**Estimation of conditional hazard function by the method of k nearest neighbor with -mixing data.**

mohammed kadi Attouch

**Abstract** The aim of this article is to study the k nearest neighbor method of the conditional hazard function for -mixing data is to prove consistency the almost complete convergence and the almost complete convergence rate of estimator of this function.

**Keywords :** Functional data; nonparametric regression, k-NN estimator, the conditional hasard function,rate of convergence, random bandwidth.
**AMS Subject Classification 2010 :** 62G05 62G08 62G20 62G35.

# Introduction

In many applicability, as in econometrie or seismology, the assumption of independence between the variables observed is not realistic. For that, it is necessary to introduce a probabilistic structure which makes it possible to control the dependence between the variables, the method is to make the assumption of mixture.

in this paper, we consider the method of K nearest neighbor for the estimate of the conditional hazard function in case mixing data which depend of number of neighbors at the point of interest at with we want to make a prediction. On the contrary, the traditional estimate of the function depend on the real valued non random bandwidth sequence (see ).

We introduce in this article the method and assumptions used in section 2, after we propose in section 3 the almost complete convergence and the almost complete convergence rate. We present in section 4, some technical tools, finally we show the proof in section 5.

# Method and assumptions

Let be dependent sequence identically distributed as , the latter being a random pair with values in the measurable space . Where is a semi metric space and is the algebra generated bu the topology of that is defined by by the semi-metric d, and B is the Borel algebra.

For , we define the K-nn kernel estimate of the conditional hazard function by:

with: : is the kernel estimator of the conditional distribution function given by:

where is an asymmetrical kernel, is the bandwith that is defined as:

 is a distribution function and is a sequence of positive real numbers (depending on n).

 is the kernel estimate of the conditional density function define by:

To prove the almost complete convergence of the K-nn estimator of the conditional hazard function and to emphasis difference between the K-nn method and the traditional kernel approach, we need some results of Ferraty et al by:

with:

and:

Where is a kernel, is a distribution function and is a non-random bandwith.

Before giving the main asymptotic result, we need some assumptions. The first one is about the probability of observing the functional random variable X around x:

* We also need to kernel :
 is a kernel of type I, so that: there exist two real constants ; such that;
* is a kernel of type II, so that: the support of is and if its derivative exists on and satisfies; for two real constants ;
* If is a kernel of type II and if satisfies: :
* Assume that the conditional moment of the response random variable is bounded:
* with continuous on x.
* The sequence is said to be -mixing if:
* where: is the algebra generated by .
* is **arithmetically** (or **equivalently** algebraically ) -mixing with rate if:
* We note the term of covariance by:
* where:
* By lemma 10.3, see , we remark, if the are -mixing then also the are -mixing.
* Assume that for , the rate of the -mixing b: there exists such that
* We need the following additional assumptions on the distribution of two distinct pair and , we assume that
* and the joint distribution function:
* satisfy:
* where:
* Define , and with is the mixing coefficient. Then assume that:

# Asymptotic properties of K-nn method

We are interested in this section, by the almost complete convergence [[1]](#footnote-24) and rate of convergence [[2]](#footnote-25) of the functional kernel estimator of the conditional hazard function , where here is a sequence of -mixing random variables (see ).
Before studying the K-nn estimator, we remind asymptotic properties of define in (???), Rabhi et al. , proved the almost complete convergence of this estimator:

**Theorem** Under the “continuity type”, suppose (H1), (H2), (H6), (H8), then we have:

**Theorem** Under the “Lipschitz”Lipschitz type" model and the hypotheses (H1), (H2), (H3), (H5),(H8), we have:

Now, we state the almost complete convergence result for nonparametric K-nn method estimate define in (???) which the proof is postponed to the appendix.

**Theorem**

Under the “continuity type” model and the hypotheses(H1), (H2), (H4), (H8), then we have:

**Theorem**

Assume that (H1), (H2), (H4), (H8) are verified,and under the “continuity type” then:

# General technical tools

Let be a sequence of random variables with values in , not necessarily identically distributed or independent, let a measurable function such that:

Collomb 1980-Burba 2008
Let be a sequence of real random variables and be a decreasing sequence of positive numbers.
If , and if, for all increasing sequences , there exist two sequences of real random variables and :

* : , and a.co.
* :
* : Assume there exists a real positive number such that:

Then:

If l=0 and if (L1), (L2), (L3) hold for any increasing sequence with limit 1, the same result holds.

Burba 2008
Let be a sequence of real random variables and be a decreasing positive sequence.
If , and if, for all increasing sequences , there exist two sequences of real random variables and :

* : , and a.co.
* :
* : Assume there exists a real positive number such that:

Then:

If l’=0 and if (L’1), (L’2), (L’3) hold for any increasing sequence with limit 1, the same result holds.

Burba use in their consistency proof of the k.n.n. kernel estimate for independent data Chernoff-type exponential inequality to check conditions or .

