Frequentist properties of Bayesian procedures for infinite-dimensional parameters

Aad van der Vaart
Vrije Universiteit Amsterdam

Forum Lectures
European Meeting of Statisticians
Toulouse, 2009

Contents

Co-authors


Ismael Castillo


Subhashis Ghosal


Bas Kleijn


Willem Kruijer

Jyri Lember


Frank van der Meulen


Judith Rousseau


Harry van Zanten

Co-authors


Ismael Castillo


Subhashis Ghosal


Bas Kleijn


Willem Kruijer

Jyri Lember


Frank van der Meulen


Judith Rousseau


Harry van Zanten

Bayesian inference

The Bayesian paradigm

- A parameter Θ is generated according to a prior distribution Π.
- Given $\Theta = \theta$ the data X is generated according to a density p_θ.

This gives a joint distribution of (X, Θ).

- Given observed data x the statistician computes the conditional distribution of Θ given $X = x$, the posterior distribution.

$$d\Pi(\theta|\, X) \propto p_\theta(X)\, d\Pi(\theta)$$

The Bayesian paradigm

- A parameter Θ is generated according to a prior distribution Π.
- Given $\Theta = \theta$ the data X is generated according to a density p_θ.

This gives a joint distribution of (X, Θ).

- Given observed data x the statistician computes the conditional distribution of Θ given $X = x$, the posterior distribution.

$$\Pi(\Theta \in B | \, X) = \frac{\int_B p_\theta(X) \, d\Pi(\theta)}{\int_\Theta p_\theta(X) \, d\Pi(\theta)}$$

The Bayesian paradigm

- A parameter Θ is generated according to a prior distribution Π.
- Given $\Theta = \theta$ the data X is generated according to a density p_θ.

This gives a joint distribution of (X, Θ).

- Given observed data x the statistician computes the conditional distribution of Θ given $X = x$, the posterior distribution.

$$d\Pi(\theta|\, X) \propto p_\theta(X)\, d\Pi(\theta)$$

Reverend Thomas

Thomas Bayes (1702–1761, 1763) followed this argument with Θ possessing the *uniform* distribution and X given $\Theta = \theta$ *binomial* (n, θ).

Using his famous rule he computed that the posterior distribution is then *Beta*$(X + 1, n - X + 1)$.

Reverend Thomas

Thomas Bayes (1702–1761, 1763) followed this argument with Θ possessing the *uniform* distribution and X given $\Theta = \theta$ *binomial* (n, θ).

Using his famous rule he computed that the posterior distribution is then *Beta*$(X + 1, n - X + 1)$.

$$\mathrm{P}(a \leq \Theta \leq b) = b - a, \qquad 0 < a < b < 1,$$

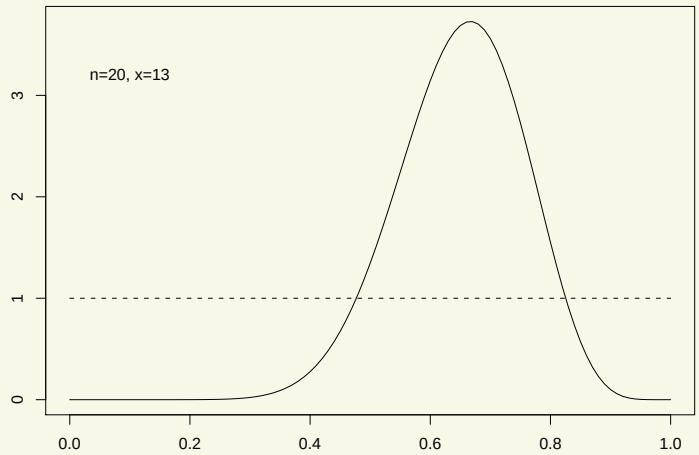
$$\mathrm{P}(X = x | \, \Theta = \theta) = \binom{n}{x} \theta^x (1 - \theta)^{n-x}, \qquad x = 0, 1, \ldots, n,$$

$$d\Pi(\theta | \, X) \propto \theta^X (1 - \theta)^{n-X} \cdot 1.$$

Reverend Thomas

Thomas Bayes (1702–1761, 1763) followed this argument with Θ possessing the *uniform* distribution and X given $\Theta = \theta$ *binomial* (n, θ).

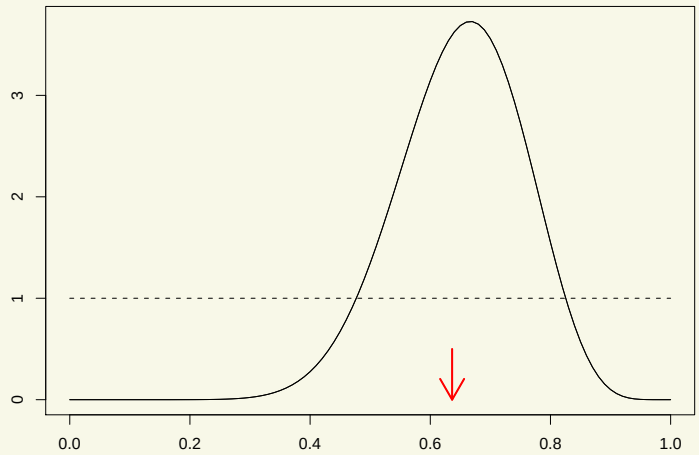
Using his famous rule he computed that the posterior distribution is then *Beta*$(X + 1, n - X + 1)$.



Reverend Thomas

Thomas Bayes (1702–1761, 1763) followed this argument with Θ possessing the *uniform* distribution and X given $\Theta = \theta$ *binomial* (n, θ).

Using his famous rule he computed that the posterior distribution is then *Beta*$(X + 1, n - X + 1)$.

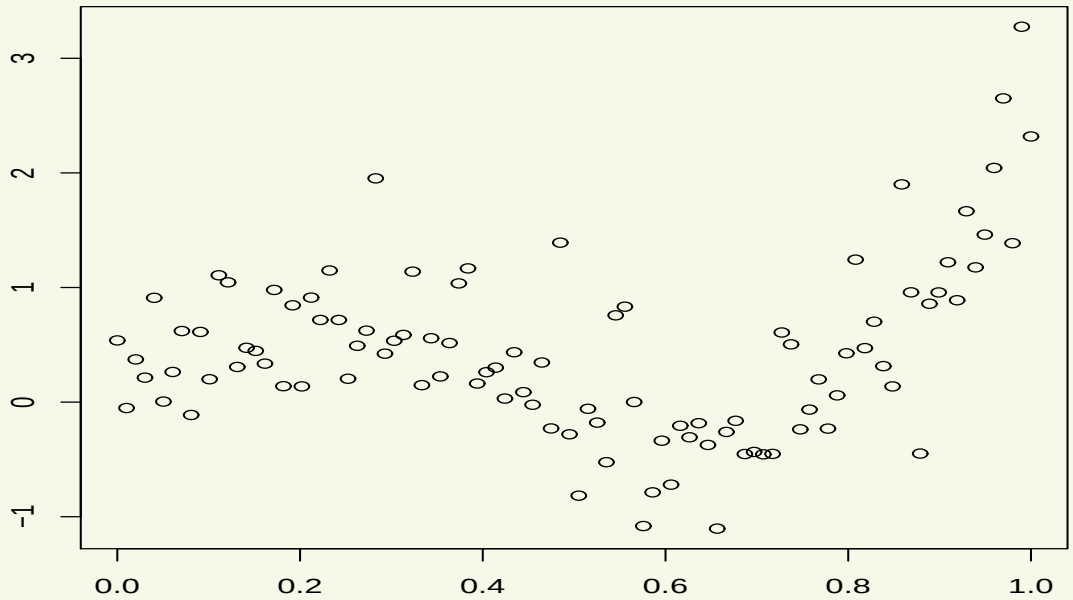


Nonparametric Bayes

If the parameter θ is a function, then the prior is a probability distribution on an function space. So is the posterior, given the data. Bayes' formula does not change:

$$d\Pi(\theta|\, X) \propto p_\theta(X)\, d\Pi(\theta).$$

Prior and posterior can be visualized by plotting functions that are simulated from these distributions.

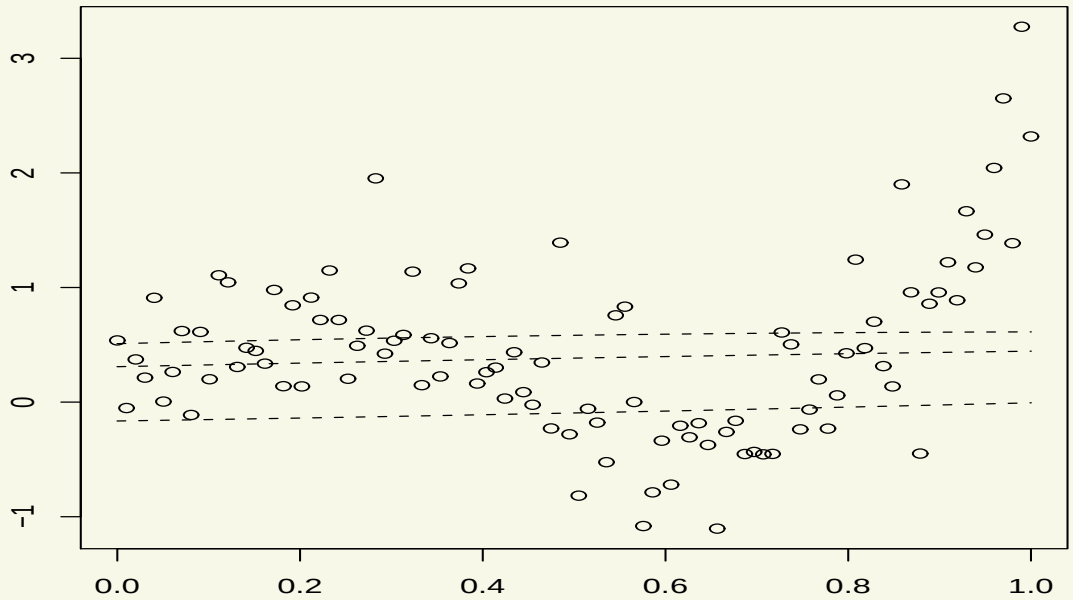


Nonparametric Bayes

If the parameter θ is a function, then the prior is a probability distribution on an function space. So is the posterior, given the data. Bayes' formula does not change:

$$d\Pi(\theta|\, X) \propto p_\theta(X)\, d\Pi(\theta).$$

Prior and posterior can be visualized by plotting functions that are simulated from these distributions.

