Introduction to Nonparametric Regression

Sana Louhichi

Journées de Statistique Mathématique et Data Science

Hammamet 01-04 Novembre 2019

Learning from data: Purposes and Examples

Book: T. Hastie, R. Tibshirani, J. Friedman (2009). "The Elements of Statistical Learning - Stanford University"

- Predict whether a patient, hospitalized due to a heart attack, will have a second heart attack. The prediction is to be based on demographic, diet and clinical measurements for that patient.
- Predict the price of a stock in 6 months from now, on the basis of company performance measures and economic data.
- Identify the numbers in a handwritten ZIP code, from a digitized image.
- Estimate the amount of glucose in the blood of a diabetic person, from the infrared absorption spectrum of that person's blood.
- Identify the risk factors for prostate cancer, based on clinical and demographic variables.

Example of Data: "Caravan"

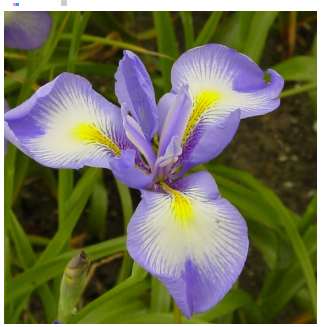
The data contains 5822 real customer records. Each record consists of 86 variables, containing sociodemographic data (variables 1-43) and product ownership (variables 44-86). The sociodemographic data is derived from zip codes. All customers living in areas with the same zip code have the same sociodemographic attributes. Variable 86 (Purchase) indicates whether the customer purchased a caravan insurance policy.

```
library('ISLR')
data("Caravan")
summary(Caravan)
help(Caravan)
dim(Caravan)
[1] 5822   86
names(Caravan)
 [1] "MOSTYPE"   "MAANTHUI"  "MGEMOMV"   "MGEMLEEF"  "MOSHOOFD"  "MGODRK"    "MGODPR"    "MGODOV"
 [9] "MGODGE"    "MRELGE"    "MRELSA"    "MRELOV"    "MFALLEEN"  "MFGEKIND"  "MFWEKIND"  "MOPLHOOG"
[17] "MOPLMIDD"  "MOPLLAAG"  "MBERHOOG"  "MBERZELF"  "MBERBOER"  "MBERMIDD"  "MBERARBG"  "MBERARBO
[25] "MSKA"      "MSKB1"     "MSKB2"     "MSKC"      "MSKD"      "MHHUUR"    "MHKOOP"    "MAUT1"
[33] "MAUT2"     "MAUT0"     "MZFONDS"   "MZPART"    "MINKM30"   "MINK3045"  "MINK4575"  "MINK7512
[41] "MINK123M"  "MINKGEM"   "MKOOPKLA"  "PWAPART"   "PWABEDR"   "PWALAND"   "PPERSAUT"  "PBESAUT"
[49] "PMOTSCO"   "PVRAAUT"   "PAANHANG"  "PTRACTOR"  "PWERKT"    "PBROM"     "PLEVEN"    "PPERSONG
[57] "PGEZONG"   "PWAOREG"   "PBRAND"    "PZEILPL"   "PPLEZIER"  "PFIETS"    "PINBOED"   "PBYSTAND
[65] "AWAPART"   "AWABEDR"   "AWALAND"   "APERSAUT"  "ABESAUT"   "AMOTSCO"   "AVRAAUT"   "AAANHANG
[73] "ATRACTOR"  "AWERKT"    "ABROM"     "ALEVEN"    "APERSONG"  "AGEZONG"   "AWAOREG"   "ABRAND"
[81] "AZEILPL"   "APLEZIER"  "AFIETS"    "AINBOED"   "ABYSTAND"  "Purchase"
```

Example of Data: "Iris"

This famous (Fisher's or Anderson's) iris data set gives the measurements in centimeters of the variables sepal length and width and petal length and width, respectively, for 50 flowers from each of 3 species of iris. The species are Iris setosa, versicolor, and virginica.

```
> dim(iris)
[1] 150   5
> names(iris)
[1] "Sepal.Length" "Sepal.Width"  "Petal.Length" "Petal.Width"  "Species"
> summary(iris$Species)
    setosa versicolor  virginica
        50         50         50
```



Example of Data: "Hitters"

This dataset was taken from the StatLib library which is maintained at Carnegie Mellon University. This is part of the data that was used in the 1988 ASA Graphics Section Poster Session. The salary data were originally from Sports Illustrated, April 20, 1987. The 1986 and career statistics were obtained from The 1987 Baseball Encyclopedia Update published by Collier Books, Macmillan Publishing Company, New York.

```
> data("Hitters")
> dim(Hitters)
[1] 322  20
> names(Hitters)
 [1] "AtBat"     "Hits"      "HmRun"     "Runs"      "RBI"       "Walks"     "Years"
 [8] "CAtBat"    "CHits"     "CHmRun"    "CRuns"     "CRBI"      "CWalks"    "League"
[15] "Division"  "PutOuts"   "Assists"   "Errors"    "Salary"    "NewLeague"
> x = model.matrix(Salary~.,Hitters)[,-1]
> y = Hitters$Salary
>
```

Example of Data: "Boston"

Housing Values in Suburbs of Boston. The Boston dataframe has 506 rows and 14 columns.

```
> library('MASS')
> data("Boston")
> help(Boston)
> dim(Boston)
[1] 506  14
> names(Boston)
 [1] "crim"    "zn"      "indus"   "chas"    "nox"     "rm"      "age"     "dis"
[10] "tax"     "ptratio" "black"   "lstat"   "medv"
>
```

Supervised learning: Classification-Regression: $Y \in \mathcal{Y}$

- $\mathcal{Y}$ discrete $\rightarrow$ classification. $\mathcal{Y} = \{0, 1\}$ binary classification
- $\mathcal{Y}$ continuous $\rightarrow$ Regression

Purpose:

- Training set of data $(X_i, Y_i)_{1 \leq i \leq n}$ with unknown distribution P
- $(X, Y) \sim P, \quad Y$ unseen. Predict the outcome Y
- Construct, using the data, a prediction model, a learner $f : \mathcal{X} \rightarrow \mathcal{Y}$

