University of Torino, December 2021

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 P_0, 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

Example 1 *(Savage, 1961) Consider 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

Course schedule

Lec V **Tests and posteriors**
Testing and posterior concentration, Doob's theorem

Lec VI **Frequentist validity of Bayesian limits**
Remote contiguity and frequentist limits

Lec VII **Posterior uncertainty quantification**
How confidence sets arise from credible sets

Lec VIII **Confidence sets in a sparse stochastic block model**
Exact, non-asymptotic confidence sets for community structure

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	$\Pi(\,\cdot\,\|X^n) : \mathscr{G}_n \to [0,1]$	Bayes's rule, inference

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

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 $G \in \mathscr{G}_n$

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

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,*

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

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 $\mathscr{P}$.*

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

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. $\square$

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,

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

So there exist elements in the convex hull co($\mathscr{P}$) arbitrarily close to $\hat{P}$. Conclude that $\hat{P}$ 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 confidence sets

Let $\mathscr{C}$ be a collection of subsets of Θ (*e.g.* intervals, balls, *etcetera*)

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,*

$$\inf_{\theta \in \Theta} P_\theta\Big(\theta \in C_\alpha(X)\Big) \geq 1 - \alpha$$

Definition 16 *Confidence sets $C_{\alpha,n}$ cover the truth asymptotically if*

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

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

Typically confidence sets are based on an estimator $\hat{\theta}$, or rather, on its sampling distribution on Θ.

Bayesian credible sets

Let $\mathscr{D}$ be a collection of subsets of Θ (*e.g.* intervals, balls, *etcetera*)

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, P^Π-almost-surely,*

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

Definition 18 *Credible sets $D_{\alpha,n}$ cover the truth asymptotically if*

$$\Pi\Big(\theta \in D_{\alpha,n}(X^n)\,\Big|\,X^n\Big) \to 1,$$

as $n \to \infty$, 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 19 *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

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

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

Odds ratios and Bayes factors

Definition 20 *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

Symmetric testing 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 $\theta \in B$, $\theta' \in V$

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

A test sequence (ϕ_n) is uniformly consistent if,

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

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

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

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

Minimax optimal tests

We say that (ϕ_n) is minimax optimal if,

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

Theorem 21 *(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 22 *(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 23 *(Weak tests) Let $\epsilon > 0$, $P_0 \in \mathscr{P}$ and a measurable $f : \mathscr{X}^n \to [0,1]$ be given. Define,*

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

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.

[B. Kleijn, A. van der Vaart, Electron. J. Statist. **6** (2012), 354-381]

Example Parametric regression

Questions

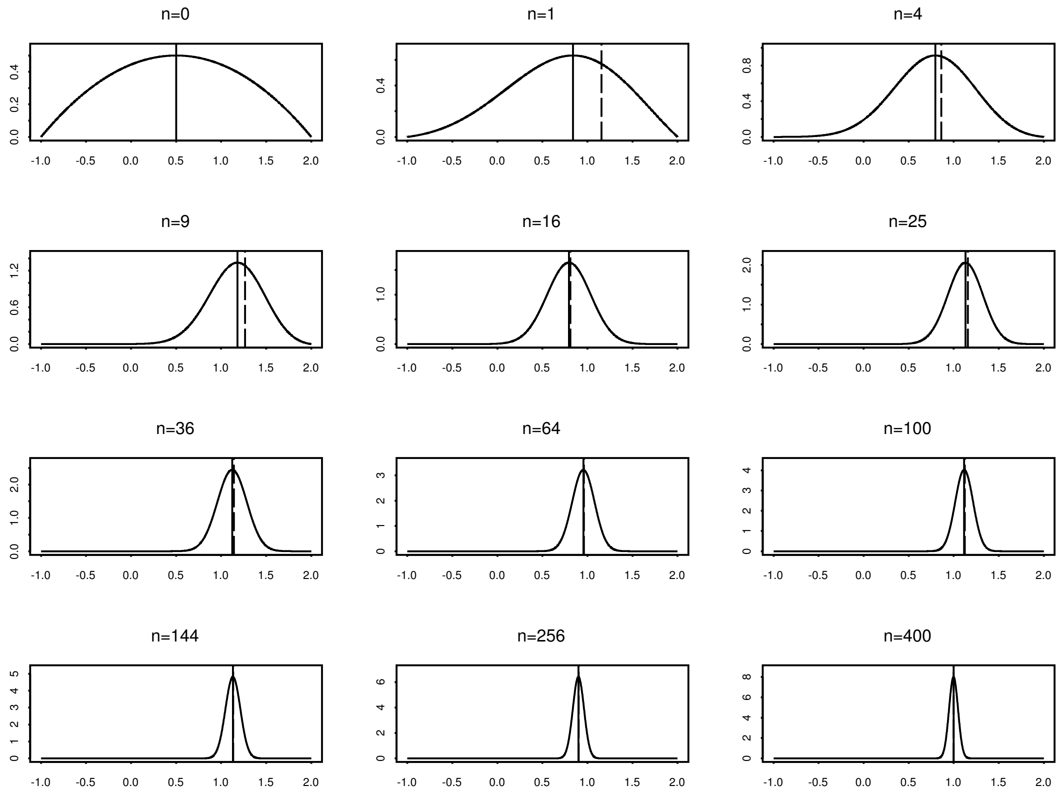
Observe *i.i.d.* $Y_1, \ldots, Y_n$, $Y_i = \theta + e_i$ (or $Y_i = \theta\, X_i + e_i$, *etcetera*) with a normally distributed error (of known variance). The density for the observation is,

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

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 24 *(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 25 *(Le Cam (1960))*

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

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

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 26 *(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 27 *An estimator sequence $\widehat{\theta}_n$ for a parameter θ_0 is said to be regular, if for every $h_n = O(1)$, with $\theta_n = \theta_0 + n^{-1/2}h_n$*

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

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

Theorem 28 *(Hájek, 1970)*
Assume that the model is LAN at θ_0 with non-singular Fisher information I_{θ_0}. Suppose $\widehat{\theta}_n$ 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 29 *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 30 *(Fisher, Cramér, Rao, Le Cam, Hájek)*

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

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

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

Remark 31 *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 32 *(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,

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

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

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

Parametric Bernstein-von Mises theorem

Theorem 33 (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 $M_n \to \infty$,

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

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 34 *With $\hat{\theta}_n$ any efficient estimator,*

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

Remark 35 *(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 36 *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 $\hat{\theta}_{3,n}$ of posterior risk is called the formal Bayes estimator associated with ℓ and Π

Theorem 37 *(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 $\sqrt{n}(\hat{\theta}_{3,n} - \theta_0)$ 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!

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

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

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

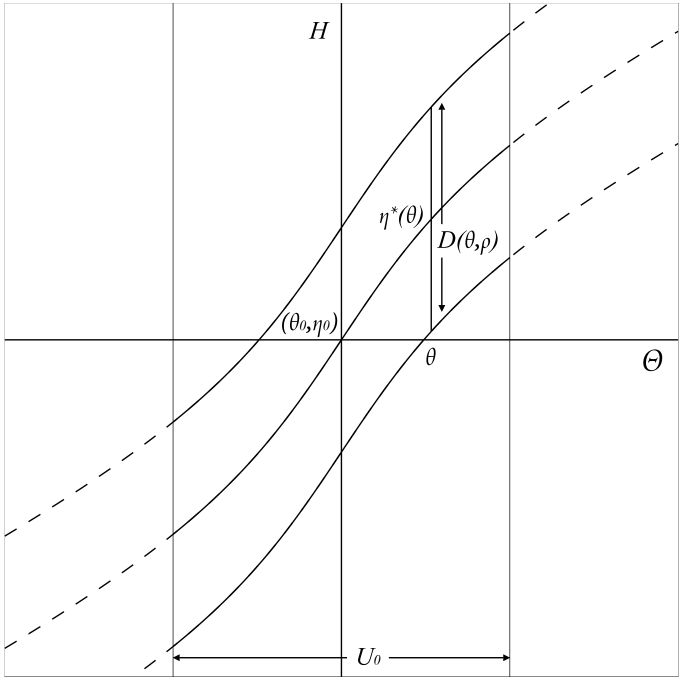
Definition 39 s_n *is said to have the ILAN property, if for every* $h_n = O_{P_0}(1)$

