Nonparametric Bayesian Methods - Lecture II

Harry van Zanten

Korteweg-de Vries Institute for Mathematics


University of Amsterdam

CRiSM Masterclass, April 4-6, 2016

Overview of Lecture II

- Frequentist asymptotics
- Parametric models: Bernstein - Von Mises
- Consistency: Doob and Schwartz
- General rate of contraction results
- Concluding remarks

Frequentist asymptotics

Illustration: nonparametric regression

Suppose we have observations

$$Y_i = f(t_i) + e_i, \qquad i = 1, \ldots, n,$$

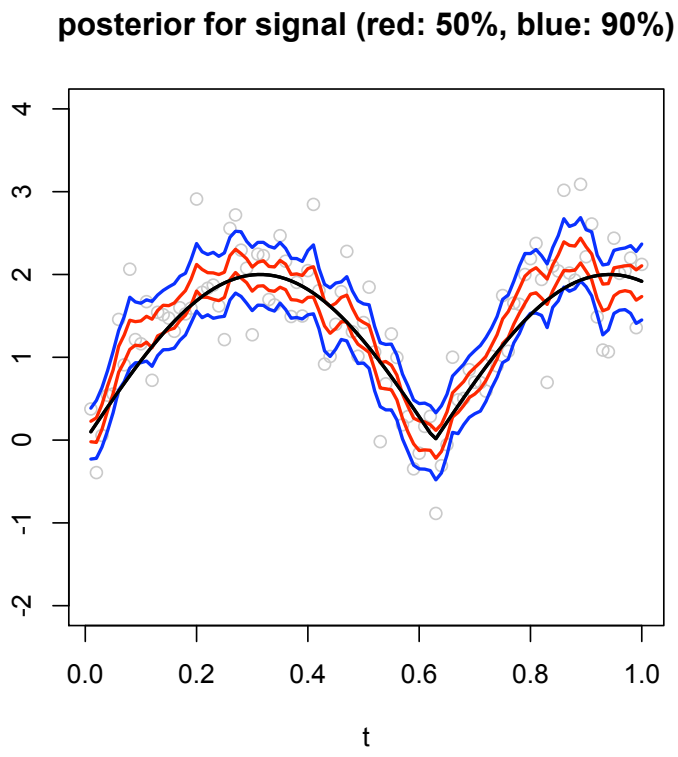
where $t_i = i/n$, f is an unknown, continuous function, e_i are independent $N(0, \sigma^2)$ for some unknown σ.

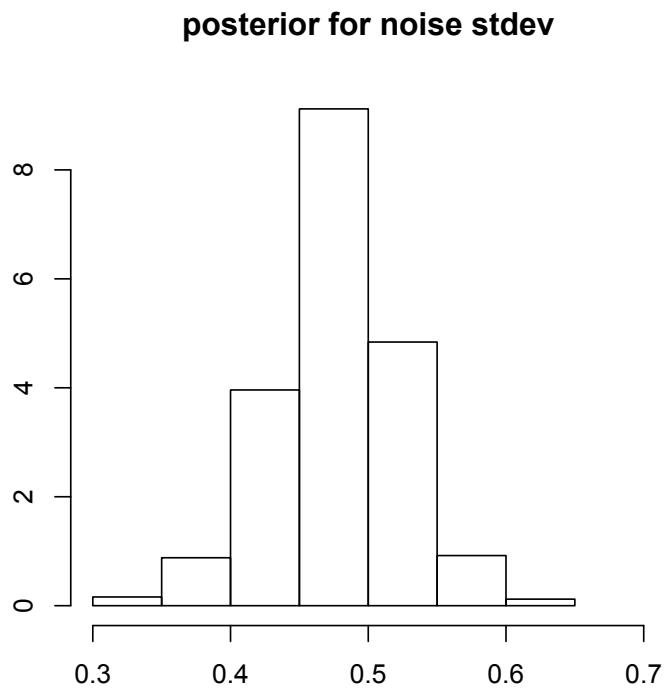
Aim: reconstruct "signal" f.

Approach:

- put priors on f and σ (for f: $\Gamma^{-1} \times$ BM, for σ: Γ^{-1}),
- numerically compute posteriors using Gibbs sampler.