Regression

Regression function of Y on X

$$f(x) = E(Y/X = x)$$

Define

$$\epsilon = Y - f(X) = Y - E(Y/X)$$

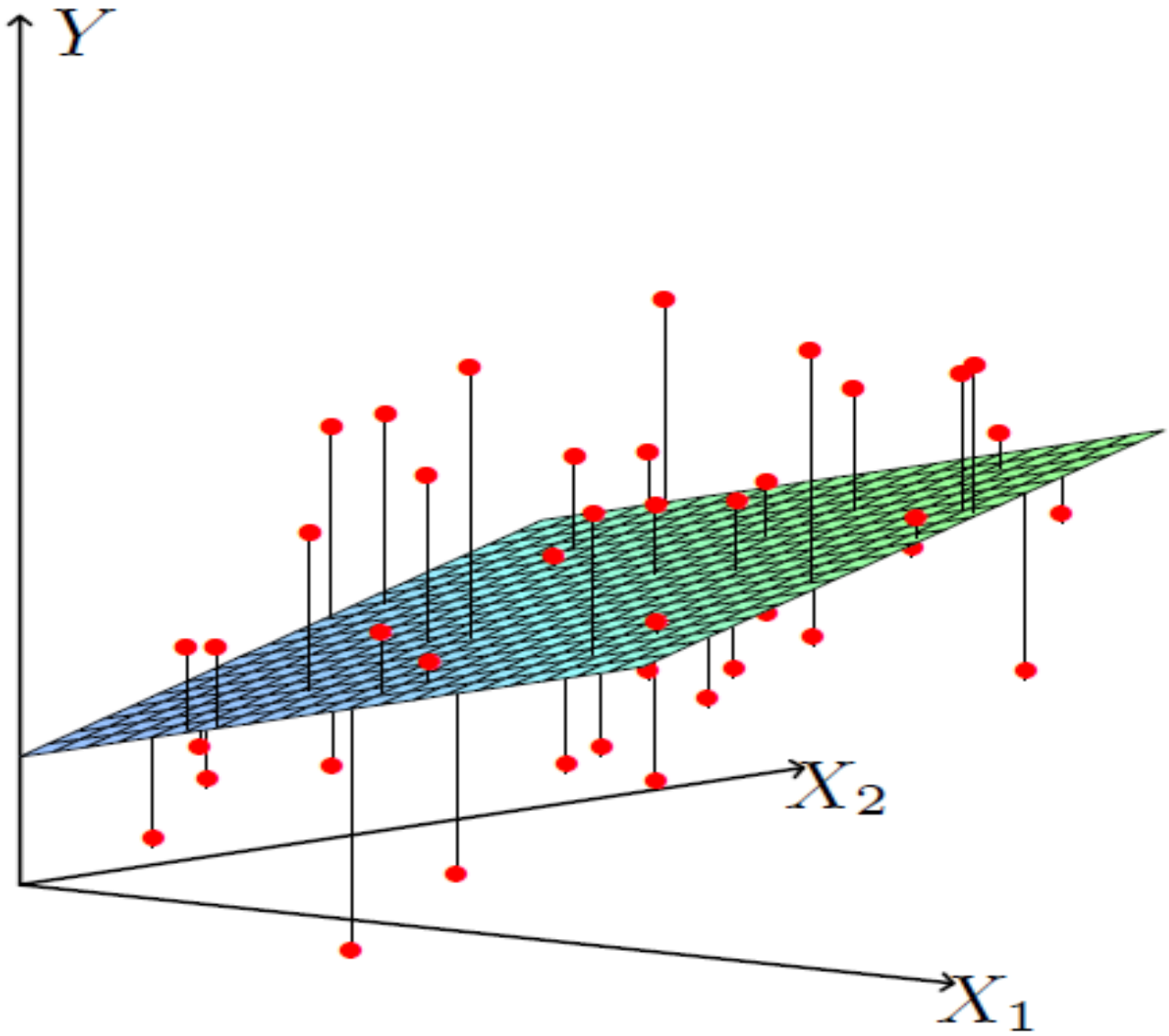
Model:

$$Y = f(X) + \epsilon, \;\; E(\epsilon/X) = 0.$$

Purpose: Estimate f based on the data set $(x_i, y_i)_{1 \leq i \leq n}$ realisations of $(X_i, Y_i)_{1 \leq i \leq n}$ iid with unknown distribution.

Parametric regression: the shape of f is known.

LS-Ridge-Lasso...



Non-parametric regression: the shape of f is unknown.

K Kernel: $\int |K(u)|du < \infty, \quad \int K(u)du = 1.$

Definition

Let $h > 0$ be a positive number, called the bandwidth. The Nadaraya-Watson kernel estimator of f is

$$\hat{f}_n(x) = \sum_{i=1}^{n} \frac{K(\frac{x-X_i}{h})}{\sum_{j=1}^{n} K(\frac{x-X_j}{h})} Y_i$$

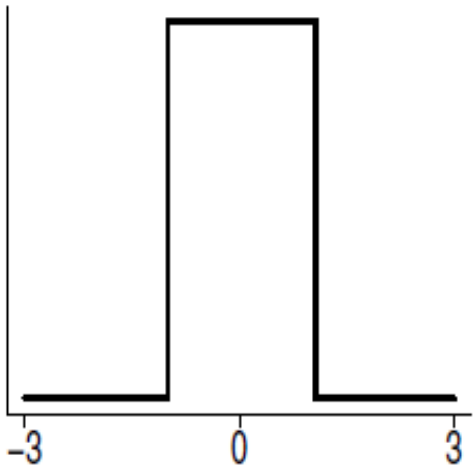
if $\sum_{j=1}^{n} K(\frac{x-X_j}{h}) \neq 0$ and 0 otherwise.

