Bayesian Regularization

Aad van der Vaart
Vrije Universiteit Amsterdam

International Congress of Mathematicians
Hyderabad, August 2010

Contents

Co-authors

Abstract result

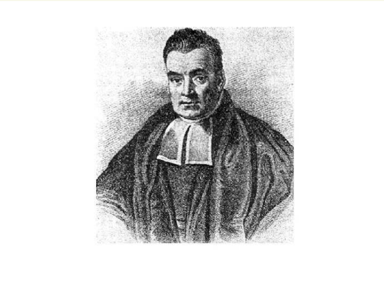
Subhashis Ghosal

Gaussian process priors

Harry van Zanten

Introduction

The Bayesian paradigm



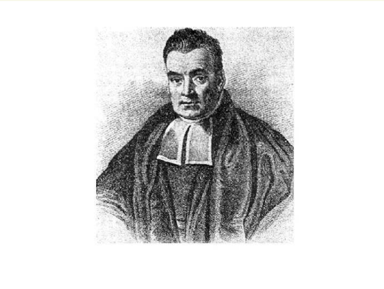
- A *parameter* Θ is generated according to a prior distribution Π.
- Given $\Theta = \theta$ the *data* X is generated according to a probability density p_θ.

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



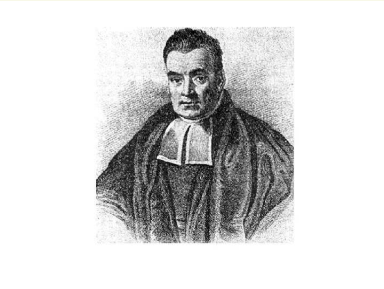
- A *parameter* Θ is generated according to a prior distribution Π.
- Given $\Theta = \theta$ the *data* X is generated according to a probability density p_θ.

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



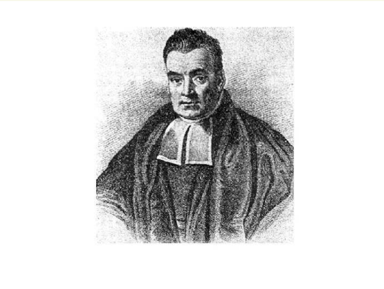
- A *parameter* Θ is generated according to a prior distribution Π.
- Given $\Theta = \theta$ the *data* X is generated according to a probability density p_θ.

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

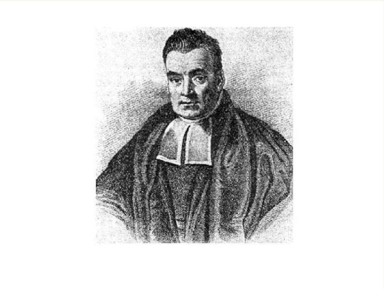


Thomas Bayes (1702–1761, 1763) followed this argument with Θ possessing the *uniform distribution* and X given $\Theta = \theta$ *binomial* (n, θ).

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

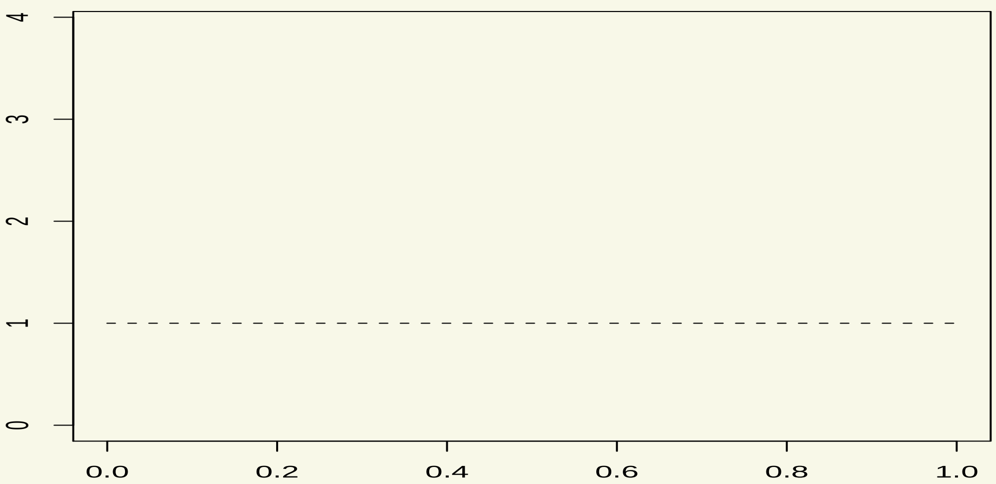
$$d\Pi_n(\theta|\, X) \propto \binom{n}{X} \theta^X (1-\theta)^{n-X} \cdot 1.$$

Reverend Thomas

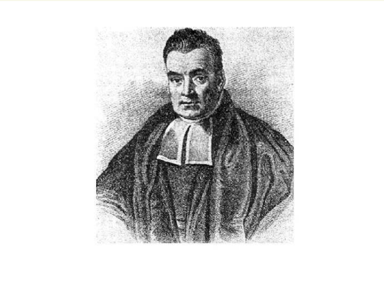


Thomas Bayes (1702–1761, 1763) followed this argument with Θ possessing the *uniform distribution* and X given $\Theta = \theta$ *binomial* (n, θ).