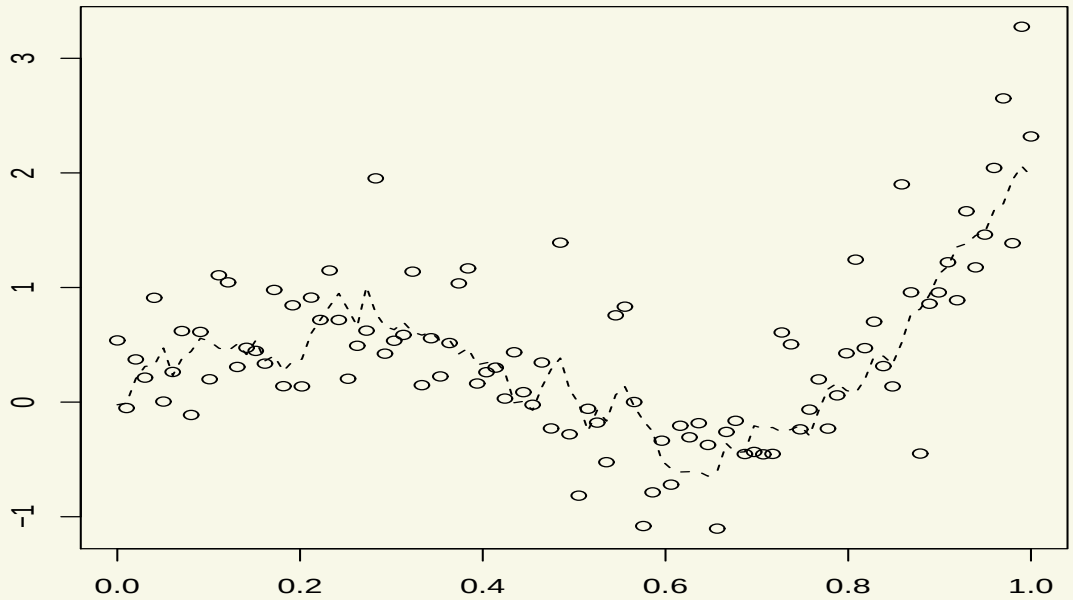


Nonparametric Bayes

If the parameter θ is a function, then the prior is a probability distribution on an function space. So is the posterior, given the data. Bayes' formula does not change:

$$d\Pi(\theta|\, X) \propto p_\theta(X)\, d\Pi(\theta).$$

Prior and posterior can be visualized by plotting functions that are simulated from these distributions.

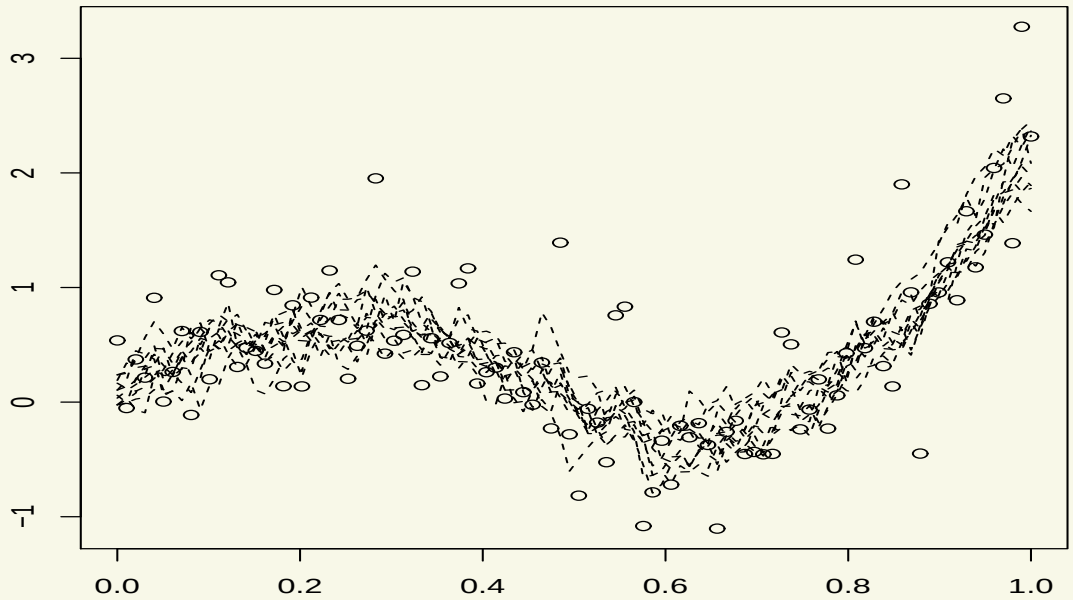


Nonparametric Bayes

If the parameter θ is a function, then the prior is a probability distribution on an function space. So is the posterior, given the data. Bayes' formula does not change:

$$d\Pi(\theta|\, X) \propto p_\theta(X)\, d\Pi(\theta).$$

Prior and posterior can be visualized by plotting functions that are simulated from these distributions.

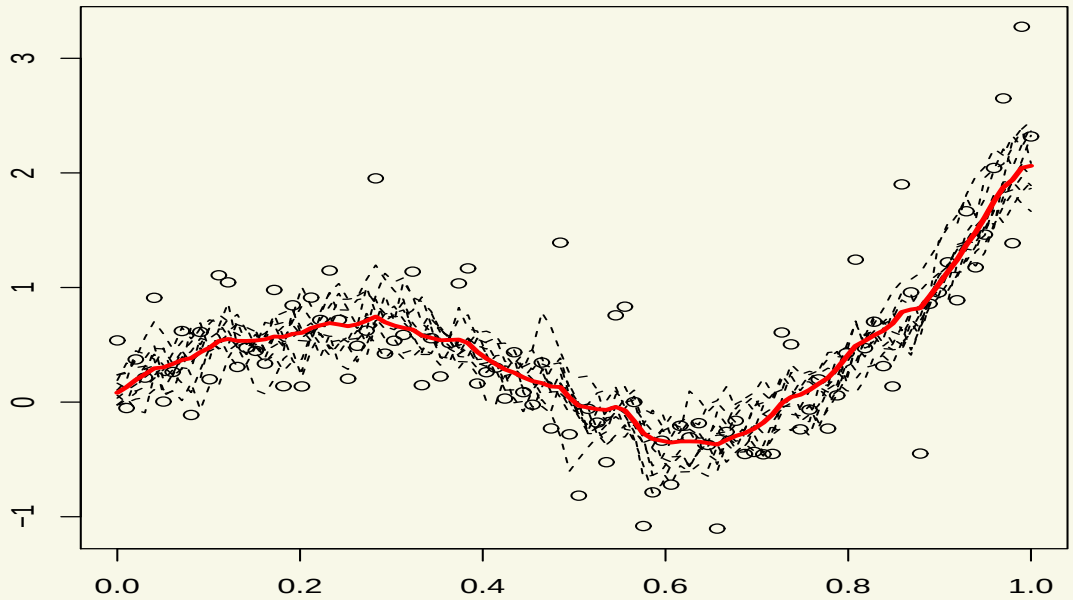


Nonparametric Bayes

If the parameter θ is a function, then the prior is a probability distribution on an function space. So is the posterior, given the data. Bayes' formula does not change:

$$d\Pi(\theta|\, X) \propto p_\theta(X)\, d\Pi(\theta).$$

Prior and posterior can be visualized by plotting functions that are simulated from these distributions.

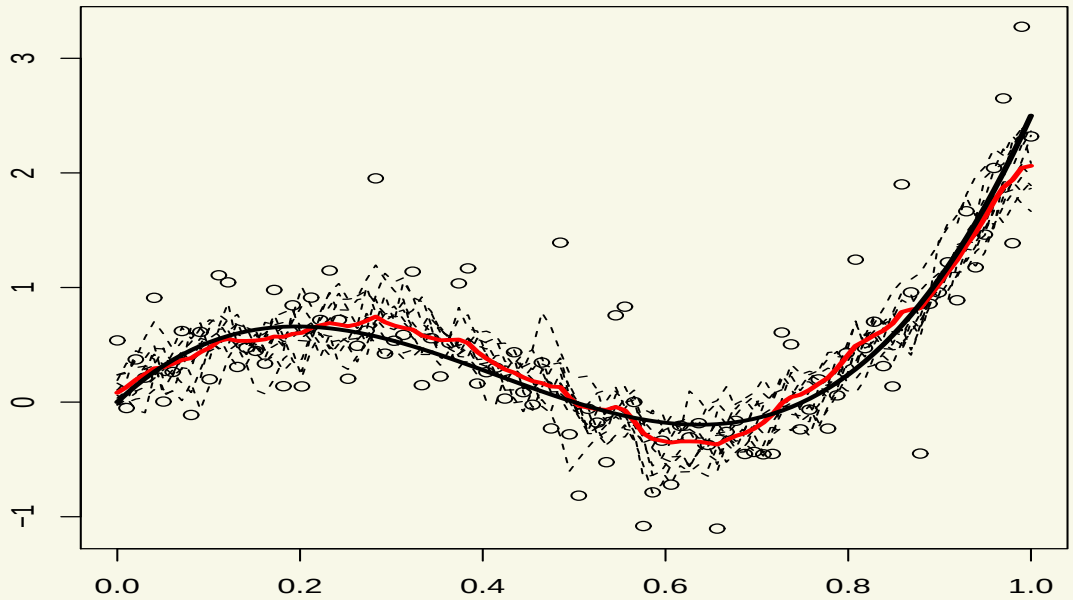


Nonparametric Bayes

If the parameter θ is a function, then the prior is a probability distribution on an function space. So is the posterior, given the data. Bayes' formula does not change:

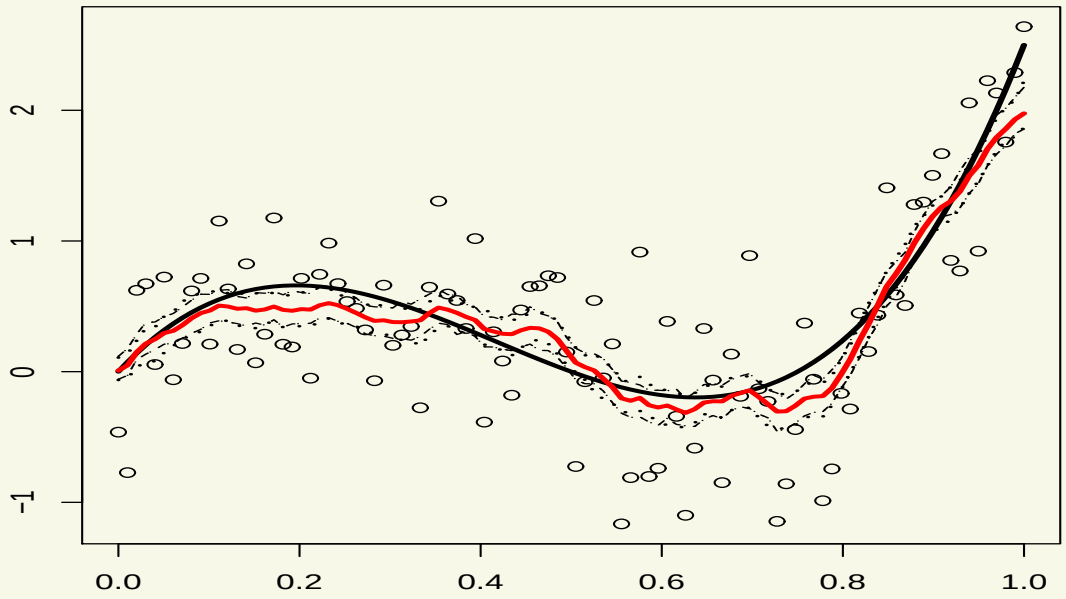
$$d\Pi(\theta|\, X) \propto p_\theta(X)\, d\Pi(\theta).$$

Prior and posterior can be visualized by plotting functions that are simulated from these distributions.



