Some Frequentist Results on Posterior Distributions on Infinite-dimensional Parameter Spaces

Aad van der Vaart
Vrije Universiteit Amsterdam

Le Cam lecture
Joint Statistical Meetings
Washington, 2009

Contents

Co-authors

Subhashis Ghosal

Harry van Zanten

PART I: Generalities

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

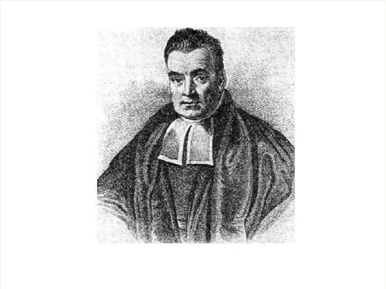
- A parameter Θ is generated according to a prior distribution Π.
- Given $\Theta = \theta$ the data X is generated according to a density p_θ.

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.

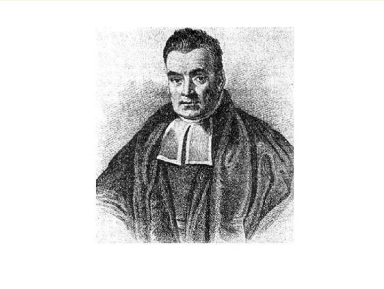
$$d\Pi(\theta|\, X) \propto p_\theta(X)\, d\Pi(\theta)$$

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

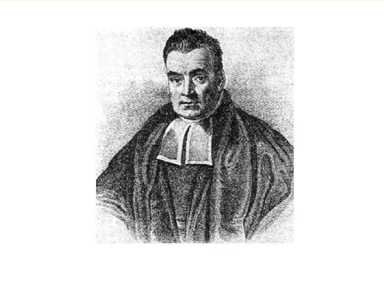


Thomas Bayes (1702–1761, 1763) followed this argument with Θ possessing the *uniform distribution* and X given $\Theta = \theta$ *binomial* (n, θ).

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

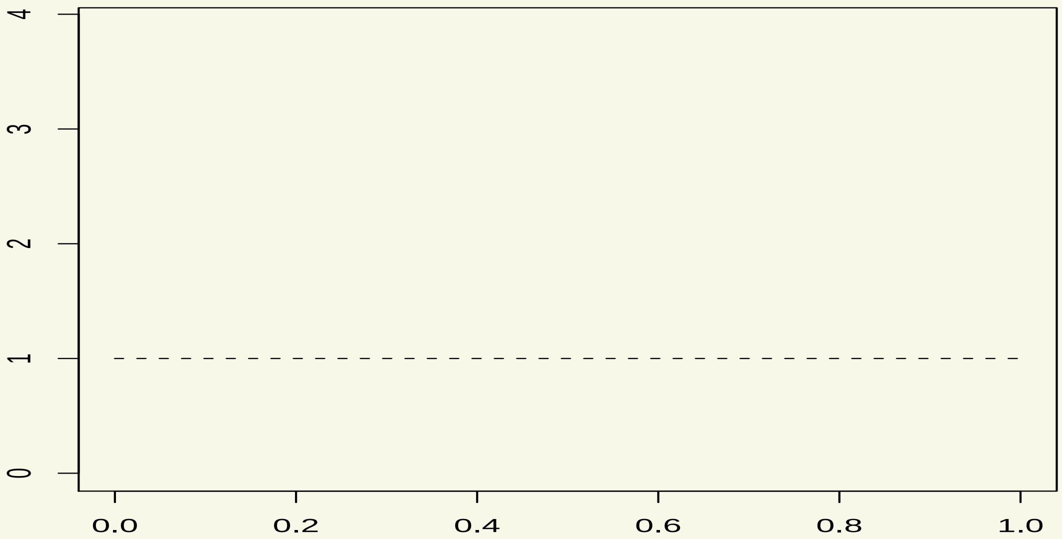
$$d\Pi_n(\theta|\, X) \propto \binom{n}{X} \theta^X (1-\theta)^{n-X} \cdot 1.$$

Reverend Thomas

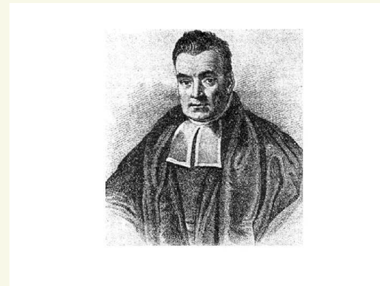


Thomas Bayes (1702–1761, 1763) followed this argument with Θ possessing the *uniform distribution* and X given $\Theta = \theta$ *binomial* (n, θ).

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

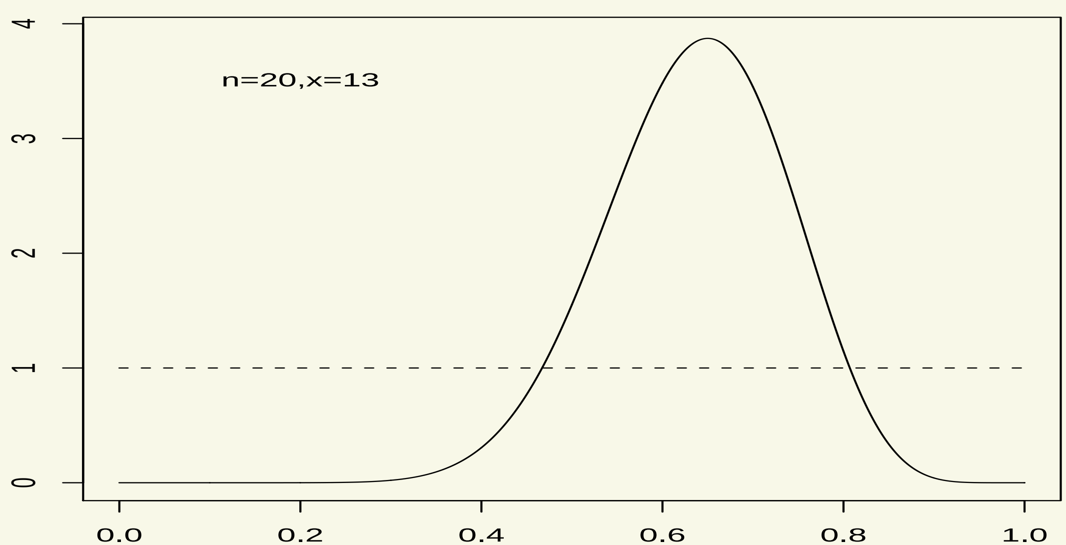


Reverend Thomas

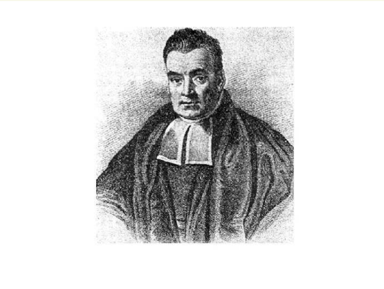


Thomas Bayes (1702–1761, 1763) followed this argument with Θ possessing the *uniform distribution* and X given $\Theta = \theta$ *binomial* (n, θ).

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

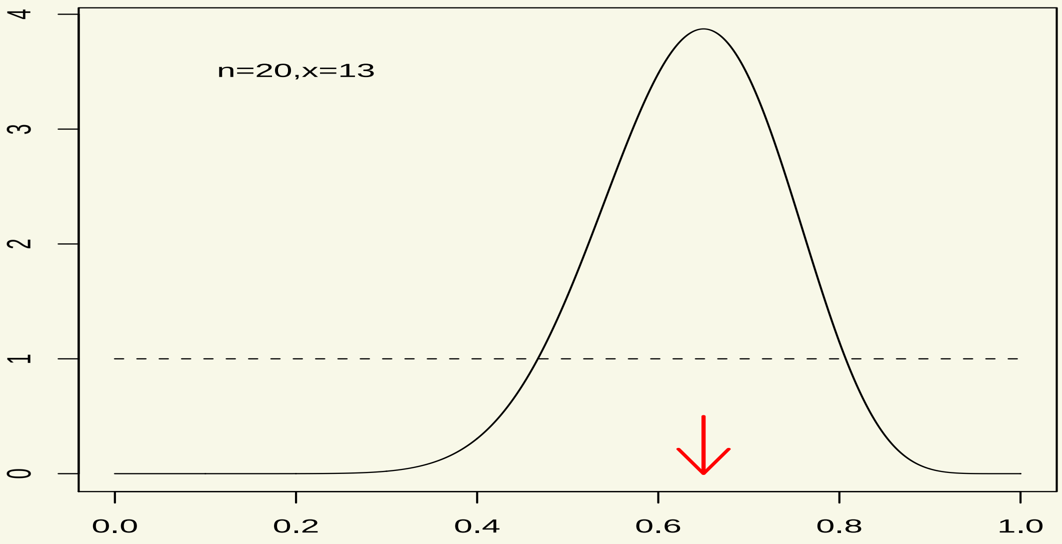


Reverend Thomas

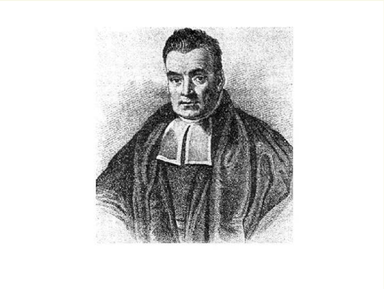


Thomas Bayes (1702–1761, 1763) followed this argument with Θ possessing the *uniform distribution* and X given $\Theta = \theta$ *binomial* (n, θ).

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

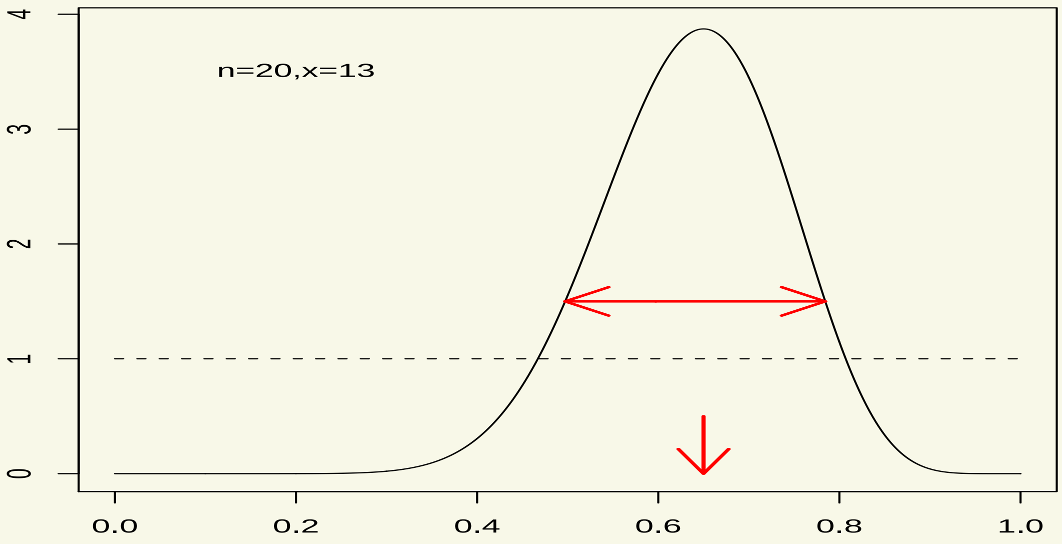


Reverend Thomas



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:

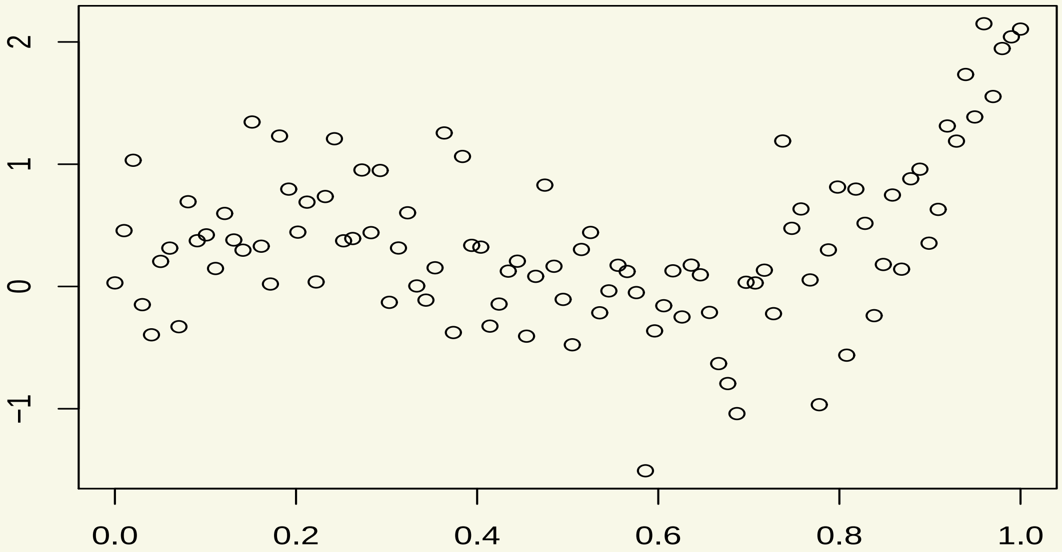
$$d\Pi(\theta|\, X) \propto p_\theta(X)\, d\Pi(\theta).$$

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.

