Nonparametric estimation for multivariate data streams

Aboubacar AMIRI *

Université Lille 3, LEM-CNRS (UMR 9221)

Journée lilloise de Proba-Stat

6 janvier 2017

*. *aboubacar.amiri@univ-lille3.fr*

Outline

Equipe Lille 3 (1/2)

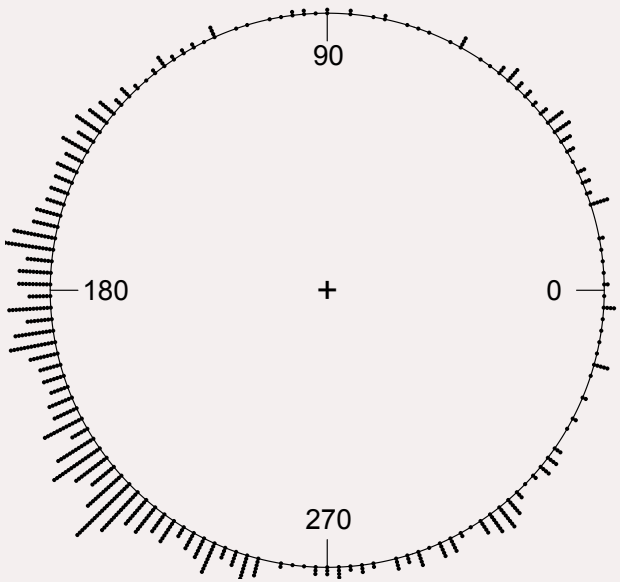
- Laurence Broze : "Séries Temporelles et Econométrie ".
- Sophie Dabo-Niang : "Statistique non-paramétrique, Statistique fonctionnelle, Statistique spatiale".
- Christian Francq : "Séries Temporelles, Econométrie, Finance"
- Jean Michel Zakoan : "Séries Temporelles, Econométrie, Finance".
- Antony Gautier : "Séries Temporelles, Econométrie".
- Olivier Torres : "Econométrie" .
- Emmanuel Thilly : "Processus Stochastiques, Processus Gaussiens, Processus de Levy,
- Camille Sabbah : "Statistique non-paramétrique, Statistique directionnelle".
- Baba Thiam : "Statistique non-paramétrique, Statistique spatiale, Statistique directionnelle".

Data streams

1. Simulation 1
2. Simulation 2

- Data streams are massive data arriving in streams, and if they are not analyzed immediately or stored, then they are lost forever.
- In many scientific and real applications, large amount of raw data can be collected extremely easily so that experiments typically yield to a huge number of data points.
- In those situations, the data arrive so rapidly that it is impossible for the user to store them all in disk (as a traditional database), and then interact with them at the time of our choosing.

Wind directions data.



- *A dataset of wind directions recorded minutely in June 2012 in Mourela, the north of Spain.*
- *Many observed data points can be available in a very short period of time.*
- *The computational time of many classical methods can quickly become large.*

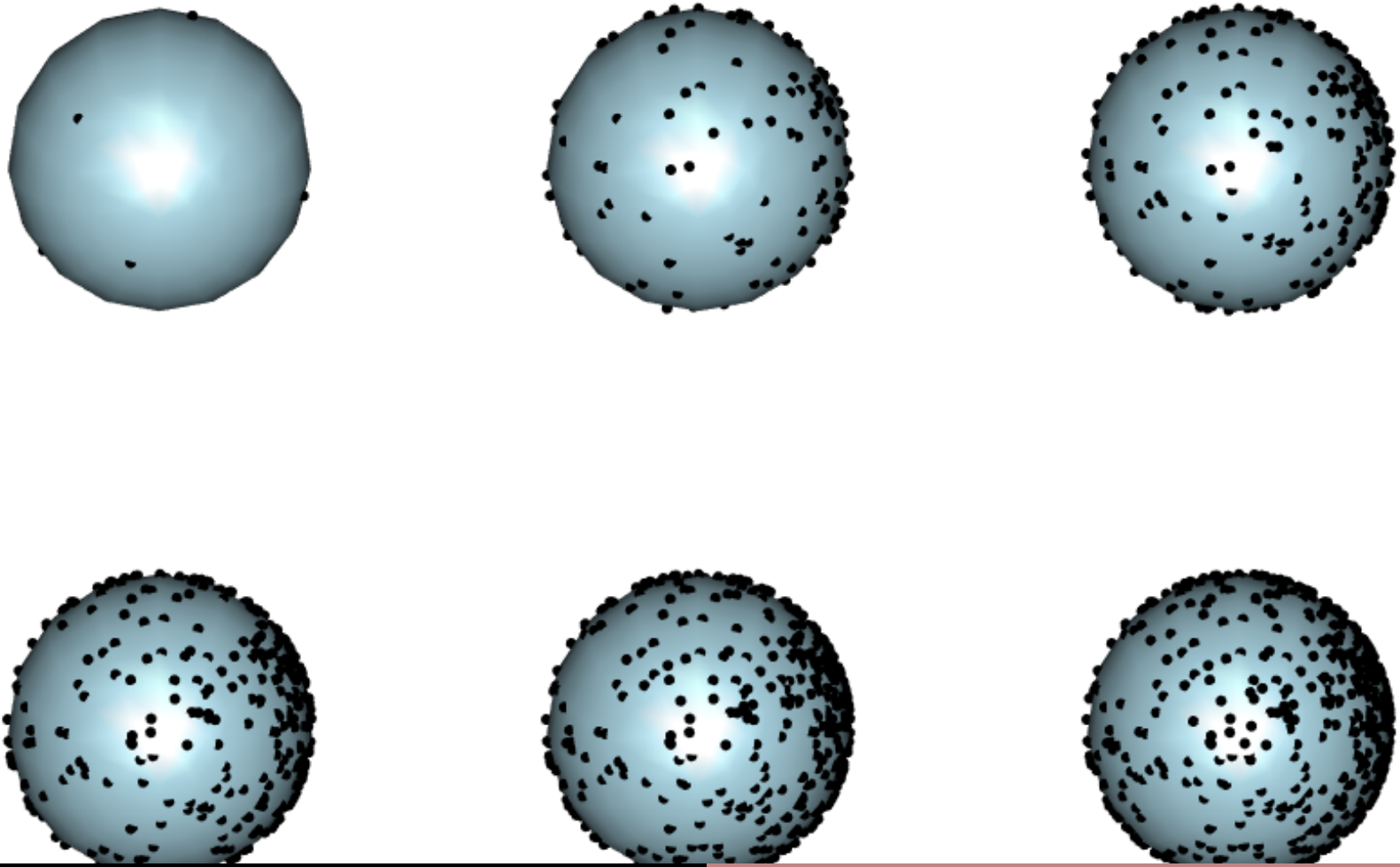
- *If the wind directions are recorded continuously (and minutely), a single week yields around 10080 observations on the unit circle.*

NASA MAGSAT dataset †.

- The NASA MAGSAT dataset consists in directions of the magnetic field.
- Measurements were made by the NASA's MAGSAT spacecraft between Nov. 2, 1979 and May 6,1980.

†. Available from : http ://omniweb.gsfc.nasa.gov/ftpbrowser/magsat.html. This home page provides listing for magnetic field vectors, with a resolution around one observation per 0.5 second.

A scatterplot of directional data streams :



Listing for magsat data from 198001010000 to 198001012359

Selected parameters :
1 Geocentric Lat.
2 Geocentric Long.
3 Radial distance

MILLISEC	1	2	3
14181	68.296	-111.378	6881.902
14672	68.326	-111.406	6881.914
15164	68.355	-111.435	6881.922
15655	68.384	-111.464	6881.934
16147	68.414	-111.493	6881.949
16638	68.443	-111.523	6881.961
17130	68.472	-111.552	6881.969
17621	68.502	-111.581	6881.980
18604	68.560	-111.640	6882.004
19096	68.589	-111.669	6882.016
19587	68.619	-111.699	6882.027
20079	68.648	-111.728	6882.035
18604	68.560	-111.640	6882.004
19096	68.589	-111.669	6882.016
19587	68.619	-111.699	6882.027
20079	68.648	-111.728	6882.035
20571	68.677	-111.758	6882.051
21062	68.707	-111.788	6882.059
21554	68.736	-111.818	6882.070
22045	68.765	-111.848	6882.082
22537	68.794	-111.878	6882.094
23028	68.824	-111.908	6882.105
23520	68.853	-111.938	6882.113
24011	68.882	-111.968	6882.125
24503	68.911	-111.998	6882.137
24994	68.941	-112.029	6882.148
25486	68.970	-112.059	6882.160
25978	68.999	-112.090	6882.168
26469	69.028	-112.120	6882.184
26961	69.057	-112.151	6882.195
27452	69.087	-112.182	6882.203
27944	69.116	-112.212	6882.215
28435	69.145	-112.243	6882.227
⋮	⋮	⋮	⋮

170604 obs/day.

- Many observed data points can be available in a very short period of time.
- The sample size can rapidly becomes huge.
- For example, if the data are recorded continuously, a single week yields around 1194228 observations.

Intel Lab dataset ‡

- Data collected from 54 sensors deployed in the Intel Berkeley Research Laboratory.