Credibility bands

Posterior gives measure of uncertainty.



75 % pointwise central posterior regions

Computation

Analytical computation of a posterior is rarely possible, but clever algorithms allow to simulate from it.

Markov Chain Monte Carlo (MCMC) produces a Markov chain $\theta_1, \theta_2, \ldots$ that has the posterior as its stationary distribution.

After discarding $\theta_1, \ldots, \theta_k$,

- the average of $\theta_{k+1}, \ldots, \theta_{k+l}$ is taken as estimate of the posterior mean
- the fraction of $\theta_{k+1}, \ldots, \theta_{k+l}$ that falls in a set B is taken as estimate of the posterior mass of B.

Computation

Analytical computation of a posterior is rarely possible, but clever algorithms allow to simulate from it.

Markov Chain Monte Carlo (MCMC) produces a Markov chain $\theta_1, \theta_2, \ldots$ that has the posterior as its stationary distribution.

After discarding $\theta_1, \ldots, \theta_k$,

- the average of $\theta_{k+1}, \ldots, \theta_{k+l}$ is taken as estimate of the posterior mean
- the fraction of $\theta_{k+1}, \ldots, \theta_{k+l}$ that falls in a set B is taken as estimate of the posterior mass of B.

Time-consuming, must be tuned properly, many short-cuts suggested, but feasible (?).

Computation — hierarchical priors

Many priors are defined by a hierarchy of the type:

- $\alpha \sim \Pi_\alpha$
- $\beta|\,\alpha \sim \Pi_{\beta|\alpha}$
- $\gamma|\,\alpha, \beta \sim \Pi_{\gamma|\alpha,\beta}$
- $\ldots$
- $\theta|\,\alpha, \beta, \cdots \sim \Pi_{\theta|\alpha,\beta,\ldots}$.

The prior for θ is a certain mixture of the priors $\Pi_{\theta|\alpha,\beta,\ldots}$ over $\alpha, \beta, \ldots$.

MCMC may simulate a Markov chain $(\alpha_1, \beta_1, \ldots, \theta_1), (\alpha_2, \beta_2, \ldots, \theta_2), \ldots$, and next forget the α's, β's, etc.

Computation — posterior mode

Computing the full posterior is the aim, but can be hard.

Finding the centre and/or spread of the posterior may be a substitute.

There are several general methods (e.g. expectation propagation, Laplace expansion, ..) and many special ones.

Posterior mode — regularization

By Bayes' rule the posterior corresponding to observing $X \sim p_\theta$ has density

$$\pi(\theta|\, X) \propto p_\theta(X)\pi(\theta).$$

The posterior mode maximizes

$$\theta \mapsto \log p_\theta(X) + \log \pi(\theta).$$

The log prior acts as a regularization penalty attached to the log likelihood.

Bayesian thinking suggests penalties.

The Bayesian choice

A true Bayesian may view priors, and hierarchies of priors, as just a way of modelling reality.

- $\alpha \sim \Pi_\alpha$
- $\beta | \, \alpha \sim \Pi_{\beta|\alpha}$
- $\gamma | \, \alpha, \beta \sim \Pi_{\gamma|\alpha,\beta}$
- $\cdots$
- $\theta | \, \alpha, \beta, \cdots \sim \Pi_{\theta|\alpha,\beta,\ldots}$
- $X | \, \theta \sim p_\theta$

The Bayesian choice

A true Bayesian may view priors, and hierarchies of priors, as just a way of modelling reality.

- $\alpha \sim \Pi_\alpha$
- $\beta | \, \alpha \sim \Pi_{\beta|\alpha}$
- $\gamma | \, \alpha, \beta \sim \Pi_{\gamma|\alpha,\beta}$
- $\cdots$
- $\theta | \, \alpha, \beta, \cdots \sim \Pi_{\theta|\alpha,\beta,\ldots}$

  ——————————————- nonBayesian

- $X | \, \theta \sim p_\theta$

The Bayesian choice

A true Bayesian may view priors, and hierarchies of priors, as just a way of modelling reality.

- $\alpha \sim \Pi_\alpha$
- $\beta | \, \alpha \sim \Pi_{\beta|\alpha}$
- $\gamma | \, \alpha, \beta \sim \Pi_{\gamma|\alpha,\beta}$
- $\cdots$

——————————————- empirical Bayes

- $\theta | \, \alpha, \beta, \cdots \sim \Pi_{\theta|\alpha,\beta,\ldots}$
- $X | \, \theta \sim p_\theta$

The Bayesian choice

A true Bayesian may view priors, and hierarchies of priors, as just a way of modelling reality.

Even a true non-Bayesian may like Bayesian methods, because:

- they are elegant
- they allow to incorporate prior information better
- they may be easier to implement

A true non-Bayesian would like to know their performance in a non-Bayesian framework.

Frequentist Bayesian theory

Frequentist Bayesian

Assume that the data X is generated according to a given parameter θ_0 and consider the posterior $\Pi(\theta \in \cdot \,|\, X)$ as a random measure on the parameter set, dependent on X.

We like $\Pi(\theta \in \cdot \,|\, X)$ to put "most" of its mass near θ_0 for "most" X.

Frequentist Bayesian

Assume that the data X is generated according to a given parameter θ_0 and consider the posterior $\Pi(\theta \in \cdot \,|\, X)$ as a random measure on the parameter set, dependent on X.

We like $\Pi(\theta \in \cdot \,|\, X)$ to put "most" of its mass near θ_0 for "most" X.

Asymptotic setting: data X^n where the information increases as $n \to \infty$.
We like the posterior $\Pi_n(\cdot \,|\, X^n)$ to contract to $\{\theta_0\}$, at a good rate.

Three desirable properties:

- Consistency + rate
- Adaptation
- Distributional convergence

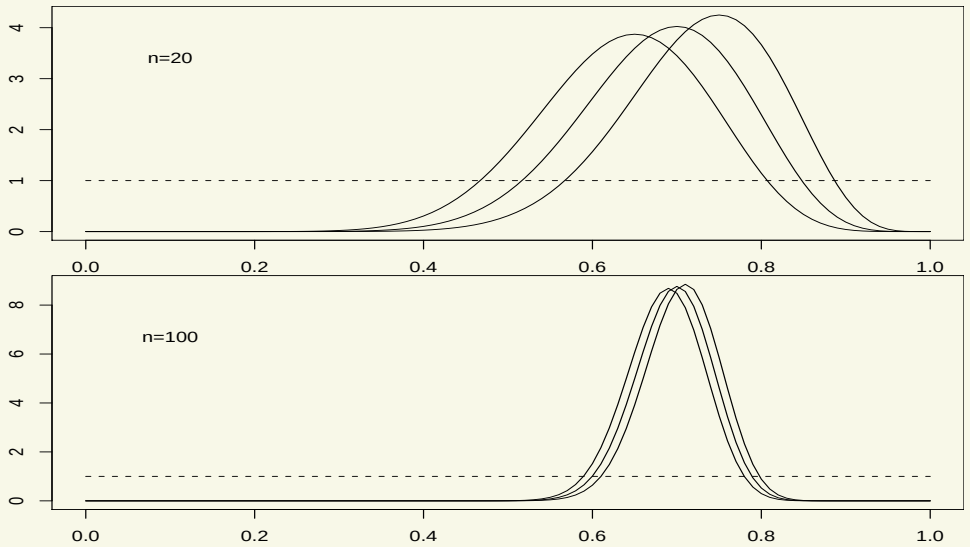
Parametric models

Suppose the data are a random sample $X_1, \ldots, X_n$ from a density $x \mapsto p_\theta(x)$ that is smoothly and identifiably parametrized by $\theta \in \mathbb{R}^d$.

THEOREM [Bernstein, von Mises, LeCam,..]
Under $P_{\theta_0}^n$-probability, for any prior with density that is positive around θ_0,

$$\left\| \Pi(\cdot \,|\, X_1, \ldots, X_n) - N_d\big(\tilde{\theta}_n, \frac{1}{n} I_{\theta_0}^{-1}\big)(\cdot) \right\| \to 0,$$

where $\tilde{\theta}_n$ is any efficient estimator of θ.



Parametric models

Suppose the data are a random sample $X_1, \ldots, X_n$ from a density $x \mapsto p_\theta(x)$ that is smoothly and identifiably parametrized by $\theta \in \mathbb{R}^d$.

THEOREM [Bernstein, von Mises, LeCam,..]
Under $P_{\theta_0}^n$-probability, for any prior with density that is positive around θ_0,

$$\Big\| \Pi(\cdot \,|\, X_1, \ldots, X_n) - N_d\big(\tilde{\theta}_n, \frac{1}{n} I_{\theta_0}^{-1}\big)(\cdot) \Big\| \to 0,$$

where $\tilde{\theta}_n$ is any efficient estimator of θ.

The posterior distribution concentrates most of its mass on balls of radius $O(1/\sqrt{n})$ around θ_0, and the Bayesian credible interval is a standard confidence interval.

