Department of Decision Sciences, Bocconi University, November 2018

Frequentist limits from Bayesian statistics

Bas Kleijn

KdV Institute for Mathematics

UNIVERSITEIT VAN AMSTERDAM

Frequentist and Bayesian philosophies

Bayesians and frequentists have different perspectives on data $X \in \mathscr{X}$ and model $\mathscr{P}$.

Starting points

Frequentist assume a true, underlying distribution P_0 that has generated the data.

Bayesian formulate belief concerning the distribution that has generated the data.

Mathematical expression

Frequentist choose a map $\hat{P} : \mathscr{X} \to \mathscr{P}$, to estimate, with a sampling distribution to test and quantify uncertainty.

Bayesian choose a prior $\Pi(\,\cdot\,)$ and condition on X to obtain a posterior $\Pi(\,\cdot\,|X)$ on the model, to estimate, test and quantify uncertainty.

A distinguishing example (Savage, 1961)

Example 1 *Consider the following three statistical experiments:*

A lady who drinks milk in her tea claims to be able to tell which was poured first, the tea or the milk. In ten trials, she is correct every time

A music expert claims to be able to tell whether a page of music was written by Haydn or by Mozart. In ten trials, he correctly determines the composer every time.

A drunken friend says that he can predict heads or tails of a fair coin-flip. In ten trials, he is right every time.

Frequentist analysis

We analyse the Bayesian procedure from a frequentist perspective.

Assumption samples X^n are $P_{0,n}$-distributed

We shall concentrate on the large-sample behaviour of the posterior. Typical questions

- Consistency Does the posterior concentrate around the point P_0?
- Rate of convergence How fast does concentration occur?
- Limiting shape Which shape does a concentrating posterior have?
- Model selection Is the Bayes factor consistent?
- Uncertainty quantification Do credible sets have coverage?

in the limit $n \to \infty$.

Goal

The question

Given the model, which priors give rise
to posteriors with good
frequentist convergence properties?

The answer

To formulate theorems that assert
asymptotic properties of the posterior,
under conditions on model, prior and $(P_{0,n})$.

Course schedule

Lec I Bayesian Basics

Frequentist/Bayesian formalisms, estimation, coverage, testing

Lec II The Bernstein-von Mises theorem

Limit shape in smooth parametric models, semi-parametrics

Lec III Bayes and the Infinite

Consistency, Doob's theorem, Schwartz's theorem

Lec IV Posterior contraction

Barron, Walker, Ghosh-Ghosal-van der Vaart theorems

Lec V Errors-in-variables regression

Consistency and rates of posterior convergence

Lec VI Tests and posteriors

Testing and posterior concentration, Doob's theorem

Course schedule

Lec VII	Frequentist validity of Bayesian limits	
	Remote contiguity and frequentist limits	
Lec VIII	Posterior uncertainty quantification	
	How confidence sets arise from credible sets	
Lec IX	Exact recovery and detection of communities	
	Consistency in the planted bi-section model	
Lec X	Uncertainty quantification for communities	
	Confidence sets for community assignment	
Lec XI	Uniform and pointwise tests	
	Which hypotheses are asymptotically testable and which are not?	
Lec XII	Bayesian tests and posterior model selection	
	Model selection with posteriors, some examples	

References

T. Bayes, *An essay towards solving a problem in the doctrine of chances,* Phil. Trans. Roy. Soc. **53** (1763), 370-418.

J. Berger, *Statistical decision theory and Bayesian analysis,* Springer, New York (1985).

L. Le Cam, G. Yang, *Asymptotics in statistics,* Springer, New York (1990).

A. van der Vaart, *Asymptotic statistics,* Cambridge university press (1998).

J. Ghosh, R. Ramamoorthi, *Bayesian nonparametrics,* Springer, New York (2003).

S. Ghosal, A. van der Vaart, *Foundations of Bayesian statistics,* Cambridge Univ Press, Cambridge (2018).

B. Kleijn, *The frequentist theory of Bayesian statistics,* Springer, New York (201?).

Lecture I

Bayesian Basics

In the first lecture, the basic formalism of Bayesian statistics is introduced and its formulation as a frequentist method of inference is given. We discuss such notions as the prior and posterior, Bayesian point estimators like the posterior mean and MAP estimators, credible intervals, odds ratios and Bayes factors. All of these are compared to more common frequentist inferential tools, like the MLE, confidence sets and Neyman-Pearson tests.

Bayesian and Frequentist statistics

sample space	$(\mathscr{X}_n, \mathscr{B}_n)$	measurable space
i.i.d. data	$X^n \in \mathscr{X}^n$	frequentist/Bayesian
models	$(\mathscr{P}_n, \mathscr{G}_n)$	model subsets $B, V \in \mathscr{G}$
parametrization	$\Theta \to \mathscr{P}_n : \theta \mapsto P_{\theta,n}$	model distributions
priors	$\Pi_n : \mathscr{G}_n \to [0,1]$	probability measure
posterior	$\Pi(\,\cdot\,\vert X^n) : \mathscr{G}_n \to [0,1]$	Bayes's rule, inference

Frequentist	assume there is P_0	$X^n \sim P_{0,n}$
Bayes	assume $P \sim \Pi$	$X^n \,\vert\, P_n \sim P_n$

Bayes's Rule and Disintegration

Definition 2 *Fix $n \geq 1$. Assume that $P \mapsto P_n(A)$ is $\mathscr{G}_n$-measurable. Given prior Π_n, a posterior is any $\Pi(\cdot\,|X^n = \cdot) : \mathscr{G}_n \times \mathscr{X}_n \to [0,1]$ s.t.*

(i) For any $G \in \mathscr{G}_n$, $x^n \mapsto \Pi(G|X^n = x^n)$ is $\mathscr{B}^n$-measurable

(ii) (Disintegration) For all $A \in \mathscr{B}^n$ and $G \in \mathscr{G}_n$

$$\int_A \Pi(G|X^n)\, dP_n^\Pi = \int_G P_n(A)\, d\Pi_n(P_n) \tag{1}$$

where $P_n^\Pi = \int P_n\, d\Pi_n(P_n)$ is the prior predictive distribution

Remark 3 *For frequentists $X^n \sim P_{0,n}$, so assume*

$$P_{0,n} \ll P_n^\Pi$$

Posteriors in dominated models

Theorem 4 *Assume* $\mathscr{P}_n = \{P_{\theta,n} : \theta \in \Theta\}$ *is dominated by a* σ*-finite* μ_n *on* $(\mathscr{X}_n, \mathscr{B}_n)$ *with densities* $p_{\theta,n} = dP_{\theta,n}/d\mu_n$. *Then,*

$$\Pi(\,\theta \in G \,|\, X^n\,) = \int_G p_{\theta,n}(X^n)\, d\Pi_n(\theta) \;\Big/\; \int_\Theta p_{\theta,n}(X^n)\, d\Pi_n(\theta), \qquad (2)$$

for all $G \in \mathscr{G}$.

Example 5 *i.i.d. data Consider* $X^n = (X_1, \ldots, X_n) \in \mathscr{X}^n$, $X^n \sim P^n$ *Choose* $\mathscr{X}_n = \mathscr{X}^n$, $\Theta = \mathscr{P} \ll \mu$, $P \mapsto P_n = P^n$ *and* $\Pi_n = \Pi$ *on* $\mathscr{P}$.

$$\Pi(\,P \in G \,|\, X^n\,) = \int_G \prod_{i=1}^n p(X_i)\, d\Pi(P) \;\Big/\; \int_{\mathscr{P}} \prod_{i=1}^n p(X_i)\, d\Pi(P),$$

Proof

Fix n (and suppress it in notation)

Fubini Prior predictive has a density with respect to μ,

$$P^{\Pi}(B) = \int_{\Theta} \int_{B} p_\theta(x)\, d\mu(x)\, d\Pi(\theta) = \int_{B} \left(\int_{\Theta} p_\theta(x)\, d\Pi(\theta) \right) d\mu(x).$$

That density $p^{\Pi} : \mathscr{X} \to \mathbb{R}$ is the denominator of the posterior. Note,

$$\int_{B} \Pi(G|X=x)\, dP^{\Pi}(x) = \int_{B} \left(\int_{G} p_\theta(x)\, d\Pi(\theta) \Big/ \int_{\Theta} p_\theta(x)\, d\Pi(\theta) \right) dP^{\Pi}(x)$$
$$= \int_{B} \int_{G} p_\theta(x)\, d\Pi(\theta)\, d\mu(x) = \int_{G} P_\theta(B)\, d\Pi(\theta),$$

so disintegration is valid.

σ-additivity of the posterior

Proposition 6 *The posterior (2) is σ-additive, P^{Π}-a.s.*

Proof Since $P^{\Pi}(p^{\Pi} > 0) = 1$, the denominator is non-zero and the posterior is well-defined P^{Π}-a.s. For x such that $p^{\Pi}(x) > 0$ and disjoint (G_n)

$$\Pi\Big(\theta \in \bigcup_{n\geq 1} G_n \;\Big|\; X = x \Big) = C(x) \int_{\cup_n G_n} p_\theta(x)\, d\Pi(\theta)$$

$$= C(x) \int \sum_{n\geq 1} 1_{\{\theta \in G_n\}}\, p_\theta(x)\, d\Pi(\theta)$$

$$= \sum_{n\geq 1} C(x) \int_{G_n} p_\theta(x)\, d\Pi(\theta) = \sum_{n\geq 1} \Pi(\, \theta \in G_n \,|\, X = x\,),$$

by monotone convergence. □

Prior to posterior

The Bayesian procedure consists of the following steps

(i) Based on the background of the data X, choose a model $\mathscr{P}$, usually with parameterization $\Theta \to \mathscr{P} : \theta \mapsto P_\theta$.

(ii) Also choose a prior measure Π on $\mathscr{P}$ (reflecting "belief"). Usually a measure on Θ is defined, inducing a measure on $\mathscr{P}$.

(iii) Calculate the posterior as a function of the data X.

(iv) Observe a realization of the data $X = x$, substitute in the posterior and do statistical inference.

Posterior predictive distribution

Definition 7 *Consider data X from $(\mathscr{X}, \mathscr{B})$, a model $\mathscr{P}$ and prior Π. Assume that the posterior $\Pi(\cdot|X)$ is a prob msr. The posterior predictive distribution is defined,*

$$\hat{P}(B) = \int_{\mathscr{P}} P(B)\ d\Pi(\,P\,|\,X\,),$$

for every event $B \in \mathscr{B}$.

Lemma 8 *The posterior predictive distribution is a probability measure, almost surely.*

Proposition 9 *Endow $\mathscr{P}$ with the topology of total variation and a Borel prior Π. Suppose, either, that $\mathscr{P}$ is relatively compact, or, that Π is Radon. Then $\hat{P}$ lies in the closed convex hull of $\mathscr{P}$, almost surely.*

Proof

Let $\epsilon > 0$ be given. There exist $\{P_1, \ldots, P_N\} \subset \mathscr{P}$ such that the balls $B_i = \{P' \in \mathscr{P} : \|P' - P_i\| < \epsilon\}$ cover $\mathscr{P}$. Define $C_{i+1} = B_{i+1} \setminus \cup_{j=1}^{i} B_j$, $(C_1 = B_1)$, then $\{C_1, \ldots, C_N\}$ is a partition of $\mathscr{P}$. Define $\lambda_i = \Pi(\, C_i \,|\, X\,)$ (almost surely) and note,

$$\begin{aligned}
\|\widehat{P} - \sum_{i=1}^{N} \lambda_i \, P_i\| &= \sup_{B \in \mathscr{B}} \left| \sum_{i=1}^{N} \int_{C_i} (P(B) - P_i(B)) \, d\Pi(\, P \,|\, X = x\,) \right| \\
&\leq \sum_{i=1}^{N} \int_{C_i} \sup_{B \in \mathscr{B}} \left| P(B) - P_i(B) \right| d\Pi(\, P \,|\, X = x\,) \\
&\leq \epsilon \sum_{i=1}^{N} \Pi(\, C_i \,|\, X\,) = \epsilon
\end{aligned}$$

So there exist elements in the convex hull $\mathrm{co}(\mathscr{P})$ arbitrarily close to $\widehat{P}$. Conclude that $\widehat{P}$ lies in its TV-closure.

Posterior mean

Definition 10 *Let $\mathscr{P}$ be a model parameterized by a closed, convex Θ, subset of $\mathbb{R}^d$. Let Π be a Borel prior. If θ is integrable with respect to the posterior, the posterior mean is defined*

$$\hat{\theta}_1(Y) = \int_\Theta \theta \; d\Pi(\,\theta \,|\, Y\,) \in \Theta,$$

almost-surely.

Remark 11 *Convexity of Θ is necessary for interpretation $P_{\hat{\theta}_1}$*

Remark 12 *Caution!*

$$\hat{P}(B) \neq P_{\hat{\theta}_1}(B)$$

and different parametrizations have different $P_{\hat{\theta}_1}$

Maximum-a-posteriori estimator

Definition 13 *Let the parametrized model $\Theta \to \mathscr{P}$ and prior Π be given. Assume that the posterior is dominated with density $\theta \mapsto \pi(\theta|X)$. The maximum-a-posteriori (MAP) estimator $\hat{\theta}_2$ is defined as*

$$\pi(\hat{\theta}_2|X) = \sup_{\theta \in \Theta} \pi(\theta|X).$$

Provided that such a point exists and is unique, the MAP-estimator is defined almost-surely.

Example 14 *i.i.d.data Assume that the prior is dominated with density $\theta \mapsto \pi(\theta)$. the MAP-estimator maximizes*

$$\Theta \to \mathbb{R} \,:\, \theta \mapsto \prod_{i=1}^{n} p_\theta(X_i)\, \pi(\theta),$$

which is equivalent to log-likelihood maximization with penalty $\log \pi(\theta)$.

Frequentist coverage

Let $\mathscr{C}$ denote a collection of subsets of Θ

Definition 15 *Assume that $X \sim P_{\theta_0}$ for some $\theta_0 \in \Theta$. Choose a confidence level $\alpha \in (0,1)$. A map $C_\alpha : \mathscr{X} \to \mathscr{C}$ is a level-α confidence set if,*

$$P_\theta\Big(\theta \in C_\alpha(X)\Big) \geq 1 - \alpha,$$

for all $\theta \in \Theta$.

Definition 16 *Asymptotic coverage whenever*

$$P_{\theta,n}\Big(\theta \in C_{\alpha,n}(X)\Big) \to 1,$$

as $n \to \infty$, for all $\theta \in \Theta$

Typically confidence sets are based on an estimator $\widehat{\theta}$, or rather, on the distribution $\widehat{\theta}$ has (the so-called sampling distribution).

Credible sets

Let $\mathscr{D}$ denote a collection of subsets of Θ

Definition 17 *Let the parametrized model $\Theta \to \mathscr{P}$ and prior Π be given. Choose a confidence level $\alpha \in (0,1)$. A map $D_\alpha : \mathscr{X} \to \mathscr{D}$ is a level-α credible set if,*

$$\Pi\Big(\theta \in D_\alpha(X) \,\Big|\, X\Big) \geq 1 - \alpha,$$

P^Π-almost-surely.

Typically, credible sets in parametric models are level sets of the posterior density, the so-called HPD-credible sets.

Randomized testing

Definition 18 *Let $\mathscr{P} = \{P_\theta : \theta \in \Theta\}$ be a model for data X. Assume given a null-hypothesis H_0 and alternative hypothesis H_1 for θ,*

$$H_0 : \quad \theta_0 \in \Theta_0, \qquad H_1 : \quad \theta_0 \in \Theta_1.$$

($\{\Theta_0, \Theta_1\}$ partition of Θ). A test function ϕ is a map $\phi : \mathscr{X} \to [0,1]$. Randomized test: reject H_0 with probability $\phi(X)$.

Type-I testing power $P \mapsto P\phi(X)$ for $\theta \in \Theta_0$
Type-II testing power $P \mapsto P(1-\phi(X))$ for $\theta \in \Theta_1$

The Neyman-Pearson lemma proves optimality of

$$\phi(y) = \begin{cases} 1 & \text{if } p_{\theta_1}(y) > c p_{\theta_0}(y) \\ \gamma(x) & \text{if } p_{\theta_1}(y) = c p_{\theta_0}(y) \\ 0 & \text{if } p_{\theta_1}(y) < c p_{\theta_0}(y) \end{cases},$$

for simple hypotheses $H_0 : P = P_{\theta_0}$ versus $H_1 : P = P_{\theta_1}$.

Odds ratios and Bayes factors

Definition 19 *Let the parametrized model $\Theta \to \mathscr{P}$ and prior Π be given. Let $\{\Theta_0, \Theta_1\}$ be a partition of Θ such that $\Pi(\Theta_0) > 0$ and $\Pi(\Theta_1) > 0$. The prior odds ratio and posterior odds ratio are defined by $\Pi(\Theta_0)/\Pi(\Theta_1)$ and $\Pi(\Theta_0|Y)/\Pi(\Theta_1|Y)$. The Bayes factor for Θ_0 versus Θ_1 is defined,*

$$B = \frac{\Pi(\Theta_0|Y)}{\Pi(\Theta_1|Y)} \frac{\Pi(\Theta_1)}{\Pi(\Theta_0)}.$$

Subjectivist Accept H_0 if the posterior odds are greater than 1

Objectivist Accept H_0 if the Bayes factor is greater than 1

Test sequences and asymptotics

Data X^n, modelled with $\mathscr{P}_n = \{P_{\theta,n} : \theta \in \Theta\}$ and hypotheses $H_0 : \theta \in B$ and $H_1 : \theta \in V$ for subsets $B, V \subset \Theta$ s.t. $B \cap V = \varnothing$.

A test sequence (ϕ_n) is pointwise consistent if for all $\theta \in B$, $\theta' \in V$

$$P_{\theta,n}\phi_n \to 0 \text{ and } Q_{n,\theta'}(1-\phi_n) \to 0,$$

A test sequence (ϕ_n) is uniformly consistent if,

$$\sup_{\theta \in B} P_{\theta,n}\phi_n \to 0 \text{ and } \sup_{\theta' \in V} Q_{n,\theta'}(1-\phi_n) \to 0,$$

A test sequence (ϕ_n) is Π-a.s. consistent if,

$$P_{\theta,n}\phi_n \to 0 \text{ and } Q_{n,\theta'}(1-\phi_n) \to 0,$$

for all Π-almost-all $\theta \in B$, $\theta' \in V$.

Minimax optimal tests

We say that (ϕ_n) is minimax optimal if,

$$\sup_{\theta\in\Theta 0} P_{\theta,n}\phi_n + \sup_{\theta\in\Theta_1} P_{\theta,n}(1-\phi_n) = \inf_{\psi}\left(\sup_{\theta\in\Theta 0} P_{\theta,n}\psi + \sup_{\theta\in\Theta_1} P_{\theta,n}(1-\psi)\right),$$

Theorem 20 *(Sion (1958))* *Assume that Φ and Θ are convex, that $\phi \mapsto R(\theta,\phi)$ is convex for every θ and that $\theta \mapsto R(\theta,\phi)$ is concave for every ϕ. Futhermore, suppose that Φ is compact and $\phi \mapsto R(\theta,\phi)$ is continuous for all θ. Then there exists a minimax optimal test ϕ^* s.t.*

$$\sup_{\theta\in\Theta} R(\theta,\phi^*) = \inf_{\phi\in\Phi}\sup_{\theta\in\Theta} R(\theta,\phi) = \sup_{\theta\in\Theta}\inf_{\phi\in\Phi} R(\theta,\phi).$$

Examples of uniform test sequences

In the following, fix $n \geq 1$ and consider *i.i.d.* data $X^n = (X_1, \ldots, X_n) \sim P^n$ for some $P \in \mathscr{P}$.

Lemma 21 *(Minimax Hellinger tests) Let $B, V \subset \mathscr{P}$ be convex with $H(B, V) > 0$. There exist a uniform test sequence (ϕ_n) s.t.*

$$\sup_{P \in B} P^n \phi_n \leq e^{-\frac{1}{2} n H^2(B,V)}, \quad \sup_{Q \in V} Q^n (1 - \phi_n) \leq e^{-\frac{1}{2} n H^2(B,V)}.$$

Proof

Minimax risk $\pi(B, V)$ for testing B versus Q is

$$\pi(B, V) = \inf_{\phi} \sup_{(P,Q)\in B\times V} (P\phi + Q(1-\phi))$$

According to the minimax theorem,

$$\inf_{\phi} \sup_{P,Q} (P\phi + Q(1-\phi)) = \sup_{P,Q} \inf_{\phi} (P\phi + Q(1-\phi))$$

On the *r.h.s.* ϕ can be chosen (P, Q)-dependently; minimal for $\phi = 1\{p < q\}$ (remember the Neyman-Pearson test) so

$$\pi(B, V) = \sup_{P,Q} (P(p < q) + Q(p \geq q))$$

Proof

Note that:

$$\begin{aligned} P(p < q) + Q(p \geq q) &= \int_{p<q} p \, d\mu + \int_{p \geq q} q \, d\mu \\ &\leq \int_{p<q} p^{1/2} q^{1/2} \, d\mu + \int_{p \geq q} p^{1/2} q^{1/2} \, d\mu \\ &= \int p^{1/2} q^{1/2} \, d\mu = 1 - \tfrac{1}{2} \int \left(p^{1/2} - q^{1/2} \right)^2 d\mu \\ &= 1 - \tfrac{1}{2} H^2(P, Q) \leq e^{-\frac{1}{2} H^2(P,Q)}. \end{aligned}$$

This relates minimax testing power to the Hellinger distance between P and Q. For product measures, n-th power.

$$\pi(P^n, Q^n) \leq e^{-\frac{1}{2} n \, H^2(P,Q)}.$$

Weak tests

In the following, fix $n \geq 1$ and consider *i.i.d.* data $X^n = (X_1, \ldots, X_n)$. The model $\mathscr{P}$ contains probability measures P s.t. $X^n \sim P^n$.

Lemma 22 *(Weak tests) Let $\epsilon > 0$, $P_0 \in \mathscr{P}$ and a measurable $f : \mathscr{X}^n \to [0,1]$ be given. Define,*

$$B = \left\{P \in \mathscr{P} : |(P^n - P_0^n)f| < \epsilon\right\}, \quad V = \left\{P \in \mathscr{P} : |(P^n - P_0^n)f| \geq 2\epsilon\right\}.$$

There exist a $D > 0$ and uniformly consistent test sequence (ϕ_n) s.t.

$$\sup_{P \in B} P^n \phi_n \leq e^{-nD}, \quad \sup_{Q \in V} Q^n (1 - \phi_n) \leq e^{-nD}.$$

Proof relies on Hoeffding's inequality

Lecture II

The Bernstein-Von Mises theorem

The second lecture is devoted to regular estimation problems and the Bernstein-von Mises theorem, both parametrically and semi-parametrically. We discuss regularity, local asymptotic normality, efficiency and the consequences and applications of the parametric Bernstein-von Mises theorem. We then turn to semiparametrics, considering consistency under perturbation, integral LAN and the semi-parametric Bernstein-von Mises theorem. Semi-parametric bias is mentioned as a major obstacle.

Example Parametric regression

Questions

Observe *i.i.d.* $Y_1, \ldots, Y_n$, $Y_i = \theta + e_i$ (or $Y_i = \theta\, X_i + e_i$, *etcetera*) with a normally distributed error (of known variance). The density for the observation is,

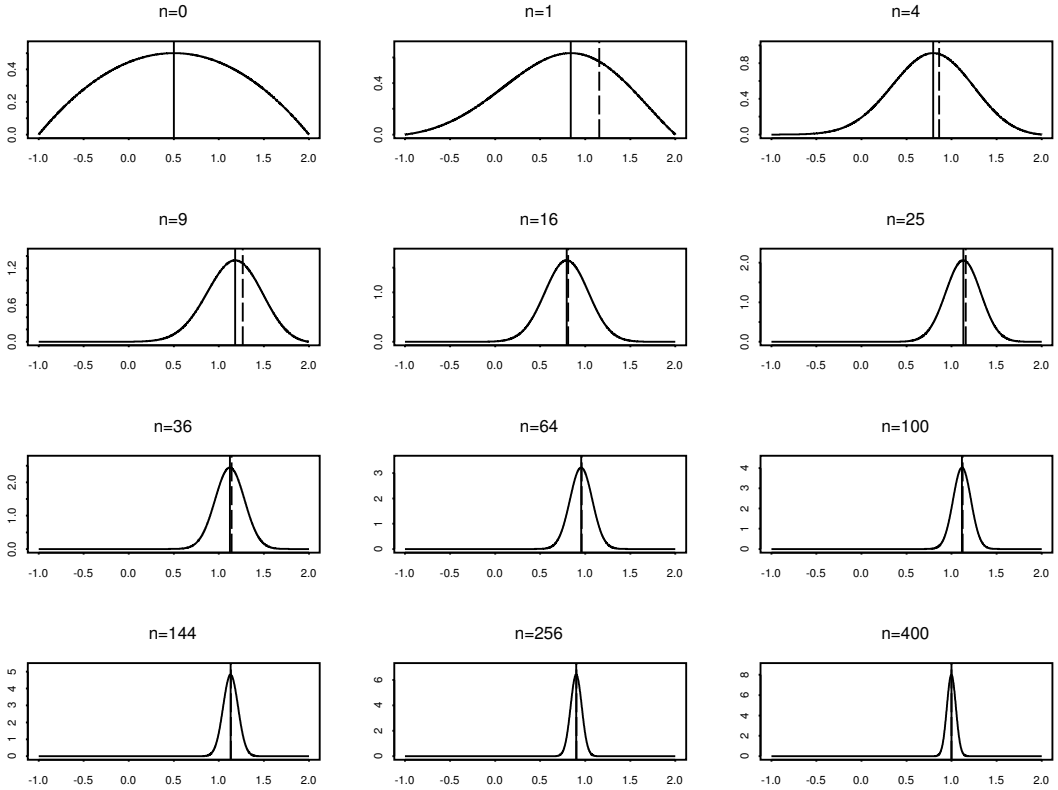
$$p_{\theta_0}(x) = \phi(x - \theta_0),$$

where ϕ is the density for the relevant normal distribution. Note the Fisher information for location is non-singular.

What should we expect of the posterior for θ in this model?

If we generalize to include non-parametric modelling freedom, what can be said about the (marginal) posterior for θ?

Convergence of the posterior



Convergence of a posterior distribution with growing sample size $n = 0, 1, 4, \ldots, 400$. Note: concentration at correct θ_0, at parameteric rate $\sqrt{n}$ and variance is the inverse Fisher information.)

Local Asymptotic Normality LAN

Definition 23 *(Le Cam (1960))*

There is a $\dot{\ell}_{\theta_0} \in L_2(P_{\theta_0})$ *with* $P_{\theta_0}\dot{\ell}_{\theta_0} = 0$ *s.t. for any* $(h_n) = O(1)$,

$$\prod_{i=1}^{n} \frac{p_{\theta_0 + n^{-1/2}h_n}}{p_{\theta_0}}(X_i) = \exp\Big(h_n^T \Delta'_{n,\theta_0} - \tfrac{1}{2} h_n^T I_{\theta_0} h_n + o_{P_{\theta_0}}(1)\Big),$$

where Δ'_{n,θ_0} *is given by,*

$$\Delta'_{n,\theta_0} = \frac{1}{\sqrt{n}} \sum_{i=1}^{n} \dot{\ell}_{\theta_0}(X_i) \xrightarrow{P_{\theta_0}\text{-w.}} N(0, I_{\theta_0}),$$

and $I_{\theta_0} = P_{\theta_0}\dot{\ell}_{\theta_0}\dot{\ell}_{\theta_0}^T$ *is the Fisher information.*

Differentiability in quadratic mean (DQM)

Definition 24 *(Le Cam (1960))*

A model $\mathscr{P}$ is differentiable in quadratic mean at θ_0 with score $\dot{\ell}_{\theta_0}$ if

$$\int \left(p_\theta^{1/2} - p_{\theta_0}^{1/2} - \tfrac{1}{2}(\theta - \theta_0)\, \dot{\ell}_{\theta_0}\, p_{\theta_0}^{1/2} \right)^2 d\mu = o\left(\|\theta - \theta_0\|^2 \right).$$

Then $P_0 \dot{\ell}_{\theta_0} = 0$, $\dot{\ell}_{\theta_0} \in L_2(P_{\theta_0})$ and $I_{\theta_0} = P_0 \dot{\ell}_{\theta_0} \dot{\ell}_{\theta_0}$ is the Fisher information.

Lemma 25 *(Le Cam (1960))*

The model $\mathscr{P}$ is DQM at θ_0 if and only if $\mathscr{P}$ is LAN at θ_0.

Regularity and the convolution theorem

Definition 26 *An estimator sequence $\hat{\theta}_n$ for a parameter θ_0 is said to be regular, if for every $h_n = O(1)$, with $\theta_n = \theta_0 + n^{-1/2}h_n$*

$$\sqrt{n}(\hat{\theta}_n - \theta_n) \xrightarrow{P_{\theta_n}\text{-w.}} L_{\theta_0}$$

for some (h_n)-independent limit distribution L_{θ_0}.

