BML lecture #5
Bayesian nonparametrics: asymptotics

http://github.com/rbardenet/bml-course

Julyan Arbel

Statify team, Inria Grenoble Rhône-Alpes & Univ. Grenoble-Alpes, France

Outline

Outline

What comes to *your* mind when you hear "Asymptotics"?

Why Asymptotics

- Construction of a prior on a nonparametric space is difficult
- We cannot hope to cover all the space of density (for example) with our prior (the prior does not have full support)
- We need to check that our inference is not completely off!

Parametric setting

We have the celebrated Bernstein-von-Mise theorem that implies that the effect of the prior on the posterior inference vanishes when the amount of information grows.

This is not true anymore in a non parametric setting

Why Asymptotics

A first order approximation is to consider the asymptotic setting.

- Adopt a Frequentist point of view: "There exists a *true*" parameter θ_0, and we study the posterior distribution with data generated w.r.t. θ_0.
- Ideally, the posterior distribution will concentrate around θ_0 when $n \to \infty$.

References

- J. K. Ghosh and R. V. Ramamoorthi. *Bayesian Nonparametrics*. New York: Springer, 2003
- Nils Lid Hjort et al. *Bayesian nonparametrics*. Vol. 28. Cambridge University Press, Apr. 2010. URL: http://www.cambridge.org/us/academic/subjects/statistics-probability/statistical-theory-and-methods/bayesian-nonparametrics
- Subhashis Ghosal and Aad Van der Vaart. *Fundamentals of nonparametric Bayesian inference*. Vol. 44. Cambridge University Press, 2017

Outline

Consistency

- $\forall n \in \mathbb{N}$, $\mathbf{X}^n$ be some observation in a sample space $\{\mathcal{X}^n, \mathcal{A}^n\}$ with distribution P_θ
- $\theta \in \Theta$ with (Θ, d) a (semi-)metric space

Let Π be a prior distribution on Θ and $\Pi(\cdot|\mathbf{X}^n)$ a version of its posterior distribution.

Definition (Consistency)

The posterior distribution $\Pi(\cdot|\mathbf{X}^n)$ is said to be (weakly) consistent at θ_0 if for all $\varepsilon > 0$

$$\Pi\left(d(\theta, \theta_0) > \varepsilon | \mathbf{X}^n\right) \xrightarrow[n\to\infty]{P_{\theta_0}} 0.$$

If the convergence is a.s. then the posterior is said to be strongly consistent

Consistency

Other view of consistency

- Consistency can be summarized by saying that the full posterior distribution converge weakly to a Dirac mass at θ_0 in P_0 probability or P_0 almost surely.
- Posterior consistency can also be characterized through the posterior distribution of $\psi(\theta)$ for some tests function $\psi \in \Psi$.

Naturally one will hope that posterior consistency implies that some summary of the posterior location would be a consistent estimator.

Theorem

Let $\Pi(\cdot|\mathbf{X}^n)$ be a posterior distribution on Θ and suppose that it is consistent at θ_0 relative to a metric d on Θ. For $\alpha \in (0,1)$, define $\hat{\theta}_n$ as the centre of the smallest ball containing at least α of the posterior mass then

$$d(\hat{\theta}_n, \theta_0) \xrightarrow[n\to\infty]{P_0, \text{ or } P_0 \text{ a.s.}} 0$$

Extra notes I

Take $\alpha = 1/2$ for simplicity and consistency in probability. Define $B(\theta, r)$ the closed ball of radius r centred around θ, and let

$$\hat{r}(\theta) = \inf\{r, \Pi(B(\theta, r)|\mathbf{X}^n) \geqslant 1/2\}$$

(and inf over the empty set is ∞). Now let $\hat{\theta}_n$ be such that

$$\hat{r}(\hat{\theta}_n) \leqslant \inf_{\theta \in \Theta} r(\theta) + 1/n$$

Consistency implies that $\Pi(B(\theta_0, \varepsilon)|\mathbf{X}^n) \to 1$ so $\hat{r}(\theta_0) \leqslant \varepsilon$ with probability tending to 1. Furthermore, $\hat{r}(\hat{\theta}_n) \leqslant \hat{r}(\theta_0) + 1/n$ thus $\hat{r}(\hat{\theta}_n) \leqslant \varepsilon + 1/n$ with probability tending to 1.

