Frame-work in Learning theory

Dominique Picard

Laboratoire Probabilités et Modèles Aléatoires
Universités Paris VII
Joint work with G. Kerkyacharian (LPMA)

Texas A&M- October 2007.

http ://www.proba.jussieu.fr/mathdoc/preprints/index.html

Bounded regression/learning problem : Model

1. $Y_i = f_\rho(X_i) + \epsilon_i,\ i = 1 \ldots n$
2. $\epsilon_i's$, i.i.d. bounded random variables
3. $X_i's$ i.i.d. random variables on a set $\mathbb{X}$ = compact domain of $\mathbb{R}^d$. Let ρ be the common (unknown) law of the vector $Z = (X, Y)$
4. f_ρ is a bounded unknown function.
5. Two kind of hypotheses
   (a) $f_\rho(X_i)$ orthogonal to ϵ_i (learning)
   (b) $X_i \perp\!\!\!\perp \epsilon_i$ (bounded regression theory)

Cucker and Smale, Poggio and Smale,..

Aim of the game

1. Minimize among 'estimators' $\hat{f} = \hat{f}(x, (X, Y)_1^n)$

$$\mathcal{E}(\hat{f}) := \mathcal{E}_\rho(\hat{f}) := \int_{\mathbb{X}\times\mathbb{R}} (\hat{f}(x) - y)^2 d\rho(x, y)$$

2. $$f_\rho(x) = \int y d\rho(y|x)$$

3. $$\mathcal{E}(\hat{f}) = \|\hat{f} - f_\rho\|_{\rho_X}^2 + \mathrm{err}(f_\rho)$$

4. $\mathcal{E}(\hat{f}) = \int_{\mathbb{X}} (\hat{f}(x) - f_\rho(x))^2 d\rho_X(x) + \int_{\mathbb{X}\times\mathbb{R}} (f_\rho(x) - y)^2 d\rho(x, y)$

Measuring the risk

1. Mean square error : $\mathbb{E}_{\rho^{\otimes n}} \|\hat{f}((X,Y)_1^n) - f_\rho\|_{\rho_X}$

2. Probability bounds : $\mathbb{P}_{\rho^{\otimes n}}\{\|\hat{f}((X,Y)_1^n) - f_\rho\|_{\rho_X} > \eta\}$

Mean square Errors and Probability bounds

– Assume f_ρ belongs to a set Θ, $\rho \in \mathcal{M}(\Theta)$ consider the Accuracy Confidence Function :

–

$$\mathbf{AC}_n(\Theta, \eta) := \inf_{\hat{f}} \sup_{\rho \in \mathcal{M}(\Theta)} \mathbb{P}_{\rho^{\otimes n}}\{\|f_\rho - \hat{f}\|_{\rho X} > \eta\}$$

–

$$\mathbf{AC}_n(\Theta, \eta) \geq C\{ \begin{array}{ll} e^{-cn\eta^2}, & \eta \geq \eta_n, \\ 1, & \eta \leq \eta_n, \end{array}$$

DeVore, Kerkyacharian, P, Temlyakov

- $$\mathbf{AC}_n(\Theta, \eta) \geq C\{ \begin{array}{ll} e^{-cn\eta^2}, & \eta \geq \eta_n, \\ 1, & \eta \leq \eta_n, \end{array}$$
- $\ln \bar{N}(\Theta, \eta_n) \sim c^2 n \eta_n^2$
- $$\bar{N}(\Theta, \delta) := \sup\{N : \exists\ f_0, f_1, ...f_N \in \Theta, \quad \text{with}$$
$$c_0\delta \leq \|f_i - f_j\|_{L_2(\rho_X)} \leq c_1\delta,\ \forall i \neq j\}.$$

–

$$\inf_{\hat{f}} \sup_{\rho \in \mathcal{M}(\Theta)} \mathbb{P}_{\rho^{\otimes n}}\{\|f_\rho - \hat{f}\| > \eta\} \geq C\{ \begin{array}{ll} e^{-cn\eta^2}, & \eta \geq \eta_n, \\ 1, & \eta \leq \eta_n, \end{array}$$

– $\eta_n = n^{-\frac{s}{2s+d}}$ for the Besov space $B^s_q(L_\infty(\mathbb{R}^d))$

– In statistics, minimax results

$$\inf_{\hat{f}} \sup_{\rho \in \mathcal{M}'(B^s_q(L_\infty(\mathbb{R}^d)))} \mathbb{E}\|f_\rho - \hat{f}\|_{dx} \geq cn^{-\frac{s}{2s+d}}$$

Ibraguimov, Hasminski, Stone 80-82

Mean square estimates

$$\hat{f} = \text{Argmin}\{\frac{1}{n}\sum_{i=1}^{n}(Y_i - f(X_i))^2,\ f \in \mathcal{H}_n\}$$

1. 2 important problems :
   (a) Not always easy to implement
   (b) depending on Θ : Search for 'Universal' estimates : working for a class of spaces Θ

Oracle Case

$$(P) : \frac{1}{n}\sum_{i=1}^{n} K_k(X_i)K_l(X_i) = \delta_{kl}$$

((K_k) o.n.b. for the empirical measure on the $X_i's$)

