24.10.2022. Part 2. **Nonparametric Estimators**

- Sketch of the lecture:

  Introduction to Motivating examples.

  Chapter 1. Density Estimators. (univariate ← $\mathbb{R}^1$) Ichi ∞! X .

  Chapter 2. Regression. (hồi quy).

  Chapter 3. Multivariate and Functional Cases. ($\mathbb{R}^k$)

Introduction.

Motivating examples:

In this lecture we assume that we observe for each individual either:

- A variable X (density)
- A pair (X, Y) where
  - Y response variable
  - X exploratory

) Regression.

We assume we get i.i.d random samples

1) $(X_1, \dots, X_n) \overset{i.i.d}{\sim} X$

or 2) $((X_1, Y_1), \dots, (X_n, Y_n)) \overset{i.i.d}{\sim} (X, Y)$

(A) Galaxies dataset.

library(MASS) | data(galaxies). | hist(galaxies, freq = F, nclass = 100)

vẽ theo tần suất — số ô chia

Ex. 1) get the data.

2) How to represent the distribution of these data.

3) Pros and cons.

ưu đ? Nhược đ?

Pros	Cons.
. non parametric.	. stepwise | not continuous constant on classes.
hàm mật độ xs ko phụ thuộc vào tham số. → no parametric density assumed.	. Strongly depending on the number and the nature of the classes.

Aim: Define a new estimator being:

- smooth.
- still non parametric.
- depending on less estimators.

↳ Discuss the non parametric estimators (ở biết phân phối gì).

. Bắt đầu làm với non-parametric → nhìn dc mô hình parametric → đưa về parametric estimators.

	Pros	Cons.
p	. Great speed of convergence . Easy to interprate.	. No flexibility. . A priori needed.
np	. Flexibility. . No a priori	. Low convergence rate. . more difficult to interprate. (giải thích)

(8) Pressure data.

Ex. 1) Get data. (from ...)

2) Draw the scatter plot of pressure w.r.t temperature.

3) How to estimate the link function.

. data (pressure)

. attach (pressure)

$\begin{cases} p = \text{pressure} \ \$\ \text{pressure} \\ t = \_\_\_\_\_ \ \$\ \text{temperature} \end{cases}$

. plot (t, p) ⟶ vẽ biểu đồ giữa t và p.

↳ non linear.

. We are focusing on a regression model.

non parametric: $Y = r(X) + \varepsilon$, with $\mathbb{E}[\varepsilon | X] = 0$.

- Y: response variable.
- r(X): explanatory variable.
- ε: error

aim: estimate function r.

. Remark.

$$\mathbb{E}[Y|X] = \mathbb{E}[r(X) + \varepsilon | X]$$
$$= \mathbb{E}[r(X) | X] + \underbrace{\mathbb{E}[\varepsilon|X]}_{0}$$
$$= r(X).$$

where r is an unknown function we want to estimate.

Aim: Get a new estimator

- non parametric (no a priori on the nature)
- smooth.
- very flexible.

(C) Doping data.

1) get the data

add file via R : Session → Set working Directory → Choose Directory

2) If a new cyclist comes and we observe his hematocrit rate x_0, could we give the class y_0 corresponding to that cyclist?

- **Discrimination problem.** Y: class $\{+,-\}$.

  X: explanatory

  Aim: Get a classifier (plan bai) constructed from the data that will propose a relevant class for any new value of X.

- **Bayes classifier.**

  unknown → estimate

$$\hat{y}_0 = \underset{y \in \mathcal{Y}}{\operatorname{argmax}} \; \widehat{\mathbb{P}(Y=y \mid X=x_0)}.$$

↳ déf optimize ssi $\mathcal{Y}$ is finite.

Problem: $y \mapsto \underbrace{\mathbb{P}(Y=y \mid X=x_0)}_{r_y(x_0)}$ are unknown.

How to estimate r_y?

$$Z_y = \mathbb{1}_{\{Y=y\}}$$

$$\mathbb{E}[Z_y \mid X] = \mathbb{E}[\mathbb{1}_{\{Y=y\}} \mid X] = \mathbb{P}(Y=y \mid X) = r_y(x)$$

Note: $Y = r(x) + \varepsilon$

$r(x) = \mathbb{E}[Y \mid X]$

↳ $Z_y = r_y(x) + \varepsilon_y$.

- **Estimated Bayes classifier.**

$$\hat{y}_0 = \underset{y \in \mathcal{Y}}{\operatorname{argmax}}\ \hat{r}_y(x_0),$$

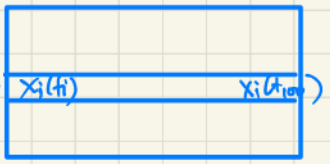
where $\hat{r}_y$ is a non parametric estimator of r_y.

① **Tecator data.**

- tec = read.table(' .dat ')

X = tec [, 1:100]

Y = tec [, 101]



- install.packages(' ggplot2 ')

library(' ggplot2 ')

wl = seq(750, 948, 2)

plot(wl, X[1,], type = 'l', ylim = range(x), xlab = 'wavelength', ylab = ' absorbance ')

(ylim: graphics function $\mathcal{O}_y$; type 'l': line)

```
for (i in 2:213)
  {
    lines (wl, X[i, ])
  }
```

. Regression on functional variable: $Y = r(X) + \varepsilon$.

Y: real valued

X: functional variable.

(E) Phoneme dataset

unique (PHONEME)

CURVES

PHONEME.

. par (mfrow = c(2,3)) ← vẽ CURVES theo cột.

(plot (CURVES [1, _] , type = 'l' , main = 'SH' , ylim = range (CURVES))

(for (i in 2: 250) { lines (CURVES [i,]) } .

↳ (_____ [201, _____ = 'IY' , _____

_____ 202 : 400 _____

(_____ [401 _____ = 'DCL' , _____

_____ 402 - 600 _____

Dismination problem with X functional.

Chapter 1. Density estimation.

$X_1, X_2, \dots, X_n \overset{i.i.d}{\sim} X$ Aim f_X

A) History.

- Parametric model

Assume $f_X = f_\theta$, $\theta \in \Theta \subseteq \mathbb{R}^k$.

$\hat{f}_X = f_{\hat{\theta}}$

- Non parametric model:

1) histogram: thông qua biểu đồ.
properties discussed in the introduction part.

2) moving window estimator.