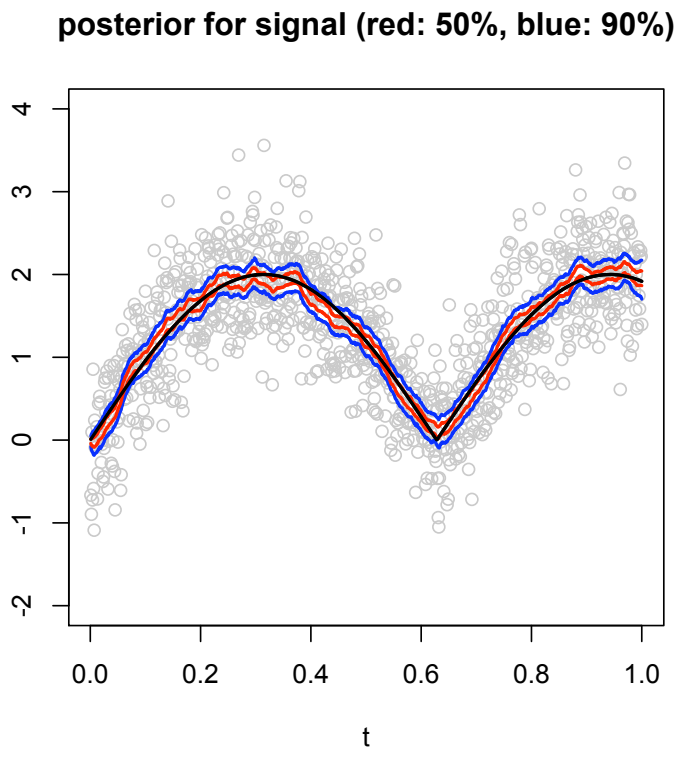
Illustration: nonparametric regression

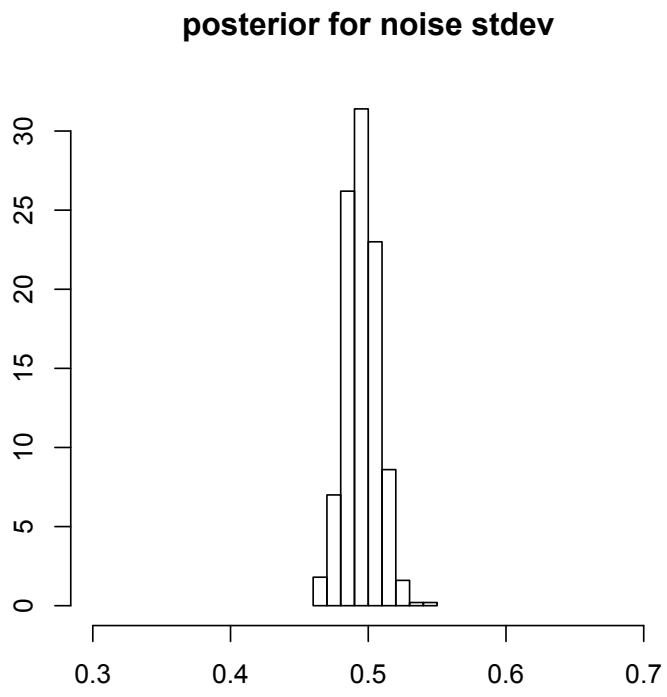




100 observations

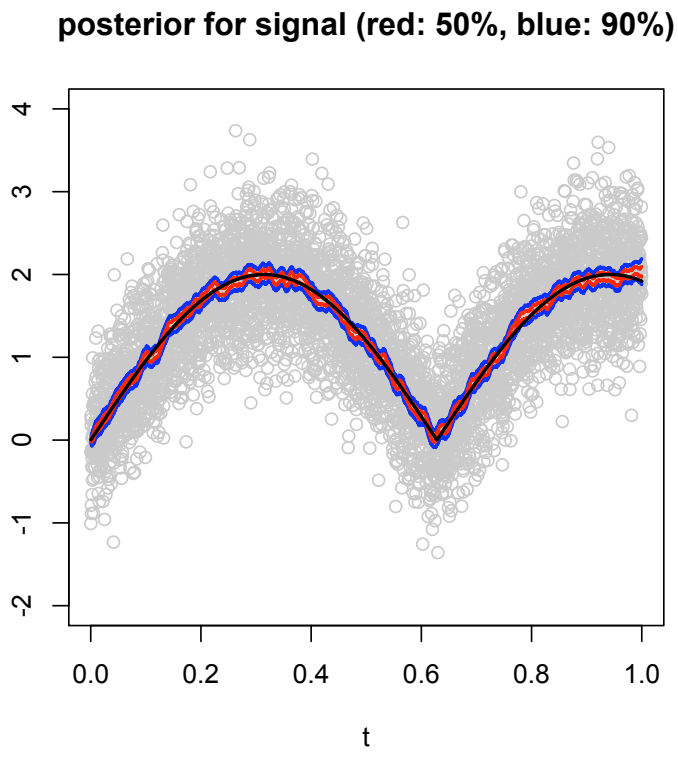
Illustration: nonparametric regression

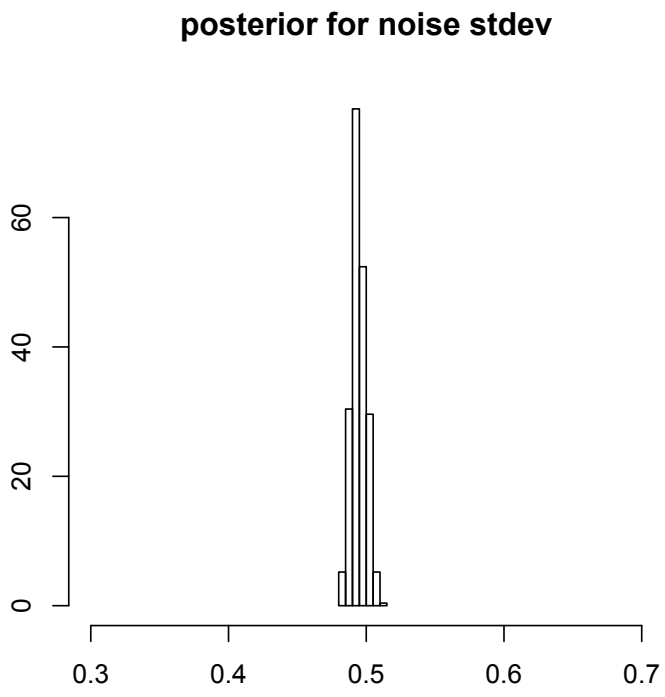




1000 observations

Illustration: nonparametric regression





5000 observations

Illustration: nonparametric regression

Questions we are interested in:

- Why does this work?
- How fast is the convergence to the unknown signal?
- Is this procedure optimal, or can we do better?
- Does this work in other statistical settings as well?
- ...

Frequentist asymptotics - 1

Suppose there is a true parameter θ_0 $(= (f_0, \sigma_0))$ generating the data. Say $\theta_0 \in \Theta$, for a parameter space Θ.

Consider a prior Π on the space Θ, compute the corresponding posterior $\Pi(\cdot \mid Y_1, \ldots, Y_n)$. (Usually, Θ is defined indirectly, as the support of the prior Π).

Say there is some natural distance d on the space Θ. Want to understand if for "large n", "most" of the posterior mass is concentrated "close to" the true parameter θ_0.

Frequentist asymptotics - 2

Main questions, more precisely:

- Consistency: does the posterior contract around θ_0 as $n \to \infty$? I.e., for all $\varepsilon > 0$, is it true that

$$\Pi(\theta \in \Theta : d(\theta, \theta_0) > \varepsilon \,|\, Y_1, \ldots, Y_n) \overset{P_{\theta_0}}{\to} 0?$$

- Contraction rate: how fast can we let $\varepsilon_n \downarrow 0$ such that still

$$\Pi(\theta \in \Theta : d(\theta, \theta_0) > M\varepsilon_n \,|\, Y_1, \ldots, Y_n) \overset{P_{\theta_0}}{\to} 0$$

for all $M > 0$ large enough?

Frequentist asymptotics: a short history

Parametric models: Bernstein-Von Mises theorem (Laplace (1800), Von Mises (30's), Le Cam (80's))

Nonparametrics:

- 40's: Doob's consistency theorem: identifiability implies consistency for Π-almost all θ.
- 60's: negative consistency examples of David Freedman
- '65: Schwartz consistency theorem: prior mass condition, testing condition
- 80's: more negative consistency examples by Diaconis and Freedman
- '99: negative Bernstein-Von Mises examples by Freedman
- '98/'99: Important extensions of Schwartz by Barron, Schervish & Wasserman, and Ghosal, Ghosh & Ramamoorthi
- '00/'01: Rate of contraction results by Ghosal, Ghosh & Van der Vaart and Shen & Wasserman.

Asymptotics for parametric models:

Bernstein - Von Mises

BvM - 1

Let $X_1, \ldots, X_n$ be a sample from a density p_{θ_0}, $\theta_0 \in \Theta \subset \mathbb{R}$. Suppose $\theta \mapsto p_\theta$ is "smooth". Consider MLE and Fisher info:

$$\hat{\theta}_n = \operatorname*{argmax}_{\theta \in \Theta} \prod_{i=1}^{n} p_\theta(X_i),$$

$$i_\theta = \mathbb{V}\mathrm{ar}_\theta \frac{\partial \log p_\theta(X_1)}{\partial \theta}$$

Then under "regularity conditions",

$$\sqrt{n}(\hat{\theta}_n - \theta_0) \Rightarrow N(0, i_{\theta_0}^{-1})$$

under P_{θ_0} as $n \to \infty$.

BvM - 2

Consider a prior Π on Θ with Lebesgue density π. Posterior:

$$\Pi(B \mid X_1, \ldots, X_n) = \frac{\int_B \prod p_\theta(X_i)\pi(\theta)\, d\theta}{\int_\Theta \prod p_\theta(X_i)\pi(\theta)\, d\theta}.$$

BvM: If π is positive and continuous at θ_0, then, under "regularity conditions"

$$\left\| \Pi(\cdot \mid X_1, \ldots, X_n) - N(\hat{\theta}_n, (n i_{\theta_0})^{-1}) \right\|_{TV} \overset{P_{\theta_0}}{\to} 0$$

as $n \to \infty$.