Theorem 27 *(Hájek, 1970)*
Assume that the model is LAN at θ_0 with non-singular Fisher information I_{θ_0}. Suppose $\hat{\theta}_n$ is a regular estimator for θ_0 with limit L_{θ_0}. Then there exists a probability kernel M_{θ_0} s.t.

$$L_\theta = N(0, I_{\theta_0}^{-1}) * M_{\theta_0}.$$

Regular estimation and efficiency

Definition 28 *Given an estimation problem with i.i.d.-P_0 data and non-singular Fisher information I_0, the influence functions Δ_n are,*

$$\Delta_n = I_0^{-1}\Delta_n' = \frac{1}{\sqrt{n}}\sum_{i=1}^{n} I_0^{-1}\dot{\ell}_{\theta_0}(X_i) \xrightarrow{P_0\text{-w.}} N(0, I_0^{-1})$$

Theorem 29 *(Fisher, Cramér, Rao, Le Cam, Hájek)*

An estimator $\widehat{\theta}_n$ is efficient if and only if it is asymptotically linear:

$$\sqrt{n}(\widehat{\theta}_n - \theta_0) = \Delta_{n,\theta_0} + o_{P_0}(1),$$

for some influence function $\Delta_{n,\theta_0} \xrightarrow{P_{\theta_0}\text{-w.}} N(0, I_{\theta_0}^{-1})$.

Remark 30 *asymptotic bias equals zero because $P_{\theta_0}\dot{\ell}_{\theta_0} = 0$.*

Efficiency of the maximum likelihood estimator

For all $n \geq 1$, let $X_1, \ldots, X_n$ denote *i.i.d.* data with marginal P_0.

Theorem 31 *(see van der Vaart (1998))*
Assume that $\mathscr{P} = \{P_\theta : \theta \in \Theta\}$ *with* Θ *open in* $\mathbb{R}^k$ *and* $\theta_0 \in \Theta$ *s.t.* $P_0 = P_{\theta_0}$. *Furthermore, assume that* $\mathscr{P}$ *is LAN at* θ_0 *and that* I_{θ_0} *is non-singular. Also assume there exists an* $L^2(P_{\theta_0})$*-function* $\dot{\ell}$ *s.t. for any* θ, θ' *in a neighbourhood of* θ_0 *and all* x,

$$\Big| \log p_\theta(x) - \log p_{\theta'}(x) \Big| \leq \dot{\ell}(x) \, \|\theta - \theta'\|,$$

If the ML estimate $\widehat{\theta}_n$ *is consistent, it is efficient,*

$$\sqrt{n}(\widehat{\theta}_n - \theta_0) \xrightarrow{P_{\theta_0}\text{-w.}} N(0, I_{\theta_0}^{-1}).$$

Parametric Bernstein-von Mises theorem

Theorem 32 (Le Cam (1953), Le Cam-Yang (1990), $h = \sqrt{n}(\theta - \theta_0)$)
Let $\mathscr{P} = \{P_\theta : \theta \in \Theta \subset \mathbb{R}^d\}$ *with thick prior* Π_Θ *be LAN at* θ_0 *with non-singular* I_{θ_0}. *Assume that for every sequence of radii* $M_n \to \infty$,

$$\Pi\Big(\|h\| \leq M_n \;\Big|\; X_1, \ldots, X_n \Big) \xrightarrow{P_0} 1$$

Then the posterior converges to normality as follows

$$\sup_B \Big| \Pi\Big(h \in B \;\Big|\; X_1, \ldots, X_n \Big) - N_{\Delta_{n,\theta_0}, I_{\theta_0}^{-1}}(B) \Big| \xrightarrow{P_0} 0$$

Remark 33 *With* $\widehat{\theta}_n$ *any efficient estimator,*

$$\sup_B \Big| \Pi\Big(\theta \in B \;\Big|\; X_1, \ldots, X_n \Big) - N_{\widehat{\theta}_n, (nI_{\theta_0})^{-1}}(B) \Big| \xrightarrow{P_0} 0$$

Remark 34 *(BK and van der Vaart, 2012) There's a version for the misspecified situation* ($P_0 \notin \mathscr{P}$).

Consequences and applications

i. Bayesian point estimators are efficient

ii. Confidence intervals based on the sampling distribution of an efficient estimator and credible sets coincide asymptotically

Model selection with the Bayesian Information Criterion (BIC). Consider parameter spaces $\Theta_k \subset \mathbb{R}^k$, $(k \geq 1)$ with models $\mathscr{P}_k$ for *i.i.d.* data $X_1, \ldots, X_n$. Define,

$$\text{BIC}(\theta, k) = -2 \log L_n(X_1, \ldots, X_n; \theta_1, \ldots, \theta_k) + k \log(n)$$

Minimization of $\text{BIC}(\theta_1, \ldots, \theta_k; k)$ with respect to θ and k is penalized ML estimate that selects a value of k. Closely related to AIC, RIC, MDL and other model selection methods.

Efficiency of formal Bayes estimators

Definition 35 *Let X, $\mathscr{P}$, Π be like before and let $\ell : \mathbb{R}^k \to [0, \infty)$ be a loss function. The posterior risk is defined almost-surely,*

$$t \mapsto \int_{\Theta} \ell\Big(\sqrt{n}(t-\theta)\Big)\, d\Pi(\theta|X).$$

A minimizer $\widehat{\theta}_{3,n}$ of posterior risk is called the formal Bayes estimator associated with ℓ and Π

Theorem 36 *(Le Cam (1953,1986) and van der Vaart (1998))*
Assume that the BvM theorem holds and that ℓ is non-decreasing and $\ell(h) \leq 1 + \|h\|^p$ for some $p > 0$ such that $\int \|\theta\|^p \, d\Pi(\theta) < \infty$. Then $\sqrt{n}(\widehat{\theta}_{3,n} - \theta_0)$ converges weakly to the minimizer of $\int \ell(t-h)\, dN_{Z, I_{\theta_0}^{-1}}(h)$ where $Z \sim N(0, I_{\theta_0}^{-1})$.

Example Semiparametric regression

New Question

Observe *i.i.d.* $X_1, \ldots, X_n$, $X_i = \theta + e_i$ (or $Y_i = \theta\, X_i + e_i$, *etcetera*) with a symmetrically distributed error. Density for X's is,

$$p_{\theta_0,\eta_0}(x) = \eta_0(x - \theta_0),$$

where $\eta \in H$ is a symmetric Lebesgue density on $\mathbb{R}$. We assume that η is smooth and that the Fisher information for location is non-singular.

Adaptivity Stein (1956), Bickel (1982)

For inference on θ_0 it does not matter whether we know η_0 or not!

$$\sqrt{n}(\widehat{\theta}_n - \theta_0) \xrightarrow{P_{\theta_0,\eta_0}\text{-w.}} N(0, I^{-1}_{\theta_0,\eta_0})$$

where I_{θ_0,η_0} is the Fisher information.

Parametric/Semi-parametric analogy

Parametric posterior

The posterior density $\theta \mapsto d\Pi(\theta|X_1,\ldots,X_n)$

$$\prod_{i=1}^{n} p_\theta(X_i)\, d\Pi(\theta) \Big/ \int_\Theta \prod_{i=1}^{n} p_\theta(X_i)\, d\Pi(\theta)$$

with LAN requirement on the likelihood.

Semiparametric analog

The marginal posterior density $\theta \mapsto d\Pi(\theta|X_1,\ldots,X_n)$

$$\int_H \prod_{i=1}^{n} p_{\theta,\eta}(X_i)\, d\Pi_H(\eta)\, d\Pi_\Theta(\theta) \Big/ \int_\Theta \int_H \prod_{i=1}^{n} p_{\theta,\eta}(X_i)\, d\Pi_H(\eta)\, d\Pi_\Theta(\theta)$$

with integral LAN requirement on Π_H-integrated likelihood.

Integral local asymptotic normality ILAN

Definition 37 *Given a nuisance prior* Π_H, *the localized integrated likelihood is,*

$$s_n(h) = \int_H \prod_{i=1}^{n} \frac{p_{\theta_0+n^{-1/2}h,\eta}}{p_{\theta_0,\eta_0}}(X_i)\, d\Pi_H(\eta),$$

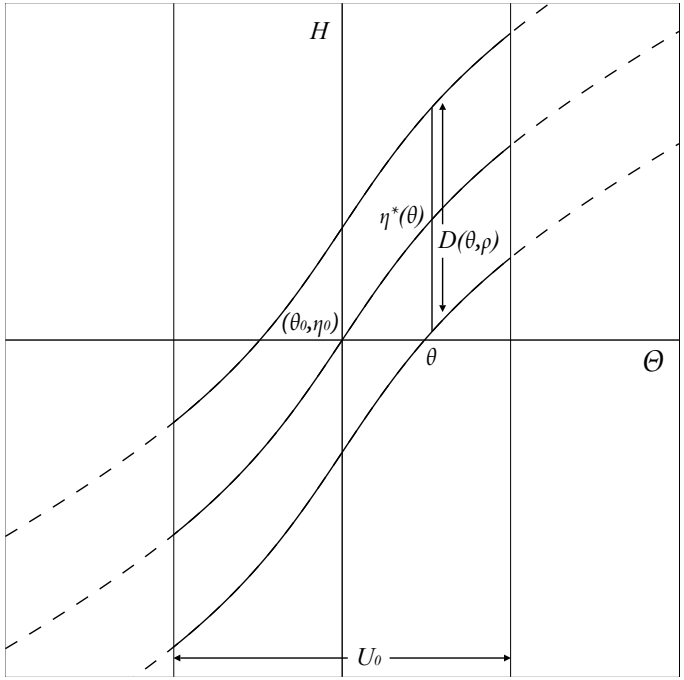
Definition 38 s_n *is said to have the ILAN property, if for every* $h_n = O_{P_0}(1)$

$$\log \frac{s_n(h_n)}{s_n(0)} = h_n^T\, \tilde{\Delta}'_{n,\theta_0,\eta_0} - \tfrac{1}{2} h_n^T \tilde{I}_{\theta_0,\eta_0} h_n + o_{P_0}(1),$$

where the efficient $\tilde{\Delta}'_{n,\theta_0,\eta_0}$ *is given by*

$$\tilde{\Delta}'_{n,\theta_0,\eta_0} = \frac{1}{\sqrt{n}} \sum_{i=1}^{\infty} \tilde{\ell}_{\theta_0,\eta_0} \xrightarrow{P_{\theta_0,\eta_0}\text{-w.}} N(0, \tilde{I}_{\theta_0,\eta_0})$$

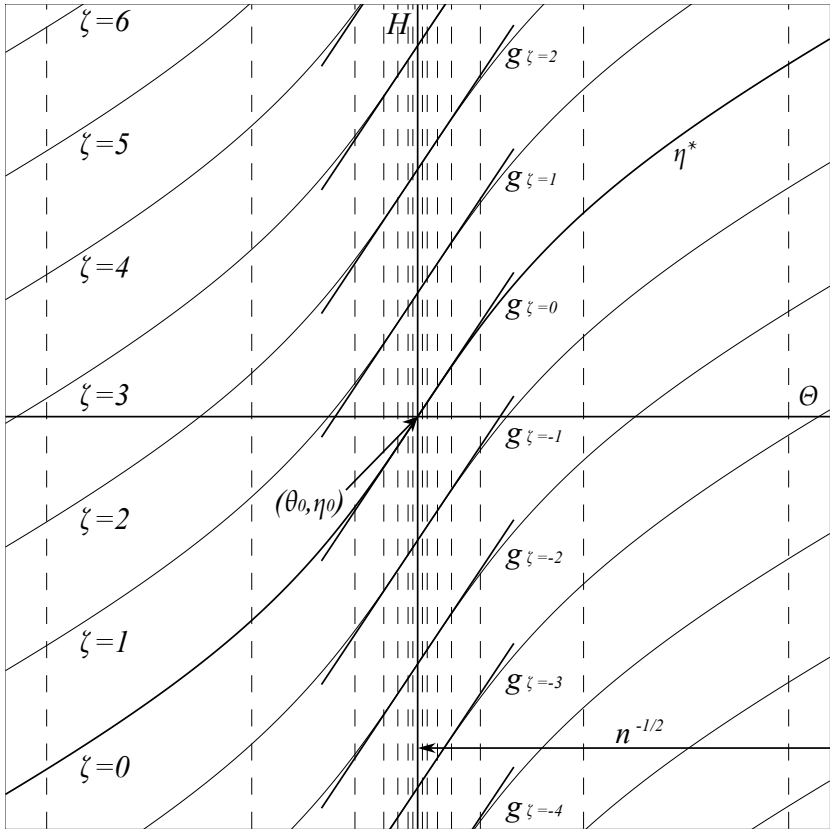
Consistency under $\sqrt{n}$-perturbation



Given $\rho_n \downarrow 0$ we speak of *consistency under $n^{-1/2}$-perturbation at rate ρ_n*, if for all $h_n = O_{P_0}(1)$.

$$\Pi_n\Big(D(\theta, \rho_n) \;\Big|\; \theta = \theta_0 + n^{-1/2} h_n \,;\, X_1, \ldots, X_n \Big) \xrightarrow{P_0} 1$$

Integral LAN



reparametrize $(\theta, \zeta) \mapsto (\theta, \eta^*(\theta) + \zeta)$

Semiparametric Bernstein-von Mises theorem

Theorem 39 (Bickel and BK (2012))

Let $\mathscr{P} = \{P_{\theta,\eta} : \theta \in \Theta, \eta \in H\}$ *with thick prior* Π_Θ *and nuisance prior* Π_H. *Assume ILAN at* P_{θ_0,η_0} *with non-singular* $\tilde{I}_{\theta_0,\eta_0}$. *Assume that for every sequence of radii* $M_n \to \infty$,

$$\Pi\Big(\|h\| \leq M_n \;\Big|\; X_1, \ldots, X_n \Big) \xrightarrow{P_0} 1$$

Then the posterior converges marginally to normality as follows

$$\sup_B \Big| \Pi\Big(h \in B \;\Big|\; X_1, \ldots, X_n \Big) - N_{\tilde{\Delta}_{n,\theta_0,\eta_0}, \tilde{I}^{-1}_{\theta_0,\eta_0}}(B) \Big| \xrightarrow{P_0} 0$$

BOTH ILAN and $\sqrt{n}$-consistency are sensitive to semiparametric bias!

Semiparametric bias

An estimator $\widehat{\theta}_n$ for θ_0 is regular but asymptotically biased if,

$$\sqrt{n}(\widehat{\theta}_n - \theta_0) = \tilde{\Delta}_{n,\theta_0,\eta_0} + \mu_{n,\theta_0,\eta_0} + o_{P_0}(1),$$

with $\tilde{\Delta}_{n,\theta_0,\eta_0} \xrightarrow{P_0\text{-w.}} N(0, \tilde{I}^{-1}_{\theta_0\eta_0})$ and $\mu_{n,\theta_0,\eta_0} = O(1)$ or worse. Typically,

$$\left| \mu_{n,\theta_0,\eta_0} \right| \leq n^{-1/2} \sup_{\eta \in D_n} \left| \tilde{I}^{-1}_{\theta_0,\eta_0} P_{\theta_0,\eta} \tilde{\ell}_{\theta_0,\eta_0} \right|$$

where D_n describes some form of localization for $\eta \in H$ around η_0.

Theorem 40 *(approximate, see Schick (1986), Klaassen (1987))*
An efficient estimator for θ_0 exists if and only if there exists an estimator $\widehat{\Delta}_n$ for the influence function, whose asymptotic bias vanishes at a rate strictly faster than $\sqrt{n}$,

$$P^n_{\theta_n,\eta} \widehat{\Delta}_n = o(n^{-1/2}),$$

Example Regression with symmetric errors

Theorem 41 *(Chae, Kim and BK (2018))*

Let $X_1, \ldots, X_n$ *be i.i.d.-*P_{θ_0,η_0}*, i.e.* $X_i = \theta_0 + e_i$ *with* e *distributed as a symmetric normal location mixture* η_0 *from* H *of the form,*

$$\eta(x) = \int \phi(x - z)\, dF(z)$$

(where F *is symmetric and* ϕ *denotes the standard normal density). With thick prior* Π_Θ *and nuisance prior* Π_H *that has full weak support, the posterior converges marginally to normality*

$$\sup_B \left| \Pi\Big(h \in B \;\Big|\; X_1, \ldots, X_n \Big) - N_{\tilde{\Delta}_{n,\theta_0,\eta_0}, \tilde{I}^{-1}_{\theta_0,\eta_0}}(B) \right| \xrightarrow{P_0} 0$$

where $\tilde{\ell}_{\theta_0,\eta_0}(X) = \dot{p}_{\theta_0,\eta_0}/p_{\theta_0,\eta_0}(X)$ *and* $\tilde{I}_{\theta_0,\eta_0} = P_0 \tilde{\ell}^2_{\theta_0,\eta_0}$.

Lecture III

Bayes and the Infinite

In the third lecture we consider application of Bayesian methods in non-parametric models: we do not focus on the construction of non-parametric priors but on the requirements for such priors to lead to consistent posteriors. After a review of the consequences of posterior consistency, we turn to Doob's theorem and Schwartz's theorem, which we prove. We also point out limitations of Schwartz's theorem.

Frequentist consistency

Let $X_1, \ldots, X_n$ be *i.i.d.*-P_{θ_0}-distributed

Consider a point-estimator $\widehat{\theta}_n(X)$.

An estimator is said to be (strongly) consistent if

$$\widehat{\theta}_n \xrightarrow{P_{\theta_0}} \theta_0.$$

E.g. if the topology is metric, a consistent estimator $\widehat{\theta}_n$ is found at a distance from θ_0 greater than some $\epsilon > 0$ with $P_{\theta_0,n}$-probability arbitrarily small, if we make the sample large enough.

Since θ_0 is unknown, we have to prove this *for all* $\theta \in \Theta$ before it is useful.

Frequentist rate of convergence

Next, suppose that $\widehat{\theta}_n \xrightarrow{P_{\theta_0}} \theta_0$. Let (r_n) be a sequence $r_n \downarrow 0$.

We say that $\widehat{\theta}_n$ converges to θ_0 at rate r_n if

$$r_n^{-1} \|\widehat{\theta}_n - \theta_0\| = O_{P_{\theta_0}}(1)$$

So r_n compensates the decrease in distance between $\widehat{\theta}_n$ and θ_0, such that the fraction is bounded in probability.

Or: the r_n are the radii of balls around $\widehat{\theta}_n$ that shrink (just) slowly enough to still capture θ_0 with high probability.

Frequentist limit distribution

Suppose that $\widehat{\theta}_n$ converges to θ_0 at rate r_n.

Let L_{θ_0} be a non-degenerate but tight distribution. If

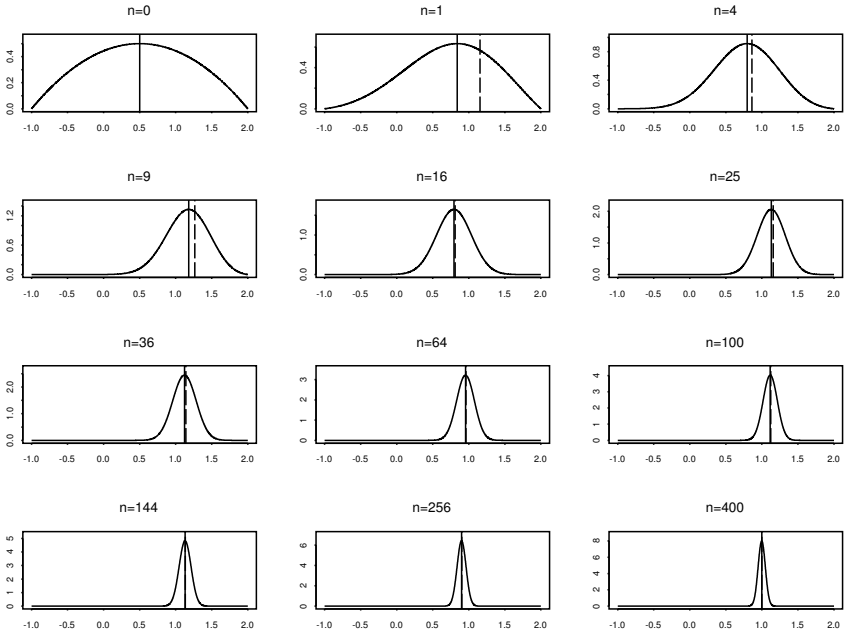
$$r_n^{-1}(\widehat{\theta}_n - \theta_0) \xrightarrow{P_{\theta_0}\text{-w.}} L_{\theta_0},$$

we say that $\widehat{\theta}_n$ converges to θ_0 at rate r_n with limit-distribution L_{θ_0}.

So if we blow up the difference between $\widehat{\theta}_n$ and θ_0 by exactly the right factors r_n^{-1}, we keep up with convergence and arrive at a stable distribution L_{θ_0}.

Posterior consistency

Given P_0-*i.i.d.* X^n, $\mathscr{P}$ with prior Π, do posteriors concentrate on P_0?



Definition 42 *Given a model $\mathscr{P}$ with Borel prior Π, the posterior is (strongly) consistent at $P \in \mathscr{P}$ if for every neighbourhood U of P*

$$\Pi(U|X^n) \xrightarrow{P(-a.s.)} 1 \tag{3}$$

Consistency is Prokhorov's weak convergence

Theorem 43 *Let $\mathscr{P}$ be a uniform model with Borel prior Π. The posterior is strongly consistent, if and only if, for every bounded, continuous $f : \mathscr{P} \to \mathbb{R}$,*

$$\int f(P)\, d\Pi(P|X^n) \xrightarrow{P_0} f(P_0), \tag{4}$$

which we denote by $\Pi(\,\cdot\,|X_1, \ldots, X_n) \xrightarrow{w} \delta_{P_0}$.

Remark 44 *All weak, polar and metric topologies are uniform:*

$$U = \{P \in \mathscr{P} : |(P - P_0)f| < \epsilon\}, V = \{P \in \mathscr{P} : \sup_{f \in B} |(P - P_0)f| < \epsilon\},$$

$$W = \{P \in \mathscr{P} : d(P, P_0) < \epsilon\},$$

for $\epsilon > 0$ and functions $0 \leq f \leq 1$ measurable (or smaller class).

Proof

Assume (3). $f : \mathscr{P} \to \mathbb{R}$ is bounded ($|f| \leq M$) and continuous. Let $\eta > 0$ be given. Let U be a neighbourhood of P_0 s.t. $|f(P) - f(P_0)| \leq \eta$ for all $P \in U$.

Integrate f with respect to the posterior and to δ_{P_0}:

$$\begin{aligned}
&\left| \int_{\mathscr{P}} f(P)\, d\Pi_n(P|X_1, \ldots, X_n) - f(P_0) \right| \\
&\qquad \leq \int_{\mathscr{P}\setminus U} |f(P) - f(P_0)|\, d\Pi_n(P|X_1, \ldots, X_n) \\
&\qquad\quad + \int_{U} |f(P) - f(P_0)|\, d\Pi_n(P|X_1, \ldots, X_n) \\
&\qquad \leq 2M\, \Pi_n(\, \mathscr{P} \setminus U \,|\, X_1, X_2, \ldots, X_n\,) \\
&\qquad\quad + \sup_{P \in U} |f(P) - f(P_0)|\, \Pi_n(\, U \,|\, X_1, X_2, \ldots, X_n\,) \\
&\qquad \leq \eta + o_{P_0}(1).
\end{aligned}$$

Proof

Conversely, assume (4) holds. Let U be an open neighbourhood of P_0. Because $\mathscr{P}$ is completely regular, there exists a continuous $f : \mathscr{P} \to [0,1]$ that separates $\{P_0\}$ from $\mathscr{P} \setminus U$, *i.e.* $f = 1$ at $\{P_0\}$ and $f = 0$ on $\mathscr{P} \setminus U$.

$$\Pi_n(\, U \,|\, X_1, X_2, \ldots, X_n \,) = \int_{\mathscr{P}} 1_U(P)\, d\Pi_n(P|X_1, \ldots, X_n)$$

$$\geq \int_{\mathscr{P}} f(P)\, d\Pi_n(P|X_1, \ldots, X_n) \xrightarrow{P_0} \int_{\mathscr{P}} f(P)\, d\delta_{P_0}(P) = 1,$$

Consequently, (3) holds.

Consistency of Bayesian point estimators

Theorem 45 *Suppose that $\mathscr{P}$ is a is endowed with the topology of total variation. Assume that the posterior is (strongly) consistent. Then the posterior mean $\hat{P}_n$ is a (P_0-almost-surely) consistent point-estimator in total-variation.*

Proof

Extend $P \mapsto \|P - P_0\|$ to the convex hull of $\mathscr{P}$. Since $P \mapsto \|P - P_0\|$ is convex, Jensen says,

$$\begin{aligned} \|\hat{P}_n - P_0\| &= \left\| \int_{\mathscr{P}} P \, d\Pi_n(P \,|\, X_1, \ldots, X_n) - P_0 \right\| \\ &\leq \int_{\mathscr{P}} \|P - P_0\| \, d\Pi_n(P \,|\, X_1, \ldots, X_n). \end{aligned}$$

Since $P \xrightarrow{\Pi_n\text{-w.}} P_0$ under $\Pi_n = \Pi_n(\cdot \,|\, X_1, \ldots, X_n)$ and $P \mapsto \|P - P_0\|$ is bounded and continuous, the *r.h.s.* converges to the expectation of $\|P - P_0\|$ under the limit δ_{P_0}, which equals zero. Hence

$$\hat{P}_n \xrightarrow{P_0} P_0,$$

in total variation.

Doob's theorem

Theorem 46 *(Doob (1948))*
Suppose that the parameter space Θ *and the sample space* $\mathscr{X}$ *are Polish spaces endowed with their respective Borel* σ*-algebras. Assume that* $\Theta \to \mathscr{P} : \theta \mapsto P_\theta$ *is one-to-one. Then for any prior* Π *on* Θ *the posterior is consistent,* Π*-almost-surely.*

Proof An application of Doob's martingale convergence theorem, combined with a difficult argument on existence of a measurable $f : \mathscr{X}^\infty \to \Theta$ s.t. $f(X_1, X_2, \ldots) = \theta$, $P_\theta^\infty - a.s.$ for all $\theta \in \Theta$ (Le Cam's accessibility (Breiman, Le Cam, Schwartz (1964), Le Cam (1986)). □

Freedman's point

Remark 47 *Doob's theorem says nothing about specific points: it is always possible that P_0 belongs to the null-set for which inconsistency occurs.*

Remark 48 *(Non-parametric counterexamples)*
Schwartz (1961), Freedman (1963,1965), Diaconis and Freedman (1986), Cox (1993), Freedman and Diaconis (1998). Basically what is shown is that Doob's null-set of inconsistency can be rather large.

Schwartz's theorem

Theorem 49 *(Schwartz (1965)) Assume that*

(i) For every $\epsilon > 0$*, there is a uniform test sequence* (ϕ_n) *such that*

$$P_0^n \phi_n \to 0, \qquad \sup_{\{P: d(P,P_0) > \epsilon\}} P^n(1 - \phi_n) \to 0.$$

(ii) Let Π *be a KL-prior, i.e. for every* $\eta > 0$,

$$\Pi\left(P \in \mathscr{P} \,:\, -P_0 \log \frac{p}{p_0} \leq \eta \right) > 0,$$

Then the posterior is consistent at P_0.

Corollary 50 *Let* $\mathscr{P}$ *be Hellinger totally bounded and let* Π *a KL-prior. Then the posterior is Hellinger consistent at* P_0.

Proof of Schwartz's theorem (I)

Let $\epsilon, \eta > 0$ be given. Define

$$V = \{ P \in \mathscr{P} : d(P, P_0) \geq \epsilon \}.$$

Split the n-th posterior (of V) with the test functions ϕ_n

$$\begin{aligned} \limsup_{n\to\infty} \Pi_n(V|X_1, \ldots, X_n) \leq & \limsup_{n\to\infty} \Pi_n(V|X_1, \ldots, X_n)(1 - \phi_n) \\ & + \limsup_{n\to\infty} \Pi_n(V|X_1, \ldots, X_n)\phi_n. \end{aligned} \tag{5}$$

Define $K_\eta = \{P \in \mathscr{P} : -P_0 \log(p/p_0) \leq \eta\}$. For every $P \in K_\eta$, LLN

$$\left| \frac{1}{n} \sum_{i=1}^{n} \log \frac{p}{p_0} - P_0 \log \frac{p}{p_0} \right| \to 0, \quad (P_0 - a.s.).$$

Proof of Schwartz's theorem (II)

So for every $\alpha > \eta$ and all $P \in K_\eta$ and large enough n,

$$\prod_{i=1}^{n} \frac{p}{p_0}(X_i) \geq e^{-n\alpha},$$

P_0^n-almost-surely. Use this to lower-bound the denominator

$$\liminf_{n\to\infty} e^{n\alpha} \int_{\mathscr{P}} \prod_{i=1}^{n} \frac{p}{p_0}(X_i)\, d\Pi(P) \geq \liminf_{n\to\infty} e^{n\alpha} \int_{K_\eta} \prod_{i=1}^{n} \frac{p}{p_0}(X_i)\, d\Pi(P)$$

$$\geq \int_{K_\eta} \liminf_{n\to\infty} e^{n\alpha} \prod_{i=1}^{n} \frac{p}{p_0}(X_i)\, d\Pi(P) \geq \Pi(K_\eta) > 0.$$

Proof of Schwartz's theorem (III)

The first term in (5) can be bounded as follows

$$\limsup_{n\to\infty} \Pi(V|X_1,\ldots,X_n)(1-\phi_n(X_1,\ldots,X_n))$$

$$\leq \frac{\limsup_{n\to\infty} e^{n\alpha} \int_V \prod_{i=1}^n (p/p_0)(X_i)\,(1-\phi_n(X_1,\ldots,X_n))\,d\Pi(P)}{\liminf_{n\to\infty} e^{n\alpha} \int_{\mathscr{P}} \prod_{i=1}^n (p/p_0)(X_i)\,d\Pi(P)} \tag{6}$$

$$\leq \frac{1}{\Pi(K_\eta)} \limsup_{n\to\infty} f_n(X_1,\ldots,X_n),$$

where we use the (non-negative)

$$f_n(X_1,\ldots,X_n) = e^{n\alpha} \int_V \prod_{i=1}^n \frac{p}{p_0}(X_i)\,(1-\phi_n)(X_1,\ldots,X_n)\,d\Pi(P).$$

Proof of Schwartz's theorem, interlude

At this stage in the proof we need the following lemma, which says that uniform consistency of testing can be assumed to be of exponential power without loss of generality.

