Neuvièmes journées de Statistique Mathématique : Learning (and) statistics with Talagrand.

A story of $\nu_n(h)$ through nonparametric inference for regression and diffusions.

Fabienne Comte *MAP5, CNRS 8145, Université Paris Cité*

7-8-9 Janvier 2026

Plan of the talk

A twenty-years small story.

1. Regression and diffusion models and their least-squares contrasts
2. Norm equivalence: from Bernstein Inequality to Tropp Chernov deviation (2012)
3. Risk bound for adaptive estimator and the Talagrand (1996) deviation Inequality.

Projection estimators and model selection

Choose a basis φ_j such that $\langle \varphi_j, \varphi_k \rangle = \int_A \varphi_j(x)\varphi_k(x)dx = \delta_{j,k}$ and set

$$S_m = \mathrm{Vect}(\varphi_1, \dots, \varphi_m), \quad \mathrm{Support}(\varphi_j) = A \subset \mathbb{R}$$

Let b denote the function we want to estimate on A ($b_A = b\mathbf{1}_A$).

- Define

$$\boxed{\widehat{b}_m = \sum_{j=1}^{m} \widehat{a}_j \varphi_j}, \quad \widehat{a}_j \text{ computed from the observations.}$$

  and prove a bound on $\mathbb{E}\left(\|\widehat{b}_m - b_A\|_n^2\right) := \mathbb{E}(m)$ and on $\mathbb{E}\left(\|\widehat{b}_m - b_A\|_f^2\right)$,
- Next, choose m from the observations:

$$\widehat{m} = \arg\min_{m \in \mathcal{M}_n} \mathrm{Crit}(m), \quad \mathcal{M}_n \subset \mathbb{N}.$$

  and prove a bound on $\mathbb{E}\left(\|\widehat{b}_{\widehat{m}} - b_A\|_n^2\right)$ of order

$$C \inf_{m \in \mathcal{M}_n} \mathbb{E}(m) + \text{ negligible terms.}$$

A cascade of models

1. Usual (homoscedatic) regression model

$$\mathbf{Y_i} = \mathbf{b}(\mathbf{X_i}) + \sigma\varepsilon_\mathbf{i}, \; i = 1, \ldots, n, \text{ with } X_i \text{ i.i.d.}, \varepsilon_i \text{ i.i.d. } (0, 1),$$

and $(X_i)_i \perp (\varepsilon_i)_i$. **Observations** $(X_i, Y_i)_{1 \leqslant i \leqslant n}$.

2. Autoregressive model.

$$\mathbf{X_{i+1}} = \mathbf{b}(\mathbf{X_i}) + \varepsilon_\mathbf{i}, \; i = 1, \ldots, n, \text{ with } \varepsilon_i \text{ i.i.d. } (0, 1).$$

The X_i's are not independent and neither the sequences $(X_i)_i$ and (ε_i). **Observations** $(X_i)_{1 \leqslant i \leqslant n+1}$.

3. Diffusion model

$$\mathbf{dX_t} = \mathbf{b}(\mathbf{X_t})\mathbf{dt} + \sigma(\mathbf{X_t})\mathbf{dW_t} \text{ with } X_0 \sim \mu$$

and W_t a standard Brownian motion. **Observations** $(X_{i\Delta})_{1 \leqslant i \leqslant n}$.
Δ small and $n\Delta$ large (high frequency data)

4. Diffusion models in FDA spirit.

$$\mathbf{dX_t^{(i)}} = \mathbf{b}(\mathbf{X_t^{(i)}})\mathbf{dt} + \sigma(\mathbf{X_t^{(i)}})\mathbf{dW_t^{(i)}}, \; X_0^{(i)} = x_0,$$

$W^{(i)}$ independent standard brownian motions, T fixed. **Observations** n independent paths ; $(X_t^{(i)})_{t \in [0,T]}, i = 1, \ldots, n$.

Link with regression (1)

Model 1 = standard regression, i.i.d. variables, unbounded noise.

Model 2 = autoregression,

- variables X_i can be identically distributed
- no independence between the X_i's $\Rightarrow$ mixing to handle dependency,
- Sequences $(\varepsilon_i)_i$ and $(X_i)_i$ no longer independent $\Rightarrow$ conditioning by $X = x$ no longer possible,
- martingale properties
- unbounded noise.

Link with regression (2)

Model 3. We define

$$Y_{i,\Delta} = \frac{X_{(i+1)\Delta} - X_{i\Delta}}{\Delta},$$

and we have

1. Approximation 1

$$Y_{i,\Delta} = b(X_{i\Delta}) + \underbrace{\frac{1}{\Delta}\int_{i\Delta}^{(i+1)\Delta} \sigma(X_s)dW_s}_{:=Z_{i,\Delta}} + \underbrace{\frac{1}{\Delta}\int_{i\Delta}^{(i+1)\Delta} [b(X_s) - b(X_{i\Delta})]ds}_{R^{(1)}_{i,\Delta}}.$$

2. Approximation 2

$$Y_{i,\Delta} = b(X_{i\Delta}) + \underbrace{\sigma(X_{i\Delta})\frac{W_{(i+1)\Delta} - W_{i\Delta}}{\Delta}}_{X,W \text{ separated}} + \underbrace{\frac{1}{\Delta}\int_{i\Delta}^{(i+1)\Delta} [b(X_s) - b(X_{i\Delta})]ds}_{R^{(1)}_{i,\Delta}}$$

$$+ \underbrace{\frac{1}{\Delta}\int_{i\Delta}^{(i+1)\Delta} [\sigma(X_s) - \sigma(X_{i\Delta})]dW_s}_{:=R^{(2)}_{i,\Delta}}.$$

Link with regression (3)

Model 3.