2a) non parametric estimation of the cumulative distribution function:

$$F_X(x) = \mathbb{P}(X \le x)$$

for all $x \in \mathbb{R}$, $\hat{F}_X(x) = \frac{1}{n} \cdot \sum_{i=1}^{n} \mathbb{1}_{\{X_i \le x\}}$

- Remark. $\mathbb{E}\left[\mathbb{1}_{\{X \le x\}}\right] = \mathbb{P}(X \le x) = F_X(x)$.

↳ unbiased estimator for F_X.

- Properties

1) for all $x \in \mathbb{R}$, $\hat{F}_X(x)$ is an unbiased estimator of $F_X(x)$.

2) ——— , $\operatorname{var}\left(\hat{F}_X(x)\right) = \dfrac{F_X(x)(1-F_X(x))}{n}$ ← going to 0

3) ——— , $\hat{F}_X(x) \xrightarrow[n\to+\infty]{L^2} F_X(x)$

4) ——— , $\hat{F}_X(x) \xrightarrow[n\to+\infty]{a.s.} F_X(x)$ (LLN)

5) ——— , $\sqrt{n}\left(\hat{F}_X(x) - F_X(x)\right) \xrightarrow[n\to+\infty]{\mathcal{L}} \mathcal{N}\left(0, \sqrt{F_X(x)(1-F_X(x))}\right)$

2b) Moving window estimator.

• Remark. $f_X(x) = \lim_{h\to 0^+} \dfrac{F_X(x+h) - F_X(x-h)}{2h}$

Proof.

$$\frac{F_X(x+h) - F_X(x-h)}{2h} = \frac{F_X(x+h) - F_X(x) + F_X(x) - F_X(x-h)}{2h}$$

$$= \frac{F_X(x+h) - F_X(x)}{2h} + \frac{F_X(x) - F_X(x-h)}{2h}$$

$$= \frac{1}{2}\left[\underbrace{\frac{F_X(x+h) - F_X(x)}{h}} + \frac{F_X(x) - F_X(x-h)}{h}\right]$$

$$\xrightarrow{h\to 0^+} \frac{1}{2}\cdot\left[f_X(x) + f_X(x)\right] = f_X(x).$$

Hence, for h small

$$f_X(x) \simeq \frac{F_X(x+h) - F_X(x-h)}{2h}$$

Hence a natural estimator could be

$$\tilde{f}_X(x) = \frac{\hat{F}_X(x+h) - \hat{F}_X(x-h)}{2h}$$

$$\hookrightarrow \tilde{f}_X(x) = \frac{\frac{1}{n}\sum_{i=1}^{n}\left[\mathbb{1}\{X_i \le x+h\} - \mathbb{1}\{X_i \le x-h\}\right]}{2h}$$

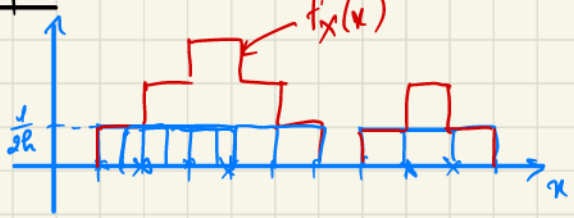
$$= \frac{\frac{1}{n}\sum_{i=1}^{n}\mathbb{1}\{x-h < X_i \le x+h\}}{2h}$$

$$= \frac{1}{nh}\sum_{i=1}^{n}\frac{1}{2}\cdot\mathbb{1}_{]-1,1]}\left(\frac{X_i - x}{h}\right)$$

$$= \frac{1}{nh}\cdot\sum_{i=1}^{n} w\left(\frac{X_i - x}{h}\right),$$

where $w(t) = \frac{1}{2}\cdot\mathbb{1}_{]-1,1]}(t)$: kernel = 'r'.

• Example.



$h = 1$.

Pros	Cons
depends on less parameter	- still stepwise (non smooth and constant on each interval)

• Remark.

$$\tilde{f}_X(x) = \frac{1}{nh}\sum_{i=1}^{n}\frac{1}{2}\mathbb{1}_{[-1,1]}\left(\frac{X_i - x}{h}\right)$$

almost surely. $\sim \mathcal{U}[-1,1]$

Comeback to galaxies.

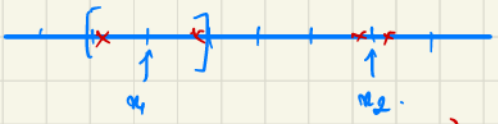
plot (density (galaxies, kernel = 'r', bw = 150))

In file Commandes _ Cours NP. R.

X = emelange
Y = dmelange

$f(x) = \alpha f_1(x, p_1) + (1-\alpha) f_0(x, p_0)$.

- **Comment:**



$[x-h, x+h]$

$h = 1$

$\tilde{f}_X(x_1) = \frac{1}{2n} \cdot 2 = \frac{1}{n}$ ← có 2 dữ liệu nằm trong khoảng $[x-1; x+1]$

$\tilde{f}_X(x_2) = \frac{1}{2n} \cdot 2 = \frac{1}{n}$

We would like to improve our estimator in order to give more weight to observation that are closer to x than the other.

- **Recall** Properties of estimators of a function g

1) Pointwise properties. (từng x^2)

a) Bias, for all $x \in \mathbb{R}$, $\text{Bias}(\hat{g}(x)) = \mathbb{E}[\hat{g}(x)] - g(x)$.

b) Variance ——— $\text{Var}(\hat{g}(x)) = \mathbb{E}[(\hat{g}(x) - \mathbb{E}(\hat{g}(x))^2]$

c) Quadratic error —— $\mathbb{E}[(\hat{g}(x) - g(x))^2] = \text{Var}(\hat{g}(x)) + (\text{Bias}(\hat{g}(x))^2$

2) Global (tổng quát cho hàm $\hat{g}$)

a) Sum of squared error : $SSE(\hat{g}) = \sum_{i=1}^{n} \left(\hat{g}(x_i) - g(x_i)\right)^2$.

b) Mean _____ : $MSE(\hat{g}) = \mathbb{E}\left[\sum_{i=1}^{n} \left(\hat{g}(x_i) - g(x_i)\right)^2\right]$

c) Integrated Squared Error : $ISE(\hat{g}) = \int \left[\hat{g}(x) - g(x)\right]^2 dx$

d) Mean Integrated Squared Error :

$$MISE(\hat{g}) = \int \mathbb{E}\left[\hat{g}(x) - g(x)\right]^2 dx$$

- **Property**

1) If $h = h_n \xrightarrow{n \to +\infty} 0$ then for all $x \in D_{f_X}$, $\tilde{f}_X(x)$ is an asymptotically unbiased estimator of $f_X(x)$.

