Asymptotic aspects of non-parametric Bayesian Statistics

Introduction

Bas Kleijn, University of Amsterdam

14th Meeting of AiO's in Stochastics, Hilversum, 8-10 May 2006

Bayesian philosophy

Bayesian school of statistics differs from the Frequentist school.

Bayesians have a different perspective on data and models.
In particular, no 'true, underlying distribution of the data'.

Bayesians have a 'belief' concerning the mechanisms generating the data. The data itself is used to correct this belief.

Mathematically
The *belief* is represented by a prior measure on the model.
The *data* is incorporated by conditioning, resulting in a posterior measure on the model.

Frequentist analysis

We shall analyse the Bayesian procedure from a frequentist perspective.

Assumption sample $X_1, \ldots, X_n$ i.i.d. P_0-distributed.

Choose a prior Π on the model $\mathscr{P}$; calculate the posterior

We shall study the large-sample behaviour of the posterior, e.g.

1. Consistency,
2. Rate of convergence,
3. Limiting shape.

in the limit $n \to \infty$.

Goal

Only freedom: choice of model and prior measure

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 the prior and the model.

Course schedule

Lecture 1 Introduction

Setting the stage: bayesian and asymptotic statistics

Lecture 2 Posterior consistency

Doob's and Schwartz' consistency theorems, test-sequences

Lecture 3 Posterior rate of convergence

The Ghosh-Ghosal-van der Vaart-theorem

Lecture 4 Posterior limiting shape

The Bernstein-Von Mises theorem

Lecture 5 Research topics

Misspecification, semiparametric Bernstein-Von Mises

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).

Asymptotic aspects of non-parametric Bayesian Statistics

Lecture 1 Bayes and the Infinite

Bas Kleijn, University of Amsterdam

14th Meeting of AiO's in Stochastics, Hilversum, 8-10 May 2006

Frequentist estimation

Choose Model $\{P_\theta : \theta \in \Theta\}$

Assume Observation $Y \in \mathcal{Y}$ is random variable, and

$$Y \sim P_{\theta_0}.$$

for some, unknown $\theta_0 \in \Theta$.

Procedure point-estimator $\hat{\theta}(Y)$

Goal Choose $\widehat{Y}$ such that it is 'close to' θ_0 with high P_{θ_0}-probability.

Bayesian estimation: prior

Assume Observation Y and parameter $\bar{\theta}$ are random variables; joint distribution Π on $\mathscr{Y} \times \Theta$.

$$(Y, \bar{\theta}) \sim \Pi$$

Choose The model arises as

$$P_\theta = \Pi_{Y|\bar{\theta}=\theta}$$

Choose The marginal for $\bar{\theta}$ (together with the conditionals $\{P_\theta : \theta \in \Theta\}$) specify Π completely.

Marginal Π_θ on Θ is called the prior. Model is probablity space

$$(\Theta, \mathscr{G}, \Pi_\theta)$$

Bayesian estimation: posterior

The other conditional distribution

$$\Pi_{\theta|Y}$$

is called the posterior. Model becomes new probability space

$$(\Theta, \mathscr{G}, \Pi_{\theta|Y})$$

Bayes' Rule Posterior in terms of P_θ and Π_θ

$$\Pi_{\theta|Y}(G|B) = \frac{\Pi_{Y|\theta}(B|G)\Pi_\theta(G)}{\Pi_Y(B)} = \frac{\Pi_{Y|\theta}(B|G)\Pi_\theta(G)}{\Pi_{Y|\theta}(B|\Theta)\Pi_\theta(\Theta)}$$

Bayesian estimation: posterior density

Bayes' Rule for densities

$$d\Pi_{\theta|Y}(\theta|Y) = \frac{\pi_{Y|\theta}(Y|\theta)\, d\Pi_\theta(\theta)}{\int_\Theta \pi_{Y|\theta}(Y|\theta)\, d\Pi_\theta(\theta)}$$

If $Y = (X_1, \ldots, X_n)$ exchangeable then

$$d\Pi_{\theta|X}(\theta|X_1, \ldots, X_n) = \frac{\prod_{i=1}^{n} p_\theta(X_i)\, d\Pi(\theta)}{\int_\Theta \prod_{i=1}^{n} p_\theta(X_i)\, d\Pi(\theta)}$$

'True' distribution?

Nowhere $Y \sim P_{\theta_0}$ for some $\theta_0 \in \Theta$.

Closest to this role: marginal of $X_1, \ldots, X_n$.

$$d\Pi_X(x_1, \ldots, x_n) = \int_\Theta \prod_{i=1}^n p_\theta(x_i)\, d\Pi(\theta).$$

called prior predictive distribution. De Finetti's theorem

Hence, True Bayesians don't accept the frequentist notion of an 'underlying, true distribution of the data'.

To the pure Bayesian, this fact invalidates almost all questions concerning asymptotic behaviour of the posterior!

See, however, Blackwell and Dubins (1962), concerning the merging of *posterior predictive distributions*.

Hybrid perspective (I)

The Bayesian procedure can be considered from a frequentist perspective.

Choose Model with a prior Π

$$(\Theta, \mathscr{G}, \Pi)$$

Assume observation $Y \in \mathscr{Y}$ is distributed

$$Y \sim P_{\theta_0}.$$

for some, unknown $\theta_0 \in \Theta$.

Hybrid perspective (II)

Procedure *Define* the posterior

$$d\Pi_{\theta|X}(\theta|X_1,\ldots,X_n) = \frac{\prod_{i=1}^{n} p_\theta(X_i)\, d\Pi(\theta)}{\int_\Theta \prod_{i=1}^{n} p_\theta(X_i)\, d\Pi(\theta)}$$

Study the random posterior measure on Θ under the assumption:

$$(X_1,\ldots,X_n) \sim P_{\theta_0}^n.$$

Alternative explanation: we develop the point of view that the likelihood is an (unnormalized) density with respect to a measure (the prior).

Bayesian point estimators

The posterior mean

$$\widehat{\theta}_n(X) = \int_\Theta \theta \, d\Pi(\theta|X_1, \ldots, X_n).$$

If posterior dominated by μ on Θ, the posterior mode (or MAP-estimator)

$$\widehat{\theta}_n(X) = \arg\max_{\theta\in\Theta} \pi(\theta|X_1, \ldots, X_n)$$

If the model is metric space, $B(\theta, \epsilon) = \{\theta' : d(\theta', \theta) < \epsilon\}$,

$$\widehat{\theta}_n(X) = \arg\max_{\theta\in\Theta} \Pi(B(\theta, \epsilon)|X_1, \ldots, X_n)$$

Use of a loss-function ℓ leads to minimization of $\int \ell(\theta, \theta') \, d\Pi(\theta'|X)$, formal Bayes estimators (Le Cam (1986))

Frequentist consistency

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

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

An estimator is said to be consistent if

$$\|\widehat{\theta}_n - \theta_0\| \xrightarrow{P_{\theta_0}} 0.$$

So for a consistent estimator, the $P_{\theta_0}^n$-probability of finding $\widehat{\theta}_n$ at any distance $\epsilon > 0$ (or more) from θ_0 becomes arbitrarily small, if we make the sample large enough.

