Nonparametric regression function estimation with non compactly supported bases.

Fabienne Comte
MAP5, Université Paris Descartes

joint with: **V. Genon-Catalot** (MAP5, Université Paris Descartes)

Freiburg, Feb.27-Mar 2 2018, 13th Probability and Statistics Days

The problem

Standard regression model

$$Y_i = b(X_i) + \varepsilon_i, \quad i = 1, \ldots, n \tag{1.1}$$

with

$$(\varepsilon_i)_{1\le i\le n} \text{ i.i.d. centered with variance } \sigma_\varepsilon^2 \quad \textbf{(noise)}$$

and

$$(X_i)_{1\le i\le n} \text{ i.i.d. with density } f, \quad \textbf{(explanatory)}$$

and

$$(X_i)_{1\le i\le n} \text{ independent of } (\varepsilon_i)_{1\le i\le n}$$

Observations $(Y_i, X_i)_{1\le i\le n}$.

Aim: **Nonparametric estimation of $b(.)$**

Projection estimator

Let $(\varphi_j)_{0\leq j\leq m-1}$ an orthonormal basis in $\mathbb{L}^2(A, dx)$, $A \subset \mathbb{R}$,

$$\langle \varphi_j, \varphi_k \rangle = \int_A \varphi_j(x)\varphi_k(x)dx = \delta_{j,k}.$$

Look for

$$\widehat{b}_m = \sum_{j=0}^{m-1} \widehat{a}_j \varphi_j$$

where

$(\widehat{a}_j)_{0\leq j\leq m-1}$ are computed **from the observations** $(Y_i, X_i)_{1\leq i\leq n}$.

Quotient estimators

Nadaraya-Watson or quotient estimators are not exactly of this type, $r = bf$, principle

$$\widetilde{b}_{m,m'} = \frac{\widehat{r}_m}{\widehat{f}_{m'}}$$

$$\widehat{r}_m = \sum_{j=0}^{m-1} \widehat{c}_j \varphi_j,\ \widehat{c}_j = \frac{1}{n}\sum_{i=1}^{n} Y_i \varphi_j(X_i), \quad \widehat{f}_{m'} = \sum_{j=0}^{m'-1} \widehat{d}_j \varphi_j,\ \widehat{d}_j = \frac{1}{n}\sum_{i=1}^{n} \varphi_j(X_i)$$

Quotient can be performed coefficient by coefficient:

$$\widetilde{\widetilde{b}}_{m,m'} = \sum_{j=0}^{m-1} \widetilde{a}_j \varphi_j,\ \widetilde{a}_j = \frac{1}{n}\sum_{i=1}^{n} \frac{Y_i \varphi_j(X_i)}{\widehat{f}_{m'}(X_i)}$$

Least squares estimator

Let

$$S_m = \mathrm{span}(\varphi_0, \dots, \varphi_{m-1}) = \left\{ t = \sum_{j=0}^{m-1} a_j \varphi_j, a_j \in \mathbb{R} \right\},$$

and consider the **least squares estimator**

$$\widehat{b}_m = \arg\min_{t \in S_m} \gamma_n(t), \quad \gamma_n(t) = \mathbf{\frac{1}{n} \sum_{i=1}^{n} [Y_i - t(X_i)]^2}.$$

Works as if $a_0, \dots, a_{m-1}$ parameters in the linear model

$$Y_i \approx a_0 \varphi_0(X_i) + \cdots + a_{m-1} \varphi_{m-1}(X_i) + \varepsilon_i$$

for which you compute the least squares estimator with classical formula.

Formula of the LS estimator

$$\widehat{b}_m = \sum_{j=0}^{m-1} \widehat{a}_j \varphi_j, \quad \widehat{\vec{a}}_{(m)} := \begin{pmatrix} \widehat{a}_0 \\ \vdots \\ \widehat{a}_{m-1} \end{pmatrix} = \left({}^t\widehat{\Phi}_m \widehat{\Phi}_m \right)^{-1} \widehat{\Phi}_m \vec{\mathscr{Y}}, \quad (1.2)$$

where

$$\vec{\mathscr{Y}} = \begin{pmatrix} Y_1 \\ \vdots \\ Y_n \end{pmatrix}, \quad \widehat{\Phi}_m = (\varphi_j(X_i))_{1 \le i \le n, 0 \le j \le m-1} \quad n \times m \text{ matrix}$$

provided that

${}^t\widehat{\Phi}_m \widehat{\Phi}_m$ **is invertible**.

