Nonparametric Regression by Projection on Non-compactly Supported Bases

Florian Dussap

9^{e} Rencontre des Jeunes Statisticien·ne·s
Avril 2022

Regression model with random design

Let $A \subset \mathbb{R}^p$, we observe $n \geqslant 1$ r.v. $(\boldsymbol{X}_i, Y_i) \in A \times \mathbb{R}$ given by:

$$Y_i = b(\boldsymbol{X}_i) + \varepsilon_i,$$

where:

- $(\boldsymbol{X}_i)$ are i.i.d. with unknown distribution μ.
- (ε_i) are i.i.d. with zero mean and known variance σ^2.
- $(\boldsymbol{X}_i)$ and (ε_i) are independent.

Our goal is to estimate the regression function $b \colon A \to \mathbb{R}$. To quantify the error of an estimator, we consider two norms:

$$\|t\|_n^2 := \frac{1}{n} \sum_{i=1}^{n} t(\boldsymbol{X}_i)^2, \quad \|t\|_\mu^2 := \int_A t(\boldsymbol{x})^2 \, \mathrm{d}\mu(\boldsymbol{x}).$$

The error relative to the norm $\|\cdot\|_\mu$ can be viewed as a prediction error:

$$\forall \hat{b} \text{ estimator}, \ \|b - \hat{b}\|_\mu^2 = \mathbb{E}_{\boldsymbol{X} \sim \mu}\left[\left(b(\boldsymbol{X}) - \hat{b}(\boldsymbol{X}) \right)^2 \,\middle|\, \boldsymbol{X}_1, \dots, \boldsymbol{X}_n \right].$$

Assumptions

1. We assume that $\mu \ll \nu$ for a fixed measure ν, and that $\frac{\mathrm{d}\mu}{\mathrm{d}\nu}$ is bounded on A. Hence, we have $\mathrm{L}^2(A, \mu) \subset \mathrm{L}^2(A, \nu)$.

2. If A is compact, we assume that:

$$\forall \boldsymbol{x} \in A, \quad \frac{\mathrm{d}\mu}{\mathrm{d}\nu}(\boldsymbol{x}) \geqslant f_0 > 0.$$

   Hence, the norms $\|\cdot\|_\mu$ and $\|\cdot\|_\nu$ are equivalent, and we have $\mathrm{L}^2(A, \mu) = \mathrm{L}^2(A, \nu)$.

3. We assume that $b \in \mathrm{L}^{2r}(A, \mu)$ for some $r \in (1, +\infty]$. We consider $r' \in [1, +\infty)$ such that $\frac{1}{r} + \frac{1}{r'} = 1$.

4. We assume that $A = A_1 \times \cdots \times A_p$ and that $\nu = \nu_1 \otimes \cdots \otimes \nu_p$.

Projection estimator I

Let $(\varphi^i_k)_{k\in\mathbb{N}}$ be an orthonormal basis of $\mathrm{L}^2(A_i,\nu_i)$. We construct a basis of $\mathrm{L}^2(A,\nu)$ by tensorization. For all $\boldsymbol{k} = (k_1,\dots,k_p) \in \mathbb{N}^p$ we define:

$$\varphi_{\boldsymbol{k}}(\boldsymbol{x}) \coloneqq (\varphi^1_{k_1} \otimes \cdots \otimes \varphi^p_{k_p})(\boldsymbol{x}) \coloneqq \varphi^1_{k_1}(x_1) \times \cdots \times \varphi^p_{k_p}(x_p).$$

For $\boldsymbol{m} \in \mathbb{N}^p_+$, we consider the model:

$$S_{\boldsymbol{m}} \coloneqq \mathrm{Span}\left(\varphi_{\boldsymbol{k}} : \forall i,\ 0 \leqslant k_i < m_i\right), \quad D_{\boldsymbol{m}} \coloneqq \dim(S_{\boldsymbol{m}}) = m_1 \cdots m_p,$$

and we estimate b by a least squares minimization on $S_{\boldsymbol{m}}$:

$$\hat{b}_{\boldsymbol{m}} \coloneqq \operatorname*{arg\,min}_{t\in S_{\boldsymbol{m}}} \frac{1}{n}\sum_{i=1}^{n}\left[Y_i - t(\boldsymbol{X}_i)\right]^2.$$

Example

1. For $A = [-\pi,\pi]$ and $\nu = \mathrm{Leb}$, we choose the trigonometric basis.
2. For $A = \mathbb{R}$ and $\nu = \mathrm{Leb}$, we choose $\varphi_k(x) = c_k H_k(x)\mathrm{e}^{-x^2/2}$ with H_k the k-th Hermite polynomial.