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

Frequentist rate of convergence

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

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

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

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

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

Frequentist limit distribution

Finally, suppose that $\widehat{\theta}_n$ converges to θ_0 at rate r_n. According to Prohorov's lemma: weakly convergent subsequence!

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

$$r_n^{-1}(\widehat{\theta}_n - \theta_0) \overset{P_{\theta_0}}{\rightsquigarrow} L_{\theta_0},$$

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}.

Typical example

Suppose that T is such that $P_\theta T(X) = \theta$, for all $\theta \in \Theta$.

Law of large numbers

$$\widehat{\theta}_n(X) = \mathbb{P}_n T \overset{P_{\theta_0}-a.s.}{\longrightarrow} P_{\theta_0} T = \theta_0.$$

So $\widehat{\theta}_n$ is consistent for estimation of θ_0.

Assume $P_{\theta_0} T(X)^2 < \infty$, the central limit theorem

$$\mathbb{G}_n T = \sqrt{n}(\mathbb{P}_n - P_{\theta_0})T \overset{P_{\theta_0}}{\rightsquigarrow} N(0, \mathrm{Var}_{\theta_0} T)$$

$\widehat{\theta}_n$ converges to θ_0 at rate $n^{-1/2}$, with limit distribution $N(0, \mathrm{Var}_{\theta_0} T)$.

Questions of asymptotic optimality concern (identifiability), minimax rates-of-convergence and minimal-variance limit distributions.

Bayesian asymptotics

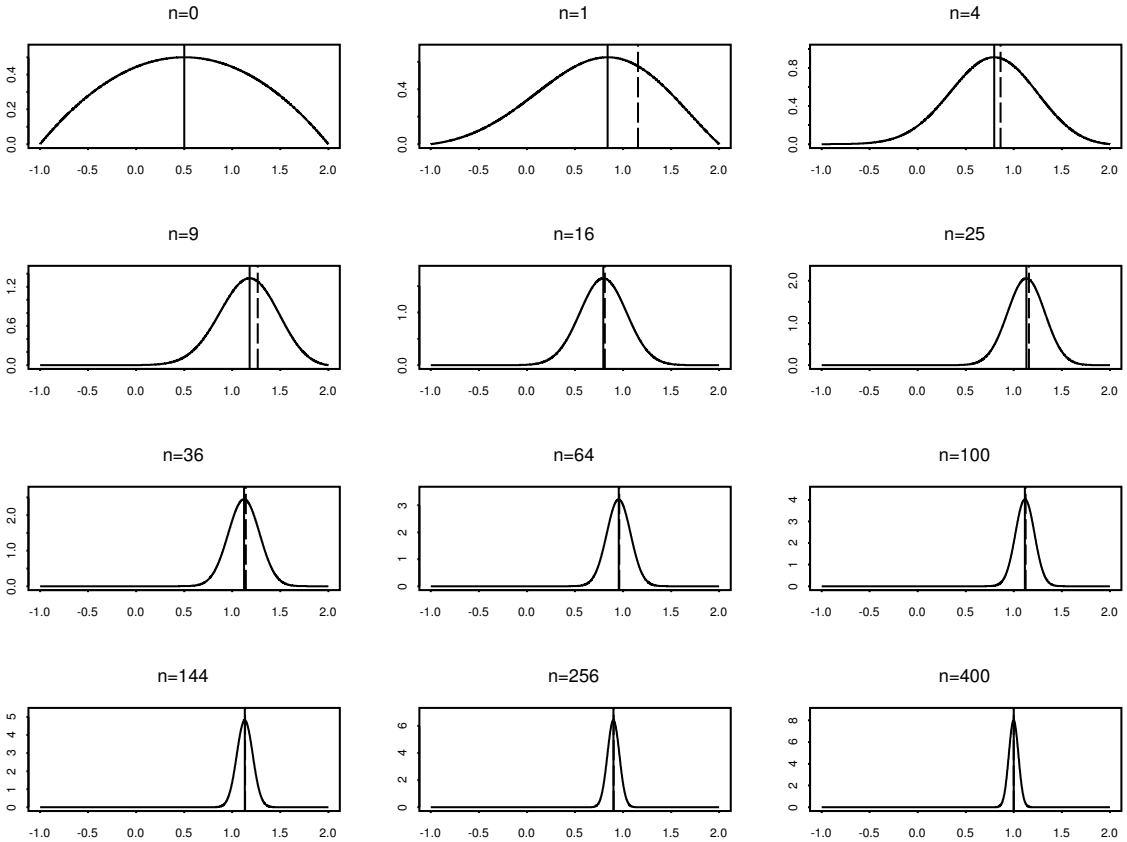
What signifies consistency for the posterior?
What prior measures give rise to consistent posteriors?

Can we determine rates for the posterior convergence?
How does this rate relate to minimax-rates? What does it mean for the prior?

Is there something like a posterior limitshape?
What is it centered on? What is the shape? What conditions for model and prior?

We shall answer the first two questions in great generality (for non-parametric models). The third question is harder. We answer it for smooth parametric models.

Simulation



Posterior density for growing n on a normal location model

Asymptotic aspects of non-parametric Bayesian Statistics

Lecture 2 Posterior consistency

Bas Kleijn, University of Amsterdam

14th Meeting of AiO's in Stochastics, Hilversum, 8-10 May 2006

Setting the stage

Non-parametric model

Discard with the parametrization $\theta \mapsto P_\theta$. The model $\mathscr{P}$ contains probability measures P and $\mathscr{P}$ is a probability space with prior Π.

Domination

For notational convenience , we assume that all $P \in \mathscr{P}$ are dominated by a σ-finite measure μ: $p = dP/d\mu$.

Metric

Furthermore, we assume that $\mathscr{P}$ is a metric space (metric d). The corresponding Borel σ-algebra is in the domain of Π.

Asymptotics

The sample is $X_1, \ldots, X_n$, i.i.d.-P_0 distributed, for some $P_0 \in \mathscr{P}$. We study the large-sample limit $n \to \infty$.

Formulations of consistency

Model $\mathscr{P}$ with topology $\mathscr{T}$ and prior Π on the Borel σ-algebra $\mathscr{B}_{\mathscr{T}}$.

