Bayesian nonparametrics

Posterior contraction and limiting shape

Ismaël Castillo

LPMA Université Paris VI

Berlin, September 2016

Part I

Introduction

Standard frequentist framework

Statistical experiment

- X random object = data
- $\mathcal{P}$ model

$$\mathcal{P} = \{P_\theta,\ \theta \in \Theta\}.$$

Frequentist assumption

$$\exists\ \theta_0 \in \Theta, \quad X \sim P_{\theta_0}$$

Standard frequentist framework

Statistical experiment

- X random object = data
- $\mathcal{P}$ model

$$\mathcal{P} = \{P_\theta,\ \theta \in \Theta\}.$$

Frequentist assumption

$$\exists\ \theta_0 \in \Theta, \quad X \sim P_{\theta_0}$$

Estimator a measurable function $\hat{\theta}(X) \in \Theta$

$\hat{\theta}(X)$ is a random point in Θ

one studies $\hat{\theta}(X)$ under $X \sim P_{\theta_0}$

example $\quad \hat{\theta}^{MLE}(X) = \underset{\theta\in\Theta}{\operatorname{argmax}}\ p_\theta(X)$

Bayesian framework

Statistical experiment

- X random object = data
- $\mathcal{P} = \{P_\theta,\ \theta \in \Theta\}$ model

Bayesian setting [Do not know θ? View it as random!]

a) $\theta \sim \Pi$ proba measure on Θ the prior distribution

b) View P_θ as law of $X|\,\theta$

$\Rightarrow$ joint distribution of (θ, X) is specified

c) law of $\theta|\,X$ is posterior distribution denoted $\Pi[\cdot|\,X]$

Bayesian estimator $\Pi(\cdot|\,X) \in \mathcal{M}_1(\Theta)$

$\Pi(\cdot|\,X)$ is a data–dependent measure on Θ

E0 – Example 0

$$X = (X_1, \ldots, X_n)$$

$$\mathcal{P} = \{\mathcal{N}(\theta, 1)^{\otimes n},\ \theta \in \mathbb{R}\}$$

Frequentist estimator

$$\hat{\theta}^{MLE}(X) = \overline{X}_n$$

Bayesian setting

a) $\theta \sim \mathcal{N}(0,1) = \Pi$ prior (say)

b) $X | \theta \sim \mathcal{N}(\theta, 1)^{\otimes n}$

c) $\theta | X \sim \mathcal{N}\left(\frac{n\overline{X}_n}{n+1}, \frac{1}{n+1}\right) = \Pi[\cdot | X]$ posterior

$$\Pi[\cdot | X] = \mathcal{N}\left(\frac{n\overline{X}_n}{n+1}, \frac{1}{n+1}\right)$$

Bayesian framework

Bayesian setting

a) $\theta \sim \Pi$ prior

b) $X | \theta \sim P_\theta$

c) $\theta | X \sim: \Pi[\cdot | X]$ posterior

All this produces a data–dependent measure $\Pi[\cdot | X]$

Bayesian framework

Bayesian setting

a) $\theta \sim \Pi$ prior

b) $X|\,\theta \sim P_\theta$

c) $\theta|\,X \sim: \Pi[\cdot|\,X]$ posterior

All this produces a data–dependent measure $\Pi[\cdot|\,X]$

And what if ...

... one would forget a)+b)+c) ...

... and study $\Pi[\cdot|\,X]$ *as a 'standard' estimator??*

Frequentist analysis of Bayesian procedures

Posterior distribution $\Pi[\cdot \mid X]$

Frequentist assumption

$$\exists\ \theta_0 \in \Theta, \quad X \sim P_{\theta_0}$$

Frequentist analysis of Bayesian procedures

Posterior distribution $\Pi[\cdot \mid X]$

Frequentist assumption

$$\exists\ \theta_0 \in \Theta, \quad X \sim P_{\theta_0}$$

E0 - Example 0

$$X \mid \theta \sim \mathcal{N}(\theta, 1)^{\otimes n}$$
$$\theta \sim \mathcal{N}(0, 1) = \Pi$$

$$\Pi[\cdot \mid X] = \mathcal{N}\left(\frac{n\overline{X}_n}{n+1}, \frac{1}{n+1}\right) \sim \overline{\theta}(X) + \frac{1}{\sqrt{n+1}}\mathcal{N}(0,1)$$

- centered close to $\hat{\theta}^{MLE}(X) = \overline{X}_n$
- *converges* at rate $1/\sqrt{n}$ towards θ_0 [see below]

Nonparametric models – regression

E1 – Gaussian white noise

$$dX^{(n)}(t) = f(t)dt + \frac{1}{\sqrt{n}}dB(t), \quad t \in [0,1]$$
$$f \in L^2[0,1], \quad B \text{ Brownian motion}$$

E1 – Fixed design regression

$$X_i = f(i/n) + \varepsilon_i$$
$$f \in \mathcal{C}^0[0,1], \quad \varepsilon_1, \ldots, \varepsilon_n \text{ iid } \mathcal{N}(0,1)$$

Unknow parameter f is a *function*, element of functional space

Sparsity

E2 – Sparse gaussian sequence model

$$X_i = \theta_i + \varepsilon_i, \quad 1 \leq i \leq n$$

$$\theta \in \ell_0[s] = \{\theta \in \mathbb{R}^n, \; |\{i : \; \theta_i \neq 0\}| \leq s\}$$

Unknown θ is a sparse vector

Nonparametric models – sampling models

E3 – Sampling from unknown distribution

$$X^{(n)} = (X_1, \ldots, X_n) \text{ i.i.d. } P \text{ on } [0,1]$$
$$P \text{ probability measure on } [0,1]$$

E3 – Density estimation

$$X^{(n)} = (X_1, \ldots, X_n) \text{ i.i.d. } f \text{ density on } [0,1]$$
$$f \text{ density on } [0,1]$$

Unknown P measure with total mass 1
Unknown f density: $f \geq 0$, $\int f = 1$

Nonparametric models – functionals

E4 – linear functional

$$\psi(f) = \int_0^1 a(u) f(u) du$$

a bounded on $[0,1]$

E5 – distribution function

$$F(\cdot) = \int_0^{\cdot} f(u) du$$

f density

Many other semiparametric functionals are also of interest ...

Prior on functions on $[0, 1]$

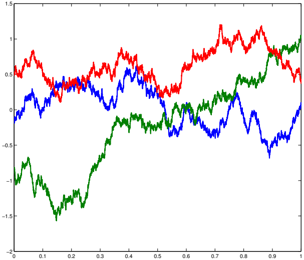
Consider the law of a process Z with a.s. continuous sample paths

Prior on functions on $[0,1]$

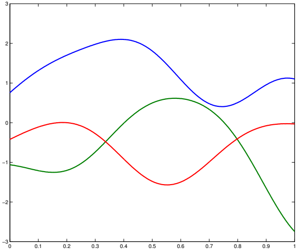
Consider the law of a process Z with a.s. continuous sample paths

Example (centered Gaussian process) $K(x,y) = E(Z_x Z_y)$ covariance of Z GP

$$K(x,y) = 1 + x \wedge y \qquad\qquad K(x,y) = e^{-(x-y)^2}$$



Brownian motion released at 0



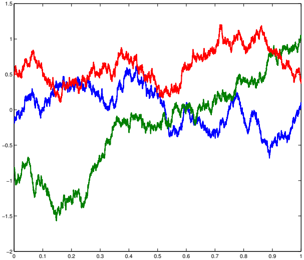
Squared-exponential GP

Prior on functions on [0, 1]

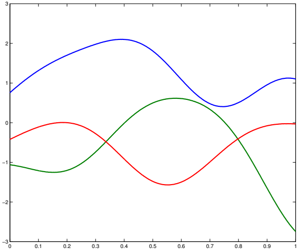
Consider the law of a process Z with a.s. continuous sample paths

Example (centered Gaussian process) $K(x,y) = E(Z_x Z_y)$ covariance of Z GP

$$K(x,y) = 1 + x \wedge y \qquad\qquad K(x,y) = e^{-(x-y)^2}$$



Brownian motion released at 0



Squared-exponential GP

Example (series expansions point of view)

$$Z(x) = \sum_{k \geq 1} \gamma_k \alpha_k \varepsilon_k(x)$$

Prior on sparse vectors

Define Π prior on $\ell_0[s] = \{\theta \in \mathbb{R}^n, \ |\{i: \ \theta_i \neq 0\}| \leq s\}$

- draw $k \sim \pi_n$ distribution on $\{0, \ldots, n\}$
- given k, pick $S \subset \{1, \ldots, n\}$ of size $|S| = k$ uniformly at random

$$\Pi_n(S|\, k) = 1/\binom{n}{k}$$

- given S, set

$$\begin{aligned} \theta_{S^c} &= 0 \\ \theta_S &\sim \bigotimes_{i \in S} g, \quad g \text{ density on } \mathbb{R} \end{aligned}$$

Sparsity

Example [special case of π_n = binomial] ['coin–flipping' prior and e.g. $\alpha_n = 1/n$]

$$\begin{array}{lll} k & \sim & \mathrm{Bin}(n, \alpha_n) = \pi_n \\ g & & \text{density on } \mathbb{R} \end{array}$$

$$\updownarrow$$

$$\Pi \sim \bigotimes_{i=1}^{n} (1 - \alpha_n)\delta_0 + \alpha_n g$$

[Remark. The posterior median can be shown to be a strict thresholding rule

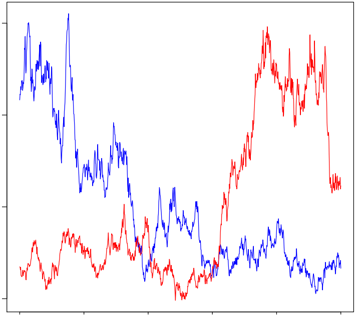
$$\hat{\theta}_i^{med}(X_i) = 0 \Leftrightarrow |X_i| \leq t(\alpha_n)$$

for some threshold $t(\alpha_n)$. For $\alpha_n = 1/n$, have $t(\alpha_n) \sim \sqrt{2\log(n)}$.]

Prior on densities on $[0, 1]$

Example – renormalisation

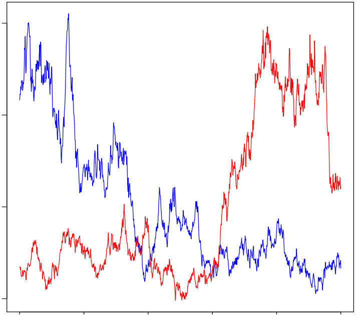
$$[0,1] \ni x \to \frac{e^{Z_x}}{\int_0^1 e^{Z_u} du}$$



Prior on densities on $[0, 1]$

Example – renormalisation

$$[0,1] \ni x \to \frac{e^{Z_x}}{\int_0^1 e^{Z_u} du}$$



Example – [more to follow!]

Priors on probability measures

- Prior on probability measures on $\mathbb{R}$

Dirichlet process $DP(M, G_0)$ on $\mathbb{R}$

G_0 probability measure on $\mathbb{R}$ mean and $M > 0$ concentration parameter

There exists a random probability measure on $\mathbb{R}$ such that

For any finite partition $(B_1, \ldots, B_r)$ of $\mathbb{R}$,

$$(G(B_1), \ldots, G(B_r)) \sim Dir(MG_0(B_1), \ldots, MG_0(B_r))$$

with $Dir(a_1, \ldots, a_r)$ the standard Dirichlet distribution on the r-simplex.

[Ferguson 1973]

Priors on probability measures, Dirichlet process

$G \sim DP(M, G_0)$ What does it look like ?

G is a discrete distribution almost surely

We have the representation [Sethuraman 94]

$$G \sim \sum_{k=1}^{\infty} \pi_k \delta_{\theta_k}$$

- $\theta_k \sim G_0(\cdot)$ i.i.d.
- π_k weights given by stickbreaking
  - $\pi_k = V_k \prod_{i=1}^{k-1}(1 - V_i)$
  - $V_1, V_2, \ldots \sim \text{Beta}(1, M)$ i.i.d.

Priors on probability measures, Dirichlet process

$G \sim DP(M, G_0)$ What does it look like ?

G is a discrete distribution almost surely

We have the representation [Sethuraman 94]

$$G \sim \sum_{k=1}^{\infty} \pi_k \delta_{\theta_k}$$

- $\theta_k \sim G_0(\cdot)$ i.i.d.
- π_k weights given by stickbreaking
  - $\pi_k = V_k \prod_{i=1}^{k-1}(1 - V_i)$
  - $V_1, V_2, \ldots \sim \text{Beta}(1, M)$ i.i.d.