Existing results

Questions are :

- Risk of the estimator for **fixed** m
- Selection of adequate model m from the data, $\widehat{m}$
- Risk of the adaptive estimator, $\widehat{b}_{\widehat{m}}$

Baraud (2000) **fixed design**, Baraud (2002) **random design** studied these questions but for compactly supported bases, assumption

$$\forall x \in A, \quad 0 < f_{\min} \leq f(x) \leq f_{\max} < +\infty.$$

Three norms in the problem:

- Empirical norm $\|t\|_n^2 = \frac{1}{n}\sum_{i=1}^{n} t^2(X_i)$,
- $\mathbb{L}^2(A, f(x)dx)$-norm, $\|t\|_f^2 = \int_A t^2(x)f(x)dx = \mathbb{E}[\|t\|_n^2]$ for t with support A,
- $\mathbb{L}^2(A, dx)$-norm, $\|t\|^2 = \int_A t^2(x)dx$.

Associated models

- Extension to **dependent models** : $X_{i+1} = b(X_i) + \varepsilon_{i+1}$, Baraud et al. (2001).
- Extension to **drift estimation** : $dX_t = b(X_t)dt + \sigma(X_t)dW_t$, Comte, Genon-Catalot and Rozenholc (2007)
- Extension of regression strategy to **other models**
  - **Survival function** estimation in presence of interval censoring
  - **Hazard rate** estimation in presence of right censoring,
  - Conditional density estimation ...

Non compactly supported bases

What happens if the basis has support $\mathbb{R}$ or $\mathbb{R}^+$?

Non compact support, what for?

- Laguerre and Hermite basis are convenient, with nice properties,
- Laguerre basis is natural if $X_i \geq 0$, Hermite basis is natural for diffusion models,
- For extension to indirect problem, where support is unknown,
- For extension to error-in-variables models, as compactly supported bases lead to non integrable Fourier transforms

Examples of compactly supported bases, $A = [0,1]$

Classical compactly supported bases are:

- Histograms $pp_j^{(0)}(x) = \sqrt{m}\mathbf{1}_{[j/m,(j+1)/m[}(x)$, for $j = 0,\dots,m-1$; piecewise polynomials with degree r;
- Compactly supported wavelets;
- Trigonometric basis with odd dimension m, $t_0(x) = \mathbf{1}_{[0,1]}(x)$ and $t_{2j-1}(x) = \sqrt{2}\cos(2\pi jx)\mathbf{1}_{[0,1]}(x)$, and $t_{2j}(x) = \sqrt{2}\sin(2\pi jx)\mathbf{1}_{[0,1]}(x)$ for $j = 1,\dots,(m-1)/2$.

All these collections satisfy $\|\sum_{j=0}^{m-1} \varphi_j^2\|_\infty \leq c_\varphi^2 m$

($c_\varphi^2 = 1$ for histograms and trigonometric basis, $c_\varphi^2 = r+1$ for p.p.)
Nested (in general or for $m = 2^k$ for increasing values of k).

Laguerre basis, $A = \mathbb{R}^+$.

Laguerre polynomials (L_j) and Laguerre functions (ℓ_j) are given by

$$L_j(x) = \sum_{k=0}^{j} (-1)^k \binom{j}{k} \frac{x^k}{k!}, \qquad \ell_j(x) = \sqrt{2} L_j(2x) e^{-x} \mathbf{1}_{x \geq 0}, \quad j \geq 0.$$

The collection $(\ell_j)_{j \geq 0}$ constitutes a complete orthonormal system on $\mathbb{L}^2(\mathbb{R}^+)$, and is such that (see Abramowitz and Stegun (1964)):

$$\forall j \geq 0,\ \forall x \in \mathbb{R}^+,\ |\ell_j(x)| \leq \sqrt{2}. \tag{2.3}$$

$(S_m = \mathrm{span}\{\ell_0, \dots, {}_{m-1}\})_m$ is nested,

(2.3) implies that $\| \sum_{j=0}^{m-1} \ell_j^2 \|_\infty \leq c_\varphi^2 m$ with $c_\varphi^2 = 2$.

Hermite basis, $A = \mathbb{R}$.

