ESAIM: PROCEEDINGS AND SURVEYS, 2020, Vol. 68, p. 20-34
Hervé Cardot & Pierre Calka

FROM REGRESSION FUNCTION TO DIFFUSION DRIFT ESTIMATION IN NONPARAMETRIC SETTING.

Fabienne Comte[1]

Abstract. We consider a diffusion model $dX_t = b(X_t)dt + \sigma(X_t)dW_t, X_0 = \eta$, under conditions ensuring existence, stationarity and geometrical β-mixing of the process solution. We assume that we observe a sample $(X_{k\Delta})_{0\leq k\leq n+1}$. Our aim is to study nonparametric estimators of the drift function $b(.)$, under general conditions. We propose projection estimators based on a least-squares type contrast and, in order to generalize existing results, we want to consider possibly non compactly supported projection bases and possibly non bounded volatility. To that aim, we relate the model with a simpler regression model, then to a more elaborate heteroscedastic model, plus some residual terms. This allows to see the role of heteroscedasticity first and the role of dependency between the variables and to present different probabilistic tools used to face each part of the problem. For each step, we try to see the "price" of each assumption.
This is the developed version of the talk given in August 2018 in Dijon, Journées MAS.

Résumé. Nous considérons un modèle de diffusion $dX_t = b(X_t)dt + \sigma(X_t)dW_t, X_0 = \eta$, sous des conditions garantissant l'existence, la stationarité et le β-mélange géométrique du processus solution. Nous supposons que nous disposons d'observations $(X_{k\Delta})_{0\leq k\leq n}$. Notre objectif est d'étudier un estimateur nonparamétrique de la fonction $b(.)$, sous des hypothèses générales. Nous proposons des estimateurs par projection, basés sur un contraste des moindres carrés. Afin de généraliser les résultats existants, nous voulons des jeux d'hypothèses autorisant des bases de projection à support non compact, ainsi que des fonctions de volatilité non bornées. Ainsi, nous relions le modèle de diffusion à un modèle plus simple de régression, puis à un modèle hétéroscédastique, plus des termes de reste. Cela nous permet de détailler le rôle de l'hétéroscédasticité puis de la dépendance, et de présenter les différents outils probabilistes utilisés pour affronter chaque problème. A chaque étape, nous étudions le prix des hypothèses.
Ceci est la version développée de l'exposé présenté en Août 2018 à Dijon, lors des Journées MAS.

Introduction

This is a version of my talk about nonparametric drift estimation of a discretely observed continuous-time diffusion. The results are proved in a series of three papers jointly written with V. Genon-Catalot, [9], [8], [10] and thus, proofs are not repeated here. The intention is rather to present an overview of the topic, and to explain the link between diffusions and heteroscedastic regression results from nonparametric estimation point of view: link means of course similarities and differences.

The three aforementioned works propose a new visit of the topic, formerly studied in several other papers: Baraud (2000, 2002) studied nonparametric regression estimation, while Baraud *et al.* (2001a, 2001b) considered the same type of questions in dependent context. Then Comte *et al.* (2007) applied the regression strategy to

[1] Université Paris Descartes, Laboratoire MAP5, email: fabienne.comte@parisdescartes.fr

Article published online by EDP Sciences and available at https://www.esaim-proc.org or https://doi.org/10.1051/proc/202068002

diffusion models. All these references rely on compactly supported bases and domains of estimation, and assume the density of the random regressors to be lower bounded on the compact estimation set. Moreover, the density lower bound appears in several denominators, making potentially some constants explode, since this term can obviously be very small. When heteroscedasticity is present is these works, the volatility function is also systematically assumed to be upper bounded: the assumption is not strong if the support is compact, but can be undesirable for classical diffusion models when considering non compact support: for instance, in the square-root (or Cox-Ingersoll-Ross) model $\sigma(x) = \sigma\sqrt{x}$, is not bounded on $\mathbb{R}^+$.

This is the reason why we re-consider these models, aiming at extending the result to the non compact support context, which means avoiding several classical lower or upper bound assumptions. This new visit was permitted by new probabilistic tools as Tropp's (2012) matricial Chernoff and Bernstein inequalities. The relevance of this powerful result for the regression setting has already been noticed by Cohen *et al.* (2013) who proposed stability conditions: we reformulate them in a way allowing to deal with the general regression estimator. Moreover, we re-define the regression estimator in a truncated version involving a matricial random cutoff, and this is new. We emphasize that, from several aspects, the regression problem seen under this general setting turns out to have similarities with inverse problems. This is especially the case for model selection results, where the new penalties and the collection of models which appear now are rather elaborate and not easy to deal with from theoretical point of view (while easy to implement in practice).

The presentation that follows relies on all the papers mentioned above, which of course do not constitute an exhaustive review of the domain.

1. From diffusion to regression

We observe with sampling interval Δ, the random variables $(X_{i\Delta})_{0\le i\le n+1}$, from the diffusion process $(X_t)_{t\ge 0}$,

$$dX_t = b(X_t)dt + \sigma(X_t)dW_t, \quad X_0 = \eta, \quad t \ge 0, \tag{1}$$

where $(W_t)_{t\ge 0}$ is a standard Brownian motion independent of η. Precise assumptions will be given later on, but let us just say that they are such that equation (1) admits a unique solution; moreover the initial condition η is chosen such that the process is stationary.

Now, we set

$$Y_{i\Delta} = \frac{X_{(i+1)\Delta} - X_{i\Delta}}{\Delta}, \quad Z_{i\Delta} = \frac{1}{\sqrt{\Delta}}\sigma(X_{i\Delta})\left(\frac{W_{(i+1)\Delta} - W_{i\Delta}}{\sqrt{\Delta}}\right),$$

and see that (1) can be re-written

$$\underbrace{Y_{i\Delta} = b(X_{i\Delta}) + Z_{i\Delta}}_{\text{regression equation}} + R_{i\Delta}, \tag{2}$$

where $R_{i,\Delta} = R_{i\Delta,1} + R_{i\Delta,2}$

$$R_{i\Delta,1} = \frac{1}{\Delta}\int_{i\Delta}^{(i+1)\Delta}(b(X_s) - b(X_{i\Delta}))ds, \quad R_{i\Delta,2} = \frac{1}{\Delta}\int_{i\Delta}^{(i+1)\Delta}(\sigma(X_s) - \sigma(X_{i\Delta}))dW_s.$$

Equation (2) is almost a regression equation, with nonparametric regression function $b(.)$. In this model, the process $Z_{i\Delta}$ plays the role of an heteroscedastic noise, and $R_{i\Delta}$ is an additional residual term that we take into account. We note that, among the difficulties, only one process is observed $X_{i\Delta}$, $i = 0, 1, \ldots, n+1$.