Lemma 51 *Let P_0 and V with $P_0 \notin V$ be given. Suppose that there exists a sequence of tests (ϕ_n) such that:*

$$P_0^n \phi_n \to 0, \qquad \sup_{P \in V} P^n(1 - \phi_n) \to 0,$$

Then there exists a sequence of tests (ω_n) and positive constants C, D such that:

$$P_0^n \omega_n \leq e^{-nC}, \qquad \sup_{P \in V} P^n(1 - \omega_n) \leq e^{-nD} \tag{7}$$

Proof of Schwartz's theorem (IV)

The previous lemma guarantees that there exists a constant $\beta > 0$ such that for large enough n,

$$P_0^\infty f_n = P_0^n f_n = e^{n\alpha} \int_V P_0^n \Big(\prod_{i=1}^n \frac{p}{p_0}(X_i)\,(1-\phi_n)(X_1,\ldots,X_n) \Big)\, d\Pi(P)$$
$$\leq e^{n\alpha} \int_V P^n (1-\phi_n)\, d\Pi(P) \leq e^{-n(\beta-\alpha)}. \tag{8}$$

Choose $\eta < \beta$ and α such that $\eta < \alpha < \frac{1}{2}(\beta+\eta)$. Markov's inequality

$$P_0^\infty \Big(f_n > e^{-\frac{n}{2}(\beta-\eta)} \Big) \leq e^{\frac{n}{2}(\beta-\eta)}\, P_0^\infty f_n \leq e^{n(\alpha-\frac{1}{2}(\beta+\eta))}.$$

Proof of Schwartz's theorem (V)

Hence $\sum_{n=1}^{\infty} P_0^{\infty}(f_n > \exp -\frac{n}{2}(\beta - \eta))$ converges. Borel-Cantelli

$$0 = P_0^{\infty}\Big(\bigcap_{N=1}^{\infty} \bigcup_{n \geq N} \Big\{ f_n > e^{-\frac{n}{2}(\beta-\eta)} \Big\}\Big) \geq P_0^{\infty}\Big(\limsup_{n\to\infty}\Big(f_n - e^{-\frac{n}{2}(\beta-\eta)}\Big) > 0\Big)$$

So $f_n \xrightarrow{P_0\text{-a.s.}} 0$ and hence

$$\Pi(V|X_1,\ldots,X_n)\,(1-\phi_n)(X_1,\ldots,X_n) \xrightarrow{P_0\text{-a.s.}} 0.$$

The other term in (5) $P_0^n \Pi(V|X_1,\ldots,X_n)\phi_n \leq P_0^n \phi_n \leq e^{-nC}$ so that

$$\Pi(V|X_1,\ldots,X_n)\,\phi_n(X_1,\ldots,X_n) \xrightarrow{P_0\text{-a.s.}} 0. \qquad (9)$$

Combination of (6) and (9) proves that (5) equals zero.

... but there are very nasty examples

Example 52 *Consider P_0 on $\mathbb{R}$ with Lebesgue density p_0 supported on an interval of width one but unknown location. With $\eta(x) > 0$, if $x \in (0,1)$ and $\eta(x) = 0$ otherwise, and $\theta \in \mathbb{R}$:*

$$p_\theta(x) = \eta(x-\theta)\, 1_{[\theta,\theta+1]}(x)$$

Note that if $\theta \neq \theta'$,

$$-P_{\theta,\eta} \log \frac{p_{\theta',\eta}}{p_{\theta,\eta}} = \infty$$

Kullback-Leibler neighbourhoods are singletons: no prior can be a Kullback-Leibler prior in this model!

Lecture IV

Posterior contraction

In the fourth lecture, we delve deeper into the theory on posterior convergence, motivated by examples that show the limitations of Schwartz's prior mass condition. We prove an alternative consistency theorem that does not rely on KL-priors. We also make contact with Barron's theorem, Walker's theorem and the Ghosal-Ghosh-van der Vaart theorem on the rate of posterior convergence. We derive a theorem on posterior rates of convergence with a KL-type prior-mass condition.

[arxiv: 1308.1263v3]

Recall Schwartz

Theorem 53 *(Schwartz (1965))*
Let $\mathscr{P}$ be Hellinger totally bounded and let Π a KL-prior, i.e. for $\eta > 0$,

$$\Pi\left(P \in \mathscr{P} \,:\, -P_0 \log \frac{p}{p_0} \leq \eta \right) > 0,$$

Then the posterior is Hellinger consistent at P_0.

Example 54 *Consider P_0 on $\mathbb{R}$ with density,*

$$p_\theta(x) = \eta(x-\theta)\, 1_{[\theta,\theta+1]}(x),$$

for some $\theta \in \mathbb{R}$. Note that if $\theta \neq \theta'$,

$$-P_{\theta,\eta} \log \frac{p_{\theta',\eta'}}{p_{\theta,\eta}} = \infty$$

no prior can be a Kullback-Leibler prior in this model!

Walker's theorem

Theorem 55 *(Walker (2004))*
Let $\mathscr{P}$ be Hellinger separable. Let $\{V_i : i \geq 1\}$ be a countable cover of $\mathscr{P}$ by balls of radius ϵ. If Π is a Kullback-Leibler prior and,

$$\sum_{i \geq 1} \Pi(V_i)^{1/2} < \infty$$

then $\Pi(\, H(P, P_0) > \epsilon \,|\, X_1, \ldots, X_n \,) \xrightarrow{P_0\text{-a.s.}} 0$.

The Ghosal-Ghosh-van der Vaart theorem

Theorem 56 *(Ghosal, Ghosh and van der Vaart, 2000)*
Let (ϵ_n) *be such that* $\epsilon_n \downarrow 0$ *and* $n\epsilon_n^2 \to \infty$. *Let* $C > 0$ *and* $\mathscr{P}_n \subset \mathscr{P}$ *be such that, for large enough* n,

(i) $N(\epsilon_n, \mathscr{P}_n, H) \leq e^{-n\epsilon_n^2}$

(ii) $\Pi(\mathscr{P} \setminus \mathscr{P}_n) \leq e^{-n\epsilon_n^2(C+4)}$

(iii) the prior Π *is a GGV-prior, i.e.*

$$\Pi\left(P \in \mathscr{P} \,:\, -P_0 \log \frac{dP}{dP_0} < \epsilon_n^2,\, P_0\left(\log \frac{dP}{dP_0}\right)^2 < \epsilon_n^2 \right) \geq e^{-Cn\epsilon_n^2}$$

Then, for some $M > 0$,

$$\Pi(\, P \in \mathscr{P} \,:\, H(P, P_0) > M\epsilon_n \,|\, X_1, \ldots, X_n \,) \xrightarrow{P_0} 0$$

... but here's another tricky example

Example 57 *Consider the distributions* P_a, $(a \geq 1)$, *defined by,*

$$p_a(k) = P_a(X = k) = \frac{1}{Z_a} \frac{1}{k^a (\log k)^3}$$

for all $k \geq 2$, *with* $Z_a = \sum_{k \geq 2} k^{-a} (\log k)^{-3} < \infty$. *For* $a = 1$, $b > 1$,

$$-P_a \log \frac{p_b}{p_a} < \infty, \quad P_a \left(\log \frac{p_b}{p_a} \right)^2 = \infty$$

Schwartz's KL-condition for the prior for the parameter a *can be satisfied but GGV priors do not exist.*

Remark 58 *With* $(\log k)^2$ *instead of* $(\log k)^3$, *KL-priors also fail.*

Posterior convergence

Recall the prior predictive distribution $P_n^{\Pi}(A) = \int_{\mathscr{P}} P^n(A)\, d\Pi(P)$.

Theorem 59 *Assume that* $P_0^n \ll P_n^{\Pi}$ *for all* $n \geq 1$. *Let* $V_1, \ldots, V_N$ *be a finite collection of model subsets. If there exist constants* $D_i > 0$ *and test sequences* $(\phi_{i,n})$ *for all* $1 \leq i \leq N$ *such that,*

$$P_0^n \phi_{i,n} + \sup_{P \in V_i} P_0^n \frac{dP^n}{dP_n^{\Pi}}(1 - \phi_{i,n}) \leq e^{-nD_i}, \tag{10}$$

for large enough n, *then any* $V \subset \bigcup_{1 \leq i \leq N} V_i$ *receives posterior mass zero asymptotically,*

$$\Pi(\, V | X_1, \ldots, X_n \,) \xrightarrow{P_0\text{-a.s.}} 0. \tag{11}$$

Proof

If $\Pi(V_i|X_1,\ldots,X_n) \xrightarrow{P_0\text{-a.s.}} 0$ for all $1 \leq i \leq N$ then the assertion is proved. So pick some i and consider,

$$P_0^n \Pi(V_i|X_1,\ldots,X_n) \leq P_0^n \phi_n + P_0^n \Pi(V_i|X_1,\ldots,X_n)(1-\phi_n)$$

By Fubini,

$$P_0^n \Pi(V_i|X_1,\ldots,X_n)(1-\phi_n) = \int_{V_i} \frac{dP^n}{P_n^\Pi}(1-\phi_n)\, d\Pi(P)$$

$$\leq \Pi(V_i) \sup_{P \in V_i} P_0\left(\frac{dP^n}{dP_n^\Pi}\right)(1-\phi_n) \leq e^{-nD_i}$$

Apply Markov and Borel-Cantelli to conclude that,

$$\limsup_{n\to\infty} \Pi(V_i|X_1,\ldots,X_n) = 0.$$

Minimax test sequence

Lemma 60 *Let $V \subset \mathscr{P}$ be given and assume that $P_0^n(dP^n/dP_n^\Pi) < \infty$ for all $P \in V$. For every B there exists a test sequence (ϕ_n) such that,*

$$P_0^n\phi_n + \sup_{P\in V} P_0^n \frac{dP^n}{dP_n^\Pi}(1-\phi_n) \leq \inf_{0\leq\alpha\leq 1} \Pi(B)^{-\alpha} \int \left(\sup_{P\in \mathrm{co}(V)} P_0\left(\frac{dP}{dQ}\right)^\alpha \right)^n d\Pi(Q|B).$$

i.e. testing power is bounded in terms of Hellinger transforms.

The construction is technically close to that needed for the analysis of posteriors for misspecified models, *i.e.* when $P_0 \notin \mathscr{P}$ (see, Kleijn and van der Vaart (2006)).

Sketch of the proof

Let $Q_n^{\Pi}(A)$ be the prior predictive with $\Pi(\,\cdot\,|B)$: $P_n^{\Pi}(A) \geq \Pi(B)\, Q_n^{\Pi}(A)$ and using Jensen's inequality, for $P_n \in \mathrm{co}(V^n)$

$$\begin{aligned} P_0^n\left(\frac{dP_n}{dP_n^{\Pi}}\right)^{\alpha} &\leq \Pi(B)^{-\alpha}\, P_0^n\left(\frac{dP_n}{dQ_n^{\Pi}}\right)^{\alpha} \\ &\leq \Pi(B)^{-\alpha}\, P_0^n \int \left(\frac{dP_n}{dQ^n}\right)^{\alpha} d\Pi(Q|B), \end{aligned}$$

Hellinger transforms "sub-factorize" over convex hulls of products

$$\begin{aligned} \sup_{P_n \in \mathrm{co}(V^n)} \int P_0^n\left(\frac{dP_n}{dQ^n}\right)^{\alpha} d\Pi(Q|B) &\leq \int \sup_{P_n \in \mathrm{co}(V^n)} P_0^n\left(\frac{dP_n}{dQ^n}\right)^{\alpha} d\Pi(Q|B) \\ &\leq \int \left(\sup_{P \in V} P_0\left(\frac{dP}{dQ}\right)^{\alpha}\right)^n d\Pi(Q|B). \end{aligned}$$

(see Le Cam (1986), or lemma 3.14 in Kleijn (2003))

A new consistency theorem

For $\alpha \in [0,1]$, model subsets B, W and a given P_0, define,

$$\pi_{P_0}(W, B; \alpha) = \sup_{P \in W} \sup_{Q \in B} P_0 \left(\frac{dP}{dQ} \right)^{\alpha}$$

Theorem 61 *Assume that* $P_0^n \ll P_n^{\Pi}$ *for all* $n \geq 1$. *Let* $V_1, \ldots, V_N$ *be model subsets. If there exist subsets* $B_1, \ldots, B_N$ *such that* $\Pi(B_i) > 0$,

$$\pi_{P_0}(\operatorname{co}(V_i), B_i) < 1$$

and $\sup_{Q \in B_i} P_0(dP/dQ) < \infty$ *for all* $P \in V_i$, *then,*

$$\Pi(V \mid X_1, \ldots, X_n) \xrightarrow{P_0\text{-a.s.}} 0$$

for any $V \subset \bigcup_{1 \leq i \leq N} V_i$.

With theorem 61 consistency in example 54 is demonstrated without problems.

Flexibility

Given a consistency question, *i.e.* given $\mathscr{P}$ and V, the approach is uncommitted regarding the prior and B. We look for neighbourhoods B of P_0 (of course such that $\sup_{Q \in B} P_0(dP/dQ) < \infty$ for all $P \in V$), which

(i) allow (uniform) control of $P_0(p/q)^\alpha$,

(ii) allow convenient choice of a prior such that $\Pi(B) > 0$.

The two requirements on B leave room for a trade-off between being 'small enough' to satisfy *(i)*, but 'large enough' to enable a choice for Π that leads to *(ii)*.

Relation with Schwartz's KL condition

Lemma 62 *Let $P_0 \in B \subset \mathscr{P}$ and $W \subset \mathscr{P}$ be given. Assume there is an $a \in (0,1)$ such that for all $Q \in B$ and $P \in W$, $P_0(dP/dQ)^a < \infty$. Then,*

$$\pi_{P_0}(W, B) < 1$$

if and only if,

$$\sup_{Q \in B} -P_0 \log \frac{dQ}{dP_0} < \inf_{P \in W} -P_0 \log \frac{dP}{dP_0}$$

Consistency in KL-divergence

Theorem 63 *Let* Π *be a Kullback-Leibler prior. Define* $V = \{P \in \mathscr{P} : -P_0 \log(dP/dP_0) \geq \epsilon\}$ *and assume that for some KL neighbourhood* B *of* P_0, $\sup_{Q \in B} P_0(dP/dQ) < \infty$ *for all* $P \in V$. *Also assume that* V *is covered by subsets* $V_1, \ldots, V_N$ *such that,*

$$\inf_{P \in \mathrm{co}(V_i)} -P_0 \log \frac{dP}{dP_0} > 0$$

for all $1 \leq i \leq N$. *Then,*

$$\Pi(\, -P_0 \log(dP/dP_0) < \epsilon \,|\, X_1, \ldots, X_n \,) \xrightarrow{P_0\text{-a.s.}} 1$$

Relation with priors that charge metric balls

Note that if we choose $\alpha = 1/2$,

$$
\begin{aligned}
P_0\left(\frac{p}{q}\right)^{1/2} &= \int \left(\frac{p_0}{q}\right)^{1/2} p_0^{1/2} p^{1/2}\, d\mu \\
&= \int p_0^{1/2}\, p^{1/2}\, d\mu + \int \left(\left(\frac{p_0}{q}\right)^{1/2} - 1\right) \left(\frac{p_0}{q}\right)^{1/2} \left(\frac{p}{q}\right)^{1/2} dQ \\
&\leq 1 - \frac{1}{2} H(P_0, P)^2 + H(P_0, Q) \left\| \frac{p_0}{q} \right\|_{2,Q}^{1/2} \left\| \frac{p}{q} \right\|_{2,Q}^{1/2}.
\end{aligned}
$$

So if $\|p/q\|_{2,Q}$ is bounded, a lower bound to $H(\mathrm{co}(V), P_0)$ and an upper bound for $H(Q, P_0)$ guarantee $\pi(\mathrm{co}(V), B; \frac{1}{2}) < 1$.

Borel priors of full support

Theorem 64 *Suppose that $\mathscr{P}$ is Hellinger totally bounded. Assume an $L > 0$ and a Hellinger ball B' centred on P_0 such that,*

$$\left\| \frac{p}{q} \right\|_{2,Q} = \left(\int \frac{p^2}{q} \, d\mu \right)^{1/2} < L, \quad \text{for all } P \in \mathscr{P} \text{ and } Q \in B'$$

If $\Pi(B) > 0$ for all Hellinger neighbourhoods of P_0, the posterior is Hellinger consistent, P_0-almost-surely.

Lemma 65 *If the KL divergence $\mathscr{P} \to \mathbb{R} : Q \mapsto -P \log(dQ/dP)$ is continuous, then a Borel prior of full support is a KL prior.*

Separable models and Barron's sieves

Theorem 66 *Let V be given. Assume that there are $K, L > 0$, submodels $(\mathscr{P}_n)_{n\geq 1}$ and a B with $\Pi(B) > 0$, such that,*

(i) there is a cover $V_1, \ldots, V_{N_n}$ for $V \cap \mathscr{P}_n$ of order $N_n \leq \exp(\frac{1}{2}Ln)$, such that for every $1 \leq i \leq N_n$,

$$\pi_{P_0}(\,\mathrm{co}(V_i), B\,) \leq e^{-L}$$

and $\sup_{Q\in B} P_0(dP/dQ) < \infty$ for all $P \in V_i$;

(ii) $\Pi(\mathscr{P} \setminus \mathscr{P}_n) \leq \exp(-nK)$ and,

$$\sup_{P\in V\setminus\mathscr{P}_n} \sup_{Q\in B} P_0\left(\frac{dP}{dQ}\right) \leq e^{\frac{K}{2}}$$

Then $\Pi(\,V \,|\, X_1, \ldots, X_n\,) \xrightarrow{P_0\text{-a.s.}} 0$.

A new theorem for separable models

Theorem 67 *Assume that* $P_0^n \ll P_n^\Pi$ *for all* $n \geq 1$. *Let* V *be a model subset with a countable cover* $V_1, V_2, \ldots$ *and* $B_1, B_2, \ldots$ *such that* $\Pi(B_i) > 0$ *and for* $P \in V_i$, *we have* $\sup_{Q \in B_i} P_0(dP/dQ) < \infty$. *Then,*

$$P_0^n \Pi(V|X_1, \ldots, X_n) \leq \sum_{i \geq 1} \inf_{0 \leq \alpha \leq 1} \frac{\Pi(V_i)^\alpha}{\Pi(B_i)^\alpha} \pi\Big(\mathrm{co}(V_i), B_i; \alpha\Big)^n.$$

Relation with Walker's condition

Corollary 68 *Assume that* $P_0^n \ll P_n^\Pi$ *for all* $n \geq 1$. *Let* V *be a subset with a countable cover* $V_1, V_2, \ldots$. *and a* B *such that* $\Pi(B) > 0$ *and for all* $i \geq 1$, $P \in V_i$, $\sup_{Q \in B} P_0(dP/dQ) < \infty$. *Also assume,*

$$\sup_{i \geq 1} \pi_{P_0}\Big(\mathrm{co}(V_i), B\Big) < 1$$

If the prior satisfies Walker's condition,

$$\sum_{i \geq 1} \Pi(V_i)^{1/2} < \infty$$

Then $\Pi(V|X_1, \ldots, X_n) \xrightarrow{P_0\text{-a.s.}} 0$.

Posterior rates of convergence

Theorem 69 *Assume that* $P_0^n \ll P_n^\Pi$ *for all* $n \geq 1$. *Let* (ϵ_n) *be s.t.* $\epsilon_n \downarrow 0$ *and* $n\epsilon_n^2 \to \infty$. *Define* $V_n = \{P \in \mathscr{P} : d(P, P_0) > \epsilon_n\}$, *submodels* $\mathscr{P}_n \subset \mathscr{P}$ *and subsets* B_n *s.t.* $\sup_{Q \in B_n} P_0(p/q) < \infty$ *for all* $P \in V_n$. *Assume that,*

(i) there is an $L > 0$ *such that* $V_n \cap \mathscr{P}_n$ *has a cover* $V_{n,1}, V_{n,2}, \ldots, V_{n,N_n}$ *of order* $N_n \leq \exp(\frac{1}{2}Ln\epsilon_n^2)$, *such that,*

$$\pi_{P_0}\Big(\mathrm{co}(V_{n,i}), B_n\Big) \leq e^{-Ln\epsilon_n^2}$$

for all $1 \leq i \leq N_n$.

(ii) there is a $K > 0$ *such that* $\Pi(\mathscr{P} \setminus \mathscr{P}_n) \leq e^{-Kn\epsilon_n^2}$ *and* $\Pi(B_n) \geq e^{-\frac{K}{2}n\epsilon_n^2}$, *while also,*

$$\sup_{P \in \mathscr{P} \setminus \mathscr{P}_n} \sup_{Q \in B_n} P_0\left(\frac{dP}{dQ}\right) < e^{\frac{K}{4}\epsilon_n^2}$$

Then $\Pi(\, P \in \mathscr{P} : d(P, P_0) > \epsilon_n \,|\, X_1, \ldots, X_n \,) \xrightarrow{P_0} 0$.

Posterior rates with Schwartz's KL priors

Theorem 70 *Let ϵ_n be such that $\epsilon_n \downarrow 0$ and $n\epsilon_n^2 \to \infty$. For $M > 0$, define $V_n = \{P \in \mathscr{P} : H(P_0, P) > M\epsilon_n\}$, $B_n = \{Q \in \mathscr{P} : -P_0 \log(dQ/dP_0) < \epsilon_n^2\}$. Assume that,*

(i) for all $P \in V_n$, $\sup\{P_0(dP/dQ) : Q \in B_n\} < \infty$

(ii) there is an $L > 0$, such that $N(\epsilon_n, \mathscr{P}, H) \leq e^{Ln\epsilon_n^2}$

(iii) there is a $K > 0$, such that for large enough $n \geq 1$,

$$\Pi\left(P \in \mathscr{P} : -P_0 \log \frac{dP}{dP_0} < \epsilon_n^2 \right) \geq e^{-Kn\epsilon_n^2}$$

then $\Pi(\, P \in \mathscr{P} : H(P, P_0) > M\epsilon_n \,|\, X_1, \ldots, X_n \,) \xrightarrow{P_0} 0$, for some $M > 0$.

With theorem 70 $\sqrt{n}$-consistency in the heavy-tailed example 57 obtains (for uniform priors on bounded intervals in $\mathbb{R}$).

Estimation of support boundary I: model

Model

Define $\Theta = \{(\theta_1, \theta_2) \in \mathbb{R}^2 : 0 < \theta_2 - \theta_1 < \sigma\}$ (for some $\sigma > 0$) and let H be a convex collection of Lebesgue probability densities $\eta : [0,1] \to [0,\infty)$ with a function $f : (0,a) \to \mathbb{R}$, $f > 0$ such that,

$$\inf_{\eta \in H} \min\left\{ \int_0^\epsilon \eta \, d\mu, \int_{1-\epsilon}^1 \eta \, d\mu \right\} \geq f(\epsilon), \quad (0 < \epsilon < a)$$

The semi-parametric model $\mathscr{P} = \{P_{\theta,\eta} : \theta \in \Theta, \eta \in H\}$,

$$p_{\theta,\eta}(x) = \frac{1}{\theta_2 - \theta_1} \eta\left(\frac{x - \theta_1}{\theta_2 - \theta_1}\right) 1_{\{\theta_1 \leq x \leq \theta_2\}}.$$

Question

We are interested in marginal consistency for θ. Define the pseudo-metric $d : \mathscr{P} \times \mathscr{P} \to [0,\infty)$,

$$d(P_{\theta,\eta}, P_{\theta',\eta'}) = \max\left\{|\theta_1 - \theta_1'|, |\theta_2 - \theta_2'|\right\}.$$

We want posterior consistency with $V = \{P_{\theta,\eta} : d(P, P_0) \geq \epsilon\}$.

Estimation of support boundary II: construction

Lemma 71 *Suppose that* $P_0(p/q) < \infty$. *Then*

$$P_0(p/q)^\alpha|_{\alpha=0} = P_0(p > 0), \quad P_0(p/q)^\alpha|_{\alpha=1} = \int \frac{p_0}{q} 1_{\{p_0>0\}}\, dP.$$

Take $B = \{Q : \|(p_0/q) - 1\|_\infty < \delta\}$,

$$\inf_{0\le\alpha\le 1} P_0\left(\frac{p}{q}\right)^\alpha \le (1+\delta)\min\Big\{P_0(p>0), P(p_0>0)\Big\}$$

The supports of p and p_0 differ by an interval of length $\ge \epsilon$,

$$\min\Big\{P_0(p>0), P(p_0>0)\Big\} \le 1 - \frac{f(\epsilon)}{\sigma}.$$

Conclude: for every $\epsilon, \delta > 0$,

$$\sup_{Q\in B}\sup_{P\in V}\inf_{0\le\alpha\le 1} P_0\left(\frac{p}{q}\right)^\alpha \le (1+\delta)\left(1 - \frac{f(\epsilon)}{\sigma}\right) < 1.$$

Estimation of support boundary III: theorem

Theorem 72 *Let* $\Theta = \{(\theta_1, \theta_2) \in \mathbb{R}^2 : 0 < \theta_2 - \theta_1 < \sigma\}$ *(for some* $\sigma > 0$*) and convex* H *with associated* f *be given. Let* Π *be a prior on* $\Theta \times H$ *such that,*

$$\Pi\Big(Q : \|(p_0/q) - 1\|_\infty < \delta\Big) > 0,$$

for all $\delta > 0$*. If* $X_1, X_2, \ldots$ *form an i.i.d.-*P_0 *sample, where* $P_0 = P_{\theta_0,\eta_0}$*, then,*

$$\Pi\Big(\|\theta - \theta_0\| < \epsilon \,\Big|\, X_1, \ldots, X_n \Big) \xrightarrow{P_0\text{-a.s.}} 1,$$

for every $\epsilon > 0$*.*

Remark 73 *The* σ*-restriction on* $\theta_1 - \theta_2$ *can be eliminated with theorem 66.*

Lecture V

Errors-in-variables regression

This lecture presents an analysis of posterior behaviour in the non-parametric structural errors-in-variables regression model, which combines smoothness classes with mixture modelling. The goal is to demonstrate some of the standard techniques that go into a proof of posterior convergence, including rates of convergence. Central in the discussion are nets and entropy numbers of parametrizing spaces, allowing for the construction of tests as well as suitable priors.

Structural errors-in-variables regression

Observe *i.i.d.* pairs $(X_i, Y_i) \in \mathbb{R}^2$ $(i \geq 1)$, assumed to obey

$$X = Z + e_1,$$
$$Y = aZ + b + e_2,$$

(e_1, e_2) and Z are independent

Compare with the simpler model

$$Y = aX + b + e$$

where errors in X bias estimation of a towards 0.

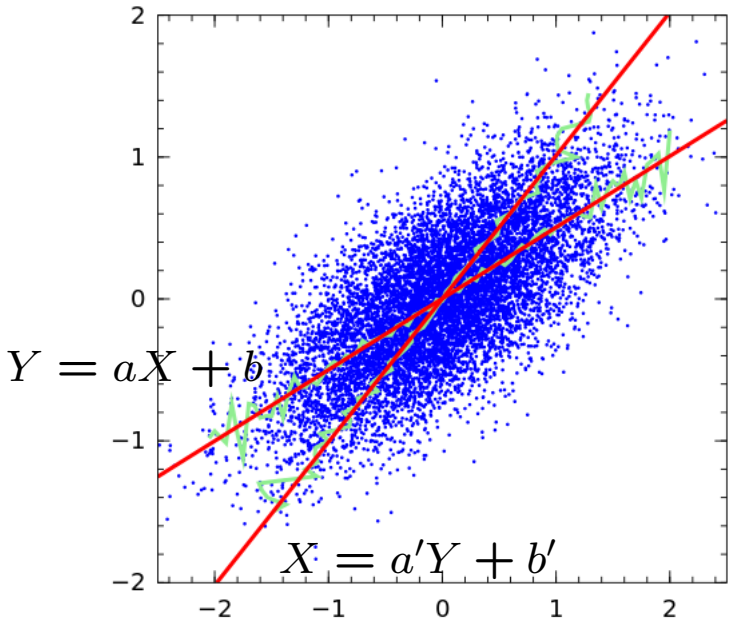


Fig. X Regression dilution regression lines for $Y = aX + b + e$ and $X = a'Y + b' + e$

Errors-in-variables references

M. Kendall, A. Stuart, *The advanced theory of statistics, Vol. 2*, (4th edition), Griffin, London (1979).

T. Anderson, *Estimating linear statistical relationships*, Ann. Statist. 12 (1984), 1–45.

P. Bickel, Y. Ritov, *Efficient estimation in the errors in variables model*, Ann. Statist. 15 (1987), 513–540.

J. Fan, Y. Truong, *Nonparametric regression with errors in variables*, Ann. Statist. 21 (1993), 1900–1925.

A. van der Vaart, *Efficient maximum-likelihood estimation in semiparametric mixture models*, Ann. Statist. 24 (1996), 862–878.