In particular: rate of contraction in the parametric case is $n^{-1/2}$ under mild conditions.

BvM - "proof" - 1

Set $h = \sqrt{n}(\theta - \theta_0)$. Have

$$\Pi(h \in B \,|\, X_1, \ldots, X_n) = \frac{\int_{\theta:\sqrt{n}(\theta-\theta_0)\in B} e^{\sum \ell_\theta(X_i)} \pi(\theta)\, d\theta}{\int_{\mathbb{R}} e^{\sum \ell_\theta(X_i)} \pi(\theta)\, d\theta},$$

with $\ell_\theta(x) = \log p_\theta(x)$. By Taylor,

$$\ell_\theta(x) - \ell_{\theta_0}(x) \approx (\theta - \theta_0)\dot{\ell}_{\theta_0}(x) + \frac{1}{2}(\theta - \theta_0)^2 \ddot{\ell}_{\theta_0}(x).$$

By the LLN, we have $\mathbb{P}_{\theta_0}$-a.s.

$$-\frac{1}{n}\sum_{i=1}^{n} \ddot{\ell}_{\theta_0}(X_i) \to -\mathbb{E}_{\theta_0}\ddot{\ell}_{\theta_0}(X_1) = \mathbb{V}\mathrm{ar}_{\theta_0}\dot{\ell}_{\theta_0}(X_1) = i_{\theta_0}.$$

Hence,

$$\begin{aligned} e^{\sum(\ell_\theta - \ell_{\theta_0})(X_i)} &\approx e^{-\frac{1}{2} i_{\theta_0}(n(\theta-\theta_0)^2 - 2\sqrt{n}(\theta-\theta_0)\Delta_n)} \\ &= e^{-\frac{1}{2} i_{\theta_0}(h-\Delta_n)^2} e^{\frac{1}{2} i_{\theta_0}\Delta_n^2}, \end{aligned}$$

BvM - "proof" - 2

where

$$\Delta_n = \frac{1}{i_{\theta_0}\sqrt{n}} \sum_{i=1}^{n} \dot{\ell}_{\theta_0}(X_i).$$

We get

$$\Pi(\sqrt{n}(\theta - \theta_0) \in B \,|\, X_1, \ldots, X_n) \approx \frac{\int_B e^{-\frac{1}{2} i_{\theta_0}(h-\Delta_n)^2} \pi(\theta_0 + h/\sqrt{n})\, dh}{\int e^{-\frac{1}{2} i_{\theta_0}(h-\Delta_n)^2} \pi(\theta_0 + h/\sqrt{n})\, dh}.$$

Now let $n \to \infty$ and conclude that

$$\Pi(B \,|\, X_1, \ldots, X_n) \approx N\Big(\theta_0 + \frac{\Delta_n}{\sqrt{n}}, \frac{1}{n i_{\theta_0}}\Big)(B).$$

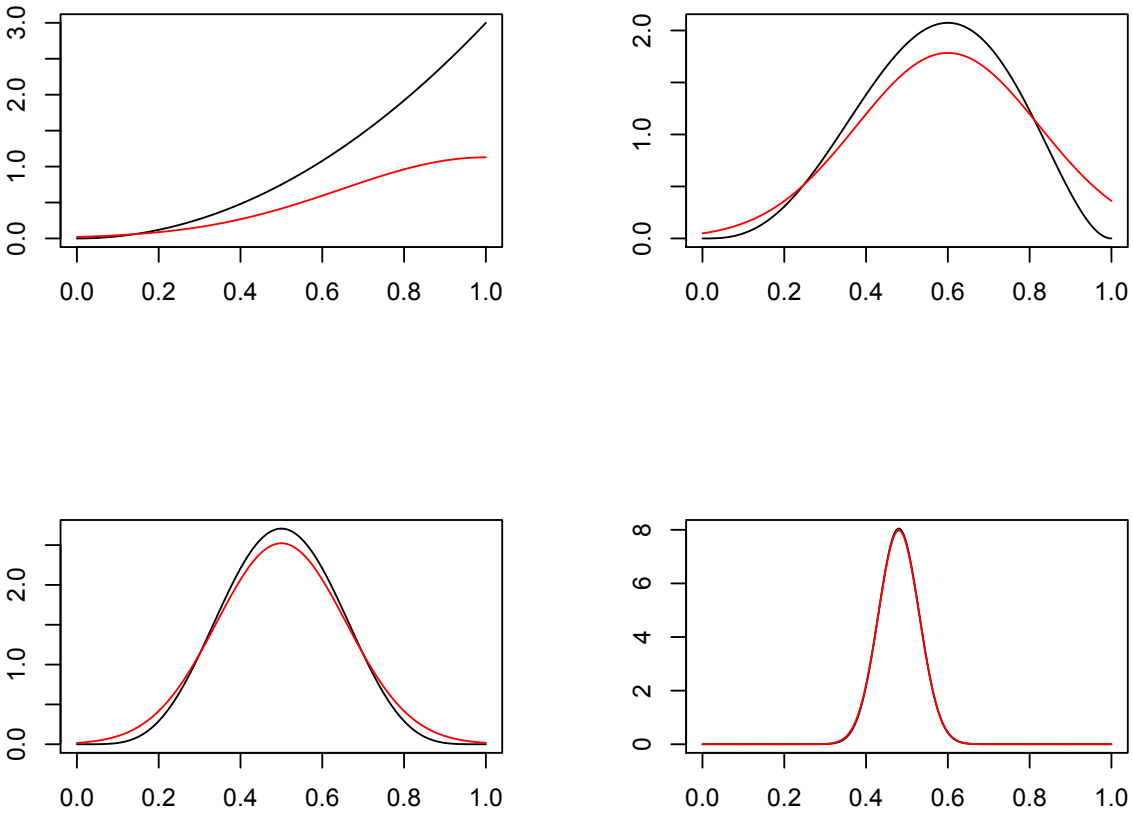
The LAN expansion also implies that

$$\hat{\theta}_n \approx \theta_0 + \frac{\Delta_n}{\sqrt{n}}.$$

Combining the last two displays completes the "proof". □

BvM - numerical illustration

Consider $X \sim \text{Bin}(n, p)$, uniform prior on p. Here $\hat{p} = X/n$, $I_p = n/(p(1-p))$. Posterior is $\text{Beta}(X+1, n-X+1)$.



n = 2, 5, 10, 100

Consistency:

Doob and Schwartz

Doob's theorem

Consider i.i.d. $X_1, \ldots, X_n \sim P_\theta$, $\theta \in \Theta$, for a "nice" metric space (Θ, d). Assume that $\theta \mapsto P_\theta$ is appropriately measurable. Let Π be a prior on (the Borel sets of) Θ.

Theorem.
Suppose that if $\theta_1 \neq \theta_2$, then $P_{\theta_1} \neq P_{\theta_2}$ (identifiability). Then for Π-almost all $\theta_0 \in \Theta$ and all $\varepsilon > 0$:

$$\Pi(\theta : d(\theta, \theta_0) > \varepsilon \,|\, X_1, \ldots X_n) \to 0,$$

$\mathbb{P}_{\theta_0}$-a.s..

This is called (strong) posterior consistency, or consistency at θ_0.