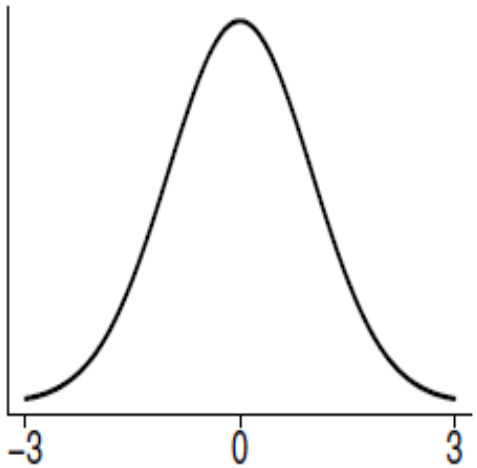
Kernel method

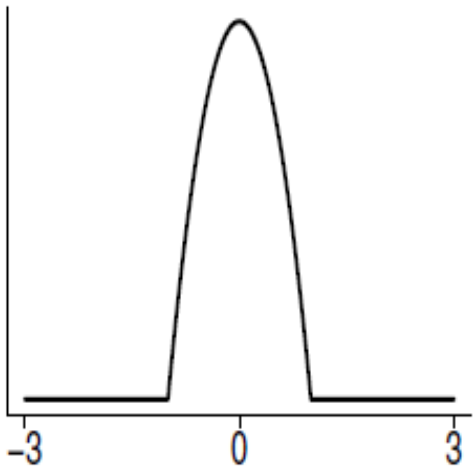
- X_1 has a known density g,

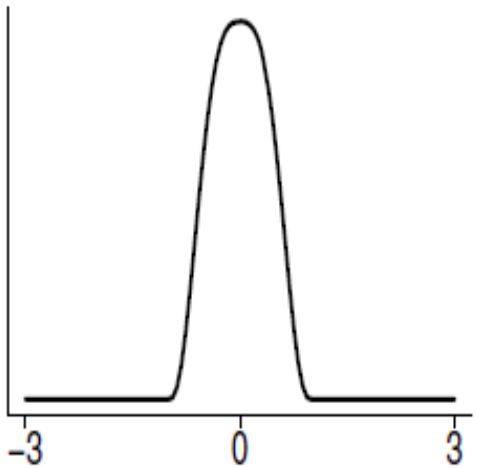
$$\hat{f}_n(x) = \frac{1}{nhg(x)} \sum_{i=1}^{n} K(\frac{x - X_i}{h}) Y_i$$

Kernel K









Kernel effects

The choice of kernel K is not too important. Estimates obtained by using different kernels are usually numerically very similar. This observation is confirmed by theoretical calculations which show that the risk is very insensitive to the choice of kernel; see Section 6.2.3 of Scott (1992).

L. Wasserman (2010). "All of Nonparametric Statistics". Springer.

h effects

K rectangular kernel

- h small, so that $|\frac{X_i - x}{h}|$ is either 0 or > 1

$$\hat{f}_n(x) = 0, \ \forall \ x \neq X_i \ \forall \ i \ \textit{and} \ \hat{f}_n(X_i) = Y_i$$

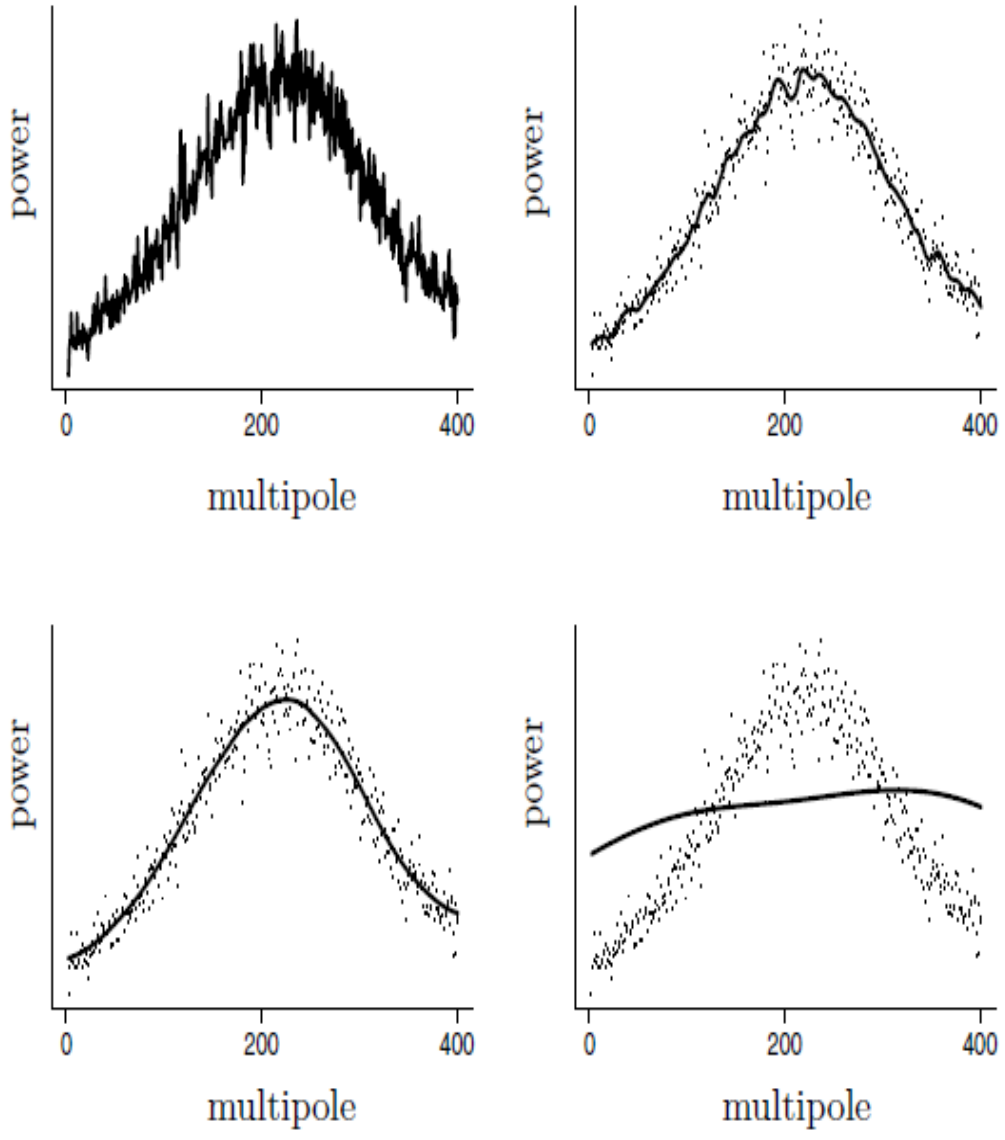
  undersmoothing, big variance

- h big

$$\hat{f}_n(x) \sim \bar{Y}$$

Oversmoothing, big bias

Impact of h


power
0
200
400
multipole
power
0
200
400
multipole
power
0
200
400
multipole
power
0
200
400
multipole

Kernel method: How to choose h?