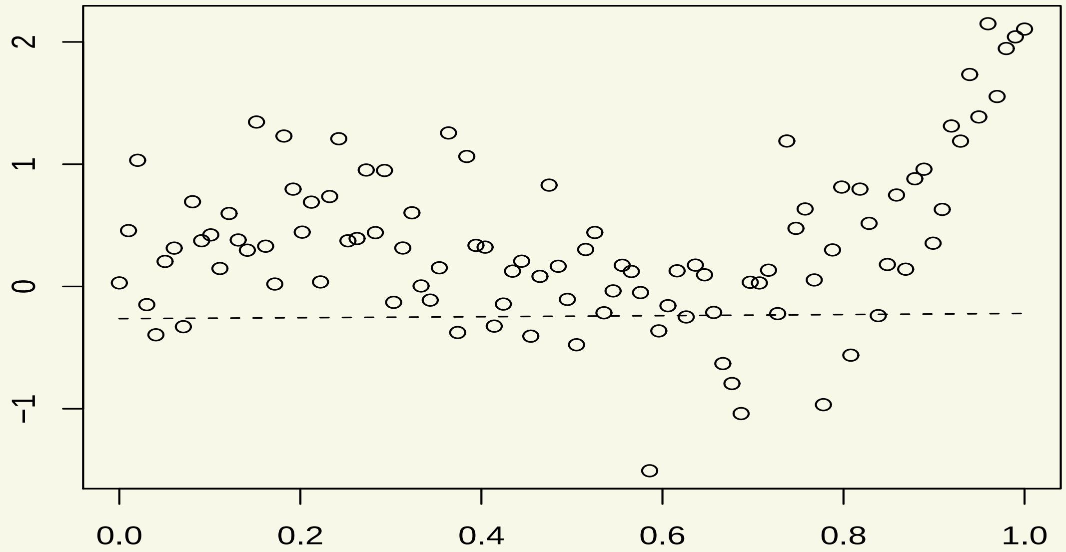


Nonparametric Bayes

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.



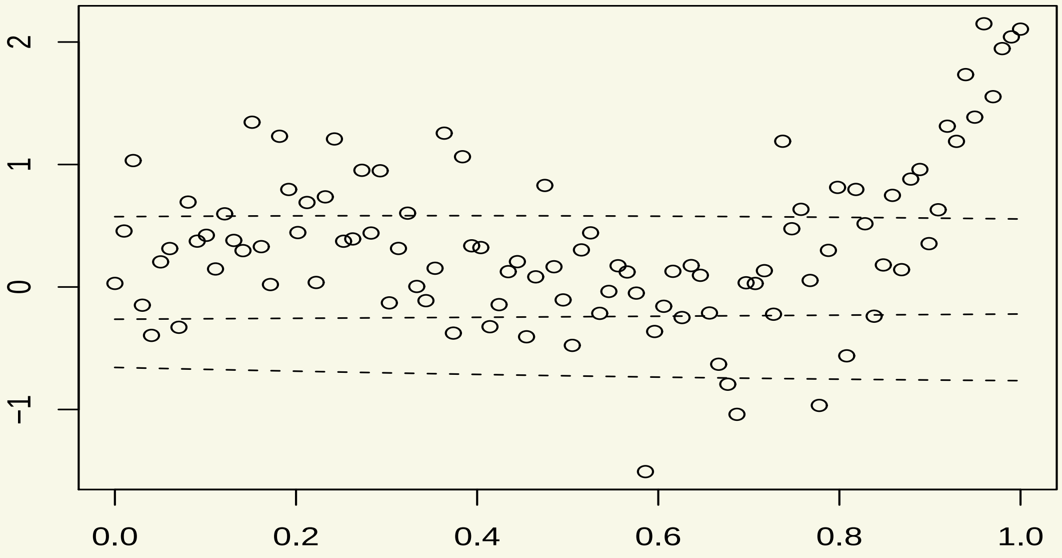
prior

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.



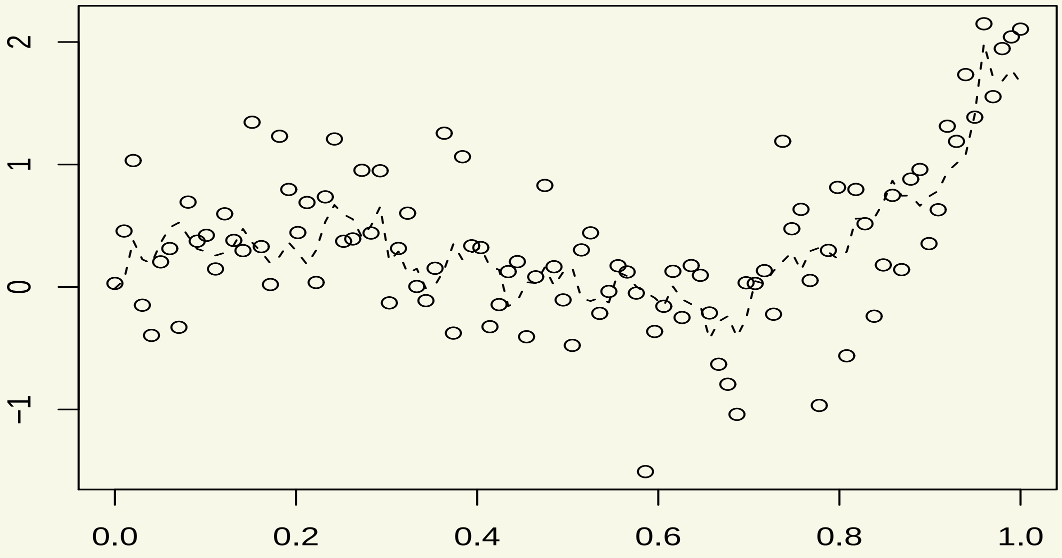
prior 3x

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.



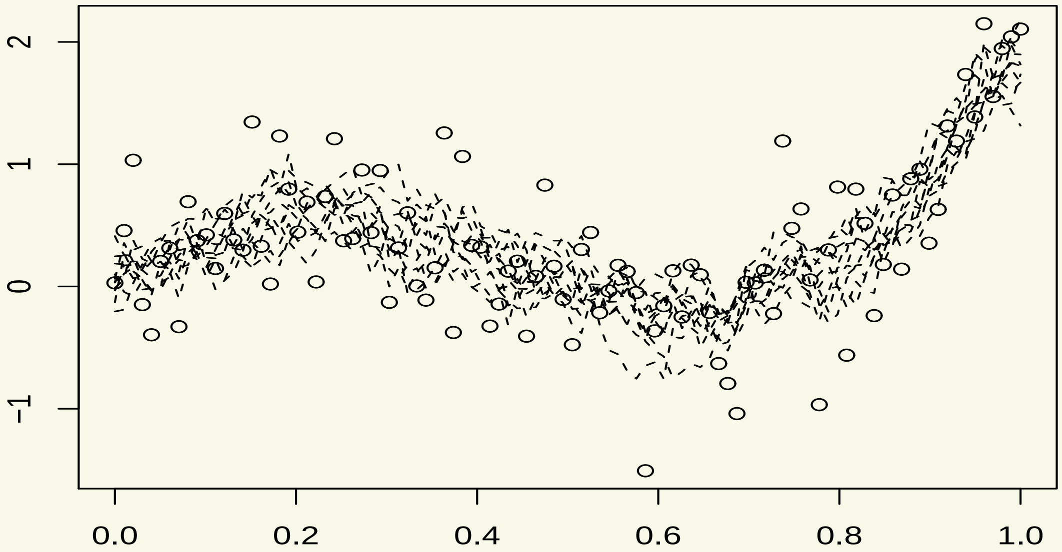
posterior

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.

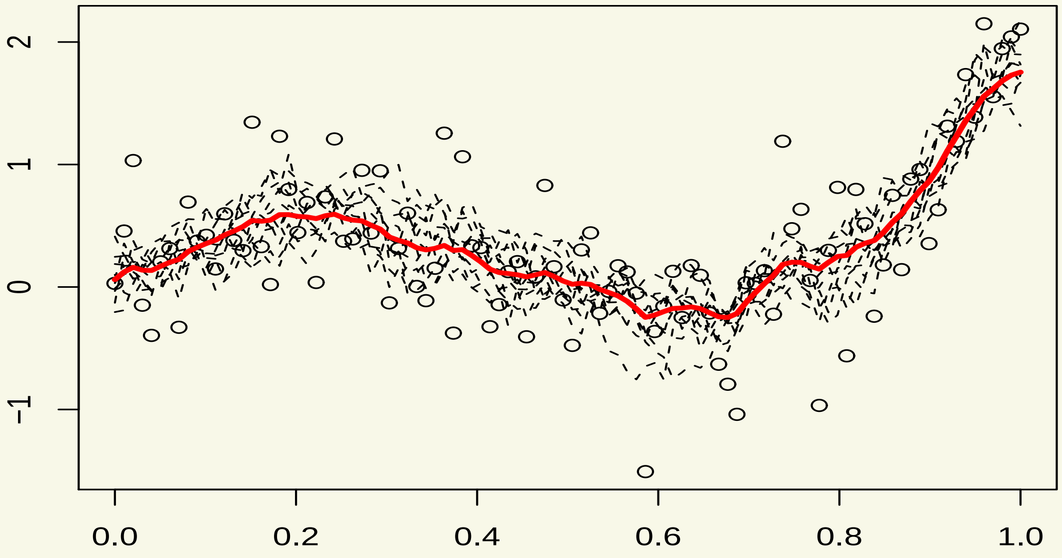


posterior 11x

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.