Parametric models

Suppose the data are a random sample $X_1, \ldots, X_n$ from a density $x \mapsto p_\theta(x)$ that is smoothly and identifiably parametrized by $\theta \in \mathbb{R}^d$.

THEOREM [Bernstein, von Mises, LeCam,..]
Under $P_{\theta_0}^n$-probability, for any prior with density that is positive around θ_0,

$$\Big\| \Pi(\cdot \,|\, X_1, \ldots, X_n) - N_d\big(\tilde{\theta}_n, \frac{1}{n} I_{\theta_0}^{-1}\big)(\cdot) \Big\| \to 0,$$

where $\tilde{\theta}_n$ is any efficient estimator of θ.

The posterior distribution concentrates most of its mass on balls of radius $O(1/\sqrt{n})$ around θ_0, and the Bayesian credible interval is a standard confidence interval.

The prior washes out completely.

Similar results for nonregular models and non-iid data.

Complete class theorem

According to the complete class theorem [Le Cam (1964)] the set of Bayes procedures is sufficiently rich to dominate every statistical procedure.

Complete class theorem

According to the complete class theorem [Le Cam (1964)] the set of all limits of Bayes procedures is sufficiently rich to dominate every statistical procedure.

Complete class theorem

According to the complete class theorem [Le Cam (1964)] the set of all limits of Bayes procedures is sufficiently rich to dominate every statistical procedure.

Which priors?

Complete class theorem

According to the complete class theorem [Le Cam (1964)] the set of all limits of Bayes procedures is sufficiently rich to dominate every statistical procedure.

Which priors?

Most priors do not work!

[Freedman and Diaconis (1980s)]

Rate of contraction

Assume X^n is generated according to a given parameter θ_0 where the information increases as $n \to \infty$.

- Posterior is consistent if $\mathrm{E}_{\theta_0}\Pi\big(\theta\colon d(\theta,\theta_0) < \varepsilon \big| X^n\big) \to 1$ for every $\varepsilon > 0$.
- Posterior contracts at rate at least ε_n if $\mathrm{E}_{\theta_0}\Pi\big(\theta\colon d(\theta,\theta_0) < \varepsilon_n \big| X^n\big) \to 1$.

Basic results on consistency were proved by Doob (1948) and Schwarz (1965). Interest in rates is recent.

Minimaxity and adaptation

To a given model Θ_α is attached an optimal rate of convergence defined by the minimax criterion

$$\varepsilon_{n,\alpha} = \inf_T \sup_{\theta \in \Theta_\alpha} \mathrm{E}_\theta d\big(T(X), \theta\big).$$

This criterion has nothing to do with Bayes. A prior is good if the posterior contracts at this rate.

Given a scale of regularity classes $(\Theta_\alpha \colon \alpha \in A)$, we like the posterior to adapt: if the true parameter belongs to Θ_α, then we like the contraction rate to be the minimax rate for the α-class.

Minimaxity and adaptation: regression

Consider estimating a function $\theta\colon [0,1]^d \to \mathbb{R}$ based on data $(x_1, Y_1), \ldots, (x_n, Y_n)$, with

$$Y_i = \theta(x_i) + e_i, \qquad \mathrm{E}\varepsilon_i = 0, i = 1, \ldots, n,$$

A standard scale of model classes are the Hölder spaces $C^\alpha[0,1]^d$, defined by the norms

$$\|\theta\|_{C_\alpha} = \sup_x \left|\theta(x)\right| + \sup_{x \neq y} \max_{k=\underline{\alpha}} \frac{\left|D^k\theta(x) - D^k\theta(y)\right|}{|x-y|^{(\alpha-\underline{\alpha})}}.$$

The square minimax rate in $L_2(0,1)$ over these classes is given by

$$\varepsilon_{n,\alpha}^2 = \inf_T \sup_{\|\theta\|_{C_\alpha} \leq 1} \mathrm{E}_\theta \int_{[0,1]^d} \left|T(x_1, Y_1, \ldots, x_n, Y_n)(s) - \theta(s)\right|^2 ds \asymp \left(\frac{1}{n}\right)^{2\alpha/(2\alpha+d)}.$$

Minimaxity and adaptation: other models

For other statistical models (density estimation, classification,...) and types of data (dependence, stochastic processes,..) and other model scales (Sobolev, Besov,..) and distances similar results are valid.

In many examples the minimax rate is $n^{-\alpha/(2\alpha+d)}$ if w_0 is a function of d arguments with partial derivatives of order α bounded by a constant.

Distributional convergence

The posterior of a "parameter" $\phi(\theta)$ is obtained from the posterior for θ by marginalization. For $\phi(\theta) \in \mathbb{R}$ we desire semiparametric Bernstein - von Mises approximations:

$$\Pi\big(\phi(\theta) \in \cdot \,|\, X^{(n)}\big) - N\Big(\hat{\phi}_n, \frac{1}{n\tilde{I}_\phi}\Big)(\cdot) \xrightarrow{P} 0,$$

where $\hat{\phi}_n$ is a (semiparametrically) efficient estimator and $\tilde{I}_\phi$ the efficient Fisher information.

For nonregular parameters we expect nonnormal distributions.

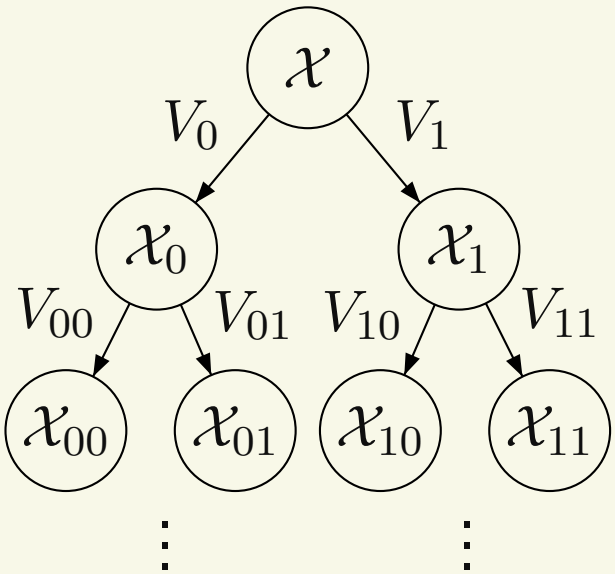
Examples

Priors on distributions: Polya trees

Given a sequence of binary partitions:

$$\mathcal{X} = \mathcal{X}_0 \cup \mathcal{X}_1 = (\mathcal{X}_{00} \cup \mathcal{X}_{01}) \cup (\mathcal{X}_{10} \cup \mathcal{X}_{11}) = \cdots,$$

assign the total mass 1 by splitting it randomly over the partitioning sets using independent Beta variables $V_0, V_{00}, V_{10}, \cdots$.



Parameters of Betas determine properties.

Dirichlet process

The Dirichlet process prior is the Polya tree prior with the parameters of V_ε equal to $\big(\alpha(\mathcal{X}_{\varepsilon 0}), \alpha(\mathcal{X}_{\varepsilon 1})\big)$ for a fixed measure α, the mean measure.

It puts mass on discrete measures only, and is a true nonparametric prior.

THEOREM
If $X_1, \ldots, X_n$ iid P_0 and the prior for P is Dirichlet (α), then, uniformly in sets B,

$$\Pi_n\Big(P\colon P(B) \in \cdot \,|\, X_1, \ldots, X_n\Big) - N\Big(\mathbb{P}_n(B), \frac{P_0(B)(1-P_0(B))}{n}\Big)(\cdot) \xrightarrow{P} 0,$$

(The posterior is Dirichlet $(\alpha + n\mathbb{P}_n)$.)

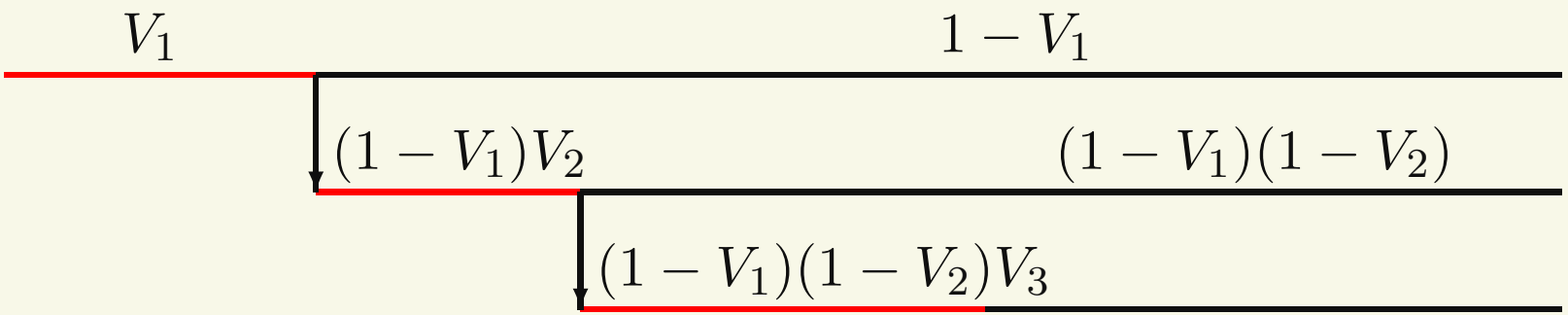
[Ferguson (1973, 74), Lo]

Priors on distributions: stick-breaking

A random discrete probability measure is obtained as

$$P = \sum_{i=1}^{\infty} W_i \delta_{Z_i}, \qquad W_i = (1 - V_1) \cdots (1 - V_{i-1}) V_i,$$

for independent $V_1, V_2, \ldots$ in $(0, 1)$ independent of iid $Z_1, Z_2, \ldots,$.



(By making the V_i or Z_i dependent on a covariate can model conditional distributions.)

Priors on distributions: stick-breaking

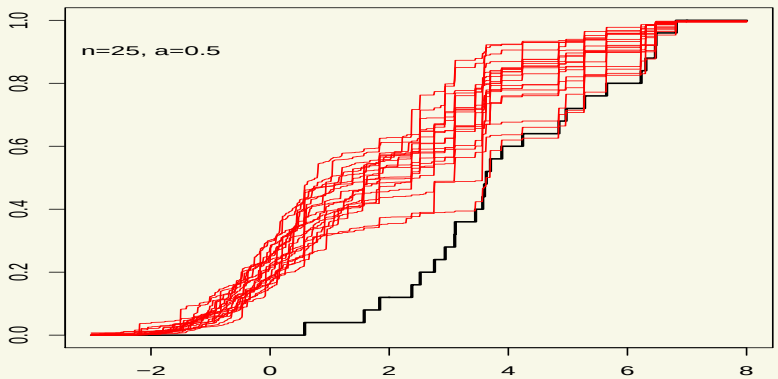
A random discrete probability measure is obtained as

$$P = \sum_{i=1}^{\infty} W_i \delta_{Z_i}, \qquad W_i = (1 - V_1) \cdots (1 - V_{i-1}) V_i,$$

for independent $V_1, V_2, \ldots$ in $(0, 1)$ independent of iid $Z_1, Z_2, \ldots,$.