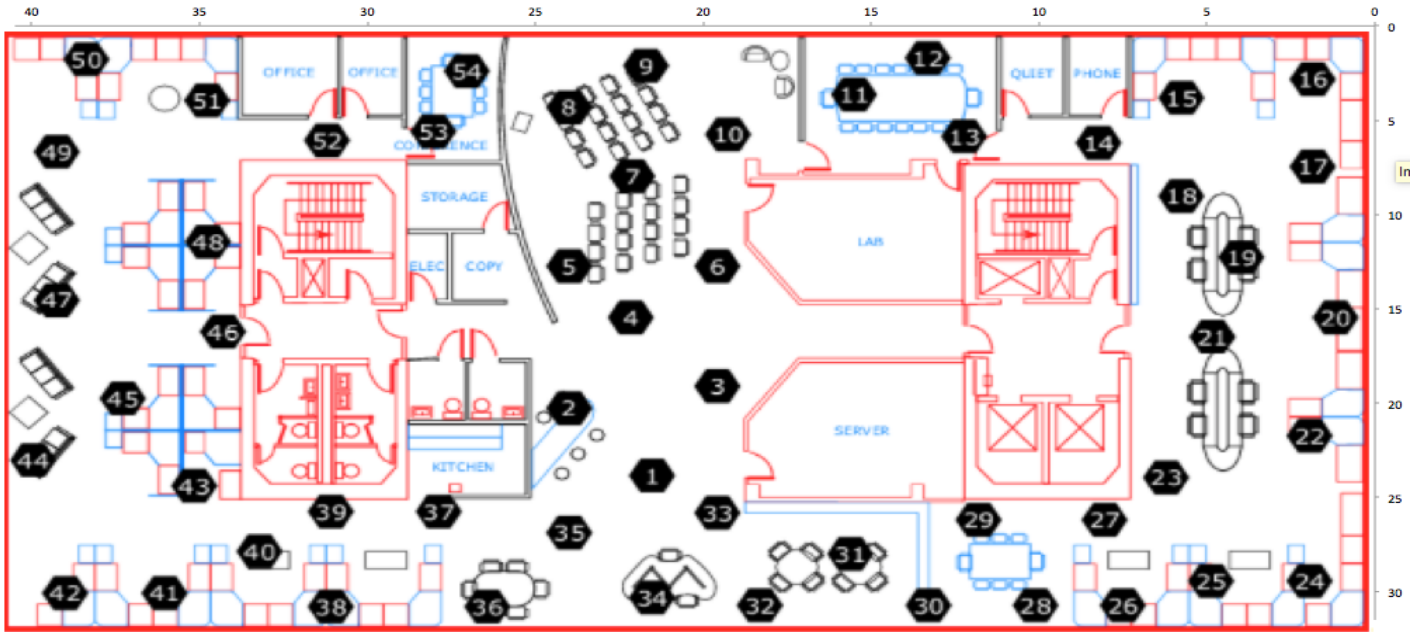


Figure: Spatial positions of the 54 sensors in the Intel Berkeley Research Laboratory

‡. http ://db.csail.mit.edu/labdata/labdata.html.

Intel Lab dataset

- The data consist of humidity, temperature, light and voltage measurements recorded every 31 seconds between February 28th and April 5th, 2004.
- Some data might be missing for specific times due to sensor failures.
- A single day yields a number of observations ranging from 0 to 2615 over one sensor.

- The computational time of many classical methods can quickly become large, so that a practitioner will not be able to obtain a prediction rapidly.
- After a certain period of time, the statistician who has to perform estimation based on traditional techniques can be tempted by throwing away a part of the sample since the estimation will start to take too much time to be updated.

- Consequently, to deal with such massive data, the traditional nonparametric techniques rapidly require a lot of time to be computed and therefore become useless in practice if real time forecasts are expected.
- How to process and analyze these data streams effectively and efficiently ?

Aggarwal (Springer 2007), Domingos and Hulten (JCGS 2003), Cao et al. (IEEE TNNLS 2012), Xu et al. (Front Comput Sci 2014)

Density estimation for directional data streams

- Spherical or directional data are concerned with multivariate data for which only the directions (and not the magnitudes) are observed.
- The resulting data points belong to the unit sphere

$$\mathcal{S}^{p-1} := \{\mathbf{v} \in \mathbb{R}^p, \mathbf{v}'\mathbf{v} = 1\}$$

of $\mathbb{R}^p$.

- Assume that we sequentially observe independent random matrices (called windows)

$$\mathbf{W}_t := (\mathbf{X}_{t1}, \ldots, \mathbf{X}_{tN_t})_{t \geq 1},$$

such that the columns $\mathbf{X}_{11}, \ldots, \mathbf{X}_{1N_1} \cdots \mathbf{X}_{n1} \ldots \mathbf{X}_{nN_n}$ of $\mathbf{W}_1, \ldots, \mathbf{W}_n$ are i.i.d. absolutely continuous (with respect to the usual surface area measure ω_p on $\mathcal{S}^{p-1}$) random vectors on $\mathcal{S}^{p-1}$ with density f.

- Letting $\hat{f}(\mathbf{W}_1, \ldots, \mathbf{W}_n)$ stand for a kernel density estimator computed from the first n windows.

- How to provide an estimator $\hat{f}$ such that $\hat{f}(\mathbf{W}_1, \ldots, \mathbf{W}_n, \mathbf{W}_{n+1})$ can be computed extremely quickly from $\hat{f}(\mathbf{W}_1, \ldots, \mathbf{W}_n)$ (on-line estimation) while keeping nice efficiency properties with respect to its natural competitors ?

R.M. algorithm for the kernel density estimation

- Letting $\widehat{f}_t$ stand for the estimator constructed using the windows $\mathbf{W}_1, \ldots, \mathbf{W}_t$.
- The Robbins-Monro recursivity yields to consider

$$\widehat{f}_t(\mathbf{x}) = \widehat{f}_{t-1}(\mathbf{x}) + \gamma_t \left(\tilde{f}_t(\mathbf{x}) - \widehat{f}_{t-1}(\mathbf{x}) \right),$$

where $\tilde{f}_t(\mathbf{x})$ is an appropriate estimator of $f(\mathbf{x})$ computed from $\mathbf{W}_t$ only.

Robbins and Monro (AOS 1951).

- Hall et al. (1987), kernel-based estimator :

$$\tilde{f}_t(\mathbf{x}) = \frac{c_0(h_t)}{N_t} \sum_{j=1}^{N_t} K_{h_t^2}\left(\mathbf{x}, \mathbf{X}_{tj}\right),$$

where

$$K_h(\mathbf{u}, \mathbf{v}) := K\left(\frac{1 - \mathbf{u}'\mathbf{v}}{h}\right),$$

with K is the directional kernel, $h_t > 0$ a sequence of bandwidth parameters and $c_0(.)$ a normalizing constant defined by

$$c_0(h)^{-1} = \int_{\mathcal{S}_{p-1}} K_{h^2}(\mathbf{x}, \mathbf{y}) \omega_p(d\mathbf{y}).$$

- The choice of step sizes :

$$\gamma_t := N_t \Big/ \sum_{s=1}^{t} N_s, \quad t = 1, \ldots, n$$

leads to the centroid form of the estimator of f based on all the information available at the time n

$$\widehat{f}_n(\mathbf{x}) = \frac{\widehat{f}_{n-1}(\mathbf{x}) \sum_{s=1}^{n-1} N_s + N_n \tilde{f}_n(\mathbf{x})}{\sum_{s=1}^{n} N_s}.$$

- As a direct consequence, we get :

$$\widehat{f}_n(\mathbf{x}) = \frac{1}{\sum_{s=1}^{n} N_s} \sum_{s=1}^{n} c_0(h_s) \sum_{j=1}^{N_s} K_{h_s^2}(\mathbf{x}, \mathbf{X}_{sj}).$$

- When the $\mathbf{W}_t$'s have widths $N_t = 1$ and if a single bandwith parameter is used ($h_t = h$), for all $t = 1, \ldots, n$, we get

$$\bar{f}_n(\mathbf{x}) = \frac{c_0(h)}{n} \sum_{t=1}^{n} K_{h^2}(\mathbf{x}, \mathbf{W}_t),$$

which is nothing more than the kernel density estimator of Hall et al. (1987).

