Bayesian statistics: posterior contraction

Young researchers' days - June 2025

Maël Duverger

Statiscal problem

- A model is a family of probability distribution :

$$\{\mathbb{P}_f^n, \quad f \in \mathcal{F}\}$$

and $\mathcal{F}$ is a subset of a vector space that can be **infinite-dimensional**.

- We assume that there is measure μ^n that dominates $\mathbb{P}_f^n$ for all $f \in \mathcal{F}$, $d\mathbb{P}_f^n / d\mu^n = p_f^n$.

- There is a "true" parameter $f_0 \in \mathcal{F}$ and we observe a realization Y^n of the law $\mathbb{P}_{f_0}^n$.
$\hookrightarrow$ n is the amount of information (*examples*)

- The **likelihood** of a function f given the observation Y^n is $\boldsymbol{p_f^n(Y^n)}$. Roughly, it's the probability, given the observation Y^n, that the data are generated by f.

- **Goal :** estimate f_0 from Y^n and obtain some guarantees when $n \to +\infty$.

Example

- **Regression model** :

$$Y_i = f_0(i/n) + \epsilon_i \ , \qquad i = 1, ..., n$$

with $f_0 \in \mathcal{F} \subset \mathbb{L}_2[0;1]$ and ϵ is the **noise**. Assume that $(\epsilon_i)_i \overset{iid}{\sim} \mathcal{N}(0,1)$, so the $(Y_i)_i$ are independent and $Y_i \sim \mathcal{N}(f_0(i/n), 1)$. The model is :

$$\left\{ \bigotimes_{i=1}^{n} \mathcal{N}\big(f(i/n), 1\big) \ , \quad f \in \mathcal{F} \right\}$$

- Given $Y^n = (Y_1, ..., Y_n)$, the likelihood of a function $f \in \mathcal{F}$ is :

$$p_f^n(Y^n) = \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi}} \exp\left(\frac{-\big(Y_i - f(i/n)\big)^2}{2} \right)$$

Prior distribution

- We put a probability distribution on $\mathcal{F}$ (or on a set that approximates $\mathcal{F}$) denoted **Π and called the prior distribution**.

- **Examples :**
  - ◇ Take a family of linearly independent functions $(b_i)_i$ of $\mathbb{L}_2$ and then :

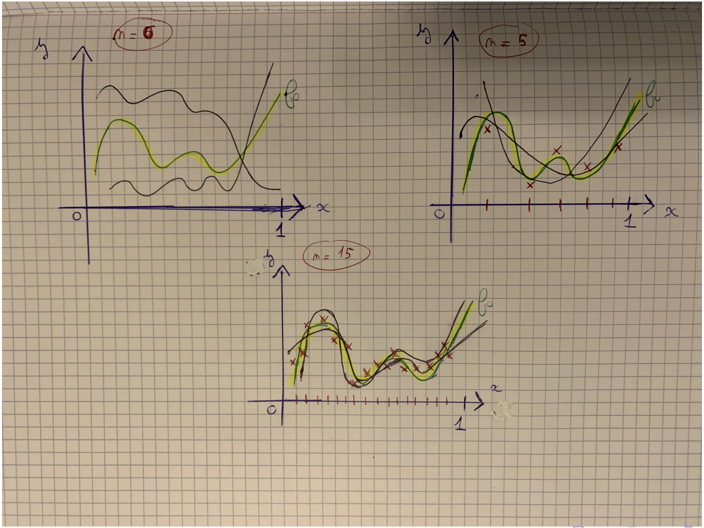
$$h = \sum_i \zeta_i b_i \ , \quad (\zeta_i)_i \overset{iid}{\sim} \mathcal{N}(0,1)$$

  - ◇ Consider a Brownian motion, one **trajectory is a continuous function** (even γ-Hölder for $\gamma < 1/2$). For more regularity, one can consider smoother Gaussian process (or other continuous-time stochastic process).