- Approximation 1 with martingale tools only,
- Approximation 2 with Talagrand-type deviation.

Model 4. Back to independence: n independent complete paths are available.

Definition of the estimator

$$\widehat{b}_m = \arg\min_{h\in S_m} \gamma_n(h)$$

with

1. **Model 1.** $\gamma_n(h) = \underbrace{\frac{1}{n}\sum_{i=1}^{n} h^2(X_i)}_{:=\|h\|_n^2} - \frac{2}{n}\sum_{i=1}^{n} Y_i h(X_i)$
2. **Model 2.** $\gamma_n(h) = \underbrace{\frac{1}{n}\sum_{i=1}^{n} h^2(X_i)}_{:=\|h\|_n^2} - \frac{2}{n}\sum_{i=1}^{n} X_{i+1} h(X_i)$
3. **Model 3.** $\gamma_n(h) = \underbrace{\frac{1}{n}\sum_{i=1}^{n} h^2(X_{i\Delta})}_{:=\|h\|_n^2} - \frac{2}{n}\sum_{i=1}^{n} Y_{i,\Delta} h(X_{i\Delta})$
4. **Model 4.** $\gamma_n(h) = \underbrace{\frac{1}{nT}\sum_{i=1}^{n} \int_0^T h^2(X_s^{(i)})ds}_{:=\|h\|_n^2} - \frac{2}{n}\sum_{i=1}^{n} \int_0^T h(X_s^{(i)})dX_s^{(i)}.$

Definition of the estimator

Let $\widehat{\Psi}_m = (\langle \varphi_j, \varphi_k \rangle_n)_{1 \leqslant j,k \leqslant m} = \frac{1}{n} \widehat{\Phi}_m^T \widehat{\Phi}_m, \quad \widehat{\Phi}_m := (\varphi_j(X_i))_{1 \leqslant i \leqslant n, 1 \leqslant j \leqslant m}$

$$\langle \varphi_j, \varphi_k \rangle_n = \frac{1}{n} \sum_{i=1}^{n} \varphi_j(X_i) \varphi_k(X_i).$$

Assume that $\widehat{\Psi}_m$ is invertible, then

$$\widehat{b}_m = \sum_{j=1}^{m} \widehat{a}_j \varphi_j, \quad \boxed{\widehat{\mathbf{a}}_m := \begin{pmatrix} \widehat{a}_1 \\ \vdots \\ \widehat{a}_m \end{pmatrix} = \widehat{\Psi}_m^{-1} \mathbf{Z}_m}$$

and

$$\mathbf{Z}_m = \begin{cases} \frac{1}{n} \widehat{\Phi}_m^T \mathbf{Y}, \quad \mathbf{Y} = (Y_1, \dots, Y_n)^T \\ \\ \frac{1}{nT} \sum_{i=1}^{n} \int_0^T \varphi_j(X_s^{(i)}) dX_s^{(i)} \end{cases}$$

Risk bound in empirical norm on a fixed model

Sometimes almost "free", with projection arguments.

Model 1. Let $(X_i, Y_i)_{1\leqslant i\leqslant n}$ be observations drawn from model (1) and set $b_A = b\mathbf{1}_A$. Assume that $b_A \in \mathbb{L}^2(A, f(x)dx)$ and that $\widehat{\Psi}_m$ is a.s. invertible. Consider the least squares estimator $\widehat{b}_m$ of b, defined as the contrast minimizer. Then

$$\begin{aligned}\mathbb{E}\big[\|\widehat{b}_m - b_A\|_n^2\big] &= \mathbb{E}\left(\inf_{h\in S_m}\|h-b_A\|_n^2\right) + \sigma_\varepsilon^2\frac{m}{n}, \\ &\leqslant \inf_{h\in S_m}\underbrace{\left[\int (b_A-h)^2(x)f(x)dx\right]}_{\|b_A-h\|_f^2} + \sigma_\varepsilon^2\frac{m}{n}.\end{aligned}$$

Questions arise to handle $\mathbb{E}\big[\|\widehat{b}_m - b_A\|_f^2\big]$ and for $\mathbb{E}\big[\|\widehat{b}_{\widehat{m}} - b_A\|_n^2\big]$.

Tropp Chernov Deviation inequality

Risk bound in integrated norm on a fixed model

Recall hat f denotes the density of X_1 and write for $b_m \in S_m$,

$$\mathbb{E}\big[\|\widehat{b}_m - b_A\|_f^2\big] \leqslant 2(\big[\|b_m - b_A\|_f^2\big] + \mathbb{E}\big[\|\widehat{b}_m - b_m\|_f^2\big]).$$

To handle $\mathbb{E}\big[\|\widehat{b}_m - b_m\|_f^2\big]$, we need a comparison of

$$\|h\|_n^2 \quad \text{and} \quad \int h^2(x)f(x)dx = \|h\|_f^2, \quad \boxed{\text{for } t \in S_m},$$

or equivalently of

$$\widehat{\Psi}_m = (\langle \varphi_j, \varphi_k \rangle_n)_{1 \leqslant j,k \leqslant m} \quad \text{to} \quad \Psi_m = \mathbb{E}(\widehat{\Psi}_m) = (\langle \varphi_j, \varphi_k \rangle_f)_{1 \leqslant j,k \leqslant m}.$$

First idea:

$$\|h\|_n^2 - \|h\|_f^2 = \frac{1}{n}\sum_{i=1}^{n}\left[h^2(X_i) - \mathbb{E}(h^2(X_i))\right].$$

Looks like a centered empirical process.