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

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

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

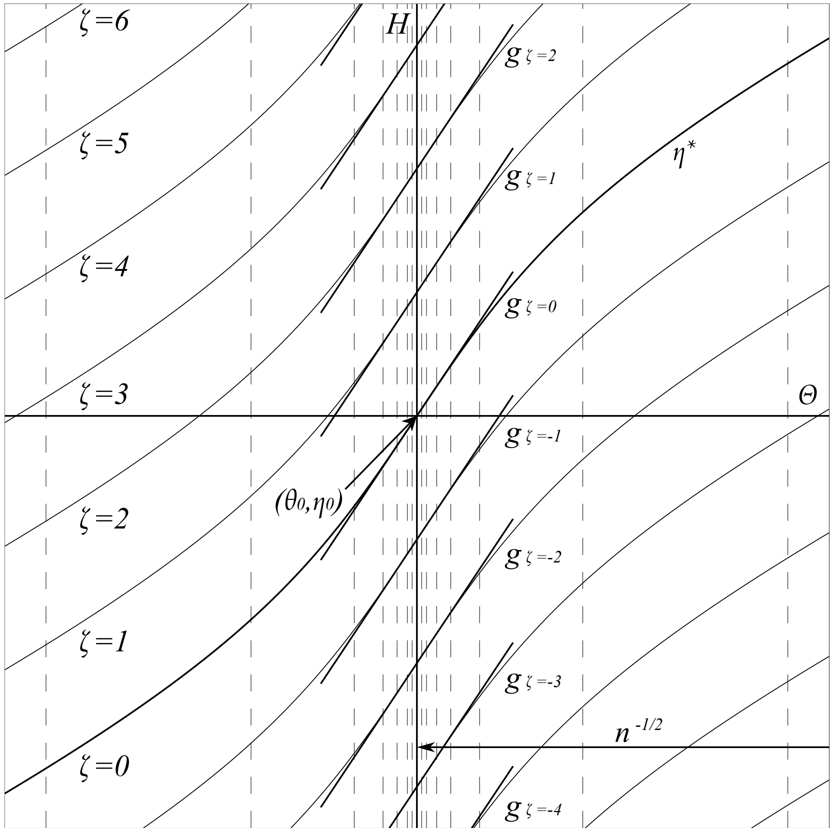
Consistency under $\sqrt{n}$-perturbation



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 40 (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 $M_n \to \infty$,

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

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 $\hat{\theta}_n$ for θ_0 is regular but asymptotically biased if,

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

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

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

Example Regression with symmetric errors

Theorem 42 *(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 $\hat{\theta}_n(X^n)$.

An estimator is said to be consistent if

$$\hat{\theta}_n(X^n) \xrightarrow{P_{\theta_0,n}} \theta_0.$$

E.g. if the topology is metric, a consistent estimator $\hat{\theta}_n(X^n)$ 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 $\widehat{\theta}_n(X^n) \xrightarrow{P_{\theta_0,n}} \theta_0$. Let (r_n) be a sequence $r_n \downarrow 0$.

We say that $\widehat{\theta}_n(X^n)$ converges to θ_0 at rate r_n if

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

r_n is such that it compensates the decrease in distance between $\widehat{\theta}_n(X^n)$ and θ_0, such that the fraction is non-degenerate and bounded in probability.

Intuitively the r_n are the radii of balls around $\widehat{\theta}_n(X^n)$ 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

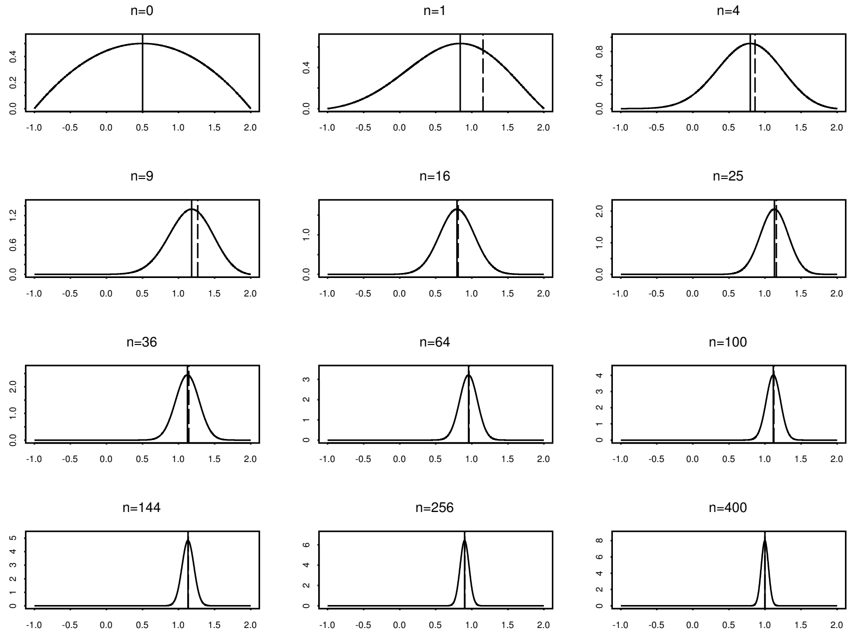
$$r_n^{-1}(\widehat{\theta}_n - \theta_0) \xrightarrow{P_{\theta_0}\text{-w.}} L_{\theta_0},$$

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

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

A posterior is consistent if it is consistent for all $P \in \mathscr{P}$.

Consistency is Prokhorov's weak convergence

Theorem 44 *Let $\mathscr{P}$ be a uniform model with Borel prior Π. The posterior is 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 45 *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 46 *Suppose that $\mathscr{P}$ is a is endowed with the topology of total variation. Assume that the posterior is consistent. Then the posterior mean $\widehat{P}_n$ is a 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's inequality 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 47 *(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 Borel 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)). $\square$

Freedman's counterexamples

Remark 48 *Doob's theorem says nothing about specific points: it is always possible that the frequentist's P_0 belongs to the null-set for which inconsistency occurs.*

Remark 49 *(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 50 *(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* $\eta > 0$,

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

Then the posterior is consistent at P_0.

Corollary 51 *Let* $\mathscr{P}$ *be Hellinger totally bounded and let* Π *a KL-prior. Then the posterior is Hellinger consistent at* P_0 *for the metric* d.

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

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

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

$$
\begin{aligned}
&\limsup_{n\to\infty} \Pi(V|X_1,\ldots,X_n)(1-\phi_n(X_1,\ldots,X_n)) \\
&\quad \le \frac{\limsup_{n\to\infty} e^{n\alpha} \int_V \prod_{i=1}^n (p/p_0)(X_i)\,(1-\phi_n(X_1,\ldots,X_n))\,d\Pi(P)}{\liminf_{n\to\infty} e^{n\alpha} \int_{\mathscr{P}} \prod_{i=1}^n (p/p_0)(X_i)\,d\Pi(P)} \qquad (6) \\
&\quad \le \frac{1}{\Pi(K_\eta)} \limsup_{n\to\infty} f_n(X_1,\ldots,X_n),
\end{aligned}
$$

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 52 *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

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

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

... but there are very nasty examples

Example 53 *Consider P_0 on $\mathbb{R}$ with Lebesgue density p_0 supported on an interval of width one but unknown location. For some $\eta : \mathbb{R} \to (0, \infty)$ 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.

[B. Kleijn, Y. Y. Zhao, Electron. J. Statist. **13.2** (2019), 4709–4742]

Recall Schwartz

Theorem 54 *(Schwartz (1965))*
Let $\mathscr{P}$ be Hellinger totally bounded and let Π *a KL-prior, i.e. for* $\eta > 0$,

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

Then the posterior is Hellinger consistent at P_0.

Example 55 *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 56 *(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 $\Pi\big(\, H(P, P_0) > \epsilon \,\big|\, X_1, \ldots, X_n \,\big) \xrightarrow{P_0\text{-a.s.}} 0$.