The Pitman-Yor process is the special case with $V_i \sim$ Beta$(1 - \alpha, \theta + i\alpha)$.
The Dirichlet the subcase with $\alpha = 0$.

THEOREM [James (2008)]
The Bernstein-von Mises theorem holds only if $\alpha = 0$.



Dirichlet mixtures

A prior on densities can be obtained by putting the Dirichlet on the mixing distribution P in

$$x \mapsto \int \frac{1}{\sigma}\phi\Big(\frac{x-z}{\sigma}\Big)\, dP(z),$$

with ϕ e.g. the normal density. We can also put a prior on the scale σ.

Dirichlet mixtures — computation

- $P \sim$ Dirichlet(α).
- $Z_1, \ldots, Z_n | \, P \sim$ iid P.
- $\varepsilon_1, \ldots, \varepsilon_n | \, P, Z_1, \ldots Z_n$ iid $\sim N(0,1)$.
- Observations $X_i = Z_i + \varepsilon_i$.

Then $Z_i | \, Z_j : j \neq i, X_1, \ldots, X_n \sim$ mixture of empirical of $(Z_j : j \neq i)$ and α.
"Gibbs sampler for simulating $Z_1, \ldots, Z_n | \, X_1, \ldots, X_n$ is partial bootstrap".

Also $P | \, Z_1, \ldots, Z_n, X_n \ldots, Z_n \sim$ Dirichlet $(\alpha + \sum \delta_{Z_i})$.

```
for (i in 1:n){  # GIBSS LOOP
        weights <- dnorm(x[i]-z,0,sigma)
        weights[i] <- 0
        wold <- sum(weights)
        if (runif(1)< wold/(wold+wnew[i])){
                j <- sample(1:n,size=1,prob=weights)
                z[i] <- z[j]
        }
        else {
        z[i] <- rnorm(1,x[i]*ts/sigma^2,sqrtts)
        }
} # END GIBBS LOOP
```

[Escobar & West (1995)]

Dirichlet mixtures of normal

$$p_{F,\sigma}(x) = \int \frac{1}{\sigma}\phi\Big(\frac{x-z}{\sigma}\Big)\, dF(z).$$

Observe a random sample of size n from density p_0 on $\mathbb{R}$. Put Dirichlet prior on F, and positive prior on $\sigma \in (a, b) \subset (0, \infty)$.

THEOREM

If $p_0 = p_{F_0,\sigma_0}$ for F_0 with subGaussian tails and $\sigma_0 \in (a, b)$, then the rate of contraction relative to Hellinger distance is $(\log n)^{\kappa}/\sqrt{n}$.

THEOREM

If p_0 is C^2 and has subGaussian tails, and the prior on σ shrinks at rate $n^{-1/5}$, then the rate of contraction is $(\log n)^{\lambda}/n^{-2/5}$.

Dirichlet mixtures of normal

$$p_{F,\sigma}(x) = \int \frac{1}{\sigma}\phi\Big(\frac{x-z}{\sigma}\Big)\, dF(z).$$

Observe a random sample of size n from density p_0 on $\mathbb{R}$. Put Dirichlet prior on F, and positive prior on $\sigma \in (a, b) \subset (0, \infty)$.

THEOREM

If $p_0 = p_{F_0,\sigma_0}$ for F_0 with subGaussian tails and $\sigma_0 \in (a, b)$, then the rate of contraction relative to Hellinger distance is $(\log n)^\kappa/\sqrt{n}$.

THEOREM

If p_0 is C^2 and has subGaussian tails, and the prior on σ shrinks at rate $n^{-1/5}$, then the rate of contraction is $(\log n)^\lambda/n^{-2/5}$.

Conjecture: if $p_0 \in C^\alpha$ with small tails and the prior on σ is inverse Gamma, then rate is $(\log n)^\lambda/n^{-\alpha/(2\alpha+1)}$. [Kruijer& Rousseau (2009)].

Bayes is smarter than kernel methods.

Dirichlet mixtures of Beta

For densities on $[0, 1]$ it is natural to take mixtures of beta densities.

$$x \mapsto \int c_{\alpha,\beta}\, x^{\alpha}(1-x)^{\beta}\, dF \circ \psi(\alpha, \beta).$$

Similar results hold. A reparameterization of $(\alpha, \beta) \mapsto \psi(\alpha, \beta)$ is necessary for the best results.

[Petrone, Ghosal, Kruijer, Rousseau]

Gaussian priors

The law of a stochastic process $(W_t : t \in T)$ is a prior distribution on the space of functions $w : T \to \mathbb{R}$.

Gaussian processes have been found useful, because

- they offer great variety
- they are easy (?) to understand through their covariance function $(s, t) \mapsto \mathrm{E} W_s W_t$
- they can be computationally attractive

Brownian density estimation

For W Brownian motion use as prior on a density p on $[0,1]$:

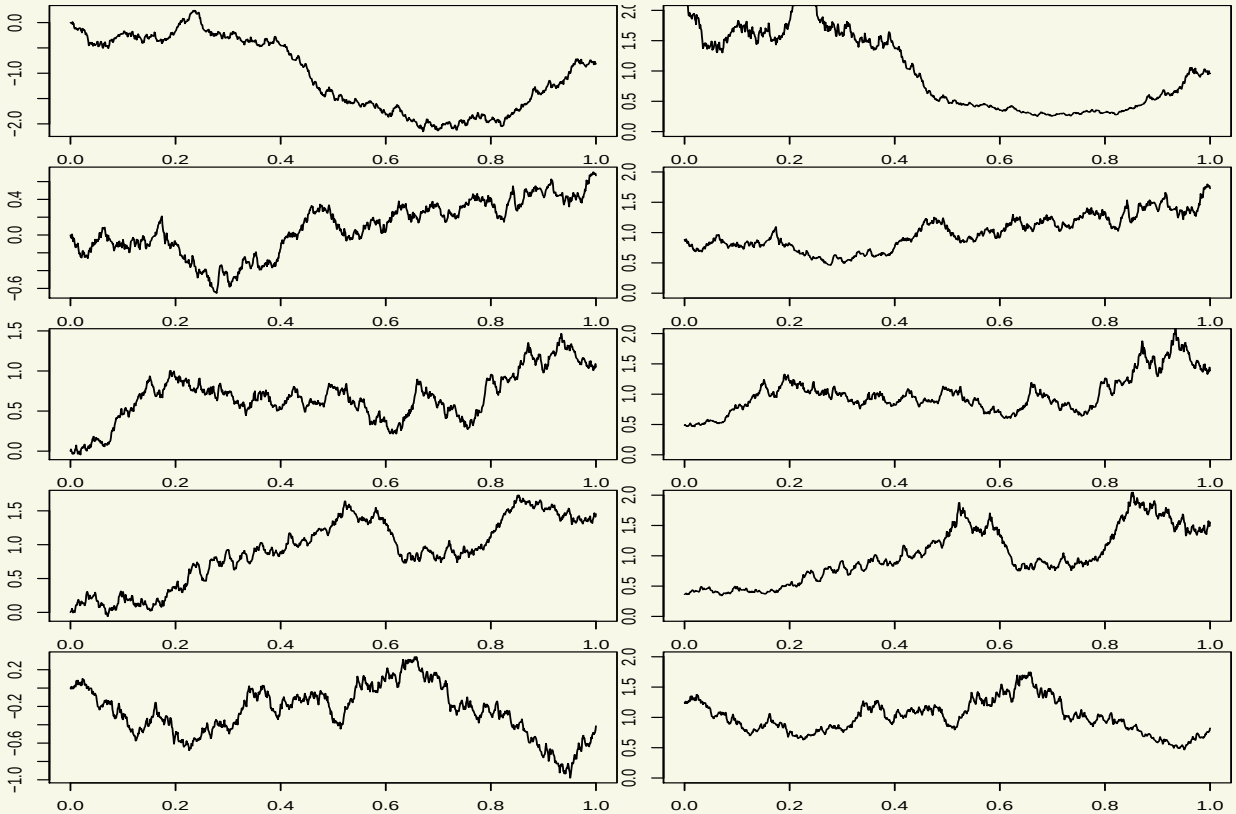
$$x \mapsto \frac{e^{W_x}}{\int_0^1 e^{W_y}\, dy}.$$

[Leonard, Lenk, Tokdar & Ghosh]

Brownian density estimation

For W Brownian motion use as prior on a density p on $[0,1]$:

$$x \mapsto \frac{e^{W_x}}{\int_0^1 e^{W_y}\, dy}.$$



Brownian motion $t \mapsto W_t$ — Prior density $t \mapsto c\exp(W_t)$

Brownian density estimation

Let $X_1, \ldots, X_n$ be iid p_0 on $[0,1]$ and let W Brownian motion. Let the prior be

$$x \mapsto \frac{e^{W_x}}{\int_0^1 e^{W_y}\, dy}$$

THEOREM

If $w_0 := \log p_0 \in C^\alpha[0,1]$, then L_2-rate is: $n^{-1/4}$ if $\alpha \geq 1/2$;
$n^{-\alpha/2}$ if $\alpha \leq 1/2$.