We emphasize that the decomposition proposed in (2) is different from the one in Comte *et al.* (2007), where the noise term was $\Delta^{-1}\int_{i\Delta}^{(i+1)\Delta}\sigma(X_s)dW_s$ and residual term $R_{i\Delta,1}$: this change, which at first sight seems minor, is in fact important because it allows to use the Talagrand deviation inequality in the model selection part, together with coupling methods, while in the former paper, we had to apply martingale deviations and chaining methods. With such probabilistic tools, the assumption that σ is bounded seemed unavoidable, while we can handle this with the new definition of $Z_{i\Delta}$ in (2).

Under conditions on $b(\cdot)$, $\sigma(\cdot)$ and η (see Assumptions **(A1)-(A4)** in section 4), there is a unique strictly stationary solution, with stationary density denoted by π. Moreover, we assume that the process $(X_t)_{t\geq 0}$ is geometrically β-mixing (and it would be interesting to study the influence of arithmetical mixing). It is clear from (2) that we have reduced model (1) to a regression model, involving an heteroscedastic noise $Z_{i\Delta}$, and dependency between the observations. So, let us start by simpler regression, before studying the impact of the presence of the volatility function and lastly, the price of dependency between the variables.

2. The origins − Standard regression model

The standard regression model writes

$$Y_i = b(X_i) + \varepsilon_i, \quad i = 1, \dots, n \tag{3}$$

with noise variables $(\varepsilon_i)_{1\leq i\leq n}$ i.i.d. centered with variance σ_ε^2, explanatory variables $(X_i)_{1\leq i\leq n}$ i.i.d. with density π, and the assumption that the sequence $(X_i)_{1\leq i\leq n}$ is independent of $(\varepsilon_i)_{1\leq i\leq n}$. Here, the observations are the couples $(Y_i, X_i)_{1\leq i\leq n}$ and our aim is nonparametric estimation of $b(.)$.

2.1. Projection estimators

We intend to build a projection estimator of $b(.)$. For that purpose, we define $(\varphi_j)_{0\leq j\leq m-1}$ an orthonormal basis in $\mathbb{L}^2(A, dx)$, where $A \subseteq \mathbb{R}$. In other words, the $(\varphi_j)_j$ satisfy

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

where $\delta_{j,k}$ denotes the Kronecker symbol. Then, we look for an estimator that may be written

$$\hat{b}_m = \sum_{j=0}^{m-1} \hat{a}_j \varphi_j, \tag{4}$$

where the coefficient estimates $(\hat{a}_j)_{0\leq j\leq m-1}$ should be computed from the observations $(Y_i, X_i)_{1\leq i\leq n}$.

2.2. Quotient estimators

Let us mention that we do not want to handle *Nadaraya-Watson* or quotient estimators, because we consider that they are not exactly of projection type as given by (4). The principle of such estimates is to define the function

$$r = b\pi$$

where b is the regression function of interest and π denotes the density of the design $(X_i)_{1\leq i\leq n}$. Indeed, this function is often simple to estimate. Then a quotient estimator is defined by

$$\widetilde{b}_{m,m'} = \frac{\hat{r}_{m'}}{\hat{\pi}_m}$$

where

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

Separately, $\hat{r}_m$ and $\hat{\pi}_{m'}$ are projection estimators and are easy to study. Note that the quotient can be performed coefficient by coefficient:

$$\widetilde{\widetilde{b}}_{m,m'} = \sum_{j=0}^{m'-1} \tilde{a}_j\varphi_j, \quad \tilde{a}_j = \frac{1}{n}\sum_{i=1}^{n} \frac{Y_i\varphi_j(X_i)}{\hat{\pi}_m(X_i)}.$$

These estimators can work, and in some cases, nothing else can be theoretically justified. However, the consequences are the same for both $\widetilde{b}_{m,m'}$ and $\widetilde{\widetilde{b}}_{m,m'}$: they do not provide a development in a basis. Their study leads to risk bounds involving the risk of both numerator and denominator estimators: each separately is rather easy, but final rates depend on the regularity of functions r and π and not only on the regularity of b. Moreover, making the ratio requires a careful choice of a cutoff to avoid that the denominator gets too small; lastly, it depends on two dimension parameters which have to be selected. It is not clear if the best selection of each separately gives the best final quotient: maybe a joint selection should be studied.

2.3. **Least squares estimator**

To define our projection estimator, let us consider the m-dimensional linear space

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

In the sequel, the **least squares estimator** is

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

To compute the coefficients, it is enough to see that everything works as if $a_0, \dots, a_{m-1}$ where the parameters in the linear model

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

and thus, we know how to get the least squares estimator with classical formula:

$$\hat{b}_m = \sum_{j=0}^{m-1} \hat{a}_j \varphi_j, \quad \hat{\vec{a}}_{(m)} := \begin{pmatrix} \hat{a}_0 \\ \vdots \\ \hat{a}_{m-1} \end{pmatrix} = \left({}^t\widehat{\Phi}_m \, \widehat{\Phi}_m \right)^{-1} {}^t\widehat{\Phi}_m \vec{\mathcal{Y}}, \tag{5}$$

where

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

provided that

$${}^t\widehat{\Phi}_m \, \widehat{\Phi}_m \text{ is invertible a.s.}$$

We want to study the risk of the estimator for fixed m, to select an adequate model $\hat{m}$ from the data and to bound the risk of the final estimator, $\hat{b}_{\hat{m}}$.

It happens that Baraud (2000, 2002) studied these questions, for bases with compact support A. Such a context allows the following assumption

$$\forall x \in A, \quad 0 < \pi_{\min} \le \pi(x) \le \pi_{\max} < +\infty. \tag{6}$$

And indeed, in several computations, the constant $\pi_{\min}$ is crucial and more precisely $1/\pi_{\min}$ is involved in several bounds. Obviously, this assumption **can not hold on a non compact** A.

Assumption (6) is useful to relate conveniently the three norms appearing in the problem:

- the empirical norm $\|t\|_n^2 = \dfrac{1}{n} \sum_{i=1}^{n} t^2(X_i)$, $\|t\|_n^2 = \dfrac{1}{n} \sum_{i=1}^{n} t^2(X_{i\Delta})$,
- the $\mathbb{L}^2(A, \pi(x)dx)$-norm, $\|t\|_\pi^2 = \int_A t^2(x)\pi(x)dx = \mathbb{E}[\|t\|_n^2]$ for t with support A,

- the $\mathbb{L}^2(A,dx)$-norm, $\|t\|^2 = \int_A t^2(x)dx$.