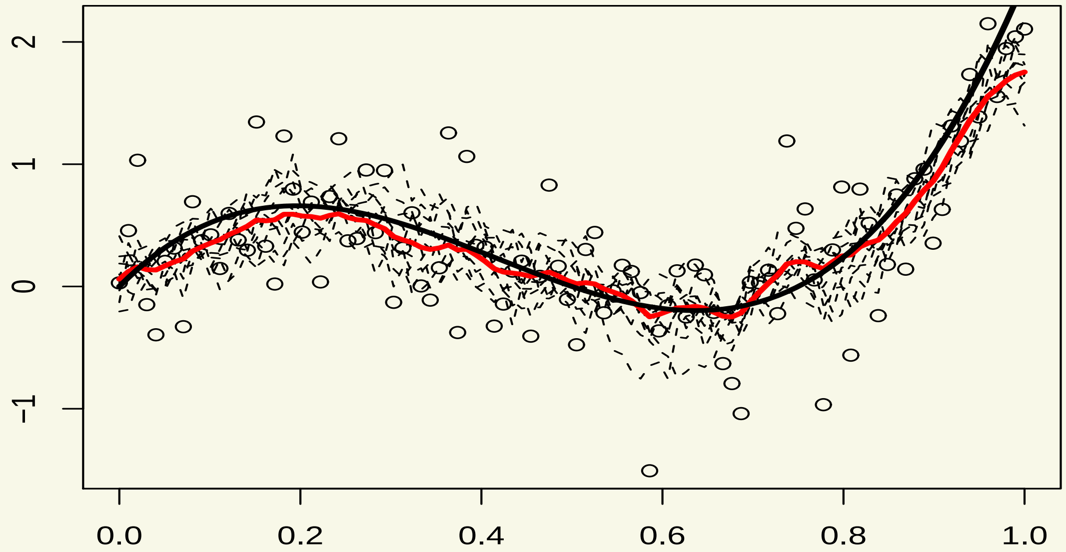
posterior mean

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.

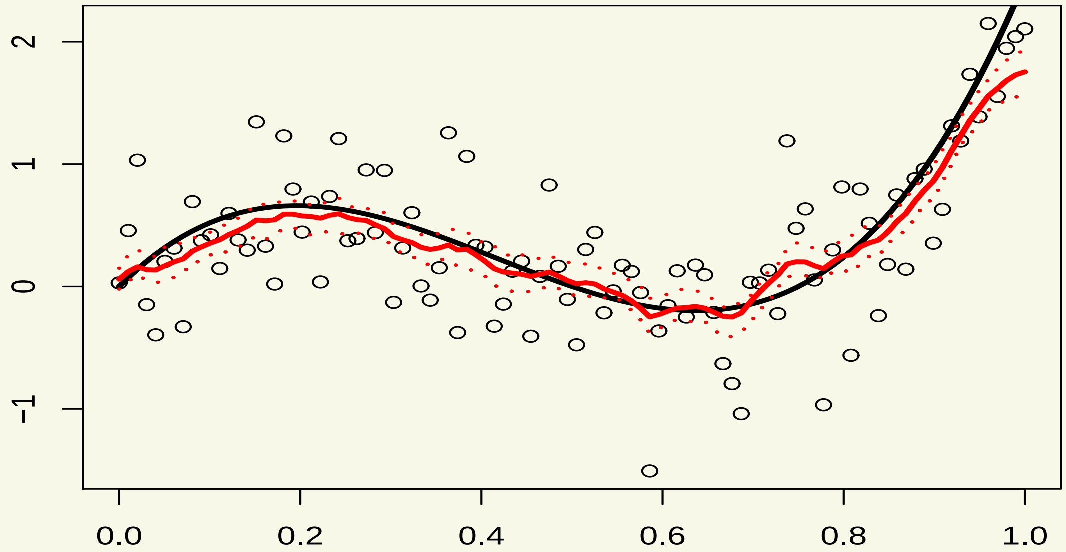


Nonparametric Bayes

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.



75 % pointwise central posterior regions

Computation

Analytical computation of a posterior is rarely possible, but clever algorithms allow to simulate from it (MCMC, ...), or compute the centre and spread (expectation propagation, Laplace expansion, ...).

Most research has focused on these algoritms.

In this talk we consider the properties of the posterior.

Frequentist Bayesian theory

Frequentist Bayesian

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

We like $\Pi(\theta \in \cdot \mid X)$ 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 $\Pi(\theta \in \cdot \mid X)$ as a *random measure* on the parameter set.

We like $\Pi(\theta \in \cdot \mid X)$ 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 $\Pi_n(\cdot \mid X^n)$ to contract to $\{\theta_0\}$, at a good rate.

Two desirable properties:

- Consistency + rate
- Adaptation

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

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

where $\tilde\theta_n$ is any efficient estimator of θ.

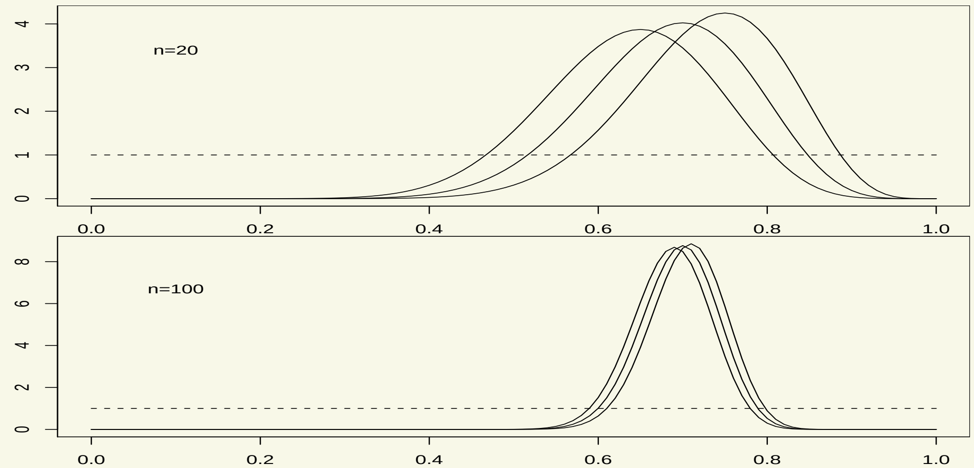
Parametric models

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, ...]
Under $P^n_{\theta_0}$-probability, for any prior with density that is positive around θ_0,

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

where $\tilde\theta_n$ is any efficient estimator of θ.



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, ...]
Under $P^n_{\theta_0}$-probability, for any prior with density that is positive around θ_0,

$$\Bigl\| \Pi_n(\cdot|\, X_1, \ldots, X_n) - N_d\bigl(\tilde\theta_n, \frac{1}{n} I_{\theta_0}^{-1}\bigr)(\cdot) \Bigr\| \to 0,$$

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

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

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. The Bayesian credible interval is a standard confidence interval.

The prior washes out completely.

Similar results for nonregular models and non-iid data.

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 $\theta \in \mathbb{R}^d$. (DQM with nonsingular Fisher information and existence of uniformly consistent tests of θ_0 versus $\{\theta\colon \|\theta - \theta_0\| > r\}$ suffice.)

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

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

where $\tilde\theta_n$ is any efficient estimator of θ.

Nonparametric and semiparametric models

For infinite-dimensional parameters the situation is very different.

Nonparametric and semiparametric models

For infinite-dimensional parameters the situation is very different.

- Most priors are inconsistent. [Freedman and Diaconis (1980s)]
- The rate of contraction often depends on the prior.
- For estimating a functional the prior is less critical, but still plays a role.

The prior does not (completely) wash out as $n \to \infty$.

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

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.

Minimaxity and adaptation

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

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

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

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

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

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

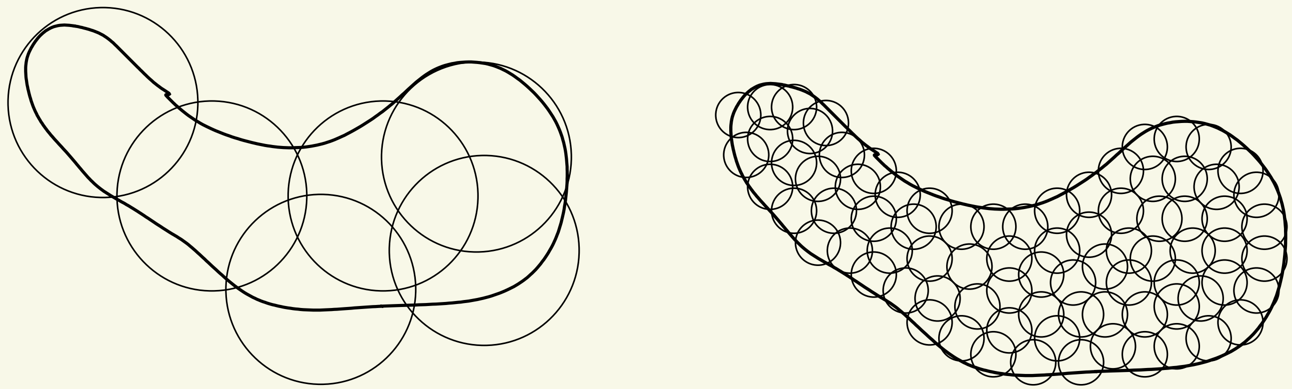
Given a scale of regularity classes $(\Theta_\alpha \colon \alpha \in A)$, we like the posterior to adapt: if the true parameter belongs to Θ_α, then we like the contraction rate to be the minimax rate for the α-class.

For instance, in typical examples $n^{-\alpha/(2\alpha+d)}$ if Θ_α is a set of functions of d arguments with partial derivatives of order α bounded by a constant (i.e. regularity α/d).

Rates

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 $\qquad$ ε small

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

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.

h is the Hellinger distance : $h^2(p,q) = \int (\sqrt{p} - \sqrt{q})^2\, d\mu$.

$B_{KL}(p_0, \varepsilon)$ is a 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.

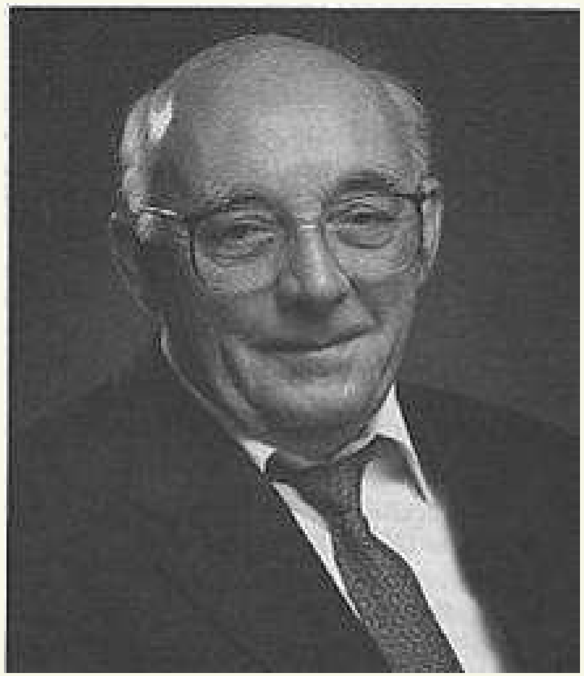
(2) $\Pi(B_{KL}(p_0, \varepsilon_n)) \geq e^{-n\varepsilon_n^2}$. prior mass.

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

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

The revised form is much improved, although probably not the final word on the subject [...] However it is awfully long, and minor corrections would make it easier to read. (Anonymous referee of Ghosal, Ghosh, vdVaart (2000).)

The revised form is much improved, although probably not the final word on the subject [...] However it is awfully long, and minor corrections would make it easier to read. (Anonymous referee of Ghosal, Ghosh, vdVaart (2000).)



Entropy

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.

Entropy

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

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

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.