The Ghosal-Ghosh-van der Vaart theorem

Theorem 57 *(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.*

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

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 58 *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 $a = 1$, $b > 1$,

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

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

Remark 59 *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 60 *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 $1 \le i \le N$ then the assertion is proved. So pick some i and consider,

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

By Fubini,

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

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

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 61 *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,*

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

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 $P_n \in \mathrm{co}(V^n)$

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

Hellinger transforms “sub-factorize” over convex hulls of products

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

(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,

$$\pi_{P_0}(W,B) = \inf_{0\leq\alpha\leq 1} \sup_{P\in W} \sup_{Q\in B} P_0\left(\frac{dP}{dQ}\right)^{\alpha}$$

Theorem 62 *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* $\Pi(B_i) > 0$*,*

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

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

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

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

With theorem 62 consistency in example 55 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 63 *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 64 *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,*

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

Relation with priors that charge metric balls

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

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

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

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

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

Lemma 66 *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 67 *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,

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

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

A new theorem for separable models

Theorem 68 *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 69 *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 70 *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,*

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

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

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

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

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 71 *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 71 $\sqrt{n}$-consistency in the heavy-tailed example 58 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,

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

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

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

Question

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

$$d(P_{\theta,\eta}, P_{\theta',\eta'}) = \max\Big\{|\theta_1 - \theta'_1|, |\theta_2 - \theta'_2|\Big\}.$$

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

Estimation of support boundary II: construction

Lemma 72 *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 $B = \{Q : \|(p_0/q) - 1\|_\infty < \delta\}$,

$$\inf_{0\le\alpha\le1} P_0\Big(\frac{p}{q}\Big)^\alpha \le (1+\delta)\min\Big\{P_0(p > 0), P(p_0 > 0)\Big\}$$

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 $\epsilon, \delta > 0$,

$$\sup_{Q\in B}\sup_{P\in V}\inf_{0\le\alpha\le1} P_0\Big(\frac{p}{q}\Big)^\alpha \le (1+\delta)\Big(1 - \frac{f(\epsilon)}{\sigma}\Big) < 1.$$

Estimation of support boundary III: theorem

Theorem 73 *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 74 *The* σ*-restriction on* $\theta_1 - \theta_2$ *can be eliminated with theorem 67.*

Lecture V

Tests and posteriors

The existence of Bayesian test sequences implies concentration of the posterior distribution and vice versa. By implication, distinctions between model subsets are asymptotically testable if and only if also expressed through posterior convergence. In a Bayesian sense this leads to a form of posterior concentration 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.

[B. Kleijn, Ann. Statist. **49.1** (2021), 182–202]

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

Theorem 75 *(Bayesian consistency, 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 76 *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, 1986a, 1986b, 1998, . . .)

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

Theorem 77 *(Frequentist, Schwartz (1965))*
Let $X_1, X_2, \ldots$ be i.i.d.-P_0 for some $P_0 \in \mathscr{P}$. Let $U \subset \mathscr{P}$ be given. If,

(i) there are $\phi_n : \mathscr{X}_n \to [0,1]$, s.t.

$$P_0^n \phi_n = o(1), \qquad \sup_{Q \in U^c} Q^n(1 - \phi_n) = o(1), \tag{12}$$

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

$$\Pi\Big(P \in \mathscr{P} : -P_0 \log \frac{dP}{dP_0} < \delta \Big) > 0, \tag{13}$$

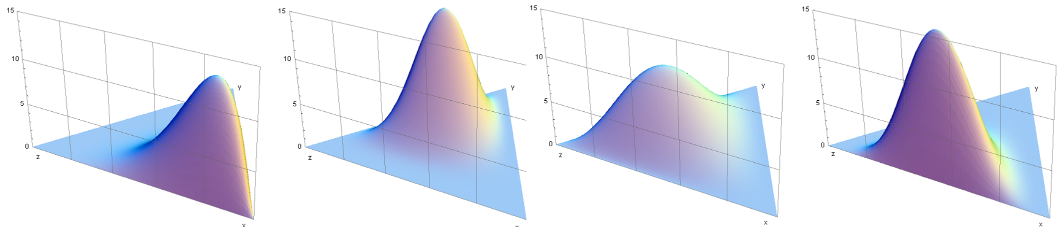
then $\Pi(U|X^n) \xrightarrow{P_0\text{-a.s.}} 1$.

Condition (i) implies $P_0 \in U$, but it is not necessary that U is a neighbourhood of P_0; *only* the existence of the test is required.

The Dirichlet process

Definition 78 *(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 79 *(Dirichlet process, Ferguson 1973,1974)*
Let μ *be a finite base measure on* $(\mathscr{X}, \mathscr{B})$. *The Dirichlet process* $P \sim D_\mu$ *is defined by random histograms: for partitions* $A_1, \ldots, A_k$ *of* $\mathscr{X}$,

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

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

Theorem 80 *(Frequentist, Dirichlet consistency)*
Let $X_1, X_2, \ldots$ be an i.i.d.-sample from P_0 With a Dirichlet prior D_μ with finite base measure μ such that $\mathrm{supp}(P_0) \subset \mathrm{supp}(\mu)$, the posterior is consistent at P_0 in Prokhorov's weak topology.

Remark 81 *(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 82 *$X^n \mapsto \Pi(P(A)|X^n)$ 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 83 *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 84 *For* $B \in \mathscr{G}$ *such that* $\Pi_n(B) > 0$*, the local prior predictive distribution is defined, for every* $A \in \mathscr{B}_n$*,*

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

Corollary 85 *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,*

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

Proof

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

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

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

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

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

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

Proof

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

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

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

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

$$\leq \frac{1}{\Pi(B)}\Big(\int_B P\phi(X)\,d\Pi(P) + \int_V Q(1-\phi)(X)\,d\Pi(Q)\Big)$$

Bayesian testability –is– posterior convergence

Proposition 86 *Let* $(\Theta, \mathscr{G}, \Pi)$ *be given. For any* $B, V \in \mathscr{G}$*, the following are equivalent,*

(i) There exist tests (ϕ_n) *such that for* Π*-almost-all* $\theta \in B$*,* $\theta' \in V$*,*

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

(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* $\theta \in B$*,* $\theta' \in V$*,*

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

Remark 87 *Interpretation distinctions between model subsets are Bayesian testable, iff they are picked up by the posterior asymptotically, iff, posterior odds for B versus V are 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,

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

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 88 *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* $\theta \in \Theta$.

Corollary 89 (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* $\mathscr{P} \setminus V$.

. . . which implies (but proves more than) Doob's 1948 consistency theorem

Examples: prior-almost-sure inconsistency (I)

Example 90 (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 91 *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 92 *(Freedman (1965))*
Let $X_1, X_2, \ldots$ be i.i.d. integers, Endow $\pi(\Lambda)$ with Prokhorov'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 equal zero for all $P \in \Lambda_k$ 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 VI

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.

[B. Kleijn, Ann. Statist. **49.1** (2021), 182–202]

Le Cam's inequality

Definition 93 *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 94 *(Le Cam, unpublished (197X) and (1986))*
Rewrite the posterior concentration inequality

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

Remark 95 *Useful in parametric models (e.g. BvM) but "a considerable nuisance" [sic, Le Cam (1986)] in non-parametric context*

Schwartz's theorem revisited

Remark 96 *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

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

then for large enough m

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

Theorem 97 *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 98 *(Schwartz's theorem)*

Remote contiguity

Definition 99 *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 100 *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 101 *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 102 *Hellinger transform* $\psi(P,Q;\alpha) = \int p^\alpha q^{1-\alpha}\, d\mu$

Le Cam's first lemma

Lemma 103 *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) $\liminf_n \int p_n^\alpha q_n^{1-\alpha}\, d\mu \to 1$ *as* $\alpha \uparrow 1$

Criteria for remote contiguity

Lemma 104 *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*

Beyond Schwartz

Theorem 105 *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* $P_{0,n} \triangleleft a_n^{-1} \, P_n^{\Pi|B}$

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

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

Remark 106 *(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 107 *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 108 *(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 109 *Larger* B_n *are possible, under conditions on the model (see Kleijn and Zhao (201x))*

Consistent posterior odds

Theorem 110 *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 posterior odds O_n *(or Bayes factors* B_n*),*