Clearly, the empirical norm a.s. converges to the $\mathbb{L}^2(A,\pi(x)dx)$-norm, so the link between the first two norms is always possible, and specifically improved by Tropp's (2012) results. Then, the $\mathbb{L}^2(A,\pi(x)dx)$-norm and the $\mathbb{L}^2(A,dx)$-norm are equivalent for A compact if π satisfies assumption (6): such equivalence allows to indifferently use π-weighted norm, which is natural in the problem, or usual norm, which is natural due to the choice of the basis. The upper-bound part of (6) does not seem a strong constraint in all cases, but the lower bound side of (6) is crucially related to the compactness of A. A solution may be to consider a $\mathbb{L}^2(A,\pi(x)dx)$-orthonormal basis: this is done in part of the proofs, but it is not possible in practice since π is unknown.

2.4. **Non compact support, what for?**

One may wonder why we are interested in this non compact assumption, so let me explain our motivations.

1- First, we aim at generalization of existing results, if it is possible.

2- We also want to better understand the dependence of results w.r.t. $\pi_{\min}$, and to see how certain bounds may explode if this term is small.

3- We have at hand simple and convenient non compactly supported bases: Laguerre ($\mathbb{R}^+$-supported) and Hermite ($\mathbb{R}$-supported) bases, which have nice properties. The Hermite basis is especially natural for diffusion models.

4- We also have in mind to extend the regression strategy to other models with unknown support of the regressor, such as: survival function estimation in presence of interval censoring, hazard rate estimation in presence of right censoring (these two cases can be expressed as univariate regression and since nonnegative random variables are often involved in such models, the Laguerre basis is of natural use), conditional density estimation ...

5- Lastly, we believe that regression strategy may be useful for inverse problems (for instance, to handle noisy observations of X).

2.5. **The bases: compact support or not**

Let us give a quick insight of the bases we concretely have in mind.

Examples of compactly supported bases, $A = [0,1]$. Classical compactly supported bases are:

- Histograms $p_j^{(0)}(x) = \sqrt{m}\mathbf{1}_{[j/m,(j+1)/m[}(x)$, for $j = 0,\dots,m-1$; more generally, piecewise polynomials with degree r, $p_j^{(r)}$;
- 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

$$\|\sum_{j=0}^{m-1}\varphi_j^2\|_\infty \le c_\varphi^2 m$$

with $c_\varphi^2 = 1$ for histograms and trigonometric basis, $c_\varphi^2 = r+1$ for piecewise polynomials. Moreover, the associated spaces S_m are 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

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

Collection $(\ell_j)_{j\geq 0}$ is a complete orthonormal system on $\mathbb{L}^2(\mathbb{R}^+)$, with

$$\forall j \geq 0, \ \ \forall x \in \mathbb{R}^+, \ \ |\ell_j(x)| \leq \sqrt{2}. \tag{7}$$

The spaces $(S_m = \text{span}\{\ell_0, \ldots, \ell_{m-1}\})_m$ are nested, and (7) 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 $j \geq 0$ are given by:

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

We have: $(h_j, j \geq 0)$ orthonormal basis of $\mathbb{L}^2(\mathbb{R}, dx)$ and

$$\|h_j\|_\infty \leq \Phi_0, \qquad \Phi_0 \simeq 1,086435/\pi^{1/4} \simeq 0.8160, \tag{8}$$

and thus the $(S_m = \text{span}\{h_0, \ldots, h_{m-1}\})_m$ are nested and (8) implies that $\| \sum_{j=0}^{m-1} h_j^2 \|_\infty \leq c_\varphi^2 m$ with $c_\varphi^2 = \Phi_0^2$.

2.6. First risk bound in independent setting

2.6.1. *No support condition for the first basic result*

To begin with, let us state a very simple and general result, recalled e.g. in Baraud (2000).

Proposition 2.1. *Let $(X_i, Y_i)_{1\leq i\leq n}$ be observations drawn from model (3), $b_A = b\mathbf{1}_A$. Assume that $b_A \in \mathbb{L}^2(A, \pi(x)dx)$ and that ${}^t\widehat{\Phi}_m\widehat{\Phi}_m$ is invertible a.s. Consider the least squares estimator $\hat{b}_m$ of b, given by (5). Then*

$$\mathbb{E}\big[\|\hat{b}_m - b_A\|_n^2\big] \quad \leq \quad \inf_{t\in S_m} \|t - b_A\|_\pi^2 + \sigma_\varepsilon^2 \frac{m}{n}. \tag{9}$$

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

Proof of Proposition 2.1. Let Π_m be the orthogonal projection (for the scalar product of $\mathbb{R}^n$) on the sub-space $\big\{\big(t(X_1), \ldots, t(X_n)\big)', t \in S_m\big\}$ of $\mathbb{R}^n$. Then

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

Now we have, thanks to $(\varepsilon_i)_{1\leq i\leq n}$ independent of $(X_i)_{1\leq i\leq n}$, by elementary matrix computation,

$$\mathbb{E}\left[\|\hat{b}_m - \Pi_m b\|_n^2\right] = \sigma_\varepsilon^2 \frac{m}{n}. \tag{10}$$

Therefore,

$$\begin{aligned} \mathbb{E}\big[\|\hat{b}_m - b_A\|_n^2\big] &= \mathbb{E}\left[\inf_{t\in S_m} \|t - b_A\|_n^2\right] + \sigma_\varepsilon^2 \frac{m}{n} &(11)\\ &\leq \inf_{t\in S_m} \|t - b_A\|_\pi^2 + \sigma_\varepsilon^2 \frac{m}{n}. &(12) \end{aligned}$$

□

We can check that, in (9), the bias is getting small when m grows, while the variance increases. This implies that a compromise will have to be found, by relevant choice of m.
But we mainly detail this result to emphasize that

- The bound (9) is general, almost exact, 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.

This is important and not so obvious.

2.6.2. *Comparison with density estimation*

Why is it important to notice the equality (11)? This is due to comparison with density estimation. Recall that, for i.i.d. X_i with density π, the projection estimator is defined by (see section 2.2):

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

This estimator satisfies

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

For all the bases we described above, we have

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

In some cases, we even have $\sum_{j=0}^{m-1} \varphi_j^2 = c_\varphi^2\, m$: this holds for histograms and trigonometric polynomials with odd dimension, with $c_\varphi = 1$. Thus we obtain a variance bound which is exactly m/n, and thus is sharp for some bases.