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

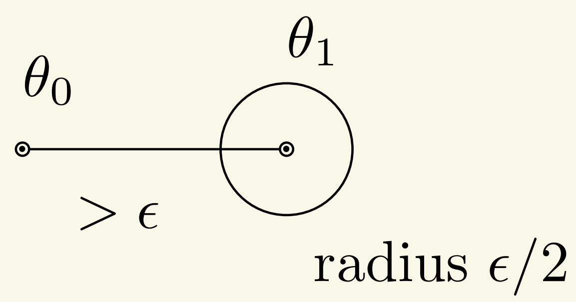
Its root is in the testing condition of Le Cam (1964).

Le Cam's testing criterion

Data X^n following statistical model $(P_\theta^n: \theta \in \Theta_n)$, metric space (Θ_n, d_n).

Assume for all $e > 0$: for all θ_1 with $d_n(\theta_1, \theta_0) > \varepsilon$ $\exists$ test ϕ_n with

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

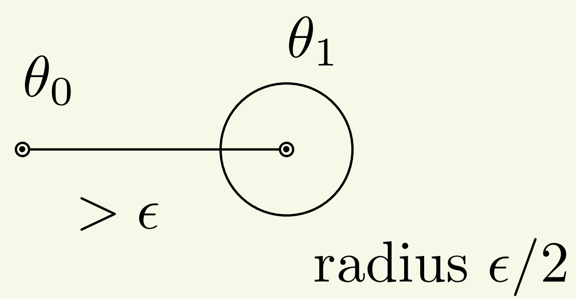


Le Cam's testing criterion

Data X^n following statistical model $(P_\theta^n : \theta \in \Theta_n)$, metric space (Θ_n, d_n).

Assume for all $e > 0$: for all θ_1 with $d_n(\theta_1, \theta_0) > \varepsilon$ $\exists$ test ϕ_n with

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

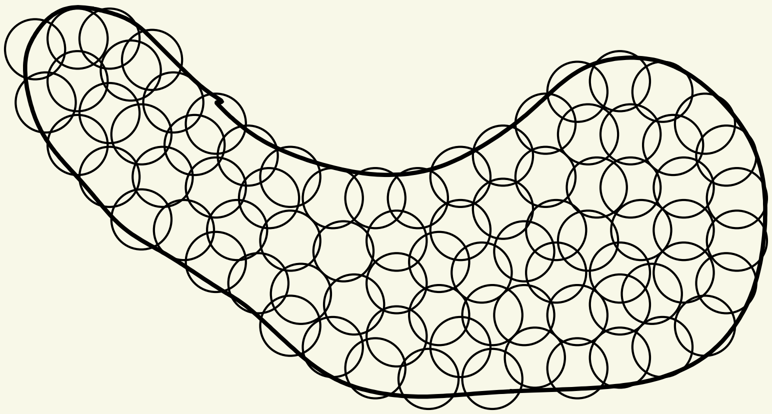


THEOREM [Le Cam (73,75,86), Birgé (83,06)]
There exist estimators $\hat{\theta}_n$ with $d_n(\hat{\theta}_n, \theta_0) = O_P(\varepsilon_n)$ if

$$\log N(\varepsilon_n, \Theta_n, d_n) \le n\varepsilon_n^2.$$

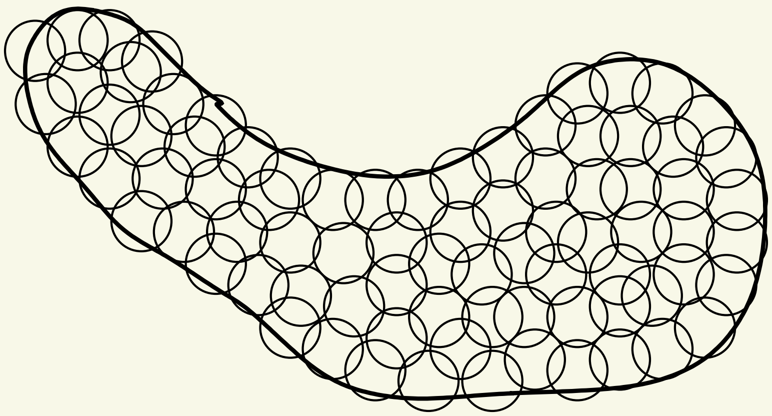
Le Cam dimension = local entropy

Instead of entropy $\log N(\varepsilon, \Theta_n, d_n)$



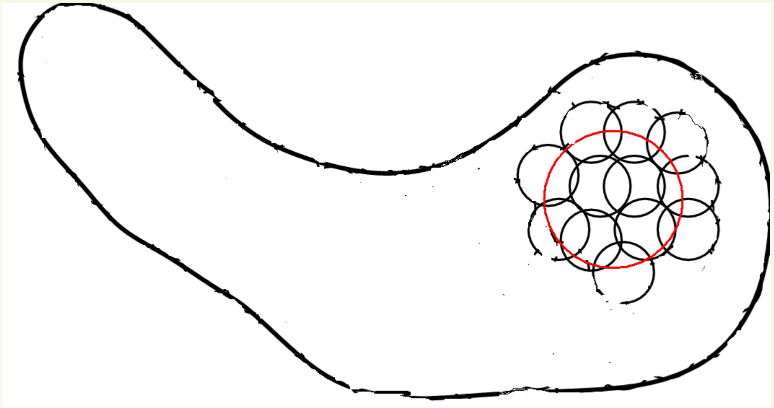
Le Cam dimension = local entropy

Instead of entropy $\log N(\varepsilon, \Theta_n, d_n)$



we can use Le Cam dimension:

$$D_n(\varepsilon, \Theta, d_n) = \sup_{\eta > \varepsilon} \log N\Big(\frac{\eta}{2}, \{\theta \in \Theta_n : d_n(\theta, \theta_0) \le \eta\}, d_n\Big).$$



Rate theorem — general

Given data X^n following P_θ^n from a model $(P_\theta^n\colon \theta \in \Theta_n)$ that satisfies Le Cam's testing criterion, and a prior Π, form posterior

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

THEOREM

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

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

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

$B_n(\theta_0, \varepsilon; k)$ is Kullback-Leibler neighbourhood of $p_{\theta_0}^n$.

The theorem can be refined in various ways. For instance, only relative prior masses matter; a further trade-off between complexity and prior mass is possible.

Settings

- iid observations (Hellinger).
- independent observations (root average square Hellinger).
- Markov chains (Hellinger transition density).
- Gaussian time series (L_2-spectral density).
- ergodic diffusions (L_2-drift/root diffusion).
- ...

Examples

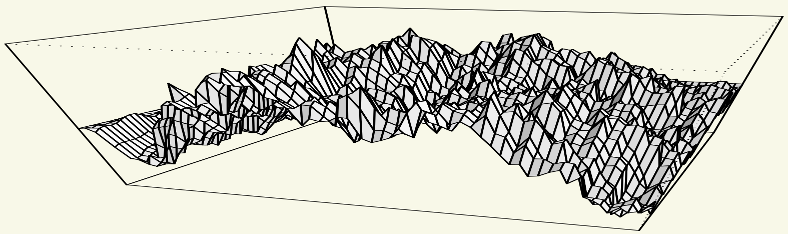
- Dirichlet mixtures of normals.
- Discrete priors.
- Mixtures of betas.
- Series priors (splines, Fourier, wavelets, ...).
- Independent increment process priors.
- Sparse priors.
- ....
- ....
- Gaussian process priors.

PART II: Gaussian process priors

Examples

Gaussian process

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



Gaussian processes have been found useful, because

- they offer great variety.
- they have a general index set T.
- 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 $[0,1]$:

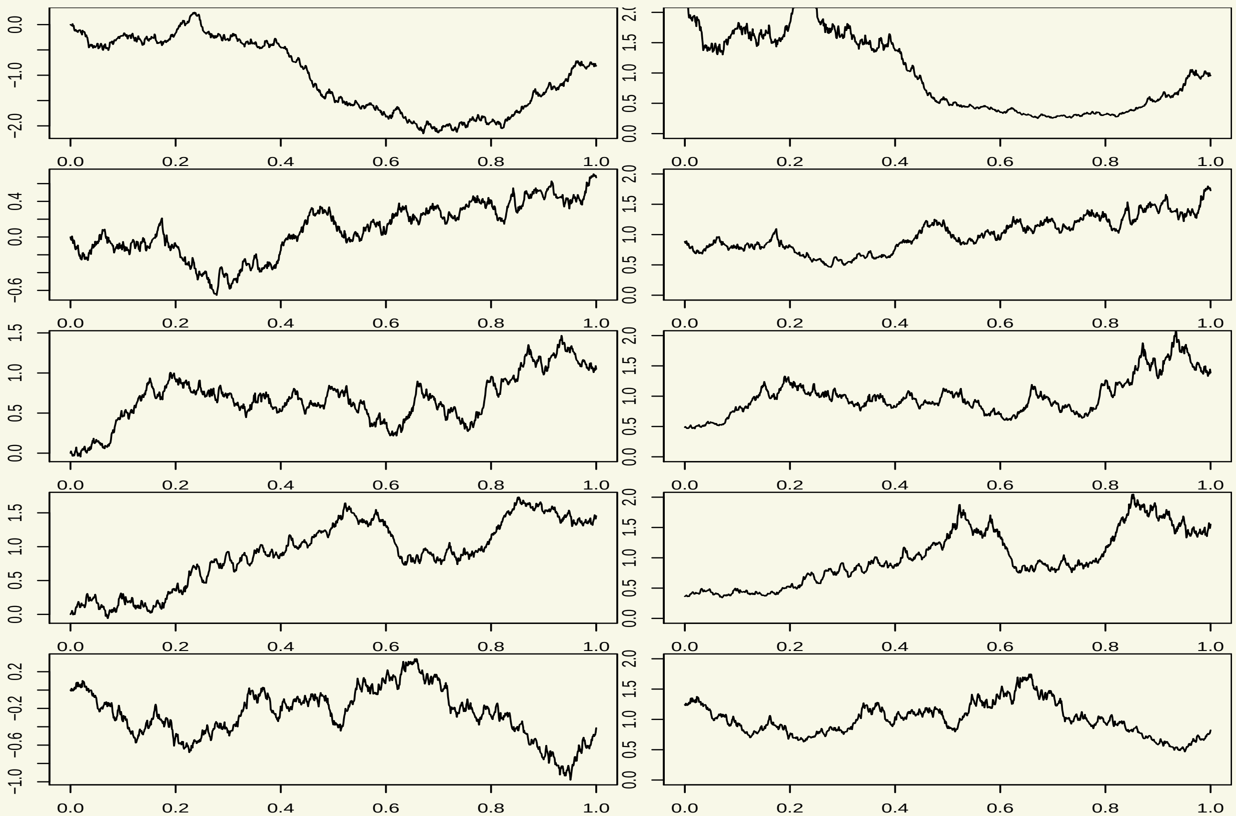
$$x \mapsto \frac{e^{W_x}}{\int_0^1 e^{W_y}\,dy}.$$

[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 $\alpha \leq 1/2$.