Bias-variance decomposition

Proposition

$$\mathbb{E}\left(\widehat{f}_n(\mathbf{x})\right) - f(\mathbf{x}) \quad = \quad h_n^2 \theta_2 \mu(K) \Psi(f, \mathbf{x}) + o(1),$$

and

$$Var\left(\widehat{f}_n(\mathbf{x})\right) \quad = \quad \frac{1}{n h_n^{p-1}} \frac{\theta_{1-p}}{r} R(K) f(\mathbf{x}) + o(1),$$

as $n \to \infty$

where : $N_n \to r$; $\quad \frac{1}{n}\sum_{s=1}^{n}\left(\frac{h_s}{h_n}\right)^q \to \theta_q$ as $n \to \infty$

$$\Psi(f,\mathbf{x}) = p^{-1}\left[\nabla^2 f(\mathbf{x}) - \mathbf{x}'\mathcal{H}(\mathbf{x})\mathbf{x}\right],$$

$$\mu(K) = \frac{\int_0^\infty v^{(p-1)/2}K(v)dv}{\left(\int_0^\infty v^{(p-3)/2}K(v)dv\right)}$$

and

$$R(K) = \frac{\int_0^\infty v^{(p-3)/2}K^2(v)dv}{2^{(p-3)/2}\omega_{p-1}\left(\int_0^\infty v^{(p-3)/2}K(v)dv\right)^2}.$$

Almost sure convergence

Proposition

$$\limsup_{n\to\infty} \sqrt{\frac{nh_n^{p-1}}{\ln\ln n}} \left(\widehat{f}_n(\mathbf{x}) - \mathbb{E}\widehat{f}_n(\mathbf{x})\right) = \sqrt{2\frac{\theta_{1-p}}{r} f(\mathbf{x}) R(K)} \quad a.s.$$

In particular, if $h_n = h_0 \left(\frac{\ln\ln n}{n}\right)^{1/(p+3)}, h_0 > 0,$ *then*

$$\limsup_{n\to\infty} \left(\frac{n}{\ln\ln n}\right)^{2/(p+3)} \left(\widehat{f}_n(\mathbf{x}) - f(\mathbf{x})\right) = \sqrt{2h_0^{1-p}\frac{\theta_{1-p}}{r} f(\mathbf{x}) R(K) + h_0^2 \theta_2 \mu(K) \Psi(f,\mathbf{x})} a.s.$$

Asymptotic normality

Proposition

If there exists $h_0 \geq 0$ *such that* $nh_n^{p+3} \to h_0$, *then*

$$\sqrt{nh_n^{p-1}}\left(\widehat{f}_n(\mathbf{x}) - f(\mathbf{x})\right) \xrightarrow{\mathcal{D}} \mathcal{N}\left(\sqrt{h_0}\theta_2\mu(K)\Psi(f,\mathbf{x}), \frac{\theta_{1-p}R(K)f(\mathbf{x})}{r}\right)$$

as $n \to \infty$.

Optimal bandwidth

- Minimizing the mean integrated square error with respect to h_n yields to the optimal choice of bandwidth

$$h_n = \left[\frac{\theta_{1-p}}{r\theta_2^2} \cdot \frac{(p-1)R(K)}{4\mu(K)^2 \int_{\mathcal{S}^{p-1}} \Psi^2(f, \mathbf{x}) \omega_p(d\mathbf{x}) n} \right]^{1/(p+3)},$$

for which we obtain that for which we obtain that

$$\theta_2 = \lim_{n\to\infty} \frac{1}{n^{1-2/(p+3)}} \sum_{s=1}^{n} s^{-2/(p+3)} = \frac{p+3}{p+1}$$

and $\theta_{1-p} = \dfrac{\theta_2}{2}$.

- For $r = 1$, we obtain that the Asymptotic Relative Error between $\bar{f}_n(\mathbf{x})$ and $\widehat{f}_n(\mathbf{x})$ is given by

$$\mathrm{ARE}(\bar{f}_n(\mathbf{x})/\widehat{f}_n(\mathbf{x})) = \theta_{1-p}{}^{\frac{4}{(p+3)}}\,\theta_2^{\frac{(2p-2)}{(p+3)}} = \left(\frac{1}{2}\right)^{\frac{4}{(p+3)}}\left(\frac{p+3}{p+1}\right)^{\frac{(2p+2)}{(p+3)}}.$$

- A plot of $\mathrm{ARE}(\bar{f}_n(\mathbf{x})/\widehat{f}_n(\mathbf{x}))$ for various values of $p \geq 2$ is provided in the next figure.

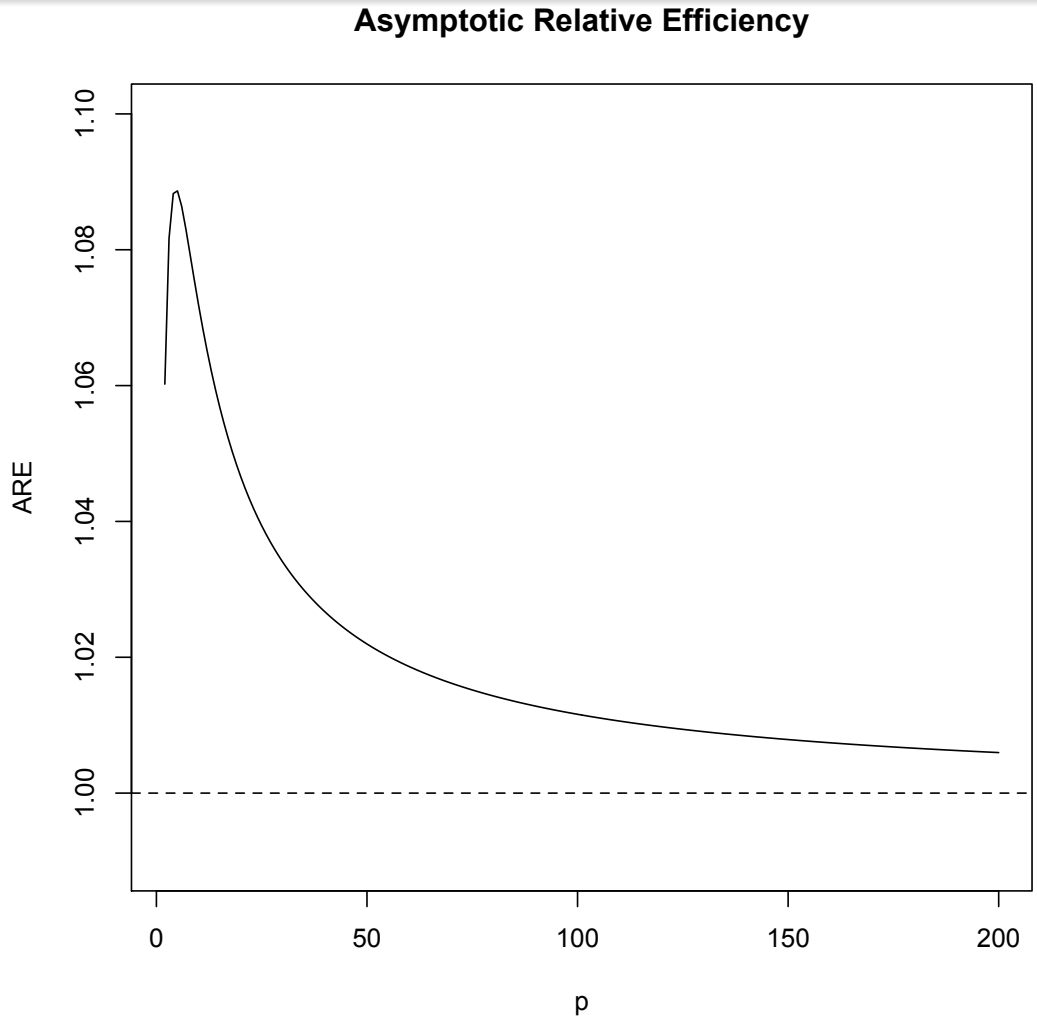


Figure: Plot of the ARE between $\bar{f}_n(\mathbf{x})$ and $\widehat{f}_n(\mathbf{x})$ $\left(\frac{\mathsf{AMISE}(\widehat{f}_n(\mathbf{x}))}{\mathsf{AMISE}(\bar{f}_n(\mathbf{x}))}\right)$ for various values of $p > 2$.

- Inspection of the Figure reveals that the non-recursive estimator is slightly more efficient than its recursive counterpart.
- The maximum value of $\mathrm{ARE}(\bar{f}_n(\mathbf{x})/\widehat{f}_n(\mathbf{x}))$ in (2.1) (viewed as a continuous function of p) is given by 1.08896 and is obtained for

$$p = (3 - e^{1-\log 2})/(e^{1-\log 2} - 1) \approx 4.56.$$

- The maximum loss of efficiency of the recursive estimator with respect to its non-recursive counterpart is quite small.
- It is also easy to show that

$$\lim_{p\to\infty} \mathrm{ARE}(\bar{f}_n(\mathbf{x})/\widehat{f}_n(\mathbf{x})) = 1$$

so that the loss of efficiency vanishes as the dimension p increases.

Regression estimation by local polynomial fitting for multivariate data streams

- $$\mathbf{W}_t := \{(\mathbf{X}_{t1}, Y_{t1}), \ldots, (\mathbf{X}_{tN_t}, Y_{tN_t})\}, \quad t = 1, \ldots, n,$$

  the sub-sample $(\mathbf{X}_{t1}, Y_{t1}), \ldots, (\mathbf{X}_{tN_t}, Y_{tN_t})$ is a sequence of random vectors identically distributed as a stationary stochastic process $(\mathbf{X}, Y)$ valued in $\mathbb{R}^d \times \mathbb{R}$ $(d \geq 1)$.
- We assume that the $(\mathbf{X}_{tj}, Y_{tj})$'s have a common joint density $f_{(\mathbf{X},Y)}(\cdot, \cdot)$.

- The goal is to provide a local polynomial estimator of the regression function

$$r(\mathbf{x}) = \mathbb{E}\left(Y|\mathbf{X} = \mathbf{x}\right)$$

and its derivatives in the data streams framework.

Ruppert and Wand (AOS 1994), Fan and Gijbels (Chapman & Hall 1996), Masry (SPA 1996), Vilar and Vilar (TEST 2000), Gu et al. (Econ. Rev. 2015),...