However, for the Laguerre basis, it is true that $\sum_{j=0}^{m-1} \varphi_j^2(0) = 2m$ and thus $\sup_{x\in\mathbb{R}^+} \sum_{j=0}^{m-1} \varphi_j^2(x) = 2m$: therefore, at first sight, we may have the same conclusion. However, for Hermite and Laguerre bases, it is proved in Comte and Genon-Catalot (2018a, Prop. 3.1), that, for some constant c,

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

Thus, for density estimation, the variance order depends on the basis. This is why we emphasize for $\hat{b}_m$, that the bound stated in Proposition 2.1 is equal to $\sigma_\varepsilon^2 m/n$ whatever the basis.

2.7. **Bound on the integrated risk**

The first step to the inverse problem is encountered when looking for a bound on the integrated risk $\mathbb{E}\left[\|\widehat{b}_m - b_A\|_\pi^2\right]$. Let us define the normalized $m \times m$ matrix

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

which is such that

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

Provided that $\widehat{\Psi}_m$ is invertible a.s., formula (5) can be re-written as

$$\hat{b}_m = \sum_{j=0}^{m-1} \hat{a}_j \varphi_j, \quad \hat{\vec{a}}_{(m)} = \widehat{\Psi}_m^{-1}(\frac{1}{n}\, {}^t\widehat{\Phi}_m \vec{\mathcal{Y}}).$$

To control the estimator, we have to study the distance between $\widehat{\Psi}_m$ and its expectation, and more precisely, to control $\|\Psi_m^{-1/2}\widehat{\Psi}_m\Psi_m^{-1/2} - \mathrm{Id}_m\|_{\mathrm{op}}$, where $\|M\|_{\mathrm{op}}^2 = \lambda_{\max}(M\, {}^tM)$, $\lambda_{\max}(M)$ denotes the largest eigenvalue of M, $\Psi_m^{-1/2}$ is a symmetric square root of Ψ_m^{-1} and Id_m denotes the $m \times m$-identity matrix. For M symmetric positive definite, $\|M\|_{\mathrm{op}} = \lambda_{\max}(M)$. The key tool for this study is given by matricial Chernoff and Bernstein deviation inequalities stated in Tropp (2012).

Let us define

$$L(m) = \sup_{x\in A} \sum_{j=0}^{m-1} \varphi_j^2(x) \ \text{ and assume } \ L(m) < +\infty. \tag{13}$$

It can be easily checked that $L(m)$ is independent of the choice of the $\mathbb{L}^2(A, dx)$-orthonormal basis of S_m. Note that we found $L(m) = c_\varphi^2 m$, for different values of c_φ, in all our examples (but other orders are possible). We also have the following useful property.

Lemma 2.1. *For nested spaces S_m and invertible Ψ_m (resp. $\widehat{\Psi}_m$), then*

$$m \mapsto \|\Psi_m^{-1}\|_{\mathrm{op}} \quad (resp. \quad m \mapsto \|\widehat{\Psi}_m^{-1}\|_{\mathrm{op}}) \ \textit{is increasing.}$$

The link between empirical and integrated-π norms is controlled on the random set defined by

$$\Omega_m(\delta) = \left\{ \sup_{t\in S_m,\, t\neq 0} \left| \frac{\|t\|_n^2}{\|t\|_\pi^2} - 1 \right| \leq \delta \right\}. \tag{14}$$

We can relate it with aforementioned distance beween $\widehat{\Psi}_m$ and Ψ_m and bound the probability of the set by using the strategy of Theorem 1 in Cohen *et al.* (2013).

Proposition 2.2. *Assume that Ψ_m is invertible and assumption (13) is satisfied. Then for all $0 \leq \delta \leq 1$,*

$$\mathbb{P}(\Omega_m(\delta)^c) \quad = \quad \mathbb{P}\left[\|\Psi_m^{-1/2}\widehat{\Psi}_m\Psi_m^{-1/2} - \mathrm{Id}_m\|_{\mathrm{op}} > \delta\right] \leq 2m \exp\left(-c(\delta)\frac{n}{L(m)(\|\Psi_m^{-1}\|_{\mathrm{op}} \vee 1)}\right),$$

where $c(\delta) = \delta + (1-\delta)\log(1-\delta)$.

As a consequence, we choose $\delta = 1/2$, define $\Omega_m := \Omega_m(1/2)$ and obtain that

$$\mathbb{P}(\Omega_m^c) \leq 2n^{-4}$$

if m is such that

$$L(m)(\|\Psi_m^{-1}\|_{\mathrm{op}} \vee 1) \leq \frac{\mathfrak{c}}{2}\frac{n}{\log(n)}, \quad \mathfrak{c} = \frac{1-\log(2)}{5}. \tag{15}$$

Condition (15) is also called *stability condition*. It is worth noting that these results are available for all possible classical bases, whether compactly supported or not. This explains why we define the trimmed estimator

$$\tilde{b}_m := \hat{b}_m \mathbf{1}_{\{L(m)\,(\|\widehat{\Psi}_m^{-1}\|_{\mathrm{op}} \vee 1) \leq \mathfrak{c}\frac{n}{\log(n)}\}}, \quad \mathfrak{c} = \frac{1-\log(2)}{5}. \tag{16}$$

This truncation, based on an empirical version of the stability condition, is mandatory to obtain a bound on the integrated risk and is the new point of our generalisation, in particular compared to the results of Baraud (2002), Cohen *et al.* (2013). This is also what makes regression study rely on methods used for inverse problems.

Proposition 2.3. *Assume $\mathbb{E}(\varepsilon_1^4) < +\infty$, $b_A \in \mathbb{L}^4(A,\pi)$. Then for all m satisfying (15), we have*

$$\mathbb{E}\big[\|\tilde{b}_m - b_A\|_\pi^2\big] \leq \left(1 + \frac{8\mathfrak{c}}{\log(n)}\right) \inf_{t\in S_m} \|b_A - t\|_\pi^2 + 8\sigma_\varepsilon^2 \frac{m}{n} + \frac{c}{n}, \tag{17}$$

where c is a constant depending on $\mathbb{E}(\varepsilon_1^4)$ and $\int b_A^4(x)\pi(x)dx$.

Note that the constant in front of the squared bias term is near of 1, especially for large n.