Let be a valued random vector in such that for some Let d a real number such that and . Then there exists a random variable such that:

* and is independent of .
* , where is the -algebra generated by .

Let a arithmetically mixing sequence in the semi metric space with , b and . Define .then we have:

where and .

# Asymptotic normality

This section contain results on the asymptotic normality of . For this, we need the followings notations:

where:

Where is a kernel, is a distribution function and is a sequence of positive real numbers. Laksaci and Mechab defined the preceding estimators by:

where:

**Theorem** Assume that (H1), (H4), (H5), (H7), (H8) hold, then for any , we have:

where

 means the convergence in distribution.
It is easy ti see that, if one impose some regularity assumptions on the real function , we can give explicitly the asymptotic behavior on the term . However, to remove the bias term from Theorem ???.

under the hypotheses of ???, we have:

**Proof of theorem???:** We consider the following decomposition:

Therefore, the theorem is a sequence of the following lemmas:

Under the hypotheses of theorem ???, we have:

where

Under the hypotheses of theorem ???, we have:

where

and

Under the hypotheses of theorem ???, we have:

## A Simulation study

In this section we will show the effectiveness of -NN method compared to the kernel estimation using simulated data. For this we considered a sample of a diffusion process on interval , and , where is the standard normal distribution and take , where is random variable Bernoulli distributed. We carried out the simulation with -sample of the curve which is represented by the following

For the scalar response variable, we took where (resp. ) is the nonlinear regression model , with is (resp. is the null function) and is an -mixing process generated by this model:

with is normally independent identically distributed random variables. We generated standard normal distributed random variable independently of .

We choose a quadratic kernel defined by:

In practice, the semi-metric choice is based on the regularity of the curves . For this we use the semi-metric defined by -distance between the derivatives of the curves. In order to evaluate the MSE (Mean Square Error) we proceed by the following algorithm:

* *Step 1.* We split our data into two subsets; the first sample, of size n=120 corresponds to the learning sample which will be used, as a sample, to compute our conditional hazard function estimators for the 80 remaining curves ( considered as the test sample).
	+ learning sample,
	+ test sample.
* *Step 2.*
	+ We use the learning sample for computing the hazard function estimator , for all .
	+ We set: .
	+ We put:
	+ for kernel method.
	 for -NN method,
	+ where:
	+ : is the nearest curve to
	+ with:
	+ and:
* *Step 3.* The error used to evaluate this comparison is the mean of square error () expressed by
* where designate the estimator used: kernel or -NN method estimation and is the true hazard function.

Consequently, the -NN method gives better results than the kernel method. This is confirmed by the MSE--NN= 2.170035 and MSE-Kernel = 13.66790.

# Real data application

To highlight the efficiency and robustness of the method of nearest neighbors with respect to the kernel method in estimating the conditional hazard function, we will test these two methods in the presence or not of heterogeneous data.

To do this, based on the study of Burba et al. (2009) which emphasizes the effect of the nature of the data (homogeneous or heterogeneous ) on the quality of the estimate, especially the superiority of the k-nearest neighbors in the presence of very heterogeneous data.

For this purpose, we apply the described algorithm used in the simulation study to El Ni time series which gives monthly Sea Surface temperatures. More information and other data-sets can be found about the phenomenon called El Ni in web site http://www.cpc.ncep.noaa.gov/data/indices
(see also:http://kdd.ics.uci.edu/databases/elnino/elnino.data.html). Our study concerns the monthly time series of the Sea Surface Temperature (SST) from June, up to May, .

A useful way to display such a time series consists in cutting it into 54 pieces or 54 “annual curves” (see Fig. 1). More precisely, let be our El Nio time series. We can build, for , the following subsequences: corresponding to the variations of the SST at the year.

The sample of size 54 was splitted into learning sample of size 44 (with all data), 30 (without the heterogeneous data, 14 values) and testing sample of size 10. Figue ??? displays the curves of learning sample for all data and the curves of learning sample without the heterogeneous one.

We plot the conditional hazard function estimated for the first 3 values of testing sample, Figure ??? depicts that the -NN method in presence of heterogeneous data give better estimation of the conditional hazard function prediction (regular function) than the kernel method estimation (non-regular function) and when the data are homogeneous the two method give the same result that can be easily seen in Figure ???.

# Appendix

*Proof of Theorem ???* We consider the following decomposition:

and our proof is finished to verify the following results:

*proof of result ???:*
To prove this result, we apply Lemma ??? with:

Choose , and such that:

Then we define

Ferraty Vieu () proved under the conditions of Theorem ??? that:

and to apply this result, we have show that the covariance term fulfills following condition: such that:

with b is the rate of mixing coefficient. If the condition on the rate b of the mixing coefficient and () hold, we have by lemma 11-5 in ():

and it is true for , then ??? is holds.
Now we can apply the result ???, then we obtain:

and

Thus condition is verified.
In parallel Ferraty and Vieu () proved that:

under the conditions (???) and (???), we have:

we get:

so that is checked.
Finally, we prove : The first part is obvious, and the second one that, :

Let , we know that;

We start by , then we have:

The second step, we centred the random variable, then the plan here is to split the data into a block schema and we applique lemma ???; we devise the set into blocks of length , set where is the Gaussian bracket [[3]](#footnote-31) and .
Let

and we define:

thus, we obtain in ???.