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

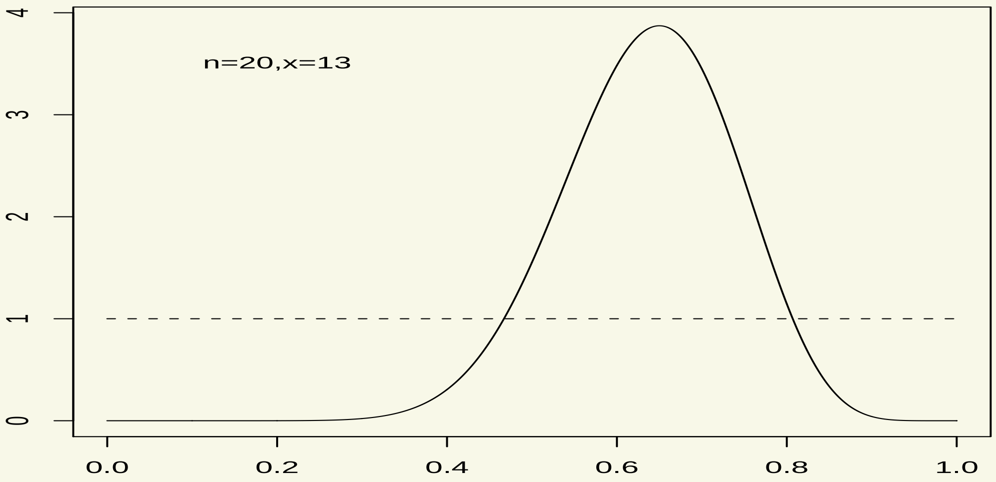


Reverend Thomas

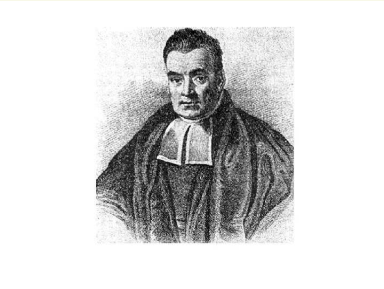


Thomas Bayes (1702–1761, 1763) followed this argument with Θ possessing the *uniform distribution* and X given $\Theta = \theta$ *binomial* (n, θ).

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

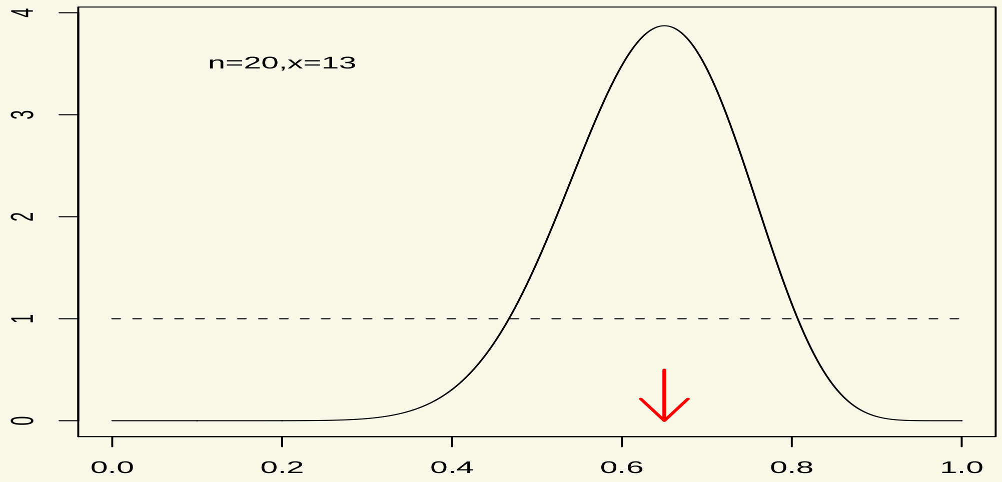


Reverend Thomas

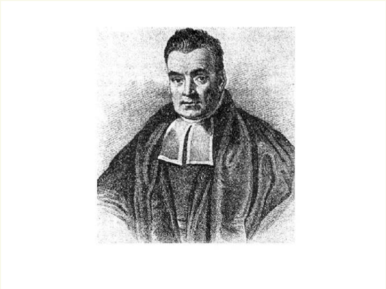


Thomas Bayes (1702–1761, 1763) followed this argument with Θ possessing the *uniform distribution* and X given $\Theta = \theta$ *binomial* (n, θ).