This estimator can be computed using hypermatrix calculus:

$$\hat{b}_{\boldsymbol{m}} = \sum_{\forall i,\, k_i < m_i} \hat{a}^{(\boldsymbol{m})}_{\boldsymbol{k}} \varphi_{\boldsymbol{k}}, \qquad \hat{\boldsymbol{a}}^{(\boldsymbol{m})} := \operatorname*{arg\,min}_{\boldsymbol{a} \in \mathbb{R}^{\boldsymbol{m}}} \left\| \boldsymbol{Y} - \hat{\boldsymbol{\Phi}}_{\boldsymbol{m}} \times_p \boldsymbol{a} \right\|^2_{\mathbb{R}^n}$$
$$= \hat{\boldsymbol{G}}_{\boldsymbol{m}}^{-1} \times_p \hat{\boldsymbol{\Phi}}^*_{\boldsymbol{m}} \times_1 \boldsymbol{Y},$$

where $\boldsymbol{Y} := (Y_1, \ldots, Y_n) \in \mathbb{R}^n$, where:

$$\hat{\boldsymbol{G}}_{\boldsymbol{m}} := \left[\langle \varphi_{\boldsymbol{j}}, \varphi_{\boldsymbol{k}} \rangle_n\right]_{\boldsymbol{j},\boldsymbol{k}} \in \mathbb{R}^{\boldsymbol{m} \times \boldsymbol{m}}, \quad \hat{\boldsymbol{\Phi}}_{\boldsymbol{m}} := \left[\varphi_{\boldsymbol{k}}(\boldsymbol{X}_i)\right]_{i,\boldsymbol{k}} \in \mathbb{R}^{n \times \boldsymbol{m}},$$

and where $\times_p$ stands for the p-contracted product:

$$\left[\boldsymbol{A} \times_p \boldsymbol{B}\right]_{\boldsymbol{j},\boldsymbol{\ell}} := \sum_{\boldsymbol{k}=(k_1,\ldots,k_p)} \boldsymbol{A}_{\boldsymbol{j},\boldsymbol{k}} \times \boldsymbol{B}_{\boldsymbol{k},\boldsymbol{\ell}}.$$

In the following, we will need to consider the expectation of $\hat{\boldsymbol{G}}_{\boldsymbol{m}}$:

$$\boldsymbol{G}_{\boldsymbol{m}} := \mathbb{E}[\hat{\boldsymbol{G}}_{\boldsymbol{m}}] = \left[\langle \varphi_{\boldsymbol{j}}, \varphi_{\boldsymbol{k}} \rangle_\mu\right]_{\boldsymbol{j},\boldsymbol{k}} \in \mathbb{R}^{\boldsymbol{m} \times \boldsymbol{m}}.$$

Basic bound on the empirical risk

We recall the classical bias-variance decomposition of the empirical risk.

Proposition

If $\hat{\boldsymbol{G}}_{\boldsymbol{m}}$ is invertible, then we have:

$$\begin{aligned}\mathbb{E}_{\boldsymbol{X}}\left[\|b - \hat{b}_{\boldsymbol{m}}\|_n^2\right] &:= \mathbb{E}\left[\|b - \hat{b}_{\boldsymbol{m}}\|_n^2 \,\middle|\, \boldsymbol{X}_1, \ldots, \boldsymbol{X}_n\right] \\ &= \inf_{t \in S_{\boldsymbol{m}}} \|b - t\|_n^2 + \sigma^2 \frac{D_{\boldsymbol{m}}}{n}.\end{aligned}$$

