Optimal adaptive estimation of the derivative of a density on $\mathbb{R}$ or $\mathbb{R}^+$

Ousmane Sacko

Joint with Fabienne Comte et Céline Duval

LABORATOIRE MAP5

Université de Paris

23/10/20

Motivations

- The derivatives of the density provide information about the slope of the curves, local extrema, saddle points...
- Mode seeking in mixture models and in data analysis, see *e.g.* Cheng (1995), Chacón and Duong (2013).
- To estimate Fisher informations, to develop statistic tests to find eventual modes (Genovese *et al* (2016)), to select the optimal bandwidth parameter for density estimation (see Silverman (1978))

Two specific examples

- Regression curves (see Park and Khang (2008)), $r(x) = \mathbb{E}(Y|X = x) = f^{(1)}(x)/f(x)$ where f is a density.
- Diffusion processes : Let $(X_t)_{t\geq 0}$ be the solution of

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

where W_t is a standard Brownian independent on η. Under additional conditions (see Schmisser (2013)), the model is stationary and admits a stationary distribution f and it holds that

$$\frac{f^{(1)}(x)}{f(x)} \propto \frac{2b(x)}{\sigma^2(x)} - 2\frac{\sigma'(x)}{\sigma(x)}.$$

If the variance σ is either a constant or known, estimating f and $f^{(1)}$ leads to an estimate of b.

1

1. Singh, R. S. (1977). Applications of estimators of a density and its derivatives to certain statistical problems. J. Roy. Statist. Soc. Ser. B, 39(3) :357–363.

Model and Assumptions

Observations : $X_1, \ldots, X_n$ i.i.d. r.v. of density f, d-times differentiable

Assumptions :

(A1) $f^{(d)} \in \mathbb{L}^2(\mathbb{R}^+)$ (Laguerre case) or $f^{(d)} \in \mathbb{L}^2(\mathbb{R})$ (Hermite case),

(A2) $\|f^{(j)}\|_\infty < +\infty$ for all $0 \leq j \leq d-1$,

(A3) $\lim_{x\to 0} f^{(j)}(x) = 0$ for all $0 \leq j \leq d-1$ (Laguerre case).

Goal : Estimate the d-th derivative of f, denoted $f^{(d)}$

Laguerre and Hermite basis

We define the Laguerre polynomials and basis by :

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

The Hermite functions $(h_j)_{j\geq 0}$ are defined from Hermite polynomials by :

$$h_j(x) = c_j H_j(x) e^{-x^2/2}, \quad c_j = (2^j j! \sqrt{\pi})^{-1/2}, \quad x \in \mathbb{R}.$$

$(\ell_j)_{j\geq 0}$ and $(h_j)_{j\geq 0}$ are orthonormal basis respectively on $\mathbb{L}^2(\mathbb{R}^+)$ or $\mathbb{L}^2(\mathbb{R})$.

Hermite and Laguerre basis, Why ?

- Laguerre and Hermite bases have nice mathematical properties,
- Hermite and Laguerre bases do not require a preliminary choice of the estimation interval,
- Laguerre basis is natural if the variables of interest are positive, Hermite basis is natural for diffusion models,
- Hermite basis have a low complexity, few coefficients are required for a good representation.

[2]

2. Belomestny, D., Comte, F., and Genon-Catalot, V. (2019). Sobolev-Hermite versus Sobolev nonparametric density estimation on R. Ann. Inst. Statist. Math., 71(1) :29–62.

Projection estimator

- Let $m \geq 1$ and $S_m = \text{span}\{\varphi_0, \dots, \varphi_{m-1}\}$, where $\varphi_j = h_j$ (Hermite case) or $\varphi_j = l_j$ (Laguerre case).

- Under (**A**1), we have $f^{(d)} = \sum_{j \geq 0} a_j \varphi_j$ and the orthogonal projection of $f^{(d)}$ on S_m is given by :

$$f_m^{(d)} = \sum_{j=0}^{m-1} a_j \varphi_j, \quad a_j = \langle f^{(d)}, \varphi_j \rangle = \int f^{(d)}(x) \varphi_j(x) dx.$$

- To estimate $f_m^{(d)}$, we build m estimators of the a_j.