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

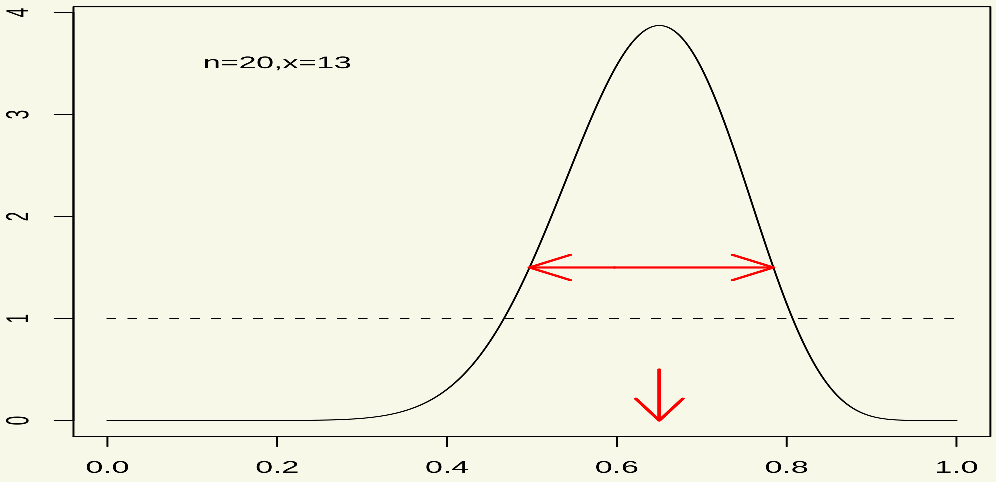


Reverend Thomas



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:

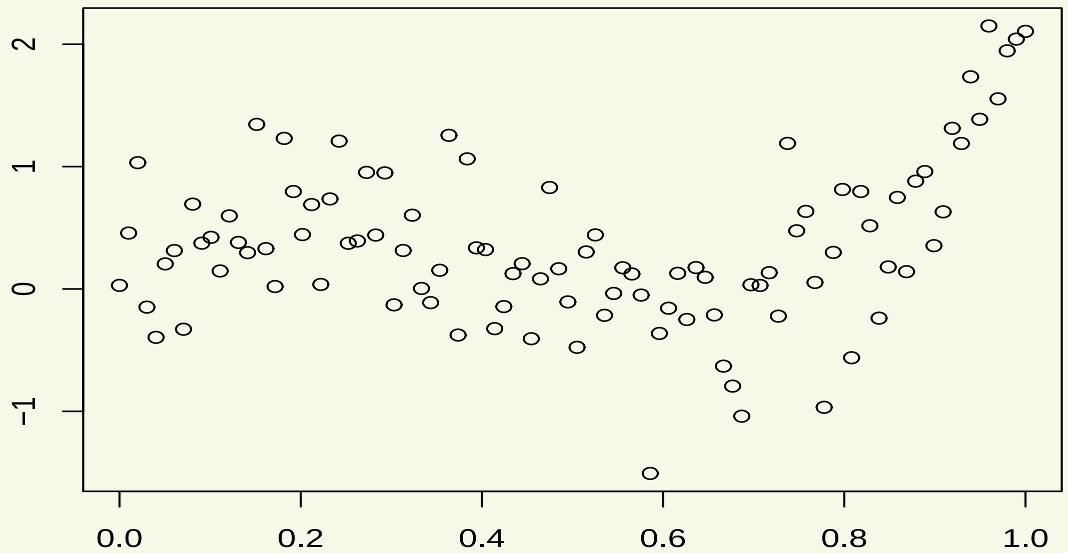
$$d\Pi(\theta|\, X) \propto p_\theta(X)\, d\Pi(\theta).$$

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.