Key set to control

First strategies: **Bernstein inequalities** for

$$\nu_n(h^2) = \frac{1}{n}\sum_{i=1}^{n}[h^2(X_i) - \mathbb{E}(h^2(X_i))] = \|h\|_n^2 - \|h\|_f^2.$$

Quadratic! $\Rightarrow$ **ugly** computations and union bounds.

Define the set where the empirical and the $\mathbb{L}^2(A, f)$ norms are equivalent for functions in S_m:

$$\Omega_m(\delta) = \left\{ \sup_{h\in S_m,\ h\neq 0} \left| \frac{\|h\|_n^2}{\|h\|_f^2} - 1 \right| \leqslant \delta \right\}, \quad \text{for } \delta \in (0,1). \tag{1}$$

It holds that for Ψ_m invertible,

$$\Omega_m(\delta) = \left\{ \|\Psi_m^{-1/2}\widehat{\Psi}_m\Psi_m^{-1/2} - \mathrm{Id}_m\|_{\mathrm{op}} \leqslant \delta \right\}.$$

On $\Omega_m(\frac{1}{2})$, for a vector $x \in \mathbb{R}^m$,

$$x^t\widehat{\Psi}_m x \leqslant (3/2)x^t\Psi_m x \text{ and } x^t\widehat{\Psi}_m^{-1}x \leqslant 2x^t\Psi_m^{-1}x.$$

Theorem (Matrix Chernoff, Tropp (2012))

Theorem

Consider a finite sequence $\{\mathbf{U}_i\}$ of independent, random, self-adjoint matrices with dimension d. Assume that each random matrix satisfies

$$\mathbf{U}_i \succcurlyeq 0 \quad \text{and} \quad \lambda_{\max}(\mathbf{U}_i) \leqslant R \ \text{ almost surely.}$$

Define $\mu_{\min} := \lambda_{\min}(\sum_k \mathbb{E}(\mathbf{U}_i))$ *and* $\mu_{\max} := \lambda_{\max}(\sum_k \mathbb{E}(\mathbf{U}_i))$. *Then*

$$\mathbb{P}\left\{\lambda_{\min}\left(\sum_i \mathbf{U}_i\right) \leqslant (1-\delta)\mu_{\min}\right\} \leqslant d\left[\frac{e^{-\delta}}{(1-\delta)^{1-\delta}}\right]^{\mu_{\min}/R} \quad \text{for } \delta \in [0,1],$$

$$\mathbb{P}\left\{\lambda_{\max}\left(\sum_i \mathbf{U}_i\right) \geqslant (1+\delta)\mu_{\max}\right\} \leqslant d\left[\frac{e^{\delta}}{(1+\delta)^{1+\delta}}\right]^{\mu_{\max}/R} \quad \text{for } \delta \geqslant 0.$$

Apply Tropp-Chernov

$$\Psi_m^{-1/2}\widehat{\Psi}_m\Psi_m^{-1/2} = \frac{1}{n}\sum_{i=1}^n \Psi_m^{-1/2}\Xi_i\Psi_m^{-1/2}, \quad \Xi_i = \left(\varphi_j(X_i)\varphi_k(X_i)\right)_{1\leqslant j,k\leqslant m}.$$

$$\mathbf{U}_i = \frac{1}{n}\Psi_m^{-1/2}\Xi_i\Psi_m^{-1/2}, \quad \mathbb{E}(\mathbf{U}_i) = \frac{1}{n}\mathbf{I}d_m.$$

$$\Rightarrow \mu_{\min} = \mu_{\max} = 1$$

and

$$R = \frac{1}{n}L(m)\|\Psi_m^{-1}\|_{\rm op}$$

where

$$\sup_{x\in A}\sum_{j=1}^m \varphi_j^2(x) \leqslant L(m)(< +\infty).$$

Applied first by Cohen *et al.* (2013), Tropp-Chernov inequality provides the adequate inequality.