Posterior is consistent if for every open neighbourhood $U \in \mathscr{T}$ of P_0

$$\Pi_n(\, U \,|\, X_1, X_2, \ldots, X_n \,) \overset{P_0-a.s.}{\longrightarrow} 1. \tag{1}$$

For a metric model $\mathscr{P}$ this is equivalent to (for every $\epsilon > 0$)

$$\Pi_n(\, d(P, P_0) \geq \epsilon \,|\, X_1, X_2, \ldots, X_n \,) \overset{P_0-a.s.}{\longrightarrow} 0, \tag{2}$$

and equivalent to

$$\Pi_n(\, \cdot \,|\, X_1, X_2, \ldots, X_n \,) \rightsquigarrow \delta_{P_0}, \qquad (P_0 - a.s.). \tag{3}$$

Proof of equivalence (I)

Assume (1) holds. Let U_k denote a decreasing sequence of open balls around P_0. Define for every $k \geq 1$ the set Ω_k such that $P_0^\infty(\Omega_k) = 1$ and the limit in (1) with $U = U_k$ holds on Ω_k. Note that $\Omega' = \cap_{k \geq 1} \Omega_k$ satisfies $P_0^\infty(\Omega') = 1$ and for all $\omega \in \Omega'$ and all $k \geq 1$:

$$\Pi_n\Big(U_k \,\Big|\, X_1(\omega), X_2(\omega), \ldots, X_n(\omega) \Big) \to 1, \qquad (n \to \infty).$$

Fix $\omega \in \Omega'$, let the open neighbourhood U of P_0 be given. Then U contains U_l for certain $l \geq 1$ and hence:

$$\Pi_n\Big(U \,\Big|\, X_1(\omega), X_2(\omega), \ldots, X_n(\omega) \Big) \geq \Pi_n\Big(U_l \,\Big|\, X_1(\omega), X_2(\omega), \ldots, X_n(\omega) \Big)$$

as $n \to \infty$. So (1) does not only hold P_0-almost-surely for each U separately, but P_0-almost-surely for all U simultaneously.

Proof of equivalence (II)

Let $f : \mathscr{P} \to \mathbb{R}$ be bounded ($|f| \leq M$) and continuous. Let $\eta > 0$ be given. Neighbourhood U of P_0 such that $|f(P) - f(P_0)| \leq \eta$ for all $P \in U$. Integrate f with respect to the posterior and to δ_{P_0}:

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

Portmanteau: $\Pi_n(\,\cdot\,|\,X_1, X_2, \ldots, X_n\,) \rightsquigarrow \delta_{P_0}$, P_0-almost-surely.

Proof of equivalence (III)

Conversely, assume $\Pi_n(\,\cdot\,|\,X_1, X_2, \ldots, X_n\,) \rightsquigarrow \delta_{P_0}$, P_0-a.s. Let U be an open neighbourhood of P_0. Based on the metric, we can construct 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 $\mathscr{P} \setminus U$.

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

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

P_0-almost-surely.

Bayesian point-estimators

Point-estimators derived from a consistent Bayesian procedure are consistent themselves under some mild conditions.

Theorem 28.1. *Suppose that the metric d is the total variation norm $\|.\|$. Assume that the posterior is consistent. Then the posterior mean $\hat{P}_n$ is a P_0-almost-surely consistent point-estimator (with respect to total-variation).*

Proof of point-estimator's consistency

Extend $P \mapsto \|P - P_0\|$ to the convex hull of $\mathscr{P}$. Since $P \mapsto \|P - P_0\|$ is convex by the triangle inequality, we apply Jensen

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

Since $P \overset{\Pi_n}{\rightsquigarrow} P_0$ 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 law δ_{P_0}, which equals zero. Hence

$$\hat{P}_n \to P_0, \qquad (P_0 - a.s.).$$

in total variation.

Remarks on Bayesian point-estimators

The above argument works for any convex metric d.

Similar arguments demonstrate consistency for other classes of point estimators derived from a consistent sequence of posterior distributions, e.g. Le Cam's *formal Bayes estimators*.

The notion of a point-estimator is not an entirely natural extension to the Bayesian framework: for example, if the model is non-convex, the expectation based on the posterior measure may lie outside the model. Similarly, perfectly well-defined posteriors may lead to ill-defined point-estimators due to integrability issues or non-existence of maximisers, which become more severe as the model becomes more complicated.

Doob's consistency theorem

Theorem 31.1. *(Doob (1948)) Suppose that both the model Θ and the sample space $\mathscr{X}$ are Polish spaces endowed with their respective Borel-σ-algebras. Assume that the map $\theta \mapsto P_\theta$ is one-to-one. Then the sequence of posteriors is consistent Π-almost-surely.*

Proof. Proof The proof of this theorem is an application of Doob's martingale convergence theorem (see e.g. Van der Vaart (1998) or Ghosh and Ramamoorthi (2003)). □

A Polish space is a complete, separable, metric space. Often needed to guarantee measurability.

The only real condition here is identifiability of θ.

Is Doob's theorem enough?

For many Bayesians, Doob's theorem is more than enough: parametric model Θ with a prior Π that dominates Lebesgue measure on Θ: inconsistency only on subsets of Lebesgue measure zero.

"the data overrides prior beliefs asymptotically".

But!

parametric objection: misspecification

Consistency only if the true distribution was not excluded from consideration in the first place by an ill-chosen prior or model. If the models does not contain the true distribution, inconsistency is guaranteed.

Freedman's point

non-parametric objection: sparsity of prior mass

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

Non-parametric counterexamples

Freedman (1963,1965), Diaconis and Freedman (1986), Cox (1993), Diaconis and Freedman (1998). Basically what is shown is that Doob's null-set of inconsistency can be rather large.

Some authors are tempted to present the above as definitive proof of the fact that Bayesian statistics is useless in non-parametric estimation problems. More precise would be the statement that not every choice of prior is suitable and some may lead to unforeseen instances of inconsistency.

Schwartz' consistency theorem

Theorem 34.1. *(Schwartz (1965)) Assume that*

(i) For every $\eta > 0$,

$$\Pi\Big(P \in \mathscr{P} \,:\, -P_0 \log \frac{p}{p_0} \leq \eta \Big) > 0, \tag{4}$$

(ii) For every $\epsilon > 0$, *there exists a sequence* (ϕ_n) *of test-functions such that:*

$$P_0^n \phi_n \to 0, \qquad \sup_{\{P : d(P,P_0) > \epsilon\}} P^n(1 - \phi_n) \to 0. \tag{5}$$

Then

$$\Pi_n\Big(d(P, P_0) \geq \epsilon \,|\, X_1, X_2, \ldots, X_n \Big) \overset{P_0 - a.s.}{\longrightarrow} 0, \tag{6}$$

for all $\epsilon > 0$.

The conditions in Schwartz' theorem

Schwartz' first condition says that all Kullback-Leibler neighbourhoods of P_0 should receive sufficient prior mass. But P_0 is unknown, so to guarantee this, we have to prove the first condition for all KL-neighbourhoods of all $P \in \mathscr{P}$. In a sense, this requires uniformity of the prior.

The second condition requires uniform testability of $\{P_0\}$ versus the complements of d-balls around P_0. One can view this in various ways. For instance, identifiability of the model in a statistical sense. Another explanation: the sequence (ϕ_n) separates $\{P_0\}$ from the complements of d-balls around P_0, in an asymptotic way.

Technically, the reasons become clear on the next few slides.

Analogous conditions will arise in the rate-of-convergence theorem.