Pitman-Yor process $\text{PY}(M, d, G_0)$

Same representation with $V_k \sim \text{Beta}(1 - d, M + kd)$

Priors on probability measures, Pólya trees on $[0, 1]$

Consider a sequence $\mathcal{I}$ of dyadic partitions of $[0, 1] = I_\varnothing$

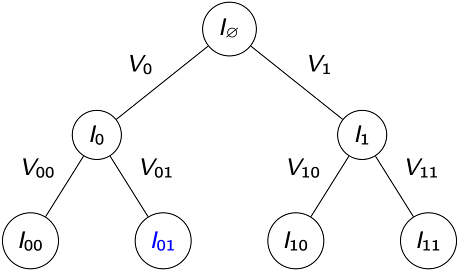
$\mathcal{I}_0 = \{[0, 1]\}$, $\mathcal{I}_1 = \{I_0, I_1\}$, $\mathcal{I}_2 = \{I_{00}, I_{01}, I_{10}, I_{11}\}$, ..., $\mathcal{I}_k = \{I_\varepsilon,\ \varepsilon \in \{0, 1\}^k\}$, ...

say for simplicity split in half $\rightarrow$ dyadic intervals $(k2^{-l}, (k+1)2^{-l}]$

Priors on probability measures, Pólya trees on $[0,1]$

Consider a sequence $\mathcal{I}$ of dyadic partitions of $[0,1] = I_\varnothing$
$\mathcal{I}_0 = \{[0,1]\}$, $\mathcal{I}_1 = \{I_0, I_1\}$, $\mathcal{I}_2 = \{I_{00}, I_{01}, I_{10}, I_{11}\}$, ..., $\mathcal{I}_k = \{I_\varepsilon,\ \varepsilon \in \{0,1\}^k\}$, ...

say for simplicity split in half $\rightarrow$ dyadic intervals $(k2^{-l}, (k+1)2^{-l}]$

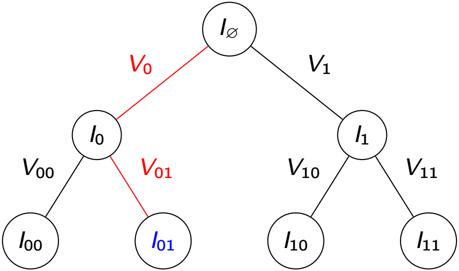


To get a probability measure, one sets $V_1 = 1 - V_0$ and more generally $V_{\varepsilon 1} = 1 - V_{\varepsilon 0}$

Priors on probability measures, Pólya trees on $[0,1]$

Consider a sequence $\mathcal{I}$ of dyadic partitions of $[0,1] = I_\varnothing$
$\mathcal{I}_0 = \{[0,1]\}$, $\mathcal{I}_1 = \{I_0, I_1\}$, $\mathcal{I}_2 = \{I_{00}, I_{01}, I_{10}, I_{11}\}$, ..., $\mathcal{I}_k = \{I_\varepsilon,\ \varepsilon \in \{0,1\}^k\}$, ...

say for simplicity split in half $\rightarrow$ dyadic intervals $(k2^{-l}, (k+1)2^{-l}]$



We set $P(I_{01}) = V_0 V_{01}$ and more generally,

$$P(I_{\varepsilon_1 \dots \varepsilon_k}) = V_{\varepsilon_1} V_{\varepsilon_1 \varepsilon_2} \cdots V_{\varepsilon_1 \varepsilon_2 \cdots \varepsilon_k} \text{ for } \varepsilon = \varepsilon_1 \varepsilon_2 \cdots \varepsilon_k$$

To get a probability measure, one sets $V_1 = 1 - V_0$ and more generally $V_{\varepsilon 1} = 1 - V_{\varepsilon 0}$

Pólya trees on $[0,1]$

A Pólya tree $PT(\mathcal{I},\alpha)$ on $[0,1]$ is the random probability measure P on $[0,1]$ defined as

- Given a collection of *independent* random variables,

$$V_{\varepsilon\mathbf{0}} \sim \mathrm{Beta}(\alpha_{\varepsilon\mathbf{0}},\alpha_{\varepsilon\mathbf{1}}), \qquad \text{any } \varepsilon\in\{0,1\}^k,\ k\geq 0$$

- Set, for any $\varepsilon=\varepsilon_\mathbf{1}\ldots\varepsilon_k\in\{0,1\}^k$, any $k\geq 0$,

$$P(I_\varepsilon)=\prod_{j=1;\ \varepsilon_j=\mathbf{0}}^{k} V_{\varepsilon_\mathbf{1}\ldots\varepsilon_{j-\mathbf{1}}\mathbf{0}} \times \prod_{j=1;\ \varepsilon_j=\mathbf{1}}^{k} (1-V_{\varepsilon_\mathbf{1}\ldots\varepsilon_{j-\mathbf{1}}\mathbf{0}})$$

Pólya trees on $[0, 1]$

A Pólya tree $PT(\mathcal{I}, \alpha)$ on $[0, 1]$ is the random probability measure P on $[0, 1]$ defined as

- Given a collection of *independent* random variables,

$$V_{\varepsilon 0} \sim \mathrm{Beta}(\alpha_{\varepsilon 0}, \alpha_{\varepsilon 1}), \qquad \text{any } \varepsilon \in \{0, 1\}^k, \ k \geq 0$$

- Set, for any $\varepsilon = \varepsilon_1 \ldots \varepsilon_k \in \{0, 1\}^k$, any $k \geq 0$,

$$P(I_\varepsilon) = \prod_{j=1;\ \varepsilon_j=0}^{k} V_{\varepsilon_1 \ldots \varepsilon_{j-1} 0} \times \prod_{j=1;\ \varepsilon_j=1}^{k} (1 - V_{\varepsilon_1 \ldots \varepsilon_{j-1} 0})$$

Special cases

- If $\alpha_\varepsilon = M\alpha(I_\varepsilon)$ for a probability measure $\alpha(\cdot)$ on $[0, 1]$, $\rightarrow$ $\mathrm{DP}(M, \alpha)$
- For $\alpha_{\varepsilon_1 \cdots \varepsilon_{k-1} 0} = \alpha_{\varepsilon_1 \cdots \varepsilon_{k-1} 1} = a_k$ and $\sum_{k=1}^{\infty} a_k^{-1} < \infty$ with $\mathcal{I}$ regular partition

$$P \ll \mathrm{Leb}[0, 1] \text{ a.s.}$$

In this case it is a prior on densities!

Part II

Rates

Bayesian dominated framework

Experiment. $X = X^{(n)}$, $\mathcal{P} = \{P_\theta^{(n)},\ \theta \in \mathcal{F}\}$, $(\mathcal{F}, \mathbb{F})$ measure space

Dominated framework. Suppose there exists a dominating measure $\mu^{(n)}$

$$dP_\theta^{(n)} = p_\theta^{(n)}(\cdot)d\mu^{(n)}$$

Bayesian setting.

a) $\theta \sim \Pi$ prior distribution

b) $X|\,\theta \sim P_\theta^{(n)}$

c) $\theta|\,X \sim: \Pi[\cdot|\,X]$ posterior

Bayes formula. For any measurable set $B \in \mathbb{F}$,

$$\Pi(B|\,X^{(n)}) = \frac{\int_B p_\theta^{(n)}(X^{(n)})d\Pi(\theta)}{\int p_\theta^{(n)}(X^{(n)})d\Pi(\theta)}.$$

Remark. $\Pi[B] = 0 \Rightarrow \Pi[B|\,X] = 0$

Consistency

Consistency. The posterior is consistent at θ_0 if, as $n \to \infty$,

$$\Pi(\cdot | X^{(n)}) \xrightarrow{w} \delta_{\theta_0}, \quad \text{in } P_{\theta_0}^{(n)}\text{-probability.}$$

In a separable metric space equipped with a distance d, for any $\varepsilon > 0$,

$$\Pi(\theta : \ d(\theta, \theta_0) < \varepsilon | \, X^{(n)}) \longrightarrow 1, \quad \text{in } P_{\theta_0}^{(n)}\text{-probab.}$$

[Notation: in the sequel, drop index 'n' and write P_{θ_0}, E_{θ_0}, X]

General consistency results

- [Doob (1949)]
- [Schwartz (1965)]
- [Barron, Schervish, Wasserman (1999)]

Convergence rate

Convergence rate. The posterior converges at rate ε_n for the distance d at θ_0 if, as $n \to \infty$,

$$E_{\theta_0} \Pi(\theta : \ d(\theta, \theta_0) \leq \varepsilon_n | X) \longrightarrow 1$$

It is an upper bound: we look for the smallest possible ε_n.

What happens in a nonparametric framework ?

- [Ghosal, Ghosh, van der Vaart 00], [Ghosal, van der Vaart 07]

Convergence rate – lower bound

Lower bound to rate. We say that ζ_n is a lower bound for the rate for d at θ_0 if

$$E_{\theta_0} \Pi(\theta : d(\theta, \theta_0) \leq \zeta_n | X) \longrightarrow 0$$

One looks for the largest possible ζ_n.

May look a little odd first that there is 'no mass' close to θ_0 ...

It just says that ζ_n is a too fast 'scaling' that captures little posterior mass

Rates, regular parametric models

E0

$$X \mid \theta \sim \mathcal{N}(\theta, 1)^{\otimes n}$$
$$\theta \sim \mathcal{N}(0, 1) = \Pi$$

$$\Pi[\cdot \mid X] = \mathcal{N}\left(\frac{n\overline{X}_n}{n+1}, \frac{1}{n+1}\right)$$

Exercise. Show that, for any $M_n \to \infty$,

$$E_{\theta_0}\Pi\left[\theta : \frac{1}{M_n\sqrt{n}} \leq |\theta - \theta_0| \leq \frac{M_n}{\sqrt{n}} \mid X\right] \to 1$$

This extends to regular parametric models and most 'reasonable' priors

Posterior and aspects of posterior

The posterior distribution $\Pi[\cdot \mid X]$ may have a well-defined

- posterior mean
$$\bar{\theta}(X) := \int \theta d\Pi(\theta \mid X)$$
- posterior mode
$$\theta^*(X) := \underset{\theta}{\operatorname{argmax}} \frac{d\Pi(\theta \mid X)}{d\mu}$$
- posterior median, posterior quantiles ...

These are estimators in the standard sense.

Often, but not always, a convergence rate for $\Pi[\cdot \mid X]$ implies the same rate for a given aspect of it

Posterior and aspects of posterior

Fact 1. Suppose $\Pi[d(\theta,\theta_0) > \varepsilon_n | X] = o_P(1)$. Let $\theta^B(X)$ be

the center of the smallest d–ball containing at least $1/2$ of the posterior mass

$$d(\theta^B(X),\theta_0) = O_P(\varepsilon_n)$$

Fact 2. Suppose

- $\Pi[d(\theta,\theta_0) > \varepsilon_n | X] = O_P(\varepsilon_n^2)$
- $\theta \to d^2(\theta,\theta_0)$ is convex and bounded [e.g. d=h]

Then

$$d(\bar{\theta}(X),\theta_0) = O_P(\varepsilon_n)$$

Posterior integrated loss

Fact 3. Suppose

$$E_{\theta_0} \int d(\theta, \theta_0)^2 d\Pi(\theta | X) \le \varepsilon_n^2$$

Then

- for any $M_n \to \infty$

$$E_{\theta_0} \Pi[\theta : \ d(\theta, \theta_0)^2 > M_n \varepsilon_n^2 | X] = o(1)$$

- If $\theta \to d^2(\theta, \theta_0)$ is convex,

$$E_{\theta_0} \left[d(\bar{\theta}(X), \theta_0)^2 \right] \le \varepsilon_n^2$$

First examples

E1 – Fixed design regression [van der Vaart, van Zanten 08, 09]

True function. Let $f_0 \in \mathcal{C}^\beta[0,1]$

Loss function. $\|g\|_n^2 = n^{-1}\sum_{i=1}^n g(t_i)^2$

Prior. Brownian motion + Gaussian

$W_t = B_t + Z_0$, with $Z_0 \sim \mathcal{N}(0,1)$

Then as $n \to \infty$,

$$E_{f_0}\Pi\left[f : \|f - f_0\|_n \leq \varepsilon_n \,|\, X\right] \to 1,$$

where

$$\varepsilon_n \sim n^{-\frac{1}{4}\wedge\frac{\beta}{2}} = \begin{cases} n^{-1/4} & \text{if } \beta \geq 1/2 \\ n^{-\beta/2} & \text{if } \beta \leq 1/2 \end{cases}$$

First examples

E1 – Fixed design regression (followed)

Prior. *Riemann-Liouville process with parameter* $\alpha > 0$

$W_t = \int_0^t (t-s)^{\alpha - 1/2} dB_s + \sum_{k=0}^{\lceil \alpha \rceil} Z_k t^k$, with $Z_k \sim \mathcal{N}(0,1)$ iid

Then

$$E_{f_0} \Pi\left[f : \|f - f_0\|_n \leq \varepsilon_n \mid X \right] \to 1,$$

where

$$\varepsilon_n \approx n^{-\frac{\alpha \wedge \beta}{2\alpha+1}} = \begin{cases} n^{-\frac{\alpha}{2\alpha+1}} & \text{if } \beta \geq \alpha \\ n^{-\frac{\beta}{2\alpha+1}} & \text{if } \beta \leq \alpha \end{cases}$$

First examples

E3 – Density estimation [van der Vaart, van Zanten 08, 09]

True density. Let $f_0 \in \mathcal{C}^\beta[0,1]$ with $f_0 > 0$.