- Given $\mathbf{x} \in \mathbb{R}^d$ and $p \in \mathbb{N}$, assume that the regression function has derivatives of total order $p+1$ at $\mathbf{x}$.
- The multivariate Taylor formula provides an approximation of $r(\mathbf{X})$ by a multivariate polynomial of total order p as :

$$r(\mathbf{X}) \simeq r(\mathbf{x}) + \sum_{\left\{\mathbf{k}\in\mathbb{N}^d \,:\, 1\leq|\mathbf{k}|\leq p\right\}} \frac{1}{\mathbf{k}!}\frac{\partial^{|\mathbf{k}|}r}{\partial\mathbf{x}^{\mathbf{k}}}(\mathbf{x})\,(\mathbf{X}-\mathbf{x})^{\mathbf{k}} \quad (3.1)$$

where :

$$|\mathbf{k}| = \sum_{i=1}^{d} k_i; \quad \mathbf{k}! = \prod_{i=1}^{d} k_i! \quad \text{and } \mathbf{x}^{\mathbf{k}} = \prod_{i=1}^{d} x_i^{k_i}.$$

- Taking into account the observations available in $\mathbf{W}_1, \ldots, \mathbf{W}_n$, we can derive the locally weighted least squares estimators of the quantities

$$\beta_{\mathbf{k}} = \frac{1}{\mathbf{k}!} \frac{\partial^{|\mathbf{k}|} r}{\partial \mathbf{x}^{\mathbf{k}}}(\mathbf{x}), \quad 0 \leq |\mathbf{k}| \leq p$$

by minimizing the objective function

$$\sum_{t=1}^{n} \sum_{j=1}^{N_t} \left\{ Y_{tj} - \sum_{\left\{\mathbf{k} \in \mathbb{N}^d \,:\, 0 \leq |\mathbf{k}| \leq p\right\}} \beta_{\mathbf{k}} \left(\mathbf{X}_{tj} - \mathbf{x}\right)^{\mathbf{k}} \right\}^2 \omega_{tj}^{(n)}(\mathbf{x}), \tag{3.2}$$

where the weights $\omega_{tj}^{(n)}$ are defined by

$$\omega_{tj}^{(n)}(\mathbf{x}) = \frac{1}{N^{(n)} h_t^d} K\left(\frac{\mathbf{X}_{tj} - \mathbf{x}}{h_t}\right), \tag{3.3}$$

Notations

- For $u = 0, \ldots, p$, let

$$L_u = \left\{ \mathbf{k} \in \mathbb{N}^d, \ |\mathbf{k}| = u \right\} \text{ and } q = \sum_{u=0}^{p} \binom{u+d-1}{d-1}.$$

Define L as the set of q d-tuples obtained by rearranging the elements of the sets $L_0, \ldots, L_p$ with respect to the lexicographic order and concatenating them as a triangular array.

- For example, if $d = 2$:

$$L = \left\{ \begin{array}{ccccc} (0,0), & & & & \\ (0,1), & (1,0), & & & \\ (0,2), & (1,1), & (2,0), & & \\ (0,3), & (1,2), & (2,1), & (3,0), & \\ \vdots & & & & \\ (0,p), & (1,p-1), & (2,p-2), & (3,p-3), & \ldots,(p,0) \end{array} \right\}.$$

- Let g be a continuous bijective function such that :

$$g : \begin{array}{l} L \longrightarrow \{0, \ldots, q-1\} \\ \mathbf{k} \longmapsto i \end{array},$$

  where i denotes the index of the d-tuples $\mathbf{k}$ in the set L.

- We note that

$$g^{-1}(i) = [i] \text{ for any } i = 0, \ldots, q-1.$$

- According to the above notation, for any $t \in \{1, \ldots, n\}$, one can define the matrices

$$\beta = \begin{pmatrix} \beta_{[0]} \\ \vdots \\ \beta_{[q-1]} \end{pmatrix} \text{ and } \mathcal{X}_t = \begin{pmatrix} 1 & (\mathbf{X}_{t1} - \mathbf{x})^{[1]} & \ldots & (\mathbf{X}_{t1} - \mathbf{x})^{[q-1]} \\ 1 & (\mathbf{X}_{t2} - \mathbf{x})^{[1]} & \ldots & (\mathbf{X}_{t2} - \mathbf{x})^{[q-1]} \\ \vdots & & & \\ 1 & (\mathbf{X}_{tN_t} - \mathbf{x})^{[1]} & \ldots & (\mathbf{X}_{tN_t} - \mathbf{x})^{[q-1]} \end{pmatrix} := \begin{pmatrix} \mathcal{X}_{t1}^{\mathsf{T}} \\ \mathcal{X}_{t2}^{\mathsf{T}} \\ \vdots \\ \mathcal{X}_{tN_t}^{\mathsf{T}} \end{pmatrix}.$$

- Finally, set

$$\mathcal{Y}_t = \left(Y_{t1}, \ldots, Y_{tN_t}\right)^{\mathsf{T}} \text{ and } \Omega_t^{(n)} = \text{diag}\left(\omega_{t1}^{(n)}, \ldots, \omega_{tN_t}^{(n)}\right), \quad t = 1, \ldots, n.$$

- Then, the derivative of (3.2) with respect to β is simply the empirical counterpart of

$$2\sum_{t=1}^{n} N_t \mathbb{E}\left(\omega_{t_1}^{(n)} \mathcal{X}_{t1} \mathcal{X}_{t1}^{\mathsf{T}} \beta - Y_{t1} \omega_{t_1}^{(n)} \mathcal{X}_{t1}\right) =: 2F(\beta).$$

- The intention of the exercise is to solve $F(\beta) = 0$.

Local polynomial regression estimation

- Let $\widehat{\beta}_n = \left(\widehat{\beta}_{[0]}^{(n)}, \cdots, \widehat{\beta}_{[q-1]}^{(n)}\right)^{\mathsf{T}}$ be an estimator of β based on $\mathbf{W}_1, \ldots, \mathbf{W}_n$.
- The the multivariate Newton-Raphson procedure yields to consider

$$\widehat{\beta}_n = \widehat{\beta}_{n-1} - D_n^{-1} \widehat{F}_n \left(\widehat{\beta}_{n-1}\right), \tag{3.4}$$

where D_n is an estimate of the matrix $\dfrac{\partial F}{\partial \beta}(\beta)$ based on $\mathbf{W}_1, \ldots, \mathbf{W}_n$ and $\widehat{F}_n(\cdot)$ is an estimator of $F(\cdot)$ based on the sub-sample $\mathbf{W}_n$ only, that is, the observations received at "time" n.

Ruppert (AOS 1985).

- Observe that :

$$\frac{\partial F}{\partial \beta}(\beta) = \sum_{t=1}^{n} N_t \mathbb{E}\left(\omega_{t_1}^{(n)} \mathcal{X}_{t1} \mathcal{X}_{t1}^{\mathsf{T}}\right).$$

- Then the empirical counterparts of $\frac{\partial F}{\partial \beta}(\beta)$ and $F(\beta)$ are respectively defined by :

$$D_n = \sum_{t=1}^{n} \mathcal{X}_t^{\mathsf{T}} \Omega_t^{(n)} \mathcal{X}_t$$

and

$$\widehat{F}_n\left(\widehat{\beta}_{n-1}\right) = \mathcal{X}_n^{\mathsf{T}} \Omega_n^{(n)} \mathcal{X}_n \left[\widehat{\beta}_{n-1} - \left(\mathcal{X}_n^{\mathsf{T}} \Omega_n^{(n)} \mathcal{X}_n\right)^{-1} \mathcal{X}_n^{\mathsf{T}} \Omega_n^{(n)} \mathcal{Y}_n\right],$$

which together with (3.4) indicate that

$$\widehat{\beta}_n = (I_q - \Gamma_n)\widehat{\beta}_{n-1} + \Gamma_n \tilde{\beta}_n, \tag{3.5}$$

where

$$\Gamma_n = D_n^{-1} \mathcal{X}_n^{\mathsf{T}} \Omega_n^{(n)} \mathcal{X}_n, \quad \tilde{\beta}_n = \left(\mathcal{X}_n^{\mathsf{T}} \Omega_n^{(n)} \mathcal{X}_n \right)^{-1} \mathcal{X}_n^{\mathsf{T}} \Omega_n^{(n)} \mathcal{Y}_n$$

and I_q is the unit matrix of size q.

- The vector $\tilde{\beta}_n$ is simply the weighted least squares estimator of β based on the bach
  $\mathbf{W}_n := \left\{ (\mathbf{X}_{n1}, Y_{n1}), \ldots, (\mathbf{X}_{nN_n}, Y_{nN_n}) \right\}$.
- The expression (3.5) bears a resemblance in its structure to the exponential smoothing scheme, except for the fact that in (3.5), the smoothing parameter is a matrix.
- This relation can be understood as a multivariate Robbins-Monro recursivity, with a step size in a matrix form.