2.8. **Implications in function of the support**

2.8.1. *Case of compact support A*

If A is compact, one usually assumes that (6) holds, and that $b_A \in \mathbb{L}^2(A,dx)$. This last condition is not very strong when A is compact. Under the upper-bound part of (6) (*i.e.* $\pi(x) \leq \pi_{\max} < +\infty, \forall x \in A$) and this integrability condition, we get that

$$\|t - b_A\|_\pi^2 \leq \pi_{\max}\|t - b_A\|^2.$$

Therefore, the bias term in bound (17) can be related to standard orders on regularity spaces associated to the bases (generally Besov spaces). In addition, using the lower-bound part of (6), we can prove

$$\pi \geq \pi_{\min} > 0 \Rightarrow \|\Psi_m^{-1}\|_{\text{op}} \leq 1/\pi_{\min}.$$

This is why under (6) and for bases such that $L(m) \leq c_\varphi^2\, m$, the stability condition (15) reduces to $m \leq c'(\pi_{\min})n/\log(n)$, which is weak and standard. No random cutoff for the estimator is needed, and the problem gets simpler.

Note that we then recover standard rates on Besov spaces (*i.e.* rates of order $n^{-2\alpha/(2\alpha+1)}$ for b belonging to Besov spaces associated with regularity α).

2.8.2. *Laguerre and Hermite bases*

As already said, Laguerre and Hermite bases have non compact support ($A = \mathbb{R}^+$ in Laguerre and $A = \mathbb{R}$ for Hermite). For these bases, we have the first favorable property:

Lemma 2.2. *For all $m \in \mathbb{N}$, Ψ_m is invertible, and for all $m \leq n$, $\widehat{\Psi}_m$ is invertible a.s.*

However, we can also prove that $\|\Psi_m^{-1}\|_{\text{op}}$ depends on m.

Proposition 2.4. *Assume that $\inf_{a\leq x\leq b} \pi(x) > 0$ 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,*

$$\|\Psi_m^{-1}\|_{\text{op}}^2 \geq c^\star m. \tag{18}$$

In fact, we believe that the order in m of $\|\Psi_m^{-1}\|_{\text{op}}$ can be much more explosive, but we can only prove the following upper bound.

Proposition 2.5. *Consider the Laguerre or the Hermite basis. Assume that*

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

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

Simulations show that $\|\Psi_m^{-1}\|_{\text{op}}$ grows very fast and therefore, the constraint $m\|\Psi_m^{-1}\|_{\text{op}} \leq \mathfrak{c}n/\log(n)$ can be strong. Indeed, if π is as in Proposition 2.5, the stability condition (15) imposes

$$m^{k+1} \lesssim \underbrace{n/\log(n)}_{\text{regression}} \quad (\text{ or } \quad \underbrace{(n\Delta)/\log(n\Delta)}_{\text{diffusion}} \ ).$$

3. Heteroscedastic regression model

The next question is: is the risk bound valid for observations from the diffusion model? Not exactly, since the discrete diffusion observations involve two additional problems:

- heteroscedasticity in the noise
- dependence between the observations.

Let us now deal with heteroscedasticity and consider observations $(X_i, Y_i)_{1\leq i\leq n}$, from the model

$$Y_i = b(X_i) + \sigma(X_i)\varepsilon_i, \quad \mathrm{Var}(\varepsilon_i) = 1. \tag{19}$$

Then if σ is bounded on A, by say $\|\sigma_A\|_\infty$, it is easy to prove that

$$\mathbb{E}\big[\|\hat{b}_m - b_A\|_n^2\big] \leq \inf_{t\in S_m} \|b_A - t\|_\pi^2 + \|\sigma_A\|_\infty^2 \frac{m}{n}.$$

Therefore, extensions of the previous results hold with σ_ε^2 replaced by $\|\sigma_A\|_\infty^2$.

Now, we study the case where we only assume $\mathbb{E}[\sigma^4(X_1)] < +\infty$. To that aim, we set

$$\Psi_{m,\sigma^2} := \left(\int \varphi_j(x)\varphi_k(x)\sigma^2(x)\pi(x)dx\right)_{0\leq j,k\leq m-1}. \tag{20}$$

Then we can prove:

Proposition 3.1. *Assume $\mathbb{E}(\sigma^4(X_1)) < +\infty, \mathbb{E}(b^4(X_1)) < +\infty$. Then, for all m satisfying (15),*

$$\mathbb{E}\big[\|\hat{b}_m - b\|_n^2\big] \leq \inf_{t\in S_m} \|b_A - t\|_\pi^2 + \frac{2}{n}\mathrm{Tr}\left[\Psi_m^{-1/2}\Psi_{m,\sigma^2}\Psi_m^{-1/2}\right] + \frac{c}{n}$$

where $c > 0$.

If $\sigma(x) = \sigma$, $\quad \mathrm{Tr}\left[\Psi_m^{-1/2}\Psi_{m,\sigma^2}\Psi_m^{-1/2}\right] = \sigma^2 m$ and we recover the homoscedastic case. But the order of the general variance term is not so obvious. To study this new quantity, the following properties are useful.

Proposition 3.2. *Let m be a nonzero integer.*

(1) *$m \mapsto \mathrm{Tr}\left[\Psi_m^{-1/2}\Psi_{m,\sigma^2}\Psi_m^{-1/2}\right]$ is nondecreasing.*

(2) *If σ is bounded on A, then $\mathrm{Tr}\left[\Psi_m^{-1/2}\Psi_{m,\sigma^2}\Psi_m^{-1/2}\right] \leq \|\sigma_A\|_\infty^2 m$.*

(3) *If $\mathbb{E}[\sigma^2(X_1)] < +\infty$ and the basis satisfies $\|\sum_{j=0}^{m-1}\varphi_j^2\|_\infty \leq c_\varphi^2 m$, then*

$$\mathrm{Tr}\left[\Psi_m^{-1/2}\Psi_{m,\sigma^2}\Psi_m^{-1/2}\right] \leq c_\varphi^2 \mathbb{E}[\sigma_A^2(X_1)]m\|\Psi_m^{-1}\|_{\mathrm{op}}.$$

Moreover, if we intend to provide a bound for the integrated risk, we still have to consider a truncated estimator: the definition of $\tilde{b}_m$ is the same as previously and given by (16), under the same stability condition.

Proposition 3.3. *Assume that $\widehat{\Psi}_m$ is invertible a.s. and that $\mathbb{E}(\sigma^4(X_1)) < +\infty, \mathbb{E}(b^4(X_1)) < +\infty$. Let m satisfy condition (15) and $\tilde{b}_m$ be given by (16), then*

$$\mathbb{E}\big[\|\tilde{b}_m - b_A\|_\pi^2\big] \quad \leq \quad \left(1 + \frac{8\mathfrak{c}}{\log(n)}\right)\inf_{t\in S_m}\|b_A - t\|_\pi^2 + \frac{8}{n}\mathrm{Tr}\left[\Psi_m^{-1/2}\Psi_{m,\sigma^2}\Psi_m^{-1/2}\right] + \frac{c'}{n},$$

where c' is a constant depending on $\mathbb{E}(\varepsilon_1^4)$ and $\int b_A^4(x)f(x)dx$.