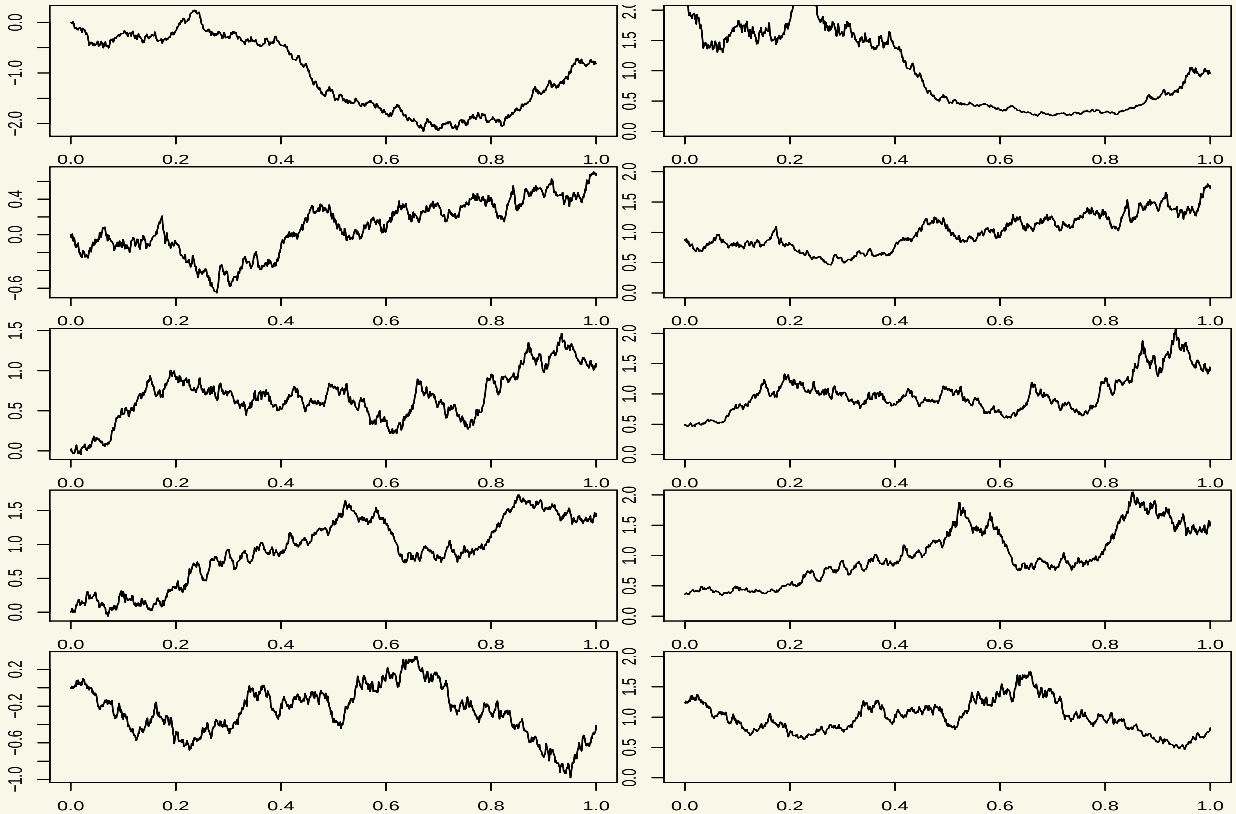
- This is optimal if and only if $\alpha = 1/2$.
- Rate does not improve if α increases from $1/2$.
- Consistency for any $\alpha > 0$.

(Lower bound: Castillo (2008).)

Brownian density estimation

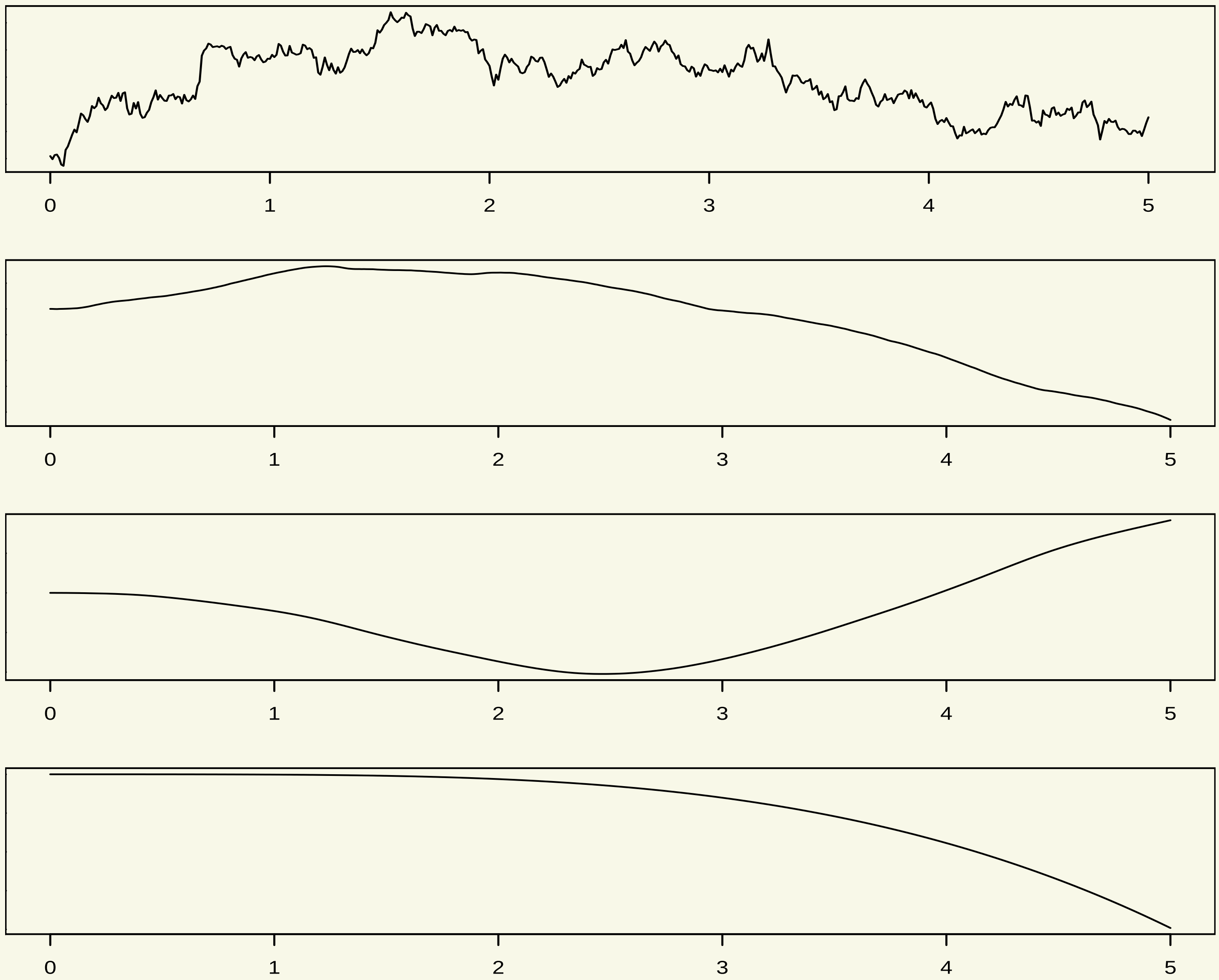
For W Brownian motion use as prior on a density p on $[0,1]$:

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



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

Integrated Brownian motion



0, 1, 2 and 3 times integrated Brownian motion

Integrated Brownian motion: Riemann-Liouville process

$(\alpha - 1/2)$-times integrated Brownian motion, released at 0

$$W_t = \int_0^t (t-s)^{\alpha-1/2}\, dB_s + \sum_{k=0}^{[\alpha]+1} Z_k t^k.$$

[B Brownian motion, $\alpha > 0$, (Z_k) iid $N(0,1)$, “fractional integral”]

THEOREM
IBM gives appropriate model for α-smooth functions: consistency for any true smoothness $\beta > 0$, but the optimal $n^{-\beta/(2\beta+1)}$ if and only if $\alpha = \beta$.

(Kimeldorf & Wahba (1970s) showed that the posterior mean for this prior on a regression function is (asymptotically) a regression spline.)

Brownian sheet

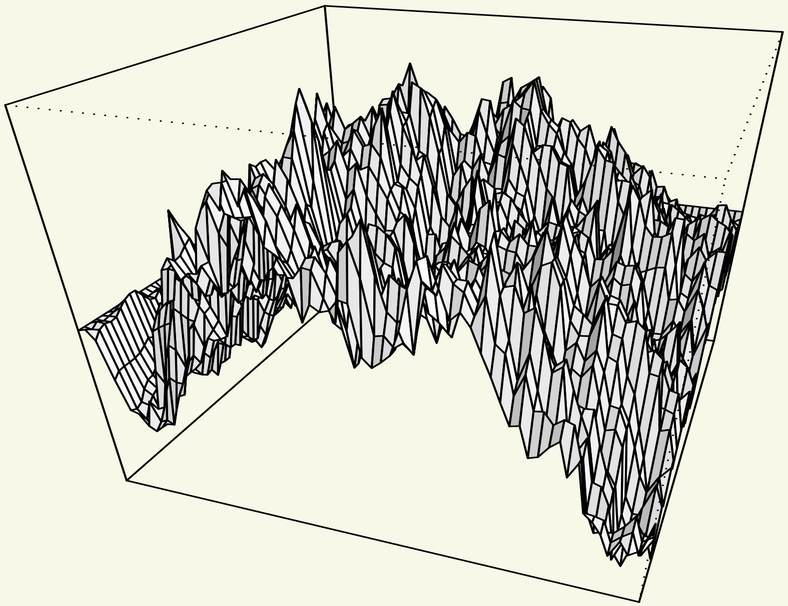
Brownian sheet $(W_t\colon t \in [0,1]^d)$ has covariance function

$$\operatorname{cov}(W_s, W_t) = (s_1 \wedge t_1) \cdots (s_d \wedge t_d).$$

BS gives rates of the order

$$n^{-1/4}(\log n)^{(2d-1)/4}$$

for sufficiently smooth w_0 $(\alpha \geq d/2)$.

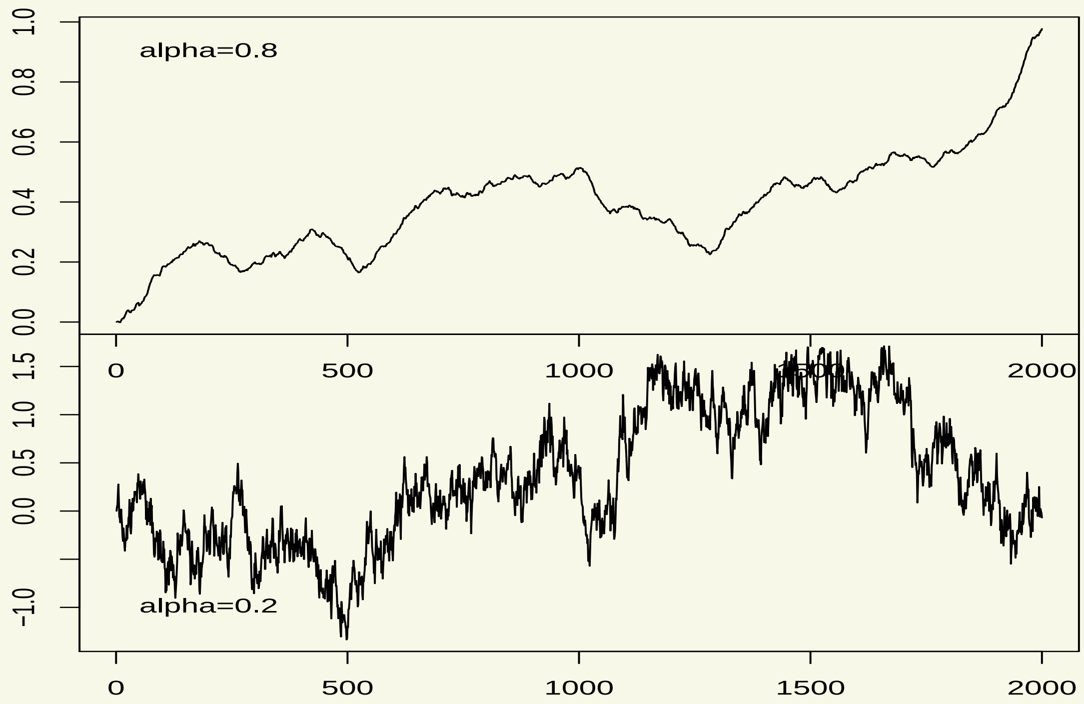


Fractional Brownian motion

W zero-mean Gaussian with (Hurst index $0 < \alpha < 1$)

$$\mathrm{cov}(W_s, W_t) = s^{2\alpha} + t^{2\alpha} - |t - s|^{2\alpha}.$$

fBM is appropriate model for α-smooth functions. Integrate to cover $\alpha > 1$.