Hermite polynomials and Hermite functions of order j for $j \geq 0$:

$$H_j(x) = (-1)^j e^{x^2} \frac{d^j}{dx^j}(e^{-x^2}), \quad h_j(x) = \mathfrak{c}_j H_j(x) e^{-x^2/2}, \quad \mathfrak{c}_j = \left(2^j j! \sqrt{\pi}\right)^{-1/2}$$

The sequence $(h_j, j \geq 0)$ is an orthonormal basis of $\mathbb{L}^2(\mathbb{R}, dx)$.
We have

$$\|h_j\|_\infty \leq \Phi_0, \qquad \Phi_0 \simeq 1,086435/\pi^{1/4} \simeq 0.8160, \tag{2.4}$$

so that the Hermite basis satisfies $\|\sum_{j=0}^{m-1} h_j^2\|_\infty \leq c_\varphi^2 m$ with $c_\varphi^2 = \Phi_0^2$.

The collection of models is nested.

No support condition for the first basic result

Proposition

Let $(X_i, Y_i)_{1\leq i\leq n}$ be observations drawn from model (1.1) and denote by $b_A = b\mathbf{1}_A$. Assume that $b_A \in \mathbb{L}^2(A, f(x)dx)$ and that $\widehat{\Psi}_m$ is invertible. Consider the least squares estimator $\widehat{b}_m$ of b, given by (1.2). Then

$$\mathbb{E}\left[\|\widehat{b}_m - b_A\|_n^2\right] \leq \inf_{t\in S_m}\left[\int (b_A - t)^2(x)f(x)dx\right] + \sigma_\varepsilon^2 \frac{m}{n}, \tag{3.5}$$

where f denotes the common density of the X_i's.

Proof of Proposition 1. Let Π_m be the orthogonal projection (for the scalar product of $\mathbb{R}^n$) on the sub-space $\left\{\left(t(X_1),\ldots,t(X_n)\right)', t\in S_m\right\}$ of $\mathbb{R}^n$. Then

$$\begin{aligned} \|\widehat{b}_m - b_A\|_n^2 &= \|\Pi_m b - b_A\|_n^2 + \|\widehat{b}_m - \Pi_m b\|_n^2 \\ &= \inf_{t\in S_m} \|t - b_A\|_n^2 + \|\widehat{b}_m - \Pi_m b\|_n^2 \end{aligned}$$

Now we have:

$$\mathbb{E}\left[\|\widehat{b}_m - \Pi_m b\|_n^2\right] = \sigma_\varepsilon^2 \frac{m}{n}. \tag{3.6}$$

By taking expectation of (3.6),

$$\mathbb{E}\left[\|\widehat{b}_m - b_A\|_n^2\right] \leq \inf_{t\in S_m} \|t - b_A\|_f^2 + \mathbb{E}\left[\|\widehat{b}_m - \Pi_m b\|_n^2\right]. \tag{3.7}$$

Plug (3.6) in (3.7), to obtain Proposition 1.

Note that:

- The result is general and holds for **any basis support**,
- The variance term is **exactly** equal to $\sigma_\varepsilon^2 m/n$, and this does not depend on the basis.

The bias tends to zero when m grows to infinity.

Lemma

If $(\varphi_j)_{j\geq 0}$ is an orthonormal basis of $\mathbb{L}^2(A, dx)$ such that, for all $j \geq 0$, $\int \varphi_j^2(x) f(x) dx < +\infty$, f is bounded on A and $\forall x \in A$, $f(x) > 0$. Then $\inf_{t\in S_m} \|b_A - t\|_f^2$ tends to 0 when m tends to infinity.

The bias is **getting small** when m grows, but the variance **increases**.
$\Rightarrow$ a compromise has to be found, by relevant choice of m.

Comparison with density estimation (1)

Why is it important to **notice the equality:**

$$\mathbb{E}(\|\widehat{b}_m - b_A\|_n^2) = \mathbb{E}(\inf_{t\in S_m} \|t - b_A\|_n^2) + \sigma_\varepsilon^2 \frac{m}{n}.$$

By comparison with **density estimation** where

$$\widehat{f}_m = \sum_{j=0}^{m-1} \widehat{c}_j \varphi_j \text{ with } \widehat{c}_j = \frac{1}{n}\sum_{i=1}^{n} \varphi_j(X_i),$$

satisfies

$$\begin{aligned}\mathbb{E}(\|\widehat{f}_m - f\|^2) &= \|f - f_m\|^2 + \frac{\sum_{j=0}^{m-1} \mathbb{E}\left[\varphi_j^2(X_1)\right]}{n} - \frac{\|f_m\|^2}{n} \\ &\leq \inf_{t\in S_m} \|f - t\|^2 + \frac{\sum_{j=0}^{m-1} \mathbb{E}\left[\varphi_j^2(X_1)\right]}{n}.\end{aligned}$$

Comparison with density estimation (2)