Note that the variance term must be replaced by an estimator in order to propose a model selection criterion (choice of m from the data), a new difficulty related to the inverse problem aspect arises.

4. Diffusion model

Now, let us get back to our original problem. The last step is about managing with dependent variables.

4.1. The price of dependence

We consider the set of assumptions

- **(A1)** $b, \sigma \in C^1(\mathbb{R})$ and there exists $L \geq 0$, such that, for all $x \in \mathbb{R}, |b'(x)| + |\sigma'(x)| \leq L$.
- **(A2)** The scale density $s(x) = \exp\left\{-2\int_0^x b(u)/\sigma^2(u)du\right\}$ satisfies $\int_{-\infty} s(x)dx = +\infty = \int^{+\infty} s(x)dx$ and the speed density $m(x) = 1/(\sigma^2(x)s(x))$ satisfies $\int_{-\infty}^{+\infty} m(x)dx = M < +\infty$.
- **(A3)** $X_0 = \eta$ has distribution $\pi(x)dx$ given by $\pi(x) = M^{-1}m(x)$.
- **(A4)** (X_t) is geometrically β-mixing.

Assumption **(A1)**, ensures that Equation (1) has a unique strong solution adapted to the filtration $(\mathcal{F}_t = \sigma(\eta, W_s, s \leq t), t \geq 0)$. The functions b, σ have linear growth:

$$\exists K, \forall x \in \mathbb{R}, |b(x)| + |\sigma(x)| \leq K(1 + |x|). \tag{21}$$

The additional assumption **(A2)**, implies that Equation (1) admits a unique invariant probability $\pi(x)dx$. Lastly, under **(A3)**, (X_t) is strictly stationary and ergodic. Assumption **(A4)** means that there exist constants $K > 0$ and $\theta > 0$ such that:

$$\beta_X(t) \leq Ke^{-\theta t}, \tag{22}$$

where $\beta_X(t)$ denotes the β-mixing coefficient of (X_t).

The previous results have to be extended by replacing X_i by $X_{i\Delta}$ and Y_i by $Y_{i\Delta} = (X_{(i+1)\Delta} - X_{i\Delta})/\Delta$. The set $\Omega_m(\delta)$ is defined as previously to compare the empirical norm to its expectation. We can prove the result:

Proposition 4.1. *Assume* **(A1)**-**(A4)** *(thus $(X_{i\Delta})_i$ is strictly stationary and geometrically β-mixing i.e. $\beta(i) = \beta_X(i\Delta) \leq Ke^{-\theta i\Delta}$ for some constants $K > 0, \theta > 0$). Assume that Ψ_m is invertible. For all $\delta \in [0, 1]$*

$$\begin{aligned}\mathbb{P}(\Omega_m(\delta)^c) &= \mathbb{P}\left[\|\Psi_m^{-1/2}\widehat{\Psi}_m\Psi_m^{-1/2} - \mathrm{Id}_m\|_{\mathrm{op}} \geq u\right] \\ &\leq 4m\exp\left(-\frac{n\Delta\theta\, c(u)}{12\log(n\Delta)\ L(m)(\|\Psi_m^{-1}\|_{\mathrm{op}} \vee 1)}\right) + \frac{\theta}{6(n\Delta)^5},\end{aligned}$$

where $c(\delta) = \delta + (1-\delta)\log(1-\delta)$.

We can see that, under geometrical mixing, the cost of dependency is a $\log(n\Delta)$ in the exponential and a negligible additive term. The extension from independent to mixing is based on coupling method and Berbee's Lemma as presented in Viennet (1997).

4.2. Diffusion risk bound

Among the difficulties, we can notice that we have to consider the truncated version of $\hat{b}_m$ for both empirical and integrated risks. Moreover, the cutoff involves additional log terms and change in constants:

$$\widetilde{b}_m = \hat{b}_m \mathbf{1}_{\{L(m)(\|\widehat{\Psi}_m^{-1}\|_{\mathrm{op}} \vee 1) \leq \mathfrak{c}^\star n\Delta/\log^2(n\Delta)\}}, \quad \mathfrak{c}^\star = \frac{\theta}{C_0} \tag{23}$$

where C_0 is a numerical constant, $C_0 \geq 72$. For geometric β-mixing, the price to pay for dependency is an additional log term in the bound, and the fact that the constant $\mathfrak{c}^\star$ is now unknown. In practice, we will take $n\Delta$ large enough, a constant equal to 1 and a cutoff equal to $n\Delta/\log^{2+\epsilon}(n\Delta)$ for some $\epsilon > 0$.

Proposition 4.2. *Let $(X_{i\Delta})_{1\le i\le n}$ drawn from the diffusion model under assumptions* **(A1)-(A4)**, $\mathbb{E}(\eta^4) < +\infty$. *Take $\Delta = \Delta_n \to 0$ and $n\Delta \to +\infty$ when $n \to +\infty$.*
Consider the estimator $\tilde{b}_m$ of b_A. Then for m such that

$$L(m)(\|\Psi_m^{-1}\|_{\rm op} \vee 1) \le \frac{\mathfrak{c}^\star n\Delta}{2\log^2(n\Delta)}, \tag{24}$$

with $\mathfrak{c}^\star$ given in (23), we have

$$\begin{aligned}
\mathbb{E}\big[\|\widetilde{b}_m - b\|_n^2\big] &\le 7 \inf_{t\in S_m} \|b_A - t\|_\pi^2 + \frac{64}{n\Delta}\mathrm{Tr}\left[\Psi_m^{-1/2}\Psi_{m,\sigma^2}\Psi_m^{-1/2}\right] + c_1\Delta + \frac{c_2}{n\Delta},\\
\mathbb{E}\big[\|\widetilde{b}_m - b\|_\pi^2\big] &\le c_3\left(\inf_{t\in S_m} \|b_A - t\|_\pi^2 + \frac{1}{n\Delta}\mathrm{Tr}\left[\Psi_m^{-1/2}\Psi_{m,\sigma^2}\Psi_m^{-1/2}\right] + \Delta + \frac{1}{n\Delta}\right),
\end{aligned}$$

where c_1, c_2, c_3 are positive constants.