In addition, $B(\theta_0, \varepsilon) \cap B(\hat{\theta}_n, \hat{r}(\hat{\theta}_n)) \neq \varnothing$ otherwise $\Pi(B(\theta_0, \varepsilon) \cup B(\hat{\theta}_n, \hat{r}(\hat{\theta}_n))|\mathbf{X}^n) = \Pi(B(\theta_0, \varepsilon)|\mathbf{X}^n) + \Pi(B(\hat{\theta}_n, \hat{r}(\hat{\theta}_n))|\mathbf{X}^n) \to 1 + 1/2$. So we have

$$d(\theta_0, \hat{\theta}_n) \leqslant \hat{r}(\theta)_n + \varepsilon \leqslant 2\varepsilon + 1/n$$

with probability that goes to 1.

Point estimator

- If Θ is a vector space, then one might want to use the posterior mean.
- The problem is that weak convergence to a Dirac does not implies convergence of moments.
- Consistency of the posterior mean holds under additional assumptions such as boundedness of posterior moments in probability or a.s. for some $p > 1$ would be sufficient.

Theorem (Posterior mean)

Assume that the balls of the metric space (Θ, d) are convex. Suppose that for any sequence $\theta_{1,n}, \theta_{2,n}$ in Θ and $\lambda_n \to 0$

$$d(\theta_{1,n}, (1-\lambda_n)\theta_{1,n} + \lambda_n \theta_{2,n}) \to 0$$

Then consistency of the posterior distribution implies consistency of the posterior mean.

Extra notes I

Let $\varepsilon > 0$ and write $\hat{\theta}_n = \int \theta \Pi(d\theta|\mathbf{X}^n)$. We decompose

$$\hat{\theta}_n = \int_{B(\theta_0,\varepsilon)} \theta \Pi(d\theta|\mathbf{X}^n) + \int_{B(\theta_0,\varepsilon)^c} \theta \Pi(d\theta|\mathbf{X}^n) = \theta_{1,n}(1-\lambda_n) + \lambda_n \theta_{2,n}$$

where $\theta_{1,n} = \int_{B(\theta_0,\varepsilon)} \theta \frac{\Pi(d\theta|\mathbf{X}^n))}{\Pi(B(\theta_0,\varepsilon)|\mathbf{X}^n)}$, $\lambda_n = \Pi(B(\theta_0,\varepsilon)|\mathbf{X}^n)$ and similarly for $\theta_{2,n}$ on the complement of $B(\theta_0,\varepsilon)$. Using Jensen inequality we have

$$d(\theta_{n,1},\theta_0) \leqslant \int_{B(\theta_0,\varepsilon)} d(\theta,\theta_0) \frac{\Pi(d\theta|\mathbf{X}^n))}{\Pi(B(\theta_0,\varepsilon)|\mathbf{X}^n)} \leqslant \varepsilon$$

In addition we have

$$d(\hat{\theta}_n,\theta_0) \leqslant d(\theta_{n,1},\theta_0) + d(\theta_{n,1},\theta_{1,n}(1-\lambda_n) + \lambda_n\theta_{2,n})$$

Using the fact that $\lambda_n \to 0$ since the the posterior is consistent, we have the desired result.

Remark

For the condition on d to hold, one can take convex and uniformly bounded metric for instance.

A first consistent posterior

Example

Assume the following model

$$X_1, \ldots, X_n \overset{iid}{\sim} P,$$
$$P \sim DP(M\alpha)$$

Consider the semi-metric $d_A(P, Q) = |P(A) - Q(A)|$ for some measurable event A on Θ, then $\Pi(\cdot|\mathbf{X}^n)$ is strongly consistent at any P_0 for d_A.

From this result we can easily obtain consistency under the weak topology. We could also obtain stronger consistency using Glivenko–Cantelli theorem.

Extra notes I

Consider $\Pi(|P(A) - P_0(A)| \geqslant \varepsilon|\mathbf{X}^n)$ and we will use the Markov inequality. We have given the results on DP that
$P|\mathbf{X}^n \sim DP(M\alpha + n\mathbb{P}_n)$, thus
$P(A)|\mathbf{X}^n \sim \beta(M\alpha(A) + n\mathbb{P}_n(A), M\alpha(A^c) + n\mathbb{P}_n(A^c))$. We thus have

$$\mathrm{E}(P(A)|\mathbf{X}^n) = \frac{M}{M+n}\alpha(A) + \frac{n}{M+n}\mathbb{P}_n(A) := \bar{P}(A)$$
$$\mathrm{var}(P(A)|\mathbf{X}^n) = \frac{\mathbb{P}_n(A)\mathbb{P}_n(A^c)}{1+n+M} \leqslant \frac{1}{4(1+n+M)}$$

Markov inequality gives

$$\Pi(|P(A)-P_0(A)| \geqslant \varepsilon|\mathbf{X}^n) \leqslant \frac{1}{\varepsilon^2}\left(|\bar{P}(A) - P_0(A)|^2 + \mathrm{var}(P(A)|\mathbf{X}^n)\right) = o(1)\,[P_0, a$$

using the law of large number on $\mathbb{P}(A)$.

