beta_2}\right)$

(A3) (a) The bandwidths h_K and a_n satisfy

$$\lim_{N\to\infty} N\phi(h_K) = +\infty, \ \lim_{N\to\infty} \frac{\log N}{N\phi(h_K)} = \lim_{n\to\infty} \frac{\log n}{n\phi(h_K)} = 0.$$

and $\exists \beta > 0, \ \lim_{n\to\infty} n^{\beta} h_H = 0.$

(b) The size n and N of the sample satisfy

$$\frac{n}{N} = \alpha \text{ and } \frac{a_n}{h_K} = O(1).$$

(A4) The kernel $K(.)$ and $W(.,.)$ are a bounded continuous lipschitz function and $\int K = 1$, such that

$$0 < C_1 \leq K \leq C_2 < \infty, \quad C_1, C_2 > 0.$$

and

$$C_3\phi(a_n) \leq E\left[W\left(\frac{d(X_j, X_i)}{a_n}, \frac{\widetilde{Y}_j - \widetilde{Y}_i}{a_n}\right)\right] \leq C_4\phi(a_n),\ C_3, C_4 > 0.$$

(A5) The distribution function $H(.)$ is a bounded continuous Lipschitz function, such that

$$\int H(t)\mathrm{d}t = 1 \text{ and } \int |t|^{A_2} H'(t)\mathrm{d}t < \infty.$$

3.4 Results

1. The quadratic error of the conditional distribution function for surrogate data

   Theorem 3.1. *Under assumptions* $(A1) - (A4)$*, we obtain*

$$\mathbb{E}\left(\widehat{F}^x(y) - F^x(y)\right)^2 = o\left(\frac{\alpha}{N\phi(h_K)^{3/2}\phi(a_n)^{3/2}}\right) + o\left(\frac{1-\alpha}{N\phi(h_k)\phi(h_H)}\right) + o\left(h_H^{\beta}h_K^{\beta}\right) + o(h_K h_H)$$

2. The asymptotic normality of the conditional distribution function for surrogate data

   Theorem 3.2. *Under assumptions* $(A1) - (A4)$*, we obtain*

$$\sqrt{Nh_N}\left(\widehat{F}^x(y) - F^x(y)\right)^2 \xrightarrow{L} \mathcal{N}\left(0, \sigma^2(x)\right)$$

   where

$$\sigma_p^2(x) = \frac{\alpha_2(x)\lambda_p(\theta(p;x);x)}{\alpha_1^2(x)}\Gamma_p^2(\theta(p;x);x) \qquad \left(\textit{with } \alpha_j(x) = K^j(1) - \int_0^1 (K^j)'(s)\beta_x(s)ds \textit{ for } j\right.$$

   and

$$\lambda_p(\theta(p;x);x) = \left(\frac{p}{1-p}\right)^2 R_+^x(\theta(p;x)) + R_-^x(\theta(p;x))$$

where

$$R_+(\theta(p;x);x) = \mathbb{E}\left[(Y_1 - \theta(p;x))^2 \mathbf{1}_{(Y_1>\theta(p;x)}|X = x\right]$$
$$R_-(\theta(p;x);x) = \mathbb{E}\left[(Y_1 - \theta(p;x))^2 \mathbf{1}_{(Y_1\leq\theta(p;x)}|X = x\right]$$

and

$$\Lambda_p(\theta(p;x);x) = A_1'(\theta(p;x);x) - \left(\frac{p}{1-p}\right) A_2'(\theta(p;x);x).$$

3.5 Appendix

Proof of Theorem 4.1.

To clarify, we put

$$K_{n,1,i}(x) = \frac{K\left(\frac{d(x,X_i)}{h_K}\right)}{\sum_N^{i=1} K\left(\frac{d(x,X_i)}{h_K}\right)} \quad \text{for} i = 1,\ldots,N$$

$$K_{n,2,j}(x) = \frac{W\left(\frac{d(x,X_j)}{a_n}, \frac{\tilde{y}-Y_j}{a_n}\right)}{\sum_{l\in V} W\left(\frac{d(x,X_l)}{a_n}, \frac{\tilde{y}-Y_l}{a_n}\right)} \quad \text{for} j \in V$$

and

$$H_i(y) = H\left(\frac{y - Y_i}{h_H}\right).$$

By a straightforward calculation we obtain

$$\begin{aligned}
\hat{F}^x(y) - F^x(y) &= \sum_{j\in\bar{V}} K_{n,1,j}(x) \sum_{i\in V} K_{n,2,i}\left(X_j, \tilde{Y}_j\right)\left[H_i(y) - u\left(X_i, \tilde{Y}_i\right)\right] && (3.2)\\
&+ \sum_{j\in\bar{V}} K_{n,1,j}(x) \sum_{i\in V} K_{n,2,i}\left(X_j, \tilde{Y}_j\right)\left[u\left(X_i, \tilde{Y}_i\right) - u\left(X_j, \tilde{Y}_j\right)\right] && (3.3)\\
&+ \sum_{i\in V} K_{n,1,i}(x)\left[F^x(Y_i) - F^x(y)\right] + \sum_{i\in V} K_{n,1,i}(x)\left[H_i(y) - F^x(Y_i)\right] && (3.4)\\
&+ \sum_{j\in\bar{V}} K_{n,1,j}(X)\left[F^x(Y_j) - F^x(y)\right] + \sum_{j\in\bar{V}} K_{n,1,j}(X)\left[u\left(X_j, \tilde{Y}_j\right) - F^x(Y_j)\right] && (3.5)\\
&:= \sum_{k=1}^{6} G_k
\end{aligned}$$

We can easily apply the Cauchy-Schwarz's inequality , there exist positive constant C, such as

$$\mathbb{E}\left[\hat{F}^x(y) - F^x(y)\right]^2 \leq C \sum_{k=1}^{6} \mathbb{E}G_k^2$$

First for $\mathbb{E}G_1^2$ with the same techniques used in Ibrahim, and the fact that $H_i(y)$ –

$u\left(X_i, \tilde{Y}_i\right)$ is bounded ,so we can also use the Cauchy-Schwarz's inequality, we get

$$\begin{aligned} \mathbb{E}G_1^2 &\leq C \sum_{i\epsilon V} \mathbb{E}\left(\sum_{j\epsilon \bar{V}} K_{n,1,j}(x) K_{n,2,i}\left(X_j, \tilde{Y}_j\right)\right)^2 \\ &\leq Cn \sum_{i\epsilon V} \sum_{j\epsilon V} \mathbb{E}^{\frac{1}{2}} K_{n,1,j}^4(x) \mathbb{E}^{\frac{1}{2}} K_{n,1,j}^4\left(X_j, \tilde{Y}_j\right) \end{aligned} \tag{3.6}$$

we have

$$K_{n,1,j}(x) = \frac{1}{N\mathbb{E}K\left(\frac{d(x,X_j)}{h_K}\right)} \frac{K\left(\frac{d(x,X_j)}{h_K}\right)}{\hat{g}_N(x)}$$

whith

$$\hat{g}_N(x) = \frac{1}{N\phi(h_K)} \sum_{j=1}^{N} K\left(\frac{d(x,X_j)}{h_K}\right)$$

then

$$\mathbb{E}\left(\frac{K^4\left(\frac{d(x,X_j)}{h_K}\right)}{\hat{g}_N^4(x)}\right) \leq C\mathbb{E}K^4\left(\frac{d(x,X_j)}{h_K}\right) + P\left(\hat{g}_N^4(x) \leq C\right)$$

We start by computing the first term, Under regularity assumptions and hypothese 3 we find

$$\begin{aligned} \mathbb{E}^{\frac{1}{2}} K_{n,1,j}^4(x) &\leq \frac{C}{N^2\phi(h_K)} \\ &= O\left(\frac{1}{N\phi(h_K)^{\frac{1}{2}}}\right) \end{aligned} \tag{3.7}$$

then for the second part we have

$$\begin{aligned} \mathbb{P}\left(\hat{g}_N^4(x) \leq C\right) &\leq P\left(|\hat{g}_N(x) - 1| \geq \frac{1}{2}\right) \\ &\leq 4var\left(\hat{g}_N(x)\right) = O\left(\frac{1}{N\phi(h_K)}\right) \end{aligned} \tag{3.8}$$

in other side ,and by same concept of $\mathbb{E}^{\frac{1}{2}} K_{n,1,j}(x)$ we obtain

$$\mathbb{E}^{\frac{1}{2}} K_{n,2,j}^4 \leq \frac{C\phi(a_n)^{\frac{1}{2}}}{n^2\phi(a_n)^2} = o\left(\frac{1}{n\phi(a_n)^{\frac{1}{2}}}\right)$$

we have

$$\mathbb{E}^{\frac{1}{2}} K_{n,2,j}^4(x_j, \tilde{y}_j) = \mathbb{E}^{\frac{1}{2}}\left(\frac{\frac{1}{n^4\phi(a_n)^4} W^4\left(\frac{d(x,X_j)}{a_n}, \frac{\tilde{y}-\tilde{Y}_j}{a_n}\right)}{g_{N'}(x,\tilde{y})^4}\right)$$

with

$$g_{N'}(x,\tilde{y}) = \frac{1}{n\phi(a_n)}\sum_{i\epsilon V} W\left(\frac{d(x,X_i)}{a_n}, \frac{\tilde{y}-\tilde{Y}_i}{a_n}\right)$$

finally by substituting $\mathbb{E}^{\frac{1}{2}} K_{n,1,j}(x)$ and $\mathbb{E}^{\frac{1}{2}} K_{n,2,j}\left(X_j, \tilde{Y}_j\right)$ in $\mathbb{E}G_1^2$ we reach at

$$\begin{aligned}\mathbb{E}G_1^2 &\leq Cn\sum_{i\epsilon V}\sum_{j\epsilon V}\frac{1}{N^2\phi(h_K)^{\frac{3}{2}}}\frac{1}{n^2\phi(a_n)^{\frac{3}{2}}} \\ &= C\frac{n}{N^2\phi(h_K)^{\frac{3}{2}}\phi(a_n)^{\frac{3}{2}}}\end{aligned} \tag{3.9}$$

under supposition(5) we get

$$\mathbb{E}G_1^2 = O\left(\frac{\alpha}{N\phi(h_K)^{\frac{3}{2}}\phi(a_n)^{\frac{3}{2}}}\right).$$