Loss function. Hellinger distance $h(f,g)^2 = \int(\sqrt{f} - \sqrt{g})^2$

Prior. Consider the distribution on continuous functions induced by

$$t \to \frac{e^{W_t}}{\int_0^1 e^{W_u} du}$$

with W_t either *Brownian motion* or *Riemann-Liouville process with parameter* $\alpha > 0$

Then, for ε_n as before,

$$E_{f_0}\Pi\left[h(f,f_0) \le \varepsilon_n | X\right] \to 1.$$

Theory: Bayesian nonparametrics

Consistency

- [Doob (1949)] Consistency up to a Π-null set
- [Schwartz (1965)] Consistency under testing and enough prior mass in Kullback-Leibler-type neighborhoods
- [Diaconis, Freedman (1986)] Example of 'innocent' prior in a semiparametric framework, for which the posterior is not consistent

Rates of convergence ?

GGV theory: first lemma

To fix ideas consider first the density estimation model

$$X = (X_1, \ldots, X_n)$$
$$\mathcal{P} = \{P_f^{\otimes n},\ dP_f = f d\mu,\ f \in \mathcal{F}\}$$

$\mathcal{F}$ some set of densities

Π prior distribution on $\mathcal{F}$

Bayes' formula

$$\Pi[B|\, X] = \frac{\int_B \prod_{i=1}^n f(X_i) d\Pi(f)}{\int \prod_{i=1}^n f(X_i) d\Pi(f)}$$

Notation

$$K(f_0, f) = \int \log \frac{f_0}{f} f_0 d\mu$$
$$V(f_0, f) = \int \left(\log \frac{f_0}{f} - K(f_0, f) \right)^2 f_0 d\mu$$

GGV theory: first lemma

Define a Kullback–Leibler type neighborhood of f_0 by

$$B_{KL}(f_0, \varepsilon_n) := \{f : \; K(f_0, f) \le \varepsilon_n^2, \; V(f_0, f) \le \varepsilon_n^2\}$$

Lemma 1. Let A_n measurable and let $n\varepsilon_n^2 \to \infty$. Suppose

$$\frac{\Pi[A_n]}{e^{-2n\varepsilon_n^2}\Pi[B_{KL}(f_0, \varepsilon_n)]} = o(1).$$

Then $\Pi[A_n | \, X] = o_P(1)$.

[Idea: very small prior mass implies small posterior mass]

Lemma 2. For any Π probability measure on $\mathcal{F}$, any $C > 0$ and $\varepsilon > 0$,

$$\int \prod_{i=1}^{n} \frac{f}{f_0}(X_i) d\Pi(f) \geq \Pi(B_{KL}(f_0, \varepsilon)) e^{-(1+C)n\varepsilon^2}$$

with P_{f_0}–probability at least $1 - 1/(C^2 n\varepsilon^2)$.

$$\int \prod_{i=1}^{n} \frac{f}{f_0}(X_i) d\Pi(f) \geq \int_B \prod_{i=1}^{n} \frac{f}{f_0}(X_i) d\bar{\Pi}(f) \Pi(B), \qquad \bar{\Pi} = \frac{\Pi|_B}{\Pi(B)}$$

Lemma 2. For any Π probability measure on $\mathcal{F}$, any $C > 0$ and $\varepsilon > 0$,

$$\int \prod_{i=1}^{n} \frac{f}{f_0}(X_i)d\Pi(f) \geq \Pi(B_{KL}(f_0,\varepsilon))e^{-(1+C)n\varepsilon^2}$$

with P_{f_0}–probability at least $1 - 1/(C^2 n\varepsilon^2)$.

$$\int \prod_{i=1}^{n} \frac{f}{f_0}(X_i)d\Pi(f) \geq \int_B \prod_{i=1}^{n} \frac{f}{f_0}(X_i)d\bar{\Pi}(f)\Pi(B), \qquad \bar{\Pi} = \frac{\Pi|_B}{\Pi(B)}$$

$$\log \int_B \prod_{i=1}^{n} \frac{f}{f_0}(X_i)d\bar{\Pi}(f) \geq \sum_{i=1}^{n} \int_B \log \frac{f}{f_0}(X_i)d\bar{\Pi}(f) \qquad [Jensen]$$

$$\geq -\sum_{i=1}^{n} \int_B \left[\log \frac{f_0}{f}(X_i) - KL(f_0,f)\right] d\bar{\Pi}(f) - n\int_B KL(f_0,f)d\bar{\Pi}(f)$$

$$\geq -\sum_{i=1}^{n} Z_i - n\varepsilon^2 \qquad \text{if } B = B_{KL}(f_0,\varepsilon)$$

$$Z_i := \int_B \left[\log \frac{f_0}{f}(X_i) - KL(f_0, f)\right] d\bar{\Pi}(f)$$

We have Z_i iid and

$$E_{f_0} Z_1 = 0 \qquad [Fubini]$$

$$\mathrm{Var}_{f_0} Z_1 \leq E_{f_0} \int_B \left[\log \frac{f_0}{f}(X_i) - KL(f_0, f)\right]^2 d\bar{\Pi}(f) \qquad [CS]$$

$$= \int_B V(f_0, f) d\bar{\Pi}(f) \leq \varepsilon^2 \bar{\Pi}(B) = \varepsilon^2 \qquad [Fubini]$$

$$Z_i := \int_B \left[\log \frac{f_0}{f}(X_i) - KL(f_0, f)\right] d\bar{\Pi}(f)$$

We have Z_i iid and

$$E_{f_0} Z_1 = 0 \qquad [Fubini]$$

$$\mathrm{Var}_{f_0} Z_1 \leq E_{f_0} \int_B \left[\log \frac{f_0}{f}(X_i) - KL(f_0, f)\right]^2 d\bar{\Pi}(f) \qquad [CS]$$

$$= \int_B V(f_0, f) d\bar{\Pi}(f) \leq \varepsilon^2 \bar{\Pi}(B) = \varepsilon^2 \qquad [Fubini]$$

By Chebyshev's inequality,

$$P_{f_0}\left[\left|\sum_{i=1}^n Z_i\right| > cn\varepsilon^2\right] \leq \frac{1}{C^2 n\varepsilon^2}.$$

On the complementary event, one thus has

$$\log \int \prod_{i=1}^n \frac{f}{f_0}(X_i) d\bar{\Pi}(f) \geq -(C+1)n\varepsilon^2$$

that is

$$\int \prod_{i=1}^n \frac{f}{f_0}(X_i) d\bar{\Pi}(f) \geq \Pi(B) e^{(C+1)n\varepsilon^2}. \quad \square$$

GGV theory: first lemma

Let $B_{KL}(f_0, \varepsilon_n) := \{f : \ K(f_0, f) \leq \varepsilon_n^2, \ V(f_0, f) \leq \varepsilon_n^2\}$

Lemma 1. Let A_n measurable and let $n\varepsilon_n^2 \to \infty$. Suppose

$$\frac{\Pi[A_n]}{e^{-2n\varepsilon_n^2}\Pi[B_{KL}(f_0, \varepsilon_n)]} = o(1).$$

Then $\Pi[A_n | X] = o_P(1)$.

[Proof of lemma 1]

Theory: Bayesian nonparametrics

Experiment. $X = X^{(n)}$, $\mathcal{P} = \{P_\theta^{(n)},\ \theta \in \mathcal{F}\}$ as before [not necessarily iid sampling]

Prior. Π prior distribution on θ

Goal. For some distance d_n and rate ε_n,

$$E_{\theta_0}\Pi\left[\theta :\ d_n(\theta,\theta_0) > M\varepsilon_n \,|\, X\right] \to 0$$

Let us assume that, for $n\varepsilon_n^2 \to \infty$,

Theory: Bayesian nonparametrics

Experiment. $X = X^{(n)}$, $\mathcal{P} = \{P_\theta^{(n)},\ \theta \in \mathcal{F}\}$ as before [not necessarily iid sampling]

Prior. Π prior distribution on θ

Goal. For some distance d_n and rate ε_n,

$$E_{\theta_0}\Pi\left[\theta :\ d_n(\theta, \theta_0) > M\varepsilon_n \mid X\right] \to 0$$

Let us assume that, for $n\varepsilon_n^2 \to \infty$,

- The prior puts enough mass on neighborhoods of θ_0

$$\Pi(B_{KL}(\theta_0, \varepsilon_n)) \geq e^{-cn\varepsilon_n^2}$$

$$B_{KL}(\theta_0, \varepsilon_n) = \left\{ \int p_{\theta_0}^{(n)} \log \frac{p_{\theta_0}^{(n)}}{p_\theta^{(n)}} \leq n\varepsilon_n^2,\ \ \int p_{\theta_0}^{(n)} \log^2 \frac{p_{\theta_0}^{(n)}}{p_\theta^{(n)}} \leq n\varepsilon_n^2 \right\}$$

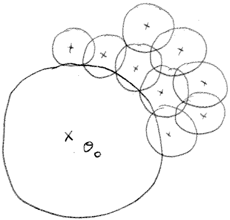
extends definition given above in iid sampling model

Theory: Bayesian nonparametrics

- Some sieve sets Θ_n capture most of the prior mass

$$\Pi(\Theta \backslash \Theta_n) \leq e^{-(c+4)n\varepsilon_n^2}$$

- Test true parameter vs. Complement of a ball **(T1)**



There exist tests ψ_n such that

$$E_{\theta_0}\psi_n \to 0$$

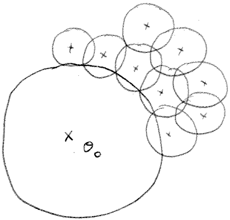
$$\sup_{\theta\in\Theta_n:\, d_n(\theta,\theta_0)>M\varepsilon_n} E_\theta(1-\psi_n) \lesssim e^{-(c+4)n\varepsilon_n^2}$$

Theory: Bayesian nonparametrics

- Some sieve sets Θ_n capture most of the prior mass

$$\Pi(\Theta \backslash \Theta_n) \leq e^{-(c+4)n\varepsilon_n^2}$$

- Test true parameter vs. Complement of a ball **(T1)**



There exist tests ψ_n such that

$$E_{\theta_0}\psi_n \to 0$$

$$\sup_{\theta \in \Theta_n:\, d_n(\theta,\theta_0) > M\varepsilon_n} E_\theta(1-\psi_n) \lesssim e^{-(c+4)n\varepsilon_n^2}$$

A sufficient condition is a control of an entropy [see below]

Theory: Bayesian nonparametrics

Theorem 1. [Ghosal, Ghosh, van der Vaart 00], [Ghosal, van der Vaart 07]

If there are sets $\Theta_n \subset \Theta$, $c, M > 0$ such that there exist tests ψ_n with

$$E_{\theta_0}\psi_n \to 0, \qquad \sup_{\theta\in\Theta_n:\, d_n(\theta,\theta_0)>M\varepsilon_n} E_\theta(1-\psi_n) \lesssim e^{-(c+4)n\varepsilon_n^2} \qquad \textit{tests}$$

$$\Pi(\Theta\backslash\Theta_n) \leq e^{-(c+4)n\varepsilon_n^2} \qquad \textit{remaining mass}$$

$$\Pi(B_{KL}(\theta_0,\varepsilon_n)) \geq e^{-cn\varepsilon_n^2} \qquad \textit{prior mass}$$

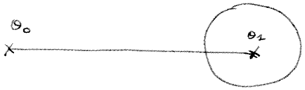
Then for $M > 0$ as above,

$$E_{\theta_0}\Pi[\theta : d_n(\theta,\theta_0) \leq M\varepsilon_n | X] \to 1$$

Testing via entropy

Tests point vs. ball for some distance d_n **(T0)**

$\exists K, \xi > 0, \ \forall \varepsilon > 0 \ \forall \theta_1 \in \Theta : d_n(\theta_1, \theta_0) > \varepsilon, \ \exists \phi_n$ test



$$\boxed{P_{\theta_0}^{(n)} \phi_n \le e^{-Kn\varepsilon^2}} \quad \text{et} \quad \boxed{\sup_{d_n(\theta, \theta_1) < \varepsilon \xi} P_{\theta}^{(n)}(1 - \phi_n) \le e^{-Kn\varepsilon^2}}$$

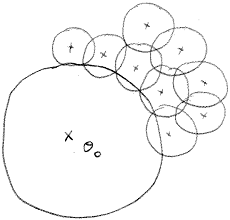
$N(\varepsilon, \Theta_n, d_n)$ = minimal number of balls of radius ε for d_n to cover Θ_n

Lemma 3. Suppose **(T0)** holds and

$$\log N(\varepsilon, \Theta_n, d_n) \le Cn\varepsilon_n^2$$

Then for any given $c > 0$ and M large enough, **(T1)** holds.