Bayesian modelling perspective

From a Bayesian point of view, a Dirac measure at θ_0 correspond to perfect knowledge of the parameter.

- Prior and posterior distribution models our knowledge about the parameter.
- Consistency thus implies that when the amount of information grows we tends toward perfect knowledge of the parameter.

A validation of Bayesian methods

The frequentist setting where there exists a *true* parameter θ_0 that generates the data can be seen as an idealized set-up.

- An experimenter feeds a Bayesian with some data using the same data-generating mechanism.
- When the number of observation grows, the Bayesian should be able to pin-point the data-generating mechanism, whatever their prior.
- A prior that does not lead to a consistent posterior should not be used.

Robustness

Two Bayesians walk into a bar... with *almost* the same prior, then their posterior inference should not differ that much.

- Let Π_1 be the prior of Bayesian number 1
- Bayesian number 2 has the polluted prior $\Pi_2 = (1-\varepsilon)\Pi_1 + \varepsilon\delta_{p_0}$ for some $p_0 \in \Theta$

The posterior of Bayesian number 2 is consistent at p_0 (to be seen later), now what if Π_1 is not consistent at p_0 ? Let d_W be the metric for the weak topology, then $d_W(\Pi_1(\cdot|\mathbf{X}^n), \Pi_2(\cdot|\mathbf{X}^n))$ would not go to 0.

Extra notes I

There exists some $\varepsilon_0 > 0$ such that

$$\Pi_{n,1}(B(\theta_0, \varepsilon_0)|\mathbf{X}^n) \not\to 0$$

Thus

$$|\Pi_{n,1}(B(\theta_0, \varepsilon_0)|\mathbf{X}^n) - \Pi_{n,2}(B(\theta_0, \varepsilon_0)|\mathbf{X}^n)| \not\to 0$$

since $\Pi_{n,2}(B(\theta_0, \varepsilon_0)|\mathbf{X}^n) \to 0$.

Outline

Doob's Theorem

Can one get general conditions on the prior to insure that it is consistent?

- A first answer: Doob's Theorem
- The posterior is consistent at every θ Π-a.s.

Consider the case of *i.i.d.* observation

Theorem

Let $\{\mathcal{X}^n, P_\theta, \Theta\}$ be a statistical model where $\{\mathcal{X}^n, \mathcal{A}^n\}$ is a Polish space with Borel σ-field and Θ a Borel subset of a Polish space. Suppose that the map $\theta \mapsto P_\theta(A)$ is Borel measurable for evey $A \in \mathcal{A}$ and $\theta \mapsto P_\theta$ is one-to-one. For any prior distribution Π on Θ, if $X_1, \ldots, X_n \overset{iid}{\sim} P_\theta$, $\theta \sim \Pi$, the posterior is strongly consistent at any θ Π-a.s.

Some remarks on Doob's Theorem

- The conditions of the theorem are extremely weak
- And no conditions on the prior
- However this is only true Π-almost surely.
- Note: the Π-null set can be quite big! we can be happy with this result only if we are confident that the parameters are on the support of the prior. In general no one can be sure that the parameter generating the data inside the support of the prior, this is a real problem in fact in general the support of the prior can be quite thin. An extreme example is the case were the prior is a Dirac on some parameter θ_0. Then Doob's theorem still holds.

Outline

Setting

Doob's approach is not enough to show consistency of the posterior. For simplicity we focus on the density estimation setting

- Θ is the set of probability density functions on $\mathcal{X}$ w.r.t. a common dominating measure ν. We denote the parameter p (instead of θ) and P the associated probability measure.
- The observation $X_1, \ldots, X_n \overset{iid}{\sim} p$ and $p \sim \Pi$.

Considering density estimation makes things easier to write without being too simplistic. The same results can be extended to nonparametric regression.

KL property

To achieve consistency, we do not want to require that the true parameter p_0 is inside the support of Π. However we still some prior mass *near* p_0.

Definition (Kullback Leibler)

Let p and p_0 be two p.d.f. with respect to a common measure such that $p_0 \ll p$ then the Kullback–Leibler divergence between p and p_0 is

$$KL(p, p_0) = \int p_0 \log(p_0/p) d\nu.$$

KL property

Definition (KL property)

We say that a prior distribution Π satisfies the Kullback–Leibler property at p_0 if for every $\varepsilon > 0$,

$$\Pi(\mathrm{KL}(p, p_0) \geqslant \varepsilon) > 0$$

We note $p_0 \in \mathrm{KL}(\Pi)$ and alternatively will say that p_0 is in the KL-support of Π.