If $\hat{\boldsymbol{G}}_{\boldsymbol{m}}$ is invertible a.s., then we have:

$$\mathbb{E}\|b - \hat{b}_{\boldsymbol{m}}\|_n^2 \leqslant \inf_{t \in S_{\boldsymbol{m}}} \|b - t\|_\mu^2 + \sigma^2 \frac{D_{\boldsymbol{m}}}{n}.$$

From the empirical norm to the design norm

We introduce the event:

$$\Omega_{\boldsymbol{m}}(\delta) := \left\{ \sup_{t \in S_{\boldsymbol{m}} \setminus \{0\}} \frac{\|t\|_\mu^2}{\|t\|_n^2} \leqslant \frac{1}{1-\delta} \right\}, \quad \delta \in (0,1).$$

Lemma

For all $\delta \in (0,1)$ and all $\boldsymbol{m} \in \mathbb{N}_+^p$, we have:

$$\mathbb{P}\left[\Omega_{\boldsymbol{m}}(\delta)^{\mathsf{c}}\right] \leqslant D_{\boldsymbol{m}} \exp\left(-h(\delta) \frac{n}{L(\boldsymbol{m}) \|\boldsymbol{G}_{\boldsymbol{m}}^{-1}\|_{\mathrm{op}}} \right),$$

where $h(\delta) := (1-\delta)\log(1-\delta) + \delta$, and where:

$$L(\boldsymbol{m}) := \left\| \sum_{\forall i \; k_i < m_i} \varphi_{\boldsymbol{k}}^2 \right\|_\infty = \sup_{t \in S_{\boldsymbol{m}} \setminus \{0\}} \frac{\|t\|_\infty^2}{\|t\|_\nu^2}.$$

Remarks on the lemma

- For the trigonometric basis, we have $L(m) \leqslant m$.
- For the Hermite basis, we have $L(m) \leqslant C\sqrt{m}$.
- If A is compact, then we have $\|\boldsymbol{G}_{\boldsymbol{m}}^{-1}\|_{\mathrm{op}} \leqslant 1/f_0$.
- If $A = \mathbb{R}$ and $(\varphi_k)_{k\in\mathbb{N}}$ is the Hermite basis, then we have $\|\boldsymbol{G}_{\boldsymbol{m}}^{-1}\|_{\mathrm{op}} \geqslant C(\mu)\sqrt{m}$ [Comte and Genon-Catalot, 2020].

Sketch of the proof of the lemma

The proof is inspired by [Cohen et al., 2013]. Let $(\phi_1, \ldots, \phi_{D_{\boldsymbol{m}}})$ be an orthonormal of $S_{\boldsymbol{m}}$ for the inner product $\langle \cdot, \cdot \rangle_\mu$, and let $\boldsymbol{H}_{\boldsymbol{m}}$ be their Gram matrix relative to the empirical inner product, that is:

$$\boldsymbol{H}_{\boldsymbol{m}} \coloneqq \left[\langle \phi_j, \phi_k \rangle_n \right]_{j,k} \in \mathbb{R}^{D_{\boldsymbol{m}} \times D_{\boldsymbol{m}}}.$$

Then, we have:

$$\sup_{t \in S_{\boldsymbol{m}} \setminus \{0\}} \frac{\|t\|_\mu^2}{\|t\|_n^2} = \|\boldsymbol{H}_{\boldsymbol{m}}^{-1}\|_{\mathrm{op}} = \frac{1}{\lambda_{\min}(\boldsymbol{H}_{\boldsymbol{m}})}.$$

Hence, we can rewrite the event as:

$$\Omega_{\boldsymbol{m}}(\delta)^{\mathrm{c}} = \left\{ \lambda_{\min}(\boldsymbol{H}_{\boldsymbol{m}}) < 1 - \delta \right\} = \left\{ \lambda_{\min}(\boldsymbol{H}_{\boldsymbol{m}}) < (1 - \delta) \lambda_{\min}(\mathbb{E} \boldsymbol{H}_{\boldsymbol{m}}) \right\},$$

since $\mathbb{E} \boldsymbol{H}_{\boldsymbol{m}} = \boldsymbol{I}_{D_{\boldsymbol{m}}}$.

We conclude using the following concentration inequality.