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

for B *versus* V *are consistent.*

Posterior odds are optimal

Proposition 111 *Let* $(\mathscr{P}, \mathscr{G})$ *be a model with prior* Π *and* $B, V \in \mathscr{G}$, $B \neq V$. *The test function* $\phi(X) = 1\{x \in \mathscr{X} : \Pi(V|X = x) \geq \Pi(B|X = x)\}$ *has optimal Bayesian testing power:*

$$\int_B P_\theta \phi \, d\Pi(\theta) + \int_V P_\theta(1 - \phi) \, d\Pi(\theta)$$
$$= \inf_\psi \Big(\int_B P_\theta \psi \, d\Pi(\theta) + \int_V P_\theta(1 - \psi) \, d\Pi(\theta) \Big).$$

Proof

Find the optimal 'decision' $\phi \in [0,1]$ for loss $\ell : \mathscr{P} \times [0,1] \to [0,1]$,

$$\ell(P,\phi) = \begin{cases} 0, & \text{if } P \notin B \cup V, \\ \bigl|\phi - 1_V(P)\bigr|, & \text{if } P \in B \cup V. \end{cases}$$

Data-driven decisions $\phi(X)$ are test functions. The Bayesian risk function,

$$r(\phi,\Pi) = \int_{\mathscr{P}} P\ell(P,\phi)\, d\Pi(P),$$

is Bayesian testing power,

$$\begin{aligned} r(\phi,\Pi) &= \int_B P\bigl|\phi - 1_V(P)\bigr|\, d\Pi(P) + \int_V Q\bigl|\phi - 1_V(Q)\bigr|\, d\Pi(Q) \\ &= \int_B P\phi\, d\Pi(P) + \int_V Q(1-\phi)\, d\Pi(Q). \end{aligned}$$

Proof

Bayes's rule if $\phi(x)$ minimizes posterior expected loss for P^{Π}-almost-all $x \in \mathscr{X}$,

$$\int_{\mathscr{P}} \ell(P, \phi(x))\, d\Pi(P|X=x) = \inf_{\psi \in [0,1]} \int_{\mathscr{P}} \ell(P, \psi)\, d\Pi(P|X=x),$$

then $\phi : \mathscr{X} \to [0,1]$ optimizes Bayesian testing power:

$$r(\phi, \Pi) = \inf\{r(\psi, \Pi)\ :\ \psi : \mathscr{X} \to [0,1]\},$$

To conclude note that,

$$\begin{aligned}
&\int_{\mathscr{P}} \ell(P, \psi(x))\, d\Pi(P|X=x) \\
&\quad = \int_B \psi(x)\, d\Pi(P|X=x) + \int_V (1-\psi(x))\, d\Pi(Q|X=x) \\
&\quad = \psi(x)\Pi(B|X=x) + (1-\psi_n(x))\Pi(V|X=x),
\end{aligned}$$

is minimal if we choose $\psi(x) = 1\{x : \Pi(V|X=x) \geq \Pi(B|X=x)\}$.

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 goodness-of-fit testing of transition probabilities with posterior odds.

Ergodic random-walks

Example 112 *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,*

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

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,

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

So lemma 104 says that

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

Remark 113 *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' $p_\alpha(k,l) = p_\alpha(k|l)\pi_\alpha(l)$,

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

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 114 *A Bayesian test sequence for H_0 versus H_1 exists based on a version of Hoeffding's 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 115 *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 posterior odds O_n *for* H_0 *versus* H_1 *are consistent.*

Lecture VII

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.

[B. Kleijn, Ann. Statist. **49.1** (2021), 182–202]

Remote contiguity in finite sample spaces

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

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

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

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

Proposition 116 *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 116 with an $0 < \epsilon < \epsilon'$ small enough guarantees exponential remote contiguity

Then theorem 105 says $\Pi(V|X^n)$ goes to zero in P_0^n-probability.

Proposition 117 *(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 118 *(Dirichlet process, Ferguson 1973,1974)*
Let μ be a finite base measure on $(\mathscr{X}, \mathscr{B})$. The Dirichlet process $P \sim D_\mu$ is defined by random histograms: for partitions $A_1, \ldots, A_k$ of $\mathscr{X}$,

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

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 23 says there are (ϕ_n)

$$\sup_{P \in B_g} P^n \phi_n \leq e^{-nD}, \quad \sup_{Q \in V_g} Q^n(1 - \phi_n) \leq e^{-nD}. \tag{14}$$

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 119 *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 $\mathscr{X}_\alpha = \{e_n : 1 \leq n \leq N_\alpha\}$

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

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 (14), implies weak consistency

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

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

Asymptotic credible and confidence sets

Definition 120 *Let* $(\Theta, \mathscr{G})$ *with priors* Π_n *and a collection* $\mathscr{D}$ *of measurable subsets of* Θ *be given. Credible sets* (D_n) *of credible levels* $1 - o(a_n)$ *are 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.*

Definition 121 *Maps* $x \mapsto C_n(x) \subset \Theta$ *are asymptotically consistent confidence sets (of levels* $1 - o(a_n)$*), if,*

$$P_{\theta,n}\Big(\theta \notin C_n(X^n)\Big) \to 0, \ (= o(a_n))$$

for all $\theta \in \Theta$. C_n *is asymptotically informative, if for all* $\theta' \neq \theta$,

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

Existence of confidence sets and tests

Theorem 122 *The following are equivalent:*

(i) For every $\theta \in \Theta$, there exist pointwise tests $\phi_{\theta,n}(X^n)$ for $\{\theta\}$ vs $\Theta \setminus \{\theta\}$ of power a_n: for all $\theta' \neq \theta$,

$$P_{\theta,n}\phi_{\theta,n} + P_{\theta',n}(1 - \phi_{\theta,n}) = o(a_n)$$

(ii) There are confidence sets $C_n(X^n)$ of levels $1 - a_n$ that are asymptotically consistent and informative: for all $\theta' \neq \theta$,

$$P_{\theta,n}\Big(\theta \notin C_n(X^n)\Big) + P_{\theta',n}\Big(\theta \in C_n(X^n)\Big) = o(a_n)$$

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

$$D_n(X^n) = B_n(\hat{\theta}_n(X^n), \hat{r}_n(X^n))$$

of credible levels $1 - o(a_n)$ *satisfy,* $P_n^{\Pi_n}$*-almost-surely,*

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

Credible sets *with* converging posteriors (II)

Suppose that (Θ_n, d_n) are metric spaces

Theorem 124 *(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 $\hat{B}_n(X^n) = B_n(\hat{\theta}_n(X^n), \hat{r}_n(X^n))$ *be level-*$1-\epsilon$ *credible balls of minimal radii. Then with high* $P_{\theta_{0,n}}$*-probability* $\hat{r}_n \leq r_n$.

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

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

Proof of theorem 124 (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 $x^n \in F_n$ and $\theta_n \in \Theta_n$, let $r_n(\theta_n, x^n) \in [0, \infty]$ denote the smallest radius of balls centred on θ_n of posterior mass at least $1 - \epsilon$.

Define $\hat{\theta}_n(x^n)$ as the centre point of a credible ball with minimal radius $\hat{r}_n(x^n) = \inf\{r_n(\theta_n, x^n) : \theta_n \in \Theta_n\}$,

$$\hat{B}_n(x^n) = B_n(\hat{\theta}_n(x^n), \hat{r}_n(x^n)),$$

of level $1 - \epsilon$. Note

$$P_{\theta_{0,n}}\Big(\Pi(\hat{B}_n(X^n)|X^n) \geq 1 - \epsilon\Big) = 1,$$

for all $n \geq 1$.