Doob's theorem

Consider i.i.d. $X_1, \ldots, X_n \sim P_\theta$, $\theta \in \Theta$, for a "nice" metric space (Θ, d). Assume that $\theta \mapsto P_\theta$ is appropriately measurable. Let Π be a prior on (the Borel sets of) Θ.

Theorem.
Suppose that if $\theta_1 \neq \theta_2$, then $P_{\theta_1} \neq P_{\theta_2}$ (identifiability). Then for Π-almost all $\theta_0 \in \Theta$ and all $\varepsilon > 0$:

$$\Pi(\theta : d(\theta, \theta_0) > \varepsilon \,|\, X_1, \ldots X_n) \to 0,$$

$\mathbb{P}_{\theta_0}$-a.s..

This is called (strong) posterior consistency, or consistency at θ_0.

Side remark: relation to consistency of estimators

Proposition.

Suppose we have posterior consistency at θ_0, relative to the metric d. Define the estimator $\hat{\theta}_n$ as the center of a ball of minimal radius that has posterior mass at least $1/2$. Then $\hat{\theta}_n$ is consistent at θ_0, i.e.

$$d(\hat{\theta}_n, \theta_0) \to 0,$$

$\mathbb{P}_{\theta_0}$-a.s..

Proof.

Let $B(\hat{\theta}_n, \hat{r})$ be a ball of minimal radius that has posterior mass at least $1/2$. For every $\varepsilon > 0$, $B(\theta_0, \varepsilon)$ asymptotically contains posterior mass 1. Hence, $\hat{r} \leq \varepsilon$. Moreover, the balls can not be disjoint. By the triangle inequality, it follows that, asymptotically, $d(\hat{\theta}_n, \theta_0) \leq \hat{r} + \varepsilon \leq 2\varepsilon$. □

Doob's theorem - "proof" - 1

Let Q be the joint distribution of θ and $X_1, X_2, \ldots$ in the Bayesian framework, i.e. under Q have $\theta \sim \Pi$ and $X_1, X_2, \ldots \mid \theta$ are i.i.d. P_θ.

The posterior is the Q-conditional distribution of $\theta \mid X_1, \ldots, X_n$.

Note:

- LLN implies that $\forall \theta$, for P_θ-almost all $X_1, X_2, \ldots$, can identify P_θ from $X_1, X_2, \ldots$:

$$\frac{1}{n} \sum_{i=1}^{n} 1_A(X_i) \to P_\theta(A), \qquad P_\theta\text{-a.s..}$$

- Identifiability assumption implies we can identify θ from P_θ.

Doob's theorem - "proof" - 2

Using some measure theory, it follows $\exists$ measurable $h : \mathbb{R}^\infty \to \Theta$:

$$h(x_1, x_2, \ldots) = \theta, \qquad \text{for } Q\text{-almost all } (\theta, x_1, x_2, \ldots).$$

Using Doob's martingale convergence theorem, we get, Q-a.s.

$$\Pi(\theta : d(\theta, \theta_0) > \varepsilon \,|\, X_1, \ldots, X_n) = \mathbb{E}_Q(1_{d(\theta,\theta_0)>\varepsilon} \,|\, X_1, \ldots, X_n)$$
$$\to \mathbb{E}_Q(1_{d(\theta,\theta_0)>\varepsilon} \,|\, X_1, X_2, \ldots) = 1_{d(\theta,\theta_0)>\varepsilon} = 1_{d(h(X_1,X_2,\ldots),\theta_0)>\varepsilon}.$$

From this, can derive that for Π-almost all θ_0, P_{θ_0}-a.s.

$$\Pi(\theta : d(\theta, \theta_0) > \varepsilon \,|\, X_1, \ldots, X_n) \to 0.$$

□

Limitations of Doob's theorem

Main issues:

- In infinite-dimensional spaces, null sets can be very large.
  $\rightarrow \exists$ examples of inconsistent procedures (take $\Pi = \delta_\theta$).
- Result is very pessimistic!
  $\rightarrow$ in many cases of interest, consistency actually holds for many more θ_0 than Doob says.

Need a different approach to obtain less pessimistic results...

Limitations of Doob's theorem

Main issues:

- In infinite-dimensional spaces, null sets can be very large.
  $\rightarrow \exists$ examples of inconsistent procedures (take $\Pi = \delta_\theta$).
- Result is very pessimistic!
  $\rightarrow$ in many cases of interest, consistency actually holds for many more θ_0 than Doob says.

Need a different approach to obtain less pessimistic results...

Schwartz' theorem - setting

Observations: sample $X_1, \ldots, X_n$ from a density $p_0 \in \mathcal{P}$, for $\mathcal{P}$ the collection of densities on the unit interval.

Prior: measure Π on $\mathcal{P}$

Posterior:

$$\Pi(B \mid X_1, \ldots, X_n) = \frac{\int_B \prod_{i=1}^n p(X_i)\, \Pi(dp)}{\int_{\mathcal{P}} \prod_{i=1}^n p(X_i)\, \Pi(dp)}.$$

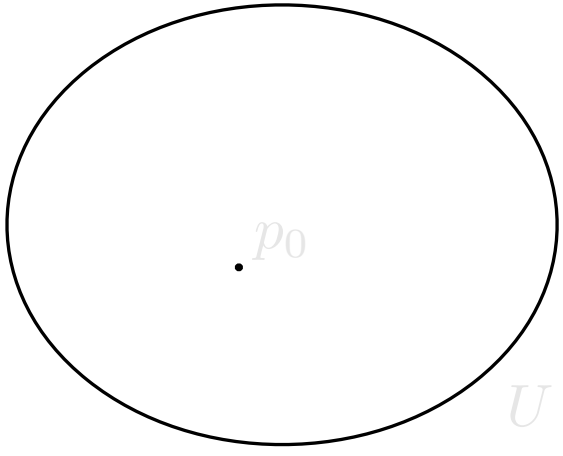
Schwartz' theorem - 1

Idea: replace identifiability by a stronger condition on testability.

Assume that for a neighborhood U of p_0: can consistently test $H_0 : p = p_0$ against $H_1 : p \in U^c$.

More precisely, assume there exist $[0, 1]$-valued tests $\varphi_n = \varphi_n(X_1, \ldots, X_n)$ such that

$$\mathbb{E}_{p_0} \varphi_n \to 0 \qquad \sup_{p \in U^c} \mathbb{E}_p (1 - \varphi_n) \to 0.$$



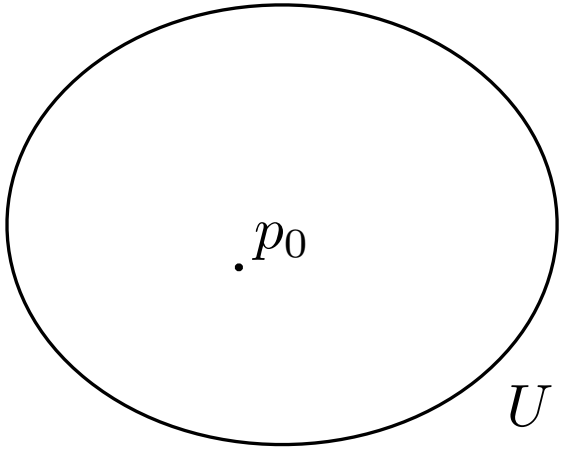
Interpretation: φ_n = probab. of rejecting $H_0 : p = p_0$.