1. $\mathcal{H}_n^{(1)} = \{f = \sum_{j=1}^{p} \alpha_j K_j\}$ (linear)
2. $\mathcal{H}_n^{(2)} = \{f = \sum_{j=1}^{p} \alpha_j K_j,\ \sum |\alpha_j| \leq \kappa\}$
   (l_1 constraint)
3. $\mathcal{H}_n^{(3)} = \{f = \sum_{j=1}^{p} \alpha_j K_j,\ \#\{|\alpha_j| \neq 0\} \leq \kappa\}$
   (sparsity)

$$\hat{\alpha}_k = \frac{1}{n}\sum_{i=1}^{n} K_k(X_i)Y_i,$$

$$\hat{\alpha}_k^{(1)} = \mathrm{sign}(\hat{\alpha}_k)|\hat{\alpha}_k - \lambda|_+, \qquad \hat{\alpha}_k^{(2)} = \hat{\alpha}_k I\{|\hat{\alpha}_k| \geq \lambda\}$$

1. $\mathcal{H}_n^{(1)} = \{f = \sum_{j=1}^{p} \alpha_j K_j\}$

   . $\qquad \hat{f} = \sum_{j=1}^{p} \hat{\alpha}_j K_j$

2. $\mathcal{H}_n^{(2)} = \{f = \sum_{j=1}^{p} \alpha_j K_j,\ \sum |\alpha_j| \leq \kappa\}$

   . $\qquad \hat{f}^{(1)} = \sum_{j=1}^{p} \hat{\alpha}_j^{(1)} K_j$

3. $\mathcal{H}_n^{(3)} = \{f = \sum_{j=1}^{p} \alpha_j K_j\ \#\{|\alpha_j| \neq 0\} \leq \kappa\}$

   . $\qquad \hat{f}^{(2)} = \sum_{j=1}^{p} \hat{\alpha}_j^{(2)} K_j$

Universality properties

$$\hat{\alpha}_k = \frac{1}{n}\sum_{i=1}^{n} K_k(X_i)Y_i,$$

$$\hat{\alpha}_k^{(1)} = \mathrm{sign}(\hat{\alpha}_k)|\hat{\alpha}_k - \lambda|_+, \qquad \hat{\alpha}_k^{(2)} = \hat{\alpha}_k I\{|\hat{\alpha}_k| \geq \lambda\}$$

$$\hat{f}^{(1)} = \sum_{j=1}^{p} \hat{\alpha}_j^{(1)} K_j, \quad \hat{f}^{(2)} = \sum_{j=1}^{p} \hat{\alpha}_j^{(2)} K_j$$

How to mimic te oracle ?

1. Condition (P) : $\frac{1}{n}\sum_{i=1}^{n} K_r(X_i)K_l(X_i) = \delta_{rl}$ is not realistic.
2. How to replace (P) by P(δ) 'δ − close' to (P) ?

Consider for instance the sparsity penalty

We want to minimize :

$$C(\alpha) := \frac{1}{n}\sum_{i=1}^{n}(Y_i - \sum_{j=1}^{p}\alpha_j K_j(X_i))^2 + \lambda\#\{\alpha_j \neq 0\}$$

$$= \frac{1}{n}\|Y - K^t\alpha\|_2^2 + \lambda\#\{\alpha_j \neq 0\}$$

$$= \frac{1}{n}\|Y - \mathrm{proj}_V(Y)\|_2^2 + \frac{1}{n}\|\mathrm{proj}_V(Y) - K^t\alpha\|_2^2 + \lambda\#\{\alpha_j \neq 0\}$$

$V = \{(\sum_{j=1}^{p} b_j K_j(X_i))_{i=1}^{n},\ b_j \in \mathbb{R}\},\ K_{ji} = K_j(X_i)$ $p \times n$ matrix

Case $\lambda = 0$

$$C(\alpha) = \frac{1}{n}\|Y - \text{proj}_V(Y)\|_2^2 + \frac{1}{n}\|\text{proj}_V(Y) - K^t\alpha\|_2^2.$$

$$K^t\hat{\alpha} = \text{proj}_V(Y)$$

$$K^t\hat{\alpha} = K^t(KK^t)^{-1}KY$$

$$\hat{\alpha} = (KK^t)^{-1}KY$$

Regression text-books

Case $\lambda \neq 0$

$$C(\alpha) = \frac{1}{n}\|Y - \text{proj}_V(Y)\|_2^2 + \frac{1}{n}\|\text{proj}_V(Y) - K^t\alpha\|_2^2 + \lambda\#\{\alpha_j \neq 0\}$$

Minimize $C(\alpha)$ equivalent to minimize $D(\alpha)$

$$\begin{aligned} D(\alpha) &= \frac{1}{n}\|\text{proj}_V(Y) - K^t\alpha\|_2^2 + \lambda\#\{\alpha_j \neq 0\} \\ &= (\alpha - \hat{\alpha})^t \frac{1}{n} KK^t(\alpha - \hat{\alpha}) + \lambda\#\{\alpha_j \neq 0\} \end{aligned}$$

Condition (P) : $\frac{1}{n}\sum_{i=1}^{n} K_r(X_i)K_l(X_i) = \delta_{rl}$

- then the $p \times p$ matrix

$$M_{np} = \frac{1}{n} K K^t = Id$$

$$(M_{np})_{kl} = (\frac{1}{n}\sum_{i=1}^{n} K_l(X_i)K_k(X_i))_{kl}$$

- $D(\alpha) = \sum_{j=1}^{p}(\alpha_j - \hat{\alpha}_j)^2 + \lambda \#\{\alpha_j \neq 0\}$

has $\hat{\alpha}_k^{(2)} = \hat{\alpha}_k I\{|\hat{\alpha}_k| \geq c\lambda\}$ as a solution.