2) If $h = h_n \xrightarrow{n \to +\infty} 0$ and $nh_n \xrightarrow{n \to +\infty} +\infty$, then for all $x \in D_{f_X}$, with $f_X(x) > 0$: $\tilde{f}_X(x) \xrightarrow[L^2]{n \to +\infty} f_X(x)$

- **Remarks**

. $\mathbb{E}\left[\tilde{f}_X(x)\right] = \dfrac{F_X(x+h) - F_X(x-h)}{2h} \xrightarrow{h \to 0^+} f_X(x)$.

. $Var\left(\tilde{f}_X(x)\right) \sim \dfrac{1}{2nh} f_X(x)$.

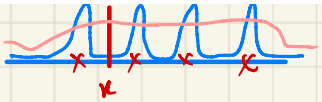
For n fixed (–) h small | h great (–)

Bias small | Bias great

Var great | Var small.

depend on việc của dữ liệu.

phụ thuộc vào dữ liệu.



26.10.2022.

B. Kernel estimator of the density function.

Either f_X is continuous at x.
or f_X is defined as F_X'

$F_X \rightsquigarrow$ có đạo hàm tại.

B1. Definition.

Idea: Replace the indicator function $w(t) = \frac{1}{2} \cdot \mathbb{1}_{[-1,1]}(t)$ by a smoother function called kernel and usually denoted K.

We assume that the kernel fulfills the following assumptions:

1) $\forall t \in \mathbb{R}, \quad K(t) \geq 0$

2) $\int K(t)\,dt = 1.$

3) $\int t.K(t)\,dt = 0.$ (có giá trị nhỏ nhất tại $t=0$)

4) K is maximum at 0 and decrease when $|t| \uparrow$.

Ex: $K(t) = \frac{1}{\sqrt{2\pi}} e^{-t^2/2} \rightsquigarrow f_X$ của $\mathcal{N}(0,1)$.

Definition. Kernel density function estimator Paizen 1964:

$$\hat{f}_X(x) = \frac{1}{nh} \sum_{i=1}^{n} K\left(\frac{x_i - x}{h}\right)$$

where K is a kernel, $h > 0$ is a smoothing parameter.

Main kernel examples

name	expression $K(t)$	graph.
uniform	$\frac{1}{2} \mathbb{1}_{[-1,1]}(t)$	
gaussian	$\frac{1}{\sqrt{2\pi}} e^{-t^2/2}$	
triangular	$(1-\lvert t\rvert)\mathbb{1}_{[-1,1]}(t)$	
Epanechnikov	$\frac{3}{4}(1-t^2)\mathbb{1}_{[-1,1]}(t)$	
Quartic	$\frac{15}{16}(1-t^2)^2$	
Triweight	$\frac{35}{32}(1-t^2)^3$	
Cosinus	$\frac{\pi}{4}\cos\left(\frac{\pi t}{2}\right)$	

Remarks

1) K is chosen to be maximal at 0 to give more weight to X_i that are close to x than to the others.

2) Using a smooth kernel will lead to a smooth estimator.

3) The role of h is exactly the same as for the moving window estimator

Comeback to R software.

1) Replace in the previous ex kernel = 'rectangular' by kernel = 'g' or 'e'
(gaussian, epanechnikov)

par(mfrow = c(2,3))

```
par(mfrow = c(1,1))
res = density(x, kernel = 'r')        # rectangular
plot(res)
lines(density(x, kernel = 'g', bw = res$bw), col = 'blue')   # Gaussian
   ______________ 'e' ______________ green)                  # epanechikov
lines(s, dmelange ______________________)   : true density.
```

- Come back to the galaxies.

```
library(MASS)
data(galaxies)
x = galaxies
```

- **Remarks**.

1) The effect of K is not so important (only for the smoothness it could be)

2) observe errors: at the points where the density is high / the curvature (ds cong.) of the density is important.

Pros	Cons.
. Smooth and non constant by part.	depend on K (no important effect) h
. weights of data depends on their proximity w.r.t x.	

2. Properties of the convergence rate.

Proposition:

1) If $K \geq 0$ then $\hat{f} \geq 0$

2) If K is a density $\hat{f}$ is a density.

3) If K is continuous then $\hat{f}$ is continuous

4) ____ C^p ____ C^p

Proof. Exercise.

Theorem — If f_X is C^2 on a neighborhood of x, a bounded (this could be relaxed if K has a compact support).

If $h_n \xrightarrow{n \to +\infty} 0$ and $nh_n \xrightarrow{n \to +\infty} +\infty$ then:

1) $\text{Bias}\left(\hat{f}(x)\right) = \frac{h_n^2}{2} f_X''(x) \mu_2(K) + o(h_n^2)$

h_n small → Bias small, plus évaluation $f_X''(x)$.

where $\mu_2(K) = \int t^2 K(t)\, dt$.

2) If $f_X(x) > 0$ then:

$$\text{MSE}\left(\hat{f}(x)\right) = \frac{h^4}{4}\left(f_X''(x)\right)^2 \mu_2^2(K) + \underbrace{\frac{1}{n.h_n} f_X(x).R(K)}_{\text{h_n small, $f_X(x)$ great} \to \text{MSE great}} + o\left(\frac{1}{n.h}\right)$$

where $R(K) = \int K^2(t)\, dt$.

↳ Cần dùng tiêu chí để lựa bộ chọn.

- **Almost sure convergence result:**

  Under some technical assumptions:

$$\hat{f}_X(x) = f_X(x) + O_{a.s}(h^2) + O_{a.s}\left(\sqrt{\frac{\ln(n)}{nh}}\right)$$

- **Theorem** If f_X is bound and C^2, $h_n \longrightarrow 0$

$$n.h_n \longrightarrow +\infty$$

$$f_X, f_X', f_X'' \in L^2(\mathbb{R})$$

then:

$$MISE\left(\hat{f}\right) = \frac{h^4}{4} \cdot \int \left(f_X''(t)\right)^2 dt \; \mu_2^2(K) + \frac{1}{n.h} R(K) + O\left(h^4 + \frac{1}{nh}\right)$$

↳ global.