Brownian density estimation

Let $X_1, \ldots, X_n$ be iid p_0 on $[0,1]$ and let W Brownian motion. Let the prior be

$$x \mapsto \frac{e^{W_x}}{\int_0^1 e^{W_y}\, dy}$$

THEOREM

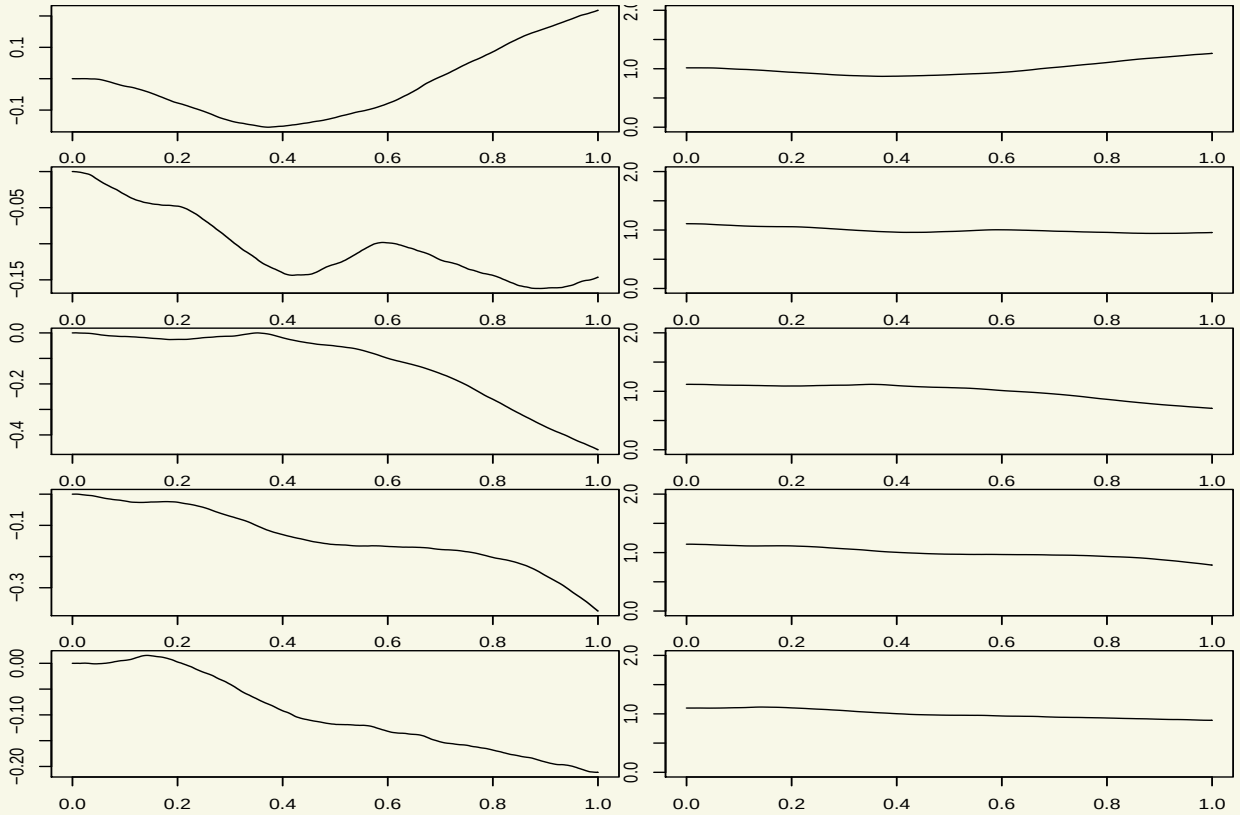
If $w_0 := \log p_0 \in C^\alpha[0,1]$, then L_2-rate is: $n^{-1/4}$ if $\alpha \geq 1/2$;
$n^{-\alpha/2}$ if $\alpha \leq 1/2$.

- This is optimal if and only if $\alpha = 1/2$.
- Rate does not improve if α increases from $1/2$.
- Consistency for any $\alpha > 0$.

[vZanten, Castillo (2008)].

Integrated Brownian density estimation

Taking a primitive smoothes.



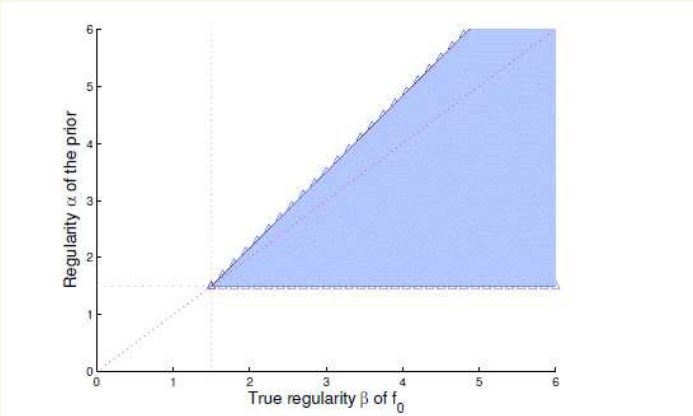
Integrated Brownian motion — Prior density

Gaussian processes — Cox Model

- $T \perp C | Z$
- hazard function $\lambda_{T|Z}(t) = \lambda(t)e^{\theta Z}$
- Observed $X = (T \wedge C, Z, \Delta = 1_{T \leq C})$

Prior on $\log \lambda$ the Riemann-Liouville process of order α (i.e. $(\alpha - 1/2)$ times integrated Brownian motion).

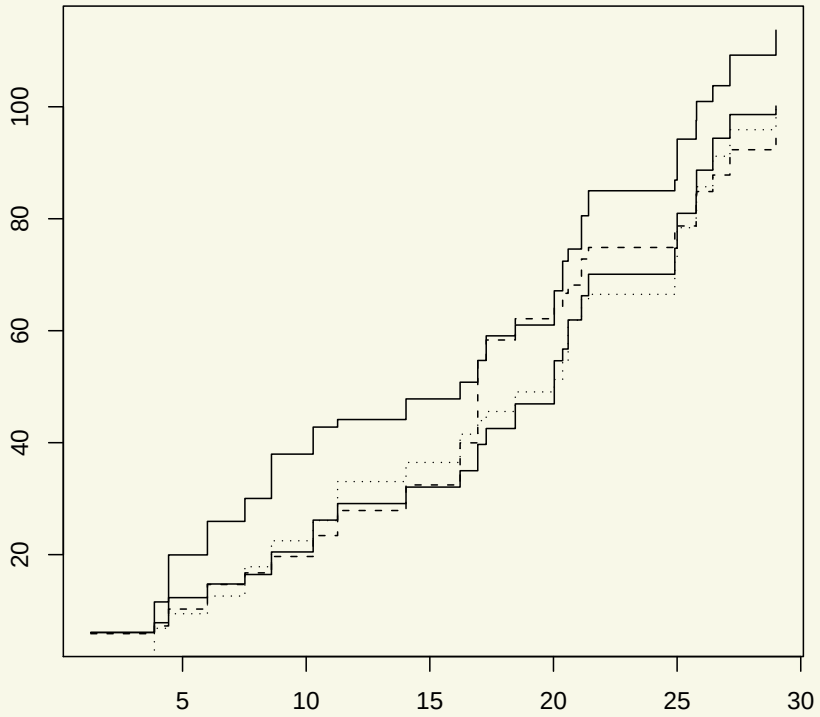
THEOREM [Castillo (2008)]
If $\log \lambda_0 \in C^\beta[0, \tau]$ for $\beta > 3/2$ and $\gamma_0 \in C^{2\beta/3}[0, \tau]$, then the Bernstein-von Mises theorem for estimating θ holds for $\alpha \in (3/2, 4\beta/3 - 1/2)$.



Independent increment processes

A prior on monotone functions can be obtained by placing randomly generated jumps at the event times of a Poisson process (a compound Poisson process).

For nonparametric priors we need more jumps, as in Lévy processes or general independent increment processes.



Independent increments processes — survival analysis

As prior a cumulative hazard function, use an IIP with jump measure

$$\nu(dt, dx) = \left(\frac{\kappa(t)}{x} + f_t(x)\right) dt\, dx, \qquad 0 \le t \le \tau, 0 \le x \le 1.$$

THEOREM [Kim & Lee (2004)]
The Bernstein - von Mises theorem holds if $0 \ll \kappa \ll \infty$ and

$$\sup_{x,t} \left| \frac{(1-x) f_t(x)}{x^\beta} \right| < \infty, \qquad \beta > 3/2.$$

Particular examples are Beta processes [Hjort (1990)] and Gamma processes. For $\beta \le 3/2$ there are counterexamples.

This extends to the Cox model [Kim (2006)], under slightly stronger assumptions.

Sparsity (1)

Observe independent $X_1, \ldots, X_n$, where X_i is $N(\theta_i, 1)$.

$$p_n := \#(1 \leq i \leq n : \theta_i \neq 0).$$

Prior on $\theta = (\theta_1, \ldots, \theta_n)$ constructed in three steps:

- Choose p from π_n on $\{1, 2, \ldots, n\}$.
- Given p choose $S \subset \{1, \ldots, n\}$ of size $|S| = p$ at random.
- Given (p, S) choose $(\theta_i : i \in S)$ from density g_S on $\mathbb{R}^p$ and set $(\theta_i : i \notin S) = 0$.

Special case: $\theta_1, \ldots, \theta_n$ iid $\alpha_n \delta_0 + (1 - \alpha_n) G$.
[George, Johnstone& Silverman]

Sparsity (1)

Observe independent $X_1, \ldots, X_n$, where X_i is $N(\theta_i, 1)$.

$$p_n := \#(1 \leq i \leq n : \theta_i \neq 0).$$

Prior on $\theta = (\theta_1, \ldots, \theta_n)$ constructed in three steps:

- Choose p from π_n on $\{1, 2, \ldots, n\}$.
- Given p choose $S \subset \{1, \ldots, n\}$ of size $|S| = p$ at random.
- Given (p, S) choose $(\theta_i : i \in S)$ from density g_S on $\mathbb{R}^p$ and set $(\theta_i : i \notin S) = 0$.

THEOREM

If $\pi_n(p) \propto e^{-p \log(n/p)}$ and g_S has heavy tails (e.g. Cauchy or Laplace), then rate of "contraction" for Euclidean norm is $p_n \log(n/p_n)$. (Gaussian priors shrink too much.)

[Castillo]

Sparsity (2)

We wish to build a prediction model for Y given $X_1, X_2, \ldots, X_p$.
The number of predictors p is large, but only few should matter.

We place prior weights on models that include various sets of X_i.
We combine these with priors on the models into an overall prior.

Conjecture: In linear regression under RIP similar results hold.

[Jiang, Yuan&Li, Johnstone&Silverman, Abramovich et al,...]

Series priors

Given a basis $e_1, e_2, \ldots$ put a prior on the coefficients $(\theta_1, \theta_2, \ldots)$ in an expansion

$$\theta = \sum_i \theta_i e_i.$$

The θ_i should decrease as $i \to \infty$. The rate of decrease determines behaviour of the posterior. Many results with Fourier series, wavelets, splines,...