Similarly to (Wang (2006) , Ibrahim(2019)) , and by the conditions imposed on W and , we have

$$|G_2| \leq \sum_{j\epsilon\bar{V}} K_{n,1,j}(x)\sum_{i\epsilon V}(\nu_j)\left\|\nu_i-\nu_j\right\|_\infty \leq a_n$$

with $\nu_k = \left(X_k, \tilde{Y}_k\right)$
move now to G_3

$$\left|\sum_{i\epsilon V} K_{n,1,i}(x)\left(F^x(Y_i) - F^x(y)\right)\right| \leq Ch_K^{\beta_1}h_H^{\beta_2}\sum_{i\epsilon V\cup\bar{V}} K_{n,1,i}(x) = Ch_K^{\beta_1}h_H^{\beta_2}$$

that's because of the kernel K is is supported into $(0,1)$ and through assumption(lipzn) we have

$$\begin{aligned}|K_{n,1,i}(x)F^x(Y_i) - F^x(y)| &\leq K_{n,1,i}(x)\sup_{t\epsilon B(x,h)}|F^x(t) - F^x(y)| \\ &\leq K_{n,1,i}(x)\sup_{t\epsilon B(x,h)} d(x,t)^{\beta_1} + |y1-y2|^{\beta_2} \leq CK_{n,1,i}(x)h_K^{\beta_1}h_H^{\beta_2}\end{aligned} \tag{3.10}$$

from that we achieve to

$$G_3 = O\left(h_K^{\beta_1}\right)O\left(h_H^{\beta_2}\right)$$

also by the same way we get $G_5 = O(h_K)O(h_H)$

it remains now to calculate G_4 and G_6

CHAPITRE 4

ASYMPTOTIC PROPERTIES OF THE KERNEL TYPE EXPECTILE REGRESSION ESTIMATOR FOR SURROGATE DATA

4.1 Introduction

Although quantile regression (QR) and expectile regression (ER) have been introduced in quantile is better known expectile . On the other hand, the the literature at almost the same time, the risk measure as and most developed in a rather rapid way than the risk measure quantiles are not always satisfactory and can be criticized for their difficulty to calculate because the corresponding loss function is not everywhere differentiable, the traditional estimation methods, for example the traditional estimation methods, for example the Gauss-Newton algorithm, are no longer applicable for generating estimators of the QR. To circumvent this problem, Koenker and Bassett reformulated the optimization problem in the framework of linear programming, (See Koenker and Bassett (1978)). The main advantage of the expectation over the quantile is its efficiency, ease of computation and estimation.

Note that it is easy to see that conditional expectations are characterized by tail conditional expectations in the same way that conditional quantiles are characterized by the conditional distribution function, for more details,We refer the interested reader to the work of Newey and Powell (1987) and Abdous (1999). In addition, the conditional expectation et al. (2003). considers sensitivity to outliers to be an important advantage, because if we important advantage because if we are measuring potential losses, we want our measure to be sensitive to extreme and outlier values. Therefore, this advantage allows for more conservative and responsive risk management. and reactive risk management.

The most popular method to involve missing data is the imputation method that fills or retrieves the missing data in the response variable Y. In this context, we can cite various works that used this technique : We can cite, Yates (1933) for the linear regression model. The kernel estimation of the mean functions is considered in Cheng (1994), the nearest neighbor imputation for the data survey is addressed in Chen and Shao (2000), the robust regression model with missing data is considered in Pérez-González et al. (2009), the asymptotic properties of the regression operator estimator when the regressor is functional and completely observed, and that missing data at random in the scalar response variable are investigated in Ferraty et al. (2013), in the case of dependent data, the reader may refer to Ling et al. (2015). In this work, we investigate the unavailability of response data because sometimes it is default or very expensive to measure some response observations ; the main idea is to recover (or fill) this missing data by a surrogate validation data set. In this context, we cite Duncan and Hill (1985), Wittes et al. (1989), Carroll and Wand (1991) and Pepe (1992). The principle of this method is to incorporate both surrogate data and the corresponding observations of the covariate X.

This paper aims to study the conditional models (conditional distribution function and the conditional quantile) for missing response by the kernel method ; we explore in this work, the aspect of missing data in the response variable. First, we consider the estimator of the conditional distribution for complete data, then by using the validation data set (see, Ibrahim et al. (2020) and Wang (2006), we build our new estimator with surrogate data and we obtain some asymptotic results for the conditional distribution and the quantiles. In the end, we realized a simulation study to improve the efficacity of our estimator.

The rest of the paper is organized as follows. We present our model in Section 2 ; the required notations and assumptions are introduced in Section 3, the main results of strong uniform consistency (with rate) and the quantile estimation as a direct consequence of our asymptotic result obtained from CFD estimation are formulated in section 4. For the numerical results, a simulation study that shows the performance of the proposed estimator is presented in Section 5.

4.2 Model and estimators

Let (X, Y) be a pair of random variable valued in $\mathcal{F} \times \mathbb{R}$, where $(\mathcal{F}, d)$ is semi-metric space equipped with a semi metric $d(.,.)$ defining a topology to me asure the proximity between two elements of $\mathcal{F}$ and which is disconnected of the definition of X in order to avoid measurability problems. In the following, we fix a point x in $\mathcal{F}$ and $p \in]0, 1[$.the conditional expectile of order p has been introduced by Newey and Powell (1987) as the minimizer of an asymmetric quadratic loss

$$\xi_p(x) = \underset{t\in\mathbb{R}}{\operatorname{argmin}}\, \delta_p(x,t),$$

where

$$\delta_p(x,t) = \mathbb{E}[p(Y-t)^2\mathbf{1}_{(Y-t)>0}{}'X = x] + \mathbb{E}[(1-p)(Y-t)^2\mathbf{1}_{(Y-t)\leqslant 0}{}'X = x],$$

where $\mathbf{1}_Z$ is the indicator function of set Z, we can show that ξ_p is the solution of

$$\frac{p}{1-p} = \frac{A_1(x,t)}{A_2(x,t)}$$

where

$$\begin{cases} A_1(x,t) = -\mathbb{E}[(Y-t)\mathbf{1}_{(Y-t)\leqslant 0}{}'X = x], \\ A_2(x,t) = \mathbb{E}(Y-t)\mathbf{1}_{(Y-t)>0}{}'X = x]. \end{cases} \tag{4.1}$$

we use the fact that the function $A(x,t) := \frac{A_1(x,t)}{A_2(x,t)}$ is an increasing function so we can express the conditional expectileξ_p of order p as follows :

$$\xi_p = \inf\{t \in \mathbb{R} : A(x,t) \geqslant \frac{p}{1-p}\}.$$

To build an estimator of the conditional expectile of order p when there are missing data in the response variable, let N and n ($n < N$) the respetive sizes of the sample set and the validation set, we assume that the observations are independent and identically distributed, V is the index set of individuals in the sampled validation set and $\bar{V} = \{1,2,\ldots,N\} - V$. Since

$$\begin{cases} \mathbb{E}[\mathbb{E}\{(Y_j-t)\mathbf{1}_{(Y_j-t)\leqslant 0}{}'X_j, \widetilde{Y}_j\}] = \mathbb{E}[(Y_j-t)\mathbf{1}_{(Y_j-t)\leqslant 0}{}'X_j = x] = A_1(x,t) \\ \mathbb{E}[\mathbb{E}\{(Y_j-t)\mathbf{1}_{(Y_j-t)>0}{}'X_j, \widetilde{Y}_j\}] = \mathbb{E}[(Y_j-t)\mathbf{1}_{(Y_j-t)>0}{}'X_j = x] = A_2(x,t) \end{cases}$$

where $\widetilde{Y}$ is a surrogate variable of Y.