Proof of lemma 3 [sketch]



- Build shells $\mathcal{C}_j$ covering the complement of the ball
- Cover each shell by a minimal number of balls
- To each center of ball, associate $\varphi_{i,j}$ test via **(T0)**
- Set

$$\psi := \max_{i,j} \varphi_{i,j}$$

- testing power $e^{-Knj\varepsilon^2}$ over shells is summable in j

Testing distances: examples

About the testing distance **(T0)**

[Le Cam 70's, Birgé 80's] $\rightarrow$ general results for testing between convex sets

Imply that **(T0)** holds in iid and some simple dependency settings for

$$d = h \quad \text{or} \quad d = \|\cdot\|_1$$

Proposition. In density estimation $P_f^{(n)} = P_f^{\otimes n}$, then **(T0)** holds for $d = \|\cdot\|_1$

Testing distances: examples

About the testing distance **(T0)**

[Le Cam 70's, Birgé 80's] $\rightarrow$ general results for testing between convex sets

Imply that **(T0)** holds in iid and some simple dependency settings for

$$d = h \quad \text{or} \quad d = \|\cdot\|_1$$

Proposition. In density estimation $P_f^{(n)} = P_f^{\otimes n}$, then **(T0)** holds for $d = \|\cdot\|_1$

[Idea of proof.] Let f_0, f_1 with $\|f_0 - f_1\|_1 > \varepsilon$. For

$$A = \{x : \ f_0(x) < f_1(x)\}, \quad \mathbb{P}_n(A) = \frac{1}{n}\sum_{i=1}^{n} 1\!\!1_{X_i \in A},$$

$$\text{consider} \quad \phi_n = 1\!\!1\left\{\mathbb{P}_n(A) > P_{f_0}(A) + \frac{\|f_0 - f_1\|_1}{3}\right\}$$

Control of $E_{f_0}\phi_n$ and $E_f(1-\phi_n)$ via Hoeffding

Theory: Bayesian nonparametrics

Theorem 1 – entropy version [GGV 00]

If $\Theta_n \subset \Theta$ and $c > 0$ such that, for d_n such that **(T0)** is verified,

$$\log N(\varepsilon_n, \Theta_n, d_n) \leq dn\varepsilon_n^2 \qquad \textit{entropy}$$

$$\Pi(\Theta \backslash \Theta_n) \leq e^{-(c+4)n\varepsilon_n^2} \qquad \textit{remaining mass}$$

$$\Pi(B_{KL}(\theta_0, \varepsilon_n)) \geq e^{-cn\varepsilon_n^2} \qquad \textit{prior mass}$$

Then for $M > 0$ large enough,

$$E_{\theta_0}\Pi[\theta : d_n(\theta, \theta_0) \leq M\varepsilon_n | X] \to 1$$

Theory, Gaussian priors

$W = (W_t : t \in T)$ centered Gaussian process taking values in Banach space $(\mathbb{B}, \|\cdot\|)$

Covariance kernel $K(s,t) = E(W_s W_t)$

Theory, Gaussian priors

$W = (W_t : t \in T)$ centered Gaussian process taking values in Banach space $(\mathbb{B}, \|\cdot\|)$

Covariance kernel $K(s,t) = E(W_s W_t)$

Reproducing Kernel Hilbert Space $\mathbb{H}$ (RKHS) associated to W.

Define a norm $\|\cdot\|_{\mathbb{H}}$ via

$$\langle \sum_{i=1}^{p} a_i K(s_i,\cdot), \sum_{j=1}^{q} b_j K(t_j,\cdot)\rangle_{\mathbb{H}} = \sum_{i,j} a_i b_j K(s_i,t_j)$$

Then one sets

$$\mathbb{H} = \overline{\mathrm{Vect}\{K(s,\cdot), \quad s \in T\}}^{\mathbb{H}}$$

Brownian motion (B_t) $\qquad \mathbb{H} = \{ \int_0^{\cdot} f(u)du, \quad f \in L^2[0,1] \}$

Series prior $\sum_{k\geq 1} \sigma_k \nu_k \varepsilon_k$ $\qquad \mathbb{H} = \{h = (h_k) \in \ell^2, \quad \sum_{k\geq 1} \sigma_k^{-2} h_k^2 < +\infty\}$

Theory, Gaussian priors

Fact. For all g in the support of W in $\mathbb{B}$, and all $\varepsilon > 0$,

$$e^{-\varphi_g(\varepsilon/2)} \leq P(\|W - g\| < \varepsilon) \leq e^{-\varphi_g(\varepsilon)}$$

where

Concentration function. Let g be in the support of W in $\mathbb{B}$. For $\varepsilon > 0$, set

$$\varphi_g(\varepsilon) = \underbrace{\inf_{h\in\mathbb{H}:\ \|h-g\|<\varepsilon} \frac{\|h\|_{\mathbb{H}}^2}{2}}_{\text{Approximation}} \underbrace{- \log P(\|W\| < \varepsilon)}_{\text{Small ball probability}}$$

Example [small ball probability] Brownian motion (B_t)

$$-\log \mathbb{P}(\|B\|_\infty < \varepsilon) \approx \varepsilon^{-2} \quad (\varepsilon \to 0)$$

Theory, Gaussian priors

[van der Vaart, van Zanten 08]

Consider a nonparametric problem with unknown function $f_0 \in \mathbb{B}$.
Prior Π = law of a Gaussian process W on $\mathbb{B}$, with RKHS $\mathbb{H}$.

Assume that

- f_0 is in the support in $\mathbb{B}$ of the prior
- the norm $\|\cdot\|$ on $\mathbb{B}$ *combines correctly* with the testing distance d

Let ε_n be a solution of the equation

$$\varphi_{f_0}(\varepsilon_n) \leq n\varepsilon_n^2$$

Then the posterior contracts at rate ε_n: for large enough M,

$$E_{f_0}\Pi(d(f, f_0) > M\varepsilon_n | X) \to 0$$

Lower bound [C 08]

Theory, Gaussian priors

Ingredients of proof :

- prior mass
  the [Fact] links $P(\|W-w\|<\varepsilon)$ and concentration function
- sieves
  [Borell 75]'s inequality
  Let $\mathbb{B}_1$ and $\mathbb{H}_1$ unit balls $\mathbb{B}$ and $\mathbb{H}$ associated to W

$$P(W\notin M\mathbb{H}_1+\varepsilon\mathbb{B}_1)\leq 1-\Phi(\Phi^{-1}(e^{-\phi_0(\varepsilon)})+M)$$

  Suggests to set $\Theta_n=\sqrt{n}\varepsilon_n\mathbb{H}_1+\varepsilon_n\mathbb{B}_1$
- entropy
  can link entropy of $\mathbb{H}_1$ and small ball probability

Example *Density estimation and Gaussian priors*

Squared-exponential covariance kernel

Centered Gaussian process Z_t with covariance

$$E(Z_t Z_s) = e^{-(s-t)^2/L}$$

[van der Vaart, van Zanten 10] show that, for fixed L, there are regular functions f_0 for which the rate is at best *logarithmic*

$$\varepsilon_n \approx (\log n)^{-\gamma(\beta)}$$

Example *Density estimation and Gaussian priors*

Squared-exponential covariance kernel

Centered Gaussian process Z_t with covariance

$$E(Z_t Z_s) = e^{-(s-t)^2/L}$$

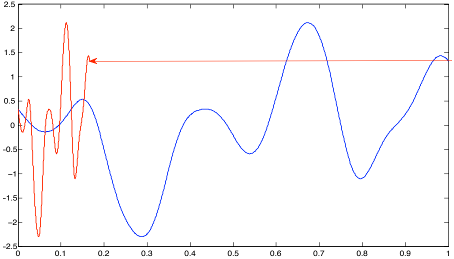
[van der Vaart, van Zanten 10] show that, for fixed L, there are regular functions f_0 for which the rate is at best *logarithmic*

$$\varepsilon_n \approx (\log n)^{-\gamma(\beta)}$$

However, those priors are used in machine learning and give very good results in practice when the parameter L is "well chosen" ...

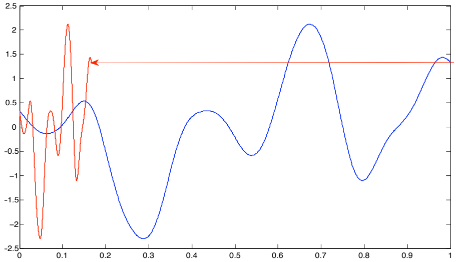
Example *Density estimation and Gaussian priors, adaptation*

[van der Vaart, van Zanten 09]



Example *Density estimation and Gaussian priors, adaptation*

[van der Vaart, van Zanten 09]



Prior Π: consider Z_{At} and set $t \to \dfrac{e^{Z_{At}}}{\int_0^1 e^{Z_{Au}}\,du}$, where

- A Gamma distribution
- $u \to Z_u$ centered GP with squared-exponential kernel

Then the posterior converges at minimax rate up to a log factor

$$E_{f_0}\Pi\left[h(f,f_0) \le M(\log n)^{\gamma(\beta)} n^{-\frac{\beta}{2\beta+1}} \mid X\right] \to 1$$

Example *Adaptation on manifolds*

[C, Kerkyacharian, Picard 14]

On $\mathcal{M}$ compact Riemannian manifold of dimension d without boundary
$\rightarrow$ Laplacian $\Delta_{\mathcal{M}}$ linear operator on functions on $\mathcal{M}$ with discrete spectrum

$$(-\Delta_{\mathcal{M}})\varphi_p = \lambda_p \varphi_p$$

$$0 \leq \lambda_1 \leq \lambda_2 \leq \cdots$$

Note that

$$\Delta_{\mathcal{M}}\left(e^{-\lambda_p t}\varphi_p\right) = -\lambda_p e^{-\lambda_p t}\varphi_p$$
$$\frac{\partial}{\partial t} e^{-\lambda_p t}\varphi_p = -\lambda_p e^{-\lambda_p t}\varphi_p$$

This is a special solution of the "heat equation" on $\mathcal{M}$

$$\Delta_{\mathcal{M}} f = \frac{\partial}{\partial t} f$$

Example *Adaptation on manifolds*

Let $\alpha_p \sim \mathcal{N}(0,1)$ iid. A GP "random solution of the heat equation" is

$$W^t := \sum_{p \geq 1} e^{-\lambda_p t/2} \alpha_p \varphi_p$$

The associated family of covariance kernels is

$$P_t(x,y) = \sum_{p \geq 1} e^{-\lambda_p t} \varphi_p(x) \varphi_p(y) \qquad \Delta_{\mathcal{M}}\text{-Heat Kernel}$$

Subgaussian estimates

$$\frac{c_1 e^{-c' \frac{\rho^2(x,y)}{t}}}{\sqrt{|B(x,\sqrt{t})||B(y,\sqrt{t})|}} \leq P_t(x,y) \leq \frac{c_2 e^{-c' \frac{\rho^2(x,y)}{t}}}{\sqrt{|B(x,\sqrt{t})||B(y,\sqrt{t})|}}$$

Example *Adaptation on manifolds*

$$dX^{(n)}(x) = f(x)dx + \frac{1}{\sqrt{n}} dZ(x), \quad x \in \mathcal{M}$$

Prior Π

- $W^T = \sum_p e^{-\lambda_p T/2} \alpha_p \varphi_p$ with $T \sim t^{-a} e^{-t^{-d/2} \log^q (1/t)}$
- W^T seen as prior on $(\mathbb{B}, \|\cdot\|) = (\mathbb{L}_2, \|\cdot\|_2)$
- Set $q = 1 + d/2$

Example *Adaptation on manifolds*

$$dX^{(n)}(x) = f(x)dx + \frac{1}{\sqrt{n}}dZ(x), \quad x \in \mathcal{M}$$

Prior Π

- $W^T = \sum_p e^{-\lambda_p T/2} \alpha_p \varphi_p$ with $T \sim t^{-a} e^{-t^{-d/2} \log^q(1/t)}$
- W^T seen as prior on $(\mathbb{B}, \|\cdot\|) = (\mathbb{L}_2, \|\cdot\|_2)$
- Set $q = 1 + d/2$

Suppose $f_0 \in B^s_{2,\infty}(\mathcal{M})$. Then for M large enough, as $n \to \infty$,

$$E_{f_0}\Pi\left[\|f - f_0\|_2 \geq M\left(\frac{\log n}{n}\right)^{s/(2s+d)} \mid X\right] \to 0$$

The rate is sharp

Example *Adaptation on manifolds*

$$dX^{(n)}(x) = f(x)dx + \frac{1}{\sqrt{n}} dZ(x), \quad x \in \mathcal{M}$$

Prior Π

- $W^T = \sum_p e^{-\lambda_p T/2} \alpha_p \varphi_p$ with $T \sim t^{-a} e^{-t^{-d/2} \log^q(1/t)}$
- W^T seen as prior on $(\mathbb{B}, \|\cdot\|) = (\mathbb{L}_2, \|\cdot\|_2)$
- Set $q = 1 + d/2$

Suppose $f_0 \in B^s_{2,\infty}(\mathcal{M})$. Then for M large enough, as $n \to \infty$,

$$E_{f_0} \Pi \left[\|f - f_0\|_2 \geq M \left(\frac{\log n}{n} \right)^{s/(2s+d)} \mid X \right] \to 0$$

The rate is sharp for small enough ρ, there exists f_0 in $B^s_{2,\infty}(\mathcal{M})$,

$$\Pi \left[\|f - f_0\|_2 \leq \rho \left(\frac{\log n}{n} \right)^{s/(2s+d)} \mid X \right] \to 0$$

Example *Sparsity*

Recall $X_i = \theta_i + \varepsilon_i$ and $\theta \in \ell_0[s]$ s–sparse vector