- Setting $V_t = N^{(n)}\Omega_t^{(n)}$, D_n can be reformulated as

$$D_n = \frac{1}{N^{(n)}} \sum_{t=1}^{n} \mathcal{X}_t^{\mathsf{T}} V_t \mathcal{X}_t.$$

- Using the relation

$$D_{n+1} = \left[1 - \frac{N_{n+1}}{N^{(n+1)}}\right] D_n + \frac{1}{N^{(n+1)}} \mathcal{X}_{n+1}^{\mathsf{T}} V_{n+1} \mathcal{X}_{n+1}$$

and the Woodbury matrix identity, we found that :

$$D_{n+1}^{-1} = \left(1 + \frac{N_{n+1}}{N(n)}\right)\left[D_n^{-1} - \frac{1}{N(n)} D_n^{-1} \mathcal{X}_{n+1}^{\mathsf{T}} V_{n+1}^{1/2} C^{-1} V_{n+1}^{1/2} \mathcal{X}_{n+1} D_n^{-1}\right],$$

where

$$C = I_{N_{n+1}} + \frac{1}{N(n)} V_{n+1}^{1/2} \mathcal{X}_{n+1} D_n^{-1} \mathcal{X}_{n+1}^{\mathsf{T}} V_{n+1}^{1/2}.$$

- Computational coast : $O\left(\sum_{s=1}^{n} N_s\right)$
- In the traditional case $N_t = 1$ for all t, our estimator $\widehat{\beta}_n$ is a sequential version of the local weighted estimator which is obtained by considering a sequence a of bandwidth parameters $h_t > 0$ rather than a single bandwidth parameter in the definition of the $\omega_{tj}^{(n)}$.

- Set :

$$H_n = \mathrm{diag}\left(1, h_n, h_n^2, \ldots, h_n^{q-1}\right),$$

$$\theta_j = \lim_{n\to\infty} \frac{1}{n} \sum_{t=1}^{n} \left(\frac{h_t}{h_n}\right)^j ; \mu_{\mathbf{i}} = \int_{\mathbb{R}^d} \mathbf{u}^{\mathbf{i}} K(\mathbf{u}) d\mathbf{u}; \ \gamma_{\mathbf{i}} = \int_{\mathbb{R}^d} \mathbf{u}^{\mathbf{i}} K^2(\mathbf{u}) d\mathbf{u};$$

- Arrange the elements of the set :

$$\mathcal{D}_{p+1} = \left\{ \frac{1}{\mathbf{k}!} \frac{\partial^{|\mathbf{k}|} r}{\partial \mathbf{x}^{\mathbf{k}}}(\mathbf{x}) : \quad |\mathbf{k}| = p+1 \right\},$$

  using the lexicographic order and refer to them as a column vector $b_{p+1}(\mathbf{x})$.

- Set $Q = \#\mathcal{D}_{p+1} + q + 1$ and define the matrix A whose (i,j)-th component is

$$a_{ij} = \theta_{[i-1]+[j-1]} \mu_{[i-1]+[j-1]} \text{ with } 1 \leq i \leq q \text{ and } q+1 \leq j \leq Q.$$

- Let B and V represent $q \times q$ matrices defined by the entries :

$$b_{ij} = \theta^2_{[i-1]+[j-1]} \mu^2_{[i-1]+[j-1]} \quad \text{and } v_{ij} = \theta_{[i-1]+[j-1]-d\gamma_{[i-1]+[j-1]}};$$

Under weak assumptions, we have

$$\sqrt{nh_n^d}\left[H_n\left(\widehat{\beta}_n-\beta\right)-h_n^{p+1}B^{-1}Ab_{p+1}(\mathbf{x})\right]\xrightarrow{\mathcal{D}}\mathcal{N}\left(\mathbf{0}_q,\Sigma\right),$$

where

$$\Sigma:=\frac{\sigma_Y^2(\mathbf{x})}{\kappa f_{\mathbf{X}}(\mathbf{x})}B^{-1}VB^{-1}.$$

Algorithms for numerical computations

1. Fix an integer $n_0 \geq 1$ (resp. $n > n_0$) as the starting (resp. the ending) time of the estimation procedure ;
2. Choose a tolerance level $\epsilon > 0$ and a kernel K ;
3. Initialization : $k \longrightarrow n_0$

   (a) observe the windows $W_1, \ldots, W_k$;

   (b). compute $N^{(k)}$, the total number of observations available at the time k.

   (c) compute the bandwidth h_k ;

   (d) for $t = 1, \ldots, k$:

      i. compute the sample size N_t of the sub-sample W_t ;

      ii. extract the design matrix $\mathcal{X}_t$ and the response vector $\mathcal{Y}_t$;

      iii. for $j = 1, \ldots N_t$:

         compute the weights $\omega_{tj}^{(k)}(x)$

      end for

      iv. define the diagonal matrix of weights $\Omega_t^{(k)} = \operatorname{diag}\left(\omega_{t1}^{(k)}(x), \ldots, \omega_{tN_t}^{(k)}(x)\right)$;

   end for ;

   (e) concatenate the matrices of weights in a quasi-diagonal matrix $\Omega_k^{\dagger}$;

   (f) define the initial design matrix and response vector : $\mathcal{X}_k^{\dagger}$ and $\mathcal{Y}_k^{\dagger}$

   (g) compute the matrices $T_k = {\mathcal{X}_k^{\dagger}}^{\mathsf{T}} \Omega_k^{\dagger} \mathcal{Y}_k^{\dagger}$, $D_k = {\mathcal{X}_k^{\dagger}}^{\mathsf{T}} \Omega_k^{\dagger} \mathcal{X}_k^{\dagger}$ and D_k^{-1};

   (h) compute the recursive estimator as $\widehat{\beta}_k = D_k^{-1} T_k$

   i. save N_k, h_k, D_k^{-1}, $\widehat{\beta}_k$.

4. | if $k+1 < n$:

   (a) | $k \longrightarrow k+1$

   i. | observe W_k
   ii. | compute the sample size N_k of the sub-sample W_k ;
   iii. | update the total number of observations : $N^{(k)} = N^{(k-1)} + N_k$.
   iv. | extract the current design matrix $\mathcal{X}_k$ and the response vector $\mathcal{Y}_k$;
   v. | update the bandwidth : $h_{k-1} \longrightarrow h_k$;
   vi. | $t \longrightarrow k$
   | repeat steps 3(d)i – 3(d)iv to obtain $V_k = N^{(k)}\Omega_k^{(k)}$;
   vii. | compute the matrix C_k, its inverse C_k^{-1} and the matrix $T_k = \frac{1}{N^{(k)}}\mathcal{X}_k^{\mathsf{T}} V_k \mathcal{X}_k$.
   viii. update D_k^{-1} :
   ix. compute the step-size matrix $\Gamma_k = D_k^{-1} T_k$.
   x. update the local linear estimator $\widehat{\beta}_k = (I_q - \Gamma_k)\widehat{\beta}_{k-1} + \mathcal{X}_k^{\mathsf{T}}\Omega_k^{(k)}\mathcal{Y}_k$
   xi. | save N_k, $N^{(k)}$, h_k, D_k^{-1}, $\widehat{\beta}_k$.

   (b) if $\|\widehat{\beta}_k - \widehat{\beta}_{k-1}\| > \epsilon$
   | repeat 4(a)
   end if

end if

Bandwidth selector

bandwidth

1. At step 3(c) of *the preview algorithm*, compute h_k using the subsample $\mathbf{W}_k$ (by cross validation or other method) ;
2. At step 4(a)v of *the preview algorithm*, update the bandwidth using a recursive estimator given by the convex combination

$$\widehat{h}_k = \left(1 - \frac{1}{k}\right) \widehat{h}_{k-1} + \frac{1}{k} \tilde{h}(\mathbf{W}_k),$$

where $\tilde{h}(\mathbf{W}_k)$ is the bandwidth selected based on the data available in the window $\mathbf{W}_k$;

end bandwidth

Estimation of a space-varying distribution

- Let $\mathcal{I}_n$ be a surface of cardinal n, which is a finite subset of a potentially observable region $\mathcal{D} \subset \mathbb{Z}^N$, where $\mathbb{Z}^N$ is endowed with the uniform metric.
- Assume that we observe a sequence of arrays

$$W_{(\mathbf{s},t)} := \left\{ X_{(\mathbf{s},t)1}, \ldots, X_{(\mathbf{s},t)k(\mathbf{s},t)} \right\},$$

  $(\mathbf{s}, t) \in \mathcal{I}_n \times \{1, \ldots, T\} := \mathcal{D}_{n,T}$, where the sub-sample $X_{(\mathbf{s},t)1}, \ldots, X_{(\mathbf{s},t)k(\mathbf{s},t)}$ is a sequence of $\mathbb{R}^d$-valued random vectors identically distributed with distribution $G_{(\mathbf{s},t)}$ and density $g_{(\mathbf{s},t)}$ with respect to Lebesgue measure.
- $k(\mathbf{s}, t)$ may be random.

- Here $g_{(\mathbf{s},t)}$ depends on an overall density function f with distribution function F such that

$$g_{(\mathbf{s},t)}(\cdot) = \alpha_{(\mathbf{s},t)}(\cdot) f(\cdot). \tag{4.6}$$

- The aim is to provide an estimator $\widehat{f}$ of the density f based on the independent samples $W_{(\mathbf{s},t)}, \quad (\mathbf{s},t) \in \mathcal{D}_{n,T}$.

- Let

$$\widehat{n} := \sum_{(\mathbf{u},v)\in\mathcal{D}_{n,T}} k(\mathbf{s},t)$$

be the overall sample size and

$$p(\mathbf{s},t) := k(\mathbf{s},t)/\widehat{n}$$

the proportion of the observations at the site $\mathbf{s}$ and time t relative to the overall sample.