we can estimate ξ_p by

$$\widehat{\xi}_p(x) = \inf\{t \in \mathbb{R} : \widehat{A}(x,t) \geqslant \frac{p}{1-p}\}$$

we define

$$\widehat{A}(x,t) := \frac{\widehat{A}_1(x,t)}{\widehat{A}_2(x,t)}$$

with

$$\begin{cases} \widehat{A_1}(x,t) = \frac{\sum_{i\in V} K(h_K^{-1}d(x,X_i))(Y_i-t)\mathbf{1}_{(Y_i-t)\leqslant 0}+\sum_{j\in\bar{V}} u_1(X_j,Y_j)K(h_K^{-1}d(x,X_i))}{\sum_{i=1}^{N} K(h_K^{-1}d(x,X_i))} \\ \widehat{A_2}(x,t) = \frac{\sum_{i\in V} K(h_K^{-1}d(x,X_i))(Y_i-t)\mathbf{1}_{(Y_i-t)>0}+\sum_{j\in\bar{V}} u_2(X_j,Y_j)K(h_K^{-1}d(x,X_i))}{\sum_{i=1}^{N} K(h_K^{-1}d(x,X_i))} \end{cases}$$

where

$$\begin{cases} u_1(X_j,\widetilde{Y}_j) = \mathbb{E}\{(Y_j-t)\mathbf{1}_{(Y_j-t)\leqslant 0}{}'X_j,\widetilde{Y}_j\} \\ u_2(X_j,\widetilde{Y}_j) = \mathbb{E}\{(Y_j-t)\mathbf{1}_{(Y_j-t)>0}{}'X_j,\widetilde{Y}_j\} \end{cases}$$

for $j \in \bar{V}$

K is kernel function and h_n is a bandwidth sequence tending to zero when N goes to infinity. Recall that the functions $u_1(.,.)$ and $u_2(.,.)$ are unknown,so we estimate those functions by validation data set :

$$\begin{cases} \widehat{u_1}(X_j,\widetilde{Y}_j) = \frac{\sum_{i\in V}(Y_i-t)\mathbf{1}_{(Y_i-t)\leqslant 0}W(\frac{d(X_j,X_i)}{a_n},\frac{\widetilde{Y}_j-\widetilde{Y}_i}{a_n})}{\sum_{i\in V} W(\frac{d(X_j,X_i)}{a_n},\frac{\widetilde{Y}_j-\widetilde{Y}_i}{a_n})} \\ \widehat{u_2}(X_j,\widetilde{Y}_j) = \frac{\sum_{i\in V}(Y_i-t)\mathbf{1}_{(Y_i-t)>0}W(\frac{d(X_j,X_i)}{a_n},\frac{\widetilde{Y}_j-\widetilde{Y}_i}{a_n})}{\sum_{i\in V} W(\frac{d(X_j,X_i)}{a_n},\frac{\widetilde{Y}_j-\widetilde{Y}_i}{a_n})}, \qquad \text{for} \quad j\in\bar{V} \end{cases}$$

$W(.,.)$ is a kernel function which is define on $\mathbb{R}^2$ and a_n is a sequence of real number which tend to zero when n tend to infinity.

4.3 Asymptotic properties of the estimator

4.3.1 The uniform almost-complete convergence

To establish the uniform almost-complete convergence of the estimator $\widehat{\xi_p}(x)$, over a fixed subset $S_{\mathcal{F}}$ of $\mathcal{F}$ For this, we denote by $\psi_{S_{\mathcal{F}}}(.)$ the Kolmogorov's $\epsilon-$entropy of $S_{\mathcal{F}}$, we introdus the closed ball centered at x with radius h by $B(x,h) = \{z \in \mathcal{F} : d(z,x) \leqslant h\}$ Our assumptions are gathered below for easy references.

(A1) $\forall h > 0, \mathbb{P}(X \in B(x,h)) = \phi(h) > 0$

(A2) For $i = 1,2$, the operators A_i and u_i are differentiable in $\mathbb{R}$ and satisfies the following Lipschitz's condition : for all $(t_1,t_2) \in \mathbb{R}$ and for all $x_1,x_2 \in \mathcal{F}$

$$|M_i(x_1,t_1) - M_i(x_2,t_2)| \leqslant C\left(d^{k_i}(x_1,x_2) + |t_1-t_2|^{\zeta_i}\right)$$

for some $\zeta_i, k_i > 0$.

(A3) For all $m \geqslant 2, \varphi_m(Y^-)\mathbb{E}[|Y^-|^m|X=x] \leqslant C < \infty$ a.s., With$Y^- = (Y-t)\mathbf{1}_{(Y_i-t)\leqslant 0}$

(A4) The bandwidths h_K and a_n satisfy

$$\lim_{N\to\infty} h = \lim_{n\to\inf} a_n = 0 \quad \text{and} \quad \lim_{N\to\infty} N\phi(h) = +\infty$$

and

$$\lim_{N\to\infty} \frac{\log N}{N\phi(h_K)} = 0 \quad \text{and} \quad \lim_{n\to\infty} \frac{\log n}{n\phi(a_n)} = 0.$$

(A5) the functions ϕ_x and $\psi_{S_\mathcal{F}}$ are such that :

($A5_1$) $\exists C > 0, \exists \eta_0 > 0, \forall \eta < \eta_0, \phi'_x(\eta) < C$, and if $K(1) = 0$, the function $\phi_x(.)$ has to fulfill the additional condition :

$$\exists C > 0, \exists \eta_0 > 0, \forall 0 < \eta < \eta_0, \int_0^\eta \phi_x(u)du > C\eta\phi_x(\eta).$$

($A5_2$) For N large enough,

$$\frac{(\log N)^2}{N\phi_x(h)} < \psi_{S_\mathcal{F}}\left(\frac{\log N}{N}\right) < N\frac{\phi_x(h)}{\log N}.$$

Kolmogorov's $\epsilon-$entropy of $S_\mathcal{F}$ satisfies

$$\sum_{n=1}^{\infty} n^\beta \exp\left\{(1-\eta)\,\psi_{S_\mathcal{F}}\left(\frac{\log N}{N}\right)\right\} < \infty \text{ for some } \beta > 0 \quad \text{and} \quad \eta > 1.$$

(A6) The kernel $K(.)$ is a continuous function from $\mathbb{R}$ into $\mathbb{R}^+$, such that $\int K = 1$, and there exist some positive constants C and C' such that

$$C\mathbf{1}_{(0,1)} \le K \le C'\mathbf{1}_{(0,1)} \tag{4.2}$$

where $\mathbf{1}_A$ denotes the indicator function on the set A.
We assume the two-dimensional kernel $W(x,y) = W_1(x)W_2(y)$ is a continuous function with a compact support satisfies (4.2), however, there exist positive finite real constants C_3 and C_4, such that

$$C\phi(a_n) \le \mathbb{E}\left[W\left(\frac{d(X_j,X_i)}{a_n}, \frac{\widetilde{Y}_j - \widetilde{Y}_i}{a_n}\right)\right] \le C\phi(a_n).$$

Remark 4.3.1.

Theorem 4.1. *Under the assumptions (A1)-(A5), and if in addition*

$$\frac{\partial A(x,\xi_p(x))}{\partial t} > 0$$

then

$$\sup_{x\in\mathcal{F}}|\widehat{\xi_p}(x)-\xi_p(x)| = O(h^{k_{l_K}}) + O(a^{\min k_{l_n}})$$

$$+O_{a.co.}\left(\sqrt{\frac{\log n}{n\phi(a_n)}}\right) + O_{a.co.}\left(\sqrt{\frac{\log N}{N\phi(h_K)}}\right).$$

Proof of Theorem :

for $\tau > 0$ small enough, we use the fact that the function $A(x,t)$ is an increasing function and has a strictly positive in the nieghbourhood of $\widehat{\xi_p}(x)$, we have

$$\sum_n \mathbb{P}\left(\sup_{x\in\mathcal{F}}|\widehat{\xi_p}(x)-\xi_p(x)| > \tau\right)$$

$$\leqslant \sum_n \mathbb{P}\left(|\widehat{A}(x,\xi_p(x)-\tau) - A(x,\xi_p(x)-\tau)| \geqslant C\tau\right)$$

$$+ \leqslant \sum_n \mathbb{P}\left(|\widehat{A}(x,\xi_p(x)+\tau) - A(x,\xi_p(x)+\tau)| \geqslant C\tau\right)$$

so, the proof the theoreme 1 is based on the following proposition

Proposition 4.1. *Under the assumptions (A1)–(A5), we have,for certain* $\delta > 0$

$$\sup_{x\in\mathcal{F}}\sup_{t\in[\xi_p(x)-\delta,\xi_p(x)+\delta]}|\widehat{A}(x,t)-A(x,t)| = O(h_K^{k_l}) + O(a_n^{k_l})$$

$$+O_{a.co.}\left(\sqrt{\frac{\psi_{S_{\mathcal{F}}}\left(\frac{\log N}{N}\right)}{n\phi(a_n)}}\right) + O_{a.co.}\left(\sqrt{\frac{\log N}{N\phi(h_K)}}\right).$$

Proof of Proposition :

the proof of this proposition is based on the following decomposition

$$\widehat{A}(x,t)-A(x,t) = \frac{1}{\widehat{A}_2(x,t)}\left[\widehat{A}_1(x,t)-A_1(x,t)\right] + \frac{A(x,t)}{\widehat{A}_2(x,t)}\left[A_2(x,t)-\widehat{A}_2(x,t)\right]$$

The proof of proposition becomes a straightforward consequence of the following lemmas.