- Given the observation of Y^n, the Bayes formula gives **the posterior distribution** :

$$\Pi(B|Y^n) = \frac{\int_B p_f^n(Y^n)\ d\Pi(f)}{\int_{\mathcal{F}} p_f^n(Y^n)\ d\Pi(f)} = \frac{\int_B p_f^n(Y^n)/p_{f_0}^n(Y^n)\ d\Pi(f)}{\int_{\mathcal{F}} p_f^n(Y^n)/p_{f_0}^n(Y^n)\ d\Pi(f)}$$

Desired behavior of the posterior



Posterior contraction

• We say that **we have a posterior contraction at rate $\varepsilon_n \xrightarrow[n\to+\infty]{} 0$** for a metric d when

$$\Pi\left(B_d(f_0,\varepsilon_n)\,\middle|\,Y^n\right) \xrightarrow[n\to+\infty]{\mathbb{P}_{f_0}} 1$$

• What are right conditions on the prior distribution Π and on the parameter space $\mathcal{F}$ to obtain posterior contraction ? **Two remarks :**

- ◇ let B be a measurable set, if $\Pi(B) = 0$, then $\Pi(B|Y^n) = 0$ for all n
  $\hookrightarrow$ the prior have to put some mass around the true parameter f_0
- ◇ The space of parameters has to do not be too large
  $\hookrightarrow$ we have to consider "small" spaces : Hölder or Sobolev spaces for examples.

Kullback-Leibler divergence

Recall that :

$$\Pi(B|Y^n) = \frac{\int_B p_f^n(Y^n)/p_{f_0}^n(Y^n)\ d\Pi(f)}{\int_{\mathcal{F}} p_f^n(Y^n)/p_{f_0}^n(Y^n)\ d\Pi(f)}$$

The likelihood ratio $p_f^n(Y^n)/p_{f_0}^n(Y^n)$ (under the law P_{f_0}) is closely related to the **Kullback-Leibler divergence $KL(\mathbb{P}_{f_0}, \mathbb{P}_f)$** :

$$KL(\mathbb{P}_{f_0}, \mathbb{P}_f) = \int \log\left(\frac{p_{f_0}}{p_f}\right) p_{f_0}\, d\mu$$

We define "KL-neighborhood" :

$$BK(f_0, \varepsilon) = \left\{ f \in \mathcal{F},\ KL(\mathbb{P}_{f_0}, \mathbb{P}_f) \leq n\varepsilon^2,\ \int \left(\log\left(\frac{p_{f_0}}{p_f}\right) - KL(\mathbb{P}_{f_0}; \mathbb{P}_f) \right)^2 d\mathbb{P}_{f_0} \leq n\varepsilon^2 \right\}$$

Prior mass

Lemma 1 (not proved)

For any probability distribution Π on $\mathcal{F}$, for any $C, \varepsilon > 0$, with $P_{f_0}-$probability at least $1 - 1/C^2 n\varepsilon^2$,

$$\int \frac{p_f^n(Y^n)}{p_{f_0}^n(Y^n)} d\Pi(f) \geq \Pi\Big(BK(f_0, \varepsilon)\Big) \times \exp\big(-(1+C)n\varepsilon^2\big)$$

Lemma 2

Let $(\varepsilon_n)_n$ a sequence cuh that $n\varepsilon_n^2 \to +\infty$. Let $(A_n)_n$ be a sequence of measurable sets such that :

$$\frac{\Pi(A_n)}{\exp(-2n\varepsilon_n^2)\Pi\big(BK(f_0, \varepsilon_n)\big)} \xrightarrow[n\to+\infty]{} 0$$

Then,

$$\Pi(A_n | Y^n) \xrightarrow[n\to+\infty]{\mathbb{P}_{f_0}} 0$$

$\hookrightarrow$ proof on board

Test

(Very) Roughly speaking :

• Let's say we want to determine whether the data have been generated by f_0 or by $f \in B_d(f_1, \alpha\varepsilon)$ with $\alpha < 1$ and $d(f_0, f_1) > \varepsilon$.

• Roughly, a **test** $\boldsymbol{\phi_n(Y^n)} \in \{\mathbf{0}, \mathbf{1}\}$ says 0 if given the data he thinks f_0 is more likely, otherwise he says 1.

• From a **statistical point of view**, the space $\mathcal{F}$ and the model are **not too complicated** if we can construct a test such that

- ◇ $\mathbb{E}_{\mathbb{P}_{f_0}}[\phi_n(Y^n)] \approx 0$
- ◇ For $f \in B_d(f_1, \alpha\varepsilon)$, $\mathbb{E}_{\mathbb{P}_f}[\phi_n(Y^n)] \approx 1 \iff \mathbb{E}_{\mathbb{P}_f}[1 - \phi_n(Y^n)] \approx 0$

Covering number

- So we have a test of "f_0 against the ball $B_d(f_1, \alpha\varepsilon)$" with $d(f_1, f_0) > \varepsilon$.