This extends quite a lot the parameters at which the posterior can be consistent at.

Existence of tests

The other requirement would be that the parameter set is not too complex.

Definition (Exponentially consistent tests)

We say that a sequence of tests φ_n for $H_0 : p = p_0$ versus $H_1 : p \in U^c$ is exponentially consistent if

$$P_0^n(\varphi_n) \lesssim e^{-Cn}, \quad \sup_{p \in U^c} P^n(1 - \varphi_n) \lesssim e^{-Cn}$$

A test is understood as a measurable map $\mathcal{X}^n \to [0, 1]$ and the corresponding statistic $\varphi_n(X_1, \ldots, X_n)$. φ_n is interpreted as the probability that the null is rejected.

Extra notes I

The existence of tests means that we can differentiate between p_0 and parameter in U^c.

It is enough to have uniformly consistent sequence of test

$$P_0(\varphi_n) \to 0, \quad \sup_{p \in U^c} P(1 - \varphi_n) \to 0.$$

Since the test is uniformly consistent then there exists $k \in \mathbb{N}$ such that $P_0^k(\varphi_k) \leqslant 1/4$, $P^k(1 - \varphi_k) \leqslant 1/4$. Now for n large, write $n = mk + r$. Slice $\mathbf{X}^n = (X_1, \ldots, X_n)$ into m sub-sample of size k $\mathbf{X}_l^n = (X_{(l-1)k+1}, \ldots, X_{lk})$ and define $Y_{l,n} = \varphi_k(\mathbf{X}_l^n)$. Now create a new test $\psi_n = \mathbb{I}\{\bar{Y}_m > 1/2\}$. We have for every $p \in U^c$, $P(1 - Y_j) \leqslant 1/4$

$$P(\psi_n) = P(\bar{Y} \leqslant 1/2) = P(1 - \bar{Y} \geqslant 1/2) = P(1 - \bar{(}Y) \geqslant 1/2) \leqslant e^{-2m/16} \lesssim e^{-Cn}$$

Using Hoeffding inequality

Extra notes II

$$\mathbb{P}(\bar{X} - \mathsf{E}(X) \geqslant \varepsilon) \leqslant \exp\left\{-2\varepsilon^2 m\right\}$$

Schwartz Theorem

Theorem

Let Π be a prior distribution on Θ such that $p_0 \in KL(\Pi)$. Let U be a neighbourhood of p_0 such that there exists an exponentially consistent sequence of tests for p_0 against U^c, then

$$\Pi(U^c|\mathbf{X}^n) \to 0 \ P_0[a.s].$$

Nothing to do we Herman Schwarz (without t!) or even Laurent Schwartz the Fields Medalist! she is Lorraine Schwartz, former student of Lucien Le Cam.

Extra notes I

$$\Pi(U^c|\mathbf{X}^n) = \frac{\int_{U^c}\prod_{i=1}^n \frac{p}{p_0}(X_i)d\Pi(p)}{\int_{U^c}\prod_{i=1}^n \frac{p}{p_0}(X_i)d\Pi(p)} := \frac{N_n}{D_n}.$$

We first show $\liminf D_n e^{n\varepsilon}/\Pi(KL(p,p_0)>\varepsilon) \geqslant 1$, $P_0[a.s.]$. Let $\Pi_0(\cdot) = \Pi(\cdot \cap KL(p,p_0)>\varepsilon)/\Pi(KL(p,p_0)>\varepsilon)$. Then

$$\begin{aligned}
\log(D_n) &\geqslant \log\left(\int_{KL(p,p_0)>\varepsilon} \frac{p}{p_0}(X_i)d\Pi_0(p)\right) + \log(\Pi(KL(p,p_0)<\varepsilon)) \\
&\geqslant \int_{KL(p,p_0)>\varepsilon} \log\left(\prod_{i=1}^n \frac{p}{p_0}(X_i)\right) d\Pi_0(p) + \log(\Pi(KL(p,p_0)<\varepsilon)) \\
&= \sum_{i=1}^n \int \log\frac{p}{p_0}(X_i)d\Pi_0(p) + \log(\Pi(KL(p,p_0)<\varepsilon))
\end{aligned}$$