- Define

$$\widehat{G}_{(\mathbf{s},t)}(x) := \frac{1}{k(\mathbf{s},t)} \sum_{j=1}^{k(\mathbf{s},t)} \mathbb{1}_{\{X_{(\mathbf{s},t)j}\leq x\}}$$

the local empirical distribution.

- From (4.6), one may write :

$$f(x) = \frac{\sum_{(\mathbf{s},t)\in\mathcal{D}_{n,T}} p(\mathbf{s},t) g_{(\mathbf{s},t)}(x)}{\sum_{(\mathbf{s},t)\in\mathcal{D}_{n,T}} p(\mathbf{s},t) \alpha_{(\mathbf{s},t)}(x)} \tag{4.7}$$

or

$$f(x) = \sum_{(\mathbf{s},t)\in\mathcal{D}_{n,T}} \frac{p(\mathbf{s},t) g_{(\mathbf{s},t)}(x)}{\alpha_{(\mathbf{s},t)}(x)} \tag{4.8}$$

since

$$\sum_{(\mathbf{s},t)\in\mathcal{D}_{n,T}} p(\mathbf{s},t) = 1.$$

- Let us focus to (4.7), and write :

$$F(u) = \int_{-\infty}^{u} f(x)dx = \int_{-\infty}^{u} \frac{\sum_{(\mathbf{s},t)\in\mathcal{D}_{n,T}} p(\mathbf{s},t) g_{(\mathbf{s},t)}(x)}{\sum_{(\mathbf{s},t)\in\mathcal{D}_{n,T}} p(\mathbf{s},t)\alpha_{(\mathbf{s},t)}(x)} dx$$

$$= \sum_{(\mathbf{s},t)\in\mathcal{D}_{n,T}} p(\mathbf{s},t)\mathbb{E}\left[\frac{\mathbb{1}_{\{X_{(\mathbf{s},t)1}\leq u\}}}{\sum_{(\mathbf{s},t)\in\mathcal{D}_{n,T}} p(\mathbf{s},t)\alpha_{(\mathbf{s},t)}\left(X_{(\mathbf{s},t)1}\right)}\right]. \quad (4.9)$$

- Replacing (4.9) by its empirical conterpart, we get

$$\widehat{F}(u) = \frac{1}{\widehat{n}} \sum_{(\mathbf{s},t)\in\mathcal{D}_{n,T}} \sum_{j=1}^{k(\mathbf{s},t)} \frac{\mathbb{1}_{\{X_{(\mathbf{s},t)j}\leq u\}}}{\sum_{(\mathbf{s},t)\in\mathcal{D}_{n,T}} p(\mathbf{s},t)\alpha_{(\mathbf{s},t)}\left(X_{(\mathbf{s},t)j}\right)}.$$

- The indicator function $\mathbb{1}_{\{u-X_{(\mathbf{s},t)j}\geq 0\}}$ may be modified by a smoothed asymptotically unbiased replacement

$$H\left(\frac{u-X_{(\mathbf{s},t)j}}{b}\right) \text{ such that } H\left(\frac{v}{b}\right) \to \mathbb{1}_{\{v\geq 0\}} \text{ as } b\to 0^{+}.$$

- So that if $b := b_{(\mathbf{s},t)j}$ (a single bandwidth is traditionally used in temporal case)

$$\widehat{F}(u) = \sum_{(\mathbf{s},t)\in\mathcal{D}_{n,T}} \sum_{j=1}^{k(\mathbf{s},t)} \frac{H\left(\dfrac{u - X_{(\mathbf{s},t)j}}{b_{(\mathbf{s},t)j}}\right)}{\displaystyle\sum_{(\mathbf{s},t)\in\mathcal{D}_{n,T}} k(\mathbf{s},t)\alpha_{(\mathbf{s},t)}\left(X_{(\mathbf{s},t)j}\right)}.$$

- Take for instance,

$$H(v) = \int_{-\infty}^{v} K(x)dx \text{ with } \int_{-\infty}^{+\infty} K(x)dx = 1.$$

- Consequently :

$$\widehat{f}(u) = C^{-1} \sum_{(\mathbf{s},t)\in\mathcal{D}_{n,T}} \sum_{j=1}^{k(\mathbf{s},t)} \frac{b_{(\mathbf{s},t)j}^{-d} K\left(\dfrac{u - X_{(\mathbf{s},t)j}}{b_{(\mathbf{s},t)j}}\right)}{\displaystyle\sum_{(\mathbf{s},t)\in\mathcal{D}_{n,T}} k(\mathbf{s},t)\alpha_{(\mathbf{s},t)}\left(X_{(\mathbf{s},t)j}\right)} \tag{4.10}$$

where

$$C = \sum_{(\mathbf{s},t)\in\mathcal{D}_{n,T}} \sum_{j=1}^{k(\mathbf{s},t)} \left[\sum_{(\mathbf{s},t)\in\mathcal{D}_{n,T}} k(\mathbf{s},t)\alpha_{(\mathbf{s},t)}\left(X_{(\mathbf{s},t)j}\right)\right]^{-1}.$$

- Following the same idea as in Amiri (2012), the quantity $b_{(\mathbf{s},t)j}^{-d}$ in (4.10) can be substituted by

$$b_{\widehat{n}}^{d(\ell-1)} b_{(\mathbf{s},t)j}^{-d\ell}, \text{ with } \ell \in [0,1], \quad b_{\widehat{n}} > 0.$$

- The parameter ℓ plays a role of regulation in quality improvement of the estimator regarding the variance and the estimation errors.

Amiri (JNPS, 2012)

- In this case,

$$\widehat{f}(u) = C_\ell^{-1} \sum_{(\mathbf{s},t)\in\mathcal{D}_{n,T}} \sum_{j=1}^{k(\mathbf{s},t)} \frac{b_{(\mathbf{s},t)j}^{-d\ell} K\left(\dfrac{u - X_{(\mathbf{s},t)j}}{b_{(\mathbf{s},t)j}}\right)}{\displaystyle\sum_{(\mathbf{s},t)\in\mathcal{D}_{n,T}} k(\mathbf{s},t)\alpha_{(\mathbf{s},t)}\left(X_{(\mathbf{s},t)j}\right)} \tag{4.11}$$

  and the normalization constant is :

$$C_\ell = \sum_{(\mathbf{s},t)\in\mathcal{D}_{n,T}} \sum_{j=1}^{k(\mathbf{s},t)} \left[b_{(\mathbf{s},t)j}^{d(1-\ell)} \sum_{(\mathbf{s},t)\in\mathcal{D}_{n,T}} k(\mathbf{s},t)\alpha_{(\mathbf{s},t)}\left(X_{(\mathbf{s},t)j}\right) \right]^{-1}$$

- Eq. (4.10) corresponds to the case $\ell = 1$.

- To simplify, the notations, let us consider the simple case $\alpha_{(\mathbf{s},t)}(\cdot) = \alpha_0$: case of data stream without concept drift.
- For $\ell \in [0,1]$, set

$$\sigma_{\mathbf{s},t}^{[\ell]} := \sum_{j=1}^{k(\mathbf{s},t)} b_{(\mathbf{s},t)j}^{d(1-\ell)}; \quad \sigma_{\cdot t}^{[\ell,n]} := \sum_{\mathbf{s}\in\mathcal{I}_n} \sigma_{\mathbf{s},t}^{[\ell]};$$

$$\sigma_{\mathbf{s}\cdot}^{[\ell,T]} := \sum_{t=1}^{T} \sigma_{\mathbf{s},t}^{[\ell]}; \quad \sigma_{\cdot\cdot}^{[\ell,n,T]} := \sum_{(\mathbf{s},t)\in\mathcal{D}_{n,T}} \sigma_{\mathbf{s},t}^{[\ell]}.$$

Then from (4.11), the estimator of f is of the form :

$$\widehat{f}_{n,T}^{[\ell]}(x) = \frac{1}{\sigma_{..}^{[\ell,n,T]}} \sum_{(\mathbf{s},t)\in\mathcal{D}_{n,T}} \sigma_{\mathbf{s},t}^{[\ell]} \widetilde{f}_{(\mathbf{s},t)}^{[\ell]}(x), \tag{4.12}$$

where

$$\widetilde{f}_{(\mathbf{s},t)}^{[\ell]}(x) = \frac{1}{\sigma_{\mathbf{s},t}^{[\ell]}} \sum_{j=1}^{k(\mathbf{s},t)} \frac{1}{b_{(\mathbf{s},t)j}^{d\ell}} K\left(\frac{x - X_{(\mathbf{s},t)j}}{b_{(\mathbf{s},t)j}}\right) \tag{4.13}$$

- Easy computations show that from (4.12), the overall density estimator at the time T can be computed recursively via a Robbins-Monro stochastic algorithm as

$$f_{n,T}^{[\ell]}(x) = \gamma_T f_{n,T-1}^{[\ell]}(x) + (1 - \gamma_T)\, f_{\cdot T}^{\star}, \tag{4.14}$$

where

$$\gamma_T = \frac{\sigma_{\cdot\cdot}^{[\ell,n,T-1]}}{\sigma_{\cdot\cdot}^{[\ell,n,T]}}$$

is the step-size and

$$f_{\cdot T}^{\star} = \frac{1}{\sigma_{\cdot T}^{[\ell,n]}} \sum_{\mathbf{s}\in\mathcal{I}_n} \sigma_{\mathbf{s},T}^{[\ell]} \widetilde{f}_{(\mathbf{s},T)}^{[\ell]}(x)$$

is the density estimator based on the observations recorded over the spatial domain $\mathcal{I}_n$ at time T.