Adaptation

Given a countable collection of models indexed by $\alpha \in A_n$, each with its own rate $\varepsilon_{n,\alpha}$ and prior $\Pi_{n,\alpha}$, form the hierarchical prior:

- choose α with weights $w_{n,\alpha} \propto \mu_\alpha e^{-Cn\varepsilon_{n,\alpha}^2}$.
- choose parameter θ according to $\Pi_{n,\alpha}$.

THEOREM
Under general conditions the posterior rate is approximately $\varepsilon_{n,\beta}$ if the true parameter belongs to model β.

Under more complicated conditions or special models similar results hold for more general weights $w_{n,\alpha}$. There are also elegant special constructions [e.g. Lecture II.]

[Scriciollo, T. Huang, Belitser, Lember,...]

Misspecification

If the true parameter is outside the support of the prior, then the posterior cannot contract to it.

THEOREM
Under general conditions the posterior contracts to the parameter “in the support” at minimal Kullback-Leibler divergence to the true parameter, at a rate as if it were “in the support”.

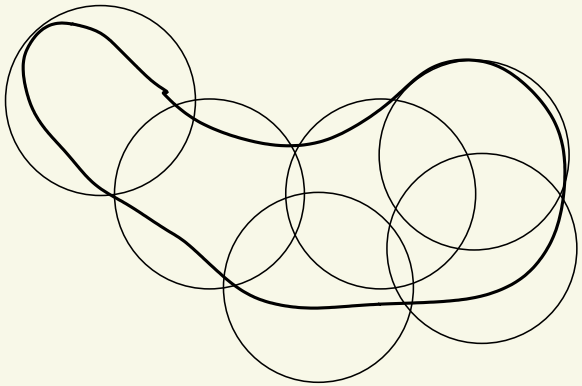
For example, a Bayesian may misrepresent the error in nonparametric regression as Gaussian, but still get consistency for the regression function.

[Klein (2006)]

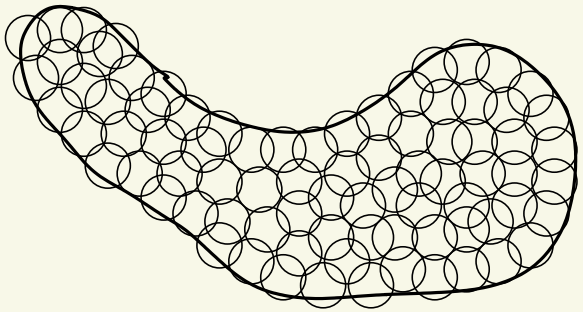
Rates — iid

Entropy

The covering number $N(\varepsilon, \Theta, d)$ of a metric space (Θ, d) is the minimal number of balls of radius ε needed to cover Θ.



ε big



ε small

Entropy is the logarithm $\log N(\varepsilon, \Theta, d)$

Entropy

The covering number $N(\varepsilon, \Theta, d)$ of a metric space (Θ, d) is the minimal number of balls of radius ε needed to cover Θ.

Entropy of a set of densities Θ relative to the Hellinger distance h characterizes the minimax rate of convergence for density estimation relative to h by the equation [Le Cam (73,75,86), Birgé (83,06)]

$$\log N(\varepsilon_n, \Theta, h) \asymp n\varepsilon_n^2.$$

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

Rate — iid observations

Given a random sample $X_1, \ldots, X_n$ from a density p_0 and a prior Π on a set $\mathcal{P}$ of densities consider the posterior

$$d\Pi_n(p|X_1, \ldots, X_n) \propto \prod_{i=1}^{n} p(X_i)\, d\Pi(p).$$

THEOREM

The Hellinger contraction rate is ε_n if there exist $\mathcal{P}_n \subset \mathcal{P}$ such that

(1) $\log N(\varepsilon_n, \mathcal{P}_n, h) \leq n\varepsilon_n^2$ and $\Pi(\mathcal{P}_n) = 1 - o(e^{-3n\varepsilon_n^2})$. entropy

(2) $\Pi(B_{KL}(p_0, \varepsilon_n)) \geq e^{-n\varepsilon_n^2}$. prior mass.

$B_{KL}(p_0, \varepsilon)$ is Kullback-Leibler neighborhood of p_0.

Rate — iid observations

Given a random sample $X_1, \ldots, X_n$ from a density p_0 and a prior Π on a set $\mathcal{P}$ of densities consider the posterior

$$d\Pi_n(p|X_1, \ldots, X_n) \propto \prod_{i=1}^{n} p(X_i)\, d\Pi(p).$$

THEOREM

The Hellinger contraction rate is ε_n if there exist $\mathcal{P}_n \subset \mathcal{P}$ such that

(1) $\log N(\varepsilon_n, \mathcal{P}_n, h) \leq n\varepsilon_n^2$ and $\Pi(\mathcal{P}_n) = 1 - o(e^{-3n\varepsilon_n^2})$. entropy

(2) $\Pi(B_{KL}(p_0, \varepsilon_n)) \geq e^{-n\varepsilon_n^2}$. prior mass.

We need $N(\varepsilon_n, \mathcal{P}_n, h) \approx e^{n\varepsilon_n^2}$ balls to cover the model. If the mass is uniformly spread, then every ball has mass

$$\frac{1}{N(\varepsilon_n, \mathcal{P}_n, h)} \approx e^{-n\varepsilon_n^2}.$$

Rate — iid observations

Given a random sample $X_1, \ldots, X_n$ from a density p_0 and a prior Π on a set $\mathcal{P}$ of densities consider the posterior

$$d\Pi_n(p|X_1, \ldots, X_n) \propto \prod_{i=1}^{n} p(X_i)\, d\Pi(p).$$

THEOREM

The Hellinger contraction rate is ε_n if there exist $\mathcal{P}_n \subset \mathcal{P}$ such that

(1) $\log N(\varepsilon_n, \mathcal{P}_n, h) \leq n\varepsilon_n^2$ and $\Pi(\mathcal{P}_n) = 1 - o(e^{-3n\varepsilon_n^2})$. entropy

(2) $\Pi(B_{KL}(p_0, \varepsilon_n)) \geq e^{-n\varepsilon_n^2}$. prior mass.

We need $N(\varepsilon_n, \mathcal{P}_n, h) \approx e^{n\varepsilon_n^2}$ balls to cover the model. If the mass is uniformly spread, then every ball has mass

$$\frac{1}{N(\varepsilon_n, \mathcal{P}_n, h)} \approx e^{-n\varepsilon_n^2}.$$

Other results

THEOREM

The Hellinger contraction rate is ε_n if there exist $\mathcal{P}_n \subset \mathcal{P}$ such that

(1) $\log N(\varepsilon_n, \mathcal{P}_n, h) \leq n\varepsilon_n^2$ and $\Pi(\mathcal{P}_n) = 1 - o(e^{-3n\varepsilon_n^2})$. entropy

(2) $\Pi(B_{KL}(p_0, \varepsilon_n)) \geq e^{-n\varepsilon_n^2}$. prior mass.

The entropy condition ensures that the likelihood is not too variable, so that it cannot be large by pure randomness.

In the Bayesian set-up the likelihood is downweighted by the prior, and a more refined trade-off can be made between prior and complexity. There are several versions of more refined trade-offs. [Ghosal& vdV, Zhang, Xing (2009)].

There are also results only in terms of the prior, but these require stronger conditions ([Lijoi, Pruenster, Walker]).

Rates — general

Setting

For $n = 1, 2, \ldots$

- $(P_\theta^n : \theta \in \Theta_n)$ experiment
- (Θ_n, d_n) metric space
- X^n observation, law $P_{\theta_0}^n$

Given prior Π_n on Θ_n form posterior

$$d\Pi_n(\theta|X^n) \propto p_\theta^n(X^n)\, d\Pi_n(\theta)$$

Setting

For $n = 1, 2, \ldots$

- $(P_\theta^n : \theta \in \Theta_n)$ experiment
- (Θ_n, d_n) metric space
- X^n observation, law $P_{\theta_0}^n$

Given prior Π_n on Θ_n form posterior

$$d\Pi_n(\theta|X^n) \propto p_\theta^n(X^n)\, d\Pi_n(\theta)$$

Rate of contraction is at least ε_n if $\forall M_n \to \infty$

$$P_{\theta_0}^n \Pi_n(\theta \in \Theta_n : d_n(\theta, \theta_0) \geq M_n \varepsilon_n | X^n) \to 0$$

Setting — Le Cam's testing criterion

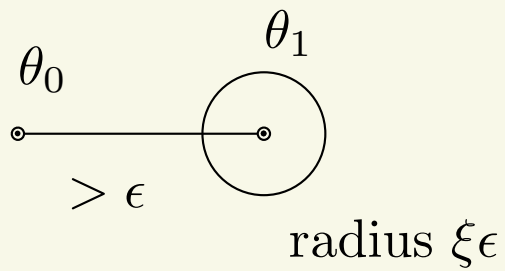
For $n = 1, 2, \ldots$

- $(P_\theta^n : \theta \in \Theta_n)$ experiment
- (Θ_n, d_n) metric space
- X^n observation, law $P_{\theta_0}^n$

Assume $\forall n \; \exists$ metric $\bar{d}_n \geq d_n$ such that $\forall \varepsilon > 0$:

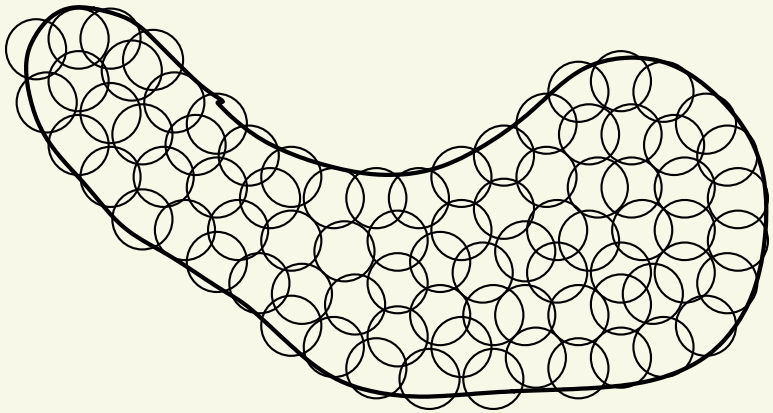
$\forall \theta_1 \in \Theta_n$ with $d_n(\theta_1, \theta_0) > \varepsilon \; \exists$ test ϕ_n with

$$P_{\theta_0}^n \phi_n \leq e^{-n\varepsilon^2}, \qquad \sup_{\theta \in \Theta_n : \bar{d}_n(\theta, \theta_1) < \varepsilon/2} P_\theta^n (1 - \phi_n) \leq e^{-n\varepsilon^2}$$