For all the bases,

$$\|\sum_{j=0}^{m-1} \varphi_j^2\|_\infty \leq c_\varphi^2 m \quad \Rightarrow \quad \sum_{j=0}^{m-1} \mathbb{E}\left[\varphi_j^2(X_1)\right] \leq c_\varphi^2 m$$

$\sum_{j=0}^{m-1} \varphi_j^2 = m$, for histograms and trigonometric polynomials with odd dimension,

$$\Rightarrow \text{ the bound can be } \textbf{sharp}.$$

For the Laguerre basis: $\sum_{j=0}^{m-1} \varphi_j^2(0) = 2m$

However, it holds for **Hermite and Laguerre** bases that

$$\sum_{j=0}^{m-1} \mathbb{E}\left[\varphi_j^2(X_1)\right] \leq c_\varphi^2 \sqrt{\mathbf{m}}.$$

Questions to solve

- Bound an integrated $\mathbb{L}^2(A, f(x)dx)$-risk instead of the empirical risk
- Model selection and use of the bases properties.

Starts with a control of $\|\widehat{\Psi}_m - \Psi_m\|_{\rm op}$

$$\text{where} \qquad \widehat{\Psi}_m = \frac{1}{n}\,{}^t\widehat{\Phi}_m \widehat{\Phi}_m = \left(\frac{1}{n}\sum_{i=1}^{n} \varphi_j(X_i)\varphi_k(X_i)\right)_{0\le j,k\le m-1},$$

$$\Psi_m := \mathbb{E}\left(\widehat{\Psi}_m\right) = (\langle \varphi_j, \varphi_k\rangle_f)_{0\le j,k\le m-1} = \left(\int_A \varphi_j(x)\varphi_k(x)f(x)dx\right)_{0\le j,k\le m-1}$$

and, for M a matrix, $\|M\|_{\rm op}$ is the operator norm, $\|M\|_{\rm op}^2 = \lambda_{\max}(MM')$.
If M is symmetric positive definite, $\|M\|_{\rm op} = \lambda_{\max}(M)$.

Deviation result

Key tool: **matricial Bernstein deviation inequality** from Tropp (2015).

Proposition

Let $X_1, \dots, X_n$ be i.i.d. with common density f such that $\|f\|_\infty < \infty$. Assume that the $(\varphi_j)_{0 \le j \le m-1}$ are such that $\|\sum_{j=0}^{m-1} \varphi_j^2\|_\infty \le c_\varphi^2 m$. Then for all $u > 0$

$$\mathbb{P}\left[\|\Psi_m - \widehat{\Psi}_m\|_{\mathrm{op}} \ge u\right] \le 2m \exp\left(-\frac{n u^2/2}{c_\varphi^2 m (\|f\|_\infty + u/3)}\right).$$

The result encompasses all possible classical bases, whether compactly supported or not.

Selection of m

Collection of nested spaces S_m: $S_m \subset S_{m'}$ for $m \leq m'$ such that, for each m, the basis $(\varphi_0, \dots, \varphi_{m-1})$ of S_m satisfies

$$\| \sum_{j=0}^{m-1} \varphi_j^2 \|_\infty \leq c_\varphi^2 m \quad \text{for} \quad c_\varphi^2 > 0 \quad \text{a constant.}$$

$\widehat{\mathscr{M}}_n$ is a **random** collection of models defined by

$$\widehat{\mathscr{M}}_n = \left\{ m \in \{1, 2, \dots, n\}, m(\|\widehat{\Psi}_m^{-1}\|_{\mathrm{op}}^2 \vee 1) \leq 4\mathfrak{c} \frac{n}{\log(n)} \right\}, \qquad (4.8)$$

$$\text{with } \mathfrak{c} = \left(6 \wedge \frac{1}{\|f\|_\infty}\right) \frac{1}{48\, c_\varphi^2}.$$

Theoretical counterpart

$$\mathscr{M}_n = \left\{ m \in \{1, 2, \dots, n\}, m\left(\|\Psi_m^{-1}\|_{\mathrm{op}}^2 \vee m\right) \leq \mathfrak{c} \frac{n}{\log(n)} \right\}, \qquad (4.9)$$

Selecting

$$\widehat{m} = \arg\min_{m \in \widehat{\mathscr{M}}_n} \left\{ -\|\widehat{b}_m\|_n^2 + \kappa \sigma_\varepsilon^2 \frac{m}{n} \right\}$$

follows from standard ideas:

- **Squared bias term** $\|b_A - b_m^f\|_f^2 = \|b_A\|_f^2 - \|b_m^f\|_f^2$ where b_m^f is the $\mathbb{L}^2(A, f(x)dx)$-orthogonal projection of b on S_m.
  - $\|b_A\|_f^2$ unknown but does not depend on m;
  - $\|b_m^f\|_f^2 = \mathbb{E}[\|b_m^f\|_n^2]$.