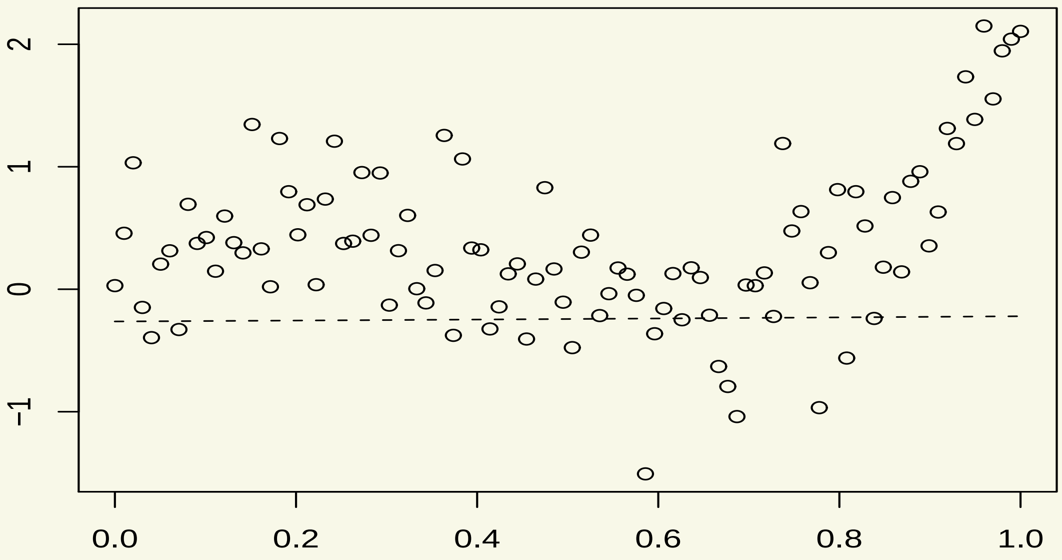
Nonparametric Bayes

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.

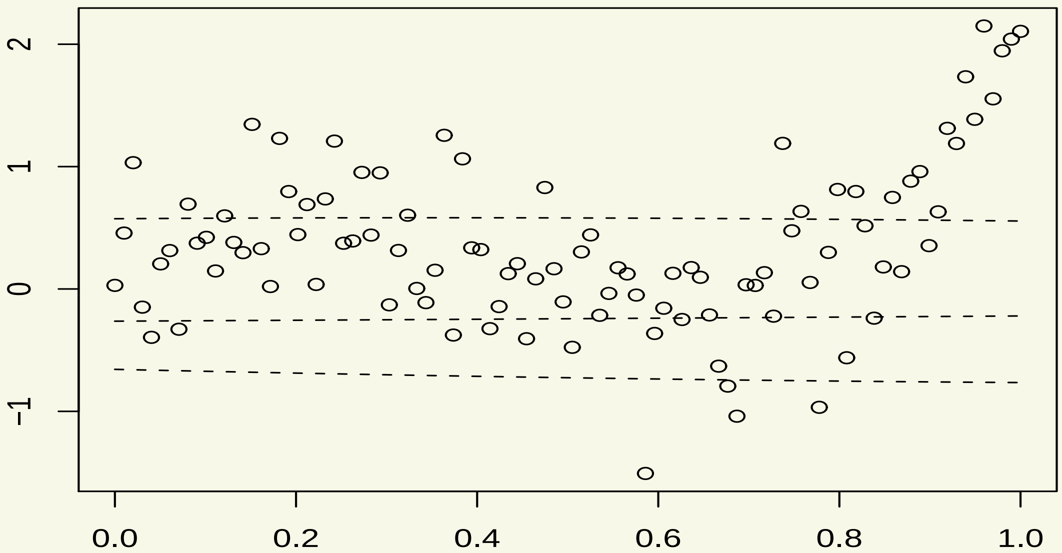


Nonparametric Bayes

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:

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

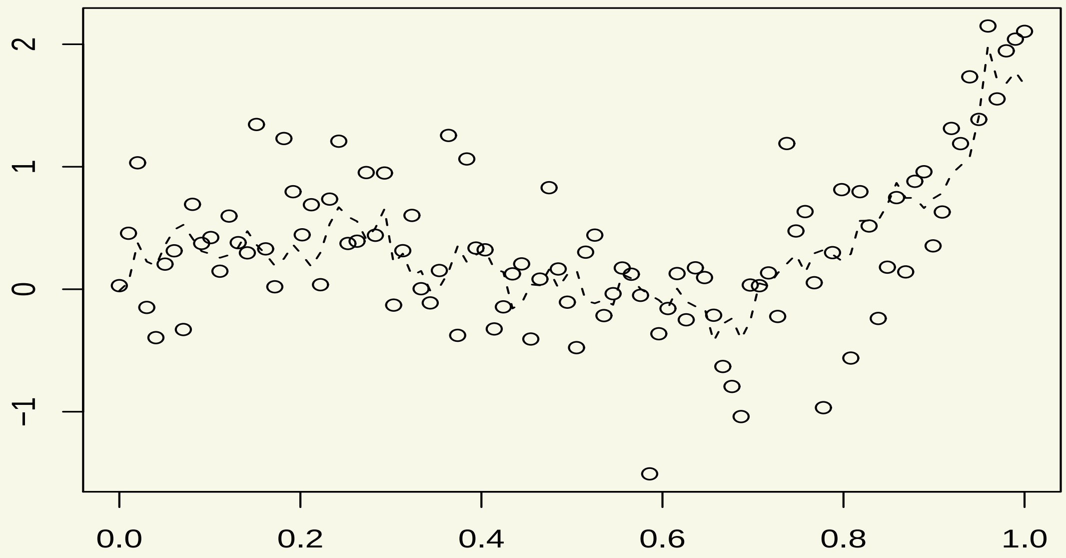
Prior and posterior can be visualized by plotting functions that are simulated from these distributions.