Proposition

Let $\widehat{\Psi}_m$, Ψ_m *be the* $m \times m$ *matrices defined in above and assume that* Ψ_m *is invertible. Then for all* $0 < \delta \leqslant 1$,

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

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

Stability condition

Fix $\delta = 1/2$. As a consequence, if m is such that **(stability condition)**:

$$\boxed{L(m)(\|\Psi_m^{-1}\|_{\rm op} \vee 1) \leqslant \frac{\mathfrak{c}(p)}{2}\frac{n}{\log(n)},} \tag{2}$$

with

$$\mathfrak{c}(p) = \frac{3\log(3/2) - 1}{p+1},$$

we obtain

$$\mathbb{P}\left[(\Omega_m(\frac{1}{2}))^c\right] \leqslant \frac{2}{n^p}.$$

Condition (2) refers to a deterministic matrix, suggests a cutoff

$$\tilde{b}_m = \hat{b}_m \mathbf{1}_{\hat{\Lambda}_m}, \quad \hat{\Lambda}_m = \left\{L(m)(\|\hat{\Psi}_m^{-1}\|_{\rm op} \vee 1) \leqslant \mathfrak{c}(p)\frac{n}{\log(n)}\right\}.$$

Proposition

Assume that $\mathbb{E}(\varepsilon_1^4) < +\infty$ *and* $b_A \in \mathbb{L}^4(A, f(x)dx)$*. Then for any* m *satisfying (2), we have*

$$\mathbb{E}\big[\|\widetilde{b}_m - b_A\|_f^2\big] \leqslant \left(1 + \frac{8\mathfrak{c}}{\log(n)}\right) \inf_{t \in S_m} \|b_A - t\|_f^2 + 8\sigma_\varepsilon^2 \frac{m}{n} + \frac{c}{n}, \qquad (3)$$

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

"Compact support" case

Case A compact and f lower bounded on A,

$$\forall x \in A, \quad f(x) \geqslant f_0.$$

Then

$$\|\Psi_m^{-1}\|_{\rm op} \leqslant \frac{1}{f_0}.$$

Condition (2) (stability) is fulfilled if

$$L(m) \leqslant f_0 \frac{\mathfrak{c}(p)}{2} \frac{n}{\log(n)}.$$

Better than first constraints given with $L(m) = m$, Baraud (2002), Baraud et al (2001).

$\|\widehat{\Psi}_m^{-1}\|_{\rm op}$ is an empirical version that avoids to estimate f_0!

What about models 2, 3, 4?

The matrix $\widehat{\Psi}_m$ only depends on the X_i or $X_{i\Delta}$ or $(X_t^{(i)})_{t\in[0,T]}$.

$\Rightarrow$ Possible to extend to dependent data, with coupling methods.

The story:

- First attempts in Baraud (2000), (2002) around the Gram matrix, Bernstein deviation, deterministic collection and compactly supported bases,
- Dependent case, Baraud *et al* (2000, 2001), SDEs Comte *et al.* (2007),
- Stability condition Cohen *et al.* (2013), (2019)
- Random collections for possibly non compact cases, Comte and Genon-Catalot (2020), and Tropp for dependent variables (2021). Recently improved by Yichuan Huang.

For the future:

- Tropp applied for **regular** collections of models, what if $L(m) = n$ and the cardinal of the collection is exponential?

- The non bounded case.
  Questions arise in FDA

$$\widehat{\Psi}_m = \left(\frac{1}{n} \sum_{i=1}^{n} \langle \varphi_j, X_i \rangle \langle \varphi_k, X_i \rangle \right)_{1 \leqslant j,k \leqslant m}$$

  No longer bounded (Computing R).

Talagrand deviation Inequality

Collection of models

Condition (2) also defines a collection of models, which will become random $(\Psi_m \longrightarrow \widehat{\Psi}_m)$:

$$\mathcal{M}_n = \left\{ m \in \{1, \ldots, n\}, L(m)(\|\Psi_m^{-1}\|_{\rm op} \vee 1) \leqslant \frac{\mathfrak{c}(p)}{2} \frac{n}{\log(n)} \right\}$$

$$\mathcal{M}_n^+ = \left\{ m \in \{1, \ldots, n\}, L(m)(\|\Psi_m^{-1}\|_{\rm op} \vee 1) \leqslant 2\mathfrak{c}(p) \frac{n}{\log(n)} \right\}$$

$$\widehat{\mathcal{M}}_n = \left\{ m \in \{1, \ldots, n\}, L(m)(\|\widehat{\Psi}_m^{-1}\|_{\rm op} \vee 1) \leqslant \mathfrak{c}(p) \frac{n}{\log(n)} \right\}$$

$$\widehat{m} = \arg \min_{m \in \widehat{\mathcal{M}}_n} \left(\gamma_n(\widehat{b}_m) + \mathrm{pen}(m) \right).$$

Decomposition of the contrast

$$\gamma_n(h) = \|h\|_n^2 - \frac{2}{n}\sum_{i=1}^n \underbrace{Y_i h(X_i)}_{Model1} \ / \ \underbrace{X_{i+1}h(X_i)}_{Model2} \ / \ \underbrace{Y_{i\Delta}h(X_{i\Delta})}_{Model3}$$

leading to

$$\gamma_n(h) - \gamma_n(\ell) = \|h-b\|_n^2 - \|\ell - b\|_n^2 - 2\nu_n(h-\ell) + \ \ residual,$$

$$\boxed{\nu_n(h) = \frac{1}{n}\sum_{i=1}^n \varepsilon_i h(X_i)}$$ $\Rightarrow h \mapsto \nu_n(h)$ is now linear.