⟶ aim: minimize MISE

- **Remarks.** from the asymptotic expression of MSE and MISE, we can try to get optimal values of h.

$$h \longmapsto AMISE(h) = \frac{h^4}{4} \int \left(f_X''(t)\right)^2 dt \, \mu_2^2(K) + \frac{1}{nh} R(K)$$

. Then we try to optimize AMISE w.r.t h:

$$AMISE'(h) = h^3 \int \left(f_X''(t)\right)^2 dt \, \mu_2^2(K) - \frac{1}{nh^2} R(K).$$

$$. \; AMISE'(h) = 0 \Rightarrow h^3 \int \left(f_X''(t)\right)^2 dt \, \mu_2^2(K) = \frac{1}{nh^2} R(K)$$

$$\Leftrightarrow h^5 = \frac{n^{-1} . R(K)}{\int \left(f_X''(t)\right)^2 dt \, \mu_2^2(K)}$$

$$\Leftrightarrow h = \left(\frac{R(K)}{R(f_X'') \, \mu_2^2(K)}\right)^{1/5} . n^{-1/5}.$$

↳ check that $n.h \longrightarrow +\infty$ and $h \longrightarrow 0$.

- $h_{AMISE} = \left(\frac{R(K)}{R(f_x'') \, \mu_2^2(K)} \right)^{1/5} \cdot n^{-1/5}$: global (optimum ẽvão x)

- On the same way :

$$h_{AMISE} = \left[\frac{f_x(x) \cdot R(K)}{\left(f_x''(x)\right)^2 \cdot \mu_2^2(K)} \right]^{1/5} \cdot n^{-1/5}$$: local (optimum ẽvão x)

$$AMISE\left(h_{AMISE}\right) = \frac{1}{4} \cdot \left[\frac{R(K)}{R(f_x'') \mu_2^2(K)} \right]^{4/5} \cdot R\left(f_x''\right) \mu_2^2(K)$$

$$+ n^{-1} \cdot n^{+1/5} \left[\frac{R(K)}{R(f_x'') \cdot \mu_2^2(K)} \right]^{-1/5} \cdot R(K) \cdot$$

$$= \frac{5}{4} \cdot \left[R^4(K) \cdot R(f_x'') \cdot \mu_2^2(K) \right]^{1/5} n^{-4/5} \cdot \text{(min)} .$$

- **Remarks**

  - in the parametric case usually : $\mathbb{E}_\theta\left[(\hat{\theta} - \theta)^2\right] \sim \frac{c}{n}$.

    if in addition :

    1) for all x : $\theta \longrightarrow f_\theta(x)$ is C^2.

    2) $\exists h \in L^2(\mathbb{R})$, $\left| \frac{\partial^2 f_\theta(x)}{\partial \theta^2} \right| \leq h(x)$. then

$$MISE\left(f_{\hat{\theta}}\right) \sim \frac{C}{n} .$$

  - for a histogram, with all the classes of the same amplitude h (biên độ)

    we may get $MISE(\text{hist}) \sim C_2 \cdot n^{-2/3}$

    ↑ histogram. → parametric estimator hội tụ nhanh hơn.

↳ 2 parameters < kernel K / smoothing h.

• ≥ dùng h_{AMISE} dc vì có f_x bên trong → tìm cách estimate.

B₃. Optimal choice of the kernel K.

Idea: the optimal expression of AMISE (h_{AMISE}) depends on:

$$C(K) = \left[R^4(K) \cdot \mu_2^2(K) \right]^{1/5}$$

We naturally want to minimize $C(K)$ w.r.t K under the constraint that K is a kernel.

Result: The winner is the Epanechnikov kernel (1969)

$$K_{opt}(t) = \frac{3}{4} \cdot (1 - t^2) \cdot \mathbb{1}_{[-1,1]}(t)$$

The efficiency of K is defined:

$$eff(K) = \left[\frac{C(K_{opt})}{C(K)} \right]^{5/4}$$

What does it mean?

nếu $K \neq K_{opt}$:

$$AMISE(n, K) = AMISE(n_0, K_{opt}) \longrightarrow$$ phải chọn n lớn hơn n_0

$$\Leftrightarrow \quad n^{-4/5} \cdot C(K) = n_0^{-4/5} C(K_{opt})$$

$$\Leftrightarrow \quad \left(\frac{n_0}{n} \right)^{4/5} = \frac{C(K_{opt})}{C(K)}$$

$$\Rightarrow \quad \frac{n_0}{n} = eff(K) < 1 \quad \left(\text{since } \frac{C(K_{opt})}{C(K)} < 1 \right)$$

$$\Rightarrow \quad n_0 = eff(K) \cdot n.$$

$\Leftrightarrow \quad n = n_0 \cdot \frac{1}{\text{eff}(K)}$

↳ nếu $k \neq k_{opt}$ → cần nhiều dữ liệu hơn n_0.

values eff(K)	K(t).
1.	Epanechnikov $\frac{3}{4}(1-t^2) \cdot \mathbb{1}_{[-1,1]}(t)$
√ 0,994	quartic $\frac{15}{16}(1-t^2)^2 \cdot \mathbb{1}_{[-1,1]}(t)$
0,987.	Triweight $\frac{35}{32}(1-t^2)^3$
0,951.	Gaussian.
? 0,986	Triangular $(1-\lvert t\rvert)$
0,93.	Uniform

tối ưu ☺

B4) Automatic choice of the smoothing parameter.

1) The normal reference.

2) The plug-in method.

3) The cross-validation method.