Example [coin–flipping prior with fixed $\alpha_n = 1/n$]

$$\begin{array}{ll} k & \sim \mathrm{Bin}(n, \alpha_n) = \pi_n \\ g & \quad \text{density on } \mathbb{R} \end{array}$$

$$\updownarrow$$

$$\Pi \sim \bigotimes_{i=1}^{n} (1 - \alpha_n)\delta_0 + \alpha_n g$$

Example *Sparsity*

Example [coin–flipping prior with random α]

$$\begin{aligned} \alpha &\sim \mathrm{Beta}(1, n) \\ k|\,\alpha &\sim \mathrm{Bin}(n, \alpha) = \pi_n \\ g &\quad \text{density on } \mathbb{R} \end{aligned}$$

$$\updownarrow$$

$$\begin{aligned} \alpha &\sim \mathrm{Beta}(1, n) \\ \Pi|\,\alpha &\sim \bigotimes_{i=1}^{n} (1-\alpha)\delta_0 + \alpha g \end{aligned}$$

Example *Sparsity*

Theorem. Let Π be the Bayesian coin–flipping prior with random α with

- $\alpha \sim \text{Beta}(1, n)$
- g the Laplace (or a heavier tailed) density

Then for M large enough,

$$\sup_{\theta_0 \in \ell_0[s_n]} E_{\theta_0} \Pi \left[\theta : \|\theta - \theta_0\| > M s_n \log(n/s_n) \,|\, X\right] \to 0.$$

Example *Mixtures*

[Rousseau 10]

Observations $X_1, \ldots, X_n$ i.i.d. density $f_0 > 0$ and Hölder $\mathcal{C}^\beta[0,1]$

Beta densities $\quad g(x|a,b) = \frac{x^{a-1}(1-x)^{b-1}}{B(a,b)} \quad g_{\alpha,\varepsilon}(x) = g(x|\frac{\alpha}{1-\varepsilon}, \frac{\alpha}{\varepsilon})$

usual parametrisation reparametrisation

Prior Π = hierarchical mixture of Beta densities

$$g_{\alpha,k,\mathbf{p}^k,\epsilon^k}(\cdot) = \sum_{j=1}^{k} p_j g_{\alpha,\varepsilon_j}(\cdot)$$

Example *Mixtures*

[Rousseau 10]

Observations $X_1, \ldots, X_n$ i.i.d. density $f_0 > 0$ and Hölder $\mathcal{C}^\beta[0,1]$

Beta densities $\quad g(x|a,b) = \frac{x^{a-1}(1-x)^{b-1}}{B(a,b)} \qquad g_{\alpha,\varepsilon}(x) = g(x|\frac{\alpha}{1-\varepsilon}, \frac{\alpha}{\varepsilon})$

usual parametrisation $\qquad$ reparametrisation

Prior Π = hierarchical mixture of Beta densities

$$g_{\alpha,k,\mathbf{p}^k,\epsilon^k}(\cdot) = \sum_{j=1}^{k} p_j g_{\alpha,\varepsilon_j}(\cdot)$$

- $\alpha \sim \pi_\alpha$
- $k \sim \pi_k$
- $\mathbf{p}^k \,|\, k \sim (p_1, \ldots, p_k)$ law on the canonical simplex of $\mathbb{R}^k$
- $\epsilon^k \,|\, k \sim (\varepsilon_1, \ldots, \varepsilon_k)$ law in $(0,1)^k$

Then the posterior concentrates at rate

$$E_{f_0}\Pi\left[h(f,f_0) \leq M(\log n)^{\rho(\beta)} n^{-\frac{\beta}{2\beta+1}} \mid X\right] \to 1$$

Posterior measure and aspects of the measure (I)

MESSAGE Full posterior measure and aspects of it may differ significantly !

Posterior measure and aspects of the measure (I)

MESSAGE Full posterior measure and aspects of it may differ significantly !

$$Y \;=\; I_{n\times n}\theta \;+\; \epsilon \qquad [seq.model]$$

Let $\bar{\Pi}_\lambda \sim \otimes_{i=1}^{p}$ Laplace(λ) [without point masses at zero]

- $\hat{\theta}_\lambda^{LASSO} = \arg\max_\theta \left[-\|Y - X\theta\|_2^2 - 2\lambda\|\theta\|_1\right]$ posterior mode of $\bar{\Pi}_\lambda[\cdot \mid Y]$

Posterior measure and aspects of the measure (I)

MESSAGE Full posterior measure and aspects of it may differ significantly !

$$Y = I_{n\times n}\theta + \epsilon \qquad [seq.model]$$

Let $\bar{\Pi}_\lambda \sim \otimes_{i=1}^p$ Laplace(λ) [without point masses at zero]

- $\hat{\theta}_\lambda^{LASSO} = \arg\max_\theta \left[-\|Y - X\theta\|_2^2 - 2\lambda\|\theta\|_1 \right]$ posterior mode of $\bar{\Pi}_\lambda[\cdot \mid Y]$

Lemma [C, Schmidt-Hieber, van der Vaart 15] There exists $\delta > 0$ such that, as $n \to \infty$,

$$E_{\theta^0=0}\bar{\Pi}_{\lambda_n}\left(\theta : \|\theta\|_2 \leq \delta \frac{\sqrt{n}}{\lambda_n} \mid Y\right) \to 0 \quad \text{with } \lambda_n = \lambda_n^{LASSO} = C \log n.$$

Posterior measure and aspects of the measure (I)

MESSAGE Full posterior measure and aspects of it may differ significantly !

$$Y \;=\; I_{n\times n}\theta \;+\; \epsilon \qquad [seq.model]$$

Let $\bar{\Pi}_\lambda \sim \otimes_{i=1}^{p}$ Laplace(λ) [without point masses at zero]

- $\hat{\theta}_\lambda^{LASSO} = \arg\max_\theta \left[-\|Y - X\theta\|_2^2 - 2\lambda\|\theta\|_1 \right]$ posterior mode of $\bar{\Pi}_\lambda[\cdot \mid Y]$

Lemma [C, Schmidt-Hieber, van der Vaart 15] There exists $\delta > 0$ such that, as $n \to \infty$,

$$E_{\theta^0=0}\bar{\Pi}_{\lambda_n}\left(\theta : \|\theta\|_2 \leq \delta\frac{\sqrt{n}}{\lambda_n} \mid Y\right) \to 0 \quad \text{with } \lambda_n = \lambda_n^{LASSO} = C\log n.$$

Note $\sqrt{n}/\log n$ suboptimal convergence rate!

Posterior measure and aspects of the measure (II)

$$Y \;=\; I_{n\times n}\theta \;+\; \varepsilon$$

Prior Π on θ Bayesian coin–flipping with random α and g (e.g.) the Laplace density

Consider estimation of θ under ℓ^q metric, $0 < q \leq 2$,

$$d_q(\theta, \theta') = \sum_{i=1}^{n} |\theta_i - \theta_i'|^q.$$

Posterior measure and aspects of the measure (II)

$$Y \;=\; I_{n\times n}\theta \;+\; \varepsilon$$

Prior Π on θ Bayesian coin–flipping with random α and g (e.g.) the Laplace density

Consider estimation of θ under ℓ^q metric, $0 < q \leq 2$,

$$d_q(\theta, \theta') = \sum_{i=1}^{n} |\theta_i - \theta_i'|^q.$$

Theorem [C, van der Vaart 12] For large enough M, as $n \to \infty$,

$$\sup_{\theta_0 \in \ell_0[s_n]} E_{\theta_0} \Pi\left[\theta : d_q(\theta, \theta_0) > M r_{n,q}^* \,|\, Y\right] \to 0$$

- Posterior measure is rate-optimal for any $0 < q \leq 2$
- Posterior mean is suboptimal (!) for $q < 1$ [Johnstone, Silverman 04]
- Posterior median is rate-optimal for any $0 < q \leq 2$ [Johnstone, Silverman 04], [C, van der Vaart 12]

Rates: conclusion and further topics

The previous general rate theorem applies to a variety of problems

- in i.i.d., non-i.i.d. Markov, hidden Markov, ...
- as soon as one can find a suitable testing distance d_n
- and verify prior–mass type conditions

Sometimes refinements can be used to tackle specific problems
[e.g. more precise assumptions using similar ideas]

Other formulations/approaches are also possible: PAC Bayes, non–asymptotic versions etc.

Yet, with these techniques, unclear how to obtain results for other distances/loss functions that do not satify testing (T0)

Part III

Shape

Gaussian example

E0 $\quad X | \theta \sim \mathcal{N}(\theta, 1)^{\otimes n}; \quad \theta \sim \mathcal{N}(0,1) = \Pi$

$$\Pi[\cdot | X] = \mathcal{N}\left(\frac{n\overline{X}_n}{n+1}, \frac{1}{n+1}\right)$$

Is $\Pi[\cdot | X]$, after recentering and rescaling, close to $\mathcal{N}(0,1)$?

Gaussian example

E0 $X|\theta \sim \mathcal{N}(\theta,1)^{\otimes n}; \quad \theta \sim \mathcal{N}(0,1) = \Pi$

$$\Pi[\cdot\,|\,X] = \mathcal{N}\left(\frac{n\overline{X}_n}{n+1}, \frac{1}{n+1}\right)$$

Is $\Pi[\cdot\,|\,X]$, after recentering and rescaling, close to $\mathcal{N}(0,1)$?

- *recentering and rescaling.* Set

$$\tau_a : x \to \sqrt{n}(x-a)$$

- *comparing distributions.* $\|P-Q\|_{TV} = \sup_A |P(A)-Q(A)| = \frac{1}{2}\|P-Q\|_1$

The recentering choice $a = \bar{X}_n$ seems natural. Then

$$\Pi[\cdot\,|\,X] \circ \tau_{\bar{X}_n}^{-1} = -\frac{\sqrt{n}\bar{X}_n}{n+1} + \sqrt{\frac{n}{n+1}}\mathcal{N}(0,1).$$

Gaussian example

E0 $\quad X|\,\theta \sim \mathcal{N}(\theta,1)^{\otimes n}; \quad \theta \sim \mathcal{N}(0,1) = \Pi$

$$\Pi[\cdot\,|\,X] = \mathcal{N}\left(\frac{n\overline{X}_n}{n+1}, \frac{1}{n+1}\right)$$

Set

$$\tau_{\bar{X}_n} : x \to \sqrt{n}(x - \bar{X}_n)$$

Proposition [convergence in Gaussian conjugate case]

$$E_{\theta_0} \left\| \Pi\left[\cdot\,|\,X\right] \circ \tau_{\bar{X}_n}^{-1} - \mathcal{N}(0,1) \right\|_{TV} \to 0$$

Exercise. Prove it.

[For instance: use $\|\cdot\|_1^2 \leq KL$ and compute KL distance between gaussians]

Historical example

Laplace [towards 1810] considered the setting

$$X|\,\theta \sim \mathrm{Bin}(n,\theta)$$
$$\theta \sim \mathrm{Unif}[0,1] = \Pi$$

with associated posterior [a fact noted by Thomas Bayes a few decades before]

$$\Pi[\cdot\,|\,X] \sim \mathrm{Beta}(X+1, n-X+1)$$

Laplace noticed that, with $\tau_{X/n} : x \to \sqrt{n}(x - X/n)$,

$$\Pi[\cdot\,|\,X] \circ \tau_{X/n}^{-1} \approx \mathcal{N}(0,1)$$

Regular parametric models

$\mathcal{P} = \{P_\theta,\ \theta \in \Theta \subset \mathbb{R}^k\}$

- $dP_\theta(x) = p_\theta(x)dx, \quad \ell_\theta := \log p_\theta$
- Suppose $\dot{\ell}_\theta = \partial\ell_\theta/\partial_\theta$ exists and $0 < \mathcal{I}_\theta := E_\theta[\dot{\ell}_\theta\dot{\ell}_\theta^T] < \infty$.