Series priors

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

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

For instance: $\theta_1, \theta_2, \ldots$ independent with $\theta_i \sim N(0, \sigma_i^2)$.

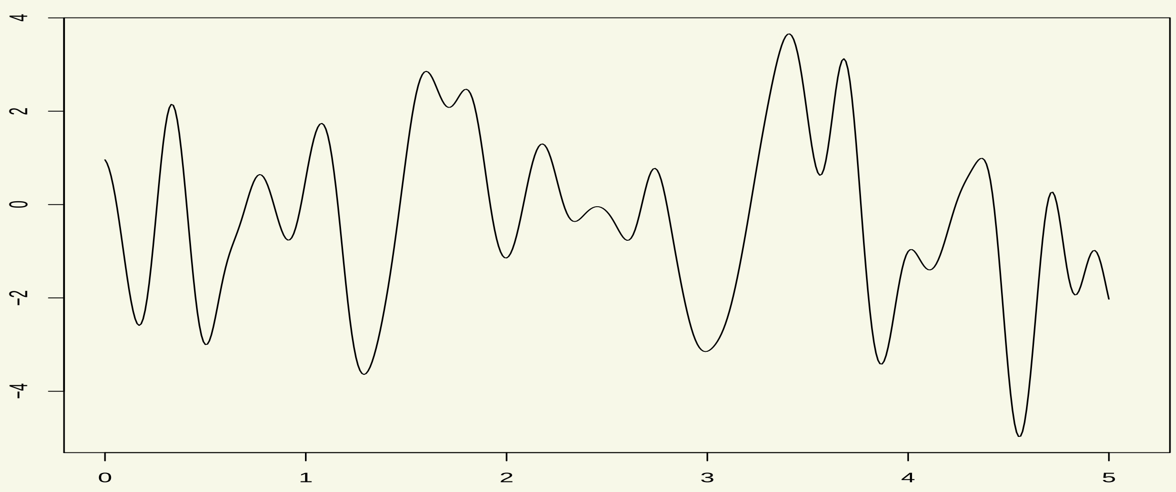
Appropriate decay of σ_i gives proper model for α-smooth functions.(E.g. with wavelets put fixed, equal prior variance on levels up to usual truncation level.)

Stationary processes

A stationary Gaussian field $(W_t: t \in \mathbb{R}^d)$ is characterized through a spectral measure μ, by

$$\mathrm{cov}(W_s, W_t) = \int e^{i\lambda^T(s-t)}\, d\mu(\lambda).$$

Smoothness of $t \mapsto W_t$ is controlled by the tails of μ. For instance, exponentially small tails give infinitely smooth sample paths; Matérn gives α-regular functions.



Stationary processes

A stationary Gaussian field $(W_t \colon t \in \mathbb{R}^d)$ is characterized through a spectral measure μ, by

$$\operatorname{cov}(W_s, W_t) = \int e^{i\lambda^T(s-t)}\, d\mu(\lambda).$$

Smoothness of $t \mapsto W_t$ is controlled by the tails of μ. For instance, exponentially small tails give infinitely smooth sample paths; Matérn gives α-regular functions.

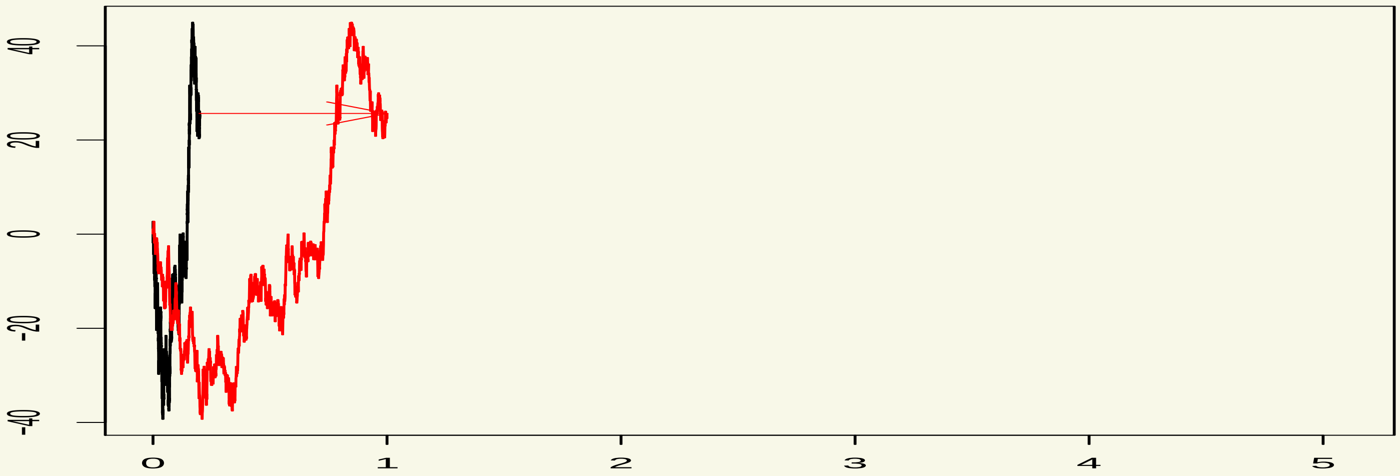
THEOREM If $\int e^{\|\lambda\|} |\hat{w}_0(\lambda)|^2\, d\lambda < \infty$, then the Gaussian spectral measure gives a near $1/\sqrt{n}$-rate of contraction; it gives consistency but suboptimal rates for Hölder smooth functions.

Conjecture: Matérn gives good results for Sobolev spaces.

Rescaling

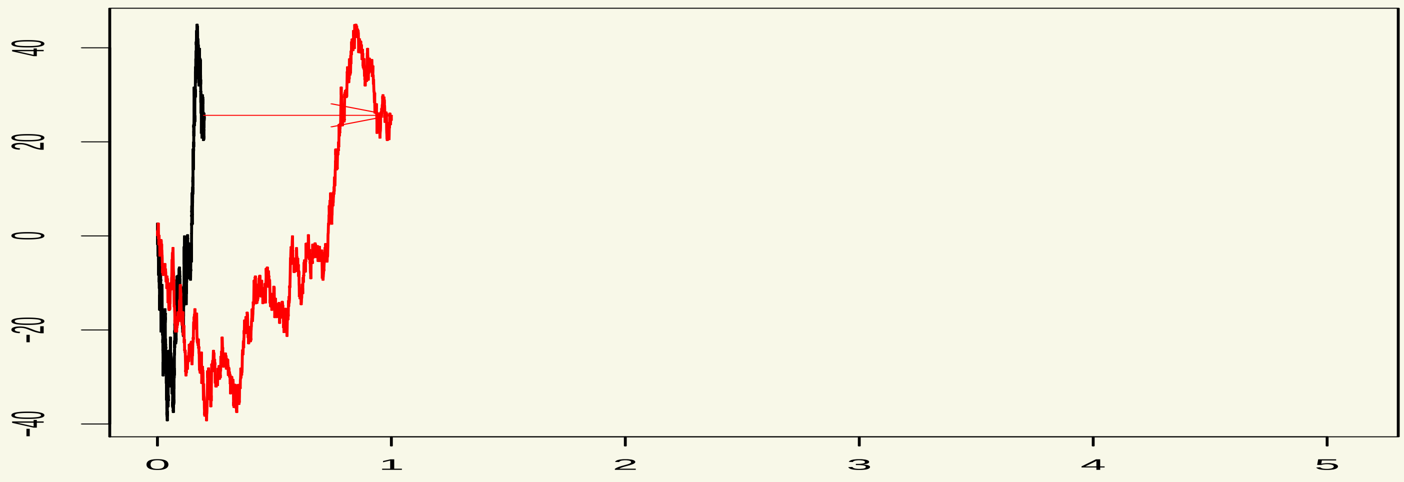
Stretching or shrinking

Sample paths can be smoothed by stretching

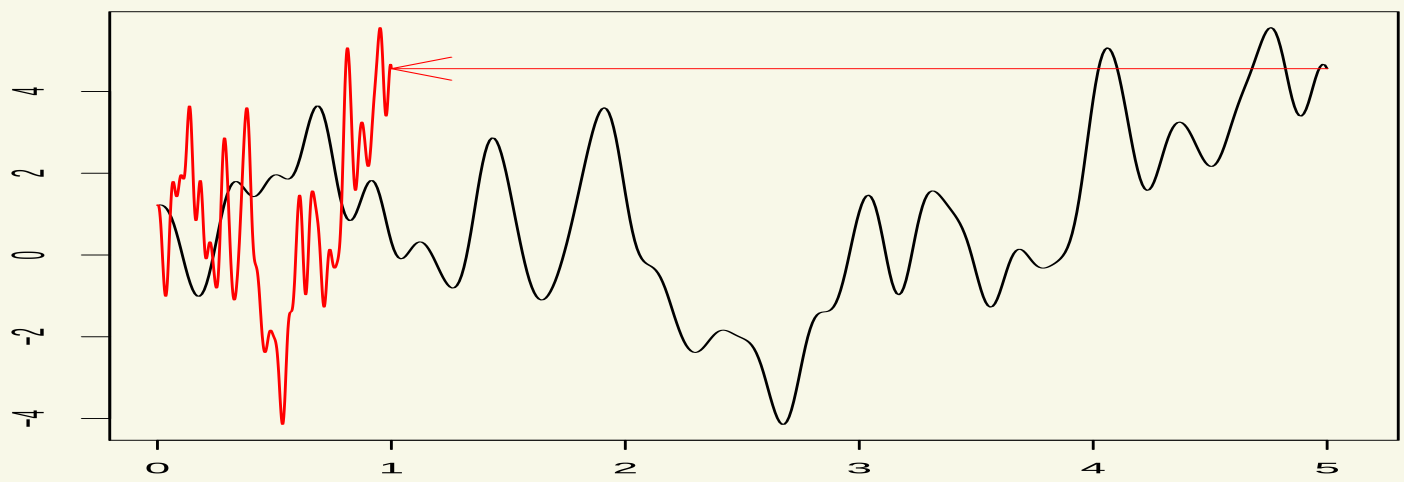


Stretching or shrinking

Sample paths can be smoothed by stretching



and roughened by shrinking



Rescaled Brownian motion

$W_t = B_{t/c_n}$ for B Brownian motion, and $c_n \sim n^{(2\alpha-1)/(2\alpha+1)}$

- $\alpha < 1/2$: $c_n \to 0$ (shrink).
- $\alpha \in (1/2, 1]$: $c_n \to \infty$ (stretch).

THEOREM
The prior $W_t = B_{t/c_n}$ gives optimal rate for $w_0 \in C^\alpha[0,1]$, $\alpha \in (0,1]$.

Surprising? (Brownian motion is self-similar!.)

Rescaled Brownian motion

$W_t = B_{t/c_n}$ for B Brownian motion, and $c_n \sim n^{(2\alpha-1)/(2\alpha+1)}$

- $\alpha < 1/2$: $c_n \to 0$ (shrink).
- $\alpha \in (1/2, 1]$: $c_n \to \infty$ (stretch).