Then

$$\widehat{m} = \arg\min_{m\in\widehat{\mathcal{M}}_n}\left(\gamma_n(\widehat{b}_m) + \text{pen}(m)\right)$$

implies that for all $m \in \widehat{\mathcal{M}}_n$

$$\gamma_n(\widehat{b}_{\widehat{m}}) + \text{pen}(\widehat{m}) \leqslant \gamma_n(\widehat{b}_m) + \text{pen}(m)$$

For all $m \in \widehat{\mathcal{M}}_n$

$$\|\widehat{b}_{\widehat{m}} - b\|_n^2 \leqslant \|\widehat{b}_m - b\|_n^2 + \text{pen}(m) + 2\nu_n(\widehat{b}_{\widehat{m}} - \widehat{b}_m) - \text{pen}(\widehat{m}).$$

Let $\Omega_n = \bigcap_m \Omega_m$. Then

$$\Xi_n = \left\{\mathcal{M}_n \subset \widehat{\mathcal{M}}_n \subset \mathcal{M}_n^+\right\} \text{ satisfies } \Omega_n \subset \Xi_n.$$

So, on Ω_n, for all $m \in \mathcal{M}_n$,

$$\|\widehat{b}_{\widehat{m}} - b\|_n^2 \leqslant \|\widehat{b}_m - b\|_n^2 + \text{pen}(m) + 2\nu_n(\widehat{b}_{\widehat{m}} - \widehat{b}_m) - \text{pen}(\widehat{m}),$$

and $\widehat{m} \in \mathcal{M}_n^+$. Then prove

$$\mathbb{E}\left[\left(\sup_{h \in S_{\widehat{m}}, \|h\|_f = 1} \nu_n^2(h) - p(\widehat{m})\right)_+ \mathbf{1}_{\Omega_n}\right] \lesssim \frac{C}{n}$$

with

$$\mathbb{E}\left[\left(\sup_{h \in S_{\widehat{m}}, \|h\|_f = 1} \nu_n^2(h) - p(\widehat{m})\right)_+ \mathbf{1}_{\Omega_n}\right] \leqslant \sum_{m \in \mathcal{M}_n^+} \mathbb{E}\left[\left(\sup_{h \in S_m, \|h\|_f = 1} \nu_n^2(h) - p(\widehat{m})\right)_+\right].$$

Theorem (Talagrand, Klein and Rio)

Theorem

Consider $\mathcal{F}$ a class at most countable of measurable functions, and $(X_i)_{i\in\{1,\dots,n\}}$ a indep random variables. For $f \in \mathcal{F}$, let

$$\nu_n(f) = \frac{1}{n}\sum_{i=1}^{n}(f(X_i) - \mathbb{E}[f(X_i)]).$$

Then for all $\epsilon > 0$,

$$\mathbb{E}\left[\left(\sup_{f\in\mathcal{F}}|\nu_n(f)|^2 - 2(1+2\epsilon)H^2\right)_+\right]$$

$$\leqslant \quad \frac{4}{b}\left[\frac{v}{n}\exp\left(-b\epsilon\frac{nH^2}{v}\right) + \frac{49M_1^2}{bC^2(\epsilon)n^2}\exp\left(-\frac{\sqrt{2}bC(\epsilon)\sqrt{\epsilon}}{7}\frac{nH}{M_1}\right)\right]$$

$$\sup_{f\in\mathcal{F}}\|f\|_\infty \leqslant M_1,\ \mathbb{E}[\sup_{f\in\mathcal{F}}|\nu_n(f)|] \leqslant H,\ \text{and}\ \sup_{f\in\mathcal{F}}\frac{1}{n}\sum_{i=1}^{n}\mathrm{Var}(f(X_i)) \leqslant v.$$

and $C(\epsilon) = (\sqrt{1+\epsilon}-1)\wedge 1$, and $b = \frac{1}{6}$.

What is needed for applying Talagrand Inequality?

- A **countable** set of functions.
  By density arguments, $\mathcal{F} \to$ unit ball of a linear normed space, as $f \to \nu_n(f)$ is continuous and $\mathcal{F}$ contains a countable dense family.

- $f(e, x) = eh(x)$ should be **bounded**, and this requires the noise ε to be bounded.

- The variables are **independent**.
  They may be so, or not...
  Variance inequalities for mixing sequences : Doukhan (1994), Doukhan, Massart, Rio (1995), Rio (2000).
  Stationary dependency with β-mixing can be handled by coupling methods (Berbee (1979), Viennet (1997)).

Application in regression with bounded noise.

Model 1. Barron, Birgé and Massart (1999), bounded noise:

$$\forall i \in \{1, \dots, n\}, \quad |\varepsilon_i| \leqslant K.$$

and compact set, to obtain a result.

$$\mathbb{H}^2 = \sigma_\varepsilon^2 \frac{m}{n} \ \propto \ \text{pen}(m), \quad v = \sigma_\varepsilon^2, \quad M_1 = K\sqrt{m}/\sqrt{f_0}.$$

$$\mathbb{E}\left[\left(\sup_{h \in S_m, \|h\|_f = 1} \nu_n^2(h) - 2(1+2\epsilon)\mathbb{H}^2\right)_+\right] \leqslant \frac{C_1}{n}\left(e^{-C_2 m} + \frac{K^2 m}{f_0 n} e^{-C_3\sqrt{f_0}\sqrt{n}/K}\right).$$

Sum over $m \in \mathcal{M}_n$ are of order $1/n$.

Baraud (2000) proposes an integration of the deviation inequality.