Lemma 4.3.1. *under the assumptions (A1)-(A5) we have*

$$\sup_{x\in\mathcal{F}}\sup_{t\in[\xi_p(x)-\delta,\xi_p(x)+\delta]}|\widehat{A}_1(x,t)-\mathbb{E}[\widehat{A}_1(x,t)]| = O_{a.co.}\left(\sqrt{\frac{\log N}{N\phi_x(h)}}\right)$$

Proof of Lemma1 :

We have for all $x \in S_{\mathcal{F}}$, we set $k(x) = \underset{k\in 1,\ldots,d_n}{\operatorname{argmin}}|x-x_k|$, and since $[\xi_p(x)-\delta,\xi_p(x)+\delta]$ is a compact subst of $\mathbb{R}$ it can covered by a finite number s_n of intervals of length l_n at some

point $t_\kappa, \kappa = 1, \dots, S_n, \quad i.e \quad [\xi_p(x) - \delta, \xi_p(x) + \delta] \subset \bigcup_{\kappa=1}^{s_n}$ where $I_\kappa = [t_\kappa - l_n, t_\kappa + l_n]$. The subset is bounded, then there exists a constant $M_0 < \infty$, such that $s_n l_n \leqslant M_0$. Let $t_{s(t)} = \underset{s \in l_1, \dots, l_{q_n}}{\operatorname{argmin}} |t - t_{s(t)}|$.
Using the triangle inequality we get for $\epsilon > 0$

$$
\begin{aligned}
&\mathbb{P}\left(\sup_{x\in\mathcal{F}}\sup_{t\in[\xi_p(x)-\delta,\xi_p(x)+\delta]}|\widehat{A}_1(x,t)-\mathbb{E}[\widehat{A}_1(x,t)]|>\epsilon\right)\\
&\leqslant \mathbb{P}\left(\sup_{x\in\mathcal{F}}\sup_{t\in[\xi_p(x)-\delta,\xi_p(x)+\delta]}|\widehat{A}_1(x,t)-\widehat{A}_1(x_{k(x)},t)|>\epsilon\right)\\
&+\mathbb{P}\left(\sup_{x\in\mathcal{F}}\sup_{t\in[\xi_p(x)-\delta,\xi_p(x)+\delta]}|\widehat{A}_1(x_{k(x)},t)-\widehat{A}_1(x_{k(x)},t_{s(t)})|>\epsilon\right)\\
&+\mathbb{P}\left(\sup_{x\in\mathcal{F}}\sup_{t\in[\xi_p(x)-\delta,\xi_p(x)+\delta]}|\widehat{A}_1(x_{k(x)},t_{s(x)})-\mathbb{E}[\widehat{A}_1(x_{k(x)},t_{s(t)})]|>\epsilon\right)\\
&+\mathbb{P}\left(\sup_{x\in\mathcal{F}}\sup_{t\in[\xi_p(x)-\delta,\xi_p(x)+\delta]}|\mathbb{E}[\widehat{A}_1(x_{k(x)},t_{s(x)})]-\mathbb{E}[\widehat{A}_1(x_{k(x)},t)]|>\epsilon\right)\\
&+\mathbb{P}\left(\sup_{x\in\mathcal{F}}\sup_{t\in[\xi_p(x)-\delta,\xi_p(x)+\delta]}|\mathbb{E}[\widehat{A}_1(x_{k(x)},t)]-\mathbb{E}[\widehat{A}_1(x,t)]|>\epsilon\right)\\
&=: L_1 + L_2 + L_3 + L_4 + L_5
\end{aligned}
$$

if we suppose tha$\Delta_i(x) = \frac{K\left(\frac{d(x,X_i)}{h}\right)}{\mathbb{E}K\left(\frac{d(x,X_i)}{h}\right)}$, we define $\widehat{A}_1$ and $\widehat{A}_2$ by

$$
\begin{cases}
\widehat{A}_1(x,t) = \sum_{i\in V} Y_i^- \Delta_i(x) + \sum_{j\in\bar{V}} \widehat{u}_1(X_j, \tilde{Y}_j)\Delta_j(x)\\
\\
\widehat{A}_2(x,t) = \sum_{i\in V} Y_i^+ \Delta_i(x) + \sum_{j\in\bar{V}} \widehat{u}_2(X_j, \tilde{Y}_j)\Delta_j(x)
\end{cases}
$$

where $Y_i^- = (Y_i - t)\mathbf{1}_{(Y_i-t)\leqslant 0}$ and $Y_i^+ = (Y_i - t)\mathbf{1}_{(Y_i-t)>0}$.

1. For L_1 and L_5, a direct consequence of the assumption (H5)

   therefore,

$$
\begin{aligned}
L_1 &\leqslant \frac{1}{n}\sup_x\sup_t\left(\sum_{i\in V}|\Delta_i(x)Y_i^- - \Delta_i(x_k)Y_i^-| + \sum_{j\in\bar{V}}|\widehat{u}_1(X_j,\tilde{Y}_j) - \widehat{u}_1(X_{kj},\tilde{Y}_j)|\right)\\
&\leqslant L_{1,1} + L_{1,2}.
\end{aligned}
$$

   For the first side we use the fact that :

$$C\phi_x(h) \leqslant \mathbb{E}[K_1(x)] \leqslant C'\phi_x(h)$$

$$L_{1,1} \leqslant \frac{C}{n\phi_x(h_n)} \sup_x \sup_t \left(\sum_{i\in V} |K_i(x) - K_i(x_k)| Y_i^- \mathbf{1}_{B(x,h)\cup B(x_k,h)(X_i)} \right)$$
$$\leqslant C \sup_x \sup_t (l_1 + l_2 + l_3),$$

with

$$l_1 = \frac{1}{n\phi_x(h)} \sum_{i\in V} |K_i(x) - K_i(x_k)| Y_i^- \mathbf{1}_{B(x,h)\cap B(x_k,h)}(X_i),$$

$$l_2 = \frac{1}{n\phi_x(h)} \sum_{i\in V} K_i(x) Y_i^- \mathbf{1}_{B(x,h)\cap \overline{B(x_k,h)}}(X_i),$$

$$l_3 = \frac{1}{n\phi_x(h)} \sum_{i\in V} K_i(x_k) Y_i^- \mathbf{1}_{\overline{B(x,h)}\cap B(x_k,h)}(X_i),$$

we have that the kernel K is a Lipschitzian function on $(0,1)$ then we can write

$$l_1 \leqslant \sup_x \sup_t \frac{C}{n} \sum_{i=1}^{n} Z_i \quad \text{with} \quad Z_i = \frac{\epsilon}{h\phi_x(h)} Y_i^- \mathbf{1}_{B(x,h)\cap B(x_k,h)}(X_i)$$

we apply the concentration inequality (see corolaire A.8 Ferraty and Vieu(2006)) with $a^2 = \epsilon(h\phi_x(h)^{-1}$ we have :

$$l_1 = O_{a.co}\left(\sqrt{\frac{\epsilon \log N}{Nh\phi_x(h)}} \right)$$

and the combination of conditions $(A5_1)$ and $(A5_2)$ allows to simplify the convergence rate and to get

$$l_1 = O_{a.co}\left(\sqrt{\frac{\epsilon \log N}{N\phi_x(h)}} \right).$$

by the same concept we get

$$l_2 = l_3 = O_{a.co}\left(\sqrt{\frac{\epsilon \log N}{N\phi_x(h)}} \right)$$

2. concerning the term L_2, by using Lipschitz's condition, we can write

$$\widehat{A}_1(x_{k(x)}, t) - \widehat{A}_1(x_{k(x)}, t_{s(t)}) \leqslant C \frac{1}{nh\phi(h)} \sum_{i\in V} K_i(x_k) |t - t_{s(x)}|$$

$$\leqslant \frac{C}{n} \sum_{i\in V} Z_i$$

where $Z_i = \frac{I_n K_i(x_k)}{h^2\phi_x(h)}$ by the standard exponential inequality for a sum of bounded