- Simplicity of calculation : $\hat{\alpha} = (KK^t)^{-1}KY = \frac{1}{n}KY$

$$\hat{\alpha}_j = \frac{1}{n}\sum_{j=1}^{p} K_j(X_i)Y_i$$

δ-Near Identity property

$$M_{np} = \frac{1}{n} K K^t$$

$$(1-\delta)\sum_{j=1}^{p} x_j^2 \leq x^t M_{np} x \leq (1-\delta)\sum_{j=1}^{p} x_j^2$$

$$(1-\delta)\sup_{j=1}^{p}|x_j| \leq \sup_{j=1}^{p}|(M_{np}x)_j| \leq (1-\delta)\sup_{j=1}^{p}|x_j|$$

Estimation procedure

$$
\begin{aligned}
t_n &= \frac{\log n}{n}, \quad \lambda_n = T\sqrt{t_n}, \quad p = [\frac{n}{\log n}]^{\frac{1}{2}} \\
z &= (z_1, \ldots, z_p)^t = (KK^t)^{-1}KY, \\
\tilde{z}_l &= z_l \mathbb{I}\{|z_l| \geq \lambda_n\}
\end{aligned}
$$

$$
\hat{f} = \sum_{l=1}^{p} \tilde{z}_l K_l(\cdot)
$$

Results

1. If f_ρ is sparse i.e. $\exists\, 0 < q < 2,\ \forall\, p,\ \exists(\alpha_1, \ldots, \alpha_p)$
   (a) $\|f_\rho - \sum_{j=1}^{p} \alpha_j K_j\|_\infty \le C p^{-1}$
   (b) $\forall\, \lambda > 0,\ \#\{|\alpha_l| \ge \lambda\} \le C\lambda^{-q}$,

$$\eta_n = [\frac{\log n}{n}]^{\frac{1}{2}-\frac{q}{4}}.$$

$$\rho\{\|f_\rho - \widehat{f}\|_{\hat{\rho}} > (1-\delta)^{-1}\eta\} \le T\{ \begin{array}{ll} e^{-cnp^{-1}\eta^2} \wedge n^{-\gamma}, & \eta \ge D\eta_n, \\ 1, & \eta \le D\eta_n, \end{array}$$

Quasi-optimality

1. Our conditions depend on the family of functions $\{K_j,\ j \geq 1\}$.
2. If the K_j's can be tensor products of wavelet bases for instance then for
$$s := \frac{d}{q} - \frac{d}{2}$$
$f \in B_r^s(L_\infty(\mathbb{R}^d))$ implies the conditions above and $\eta_n = n^{-\frac{s}{2s+d}}$.

Near Identity property : How to make it work ?

$d = 1$

1. Take $\{\phi_k, k \geq 1\}$ be a smooth orthonormal basis of $L_2[0,1](dx)$
2. H with $H(X_i) = \frac{i}{n}$
3. Change the time scale : $K_k = \phi_k(H)$
4. $P_n(k,l) = \frac{1}{n}\sum_{i=1}^{n} K_k(X_i)K_l(X_i) = \frac{1}{n}\sum_{i=1}^{n} \phi_k(\frac{i}{n})\phi_l(\frac{i}{n}) \sim \delta_{kl}$