M. Taupin, *Semi-parametric estimation in the nonlinear structural errors-in-variables model*, Ann. Statist. 29 (2001), 66–93.

Model definition

Structural errors-in-variables model

$$X = Z + e_1, \quad Y = f(Z) + e_2, \tag{12}$$

$(e_1, e_2) \sim \Phi_\sigma \times \Phi_\sigma$, $Z \sim F$ (e_1, e_2) and Z are independent
Parameter $(\sigma, f, F) \in I \times \mathscr{F} \times D$, where

$I \subset (0, \infty)$ is a closed interval
$D = \mathscr{D}[-A, A]$ with Stieltjes functions F.
$\mathscr{F}$ bounded family of continuous functions $f : [-A, A] \to [-B, B]$.

Density for $P_{\sigma,f,F}$

$$p_{\sigma,f,F}(x, y) = \int_{-A}^{A} \varphi_\sigma\Big(x - z\Big)\varphi_\sigma\Big(y - f(z)\Big)\, dF(z), \tag{13}$$

Smoothness classes $\mathscr{F}$

$\mathscr{F} \subset (C_B[-A, A], \|\cdot\|)$ all continuous functions $f : [-A, A] \to [-B, B]$

$\text{Lip}_M(\alpha)$, $M > 0$, $0 < \alpha \leq 1$ all $f \in C_B[-A, A]$ with

$$\left|f(z) - f(z')\right| \leq M|z - z'|^\alpha,$$

$D_{\alpha,M}(q)$ $M > 0$, $0 < \alpha \leq 1$, $q \geq 1$ all q-diff $f \in C_B[-A, A]$ with

$$\left|f^{(q)}(z) - f^{(q)}(z')\right| \leq M|z - z'|^\alpha,$$

$\mathscr{F}_\Theta = \{f_\theta : \theta \in \Theta\} \subset \text{Lip}_M(\alpha)$, with bounded, open $\Theta \subset \mathbb{R}^k$ and

$$\|f_{\theta_1} - f_{\theta_2}\| \leq L\|\theta_1 - \theta_2\|^\rho_{\mathbb{R}^k},$$

Take first two cases together, in $(C_\beta[-A, A], \|\cdot\|_\beta)$

$$\|f\|_\beta = \max_{k \leq \lfloor\beta\rfloor} \|f^{(k)}\| + \sup_{z_1, z_2} \frac{\left|f^{(\lfloor\beta\rfloor)}(z_1) - f^{(\lfloor\beta\rfloor)}(z_2)\right|}{|z_1 - z_2|^{\beta - \lfloor\beta\rfloor}},$$

and $C_{\beta,L}[-A, A] = \{f \in C_\beta[-A, A] : \|f\|_\beta \leq L\}$

Recall the GGV theorem

Theorem 74 *(Ghosal, Ghosh and van der Vaart, 2000)*
Let (ϵ_n) *be such that* $\epsilon_n \downarrow 0$ *and* $n\epsilon_n^2 \to \infty$. *Let* $C > 0$ *and* $\mathscr{P}$ *be such that, for large enough* n,

(i) $N(\epsilon_n, \mathscr{P}, H) \leq e^{-n\epsilon_n^2}$

(ii) the prior Π *is a GGV-prior, i.e.*

$$\Pi\left(P \in \mathscr{P} \,:\, -P_0 \log \frac{dP}{dP_0} < \epsilon_n^2,\, P_0\left(\log \frac{dP}{dP_0}\right)^2 < \epsilon_n^2 \right) \geq e^{-Cn\epsilon_n^2}$$

Then $\Pi(\, P \in \mathscr{P} \,:\, H(P, P_0) > M\epsilon_n \,|\, X_1, \ldots, X_n \,) \xrightarrow{P_0} 0$ *for some* $M > 0$.

Entropy calculation (I)

Note

$$
\begin{aligned}
& H(P_{\sigma,f,F}, P_{\tau,g,F'}) \\
& \qquad \le H(P_{\sigma,f,F}, P_{\tau,f,F}) + H(P_{\tau,f,F}, P_{\tau,g,F}) + H(P_{\tau,g,F}, P_{\tau,g,F'}) \\
& \qquad \le \|p_{\sigma,f,F} - p_{\tau,f,F})\|_{1,\mu}^{1/2} + K\|f-g\| + \|p_{\tau,g,F} - p_{\tau,g,F'}\|_{1,\mu}^{1/2} \\
& \qquad \le K_1|\sigma-\tau|^{1/2} + K_2\|f-g\| + \|p_{\tau,g,F} - p_{\tau,g,F'}\|_{1,\mu}^{1/2},
\end{aligned}
\tag{14}
$$

If I' and D' are ϵ^2-nets in I and D, and $\mathscr{F}'$ is an ϵ-net in $\mathscr{F}$, then

$$
\left\{P_{\sigma,f,F} : \sigma \in I', f \in \mathscr{F}', F \in D'\right\}
$$

is a $K\epsilon$-net in $\mathscr{P}$ for the Hellinger metric (for some $K > 0$)

Entropy calculation (II)

Therefore,

$$\begin{aligned}
&\log N(K\epsilon^{\alpha/2}, \mathscr{P}, H) \\
&\qquad \leq \log N(\epsilon^{\alpha}, I, |\,.\,|) + \log N(\epsilon^{\alpha/2}, \mathscr{F}, \|\,.\,\|) + \log N(\mathscr{Q}^{\epsilon}) \\
&\qquad \leq L''' \log \frac{1}{\epsilon} + \log N(\epsilon^{\alpha/2}, \mathscr{F}, \|\,.\,\|) + L''\left(\log \frac{1}{\epsilon}\right)^3
\end{aligned}$$

The Hellinger covering number is bounded by two contributions

$$\log N(\epsilon, \mathscr{P}, H) \leq L_0\left(\log \frac{1}{\epsilon}\right)^3 + \log N(L\epsilon, \mathscr{F}, \|\,.\,\|), \qquad (15)$$

The 'parametric' case

Lemma 75 *If there exists a constant* $L_1 > 0$ *such that:*

$$\log N(\epsilon, \mathscr{F}, \|\,.\,\|) \leq L_1 \left(\log \frac{1}{\epsilon}\right)^3, \tag{16}$$

for small enough $\epsilon > 0$*, then the entropy condition of the GGV thm is satisfied with:*

$$\epsilon_n = n^{-1/2} (\log n)^{3/2}, \tag{17}$$

for large enough n*.*

Proof of lemma 75

$\log N(\epsilon, \mathscr{P}, H)$ is upper bounded by the first term in (15) (with a larger choice for L_0). With $\epsilon_n = n^{-1/2}(\log n)^{3/2}$, $\epsilon_n \downarrow 0$ and $n\epsilon_n^2 \to \infty$. Note $\epsilon_n \geq 1/n$ for large enough n so that for some $L > 0$,

$$\log N(\epsilon_n, \mathscr{F}, \|\,.\,\|) \leq \log N(1/n, \mathscr{F}, \|\,.\,\|) \leq L(\log n)^3,$$

while $n\epsilon_n^2 = (\log n)^3$, so ϵ_n satisfies the entropy condition of the GGV thm.

The 'non-parametric' case

Lemma 76 *For an errors-in-variables model $\mathscr{P}$ with $\mathscr{F} = C_{\beta,M}[-A, A]$, the entropy condition of the GGV thm is satisfied by*

$$\epsilon_n = n^{-\frac{\beta}{2\beta+1}}, \tag{18}$$

Jackson's approximation theorem (Jackson (1930)) if $f \in \mathsf{Lip}_M(\alpha)$ there exists an n-th order polynomial p_n such that

$$\|f - p_n\| \leq \frac{K}{n^\alpha} \tag{19}$$

if $f \in D_{\alpha,M}(q)$, there exists n-th order polynomial p_n such that

$$\|f - p_n\| \leq \frac{K'}{n^{q+\alpha}}, \tag{20}$$

Proof of lemma 76

Lemma 77 *(Kolmogorov, Tikhomirov (1961)) Let $\beta > 0$, $M > 0$ be given. There is a $K > 0$ such that,*

$$\log N\Big(\epsilon, C_{\beta,M}[-A, A], \|\cdot\|\Big) \leq K\left(\frac{1}{\epsilon}\right)^{1/\beta}, \tag{21}$$

for all $\epsilon > 0$.

Therefore

$$\log N(\epsilon, \mathscr{P}, H)) \leq \frac{K}{\epsilon^{1/\beta}},$$

for small enough ϵ. The sequence $\epsilon_n = n^{-\frac{\beta}{2\beta+1}}$ satisfies $\epsilon_n \downarrow 0$ and $n\epsilon_n^2 \to \infty$ and

$$\log N(\epsilon_n, \mathscr{P}, H) \leq K\, n^{1/(2\beta+1)} = Kn \cdot n^{-\frac{2\beta}{2\beta+1}} = Kn\epsilon_n^2,$$

for large enough n.

Definition of the prior Π

Choose priors Π_I, $\Pi_{\mathscr{F}}$ and Π_D on the parameter spaces I, $\mathscr{F}$ and D

$$\Pi(A) = (\Pi_I \times \Pi_{\mathscr{F}} \times \Pi_D)(P^{-1}(A))$$

where $P : I \times \mathscr{F} \times D \to \mathscr{P} : (\sigma, f, F) \mapsto P_{\sigma,f,F}$.

Lemma 78 *View $(I, |\cdot|)$ and $(\mathscr{F}, \|\cdot\|)$ and D with Prokhorov's weak topology as metric spaces. Then the map $\widehat{p} : I \times \mathscr{F} \times D \to L_1(\mu)$ is continuous for the product topology.*

Π_I any prior with a continuous and strictly positive Lebesgue density on I. Π_D Dirichlet with base measure with a continuous and strictly positive Lebesgue density on $[-A, A]$.

Proof of lemma 78

Let $(\sigma_n, f_n, F_n) \to (\sigma, f, F)$ be given. From (14) we know

$$\left\| p_{\sigma_n, f_n, F_n} - p_{\sigma, f, F} \right\|_{1,\mu} \leq K_1 |\sigma_n - \sigma| + K_2 \|f_n - f\| + \left\| p_{\sigma, f, F_n} - p_{\sigma, f, F} \right\|_{1,\mu},$$

for some constants $K_1, K_2 > 0$.

Since $F_n \xrightarrow{w.} F$, f is cont and φ_σ is bnd cont,

$$\int_{[-A,A]} \varphi_\sigma\big(x-z\big)\varphi_\sigma\big(y-f(z)\big)\, dF_n(z) \to \int_{[-A,A]} \varphi_\sigma\big(x-z\big)\varphi_\sigma\big(y-f(z)\big)\, dF(z)$$

So pointwise $p_{\sigma,f,F_n}(x,y) \to p_{\sigma,f,F}(x,y)$ and Scheffé's lemma gives

$$\left\| p_{\sigma, f, F_n} - p_{\sigma, f, F} \right\|_{1,\mu} \to 0,$$

Prior-mass lower bound in the GGV thm

Define

$$B\Big(\epsilon\Big) = \Big\{ P \in \mathscr{P} \,:\, -P_0 \log \frac{dP}{dP_0} < \epsilon^2,\, P_0\Big(\log \frac{dP}{dP_0}\Big)^2 < \epsilon^2 \Big\}$$

Theorem 79 *Suppose that* $\mathscr{F}$ *is* $\mathrm{Lip}_M(\alpha)$, $D_{\alpha,M}(q)$ *or* $\mathscr{F}_{\ominus}$. *Define a prior* Π *on* $\mathscr{P}$ *of product form. Then there exist constants* $K, c, C > 0$ *such that*

$$\Pi\Big(B\Big(K\delta \log(1/\delta)\Big)\Big) \geq C \exp\Big(-c\Big(\log(1/\delta)\Big)^3\Big)\, \Pi_{\mathscr{F}}\Big(f \in \mathscr{F} : \|f - f_0\| \leq \delta\Big),$$

for small enough δ.

Net priors on regression classes

$\mathscr{F}$ equicontinuous Let $(a_m)_{m\geq 1}$, $a_m \downarrow 0$ be given. For $m \geq 1$ there is a a_m-net $\{f_i \in \mathscr{F} : i = 1, \ldots, N_m\}$ in $\mathscr{F}$, where $N_m = N(a_m, \mathscr{F}, \|\,.\,\|) < \infty$. Define finitely supported probability measures Π_m,

$$\Pi_m(A) = \sum_{i=1}^{N_m} \frac{1}{N_m} \delta_{f_i}(A).$$

With any (b_m) such that $b_m \geq 0$ and $\sum_{m=1}^{\infty} b_m = 1$, also define

$$\Pi_{\mathscr{F}}(A) = \sum_{m=1}^{\infty} b_m \Pi_m(A) \tag{22}$$

Note for every $f_0 \in \mathscr{F}$ and all $\delta > 0$, we have:

$$\Pi_{\mathscr{F}}\Big(\|f - f_0\| \leq \delta\Big) \geq \frac{b_m}{N_m},$$

if $a_m \leq \delta$, *i.e.* for all m large enough.

Lower bounds for net prior mass in small balls

Lemma 80 *Let $\beta > 0$ and $M > 0$ be given and take $\mathscr{F} = C_{\beta,M}[-A, A]$. There exists a net prior $\Pi_{\mathscr{F}}$ and a constant $K > 0$ such that*

$$\log \Pi_{\mathscr{F}}\Big(f \in \mathscr{F} : \|f - f_0\| \leq \delta\Big) \geq -K \frac{1}{\delta^{1/\beta}}, \tag{23}$$

for small enough δ.

Lemma 81 *Let parametric $\mathscr{F} = \mathscr{F}_\Theta$ be given. If Π_Θ is Lebesgue-abs-cont with continuous and strictly positive density then for a constant $K > 0$,*

$$\log \Pi_{\mathscr{F}}\Big(f \in \mathscr{F} : \|f - f_0\| \leq \epsilon\Big) \geq -K \log \frac{1}{\epsilon}, \tag{24}$$

for small enough $\epsilon > 0$.

Proof of lemma 80

Choose $a_m = m^{-\beta}$. Then N_m satisfies for some constant $K' > 0$:

$$\log N_m = \log N(a_m, \mathscr{F}, \|\,.\,\|) \leq K' a_m^{-1/\beta} = K' m,$$

c.f. lemma 77. Choose $b_m = (1/2)^m$. Let M be an integer s.t.

$$\frac{1}{\delta^{1/\beta}} \leq M \leq \frac{1}{\delta^{1/\beta}} + 1.$$

Then for all $m \geq M$, $a_m \leq \delta$ and the net prior $\Pi_{\mathscr{F}}$ satisfies:

$$\begin{aligned}\Pi_{\mathscr{F}}\Big(f \in \mathscr{F} \colon \|f - f_0\| \leq \delta\Big) &\geq \sum_{m \geq M} b_m \Pi_m\Big(\|f - f_0\| \leq \delta\Big) \\ &\geq \sum_{m \geq M} \left(\frac{e^{-K'}}{2}\right)^m \geq \tfrac{1}{2} e^{-K' M} \geq \tfrac{1}{2} e^{-K'(\delta^{-1/\beta}+1)} \geq \tfrac{1}{2} e^{-2K' \delta^{-1/\beta}},\end{aligned} \tag{25}$$

for small enough δ.

Rates of posterior convergence

Theorem 82 *The posterior converges at rate* ϵ_n.

$\mathscr{F} = \mathrm{Lip}_M(\alpha)$ *with a net prior, prior-mass in neighbourhoods of* f_0 *determines* ϵ_n,

$$\epsilon_n = n^{-\frac{\alpha}{2\alpha+1}} (\log n)^{\frac{1}{2\alpha}}.$$

$\mathscr{F} = D_{\alpha,M}(q)$ *with a net prior, prior-mass in neighbourhoods of* f_0 *determines* ϵ_n,

$$\epsilon_n = n^{-\frac{q+\alpha}{2q+2\alpha+1}} (\log n)^{\frac{1}{2q+2\alpha}}.$$

$\mathscr{F} = \mathscr{F}_{\ominus}$ *with Lebesgue prior with continuous and strictly positive density, prior-mass in neighbourhoods of* F_0 *determines* ϵ_n

$$\epsilon_n = n^{-1/2} (\log n)^{3/2}.$$

Lecture VI

Tests and posteriors

The existence of Bayesian test sequences implies convergence of the posterior distribution. By implication any distinction between model subsets that is asymptotically testable is also expressed through posterior convergence. In a Bayesian sense, this leads to an equivalence that implies Doob's theorem. By contrast frequentist convergence is by no means settled, and counterexamples abound, while Schwartz's theorem formulates a very sharp sufficient condition.

arxiv:1611.08444 (MATH.ST)

The i.i.d. consistency theorems (I)

Theorem 83 *(Bayesian, Doob (1948))*
Assume that $X^n = (X_1, \ldots, X_n)$ *are i.i.d. Let* $\mathscr{P}$ *and* $\mathscr{X}$ *be Polish spaces and let* Π *be a Borel prior. Then the posterior is consistent at* P*, for* Π*-almost-all* $P \in \mathscr{P}$

Example 84 *For some* $Q \in \mathscr{P}$*, take* $\Pi = \delta_Q$*. Then* $\Pi(\cdot|X^n) = \delta_Q$ *as well,* P_n^Π*-almost-surely. If* $X_1, \ldots, X_n \sim P_0^n$ *(require* $P_0^n \ll P_n^\Pi = Q^n$*), the posterior is not frequentist consistent.*

Non-trivial counterexamples are due to Schwartz (1961) and Freedman (1963,1965,. . .)

The i.i.d. consistency theorems (II)

Theorem 85 *(Frequentist, Schwartz (1965))*
Let $X_1, X_2, \ldots$ *be i.i.d.-*P_0 *for some* $P_0 \in \mathscr{P}$. *If,*

(i) For every nbd U *of* P_0, *there are* $\phi_n : \mathscr{X}_n \to [0,1]$, *s.t.*

$$P_0^n \phi_n = o(1), \qquad \sup_{Q \in U^c} Q^n (1 - \phi_n) = o(1), \tag{26}$$

(ii) and Π *is a Kullback-Leibler prior, i.e. for all* $\delta > 0$,

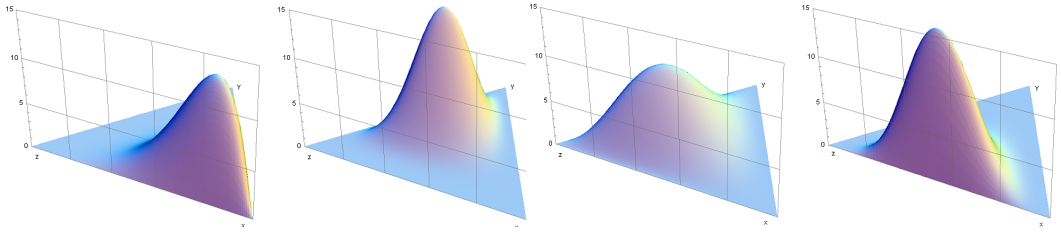
$$\Pi\left(P \in \mathscr{P} \,:\, -P_0 \log \frac{dP}{dP_0} < \delta \right) > 0, \tag{27}$$

then $\Pi(U|X^n) \xrightarrow{P_0\text{-a.s.}} 1$.

The Dirichlet process

Definition 86 *(Dirichlet distribution)*
A $p = (p_1, \ldots, p_k)$ $p_l \geq 0$ *and* $\sum_l p_l = 1$ *is Dirichlet distributed with parameter* $\alpha = (\alpha_1, \ldots, \alpha_k)$, $p \sim D_\alpha$, *if it has density*

$$f_\alpha(p) = C(\alpha) \prod_{l=1}^{k} p_l^{\alpha_l - 1}$$



Definition 87 *(Dirichlet process, Ferguson 1973,1974)*
Let μ *be a finite Borel msr on* $(\mathscr{X}, \mathscr{B})$. *The Dirichlet process* $P \sim D_\mu$ *is defined by,*

$$\Big(P(A_1), \ldots, P(A_k)\Big) \sim D_{(\mu(A_1), \ldots, \mu(A_k))}$$

The i.i.d. consistency theorems (III)

Theorem 88 *(Frequentist, Dirichlet consistency)*
Let $X_1, X_2, \ldots$ be an i.i.d.-sample from P_0 If Π is a Dirichlet prior D_α with finite α such that $\text{supp}(P_0) \subset \text{supp}(\alpha)$, the posterior is consistent at P_0 in Prohorov's weak topology

Remark 89 *(Freedman (1963))*
Dirichlet priors are tailfree: if A' refines A and $A'_{i1} \cup \ldots \cup A'_{il_i} = A_i$, then $(P(A'_{i1}|A_i), \ldots, P(A'_{il_i}|A_i) : 1 \leq i \leq k)$ is independent of $(P(A_1), \ldots, P(A_k))$.

Remark 90 *$X^n \mapsto \Pi(P(A)|X^n)$ is $\sigma_n(A)$-measurable where $\sigma_n(A)$ is generated by products of the form $\prod_{i=1}^n B_i$ with $B_i = \{X_i \in A\}$ or $B_i = \{X_i \notin A\}$.*

A posterior concentration inequality (I)

Lemma 91 *Let $(\mathscr{P}, \mathscr{G})$ be given. For any prior Π, any test function ϕ and any $B, V \in \mathscr{G}$,*

$$\int_B P\Pi(V|X)\, d\Pi(P) \leq \int_B P\phi\, d\Pi(P) + \int_V Q(1-\phi)\, d\Pi(Q)$$

Definition 92 *For $B \in \mathscr{G}$ such that $\Pi_n(B) > 0$, the local prior predictive distribution is defined, for every $A \in \mathscr{B}_n$,*

$$P_n^{\Pi|B}(A) = \int P_{\theta,n}(A)\, d\Pi_n(\theta|B) = \frac{1}{\Pi(B)} \int_B P_{\theta,n}(A)\, d\Pi_n(\theta).$$

Corollary 93 *Consequently, for any sequences (Π_n), (B_n), (V_n) such that $B_n \cap V_n = \varnothing$ and $\Pi_n(B_n) > 0$, we have,*

$$\begin{aligned} P_n^{\Pi|B_n}\Pi(V_n|X^n) &:= \int P_{\theta,n}\Pi(V_n|X^n)\, d\Pi_n(\theta|B_n) \\ &\leq \frac{1}{\Pi_n(B_n)}\left(\int_{B_n} P_{\theta,n}\phi_n\, d\Pi_n(\theta) + \int_{V_n} P_{\theta,n}(1-\phi_n)\, d\Pi_n(\theta)\right) \end{aligned}$$

Proof

Disintegration: for all $A \in \mathscr{B}^n$ and $V \in \mathscr{G}$,

$$\int_{\mathscr{X}} 1_A(X)\Pi(V|X)\, dP^{\Pi} = \int_V \int_{\mathscr{X}} 1_A(X)\, dP\, d\Pi(P)$$

So for any $\mathscr{B}^n$-measurable, simple $f(X) = \sum_{j=1}^J c_j\, 1_{A_j}(X)$,

$$\int_{\mathscr{X}} f(X)\Pi(V|X)\, dP^{\Pi} = \int_V \int_{\mathscr{X}} f(X)\, dP\, d\Pi(P)$$

Taking monotone limits, we see this equality also holds for any positive, measurable $f : \mathscr{X} \to [0, \infty]$. In particular, with $f(X) = (1 - \phi(X))$,

$$\int_{\mathscr{P}} P\Big((1 - \phi(X))\Pi(V|X)\Big)\, d\Pi(P) = \int_V P(1 - \phi(X))\, d\Pi(P)$$

Proof

Since $B \subset \mathscr{P}$ and the integrand is positive,

$$\int_B P\Big((1-\phi)(X)\Pi(V|X)\Big)\, d\Pi(P)$$
$$\leq \int_{\mathscr{P}} P\Big((1-\phi(X))\Pi(V|X)\Big)\, d\Pi(P) = \int_V P(1-\phi(X))\, d\Pi(P)$$

bring the 2nd term on the *l.h.s.* to the *r.h.s.* and divide by $\Pi(B) > 0$,

$$\int P\Pi(V|X)\, d\Pi(P|B)$$
$$\leq \frac{1}{\Pi(B)}\left(\int_B P\phi(X)\Pi(V|X)\, d\Pi(P) + \int_V P(1-\phi)(X)\, d\Pi(P)\right)$$
$$\leq \frac{1}{\Pi(B)}\left(\int_B P\phi(X)\, d\Pi(P) + \int_V P(1-\phi)(X)\, d\Pi(P)\right)$$

Martingale convergence

Proposition 94 *Let* $(\Theta, \mathscr{G}, \Pi)$ *be given. For any* $B, V \in \mathscr{G}$*, the following are equivalent,*

(i) There exist Bayesian tests (ϕ_n) *for* B *versus* V*;*

(ii) There exist tests (ϕ_n) *such that,*

$$\int_B P_{\theta,n}\phi_n \, d\Pi(\theta) + \int_V P_{\theta,n}(1-\phi_n) \, d\Pi(\theta) \to 0,$$

(iii) For Π*-almost-all* $\theta \in B$*,* $\eta \in V$*,*

$$\Pi(V|X^n) \xrightarrow{P_{\theta,n}} 0, \quad \Pi(B|X^n) \xrightarrow{P_{\eta,n}} 0$$

Remark 95 *Interpretation distinctions between model subsets are Bayesian testable, iff they are picked up by the posterior asymptotically, iff, the Bayes factor for B versus V is consistent*

Proof

Condition *(i)* implies *(ii)* by dominated convergence. Assume *(ii)* and note that by the previous lemma,

$$\int P^n \Pi(V|X^n)\, d\Pi(P|B) \to 0.$$

Martingale convergence (in $L^1(\mathscr{X}^\infty \times \mathscr{P})$) implies that there is a $g : \mathscr{X}^\infty \to [0,1]$ such that,

$$\int P^\infty \left| \Pi(V|X^n) - g(X^\infty) \right| d\Pi(P|B) \to 0,$$

So $\int P^\infty g\, d\Pi(P|B) = 0$, so $g = 0$, P^∞-almost-surely for Π-almost-all $P \in B$. Using martingale convergence again (now in $L^\infty(\mathscr{X}^\infty \times \mathscr{P})$), conclude $\Pi(V|X^n) \to 0$ P^∞-almost-surely for Π-almost-all $P \in B$, *i.e.* *(iii)* follows.

Choose $\phi(X^n) = \Pi(V|X^n)$ to conclude that *(i)* follows from *(iii)*.

Prior-almost-sure consistency

Corollary 96 *Let Hausdorff completely regular Θ with Borel prior Π be given. Then the following are equivalent,*

(i) for Π-almost-all $\theta \in \Theta$ and any nbd U of θ there exist a msb $B \subset U$ with $\Pi(B) > 0$ and Bayesian tests (ϕ_n) for B vs $V = \Theta \setminus U$,

(ii) the posterior is consistent at Π-almost-all $\theta \in \Theta$.

Remark 97 (Doob (1948))
Let $\mathscr{P}$ be a Polish space and assume that all $P \mapsto P^n(A)$ are Borel measurable. Then, for any prior Π, any Borel set $V \subset \mathscr{P}$ is Bayesian testable versus $\mathscr{P} \setminus V$.

Corollary 98 *(More than) Doob's 1948 theorem (see theorem 183)*

Examples: prior-almost-sure inconsistency (I)

Example 99 (Freedman (1963))

Let $X_1, X_2, \ldots$ *be i.i.d.positive integers.*

$\Lambda \subset \ell^1$ *the space of all prob dist on* $\mathbb{N}$ $(P_0 \in \Lambda)$: $p(i) = P(\{X = i\})$.

Schur's property Total-variational and weak topologies on Λ *equivalent*

$$P \to Q \text{ means } p(i) \to q(i) \text{ for all } i \geq 1.$$

Goal is a prior with P_0 *in its support while posterior concentrates around some* $Q \in \Lambda \setminus \{P_0\}$.

Examples: prior-almost-sure inconsistency (II)

Consider sequences (P_m) and (Q_m) such that

$$Q_m \to Q, \quad P_m \to P_0, \text{ as } m \to \infty$$

Prior Π places masses $\alpha_m > 0$ at P_m and $\beta_m > 0$ at Q_m $(m \geq 1)$, so that P_0 lies in the support of Π.

First step construct (P_0-dependently) Q_m, leads to a posterior with,

$$\frac{\Pi(\{Q_m\}|X^n)}{\Pi(\{Q_{m+1}\}|X^n)} \xrightarrow{P_0\text{-a.s.}} 0,$$

forcing all posterior mass that resides in $\{Q_m : m \geq 1\}$ into arbitrary tails $\{Q_m : m \geq M\}$, *i.e.* arbitrarily small neighbourhoods of Q.

Examples: prior-almost-sure inconsistency (III)

Second step choose (P_m) and (α_m) such that posterior mass in $\{P_m : m \geq 1\}$ also accumulates in tails.

But if ratios α_m/β_m decrease to zero very fast with m,

$$\frac{\Pi(\{P_m : m \geq M\}|X^n)}{\Pi(\{Q_m : m \geq M\}|X^n)} < \epsilon,$$

P_0-a.s. for large enough M.

Conclusion for every neighbourhood U_Q of Q,

$$\Pi(U_Q|X^n) \xrightarrow{P_0\text{-a.s.}} 1,$$

so the posterior is inconsistent.

Remark 100 *Other choices of the weights* (α_m) *with more prior mass in the tails* do have *consistent posteriors*.