Naive way:

$$\nu_n(h) = \nu_{n,1}(h) + \nu_{n,2}(h), \quad \nu_{n,1}(h) = \frac{1}{n}\sum_{i=1}^{n} \varepsilon_{i,1} h(X_i)$$

with

$$\varepsilon_{i,1} = \varepsilon_i \mathbf{1}_{|\varepsilon_i| \leqslant K} - \mathbb{E}(\varepsilon_i \mathbf{1}_{|\varepsilon_i| \leqslant K}).$$

Take $K = K(n)$, and $K(n) = \mathbf{c}\sqrt{n}/\log(n)$, for collection of models with cardinality n, in the previous

$$\mathbb{E}\left[\left(\sup_{h\in S_m, \|h\|_f=1} \nu_{n,1}^2(h) - \underbrace{2(1+2\epsilon)\mathbb{H}^2}_{\sim \text{pen}(m)}\right)_+\right] \leqslant \frac{C_1}{n}\left(e^{-C_2 m} + \frac{K^2 m}{f_0 n} e^{-C_3\sqrt{f_0 n}/K}\right).$$

$\Rightarrow$ keep an order $1/n$.

$$\mathbb{E}\left(|\varepsilon_1|^2 \mathbf{1}_{|\varepsilon_1| > K(n)}\right) \leqslant \frac{\mathbb{E}\left(|\varepsilon_1|^{2+q}\right)}{K(n)^q} = \frac{\log^q(n)}{\mathbf{c}^q n^q} \mathbb{E}\left(|\varepsilon_1|^{2+q}\right).$$

Choose q, deduce the required moment condition on the noise.

Non compact case: replace f_0^{-1} by $\|\Psi_m^{-1}\|_{\rm op}$ with stability constraint

$$L(m)(\|\Psi_m^{-1}\|_{\rm op} \vee 1) \leqslant \frac{\mathfrak{c}(p)}{2} \frac{n}{\log(n)}$$

and get

$$\mathbb{E}\left[\left(\sup_{h\in S_m, \|h\|_f=1} \nu_{n,1}^2(h) - 2(1+2\epsilon)\mathbb{H}^2\right)_+\right]$$
$$\leqslant \quad \frac{C_1}{n}\left(e^{-C_2 m} + \frac{K^2}{\log(n)} e^{-C_3\sqrt{mL(m)}\sqrt{\log(n)/K^2}}\right).$$

Need K of order $\sqrt{\log(n)}$ (or $\log(n)$) $\Rightarrow$ **Subgaussian** types conditions on the ε_i's.

Dependent variables.

For $\mathcal{U}$ and $\mathcal{V}$ two σ-fields,

$$\beta(\mathcal{U},\mathcal{V}) = \frac{1}{2}\sup\left\{\sum_i\sum_j |\mathbb{P}(U_i)\mathbb{P}(V_j) - \mathbb{P}(U_i\cap V_j)|\right\}$$

where the supremum is taken over all finite partitions $(U_i)_{i\in I}$ and $(V_j)_{j\in J}$ of Ω, which are respectively $\mathcal{U}$ and $\mathcal{V}$ measurable.

For Y a strictly stationary sequence,

$$\beta_k(Y) = \beta(\mathcal{F}_0, \mathcal{G}_k) \text{ with } \mathcal{F}_0 = \sigma(Y_i, i \leqslant 0) \text{ and } \mathcal{G}_k = \sigma(Y_i, i \geqslant k)$$

First example of handling dependent (β-mixing) variables for model selection: Viennet (1997) for density estimation, empirical process

$$\frac{1}{n}\sum_{i=1}^n [f(X_i) - \mathbb{E}(f(X_i))].$$

Penalization from a variance bound for $\mathbb{H}^2$ involves the mixing coefficients.

Dependent variables in autoregression.

Here $X_{i+1} = b(X_i) + \varepsilon_i$, the dependent variables are

$$u_i := (X_i, \varepsilon_i),$$

with (X_i and ε_i) independent for fixed i.

Under conditions on b and the initial value, the sequence X_i and then u_i is stationary and β-mixing, i.e. $\beta_k(u), \beta_k(X) \to 0$ when $k \to +\infty$.

What is coupling, for β-mixing sequences?
It is a method for building a twin sequence $u_i^\star = (X_i^\star, \varepsilon_i^\star)$ independent by blocks of size $q = q(n)$ with a price of substitution of order the mixing coefficient β_q. Talagrand applied to the $u_i^\star$.

Specificity of regression: the variance term in the Talagrand $\mathbb{H}^2$ is computed without the mixing inequalities, so with the same penalty as in the independent case.

Martingale strategy

Useful remark:

$$\sum_{i=1}^{n} \varepsilon_i h(X_i) \text{ is a } \mathcal{F}_n\text{-martingale}$$

with $\mathcal{F}_n = \sigma(X_i, i \leqslant n+1, \varepsilon_i, i \leqslant n)$.

Sub-gaussian noise assumption: $\mathbb{E}(e^{u\varepsilon_1}) \leqslant \exp\left(\frac{u^2 s^2}{2}\right), \forall u \in \mathbb{R}$.

$$\mathbb{P}\left(\sum_{i=1}^{n} \varepsilon_i h(X_i) \geqslant n\mathtt{a}, \;\; \|h\|_n^2 \leqslant v^2\right) \leqslant \exp\left(-\frac{n\mathtt{a}^2}{2s^2v^2}\right).$$

The deviation of the supremum is deduced by a $\mathbb{L}_2 - \mathbb{L}_\infty$-chaining method, which became universal.