Note that condition (24) is similar to condition (15) with n replaced by $n\Delta$ and a log term which becomes $\log^2$ due to dependency. Still compared to the independent case, there is also a loss in the multiplicative constants of the two bounds.

Let us comment on the diffusion bounds. There are three main differences in comparison with our 2007-result (Comte *et al* (2007)):

- the estimator is truncated with a random cutoff,
- the variance order is new and more general,
- the stability condition (24) is new and expressed with respect to the basis at hand. This is what allows to estimate the constraint to define the truncation in (23).

Of course, we can distinguish special cases

- If σ bounded on A, then the bounds on $\mathrm{Tr}\left[\Psi_m^{-1/2}\Psi_{m,\sigma^2}\Psi_m^{-1/2}\right]$ given in Proposition (3.2) lead to a variance of order $m/(n\Delta)$ as in Comte *et al.* (2007).
- Otherwise we have, under $L(m) \le c_\varphi^2\, m$,

$$\mathrm{Tr}\left[\Psi_m^{-1/2}\Psi_{m,\sigma^2}\Psi_m^{-1/2}\right] \le c_\varphi^2 \mathbb{E}[\sigma_A^2(X_1)] m \|\Psi_m^{-1}\|_{\rm op}.$$

- As in the independent case, if A compact and $\pi(x) \ge \pi_{\min}$ for all $x \in A$, we have $\|\Psi_m^{-1}\|_{\rm op} \le 1/\pi_{\min}$ and the restriction (24) on m reduces to the simpler condition $m \le cn\Delta/\log^2(n\Delta)$, as in Comte *et al.* (2007).

5. Model selection and adaptive result

For sake of brevity, we present the model selection procedure in the diffusion context directly.

5.1. Procedure of selection

Now, we aim at selecting an adequate value for the dimension of the projection space, and the procedure should rely on avalaible data. For this, we consider the following collection of models

$$\mathcal{M}_{n\Delta} = \left\{ m \in \mathbb{N}, \quad c_\varphi^2 m\, (\|\Psi_m^{-1}\|_{\rm op}^2 \vee 1) \le \frac{\mathfrak{d}}{4}\frac{n\Delta}{\log^2(n\Delta)} \right\}, \text{ where } \mathfrak{d} = \frac{\theta}{C_0\,(\|\pi\|_\infty + \frac{1}{3})} \tag{25}$$

and C_0 is a numerical constant.

The stability condition here has to be reinforced: this is due to the fact that the problem is very difficult to handle, from the theoretical point of view. Moreover, we have to obtain an additional control of $\|\widehat{\Psi}_m - \Psi_m\|_{\text{op}}$, and this requires more constraints than the bound on $\mathbb{P}(\Omega_m(\delta)^c)$.

Proposition 5.1. (i) *Independent case. Let $X_1, \dots, X_n$ be i.i.d. with common density π such that $\|\pi\|_\infty < \infty$. Then for all $u > 0$*

$$\mathbb{P}\left[\|\Psi_m - \widehat{\Psi}_m\|_{\text{op}} \geq u\right] \leq 2m \exp\left(-\frac{n\,u^2/2}{c_\varphi^2\, m\, (\|\pi\|_\infty + u/3)}\right).$$

(ii) *Diffusion (dependent) case. Assume* **(A1)**-**(A4)** *(thus $(X_{i\Delta})_i$ is strictly stationary and geometrically β-mixing i.e. $\beta(i) = \beta_X(i\Delta) \leq Ke^{-\theta i\Delta}$ for some constants $K > 0, \theta > 0$). If in addition π is upper bounded (i.e. $\|\pi\|_\infty < +\infty$), then, for all $u > 0$,*

$$\mathbb{P}\left[\|\Psi_m - \widehat{\Psi}_m\|_{\text{op}} \geq u\right] \leq 4m \exp\left(-\frac{n\Delta\theta u^2/2}{12L(m)\log(n\Delta)\,(\|\pi\|_\infty \vee 1 + 2u/3)}\right) + \frac{\theta}{6(n\Delta)^5}.$$

For model selection we work with the collection of nested spaces S_m and assume that the basis $(\varphi_0, \dots, \varphi_{m-1})$ of S_m satisfies

$$\|\sum_{j=0}^{m-1} \varphi_j^2\|_\infty \leq c_\varphi^2 m \quad \text{for} \quad c_\varphi^2 > 0 \quad \text{a constant.} \tag{26}$$

Now, we can define the final estimator. We consider the empirical counterpart of $\mathcal{M}_{n\Delta}$, namely $\widehat{\mathcal{M}}_{n\Delta}$, a random collection of models defined by

$$\widehat{\mathcal{M}}_{n\Delta} = \left\{m \in \mathbb{N}, \quad c_\varphi^2 m\,(\|\widehat{\Psi}_m^{-1}\|_{\text{op}}^2 \vee 1) \leq \mathfrak{d}\frac{n\Delta}{\log^2(n\Delta)}\right\},$$

with $\mathfrak{d}$ defined in (25). We select

$$\hat{m} = \arg\min_{m \in \widehat{\mathcal{M}}_{n\Delta}} \left\{-\|\hat{b}_m\|_n^2 + \kappa\mathbb{E}[\sigma^2(X_0)]\frac{c_\varphi^2 m\|\widehat{\Psi}_m^{-1}\|_{\text{op}}}{n\Delta}\right\}$$

and study $\hat{b}_{\hat{m}}$.

The proposed criterion follows from standard approximations, which make the procedure likely to perform an automatic bias-variance tradeoff:

- The squared bias term $\|b_A - b_m^\pi\|_\pi^2 = \|b_A\|_\pi^2 - \|b_m^\pi\|_\pi^2$ where b_m^π is the $\mathbb{L}^2(A, \pi(x)dx)$-orthogonal projection of b on S_m. This implies that $-\|\hat{b}_m\|_n^2$ approximates the squared bias $(-\|b_m^\pi\|_\pi^2)$, up to an additive constant $(\|b_A\|_\pi^2)$.
- The term $\mathbb{E}[\sigma^2(X_0)](m\|\widehat{\Psi}_m^{-1}\|_{\text{op}})/(n\Delta)$ has the order of an upper bound on the variance term, which can be estimated.

In the independent regression cases, the term $\|\hat{b}_m\|_n^2$ is unchanged, $n\Delta$ is replaced by n and squares on log terms disappear. In the homoscedastic case, the penalty is equal to $\sigma_\varepsilon^2 m/n$, that is simpler (no random matrix) and however sharper.

5.2. General risk bound

We can prove the following result.