Theorem ([Gittens and Tropp, 2011], [Tropp, 2012])

Let $\boldsymbol{Z}_1, \ldots, \boldsymbol{Z}_n$ be independent random self-adjoint positive semi-definite matrices with dimension d, such that $\sup_k \lambda_{\max}(\boldsymbol{Z}_k) \leqslant R$ a.s. If we define:

$$\mu_{\min} := \lambda_{\min}\left(\sum_{k=1}^{n} \mathbb{E}[\boldsymbol{Z}_k]\right),$$

then we have:

$$\mathbb{P}\left[\lambda_{\min}\left(\sum_{k=1}^{n} \boldsymbol{Z}_k\right) \leqslant (1-\delta)\mu_{\min}\right] \leqslant d \times \left(\frac{e^{-\delta}}{(1-\delta)^{(1-\delta)}}\right)^{\mu_{\min}/R},$$

$$\mathbb{P}\left[\lambda_{\min}\left(\sum_{k=1}^{n} \boldsymbol{Z}_k\right) \geqslant (1+\delta)\mu_{\min}\right] \leqslant \left(\frac{e^{\delta}}{(1+\delta)^{(1+\delta)}}\right)^{\mu_{\min}/R}.$$

Bound on the prediction risk

Let us consider the collection:

$$\mathcal{M}_{n,\alpha} := \left\{ \boldsymbol{m} \in \mathbb{N}_+^p \;\middle|\; L(\boldsymbol{m})\big(\|\boldsymbol{G}_{\boldsymbol{m}}^{-1}\|_{\mathrm{op}} \vee 1\big) \leqslant \alpha \frac{n}{\log n} \right\}.$$

If $\boldsymbol{m} \in \mathcal{M}_{n,\alpha}$, then we have $\mathbb{P}[\Omega_{\boldsymbol{m}}(\delta)^{\mathrm{c}}] \leqslant D_{\boldsymbol{m}}\, n^{-\alpha} \leqslant n^{-\alpha+1}$.

Theorem

For all $\alpha \in (0, \frac{1}{2r'+1})$ and for all $\boldsymbol{m} \in \mathcal{M}_{n,\alpha}$ we have:

$$\mathbb{E}\|b - \hat{b}_{\boldsymbol{m}}\|_\mu^2 \leqslant C_n(\alpha, r') \inf_{t \in S_{\boldsymbol{m}}} \|b - t\|_\mu^2 + C'(\alpha, r')\, \sigma^2 \frac{D_{\boldsymbol{m}}}{n} + R_n,$$

with:

$$R_n = \frac{C''\big(\|b\|_{\mathrm{L}^{2r}(\mu)}, \sigma^2, \alpha\big)}{n \log n}.$$

A model selection result in a fixed design setting

Let $\widehat{\mathcal{M}}_n$ a model collection that may depend on the $(\boldsymbol{X}_i)$, and let:

$$\hat{\boldsymbol{m}} := \underset{\boldsymbol{m} \in \widehat{\mathcal{M}}_n}{\arg\min} \left(-\|\hat{b}_{\boldsymbol{m}}\|_n^2 + \mathrm{pen}(\boldsymbol{m}) \right), \ \mathrm{pen}(\boldsymbol{m}) := (1+\theta)\sigma^2 \frac{D_{\boldsymbol{m}}}{n}.$$

Theorem ([Baraud, 2000])

If $\mathbb{E}|\varepsilon_1|^q$ is finite for some $q > 4$, then the following upper bound holds:

$$\mathbb{E}_{\boldsymbol{X}} \|b - \hat{b}_{\hat{\boldsymbol{m}}}\|_n^2 \leqslant C(\theta) \inf_{\boldsymbol{m} \in \widehat{\mathcal{M}}_n} \left(\inf_{t \in S_{\boldsymbol{m}}} \|b - t\|_n^2 + \sigma^2 \frac{D_{\boldsymbol{m}}}{n} \right) + \sigma^2 \frac{\Sigma_n(\theta, q)}{n},$$

with $\Sigma_n(\theta, q) := C'(\theta, q) \frac{\mathbb{E}|\varepsilon_1|^q}{\sigma^q} \sum_{\boldsymbol{m} \in \widehat{\mathcal{M}}_n} D_{\boldsymbol{m}}^{-(\frac{q}{2}-2)}$.