Examples: prior-almost-sure inconsistency (IV)

Objection knowledge of P_0 is required to construct the prior (unfortunate but of no concern in any generic sense).

$\pi(\Lambda)$ the space of all Borel distributions on Λ. Since Λ is Polish, so are $\pi(\Lambda)$ and $\Lambda \times \pi(\Lambda)$.

Theorem 101 *(Freedman (1965))*
Let $X_1, X_2, \ldots$ be i.i.d. integers, Endow $\pi(\Lambda)$ with Prohorov's weak topology. The set of $(P_0, \Pi) \in \Lambda \times \pi(\Lambda)$ such that for all open $U \subset \Lambda$,

$$\limsup_{n \to \infty} P_0^n \Pi(U | X^n) = 1,$$

is residual.

The set of $(P_0, \Pi) \in \Lambda \times \pi(\Lambda)$ for which the limiting behaviour of the posterior is acceptable to the frequentist, is meagre in $\Lambda \times \pi(\Lambda)$.

Examples: prior-almost-sure inconsistency (V)

The proof relies on the following (see also Le Cam (1986), 17.7)

for every $k \geq 1$ Λ_k is all prob dist P on $\mathbb{N}$ with $P(X = k) = 0$

$\Lambda_0 = \cup_{k \geq 1} \Lambda_k$ Pick $P_0, Q \in \Lambda \setminus \Lambda_0$ such that $P_0 \neq Q$.

Place a prior Π_0 on Λ_0 and choose $\Pi = \frac{1}{2}\Pi_0 + \frac{1}{2}\delta_Q$.

Because Λ_0 is dense prior Π has full support

Examples: prior-almost-sure inconsistency (VI)

P_0 has full support in $\mathbb{N}$ so for every $k \in \mathbb{N}$, $P_0^\infty(\exists_{m \geq 1} : X_m = k) = 1$

If we observe $X_m = k$ likelihoods for $n \geq m$ equal zero on Λ_k so

$$\Pi(\Lambda_k | X^n) = 0$$

for all $n \geq m$, P_0^∞-almost-surely.

Freedman shows that this implies

$$\Pi(\Lambda_0 | X^n) \xrightarrow{P_0\text{-a.s.}} 0$$

forcing all posterior mass onto the point $\{Q\}$.

$$\Pi(\{Q\} | X^n) \xrightarrow{P_0\text{-a.s.}} 1$$

Lecture VII

Frequentist validity of Bayesian limits

Remote contiguity is the extra property that lends validity to Bayesian limits for the frequentist. It is required that the prior is such that locally-averaged likelihoods are indistinguishable from the likelihoods associated with true distributions of the data in a specific way that generalizes Le Cam's property of contiguity.

arxiv:1611.08444 (MATH.ST)

Le Cam's inequality

Definition 102 *For $B \in \mathscr{G}$ such that $\Pi_n(B) > 0$, the local prior predictive distribution is $P_n^{\Pi|B} = \int P_{\theta,n}\, d\Pi_n(\theta|B)$.*

Remark 103 *(Le Cam, unpublished (197X) and (1986))*
Rewrite the posterior concentration inequality

$$P_0^n \Pi(V_n|X^n) \leq \left\| P_0^n - P_n^{\Pi|B_n} \right\| + \int P^n \phi_n \, d\Pi(P|B_n) + \frac{\Pi(V_n)}{\Pi(B_n)} \int Q^n (1-\phi_n) \, d\Pi(Q|V_n)$$

Remark 104 *Useful in parametric models (e.g. BvM) but "a considerable nuisance" [sic, Le Cam (1986)] in non-parametric context*

Schwartz's theorem revisited

Remark 105 *Suppose that for all $\delta > 0$, there is a B s.t. $\Pi(B) > 0$ and for Π-almost-all $\theta \in B$ and large enough n*

$$P_0^n \Pi(V|X^n) \leq e^{n\delta} P_{\theta,n} \Pi(V|X^n)$$

then (by Fatou) for large enough m

$$\limsup_{n\to\infty} \left[(P_0^n - e^{n\delta} P_n^{\Pi|B}) \Pi(V|X^n) \right] \leq 0$$

Theorem 106 *Let $\mathscr{P}$ be a model with KL-prior Π; $P_0 \in \mathscr{P}$. Let $B, V \in \mathscr{G}$ be given and assume that B contains a KL-neighbourhood of P_0. If there exist Bayesian tests for B versus V of exponential power then*

$$\Pi(V|X^n) \xrightarrow{P_0 - a.s.} 0$$

Corollary 107 *(Schwartz's theorem)*

Remote contiguity

Definition 108 *Given* (P_n), (Q_n), Q_n *is contiguous w.r.t.* P_n ($Q_n \lhd P_n$), *if for any msb* $\psi_n : \mathscr{X}^n \to [0,1]$

$$P_n\psi_n = o(1) \quad \Rightarrow \quad Q_n\psi_n = o(1)$$

Definition 109 *Given* (P_n), (Q_n) *and a* $a_n \downarrow 0$, Q_n *is* a_n*-remotely contiguous w.r.t.* P_n ($Q_n \lhd a_n^{-1}P_n$), *if for any msb* $\psi_n : \mathscr{X}^n \to [0,1]$

$$P_n\psi_n = o(a_n) \quad \Rightarrow \quad Q_n\psi_n = o(1)$$

Remark 110 *Contiguity is stronger than remote contiguity*
note that $Q_n \lhd P_n$ *iff* $Q_n \lhd a_n^{-1}P_n$ *for all* $a_n \downarrow 0$.

Definition 111 *Hellinger transform* $\psi(P,Q;\alpha) = \int p^\alpha q^{1-\alpha}\, d\mu$

Le Cam's first lemma

Lemma 112 *Given* (P_n), (Q_n) *like above,* $Q_n \lhd P_n$ *iff:*

(i) *If* $T_n \xrightarrow{P_n} 0$, *then* $T_n \xrightarrow{Q_n} 0$

(ii) *Given* $\epsilon > 0$, *there is a* $b > 0$ *such that* $Q_n(dQ_n/dP_n > b) < \epsilon$

(iii) *Given* $\epsilon > 0$, *there is a* $c > 0$ *such that* $\|Q_n - Q_n \wedge c\, P_n\| < \epsilon$

(iv) *If* $dP_n/dQ_n \xrightarrow{Q_n\text{-w.}} f$ *along a subsequence*, *then* $P(f > 0) = 1$

(v) *If* $dQ_n/dP_n \xrightarrow{P_n\text{-w.}} g$ *along a subsequence*, *then* $Eg = 1$

(vi) $\liminf_n \psi(P_n, Q_n; \alpha) \to 1$ *as* $\alpha \uparrow 1$

Criteria for remote contiguity

Lemma 113 *Given* (P_n), (Q_n), $a_n \downarrow 0$, $Q_n \lhd a_n^{-1} P_n$ *if any of the following holds:*

(i) *For any bnd msb* $T_n : \mathscr{X}^n \to \mathbb{R}$, $a_n^{-1} T_n \xrightarrow{P_n} 0$, *implies* $T_n \xrightarrow{Q_n} 0$

(ii) *Given* $\epsilon > 0$, *there is a* $\delta > 0$ *s.t.* $Q_n(dP_n/dQ_n < \delta\, a_n) < \epsilon$ *f.l.e.n.*

(iii) *There is a* $b > 0$ *s.t.* $\liminf_{n\to\infty} b\, a_n^{-1}\, P_n(dQ_n/dP_n > b\, a_n^{-1}) = 1$

(iv) *Given* $\epsilon > 0$, *there is a* $c > 0$ *such that* $\|Q_n - Q_n \wedge c\, a_n^{-1} P_n\| < \epsilon$

(v) *Under* Q_n, *every subsequence of* $(a_n (dP_n/dQ_n)^{-1})$ *has a weakly convergent subsequence*

[(vi) $\lim_{\alpha\uparrow 1} \liminf_n a_n{}^{-\alpha}\, \psi(P_n, Q_n; \alpha) > 0]$

Beyond Schwartz

Theorem 114 *Let* $(\Theta, \mathscr{G}, \Pi)$ *and* $(X_1, \ldots, X_n) \sim P_{0,n}$ *be given. Assume there are* $B, V \in \mathscr{G}$ *with* $\Pi(B) > 0$ *and* $a_n \downarrow 0$ *s.t.*

(i) There exist Bayesian tests for B *versus* V *of power* a_n*,*

$$\int_B P_{\theta,n}\phi_n \, d\Pi(\theta) + \int_V P_{\theta,n}(1-\phi_n) \, d\Pi(\theta) = o(a_n)$$

(ii) The sequence $(P_{0,n})$ *satisfies* $P_{0,n} \lhd a_n^{-1} P_n^{\Pi|B}$

Then $\Pi(V|X^n) \xrightarrow{P_0} 0$

Application to i.i.d. consistency (I)

Remark 115 *(Schwartz (1965))*
Take $P_0 \in \mathscr{P}$, and define

$$V_n = \{P \in \mathscr{P} : H(P, P_0) \geq \epsilon\}$$

$$B_n = \{P : -P_0 \log dP/dP_0 < \tfrac{1}{2}\epsilon^2\}$$

With $N(\epsilon, \mathscr{P}, H) < \infty$, and a_n of form $\exp(-nD)$ the theorem proves Hellinger consistency with KL-priors.

Consistency with n-dependence

Theorem 116 *Let $(\mathscr{P}, \mathscr{G})$ with priors (Π_n) and $(X_1, \ldots, X_n) \sim P_{0,n}$ be given. Assume there are $B_n, V_n \in \mathscr{G}$ and $a_n, b_n \geq 0$, $a_n = o(b_n)$ s.t.*

(i) There exist Bayesian tests for B_n versus V_n of power a_n,

$$\int_{B_n} P_{\theta,n}\phi_n \, d\Pi_n(\theta) + \int_{V_n} P_{\theta,n}(1-\phi_n) \, d\Pi_n(\theta) = o(a_n)$$

(ii) The prior mass of B_n is lower-bounded by b_n, $\Pi_n(B_n) \geq b_n$

(iii) The sequence $(P_{0,n})$ satisfies $P_0^n \lhd b_n a_n^{-1} P_n^{\Pi_n|B_n}$

Then $\Pi_n(V_n|X^n) \xrightarrow{P_0} 0$

Application to i.i.d. consistency (II)

Remark 117 *(Barron-Schervish-Wasserman (1999), Ghosal-Ghosh-vdVaart (2000), Shen-Wasserman (2001))*
Take $P_0 \in \mathscr{P}$, and define

$$V_n = \{P \in \mathscr{P} : H(P, P_0) \geq \epsilon_n\}$$

$$B_n = \{P : -P_0 \log dP/dP_0 < \tfrac{1}{2}\epsilon_n^2,\ P_0 \log^2 dP/dP_0 < \tfrac{1}{2}\epsilon_n^2\}$$

With $\log N(\epsilon_n, \mathscr{P}, H) \leq n\epsilon_n^2$, and a_n and b_n of form $\exp(-Kn\epsilon_n^2)$ the theorem proves Hellinger consistency at rate ϵ_n

Remark 118 *Larger B_n are possible, under conditions on the model (see Kleijn and Zhao (201x))*

Consistent Bayes factors

Theorem 119 *Let the model* $(\mathscr{P}, \mathscr{G})$ *with priors* (Π_n) *be given. Given* $B, V \in \mathscr{G}$ *with* $\Pi(B), \Pi(V) > 0$ *s.t.*

(i) There are Bayesian tests for B *versus* V *of power* $a_n \downarrow 0$,

$$\int_B P_{\theta,n}\phi_n \, d\Pi_n(\theta) + \int_V P_{\theta,n}(1-\phi_n) \, d\Pi_n(\theta) = o(a_n)$$

(ii) For all $\theta \in B$, $P_{\theta,n} \lhd a_n^{-1} \, P_n^{\Pi_n|B}$; *for all* $\eta \in V$, $P_{\eta,n} \lhd a_n^{-1} \, P_n^{\Pi_n|V}$

Then or Bayes factors (or posterior odds),

$$B_n = \frac{\Pi(B|X^n)}{\Pi(V|X^n)} \frac{\Pi(V)}{\Pi(B)}$$

for B *versus* V *are consistent.*

Random-walk goodness-of-fit testing (I)

Given $(S, \mathscr{S})$ state space for a discrete-time, stationary Markov process with transition kernel $P(\cdot|\cdot) : \mathscr{S} \times S \to [0, 1]$, the data consists of random walks X^n.

Choose a finite partition $\alpha = \{A_1, \ldots, A_N\}$ of S and 'bin the data': Z^n in finite state space S_α. Z^n is stationary Markov chain on S_α with transition probabilities

$$p_\alpha(k|l) = P(X_i \in A_k | X_{i-1} \in A_l),$$

We assume that p_α is ergodic with equilibrium distribution π_α.

We are interested in Bayes factors for goodness-of-fit testing of transition probabilities.

Ergodic random-walks

Example 120 *Assume that* $p_0 \in \Theta$ *generates an ergodic Markov chain* Z^n. *Denote* $Z^n \sim P_{0,n}$ *and equilibrium distribution* π_0

For given $\epsilon > 0$, *define,*

$$B' = \left\{ p_\alpha \in \Theta : \sum_{k,l=1}^{N} -p_0(l|k)\pi_0(k) \log \frac{p_\alpha(l|k)}{p_0(l|k)} < \epsilon^2 \right\}.$$

Assume $\Pi(B') > 0$.

According to the ergodic theorem,

$$\frac{1}{n} \sum_{i=1}^{n} \log \frac{p_\alpha(Z_i|Z_{i-1})}{p_0(Z_i|Z_{i-1})} \xrightarrow{P_{0,n}\text{-a.s.}} \sum_{k,l=1}^{N} p_0(l|k)\pi_0(k) \log \frac{p(l|k)}{p_0(l|k)},$$

Remote contiguity of ergodic random-walks

so for every $p_\alpha \in B'$ and large enough n, $P_{0,n}$-almost-surely

$$\frac{dP_{\alpha,n}}{dP_{0,n}}(Z^n) = \prod_{i=1}^{n} \frac{p_\alpha(Z_i|Z_{i-1})}{p_0(Z_i|Z_{i-1})} \geq e^{-\frac{n}{2}\epsilon^2}$$

Fatou's lemma implies remote contiguity because,

$$P_{0,n}\left(\int \frac{dP_{\alpha,n}}{dP_{0,n}}(Z^n)\, d\Pi(p_\alpha|B') < e^{-\frac{n}{2}\epsilon^2}\right) \to 0.$$

So lemma 113 says that

$$P_{0,n} \lhd \exp(\tfrac{n}{2}\epsilon^2) P_n^{\Pi|B'}$$

Remark 121 *Exponential remote contiguity is* not enough *for goodness-of-fit tests below. Instead we use to* local asymptotic normality *for a sharper result.*

Random-walk goodness-of-fit testing (II)

Fix $P_0, \epsilon > 0$ and hypothesize on 'bin probabilities' $p_\alpha(k,l) = p_\alpha(k|l)\pi_\alpha(l)$,

$$H_0 : \max_{k,l} \Big| p_\alpha(k,l) - p_0(k,l) \Big| < \epsilon, \quad H_1 : \max_{k,l} \Big| p_\alpha(k,l) - p_0(k,l) \Big| \geq \epsilon,$$

Define, for $\delta_n \downarrow 0$,

$$B_n = \{p_\alpha \in \Theta : \max_{k,l} \Big| p_\alpha(k,l) - p_0(k,l) \Big| < \epsilon - \delta_n\}$$

$$V_{k,l} = \{p_\alpha \in \Theta : \Big| p_\alpha(k,l) - p_0(k,l) \Big| \geq \epsilon\},$$

$$V_{+,k,l,n} = \{p_\alpha \in \Theta : \ p_\alpha(k,l) - p_0(k,l) \ \geq \epsilon + \delta_n\},$$

$$V_{-,k,l,n} = \{p_\alpha \in \Theta : \ p_\alpha(k,l) - p_0(k,l) \ \leq -\epsilon - \delta_n\}.$$

Remark 122 *A Bayesian test sequence for H_0 versus H_1 exists based on a version of Hoeffdings inequality for random walks (Glynn and Ormoneit (2002), Meyn and Tweedie (2009))*

Random-walk goodness-of-fit testing (III)

Choquet $p_\alpha(k|l) = \sum_{E \in \mathscr{E}} \lambda_E E(k|l)$ where the N^N transition kernels E are deterministic. Define,

$$S_n = \left\{ \lambda_{\mathscr{E}} \in S^{N^N} : \lambda_E \geq \lambda_n / N^{N-1}, \text{for all } E \in \mathscr{E} \right\},$$

for $\lambda_n \downarrow 0$.

Theorem 123 *Choose a prior $\Pi \ll \mu$ on S^{N^N} with continuous, strictly positive density. Assume that,*

(i) $n\lambda_n^2 \delta_n^2 / \log(n) \to \infty$,

(ii) $\Pi(B \setminus B_n), \Pi(\Theta \setminus S_n) = o(n^{-(N^N/2)})$,

(iii) $\Pi(V_{k,l} \setminus (V_{+,k,l,n} \cup V_{-,k,l,n})) = o(n^{-(N^N/2)})$, *for all* $1 \leq k, l \leq N$.

Then the Bayes factors F_n for H_0 versus H_1 are consistent.

Lecture VIII

Posterior uncertainty quantification

As we have seen in Lecture II the Bernstein-von-Mises limit allows us to identify credible sets and confidence sets in the large-sample limit. This identification extends much further: in this lecture we consider various ways in which credible sets and their enlargements serve as confidence sets. Before we turn to posterior uncertainty quantification, we look in detail at the proof of frequentist posterior consistency with the Dirichlet prior.

arxiv:1611.08444 (MATH.ST)

Remote contiguity in finite sample spaces

Observe an *i.i.d.* sample $X_1, X_2, \ldots$ taking values in a space $\mathscr{X}$ of finite order N. Let M denote the space of all probability measures on $\mathscr{X}$.

$(M, \|\cdot\|)$ is isometric to the simplex,

$$S_N = \left\{p = (p(1), \ldots, p(N)) : \min_k p(k) \geq 0, \Sigma_i\, p(i) = 1\right\},$$

with ℓ^1-norm: $\|p - q\| = \Sigma_k\, |p(k) - q(k)|$.

Proposition 124 *If i.i.d. $X_1, X_2, \ldots$ are $\mathscr{X}$-valued, then for any $n \geq 1$, any Borel prior Π of full support on M, any $P_0 \in M$ and any ball B around P_0, there exists an $\epsilon' > 0$ such that,*

$$P_0^n \lhd e^{\frac{1}{2}n\epsilon^2} P_n^{\Pi|B},$$

for all $0 < \epsilon < \epsilon'$.

Consistency with finite sample spaces

Given $\delta > 0$, consider

$$B = \{P \in M : \|P - P_0\| < \delta\},\ V = \{Q \in M : \|Q - P_0\| > 2\delta\}.$$

M is compact $N(\delta, M, \|\cdot\|) < \infty$ for all δ and there exist uniform tests for B versus V (with power e^{-nD}, $D > 0$).

Proposition 124 with an $0 < \epsilon < \epsilon'$ small enough guarantees exponential remote contiguity

Then theorem 114 says $\Pi(V|X^n)$ goes to zero in P_0^n-probability.

Proposition 125 *(Freedman, 1965) A posterior resulting from a prior Π of full support on M is consistent in total variation.*

Weak consistency with Dirichlet process priors

Recall

Definition 126 *(Dirichlet process, Ferguson 1973,1974)*
Let μ *be a finite Borel msr on* $([0,1], \mathscr{B}[0,1])$. *The Dirichlet process* $P \sim D_\mu$ *is defined by,*

$$\Big(P(A_1), \ldots, P(A_k)\Big) \sim D_{(\mu(A_1),\ldots,\mu(A_k))}$$

Define Prokhorov's weak neighbourhoods $f : [0,1] \to [0,1]$ continuous

$$U_f = \Big\{P \in M^1[0,1] \,:\, |(P - P_0)f| < \epsilon\Big\}$$

$V_f = M^1[0,1] \setminus U_f$ We want to show $P_0^n \Pi(V_f | X^n) = o(1)$.

Suitable weak tests

For continuous $f : [0,1] \to [0,1]$ and

$$B_f = \Big\{P \,:\, |(P - P_0)f| < \epsilon\Big\}, \quad V_f = \Big\{P \,:\, |(P - P_0)f| \geq 4\epsilon\Big\}.$$

Any cont $x \mapsto f(x)$ is ϵ-uniformly approximated by some g

$$g(x) = \sum_{n=1}^{N} g_n \, 1_{A_n}(x)$$

on a partition in intervals $A_1, \ldots, A_N$

$$B_g = \Big\{P \,:\, |(P - P_0)g| < 2\epsilon\Big\}, \quad V_g = \Big\{P \,:\, |(P - P_0)g| \geq 3\epsilon\Big\}.$$

$B_f \subset B_g$, $V_f \subset V_g$ and Lemma 22 says there are (ϕ_n)

$$\sup_{P \in B_g} P^n \phi_n \leq e^{-nD}, \quad \sup_{Q \in V_g} Q^n (1 - \phi_n) \leq e^{-nD}. \tag{28}$$

Remote contiguity in restricted form

For given f and $\epsilon > 0$, construct g on some α.

Define sub-σ-algebra $\sigma_{\alpha,n} = \sigma(\alpha^n)$ on $\mathscr{X}_n = [0,1]^n$.

Remark 127 *Tailfreeness (Freedman, 1965)*

$$\mathscr{X}_n \to [0,1] \,:\, X^n \mapsto \Pi(V_g \,|X^n) \text{ is } \sigma_{\alpha,n}\text{-measurable}$$

Remote contiguity,

$$P_n^{\Pi|B_g}\psi_n(X^n) = o(\rho_n) \quad \Rightarrow \quad P_0^n\psi_n(X^n) = o(1),$$

only for $\sigma_{\alpha,n}$-measurable $\psi_n : \mathscr{X}^n \to [0,1]$

Partitions and projections

Project $[0,1]$ onto $\mathscr{X}_\alpha = \{e_n : 1 \le n \le N_\alpha\}$

$$\varphi_\alpha(x) = \Big(1\{x \in A_1\}, \ldots, 1\{x \in A_{N_\alpha}\}\Big).$$

and consider $\varphi_{*\alpha} : M^1[0,1] \to S_{N_\alpha}$,

$$\varphi_{*\alpha}(P) = \Big(P(A_1), \ldots, P(A_{N_\alpha})\Big),$$

Remote contiguity and testing happen equivalently in S_{N_α}

Full support of Π_α guarantees remote contiguity with exponential rates. Together with tests (28), implies weak consistency

$$\Pi(V_f|X^n) \le \Pi(V_g|X^n) \xrightarrow{P_0} 0$$

Dirichlet process prior full support of the base measure μ implies full support for all Π_α, if $\mu(A_i) > 0$ for all $1 \le i \le N_\alpha$. Particularly, we require $P_0 \ll \mu$ for consistent estimation.

Credible sets and confidence sets

Let $\mathscr{D}$ denote a collection of measurable subsets of Θ

Definition 128 *Let $(\Theta, \mathscr{G})$ with priors Π_n be given. Denote the sequence of posteriors by $\Pi(\cdot|\cdot) : \mathscr{G} \times \mathscr{X}_n \to [0,1]$. A sequence of credible sets (D_n) of credible levels $1 - a_n$ (with $a_n \downarrow 0$) is a sequence of set-valued maps $D_n : \mathscr{X}_n \to \mathscr{D}$ such that,*

$$\Pi(\Theta \setminus D_n(X^n)|X^n) = o(a_n),$$

$P_n^{\Pi_n}$-almost-surely (or in $P_n^{\Pi_n}$-probability).

Definition 129 *A sequence of maps $x \mapsto C_n(x) \subset \Theta$ forms an asymptotically consistent sequence of confidence sets (of credible levels $1 - o(a_n)$), if,*

$$P_{\theta_0,n}\Big(\theta_0 \notin C_n(X^n)\Big) = o(a_n)$$

for all $\theta_0 \in \Theta$.

Credible sets *with* converging posteriors (I)

Distinguish theorems with posteriors convergence as a condition and theorems without such conditions.

We assume that (Θ_n, d_n) are metric spaces. Denote balls,

$$B_n(\theta_n, r_n) = \{\theta'_n \in \Theta_n : d_n(\theta', \theta_n) \leq r_n\},$$

where both θ_n and r_n may be random.

Definition 130 *Let (Θ_n, d_n) with priors Π_n be given. A sequence of credible balls*

$$D_n = B_n(\widehat{\theta}_n, \widehat{r}_n)$$

of credible levels $1 - o(a_n)$ satisfy,

$$\Pi(\Theta \setminus D_n(X^n)|X^n) = \Pi(d_n(\theta_n, \widehat{\theta}_n) \leq \widehat{r}_n|X^n) = o(a_n),$$

$P_n^{\Pi_n}$-almost-surely (or in $P_n^{\Pi_n}$-probability).

Credible sets *with* converging posteriors (II)

Suppose that (Θ_n, d_n) are metric spaces

Theorem 131 *(van Waaij, BK, 2018/19)*
Suppose that $0 < \epsilon \leq 1$, $P_{\theta_{0,n}} \ll P_n^{\Pi_n}$ *and*

$$\Pi\Big(d_n(\theta_n, \theta_{0,n}) \leq r_n \,\Big|\, X^n \Big) \xrightarrow{P_{\theta_{0,n}}} 1$$

Let $\widehat{B}_n = B_n(\hat{\theta}_n, \hat{r}_n)$ *be level-*$1-\epsilon$ *credible balls of minimal radii.*

Then with high $P_{\theta_{0,n}}$*-probability* $\hat{r}_n \leq r_n$

And $C_n(X^n) = B_n(\hat{\theta}_n, \hat{r}_n + r_n) \subset B_n(\hat{\theta}_n, 2r_n)$ *have asymptotic coverage,*

$$P_{\theta_{0,n}}\Big(\theta_{0,n} \in C_n(X^n) \Big) \to 1,$$

Credible sets *with* converging posteriors (III)

Theorem 132 *(van Waaij, BK, 2018/19)*
Suppose that $0 < \epsilon \leq 1$, $P_{\theta_{0,n}} \ll P_n^{\Pi_n}$ *and*

$$\Pi\Big(d_n(\theta_n, \theta_{0,n}) \leq r_n \,\Big|\, X^n \Big) \xrightarrow{P_{\theta_{0,n}}} 1$$

Let $B_n(\hat{\theta}_n, \hat{r}_n)$ *be level-*$1-\epsilon$ *credible balls of near-minimal radii*

$$\hat{r}_n = (1 + o(1))\hat{r}_{\epsilon,n}$$

where $\hat{r}_{\epsilon,n}$ *denote the infimal radii of level-*$1-\epsilon$ *credible balls.*

Then with high $P_{\theta_{0,n}}$*-probability* $\hat{r}_n \leq (1 + o(1))r_n$

And $C_n(X^n) = B_n(\hat{\theta}_n, \hat{r}_n + r_n) \subset B_n(\hat{\theta}_n, 2(1+o(1))r_n)$ *have asymptotic coverage,*

$$P_{\theta_{0,n}}\Big(\theta_{0,n} \in C_n(X^n) \Big) \to 1,$$

Proof of theorem 132 (I)

Let $n \geq 1$ be given

The posterior $\Pi(\cdot|X^n = x^n)$ is defined for all x^n in an event F_n such that $P_n^{\Pi_n}(F_n) = 1$, and because $P_{0,n} \ll P_n^{\Pi_n}$, also $P_{\theta_{0,n}}(F_n) = 1$.

For $x^n \in F_n$, let $r_n(\theta_n, x^n)$ denote the infimal radius of balls in Θ_n centred on θ_n of posterior mass at least $1 - \epsilon$.

Define $\widehat{\theta}_n(x^n)$ as the centre point of a credible ball $\widehat{B}_n = B_n(\widehat{\theta}_n, \widehat{r}_n)$ of level $1 - \epsilon$, with near-minimal radius $\widehat{r}_n$, *i.e.*

$$\widehat{r}_n(x^n) \leq (1 + o(1)) \inf\{r_n(\theta_n, x^n) : \theta_n \in \Theta_n\}$$

Note

$$P_{\theta_0,n}\Big(\Pi(\widehat{B}_n|X^n) \geq 1 - \epsilon\Big) = 1,$$

for all $n \geq 1$.

Proof of theorem 132 (II)

Posterior convergence the ball $B_n(\theta_{0,n}, r_n)$ is a credible ball of level $1-\epsilon$ for large enough n. Therefore, with high $P_{0,n}$-probability

$$\widehat{r}_n(X^n) \leq (1+o(1))\, r_n(\theta_{0,n}, X^n) \leq (1+o(1))\, r_n.$$

Posterior convergence the balls $B_n(\theta_{0,n}, r_n)$ satisfy

$$P_{\theta_{0,n}}\Big(\Pi(B_n(\theta_{0,n}, r_n)|X^n) > \epsilon \Big) \to 1.$$

Conclude that, with high $P_{\theta_{0,n}}$-probability,

$$B_n(\theta_{0,n}, r_n) \cap B_n(\widehat{\theta}_n(X^n), \widehat{r}_n(X^n)) \neq \varnothing,$$

implying asymptotic coverage of $\theta_{0,n}$ for $C_n(X^n)$.