Proof of theorem 124 (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

$$\hat{r}_n(X^n) \leq r_n(\theta_{0,n}, X^n) \leq r_n.$$

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,

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

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

Remark 125 *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.*

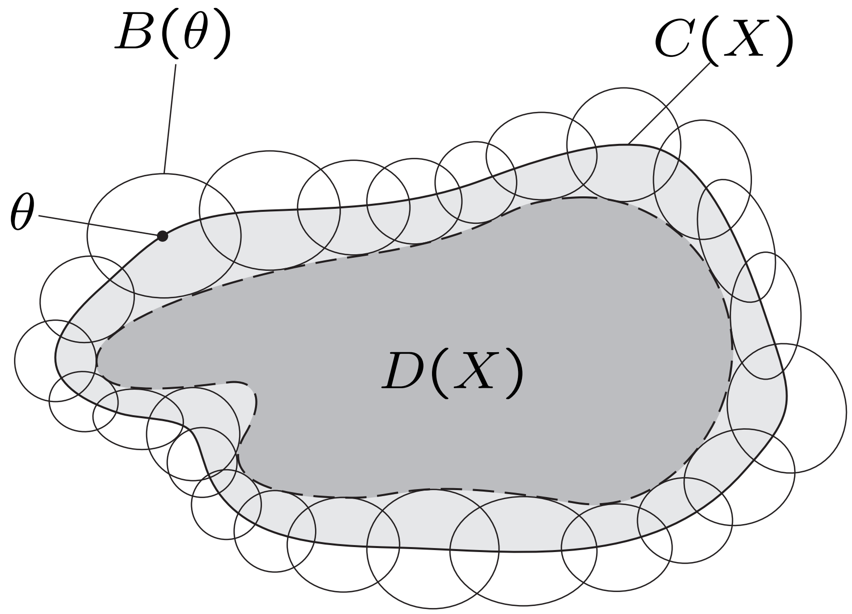
Credible sets *without* converging posteriors

Definition 126 *Let $D(X)$ be a credible set in Θ and let B denote a set function $\theta \mapsto B(\theta) \subset \Theta$. A model subset $C(X)$ is said to be a confidence set associated with $D(X)$ under B, if for all $\theta \in \Theta \setminus C(X)$,*

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

Definition 127 *The intersection $C_0(X)$ of all $C(X)$ like above is a confidence set associated with $D(X)$ under B, called the minimal confidence set associated with $D(X)$ under B.*

B-Enlargement of credible sets



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

B-Enlarged credible sets are confidence sets

Theorem 128 *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* $D_n(X^n)$ *denote level-*$(1 - o(a_n))$ *credible sets. Furthermore, for all* $\theta \in \Theta$, *let* $\theta \mapsto B_n(\theta)$ *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 $C_n(X^n)$ *associated with the credible sets* $D_n(X^n)$ *under* B_n *are asymptotically consistent, that is,*

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

Proof of theorem 128 (I)

Let D_n denote credible sets of levels $1-o(a_n)$, defined for all $x^n \in F_n \subset \mathscr{X}_n$ such that $P_n^{\Pi_n}(F_n) = 1$. For any $x^n \in F_n$, $C_n(x^n)$ is a confidence set associated with $D_n(x^n)$ under B.

Note that by definition of $C_n(x^n)$,

$$\theta_0 \in \Theta \setminus C_n(x^n) \quad \Rightarrow \quad B_n(\theta_0) \cap D_n(x^n) = \varnothing.$$

Then $\Pi(B_n(\theta_0)|x^n) = o(a_n)$.

So for all $x^n \in F_n$ the functions $x \mapsto 1\{\theta_0 \in \Theta \setminus C_n(x^n)\} \, \Pi(B(\theta_0)|x^n)$ are $o(a_n)$.

Proof of theorem 128 (II)

Integrate with respect to $P_n^{\Pi_n}$ and divide by $\Pi_n(B_n(\theta_0))$ to find,

$$\frac{1}{\Pi_n(B_n(\theta_0))}\int 1\{\theta_0 \in \Theta \setminus C_n(x^n)\}\,\Pi(B_n(\theta_0)|x^n)\,dP_n^{\Pi_n} = o(a_n b_n^{-1}).$$

By Bayes's rule in the form (1),

$$\begin{aligned}
&P_n^{\Pi_n|B_n(\theta_0)}\Big(\theta_0 \in \Theta \setminus C_n(X^n)\Big) \\
&\qquad\qquad = \int P_{\theta,n}\Big(\theta_0 \in \Theta \setminus C_n(X^n)\Big)\,d\Pi_n(\theta|B_n) = o(a_n b_n^{-1}).
\end{aligned}$$

Since $P_{\theta_0,n} \lhd b_n a_n^{-1}\, P_n^{\Pi_n|B_n(\theta_0)}$ this implies asymptotic coverage.

Methodology: confidence sets from posteriors (I)

Corollary 129 *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* $D_n(X^n)$ *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 130 *Let* Π_n *denote Borel priors on* $\mathscr{P}$, *with constant* $C > 0$ *and rate sequence* $\epsilon_n \downarrow 0$ *such that:*

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

Given credible sets $D_n(X^n)$ *of level* $1 - o(\exp(-C'n\epsilon_n^2))$, *for some* $C' > C$. *Then radius-*ϵ_n *Hellinger-enlargements* $C_n(X^n)$ *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 131 *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 VIII

Confidence sets in a sparse stochastic block model

In a sparse stochastic block model with two communities of unequal sizes we derive two posterior concentration inequalities, for (1) posterior (almost-)exact recovery of the community structure; (2) a construction of confidence sets for the community assignment from credible sets with finite graph sizes, enabling *exact frequentist uncertain quantification with Bayesian credible sets at non-asymptotic graph sizes*. It is argued that a form of early stopping applies to MCMC sampling of the posterior to enable the computation of confidence sets at larger graph sizes.

[B. Kleijn and J. van Waaij, arXiv:1810.09533, 2108.07078 [math.ST]]

Part I

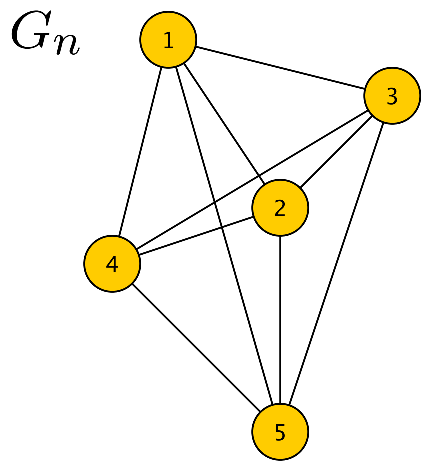
Sparse stochastic block models

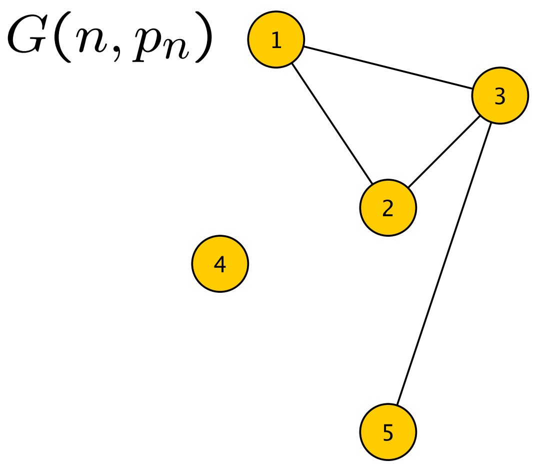
Erdös-Rényi random graphs

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

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

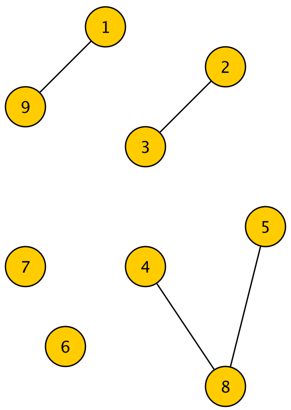
Result random graph $G(n, p_n) = (V_n, E'_n)$ (Erdös, Rényi (1959, 1961)).