The law of large numbers implies

$$\frac{1}{n}\sum_{i=1}^n \int \log\frac{p}{p_0}(X_i)d\Pi_0(p) \to P_0\int \frac{p}{p_0}(X_i)d\Pi_0(p),\ P_0[a.s.]$$

Extra notes II

which is $-\int KL(p, p_0) d\Pi_0(p) > -\varepsilon$. Thus

$$\liminf D_n e^{n\varepsilon} / \Pi(KL(p, p_0) > \varepsilon) \geqslant 1, \ P_0[a.s.]$$

For n large enough we have the following $P_0[a.s.]$

$$\begin{aligned}\Pi(U^c|\mathbf{X}^n) &\leqslant \varphi_n + (1-\varphi_n)\frac{N_n}{D_n} \\ &\leqslant \varphi_n + (1-\varphi_n) N_n e^{\varepsilon n} \Pi(KL(p, p_0) > \varepsilon)\end{aligned}$$

Furthermore we have that

$$\begin{aligned}P_0^n N_n (1-\varphi_n) &= P_0^n \int_{U^c} (1-\varphi_n) \prod_{i=1}^n \frac{p}{p_0}(X_i) \Pi(dp) \\ &= \int_{U^c} P^n (1-\varphi_n) \Pi(dp) \leqslant e^{-Cn}\end{aligned}$$

We thus get $P_0\Pi(U^c|\mathbf{X}^n) \leqslant e^{-C'n}$ for $\varepsilon < C$ and for $C' = C - \varepsilon$. Using Borel–Cantelli we get that $\Pi(U^c|\mathbf{X}^n) \to 0 P_0[a.s.]$.

Schwartz Theorem

- Need to test away all densities in U^c
- Might not be possible for strong neighbourhood of p_0 (L_1 metrics)

Extension of Schwartz theorem

The idea is that not *all* functions in U^c matters and we can discard function with very low prior probabilities.

Theorem

The results of the previous theorem are still valid if we replace the assumption on the existence of tests by:

$$\Theta_n \subset \Theta$$

$$\Pi(\Theta_n) \leqslant e^{-Cn},\ P_0^n \varphi_n \leqslant e^{-Cn},\ \sup_{p \in U^c \cap \Theta_n} P(1-\varphi_n) \leqslant e^{-Cn}$$

Existence of tests

Schwartz's theorem require the existence of exponentially consistent tests

- We can differentiate between θ_0 and U^c
- The model is not too complex

Question

When do such tests exists ?

Let see the example of iid observations.

Sketch of the proof

- Cannot directly construct test against $U^c = \{p, d(p, p_0) > \varepsilon\}$...
- Construct an exponentially consistent test against a generic ball that is at least at distance ε
- Cover U^c with N of these balls, and construct a test from the N corresponding tests.

Consistency under Entropy bound

We combine the preceding results to get general conditions on the prior and on the model, that insure consistency.

Theorem

The posterior is strongly consistent relative to the L_1 distance at every p_0 in the KL-support of the prior if for every $\varepsilon > 0$ there exist Θ_n such that for $C > 0$ and $0 < c < 1/2$

$$\Pi(\Theta_n^c) \leqslant e^{-Cn}, \ \log N(\varepsilon, \Theta_n, ||\cdot||_1) \leqslant cn\varepsilon_n^2,$$

for n large enough.

Outline

Definition

Contraction rates is a refinement of establishing posterior consistency

- How fast posterior concentrates its mass around the true parameter
- Helps to see how much the prior influences the posterior

Definition

Let ε_n be a positive sequence. The posterior contracts at the rate ε_n at θ_0 if for any $M_n \to \infty$

$$\Pi(d(\theta, \theta_0) > M_n \varepsilon_n | \mathbf{X}^n) \xrightarrow[n \to \infty]{P_{\theta_0}} 0$$

If all the experiments share the same probability space and the convergence is $P_{\theta_0}[a.s]$ we say that the posterior contracts in the strong sense.

Remarks

- Any slower rate than ε_n also fits the definition so we will say *a* posterior contraction rate
- We will naturally try to find the fastest possible rate!

Regarding M_n

- The sequence M_n plays virtually no role in the posterior rate. In many cases it can be fixed to a constant M.
- For finite dimensional models M_n must be allowed to grow to obtain the usual $n^{-1/2}$ rate in smooth models.

Consequences of posterior contraction

Point Estimator

- Let $\hat{\theta}_n =$ centre of the smallest balls that contains at least $1/2$ of the posterior mass.
- Assume that the posterior contracts at θ_0 with rate ε_n for the metric d

Then $d(\hat{\theta}_n, \theta) = O_P(\varepsilon_n)$ in P_0 probability (or a.s. if strong contraction).