Remark 133 *Proof does not lead to automatic* rate-adaptivity *(Hengartner (1995), Cai, Low and Xia (2013), Szabó, vdVaart, vZanten (2015)) when $r_n = r_n(P_{0,n})$: estimation of r_n is problematic.*

Uncertainty quantification in the EIV model

Example 134 *Theorem 82 says the posterior converges at rate* ϵ_n. *Let* $\hat{B}_n = B_n(\hat{\theta}_n, \hat{r}_n)$ *be level-*$1-\epsilon$ *credible balls of minimal radii and* $C_n = B_n(\hat{\theta}_n, \hat{r}_n + \epsilon_n) \subset B_n(\hat{\theta}_n, 2(1+o(1))\epsilon_n)$.

$\mathscr{F} = \mathsf{Lip}_M(\alpha)$ *with a net prior, the sets* $C(X^n)$ *have asymptotic coverage, and shrink like,*

$$\epsilon_n = n^{-\frac{\alpha}{2\alpha+1}} (\log n)^{\frac{1}{2\alpha}}.$$

$\mathscr{F} = D_{\alpha,M}(q)$ *with a net prior, the sets* $C(X^n)$ *have asymptotic coverage, and shrink like,*

$$\epsilon_n = n^{-\frac{q+\alpha}{2q+2\alpha+1}} (\log n)^{\frac{1}{2q+2\alpha}}.$$

$\mathscr{F} = \mathscr{F}_{\ominus}$ *with Lebesgue prior with continuous and strictly positive density, the sets* $C(X^n)$ *have asymptotic coverage, and shrink like,*

$$\epsilon_n = n^{-1/2} (\log n)^{3/2}.$$

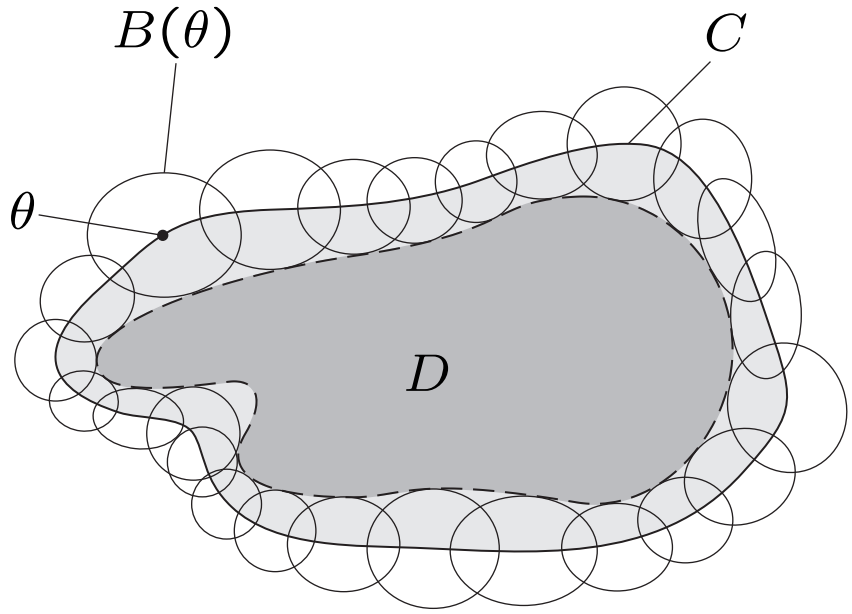
Credible sets *without* converging posteriors

Definition 135 *Let D be a credible set in Θ and let B denote a set function $\theta \mapsto B(\theta) \subset \Theta$. A model subset C is said to be a confidence set associated with D under B, if for all $\theta \in \Theta \setminus C$,*

$$B(\theta) \cap D = \varnothing$$

Definition 136 *The intersection C_0 of all C like above is a confidence set associated with D under B, called the minimal confidence set associated with D under B.*

B-Enlargement of credible sets



A credible set D and its associated confidence set C under B in terms of Venn diagrams: additional points $\theta \in C \setminus D$ are characterized by non-empty intersection $B(\theta) \cap D \neq \varnothing$.

B-Enlarged credible sets are confidence sets

Theorem 137 *Let* $0 \leq a_n \leq 1$, $a_n \downarrow 0$ *and* $b_n > 0$ *such that* $a_n = o(b_n)$ *be given and let* D_n *denote level-*$(1-o(a_n))$ *credible sets. Furthermore, for all* $\theta \in \Theta$, *let* B_n *be set functions such that,*

(i) $\Pi_n(B_n(\theta_0)) \geq b_n$,

(ii) $P_{\theta_0,n} \lhd b_n a_n^{-1} \, P_n^{\Pi_n | B_n(\theta_0)}$.

Then any confidence sets C_n *associated with the credible sets* D_n *under* B_n *are asymptotically consistent, that is,*

$$P_{\theta_0,n}\Big(\theta_0 \in C_n(X^n) \Big) \to 1.$$

Methodology: confidence sets from posteriors (I)

Corollary 138 *Given* $(\Theta, \mathscr{G})$, (Π_n) *and* (B_n) *with* $\Pi_n(B_n) \geq b_n$ *and* $P_{\theta,n} \lhd P_n^{\Pi_n|B_n}$, *any credible sets* D_n *of level* $1 - a_n$ *with* $a_n = o(b_n)$ *have associated confidence sets under* B_n *that are asymptotically consistent.*

Next, assume that $(X_1, X_2, \ldots, X_n) \in \mathscr{X}^n \sim P_0^n$ for some $P_0 \in \mathscr{P}$.

Corollary 139 *Let* Π_n *denote Borel priors on* $\mathscr{P}$, *with constant* $C > 0$ *and rate sequence* $\epsilon_n \downarrow 0$ *such that:*

$$\Pi_n\left(P \in \mathscr{P} \,:\, -P_0 \log \frac{dP}{dP_0} < \epsilon_n^2,\, P_0\left(\log \frac{dP}{dP_0}\right)^2 < \epsilon_n^2 \right) \geq e^{-Cn\epsilon_n^2}.$$

Given credible sets D_n *of level* $1 - o(\exp(-C'n\epsilon_n^2))$, *for some* $C' > C$. *Then radius-*ϵ_n *Hellinger-enlargements* C_n *are asymptotically consistent confidence sets.*

Methodology: confidence sets from posteriors (II)

Note the relation between Hellinger diameters,

$$\text{diam}_H(C_n(X^n)) = \text{diam}_H(D_n(X^n)) + 2\epsilon_n.$$

If, in addition, tests satisfying

$$\int_{B_n} P_{\theta,n}\phi_n(X^n)\, d\Pi_n(\theta) + \int_{V_n} P_{\theta,n}(1-\phi_n(X^n))\, d\Pi_n(\theta) = o(a_n),$$

with $a_n = \exp(-C'n\epsilon_n^2)$ exist, the posterior is Hellinger consistent at rate ϵ_n, so that $\text{diam}_H(D_n(X^n)) \leq M\epsilon_n$ for some $M > 0$.

If ϵ_n is the minimax rate of convergence for the problem, the confidence sets $C_n(X^n)$ are rate-optimal (Low, (1997)).

Remark 140 *Rate-adaptivity (Hengartner (1995), Cai, Low and Xia (2013), Szabó, vdVaart, vZanten (2015)) is not possible like this because a definite choice for the sets in B_n is required.*

Lecture IX

Exact recovery and detection of communities

The planted bi-section model describes random graphs X^n of $2n$ vertices, divided in two (unobserved) communities, with given within-community p_n and between-community q_n edge-probabilities. Can we estimate the communities consistently as the graph size increases? In this lecture we show that posteriors recover communities exactly or concentrates on assignments with a controlled number of mis-assignments, depending on the Erdös-Rényi phase of the sequence of graphs.

arxiv:1810.09533 (MATH.ST)

Erdös-Rényi random graphs

Fix $n \geq 1$, denote $G_n = (V_n, E_n)$ complete graph with n vertices and percolate edges (with $p_n \in (0,1)$),

For every $e \in E_n$ independently, include e in $E'_n \subset E_n$ wp. p_n.

Result random graph $G(n, p_n) = (V_n, E'_n)$ (Erdös, Rényi (1959–1961)).

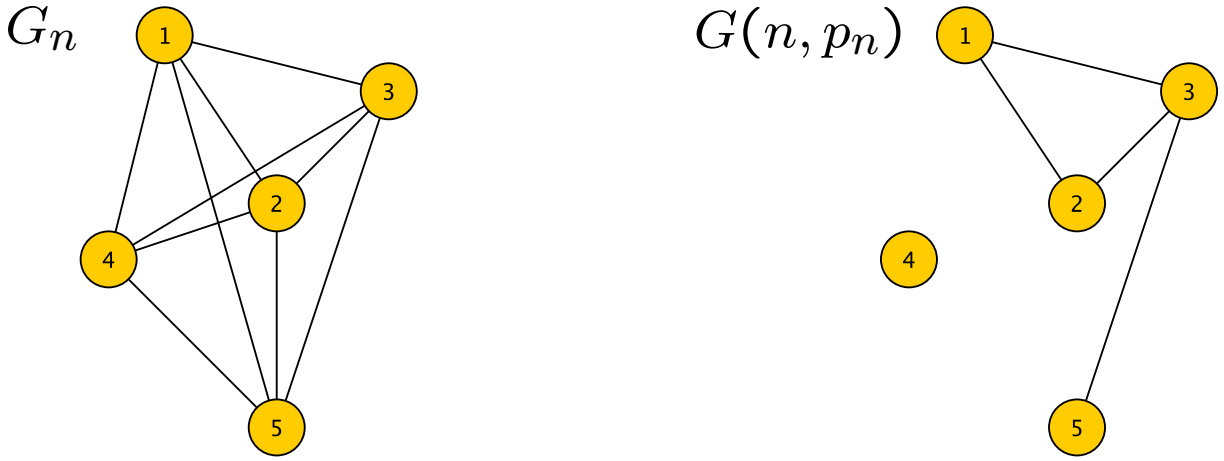


Fig. 1. Complete graph and edge-percolated ER-graph

Phases of the Erdös-Rényi graph

Like (most) physical materials, ER-graphs have three phases

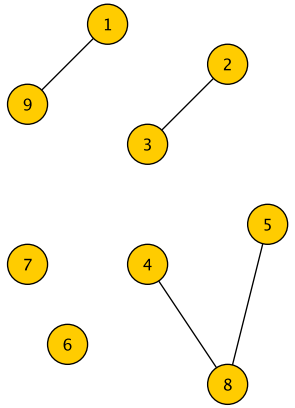


Fig 2a. Fragmented

sizes $< O(\log(n))$

$p_n < 1/n$

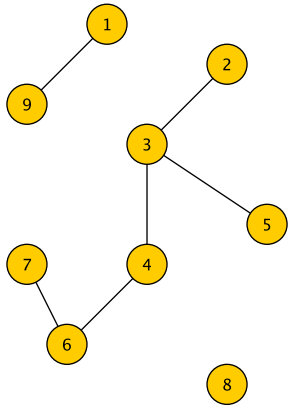


Fig 2b. Giant comp

size $\sim O(n)$

$1/n < p_n < \log(n)/n$

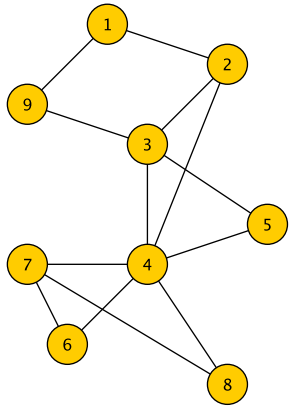


Fig 2c. Connected

$p_n > \log(n)/n$

The planted bi-section model (I)

Consider $G_{2n} = (V_{2n}, E_{2n})$ with class assignment $\theta_n \in \{0,1\}^{2n}$. Write $V_{2n} = Z_0(\theta_n) \cup Z_1(\theta_n)$, split in class zero $Z_0(\theta_n)$ and class one $Z_1(\theta_n)$.

Edge-percolate with some $p_n, q_n \in [0,1]$,

For every $e \in E_{2n}$ independently,

include e in $E'_{2n} \subset E_{2n}$ wp. $\begin{cases} p_n, & \text{if } e \text{ lies within } Z_0 \text{ or } Z_1, \\ q_n, & \text{if } e \text{ connects } Z_0 \text{ and } Z_1. \end{cases}$

Result random graph $X^n = (V_{2n}, E'_{2n})$

Community detection

Example PBM with $n = 12$, $0 < q_6 \ll p_6 < 1$, $\theta_n = 000000111111$

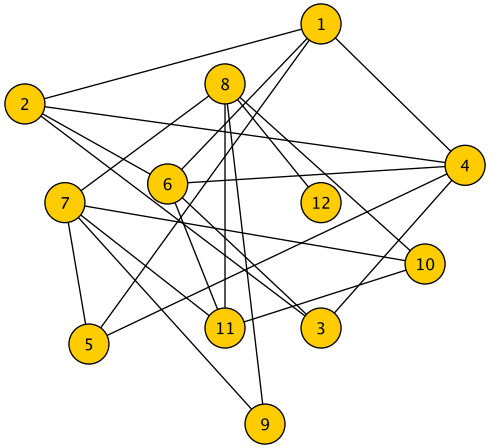


Fig 3a. Observed

Data X^n

PB graph

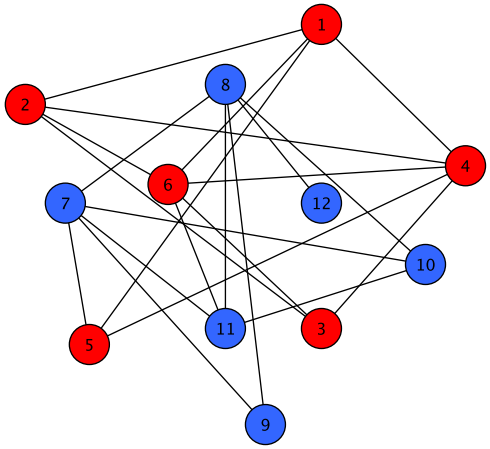


Fig 3b. Unobserved

Communities of θ_n

$Z_0(\theta_n), Z_1(\theta_n)$

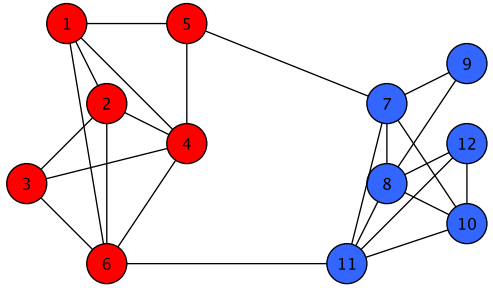


Fig 3c. Detection

Estimate with

$\hat{Z}_0(X^n), \hat{Z}_1(X^n)$

Methods of community detection (I)

General

Abbe, E. (2018). Community Detection and Stochastic Block Models: Recent Developments. J. Machine Learning Research 18.177, 1-86.

Spectral clustering

Krzakala, F. et al. (2013). Spectral redemption in clustering sparse networks. PNAS 110.52, 20935-20940.

Maximization of the likelihood and other modularities

Girvan, M. and M. E. J. Newman (2002). Community structure in social and biological networks. PNAS 99.12, 7821-7826.

Bickel, P. J. and A. Chen (2009). A nonparametric view of network models and Newman-Girvan and other modularities. PNAS 106.50, 21068-21073.

Choi, D. S., P. J. Wolfe, and E. M. Airoldi (2012). Stochastic blockmodels with a growing number of classes. Biometrika 99.2, 273-284.

Amini, A. A. et al. (2013). Pseudo-likelihood methods for community detection in large sparse networks. Ann. Statist. 41.4, 2097-2122.

Methods of community detection (II)

Semi-definite programming

Hajek, B., Y. Wu, and J. Xu (2016). Achieving Exact Cluster Recovery Threshold via Semidefinite Programming. *IEEE Trans. Inf. Theor. 62.5, 2788-2797.*

Guédon, O. and R. Vershynin (2016). Community detection in sparse networks via Grothendiecks inequality. *PTRF 165.3, 1025-1049.*

Penalized ML detection, minimax misclassification

Zhang, A. Y. and H. H. Zhou (2016). Minimax rates of community detection in stochas- tic block models. *Ann. Statist. 44.5, 2252-2280.*

Gao, C. et al. (2017). Achieving Optimal Misclassification Proportion in Stochastic Block Models. *J. Machine Learning Research 18.60, 1-45.*

Methods of community detection (III)

Bayesian methods

Nowicki, K. and T. A. B. Snijders (2001). Estimation and Prediction for Stochastic Blockstructures. *JASA 96.455, 1077-1087.*

Decelle, A. et al. (2011a). Asymptotic analysis of the stochastic block model for modular networks and its algorithmic applications. *Phys. Rev. E 84 (6), p. 066106.*

Decelle, A. et al. (2011b). Inference and Phase Transitions in the Detection of Modules in Sparse Networks. *Phys. Rev. Lett. 107 (6), p. 065701.*

Suwan, S. et al. (2016). Empirical Bayes estimation for the stochastic blockmodel. *Electron. J. Statist. 10.1, 761-782.*

Mossel, E., J. Neeman, and A. Sly (2016b). Consistency thresholds for the planted bisection model. *Electron. J. Probab. 21, 24 pp.*

PBM phases: Kesten-Stigum phase

Estimators $\widehat{\theta}_n$ detect the class assignment $\theta_{0,n}$

$$\frac{1}{2n}\left|\sum_{i=1}^{2n}(-1)^{\widehat{\theta}_{n,i}}\,(-1)^{\theta_{0,n,i}}\right| \xrightarrow{P_{\theta_{0,n}}} 1,$$

Decelle, A. et al. (2011) Argue that detection is possible if

$$(c_n - d_n)^2 > 2(c_n + d_n), \tag{29}$$

with $p_n = c_n/n$, $q_n = d_n/n$.

Mossel, Neeman, Sly (2015–) Prove detection is possible, iff,

$$\frac{n(p_n - q_n)^2}{p_n + q_n} \to \infty, \tag{30}$$

PBM phases: Chernoff-Hellinger phase

Estimators $\widehat{\theta}_n$ recover the class assignment exactly

$$P_{\theta_{0,n}}\left(\widehat{\theta}_n(X^n) = \theta_{0,n}\right) \to 1,$$

Dyer and Frieze (1989) Exact recovery possible, whenever,

$$p_n - q_n \geq A\frac{\log n}{n}, \tag{31}$$

Mossel, Neeman, Sly (2016) Exact recovery possible, iff,

$$(a_n + b_n - 2\sqrt{a_n b_n} - 1)\log n + \tfrac{1}{2}\log\log n \to \infty. \tag{32}$$

with $p_n = a_n \log(n)/n$,$q_n = b_n \log(n)/n$ and there is $C > 0$ s.t. $C^{-1} \leq a_n, b_n \leq C$.

Planted bi-section model (II)

Let $(\theta_{0,n})$ and $(p_n), (q_n)$ be given; for all $n \geq 1$,

Edge probabilities

$$Q_{ij}(\theta) := P_{\theta,n}(X_{ij} = 1) = \begin{cases} p_n, & \text{if } \theta_{n,i} = \theta_{n,j}, \\ q_n, & \text{if } \theta_{n,i} \neq \theta_{n,j}, \end{cases}$$

Likelihood

$$p_{\theta,n}(X^n) = \prod_{i<j} Q_{i,j}(\theta)^{X_{ij}} (1 - Q_{i,j}(\theta))^{1-X_{ij}}.$$

Posterior

$$\Pi(A|X^n) = \sum_{\theta_n \in A} p_{\theta,n}(X^n)\, \pi_n(\theta_n) \Big/ \sum_{\theta_n \in \Theta_n} p_{\theta,n}(X^n)\, \pi_n(\theta_n),$$

Testing and posterior convergence

$k(\theta_{1,n}, \theta_{2,n})$ minimal number of pair-exchanges to take $\theta_{1,n}$ into $\theta_{2,n}$.
Define $V_{n,k} = \{\theta_n : k(\theta_n, \theta_{0,n}) = k\}$ and $V_n = \cup\{V_{n,k} : k = k_n, \ldots, \lfloor n/2 \rfloor\}$

With $P_{\theta_0,n}\phi_{\theta_n,n}(X^n) + P_{\theta_n,n}(1 - \phi_{\theta_n,n}(X^n)) \leq a_{n,k}$,

$$P_{\theta_0,n}\Pi(k(\theta_n, \theta_{0,n}) \geq k_n | X^n) = \sum_{k=k_n}^{\lfloor n/2 \rfloor} P_{\theta_0,n}\Pi(V_{n,k}|X^n)$$

$$\leq \sum_{k=k_n}^{\lfloor n/2 \rfloor} \left(P_{\theta_0,n}\phi_{k,n}(X^n) + \sum_{\theta_n \in V_{n,k}} P_{\theta_n,n}(1 - \phi_{k,n}(X^n)) \right)$$

$$\leq \sum_{k=k_n}^{\lfloor n/2 \rfloor} \sum_{\theta_n \in V_{n,k}} \left(P_{\theta_{0,n},n}\phi_{\theta_n,n}(X^n) + P_{\theta_n,n}(1 - \phi_{\theta_n,n}(X^n)) \right)$$

$$\leq \sum_{k=k_n}^{\lfloor n/2 \rfloor} \binom{n}{k}^2 a_{k,n}.$$

Posterior exact recovery

Theorem 141 *For some $\theta_{0,n} \in \Theta_n$, assume that $X^n \sim P_{\theta_{0,n}}$, $(n \geq 1)$. With uniform priors and (p_n) and (q_n) such that,*

$$\left(1 + \left(1 - p_n - q_n + 2p_n\, q_n + 2\sqrt{p_n(1-p_n)q_n(1-q_n)}\right)^{n/2}\right)^{2n} \to 1, \tag{33}$$

the posterior succeeds in exact recovery, i.e.

$$\Pi(\theta_n = \theta_{0,n}|X^n) \xrightarrow{P_{\theta_{0,n}}} 1, \tag{34}$$

as $n \to \infty$.

Sharpness of sparsity bound for exact recovery

$$\Big(1+\big(1-p_n-q_n+2p_n\,q_n+2\sqrt{p_n(1-p_n)q_n(1-q_n)}\big)^{n/2}\Big)^{2n}$$

$$= \left(1+\left(1-\left(a_n+b_n-2\sqrt{a_nb_n}+o(n^{-1}\log n)\right)\frac{\log n}{n}\right)^{n/2}\right)^{2n}$$

$$\approx \left(1+n^{-\frac{1}{2}(a_n+b_n-2\sqrt{a_nb_n}))}\right)^{2n} = \left(1+\frac{1}{n}n^{-\frac{1}{2}(a_n+b_n-2\sqrt{a_nb_n}-2)}\right)^{2n}$$

$$\approx \exp\Big(2e^{-\frac{1}{2}(a_n+b_n-2\sqrt{a_nb_n}-2)\log n}\Big)$$

In the Chernoff-Hellinger phase,

$$(a_n+b_n-2\sqrt{a_nb_n}-2)\log n\to\infty, \tag{35}$$

is sufficient for exact posterior recovery.

Exact recovery with ML/MAP-estimators

Corollary 142 *Under the conditions of theorem 141, the MAP-/ML-estimator $\widehat{\theta}_n$ recovers $\theta_{0,n}$ exactly*

$$P_{\theta_{0,n}}(\widehat{\theta}_n(X^n) = \theta_{0,n}) \to 1.$$

Proof Uniformity of the priors maximization of the posterior density (with respect to the counting measure) on Θ_n, is the same as maximization of the likelihood. Due to eq. (34), the posterior density in the point $\theta_{0,n}$ in Θ_n converges to one in $P_{\theta_{0,n}}$-probability. Accordingly, the point of maximization is $\theta_{0,n}$ with high probability. □

Testing power in the PBM (I)

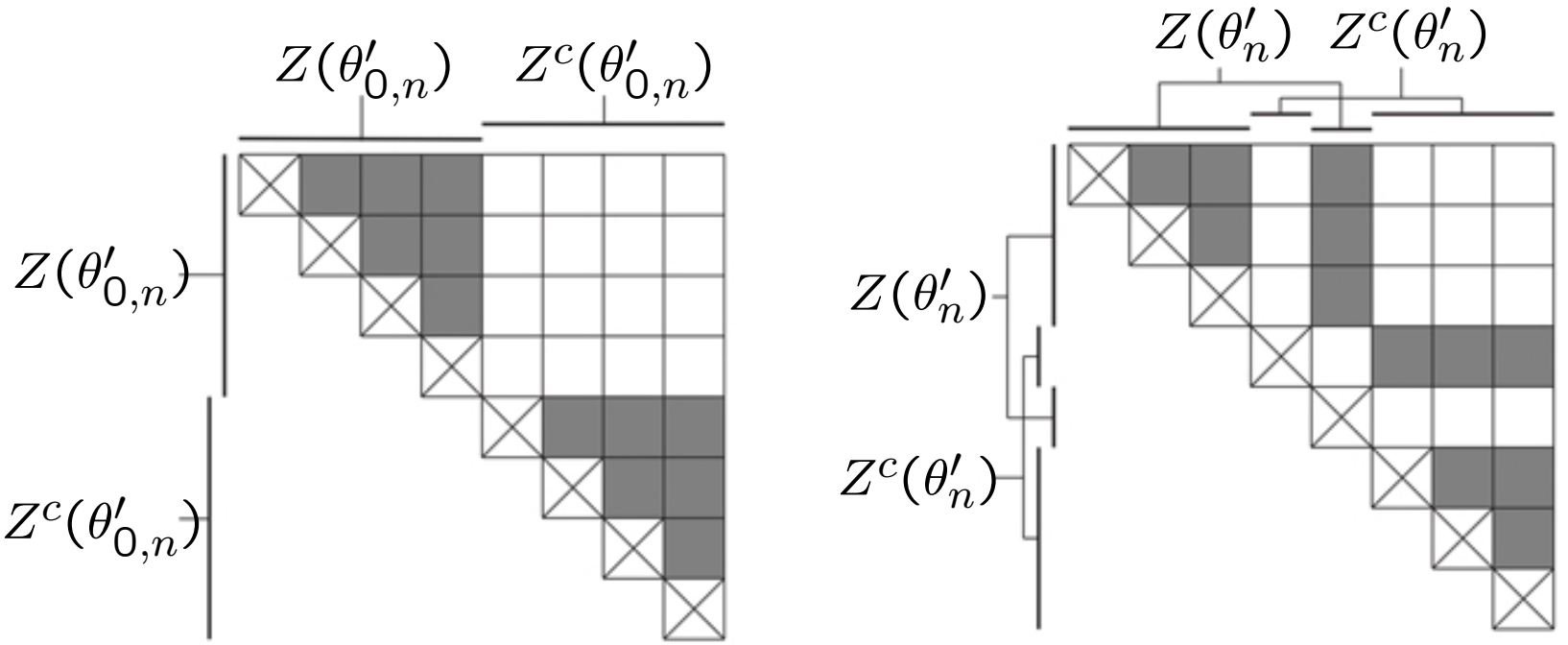


Fig 4. Class assignments $\theta'_{0,n}$ and to θ'_n for $n = 4$ and $k = 1$. Vertex sets $Z(\cdot)$ and $Z^c(\cdot)$ are zero- and one-classes. Dark are edges occuring wp p_n and light wp q_n.

Testing power in the PBM (II)

As we have seen the likelihood ratio test satisfies,

$$P_{\theta_0,n}\phi_n(X^n) + P_{\theta,n}(1-\phi_n(X^n)) \leq P_{\theta_0,n}\left(\frac{p_{\theta,n}}{p_{\theta_0,n}}(X^n)\right)^{1/2},$$

and the likelihood ratio can be written as,

$$\frac{p_{\theta,n}}{p_{\theta_0,n}}(X^n) = \left(\frac{1-p_n}{p_n}\frac{q_n}{1-q_n}\right)^{S_n-T_n}$$

where,

$$(S_n,T_n) \sim \begin{cases} \mathsf{Bin}(2k(n-k),p_n) \times \mathsf{Bin}(2k(n-k),q_n), & \text{if } X^n \sim P_{\theta_0,n}, \\ \mathsf{Bin}(2k(n-k),q_n) \times \mathsf{Bin}(2k(n-k),p_n), & \text{if } X^n \sim P_{\theta,n}. \end{cases} \tag{36}$$

Testing power in the PBM (III)

$$P_{\theta_0,n}\left(\frac{p_n}{1-p_n}\,\frac{1-q_n}{q_n}\right)^{\frac{1}{2}(T_n-S_n)} = Pe^{\frac{1}{2}\lambda_n S_n}\, Pe^{-\frac{1}{2}\lambda_n T_n}$$

with $\lambda_n := \log(1-p_n) - \log(p_n) + \log(q_n) - \log(1-q_n)$. Conclude,

$$\begin{aligned}
&P_{\theta_0,n}\left(\frac{p_{\theta,n}}{p_{\theta_0,n}}(X^n)\right)^{1/2} \\
&\quad = \left(\left(1-p_n+p_n\left(\frac{1-p_n}{p_n}\,\frac{q_n}{1-q_n}\right)^{1/2}\right)\left(1-q_n+q_n\left(\frac{p_n}{1-p_n}\,\frac{1-q_n}{q_n}\right)^{1/2}\right)\right)^{2k(n-k)} \\
&\quad = \left(\left((1-p_n)+p_n^{1/2}q_n^{1/2}\left(\frac{1-p_n}{1-q_n}\right)^{1/2}\right)\left((1-q_n)+p_n^{1/2}q_n^{1/2}\left(\frac{1-q_n}{1-p_n}\right)^{1/2}\right)\right)^{2k(n-k)} \\
&\quad = \left((1-p_n)(1-q_n)+2p_n^{1/2}q_n^{1/2}(1-p_n)^{1/2}(1-q_n)^{1/2}+p_nq_n\right)^{2k(n-k)},
\end{aligned}$$

Proof of theorem 141

Proof For every $n \geq 1$, $k \geq 1$ and given $\theta_{0,n}$, there exists a test sequence with $a_{n,k} = (1-\mu_n)^{2k(n-k)}$ and

$$\mu_n = p_n + q_n - 2p_n\,q_n - 2(p_n(1-p_n)q_n(1-q_n))^{1/2} \in [0,1].$$

With $z_n = (1-\mu_n)^{n/2}$,

$$P_{\theta_0,n}\Pi(V_n|X^n) \leq \sum_{k=1}^{\lfloor n/2 \rfloor} \binom{n}{k}^2 (1-\mu_n)^{2k(n-k)} \leq \sum_{k=1}^{\lfloor n/2 \rfloor} \binom{n}{k}^2 (1-\mu_n)^{nk}$$

$$\leq \sum_{k=1}^{\lfloor n/2 \rfloor} \binom{2n}{2k} (1-\mu_n)^{nk} \leq \sum_{l=1}^{2n} \binom{2n}{l} z_n^l = (1+z_n)^{2n} - 1$$

The right-hand side goes to zero if (33) is satisfied. □

Posterior detection at rate $k_n = o(n)$

Theorem 143 *For some $\theta_{0,n} \in \Theta_n$, let $X^n \sim P_{\theta_{0,n}}$ for every $n \geq 1$. If we equip all Θ_n with uniform priors and (p_n) and (q_n) are such that,*

$$\frac{n}{k_n}\left(1 - p_n - q_n + 2p_n\, q_n + 2\sqrt{(p_n(1-p_n)q_n(1-q_n))}\right)^{n/2} \to 0, \quad (37)$$

as $n \to \infty$, then,

$$\Pi(W_n|X^n) \xrightarrow{P_0} 0, \text{ as } n \to \infty, \quad (38)$$

i.e. the posterior detects $\theta_{0,n}$ at rate k_n.