R : risk

$$h_{opt} = argmin_h R(f, \hat{f}_n)$$

- MSE= $\mathbb{E}[(f(x) - \hat{f}_n(x))^2]$
- IMSE = $\int \mathbb{E}[(f(x) - \hat{f}_n(x))^2]dx$.

- MASE= $\frac{1}{n}\sum_{i=1}^{n} u(x_i)\mathbb{E}[(f(x_i) - \hat{f}_n(x_i))^2]$
- ASE = $\frac{1}{n}\sum_{i=1}^{n} u(x_i)(f(x_i) - \hat{f}_n(x_i))^2$.

How to choose h? Minimizer of the MASE

$f \in \mathcal{C}^2$, $x_i = \frac{i}{n}$, $Y_i = f(x_i) + \epsilon_i$

Lemma

Define,

$$D_n(h) = \frac{h^4}{4} \int_0^1 u(x) f''^2(x) dx \left(\int_{-1}^1 t^2 K(t) dt \right)^2$$
$$+ \frac{\sigma^2}{nh} \left(\int_0^1 u(x) dx \right) \int_{-1}^1 K^2(y) dy.$$

Then for any $n \geq 1$ and $h \in]0, \epsilon[$,

$$MASE(h) = D_n(h) + O(\frac{1}{n}) + O(h^5) + O(\frac{1}{n^2 h^4}) + \frac{\epsilon(h)}{nh},$$

where O is uniformly on n and h, $\epsilon(h)$ depends on h (but not on n) and tends to 0 when h tends to 0.

How to choose h? Minimizer of the MASE

Let $h_n^* = argmin_{h>0} D_n(h)$. Clearly, if $\int_0^1 u(x) f''^2(x) dx \neq 0$ then

$$h_n^* = n^{-1/5} \left(\frac{(\int_0^1 u(x)dx) \int_{-1}^1 K^2(y)dy\sigma^2}{\int_0^1 u(x)f''^2(x)dx(\int_{-1}^1 t^2 K(t)dt)^2} \right)^{1/5} =: cn^{-1/5}.$$

Problem ?

How to choose h? Minimizer of the ASE

$$ASE(h) = \frac{1}{n}\sum_{i=1}^{n} u(x_i)(\hat{f}(x_i) - f(x_i))^2,$$

$$\hat{h}_n = argmin_h ASE(h)$$

Problem ?

How to choose h?

$$\frac{1}{n}\sum_{i=1}^{n} u(x_i)(Y_i - \hat{f}(x_i))^2$$

Problem ?

A poor estimate of MASE(h): it is biased downwards and typically leads to overfitting. The reason is that we are using the data twice: to estimate the function and to estimate the risk.

How to choose h? Cross Validation

Definition

$$CV(h) = \frac{1}{n}\sum_{i=1}^{n} u(x_i)(Y_i - \hat{f}^{-i}(x_i))^2$$

where $\hat{f}^{-i}$ is the estimator obtained by omitting the ith pair (x_i, Y_i).

$$\hat{f}(x) = \sum_{j=1}^{n} l_{j,n}(x) Y_j \quad \hat{f}^{-i}(x) = \sum_{j=1}^{n} l_{j,n}^{-i}(x) Y_j,$$

$$l_{j,n}^{-i}(x) = \frac{l_{j,n}(x)}{\sum_{j,\, j\neq i} l_{j,n}(x)} \mathbb{1}_{j\neq i}.$$

How to choose h? Generalized Cross Validation

Lemma

$$CV(h) = \frac{1}{n}\sum_{i=1}^{n} u(x_i)\frac{(Y_i - \hat{f}(x_i))^2}{(1 - l_i(x_i))^2}$$

Definition

$$GCV(h) = \frac{1}{n}\sum_{i=1}^{n} u(x_i)\frac{(Y_i - \hat{f}(x_i))^2}{(1 - \nu/n)^2}$$

$$\nu = tr(L) = \sum_{i=1}^{n} l_i(x_i) = \frac{1}{nh}K(0)$$

$$(1-x)^{-2} \sim 1 + 2x \quad x \sim 0.$$

How to choose h? Mallows criterion

Definition

$$C_p := C_p(h) = \frac{1}{n}\sum_{i=1}^{n} u(x_i)(Y_i - \hat{f}(x_i))^2 + 2\frac{\nu}{n}\hat{\sigma}^2,$$

where,

$$\hat{\sigma}^2 = \frac{1}{n}\sum_{i=1}^{n} u(x_i)(Y_i - \hat{f}(x_i))^2, \quad \nu = tr(L) = \sum_{i=1}^{n} l_i(x_i) = \frac{1}{nh}K(0).$$

$h_n = argmin_h MASE(h) \quad \hat{h}_n = argmin_h ASE(h), \quad \hat{h}_{ML} = argmin_h C_p(h).$

All those windows are nearly equivalent in probability.

Paper: W. Härdle, P. Hall and J. S. Marron (1988). How far are automatically chosen regression smoothing parameters from their optimum? Journal of the American Statistical Association 83, 86-95.

Part 2: Nonparametric statistics for dependent data

Dependent data

- Discrete time dependent data
  - $AP(p)$: Autoregressive model of order p

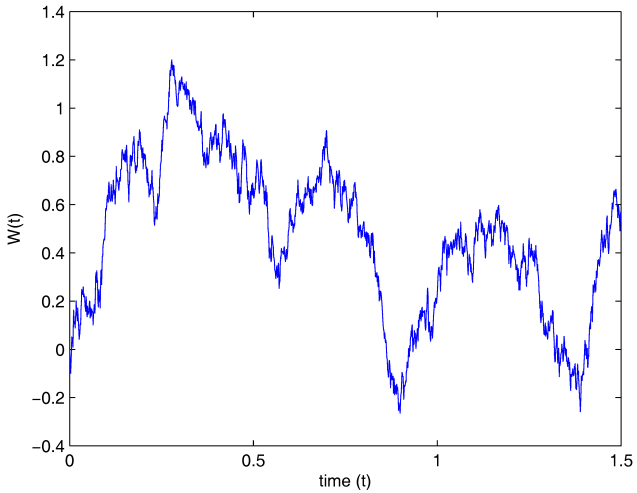
$$\epsilon_n = \sum_{j=1}^{p} \alpha_j \epsilon_{n-j} + \alpha_0 + \eta_n$$

  - $ARCH(p)$: Autoregressive Conditional Heteroscedasticity of order p,

$$\epsilon_n = \sqrt{\sum_{j=1}^{p} \alpha_j \epsilon_{n-j}^2 + \alpha_0} + \eta_n$$

Dependent data

- Continuous time.
  $g(t)$ the digoxin plasma concentration after an oral dosage.
  Data: $g(t_1), \cdots, g(t_n)$
  - Wiener process $(W_t)_{t\geq 0}$ $\mathrm{Cov}(W_t, W_s) = \min(s, t)$.