Schwartz' theorem - 1

Idea: replace identifiability by a stronger condition on testability.
Assume that for a neighborhood U of p_0: can consistently test $H_0 : p = p_0$ against $H_1 : p \in U^c$.

More precisely, assume there exist $[0, 1]$-valued tests $\varphi_n = \varphi_n(X_1, \ldots, X_n)$ such that

$$\mathbb{E}_{p_0} \varphi_n \to 0 \qquad \sup_{p \in U^c} \mathbb{E}_p (1 - \varphi_n) \to 0.$$



Interpretation: φ_n = probab. of rejecting $H_0 : p = p_0$.

Schwartz' theorem - 2

Consistency for which p_0?

Define the Kulback-Leibler divergence

$$K(p,q) = \int p(x) \log \frac{p(x)}{q(x)}\, dx.$$

Consistency for Π-almost all p_0 will be replaced by consistency for all p_0 in the KL-support of Π: say that $p_0 \in KL(\Pi)$ if for all $\varepsilon > 0$,

$$\Pi(p : K(p_0, p) < \varepsilon) > 0.$$

Hence, get consistency if prior puts mass "arbitrarily close" to p_0 (in KL-sense) $\rightarrow$ enter approximation theory.

Schwartz' theorem - 2

Consistency for which p_0?

Define the Kulback-Leibler divergence

$$K(p,q) = \int p(x) \log \frac{p(x)}{q(x)}\, dx.$$

Consistency for Π-almost all p_0 will be replaced by consistency for all p_0 in the KL-support of Π: say that $p_0 \in KL(\Pi)$ if for all $\varepsilon > 0$,

$$\Pi(p : K(p_0, p) < \varepsilon) > 0.$$

Hence, get consistency if prior puts mass "arbitrarily close" to p_0 (in KL-sense) $\rightarrow$ enter approximation theory.

Schwartz' theorem - 2

Consistency for which p_0?

Define the Kulback-Leibler divergence

$$K(p, q) = \int p(x) \log \frac{p(x)}{q(x)} \, dx.$$

Consistency for Π-almost all p_0 will be replaced by consistency for all p_0 in the KL-support of Π: say that $p_0 \in KL(\Pi)$ if for all $\varepsilon > 0$,

$$\Pi(p : K(p_0, p) < \varepsilon) > 0.$$

Hence, get consistency if prior puts mass "arbitrarily close" to p_0 (in KL-sense) $\rightarrow$ enter approximation theory.

Schwartz' theorem - 2

Consistency for which p_0?

Define the Kulback-Leibler divergence

$$K(p, q) = \int p(x) \log \frac{p(x)}{q(x)} \, dx.$$

Consistency for Π-almost all p_0 will be replaced by consistency for all p_0 in the KL-support of Π: say that $p_0 \in KL(\Pi)$ if for all $\varepsilon > 0$,

$$\Pi(p : K(p_0, p) < \varepsilon) > 0.$$

Hence, get consistency if prior puts mass "arbitrarily close" to p_0 (in KL-sense) $\rightarrow$ enter approximation theory.

Schwartz' theorem - 1

Theorem.
Suppose that p_0 is in the KL-support of the prior and that for a neighborhood $U \subset \mathcal{P}$ of p_0, there exist tests such that $\mathbb{E}_{p_0}\varphi_n \to 0$ and $\sup_{p \in U^c} \mathbb{E}_p(1-\varphi_n) \to 0$. Then

$$\Pi(U^c \mid X_1, \ldots, X_n) \to 0$$

$\mathbb{P}_0$-a.s..

Hence we have (strong) consistency at p_0 if (i) tests exist for *every* neighborhood U of p_0 and (ii) p_0 is in the KL-support of the prior.

Schwartz' theorem - 1

Theorem.
Suppose that p_0 is in the KL-support of the prior and that for a neighborhood $U \subset \mathcal{P}$ of p_0, there exist tests such that $\mathbb{E}_{p_0}\varphi_n \to 0$ and $\sup_{p \in U^c} \mathbb{E}_p(1-\varphi_n) \to 0$. Then

$$\Pi(U^c \mid X_1, \ldots, X_n) \to 0$$

$\mathbb{P}_0$-a.s..

Hence we have (strong) consistency at p_0 if (i) tests exist for *every* neighborhood U of p_0 and (ii) p_0 is in the KL-support of the prior.

Schwartz' theorem - "proof"

Write

$$\Pi(U^c \mid X_1, \ldots, X_n) \leq \varphi_n + \frac{\int_{U^c}(1-\varphi_n)\frac{d\mathbb{P}^n}{d\mathbb{P}_0^n}\,\Pi(dp)}{\int_{\mathcal{P}}\frac{d\mathbb{P}^n}{d\mathbb{P}_0^n}\,\Pi(dp)}.$$

Denominator: restrict integral to KL ball + Jensen+ LLN, get

$$\text{denominator} \geq e^{-n\varepsilon}\Pi(p_0 : KL(p_0, p) < \varepsilon).$$

Combine with testing assumptions. Use that by Hoeffding, can assume tests have exponential power. □

Schwartz' theorem - weak topology

Note: the testing condition depends on the topology.

Example.

If U is a weak neighborhood of the form

$$U = \left\{ p : \mathbb{E}_p \psi(X_1) < \mathbb{E}_{p_0} \psi(X_1) + \varepsilon \right\}$$

for a bounded, continuous ψ and $\varepsilon > 0$, then required tests always exist, by Hoeffding.

Hence, always have consistency relative to the weak topology for every p_0 in the KL-support of the prior.

For consistency in stronger topologies (e.g. Hellinger, L^1), more is needed.

Extended Schwartz theorem

More useful version:

Theorem.
Suppose that p_0 is in the KL-support of the prior and that for a neighborhood $U \subset \mathcal{P}$ of p_0, there exist tests φ_n and $C > 0$ and $\mathcal{P}_n \subset \mathcal{P}$ such that

$$\mathbb{E}_{p_0}\varphi_n \leq e^{-Cn}, \qquad \sup_{p \in U^c \cap \mathcal{P}_n} \mathbb{E}_p(1-\varphi_n) \leq e^{-Cn}, \quad \Pi(\mathcal{P}_n^c) \leq e^{-Cn}.$$

Then

$$\Pi(U^c \mid X_1, \ldots, X_n) \to 0$$

$\mathbb{P}_0$-a.s..

$\mathcal{P}_n$: sieves. Idea: sets with very little prior mass (like $\mathcal{P}_n^c$) get no posterior mass, asymptotically.