Using an integration by part, we have

$$a_j = \left[\sum_{k=0}^{d-1} (-1)^k f^{(d-1-k)}(x) \varphi_j^{(k)}(x) \right]_{-\infty}^{+\infty} + (-1)^d \int_{\mathbb{R}} \varphi_j^{(d)}(x) f(x) dx.$$

Under (**A**2), (**A**3), we derive

$$a_j = (-1)^d \int_{\mathbb{R}} \varphi_j^{(d)}(x) f(x) dx = (-1)^d \mathbb{E}\left[\varphi_j^{(d)}(X_1) \right].$$

Then, we define our estimator as follows :

$$\widehat{f}_{m,(d)} = \sum_{j=0}^{m-1} \widehat{a}_j^{(d)} \varphi_j, \qquad \widehat{a}_j^{(d)} = \frac{(-1)^d}{n} \sum_{i=1}^{n} \varphi_j^{(d)}(X_i).$$

- For $d = 0$, we obtain an estimator of f.

Question ?

Questions are :

- Rate of convergence and optimality
- Selection of model and risk of the adaptive estimator

Risk bound for m fixed

Proposition

*Under (**A**1), ..., (**A**3) and if in addition*

$\mathbb{E}[X_1^{-d-1/2}] < +\infty$ *(Laguerre case) or* $\mathbb{E}[|X_1|^{2/3}] < +\infty$ *(Hermite case).*

Then, for sufficiently large $m \geq d$, it holds that

$$\mathbb{E}\left[\|\widehat{f}_{m,(d)} - f^{(d)}\|^2\right] \leq \|f_m^{(d)} - f^{(d)}\|^2 + C\frac{m^{d+\frac{1}{2}}}{n} - \frac{\|f_m^{(d)}\|^2}{n}, \qquad (3.1)$$

for a positive constant C depending on the moments in condition (1) (but not on m nor n).

- 1^{st} term : $\|f_m^{(d)} - f^{(d)}\|^2 = \sum_{j\geq m} a_j^2$ bias term decreases with m.
- 2^{nd} term : variance term increases with m.

The optimal choice of m requires a bias-variance compromise.

Proposition

Under the Assumptions of Proposition 1, it holds, for some constant $c > 0$, that

$$\mathbb{E}\left[\|\widehat{f}_{m,(d)} - f^{(d)}\|^2\right] \geq \|f_m^{(d)} - f^{(d)}\|^2 + c\frac{m^{d+\frac{1}{2}}}{n} - \frac{\|f_m^{(d)}\|^2}{n}.$$

The next step is to compute the rate of convergence.

Sobolev-Hermite or Laguerre regularity spaces

Define $W^s(D)$ (see Bongioanni and Torrea (2006)) by :

$$W^s(D) = \{\theta \in \mathbb{L}^2(\mathbb{R}), |\theta|_s = \sum_{k \geq 0} k^s a_k^2(\theta) \leq D\}, \quad D > 0,$$

where

$$a_k(\theta) = \int \theta(u)\varphi_k(u)du,$$

$W^s(D) = W^s_H(D)$ (Hermite case) or $W^s(D) = W^s_L(D)$, (Laguerre case).[3]

3. Bongioanni, B. and Torrea, J. L. (2006). Sobolev spaces associated to the harmonic oscillator. Proc. Indian Acad. Sci. Math. Sci., 116(3) :337–360.

Rate of convergence

Theorem

*Assume that (**A**1), . . ., (**A**3) hold and*

$\mathbb{E}[X_1^{-d-1/2}] < +\infty$ *(Laguerre case) or* $\mathbb{E}[|X_1|^{2/3}] < +\infty$ *(Hermite case).*

Then, for sufficiently large $m \geq d$ *and* $m_{opt} = [n^{(2/2s+1)}]$*, it holds :*

$$\sup_{f \in W^s(D)} \mathbb{E}\left[\|\widehat{f}_{m_{opt},(d)} - f^{(d)}\|^2\right] \leq C(s,d,D) n^{-\frac{2(s-d)}{2s+1}}.$$

- Same rate as the one obtained by Rao (1996) (i.i.d), Giné and Nick (2016) Schmisser (2013) (dependent case).
- m_{opt} does not depend on d and is specific here : the role of m is played here by $\sqrt{m}$.
- Better when s increases, slower when d increases (most difficult and inverse problem).