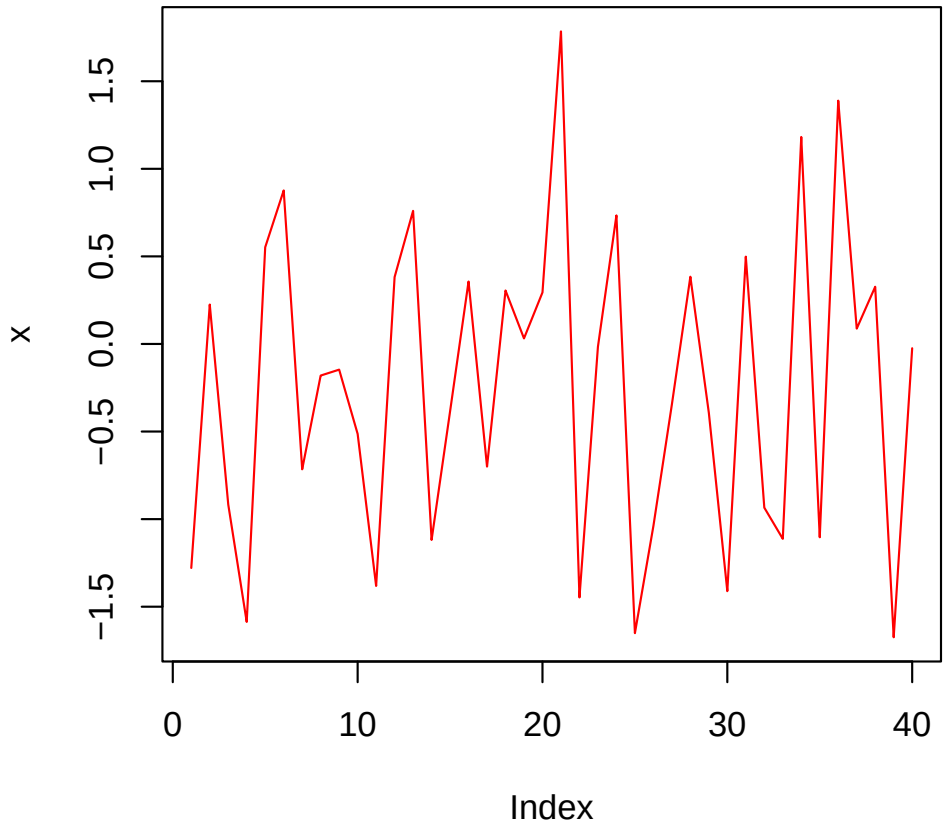


FIG. 1 – Ordering by arrival times

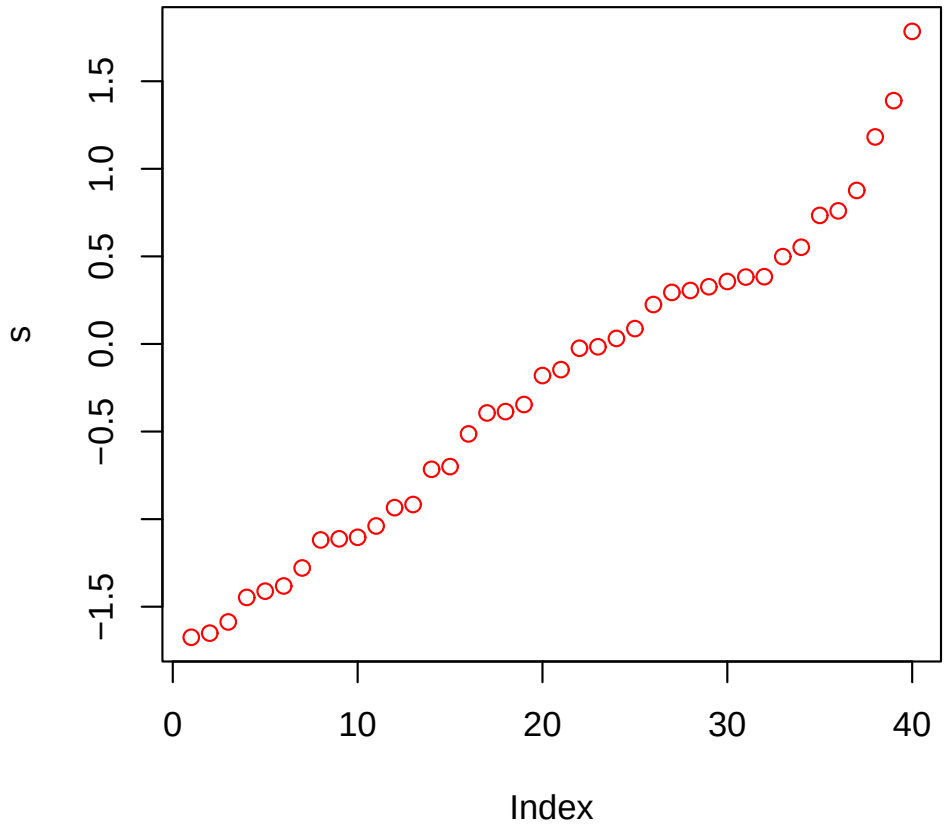


FIG. 2 – Sorting

Choosing H

- Ordering the $X_i's$: $(X_1, \ldots, X_n) \rightarrow (X_{(1)} \leq \ldots \leq X_{(n)})$
- Consider $\hat{G}_n(x) = \frac{1}{n} \sum_{i=1}^{n} I\{X_i \leq x\}$
- $\hat{G}_n(X_{(i)}) = \frac{i}{n}$
- $H = \hat{G}_n$ is stable (i.e. close to $G(x) = \rho(X \leq x)$)
- $\phi_l(\hat{G}_n) \sim \phi_l(G)$

Near Identity property

$$d \geq 2$$

Finding H such that $H(X_i) = (\frac{i_1}{n}, \ldots, \frac{i_d}{n})$, for instance in a 'stable way' is a difficult problem.

Near Identity property

$K_1, \ldots, K_p$ NIP

if there exist a measure μ and cells $C_1, \ldots, C_N$ such that :

$$|\int K_l(x)K_r(x)d\mu(x) - \delta_{lr}| \leq \delta_1(l,r)$$

$$|\frac{1}{N}\sum_{i=1}^{N} K_l(\xi_i)K_r(\xi_i) - \int K_l(x)K_r(x)d\mu(x)| \leq \delta_2(l,r),$$

$$\forall \xi_1 \in C_1, \ldots, \xi_N \in C_N$$

$$\sum_{r=1}^{p} [\delta_1(l,r) + \delta_2(l,r)] \leq \delta$$

Examples : Tensor products of bases, uniform cells

1. $d = 1$, μ Lebesgue measure, on $[0,1]$, $K_1, \ldots, K_p$ is a smooth orthonormal basis (Fourier, wavelet,...) $\delta_1 = 0$, $\delta_2(l,r) = \frac{p}{N}$.
   • $\sum_{r=1}^{p} \delta_2(l,r) \leq \frac{p^2}{N} \leq c\frac{1}{\log N} := \delta$ for $p = [\frac{N}{\log N}]^{\frac{1}{2}}$
   ($p \leq \sqrt{\delta N}$ is enough)

2. $d > 1$, μ Lebesgue measure, on $[0,1]^d$ $K_1, \ldots, K_p$ tensor products of the previous basis. $N = m^d$, $p = \Gamma^d$.
   $\delta_1 = 0$, $\delta_2(l,r) = [\frac{p}{N}]^{\frac{\sup(1, H(l,r))}{d}}$

   $l = (l_1, \ldots, l_d)$, $r = (r_1, \ldots, r_d)$, $\quad H(l,r) = \sum_{i \leq d} I\{l_i \neq r_i\}$