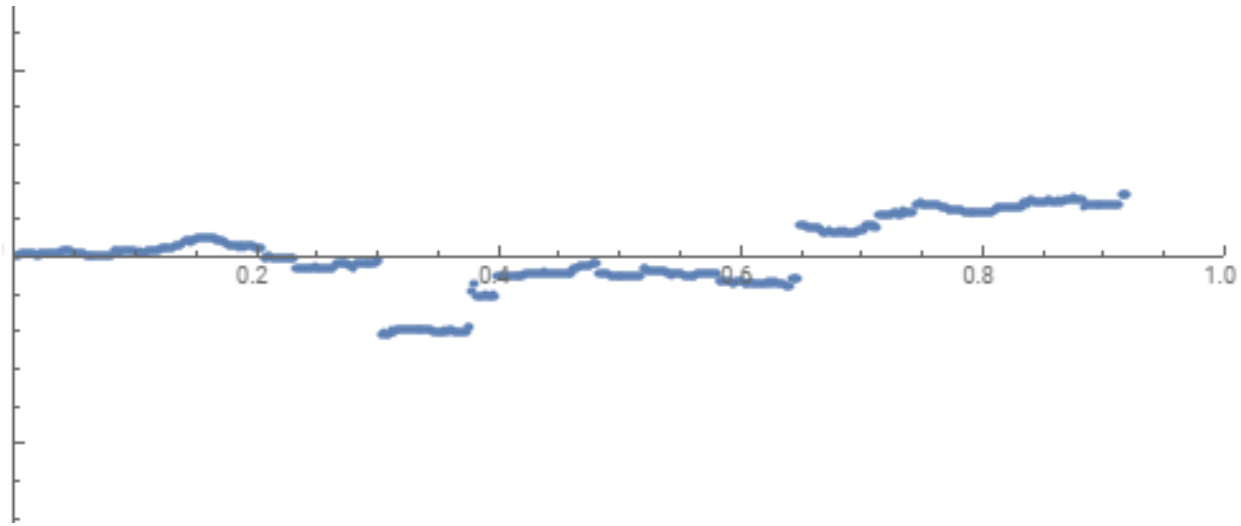


  - Ornstein-Uhlenbeck process $(X_t)_{t\geq 0}$, $\mathrm{Cov}(X_t, X_s) = e^{-|t-s|}$.

$$dX_t = \sigma dW_t + (b - X_t)dt.$$

Dependent data

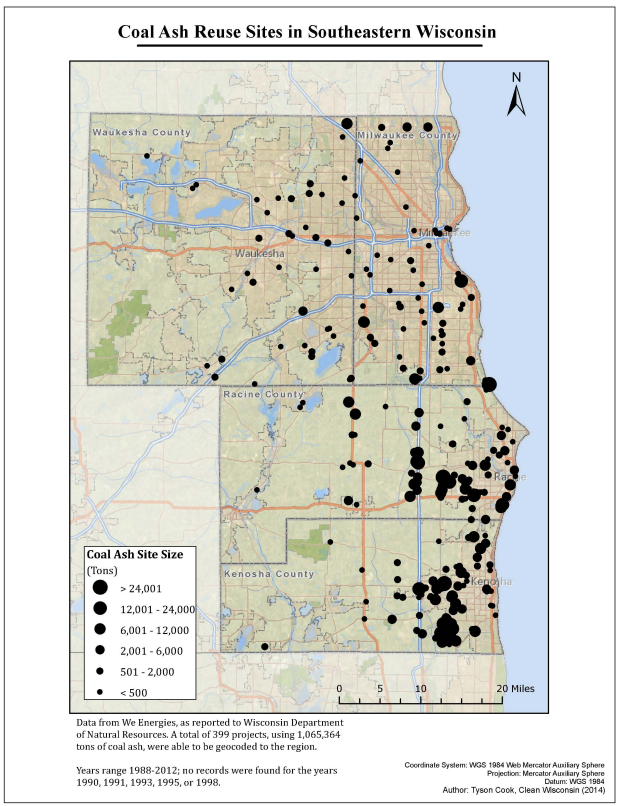
- Continuous time



Dependent data

- Spatial dependence $(\epsilon_x)_{x \in S}$ $S \subset R^d$.

library(gstat) data(coalash)
The coalash data frame is a spatial data set with 208 rows and 3 columns: x, y and coal.



Dependent data

- Spatio-temporal dependence $(\epsilon_x(t))_{x \in S,\ t \geq 0}$ $S \subset R^d$.