Is this the optimal rate ?

Lower bound

Using the scheme given in Tsybakov (2009), we prove

Theorem

Let $s \geq d$ be an integer and $\widetilde{f}_{n,d}$ be any estimator of $f^{(d)}$. Then for n large enough, we have

$$\inf_{\widetilde{f}_{n,d}} \sup_{f \in W^s(D)} \mathbb{E}[\|\widetilde{f}_{n,d} - f^{(d)}\|^2] \geq c(s,d) n^{-\frac{2(s-d)}{2s+1}},$$

where $W^s(D)$ stands either for $W_L^s(D)$ or for $W_H^s(D)$.

But $m_{opt} = n^{1/(2s+1)}$ is not feasible ?

[4]

4. Tsybakov, A. B. (2009). Introduction to nonparametric estimation. Springer Series in Statistics. Springer, New York. Revised and extended from the 2004 French original, Translated by Vladimir Zaiats.

Model selection

Goal : To propose an automatic choice of the dimension m. We look for m such that

$$\mathbb{E}\left[\|\widehat{f}_{m,(d)} - f^{(d)}\|^2\right] = \|f_m^{(d)} - f^{(d)}\|^2 + \frac{1}{n}\sum_{j=0}^{m-1} \mathrm{Var}\left[\varphi_j^{(d)}(X_1)\right].$$

is minimal.

- Remark $\|f_m^{(d)} - f^{(d)}\|^2 = \|f^{(d)}\|^2 - \|f_m^{(d)}\|^2$
- set $V_{m,d} = \sum_{j=0}^{m-1} \mathbb{E}[(\varphi_j^{(d)}(X_1))^2]$

Replacing $\|f_m^{(d)}\|^2$ and $V_{m,d}$ by their estimator, we select m by :

$$\widehat{m}_n := \underset{m\in\mathscr{M}_{n,d}}{\mathrm{argmin}}\{-\|\widehat{f}_{m,(d)}\|^2 + \widehat{\mathrm{pen}}_d(m)\}, \quad \widehat{\mathrm{pen}}_d(m) = \kappa\frac{\widehat{V}_{m,d}}{n},$$

where $\widehat{V}_{m,d} = \frac{1}{n}\sum_{i=0}^{n}\sum_{j=0}^{m-1}(\varphi_j^{(d)}(X_i))^2$, $\kappa > 0$ must be calibrated.

Final result

Theorem

*Let $\mathscr{M}_{n,d} := \{d, \dots, m_n(d)\}$, where $m_n(d) \geq d$. Assume that (**A**1) and (**A**2) hold, and that (**A**3) holds in the Laguerre case, and that $\|f\|_\infty < +\infty$.*

AL. *Set $m_n(d) = \lfloor (n/\log^3(n))^{\frac{2}{2d+1}} \rfloor$, assume that $\sup_{x\in\mathbb{R}^+} \frac{f(x)}{x^d} < +\infty$ in the Laguerre case,*

AH. *Set $m_n(d) = \lfloor n^{\frac{2}{2d+1}} \rfloor$ in the Hermite case.*

Then, for any $\kappa \geq \kappa_0 := 32$ it holds that

$$\mathbb{E}\left[\|\widehat{f}_{\widehat{m}_n,(d)} - f^{(d)}\|^2\right] \leq C \inf_{m\in\mathscr{M}_{n,d}} \left(\|f_m^{(d)} - f^{(d)}\|^2 + \mathrm{pen}_d(m)\right) + \frac{C'}{n},$$

where $\mathrm{pen}_d(m) = \kappa \frac{V_{m,d}}{n}$, C is a universal constant ($C = 3$ suits) and C' is a constant depending on $\sup_{x\in\mathbb{R}^+} \frac{f(x)}{x^d} < +\infty$ and $\mathbb{E}[X_1^{-d-1/2}] < +\infty$ (Laguerre case) or $\|f\|_\infty$ (Hermite case).