   • $\sum_{r=1}^{p} \delta_2(l,r) \leq [\frac{p^2}{N}]^{\frac{1}{d}} = \frac{c}{[\log N]^{\frac{1}{d}}} := \delta$ for $p \sim [\frac{N}{\log N}]^{\frac{1}{2}}$
   ($p \leq \sqrt{\delta^d N}$ is enough)

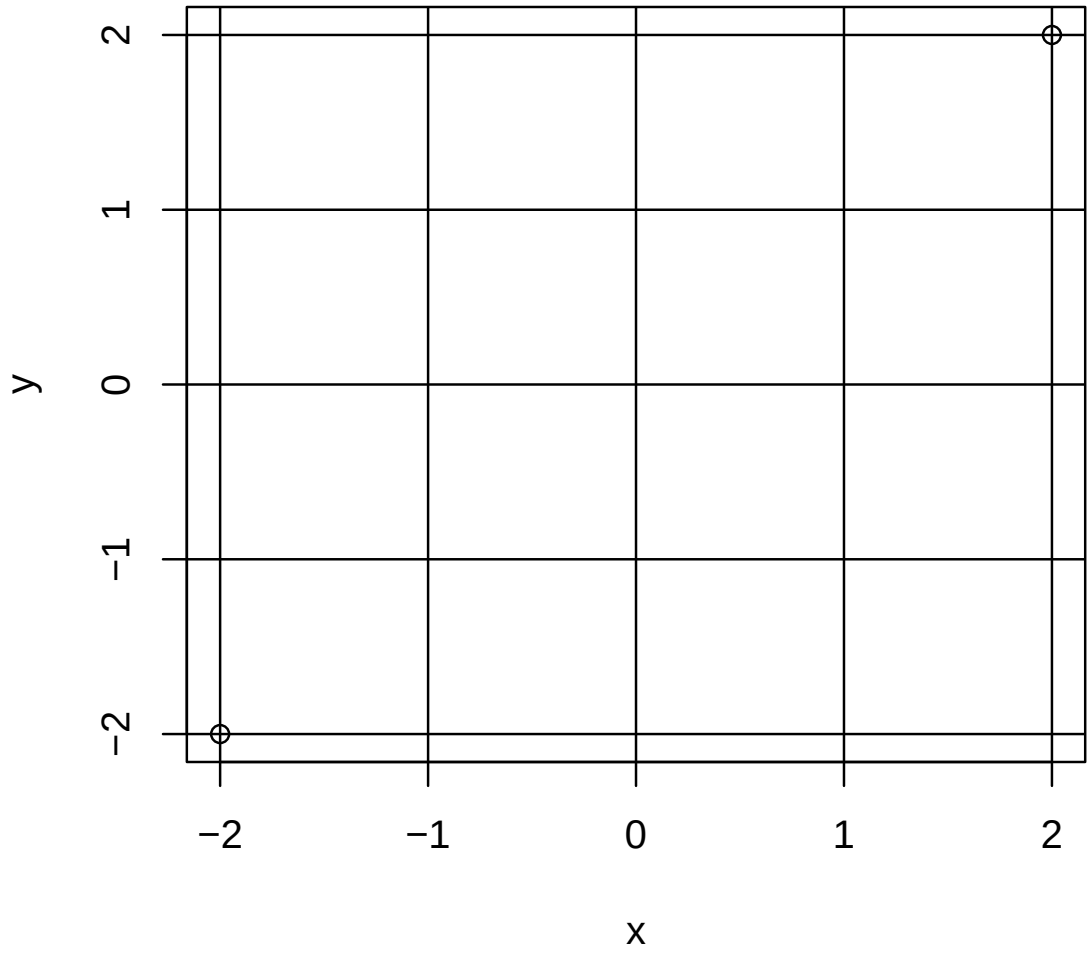
How to relate these assumptions with the near Identity condition ?

What we have here :

$$\frac{1}{N}\sum_{i=1}^{N} K_l(\xi_i)K_r(\xi_i) \quad \xi_1 \in C_1, \ldots, \xi_N \in C_N \text{ 'not too far from' } \delta_{lr}$$

What we want

$$\frac{1}{n}\sum_{i=1}^{n} K_l(X_i)K_r(X_i) \text{ 'not too far from' } \delta_{lr}$$