Interest: handle large (irregular) collections of models. No mixing coefficients here (but still for Tropp).

Diffusion models

Recall that diffusion model are defined by

$$\mathbf{dX_t = b(X_t)dt} + \sigma(\mathbf{X_t})\mathbf{dW_t} \text{ with } X_0 \sim \mu$$

and W_t a standard Brownian motion. **Observations** $(X_{i\Delta})_{1\leqslant i\leqslant n}$.
Δ small and $n\Delta$ large (high frequency data)
We define

$$Y_{i,\Delta} = \frac{X_{(i+1)\Delta} - X_{i\Delta}}{\Delta}.$$

Approximation 1

$$Y_{i,\Delta} = b(X_{i\Delta}) + \underbrace{\frac{1}{\Delta}\int_{i\Delta}^{(i+1)\Delta} \sigma(X_s)dW_s}_{:=Z_{i,\Delta}} + \underbrace{\frac{1}{\Delta}\int_{i\Delta}^{(i+1)\Delta} (b(X_s) - b(X_{i\Delta})ds}_{R^{(1)}_{i,\Delta}}.$$

The "noise" $Z_{i,\Delta}$ depends on X_s so coupling is uneasy and looses the independence structure $\Rightarrow$ Martingale method.

Martingale method for diffusions

Gaussian assumption on the noise free! (Brownian motion).

$$\nu_n(h) = \sum_{i=1}^{n} h(X_{i\Delta}) \int_{i\Delta}^{(i+1)\Delta} \sigma(X_s) dW_s = \int_0^{n\Delta} \widetilde{h}(X_s)\sigma(X_s) dW_s$$

with $\widetilde{h}(X_s) = h(X_{i\Delta})$ for $i\Delta \leqslant s < (i+1)\Delta$.

$$\mathbb{P}\left(\sum_{i=1}^{n} h(X_{i\Delta}) Z_{i\Delta} \geqslant n\mathsf{a}, \ \ \|h\|_n^2 \leqslant v^2\right) \leqslant \exp\left(-\frac{n\Delta \mathsf{a}^2}{2\|\sigma\|_\infty^2 v^2}\right).$$

Under σ bounded due to $\langle M\rangle_{(n+1)\Delta} = \sum_{i=1}^{n} t^2(X_{i\Delta}) \int_{i\Delta}^{(i+1)\Delta} \sigma^2(X_s) ds.$

Residual terms to bound.

Least squares penalized estimator

$$\widehat{m} = \arg \min_{m \in \widehat{\mathcal{M}}_n} \left[\gamma_n(\widehat{b}_m) + \text{pen}(m) \right]$$

under conditions and on compact support is such that

$$\mathbb{E}(\|\widehat{b}_{\widehat{m}} - b_A\|_n^2) \leqslant C \inf_{m \in \mathcal{M}_n} \left(\|b_m - b_A\|_f^2 + \frac{\sigma_1^2 D_m}{n\Delta} \right) + K'\Delta + \frac{K"}{n\Delta}$$

for

$$\text{pen}(m) \geqslant \kappa \sigma_1^2 \frac{D_m}{n\Delta},$$

where σ_1 is an upper bound on σ.

$\Rightarrow$ Asymptotic is with respect to $n\Delta$.

Coupling method for diffusions

Works with the approximation

$$Y_{i,\Delta} = b(X_{i\Delta}) + \underbrace{\sigma(X_{i\Delta})\frac{W_{(i+1)\Delta} - W_{i\Delta}}{\Delta}}_{X,W \text{ separated}} \quad + \quad \underbrace{\frac{1}{\Delta}\int_{i\Delta}^{(i+1)\Delta} [b(X_s) - b(X_{i\Delta})]ds}_{R^{(1)}_{i,\Delta}}$$

$$+ \quad \underbrace{\frac{1}{\Delta}\int_{i\Delta}^{(i+1)\Delta} [\sigma(X_s) - \sigma(X_{i\Delta})]dW_s}_{:=R^{(2)}_{i,\Delta}}\,.$$

Price: one more residual term to control.

Tropp-Chernov generalization to dependent variables is possible
Non compact support results.

Diffusions in FDA: back to independence

$$\gamma_n(h) = \frac{1}{nT}\sum_{i=1}^{n}\left(\int_0^T h^2(X_i(u))du - 2\int_0^T h(X_i(u))dX_i(u)\right)$$

$$\widehat{Z}_m = \left(\frac{1}{nT}\sum_{i=1}^{n}\int_0^T \varphi_j(X_i(u))dX_i(u)\right)_{j=0,\dots,m-1}$$

and the $m \times m$-matrix

$$\widehat{\Psi}_m = \left(\frac{1}{nT}\sum_{i=1}^{n}\int_0^T \varphi_j(X_i(u))\,\varphi_\ell(X_i(u))du\right)_{j,\ell=0,\dots,m-1}.$$

Then, provided that $\widehat{\Psi}_m$ is a.s. invertible,

$$\hat{\mathbf{a}}_m = \widehat{\Psi}_m^{-1}\widehat{Z}_m.$$

$\widehat{\Psi}_m$ is no longer of the form $(1/n)\widehat{\Phi}_m^\top\widehat{\Phi}_m$.