Extended Schwartz theorem - Hellinger topology - 1

Define the Hellinger distance by

$$h^2(p,q) = \int (\sqrt{p(x)} - \sqrt{q(x)})^2 \, dx.$$

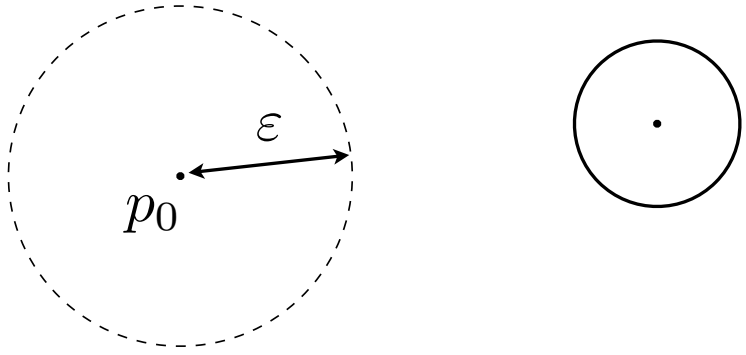
Theorem.
For every *convex* $\mathcal{Q} \subset \mathcal{P}$ such that $h(p_0, p) > \varepsilon$ for all $p \in \mathcal{Q}$, there exists a test φ_n s.t.

$$\mathbb{E}_{p_0} \varphi_n \leq e^{-n\varepsilon^2/2}, \qquad \sup_{p \in \mathcal{Q}} \mathbb{E}_p (1 - \varphi_n) \leq e^{-n\varepsilon^2/2}.$$

"Proof."
Minimize $\varphi_n \mapsto \mathbb{E}_{p_0} \varphi_n + \sup_{p \in \mathcal{Q}} \mathbb{E}_p (1 - \varphi_n)$ over all tests φ_n using the minimax theorem. □

Extended Schwartz theorem - Hellinger topology - 2

Theorem gives tests for $H_0 : p = p_0$ against the hypothesis H_1 that p is in a ball at Hellinger distance at least ε from p_0:

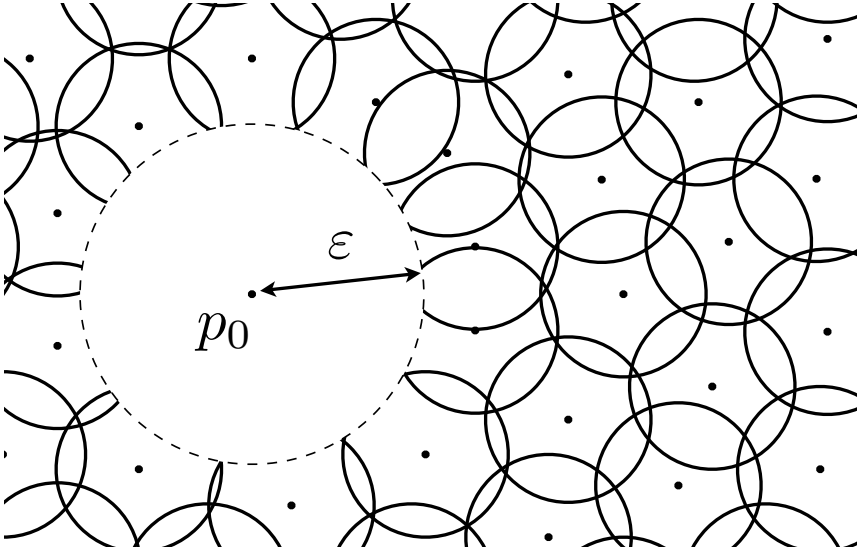


For consistency relative to the Hellinger distance, need test for p_0 against the complement of the ε-ball around p_0 (intersected with a sieve $\mathcal{P}_n$).

Extended Schwartz theorem - Hellinger topology - 3

Idea:

- Cover the complement of the ε-ball with small balls.
- For every such small ball, have a local test for p_0 against that small ball.
- Make a new global test that rejects $H_0 : p = p_0$ if any of the local tests rejects it.



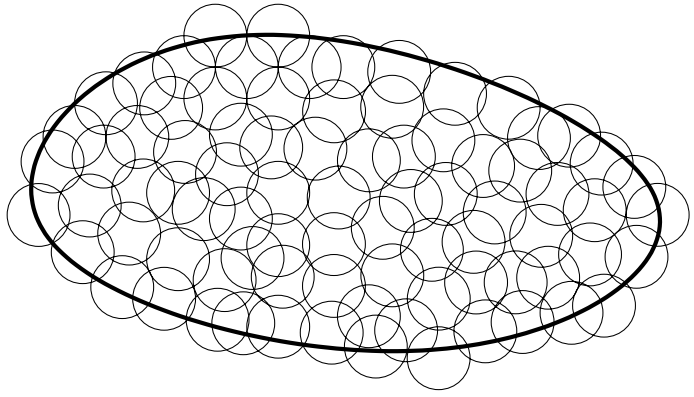
Extended Schwartz theorem - Hellinger topology - 4

Power of the new global test depends on the number of small balls needed.

Covering number: for $\mathcal{Q} \subset \mathcal{P}$ and $\varepsilon > 0$, define

$N(\varepsilon, \mathcal{Q}, h) =$

minimum number of h-balls of radius ε needed to cover $\mathcal{Q}$.



We call $\log N(\varepsilon, \mathcal{Q}, h)$ the metric entropy of $\mathcal{Q}$ w.r.t. h.

Extended Schwartz theorem - Hellinger topology - 5

Theorem.
Suppose that $p_0 \in KL(\Pi)$ and that for every $\varepsilon > 0$, there exist $\mathcal{P}_n \subset \mathcal{P}$ and constants $C < 6$ and $D > 0$ such that $N(\varepsilon, \mathcal{P}_n, h) \leq \exp(Cn\varepsilon^2)$ and $\Pi(\mathcal{P}_n^c) \leq \exp(-Dn)$. Then we have posterior consistency w.r.t. the Hellinger distance.

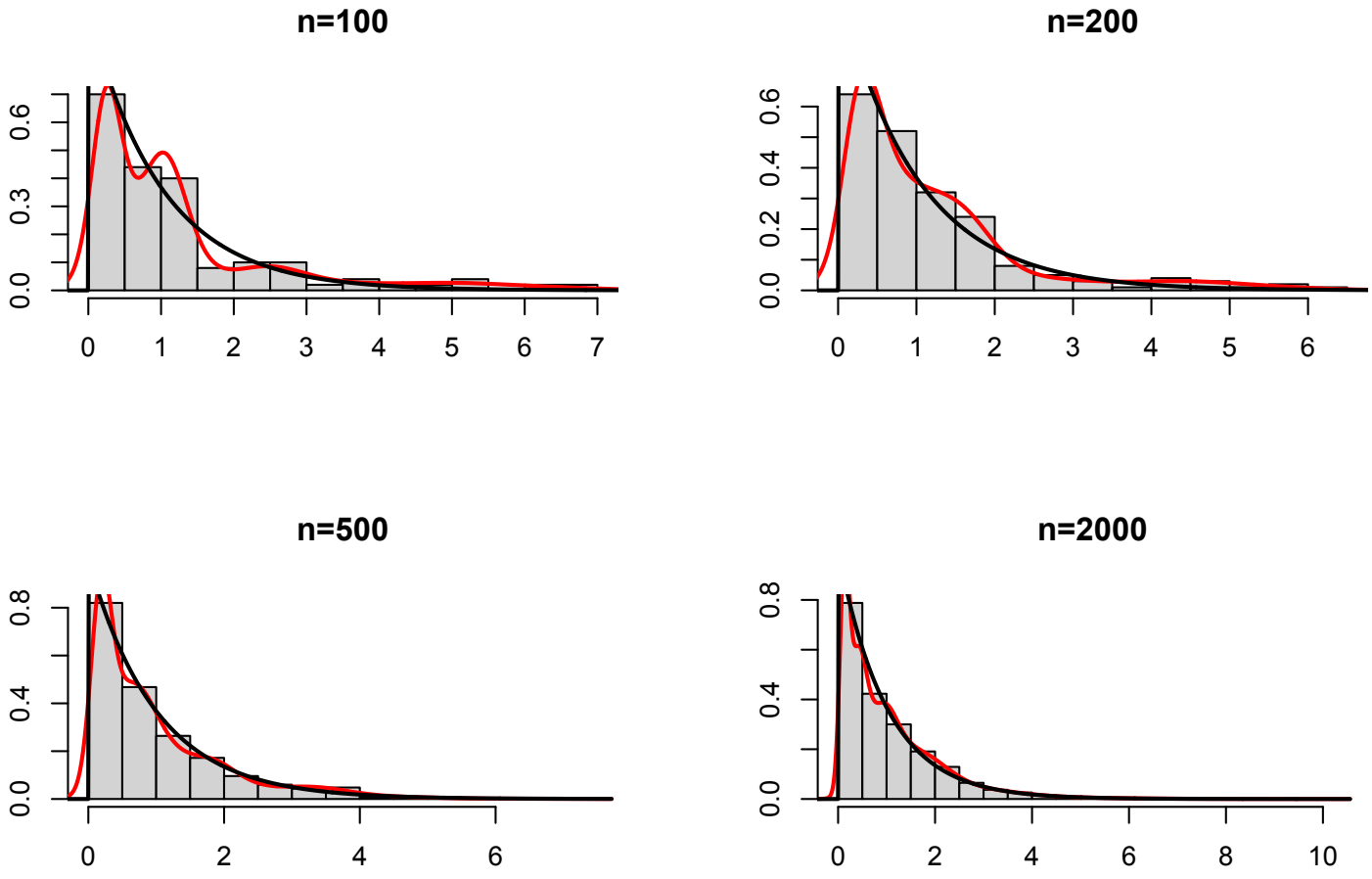
Conditions essentially:

- Should have true density in KL-support of the prior.
- All but a negligible amount of prior mass should be should be concentrated on a set whose "size", or "complexity" is not too large.

Concrete examples: many, but case-by-case analysis...

Consistency - example

Truth: exponential, prior: Dirichlet mixture of normals



[Ghosal, Ghosh, Ramamoorthi (1999)]

General rate of contraction results

Posterior contraction

Consistency: all posterior mass is ultimately contained in arbitrarily small neighborhoods of the true parameter.

Posterior contraction: how fast can we let these neighborhoods shrink, while still capturing all the posterior mass in the limit?

In other words: find the fastest converging $\varepsilon_n \downarrow 0$ such that asymptotically, all posterior mass in located in balls around θ_0 with radius of the order ε_n.

Definition: we say the posterior contracts around θ_0 at the rate ε_n if

$$\Pi(\theta \in \Theta : d(\theta, \theta_0) > M\varepsilon_n \,|\, X_1, \ldots, X_n) \xrightarrow{P_{\theta_0}} 0$$

for all $M > 0$ large enough.

Posterior contraction

Consistency: all posterior mass is ultimately contained in arbitrarily small neighborhoods of the true parameter.

Posterior contraction: how fast can we let these neighborhoods shrink, while still capturing all the posterior mass in the limit?

In other words: find the fastest converging $\varepsilon_n \downarrow 0$ such that asymptotically, all posterior mass in located in balls around θ_0 with radius of the order ε_n.

Definition: we say the posterior contracts around θ_0 at the rate ε_n if

$$\Pi(\theta \in \Theta : d(\theta, \theta_0) > M\varepsilon_n \,|\, X_1, \ldots, X_n) \overset{P_{\theta_0}}{\to} 0$$

for all $M > 0$ large enough.

Posterior contraction

Consistency: all posterior mass is ultimately contained in arbitrarily small neighborhoods of the true parameter.

Posterior contraction: how fast can we let these neighborhoods shrink, while still capturing all the posterior mass in the limit?

In other words: find the fastest converging $\varepsilon_n \downarrow 0$ such that asymptotically, all posterior mass in located in balls around θ_0 with radius of the order ε_n.

Definition: we say the posterior contracts around θ_0 at the rate ε_n if

$$\Pi(\theta \in \Theta : d(\theta, \theta_0) > M\varepsilon_n \,|\, X_1, \ldots, X_n) \xrightarrow{P_{\theta_0}} 0$$

for all $M > 0$ large enough.

Posterior contraction

Consistency: all posterior mass is ultimately contained in arbitrarily small neighborhoods of the true parameter.

Posterior contraction: how fast can we let these neighborhoods shrink, while still capturing all the posterior mass in the limit?

In other words: find the fastest converging $\varepsilon_n \downarrow 0$ such that asymptotically, all posterior mass in located in balls around θ_0 with radius of the order ε_n.

Definition: we say the posterior contracts around θ_0 at the rate ε_n if

$$\Pi(\theta \in \Theta : d(\theta, \theta_0) > M\varepsilon_n \,|\, X_1, \ldots, X_n) \overset{P_{\theta_0}}{\to} 0$$

for all $M > 0$ large enough.

Relation to convergence rates of estimators - 1

Proposition.

Suppose we have posterior contraction around θ_0 at the rate ε_n, relative to the metric d. Define the estimator $\hat{\theta}_n$ as the center of a ball of minimal radius that has posterior mass at least $1/2$. Then $d(\hat{\theta}_n, \theta_0) = O_{P_{\theta_0}}(\varepsilon_n)$, i.e. for all $\varepsilon > 0$, there exists $M > 0$ s.t.

$$\mathbb{P}_{\theta_0}(d(\hat{\theta}_n, \theta_0) > M\varepsilon_n) \leq \varepsilon.$$

Proof.

Let $B(\hat{\theta}_n, \hat{r})$ a ball of minimal radius that has posterior mass at least $1/2$. For every $\varepsilon > 0$, there exists an $M > 0$ s.t. $B(\theta_0, M\varepsilon_n)$ asymptotically contains posterior mass $1 - \varepsilon$ w.p. $1 - \varepsilon$. Hence on that event, $\hat{r} \leq M\varepsilon_n$. Moreover, the balls can not be disjoint. By the triangle inequality, it follows that, asymptotically w.p. $1 - \varepsilon$, $d(\hat{\theta}_n, \theta_0) \leq \hat{r} + M\varepsilon_n \leq 2M_n\varepsilon_n$. □

Relation to convergence rates of estimators - 2

So if we have posterior contraction at rate ε_n, then there exists an estimator with convergence rate ε_n.

Important consequence: posterior contraction rates are limited by the best possible convergence rates of frequentist estimators.

For many statistical problems, lower bounds (e.g. of minimax type) for convergence rates of estimators are known. Posteriors can never contract faster.

Under regularity conditions, best possible rate at which we can estimate a β-smooth function of d variables is typically $n^{-\beta/(d+2\beta)}$.

Q: which priors produce posteriors that contract at this optimal rate?

Relation to convergence rates of estimators - 2

So if we have posterior contraction at rate ε_n, then there exists an estimator with convergence rate ε_n.

Important consequence: posterior contraction rates are limited by the best possible convergence rates of frequentist estimators.

For many statistical problems, lower bounds (e.g. of minimax type) for convergence rates of estimators are known. Posteriors can never contract faster.

Under regularity conditions, best possible rate at which we can estimate a β-smooth function of d variables is typically $n^{-\beta/(d+2\beta)}$.