Proof of Schwartz (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 and take the lim sup:

$$\begin{aligned} \limsup_{n\to\infty} \Pi_n(V|X_1, \ldots, X_n) \leq & \limsup_{n\to\infty} \Pi_n(V|X_1, \ldots, X_n)(1 - \phi_n) \\ & + \limsup_{n\to\infty} \Pi_n(V|X_1, \ldots, X_n)\phi_n. \end{aligned} \tag{7}$$

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

$$\left| \mathbb{P}_n \log \frac{p}{p_0} - P_0 \log \frac{p}{p_0} \right| \to 0, \quad (P_0 - a.s.).$$

Proof of Schwartz (II)

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

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

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 (III)

The first term in (7) can be bounded as follows

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

$$\leq \frac{\limsup_{n\to\infty} e^{n\alpha} \int_V \prod_{i=1}^n (p/p_0)(X_i)\,(1-\phi_n)(X_1,\ldots,X_n)\,d\Pi(P)}{\liminf_{n\to\infty} e^{n\alpha} \int_{\mathscr{P}} \prod_{i=1}^n (p/p_0)(X_i)\,d\Pi(P)} \tag{8}$$

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

where we use the (non-negative)

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

Exponential testing power

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 39.1. *Suppose that for given $\epsilon > 0$ there exists a sequence of tests $(\phi_n)_{n\geq 1}$ such that:*

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

where $V_\epsilon = \{P \in \mathscr{P} : d(P, P_0) \geq \epsilon\}$. Then there exists a sequence of tests $(\omega_n)_{n\geq 1}$ and positive constants C, D such that:

$$P_0^n \omega_n \leq e^{-nC}, \qquad \sup_{P \in V_\epsilon} P^n(1-\omega_n) \leq e^{-nD} \tag{9}$$

Proof of Schwartz (IV)

Lemma 39.1) guarantees that there exists a constant $\beta > 0$ such that for large enough n,

$$P_0^\infty f_n = P_0^n f_n = e^{n\alpha} \int_V P_0^n \Big(\prod_{i=1}^n \frac{p}{p_0}(X_i)\,(1-\phi_n)(X_1,\ldots,X_n) \Big)\, d\Pi(P)$$
$$\leq e^{n\alpha} \int_V P^n(1-\phi_n)\, d\Pi(P) \leq e^{-n(\beta-\alpha)}. \tag{10}$$

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 (V)

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

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

So $f_n \to 0$, $(P_0 - a.s.)$ and hence

$$\Pi_n(V|X_1, \ldots, X_n)\,(1 - \phi_n)(X_1, \ldots, X_n) \overset{P_0-a.s.}{\longrightarrow} 0.$$

The other term in (7) is treated similarly: $P_0^n \Pi(V|X_1, \ldots, X_n)\phi_n \leq P_0^n \phi_n \leq e^{-nC}$ (lemma 39.1); use Markov's inequality and the first Borel-Cantelli lemma again to show that:

$$\Pi(V|X_1, \ldots, X_n)\,\phi_n(X_1, \ldots, X_n) \overset{P_0-a.s.}{\longrightarrow} 0. \qquad (11)$$

Combination of (8) and (11) proves that (7) equals zero.

Asymptotic aspects of non-parametric Bayesian Statistics

Lecture 3 Posterior rate

Bas Kleijn, University of Amsterdam

14th Meeting of AiO's in Stochastics, Hilversum, 8-10 May 2006

Setting the stage

Non-parametric model

Again, let $\mathscr{P}$ be a model with metric d, prior Π, where it is assumed that the σ-algebra on the model contains the Borel σ-algebra.

Frequentist sample

Assume that $X_1, X_2, \ldots$ is an infinite *i.i.d.* sample from an unknown distribution $P_0 \in \mathscr{P}$.

KL-neighbourhoods

We shall need a particular variant of the Kullback-Leibler neighbourhood used in Schwartz' theorem (theorem 34.1): for every $\epsilon > 0$

$$B(\epsilon) = \left\{ P \in \mathscr{P} : -P_0 \log \frac{p}{p_0} \leq \epsilon^2, \, P_0 \left(\log \frac{p}{p_0} \right)^2 \leq \epsilon^2 \right\}. \qquad (12)$$

Definition of posterior rate

Conceptually

Define the posterior rate of convergence as the fastest rate (ϵ_n) at which we can let balls $D_d(P_0, \epsilon_n) = \{P \in \mathscr{P} : d(P, P_0) < M\epsilon_n\}$ shrink to radius zero, while still capturing posterior mass (arbitrarily close to) one in the limit $n \to \infty$.

Mathematically

Let (ϵ_n) be such that $\epsilon_n > 0$, $\epsilon_n \downarrow 0$. We say that the posterior $\Pi(\cdot \,|\, X_1, X_2, \ldots, X_n)$ converges to P_0 at rate ϵ_n, if for some $M > 0$:

$$\Pi_n\Big(d(P, P_0) \geq M\epsilon_n \;\Big|\; X_1, X_2, \ldots, X_n \Big) \xrightarrow{P_0} 0, \tag{13}$$

Rates for Bayesian point-estimators

Point-estimators derived from a posterior that converges at rate ϵ_n converge at rate ϵ_n themselves under some mild conditions.

Assume that posterior satisfies (13) with rate (ϵ_n) and constant $M > 0$. Define point estimators $\tilde{P}_n(X)$ as (near-)maximisers of

$$P \mapsto \Pi_n\Big(D_d\Big(P, M\epsilon_n\Big) \;\Big|\; X_1, \ldots, X_n \Big),$$

Lemma 45.1. *For some $M > 0$, the estimator sequence $\tilde{P}_n$ satisfies*

$$P_0^n\Big(\epsilon_n^{-1}\, d(\tilde{P}_n, P_0) \leq 2M \Big) \to 1 \tag{14}$$

As a result, ϵ_n is an upper bound for the rate at which $\tilde{P}_n$ converges to P_0 with respect to d, *i.e.* $\epsilon_n^{-1} d(\tilde{P}_n, P_0) = O_{P_0}(1)$.

Proof of rate for point-estimators

By definition of a near-maximiser:

$$\Pi_n(\, B(\tilde{P}_n, M\epsilon_n) \,|\, X_1, \ldots, X_n \,)$$

$$\geq \sup_{P \in \mathscr{P}} \Pi(\, B(P, M\epsilon_n) \,|\, X_1, \ldots, X_n \,) - o_{P_0}(1)$$

$$\geq \Pi(\, B(P_0, M\epsilon_n) \,|\, X_1, \ldots, X_n \,) - o_{P_0}(1).$$

First term on the *r.h.s.* converges to one by assumption, so the *l.h.s.* converges to one in P_0-probability. If $d(\tilde{P}_n, P_0) > 2M\epsilon_n$

$$B(\tilde{P}_n, M\epsilon_n) \cap B(P_0, M\epsilon_n) = \emptyset$$

The total posterior mass does not exceed one, so $d(\tilde{P}_n, P_0) \leq 2M\epsilon_n$ with P_0-probability growing to one.

Optimal rates

The possibility to construct point-estimators from posteriors converging at the same rate, implies that limitations on the rate of convergence derived for point estimators, apply also to Bayesian rates.

This argument applies to other asymptotic optimality criteria as well.

In particular, minimax rates are optimal in many problems, e.g. in density estimation.