variables and for $I_n = N^{-\beta}$ we get

$$L_2 = L_4 = O_{\text{a.co.}}\left(\sqrt{\frac{\log N}{Nh\phi(h_n)}}\right)$$

3. For the third term we have :

$$\begin{aligned} L_3 &= \sup_{x\in\mathcal{F}} \sup_{t\in[\xi_p(x)-\delta,\xi_p(x)+\delta]} \left| \frac{1}{N}\left(\sum_{i\in V} \frac{Y_i^- K\left(x_{k(x)}\right)}{\mathbb{E}[K_1]} - \mathbb{E}\left[\sum_{i\in\bar{V}} \frac{Y_i^- K\left(x_{k(x)}\right)}{\mathbb{E}[K_1]} \right] \right) \right| \\ &+ \sup_{x\in\mathcal{F}} \sup_{t\in[\xi_p(x)-\delta,\xi_p(x)+\delta]} \left| \left(\sum_{i\in\bar{V}} \frac{K\left(x_{k(x)}\right)}{\mathbb{E}[K_1]} - \mathbb{E}\left[\sum_{i\in V} \frac{K\left(x_{k(x)}\right)}{\mathbb{E}[K_1]} \right] \right) \right| \\ &:= L_{3,1} + L_{3,2} \end{aligned} \tag{4.3}$$

then, for $L_{3,1}$

$$\mathbb{P}\left(L_{3,1} > \tau\sqrt{\frac{\log n}{n\phi(h)}} \right) \leqslant q_n d_n \max_{x\in\mathcal{F}} \max_{t\in[\xi_p(x)-\delta,\xi_p(x)+\delta]} \mathbb{P}\left(\left| \frac{1}{N}\sum_{i\in V} \Lambda_i \right| > \eta \right)$$

with

$$\Lambda_i = \frac{1}{\mathbb{E}[K_1]}\left(K_i\left(x_{k(x)}\right) Y_i^-) - \mathbb{E}\left[Y_i^- K_i\left(x_{k(x)}\right) \right] \right)$$

We use of the Bernstein's exponential inequality, but first we have to evaluate asymptotically the m-th order moment of Λ_i :

$$\begin{aligned} \mathbb{E}\left(\left| \frac{1}{\mathbb{E}[K_1(x)]} K_1 Y_1^- \right|^m \right) &= \frac{1}{|\mathbb{E}[K_1(1)]|^m} \mathbb{E}\left[|K_1|^m |Y_1^-|^m \right] \\ &= \frac{1}{|\mathbb{E}[K_1(1)]|^m} \mathbb{E}\left[\mathbb{E}\left[|Y_1^-|^m | X = x \right] K_1^m \right] \\ &= \mathbb{E}\left[\varphi_m(Y_1^-)(\widehat{K}_1)^m \right] \end{aligned} \tag{4.4}$$

where we have used the notation

$$\widehat{K}_1(.) = \frac{K_i(.)}{\mathbb{E}[K_i(.)]} \quad \text{and recall that } \varphi_m(Y_1^-) = \mathbb{E}[|Y_1^-|^m | X = x]$$

By using the assumpions (A1),(A3) and (A5), we find

$$\mathbb{E}\left| (\widehat{K}_1 Y_1^-) \right|^m \leqslant C_3(\phi(h))^{-m+1}$$

which implies that

$$\mathbb{E}\left| (\widehat{K}_1 Y_1^-) \right|^m = O\left((\phi(h))^{-m+1} \right)$$

on the other hand by using the Newton's binomial formula, we obtaine that

$$\begin{aligned}\mathbb{E}\left|Y_1^- K_1(x) - \mathbb{E}\left[Y_1^- K_1(x)\right]\right|^m &= \mathbb{E}\left|\sum_{l=0}^{m} C_l^m \left(Y_1^- K_1\right)^l \left(-\mathbb{E}\left[Y_1^- K_1\right]\right)^{m-l}\right| \\ &\leqslant \sum_{l=0}^{m} C_l^m \mathbb{E}\left|Y_1^- K_1\right|^l \mathbb{E}\left|\mathbb{E}\left[Y_1^- K_1\right]\right|^{m-l} \\ &\leqslant \sum_{l=0}^{m} C_l^m \mathbb{E}\left|Y_1^- K_1\right|^l \left|\mathbb{E}\left[\mathbb{E}\left[Y_1^- K_1\right]|X_1\right]\right|^{m-l} \\ &\leqslant C\sum_{l=0}^{m} C_l^m (\phi(h))^{-l+1}\left|A_1^{X_1}(t)\right|^{m-l} \\ &\leqslant C\max_{0\leqslant l\leqslant m}(\phi(h))^{-l+1} = C(\phi(h))^{-m+1}.\end{aligned}$$

then we get

$$\mathbb{E}\left|Y_1^- K_1(x) - \mathbb{E}\left[Y_1^- K_1(x)\right]\right|^m = O\left((\phi(h))^{-m+1}\right).$$

to achieve the proof we apply the Bernstien's inequality for Λ_i, we take $q_n = O(l_n^{-1}), C\tau^2 = 2\beta + 1$, such that

$$\begin{aligned}q_n \max_{t_s}\mathbb{P}\left(\left|\frac{1}{N}\sum_{i\in V}\Lambda_i\right| > \tau\sqrt{\frac{\log N}{N\phi(h)}}\right) &\leqslant q_n 2\exp -C\tau^2\log N \\ &\leqslant CN^{\beta}N^{-2\beta-1},\end{aligned}$$

so

$$\mathbb{P}\left(L_{3,1} > \tau\sqrt{\frac{\log N}{N\phi(h)}}\right) \leqslant CN^{-\beta-1},$$

Concerning the second term, we follow the same steps as for $L_{3,1}$ by taking $(Y_i - t_s)\mathbf{1}_{(Y_i - t)\leqslant 0} = 1$ we obtain $L_{3,2}$.

So

$$L_3 = O_{a.co}\left(\sqrt{\frac{\log N}{N\phi(h)}}\right)$$

Lemma 4.3.2. *under the assumptions (A1)-(A6) we have*

$$\sup_{x\in\mathcal{F}}\sup_{t\in[\xi_p(x)-\delta,\xi_p(x)+\delta]}\left|\mathbb{E}[\widehat{A}_1(x,t)] - A_1(x,t)\right| = O(h_K^{k_l}) + O(a_n^{k_l}) + O_{a.co.}\left(\sqrt{\frac{\psi_{S_{\mathcal{F}}}\left(\frac{\log N}{N}\right)}{n\phi(a_n)}}\right)$$

Proof of lemma 4.3.2 :

Notice that we have

$$\mathbb{E}[\widehat{A}_1(x,t)] - A_1(x,t)] = n\mathbb{E}\left[\frac{Y_i^- K_1}{\mathbb{E}[K_1]}\right] - (N-n)\mathbb{E}\left[\frac{\hat{u}_1(X_j,\tilde{Y}_j)K_1}{\mathbb{E}[K_1]}\right] - (-\mathbb{E}[Y_i^-|X=x])$$
$$= T_1 + T_2.$$

- Concerning the first term :

$$\mathbb{E}[Y_i^-|X=x] - \mathbb{E}\left[\frac{Y_i^- K_1}{\mathbb{E}[K_1]}\right] = \frac{1}{\mathbb{E}[K_1(x)]}\{\mathbb{E}\left[K_1(x)\mathbb{E}[Y_i^-|X=x]\right] - \mathbb{E}\left[Y_i^-|X_1\right]\}$$
$$= \frac{1}{\mathbb{E}[K_1(x)]}\{\mathbb{E}\left[K_1(x)(A_1^{X_1}(t) - A_1(x,t))\right]\}.$$

We have

$$\mathbb{E}\left[K_1(x)(A_1(X_1,t) - A_1(x,t))\right] = \mathbb{E}\left(K\left(\frac{d(x,X_1)}{h_n}\right)(A_1(X_1,t) - A_1(x,t))\right)$$
$$= \mathbb{E}[K_1(x)(A_1(X_1,t) - A_1(x,t))\mathbf{1}_{B(x,h_n)}].$$

By using the Lipschitz's condition, we obtain that

$$\mathbb{E}[Y_i^-|X=x] - \mathbb{E}\left[\frac{Y_i^- K_1}{\mathbb{E}[K_1]}\right] = \frac{1}{[K_1(x)}|\mathbb{E}[K_1(x)(A_1(X_1,t) - A_1(x,t))\mathbf{1}_{B(x,h_n)}]| \leqslant C_1 h_n^{k_1}.$$