  $\Rightarrow -\|\widehat{b}_m\|_n^2$ approximates the squared bias, up to an additive constant,
- $\sigma_\varepsilon^2 m/n$ has the **variance order**.

The procedure aims at performing an automatic **bias-variance tradeoff**.

Theorem

Let $(X_i, Y_i)_{1\leq i\leq n}$ be observations from model (1.1). Assume that:

- *for each m, the basis of S_m satisfies $\|\sum_{j=0}^{m-1}\varphi_j^2\|_\infty \leq c_\varphi^2 m$ for $c_\varphi^2 > 0$ a constant.*
- $\|f\|_\infty < +\infty$,
- $\mathbb{E}(\varepsilon_1^6) < +\infty$ *and* $\mathbb{E}\left[b^4(X_1)\right] < +\infty$.

Then, there exists a numerical constant κ_0 such that for $\kappa \geq \kappa_0$, we have

$$\mathbb{E}\left[\|\widehat{b}_{\widehat{m}} - b_A\|_f^2\right] \leq C \inf_{m\in\mathscr{M}_n}\left(\inf_{t\in S_m}\|b_A - t\|_f^2 + \sigma_\varepsilon^2\frac{m}{n}\right) + \frac{C'}{n}$$

where C is a numerical constant and C' is a constant depending on f, b, σ_ε.

What is new here?

- General result with **no support constraint**
- Standard moment conditions
- **Random** collection of models $\widehat{\mathscr{M}}_n$

Remark: $\widehat{\mathscr{M}}_n \Rightarrow$ Limitation of the models considered in the collection for selection, corresponds to a kind of **cutoff for inversion of** $\widehat{\Psi}_m$.
$\mathscr{M}_n$ limitation in **reachable values of** m.

Remains to be done:

- Estimate σ_ε^2 in the penalty,
- Estimate $\|f\|_\infty$ in the collection of models
- Calibration of κ.

Application to compactly supported bases

If A compact, one can assume

$$\forall x \in A, 0 < f_0 \leq f(x) \leq f_1 < +\infty$$

- $b \in \mathbb{L}^2(A, dx)$ can be assumed (not so strong as A compact) and

$$f \leq f_1 \Rightarrow \|t - b_A\|_f^2 \leq f_1 \|t - b_A\|^2.$$

- We can prove

$$f \geq f_0 > 0 \Rightarrow \|\Psi_m^{-1}\|_{\text{op}} \leq 1/f_0.$$

So we can take

$$\mathscr{M}_n = \{m \in \{1, \dots, n\}, m \leq c'(f_0)n/\log(n)\} = \widehat{\mathscr{M}_{\mathbf{n}}}.$$

Weak constraint on $m \in \mathscr{M}_n$ and standard rates on Besov spaces $(n^{-2\alpha/(2\alpha+1)})$.

Application to non compact A

$A = \mathbb{R}^+$ and Laguerre basis and $A = \mathbb{R}$ and Hermite basis.
We still have

Lemma

For all $m \in \mathbb{N}$, Ψ_m is invertible, and for all $m \leq n$, $\widehat{\Psi}_m$ is invertible.

but:

Proposition

Assume that $\inf_{a \leq x \leq b} f(x) > 0$ for some interval $[a, b]$ in the Hermite case and with $0 < a < b$ in the Laguerre case. Then there exists a constant $c^\star$ such that, for all m,

$$\|\Psi_m^{-1}\|_{\mathrm{op}}^2 \geq c^\star m. \tag{6.10}$$

Proposition

Consider the Laguerre or the Hermite basis. Assume that

- $f(x) \geq c/(1+x)^k$ *for* $x \geq 0$ *in the Laguerre case;*
- *or* $f(x) \geq c/(1+x^2)^k$ *for* $x \in \mathbb{R}$ *in the Hermite case.*

Then for m large enough, $\|\Psi_m^{-1}\|_{\mathrm{op}} \leq Cm^k$.

Simulations show that $\|\Psi_m^{-1}\|_{\mathrm{op}}$ **grows very fast** and $\widehat{\mathscr{M}_n}$ is **small**.