Hellinger metric

$$H(P,Q) = \left(\int \left(p^{1/2} - q^{1/2} \right)^2 d\mu \right)^{1/2}.$$

Ghosal-Ghosh-van der Vaart theorem

Theorem 48.1. *(Ghosal-Ghosh-van der Vaart (2000))*
Suppose that for (ϵ_n) *such that* $\epsilon_n > 0$, $\epsilon_n \downarrow 0$ *and* $n\epsilon_n^2 \to \infty$, *two conditions hold:*

(i) There exists a constant $C > 0$ *such that:*

$$\Pi\Big(B(\epsilon_n)\Big) \geq e^{-nC\epsilon_n^2}. \tag{15}$$

(ii) There exists a sequence ϕ_n *of test-functions* ϕ_n *and a constant* $L > 0$ *such that:*

$$P_0^n \phi_n \to 0, \qquad \sup_{P: d(P,P_0) \geq \epsilon_n} P^n(1-\phi_n) \leq e^{-nL\epsilon_n^2}. \tag{16}$$

Then for a sufficiently large $M > 0$,

$$P_0^n \Pi_n(\, d(P, P_0) \geq M\epsilon_n \,|\, X_1, \ldots, X_n \,) \to 0. \tag{17}$$

Proof of GGV theorem (I)

Let $\eta > 0$; define $A(\eta) = \{P \in \mathscr{P} : d(P, P_0) \geq \eta\}$. The expectation in (17) is decomposed using the tests ϕ_n; for every $n \geq 1$ and every $M > 1$, we have:

$$\begin{aligned} & P_0^n \Pi_n\Big(A(M\epsilon_n) \,\Big|\, X_1, \ldots, X_n \Big) \\ & \quad = P_0^n \, \Pi_n\Big(A(M\epsilon_n) \,\Big|\, X_1, \ldots, X_n \Big) \, \phi_n(X) \\ & \qquad + P_0^n \, \Pi_n\Big(A(M\epsilon_n) \,\Big|\, X_1, \ldots, X_n \Big) (1 - \phi_n)(X). \end{aligned}$$

We estimate the terms on the right-hand side separately. Due to the first inequality in (16):

$$P_0^n \, \Pi_n\Big(A(M\epsilon_n) \,\Big|\, X_1, \ldots, X_n \Big) \, \phi_n(X) \leq P_0^n \, \phi_n(X) \to 0,$$

the first term converges to zero.

Proof of GGV theorem (II)

To estimate the second term, we use the definition of the posterior to obtain:

$$P_0^n \Pi_n\Big(A(M\epsilon_n) \,\Big|\, X_1, \ldots, X_n \Big)(1-\phi_n)(X)$$
$$= P_0^n\Big[\int_{A(M\epsilon_n)} \prod_{i=1}^n \frac{p}{p_0}(X_i)\,(1-\phi_n)\,d\Pi(P) \Big/ \int_{\mathscr{P}} \prod_{i=1}^n \frac{p}{p_0}(X_i)\,d\Pi(P) \Big] \tag{18}$$

First we concentrate on the denominator, using assumption (15).

Proof of GGV theorem (III)

Lemma 51.1. *Let $\epsilon > 0$ be given. If $\Pi(B(\epsilon)) > 0$, then for every $K > 0$:*

$$P_0^n\left(\int_{B(\epsilon)} \prod_{i=1}^{n} \frac{p}{p_0}(X_i)\, d\Pi(P) \leq e^{-n\epsilon^2(1+K)}\Pi(B(\epsilon))\right) \leq \frac{1}{nK^2\epsilon^2}. \quad (19)$$

Proof. Proof The proof of this lemma can be found as lemma 8.1 in Ghosal-Ghosh-van der Vaart (2000). □

Recall

$$B(\epsilon) = \left\{ P \in \mathscr{P} : -P_0 \log \frac{p}{p_0} \leq \epsilon^2,\ P_0\left(\log \frac{p}{p_0}\right)^2 \leq \epsilon^2 \right\}.$$

Proof of GGV theorem (III)

Let Ω_n be the subset of $\mathscr{X}^n$ for which the inequality between left- and right-hand sides in the following display holds:

$$\int_{\mathscr{P}} \prod_{i=1}^{n} \frac{p}{p_0}(X_i)\, d\Pi(P) \geq \int_{B(\epsilon_n)} \prod_{i=1}^{n} \frac{p}{p_0}(X_i)\, d\Pi(P) \geq e^{-(1+K)n\epsilon_n^2} \Pi(B(\epsilon_n)), \tag{20}$$

(with $K > 0$ as yet unspecified). Decompose the P_0^n-expectation in (18) into separate integrals over Ω_n and $\mathscr{X}^n \setminus \Omega_n$

$$\begin{aligned} & P_0^n \Pi_n\Big(A(M\epsilon_n) \,\Big|\, X_1, \ldots, X_n \Big)(1-\phi_n) \\ & \qquad \leq P_0^n \Pi_n\Big(A(M\epsilon_n) \,\Big|\, X_1, \ldots, X_n \Big)(1-\phi_n) 1_{\Omega_n} + P_0^n(\mathscr{X}^n \setminus \Omega_n). \end{aligned}$$

Now $P_0^n(\mathscr{X}^n \setminus \Omega_n) = o(1)$ as $n \to \infty$ according to (19).

Proof of GGV theorem (IV)

The first term is estimated as follows:

$$
\begin{aligned}
& P_0^n \Pi_n\Big(A(M\epsilon_n) \,\Big|\, X_1, \ldots, X_n \Big)(1-\phi_n)(X)\, 1_{\Omega_n} \\
& \quad \le \frac{e^{(1+K)n\epsilon_n^2}}{\Pi(B(\epsilon_n))} P_0^n \Big(\int_{A(M\epsilon_n)} \prod_{i=1}^n \frac{p}{p_0}(X_i)\,(1-\phi_n)(X)\, d\Pi(P) \Big) \\
& \quad \le \frac{e^{(1+K)n\epsilon_n^2}}{\Pi(B(\epsilon_n))} \int_{A(M\epsilon_n)} P^n(1-\phi_n)(X)\, d\Pi(P) \\
& \quad \le e^{(1+K)n\epsilon_n^2} \frac{\Pi(A(M\epsilon_n))}{\Pi(B(\epsilon_n))} \sup_{P \in A(M\epsilon_n)} P^n(1-\phi_n),
\end{aligned}
\tag{21}
$$

where we have substituted (20) and used the positivity of the integrand, applied Fubini's theorem and bounded the integrand by its supremum over $A(M\epsilon_n)$.

Proof of GGV theorem (V)

Application of the second inequality in (16) gives:

$$P_0^n \Pi_n\Big(A(M\epsilon_n) \,\Big|\, X_1, \ldots, X_n \Big)(1-\phi_n) \leq e^{(1+K+C-M^2L)n\epsilon_n^2} + o(1).$$

Hence, for all $K > 0$ there exists a constant $M > 0$ such that the above expression converges to zero. This leads us to conclude that:

$$P_0^n \, \Pi_n\Big(A(M\epsilon_n) \,\Big|\, X_1, \ldots, X_n \Big) \to 0, \qquad (n \to \infty).$$

for sufficiently large $M > 0$.