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.

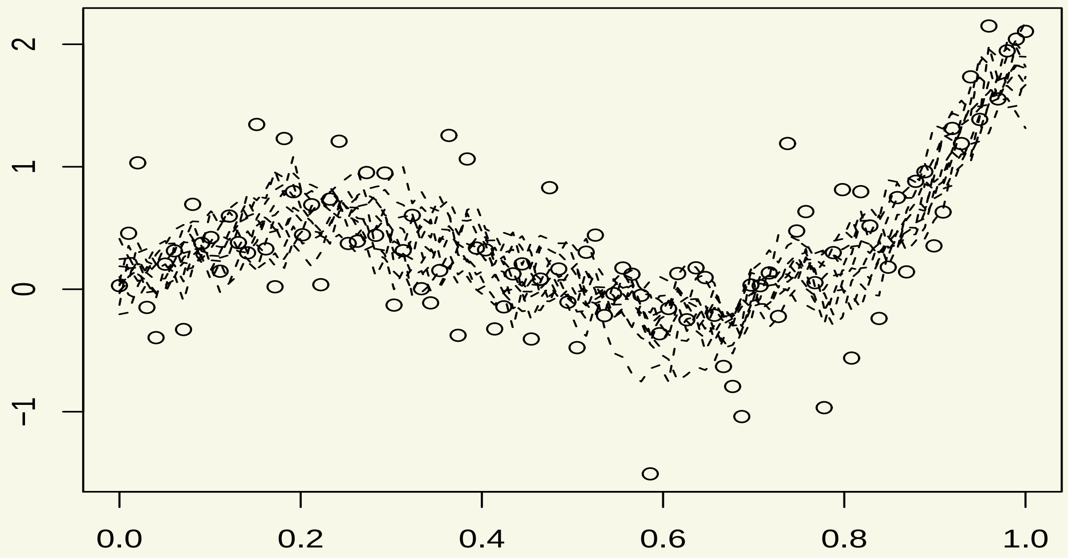


Nonparametric Bayes

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.

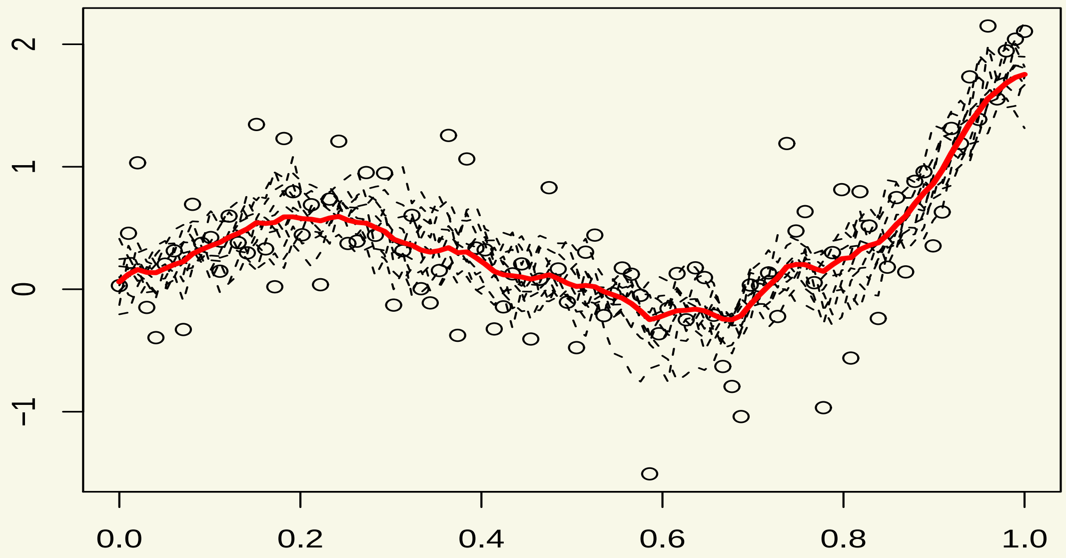


Nonparametric Bayes

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.