Now, we apply Lemma ??? in the first term of ??? with:

* lead to

and we can construct such that:

* the random variables are independent
* have a same distribution as .

then:

with this leads to:

and by applying lemma ??? in the second term of ???, and some calculation we obtain:

we choose the sequence such that:

with: is a positive constant, , and by the condition of the mixing coefficient b, we obtain:

consequently, we get the second term of ??? is finite,finite, then:

Let turn to the first term of ??? and apply Markov inequality for some , we obtain:

Applying:

we get:

we evaluate the expectation, for this note that:

and after the calculations while we applying:

then

We apply now Lemma ??? with: and , then:
under :

and under :

finally, we obtain from ??? with and

because:

and thereafter of ???, we conclude that:

Same way, we can show that:

Now,it is enough for us to show:

We know that:

and by the definition of in ??? and we obtain

Consequently we have:

While following the same step to prove that

then is verified.

*Proof of result ???:*
To prove this result, we use the steps of proof of result ??? with:

and the result proved by Ferraty and Vieu()

*Proof of result ???:*
It is clear that:

turning now, to the term of probability, we obtain:

For the second term, by result ???, we have

then; for , we obtain:

*Proof of Theorem ???*
We consider the decomposition (???), and the proof of this Theorem is a consequence of this results:

*ProofProof of result (???):*
To prove this result, we use Lemme 1. Choose as an increasing sequence in with limit 1. Furthermore, we choose and under (???), Ferraty and Vieu proved under the conditions of Theorem 3 that:

with

then

If holds then:

the same is true for the bandwidth .

And if holds:

under the same choice of and as above, we have:

we get:

so that, is checked. Now wear able to apply Theorem ???, we obtain:

and is verified.

*Proof of result ???:*
To prove this result, we use Lemma ???. Choose as an increasing sequence in with limit 1.
Furthermore, we choose and under (???) Ferraty Vieu ((missing citation)) proved under the conditions of Theorem ??? that:

with

then

Under the notations (???) and (???), we have:

Then we obtain:

then is checked. Now, we apply (???) under (???), we get:

which leads the condition

**Proof of lemma ???** :
We pose the following notation:

To prove this lemma, we apply Lemma 4-2 in Burba *et al.* with:

Choose , and such that:

Define

The asymptotic normality of:

was proved in Lemma 4 in by choosing the bandwidth parameter as: .

**Proof of Lemma ???** :
On the one hand, Laksaci and Mechab proved that:

with:

therefore, we use (???) under (???) and (???) and since is bounded, we obtain:

And on the other hand:

with:

Then under the notations:

and (???), (???):

**Proof of lemma ???** :
Under (???), Ferraty and Vieu in proved that:

To establish the convergence in probability of denominator, we consider the following decomposition under the notation (???):

concerning the first term; Ferraty and Vieu showed that:

and the second term, under (???), ???), observe that:

on the other hand, by applying the result (6-19) of , we obtain:

# References

Burba, F., Ferraty, F., & Vieu, P. C. de l’estimateur à.

Burba, F., Ferraty, F., & Vieu, P. K.-N. N. method in.

Collomb, G. E. de la regression par la methode des k plus.

Collomb, G. Q. propriétés de la methode du noyau pour.

Lian, H. C. on functional k-nearest neighbor regression.

Maillot, B., & Louani, D. *P. asymptotiques de quelques estimateurs non-paramétriques pour des variables vectorielles et fonctionnelles*. (2008). Paris.

Oliveira, D. de M., P. E. *Nonparametic density and regression estimate functional data*. Tech. rep.

Ramsay, J., & Silverman, 2nd, B. *Functional Data Analysis*.

Roussas, G. N. estimation of the transition distribution.

Samanta, M. N.-parametric estimation of conditional quantiles.

1. Let $(X\_{n})\_{n\in N}$ be a sequence of the real variable. We say that $(X\_{n})$ converge almost completely to (X) if and only if:$∀ϵ>0,\sum\_{n=1}^{\infty }P[|X\_{n}−X|>ϵ]<\infty $. [↑](#footnote-ref-24)
2. We say that $(X\_{n})=O(u\_{n})$ if and only if: $∃ϵ>0$, such that, $\sum\_{n=1}^{\infty }P[|X\_{n}|>ϵu\_{n}]<\infty $. [↑](#footnote-ref-25)
3. $=max\{y\in Z/z\leq x\},x\in R.$ [↑](#footnote-ref-31)