$$\|h\|_n^2 = \frac{1}{nT}\sum_{i=1}^n \int_0^T h^2(X_i(u))du, \quad \langle s,t\rangle_n = \frac{1}{nT}\sum_{i=1}^n \int_0^T t(X_i(u))s(X_i(u))du,$$

$$\boxed{\nu_n(h) = \frac{1}{nT}\sum_{i=1}^n \int_0^T h(X_i(u))\sigma(X_i(u))dW_i(u).}$$

Therefore,

$$\mathbb{E}\|h\|_n^2 = \|h\|_{f_T}^2, \quad \mathbb{E}\langle h, h^\star\rangle_n = \langle h, h^\star\rangle_{f_T}$$

and

$$\mathbb{E}\nu_n(h) = 0, \quad \mathbb{E}\nu_n^2(h) = \|h\sigma\|_{f_T}^2/nT$$

where f_T is an integral of the transition density.

Martingale + chaining strategy for deviation. T fixed.

Talagrand/Truncation strategy possible?
Discretization, Denis, Dion, Martinez (2021)

Concluding remarks.

- Improvement over time for generalizations and improvement of results and conditions!
- Question. Tropp-Chernov for unbounded variables.
- Question. Ready-to-use deviation of the empirical process under moment conditions in the non bounded case.
- Dimension $\geqslant 1$ for X: Dussap's paper in the multivariate case and the algebra of hypermatrices.
- Higher dimension: Additive model + Lasso + Model selection. In which order?

Thank you for your attention !

Deviation Inequalities.

Ledoux, M. (2001) *The concentration of measure phenomenon.* Math. Surveys Monogr., **89** American Mathematical Society, Providence, RI.

Tropp, J. A. (2012). User-friendly tail bounds for sums of random matrices. *Found. Comput. Math.*, 12(4):389–434.

Talagrand, M, 1996. New concentration inequalities in product spaces. *Invent. Math.*, **126**, 505-563.

Klein, T. and **Rio**, E. (2005). Concentration around the mean for maxima of empirical processes. *Ann. Probab.* **33**, 1060-1077.

Mixing

Berbee H.C.P., 1979. *Random walks with stationary increments and renewal theory.* Cent. Math. Tracts, Amsterdam.

Delyon, B. (1986). Mélange: exemples, applications Examples and applications of mixing. *Publ. Inst. Rech. Math. Rennes*, 42–60.

Doukhan, P., 1995 *Mixing properties and examples.* Springer-Verlag.

Rio E., 2000. *Théorie asymptotique des processus aléatoires faiblement dépendants.* Mathématiques et applications de la SMAI, 31, Springer.

Viennet G., 1997. Inequalities for absolutely regular sequences: application to density estimation. *Probab. Th. Rel. Fields*, **107**, 467-492.

Model selection

Barron, A.R., **Birgé**, L. and **Massart**, P. (1999). Risk bounds for model selection via penalization. *Probab. Theory Related Fields* **113**, 301–413.

Birgé, L. and **Massart**, P. (1997). From model selection to adaptive estimation. Festschrift for Lucien Le Cam, Springer, New York, 55-87.

Birgé, L. and **Massart**, P. (1998). Minimum contrast estimators on sieves: exponential bounds and rates of convergence. *Bernoulli* **4**, 329–375.

Massart, P. *Concentration inequalities and model selection*, volume 1896 of *Lecture Notes in Mathematics*. Springer, Berlin, 2007. Lectures from the 33rd Summer School on Probability Theory held in Saint-Flour, July 6–23, 2003, With a foreword by Jean Picard.

Nonparametric estimation for Regression and SDE

- 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 an autoregression and a geometrical β-mixing regression framework. *Ann. Statist.* **39**, 839-875.
- Baraud, Y., Comte, F. and Viennet, G. (2001b). Model selection for (auto)-regression with dependent data. *ESAIM Probab. Statist.* **5**, 33-49.
- Cohen, A., Davenport, M.A. and Leviatan, D. (2013). On the stability and accuracy of least squares approximations. *Foundations of Computational Mathematics* **13**, 819-834.
- Cohen, A., Davenport, M.A. and Leviatan, D. (2019). Correction to: On the stability and accuracy of least squares approximations. *Foundations of Computational Mathematics* **19**, 239.

References (4) *(to be continued)*

- Comte, F., Genon-Catalot, V. and Rozenholc, Y. (2007). Penalized nonparametric mean square estimation of the coefficients of diffusion processes. *Bernoulli* **13**, 514-543.
- Comte, F. and Genon-Catalot, V. (2020). Regression function estimation on non compact support as a partly inverse problem. *Ann Inst Stat Math*, **72**, 1023-1054.
- Comte, F. and Genon-Catalot, V. (2020). Nonparametric drift estimation for *i.i.d.* paths of stochastic differential equations. *The Annals of Statistics* **48**, 3336-3365.
- Denis, C., Dion-Blanc, C. and Martinez, M. (2021) A ridge estimator of the drift from discrete repeated observations of the solution of a stochastic differential equation. *Bernoulli* **27**, 2675-2713.
- Dussap, F. (2023). Nonparametric multiple regression by projection on non-compactly supported bases. *Ann. Inst. Statist. Math.* **75**, 731-771.
- Marie, N. (2025) *From nonparametric regression to statistical inference for non-ergodic diffusion processes*. Cham: Springer.