If f is as in the Proposition, then

$$m \in \mathscr{M}_n \Rightarrow m^{2k+1} \lesssim n/\log(n).$$

Consider $A = \mathbb{R}^+$, Laguerre basis and **Sobolev-Laguerre space**:

$$b_A \in W^s(R) = \{h \in \mathbb{L}^2(\mathbb{R}^+, dx), \sum_{j\geq 0} j^s a_j^2(h) \leq R\},$$

with $a_j(h) = \langle h, \ell_j \rangle$, and that $f \leq f_1$. Then

$$\inf_{t \in S_m} \|b_A - t\|_f^2 \lesssim m^{-s}.$$

Compromise: squared bias m^{-s} − variance m/n: $\mathbf{m}_{\text{opt}} = n^{1/(s+1)}$. Resulting rate $n^{-s/(s+1)}$ reached only if

$$\mathbf{m}_{\text{opt}}^{2k+1} \leq n/\log(n) \quad \text{i.e. if} \quad s > 2k.$$

Remark. If b_A is a combination of Γ functions, then rate $\log(n)/n$ can be reached by the adaptive estimator.

About the order of $\|\Psi_m^{-1}\|_{\text{op}}$

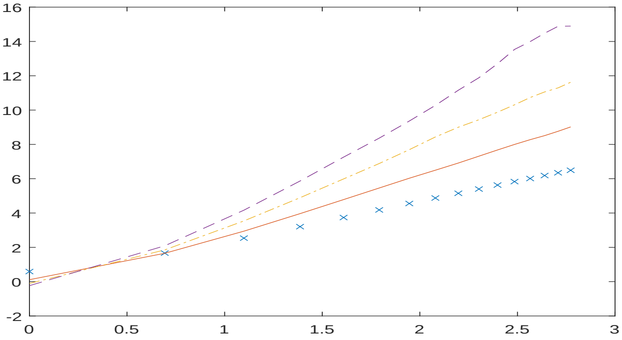


Figure: Laguerre basis. $\log(m) \mapsto \log(\|\Psi_m^{-1}\|_{\text{op}})$, density of X given by $f_k(x) = (k-1)/(1+x)^k \mathbf{1}_{x \geq 0}$, $k = 2$ (blue x marks), $k = 3$ (red solid), $k = 4$ (yellow dashdots) and $k = 5$ (purple dashed). **Estimated slope regression coefficients: 2.09 - 3.16 - 4.21 - 5.58**

Extension to dependent models

- **Autoregressive model.**

$$X_{i+1} = b(X_i) + \varepsilon_{i+1}, \quad (\varepsilon_i)_{i\geq 0} \text{ i.i.d., centered with variance } \sigma_\varepsilon^2,$$

with X_0 is independent of the sequence $(\varepsilon_i)_{i\geq 0}$.

$$\widehat{b}_m = \arg\min_{t\in S_m} \bar{\gamma}_n(t), \quad \text{with } \bar{\gamma}_n(t) = \frac{1}{n}\sum_{i=1}^{n} t^2(X_i) - 2X_{i+1}t(X_i).$$

- **Diffusion model.**

Observations with sampling interval Δ, $(X_{i\Delta})_{1\leq i\leq n}$, from the diffusion process

$$dX_t = b(X_t)dt + \sigma(X_t)dW_t, \quad X_0 \sim \eta.$$

Set

$$Y_{i\Delta} = \frac{X_{(i+1)\Delta} - X_{i\Delta}}{\Delta}, \quad Z_{i\Delta} = \frac{1}{\Delta}\int_{i\Delta}^{(i+1)\Delta} \sigma(X_s)dW_s$$

$$\text{and} \quad R_{i\Delta} = \frac{1}{\Delta}\int_{i\Delta}^{(i+1)\Delta} [b(X_s) - b(X_{i\Delta})]ds.$$

The regression equation writes:

$$\mathbf{Y}_{i\Delta} = b(X_{i\Delta}) + Z_{i\Delta} + R_{i\Delta},$$

where • $Z_{i\Delta}$ plays the role of the **noise**,

• $R_{i\Delta}$ is an additional **residual term** to take into account.

$$\widehat{b}_m = \arg\min_{t \in S_m} \left[\frac{1}{n}\sum_{i=1}^{n} t^2(X_{i\Delta}) - 2Y_{i\Delta}t(X_{i\Delta}) \right]$$

In both models

- Only **one** process is observed X_i or $X_{i\Delta}$, $i = 0, 1, \ldots, n$
- Under conditions on $b(\cdot)$ or on $b(\cdot)$ and $\sigma(\cdot)$ and on the initial condition,
    - There is a **strictly stationary** solution (with stationary density denoted by f)
    - which is **geometrically** β-mixing.