This, implies that

$$T_1 = O(h_n^{k_1})$$

- In the other hand we have :

$$A_1(x,t) - \mathbb{E}\left[\hat{u}_1(X_j,\tilde{Y}_j)\frac{K_1}{\mathbb{E}(K_1)})\right] = \mathbb{E}\left(u_1(X_j,\tilde{Y}_j) - \hat{u}_1(X_j,\tilde{Y}_j)\frac{K_1}{\mathbb{E}[K_1]}\right)$$
$$+ \mathbb{E}\left(A_1(x,t) - Y_i^-\frac{K_1}{\mathbb{E}[K_1]}\right)$$
$$+ \mathbb{E}\left((Y_i^- - u_1(X_j,\tilde{Y}_j)\frac{K_1}{\mathbb{E}[K_1]}\right).$$

first we have

$$\sup_{x\epsilon s}|\mathbb{E}\left(u_1(X_j,\tilde{Y}_j) - \hat{u}_1(X_j,\tilde{Y}_j)\frac{K_1}{\mathbb{E}[K_1]}\right)| = O\left(\sup_{x\epsilon s}|u_1(X_j,\tilde{Y}_j) - \hat{u}_1(X_j,\tilde{Y}_j)|\right)$$

we consider the following decomposition for $j \in \overline{V}$

$$\widehat{u}(X_j,\widetilde{Y}_j) - u(X_j,\widetilde{Y}_j) = -\frac{u}{\widehat{u}_D^x}(\widehat{u}_D^x - 1) + \frac{1}{\widehat{u}_D^x}\{\widehat{u}_N^x(y) - \mathbb{E}[\widehat{u}_N^x(y)] - (u - \mathbb{E}[\widehat{u}_N^x(y)])\}$$
$$:= -\frac{u}{\widehat{u}_D^x}T_{2,1} + \frac{1}{\widehat{u}_D^x}(T_{2,2} - T_{2,3})$$

then,under assumptions (A1)-(A6)

$$T_{2,1} = T_{2,2} = O_{\text{a.co.}}\left(\sqrt{\frac{\psi_{S_{\mathcal{F}}}\left(\frac{\log N}{N}\right)}{n\phi(a_n)}}\right) \text{ and } T_{2,3} = O(a_n^{k_l}) + O(h_n^{k_l}).$$

it is clear that :

$$\sup_{x\in s}|\mathbb{E}\left(u_1(X_j,\tilde{Y}_j) - \hat{u}_1(X_j,\tilde{Y}_j)\frac{K_1}{\mathbb{E}[K_1]}\right)| = O\left(\sup_{x\in s}\left|u_1(X_j,\tilde{Y}_j) - \hat{u}_1(X_j,\tilde{Y}_j)\right|\right)$$

and

$$\sup_{x\in s}\left|\mathbb{E}\left((Y_i^- - u_1(X_j,\tilde{Y}_j)\frac{K_1}{\mathbb{E}[K_1]}\right)\right| = 0$$

Lemma 4.3.3. *under the assumptions (A1)-(A6) we have*

$$\sum_n \mathbb{P}\left(\inf_{x\in S_{\mathcal{F}}}\inf_{t\in[\xi_p(x)-\delta,\xi_p(x)+\delta]}|\widehat{A}_2(x,t)]| \leqslant \epsilon'\right) < \infty \quad \textit{for certain} \quad \epsilon' > 0.$$

Proof of Lemma 3 :

By taking into account the fact that $A_2(x,t)$ is strictly positive function, then for all $t \in \mathbb{R}$, we can easily deduce that

$$\mathbb{P}\left(\inf_{x\in S_{\mathcal{F}}}\inf_{t\in[\xi_p(x)-\delta,\xi_p(x)+\delta]}\left|\widehat{A}_2(x,t)\right| \leqslant \frac{1}{2}\inf_{x\in S_{\mathcal{F}}}\inf_{t\in[\xi_p(x)-\delta,\xi_p(x)+\delta]}A_2(x,t)\right)$$
$$\leqslant \mathbb{P}\left(\inf_{x\in S_{\mathcal{F}}}\inf_{t\in[\xi_p(x)-\delta,\xi_p(x)+\delta]}\left|\widehat{A}_2(x,t) - A_2(x,t)\right| > \frac{1}{2}\inf_{x\in S_{\mathcal{F}}}\inf_{t\in[\xi_p(x)-\delta,\xi_p(x)+\delta]}A_2(x,t)\right).$$

By applying Lemma's 1 result for $\epsilon = \frac{1}{2}\inf_{x\in S_{\mathcal{F}}}\inf_{t\in[\xi_p(x)-\delta,\xi_p(x)+\delta]}A_2(x,t) > 0$, we obtain

$$\sum_{n\geqslant 0}\mathbb{P}\left(\inf_{x\in S_{\mathcal{F}}}\inf_{t\in[\xi_p(x)-\delta,\xi_p(x)+\delta]}\left|\widehat{A}_2(x,t)\right| \leqslant \frac{1}{2}\inf_{x\in S_{\mathcal{F}}}\inf_{t\in[\xi_p(x)-\delta,\xi_p(x)+\delta]}A_2(x,t)\right) < \infty$$

therefore the lemma's proof is complete.

4.3.2 Asymptotic normality

this section is devoted to the establichment of the asymtotic normality of $\widehat{\xi_p}$,

Theorem 4.2. *Under the Hypotheses (A1)-() we have when* $n \to \infty$

$$\left(\frac{n\phi_x(h_n)}{\sigma_p^2}(x)\right)^{1/2}\left(\hat{\xi}_p(x)-\xi_p(x)\right)\xrightarrow{D}\mathcal{N}(0,1)$$

where

$$\sigma_p^2(x)=\frac{\alpha_2(x)\lambda_p(\theta(p;x);x)}{\alpha_1^2(x)}\Gamma_p^2(\theta(p;x);x)\qquad\left(\textit{with }\alpha_j(x)=K^j(1)-\int_0^1(K^j)'(s)\beta_x(s)ds\textit{ for }j=1,2\right)$$

and

$$\lambda_p(\theta(p;x);x)=\left(\frac{p}{1-p}\right)^2R_+^x(\theta(p;x))+R_-^x(\theta(p;x))$$

where

$$\begin{aligned}R_+(\theta(p;x);x)&=\mathbb{E}\left[(Y_1-\theta(p;x))^2\mathbf{1}_{(Y_1>\theta(p;x)}|X=x\right]\\R_-(\theta(p;x);x)&=\mathbb{E}\left[(Y_1-\theta(p;x))^2\mathbf{1}_{(Y_1\le\theta(p;x)}|X=x\right]\end{aligned}$$

and

$$\Lambda_p(\theta(p;x);x)=A_1'(\theta(p;x);x)-\left(\frac{p}{1-p}\right)A_2'(\theta(p;x);x).$$

Proof of theorem 2 :

To simplify the notation, we denote $Z_n=\left(\sqrt{n\phi_x(h_n)}\sigma_p^{-1}(x)\right)^{1/2}(\widehat{\xi}_p(x)-\xi_p(x))$

For $z\in\mathbb{R}$, we put $u_p(z,x)=\xi_p(x)+z(n\phi_x(h))^{-1/2}\sigma_p(x))$, and $\widehat{\lambda}_p(z;x)=\left(\frac{p}{1-p}\right)\widehat{A}_2(u_p(z,x))-\widehat{A}_1(u_p(z,x))$

then

$$\begin{aligned}\mathbb{P}[Z_n\le z]=&\mathbb{P}\left[\widehat{\xi_p}(x)\le u_p(z,x)\right]\\=&\mathbb{P}\left[\frac{p}{1-p}\le\widehat{A}(u_p(z,x);x)\right]+\mathbb{P}\left[\left(\widehat{\xi_p}(x)\le u_p(z,x)\right)\cap\left(\widehat{A}(u_p(z,x))=0\right)\right]\\=&\mathbb{P}\left[\widehat{\lambda}_p(z;x)-\mathbb{E}\left[\widehat{\lambda}_p(z;x)\right]\le\mathbb{E}\left[-\widehat{\lambda}_p(z;x)\right]\right]\\&+\mathbb{P}\left[\left(\widehat{\xi_p}(x)\le u_p(z,x)\right)\cap\left(\widehat{A}(u_p(z,x))=0\right)\right]\end{aligned}$$

we used the fact that the function $\widehat{A}(.;x)$ is an increasing function and$\widehat{\xi_p}(x)$ is the unique solution of the equation $\widehat{A}(\xi;x)=\frac{p}{1-p}$.In order to achieve the proof of theorem 2 we have to prove that the second term converges to zero , and establish the convergence in distribution of the second term to a standard normal varible , as n tends to infinity.

Lemma 4.3.4. *under the assumptions (A1)-(A6) we have as* $n\to\infty$

$$\mathbb{E}\left[-\widehat{\lambda}_p(z;x)\right]=z\frac{\sigma_p(x)\Lambda_p(\xi(p;x);x)}{\sqrt{n\phi_x(h)}}+o\left((n\phi_x(h))^{-1/2}\right)$$

Proof of Lemma : by the definition of $\widehat{\lambda}_p(z;x)$ and using the fact that

$$\begin{aligned}
\mathbb{E}\left[-\widehat{\lambda}_p(z;x)\right] &= \mathbb{E}[\widehat{A}_1(u_p(z;x);x)] - \frac{p}{1-p}\mathbb{E}[\widehat{A}_2(u_p(z,x);x)] \\
&= \mathbb{E}[\widehat{A}_1(u_p(z,x);x)] - A_1(u_p(z,x);x) \\
&+ \frac{p}{1-p}(A_2(u_p(z,x);x)) - \mathbb{E}[\widehat{A}_2(u_p(z,x);x)] \\
&+ A_1(u_p(z,x);x) - A_1(\xi(p,x);x) \\
&- \frac{p}{1-p}A_2(u_2(z,x);x) - A_2(\xi(p,x);x) \\
&+ \left(A_1(\xi(p,x)) - \frac{p}{1-p}A_2(\xi(p,x);x)\right) \\
&=: I_1 + I_2 + I_3 + I_4 + I_5.
\end{aligned}$$