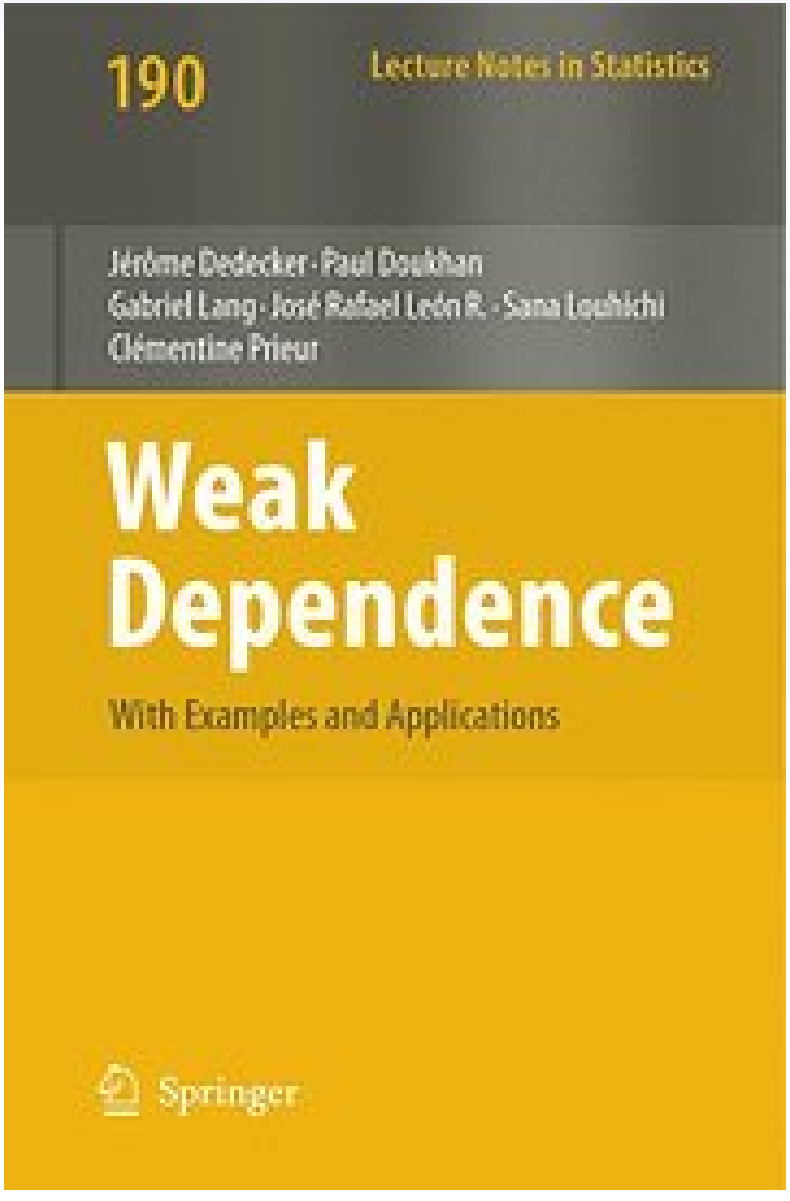
French pollution data: The data cover the French territory for the year 2014. Time series of hourly and daily observed concentrations of four pollutants (ozone, nitrogen dioxide, particulate matter PM10 and PM2.5) were retrieved from the national air quality database. They come from the continuous measurements carried out by the French associations responsible for air quality monitoring (the AASQAs) at 507 stations. Air quality is measured in different types of location characterized by the station environment (rural, suburban, urban) and the type of influence (background, traffic, industrial).

Dependent data

- Spatio-temporal dependence $(\epsilon_x(t))_{x \in S,\ t \geq 0}$ $S \subset R^d$.

STATION	TIME	PM10	PM25	NO2	O3
station1	2015-01-01	24.645	.	8.234	67.123
station1	2015-01-02	36.765	.	7.233	89.234
station1	2015-01-03	23.233	.	8.219	90.111
...					
station2	2015-01-01	32.860	35.233	.	45.222
station2	2015-01-02	33.460	12.231	.	23.433
...					

Source: Analyzing spatio-temporal data with R: Everything you always wanted to know - but were afraid to ask. Journal de la Société Française de Statistique.


190
Lecture Notes in Statistics
Jérôme Dedecker · Paul Doukhan
Gabriel Lang · José Rafael León R. · Sana Louhichi
Clémentine Prieur
Weak Dependence
With Examples and Applications
Springer

Description and measures of dependence

- **Martingale difference sequences**:
  ϵ_i est $\mathcal{F}_i$-measurable and

$$\mathbb{E}(\epsilon_i | \mathcal{F}_{i-1}) = 0.$$

- **Associated sequences**: for any f, g coordinatewise nondecreasing

$$\mathrm{Cov}(f(\epsilon_1, \cdots, \epsilon_n), g(\epsilon_1, \cdots, \epsilon_n)) \geq 0.$$

- **Markov chains**:

$$\epsilon_n = F(\epsilon_{n-1}, \eta_n), \ \eta_n \perp (\epsilon_i)_{i \leq n-1}$$

- **Mixing (strongly mixing, β-mixing, ϕ-mixing, $\cdots$)**

$$\alpha_n = \sup_k \alpha(\sigma(\epsilon_i, \ i \leq k), \sigma(\epsilon_i, \ i \geq k+n))$$

$$\alpha(\mathcal{A}, \mathcal{B}) = \sup_{A \in \mathcal{A}, B \in \mathcal{B}} |\mathrm{Cov}(\mathbb{1}_A, \mathbb{1}_B)|$$

Mathematical tools for independent and centered observations

- **Moments inequalities: Marcinkiewicz-Zygmund inequality**

$$\mathbb{E}\left|\sum_{i=1}^{n}\epsilon_i\right|^p \leq C_p n^{p/2}$$

- **Moments inequalities: Rosenthal type inequality**

$$\mathbb{E}\left|\sum_{i=1}^{n}\epsilon_i\right|^p \leq C_p\left(\left(\mathrm{Var}(\sum_{i=1}^{n}\epsilon_i)\right)^{p/2}+\sum_{i=1}^{n}\mathbb{E}(|\epsilon_i|^p)\right)$$

Mathematical tools for independent and centered observations

- **Lévy's Maximal inequality**:

$$\mathbb{P}\left(\max_{k\leq n}\sup_{s\in S}|\sum_{i=1}^{k}\epsilon_{i,s}|\geq t\right)\leq 2\mathbb{P}\left(\sup_{s\in S}|\sum_{i=1}^{n}\epsilon_{i,s}|\geq t\right),$$

where $(\epsilon_{1,s})_{s\in S},\cdots,(\epsilon_{n,s})_{s\in S}$ are independent.

Mathematical tools for independent and centered observations

- **Lévy's Maximal inequality**:

$$\mathbb{P}\left(\max_{k\leq n}\sup_{s\in S}|\sum_{i=1}^{k}\epsilon_{i,s}|\geq t\right)\leq 2\mathbb{P}\left(\sup_{s\in S}|\sum_{i=1}^{n}\epsilon_{i,s}|\geq t\right),$$

  where $(\epsilon_{1,s})_{s\in S},\cdots,(\epsilon_{n,s})_{s\in S}$ are independent.

- **Central limit theorem for quadratic form**:

$$Y_i(h)=a_i(h)\epsilon_i+\epsilon_i\sum_{j=1}^{i-1}b_{i,j}(h)\epsilon_j$$

Empirical choice of h for dependent observations

- Effect of the dependence in Mallows criterion
- Tools in the dependence context

Paper: K. Benhenni, D. Girard and S. Louhichi (2019). On smoothing parameters selection problems in nonparametric regression models with martingale difference errors. Submitted.

We suppose the following two assumptions.