The model is locally asymptotically normal (LAN) at point $\theta_0 \in \Theta$ if $\forall h \in \mathbb{R}^k$

$$\log \prod_{i=1}^n \frac{p_{\theta_0+h/\sqrt{n}}}{p_{\theta_0}}(X_i) = \frac{1}{\sqrt{n}} h^T \sum_{i=1}^n \dot{\ell}_{\theta_0}(X_i) - \frac{1}{2} h^T \mathcal{I}_{\theta_0} h + o_{P_{\theta_0}}(1).$$

An estimator $\hat{\theta} = \hat{\theta}_n(X)$ is (asymp. linear and) efficient at θ_0 if

$$\sqrt{n}(\hat{\theta} - \theta_0) = \mathcal{I}_{\theta_0}^{-1} \frac{1}{\sqrt{n}} \sum_{i=1}^n \dot{\ell}_{\theta_0}(X_i) + o_{P_{\theta_0}}(1).$$

Recall the notation $\tau_{\hat{\theta}} : x \to \sqrt{n}(x - \hat{\theta})$

Bernstein–von Mises, smooth parametric models

$X_1, \dots, X_n | \, \theta \sim P_\theta^{\otimes n}, \; \theta \sim \Pi$

Theorem [Bernstein–von Mises] [Le Cam – van der Vaart] Suppose

(S) the model is LAN at θ_0

(T) for any $\varepsilon > 0$ there exist tests ϕ_n with

$$E_{\theta_0} \phi_n = o(1), \qquad \sup_{|\theta - \theta_0| > \varepsilon} E_\theta (1 - \phi_n) = o(1)$$

(P) the prior Π has a continuous positive density at θ_0

Then for $\hat{\theta}_n$ efficient estimator at θ_0, as $n \to \infty$,

$$\left\| \Pi[\cdot \,|\, X] \circ \tau_{\hat{\theta}_n}^{-1} - \mathcal{N}\left(0, \mathcal{I}_{\theta_0}{}^{-1}\right)(\cdot) \right\|_1 \longrightarrow 0 \quad \text{under } P_{\theta_0}.$$

Bernstein–von Mises, smooth parametric models

By the Bernstein–von Mises (BvM) theorem,

$$\left\| \Pi[\cdot \mid X] \circ \tau_{\hat\theta_n}^{-1} - \mathcal{N}\left(0, {\mathcal{I}_{\theta_0}}^{-1}\right)(\cdot) \right\|_1 \longrightarrow 0 \quad \text{under } P_{\theta_0}.$$

For $\hat\theta$ efficient estimator, under P_{θ_0},

$$\sqrt{n}(\hat\theta - \theta_0) \xrightarrow{d} \mathcal{N}(0, \mathcal{I}_{\theta_0}^{-1})$$

This shows the remarkable 'duality', asymptotically,

$$\mathcal{L}^{\Pi}(\sqrt{n}(\theta - \hat\theta) \mid X) \quad \approx \quad \mathcal{L}(\sqrt{n}(\hat\theta - \theta) \mid \theta = \theta_0)$$

Bayes law $\approx$ Frequentist law

Parametric BvM implies: credible sets are confidence sets!

Suppose BvM holds for $\Theta \subset \mathbb{R}$ [dimension 1]

$$\left\| \Pi(\cdot|X) \circ \tau_{\hat{\theta}}^{-1} - \mathcal{N}\left(0, \mathcal{I}_{\theta_0}^{-1}\right)(\cdot) \right\|_1 \longrightarrow 0 \quad \text{under } P_{\theta_0}.$$

Let A_n, B_n be such that, for $\alpha = 5\%$,

$$\Pi((-\infty, A_n)|\, X) = \Pi((B_n, +\infty)|\, X) = \alpha/2$$

By definition $\Pi\left[(A_n, B_n)|\, X\right] = 1 - \alpha$ Credible set

Exercise. As $n \to \infty$,

$$P_{\theta_0}\left[\theta_0 \in (A_n, B_n)\right] \to 1 - \alpha.$$

so that (A_n, B_n) is an asymptotic confidence set at level $1 - \alpha$.

So, if the *parametric* BvM holds, credible sets are (asymptotically) confidence sets.

BvM in Gaussian model for nonconjugate prior

As a particular case of the previous BvM theorem, we have

Proposition [BvM for nonconjugate prior in Gaussian model]. Let

$$X_1, \ldots, X_n |\, \theta \sim \mathcal{N}(\theta, 1)$$

$$\theta \sim \Pi$$

with $d\Pi(\theta) = \pi(\theta)d\theta$ and π continuous and positive density on $\mathbb{R}$. Then

$$\left\| \Pi[\cdot \,|\, X] \circ \tau_{\bar{X}_n}^{-1} - \mathcal{N}(0,1)(\cdot) \right\|_1 \to^P 0.$$

BvM in Gaussian model for nonconjugate prior

As a particular case of the previous BvM theorem, we have

Proposition [BvM for nonconjugate prior in Gaussian model]. Let

$$X_1, \ldots, X_n | \, \theta \sim \mathcal{N}(\theta, 1)$$

$$\theta \sim \Pi$$

with $d\Pi(\theta) = \pi(\theta) d\theta$ and π continuous and positive density on $\mathbb{R}$. Then

$$\left\| \Pi[\cdot \,|\, X] \circ \tau_{\bar{X}_n}^{-1} - \mathcal{N}(0,1)(\cdot) \right\|_1 \to^P 0.$$

Idea of proof

$$g_n^X(u) := \frac{\exp(-u^2/2)\pi(\bar{X}_n + \frac{u}{\sqrt{n}})}{\int \exp(-u^2/2)\pi(\bar{X}_n + \frac{u}{\sqrt{n}}) du} \to \frac{\exp(-u^2/2)\pi(\theta_0)}{\sqrt{2\pi}\pi(\theta_0)}$$

use Scheffé's lemma on event $\{|\bar{X}_n - \theta_0| \leq 1/\log(n)\}$

Towards a nonparametric BvM?

And what about a nonparametric BvM?

Gaussian white noise model

$$dX^{(n)}(t) = f(t)dt + \frac{1}{\sqrt{n}}dW(t) \qquad\qquad t \in [0,1]$$

- Observe a trajectory $X^{(n)}$
- $f \in L^2[0,1] =: L^2$ is unknown = object of interest

Gaussian white noise model

$$dX^{(n)}(t) = f(t)dt + \frac{1}{\sqrt{n}}dW(t) \qquad\qquad t \in [0,1]$$

- Observe a trajectory $X^{(n)}$
- $f \in L^2[0,1] =: L^2$ is unknown = object of interest

$$\int \varphi_k(t)dX^{(n)}(t) = \int \varphi_k(t)f(t)dt + n^{-1/2}\int \varphi_k(t)dW(t), \qquad k \geq 1$$

$$X_k = f_k + n^{-1/2}\varepsilon_k, \qquad k \geq 1$$

Gaussian white noise model

$$dX^{(n)}(t) = f(t)dt + \frac{1}{\sqrt{n}}dW(t) \qquad t \in [0,1]$$

- Observe a trajectory $X^{(n)}$
- $f \in L^2[0,1] =: L^2$ is unknown = object of interest

$$\int \varphi_k(t)dX^{(n)}(t) = \int \varphi_k(t)f(t)dt + n^{-1/2}\int \varphi_k(t)dW(t), \qquad k \geq 1$$

$$X_k = f_k + n^{-1/2}\varepsilon_k, \qquad k \geq 1$$

Equivalently, writing $\mathbb{X}^{(n)}$ for the sequence $\{X_k,\ k \geq 1\}$,

$$\mathbb{X}^{(n)} = f + n^{-1/2}\mathbb{W}$$

Bernstein-von Mises, nonparametric ?

$$\mathbb{X}^{(n)} = f + n^{-1/2}\mathbb{W}$$

- Let Π be a prior on $f \in L^2$ $\longleftrightarrow$ prior on the coordinates $f_k = \langle f, \varphi_k \rangle$, $k \geq 1$
- Set $\hat{f} := E^{\Pi}[f \,|\, X^{(n)}]$ posterior mean

Two questions

1. Can one find Π in such a way that

$$\Pi(f - \hat{f} \,|\, X^{(n)}) \;\approx\; \mathcal{L}(\hat{f} - f \,|\, f) \;\approx\; \text{optimal law in some sense} \quad ?$$

2. In which sense should one interpret $\approx$?

Nonparametric BvMs, related work

- Negative results [Cox 93], [Freedman 99], [Leahu 11]

  Let Π be a Gaussian prior on f sitting on L^2

  A nonparametric BvM does not hold in L^2

- Some positive results exist for specific models/priors
  Mostly in cases where some form of conjugacy is present

  - [Lo 80s] Consider a i.i.d. sample $X_1, \ldots, X_n \sim P$ in $\mathbb{R}$ with c.d.f. F

    Dirichlet process prior on $P \rightarrow$ nonparametric-BvM for F

    $$\sqrt{n}(F - F_n) | X_1, \ldots, X_n \xrightarrow{d} \mathcal{G},$$

    in probability, with $\mathcal{G}$ law of a P-Brownian bridge.

  - [Kim and Lee 06] Neutral to the right prior on F in Survival analysis model

    $\rightarrow$ NP-BvM for A cumulative hazard function

A large enough Hilbert space

Standard Sobolev spaces For $\{\varphi_k,\ k \geq 1\}$ smooth enough orthonormal basis of L^2

$$H_2^s := \left\{ f \in L^2([0,1]) :\ \|f\|_{s,2}^2 := \sum_{k\geq 1} k^{2s} |\langle \varphi_k, f\rangle|^2 < \infty \right\}$$

A large enough Hilbert space

Standard Sobolev spaces $\{\psi_{lk}\}$ smooth enough orthonormal basis of L^2,

$$H_2^s := \left\{ f \in L^2([0,1]) : \ \|f\|_{s,2}^2 := \sum_{l \in \mathcal{L}} a_l^{2s} \sum_{k \in \mathcal{Z}_l} |\langle \psi_{lk}, f \rangle|^2 < \infty \right\}$$

1. $|\mathcal{Z}_l| = 1$, $a_l = \max(2, |l|)$ $\rightarrow$ Fourier-type basis
2. $\mathcal{L} \subset \mathbb{N}$, $a_l = |\mathcal{Z}_l| = 2^l$ $\rightarrow$ wavelet-type basis

Negative Sobolev space: for $s \in \mathbb{R}$,

$$H_2^s \equiv \left\{ f : \ \|f\|_{s,2}^2 := \sum_{l \in \mathcal{L}} a_l^{2s} \sum_{k \in \mathcal{Z}_l} |\langle \psi_{lk}, f \rangle|^2 < \infty \right\}$$

A large enough Hilbert space H

We use ‘logarithmic Sobolev spaces’ to get sharp rates. For $\delta \geq 1$ and $s = -1/2$, let

$$H := H_2^{-1/2,\delta} \equiv \left\{ f : \ \|f\|^2_{-1/2,2,\delta} := \sum_{l \in \mathcal{L}} \frac{a_l^{2(-1/2)}}{(\log a_l)^{2\delta}} \sum_{k \in \mathcal{Z}_l} |\langle \psi_{lk}, f \rangle|^2 < \infty \right\}$$

A large enough Hilbert space H

We use 'logarithmic Sobolev spaces' to get sharp rates. For $\delta \geq 1$ and $s = -1/2$, let

$$H := H_2^{-1/2,\delta} \equiv \left\{ f : \ \|f\|^2_{-1/2,2,\delta} := \sum_{l \in \mathcal{L}} \frac{a_l^{2(-1/2)}}{(\log a_l)^{2\delta}} \sum_{k \in \mathcal{Z}_l} |\langle \psi_{lk}, f \rangle|^2 < \infty \right\}$$

- $\mathbb{X}^{(n)} = f + \frac{1}{\sqrt{n}} \mathbb{W}$ white noise model in H
- Note that $\mathbb{W}$ a.s. belong to H (as well as $\mathbb{X}^{(n)}, f$), as

$$E\|\mathbb{W}\|^2_{-1/2,2,\delta} = \sum_{l \in \mathcal{L}} a_l^{-1} (\log a_l)^{-2\delta} \sum_{k \in \mathcal{Z}_l} E g_{lk}^2 < \infty.$$

[This implies that $\mathbb{W}$ takes values in H a.s. and is tight in H]

$\mathbb{W}$ is a centered Gaussian measure on H with $E\mathbb{W}(g)\mathbb{W}(h) = \langle g, h \rangle \ \ \forall f, g \in L^2$.

A large enough Hilbert space H

We use 'logarithmic Sobolev spaces' to get sharp rates. For $\delta \geq 1$ and $s = -1/2$, let

$$H := H_2^{-1/2,\delta} \equiv \left\{ f : \|f\|^2_{-1/2,2,\delta} := \sum_{l \in \mathcal{L}} \frac{a_l^{2(-1/2)}}{(\log a_l)^{2\delta}} \sum_{k \in \mathcal{Z}_l} |\langle \psi_{lk}, f \rangle|^2 < \infty \right\}$$

- $\mathbb{X}^{(n)} = f + \frac{1}{\sqrt{n}} \mathbb{W}$ white noise model in H
- Note that $\mathbb{W}$ a.s. belong to H (as well as $\mathbb{X}^{(n)}$, f), as

$$E\|\mathbb{W}\|^2_{-1/2,2,\delta} = \sum_{l \in \mathcal{L}} a_l^{-1} (\log a_l)^{-2\delta} \sum_{k \in \mathcal{Z}_l} E g_{lk}^2 < \infty.$$

[This implies that $\mathbb{W}$ takes values in H a.s. and is tight in H]

$\mathbb{W}$ is a centered Gaussian measure on H with $E\mathbb{W}(g)\mathbb{W}(h) = \langle g, h \rangle \ \ \forall f, g \in L^2$.

- Denote by $\mathcal{N}$ the law of $\mathbb{W}$ as a r.v. in H [limiting non-parametric distribution]

Weak nonparametric BvM, definition

$$\mathbb{X}^{(n)} = f + \frac{1}{\sqrt{n}}\mathbb{W}$$

- Let Π prior on $f \in L^2 \subset H$. Let $\Pi_n = \Pi(\cdot \mid \mathbb{X}^{(n)})$ be the posterior distribution on H.

Weak nonparametric BvM, definition

$$\mathbb{X}^{(n)} = f + \frac{1}{\sqrt{n}}\mathbb{W}$$

- Let Π prior on $f \in L^2 \subset H$. Let $\Pi_n = \Pi(\cdot \,|\, \mathbb{X}^{(n)})$ be the posterior distribution on H.
- Denote $\tau : \; f \mapsto \sqrt{n}(f - \mathbb{X}^{(n)})$ and $\Pi_n \circ \tau^{-1}$ be the rescaled posterior.