Posterior detection in the Kesten-Stigum phase

Corollary 144 *Under the conditions of theorem 143 with* (p_n) *and* (q_n) *such that,*

$$n\Big(p_n + q_n - 2p_n\, q_n - 2\sqrt{(p_n(1-p_n)q_n(1-q_n))}\Big) \to \infty, \tag{39}$$

as $n \to \infty$*, then there exists a sequence* $k_n = o(n)$ *such that,*

$$\Pi\Big(\, k_n(\theta_n, \theta_{0,n}) \geq k_n \,\Big|\, X^n \,\Big) \xrightarrow{P_0} 0,$$

i.e. the posterior detects $\theta_{0,n}$*.*

Sharpness of the sparsity condition for detection

Note as $p_n, q_n \to 0$, we may expand,

$$\sqrt{p_n} - \sqrt{q_n} = \frac{1}{2\sqrt{\frac{1}{2}(p_n + q_n)}}(p_n - q_n) + O(|p_n - q_n|^2).$$

which means that,

$$\mu_n = (\sqrt{p_n} - \sqrt{q_n})^2 + O(n^{-2}) = \frac{(p_n - q_n)^2}{2(p_n + q_n)} + O(n^{-2})$$

Eq. (39), $n\mu_n \to \infty$ is also a *necessary condition* for the possibility of detection, (see eq. (30)).

Lecture X

Uncertainty quantification for communities

In this lecture we delve deeper into frequentist inference with the posterior in the planted bi-section model: we explore frequentist coverage of credible sets of communities, their enlargements and their diameters. We distinguish the cases where the posterior known to converge with a certain rate of mis-assignment and where that is not known. Remote contiguity plays a role close to the Erdös-Rényi submodel.

arxiv:1810.09533 (MATH.ST)

Credible sets in the planted bi-section model

Distinguish theorems with posteriors convergence as a condition and theorems without such conditions. Recall:

Exact recovery whenever

$$\left(1+\left(1-p_n-q_n+2p_n\,q_n+2\sqrt{p_n(1-p_n)q_n(1-q_n)}\right)^{n/2}\right)^{2n}\to 1,$$

Detection at (known) rate $k_n=o(n)$ whenever

$$\frac{n}{k_n}\left(1-p_n-q_n+2p_n\,q_n+2\sqrt{(p_n(1-p_n)q_n(1-q_n))}\right)^{n/2}\to 0,$$

When $(p_n),(q_n)$ (and (k_n)) are not known, still theorems?

Credible sets of minimal order

Most natural minimal-order credible set $E_n(X^n)$, $\alpha \in [0,1]$, given X^n

calculate $\Pi(\{\theta_n\}|X^n)$ of all $\theta_n \in \Theta_n$,

order by decreasing posterior weights: $\Theta_n = \{\theta_{n,1}, \theta_{n,2}, \ldots, \theta_{n,|\Theta_n|}\}$

define $E_n(X^n) = \{\theta_{n,1}, \ldots, \theta_{n,m(X^n)}\}$,

$$m(X^n) = \min\Big\{1 \le m \le |\Theta_n| : \Pi(\{\theta_{n,1}, \ldots, \theta_{n,m}\}|X^n) \ge 1-\alpha\Big\}.$$

Show that order of $E_n(X^n)$ is upper bounded

Credible sets in the Chernoff-Hellinger phase

If the posteriors concentrate amounts of mass on $\{\theta_{0,n}\}$ arbitrarily close to one with growing n, then a sequence of credible sets of a certain, fixed level contains $\theta_{0,n}$ for large enough n.

Proposition 145 *Let $0 < \epsilon \leq 1$ be given. Suppose*

$$\Pi(\theta = \theta_{0,n} | X^n) \xrightarrow{P_{\theta_{0,n}}} 1.$$

Then, for any (D_n) of credible sets of levels $1 - \epsilon$

$$P_{\theta_{0,n}}\Big(\theta_{0,n} \in D_n(X^n) \Big) \to 1,$$

Conclude In the Chernoff-Hellinger phase credible sets of minimal order *are* $\{\theta_0\}$ with high $P_{\theta_{0,n}}$-probability.

Credible sets of minimal diameter

Metric on Θ_n $k : \Theta_n \times \Theta_n \to \{0, 1, \ldots, \lfloor n/2 \rfloor\}$ (Recall $k(\theta_n, \eta_n)$ is smallest number of pair-exchanges between two representations θ_n and η_n in Θ_n).

Spheres and Balls with $V_{n,k}(\theta_n) = \{\theta'_n \in \Theta_n : k(\theta'_n, \theta_n) = k\}$ define

$$B_n(\theta_n) = \bigcup_{k=0}^{k_n} V_{n,k}(\theta_n),$$

Enlargement C_n of $D_n \subset \Theta_n$ by B_n

$$C_n = \Big\{\theta_n \in \Theta_n : \exists_{\eta_n \in D_n},\, k(\eta_n, \theta_n) \leq k_n\Big\},$$

Diameter $\mathrm{diam}_n(C)$ of a subset $C \subset \Theta_n$

$$\mathrm{diam}_n(C) = \max\Big\{k(\theta_n, \eta_n) : \theta_n, \eta_n \in C\Big\}.$$

Credible sets in the Kesten-Stigum phase

If the posterior detects communities with rate (k_n) then the k_n-enlargements (C_n) are consistent confidence sets.

Theorem 146 *Suppose that* $0 < \epsilon \leq 1$ *and*

$$\Pi\Big(k(\theta_n, \theta_{0,n}) \leq k_n \,\Big|\, X^n \Big) \xrightarrow{P_{\theta_{0,n}}} 1$$

Let (D_n) *be credible sets of levels* $1-\epsilon$ *of minimal diameters. Then with high* $P_{\theta_{0,n}}$*-probability*

$$\mathrm{diam}_n(D_n) \leq 2k_n$$

and the k_n*-enlargements* C_n *of the* D_n *satisfy,*

$$P_{\theta_{0,n}}\Big(\theta_{0,n} \in C_n(X^n) \Big) \to 1,$$

Proof of theorem 146 (I)

Let $n \geq 1$, $\theta_n \in \Theta_n$ and $x^n \in \mathscr{X}_n$ be given

Let $k_n(\theta_n, x^n)$ denote the radius of the smallest ball in Θ_n centred on θ_n of posterior mass at least $1 - \epsilon$.

Define $\widehat{\theta}_n(x^n)$ as the centre point of a smallest credible ball $B_n(\widehat{\theta}_n(x^n))$ in Θ_n of level $1 - \epsilon$:

$$k_n(\widehat{\theta}_n(x^n)) = \min\Big\{k_n(\theta_n, x^n) : \theta_n \in \Theta_n\Big\},$$

Then

$$P_{\theta_0,n}\Big(\Pi(B_n(\widehat{\theta}_n(X^n))|X^n) \geq 1 - \epsilon\Big) = 1,$$

for all $n \geq 1$.

Proof of theorem 146 (II)

Posterior convergence the ball $B_n(\theta_{0,n})$ is a credible ball of level $1-\epsilon$ for large enough n. Therefore,

$$k_n(\widehat{\theta}_n(x^n)) \leq k_n.$$

Posterior convergence the balls $B_n(\theta_{0,n})$ of radii k_n centred on $\theta_{0,n}$, satisfy

$$P_{\theta_{0,n}}\Big(\Pi(B_n(\theta_{0,n})|X^n) > \epsilon\Big) \to 1.$$

Conclude that, with high $P_{\theta_{0,n}}$-probability,

$$B_n(\theta_{0,n}) \cap B_n(\widehat{\theta}_n(X^n)) \neq \varnothing,$$

implying asymptotic coverage of $\theta_{0,n}$ for the k_n-enlargement $C_n(X^n)$ of $B_n(\widehat{\theta}_n(X^n))$.

Summary: credible sets from converging posteriors

Proposition 145 requires exact recovery,

$$\left(1+\left(1-p_n-q_n+2p_n\,q_n+2\sqrt{p_n(1-p_n)q_n(1-q_n)}\right)^{n/2}\right)^{2n}\to 1,$$

Theorem 146 requires detection at known rate k_n,

$$\frac{n}{k_n}\left(1-p_n-q_n+2p_n\,q_n+2\sqrt{(p_n(1-p_n)q_n(1-q_n))}\right)^{n/2}\to 0,$$

Theorem 146 cannot be applied just based on detection:

$$n\left(p_n+q_n-2p_n\,q_n-2\sqrt{(p_n(1-p_n)q_n(1-q_n))}\right)\to\infty,$$

is not strong enough.

Confidence sets from credible sets

But even if testing cannot serve as a condition, the use of credible sets as confidence sets remains valid, as long as credible levels grow to one fast enough. Here $b_n = |\Theta_n|^{-1} = (\frac{1}{2}\binom{2n}{n})^{-1}$.

Proposition 147 *With some $\theta_{0,n} \in \Theta_n$, let D_n be a sequence of credible sets, such that,*

$$\Pi(D_n(X^n)|X^n) \geq 1 - a_n,$$

for some sequence (a_n) with $a_n = o(b_n)$. Then,

$$P_{\theta_{0,n}}\Big(\theta_0 \in D_n(X^n)\Big) \geq 1 - b_n^{-1} a_n.$$

Proof of proposition 147

Proof If $\theta_{0,n} \notin D_n(X^n)$ then

$$\Pi(\{\theta_{0,n}\}|X^n) \leq a_n$$

Then by Bayes's Rule (1)

$$\begin{aligned} P_{\theta_0,n}\Big(\theta_0 \in \Theta \setminus D_n(X^n)\Big) &= P_n^{\Pi|\{\theta_0\}}\Big(\theta_0 \in \Theta \setminus D_n(X^n)\Big) \\ &= b_n^{-1} \int_{\{\theta_{0,n}\}} P_{\theta,n}\Big(\theta_0 \in \Theta \setminus D_n(X^n)\Big)\, d\Pi_n(\theta) \\ &\leq b_n^{-1} \int_{\mathscr{X}_n} \Big(1\{\theta_0 \in \Theta_n \setminus D_n(x^n)\}\, \Pi(\{\theta_{0,n}\}|x^n)\, dP_n^{\Pi}(x^n) \\ &\leq b_n^{-1} a_n \end{aligned}$$

□

Credible set enlargement and remote contiguity

To mitigate the lower bound on credible levels, we shall use enlarged credible sets.

Competing influences when enlarging

- prior masses $b_n = \Pi_n(B_n(\theta_{0,n}))$ become larger relaxing the rate at which credible levels are required to go to one.
- greater diversity of likelihood ratios with random fluctuations that take them further away from one

Close to the Erdös-Rényi submodel fluctuations of likelihood ratios are small so remote contiguity is achieved relatively easily

A 'statistical phase'

Differences of within-class and between-class edge probabilitites

$$p_n - q_n = o(n^{-1}), \tag{40}$$

while satisfying also the condition that,

$$p_n^{1/2}(1-p_n)^{1/2} + q_n^{1/2}(1-q_n)^{1/2} = o(n|p_n - q_n|). \tag{41}$$

In this regime, $p_n, q_n \to 0$ or $p_n, q_n \to 1$.

If $p_n, q_n \to 0$ as in the sparse phases, (41) amounts to,

$$n(p_n^{1/2} - q_n^{1/2}) \to \infty,$$

i.e. $p_n - q_n$ may not converge to zero too fast.

Remote contiguity

Define,

$$\rho_n = \min\left\{\left(\frac{1-p_n}{p_n}\frac{q_n}{1-q_n}\right), \left(\frac{p_n}{1-p_n}\frac{1-q_n}{q_n}\right)\right\} = e^{-|\lambda_n|}$$

and,

$$\alpha_n = \frac{1}{|B_n|}\sum_{k=0}^{k_n}\binom{n}{k}^2 2k(n-k)$$

and (with any $C > 1$)

$$d_n = \rho_n{}^{C\alpha_n|p_n-q_n|},$$

Lemma 148 *Let (k_n) be given and assume (41). Then for any $\theta_{0,n} \in \Theta_n$, $B_n = B_n(\theta_{0,n})$*

$$P_{\theta_{0,n}} \lhd d_n^{-1} P_n^{\Pi|B}.$$

Remote contiguity and enlargement of credible sets

Theorem 149 *Let* $\theta_{0,n} \in \Theta_n$ *and* $0 \leq a_n \leq 1$, $b_n > 0$ *such that* $a_n = o(b_n)$ *be given. Choose priors* Π_n *and let* D_n *denote level-*$1 - a_n$ *credible sets in* Θ_n. *For all* $\theta \in \Theta$, *let* $B_n = \{B_n(\theta_n) \in \mathscr{G}_n : \theta_n \in \Theta_n\}$ *be such that,*

(i.) $\Pi_n(B_n(\theta_0)) \geq b_n$,

(ii.) for some $d_n \downarrow 0$ *such that* $b_n^{-1} a_n = o(d_n)$

$$P_{\theta_0,n} \lhd d_n^{-1} \, P_n^{\Pi|B(\theta_0)}$$

Then the B_n*-enlargements* C_n *of* D_n *are asymptotically consistent*

$$P_{\theta_0,n}\Big(\theta_0 \in C_n(X^n) \Big) \to 1. \tag{42}$$

Suitable enlarged credible sets are confidence sets

Theorem 150 *Let $\theta_{0,n} \in \Theta_n$ be given and assume that*

$$p_n - q_n = o(n^{-1}),$$

$$p_n^{1/2}(1-p_n)^{1/2} + q_n^{1/2}(1-q_n)^{1/2} = o(n|p_n - q_n|).$$

Choose (k_n) and $C > 1$ to fix (d_n). Let D_n be any credible sets of levels $1 - a_n$ with a_n such that,

$$\Pi_n(B_n(\theta_0))^{-1} a_n = o(d_n)$$

Then the B_n-enlargements $C_n(X^n)$ of $D_n(X^n)$ satisfy,

$$P_{\theta_0,n}\Big(\theta_0 \in C_n(X^n)\Big) \to 1,$$

So the enlargements C_n are asymptotic confidence sets

Credible level improvement factor (I)

Which a_n instead of $a_n = o(\binom{2n}{n}^{-1}) \approx o(4^{-n})$?

Assume $k_n = \beta\, n$ for some fixed $\beta \in (0,1)$

Stirling's approximation lower bound

$$\frac{\Pi_n(B_n)}{\Pi_n(\{\theta_{0,n}\}))} = \sum_{k=0}^{k_n} \binom{n}{k}^2 \geq \binom{n}{k_n}^2 \geq \frac{1}{2\pi n} \frac{1}{\beta(1-\beta)} f(\beta)^n,$$

where $f : (0,1) \to (1,4]$ is given by,

$$f(\beta) = (1-\beta)^{-2(1-\beta)} \beta^{-2\beta}.$$

Credible level improvement factor (II)

Approximate $\alpha_n \approx 2k_n(n-k_n)$ and assume $\lambda_n = O(1)$ Because $n|p_n - q_n| \to 0$, we also have,

$$d_n = \rho_n^{C\alpha_n|p_n-q_n|} \approx \rho_n^{2Cn^2\beta(1-\beta)|p_n-q_n|} = e^{-|\lambda_n|o(n)}.$$

d_n is sub-exponential and does not play a role.

Given X^n and $\beta \in (0,1)$, sets $D_n(X^n)$ such that

$$\Pi(D_n(X^n)|X^n) \geq 1 - a_n f(\beta)^n$$

have enlarged confidence sets $(C_n(X^n))$ that cover $\theta_{0,n}$ with high probability.

Credible levels
before $1 - a_n \approx 1 - o(4^{-n})$
improved $1 - o(c^{-n})$ for any $0 < c < 4$ $(1/2 > \beta > 0)$

Lecture XI

Uniform and pointwise tests

In this lecture we take a closer look at the exact meaning of asymptotic testability: given two alternatives (model subsets) B and V, is there a test sequence that we could use to detect whether the true distribution of the data lies in B or V? This question can be formulated from Bayesian or frequentist (uniform) points of view, and the answers take the form of topological characterizations.

Asymptotic symmetric testing

Observe *i.i.d.* $X^n \sim P^n$, model $P \in \mathscr{P}$. For disjoint $B, V \subset \mathscr{P}$,

$$H_0 : P \in B, \quad \text{or} \quad H_1 : P \in V.$$

Tests $\phi_n : \mathscr{X}^n \to [0,1]$; asymptotically, require:

$$\text{(type-I)} \quad P^n \phi_n \to 0 \text{ for } P \in B, \text{ and,}$$
$$\text{(type-II)} \quad P^n(1-\phi_n) \to 0 \text{ for } P \in V.$$

Equivalently, we want,

> A testing procedure that chooses for B or V based on X^n for every $n \geq 1$, has property (D) if it is wrong only a finite number of times with P^∞-probability one.

Property (D) is sometimes referred to as "discernibility".

Some examples and unexpected answers (I)

Consider non-parametric regression with $f : X \to \mathbb{R}$ and test for smoothness,

$$H_0 : f \in C^1(X \to \mathbb{R}), \quad H_1 : f \in C^2(X \to \mathbb{R}),$$

Consider a non-parametric density estimation with $p : \mathbb{R} \to [0, \infty)$ and test for square-integrability,

$$H_0 : \int x^2 \, p(x) \, dx < \infty, \quad H_1 : \int x^2 \, p(x) \, dx = \infty.$$

Practical problem we cannot use the data to determine with asymptotic certainty, if CLT applies with our data.

Some examples and unexpected answers (II)

Coin-flip $X^n \sim \text{Bernoulli}(p)^n$ with $p \in [0,1]$.

Consider Cover's rational mean problem: test for rationality:

$$H_0 : p \in [0,1] \cap \mathbb{Q}, \quad H_1 : p \in [0,1] \setminus \mathbb{Q}.$$

Consider also Dembo and Peres's irrational alternative:

$$H_0 : p \in [0,1] \cap \mathbb{Q}, \quad H_1 : p \in [0,1] \cap \sqrt{2} + \mathbb{Q},$$

Consider ultimately fractal hypotheses, *e.g.* with Cantor set C,

$$H_0 : p \in C, \quad H_1 : p \in [0,1] \setminus C.$$

The Le Cam-Schwartz theorem

Theorem 151 (Le Cam-Schwartz, 1960) *Let $\mathscr{P}$ be a model for i.i.d. data X^n with disjoint subsets B, V. The folllowing are equivalent:*

i. there exist (uniformly) consistent tests for B vs V,

ii. there is a sequence of $\mathscr{U}_\infty$-uniformly continuous $\psi_n : \mathscr{P} \to [0,1]$,

$$\psi_n(P) \to 1_V(P), \tag{43}$$

(uniformly) for all $P \in B \cup V$.

Topological context uniform space $(\mathscr{P}, \mathscr{U}_\infty)$.

The Dembo-Peres theorem

Theorem 152 (Dembo and Peres, 1995) *Let $\mathscr{P}$ be a model dominated by Lebesgue measure μ for i.i.d. data X^n. Model subsets B, V that are contained in disjoint countable unions of closed sets for Prokhorov's weak topology have tests with property (D). If there exists an $\alpha > 1$ such that $\int (dP/d\mu)^\alpha \, d\mu < \infty$ for all $P \in \mathscr{P}$, then the converse is also true.*

Topological context L^1-weakly compact, dominated model $\mathscr{P}$ with Prokhorov's weak topology.

Three forms of testability

Definition 153 (ϕ_n) *is a uniform test sequence for* B *vs* V*, if,*

$$\sup_{P\in B} P^n\phi_n \to 0, \quad \sup_{Q\in V} Q^n(1-\phi_n) \to 0. \tag{44}$$

Definition 154 (ϕ_n) *is a pointwise test sequence for* B *vs* V*, if,*

$$\phi_n(X^n) \xrightarrow{P} 0, \quad \phi_n(X^n) \xrightarrow{Q} 1, \tag{45}$$

for all $P \in B$ *and* $Q \in V$*.*

Definition 155 (ϕ_n) *is a Bayesian test sequence for* B *vs* V*, if,*

$$\phi_n(X^n) \xrightarrow{P} 0, \quad \phi_n(X^n) \xrightarrow{Q} 1, \tag{46}$$

for Π*-almost-all* $P \in B$ *and* $Q \in V$*.*

Questions

Existence

Existence of uniform tests?

Existence of pointwise tests?

Existence of Bayesian tests?

Construction

How does one model-select? Are there constructive solutions?

Examples

Select the correct directed, acyclical graph in a graphical model; select the right number of clusters in a clustering model.

Uniform testability has exponential power

Proposition 156 *Let $\mathscr{P}$ be a model for i.i.d. data with disjoint B and V. The following are equivalent:*

i. there exists a uniform test sequence (ϕ_n),

$$\sup_{P\in B} P^n\phi_n \to 0, \quad \sup_{Q\in V} Q^n(1-\phi_n) \to 0,$$

ii. there is a exponentially powerful uniform test sequence (ψ_n), i.e. there is a $D > 0$ such that,

$$\sup_{P\in B} P^n\psi_n \leq e^{-nD}, \quad \sup_{Q\in V} Q^n(1-\psi_n) \leq e^{-nD}.$$

The model as a uniform space (I)

Take $\mathscr{X}$ a separable metrizable space, with Borel σ-algebra $\mathscr{B}$.

The class $\mathscr{F}_n$ contains all bounded, $\mathscr{B}^n$-measurable $f : \mathscr{X}^n \to \mathbb{R}$.

For every $n \geq 1$ and $f \in \mathscr{F}_n$, define the entourage,

$$W_{n,f} = \{(P, Q) \in \mathscr{P} \times \mathscr{P} : |P^n f - Q^n f| < 1\}.$$

Defines uniformity $\mathscr{U}_n$ (with topology $\mathscr{T}_n$). Take $\mathscr{U}_\infty = \cup_{n \geq 1} \mathscr{U}_n$.

$$P \to Q \text{ in } \mathscr{T}_\infty \quad \Leftrightarrow \quad \int f \, dP^n \to \int f \, dQ^n,$$

for all $n \geq 1$ and all $f \in \mathscr{F}_n$. Note also,

$$\mathscr{U}_C \subset \mathscr{U}_1 \subset \cdots \subset \mathscr{U}_\infty \subset \mathscr{U}_{TV}.$$

The model as a uniform space (II)

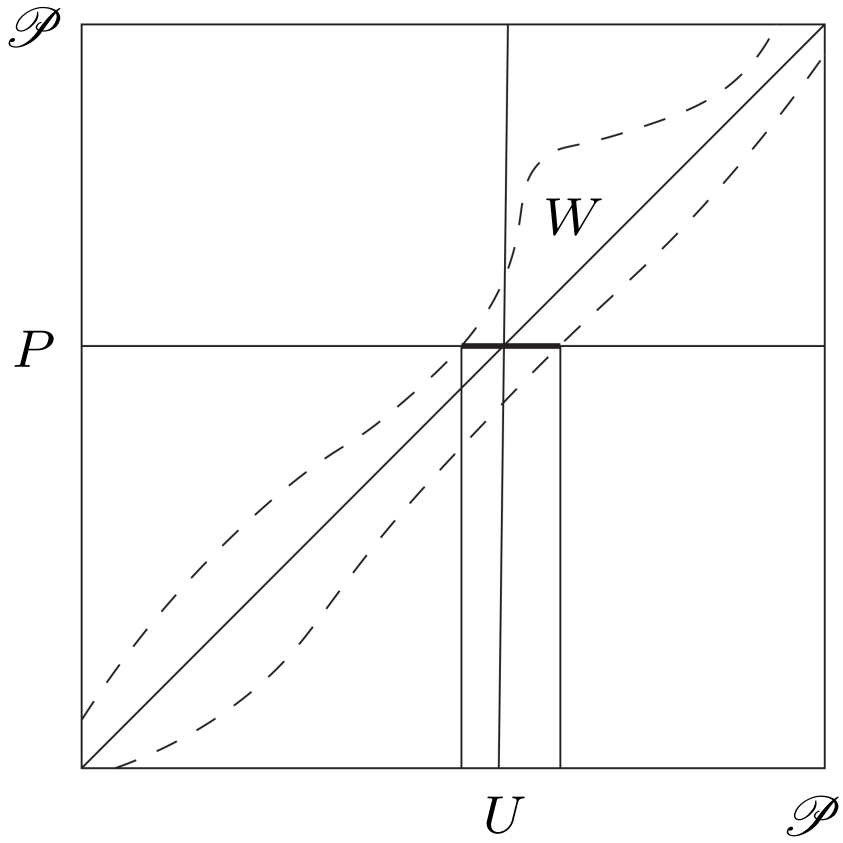


Fig 1. Let $P \in \mathscr{P}$ and entourage W be given. A neighbourhood U corresponds to $U = \{Q \in \mathscr{P} : (Q, P) \in W\}$

Uniform separation (I)

Definition 157 *Subsets $B, V \subset \mathscr{P}$ are uniformly separated by $\mathscr{U}_\infty$, if there exists an entourage $W \in \mathscr{U}_\infty$ such that,*

$$(B \times V \cup V \times B) \cap W = \varnothing.$$

In other words, there are $J, m \geq 1$, $\epsilon > 0$ and bounded, measurable functions $f_1, \ldots, f_J : \mathscr{X}^m \to [0,1]$ such that, for any $P, Q \in B \cup V$, if,

$$\max_{1 \leq j \leq J} \left| P^m f_j - Q^m f_j \right| < \epsilon,$$

then either $P, Q \in B$, or $P, Q \in V$. (If the model is $\mathscr{T}_\infty$-compact, $m = 1$ suffices).

Uniform separation (II)

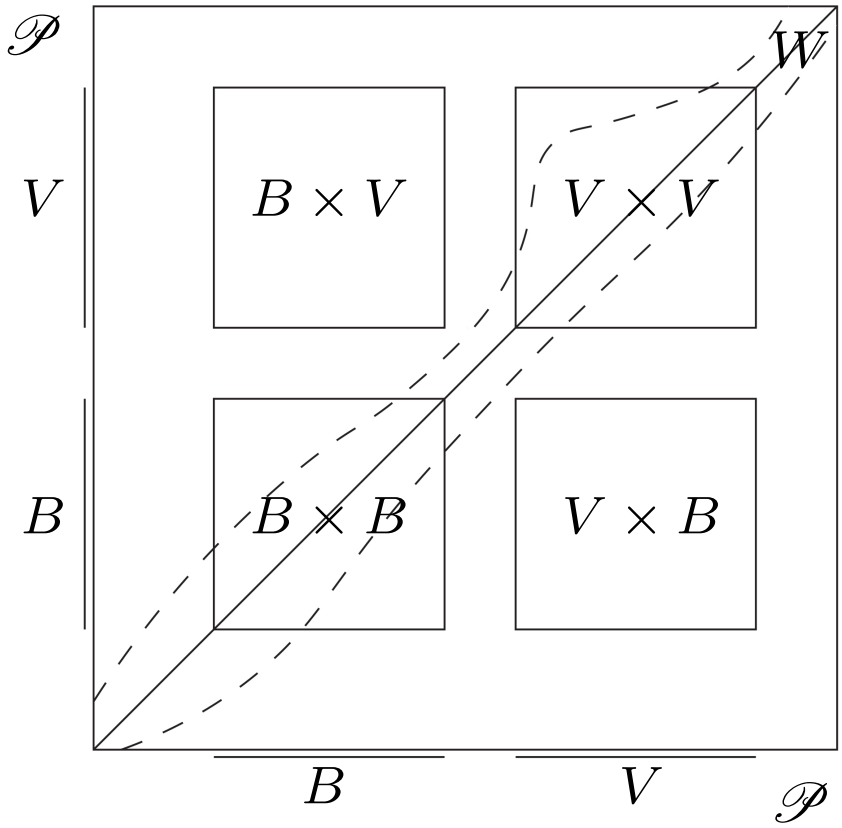


Fig 2. Let $B, V \subset \mathscr{P}$ and entourage W be given. W separates B and V if $B \times V$ and $V \times B$ do not meet W.