Complete graph and edge-percolated ER-graph

Sparsity phases of the Erdös-Rényi random graph



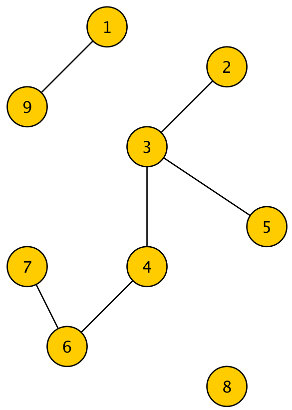
Fragmented

$p_n < 1/n$

Many fragments

clusters $\leq O(\log(n))$

$E(N_i) = O(1)$



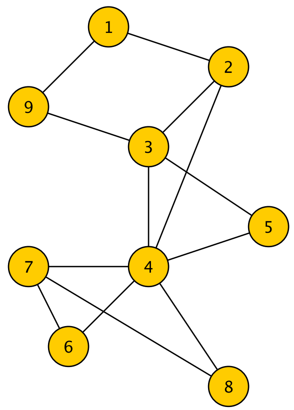
Kesten-Stigum

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

Giant component

cluster $\sim O(n)$

$E(N_i) = O(np_n)$



Chernoff-Hellinger

$p_n > \log(n)/n$

Connected

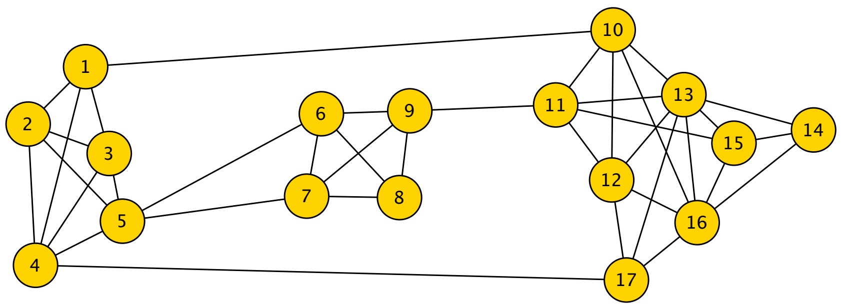
cluster $= n$

$E(N_i) = O(\log(n))$

Two-community stochastic block model

Consider $G_n = (V_n, E_n)$ with community assignment $\theta_n \in \Theta_n = \{0,1\}^n$. Split $V_n = Z_0(\theta_n) \cup Z_1(\theta_n)$. For every $e \in E_n$ independently,

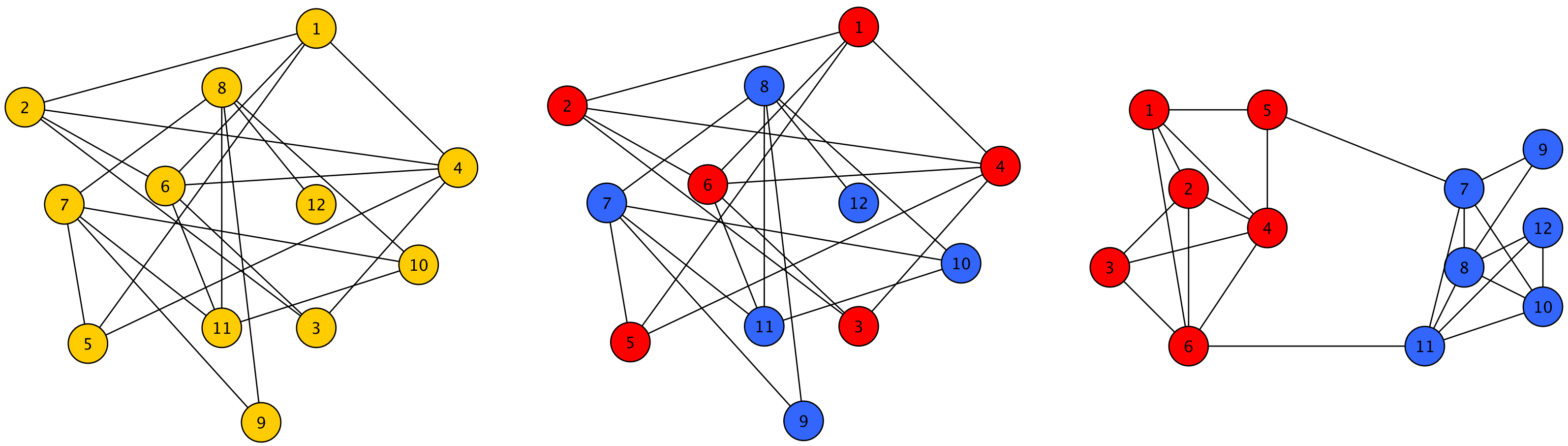
$$\text{include } e \text{ in } E'_n \subset E_n \text{ wp. } \begin{cases} p_n, & \text{if } e \text{ lies within } Z_0 \text{ or } Z_1, \\ q_n, & \text{if } e \text{ lies between } Z_0 \text{ and } Z_1. \end{cases}$$



Three-community SBM graph $X^n = (V_n, E'_n) \in \mathscr{X}_n$, $X^n \sim P_{\theta_n}$

Community detection

Example SBM with $n = 12$, $0 < q_n \ll p_n < 1$, $\theta_n = 000000111111$



Observation

Data $X^n \sim P_{\theta_n}$

Unobserved

Communities of θ_n

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

Detection

Estimate with

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

Asymptotic community detection

Definition 132 *Given community assignments θ_n for all $n \geq 1$, an estimator sequence $\hat{\theta}_n : \mathscr{X}_n \to \Theta_n$ is said to recover θ_n exactly, if,*

$$P_{\theta_n,n}\Big(\hat{\theta}_n(X^n) = \theta_n\Big) \to 1,$$

as $n \to \infty$.

Let $k : \Theta_n \times \Theta_n \to \{0, 1, \ldots, n\}$ denote the Hamming distance.

Definition 133 *Given community assignments θ_n for all $n \geq 1$ and some sequence of error rates (k_n) of order $k_n = O(n)$, an estimator sequence $\hat{\theta}_n : \mathscr{X}_n \to \Theta_n$ is said to recover θ_n almost-exactly with error rate k_n, if,*

$$P_{\theta_n,n}\Big(k(\hat{\theta}_n(X^n), \theta_n) \leq k_n\Big) \to 1,$$

as $n \to \infty$.

Part II

Posterior concentration

Posterior concentration (I)

Let,

$$\rho(p,q) = p^{1/2}q^{1/2} + (1-p)^{1/2}(1-q)^{1/2},$$

denote the Hellinger-affinity between two Bernoulli-distributions with parameters $p, q \in (0,1)$.

Theorem 134 *For fixed $n \geq 1$, suppose $X^n \sim P_{\theta_n,n}$ with $\theta_n \in \Theta_n$ and choose the uniform prior on Θ_n. Then,*

$$P_{\theta_n,n}\Pi(\{\theta_n\}|X^n) \geq 1 - \frac{n}{2}\rho(p_n,q_n)^{n/2}\, e^{n\rho(p_n,q_n)^{n/2}},$$

implying that if,

$$n\rho(p_n,q_n)^{n/2} \to 0, \tag{15}$$

then the posterior recovers the true community assignment exactly.

Exact recovery in the Chernoff-Hellinger phase

$$\text{Sparsity} \quad p_n = a_n \frac{\log(n)}{n}, \quad q_n = b_n \frac{\log(n)}{n}.$$

Corollary 135 *Assume the conditions of theorem 134. If the sequences a_n, b_n in the Chernoff-Hellinger phase satisfy,*

$$\left((\sqrt{a_n} - \sqrt{b_n})^2 - \frac{a_n b_n \log(n)}{2n} - 4 \right) \log(n) \to \infty, \tag{16}$$

then the posterior recovers the community assignments exactly.

For a_n, b_n of order $O(1)$, a simple sufficient condition for exact recovery is,

$$\left((\sqrt{a_n} - \sqrt{b_n})^2 - 4 \right) \log n \to \infty, \tag{17}$$

Posterior concentration (II)

Define the (Hamming-)metric balls,

$$B_n(\theta_n, k_n) = \{\eta_n \in \Theta_n : k(\eta_n, \theta_n) \leq k_n\}, \tag{18}$$

Theorem 136 *For fixed $n \geq 1$, suppose $X^n \sim P_{\theta_n,n}$ with $\theta_n \in \Theta_n$ and choose the uniform prior on Θ_n. For some λ_n with $0 < \lambda_n < 1/2$, let k_n be an integer such that $k_n \geq \lambda_n n$. Then,*

$$P_{\theta_n,n}\Pi\Big(B_n(\theta_n, k_n) \,\Big|\, X^n\Big) \geq 1 - \frac{1}{2}\Big(\tfrac{e}{\lambda_n}\rho(p_n, q_n)^{n/2}\Big)^{\lambda_n n}\Big(1 - \tfrac{e}{\lambda_n}\rho(p_n, q_n)^{n/2}\Big)^{-1}.$$

Recovery in the Kesten-Stigum phase (I)

$$\text{Sparsity} \quad p_n = \frac{c_n}{n}, \quad q_n = \frac{d_n}{n}.$$

Proposition 137 *If the sequences c_n, d_n and the fractions λ_n satisfy,*

$$\lambda_n n \Big(\log(\lambda_n) + \frac{1}{4}\big(\sqrt{c_n} - \sqrt{d_n}\big)^2 - 1 \Big) \to \infty, \tag{19}$$

then posteriors recover the community assignment almost-exactly with any error rate $k_n \geq \lambda_n n$.