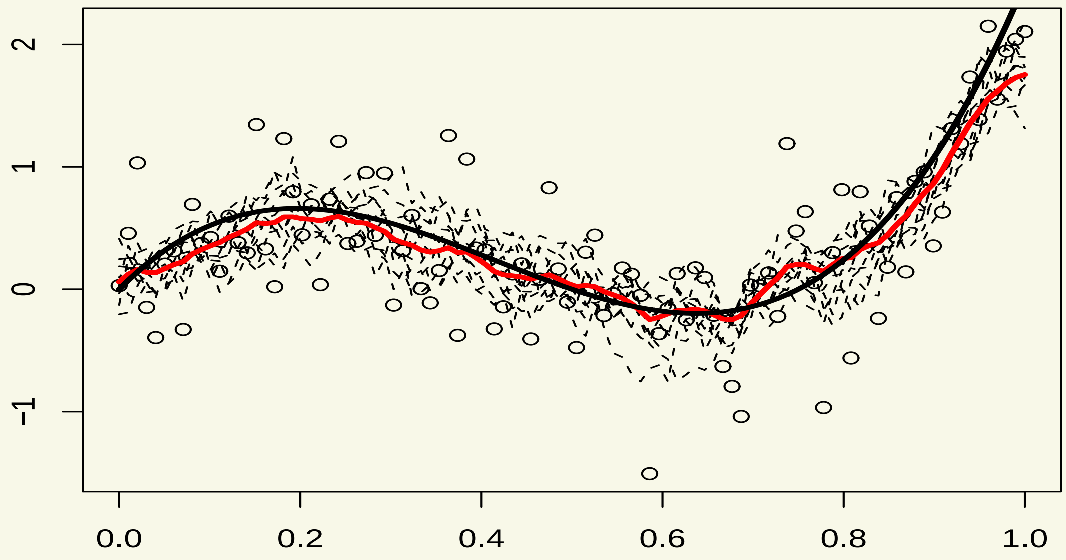


Nonparametric Bayes

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.



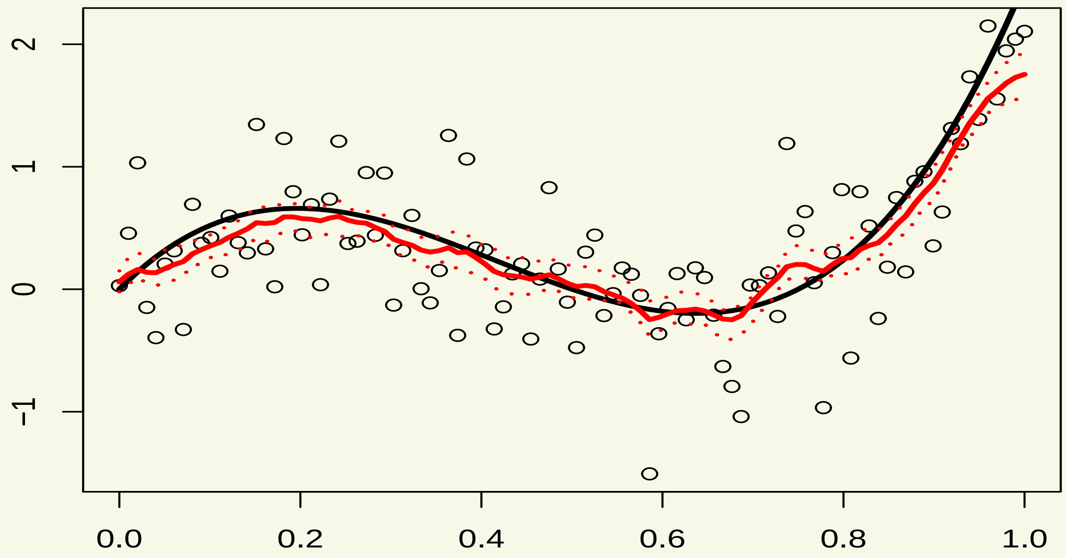
Nonparametric Bayes

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.



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.

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*.

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, for every $\varepsilon > 0$,
  $\mathrm{E}_{\theta_0}\Pi\big(\theta\colon d(\theta,\theta_0) < \varepsilon |\, X^n\big) \to 1$.