Testing and metric entropy and minimax rates

Condition existence of test-sequences (ϕ_n) such that

$$P_0^n \phi_n \to 0, \qquad \sup_{P: d(P,P_0) \geq \epsilon_n} P^n(1-\phi_n) \leq e^{-n\epsilon_n^2}.$$

If $d = H$, sufficient condition on packing numbers

$$D(\epsilon_n, \mathscr{P}, H) \leq e^{n\epsilon_n^2}. \tag{22}$$

(See Ghosal, Ghosh, van der Vaart (2000), Birgé (1983,1984) and Le Cam (1986)

The minimal ϵ_n satisfying $\log D(\epsilon_n, \mathscr{P}, H) \leq n\epsilon_n^2$ is (roughly) the fastest Hellinger-rate obtainable by any method of point-estimation. the so-called minimax-rate, Birgé (1983).

Covering and packing numbers

The packing number $D(\eta, \mathscr{X}, \rho)$ of a metric space $(\mathscr{X}, \rho)$ is defined as the maximal number of points in $\mathscr{X}$ such that the ρ-distance between all pairs is at least η.

This number is related to the so-called covering number $N(\eta, \mathscr{X}, \rho)$ which is defined as the minimal number of ρ-balls of radius η needed to cover $\mathscr{X}$.

For many models entropy numbers are well-known or can be calculated (Kolmogorov, Tikhomirov (1961), van der Vaart and Wellner (1996)).

Lemma 56.1. *For any metric space* $(\mathscr{X}, \rho)$ *and all* $\epsilon > 0$

$$N(\epsilon, \mathscr{X}, \rho) \leq D(\epsilon, \mathscr{X}, \rho) \leq N(\epsilon/2, \mathscr{X}, \rho)$$

Proof of metric entropy inequalities

Let $\epsilon > 0$ be given. By definition of $D(\epsilon, \mathscr{X}, \rho) = D$, there exists a maximal ϵ-separated set $\{x_1, \ldots, x_D\} \subset \mathscr{X}$, This means that for all $x \in \mathscr{X}$,

$$\min\{d(x, x_i) : 1 \leq i \leq D\} < \epsilon.$$

So there exists an i such that $d(x, x_i) < \epsilon$, that is, $x \in D_\rho(x_i, \epsilon)$. Hence,

$$\{ D_\rho(x_i, \epsilon) : 1 \leq i \leq D\},$$

forms a cover of $\mathscr{X}$. Since $N(\epsilon, \mathscr{X}, \rho)$ is the minimal number of balls $D_\rho(\,\cdot\,, \epsilon)$ needed to cover $\mathscr{X}$,

$$N(\epsilon, \mathscr{X}, \rho) \leq D(\epsilon, \mathscr{X}, \rho)$$

The other inequality follows from similar arguments.

Metric entropy for compact $\Theta \subset \mathscr{X}$

Lemma 58.1. *For any compact K subset of a metric space $(\mathscr{X}, d)$, for all $\epsilon > 0$*

$$\log N(\epsilon, \mathscr{X}, d) < \infty \tag{23}$$

Proof. Proof Let $\epsilon > 0$ be given. Cover K by the collection of open d-balls

$$\{D_d(x, \epsilon) : x \in K\}.$$

Since K is compact, any open cover has a finite subcover, *i.e.* there exists a finite set of points $\{x_1, \ldots, x_N\} \subset K$ such that $\{D_d(x_i, \epsilon) : 1 \leq i \leq N\}$ covers K. Since $N(\epsilon, K, \rho)$ is the minimal number of balls needed to cover K, $\log N(\epsilon, K, \rho) \leq \log N < \infty$. □

This is important for the existence of consistency tests: a test-sequence (ϕ_n) as needed in Schwartz' theorem if (23) holds.

Metric entropy for compact $\Theta \subset \mathbb{R}^d$ (I)

Lemma 59.1. *For any compact Θ subset of $\mathbb{R}^d$, there exists a constant $M > 0$ such that for small enough $\epsilon > 0$*

$$N(\epsilon, \Theta, \|\cdot\|) < \left(\frac{M}{\epsilon}\right)^d$$

Proof Θ is compact, so Θ is bounded, *i.e.* there exists a constant $M' > 0$ such that $\Theta \subset B(0, M')$. Hence,

$$N(\epsilon, \Theta, \|\cdot\|) \leq N(\epsilon, B(0, M'), \|\cdot\|).$$

Let $\epsilon > 0$ be given. Due to lemma 56.1, we see

$$N(\epsilon, B(0, M'), \|\cdot\|) \leq D(\epsilon, B(0, M'), \|\cdot\|) = D$$

Let $\theta_1, \ldots, \theta_D\}$ be maximal ϵ-separated in Θ. This implies that:

$$i \neq j \quad \Rightarrow \quad B(\theta_i, \tfrac{1}{2}\epsilon) \cap B(\theta_j, \tfrac{1}{2}\epsilon) = \varnothing.$$

Metric entropy for compact $\Theta \subset \mathbb{R}^d$ (II)

Moreover,

$$\bigcup_{1 \le i \le D} B(\theta_i, \tfrac{1}{2}\epsilon) \subset B(0, M' + \tfrac{1}{2}\epsilon)$$

We compare the Lebesgue-measures

$$\mu\Big(\bigcup_{1 \le i \le D} B(\theta_i, \tfrac{1}{2}\epsilon) \Big) \le \mu\Big(B(0, M' + \tfrac{1}{2}\epsilon) \Big)$$

Denoting the Lebesgue-measure of a d-dimensional ball of radius r by $V_d\, r^d$, we arrive at:

$$\mu\Big(\bigcup_{1 \le i \le D} B(\theta_i, \tfrac{1}{2}\epsilon) \Big) = \sum_{i=1}^{D} V_d \big(\tfrac{1}{2}\epsilon\big)^d = D\, V_d \big(\tfrac{1}{2}\epsilon\big)^d \le V_d \big(M' + \tfrac{1}{2}\epsilon\big)^d .$$

Hence,

$$D(\epsilon, B(0, M'), \|\cdot\|) \le \left(\frac{2M' + \epsilon}{\epsilon}\right)^d \le \left(\frac{3M'}{\epsilon}\right)^d ,$$

for small enough ϵ.

Constructing tests (I)

Define minimax risk $\pi(P, \mathscr{Q})$ for testing P against convex $\mathscr{Q}$

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

Under (convexity, continuity, compactness) conditions, we can apply the minimax theorem (Strasser (1985), following Le Cam (>195?))

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

On the *r.h.s.* ϕ can be chosen Q-dependently; minimal for $\phi = 1\{p < q\}$

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

Constructing tests (II)

Note that:

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

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

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

Constructing tests (III)

But this works only for convex $\mathscr{Q}$. If we want to test against non-convex alternatives (like complements of H-balls), we cover by (convex) balls and combine the corresponding tests.

Lemma 63.1. *Given two test-sequences $(\omega_{1,n})$ and $(\omega_{2,n})$, such that $(i = 1, 2)$:*

$$P_0^n \omega_{i,n} \to 0, \qquad \sup_{P \in \mathscr{Q}_i} P^n(1 - \omega_{i,n}) \to 0,$$

then there exists a test-sequence (ψ_n) such that:

$$P_0^n \psi_n \to 0, \qquad \sup_{P \in \mathscr{Q}_1 \cup \mathscr{Q}_2} P^n(1 - \psi_n) \to 0,$$

A slightly stronger version of this lemma preserves (exponential) testing power.