y
2
1
0
−1
−2
−2
−1
0
1
2
x

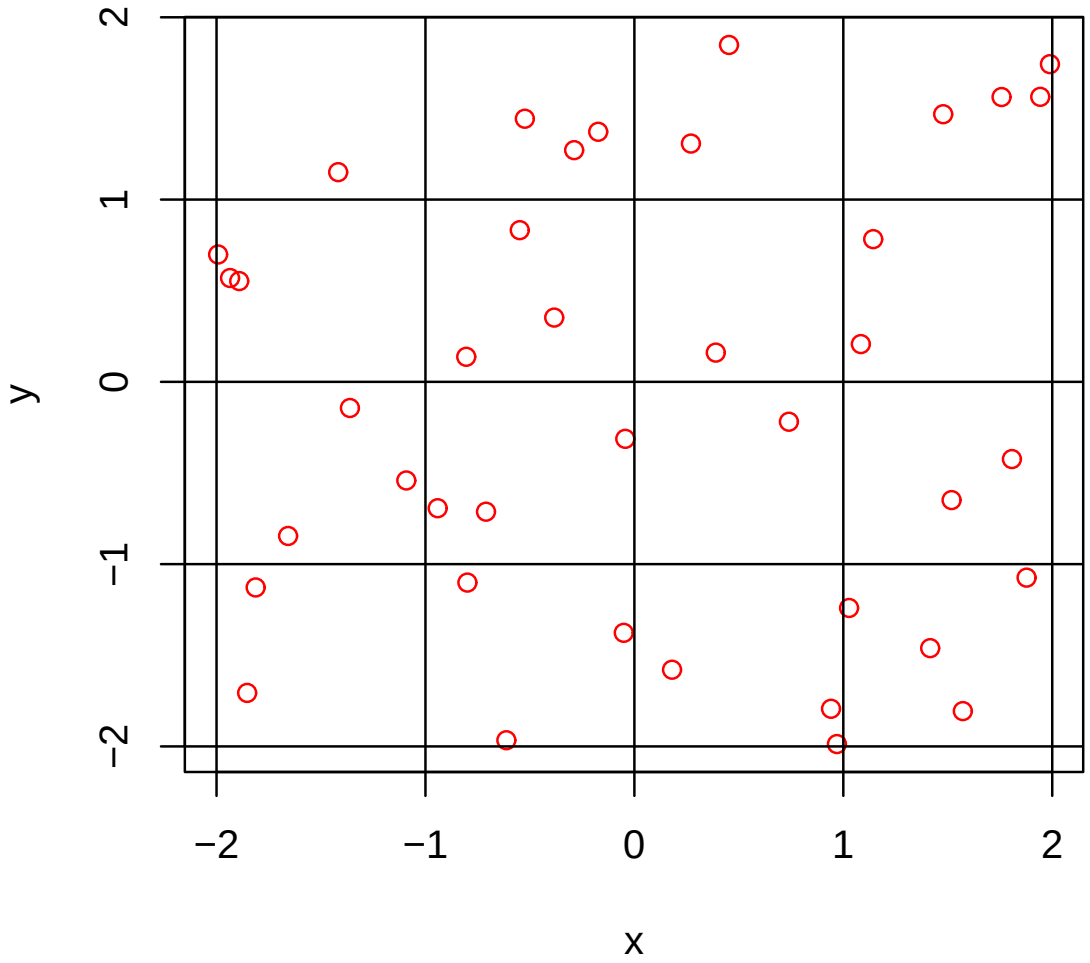
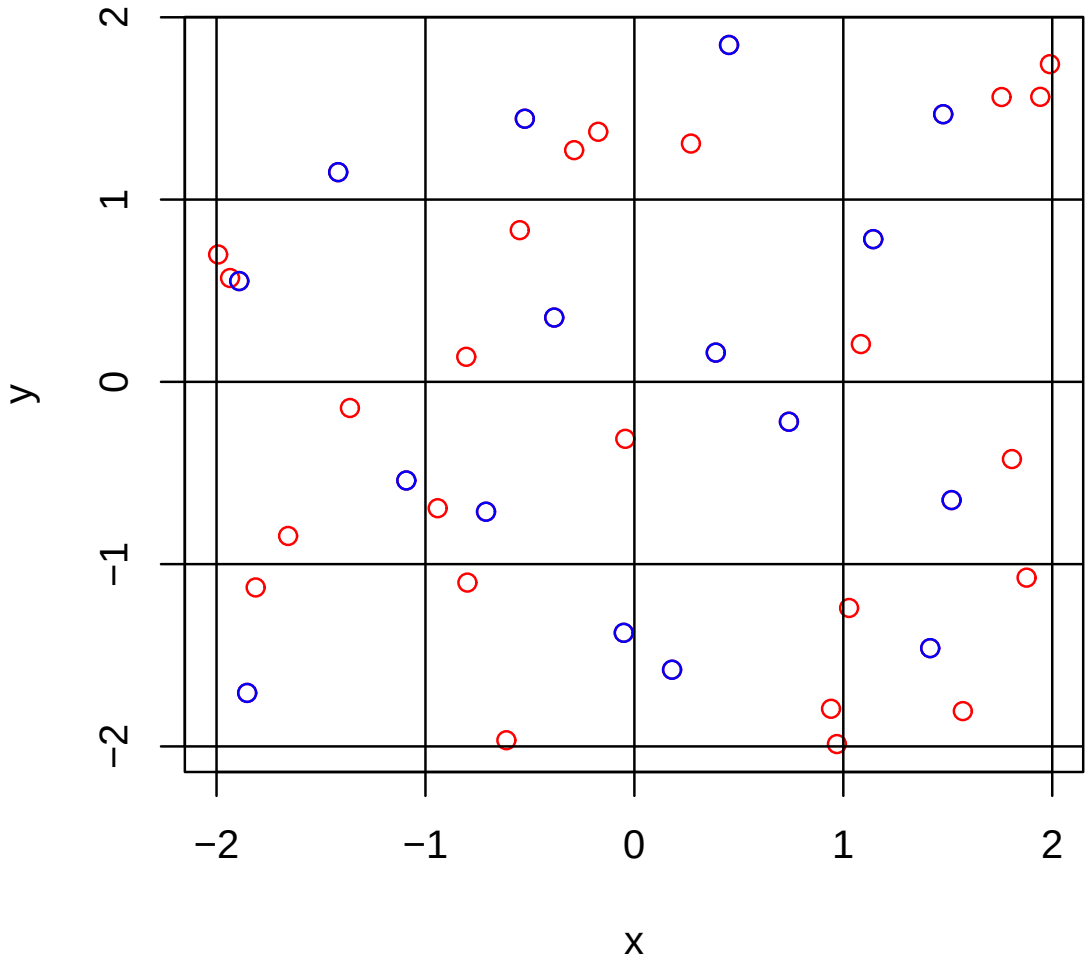
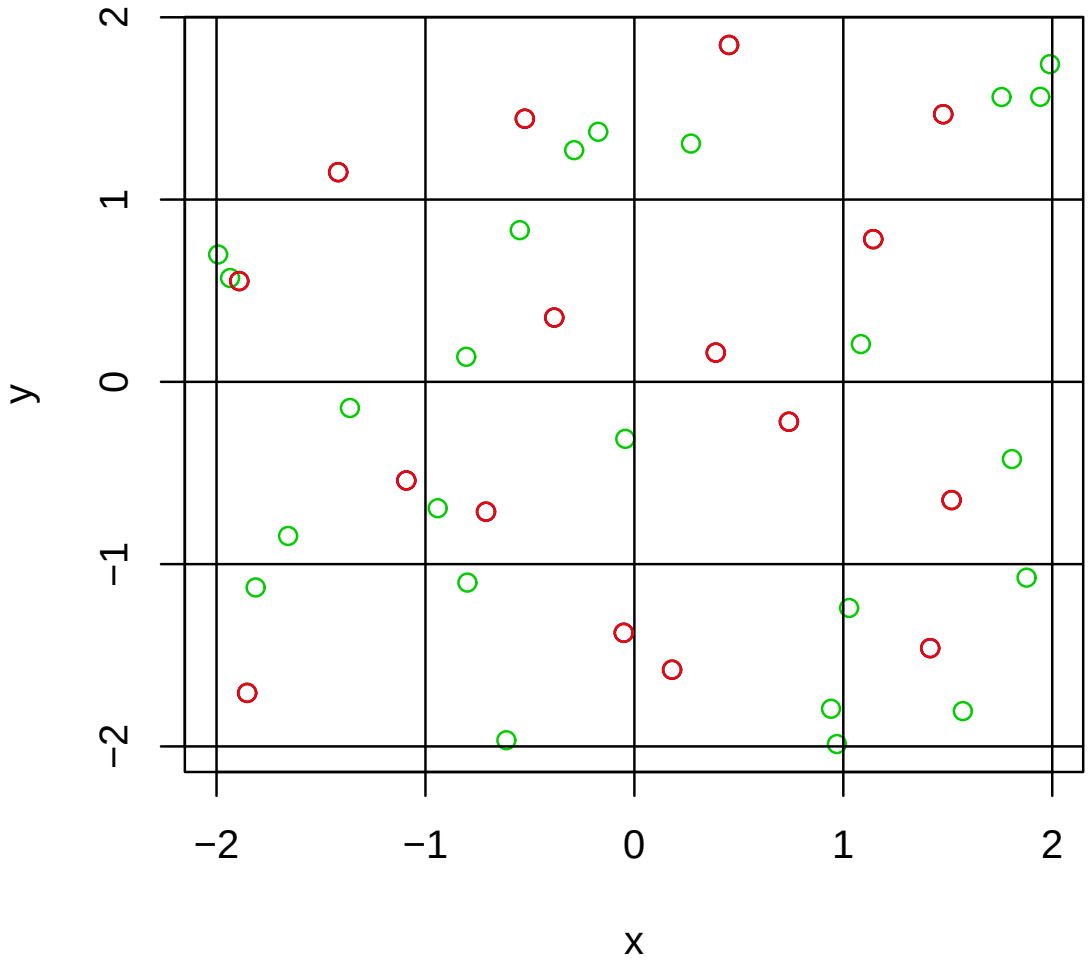