Consequences of posterior contraction

Posterior mean

If the metric d is bounded and $\theta \mapsto d^s(\theta, \theta_0)$ is convex for some $s \geqslant 1$ then the posterior mean $\tilde{\theta}_n$ satisfies

$$d(\tilde{\theta}_n, \theta_0) \leqslant M_n \varepsilon_n + \|d\|_\infty^{1/s} \Pi_n(d(\theta, \theta_0) \geqslant M_n \varepsilon_n | \mathbf{X}^n)^{1/s}.$$

- First term is the dominating term
- The second term is exponentially small in general

Some first Examples - Parametric models

- Let $X_1, \ldots, X_n | \theta \overset{iid}{\sim} \mathcal{B}(\theta)$, and $\theta \sim \beta(\alpha, \beta)$. The posterior contracts at a rate $n^{-1/2}$
- Let $X_1, \ldots, X_n | \theta \overset{iid}{\sim} \mathcal{U}([0, \theta])$ and $\pi(\theta) \propto \theta^{-a}$. The posterior contracts at a rate n^{-1}.

Parametric regular models

In fact for all regular finite dimensional models the Bernstein von-Mises theorem implies posterior rates $n^{-1/2}$.

Nonparametric example: Dirichlet Process

- $X_1, \ldots, X_n | P \overset{iid}{\sim} P$
- $P \sim DP(M\alpha)$ for α a probability measure on $\mathcal{X}$.

The posterior distribution is $P|\mathbf{X}^n \sim DP(M\alpha + n\mathbb{P}_n)$.

Local semi-metric

For A a measurable set, let $d(P, Q) = |P(A) - Q(A)|$. The posterior distribution is consistent at P_0 at a rate $n^{-1/2}$.

Global metric

For ν a σ-finite measure and F and G two c.d.f. let $d(F, G) = \|F - G\|_\nu^2 = \int (F(t) - G(t))^2 d\nu(t)$. The posterior contracts at rate $n^{-1/2}$ at P_0 for this metric.

Nonparametric example: White Noise

Consider the following model for W_t a white noise

$$X_t = f(t) + n^{-1/2} W_t.$$

Projecting this model onto the Fourier basis if $f \in L_2$, we have the equivalent formulation

$$X_{i,n} = \theta_i + n^{-1/2} \varepsilon_i, \ \ i \in \mathbb{N}^*$$

$\theta \in \ell_2(\mathbb{R})$. Assume the following prior

$$\theta_i \overset{ind.}{\sim} \mathcal{N}(0, i^{-2\alpha-1}).$$

If $\theta_0 \in \mathcal{S}^{2,2}_\beta$ then the posterior contracts at θ_0 at the rate $n^{-\min(\alpha,\beta)/(2\alpha+1)}$.

Outline

General theorem

- Result similar to Schwartz theorem ?
- We focus on the case of i.i.d observations $X_1, \ldots, X_n \overset{iid}{\sim} P$
- The parameter set Θ is the set of probability densities with respect to a common dominating measure μ.

Let Π_n be a sequence of prior, we study the sequence of posterior distributions $\Pi_n(\cdot|\mathbf{X}^n)$ under the assumption that the data are generated with $p = p_0$.

General Theorem

We follow the same steps than for Schwartz Theorem

- Existence of tests to separate p_0 from complement of balls
- KL conditions, the prior puts enough mass on neighbourhood of p_0

Define $V_{k,0}$ the kth KL variation

$$V_{2,0} = P_0 \left(\log \left(\frac{p_0}{p}(X) \right) \right)^2$$

we define the KL neighbourhood as

$$B_0(p_0, \varepsilon) = \{p, KL(p_0, p) \leqslant \varepsilon^2\}$$
$$B_k(p_0, \varepsilon) = \{p, KL(p_0, p) \leqslant \varepsilon^2, V_{k,0}(p_0, p) \leqslant \varepsilon^k\}$$

General theorem

Theorem

Let $d \leqslant h$ be a metric on Θ for which balls are convex, and let $\Theta_n \subset \Theta$. The posterior contracts at a rate ε_n for all ε_n such that $n\varepsilon_n^2 \to \infty$ and such that for positive constants c_1, c_2 and any $\underline{\varepsilon}_n \leqslant \varepsilon_n$

$$\log N(\varepsilon_n, \Theta_n, d) \leqslant c_1 n\varepsilon_n^2,$$

$$\Pi_n(B_{2,0}(p_0, \underline{\varepsilon}_n^2)) \geqslant e^{-c_2 n\underline{\varepsilon}_n^2}$$

$$\Pi(\Theta_n^c) \leqslant e^{-(c_2+3)n\underline{\varepsilon}_n^2}$$

General Theorem

- The KL condition can be refined, but the idea is basically the same
- Entropy condition is useful for the existence of tests
- Entropy condition can be replaced by a local entropy, which is more like a *dimension of* Θ_n