THEOREM
The prior $W_t = B_{t/c_n}$ gives optimal rate for $w_0 \in C^\alpha[0,1]$, $\alpha \in (0,1]$.

Surprising? (Brownian motion is self-similar!.)

Appropriate rescaling of k times integrated Brownian motion gives optimal prior for every $\alpha \in (0, k+1]$.

Rescaled smooth stationary process

A Gaussian field with infinitely-smooth sample paths is obtained for

$$\mathrm{E}G_s G_t = \exp(-\|s-t\|^2).$$

THEOREM
The prior $W_t = G_{t/c_n}$ for $c_n \sim n^{-1/(2\alpha+d)}$ gives nearly optimal rate for $w_0 \in C^\alpha[0,1]$, any $\alpha > 0$.

Messages

- Scaling changes the properties of the prior.
- Hyper parameters are important.

A smooth prior process can be scaled to achieve any desired level of "prior roughness", but a rough process cannot be smoothed much and will necessarily impose its roughness on the data.

Adaptation

Hierarchical priors

For each $\alpha > 0$ there are several priors Π_α (Riemann-Liouville, Fractional, Series, Matérn, rescaled processes,...) that are appropriate for estimating α-smooth functions.

We can combine them into a mixture prior:

- Put a prior weight $d\rho(\alpha)$ on α.
- Given α use an optimal prior Π_α for that α.

This works (nearly), provided ρ is chosen with some (but not much) care.

The weights $d\rho(\alpha) \propto e^{-n\varepsilon_{n,\alpha}^2}\, d\alpha$ always work.

[Lember, Szabo]

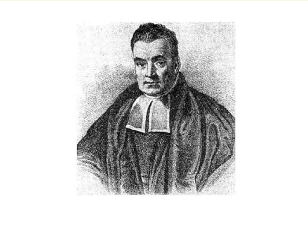
Adaptation by rescaling

- Choose A^d from a Gamma distribution.
- Choose $(G_t: t > 0)$ centered Gaussian with $\mathrm{E}G_sG_t = \exp\bigl(-\|s-t\|^2\bigr)$.
- Set $W_t \sim G_{At}$.

THEOREM

- if $w_0 \in C^\alpha[0,1]^d$, then the rate of contraction is nearly $n^{-\alpha/(2\alpha+d)}$.
- if w_0 is supersmooth, then the rate is nearly $n^{-1/2}$.

Reverend Thomas solved the bandwidth problem!?



General formulation of rates

Two ingredients

Two ingredients:

- RKHS
- Small ball exponent

Reproducing kernel Hilbert space

Think of the Gaussian process as a random element in a Banach space $(\mathbb{B}, \|\cdot\|)$.

To every such Gaussian random element is attached a certain Hilbert space $(\mathbb{H}, \|\cdot\|_{\mathbb{H}})$, called the RKHS.

$\|\cdot\|_{\mathbb{H}}$ is stronger than $\|\cdot\|$ and hence can consider $\mathbb{H} \subset \mathbb{B}$.

Reproducing kernel Hilbert space

Think of the Gaussian process as a random element in a Banach space $(\mathbb{B}, \|\cdot\|)$.

To every such Gaussian random element is attached a certain Hilbert space $(\mathbb{H}, \|\cdot\|_{\mathbb{H}})$, called the RKHS.

$\|\cdot\|_{\mathbb{H}}$ is stronger than $\|\cdot\|$ and hence can consider $\mathbb{H} \subset \mathbb{B}$.

DEFINITION
For $S\colon \mathbb{B}^* \to \mathbb{B}$ defined by

$$Sb^* = \mathrm{E} W b^*(W),$$

the RKHS is the completion of $S\mathbb{B}^*$ under

$$\langle Sb_1^*, Sb_2^* \rangle_{\mathbb{H}} = \mathrm{E} b_1^*(W) b_2^*(W).$$

Reproducing kernel Hilbert space

Think of the Gaussian process as a random element in a Banach space $(\mathbb{B}, \|\cdot\|)$.

To every such Gaussian random element is attached a certain Hilbert space $(\mathbb{H}, \|\cdot\|_{\mathbb{H}})$, called the RKHS.

$\|\cdot\|_{\mathbb{H}}$ is stronger than $\|\cdot\|$ and hence can consider $\mathbb{H} \subset \mathbb{B}$.

DEFINITION

For a process $W = (W_x : x \in \mathcal{X})$ with bounded sample paths and covariance function $K(x, y) = \mathrm{E} W_x W_y$, the RKHS is the completion of the set of functions

$$x \mapsto \sum_i \alpha_i K(y_i, x),$$

under

$$\Big\langle \sum_i \alpha_i K(y_i, \cdot), \sum_j \beta_j K(z_j, \cdot) \Big\rangle_{\mathbb{H}} = \sum_i \sum_j \alpha_i \beta_j K(y_i, z_j).$$

Reproducing kernel Hilbert space

Think of the Gaussian process as a random element in a Banach space $(\mathbb{B}, \|\cdot\|)$.

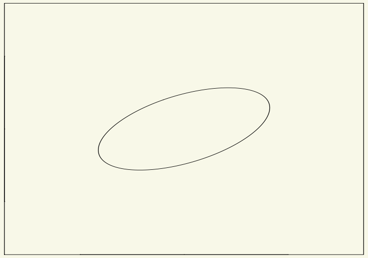
To every such Gaussian random element is attached a certain Hilbert space $(\mathbb{H}, \|\cdot\|_{\mathbb{H}})$, called the RKHS.

$\|\cdot\|_{\mathbb{H}}$ is stronger than $\|\cdot\|$ and hence can consider $\mathbb{H} \subset \mathbb{B}$.

EXAMPLE

If W is multivariate normal $N_d(0, \Sigma)$, then the RKHS is $\mathbb{R}^d$ with norm

$$\|h\|_{\mathbb{H}} = \sqrt{h^t \Sigma^{-1} h}$$



Reproducing kernel Hilbert space

Think of the Gaussian process as a random element in a Banach space $(\mathbb{B}, \|\cdot\|)$.

To every such Gaussian random element is attached a certain Hilbert space $(\mathbb{H}, \|\cdot\|_{\mathbb{H}})$, called the RKHS.

$\|\cdot\|_{\mathbb{H}}$ is stronger than $\|\cdot\|$ and hence can consider $\mathbb{H} \subset \mathbb{B}$.

EXAMPLE

Any W can be represented as

$$W = \sum_{i=1}^{\infty} \mu_i Z_i e_i,$$

for numbers $\mu_i \downarrow 0$, iid standard normal $Z_1, Z_2, \ldots$, and $e_1, e_2, \ldots \in \mathbb{B}$ with $\|e_1\| = \|e_2\| = \cdots = 1$. The RKHS consists of all $h := \sum_i h_i e_i$ with

$$\|h\|_{\mathbb{H}}^2 := \sum_i \frac{h_i^2}{\mu_i^2} < \infty.$$

Reproducing kernel Hilbert space

Think of the Gaussian process as a random element in a Banach space $(\mathbb{B}, \|\cdot\|)$.

To every such Gaussian random element is attached a certain Hilbert space $(\mathbb{H}, \|\cdot\|_{\mathbb{H}})$, called the RKHS.

$\|\cdot\|_{\mathbb{H}}$ is stronger than $\|\cdot\|$ and hence can consider $\mathbb{H} \subset \mathbb{B}$.

EXAMPLE
Brownian motion is a random element in $C[0,1]$.
Its RKHS is $\mathbb{H} = \{h\colon \int h'(t)^2\,dt < \infty\}$ with norm $\|h\|_{\mathbb{H}} = \|h'\|_2$.

Small ball probability

The small ball probability of a Gaussian random element W in $(\mathbb{B}, \|\cdot\|)$ is

$$\mathrm{P}(\|W\| < \varepsilon),$$

and the small ball exponent is

$$\phi_0(\varepsilon) = -\log \mathrm{P}(\|W\| < \varepsilon).$$

Small ball probability

The small ball probability of a Gaussian random element W in $(\mathbb{B}, \|\cdot\|)$ is

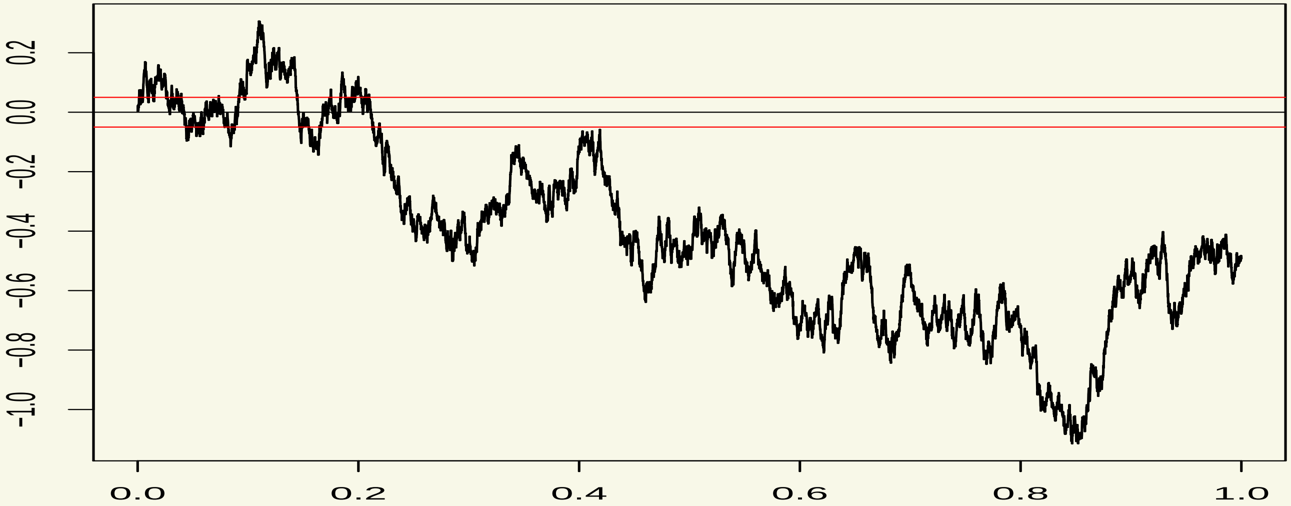
$$\mathrm{P}(\|W\| < \varepsilon),$$

and the small ball exponent is

$$\phi_0(\varepsilon) = -\log \mathrm{P}(\|W\| < \varepsilon).$$

EXAMPLE

For Brownian motion $\phi_0(\varepsilon) \asymp (1/\varepsilon)^2$ as $\varepsilon \downarrow 0$.

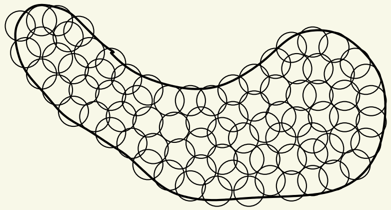


Small ball probability

Small ball probabilities can be computed either by probabilistic arguments, or analytically from the RKHS.

Small ball probability

Small ball probabilities can be computed either by probabilistic arguments, or analytically from the RKHS.



$N(\varepsilon, B, d) = \#\ \varepsilon$-balls

THEOREM [Kuelbs & Li (93)]

For $\mathbb{H}_1$ the unit ball of the RKHS (up to constants),

$$\phi_0(\varepsilon) \asymp \log N\Bigl(\frac{\varepsilon}{\sqrt{\phi_0(\varepsilon)}}, \mathbb{H}_1, \|\cdot\|\Bigr).$$

There is a big literature on small ball probabilities. (In July 2009 243 entries in database maintained by Michael Lifshits.)