FIG. 3 – Typical situation


2
1
0
−1
−2
y
−2
−1
0
1
2
x


2
1
0
−1
−2
y
−2
−1
0
1
2
x

Procedure

1. We choose cells C_l such that there exist at least one among the observation points X_i's in each cell.
2. We keep only one data point in each cell. (reducing the set of observation :

$$(X_1, Y_1), \ldots, (X_n, Y_n), \quad \rightarrow (X_1, Y_1), \ldots, (X_N, Y_N)$$

3. $n \longrightarrow N$, $\delta \sim \frac{1}{\log N}$ near identity property.
4. If ρ_X is absolutely continuous with respect to μ, with density lower and upper bounded, then $N \sim [\frac{n}{\log n}]$ with overwhelming probability.

Estimation procedure

$$
\begin{aligned}
t_N &= \frac{\log N}{N}, \quad \lambda_N = T\sqrt{t_N}, \quad p = [\frac{N}{\log N}]^{\frac{1}{2}} \\
z &= (z_1, \ldots, z_p)^t = (KK^t)^{-1}KY, \\
\tilde{z}_l &= z_l \mathbb{I}\{|z_l| \geq \lambda_N\}
\end{aligned}
$$

$$
\hat{f} = \sum_{l=1}^{p} \tilde{z}_l K_l(\cdot)
$$

1. If f_ρ is sparse i.e. $\exists\, 0 < q < 2,\ \forall\, p,\ \exists(\alpha_1, \ldots, \alpha_p)$

   (a) $\|f_\rho - \sum_{j=1}^{p} \alpha_j K_j\|_\infty \le C p^{-1}$

   (b) $\forall\, \lambda > 0,\ \#\{|\alpha_l| \ge \lambda\} \le C\lambda^{-q}$,

$$\eta_N = [\frac{\log N}{N}]^{\frac{1}{2}-\frac{q}{4}}.$$

$$\rho\{\|f_\rho - \hat{f}\| > (1-\delta)^{-1}\eta\} \le T\{ \begin{array}{ll} e^{-cNp^{-1}\eta^2} \wedge N^{-\gamma}, & \eta \ge D\eta_N, \\ 1, & \eta \le D\eta_N, \end{array}$$

$$\|f_\rho - \hat{f}\| = \|f_\rho - \hat{f}\|_{\hat{\rho}}$$

or (if $\rho_X << \mu$)

$$\|f_\rho - \hat{f}\| = \|f_\rho - \hat{f}\|_{\rho_X}$$

What to do with the remaining data ?