We choose the model collection:

$$\widehat{\mathcal{M}}_{n,\beta} := \left\{ \boldsymbol{m} \in \mathbb{N}_+^p \,\middle|\, L(\boldsymbol{m}) \left(\|\hat{\boldsymbol{G}}_{\boldsymbol{m}}^{-1}\|_{\mathrm{op}} \vee 1 \right) \leqslant \beta \frac{n}{\log n} \right\}.$$

Oracle bound for the empirical risk

Theorem

If $\mathbb{E}|\varepsilon_1|^q$ *is finite for some* $q > 6$*, then there exists a constant* $\alpha_{\beta,r'} > 0$ *such that for all* $\alpha \in (0, \alpha_{\beta,r'})$*, we have:*

$$\mathbb{E}\|b - \hat{b}_{\hat{\boldsymbol{m}}}\|_n^2 \leqslant C(\theta) \inf_{\boldsymbol{m} \in \mathcal{M}_{n,\alpha}} \left(\inf_{t \in S_{\boldsymbol{m}}} \|b - t\|_\mu^2 + \sigma^2 \frac{D_{\boldsymbol{m}}}{n} \right) + \sigma^2 \frac{\Sigma(\theta, q)}{n} + R_n,$$

where:

$$R_n := C'(\|b\|_{\mathrm{L}^{2r}(\mu)}, \sigma^2) \frac{(\log n)^{(p-1)/r'}}{n^{\kappa(\alpha,\beta)/r'}},$$

$$\Sigma(\theta, q) := C''(\theta, q) \frac{\mathbb{E}|\varepsilon_1|^q}{\sigma^q} \sum_{\boldsymbol{m} \in \mathbb{N}_+^p} D_{\boldsymbol{m}}^{-(\frac{q}{2}-2)},$$

with $\kappa(\alpha, \beta)$ *a positive constant satisfying* $\frac{\kappa(\alpha,\beta)}{r'} > 1$ *and* $\frac{\kappa(\alpha,\beta)}{r'} \to 1$ *as* $\alpha \to \alpha_{\beta,r'}$.

Oracle bound for a compact domain

Theorem

We assume that A is compact. If $\mathbb{E}|\varepsilon_1|^q$ is finite for some $q > 6$, then there exists $\beta_{f_0,r'} > 0$ such that for all $\beta \in (0, \beta_{f_0,r'})$, there exists $\alpha_{\beta,r'} > 0$ such that for all $\alpha \in (0, \alpha_{\beta,r'})$, we have:

$$\mathbb{E}\|b - \hat{b}_{\hat{\boldsymbol{m}}}\|_\mu^2 \leqslant C(\theta, \beta, r) \inf_{\boldsymbol{m} \in \mathcal{M}_{n,\alpha}} \left(\inf_{t \in S_{\boldsymbol{m}}} \|b - t\|_\mu^2 + \sigma^2 \frac{D_{\boldsymbol{m}}}{n} \right) + C'(\beta, r)\sigma^2 \frac{\Sigma(\theta, q)}{n} + R_n,$$

where the remainder term is given by:

$$R_n = C''\left(\|b\|_{\mathrm{L}^{2r}(\mu)}, \sigma^2, \beta, r\right) \left(n^{-\frac{\kappa(\alpha,\beta)}{r'}} (\log n)^{\frac{p-1}{r'}} + n^{-\lambda(\beta,r,f_0)} (\log n)^{\frac{p-1}{r'} - 1} \right)$$

with $\lambda(\beta, r, f_0) > 1$ and $\frac{\kappa(\alpha,\beta)}{r'} > 1$.

Oracle bound in the general case I

The compact case result is proven using the concentration inequalities of [Gittens and Tropp, 2011]. But the proof relies critically on the lower bound of $\frac{\mathrm{d}\mu}{\mathrm{d}\nu}$. In the general case, we use the matrix Bernstein bound instead.