First, under Hypotheses have :

$$|\mathbb{E}[\widehat{A}_1(u_p(z,x))] - A_1(u_p(z,x);x)| \leq C_1 h_n^{k_1} + C_1 a_n^{k_1}$$

and

$$\frac{p}{1-p}|\mathbb{E}[\widehat{A}_2(u_p(z,x))] - A_2(u_p(z,x);x)| \leq C_2 h_n^{k_2} + C_2 a_n^{k_2}$$

on the other hand to assess I_3 and I_4 we use the Taylor expansion of the function $A_1(.;x)$ and the first part of the Lipschitzian , we get

$$\begin{aligned}
J_3 &= A_1(u_p(z,x);x) - A_1(\xi(z,x);x) \\
&= z\sigma_p(x)(n\phi_x(h))^{-\frac{1}{2}}A_1'(\xi(z,x);x) \\
&+ o\left((n\phi_x(h))^{-\frac{1}{2}}\right)
\end{aligned}$$

and by using the same arguments we obtain

$$\begin{aligned}
J_4 &= \left(-\frac{p}{1-p}\right)A_1(u_p(z,x);x) - A_1(\xi(z,x);x) \\
&= z\sigma_p(x)(n\phi_x(h))^{-\frac{1}{2}}\left(\left(-\frac{p}{1-p}\right)A_2'(\xi(z,x);x)\right) \\
&+ o\left((n\phi_x(h))^{-\frac{1}{2}}\right)
\end{aligned}$$

recall that

$$\Lambda_p(\theta(p;x);x) = A_1'(\theta(p;x);x) - \left(\frac{p}{1-p}\right)A_2'(\theta(p;x);x).$$

and under Hypotheses achieve the proof

Lemma 4.3.5. *under the assumptions (A1)-(A6) we have as* $n \to \infty$

$$(n\phi_x(h))\mathbb{V}ar\left(\sum_{i=1}^{n}\mu_i(x)\right) \to \frac{\alpha_2\lambda_p(\xi(p,x);x)}{\alpha_1^2(x)}$$

and

$$Z_n = \left(\frac{n\phi_x(h)}{\sigma_p^2(x)}\right)^{\frac{1}{2}} (\Lambda_p(\xi(p;x);x))^{-1}\left(\widehat{\lambda}_p(p,x) - \mathbb{E}\left[\widehat{\lambda}_p(p,x)\right]\right) \xrightarrow{D} \mathcal{N}(0,1)$$

where

$$\mu_i(x) = \frac{1}{n\mathbb{E}[K_1(x)]}K_i(x)(Y_i - u_p(z,x))\left(\frac{p}{1-p}\mathbf{1}_{E_i} + \mathbf{1}_{E_i^c}\right)$$

with

$$E_i = Y_i > u_p(z,x) \quad \textit{and} \quad E_i^c = Y_i \leq u_p(z,x)$$

we can

Lemma 4.3.6. *under the assumptions (A1)-(A6) we have as* $n \to \infty$

$$\mathbb{P}\left[\left(\widehat{\xi}_p(x) \leq u_p(z,x)\right) \cap \left(\widehat{A}_2(u_p(z,x)) = 0\right)\right] \leq \exp[-n\phi_x(h)]$$

To obtain this results we companin the idea Wang (2006) and Mohammedi (2020) .

GENERAL CONCLUSION AND PROSPECTS

Conclusion

The literature on missing data is still relevant, especially with regard to the estimation of the functional parameters present in this model. Recall that one of the main motivations for the craze of nonparametric functional statistics is the solution it offers for the problem of the scourge of dimension, and the power of computers which have made it possible to process data in very large dimensions.

In this thesis, we are interested in the robust estimation of the regression operator in the presence of missing data. It is clear that the superiority of this approach over the classical method is the main motivation for this subject. In order to highlight this superiority in NPFDA, we first studied, the asymptotic properties of a nonparametric estimator of the relative error regression given a functional explanatory variable, when the scalar response is right censored, in the i.i.d. case. We establish the strong almost complete convergence rate and asymptotic normality of these estimators.

As a first idea of extension, is to establish similar results when one frees oneself from the assumption of independence. It is well known that in practice several processes have a certain dependence. The second part of this thesis is devoted to the problem of estimating the relative regression operator when the observation are $\alpha-$mixing. We establish the almost complete convergence rate of these estimators. A simulation study and real data application are performed to illustrate how this fact allows getting higher predictive performances than those obtained with standard estimates.

Finally, it seems possible to us interested in studying the robust model given a func-

tional explanatory variable, in the case of a scalar missing at random (MAR) response, for both cases, without and with unknown scale parameter. We establish, the almost complete convergence rate of our estimators in the two proposed models.

Prospects

To conclude the work of this thesis, many questions remain unanswered. We believe we will invest in the future on a few issues in order to improve and extend our results.

— We think it is possible to adapt our results to another type of dependency such as the quasi-associated and the ergodic case.

— Other issues are possible, such that extensions our estimators to the local linear ideas.

— Another possible prospect is to obtain the asymptotic normality of the robust equivariant regression for functional data with responses missing at random.

— Robust estimation with single functional index model can be approached in the missing case.

— We will be able to elaborate the asymptotic properties of our estimators based on the k nearest neighbor (k–NN) method or other methods on the bandwidth selection, because it allows the improvement of the quality of the estimator.

— We can generalize the results obtained using other models such as the aditive model or the semi-functional partial linear model.

— An important issue about the comparison of the constructed estimators when there are surrogate outputs.

BIBLIOGRAPHIE

Aigner, D. J., Amemiya, T., and Poirier, D. J. (1976). On the estimation of production frontiers : maximum likelihood estimation of the parameters of a discontinuous density function. *Internat. Econom. Rev.*, 17(2) : 377-396.

Aneiros, G. and Vieu, P. (2015). Partial linear modelling with multi-functional covariates. *Comput. Statist.*, 30(3) : 647-671.

Artzner, P., Delbaen, F., Eber, J.-M., and Heath, D. (1999). Coherent measures of risk. *Math. Finance*, 9(3) : 203-228.

Bellini, F., Klar, B., Müller, A., and Rosazza Gianin, E. (2014). Generalized quantiles as risk measures. *Insur. Math. Econ.*, 54 : 41-48.

Bellini, F. and Bignozzi, V. (2015). On elicitable risk measures. *Quant. Finance*, 15(5) : 725-733.

Bellini, F. and Di Bernardino, E. (2017). Risk management with expectiles. *The European Journal of Finance*, 23(6) : 487–506.

Bellini, F., Bignozzi, V. and Puccetti, G. (2018). Conditional expectiles, time consistency and mixture convexity properties. *Insurance Math. Econom.*, 82 : 117–123.

Bierens, H.J. (1987), Kernel Estimators of Regression Functions, in : Truman F.Bewley (ed.). *Advances in Econometrics 1985*, New York : Cambriclje University Press, 99-144

Brillinger, D. (1981), Time Series : *Data Analysis and Theory*. Holden-Day Inc., San Francisco, expanded edition.

Cai, J. and Weng, C. (2016). Optimal reinsurance with expectile. *Scandinavian Actuarial Journal*, 2016, (7) : 624–645.

Collomb. G. (1981), Estimation Non-paramétrique de la Regression : Revue Bibliographique, *International Statistical Review*. 49 : 75-93.

Collomb.G. (1985), Nonparametric Regression : An Up-to-Date Bibliography, *Statistics*. 16 : 309-324.

Crambes, C., Delsol, L., and Laksaci, A. (2008). Robust nonparametric estimation for functional data. *J. Nonparametr. Stat.*, 20(7) : 573-598.

2018)]daouia19 Daouia, A. Gijbels, I., and Stuper, G. (2019). Extremiles : A new prespective on asymmetric least squares. *Journal of the American Statistical Association*, 114(527) : 1366-1381.

Diggle, P., Liang, K., and Zeger, S. (1994). *Analysis of Longitudinal Data.* Oxford University Press.

B. Efron, Regression percentiles using asymmetric squared error loss, *Stat. Sin.* (1991), 1(1) : 93–125.

Emmer, S., Kratz, M., and Tasche, D. (2015). What is the best risk measure in practice ? a comparison of standard measures. *Journal of Risk* , 18 :2, 31-60.

Ferraty, F. and Vieu, P. (2006). *Nonparametric functional data analysis.* Springer Series in Statistics. Springer, New York. Theory and practice.

Ferraty, F., Mas, A., and Vieu, P. (2007). Nonparametric regression on functional data : inference and practical aspects. *Aust. N. Z. J. Stat.*, 49(3) : 267-286.

Fomby et Hill RC 2000 Applying Kernel and *Nonparametric Estimation to Economic Topics.* JAI Press.

Gasser T, Muller HG, Kohler W, Molinari L, Prader A. 1984. Nonparametric regression analysis of growth curves. *Annals of Statistics.* 12 : 210–229

Gasser T, Kneip A. 1995. Searching for structure in curve samples. *Journal of the American Statistical Association.* 90 : 1179–1188

Gneiting, T. (2011). Making and evaluating point forecasts. *J. Amer. Statist. Assoc.*, 106(494) : 746-762.