1) The normal reference.

$$h_{AMISE} = \left[\frac{R(K)}{R(f_X'') \cdot \mu_2^2(K)} \right]^{1/5} \cdot n^{-1/5}$$

- Idea: compute $R(f_X'')$ using the Gaussian hypothesis on the density:

$$f_X(t) = \frac{1}{\sigma\sqrt{2\pi}} e^{-(t-\mu)^2/2\sigma^2}$$

$$f_x'(t) = \frac{1}{\sigma\sqrt{2\pi}} \cdot \frac{-2(t-\mu)}{2\sigma^2} \cdot e^{-(t-\mu)^2/2\sigma^2}$$

$$= \frac{1}{\sigma\sqrt{2\pi}} \cdot \frac{-(t-\mu)}{\sigma^2} \cdot e^{-(t-\mu)^2/2\sigma^2}$$

$$f_x''(t) = \frac{1}{\sigma\sqrt{2\pi}} \left[\frac{-1}{\sigma^2} e^{-(t-\mu)^2/2\sigma^2} + \frac{(t-\mu)^2}{\sigma^4} \cdot e^{-(t-\mu)^2/2\sigma^2} \right]$$

$$= \frac{1}{\sigma\sqrt{2\pi}} \left[\frac{(t-\mu)^2}{\sigma^4} - \frac{1}{\sigma^2} \right] e^{-(t-\mu)^2/2\sigma^2} \cdot$$

$$\Rightarrow R(f_x'') = \int (f_x''(t))^2 \, dt$$

$$= \int \frac{1}{2\pi\sigma^2} \cdot \frac{1}{\sigma^4} \cdot \left[\frac{(t-\mu)^2}{\sigma^2} - 1 \right]^2 e^{-(t-\mu)^2/\sigma^2} \, dt$$

Let $z = \frac{\sqrt{2}(t-\mu)}{\sigma} \Rightarrow dz = \frac{\sqrt{2}\,dt}{\sigma}$. (đề xuất bởi chuyên)

$$\Rightarrow R(f_x'') = \int \frac{1}{2\pi\sigma^2} \cdot \frac{1}{\sigma^4} \cdot \left(\frac{z^2}{2} - 1 \right)^2 \cdot e^{-z^2/2} \frac{\sigma}{\sqrt{2}} dz \cdot$$

$$= \frac{1}{2\sqrt{\pi}\cdot\sigma^5} \int \frac{1}{\sqrt{2\pi}} \left(\frac{z^2}{2} - 1 \right)^2 e^{-z^2/2} \, dz \cdot$$

$$= \frac{1}{2\sqrt{\pi}\cdot\sigma^5} \int \frac{1}{\sqrt{2\pi}} \left(\frac{z^4}{4} - z^2 + 1 \right) e^{-z^2/2} \, dz$$

$$= \frac{1}{2\sqrt{\pi}\cdot\sigma^5} \left(\frac{E[Z^4]}{4} + 1 - \underbrace{E[Z^2]}_{1} \right) \quad \text{where } Z \sim N(0,1)$$

$$= \frac{1}{2\sqrt{\pi}\cdot\sigma^5} \cdot \frac{3}{4} = \frac{3}{8\sqrt{\pi}\cdot\sigma^5} \cdot$$

- $R(f_X'') = \dfrac{3}{8\sqrt{\pi}\cdot\hat{\sigma}^5}$, where $\hat{\sigma} = \min\left(S_{n-1}, \dfrac{R}{1.349}\right)$

$R = q_{0.75} - q_{0.25}$: tính dựa vào data

↑ phân vị thứ 0.75.

- **Remark.**

  - If $Z \sim N(0,1)$ then $q_Z(0.75) - q_Z(0.25) = 1.349$. (tính dựa vào χ^2; theory quantile)

  If $X \sim N(\mu, 1)$ then $q_X(0.75) - q_X(0.25) = 1.349$.

  If $Y \sim N(\mu, \sigma)$ then $q_Y(0.75) - q_Y(0.25) = \sigma \times 1.349$.

  Indeed, $P(Y \le t) = 0.75$

  $\Leftrightarrow P\left(\underbrace{\dfrac{Y-\mu}{\sigma}}_{Z} \le \underbrace{\dfrac{t-\mu}{\sigma}}_{q_Z(0.75)}\right) = 0.75.$

  $\Rightarrow q_Y(0.75) = \sigma \cdot q_Z(0.75) + \mu.$

  $q_Y(0.25) = \sigma \cdot q_Z(0.25) + \mu.$

  $\Rightarrow q_Y(0.75) - q_Y(0.25) = \sigma\left(q_Z(0.75) - q_Z(0.25)\right)$

  $= \sigma \times 1.349.$

- $h_{NR} = \left(\dfrac{8\sqrt{\pi}\cdot\hat{\sigma}^5}{3} \cdot \dfrac{R(K)}{\mu_2^2(K)}\right)^{1/5} \cdot n^{-1/5}.$

  normal reference

  $= \hat{\sigma} \cdot \left(\dfrac{8\sqrt{\pi}\cdot R(K)}{3\mu_2^2(K)}\right)^{1/5} \cdot n^{-1/5}$

Kernel	$\left(\frac{8\sqrt{\pi}R(K)}{3\mu_2^2(K)}\right)^{1/5}$
Epanechnikov	2.34
Gaussian	1.06
Triweight	2.78

- Pros: Easy to compute and implement.
- Cons: If our normal assumption is incorrect, we may be wrong.

2) The plug-in method

Idea: Estimate $R(f_X'')$ by $R(\hat{f}_0'')$ where $\hat{f}_0$ is a non optimal kernel estimator of f_X. A C^2 kernel is used. (Sheather-Jones, 1991)

3) Cross-validation.

$$MISE(\hat{f}) = E\left[\int [\hat{f}_X(t) - f_X(t)]^2 dt\right]$$

$$= E\left[\int \left[\hat{f}_X^2(t) - 2\hat{f}_X(t) f(t) + f_X^2(t)\right] dt\right]$$

$$= E\left[\int \hat{f}_X^2(t)\, dt\right] - 2E\left[\int \hat{f}_X(t) f_X(t)\, dt\right] + \int f_X^2(t)\, dt.$$

- Remark. minimizing $MISE(\hat{f})$, is equivalent to minimizing $MISE(\hat{f}) - \int f_X^2(t)\, dt$
- Remark. $\int \hat{f}^2(t)\, dt$ is an unbiased estimator of $E\left[\int \hat{f}(t)\, dt\right]$ $= \varphi(R)$

- $\frac{1}{n}\cdot\sum_{i=1}^{n} f^{-i}(X_i)$, where f^{-i} is the kernel estimator computed from the subsample where X_i has been removed.

$$f^{-i}(t) = \frac{1}{(n-1).h}\sum_{\substack{j=1\\ j\neq i}}^{n} K\left[\frac{X_j - t}{h}\right]$$

bỏ X_i ra để tính f^{-i}.