Lemma

For all $x > 0$ and all $\boldsymbol{m} \in \mathbb{N}_+^p$ we have:

$$\mathbb{P}\Big[\|\hat{\boldsymbol{G}}_{\boldsymbol{m}} - \boldsymbol{G}_{\boldsymbol{m}}\|_{\mathrm{op}} \geqslant x\Big] \leqslant D_{\boldsymbol{m}} \exp\left(-n \times \frac{x^2/2}{L(\boldsymbol{m})\big(\|\frac{\mathrm{d}\mu}{\mathrm{d}\nu}\|_\infty + \frac{2}{3}x\big)}\right).$$

To obtain an oracle bound, we need to restrict the model collections:

$$\mathcal{M}'_{n,\alpha} := \left\{\boldsymbol{m} \in \mathbb{N}_+^p \,\middle|\, L(\boldsymbol{m})\left(\|\boldsymbol{G}_{\boldsymbol{m}}^{-1}\|_{\mathrm{op}}^{2} \vee 1\right) \leqslant \alpha \frac{n}{\log n}\right\},$$

$$\widehat{\mathcal{M}}'_{n,\beta} := \left\{\boldsymbol{m} \in \mathbb{N}_+^p \,\middle|\, L(\boldsymbol{m})\left(\|\hat{\boldsymbol{G}}_{\boldsymbol{m}}^{-1}\|_{\mathrm{op}}^{2} \vee 1\right) \leqslant \beta \frac{n}{\log n}\right\}.$$

Oracle bound in the general case II

In the following, let $B \coloneqq (\|\frac{\mathrm{d}\mu}{\mathrm{d}\nu}\|_\infty + \frac{2}{3})^{-1}$.

Theorem

If $\mathbb{E}|\varepsilon_1|^q$ is finite for some $q > 6$, then there exists $\beta_{B,r'} > 0$ such that for all $\beta \in (0, \beta_{B,r'})$, there exists $\alpha_{\beta,r'} > 0$ such that for all $\alpha \in (0, \alpha_{\beta,r'})$, we have:

$$\mathbb{E}\|b - \hat{b}_{\hat{\boldsymbol{m}}}\|_\mu^2 \leqslant C(\theta, \beta, r) \inf_{\boldsymbol{m} \in \mathcal{M}'_{n,\alpha}} \left(\inf_{t \in S_{\boldsymbol{m}}} \|b - t\|_\mu^2 + \sigma^2 \frac{D_{\boldsymbol{m}}}{n} \right) + C'(\beta, r)\sigma^2 \frac{\Sigma(\theta, q)}{n} + R_n,$$

where the remainder term is given by:

$$R_n = C''\left(\|b\|_{\mathrm{L}^{2r}(\mu)}, \sigma^2, \beta, r\right) \left(n^{-\frac{\kappa(\alpha,\beta)}{r'}} (\log n)^{\frac{p-1}{r'}} + n^{-\lambda(\beta,r,B)} (\log n)^{\frac{p-1}{r'}-1} \right)$$

with $\lambda(\beta, r, B) > 1$ and $\frac{\kappa(\alpha,\beta)}{r'} > 1$.

Conclusion

- We obtain bounds for the empirical risk from the results for fixed design regression.
- To obtain a bound on the prediction risk, we need to study the minimum eigenvalue of a random matrix. We do so by using concentration inequalities of [Gittens and Tropp, 2011] and [Tropp, 2012].
- From these inequalities, we obtain a condition on the size of the models that entails that the prediction risk satisfies the same bound than the empirical risk.
- Even in the noiseless case ($\varepsilon_i = 0$), regularization is required [Cohen et al., 2013].

References

Baraud, Y. (2000).
Model selection for regression on a fixed design.
Probability Theory and Related Fields, 117(4):467–493.

Cohen, A., Davenport, M. A., and Leviatan, D. (2013).
On the Stability and Accuracy of Least Squares Approximations.
Foundations of Computational Mathematics, 13(5):819–834.

Comte, F. and Genon-Catalot, V. (2020).
Regression function estimation as a partly inverse problem.
Annals of the Institute of Statistical Mathematics, 72(4):1023–1054.

Gittens, A. and Tropp, J. A. (2011).
Tail bounds for all eigenvalues of a sum of random matrices.
arXiv:1104.4513 [math].

Tropp, J. A. (2012).
User-Friendly Tail Bounds for Sums of Random Matrices.
Foundations of Computational Mathematics, 12(4):389–434.