Constructing tests (IV)

Proof. Proof Define $\psi_n = \omega_{1,n} \vee \omega_{2,n}$. Then

$$P_0^n \psi_n \leq P_0^n \omega_{1,n} + P_0^n \omega_{2,n} \to 0.$$

and

$$\begin{aligned} \sup_{P \in \mathcal{Q}_1 \cup \mathcal{Q}_2} P^n(1-\psi_n) &= \sup_{P \in \mathcal{Q}_1} P^n(1-\psi_n) \vee \sup_{P \in \mathcal{Q}_2} P^n(1-\psi_n) \\ &\leq \sup_{P \in \mathcal{Q}_1} P^n(1-\omega_{1,n}) \vee \sup_{P \in \mathcal{Q}_2} P^n(1-\omega_{2,n}) \to 0. \end{aligned}$$

□

From the proof, we see that we can combine N tests into one and control the power, if the (exponential) testing power is balanced against the (exponential) growth of N.

Asymptotic aspects of non-parametric Bayesian Statistics

Lecture 4 Posterior limitshape

Bas Kleijn, University of Amsterdam

14th Meeting of AiO's in Stochastics, Hilversum, 8-10 May 2006

Setting the stage

Parametric model

Θ open subset of $\mathbb{R}^d$; $\mathscr{P} = \{P_\theta : \theta \in \Theta\}$. Dominated by μ: $p_\theta = dP_\theta/d\mu$.

Sample

The sample $X_1, X_2, \ldots$ be distributed *i.i.d.*-P_0. $P_0 = P_{\theta_0}$, for some $\theta_0 \in \Theta$.

Continuous, positive prior

Prior Π with Lebesgue-density π, continuous on neighbourhood of θ_0 and $\pi(\theta_0) > 0$.

Localization

We 'localize' the model: centre on θ_0 and rescale by $\sqrt{n}$, $H = \sqrt{n}(\underline{\theta} - \theta_0) \in \mathbb{R}^d$.

The posterior for H is $\Pi_n(H \in B | X_1, \ldots, X_n) = \Pi_n(\sqrt{n}(\underline{\theta} - \theta_0) \in B | X_1, \ldots, X_n)$.

Differentiability conditions

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.

A model $\mathscr{P}$ is locally asymptotically normal (LAN) at θ_0, if for every random sequence $h_n = O_{P_0}(1)$,

$$\log \prod_{i=1}^n \frac{p_{\theta_0 + h_n/\sqrt{n}}}{p_{\theta_0}}(X_i) = h_n^T \mathbb{G}_n \dot{\ell}_{\theta_0} - \tfrac{1}{2} h_n^T I_{\theta_0} h_n + o_{P_0}(1) \tag{24}$$

Lemma 67.1. *The model $\mathscr{P}$ is differentiable in quadratic mean at θ_0 iff $\mathscr{P}$ is LAN at θ_0.*

The Bernstein-Von Mises theorem

Define

$$\Delta_{n,\theta_0}(X) = I_{\theta_0}^{-1}\mathbb{G}_n\dot{\ell}_{\theta_0}$$

Theorem 68.1. *Let $\mathscr{P}$ be differentiable in quadratic mean at θ_0 with non-singular Fisher-information I_{θ_0}. Assume that for every sequence of balls $(K_n)_{n\geq 1} \subset \mathbb{R}^d$ with radii $M_n \to \infty$, we have:*

$$\Pi_n\Big(H \in K_n \;\Big|\; X_1, \ldots, X_n \Big) \xrightarrow{P_0} 1. \tag{25}$$

Then the posterior converges to a sequence of normal distributions in total variation:

$$\sup_B \Big| \Pi_n\Big(H \in B \;\Big|\; X_1, \ldots, X_n \Big) - N_{\Delta_{n,\theta_0}, I_{\theta_0}^{-1}}(B) \Big| \xrightarrow{P_0} 0. \tag{26}$$

Connection with the MLE

The Δ_{n,θ_0} describe the local behaviour of the maximum-likelihood estimator (up to $o_{P_0}(1)$).

Lemma 69.1. *Assume that $\mathscr{P}$ is differentiable in quadratic mean at θ_0 with non-singular Fisher information. Moreover, suppose that there exists an $L_2(P_0)$-function m such that for all pairs θ_1, θ_2 in a neighbourhood of θ_0,*

$$|\log p_{\theta_1}(X) - \log p_{\theta_2}(X)| \leq m(X)\|\theta_1 - \theta_2\|.$$

Then any consistent estimator $\widehat{\theta}_n$ such that

$$\mathbb{P}_n \log p_{\widehat{\theta}_n} \geq \sup_{\theta} \mathbb{P}_n \log p_\theta - o_{P_0}(n^{-1})$$

satisfies

$$\sqrt{n}(\widehat{\theta}_n - \theta_0) = \Delta_{n,\theta_0} + o_{P_0}(1). \tag{27}$$

Posterior and MLE

As a result

$$\left\| N_{\sqrt{n}(\hat{\theta}_n - \theta_0), I_{\theta_0}^{-1}} - N_{\Delta_{n,\theta_0}, I_{\theta_0}^{-1}} \right\| \xrightarrow{P_0} 0$$

Since the total variation norm is invariant under shifts and rescalings, the BvM-assertion can be rewritten:

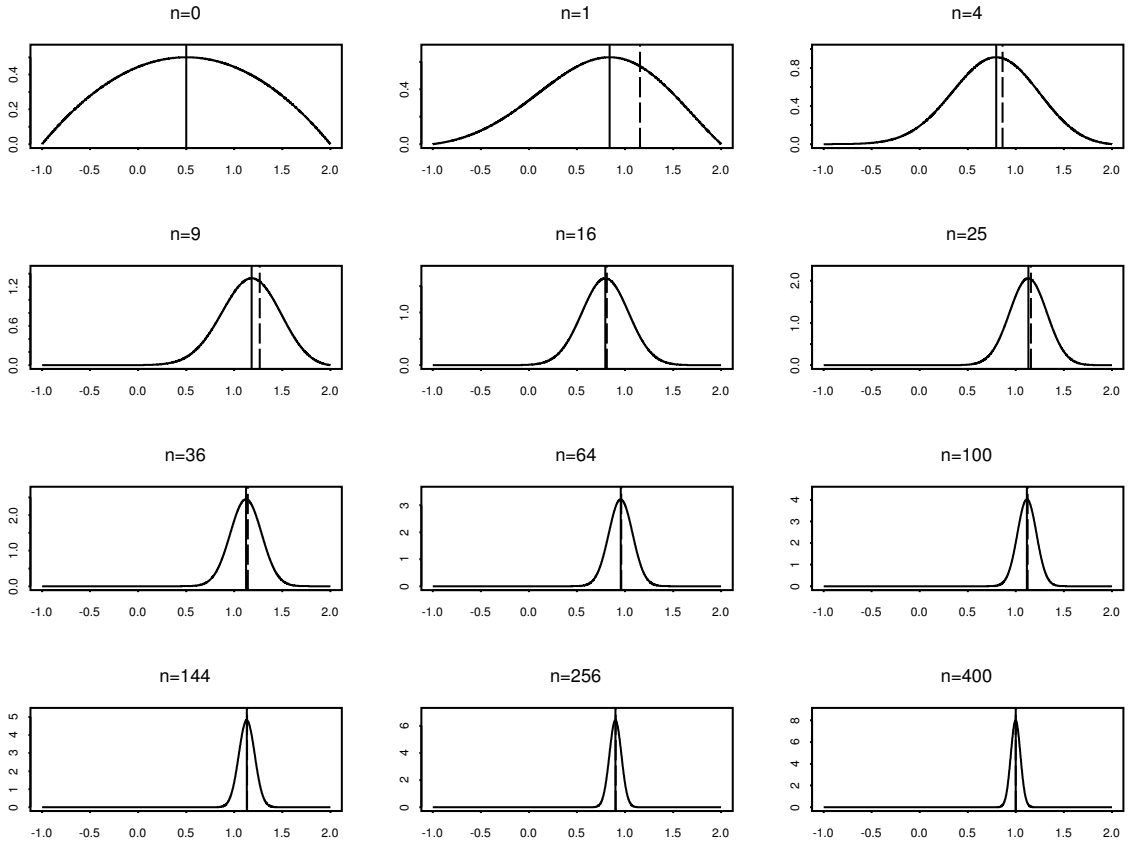
$$\left\| \Pi\Big(\bar{\theta} \in B \;\Big|\; X_1, \ldots, X_n \Big) - N_{\hat{\theta}_n, n^{-1} I_{\theta_0}^{-1}}(B) \right\| \xrightarrow{P_0} 0.$$

Usually, $\hat{\theta}_n$ is the MLE, but any best regular estimator satisfies (27) and can be used in this role.

We see that

Bayesian point-estimators and MLE coincide up to $o_{P_0}(1)$.

Simulation



Posterior density for growing n on a normal location model

Credible sets and confidence regions (I)

Bayesian

A credible set of level α is any subset B_α of the model such that:

$$\Pi(\,\bar{\theta} \in B_\alpha \,|\, X_1, \ldots, X_n\,) \geq 1 - \alpha. \tag{28}$$

Frequentist

A confidence region of level α is a (random) subset $C_\alpha(X)$ of the model such that:

$$P_0^n(\,\theta_0 \in C_\alpha(X)\,) \geq 1 - \alpha \tag{29}$$

Credible sets and confidence regions (II)

Let $\Theta \subset \mathbb{R}$. Assume that the BvM-assertion holds.

Let $\alpha > 0$. Based on (a realization of) the posterior, pick a (sample-dependent) credible interval $B_\alpha(X)$. Define $B'_\alpha(X) = \sqrt{n}(B_\alpha(X) - \theta_0)$. For every $\delta > 0$,

$$P_0^n\Big(\Big|\Pi_n(\, B'_\alpha(X) \,|\, X_1, \ldots, X_n\,) - N_{\Delta_n, I_0^{-1}}(B'_\alpha(X))\Big| > \alpha\Big) < \delta,$$

for large enough n. Using (28)

$$P_0^n\Big(N_{\Delta_n, I_0^{-1}}(B'_\alpha(X)) \leq 1 - 2\alpha\Big) < \delta,$$

Credible sets and confidence regions (III)

If $h \notin B'_\alpha(X)$ then $N_{h,I_0^{-1}}(B'_\alpha(X)) < 1/2$, so if $\alpha < 1/4$,

$$P_0^n(\, \Delta_{n,I_0^{-1}} \in B'_\alpha(X)\,) > 1 - \delta.$$

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

$$P_0^n(\, \theta_0 \in B_\alpha(X)\,) > 1 - 2\delta.$$

for large enough n (refer to 29)). We conclude that

If the BvM-assertion holds, credible sets are confidence regions asymptotically.

This is important, because it is relatively easy to compute (or rather, simulate) posterior distributions. (Markov-chain Monte-Carlo)

Semiparametric Bernstein-Von Mises theorem? (I)

The posterior limitshape in non-parametric models cannot be expected to be Gaussian. It is simply too good to be true. (Freedman (1998)).

But think of the following: we have a semiparametric model

$$\Theta \times H \to \mathscr{P} : (\theta, \eta) \mapsto P_{\theta,\eta}$$

where θ is a finite-dimensional *parameter of interest* and η is an infinite-dimensional *nuisance parameter*. We assume that the sample $X_1, X_2, \ldots$ is *i.i.d.*P_{θ_0,η_0}-distributed, for some $\theta_0 \in \Theta$ and $\eta_0 \in H$. We are interested only in estimation of θ_0,

Frequentist theory for differentiable semiparametric estimation problems is well-known (see *e.g.* van der Vaart (1988)). (Cox-model, models in survival analysis, mixture-models, errors-in-variables regression, etc).

Semiparametric Bernstein-Von Mises theorem? (II)

Think of the posterior for a parametric model:

$$d\Pi_{\theta|X}(\theta|X_1,\ldots,X_n) = \frac{\prod_{i=1}^{n} p_\theta(X_i)\, d\Pi(\theta)}{\int_\Theta \prod_{i=1}^{n} p_\theta(X_i)\, d\Pi(\theta)}$$

In semiparametric context

$$d\Pi_{\theta|X}(\theta|X_1,\ldots,X_n) = \frac{\int_H \prod_{i=1}^{n} p_{\theta,\eta}(X_i)\, d\Pi_H(\eta)\, d\Pi_\Theta(\theta)}{\int_\Theta \int_H \prod_{i=1}^{n} p_{\theta,\eta}(X_i)\, d\Pi_H(\eta)\, d\Pi_\Theta(\theta)}$$

Semiparametric Bernstein-Von Mises theorem? (III)

In the parametric Bernstein-Von Mises proof, we only use the LAN-property:

$$\log \prod_{i=1}^{n} \frac{p_{\theta_0+h_n/\sqrt{n}}}{p_{\theta_0}}(X_i) = h_n^T \mathbb{G}_n \dot{\ell}_{\theta_0} - \tfrac{1}{2} h_n^T I_{\theta_0} h_n + o_{P_0}(1)$$

In the semiparametric case, we therefore want:

$$\log \int_H \prod_{i=1}^{n} \frac{p_{\theta_0+h_n/\sqrt{n},\eta}}{p_{\theta_0}}(X_i)\, d\Pi_H(\eta)$$

$$= \log \int_H \prod_{i=1}^{n} \frac{p_{\theta_0,\eta}}{p_{\theta_0}}(X_i)\, d\Pi_H(\eta) + h_n^T \mathbb{G}_n \tilde{\ell}_{\theta_0,\eta_0} - \tfrac{1}{2} h_n^T \tilde{I}_{\theta_0,\eta_0} h_n + o_{P_0}(1)$$

In addition, $\sqrt{n}$-testability for the parameter θ id needed. The original BvM-proof stays intact; we need sufficient conditions for the above...

Work in progress