- Fact: $\mathbb{E}\left[\frac{1}{n}\cdot\sum_{i=1}^{n}\hat{f}^{-i}(X_i)\right] = \mathbb{E}\left[\int f_X(t)\cdot\hat{f}_X(t)\,dt\right]$

Proof

$$\mathbb{E}\left[\frac{1}{n}\cdot\sum_{i=1}^{n}\hat{f}^{-i}(X_i)\right] = \mathbb{E}\left[\hat{f}^{-1}(X_1)\right] \quad (\text{since } f^{-i}(X_i) \overset{i.i.d}{\sim})$$

$$= \mathbb{E}\left[\mathbb{E}\left[f^{-1}(X_1)\,|\,X_2 \dots X_n\right]\right] = \mathbb{E}\left[\int \hat{f}^{-1}(t)\,f(t)\,dt\right]$$

$$= \mathbb{E}\left[\int \frac{1}{(n-1).h}\sum_{j=2}^{n} K\left(\frac{X_j - t}{h}\right) f_X(t)\,dt\right]$$

$$= \int \frac{1}{(n-1).h}\cdot\sum_{j=2}^{n}\mathbb{E}\left[K\left(\frac{X_j - t}{h}\right)\right] f_X(t)\,dt$$

$$= \int\left(\frac{1}{h}\int K\left(\frac{y-t}{h}\right) f_X(y)\,dy\right) f_X(t)\,dt.$$

we need to prove that $= \mathbb{E}\left[\hat{f}_X(t)\right]$

$$= \int \frac{1}{h}\cdot\mathbb{E}\left[K\left(\frac{X-t}{h}\right)\right] f_X(t)\,dt.$$

- $$\mathbb{E}\left[\int \hat{f}_X(t)\,f_X(t)\,dt\right] = \int f_X(t).\,\mathbb{E}\left[\frac{1}{n.h}\sum_{i=1}^{n} K\left(\frac{X_i - t}{h}\right)\right] dt$$

$$= \int f_X(t).\frac{1}{nh}\,\mathbb{E}\left[\sum_{i=1}^{n} K\left(\frac{X_i - t}{h}\right)\right] dt$$

$$= \int f_X(t) \cdot \frac{1}{h} \cdot \mathbb{E}\left[K\left(\frac{X-t}{h}\right)\right] dt .$$

$$\Rightarrow \mathbb{E}\left[\frac{1}{n}\cdot\sum_{i=1}^{n} f^{-i}(X_i)\right] = \mathbb{E}\left[\int f_X(t)\cdot \hat{f}(t)\,dt\right] \quad \square$$

Now,

$$\hat{\varphi}(h) = \int \hat{f}^2(t)\,dt - 2\cdot\frac{1}{n}\cdot\sum_{i=1}^{n}\hat{f}^{-i}(X_i)$$

is hence an unbiased estimator of $\varphi(h)$

And we have:

$$h_{ucv} = \underset{h}{\operatorname{argmin}}\,\hat{\varphi}(h)$$

↑ cross-validation.

Back to R:

- `x = duelange.`

  `plot( density(x, bw = 'nrd0' ))` ← mặc định.

  `lines( ___(x, bw = 'nrd'), col = 'blue')`

  `_____ 'SJ' _____ 'green')`

  `:` `'ucv'`

  `lines(s, duelange ...)` : hàm chưa xác

  ↳ ucv is okay nhất.

  bw = 'nrd' : normal reference.
  = 'SJ' : plug in method
  = 'ucv' : unbiased cross-validation

- `library(sm)`

  `sm.density(x, method = 'normal')` (hoặc cho h = _)

  `_____ 'cv', add = T, col = 'blue')`

  `_____ 'sj', _____ green.`

B_5. Mode estimation. ← xap xi^2 mode.

$$\theta_0 \in \operatorname*{argmax}_{\theta \in \mathbb{R}} f_x(\theta)$$

$$\hat{\theta} \in \operatorname*{argmax}_{\theta \in \mathbb{R}} \hat{f}_x(\theta).$$

- Idea: compute on a grid $t_1, \dots, t_p$ $\hat{f}_x$

  $\hat{f}(t_1) \dots \hat{f}(t_p)$

  and get t_i giving the maximum value.

  a, b, M : $t_1, \dots t_p$ from a to b, M points.

↳ ? : change max to min?

28.10.2022.

Chapter 2. Kernel estimator for nonparametric regression.

mối liên hệ giữa

- Reminder: Regression model:

$$Y(X) = r(X) + \varepsilon \quad \text{, with} \quad \begin{cases} Y, X \text{ are real-valued.} \\ \mathbb{E}[\varepsilon | X] = 0 \end{cases}$$

$\Rightarrow r(x) = \mathbb{E}[Y|X]$.

- The function r is unknown and we want to estimate it.

- Linear regression: $r(x) = a.X + b$.

$(X_1, Y_1), \dots, (X_n, Y_n) \overset{i.i.d}{\sim} (X, Y)$

1) Definition

Nadaraya - Watson kernel estimator (1964).

$$\hat{r}(x) = \begin{cases} \dfrac{\sum_{i=1}^{n} Y_i . K\left(\frac{X_i - x}{h}\right)}{\sum_{i=1}^{n} K\left(\frac{X_i - x}{h}\right)} & \text{if } \sum_{i=1}^{n} K\left(\frac{X_i - x}{h}\right) \neq 0 \\ \overline{Y} & \text{else.} \end{cases}$$

- Remarks.

1) $\hat{r}(x) = \sum_{i=1}^{n} Y_i \, w_i(x)$ with $w_i(x) = \begin{cases} \dfrac{K\left(\frac{X_i - x}{h}\right)}{\sum_{j=1}^{n} K\left(\frac{X_j - x}{h}\right)} & \text{if } \sum K \neq 0 \\ \frac{1}{n} & \text{else.} \end{cases}$

If h is great: the weight $w_i(x)$ are all great and similar and you get an estimator of $\overline{Y}$.

h is small : the estimator is $\bar{Y}$ unless x is very closed to x_i, $1 \leq i \leq n$

and in that case, $w_i(x) \simeq 1$ and $w_j(x) \simeq 0$, $j \neq i$.

and hence $\hat{r}(x) \simeq Y_i$

2) $\hat{r}(x) = \dfrac{\hat{g}(x)}{\hat{f}(x)}$ where $\hat{g}(x) = \dfrac{1}{nh} \sum_{i=1}^{n} Y_i K\left(\dfrac{x_i - x}{h}\right)$

$\hat{f}(x) = \dfrac{1}{nh} \sum_{i=1}^{n} K\left(\dfrac{x_i - x}{h}\right)$

whenever $\hat{f}(x) \neq 0$.

2) Asymptotic properties.