- Therefore, (4.14) indicates how to update the estimator from its immediate past when n new observations are recorded over the spatial domain $\mathcal{I}_n$.

Increasing domain and infill asymptotics

- The growth of the sample in increasing domain asymptotics is a consequence of an unbounded expansion of the sample region $\mathcal{I}_n$.
- Under infill asymptotics the sample region is fixed and the growth of the sample size is due to sampling that is dense in the region $\mathcal{D}$.
- Here, we consider the increasing domain asymptotics and for simplicity the bivariate regular lattice ($N = 2$), described as : $\mathcal{D}$ is a regular lattice and

$$\mathcal{I}_n = \{\mathbf{s} = (s_1, s_2),\ 1 \leq s_j \leq n_j, j = 1, 2\}$$

is rectangular.

Cressie (Wiley 1993).

- More precisely, we have from top to bottom and right to left :

$$\mathcal{I}_n = \left\{ \begin{array}{ccc} (1,1) & (2,1) & (n_1,1), \\ (1,2) & (2,2) & (n_1,2) \\ \vdots & \vdots & \vdots \\ (1,n_2), & (2,n_2) & \ldots (n_1,n_2) \end{array} \right\}.$$

- For simplicity, renumber (using a lexicographic order) the observations $\{X_{(\mathbf{s},t)j},\ \mathbf{s} \in \mathcal{I}_n\}$ as a triangular array $\{X_{(k,t,n)j},\ k = 1, ..., n\}$.
- In this case, each site $\mathbf{s} = (s_1, s_2) \in \mathcal{I}_n$ is identified by an indice $k = n_2(i-1) + j$ in the triangular array setting.

- Equation (4.14) allows us to update the estimation whenever a new additional observation site appears.

$$f_{n,T}^{[\ell]}(x) = \underbrace{\frac{\sigma_{\cdot\cdot}^{[\ell,n-1,T-1]}}{\sigma_{\cdot\cdot}^{[\ell,n,T]}}}_{\gamma_{n,T}^{(1)}} f_{n-1,T-1}^{[\ell]}(x) + \underbrace{\frac{\sigma_{\cdot T-1}^{[\ell,n]}}{\sigma_{\cdot\cdot}^{[\ell,n,T]}}}_{\gamma_{n,T}^{(2)}} f_{\cdot T-1}^{\star}(x)$$
$$+ \underbrace{\frac{\sigma_{\mathbf{s}_n\cdot}^{[\ell,T]}}{\sigma_{\cdot\cdot}^{[\ell,n,T]}}}_{\gamma_{n,T}^{(3)}} f_{\mathbf{s}_n\cdot}^{\star}(x) + \underbrace{\frac{\sigma_{\mathbf{s}_n,T}^{[\ell]}}{\sigma_{\cdot\cdot}^{[\ell,n,T]}}}_{\gamma_{n,T}^{(4)}} \widetilde{f}_{(\mathbf{s}_n,T)}^{[\ell]}(x),$$

where $f_{\mathbf{s}_n\cdot}^{\star}(x) = \dfrac{1}{\sigma_{\mathbf{s}_n\cdot}^{[\ell,T]}} \displaystyle\sum_{t=1}^{T-1} \sigma_{\mathbf{s}_n,t}^{[\ell]} \widetilde{f}_{(\mathbf{s}_n,t)}^{[\ell]}(x).$

- Observe that $\displaystyle\sum_{k=1}^{4} \gamma_{n,T}^{(k)} = 1.$

Particular cases

- $h_{(\mathbf{s},t)j} = h_{(\mathbf{s},t)}$, for any $1 \leq j \leq k(\mathbf{s},t)$ (choice of the same value of smooth parameter for each window) :

$$\widetilde{f}^{[\ell]}_{(\mathbf{s},t)}(x) = \frac{1}{k(\mathbf{s},t) b^d_{(\mathbf{s},t)}} \sum_{j=1}^{k(\mathbf{s},t)} K\left(\frac{x - X_{(\mathbf{s},t)j}}{b_{(\mathbf{s},t)}}\right) \qquad (4.15)$$

and

$$\sigma^{[\ell]}_{\mathbf{s},t} = k(\mathbf{s},t) b^{d(1-\ell)}_{(\mathbf{s},t)}$$

- In particular, if $k(\mathbf{s},t)=1$ for all $(\mathbf{s},t)\in\mathcal{D}_{n,T}$, then we get :

$$f_{n,T}^{[\ell]}(x):=\frac{1}{\sum\limits_{(\mathbf{s},t)\in\mathcal{D}_{n,T}} b_{(\mathbf{s},t)}^{d(1-\ell)}}\sum_{(\mathbf{s},t)\in\mathcal{D}_{n,T}}\frac{1}{b_{(\mathbf{s},t)}^{d\ell}}K\left(\frac{x-X_{(\mathbf{s},t)}}{b_{(\mathbf{s},t)}}\right).$$

- Furthermore, in the special case when $b_{(\mathbf{s},t)}=b_{\widehat{n}}$

$$f_{n,T}^{[\ell]}(x):=\frac{1}{\widehat{n}b_{\widehat{n}}}\sum_{(\mathbf{s},t)\in\mathcal{D}_{n,T}}K\left(\frac{x-X_{(\mathbf{s},t)}}{b_{\widehat{n}}}\right)$$

is simply the classic Parzen- Rosenblatt kernel density estimator in the spatio-temporal framework.

Wang and Wang (JNPS, 2009), Wang et al. (JNPS, 2012).

Simulations studies

- The von Mises kernel $K(t) = e^{-t}$ is considered in the implementation of both algorithms.
- We estimated the bandwidths $h_1, \ldots, h_n$ by using the rule of thumb described previously and a recursive version of the cross validation method based on the squared-error loss.

$$\widehat{h}_t = (1 - \gamma_t)\widehat{h}_{t-1} + \gamma_t \widehat{h}_{\mathrm{CV}}(\mathbf{W}_t),$$

where $\widehat{h}_{\mathrm{CV}}(\mathbf{W}_t)$ is the bandwidth selected by a cross-validation based on the data available in the window $\mathbf{W}_t$.

$$\widehat{h}_{\mathrm{CV}}(\mathbf{W}_t) = \arg\min_h 2N_t^{-1} \sum_{j=1}^{N_t} \tilde{f}_{tj}(\mathbf{X}_{tj}) - \int_{\mathcal{S}_{p-1}} \tilde{f}_{tj}^2(\mathbf{x}) \omega_p(d\mathbf{x}),$$

where

$$\tilde{f}_{tj}(\mathbf{x}) = \frac{c_0(h_t)}{N_t - 1} \sum_{i \neq j}^{N_t} K_{h_t^2}(\mathbf{x}, \mathbf{X}_{ti})$$

is the density estimate constructed over the observations contained in $\mathbf{W}_t$ leaving out the sample value $\mathbf{X}_{tj}$.