Weak nonparametric BvM, definition

$$\mathbb{X}^{(n)} = f + \frac{1}{\sqrt{n}}\mathbb{W}$$

- Let Π prior on $f \in L^2 \subset H$. Let $\Pi_n = \Pi(\cdot \mid \mathbb{X}^{(n)})$ be the posterior distribution on H.
- Denote $\tau : \ f \mapsto \sqrt{n}(f - \mathbb{X}^{(n)})$ and $\Pi_n \circ \tau^{-1}$ be the rescaled posterior.
- Let β be the bounded Lipschitz metric for weak cv. of probability measures on H

$$\beta(\mu, \nu) = \sup_{u \in BL(1)} \left| \int_S u(s)(d\mu - d\nu)(s) \right|$$

Weak nonparametric BvM, definition

$$\mathbb{X}^{(n)} = f + \frac{1}{\sqrt{n}} \mathbb{W}$$

- Let Π prior on $f \in L^2 \subset H$. Let $\Pi_n = \Pi(\cdot \mid \mathbb{X}^{(n)})$ be the posterior distribution on H.
- Denote $\tau : \ f \mapsto \sqrt{n}(f - \mathbb{X}^{(n)})$ and $\Pi_n \circ \tau^{-1}$ be the rescaled posterior.
- Let β be the bounded Lipschitz metric for weak cv. of probability measures on H

$$\beta(\mu, \nu) = \sup_{u \in BL(1)} \left| \int_S u(s)(d\mu - d\nu)(s) \right|$$

Definition The prior Π satisfies the weak Bernstein - von Mises phenomenon in H if

$$\beta(\Pi_n \circ \tau^{-1}, \mathcal{N}) \to^{P^n_{f_0}} 0, \quad (n \to \infty).$$

Product priors and γ-smooth functions

Consider priors of the form

$$\Pi \sim \bigotimes_{l,k} \pi_{lk}$$

- defined on the coordinates of the basis $\{\psi_{lk}\}$, with
- π_{lk} probability measures with Lebesgue density φ_{lk} on $\mathbb{R}$.
- Further assume, for some fixed density φ,

$$\varphi_{lk}(\cdot) = \frac{1}{\sigma_l}\varphi\left(\frac{\cdot}{\sigma_l}\right) \quad \forall k \in \mathcal{Z}_l \quad \text{with } \sigma_l > 0.$$

For instance, to model γ-smooth functions, take $\sigma_l = 2^{-l(\gamma+1/2)}$ and set, with $g_{lk} \sim \varphi$,

$$G_\gamma = \sum_l \sum_{k \in \mathcal{Z}_l} 2^{-l(\gamma+1/2)} g_{lk}\psi_{lk}, \quad \gamma > 0$$

Main theorem - Condition (P)

Denote $f_{0,lk} = \langle f_0, \psi_{lk} \rangle$ the coordinates of the true f_0 on basis $\{\psi_{lk}\}$.

Suppose that for some $M > 0$,

$$\textbf{(P1)} \qquad \sup_{l,k} |f_{0,lk}|/\sigma_l \leq M.$$

Suppose also that φ is bounded and that for some $\tau > M$, $c_\varphi > 0$,

$$\textbf{(P2)} \qquad \varphi(x) \geq c_\varphi \ \ \forall x \in (-\tau, \tau), \quad \int_{\mathbb{R}} x^2 \varphi(x) dx < \infty.$$

For the previous wavelet prior, **(P1)** asks for

$$|f_{0,lk}| \leq M 2^{-l(\gamma+1/2)} \quad \forall l, k \qquad \Longleftrightarrow \qquad f_0 \in \mathcal{C}^\gamma.$$

Main theorem in white noise [C. and Nickl 13]

Theorem [Weak nonparametric BvM in H]

Any product prior Π and f_0 satisfying Conditions (P) verify the weak Bernstein-von Mises phenomenon in H,

$$\beta(\Pi(\cdot \mid X^{(n)}) \circ \tau^{-1}, \mathcal{N}) \to^{P^n_{f_0}} 0.$$

Moreover the posterior mean $\hat{f}_n = E^{\Pi}[f \mid X^{(n)}]$ is efficient in the sense that

$$\|\hat{f}_n - \mathbb{X}^{(n)}\|_H = o_P(1/\sqrt{n}).$$

Nonparametric BvM: Note on weak convergence

Unlike in total variation convergence one does not have uniformity in all Borel sets B in

$$|\Pi(\cdot \mid X) \circ \tau_T^{-1}(B) - \mathcal{N}(B)| \to^{P_{f_0}^n} 0.$$

One does have uniformity in all sets that have a uniformly smooth boundary for the probability measure $\mathcal{N}$.

In particular uniformity holds for all $\|\cdot\|_H$-balls

$$\{B_H(0,t) : \ 0 < t \leq M\}$$

Application: Credible ellipsoids

Suppose the weak BvM theorem holds for Π.

A natural $(1-\alpha)$-credible set is then obtained by solving for $R_n = R_n(\alpha, X^{(n)})$ such that

$$\Pi(C_n | X^{(n)}) = 1 - \alpha, \quad \text{where} \quad C_n = \left\{ f : \ \|f - \mathbb{X}^{(n)}\|_H \leq R_n/\sqrt{n} \right\}$$

C_n is the smallest ball around $\mathbb{X}^{(n)}$ charged by the posterior with probability $1-\alpha$.

Theorem [Elliptical Credible set C_n]

The random set C_n has confidence $1-\alpha$ asymptotically in that

$$P_{f_0}^n(f_0 \in C_n) \to 1 - \alpha \quad \text{and} \quad R_n = O_P(1).$$

Application: BvM and Credible sets for smooth functionals

- Linear functionals, for $g_L \in H_2^s$, $s > 1/2$,

$$L: \ f \to \int_0^1 f(t) g_L(t) dt.$$

Application: BvM and Credible sets for smooth functionals

- Linear functionals, for $g_L \in H_2^s$, $s > 1/2$,

$$L: \ f \to \int_0^1 f(t) g_L(t) dt.$$

- Smooth nonlinear functionals, e.g.

$$f \to \int_0^1 f^2(t) dt, \quad f_0 \in \mathcal{C}^\beta, \ \beta > 1/2.$$

- Self-convolutions of 1-periodic functions

$$f \to f * f$$

- $\cdots$

Leads to BvM for these functionals

Deduce confident credible sets by taking quantiles of the induced posterior

Application: Confidence sets in L^2, uniform priors

Application: Confidence sets in L^2, uniform priors

Consider first the special case of a uniform wavelet prior Π on L^2

$$U_{\gamma,M} = \sum_{l,k} 2^{-l(\gamma+1/2)} u_{lk}\psi_{lk}(\cdot), \quad u_{lk} \sim \mathcal{U}[-M, M] \text{ i.i.d.}$$

Such priors model functions in a Hölder ball of radius M, with posteriors Π_n contracting about f_0 at the L^2-minimax rate $n^{-\gamma/(2\gamma+1)}$ within log factors if $\|f_0\|_{\gamma,\infty} \leq M$.

It is natural to intersect the ellipsoid credible set C_n with the Hölderian support of Π

$$C_n' = \left\{ f : \|f - \hat{f}_n\|_H \leq R_n/\sqrt{n}, \quad \|f\|_{\gamma,\infty} \leq M \right\}$$

Application: Confidence sets in L^2, uniform priors

Consider first the special case of a uniform wavelet prior Π on L^2

$$U_{\gamma,M} = \sum_{l,k} 2^{-l(\gamma+1/2)} u_{lk}\psi_{lk}(\cdot), \quad u_{lk} \sim \mathcal{U}[-M, M] \text{ i.i.d.}$$

Such priors model functions in a Hölder ball of radius M, with posteriors Π_n contracting about f_0 at the L^2-minimax rate $n^{-\gamma/(2\gamma+1)}$ within log factors if $\|f_0\|_{\gamma,\infty} \leq M$.

It is natural to intersect the ellipsoid credible set C_n with the Hölderian support of Π

$$C_n' = \left\{ f : \|f - \hat{f}_n\|_H \leq R_n/\sqrt{n}, \quad \|f\|_{\gamma,\infty} \leq M \right\}$$

Proposition [Confidence sets for uniform product priors]

$$\Pi(C_n' | X^{(n)}) = 1 - \alpha, \quad P_{f_0}^n(f_0 \in C_n') \to 1 - \alpha$$

and the L^2-diameter $|C_n'|_2$ of C_n' satisfies, for some $\kappa > 0$,

$$|C_n'|_2 = O_P(n^{-\gamma/(2\gamma+1)}(\log n)^{\kappa}).$$

Application: Confidence sets in L^2, general product priors

Consider more generally, if $f_0 \in \mathcal{C}^\gamma$,

$$G_\gamma = \sum_{l,k} 2^{-l(\gamma+1/2)} g_{lk} \psi_{lk}(\cdot), \quad g_{lk} \sim \text{ i.i.d. } \varphi$$

Let $\delta > 0$ arbitrary. Let, for $\| \cdot \|_{\gamma,2,1}$ the γ-Sobolev norm with log correction,

$$C_n'' = \left\{ f : \|f - \hat{f}_n\|_H \leq R_n/\sqrt{n}, \quad \|f\|_{\gamma,2,1} \leq M_n + 4\delta \right\},$$

where M_n is defined as a 'quantile' for the γ-norm

Application: Confidence sets in L^2, general product priors

Consider more generally, if $f_0 \in \mathcal{C}^\gamma$,

$$G_\gamma = \sum_{l,k} 2^{-l(\gamma+1/2)} g_{lk}\psi_{lk}(\cdot), \quad g_{lk} \sim \text{ i.i.d. } \varphi$$

Let $\delta > 0$ arbitrary. Let, for $\|\cdot\|_{\gamma,2,1}$ the γ-Sobolev norm with log correction,

$$C_n'' = \left\{ f : \|f - \hat{f}_n\|_H \leq R_n/\sqrt{n}, \quad \|f\|_{\gamma,2,1} \leq M_n + 4\delta \right\},$$

where M_n is defined as a 'quantile' for the γ-norm: for any n and $\delta_n = (\log n)^{-1/4}$,

$$M_n = \inf\left\{ M > 0 : \quad \Pi_n(f : \,|\|f\|_{\gamma,2,1} - M| \leq \delta) \geq 1 - \delta_n \right\},$$

Proposition [Confidence sets for general product priors]

$$P_{f_0}^n(f_0 \in C_n'') \to 1 - \alpha, \qquad \Pi(C_n'' \,|\, X^{(n)}) = 1 - \alpha + o_P(1)$$

$$|C_n''|_2 = O_P(n^{-\gamma/(2\gamma+1)}(\log n)^\kappa).$$

Nonparametric BvMs in white noise, further questions

- Adaptation
  Prior on regularity α [Kolyan Ray] $\to$ one obtains e.g. adaptive confidence regions (modulo expected appropriate restrictions)

Nonparametric BvMs in white noise, further questions

- Adaptation
  Prior on regularity α [Kolyan Ray] $\rightarrow$ one obtains e.g. adaptive confidence regions (modulo expected appropriate restrictions)

- How about, still in the white noise model, getting confidence sets for a *different norm*, for instance the L^∞-norm instead of L^2 ?

Nonparametric BvMs in white noise, further questions

- Adaptation
  Prior on regularity α [Kolyan Ray] $\rightarrow$ one obtains e.g. adaptive confidence regions (modulo expected appropriate restrictions)

- How about, still in the white noise model, getting confidence sets for a *different norm*, for instance the L^∞-norm instead of L^2 ?

  This is possible, provided one slightly changes the definition of the large space H

  Replace Hilbert space H ($= L^2$ structure)

  $\rightarrow$ with some Besov-type space $\mathcal{M}$ (with norm related to L^∞-structure)

The density model

$$X_1, \ldots, X_n \text{ i.i.d. } \sim P \quad \text{with } dP = f d\mu, \; f \text{ unknown density on } [0,1].$$

The density model

$$X_1, \dots, X_n \text{ i.i.d. } \sim P \quad \text{with } dP = f d\mu, \ f \text{ unknown density on } [0,1].$$

Two examples of prior distributions Π on β-smooth densities

1. Random histograms

$$f(x) = 2^L \sum_{k=0}^{2^L-1} h_k \mathbb{1}_{(k2^{-L},(k+1)2^{-L})}(x), \quad 2^L = 2^{L_n} = n^{\frac{1}{1+2\beta}}$$

$$h = (h_1, \dots, h_L) \sim \text{Dirichlet}(1, \dots, 1)$$

2. Log-density priors

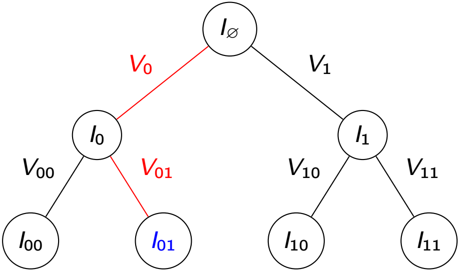
$$f(x) = \frac{e^{Z(x)}}{\int_0^1 e^{Z(u)} du}$$

$$Z(x) = \sum_{l=0}^{L} \sum_k \sigma_l \alpha_{lk} \psi_{lk}(x), \quad 2^L = 2^{L_n} = n^{\frac{1}{1+2\beta}}$$

with (e.g.) $\sigma_l = 2^{-l\beta}$ and $\alpha_{lk} \sim N(0,1)$ i.i.d.