Assumptions (A) and notation. Suppose that both the functions $h \longmapsto T_n(h) := ASE(h)$ and $h \longmapsto CL(h) := C_p(h)$ are with continuous first time differentiable, that $T_n'(\hat{h}_n) = 0$ and that $CL'(\hat{h}_{ML}) = 0$ almost surely. Suppose also that the function $h \longmapsto \mathbb{E}(T_n(h))$ is twice differentiable with continuous second derivative. Suppose that $\frac{\partial^2}{\partial^2 h}\mathbb{E}(T_n(h)) = \mathbb{E}(T_n''(h))$.

Assumptions (B). The errors $(\epsilon_i)_{i\geq 0}$ form a stationary centered MDS with respect to its natural filtration, i.e, for any $i > 0$, ϵ_i is $\mathcal{F}_i$-measurable and $\mathbb{E}(\epsilon_i|\mathcal{F}_{i-1}) = 0$ where $\mathcal{F}_i = \sigma(\epsilon_1, \cdots, \epsilon_i)$.

Our first result proves that for MDS of errors, the selected bandwidths $h_n, h_n^*, \hat{h}_n$ and $\hat{h}_{ML}$ are nearly equivalent.

Proposition

Suppose that Assumptions (A) and (B) are satisfied. Then

$$\frac{h_n^*}{h_n}, \frac{\hat{h}_n}{h_n}, \frac{\hat{h}_{ML}}{h_n}$$

converge all in probability to 1 as n tends to infinity.

Our second result gives the rate at which $\hat{h}_n - \hat{h}_{ML}$ converges in distribution to a centered normal law.

Theorem

Suppose that Assumptions (A) and (B) are satisfied. Suppose, moreover, that there exists a positive decreasing function Φ defined on $\mathbb{R}^+$ satisfying

$$\sum_{s=1}^{\infty} s^4 \Phi(s) < \infty,$$

and for any positive integer q less than 6,
$1 \leq i_1 \leq \cdots \leq i_k < i_{k+1} \leq i_q \leq n$ such that
$i_{k+1} - i_k \geq \max_{1 \leq l \leq q-1}(i_{l+1} - i_l)$

$$|\mathrm{Cov}(\epsilon_{i_1} \cdots \epsilon_{i_k}, \epsilon_{i_{k+1}} \cdots \epsilon_{i_q})| \leq \Phi(i_{k+1} - i_k). \qquad (1)$$

Then

$$n^{3/10}(\hat{h}_n - \hat{h}_{ML})$$

converges in distribution to a centered normal law with variance Σ^2 given by

$$\Sigma^2 = \frac{4\sigma^{12/5}B^{1/5}}{5A^{6/5}}\left(\int t^2K(t)dt\right)^2 \int_0^1 u^2(x)f''^2(x)dx$$
$$+\frac{16\sigma^{12/5}}{5A^{1/5}B^{4/5}} \int_0^1 u^2(x)dx \int_0^1 (K-G)^2(u)du,$$

where $\sigma^2 = \mathbb{E}(\epsilon_1^2)$, G is the function defined for any $x \in \mathbb{R}$ by $G(x) = -xK'(x)$ and

$$A = \int_0^1 u(x)f''^2(x)dx \int t^2K(t)dt, \quad B = \int_0^1 u(x)dx \int K^2(t)dt.$$

Application to ARCH(1) processes

We consider an ARCH(1) process defined, for $n \geq 1$, by the following stochastic difference equation,

$$\epsilon_n = \eta_n \sqrt{\sigma^2(1-\alpha) + \alpha\epsilon_{n-1}^2}, \ 0 \leq \alpha < 1, \ \sigma^2 > 0 \qquad (2)$$

where $(\eta_n)_{n\geq 1}$ is an iid centered sequence distributed as a standard normal law and such that η_n is independent of $(\epsilon_1, \cdots, \epsilon_{n-1})$.

Proposition

Let (ϵ_n) be a strictly stationary ARCH(1) process as defined by (2). Suppose that $\alpha^8 \prod_{i=1}^{8}(2i-1) < 1$. Then the requirements of Theorem 1 are satisfied by the sequence (ϵ_n).

Tools for martingale difference sequences

We recall the following Marcinkiewicz-Zygmund type inequality which is a simple consequence of the Minkowski and the Burkholder inequalities (see Burkholder (1988)).

Theorem

Let $(\eta_i)_{i\geq 0}$ be a stationary centered sequence of martingale difference of finite pth moment with $p \geq 2$. Then there exists a positive constant c_p such that for any positive integer n,

$$\left\| \sum_{i=1}^{n} \eta_i \right\|_p^2 \leq c_p \sum_{i=1}^{n} \|\eta_i\|_p^2.$$

An immediate consequence of Theorem 2 is the following corollary.

Corollary

Let $(\eta_i)_{i\geq 0}$ be a stationary sequence of martingale difference of finite pth moment with $p \geq 2$. Then there exists a positive constant c_p such that for any positive integer n,

$$\left\| \sum_{i=1}^{n} d_{i,n}\eta_i \right\|_p^2 \leq c_p \sum_{i=1}^{n} d_{i,n}^2,$$

and for any sequence of real numbers $(d_{i,n})_{1\leq i\leq n}$.

We also need the following proposition whose proof uses Theorem 2 above.

Proposition

Let $(\eta_i)_{i\geq 0}$ be a stationary sequence of martingale difference such that $\|\eta_i\|_{2p} < \infty$ for some $p \geq 2$. Then, there exists a positive constant c_p such that for any positive integer n,

$$\left\| \sum_{i=1}^{n} \sum_{j=1}^{i-1} b_{i,j,n} \eta_j \eta_i \right\|_p^2 \leq c_p \sum_{i=1}^{n} \sum_{j=1}^{i-1} b_{i,j,n}^2,$$

and for any sequence of real numbers $b_{i,j,n}$.

The following maximal inequality is also very needed in the proofs. Its proof needs some chaining arguments.