Empirical Bayes

(see *Johnstone and Silverman*)

- Hard thresholding (in practice) is not the best choice.
- Better choices are obtained using rules issued from Bayesian procedures using a prior of the form :

$$\omega\delta_{\{0\}} + (1-\omega)g$$

where g is a Gaussian (with large variance) or a Laplace distribution.

With the associated procedure

$$z_i^* = z_i \mathbb{I}\{|z_i| \geq t(\omega)\}$$

• the parameter ω in the a priori distribution can again be 'learned' using the observed data if the sample is divided into two pieces -one used to learn this parameter, the other one to operate the bayesian procedure itself, with the learned parameter $\hat{\omega}$,

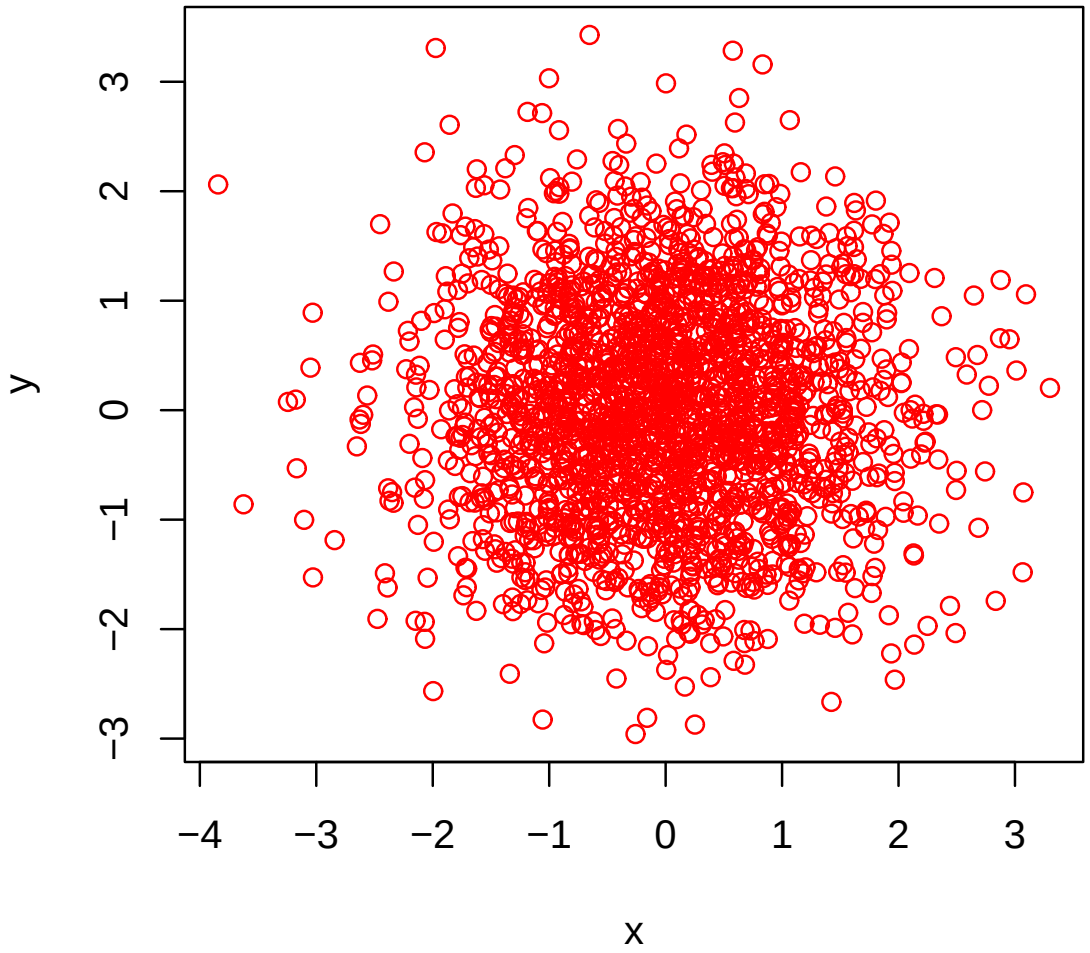
$$z_l^* = z_l \mathbb{I}\{|z_l| \geq t(\hat{\omega})\}$$

• In our context, the remaining data, naturally serve to choose the hyper parameter of the a priori distribution.

Condition under which the results are still valid

Learning $\rightarrow$ Regression :

$$Y_i = f_\rho(X_i) + \varepsilon_i, \quad X_i \perp\!\!\!\perp \varepsilon_i$$


y
3
2
1
0
−1
−2
−3
−4
−3
−2
−1
0
1
2
3
x

Examples : Wavelet frames on the sphere, Voronoi cells

Uniform cells can be replaced by Voronoi cells contructed on an N-net on the sphere (or on the ball), with an adapted basis (spherical harmonics, in the case of the sphere).