The density model

3. Pólya trees



- $P(I_{\varepsilon_1\ldots\varepsilon_k}) = V_{\varepsilon_1} V_{\varepsilon_1\varepsilon_2} \cdots V_{\varepsilon_1\varepsilon_2\cdots\varepsilon_k}$ and $V_{\varepsilon 1} = 1 - V_{\varepsilon 0}$
- $V_{\varepsilon 0} \sim \text{Beta}(\alpha_{\varepsilon 0}, \alpha_{\varepsilon 1})$
- Take

$$\alpha_\varepsilon = a_l, \quad \text{for all } \varepsilon \text{ with } |\varepsilon| = l,$$

- with

$$\alpha_l = 2^{2l/\beta}$$

NP BvM: Natural limiting distribution ($\Box$)

$$X_1, \ldots, X_n \text{ i.i.d. } \sim P \quad \text{with } dP = fd\mu$$

Suppose one wants to estimate $P\psi_{lk} = \langle f, \psi_{lk} \rangle$.

Natural estimator is $P_n\psi_{lk} = \frac{1}{n}\sum_{i=1}^{n} \psi_{lk}(X_i)$

NP BvM: Natural limiting distribution ($\square$)

$$X_1, \ldots, X_n \text{ i.i.d. } \sim P \quad \text{with } dP = fd\mu$$

Suppose one wants to estimate $P\psi_{lk} = \langle f, \psi_{lk} \rangle$.

Natural estimator is $P_n\psi_{lk} = \frac{1}{n}\sum_{i=1}^{n} \psi_{lk}(X_i)$

Under true $P_0 = P_{f_0}$,

$$\sqrt{n}(P_n - P_0)\psi_{lk} \xrightarrow{d} \mathbb{G}_{P_0}(\psi_{lk}),$$

where, for any l, k,

$$\mathbb{G}_{P_0}(\psi_{lk}) \sim N(0, \|\psi_{lk} - E_{P_0}\psi_{lk}(X)\|_{2,P_0}^2)$$

$\rightarrow$ natural limiting distribution ($\square$) $=$ $\mathbb{G}_{P_0}$, P_0-white 'bridge'

NP BvM, the convergence $\overset{??}{\to}$: a large enough space $\mathcal{M}_0$

Let w_l sequence such that $w_l/\sqrt{l} \uparrow \infty$. Call this admissible sequence.

$$\mathcal{M} := \mathcal{M}(w) = \left\{ f = \{\langle f, \psi_{lk}\rangle\}, \quad \sup_l \max_k \frac{|\langle f, \psi_{lk}\rangle|}{w_l} < \infty \right\}$$

$$\mathcal{M}_0 := \mathcal{M}_0(w) = \left\{ f = \{\langle f, \psi_{lk}\rangle\}, \quad \lim_{l\to\infty} \max_k \frac{|\langle f, \psi_{lk}\rangle|}{w_l} = 0 \right\}$$

$\mathcal{M}_0$ is a closed separable subspace of $\mathcal{M}$.

Lemma The process $\mathbb{G}_{P_0}$ a.s. belong to $\mathcal{M}_0 = \mathcal{M}_0(w)$ for any admissible w.

Remarks • $L^2 \subset \mathcal{M}_0$ • $\mathbb{G}_{P_0}$ has (Besov)-regularity essentially $-1/2$

Nonparametric BvM in space $\mathcal{M}_0$ [C. and Nickl 14]

Let $\tau_{T_n} : f \to \sqrt{n}(f - T_n)$ for some centering sequence T_n.

Definition The prior Π with posterior $\Pi_n = \Pi(\cdot | X)$ satisfies the weak BvM in $\mathcal{M}_0$ with centering T_n if

$$\Pi_n \circ \tau_{T_n}^{-1} \xrightarrow{d} \mathbb{G}_{P_0}$$

Nonparametric BvM in space $\mathcal{M}_0$ [C. and Nickl 14]

Let $\tau_{T_n} : f \to \sqrt{n}(f - T_n)$ for some centering sequence T_n.

Definition The prior Π with posterior $\Pi_n = \Pi(\cdot \mid X)$ satisfies the weak BvM in $\mathcal{M}_0$ with centering T_n if

$$\beta_{\mathcal{M}_0}(\Pi_n \circ \tau_{T_n}^{-1}, \mathbb{G}_{P_0}) \to^{P_0} 0 \quad (n \to \infty).$$

Nonparametric BvM in space $\mathcal{M}_0$ [C. and Nickl 14]

Let $\tau_{T_n} : f \to \sqrt{n}(f - T_n)$ for some centering sequence T_n.

Definition The prior Π with posterior $\Pi_n = \Pi(\cdot \mid X)$ satisfies the weak BvM in $\mathcal{M}_0$ with centering T_n if

$$\beta_{\mathcal{M}_0}(\Pi_n \circ \tau_{T_n}^{-1}, \mathbb{G}_{P_0}) \to^{P_0} 0 \quad (n \to \infty).$$

Theorem [Weak BvM for first two examples of priors] Let $f_0 \in \mathcal{C}^\beta$ for some $\beta > 1/2$. Let Π be either the histogram prior ($1/2 < \beta \leq 1$) or the log-density prior ($\beta > 1$). Then for admissible w

$$\beta_{\mathcal{M}_0(w)}(\Pi(\cdot \mid X) \circ \tau_{T_n}^{-1}, \mathbb{G}_{P_0}) \to 0,$$

with centering $T_n = P_n(L_n)$ defined by

$$\langle P_n(L_n), \psi_{lk} \rangle = \begin{cases} \langle P_n, \psi_{lk} \rangle & \text{if } 2^l \leq 2^L_n = n^{\frac{1}{1+2\beta}} \\ 0 & \text{if } 2^l > 2^L_n, \end{cases}$$

Nonparametric BvM for Pólya trees [C 16]

Theorem [nonparametric BvM for the Pólya trees]
Let $f_0 \in \mathcal{C}^\beta$ for $\beta \in (0,1]$ and suppose $\|\log f_0\|_\infty < \infty$. Let Π Pólya tree prior with

$$\alpha_\varepsilon = a_l = 2^{2l/\beta}, \quad \text{any } |\varepsilon| = l,\ l \geq 0.$$

Then for admissible w

$$\beta_{\mathcal{M}_{\mathbf{0}}(w)}(\Pi(\cdot \mid X) \circ \tau_{T_n}^{-1}, \mathbb{G}_{P_{\mathbf{0}}}) \to 0,$$

with centering $T_n = P_n(L_n)$ defined by, for P_n the empirical measure,

$$\langle P_n(L_n), \psi_{lk} \rangle = \begin{cases} \langle P_n, \psi_{lk} \rangle & \text{if } 2^l \leq 2^L_n = n^{\frac{1}{1+2\beta}} \\ 0 & \text{if } 2^l > 2^L_n, \end{cases}$$

More generally can also consider mismatched regularity $\alpha \neq \beta$

Applications

- By the continuous mapping theorem, from BvM in $\mathcal{M}_0$ one can deduce limiting shape results for *continuous functionals*

$$\begin{array}{cccc} \psi : & f & \to & \psi(f) \\ & \mathcal{M}_0 & \to & \mathcal{T} \end{array}$$

Applications

- By the continuous mapping theorem, from BvM in $\mathcal{M}_0$ one can deduce limiting shape results for *continuous functionals*

$$\begin{array}{cccc} \psi : & f & \to & \psi(f) \\ & \mathcal{M}_0 & \to & \mathcal{T} \end{array}$$

- Nonparametric confidence bands [of fixed regularity]

  The nonparametric BvM in $\mathcal{M}_0$ leads to

  - confident credible balls in $\mathcal{M}_0$
    [of size $1/\sqrt{n}$]
  - Further inserting it with a ‘regularity constraint’,
    [Rate of the order 1 in $\mathcal{C}^\alpha$]
  - it leads to L^∞ confidence bands of minimax diameter

Application 3. Donsker's theorem for posterior distributions

Define, for a density f and a centering sequence T_n,

$$F(t) := \int_0^t f(u)du, \quad \mathbb{T}_n := \int_0^t T_n(u)du.$$

Application 3. Donsker's theorem for posterior distributions

Define, for a density f and a centering sequence T_n,

$$F(t) := \int_0^t f(u)du, \quad \mathbb{T}_n := \int_0^t T_n(u)du.$$

Theorem [Donsker's theorem for posterior on F]

Suppose Π satisfies weak BvM in $\mathcal{M}_0(w)$ with centering $T_n \in L^2$ and $\sum_l w_l 2^{-l/2} < \infty$. Let $\mu_{\mathbb{T}_n} : f \to \sqrt{n}(F - \mathbb{T}_n)$. Then

$$\beta_{C^0[0,1]}(\,\Pi(\cdot\,|\,X) \circ \mu_{\mathbb{T}_n}^{-1}\ ,\ \mathcal{G}\,) \to^{P_0} 0,$$

with $\mathcal{G}$ the law of a P_0-Brownian bridge G, as well as

$$\beta_{\mathbb{R}}(\,\mathcal{L}(\sqrt{n}\|F - \mathbb{T}_n\|_\infty\,|\,X)\ ,\ \mathcal{L}(\|G\|_\infty)\,) \to^{P_0} 0.$$

Corollary. The previous examples of priors satisfy a BvM for F with centering $\mathbb{T}_n$. An extra argument shows that the 'standard' centering F_n can be used as well.

Multiscale BvM and beyond

It can be useful to decompose the study of posterior on f

$$f \quad \longleftrightarrow \quad \langle f, \psi_{lk} \rangle$$

into a collection of semiparametric functionals

- NP BvM in $\mathcal{M}_0$: A key step in the proof is to prove

$$E[\|f - T_n\|_{\mathcal{M}_0} \,|\, X] = O_{P_0}(1/\sqrt{n})$$

  Based on decomposition of the $\mathcal{M}_0$-norm along $\langle f, \psi_{lk} \rangle$
  Also need asymptotic gaussianity to check cv. of finite-dimensional distributions

- This idea can also be used for stronger distances e.g. $\| \cdot \|_\infty$ [C. 14]

Rates: multiscale approach

Given a function f and a wavelet basis $\{\psi_{lk}\}$, consider the mapping

$$f \to \langle f, \psi_{lk} \rangle_2$$

- can be viewed as a semiparametric functional
- the collection over all l, k enables to reconstruct f

Rates: multiscale approach

Given a function f and a wavelet basis $\{\psi_{lk}\}$, consider the mapping

$$f \to \langle f, \psi_{lk} \rangle_2$$

- can be viewed as a semiparametric functional
- the collection over all l, k enables to reconstruct f

For *localised* wavelet bases $\{\psi_{lk}\}$ (think of the Haar basis) $\sum_{k=0}^{2^l-1} \|\psi_{lk}\|_\infty \leq C 2^{l/2}$.

$$\|f - f_0\|_\infty \lesssim \sum_l 2^{l/2} \max_{0 \leq k \leq 2^l - 1} |\langle f - f_0, \psi_{lk} \rangle_2|$$

Various regimes may appear for the functionals $f \to \langle f, \psi_{lk} \rangle_2$

- 'Small l' $\to$ BvM-type regime
- 'Large l' The prior mostly takes over

Have to study the functionals *simultaneously*

Rates via multiscale, examples

Consider the density estimation model. Set $\varepsilon^*_{n,\alpha} = (n/\log n)^{-\alpha/(2\alpha+1)}$, $\alpha > 0$

Example Let $\sigma_l = 2^{-l(\alpha+\frac{1}{2})}$, $\alpha_{lk} \sim \text{Laplace}(1)$ i.i.d., $L_n = n^{1/(2\alpha+1)}$,

$$\Pi_1 : \qquad f = \frac{e^T}{\int_0^1 e^T}, \qquad \text{with } T(\cdot) = \sum_{l=0}^{L_n} \sum_{k=0}^{2^l-1} \sigma_l \alpha_{lk} \psi_{lk}(\cdot),$$

Theorem [C 14] Let f_0 be Hölder $\alpha > 1$ and bounded away from 0 and ∞ on $[0,1]$. For any $M_n \to \infty$,

$$E_{f_0} \Pi_1[f : \ \|f - f_0\|_\infty \le M_n \varepsilon^*_{n,\alpha} \,|\, X^{(n)}] \to 1$$

A similar result holds for the Pólya tree prior with $a_l = l 2^{2l/\beta}$ and $\alpha \le 1$ [C 16]

Conclusion

Conclusion

- Bayesian approach is useful to suggest estimators
- Allows to naturally integrate hyperparameters via hierarchies
  cf. adaptation
- We have presented tools which enable to guarantee convergence properties under E_{θ_0}

Future work

- Construction of priors and convergence in high-dimensional problems, semiparametrics, inverse problems etc.
- Machine learning: numerous interesting priors considered, few theoretical studies of convergence
- Much to do in terms of
  - limiting shape results
  - confident credible sets

Thank you!