▲ Model selection can be done similarly

▲ The main theorem can be extended to both contexts,

▲ The matricial deviation inequality can be extended in the mixing framework.

Proposition (Mixing matrix deviation inequality)

Let $(X_i)_i$ be a strictly stationary and geometrically β-mixing process:

$$\beta_k \leq ce^{-\theta k} \text{ for some constants } c > 0, \theta > 0,$$

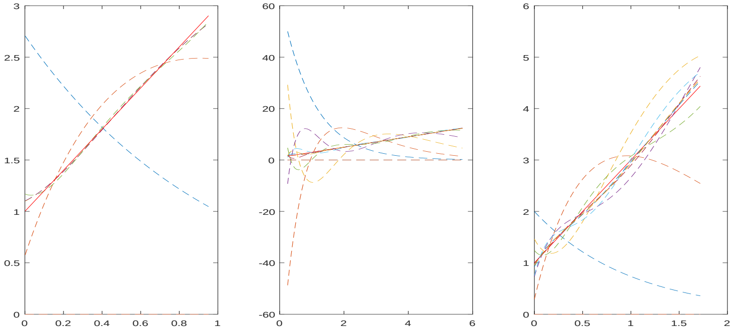
with marginal density f and assume that

- $\mathbb{E}(X_1^{8/3}) < +\infty$ *in the* **Hermite basis**,
- $\mathbb{E}(1/X_1^2) < +\infty$ *in the* **Laguerre basis**.

Then for all $u > 0$

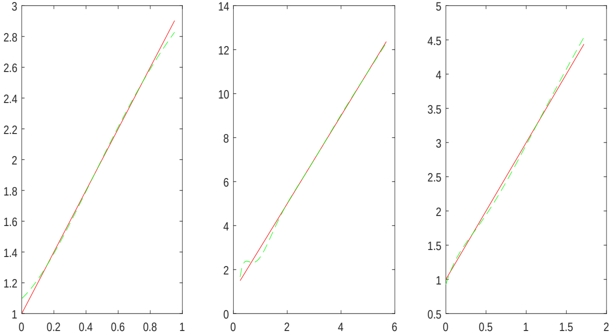
$$\mathbb{P}\left[\|\Psi_m - \widehat{\Psi}_m\|_{\text{op}} \geq u\right] \leq 2m\exp\left(-\frac{nu^2/2}{\mathbf{a}m(1+\log(n)u)}\right) + \frac{c}{\mathbf{n}^4},$$

where **a** *is a constant depending on the β_k and the moments.*



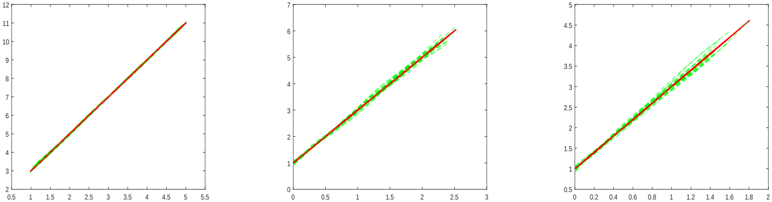
$X \sim \mathscr{U}([0,1])$ $X \sim \mathscr{N}(4,1)$ $X \sim f_3$

Figure: Beam of the proposals $\widehat{b}_m$ for $m = 1$ to $m_{\max}$ in the Laguerre basis. Function $b(x) = 2x + 1$, $n = 1000$, density $f_k(x) = (k-1)/(1+x)^k \mathbf{1}_{x \geq 0}$.



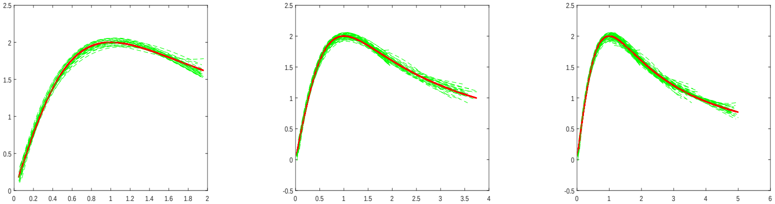
$X \sim \mathscr{U}([0,1])$ $\quad$ $X \sim \mathscr{N}(4,1)$ $\quad$ $X \sim f_3$

Figure: The estimator associated to previous beams, as selected by the procedure, $\widehat{b}_{\widehat{m}}$. Function $b(x) = 2x + 1$, $n = 1000$, density $f_k(x) = (k-1)/(1+x)^k \mathbf{1}_{x \geq 0}$.