Comments

- $\widehat{f}_{\widehat{m}_n,(d)}$ achieves automatically a bias-variance compromise
- $\widehat{f}_{\widehat{m}_n,(d)}$ performs as well as the best model in collection, up to the multiplicative C
- Non-asymptotic result
- estimator easy to implement

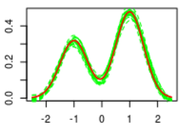
Simulation

We consider the following distributions

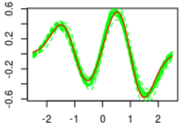
(i) Mixed Gaussian $0.4\mathcal{N}(-1,1/2)+0.6\mathcal{N}(1,1/2)$, $I=[-2.5,2.5]$, (Hermite case)

(ii) Gamma $\Gamma(5,5)/10$, $I=[0,7]$, (Laguerre case).

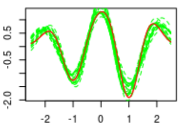
Illustration



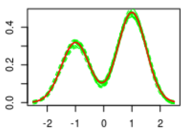
$\overline{\widehat{m}_n} = 7.65$



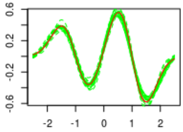
$\overline{\widehat{m}_n} = 8.15$



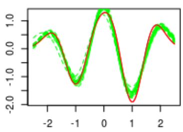
$\overline{\widehat{m}_n} = 7.85$



$\overline{\widehat{m}_n} = 9.45$



$\overline{\widehat{m}_n} = 9.70$



$\overline{\widehat{m}_n} = 8.95$

FIGURE – 20 estimates $\widehat{f}_{\widehat{m}_n,(d)}$ in the Hermite basis, with $n = 500$ (first line) and $n = 2000$ (second line). The true quantity is in bold red and the estimate in dotted lines (left $d = 0$, middle $d = 1$ and right $d = 2$).

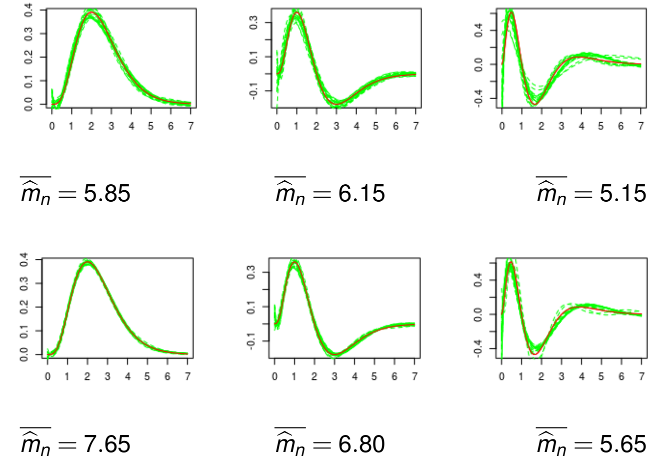


FIGURE – 20 estimates $\widehat{f}_{\widehat{m}_n,(d)}$ in the Laguerre basis, with $n = 500$ (first line), and $n = 2000$ (second line). The true quantity is in bold red and the estimate in dotted lines (left $d = 0$, middle $d = 1$ and right $d = 2$).

Conclusion and perspectives

- Conclusion
  - Projection estimator simple to implement, numerically stable, adaptive
  - Simulation results very satisfactory.
- Perspectives
  - Risque de l'estimateur adaptative pour le cas noyau.
  - Est ce possible d'étendre les résultats au cas des variables dépendantes ?

Merci pour votre attention !

Reference

Comte, F. and Genon-Catalot, V. Laguerre and Hermite bases for inverse problems. *J. Korean Statist. Soc.,* 47(3) :273–296, 2018.

Chacón, J.E. and Duong, T. Data-driven density derivative estimation, with applications to nonparametric clustering and bump hunting. Electron. J. Stat. 7 (2013), 499–532.

Rao, B. L. S. P. Nonparametric estimation of the derivatives of a density by the method of wavelets. *Bull. Inform. Cybernet.,* 28(1) :91–100, 1996.

Schmisser, E. Nonparametric estimation of the derivatives of the stationary density for stationary processes. *ESAIM Probab. Stat.,* 17 :33–69, 2013.