Interpretation

Assume that d and KL are equivalent

- We need $e^{n\varepsilon_n^2}$ balls to cover Θ_n.
- If the prior spread evenly the mass on these balls, we have $e^{-Cn\varepsilon_n^2}$ mass on each of these balls thus KL condition is satisfied
- If the spread is uneven, then KL condition might not be satisfied for some p_0.

Outline

General observations

- The previous theorem can be generalized to more general observation (like regression for instance)
- But we have to be careful with the metric we use, and the existence of test is not guaranteed!
- To be general we will have to assume that we can test away parameters

Existence of tests

Let d_n and e_n be two semi-metrics on Θ. For $\varepsilon > 0$, and for all $\theta_1 \in \Theta$ such that $d_n(\theta_0, \theta_1) > \varepsilon$ there exists φ_n

$$P^n_{\Theta_0} \varphi_n \leqslant e^{-Kn\varepsilon^2}, \quad \sup_{\theta, e_n(\theta,\theta_1) \leqslant \xi\varepsilon} P^n_\theta (1 - \varphi_n) \leqslant e^{-Kn\varepsilon^2}$$

General theorem

Define the following KL-neighbourhood

$$V_{k,0}(f,g) = \int f |\log(f/g) - KL(f,g)|^k d\mu$$

$$B_n(\theta_0, \varepsilon, k) = \left\{ \theta \in \Theta KL(p^n_{\theta_0}, p^n_\theta) \leqslant n\varepsilon^2, V_{k,0}(p^n_{\theta_0}, p^n_\theta) \leqslant n^{k/2}\varepsilon^k \right\}$$

General theorem

Theorem

Let d_n and e_n be two semi-metrics on Θ, such that tests exists, $\varepsilon_n \to 0$, $n\varepsilon_n^2 \to \infty$, $k > 1$, $\Theta_n \subset \Theta$ such that for sufficiently large $j \in \mathbb{N}$

$$\sup_{\varepsilon \geqslant \varepsilon_n} \log N\left(\frac{1}{2}\xi\varepsilon, \{\theta \in \Theta_n d_n(\theta_0, \theta) \leqslant \varepsilon\}, e_n\right) \leqslant n\varepsilon_n^2$$

$$\frac{\Pi_n(\theta \in \Theta_n, j\varepsilon_n \leqslant d_n(\theta, \theta_0) \leqslant 2j\varepsilon_n)}{\Pi_n(B_n(\theta_0, \varepsilon_n, k))} \leqslant e^{Kn\varepsilon_n^2 j^2/2}$$

$$\frac{\Pi_n(\Theta_n^c)}{\Pi_n(B_n(\theta_0, \varepsilon_n, k))} \leqslant e^{-2n\varepsilon_n}$$

then $P_{\theta_0}^n \Pi_n(d_n(\theta_0, \theta) \geqslant M_n\varepsilon_n) = o(1)$

Independent observations

- Assume that the measure $P_\theta^n = \bigotimes_{i=1}^n P_{i,\theta}$ on some product space $\bigotimes_{i=1}^n \{\mathcal{X}_i, \mathcal{A}_i\}$.
- Assume that each measures $P_{i,\theta}$ are absolutely continuous w.r.t μ_i
- Define the Root average Hellinger distance

$$d_{n,H}(\theta, \theta') = \left(\frac{1}{n} \sum_{i=1}^n \int (\sqrt{dP_{i,\theta}} - \sqrt{dP_{i,\theta'}})^2 \right)^{1/2}$$

Lemma

For all here exists tests φ_n such that

$$P_{\theta_0}^n \varphi_n \leqslant e^{-n d_{n,H}(\theta_0, \theta_1)}, \quad P_\theta^n \leqslant e^{-n d_{n,H}(\theta_0, \theta_1)}$$

for all θ such that $d_{n,H}(\theta, \theta_1) \leqslant \frac{1}{18} d_{n,H}(\theta_0, \theta_1)$

Independent observations

We can also simplify the KL condition in this case. Note that

$$KL(p^n_{\theta_0}), p^n_\theta) = \sum_{i=1}^{n} KL(p_{i,\theta_0}, p_{i,\theta})$$