- 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

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

$$\varepsilon_n = \inf_T \sup_{\theta \in \Theta} \mathrm{E}_\theta d\big(T(X), \theta\big).$$

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

Adaptation

A *model* can be viewed as an instrument to test quality.

It makes sense to use a collection $(\Theta_\alpha : \alpha \in A)$ of models simultaneously, e.g. a "scale" of *regularity classes*.

A posterior is good if it adapts: if the true parameter belongs to Θ_α, then the contraction rate is at least the minimax rate for this model.

Bayesian perspective

Any prior (and hence posterior) is appropriate per se.

In complex situations subject knowledge can be and must be incorporated in the prior.

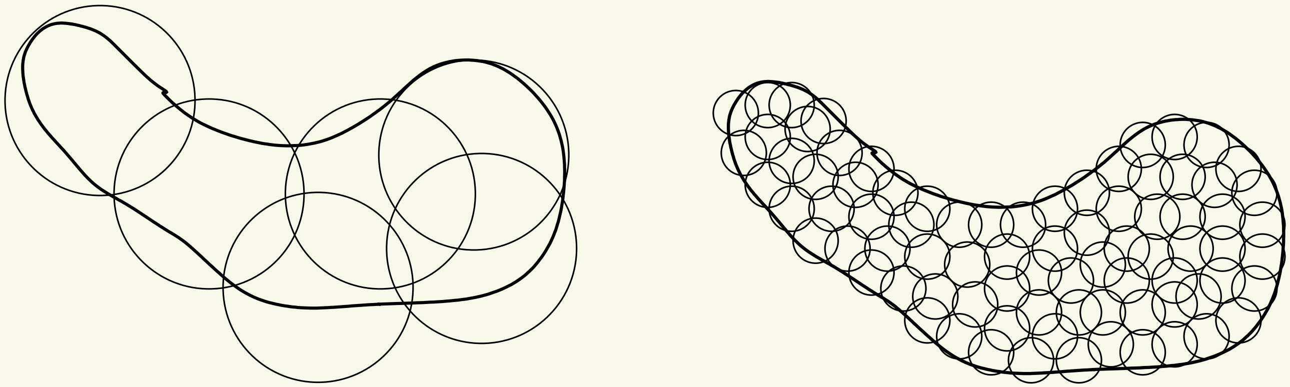
Computational ease is important for prior choice as well.

Frequentist properties reveal key properties of priors of interest.

Abstract result

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 $\qquad\qquad$ ε small

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

Rate theorem — 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 [Ghosal+Ghosh+vdV, 2000]
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) $\Pi\big(B_{KL}(p_0, \varepsilon_n)\big) \geq e^{-n\varepsilon_n^2}$. 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 theorem — 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 [Ghosal+Ghosh+vdV, 2000]
The Hellinger contraction rate is ε_n if there exist $\mathcal{P}_n \subset \mathcal{P}$ such that

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

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

The entropy condition ensures that the likelihood is not too variable, so that it cannot be large at a wrong place by pure randomness.
Le Cam (1964) showed that it gives the minimax rate.

Rate theorem — general

Given data X^n following a model $(P_\theta^n : \theta \in \Theta)$ 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 $\Theta_n \subset \Theta$ such that

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

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

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

Rate theorem — general

Given data X^n following a model $(P_\theta^n\colon \theta \in \Theta)$ 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 $\Theta_n \subset \Theta$ such that

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

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

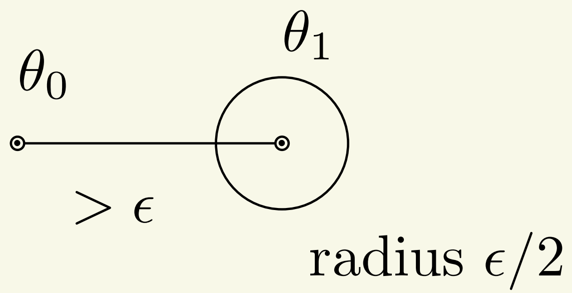
The theorem can be refined in various ways.

Le Cam's testing criterion

Statistical model $(P_\theta^n: \theta \in \Theta)$ indexed by metric space (Θ, d).

For all $\varepsilon > 0$: for all θ_1 with $d(\theta_1, \theta_0) > \varepsilon$ $\exists$ test ϕ_n with

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



This applies to *independent data, Markov chains, Gaussian time series, ergodic diffusions,* $\ldots$.

Adaptation by hierarchical prior — i.i.d.