Theorem 5.1. *Let $(X_{i\Delta})_{0 \leq i \leq n+1}$ be observations from the diffusion model under* **(A1)**-**(A4)**. *Assume that condition (26) holds, $\|\pi\|_\infty < +\infty$ and $\mathbb{E}(\eta^8) < +\infty$. Then, there exists a numerical constant κ_0 such that for*

$\kappa \geq \kappa_0$,

$$\mathbb{E}\big[\|\hat{b}_{\hat{m}} - b_A\|_\pi^2\big] \leq C \inf_{m \in \mathcal{M}_{n\Delta}} \left(\inf_{t \in S_m} \|b_A - t\|_\pi^2 + \kappa \mathbb{E}[\sigma^2(X_0)] \frac{c_\varphi^2 m \|\Psi_m^{-1}\|_{\mathrm{op}}}{n\Delta} \right) + C_1' \Delta + \frac{C_2' \log^2(n\Delta)}{n\Delta}$$

where C is a numerical constant and C_1', C_2' are constants depending on π, b, σ.

Some practical considerations are in order. In the definition of $\widehat{\mathcal{M}}_{n\Delta}$, the constant $\mathfrak{d}$ is replaced by 1, and the $\log^2(n\Delta)$ by $\log^{2+\epsilon}(n\Delta)$, which preserves the result for $n\Delta$ large enough. The term $\mathbb{E}[\sigma^2(X_0)]$ is replaced by a residual least squares estimator $(1/n)\sum_{i=1}^n [Y_{i\Delta} - \hat{b}_{m_n}(X_{i\Delta})]^2$ for m_n an arbitrary dimension. A theoretical study of this step is done in Baraud (2002). The constant κ is calibrated by preliminary simulation experiments, as usual for model selection methods.

Let us say what is new here. First, we provide a general result with no support constraint. Second, the result requires only moment conditions for σ. Lastly, the collection of models $\widehat{\mathcal{M}}_{n\Delta}$ is new and random, and contains implicitly the random truncation of the estimator.

Examples of simulation results are given in the papers [9], [8] and [10]. They show that the method works with Laguerre and Hermite bases, and that in some cases $\|\widehat{\Psi}_m^{-1}\|_{\mathrm{op}}$ can increase very fast, which considerably reduces the number of models satisfying the stability condition. It seems however that this is also associated in such cases with very good estimators even for small dimensions.

6. Concluding Remarks

We have presented a generalization of nonparametric least squares procedure for regression function estimation, which is from the theoretical point of view compatible with non compactly supported bases and non bounded volatilities, and remains from practical point of view, fast and simple. We have introduced a new random cutoff in the definition of the estimator and a new stability condition. We propose an associated model selection procedure which relies importantly on methods used to handle inverse problems.

There remains clearly open questions.

Optimality results are available for homoscedastic regression, but remains an open problem in heteroscedastic regression, under our general assumptions. Gaïffas (2005, 2007) probably initiated a possible way to explore the question.

Another question is related to penalization. In the heteroscedastic context, the variance term is proportional to $\mathrm{Tr}\left[\Psi_m^{-1/2}\Psi_{m,\sigma^2}\Psi_m^{-1/2}\right]$ divided by n in the i.i.d. context, and $n\Delta$ in the diffusion setting. This term should provide the correct value of the penalty in model selection step. For diffusion models, we used instead the bound $\mathbb{E}[\sigma^2(X_0)]c_\varphi^2 m\|\widehat{\Psi}_m^{-1}\|_{\mathrm{op}}/(n\Delta)$ but it is probably too large in practice. Numerical tests show that separating the matrices $\widehat{\Psi}_m^{-1}$ and $\widehat{\Psi}_{m,\sigma^2}$ may not be a good idea because a kind of compensation seems to happen in their product.

In the i.i.d. heteroscedastic case, further investigations lead us to finally obtain risk bounds for a penalty proportional to $m\|\widehat{\Psi}_m^{-1}\widehat{\Psi}_{m,\sigma^2}\|_{\mathrm{op}}/n$ with $[\widehat{\Psi}_{m,\sigma^2}]_{j,k} = n^{-1}\sum_{i=1}^n \varphi_j(X_i)\varphi_k(X_i)\sigma^2(X_i)$ which is numerically better but still requires the knowledge of σ. In practice, it is computed by replacing $\sigma^2(X_i)$ by $(Y_i - \hat{b}_{m^\star}(X_i))^2$ with $m^\star$ defined as $\widehat{M}_n - 2$ where $\widehat{M}_n$ is the maximal element of $\widehat{\mathcal{M}}_n$, the random collection of models considered in the regression setting. This is the beginning of further theoretical questions on the topic.

References

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

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

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

[4] Baraud, Y., Comte, F. and Viennet, G. (2001b) Model selection for (auto)-regression with dependent data. *ESAIM P& S* **5**, 33-49.

[5] Cohen, A., Davenport, M.A. and Leviatan, D. (2013). On the stability and acuracy of least squares approximations. *Found. Comput. math.* **13**, 819-834.
[6] Comte, F., Genon-Catalot, V. and Rozenholc, Y. (2007) Penalized nonparametric mean square estimation of the coefficients of diffusion processes. *Bernoulli* **13**, 514-543.
[7] Comte, F. and Genon-Catalot, V. (2018a) Laguerre and Hermite bases for inverse problems. *Journal of the Korean Statistical Society*, **47**, 273-296.
[8] Comte, F. and Genon-Catalot, V. (2018b) Drift estimation on non compact support for diffusion models. Preprint hal-01916503.
[9] Comte, F. and Genon-Catalot, V. (2019a) Regression function estimation on non compact support as a partly inverse problem. To appear in *The Annals of the Institute of Statistical Mathematics*, https://doi.org/10.1007/s10463-019-00718-2.
[10] Comte, F. and Genon-Catalot, V. (2019b) Regression function estimation on non compact support in an heteroscedastic model, Preprint hal-02009555.
[11] Gaïffas, S. (2005) Convergence rates for pointwise curve estimation with a degenerate design. *Math. Methods Statist.* **14**, 1-27.
[12] Gaïffas, S. (2007) Sharp estimation in sup norm with random design. *Statist. Probab. Lett.* **77**, 782-794.
[Tropp, 2012] Tropp, J. A. (2012). User-friendly tail bounds for sums of random matrices. *Found. Comput. Math.*, 12(4):389–434.
[13] Viennet, G. (1997). Inequalities for absolutely regular processes: application to density estimation. *Probab. Theory Relat. Fields* **107**, 467-492.