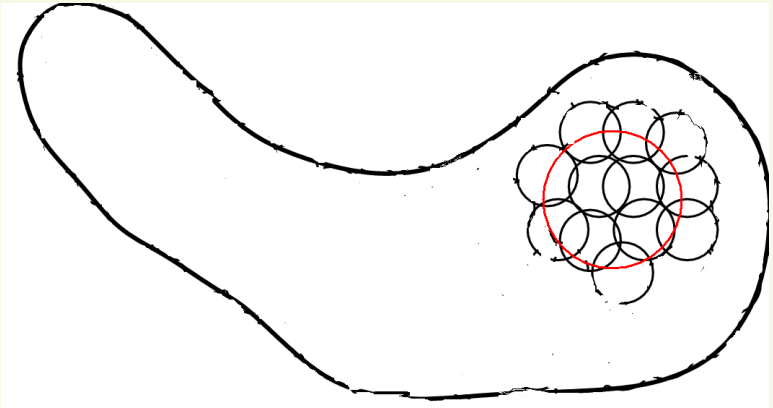
Le Cam dimension = local entropy

$N(\varepsilon, \Theta, d) =$ smallest number of balls of radius ε needed to cover Θ



Le Cam dimension = local entropy

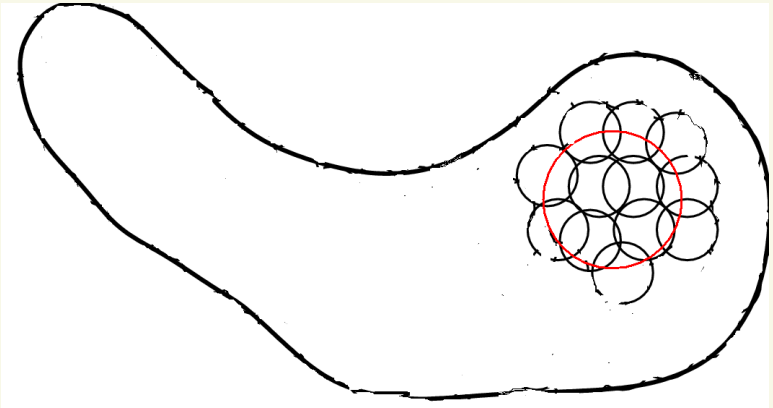
$N(\varepsilon, \Theta, d) =$ smallest number of balls of radius ε needed to cover Θ



$$D_n(\varepsilon, \Theta, d_n, \bar{d}_n) = \sup_{\eta > \varepsilon} \log N\big(\eta/2, \{\theta \in \Theta_n : d_n(\theta, \theta_0) \leq \eta\}, \bar{d}_n\big).$$

Le Cam dimension = local entropy

$N(\varepsilon, \Theta, d) =$ smallest number of balls of radius ε needed to cover Θ



$$D_n(\varepsilon, \Theta, d_n, \bar{d}_n) = \sup_{\eta > \varepsilon} \log N\big(\eta/2, \{\theta \in \Theta_n : d_n(\theta, \theta_0) \leq \eta\}, \bar{d}_n\big).$$

THEOREM [Le Cam (73,75,86), Birgé (83,06)]
There exist estimators $\hat{\theta}_n$ with $d_n(\hat{\theta}_n, \theta_0) = O_P(\varepsilon_n)$ if

$$D_n(\varepsilon_n, \Theta_n, d_n, \bar{d}_n) \leq n\varepsilon_n^2.$$

Rate theorem

THEOREM

The rate of contraction is $\varepsilon_n \gg 1/\sqrt{n}$ if there exist $\tilde{\Theta}_n \subset \Theta_n$ such that

(1) $D_n(\varepsilon_n, \tilde{\Theta}_n, d_n, \bar{d}_n) \leq n\varepsilon_n^2$ and $\Pi_n(\Theta_n - \tilde{\Theta}_n) = o(e^{-3n\varepsilon_n^2})$.

(2) $\Pi_n(B_n(\theta_0, \varepsilon_n; k)) \geq e^{-n\varepsilon_n^2}$.

$$B_n(\theta_0, \varepsilon; k) = \left\{\theta \in \Theta_n : K(p_{\theta_0}^n, p_\theta^n) \leq n\varepsilon^2, V_k(p_{\theta_0}^n, p_\theta^n) \leq n^{k/2}\varepsilon^k\right\}$$

(Kullback-Leibler neighborhood)

$$K(p,q) = P\log(p/q) \qquad V_k(p,q) = P\left|\log(p/q) - K(p,q)\right|^k$$

Sketch of proof

STEP 1: With high probability

$$\int p_\theta^n / p_{\theta_0}^n \, d\Pi_n(\theta) \geq e^{-n\varepsilon_n^2} \Pi_n\big(B_n(\theta_0, \varepsilon_n)\big).$$

STEP 2: There exist tests $\phi_n = 1_{K_n}$ with

$$P_\theta^n(K_n) \to 0, \qquad \sup_{\theta : d_n(\theta, \theta_0) > \varepsilon_n} P_\theta^n(K_n^c) \leq e^{-n\varepsilon_n^2}.$$

STEP 3: For $B = \{\theta : d_n(\theta, \theta_0) \geq \varepsilon_n\}$:

$$P_{\theta_0}^n \int_B p_\theta^n / p_{\theta_0}^n \, d\Pi_n(\theta) 1_{K_n^c} \leq \int_B P_\theta^n(K_n^c) \, d\Pi_n(\theta) \leq \sup_{\theta \in B} P_\theta^n(K_n^c).$$

Rate theorem — refined

THEOREM

The rate of contraction is ε_n if there exist $\tilde{\Theta}_n \subset \Theta_n$ such that

(1) $D_n(\varepsilon_n, \tilde{\Theta}_n, d_n, \bar{d}_n) \leq n\varepsilon_n^2$ and $\dfrac{\Pi_n(\Theta_n - \tilde{\Theta}_n)}{\Pi_n(B_n(\theta_0, \varepsilon_n; k))} = o(e^{-2n\varepsilon_n^2})$.

(2) $\dfrac{\Pi_n(\theta \in \Theta_n : d_n(\theta, \theta_0) \leq 2j\varepsilon_n)}{\Pi_n(B_n(\theta_0, \varepsilon_n; k))} \leq e^{Kn\varepsilon_n^2 j^2/2} \qquad \forall j.$

Further trade-off between complexity and prior mass possible.

I.i.d. observations

Data $X_1, \ldots, X_n$, iid with density p_θ.

MAIN RESULT HOLDS WITH

- d_n Hellinger distance h (or L_1 or L_2)
- $B_n(\theta_0, \varepsilon; 2) = \{\theta : K(\theta_0, \theta) \leq \varepsilon^2, V_2(\theta_0, \theta) \leq \varepsilon^2\}$

$$h(\theta, \theta')^2 = \int \left(\sqrt{p_\theta} - \sqrt{p_{\theta'}}\right)^2 \, d\mu$$
$$K(\theta, \theta') = P_\theta \log(p_\theta / p_{\theta'})$$
$$V_2(\theta, \theta') = P_\theta \left(\log(p_\theta / p_{\theta'})\right)^2$$

Independent observations

Data $X_1, \ldots, X_n$, independent with $X_i \sim p_{\theta,i}$.

MAIN RESULT HOLDS WITH

- $d_n^2(\theta, \theta') = \frac{1}{n}\sum_{i=1}^n h_i(\theta, \theta')^2$
- $B_n(\theta_0, \varepsilon; 2) = \{\theta\colon \frac{1}{n}\sum_{i=1}^n K_i(\theta_0, \theta) \vee \frac{1}{n}\sum_{i=1}^n V_{2,i}(\theta_0, \theta) \leq \varepsilon^2\}$

h_i, K_i and $V_{2,i}$ computed for ith observation

Markov chains

Data $(X_0, X_1, \ldots, X_n)$ for $\cdots, X_0, X_1, X_2, \cdots$ stationary Markov chain with initial density q_θ and transition density $p_\theta(\cdot|\cdot)$.

Assume $\exists$ integrable r, constants $0 < c < C$ and $k > 2$:

1. $c\, r(y) \le p_\theta(y|x) \le C\, r(y)$,
2. α-mixing, $\sum_{h=0}^{\infty} \alpha_h^{1-1/k} < \infty$

MAIN RESULT HOLDS WITH

- $d_n^2(\theta, \theta') = \iint \left[\sqrt{p_\theta(y|x)} - \sqrt{p_{\theta'}(y|x)}\right]^2 d\mu(y)\, r(x)\, d\mu(x)$
- $B_n(\theta_0, \varepsilon; k) =$
  $\left\{\theta\colon P_{\theta_0} \log \frac{p_{\theta_0}}{p_\theta}(X_1|X_0) \le \varepsilon^2, P_{\theta_0} \left|\log \frac{p_{\theta_0}}{p_\theta}(X_1|X_0)\right|^k \le \varepsilon^k\right\}$

Gaussian time series

Data $(X_0, X_1, \ldots, X_n)$ for $\cdots, X_0, X_1, X_2, \cdots$ stationary mean zero Gaussian process with spectral density $\theta \in \Theta$.

Assume

1. $\sup_{\theta \in \Theta} \| \log \theta \|_\infty < \infty$
2. $\sup_{\theta \in \Theta} \sum_{h=-\infty}^{\infty} |h| (\mathrm{E}_\theta X_h X_0)^2 < \infty$

MAIN RESULT HOLDS WITH

- d_n: L_2-norm, $\bar{d}_n$: supremum-norm,
- $B_n(\theta_0, \varepsilon; 2)$: L_2-ball.

Ergodic diffusions

Data $(X_t : 0 \leq t \leq n)$ for X solution to $dX_t = \theta(X_t)\,dt + \sigma(X_t)\,dB_t$, where B is Brownian motion.

Assume

1. stationary ergodic, state space I,
2. stationary measure μ_{θ_0}.

MAIN RESULT HOLDS WITH

- $d(\theta, \theta') = \|(\theta - \theta')1_J/\sigma\|_{\mu_{\theta_0},2} \qquad J \subset I$
- $\bar{d}(\theta, \theta') = \|(\theta - \theta')/\sigma\|_{\mu_{\theta_0},2}$
- $B(\theta_0, \varepsilon; 2)$ $\|\cdot/\sigma\|_{\mu_{\theta_0},2}$-ball

Gaussian process priors

End Lecture 1

Lecture 2 (Thursday): Gaussian priors