Corollary 138 *Recovery c.f. (Decelle et al. (2011))*
Let $0 < \lambda < 1/2$ be given. If, for some constant $C > 1$ and large enough n,

$$(\sqrt{c_n} - \sqrt{d_n})^2 > 4C(1 - \log(\lambda)), \tag{20}$$

then the posterior recovers the community assignment almost exactly with error rate $k_n = \lambda n$.

Recovery in the Kesten-Stigum phase (II)

Corollary 139 *Weak consistency (Mossel, Neeman, Sly (2016))*
If the sequences c_n *and* d_n *satisfy,*

$$\frac{(c_n - d_n)^2}{2(c_n + d_n)} \to \infty, \tag{21}$$

the posterior recovers the true community assignment almost exactly with any error rate $k_n \geq \lambda_n n$ *for some vanishing fraction* $\lambda_n \to 0$.

Corollary 140 *Let* $0 < \lambda_n < 1/2$ *be given, such that* $\lambda_n \to 0$, $\lambda_n n \to \infty$. *If, for some constant* $C > 1$,

$$(\sqrt{c_n} - \sqrt{d_n})^2 + 4C \log(\lambda_n) \to \infty, \tag{22}$$

then the posterior recovers the community assignments almost exactly with error rate $k_n = \lambda_n n$.

Part III

Uncertainty quantification

Bayesian and frequentist uncertainty quantified

Definition 141 *Given $n \geq 1$, a prior Π_n and data X^n, a credible set of credible level $1-\gamma$ is any $D(X^n) \subset \Theta_n$ such that:*

$$\Pi(D(X^n)|X^n) \geq 1-\gamma,$$

$P_n^{\Pi_n}$-almost-surely.

Definition 142 *Given $\theta_n \in \Theta_n$ and data $X^n \sim P_{\theta_n,n}$, a confidence set $C(X^n) \subset \Theta_n$ of confidence level $1-\alpha$ is defined by any $x^n \mapsto C(x^n) \subset \Theta_n$ such that,*

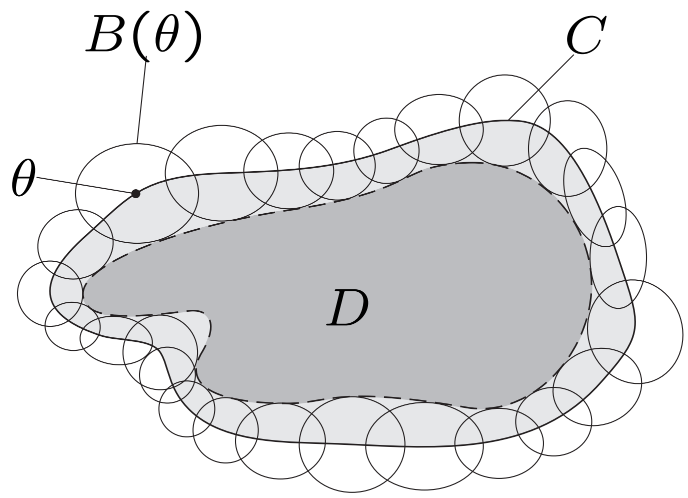
$$P_{\theta_n,n}\Big(\theta_n \in C(X^n)\Big) \geq 1-\alpha.$$

Enlargement of credible sets

Lemma 143 *Fix* $n \geq 1$, *let* $\theta_n \in \Theta_n$, $X^n \sim P_{\theta_n,n}$ *be given. For any* $B \subset \Theta_n$, $0 < \beta < 1$,

$$P_{\theta_n,n}\Pi(B|X^n) \geq 1-\beta \quad \Rightarrow \quad P_{\theta_n,n}\Big(B \cap D(X^n) \neq \varnothing\Big) \geq 1 - \frac{\beta}{1-\gamma}.$$

for any credible set $D(X^n) \subset \Theta_n$ *of credible level* $1-\gamma$.



Enlargement of D by sets $B(\theta)$ to form C

Credible sets are confidence sets (I)

Proposition 144 *For fixed $n \geq 1$, suppose $X^n \sim P_{\theta_n,n}$ with $\theta_n \in \Theta_n$. Every credible set $D(X^n)$ of credible level $1-\gamma$ is a confidence set of confidence level,*

$$P_{\theta_n,n}\Big(\theta_n \in D(X^n)\Big) \geq 1 - \frac{n}{2(1-\gamma)}\rho(p_n, q_n)^{n/2}\, e^{n\rho(p_n,q_n)^{n/2}}. \qquad (23)$$

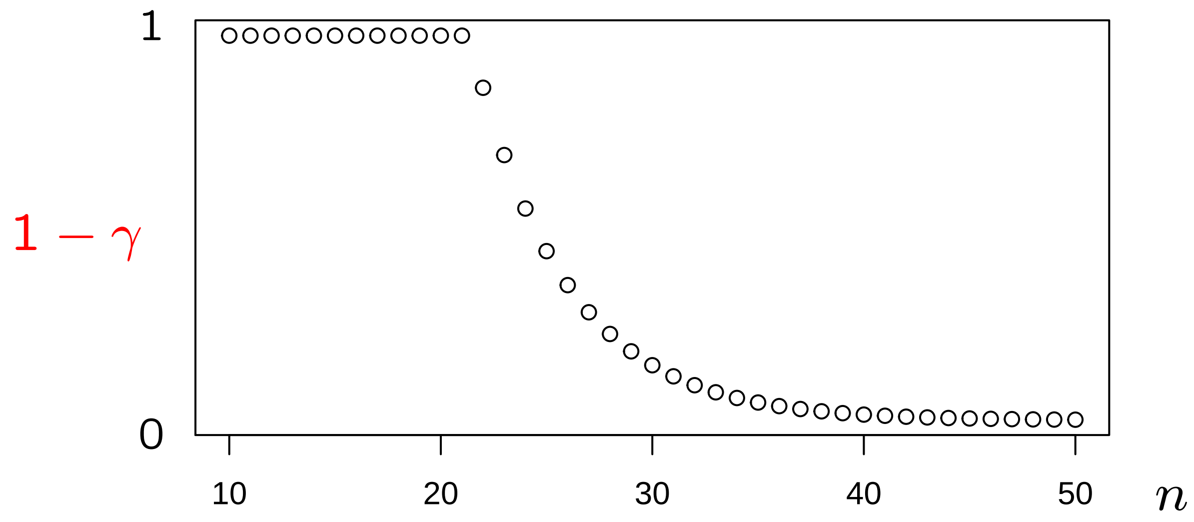
Method 18.2 For graph size n, realised graph $X^n = x^n$, known p, q and realised posterior $\Pi(\cdot|X^n = x^n)$, choose a desired confidence level $0 < 1-\alpha < 1$, we choose credible level,

$$1-\gamma = \min\{1, (n/2\alpha)\rho(p,q)^{n/2}\, e^{n\rho(p,q)^{n/2}}\}. \qquad (24)$$

Credible sets are confidence sets (II)

Example 145 *Take* $p = 0.9$, $q = 0.1$ *and confidence level* $1-\alpha = 0.95$. $\rho(p,q) = 0.6$ *and* $(n/2)\rho(p,q)^{n/2} \approx 0.0211$. *As* n *varies,*

any (unenlarged) credible set of credible level $1-\gamma$ *is a confidence set of confidence level* 0.95



Required credible level for confidence level $1-\alpha = 0.95$

Enlarged credible sets are confidence sets (I)

The k-enlargement $C(X^n)$ of $D(X^n)$ is the union of all Hamming balls of radius $k \geq 1$ that are centred on points in $D(X^n)$,

$$C(X^n) = \left\{ \theta_n \in \Theta_n : \exists_{\eta_n \in D_n(X^n)}, k(\theta_n, \eta_n) \leq k \right\},$$