Lemma

Let $(\eta_i)_{i\geq 0}$ be a sequence of stationary martingale difference with $\|\eta_i\|_p < \infty$ for some $p \geq 2$. Let $(c_{i,n}(h))_{i,n,h}$ be a sequence of weights satisfying, for any $h, h' \in H_n = [an^{-1/5}, bn^{-1/5}]$,

$$|c_{i,n}(h) - c_{i,n}(h')| \leq cst\,|h - h'|.$$

and

$$\max_{i\leq n} \sup_{h\in H_n} |c_{i,n}(h)| \leq cst\, n^{-\alpha},\ \alpha > \frac{5p-2}{10(p-1)}.$$

Then,

$$\lim_{n\to\infty} \left\| \sup_{h\in H_n} \left| \sum_{i=1}^{n} c_{i,n}(h)\eta_i \right| \right\|_p = 0.$$

Lemma

Let $(\epsilon_j)_j$ be a sequence of random variables with finite fourth moment and such that,

$$\sup_i \sum_{j=1}^{\infty} |\mathrm{Cov}(\epsilon_i^2, \epsilon_j^2)| < \infty.$$

Let for $h \in H_n = [an^{-1/5}, bn^{-1/5}]$, $(d_{j,n}(h))_{1 \leq j \leq n}$ be a sequence of real numbers satisfying for any $1 \leq j \leq n$,

$$|d_{j,n}(h)| \leq \frac{cst}{n}, \quad and, \quad |d_{j,n}(h) - d_{j,n}(h')| \leq cst\, n^{-2/5}|h - h'|.$$

Then,

$$\lim_{n \to \infty} \left\| \sup_{h \in H_n} \left| \sum_{i=1}^{n} d_{i,n}(h)(\epsilon_i^2 - \mathbb{E}(\epsilon_i^2)) \right| \right\|_2 = 0.$$

Lemma

Let $(\eta_i)_{i\geq 0}$ be a stationary sequence of martingale difference random variables with finite moment of order $2p$, for some $p > 8$. Suppose that, for any $h, h' \in H_n$

$$|b_{i,j,n}(h)| \leq \frac{cst}{n} \mathbb{1}_{|i-j|\leq 2nh},$$

$$|b_{i,j,n}(h) - b_{i,j,n}(h')| \leq cst\, n^{-4/5} |h - h'| \mathbb{1}_{|i-j|\leq 2n \max(h,h')}.$$

Then,

$$\lim_{n\to\infty} \left\| \sup_{h\in H_n} \left| \sum_{i=1}^{n} \sum_{j=1}^{i-1} b_{i,j,n}(h) \eta_j \eta_i \right| \right\|_p = 0.$$

CLT

Recall that $K - G$ is an even function, $[-1, 1]$-supported, that the window h_n is a positive sequence satisfying

$$\lim_{n\to\infty} h_n = 0, \quad \lim_{n\to\infty} nh_n = \infty.$$

Define, for $i = 1, \cdots, n$, $x_i = \frac{i}{n}$ and, for a positive constant C_K depending only on K,

$$a_{i,n}(h_n) = C_K \frac{h_n}{n} f''(x_i)u(x_i)$$

$$b_{i,j}(h_n) = \frac{1}{n^2 h_n^2}(K - G)(\frac{x_i - x_j}{h_n})$$

$$\tilde{b}_{i,j} = b_{i,j}(h_n)(u(x_i) + u(x_j)).$$

Let $(\epsilon_i)_{i\geq 0}$ be a centered sequence of stationary MD random variables with finite second moment σ^2. The purpose of this section is to prove, letting

$$Y_{i,n}(h_n) = a_{i,n}(h_n)\epsilon_i + \sum_{j=1}^{i-1} \tilde{b}_{i,j}\epsilon_i\epsilon_j, \tag{3}$$

CLT: Control of the variance

Proposition

Suppose that there exists a positive decreasing function Φ defined on $\mathbb{R}^+$ satisfying

$$\sum_{s=1}^{\infty} s^2 \Phi(s) < \infty,$$

and for any $1 \leq i_1 \leq i_2 < i_3 \leq i_4 \leq i_5 \leq n$ such that $i_3 - i_2 \geq \max(i_2 - i_1, i_4 - i_3, i_5 - i_4)$

$$|\mathrm{Cov}(\epsilon_{i_1}\epsilon_{i_2}, \epsilon_{i_3}\epsilon_{i_4})| \leq \Phi(i_3 - i_2)$$
$$|\mathrm{Cov}(\epsilon_{i_2}, \epsilon_{i_3}\epsilon_{i_4}\epsilon_{i_5})| \leq \Phi(i_3 - i_2).$$

Then

$$\mathrm{Var}\left(\sum_{i=1}^{n} Y_{i,n}(h_n)\right) = \frac{h_n^2\sigma^2}{n} C_K^2 \int u^2(x) f''^2(x) dx$$
$$+\frac{4\sigma^4}{n^2 h_n^3} \int_0^1 u^2(x) dx \int_0^1 (K-G)^2(u) du$$
$$+o(\frac{1}{n^2 h_n^3} + \frac{h_n^2}{n}).$$

Proposition

Let $(\epsilon_i)_{i\geq 0}$ be a stationary sequence of centered martingale difference random variables relative to the filtration $\mathcal{F}_i = \sigma(\epsilon_1, \cdots, \epsilon_i)$. Suppose that $\mathbb{E}(\epsilon_1^8) < \infty$. Suppose, moreover, that there exists a positive decreasing function Φ defined on $\mathbb{R}^+$ satisfying

$$\sum_{s=1}^{\infty} s^4 \Phi(s) < \infty,$$

and for any positive integer $q \leq 6$,
$1 \leq i_1 \leq \cdots \leq i_k < i_{k+1} \leq i_q \leq n$ such that
$i_{k+1} - i_k \geq \max_{1\leq l\leq k}(i_{l+1} - i_l)$

$$|\mathrm{Cov}(\epsilon_{i_1} \cdots \epsilon_{i_k}, \epsilon_{i_{k+1}} \cdots \epsilon_{i_q})| \leq \Phi(i_{k+1} - i_k).$$

Let $Y_{in}(h_n)$ be as defined in (3) with $h_n = cn^{-1/5}$.

Then

$$n^{7/10} \sum_{i=1}^{n} Y_{i,n}(h_n) \Longrightarrow \mathcal{N}(0, V),$$

where $\Longrightarrow$ denotes convergence in distribution when n tends to infinity, the variance V is defined by,

$$V = c^2 C_K^2 \sigma^2 \int_0^1 u^2(x) f''^2(x) dx$$
$$+ \frac{4}{c^3} \sigma^4 \int_0^1 u^2(x) dx \int_0^1 (K - G)^2(u) du,$$

and $\sigma^2 = \mathbb{E}(\epsilon_1^2)$.