Grenander U. 1950. Stochastic processes and statistical inference. *Arkiv for Matematik*. 1 : 195–277

Hahn, J. (1995). Bootstrapping quantile regression estimators. Econometric Theory, 11(1) : 105-121.

Hall P, Racine JS, and Li Q 2004 Cross-validation and the estimation of conditional probability densities. *Journal of the American Statistical Association*. 99 : 1015–1026.

Jones, M. C. (1994). Expectiles and M-quantiles are quantiles. *Statist. Probab. Lett.*, 20(2) : 149-153.

Kara, L.-Z., Laksaci, A., Rachdi, M., and Vieu, P. (2017). Data-driven knn estimation in nonparametric functional data analysis. *Journal of Multivariate Analysis* , 153 : 176-188.

Ling, N., Meng, S., and Vieu, P. (2019). Uniform consistency rate of knn regression estimation for functional time series data. *Journal of Nonparametric Statistics* , 31(2) : 451-468.

Loève, M. (1963). *Probability theory*. Third edition. D. Van Nostrand Co., Inc., Princeton, N.J.-Toronto, Ont.-London.

Maume-Deschamps, V., Rullière, D., and Said, K. (2017). Multivariate extensions of expectiles risk measures. *Depend. Model.*, 5(1) : 20-44.

Maume-Deschamps, V., Rullière, D., and Said, K. (2018). Asymptotics multivariate expectiles. working paper or preprint.

Mohammedi, M. *Contribution à l'estimation non paramétrique des expectiles en statistique fonctionnelle*. Thèse de doctorat, UDL Sidi Bel Abbès, (2021).

Nadaraya, E.A. (1964), On Estimating Regression, *Theory of Probability and its Applications.* 9 :141-142

Newey, W.K. and Powell, J.L., Asymmetric least squares estimation and testing, *Econometrica*. (1987), 55(4) : 819–847.

Novo, S., Aneiros, G., and Vieu, P. (2019). Automatic and location-adaptive estimation in functional singleindex regression. *Journal of Nonparametric Statistics*, 31(2), 364-392.

Parzen, E. (1962). On estimation of a probability density function and mode. *Ann. Math. Statist.*, 33(3), 1065-1076.

R. Koenker, G.W. Bassett Jr, Regression quantiles, *Econometrica*. (1978), 46(1) : 33–50.

Rachdi, M. and Vieu, P. (2007). Nonparametric regression for functional data : Automatic smoothing parameter selection. *Journal of Statistical Planning and Inference*, 137(9) : 2784-2801.

Ramsay JO. 1982. When the data are functions. *Psychometrika.* 47 : 379–396

Ramsay JO, Dalzell C. 1991. Some tools for functional data analysis. *Journal of the Royal Statistical Society. Series B (Methodological)* : 539–572

Raña, P., Aneiros, G., Vilar, J., and Vieu, P. (2016). Bootstrap confidence intervals in functional nonparametric regression under dependence. *Electron. J. Stat.*, 10(2) : 1973-1999.

Rao CR. 1958. Some statistical methods for comparison of growth curves. *Biometrics*. 14 : 1–17

Rice J, Silverman B. 1991. Estimating the mean and covariance structure nonparametrically when the data are curves. *Journal of the Royal Statistical Society. Series B (Methodological).* 53 : 233–243

Rosenblatt, M. (1956). Remarks on some nonparametric estimates of a density function. *Ann. Math. Statist.*, 27(3) : 832-837.

Ruppert D, Wand MP, and Carroll RJ 2003. *Semiparametric Regression*. Cambridge University Press.

Shang, H. L. (2014). Bayesian bandwidth estimation for a functional nonparametric regression model with mixed types of regressors and unknown error density. *J. Nonparametr. Stat.*, 26(3) : 599-615.

Taylor J.W., Estimating value at risk and expected short fall using expectiles, *J. Financial Econom*. (2008), 6(2) : 231–252.

Watson.G.S. (1964), Smooth Regression Analysis, *Sankhya A* 26, 359-372

Q. Yao, H. Tong, Asymmetric least squares regression estimation : a nonparametric approach, *J. Nonparametr. Stat.* (1996), 6(2–3) : 273–292.

Ziegel, J. F. (2016). Coherence and elicitability. *Math.Finance*, 26(4) : 901-918.

Attouch, M. Laksaci, A. and Rafaa, F. (2017). Local linear estimate of the regression operator by the kNN method. Comptes Rendus Mathematique, Vol. 355, No. 7, pp. 824–829.

Barrientos-Marin, J., Ferraty, F. and Vieu, P. (2010). Locally modelled regression and functional data. Journal of Nonparametric Statistics, Vol. 22, No. 5, pp. 617–632.

Benhenni, K., Ferraty, F., Rachdi, M. and Vieu, P. (2007). Local smoothing regression with functional data. Computational Statistics, Vol. 22, No. 3, pp. 353–369.

Bosq, D. (2000). *Linear processes in function spaces*, volume 149 of Lecture Notes in Statistics.

Carroll, R.J. and Wand, M.P. (1991). Semiparametric estimation in logistic measurement error modelss. Journal of the Royal Statistical Society : Series B (Methodological), Vol. 53, No 3, pp. 573–585.

Chen, J.H. and Shao, J. (2000). Nearest neighbor imputation for survey data. Journal of Official Statistics, Vol. 16, No. 2, pp. 113.

Cheng, P.E. (1994). Nonparametric estimation of mean functionals with data missing at random. Journal of the American Statistical Association, Vol. 89, No. 425, pp. 81–87.

Dabo-Niang, S. (2002). Estimation de la densité dans un espace de dimension infinie : Application aux diffusions. Comptes Rendus Mathematique, Vol. 334, No. 3, pp. 213–216.

Duncan, G.J. and Hill, D. H. (1985). An investigation of the extent and consequences of measurement error in labor-economic survey data. Journal of Labor Economics, Vol. 3, No. 4, pp. 508–532.

Ferraty, F., Goia, A. and Vieu, P. (2002). Functional nonparametric model for time series : a fractal approach for dimension reduction. Test, Vol. 11, No. 2, pp. 317–344.

Ferraty, F., Laksaci, A. and Vieu, P. (2006). Estimating some characteristics of the conditional distribution in nonparametric functional models. *Statistical Inference for Stochastic Processes*, Vol. 9, No. 1, pp. 47–76.

Ferraty, F., Rabhi, A. and Vieu, P. (2005). Conditional Quantiles for Dependent Functional Data with Application to the Climatic El Niñ̃o Phenomenon. *Sankhyā*, Vol.67, No. 2, pp. 378–398.

Ferraty, F., Sued, M. and Vieu, P. (2013). Mean estimation with data missing at random for functional covariables. Statistics, Vol. 47, No. 4, pp. 688–706.

Ferraty, F. and Vieu, P. (2002). The functional nonparametric model and application to spectrometric data. Computational Statistics, Vol. 17, No. 4, pp. 545–564.

Ferraty, F. and Vieu, P. (2003). Curves discrimination : a nonparametric functional approach. Computational Statistics & Data Analysis, Vol. 44, No. (1-2), pp. 161–173.

Ferraty, F., and Vieu, P. (2006). *Nonparametric functional data analysis : theory and practice*. Springer Science & Business Media.

Ibrahim, F., Hajj Hassan, A., Demongeot, J. and Rachdi, M. (2020). Regression model for surrogate data in high dimensional statistics. Communications in Statistics-Theory and Methods, Vol. 49, No. 13, 3206–3227.

Ling, N., Liang, L. and Vieu, P. (2015). Nonparametric regression estimation for functional stationary ergodic data with missing at random. Journal of Statistical Planning and Inference, Vol. 162, pp. 75–87.

Pepe, M. S. (1992). Inference using surrogate outcome data and validation sample. Biometrika, Vol. 79, No. 2, pp. 355–65.

Pérez-González, A., Vilar-Fernández, J. M. and González-Manteiga, W. (2009). Asymptotic properties of local polynomial regression with missing data and correlated errors. Annals of the Institute of Statistical Mathematics, Vol. 61, No. 1, pp. 85–109.

Ramsay, J. and Silverman, B. (1997). *Functional Data Analysis*. Springer-Verlag.

Ramsay, J. and Silverman, B. (2002). *Appied functional Data Analysis*. Springer-Verlag.

Roussas, G. G. (1968). On some properties of nonparametric estimates of probability density functions. Bull. Soc. Math. Greece (N.S.), Vol. 9, No. 9, pp. 29–43.

Wang, Q. (2006). Nonparametric regression function estimation with surrogate data and validation sampling. Journal of multivariate analysis, Vol. 97, No. 5, pp. 1142–1161.

Wittes, J., Lakatos, E. and Probstfield, J. (1989). Surrogate endpoints in clinical trails : Cardiovascular diseases. Statistics in Medicine, Vol. 8, No. 4, pp. 415-25.

Yates, F. (1933). The analysis of replicated experiments when the field results are incomplete. Empire Journal of Experimental Agriculture, Vol. 1, No. 2, pp. 129–142.