Q: which priors produce posteriors that contract at this optimal rate?

Recall: Extended Schwartz theorem

Observations: sample $X_1, \ldots, X_n$ from a density $p_0 \in \mathcal{P}$, for $\mathcal{P}$ the collection of densities on the unit interval. Prior: measure Π on $\mathcal{P}$

Theorem.
If there exist $\mathcal{P}_n \subset \mathcal{P}$ and $C < 6$, $D > 0$ such that for every $\varepsilon > 0$

$$\begin{aligned} \Pi(p : K(p_0, p) < \varepsilon) &> 0, \\ \Pi(\mathcal{P}_n^c) &\leq e^{-Dn}, \\ \log N(\varepsilon, \mathcal{P}_n, h) &\leq Cn\varepsilon^2. \end{aligned}$$

Then

$$\Pi(p : h(p, p_0) > \varepsilon \,|\, X_1, \ldots, X_n) \xrightarrow{P_0} 0$$

as $n \to \infty$.

General contraction rate theorem - 1

Distances:

$$h^2(p,q) = \int (\sqrt{p(x)} - \sqrt{q(x)})^2 \, dx \qquad \text{(Hellinger)}$$

$$K(p,q) = \int p(x) \log \frac{p(x)}{q(x)} \, dx \qquad \text{(Kulback-Leibler)}$$

$$V(p,q) = \int p(x) \Big(\log \frac{p(x)}{q(x)} \Big)^2 \, dx$$

KL-type ball:

$$B_n(p_0, \varepsilon) = \{p \in \mathcal{P} : K(p_0, p) \leq \varepsilon^2, V(p_0, p) \leq \varepsilon^2\}.$$

General contraction rate theorem - 2

Theorem.
If there exist $\mathcal{P}_n \subset \mathcal{P}$ and positive numbers ε_n such that $n\varepsilon_n^2 \to \infty$ and, for some $c > 0$,

$$\begin{aligned} \Pi(B_n(p_0, \varepsilon_n)) &\geq e^{-cn\varepsilon_n^2}, \\ \Pi(\mathcal{P}_n^c) &\leq e^{-(c+4)n\varepsilon_n^2}, \\ \log N(\varepsilon_n, \mathcal{P}_n, h) &\leq n{\varepsilon_n}^2, \end{aligned}$$

then for $M > 0$ large enough

$$\Pi(p : h(p, p_0) > M\varepsilon_n \,|\, X_1, \ldots, X_n) \xrightarrow{P_0} 0.$$

[Ghosal, Ghosh and Van der Vaart (2000)]

General contraction rate theorem - 3

- Proof: refinement of Schwartz.
- Theorem gives the "right" rates for many priors.
- Verifying the three conditions for a particular prior can be hard! [Shen, Tokdar, Ghosal (2013)]
- Versions of this theorem now exist for many nonparametric statistical settings: regression, classification, signal-in-white-noise, drift estimation for diffusions, Markov chains, time series, ...

The general theorem only becomes useful when combined with techniques tailored to specific (classes of) priors!

General contraction rate theorem - 3

- Proof: refinement of Schwartz.
- Theorem gives the "right" rates for many priors.
- Verifying the three conditions for a particular prior can be hard! [Shen, Tokdar, Ghosal (2013)]
- Versions of this theorem now exist for many nonparametric statistical settings: regression, classification, signal-in-white-noise, drift estimation for diffusions, Markov chains, time series, ...

The general theorem only becomes useful when combined with techniques tailored to specific (classes of) priors!

General contraction rate theorem - 3

- Proof: refinement of Schwartz.
- Theorem gives the "right" rates for many priors.
- Verifying the three conditions for a particular prior can be hard! [Shen, Tokdar, Ghosal (2013)]
- Versions of this theorem now exist for many nonparametric statistical settings: regression, classification, signal-in-white-noise, drift estimation for diffusions, Markov chains, time series, . . .

The general theorem only becomes useful when combined with techniques tailored to specific (classes of) priors!

General contraction rate theorem - 3

- Proof: refinement of Schwartz.
- Theorem gives the "right" rates for many priors.
- Verifying the three conditions for a particular prior can be hard! [Shen, Tokdar, Ghosal (2013)]
- Versions of this theorem now exist for many nonparametric statistical settings: regression, classification, signal-in-white-noise, drift estimation for diffusions, Markov chains, time series, ...

The general theorem only becomes useful when combined with techniques tailored to specific (classes of) priors!

General contraction rate theorem - 3

- Proof: refinement of Schwartz.
- Theorem gives the "right" rates for many priors.
- Verifying the three conditions for a particular prior can be hard! [Shen, Tokdar, Ghosal (2013)]
- Versions of this theorem now exist for many nonparametric statistical settings: regression, classification, signal-in-white-noise, drift estimation for diffusions, Markov chains, time series, ...

The general theorem only becomes useful when combined with techniques tailored to specific (classes of) priors!

Concluding remarks

Take home from Lecture II

- Frequentist notions like consistency and convergence rates can be useful to assess performance of nonparametric Bayes procedures.
- Contrary to the parametric case, performance depends crucially on the fine properties of the prior!
- We have general theorems that give conditions for consistency or contraction rates in terms of (i) the amount of mass that the prior gives to neighborhoods of the truth, (ii) the "size", or "complexity" of the sets where the prior puts all but a negligible amount of mass.

Q: how do we verify these conditions for interesting priors?

Take home from Lecture II

- Frequentist notions like consistency and convergence rates can be useful to assess performance of nonparametric Bayes procedures.
- Contrary to the parametric case, performance depends crucially on the fine properties of the prior!
- We have general theorems that give conditions for consistency or contraction rates in terms of (i) the amount of mass that the prior gives to neighborhoods of the truth, (ii) the "size", or "complexity" of the sets where the prior puts all but a negligible amount of mass.

Q: how do we verify these conditions for interesting priors?

Some references for Lecture II - 1

BvM and Doob's theorem:

- Van der Vaart, A. W. (2000). Asymptotic statistics. Cambridge university press.

Consistency:

- Doob, J. L. (1948). Application of the theory of martingales. Le calcul des probabilites et ses applications, 23-27.
- Schwartz, L. (1965). On Bayes procedures. Probability Theory and Related Fields, 4(1), 10-26.

Freedman/Diaconis:

- Diaconis, P., and Freedman, D. (1986). On the consistency of Bayes estimates. The Annals of Statistics, 1-26.
- Freedman, D. (1999). On the Bernstein-von Mises theorem with infinite-dimensional parameters. Annals of Statistics, 1119-1140.

Some references for Lecture II - 2

Extended consistency theorems:

- Barron, A., Schervish, M. J., and Wasserman, L. (1999). The consistency of posterior distributions in nonparametric problems. The Annals of Statistics, 27(2), 536-561.
- Ghosal, S., Ghosh, J. K., and Ramamoorthi, R. V. (1999). Posterior consistency of Dirichlet mixtures in density estimation. Ann. Statist, 27(1), 143-158.

Contraction rate theorems:

- Ghosal, S., Ghosh, J. K., and Van Der Vaart, A. W. (2000). Convergence rates of posterior distributions. Annals of Statistics, 28(2), 500-531.
- Shen, X., and Wasserman, L. (2001). Rates of convergence of posterior distributions. Annals of Statistics, 687-714.