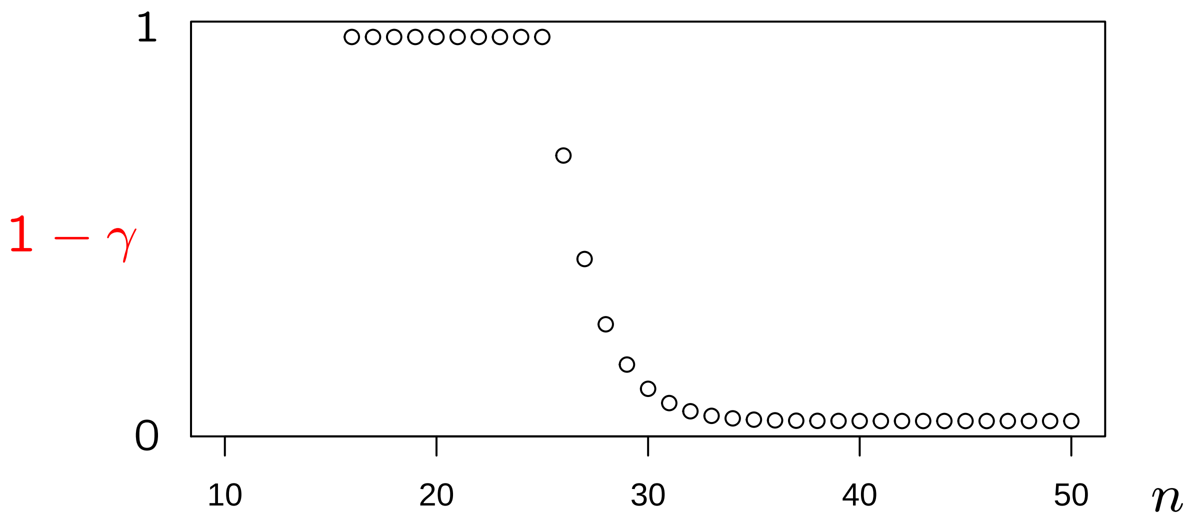
Proposition 146 *For fixed $n \geq 1$, suppose $X^n \sim P_{\theta_n,n}$ with $\theta_n \in \Theta_n$. Define $k = \lceil \lambda n \rceil$. Then the k-enlargement $C(X^n)$ of any credible set $D(X^n)$ of level $1-\gamma$ is a confidence set of confidence level,*

$$P_{\theta_n,n}\Big(\theta_n \in C(X^n)\Big) \geq 1 - \frac{1}{2(1-\gamma)} \Big(\tfrac{e}{\lambda}\rho(p_n, q_n)^{n/2}\Big)^{\lambda n} \Big(1 - \tfrac{e}{\lambda}\rho(p_n, q_n)^{n/2}\Big)^{-1}.$$

Enlarged credible sets are confidence sets (II)

Example 147 *Again $p = 0.9$, $q = 0.1$ and confidence level $1 - \alpha = 0.95$. For $\lambda = 0.05$ and varying graph size n,*

any $0.05n$-enlarged credible set of credible level $1 - \gamma$ is also a confidence set of confidence level 0.95

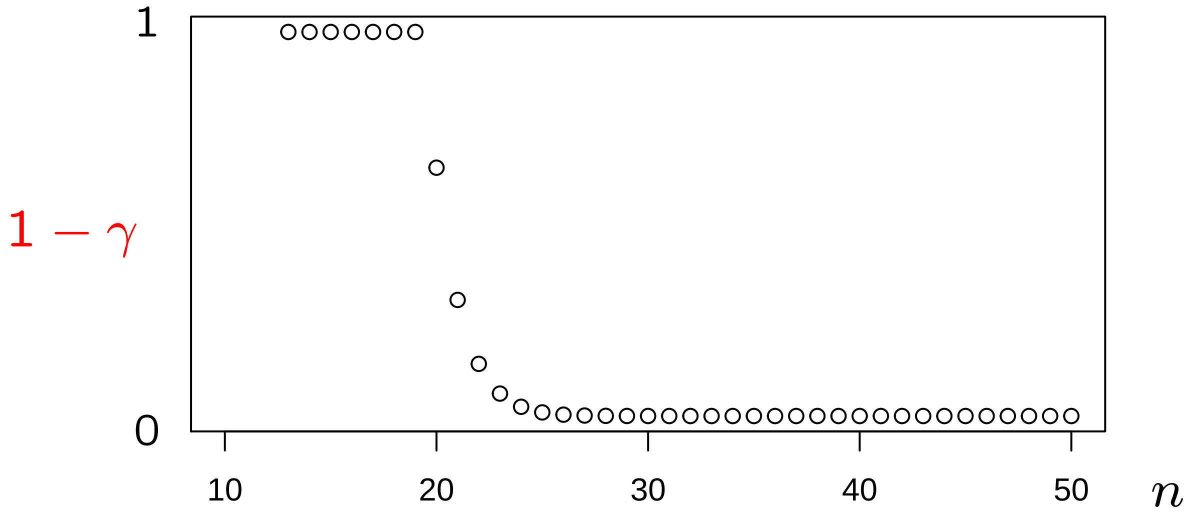


Required credible level for confidence level $1 - \alpha = 0.95$ ($\lambda = 0.05$)

Enlarged credible sets are confidence sets (III)

Example 148 *Again* $p = 0.9$*,* $q = 0.1$ *and confidence level* $1 - \alpha = 0.95$*. For* $\lambda = 0.1$ *and varying graph size* n*,*

any $0.1n$*-enlarged credible set of credible level* $1 - \gamma$ *is also a confidence set of confidence level* 0.95

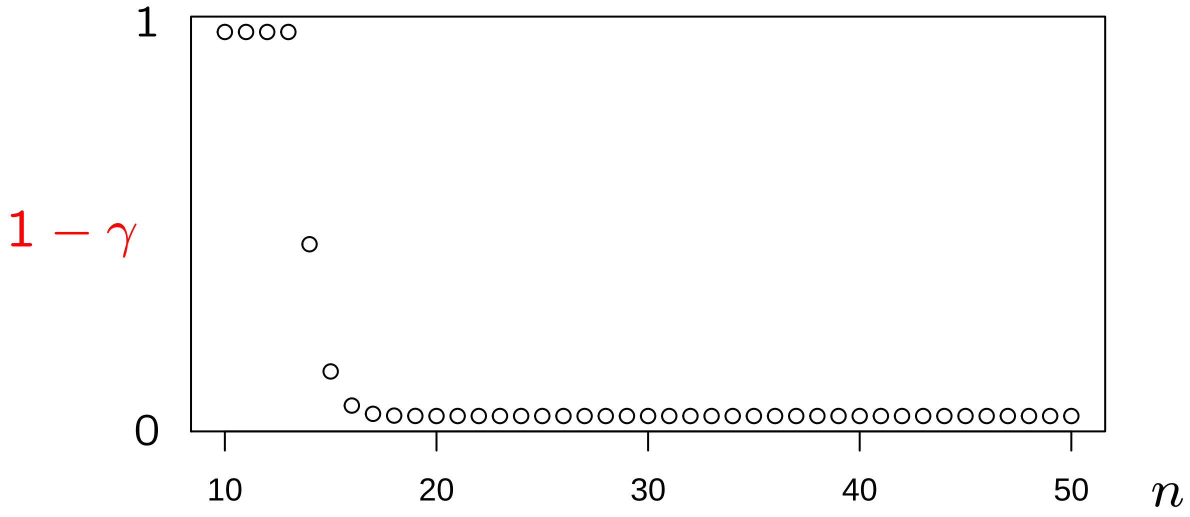


Required credible level for confidence level $1 - \alpha = 0.95$ *(*$\lambda = 0.1$*)*

Enlarged credible sets are confidence sets (IV)

Example 149 *Again* $p = 0.9$, $q = 0.1$ *and confidence level* $1 - \alpha = 0.95$. *For* $\lambda = 0.25$ *and varying graph size* n,

any $0.25n$*-enlarged credible set of credible level* $1 - \gamma$ *is also a confidence set of confidence level* 0.95



Required credible level for confidence level $1 - \alpha = 0.95$ ($\lambda = 0.25$)

Discussion

Sharpness of the bounds If posterior concentration bounds are not sharp, lower bounds for credible levels become unnecessary high and enlargement radii become unnecessarily large.

Early stopping Since only community assignments with high posterior probabilities are needed in credible sets of low credible level, small MCMC samples may not hamper the construction of confidence sets: some form of early stopping of the MCMC sequence may be justified.

Generalization and cross validation All of this generalizes and can be verified by simulation and cross validation.