Rates for Gaussian priors

Prior W is Gaussian map in $(\mathbb{B}, \|\cdot\|)$ with RKHS $(\mathbb{H}, \|\cdot\|_{\mathbb{H}})$ and small ball exponent $\phi_0(\varepsilon) = -\log \mathrm{P}(\|W\| < \varepsilon)$.

THEOREM
If statistical distances on the model combine appropriately with the norm $\|\cdot\|$ of $\mathbb{B}$, then the posterior rate is ε_n if

$$\phi_0(\varepsilon_n) \leq n{\varepsilon_n}^2 \qquad \text{AND} \qquad \inf_{h\in\mathbb{H}:\|h-w_0\|<\varepsilon_n} \|h\|_{\mathbb{H}}^2 \leq n{\varepsilon_n}^2.$$

- Both inequalities give lower bound on ε_n.
- The first depends on W and not on w_0.
- If $w_0 \in \mathbb{H}$, then second inequality is satisfied.

Example — Brownian motion

W one-dimensional Brownian motion on $[0,1]$.

- RKHS $\mathbb{H} = \{h\colon \int h'(t)^2\,dt < \infty\}, \quad \|h\|_{\mathbb{H}} = \|h'\|_2$.
- Small ball exponent $\phi_0(\varepsilon) \lesssim (1/\varepsilon)^2$.

LEMMA

If $w_0 \in C^\alpha[0,1]$ for $0 < \alpha < 1$, then $\inf_{h\in\mathbb{H}:\|h-w_0\|_\infty<\varepsilon} \|h'\|_2^2 \lesssim \Big(\frac{1}{\varepsilon}\Big)^{(2-2\alpha)/\alpha}$.

Example — Brownian motion

W one-dimensional Brownian motion on $[0,1]$.

- RKHS $\mathbb{H} = \{h: \int h'(t)^2\,dt < \infty\}, \quad \|h\|_{\mathbb{H}} = \|h'\|_2$.
- Small ball exponent $\phi_0(\varepsilon) \lesssim (1/\varepsilon)^2$.

LEMMA

If $w_0 \in C^\alpha[0,1]$ for $0 < \alpha < 1$, then $\inf_{h\in\mathbb{H}:\|h-w_0\|_\infty<\varepsilon} \|h'\|_2^2 \lesssim \left(\frac{1}{\varepsilon}\right)^{(2-2\alpha)/\alpha}$.

CONSEQUENCE:

Rate is ε_n if $(1/\varepsilon_n)^2 \le n\varepsilon_n^2$ AND $(1/\varepsilon_n)^{(2-2\alpha)/\alpha} \le n\varepsilon_n^2$.

- First implies $\varepsilon_n \ge n^{-1/4}$ for any w_0.
- Second implies $\varepsilon_n \ge n^{-\alpha/2}$ for $w_0 \in C^\alpha[0,1]$.

Examples of settings

Basic rate result

Prior W is Gaussian map in $(\mathbb{B}, \|\cdot\|)$ with RKHS $(\mathbb{H}, \|\cdot\|_{\mathbb{H}})$ and small ball exponent $\phi_0(\varepsilon)$.

THEOREM
If statistical distances on the model combine appropriately with the norm $\|\cdot\|$ of $\mathbb{B}$, then the posterior rate is ε_n if

$$\phi_0(\varepsilon_n) \leq n{\varepsilon_n}^2 \qquad \text{AND} \qquad \inf_{h\in\mathbb{H}:\|h-w_0\|<\varepsilon_n} \|h\|_{\mathbb{H}}^2 \leq n{\varepsilon_n}^2.$$

Density estimation

Data $X_1, \ldots, X_n$ iid from density on $[0,1]$,

$$p_w(x) = \frac{e^{w_x}}{\int_0^1 e^{w_t}\,dt}.$$

- Distance on parameter: Hellinger on p_w.
- Norm on W: uniform.

Density estimation

Data $X_1, \ldots, X_n$ iid from density on $[0,1]$,

$$p_w(x) = \frac{e^{w_x}}{\int_0^1 e^{w_t}\,dt}.$$

- Distance on parameter: Hellinger on p_w.
- Norm on W: uniform.

LEMMA $\forall v, w$

- $h(p_v, p_w) \leq \|v-w\|_\infty \, e^{\|v-w\|_\infty/2}$.
- $K(p_v, p_w) \lesssim \|v-w\|_\infty^2 \, e^{\|v-w\|_\infty}(1 + \|v-w\|_\infty)$.
- $V(p_v, p_w) \lesssim \|v-w\|_\infty^2 \, e^{\|v-w\|_\infty}(1 + \|v-w\|_\infty)^2$.

Classification

Data $(X_1, Y_1), \ldots, (X_n, Y_n)$ iid in $[0,1] \times \{0,1\}$

$$\mathrm{P}_w(Y = 1 | X = x) = \Psi(w_x),$$

for Ψ the logistic or probit link function.

- Distance on parameter: L_2-norm on $\Psi(w)$.
- Norm on W for logistic: $L_2(G)$, G marginal of X_i.

  Norm on W for probit: combination of $L_2(G)$ and $L_4(G)$.

Regression

Data $Y_1, \ldots, Y_n$, fixed design points $x_1, \ldots, x_n$,

$$Y_i = w(x_i) + e_i,$$

for $e_1, \ldots, e_n$ iid Gaussian mean-zero errors.

- Distance on parameter: empirical L_2-distance on w.
- Norm on W: uniform.

Ergodic diffusions

Data $(X_t: t \in [0, n])$

$$dX_t = w(X_t)\,dt + \sigma(X_t)\,dB_t.$$

Ergodic, recurrent on $\mathbb{R}$, stationary measure μ_0, "usual" conditions.

- Distance on parameter: random Hellinger h_n.
- Norm on W: $L_2(\mu_0)$.

$$h_n^2(w_1, w_2) = \int_0^n \Big(\frac{w_1(X_t) - w_2(X_t)}{\sigma(X_t)}\Big)^2\,dt \approx \|(w_1 - w_2)/\sigma\|_{\mu_0,2}^2.$$

[van der Meulen & vZ & vdV, Panzar & vZ]

Proof ingredients

Proof

Given that the relevant statistical distances translate into the Banach space norm, it follows that the posterior rate is ε_n if there exist sets $\mathbb{B}_n$ such that

(1) $\log N(\varepsilon_n, \mathbb{B}_n, d) \leq n\varepsilon_n^2$ and $\Pi_n(\mathbb{B}_n) = 1 - o(e^{-3n\varepsilon_n^2})$. entropy.

(2) $\Pi_n\bigl(w\colon \|w - w_0\| < \varepsilon_n\bigr) \geq e^{-n\varepsilon_n^2}$. prior mass.

The second condition actually implies the first.

Prior mass

W a Gaussian map in $(\mathbb{B}, \|\cdot\|)$ with RKHS $(\mathbb{H}, \|\cdot\|_{\mathbb{H}})$ and small ball exponent $\phi_0(\varepsilon)$.

$$\phi_{w_0}(\varepsilon) := \phi_0(\varepsilon) + \inf_{h \in \mathbb{H}: \|h - w_0\| < \varepsilon} \|h\|_{\mathbb{H}}^2.$$

Prior mass

W a Gaussian map in $(\mathbb{B}, \|\cdot\|)$ with RKHS $(\mathbb{H}, \|\cdot\|_{\mathbb{H}})$ and small ball exponent $\phi_0(\varepsilon)$.

$$\phi_{w_0}(\varepsilon) := \phi_0(\varepsilon) + \inf_{h \in \mathbb{H}: \|h - w_0\| < \varepsilon} \|h\|_{\mathbb{H}}^2.$$

THEOREM [Kuelbs & Li (93)]
Concentration function measures concentration around w_0:

$$\mathrm{P}(\|W - w_0\| < \varepsilon) \asymp e^{-\phi_{w_0}(\varepsilon)}.$$

(up to factors 2)

Complexity

RKHS gives the "geometry of the support of W".

THEOREM
The closure of $\mathbb{H}$ in $\mathbb{B}$ is support of the Gaussian measure (and hence posterior inconsistent if $\|w_0 - \mathbb{H}\| > 0$).

THEOREM [Borell (75)]
For $\mathbb{H}_1$ and $\mathbb{B}_1$ the unit balls of RKHS and $\mathbb{B}$

$$\mathrm{P}(W \notin M\mathbb{H}_1 + \varepsilon\mathbb{B}_1) \leq 1 - \Phi\bigl(\Phi^{-1}(e^{-\phi_0(\varepsilon)}) + M\bigr).$$

Proof

Given that the relevant statistical distances translate into the Banach space norm, it follows that the posterior rate is ε_n if there exist sets $\mathbb{B}_n$ such that

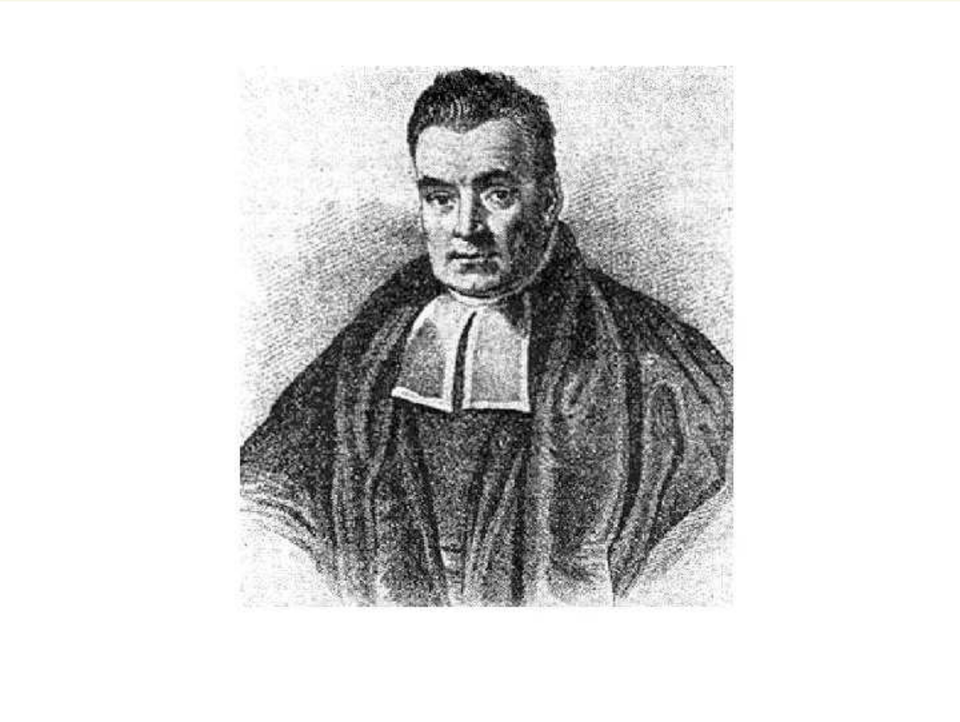
(1) $\log N(\varepsilon_n, \mathbb{B}_n, d) \leq n\varepsilon_n^2$ and $\Pi_n(\mathbb{B}_n) = 1 - o(e^{-3n\varepsilon_n^2})$. entropy.

(2) $\Pi_n\bigl(w\colon \|w - w_0\| < \varepsilon_n\bigr) \geq e^{-n\varepsilon_n^2}$. prior mass.

Take $\mathbb{B}_n = M_n\mathbb{H}_1 + \varepsilon_n\mathbb{B}_1$ for appropriate M_n.

Conclusion

Conclusion



Bayesian inference with Gaussian processes is flexible and elegant. However, priors must be chosen with some care: eye-balling pictures of sample paths or staring at the covariance function does not reveal the fine properties [David Freedman] that matter for posterior performance.