• Hypotheses

a) r and f_X are C^2 on a neighborhood of x.

b) $f_X(x) > 0$

c) $\lim\limits_{n \to +\infty} nh_n = +\infty$ and $\lim\limits_{n \to +\infty} h_n = 0$

d) Let consider ϕ : $\phi(x) = \mathbb{E}[Y^2 | X]$

Assume that ϕ is continuous at point x.

e) K is bounded, integrable, positive, symetric and with compact support.

• Theorem 1 — If hypotheses (a) → (e) hold then :

$$\text{MSE}(\hat{r}(x)) = \mathbb{E}\left[(\hat{r}(x) - r(x))^2\right] = B^2(x) \cdot h^4 + V(x) \cdot \frac{1}{nh} + O\left(h^2 + \frac{1}{nh}\right)$$

where $B(x) = \frac{1}{2 f_X(x)} \mu_2(K) \cdot \left(g^{(2)}(x) - r(x) f^{(2)}(x) \right)$

$V(x) = \frac{\Phi(x) - r^2(x)}{f_X(x)} \cdot R(K)$

- Complementary hypotheses:

$$MISE(\hat{r}) = \mathbb{E}\left[\int (\hat{r}(x) - r(x))^2 w(x) \, dx \right]$$

where f) w is a weight function, positive, bounded and with compact support C

a') r and f_X are C^2 on a neighborhood of C.

d') Φ is continuous on a _____.

- Theorem 2 — If hypotheses (a'), (b), (c), (d'), (e), (f) hold then:

$$MISE(\hat{r}) = B^2 h^4 + V \cdot \frac{1}{nh} + O\left(\frac{1}{nh} + h^4 \right)$$

where $B^2 = \int B^2(x) \cdot w(x) \, dx$ and $V = \int V(x) w(x) \, dx$.

- Other results:

. We could also get results (under technical assumptions) on almost sure convergence:

$$\hat{r}(x) - r(x) = O_{as}(h^2) + O_{as}\left(\sqrt{\frac{\ln n}{nh}} \right) \quad (3)$$

. If we assume that r and f_X are C^3 and K is a kernel of order p:

$$\hat{r}(x) - r(x) = O_{as}(h^p) + O_{as}\left(\sqrt{\frac{\ell_n n}{nh}}\right) \quad (4)$$

. **Convergence rates**

. If (4) holds : $h \sim c\left(\frac{n}{\ell_n(n)}\right)^{\frac{-1}{2p+1}}$.

$$\hat{r}(x) - r(x) = O_{as}\left[\left(\frac{n}{\ell_n n}\right)^{\frac{p}{2p+1}}\right]$$

. If condition of theorem 1 (resp. Theorem 2) hold, taking $h \sim C n^{-1/5}$

$$MSE(\hat{r}(x)) = O(n^{-4/5})$$

resp

$$MISE(\hat{r}) = O(n^{-4/5})$$

3) **Choice of the smoothing parameter in practice. (Choice of h).**

. **A popular method : cross - validation.**

Introduce $CV(h) = \sum_{i=1}^{n} (Y_i - \hat{r}^{-i}(X_i))^2 w(X_i)$

where r^{-i} is computed from all the data excepted (X_i, Y_i)

$$h_{cv} = \underset{h \in H_n}{\arg\min}\, CV(h)$$

. **Remark.**

$$CV(h) = \sum_{i=1}^{n} (r(X_i) + \varepsilon_i - \hat{r}^{-i}(X_i))^2 w(X_i)$$

lấy E thì giống MISE

$$= \sum_{i=1}^{n} \varepsilon_i^2 w(X_i) + \sum_{i=1}^{n} (r(X_i) - \hat{r}^{-i}(X_i))^2 w(X_i)$$

$$+ 2\sum \varepsilon_i (r(X_i) - \hat{r}^{-i}(X_i)) w(X_i)$$

ε_i độc lập với h.

lấy E thì = 0.

. **Property**. If condition of Theorem 2 hold and H_n contains values fulfilling (e) then (with technical conditions)

$$\frac{\inf_{R \in H_n} MISE(\hat{R})}{MISE(\hat{R}_{cv})} \xrightarrow[n \to +\infty]{as} 1 .$$

. **R programming**: data (pressure)

```
t = ...
p = ...
plot(t,p)
sm.regression(t, p, method = 'cv')
```

. **Simulated data**:

```
n = 1000
x = rexp(x, 2)
y = 10*exp(-3*x) + rnorm(n, 0, exp(-x/2))
plot(function(x) 10*exp(-3*x), 0, 3)
```

↳ trục x chạy từ 0 → 3.

```
lines(x, y, col = 'red', type = 'p')
```

← vẽ từng điểm

```
res = sm.regression(x, y, method = 'cv', add = T, col = 'blue', eval.points = sort(x))
```

'df'

'aice'

.
```
e = rnorm(n, 0, 1)
x = rnorm(n, 2, 0.5)
y = cos(2*pi*x) + 30*exp(-3*x) + 0.1*e.
plot(x, y)
```

. names(res) : xem res gồm ≃ bt gì.

cv = 0

for (i in 1:n)

{cv = cv + $\frac{1}{n}$ (y[i] − sm.regression (x[−i], y[−i], method = 'cv', display = 'none', eval.points = x[i]) $estimate)^2}

(tính tại x[i])

abline (0,1) : vẽ đt thẳng $y = 1x + 0$

• Discrimination: (bt phân loại)

Y takes value in $\mathcal{Y}$

X real valued.

. Estimated Bayes classifer:

$$\hat{y}_0 = \underset{y \in \mathcal{Y}}{\arg\max}\ \widehat{\mathbb{P}(Y = y(X))}\Big|_{X = x_0}$$

$$= \underset{y \in \mathcal{Y}}{\arg\max}\ \hat{r}_y(x_0)$$

Reminder: Take $Z_y = \mathbb{1}\{Y = y\}$

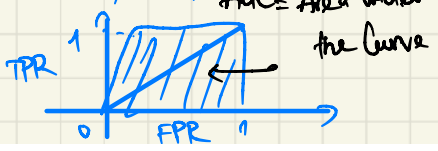
$$Z_y = r_y(X) + \varepsilon_y.$$

$$= \mathbb{P}(Y = y | X)$$

. classif_NP (Iherma, test, c(51,27,72))

dữ liệu mới cần phân loại.