Characterisation of uniform testability

Theorem 158 *Let $\mathscr{P}$ be a model for i.i.d. data with disjoint B and V. The following are equivalent:*

(i.) there exist uniform tests ϕ_n for B versus V,
(ii.) the subsets B and V are uniformly separated by $\mathscr{U}_\infty$.

Corollary 159 (Parametrised models) *Suppose $\mathscr{P} = \{p_\theta : \theta \in \Theta\}$, with (Θ, d) compact, metric space and $\theta \to P_\theta$ identifiable and $\mathscr{T}_\infty$-continuous, (that is, for every $f \in \mathscr{F}_n$, $\theta \mapsto \int f \, dP_\theta^n$ is continuous). If $B_0, V_0 \subset \Theta$ with $d(B_0, V_0) > 0$, then the images $B = \{P_\theta : \theta \in B_0\}$, $V = \{P_\theta : \theta \in V_0\}$ are uniformly testable.*

Closures are important

Proposition 160 *Let $\mathscr{P}$ be a model for i.i.d. data and let B, V be disjoint model subsets with $\mathscr{T}_\infty$-closures $\overline{B}$ and $\overline{V}$. If B, V are uniformly separated by $\mathscr{U}_\infty$, then $\overline{B} \cap \overline{V} = \varnothing$. If $\mathscr{P}$ is relatively $\mathscr{T}_\infty$-compact, the converse is also true.*

Theorem 161 (Dunford-Pettis) *Assume $\mathscr{P}$ is dominated by a probability measure Q with densities in $\mathscr{P}_Q \subset L^1(Q)$; $\mathscr{P}_Q$ is relatively weakly compact, if and only if, for every $\epsilon > 0$ there is an $M > 0$ such that,*

$$\sup_{P \in \mathscr{P}} \int_{\{dP/dQ > M\}} \frac{dP}{dQ}\, dQ < \epsilon,$$

that is, $\mathscr{P}_Q$ is uniformly Q-integrable.

Pointwise testability: equivalent formulations

Proposition 162 *Let $\mathscr{P}$ be a model for i.i.d. data and let B, V be disjoint model subsets. The following are equivalent:*

i. there are tests (ϕ_n) such that, for all $P \in B$ and $Q \in V$,

$$P^n \phi_n \to 0, \quad Q^n(1 - \phi_n) \to 0,$$

ii. there are tests (ϕ_n) such that, for all $P \in B$ and $Q \in V$,

$$\phi_n(X^n) \xrightarrow{P} 0, \quad (1 - \phi_n(X^n)) \xrightarrow{Q} 0,$$

iii. there are tests (ϕ_n) such that, for all $P \in B$ and $Q \in V$,

$$\phi_n(X^n) \xrightarrow{P\text{-a.s.}} 0, \quad (1 - \phi_n(X^n)) \xrightarrow{Q\text{-a.s.}} 0.$$

Pointwise testability from consistent estimators

Consistent estimators $\hat{P}_n : \mathscr{X}^n \to \mathscr{P}$: for all P and nbd U of P,

$$P^n\Big(\hat{P}_n(X^n) \in U\Big) \to 1, \text{ as } n \to \infty.$$

For open $B, V \subset \mathscr{P}$, define $\phi_n(X^n) = 1\{\hat{P}_n \in V\}$. For any $P \in B$, B is a neighbourhood of P so $P^n\phi_n = P^n(\hat{P}_n \in V) \leq P^n(\hat{P}_n \notin B) \to 0$. For any $Q \in V$, $Q^n(1 - \phi_n) \to 0$. So (ϕ_n) is a pointwise test sequence for B vs V.

Restrict to $\mathscr{P}' = B \cup V$, then B and V are *clopen* sets.

Proposition 163 *If $P \in \mathscr{P}$ can be estimated consistently and B is clopen, there exist pointwise tests for B vs its complement.*

Necessary conditions: pointwise non-testability

Suppose that there exist pointwise tests (ϕ_n) for B, V. Define,

$$g_n : \mathscr{P} \to [0,1] : P \mapsto P^n \phi_n,$$

which are all $\mathscr{U}_\infty$-uniformly continuous.

Proposition 164 *If there is a pointwise test (ϕ_n) for B vs V, then B, V are both G_δ- and F_σ-sets with respect to $\mathscr{T}_\infty$ (in the subspace $B \cup V$).*

Corollary 165 *Suppose $\mathscr{P} = B \cup V$ is Polish in the $\mathscr{T}_\infty$-topology. Pairs B, V that are pointwise testable, are both Polish spaces.*

Corollary 166 *If there exists a Baire subspace D of $\mathscr{P}$ in which both $D \cap B$ and $D \cap V$ are dense, then B is not testable versus V.*

Pointwise non-testability: examples (I)

Example 167 *Is Cover's rational means problem testable?*

Dunford-Pettis theorem shows that $\mathscr{P}$ is $\mathscr{T}_\infty$-compact and $[0,1] \to \mathscr{P} : p \mapsto P_p$ is a $\mathscr{T}_\infty$-homeomorphism. Since $[0,1]$ is a complete metric space, $\mathscr{P}$ is a Baire space for the $\mathscr{T}_\infty$-topology. Because both $[0,1] \cap \mathbb{Q}$ and $[0,1] \setminus \mathbb{Q}$ are dense in $[0,1]$, the images $\mathscr{P}_0 := \{P_p : p \in [0,1] \cap \mathbb{Q}\}$ and $\mathscr{P}_1 := \{P_p : p \in [0,1] \setminus \mathbb{Q}\}$ are $\mathscr{T}_\infty$-dense in $\mathscr{P}$: there is no pointwise test for $p \in [0,1] \cap \mathbb{Q}$ versus $p \in [0,1] \setminus \mathbb{Q}$.

Pointwise non-testability: examples (II)

Example 168 *Is Dembo and Peres's irrational alternative testable?*

Any countable $\mathscr{P}$ is Polish in the discrete topology. Any subset B of $\mathscr{P}$ is a countable union of closed sets ($B = \cup_{b \in B} \{b\}$), so it remains possible that there exists a pointwise test for Dembo and Peres's problem.

Example 169 *Is Cantor's fractal alternative testable?*

The interval $[0, 1]$ is Polish and $\mathscr{P}$ is homeomorphic. The Cantor set C is closed and its complement is open. Open sets in metrizable spaces are F_σ-sets. So it remains possible there exists a pointwise test for Cantor's fractal alternative.

Pointwise non-testability: examples (III)

Example 170 *Is integrability of a real-valued X, $P|X| < \infty$, testable?*

Model $\mathscr{P} = \{$all probability distributions on $\mathbb{R}\}$. $\mathscr{P}$ is Baire space for $\mathscr{T}_{TV}$. Define,

$$B = \{P \in \mathscr{P} : P|X| < \infty\}, \quad V = \{P \in \mathscr{P} : P|X| = \infty\}.$$

B cannot be tested versus V.

Namely Let $P \in B$ and $Q \in V$ be given. For any $0 < \epsilon < 1$, $P' = (1-\epsilon)P + \epsilon Q$ satisfies $\|P' - P\| = \epsilon\|(P+Q)\| \leq 2\epsilon$, but $P' \in V$. Conclude that V lies $\mathscr{T}_{TV}$-dense in $\mathscr{P}$.

Conversely, Q is tight, so for every $\epsilon > 0$, there exists an $M > 0$ such that $|Q(A) - Q(A||X| \leq M)| < \epsilon$ for all measurable $A \subset \mathbb{R}$. Since $Q(\cdot||X| \leq M) \in B$, we also see that B lies $\mathscr{T}_{TV}$-dense in $\mathscr{P}$.

Pointwise testability in dominated models

Definition 171 *The testing problem has a (uniform) representation on X, if there exists a $\mathscr{T}_\infty$-(uniformly-)continuous, surjective map $f : B \cup V \to X$ such that $f(B) \cap f(V) = \varnothing$.*

Definition 172 *The model is parametrised by Θ, if there exists a $\mathscr{T}_\infty$-continuous bijection $P_\cdot : \Theta \to \mathscr{P}$ (i.e. for every $m \geq 1$ and measurable $f : \mathscr{X}^m \to [0,1]$, the map $\theta \mapsto \int f \, dP_\theta^m$ is continuous).*

If Θ is compact, any parametrization is a homeomorphism, so the inverse gives rise to representations of testing problems in $\mathscr{P}$.

Characterisation of pointwise testability

Theorem 173 *Let $\mathscr{P}$ be a dominated model for i.i.d. data with disjoint B, V. The following are equivalent,*

i. *there exists a pointwise test for B vs V,*

ii. *the problem has a representation $f : B \cup V \to X$ on a normal space X and there exist disjoint F_σ-sets $B', V' \subset X$ such that $f(B) \subset B', f(V) \subset V'$,*

iii. *the problem has a uniform representation $\psi : B \cup V \to X$ on a separable, metrizable space X with $\psi(B), \psi(V)$ both F_σ- and G_δ-sets.*

Pointwise testability: corollaries (I)

Corollary 174 *Suppose that $\mathscr{P}$ is dominated and there exist disjoint F_σ-sets B', V' in the completion $\hat{\mathscr{P}}$ (for $\mathscr{U}_\infty$) with $B \subset B'$, $V \subset V'$. Then B is pointwise testable versus V.*

Corollary 175 *Suppose that $\mathscr{P}$ is dominated and complete (for $\mathscr{U}_\infty$) with disjoint subsets B, V. Then B is pointwise testable versus V, if and only if, there exist disjoint F_σ-sets $B', V' \subset \mathscr{P}$ with $B \subset B'$, $V \subset V'$.*

Pointwise testability: corollaries (II)

Corollary 176 *Suppose that $\mathscr{P}$ is dominated and TV-totally-bounded. Then disjoint $B, V \subset \mathscr{P}$ are pointwise testable, if and only if, B, V are both F_σ- and G_δ-sets in $B \cup V$ (for $\mathscr{T}_{TV}$).*

Corollary 177 *Suppose that $\mathscr{P}$ is dominated by a probability measure, with a uniformly integrable family of densities. Then disjoint $B, V \subset \mathscr{P}$ are pointwise testable, if and only if, B, V are both F_σ- and G_δ-sets in $B \cup V$ (for $\mathscr{T}_C$).*

Pointwise testability: examples (I)

Example 178 *Is independence of two events A and B testable?*

Let $A, B \in \mathscr{B}$ be msb subsets. Consider,

$$H_0 : P(A \cap B) = P(A)P(B), \quad H_1 : P(A \cap B) \neq P(A)P(B).$$

Define $\mathscr{U}_1$-continuous $f_i : \mathscr{P} \to [0,1]$, $(i = 1,2,3)$,

$$f_1(P) = P(A \cap B), \quad f_2(P) = P(A), \quad f_3(P) = P(B),$$

and continuous $g : [0,1]^3 \to [1,-1]$, $g(x_1,x_2,x_3) = x_1 - x_2x_3$. Now,

$$h : \mathscr{P} \to [0,1] : P \mapsto |g \circ (f_1, f_2, f_3)(P)|,$$

is $\mathscr{U}_1$-continuous. Then $B = h^{-1}(\{0\})$ is closed (for $\mathscr{T}_\infty$) and (since the complement V' is open in $[0,1]$, it is F_σ, so) $V = h^{-1}(V')$ is F_σ (for $\mathscr{T}_\infty$). So independence of events A and B is asymptotically testable.

Pointwise testability: examples (II)

Example 179 *Is independence of real-valued X and Y testable?*

Let $A_k \in \sigma_X, B_l \in \sigma_Y$ be generators. Consider,

$H_0 : \forall_{k,l} P(A_k \cap B_l) = P(A_k)P(B_l)$, $H_1 : \exists_{k,l} P(A_k \cap B_l) \neq P(A_k)P(B_l)$.

Define $\mathscr{U}_1$-continuous $f_{kl,i} : \mathscr{P} \to [0,1]$, $(i = 1,2,3)$,

$$f_{kl,1}(P) = P(A_k \cap B_l), \quad f_{k,2}(P) = P(A_k), \quad f_{l,3}(P) = P(B_l),$$

and continuous $g : [0,1]^3 \to [1,-1]$, $g(x_1,x_2,x_3) = x_1 - x_2x_3$. Now,

$$h : \mathscr{P} \to [0,1]^{\mathbb{N}} : P \mapsto (|g \circ (f_{kl,1}, f_{k,2}, f_{l,3})(P)| : k,l \geq 1),$$

is $\mathscr{U}_1$-continuous. Then $B = h^{-1}(\{0\})$ is closed (for $\mathscr{T}_\infty$) and (since the complement V' is open in $[0,1]^{\mathbb{N}}$, it is F_σ, so) $V = h^{-1}(V')$ is F_σ (for $\mathscr{T}_\infty$). So independence of X and Y is asymptotically testable.

Lecture XII

Bayesian tests and posterior model selection

Having considered the *existence* of frequentist tests in detail in the previous lecture, now we turn to *construction* of such tests. Posterior odds are *optimal* Bayesian tests and remote contiguity allows for frequentist interpretation of their asymptotic conclusions. Posterior odds or Bayes factors can be used by the frequentist to select a model in an asymptotically correct way, if the prior induces remote contiguity and Bayesian tests of compatible rates.

Bayesian testability: equivalent formulations

Theorem 180 *Let a model* $(\mathscr{P}, \mathscr{G}, \Pi)$ *with* $B, V \in \mathscr{G}$ *be given, with* $\Pi(B) > 0, \Pi(V) > 0$. *The following are equivalent,*

i. there exist Bayesian tests for B *vs* V,

ii. there are tests ϕ_n *such that for* Π*-almost-all* $P \in B, Q \in V$,

$$P^n \phi_n \to 0, \quad Q^n(1 - \phi_n) \to 0,$$

iii. there are tests $\phi_n : \mathscr{X}^n \to [0, 1]$ *such that,*

$$\int_B P^n \phi_n \, d\Pi(P) + \int_V Q^n(1 - \phi_n) \, d\Pi(Q) \to 0,$$

iv. for Π*-almost-all* $P \in B$, $Q \in V$,

$$\Pi(V|X^n) \xrightarrow{P} 0, \quad \Pi(B|X^n) \xrightarrow{Q} 0.$$

Characterisation of Bayesian testability

Definition 181 *Given model* $(\mathscr{P}, \mathscr{G}, \Pi)$. *An event* $B \in \mathscr{B}^\infty$ *is called a* Π*-zero-one set, if* $P^\infty(B) \in \{0,1\}$, *for* Π*-almost-all* $P \in \mathscr{P}$. *A model subset* $G \in \mathscr{G}$ *is called a* Π*-one set if there is a* Π*-zero-one set* B *such that* $G = \{P \in \mathscr{P} : P^\infty(B) = 1\}$.

Proposition 182 Doob (1948), Breiman-Le Cam-Schwartz (1960)) *Let* $(\mathscr{P}, \mathscr{G}, \Pi)$ *be given. Let* V *be a* Π*-one set. Then, for* Π*-almost-all* $P \in \mathscr{P}$,

$$\Pi(V|X^n) \xrightarrow{P\text{-a.s.}} 1_V(P). \tag{47}$$

Theorem 183 *(Le Cam (1986)) Let the model* $\mathscr{P}$ *be a completely regular space with Borel* σ*-algebra* $\mathscr{G}$ *and Radon prior* Π *and hypotheses* B, V. *There is a Bayesian test sequence for* B *vs* V, *if and only if,* B, V *are* $\mathscr{G}$*-measurable.*

Proof of proposition 182 (I)

Products $\Omega_n = \mathscr{P} \times \mathscr{X}^n$ with product σ-algebras $\sigma(\mathscr{G} \times \mathscr{B}^n)$ and with sub-σ-algebras $\mathscr{F}_n = \{\varnothing, \mathscr{P}\} \times \mathscr{B}^n$. The product $\Omega = \mathscr{P} \times \mathscr{X}^\infty$ with full product σ-algebra $\mathscr{F} = \sigma(\mathscr{G} \times \mathscr{B}^\infty)$ has a sub-σ-algebra $\mathscr{F}_\infty = \{\varnothing, \mathscr{P}\} \times \mathscr{B}^\infty$, and the filtration $\{\mathscr{F}_n : n \geq 1\}$ has limit $\mathscr{F}_\infty$.

Define $S : \mathscr{F} \mapsto [0, 1]$,

$$S(A \times B) = \int_A P^\infty(B)\, d\Pi(P),$$

($A \in \mathscr{G}$ and $B \in \mathscr{B}^\infty$).

The $\mathscr{F}$-msb map $(P, x_\infty) \mapsto 1_V(P)$ is such that,

$$\Pi(V|X^n) = E[\,1_V|\mathscr{F}_n]$$

Proof of proposition 182 (II)

so posteriors form an L^∞-martingale relative to S. Doob's martingale convergence theorem there exists an $\mathscr{F}_\infty$-measurable f_V such that $\Pi(V|X^n) \to f_V$, S-almost-surely.

Since V is a Π-one set, there is a $B \in \mathscr{B}^\infty$ such that $1_V(P) = 1_B(x_\infty)$, S-almost-surely.

$$\Pi(V|X^n) = E_S\Big[1_V \,\Big|\, \mathscr{F}_n\Big] = E_S\Big[1_B \,\Big|\, \mathscr{F}_n\Big] \to E_S\Big[1_B \,\Big|\, \mathscr{F}_\infty\Big] = 1_B = 1_V,$$

P-almost-surely for Π-almost-all P (by Fubini's theorem).

Bayesian testing power

Denote the density for the local prior predictive distribution $P_n^{\Pi|B}$ with respect to $\mu_n = P_n^{\Pi|B} + P_n^{\Pi|V}$ by $p_{B,n}$, and similar for $P_n^{\Pi|V}$.

Proposition 184 *Let $(\mathscr{P}, \mathscr{G}, \Pi)$ be a model with measurable B, V. There are tests ϕ_n such that,*

$$\begin{aligned} &\int_B P^n \phi_n \, d\Pi(P) + \int_V Q^n (1 - \phi_n) \, d\Pi(Q) \\ &\qquad \leq \int \Big(\Pi(B)\, p_{B,n}(x) \Big)^{\alpha} \Big(\Pi(V)\, p_{V,n}(x) \Big)^{1-\alpha} \, d\mu_n(x), \end{aligned} \tag{48}$$

for every $n \geq 1$ and any $0 \leq \alpha \leq 1$.

Posterior odds are optimal Bayesian tests

Proposition 185 *For every $n \geq 1$, the test,*

$$\phi_n(X^n) = 1\{X^n : \Pi(V|X^n) \geq \Pi(B|X^n)\},$$

based on posterior odds has optimal Bayesian testing power.

Proof Decision-theory Look for the optimal decision $\phi(X^n) \in [0,1]$ for picking B or V based on X^n loss $\ell : \mathscr{P} \times [0,1] \to [0,1]$,

$$\ell(P,\phi) = \begin{cases} 0, & \text{if } P \notin B \cup V, \\ \left|\phi - 1_V(P)\right|, & \text{if } P \in B \cup V. \end{cases}$$

Bayesian risk functions are the Bayesian testing power,

$$\begin{aligned} r_n(\phi_n, \Pi) &= \int_{\mathscr{P}} P^n \ell(P, \phi_n)\, d\Pi(P) \\ &= \int_B P^n \phi_n \, d\Pi(P) + \int_V Q^n (1 - \phi_n)\, d\Pi(Q), \end{aligned}$$

Proof of proposition 185 (I)

Bayes's rule says

$$r_n(\phi_n, \Pi) = \int_{\mathscr{P}} P^n \ell(P, \phi_n)\, d\Pi(P)$$
$$= \int_{\mathscr{X}^n} \ell(P, \phi_n(x^n))\, d\Pi(P|X^n = x^n)\, dP_n^{\Pi}(x^n)$$

Define $\phi_n(x^n)$ as a pointwise minimizer in,

$$\int_{\mathscr{P}} \ell(P, \phi_n(x^n))\, d\Pi(P|X^n = x^n) = \inf_{\psi \in [0,1]} \int_{\mathscr{P}} \ell(P, \psi)\, d\Pi(P|X^n = x^n),$$

for P_n^{Π}-almost-all $x^n \in \mathscr{X}^n$.

Proof of proposition 185 (II)

and for any $\psi_n : \mathscr{X}^n \to [0,1]$

$$\begin{aligned} r_n(\phi_n, \Pi) &= \int_{\mathscr{X}^n} \int_{\mathscr{P}} \ell(P, \psi_n(x^n))\, d\Pi(P|X^n = x^n)\, dP_n^{\Pi}(x^n) \\ &= \int_{\mathscr{X}^n} \inf_{\psi \in [0,1]} \int_{\mathscr{P}} \ell(P, \psi)\, d\Pi(P|X^n = x^n)\, dP_n^{\Pi}(x^n) \\ &\leq \inf_{\psi_n} r_n(\psi_n, \Pi), \end{aligned}$$

To conclude solve the minimization problem

$$\begin{aligned} &\int_{\mathscr{P}} \ell(P, \psi_n(x^n))\, d\Pi(P|X^n = x^n) \\ &\quad = \int_B \psi_n(x^n)\, d\Pi(P|X^n = x^n) + \int_V (1 - \psi_n(x^n))\, d\Pi(Q|X^n = x^n) \\ &\quad = \psi_n(x^n)\Pi(B|X^n = x^n) + (1 - \psi_n(x^n))\Pi(V|X^n = x^n), \end{aligned}$$

is minimal for $\psi_n(x^n) = 1\{x^n : \Pi(V|x^n) \geq \Pi(B|x^n)\}$.

Posterior odds model selection for frequentists

Johnson & Rossell (JRSSB, 2010), Taylor & Tibshirani (PNAS, 2016)

Theorem 186 *Given measurable* $B, V \subset \Theta$ $(\Pi(B), \Pi(V) > 0)$ *and,*

i. there are Bayesian tests for B *vs* V *of power* $a_n \downarrow 0$*,*

$$\int_B P^n \phi_n \, d\Pi(P) + \int_V Q^n (1 - \phi_n) \, d\Pi(Q) = o(a_n),$$

ii. and, for all $P \in B$*,* $P^n \lhd a_n^{-1} P_n^{\Pi|B}$*; for all* $Q \in V$*,* $Q^n \lhd a_n^{-1} P_n^{\Pi|V}$*,*

then posterior odds give rise to a pointwise test for B *vs* V*.*

Consistent model selection

Let $\mathscr{P}$ be a model for *i.i.d.* data $X^n \sim P^n$, $(n \geq 1)$, and suppose that $(\mathscr{P}, \mathscr{G}, \Pi)$ has finite, measurable partition,

$$P \in \mathscr{P} = \mathscr{P}_1 \cup \ldots \cup \mathscr{P}_M.$$

Model-selection Which $1 \leq i \leq M$? (such that $P \in \mathscr{P}_i$)

Theorem 187 *Assume that for all* $1 \leq i < j \leq M$,

$$\mathscr{P}_i \text{ and } \mathscr{P}_j \text{ are } \mathscr{U}_\infty\text{-uniformly separated}.$$

Let $1 \leq i \leq M$ *be such that* $P \in \mathscr{P}_i$. *If* Π *is a KL-prior, then indicators for posterior odds,*

$$\phi_n(X^n) = 1\Big\{X^n : \Pi(\mathscr{P}_i|X^n) \geq \sum_{j \neq i} \Pi(\mathscr{P}_j|X^n)\Big\},$$

are a pointwise test for $\mathscr{P}_i$ *vs* $\cup_{j \neq i} \mathscr{P}_j$.

Example: select the DAG (I)

Observe an *i.i.d.* X^n of vectors of discrete random variables $X_i = (X_{1,i}, \ldots, X_{k,i}) \in \mathbb{Z}^k$, $1 \leq i \leq n$.

Define a family $\mathscr{F}$ of kernels $p_\theta(\cdot|\cdot) : \mathbb{Z} \times \mathbb{Z}^l \to [0,1]$, for $\theta \in \Theta$, $1 \leq l \leq k$. Assume that Θ is compact and,

$$\theta \mapsto \sum_{x \in \mathbb{Z}} f(x) P_\theta(x|z_1, \ldots, z_l)$$

is continuous, for every bounded $f : \mathbb{Z} \to \mathbb{R}$ and all $z_1, \ldots, z_l \in \mathbb{Z}$.

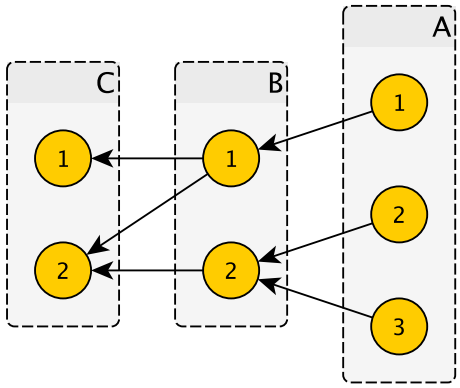
$X \sim P$ follows a graphical model,

$$P_{\mathscr{A},\theta}(X_1 \in B_1, \ldots, X_k \in B_k) = \prod_{i=1}^{k} P_{\theta_i}(X_i \in B_i | \mathscr{A}_i)$$

where $\mathscr{A}_i \subset \{1, \ldots, k\}$ denotes the *parents* of X_i (and $\mathscr{A}_{ij} = \mathscr{A}_i \cup \mathscr{A}_j$). Together, the $\mathscr{A}_i$ describe a directed, a-cyclical graph (DAG).

Example: select the DAG (II)

The DAG $\mathscr{A} = (\mathscr{A}_i : 1 \leq i \leq k)$ represents a number of conditional independence statements concerning the components $X_1, \ldots, X_k$.



$$P_{\mathscr{A},\theta}(C_1 \in \cdot, \ldots, A_3 \in \cdot)$$
$$= P_{\theta_{C,1}}(\cdot|B_1) \times P_{\theta_{C,2}}(\cdot|B_1, B_2)$$
$$\times P_{\theta_{B,1}}(\cdot|A_1) \times P_{\theta_{B_2}}(\cdot|A_2, A_3)$$
$$\times P_{\theta_{A,1}}(\cdot) \times P_{\theta_{A,2}}(\cdot) \times P_{\theta_{A,3}}(\cdot)$$

Fig 1. An small example DAG: No arrow means $X_i \perp X_j | \mathscr{A}_{ij}$. $\mathscr{A}_{C_1} = \{B_1\}$, $\mathscr{A}_{B_2} = \{A_2, A_3\}$, so given B_1, A_2 and A_3, C_1 is independent of B_2.

Example: select the DAG (III)

Define the submodels $\mathscr{P}_{\mathscr{A}} = \{P_{\mathscr{A},\theta} : \theta \in \Theta^k\}$, for all $\mathscr{A}$. Given any $\mathscr{A}' \neq \mathscr{A}$, there is a pair $X_i \perp X_j | \mathscr{A}_{ij}$ but $X_i \not\perp X_j | \mathscr{A}'_{ij}$.

Require that, for all θ, all $A, B \subset \mathbb{Z}$,

$$\left| P_{\mathscr{A}',\theta}(X_i \in A, X_j \in B | \mathscr{A}_{ij}) - P_{\mathscr{A}',\theta}(X_i \in A | \mathscr{A}_{ij})\, P_{\mathscr{A}',\theta}(X_j \in B | \mathscr{A}_{ij}) \right| > \epsilon,$$

for some $\epsilon > 0$ that depends only on $\mathscr{A}$ and $\mathscr{A}'$.

With a KL-prior posterior odds for $\mathscr{P}_{\mathscr{A}}$ select the correct DAG $\mathscr{A}$.

Example: how many clusters? (I)

Observe *i.i.d.* $X^n \sim P^n$, where P dominated with density p.

Clusters Family $\mathscr{F}$ of kernels $\varphi_\theta : \mathbb{R} \to [0, \infty)$, with parameter $\theta \in \Theta$. Assume Θ compact and,

$$\theta \mapsto \int f(x)\varphi_\theta(x)\, dx,$$

is continuous, for every bounded, measurable $f : \mathbb{R} \to \mathbb{R}$. Define $\Theta'_M = \Theta^M / \sim$.

Model Assume that there is an $M > 0$ such that p can be written as,

$$p_{\lambda,\theta}(x) = \sum_{m=1}^{M} \lambda_m p_{\theta_m}(x),$$

for some $M \geq 1$, with $\lambda \in S_M = \{\lambda \in [0,1]^M : \sum_m \lambda_m = 1\}$, $\theta \in \Theta'_M$.

Example: how many clusters? (II)

Assume M less than some known M'. Choose prior $\Pi_{\lambda,M}$ for $\lambda \in S_M$ such that, for some $\epsilon > 0$,

$$\Pi_{\lambda,M}\Big(\lambda \in S_M : \epsilon < \min\{\lambda_m\}, \max\{\lambda_m\} < 1 - \epsilon\Big) = 1.$$

For $\theta \in \Theta'_M$ also choose a prior $\Pi_{\theta,M}$ that 'stays away from the edges'. Define,

$$\Pi = \sum_{M=1}^{M'} \mu_M \, \Pi_{\lambda,M} \times \Pi_{\theta,M}.$$

(for $\sum_M \mu_M = 1$).

If Π is a KL-prior, posterior odds select the correct number of clusters M. If there are no M' and ϵ known, there are sequences $M'_n \to \infty$ and $\epsilon_n \downarrow 0$ with priors Π_n that finds the correct number of clusters.