- If the **covering number** of the space is not too large, we can construct a global test $\bar{\phi}_n$ of f_0 against $B_d(f_0, \varepsilon)^c$ that is :
  - ◇ $\mathbb{E}_{\mathbb{P}_{f_0}}[\bar{\phi}_n(Y^n)] \approx 0$
  - ◇ For $f \in B_d(f_0, \varepsilon)^c$, $\mathbb{E}_{\mathbb{P}_f}[1 - \bar{\phi}_n(Y^n)] \approx 0$

- Covering number : $\mathcal{N}(r, \mathcal{E}, d) = \min \left\{k,\ \exists(f_1, ..., f_k) \text{ such that } \mathcal{E} \subset \bigcup_{i=1}^k B_d(f_i, r)\right\}$.

Theorem

Theorem (Ghosal, Ghosh and van der Vaart - 2000)

Let ε_n such that $n\varepsilon_n^2 \to +\infty$. Suppose that :

(i) $\Pi\Big(BK(f_0, \varepsilon_n)\Big) \geq \exp(-Cn\varepsilon_n^2)$

(ii) There exists a measurable set $\mathcal{F}_n$ such that $\Pi\Big(\mathcal{F}_n^c\Big) \leq \exp(-(C+4)n\varepsilon_n^2)$

(iii) There exists $\alpha > 0$ such that for any $\varepsilon > 0$ and for any $f \in \mathcal{F}_n$ with $d(f_0, f) > \varepsilon$, we can construct a test ϕ_n of "f_0 against $B(f_1, a\varepsilon)$" that verifies :

$$\mathbb{E}_{\mathbb{P}_{f_0}}\left[\phi_n(Y^n)\right] \leq \exp(-Kn\varepsilon_n^2) \quad \text{and} \sup_{f \in_d (f, \alpha\varepsilon)} \mathbb{E}_{\mathbb{P}_f}\left[1 - \phi_n(Y^n)\right] \leq \exp(-Kn\varepsilon_n^2)$$

(iv) $\mathcal{N}\Big(\varepsilon_n, \mathcal{F}_n, d\Big) \leq \exp\left(Dn\varepsilon_n^2\right)$

Then we have posterior concentration around f_0 at rate ε_n in terms of metric d

Illustration with the example

- Recall : $Y_i = f_0(i/n) + \epsilon_i$ and the model is $\left\{ \bigotimes_{i=1}^{n} \mathcal{N}(f(i/n), 1) \;, \quad f \in \mathcal{F} \right\}$.
- We take $\mathcal{F} = \mathcal{H}(\beta), \beta \in]0;1]$,the space of functions f such that there exist $L > 0$

$$\forall x, y \in [0;1], \; |f(x) - f(y)| \leq L|x - y|^{\beta}$$

- **Prior** :
  - ◇ let $K_n(\beta) = n^{\frac{1}{2\beta+1}}$
  - ◇ $f(x) = \sum_{k=1}^{K_n(\beta)} f_k \mathbb{1}_{[\frac{k-1}{K_n(\beta)}; \frac{k}{K_n(\beta)}]}(x)$ and $(f_k)_k \overset{iid}{\sim} Laplace(1)$
- We set $\mathcal{F}_n = \{ f = \sum_{k=1}^{K_n(\beta)} f_k \mathbb{1}_{[\frac{k-1}{K_n(\beta)}; \frac{k}{K_n(\beta)}]}, \; \max_{1 \leq k \leq K_n(\beta)} |f_k| \leq n \}$.

$\hookrightarrow \lim_{n \to +\infty} \mathcal{F}_n = \bigcup_{n \geq 1} \mathcal{F}_n$ is dense in $\mathcal{F}$ (and even $dist_\infty(\mathcal{F}_n, \mathcal{F}^L) \approx Ln^{\frac{-\beta}{2\beta+1}}$)

- The prior puts most of its mass on $\mathcal{F}_n$ and $\mathcal{F}_n$ is simple and small enough (we can construct test and control its covering number).