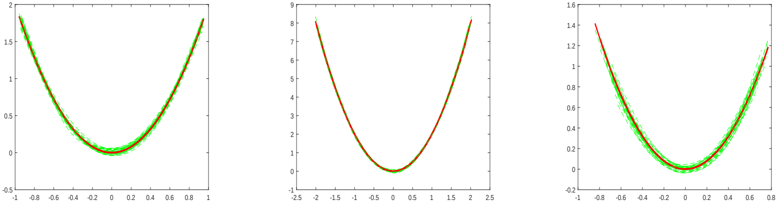
$\mathcal{N}(3,1)$	f_k with $k=4$	f_k with $k=5$
$\bar{\widehat{m}} = 7.7(0.5)$	$\bar{\widehat{m}} = 9.9(1.9)$	$\bar{\widehat{m}} = 6.7(1.1)$
$\bar{m}_{\max} = 8.8(0.4)$	$\bar{m}_{\max} = 10.4(1.7)$	$\bar{m}_{\max} = 7.6(1.0)$

Figure: 25 estimated curves in Laguerre basis (dotted -green/grey), the true in bold (red), $n = 1000$, $b(x) = 2x + 1$ and different laws for the design.



$X \sim \mathscr{U}([0,2])$	$X \sim \mathscr{E}(1)$	$X \sim \mathscr{N}(3,1)$
$\bar{\widehat{m}} = 2.5(0.8)$	$\bar{\widehat{m}} = 5.1(0.8)$	$\bar{\widehat{m}} = 5.1(0.7)$
$\bar{m}_{\max} = 5.0(0)$	$\bar{m}_{\max} = 9.4(0.6)$	$\bar{m}_{\max} = 8.9(0.3)$

Figure: 25 estimated curves in Laguerre basis (dotted -green/grey), the true in bold (red), $n = 1000$, $b(x) = 4x/(1+x^2)\mathbf{1}_{x\geq 0}$ and different laws for the design.



$X \sim \mathscr{U}([-1,1])$ $\quad$ $X \sim \mathscr{N}(0,1)$ $\quad$ $X \sim \mathit{Laplace}/4$

$\bar{\widehat{m}} = 5.1(0.4)$ $\quad$ $\bar{\widehat{m}} = 13.4(1.5)$ $\quad$ $\bar{\widehat{m}} = 6.2(1.1)$

$\bar{m}_{\max} = 7.0(0)$ $\quad$ $\bar{m}_{\max} = 15.0(0.9)$ $\quad$ $\bar{m}_{\max} = 8.1(0.6)$

Figure: 25 estimated curves in Hermite basis (dotted -green/grey), the true in bold (red), $n = 1000$, $b(x) = 2x^2$ and different laws for the design.

Conclusion

- Least squares procedure for nonparametric regression function estimation is simple and powerful
- The procedure we propose is general and rely on a random collection of models
- Laguerre and Hermite basis are of simple use but have specific properties
- Theoretical results generalize existing ones for non compactly supported bases
- Still remaining questions: can the bias be improved by the weight? Rates and optimality?

Thank you!

Conclusion

- Least squares procedure for nonparametric regression function estimation is simple and powerful
- The procedure we propose is general and rely on a random collection of models
- Laguerre and Hermite basis are of simple use but have specific properties
- Theoretical results generalize existing ones for non compactly supported bases
- Still remaining questions: can the bias be improved by the weight? Rates and optimality?

Thank you!

Selected references

Baraud, Y. (2000) Model selection for regression on a fixed design. *Probab. Theory Related Fields* **117**, 467-493.

Baraud, Y. (2002) Model selection for regression on a random design. *ESAIM Probab. Statist.* **6**, 127-146.

Baraud, Y., Comte, F. and Viennet, G. (2001a) Adaptive estimation in autoregression or β-mixing regression via model selection. *Ann. Statist.* **29**, 839-875.

Comte, F., Genon-Catalot, V. and Rozenholc, Y. (2007) Penalized nonparametric mean square estimation of the coefficients of diffusion processes. *Bernoulli* **13**, 514-543.

Tropp, J. A. (2015).
An introduction to matrix concentration inequalities.
Found. Trends Mach. Learn., 8(1-2):1–230.

Viennet, G. (1997). Inequalities for absolutely regular processes: application to density estimation. *Probab. Theory Relat. Fields* **107**, 467-492.