Furthermore for the KL-variation term we have that

$$V_{k,0}(p^n_{\theta_0}, p^n_\theta) \leqslant n^{k/2} C_k \frac{1}{n} \sum_{i=1}^{n} V_{k,0}(p_{i,\theta_0}, p_{i,\theta})$$

Thus the KL condition can be re-written

$$B^*_n(\theta_0, \varepsilon, k) = \Big\{\theta \in \Theta, \frac{1}{n}\sum_{i=1}^{n} KL(p_{i,\theta_0}, p_{i,\theta}) \leqslant \varepsilon^2, \\ \frac{1}{n}\sum_{i=1}^{n} V_{k,0}(p_{i,\theta_0}, p_{i,\theta}) \leqslant C_k\varepsilon^2\Big\}$$

NP Regression with splines

Consider the model

$$X_i = f(z_i) + \varepsilon_i, \ \ i = 1, \ldots, n$$

where $\varepsilon_i \overset{iid}{\sim} \mathcal{N}(0, \sigma^2)$ and the $z_i \in \mathbb{R}$ are known fixed covariates. For simplicity σ^2 is also assume to be known. Let $\mathbb{P}_n^z = \frac{1}{n}\sum_{i=1}^n \delta_{z_i}$ and $\|\cdot\|_n$ the $L_2(\mathbb{P}_n^z)$ norm

Lemma

We have the following results

$$KL(P_{f,i}, P_{g,i}) = \frac{1}{2\sigma^2}(f(z_i) - g(z_i))^2$$

$$V_{0,2}(P_{f,i}, P_{g,i}) = \frac{1}{\sigma^2}(f(z_i) - g(z_i))^2$$

NP Regression with splines

Assume that $f_0 \in \mathcal{H}(\alpha, L)$ such that $\|f_0\|_\infty \leqslant H$, then the $d^2_{n,H}$ and $\|\cdot\|^2_n$ are equivalent.

Spline prior

Consider $(B_j)_{j=1}^J$ the B-splines basis with J equally spaced nodes, and consider

$$f_\beta(\cdot) = \sum_{j=1}^{J} \beta_j B_j(\cdot)$$

and induce a prior on f by choosing a prior on β, $\beta_j \overset{iid}{\sim} g$.

Approximation techniques with splines gives us that for $\beta^* \in \mathbb{R}^J$ the coefficient of the projection of f_0 in $\mathrm{Span}(B_j)$,

$$\|f_{\beta^*} - f_0\|_\infty \leqslant J^{-\alpha}\|f_0\|_\alpha$$

NP Regression with splines

We also need to impose conditions on the design.Let Σ_n be such that $\Sigma_{n,i,j} = \int B_i B_j d\mathbb{P}_n^z$. We assume that

$$J^{-1}\|\beta\|^2 \asymp \beta'\Sigma_n\beta$$

so that

$$\|f_{\beta_1} - f_{\beta_2}\|_n \asymp \sqrt{J}\|\beta_1 - \beta_2\|$$

We can thus perform calculations in terms of the Euclidean norm of the coefficients.

Theorem

Assume that g is a standard Gaussian distribution, and assume that $J = J_n \asymp n^{1/(2\alpha+1)}$, then the posterior contracts at a rate $\varepsilon_n = n^{-\alpha/(2\alpha+1)}$.

- This is the minimax rate, in addition this rate is uniform over all bounded $\mathcal{H}(\alpha, L)$ functions.
- Some condition can be relaxed, in particular, g could be any distribution such that for every β^* such that $\|\beta^*\|_\infty \leqslant C$ $\Pi(\|\beta - \beta^*\| \leqslant \varepsilon) \geqslant e^{-cJ \log(1/\varepsilon)}$. Some log factor may appear in the rate.
- The boundedness condition could also be dropped by considering likelihood ratio tests for $\|\cdot\|_n$ norm.

Acknowledgements

I would like to thank Jean-Bernard Salomond for sharing his expertise and slides on asymptotic aspects of Bayesian nonparametric procedures.

References I

[1] Subhashis Ghosal and Aad Van der Vaart. *Fundamentals of nonparametric Bayesian inference*. Vol. 44. Cambridge University Press, 2017.

[2] J. K. Ghosh and R. V. Ramamoorthi. *Bayesian Nonparametrics*. New York: Springer, 2003.

[3] Nils Lid Hjort, Chris Holmes, Peter Müller, and Stephen G Walker. *Bayesian nonparametrics*. Vol. 28. Cambridge University Press, Apr. 2010. URL: http://www.cambridge.org/us/academic/subjects/statistics-probability/statistical-theory-and-methods/bayesian-nonparametrics.