Y \ $\hat{Y}$	−	+
−	TN	FP
+	FN	TP



AUC = Area under the Curve

$$\hat{Y} = \begin{cases} 1 & \text{if } \hat{P}(Y=y|X)\big|_{X=x_0} > \underbrace{0.5}_{t} \\ 0 & \text{else} \end{cases}$$

AUV càng lớn thì phân loại càng tốt

đồ thị về phụ E vào t

- $FPR(t) = \dfrac{FP(t)}{TN(t) + FP(t)}$
- $TPR(t) = \dfrac{TP(t)}{FN(t) + TP(t)}$

28. 11. 2022

Chapter 3. From univariate, to multivariate and functional data.

1) Multivariate case.

$X : \Omega \to \mathbb{R}^p$

$Y : \Omega \to \mathbb{R}$

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

$$\hat{r}(x) = \begin{cases} \sum Y_i K\left(\frac{X_i - x}{h}\right) \Big/ \sum K\left(\frac{X_i - x}{h}\right) & \text{if } \sum K\left(\frac{X_i - x}{h}\right) \neq 0 \\ \bar{Y} & \text{else.} \end{cases}$$

Ex: 1) $\hat{K}(\vec{t}) = K_1(t_1) \cdot K_2(t_2) \dots K_p(t_p)$, where $K_i : \mathbb{R} \to \mathbb{R}$.

2) $K(\vec{t}) = K_0(\|\vec{t}\|)$ $\quad K_0 : \mathbb{R} \to \mathbb{R}$.

Nd: $$K\left(\frac{X_i - x}{h}\right) = \prod_{j=1}^{p} K_j\left(\frac{X_{ij} - x}{h_j}\right)$$

$$K\left(\frac{X_i - x}{h}\right) = K_0\left(\frac{\|X_i - x\|}{h}\right)$$

We can get, under some conditions:

$$MSE(\hat{r}(x)) = B^2(x) \cdot h^4 + V(x) \cdot \frac{1}{n \cdot h^p} + \mathcal{O}\left(h^4 + \frac{1}{nh^p}\right)$$

$$MISE(\hat{r}) = B^2 \cdot h^4 + V \cdot \frac{1}{nh^p} + \mathcal{O}\left(h^4 + \frac{1}{nh^p}\right)$$

- Hence the optimal smoothing parameter:

$$h_{opt} = \left[\frac{V_p}{4B^2} \right]^{\frac{1}{p+4}} n^{\frac{-1}{p+4}}$$

Then optimal AMISE $\sim C.n^{-\frac{4}{p+4}}$ and hence the cv rate $C.n^{-2/p+4}$

For a.s convergence, the optimal rate $C.\left(\frac{n}{\ln(n)}\right)^{\frac{-2}{4+p}}$.

- **Important result.** Stone 1981 - 1982.

if f_X and r are C^2 then the optimal convergence rate for any estimator:

1) $C.n^{-\frac{k}{2k+p}}$ for the $\mathbb{L}_q$ cv $q>0$.

2) $C.\left(\frac{n}{\ln(n)}\right)^{-\frac{k}{2k+p}}$ for the $\mathbb{L}_\infty$ cv.

- **An explaination.** the curse of dimensionality.

$n = 1\,000\,000$ values $\in [[0,1]^{\otimes p}]$

. Take $h = 0.05$: $\mathbb{P}\left(\max|X_i - x_i| \leq h\right) \leq \frac{0.1^p}{1^p} = 0.1^p$

. Let N_p be the number of values in $B_h(x) = \{t, \max|t_i - x_i| \leq h\}$

$N_p \sim \mathcal{B}in(n, p_h(x))$

$\mathbb{E}[N_p] = n.p_h(x) \leq 10^6 \times (0.1)^p = 10^{6-p}$

for $p=1$	10^5
$p=2$	10^4
$p=3$	10^3
$p=4$	10^2

$p = 5$ ✓ $\quad 10'$

$p = 6$ $\quad 1$.

. **Remark** The problem does not come from the estimator method but from the nature of the problem.

. We may consider semi-parametric model:

. Additive model:

$$r(x) = \mu + \sum_{i=1}^{n} r_i(x_i)$$

. Single index model:

$$r(x) = r(\langle x, \theta \rangle)$$

. **Practice on R** : faithful data.

số điểm chia.

$e_0 =$ seq (min (eruptions), max (eruptions), len = 100)

$w_0 =$ _____ (waiting) ____ (waiting) _____)

sm. density (cbind (eruptions, waiting), eval. points = cbind (e_0, w_0)

eval. grid = T, structure = 'common')

$h_1 = h_2 = h$: 'common'.

$h_1 \neq h_2$: 'separate'

$h_i = \sigma_i h$: 'scaled'

. **Pirate data**.

t

p

x

y

sm. regression (cbind (x, y), p, structure = 'common', eval. points = cbind (x, y), method = 'cv')

2. Functional case.

- $X: \Omega \to (\mathcal{F}, d)$ d is a pseudo-distance ($d^2 \neq$ nhau có khi d khoảng cách $= 0$)

$$\hat{r}(x) = \begin{cases} \dfrac{\sum_{i=1}^{n} Y_i K\left(d(X_i, x)/h\right)}{\sum_{i=1}^{n} K\left(d(X_i, x)/h\right)} & \text{if } \sum K\left(d(X_i, x)/h\right) \neq 0 \\ \bar{Y} & \text{else.} \end{cases}$$

- **Remarks**

1) for NP models, the curse of dimensionaly cannot be avoided.

$$Y = r(X) + \varepsilon, \quad \mathbb{E}[\varepsilon | X] = 0$$

$$r_d(x) = \mathbb{E}[Y | z, d(x, z) = 0] : \text{maybe} \neq r(x)$$

2) A kind of semi-parametric model.

3) Idea: use the pseudo-distance as a tool to extract relevant information on Y from X

- **Example of pseudo-distance**

1) Given the basis $\Psi_1, \ldots, \Psi_k, \ldots$: ortho-normal basis.

$$X = \sum_{j=1}^{+\infty} \lambda_j \Psi_j$$

$$x = \sum_{j=1}^{+\infty} \alpha_j \Psi_j$$

$$d(X, x) = \sum_{i=1}^{M} (\lambda_j - \alpha_j)^2$$

2) Project on M first PCA scores and do the same.

3) $\mathbb{L}^2$ distance on derivatives:

$$d(x, z) = \int \left(x_{(t)}^{(k)} - z_{(t)}^{(k)} \right)^2 dt$$