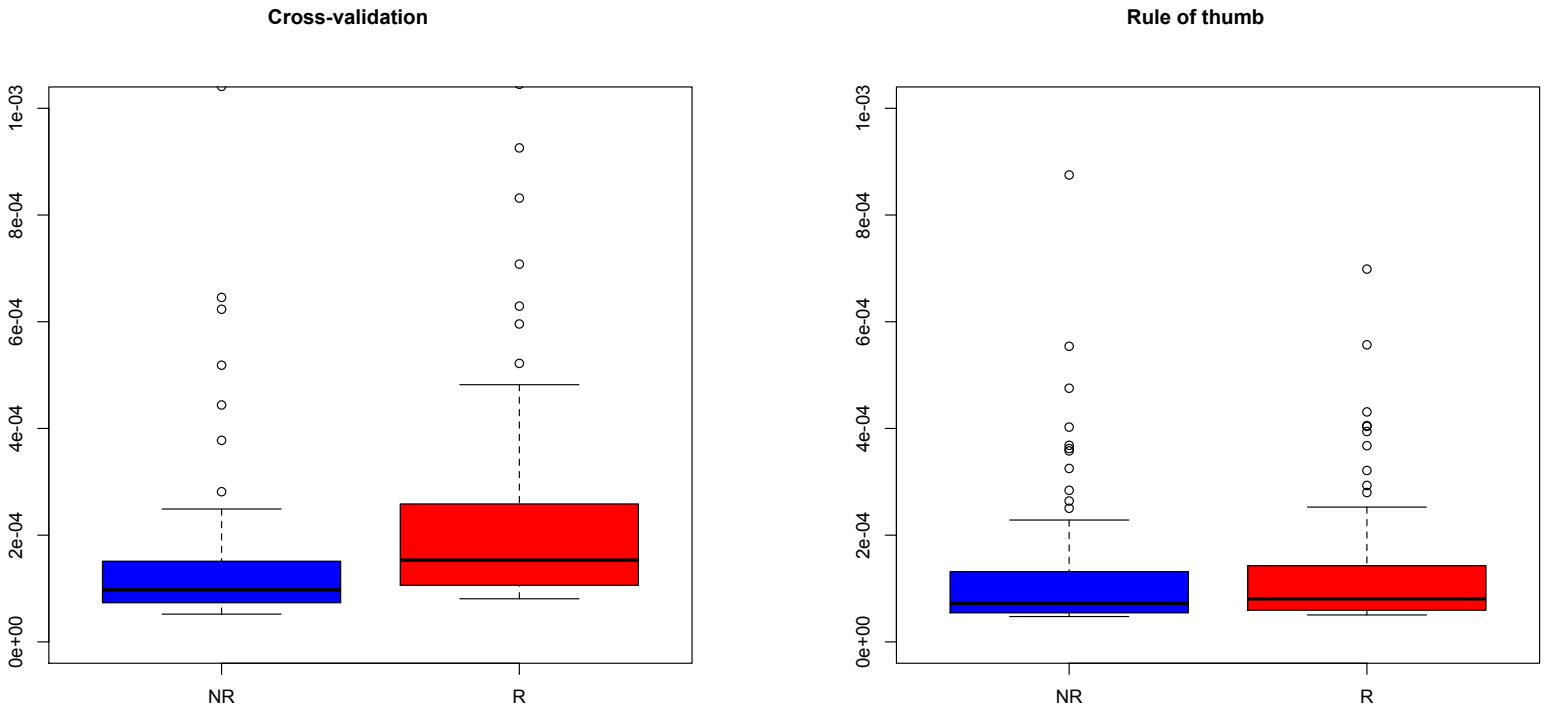


Figure: Boxplots of the average (over the M replications) mean square errors (computed at the various times $t = 1, \ldots, 100$) of the recursive (red boxes) and the non recursive (blue boxes) density estimators computed at $\mathbf{x} = (0, 1)$ using the cross-validation bandwidth selection (on the left) and the rule of thumb bandwidth selection (on the right) with observations distributed as bivariate Fisher-von Mises vectors (sampling scheme (i)) .

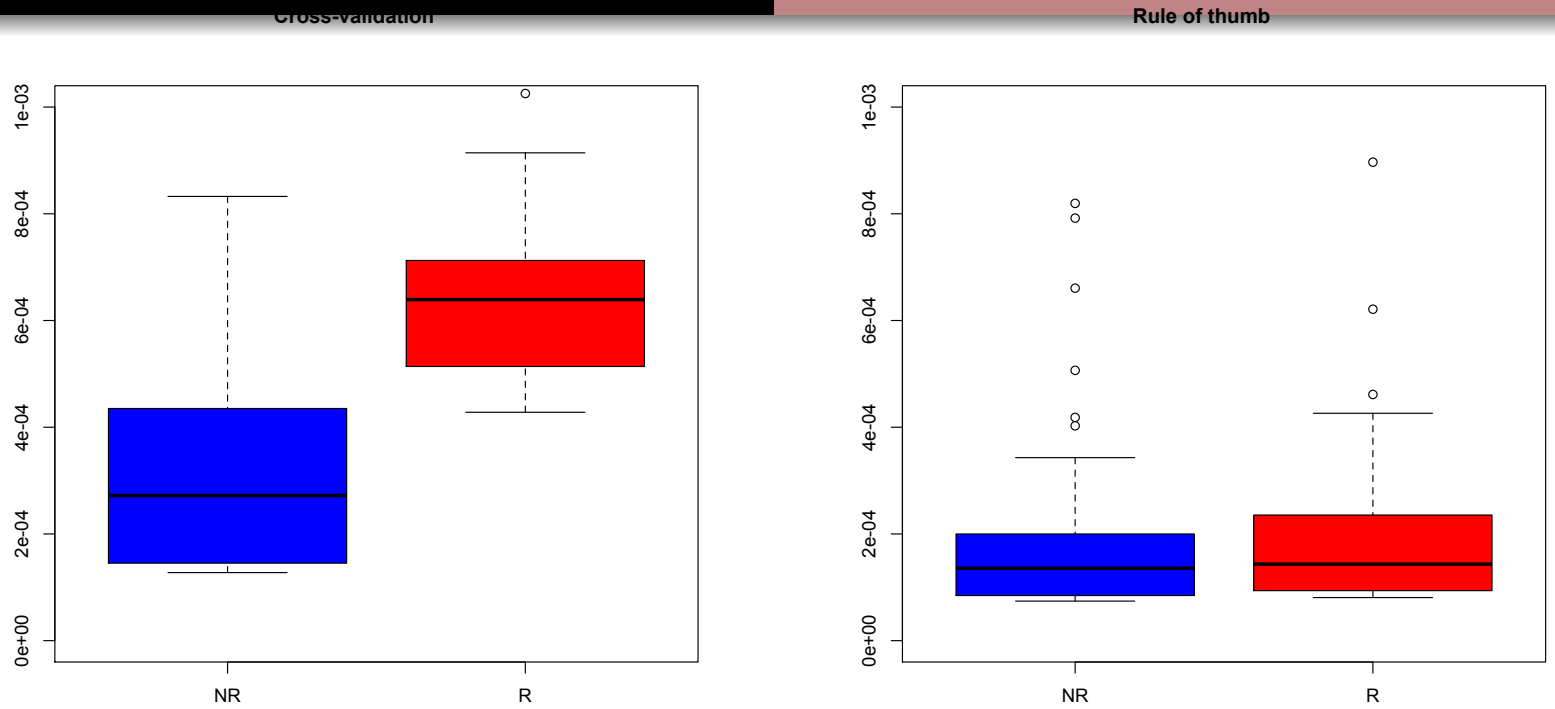


Figure: Boxplots of the average (over the M replications) mean square errors (computed at the various times $t = 1, \ldots, 100$) of the recursive (red boxes) and the non recursive (blue boxes) density estimators computed at $\mathbf{x} = (0, 0, 1)$ using the cross-validation bandwidth selection (on the left) and the rule of thumb (on the right) with observations distributed as ($p =$)3-dimensional Fisher-von Mises vectors (sampling scheme (ii)) .

A simulation study

- We generated $M = 100$ data streams from the model

$$Y_i = \sin(0.5\pi X_i) + \varepsilon_i, \quad \varepsilon_i = 0.5\varepsilon_{i-1} + \eta_i,$$

$$X_i \sim \mathcal{U}_{[-3,3]}, \quad \eta_i \sim \mathcal{B}(0.5), \quad i = 1, \ldots, 7200$$

- Once any simulated database is created and saved on disk, we opened a connection to the file where it was written and treated it as a stream [§] such that $N_t = 24$ and $n = 300$.
- Goal : estimate $\left(\beta_{[0]}, \beta_{[1]}\right)$.

§. https ://cran.r-project.org/web/packages/stream/stream.pdf

- Then, at any instant t, the row data $\mathbf{W}_t$ is replayed back.
- In this context, in which there is a time-varying sample size, our estimation procedure is started at the instant $t_0 = 1$ and the density estimator has been continuously updated with respect to the time until they reach the final instant n.
- Also, in order to avoid unnecessary calculations, the algorithm has been stopped if the absolute distance between two values of the density estimator obtained in two successive steps is less than 10^{-5}.

- Denoting by $\widehat{\beta}_t(x_i)^{[m]}$ the value of an estimator of $\beta(x_i)$ computed (from the mth replication of the Monte-Carlo procedure) at the point $x_i, i = 1, \ldots 100$ receiving the sample $\mathbf{W}_t$, $t \in \{1, \ldots, n\}$, the efficiency (at the step t) of the estimators is evaluated using the average mean square error

$$\mathrm{MSE}\left(\widehat{\beta}_t\right) = \frac{1}{100M} \sum_{m=1}^{M} \sum_{i=1}^{100} \left\| \beta(x_i) - \widehat{\beta}_t(x_i)^{[m]} \right\|^2, \quad (5.16)$$

- We provide the sequence $\mathrm{MSE}\left(\widehat{\beta}_t\right), t = 1, \ldots, 300.$

t	10	50	100	200	300
Recursive estimator	0.9238	0.2525	0.2403	0.2308	0.2301
Non recursive estimator	0.1619	0.1561	0.1537	0.1516	0.1438

Table: MSE for comparing recursive and non recursive estimators

- We also added the computational time elapsed for the estimation of the density at one point using the usual kernel estimator and its recursive version proposed here.

t	10	50	100	200	300
Recursive estimator	0.0012	0.0012	0.0012	0.0011	0.0012
Non recursive estimator	0.0241	0.236	0.9582	3.1297	7.5868

Table: Computational time in seconds for comparing recursive and non recursive estimators

Conclusion

We proposed algorithms for nonparametric estimation such that :

- the estimation procedure can only store a very limited amount of data to summarize the data stream.
- the incoming data points cannot be permanently stored
- the estimation procedure can process data points as fast as the data is arriving.
- the estimation procedure is able to deal with a data generating process which evolves over space and time (e.g., distributions change or new structure in the data appears.

Extensions

- Asymptotic results with bandwidths obtained by stochastic approximation ;
- Data stream in continuous time ;
- Other choices of the step-size matrix Γ_t
- Data stream with random batch size N_t ;
- Inference with non-stationary data streams (Concept drift) :
- ...