Proof, first step : ”global” test

On board if time

Proof, second step : isolating the main term

- By (i), (ii) and lemma 2, $\Pi(\mathcal{F}_n^c|Y^n) \xrightarrow[n\to+\infty]{} 0$.

- Let $C_n = \{f \in \mathcal{F}_n,\ d(f,f_0) \geq M\varepsilon_n\}$, we have to show that $\Pi(C_n|Y^n) \xrightarrow[\mathbb{P}_{f_0}]{n\to+\infty} 0$.

- Let also $B_n = \left\{ \int \frac{p_f^n(Y^n)}{p_{f_0}^n(Y^n)} d\Pi(f) \geq \Pi\Big(BK(f_0,\varepsilon\Big) \times \exp\left(-(1+C)n\varepsilon^2\right)\right\}$, by the first lemma, $\mathbb{P}_{f_0}(B_n^c) = o(1)$. In addition, $\mathbb{E}_{\mathbb{P}_{f_0}}[\bar{\phi}_n(Y^n)] = o(1)$.

- So, as $\Pi(C_n|Y^n) \leq 1$, we have :

$$\Pi(C_n|Y^n) = \Pi(C_n|Y^n)\mathbb{1}_{B_n}(1-\bar{\phi}_n(Y^n)) + o_{\mathbb{P}_{f_0}}(1)$$

$$= \frac{\displaystyle\int_{C_n} \frac{p_f^n(Y^n)}{p_{f_0}^n(Y^n)}(1-\bar{\phi}_n(Y^n))\ d\Pi(f)}{\displaystyle\int_{\mathcal{F}} \frac{p_f^n(Y^n)}{p_{f_0}^n(Y^n)}\ d\Pi(f)}\mathbb{1}_{B_n} + o_{\mathbb{P}_{f_0}}(1)$$

Proof, second step

Taking expectation we have :

$$\mathbb{E}_{\mathbb{P}_{f_0}}\left[\Pi(C_n|Y^n)\right] = \mathbb{E}_{\mathbb{P}_{f_0}}\left[\frac{\displaystyle\int_{C_n}\frac{p_f^n(Y^n)}{p_{f_0}^n(Y^n)}(1-\bar{\phi}_n(Y^n))\ d\Pi(f)}{\displaystyle\int_{\mathcal{F}}\frac{p_f^n(Y^n)}{p_{f_0}^n(Y^n)}\ d\Pi(f)}\mathbb{1}_{B_n}\right] + o(1)$$

$$\leq \mathbb{E}_{\mathbb{P}_{f_0}}\left[\frac{\displaystyle\int_{C_n}\frac{p_f^n(Y^n)}{p_{f_0}^n(Y^n)}(1-\bar{\phi}_n(Y^n))\ d\Pi(f)}{\Pi\Big(BK(f_0,\varepsilon)\Big)\times\exp\big(-(1+C)n\varepsilon^2\big)}\mathbb{1}_{B_n}\right] + o(1) \text{ by def. of } B_n$$

$$\leq \frac{\displaystyle\int_{C_n}\mathbb{E}_{\mathbb{P}_{f_0}}\left[\frac{p_f^n(Y^n)}{p_{f_0}^n(Y^n)}(1-\bar{\phi}_n(Y^n))\right]d\Pi(f)}{\Pi\Big(BK(f_0,\varepsilon)\Big)\times\exp\big(-(1+C)n\varepsilon^2\big)} + o(1) \text{ by Fubini}$$

$$= \frac{\displaystyle\int_{C_n}\mathbb{E}_{\mathbb{P}_f}\Big[(1-\bar{\phi}_n(Y^n))\Big]d\Pi(f)}{\Pi\Big(BK(f_0,\varepsilon)\Big)\times\exp\big(-(1+C)n\varepsilon^2\big)} + o(1) \quad \textit{details}$$

Proof, conclusion

Using the property of the test $\bar{\phi}_n$, and the "prior mass condition" we finally have :

$$\mathbb{E}_{\mathbb{P}_{f_0}}\left[\Pi(C_n|Y^n)\right] \leq \exp\left((C+2)n\varepsilon_n^2\right) \times \exp\left(-(C+4)n\varepsilon_n^2\right) + o(1) = o(1) \quad (1)$$

Reference : Convergence rates of posterior distributions - Subhashis Ghosal, Jayanta K. Ghosh, Aad W. van der Vaart - Annals of Statistics - 2000