- $\mathcal{P}_{n,\alpha}$ collection of densities with prior $\Pi_{n,\alpha}$, for $\alpha \in A$.
- Prior "weights" $\lambda_n = (\lambda_{n,\alpha} : \alpha \in A)$.
- $\Pi_n = \sum_{\alpha \in A} \lambda_{n,\alpha} \Pi_{n,\alpha}$.

THEOREM
The Hellinger contraction rate is $\varepsilon_{n,\beta}$ if the prior weights satisfies (*) below and

(1) $\log N\bigl(\varepsilon_{n,\alpha}, \mathcal{P}_{n,\alpha}, h\bigr) \le n\varepsilon_{n,\alpha}^2$, every $\alpha \in A$.

(2) $\Pi_{n,\beta}\bigl(B_{n,\beta}(p_0, \varepsilon_{n,\beta})\bigr) \ge e^{-n\varepsilon_{n,\beta}^2}$.

$B_{n,\alpha}(p_0, \varepsilon)$ is Kullback-Leibler type neighbourhood of p_0 within $\mathcal{P}_{n,\alpha}$.

Condition (*) on prior weights (simplified)

$$\sum_{\alpha<\beta} \sqrt{\frac{\lambda_{n,\alpha}}{\lambda_{n,\beta}}} e^{-n\varepsilon_{n,\alpha}^2} + \sum_{\alpha>\beta} \sqrt{\frac{\lambda_{n,\alpha}}{\lambda_{n,\beta}}} \leq e^{n\varepsilon_{n,\beta}^2},$$

$$\sum_{\alpha<\beta} \frac{\lambda_{n,\alpha}}{\lambda_{n,\beta}} \Pi_{n,\alpha}\big(C_{n,\alpha}(p_0, \varepsilon_{n,\alpha})\big) \leq e^{-4n\varepsilon_{n,\beta}^2}.$$

$\alpha < \beta$ means $\varepsilon_{n,\alpha} \gtrsim \varepsilon_{n,\beta}$.
$C_{n,\alpha}(p_0, \varepsilon)$ is Hellinger ball of radius ε around p_0 in $\mathcal{P}_{n,\alpha}$.

Condition (*) on prior weights (simplified)

$$\sum_{\alpha<\beta}\sqrt{\frac{\lambda_{n,\alpha}}{\lambda_{n,\beta}}}e^{-n\varepsilon_{n,\alpha}^2}+\sum_{\alpha>\beta}\sqrt{\frac{\lambda_{n,\alpha}}{\lambda_{n,\beta}}}\leq e^{n\varepsilon_{n,\beta}^2},$$

$$\sum_{\alpha<\beta}\frac{\lambda_{n,\alpha}}{\lambda_{n,\beta}}\Pi_{n,\alpha}\big(C_{n,\alpha}(p_0,\varepsilon_{n,\alpha})\big)\leq e^{-4n\varepsilon_{n,\beta}^2}.$$

$\alpha<\beta$ means $\varepsilon_{n,\alpha}\gtrsim\varepsilon_{n,\beta}$.
$C_{n,\alpha}(p_0,\varepsilon)$ is Hellinger ball of radius ε around p_0 in $\mathcal{P}_{n,\alpha}$.

In many situations there is much freedom in choice of weights.

The weights $\lambda_{n,\alpha}\propto\mu_\alpha e^{-Cn\varepsilon_{n,\alpha}^2}$ always work.

Model selection

THEOREM

Under the conditions of the theorem

$$\Pi_n\big(\alpha\colon \alpha < \beta | X_1, \cdots, X_n\big) \xrightarrow{P} 0,$$
$$\Pi_n\big(\alpha\colon \alpha \gtrsim \beta, h(\mathcal{P}_{n,\alpha}, p_0) \gtrsim \varepsilon_{n,\beta} | X_1, \cdots, X_n\big) \xrightarrow{P} 0.$$

Too “big” models do not get posterior weight. Neither do “small” models that are “far” from the truth.

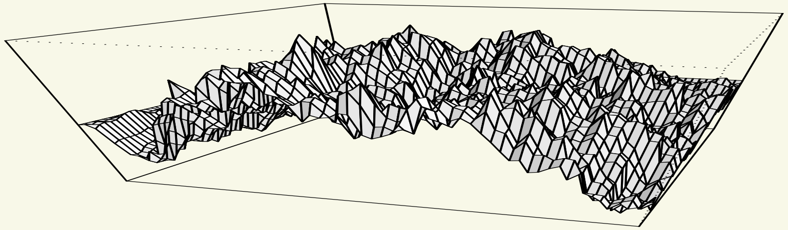
Examples of priors

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

Gaussian process priors

Gaussian process

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



Gaussian processes have been found useful, because of their variety and because of computational properties.

Every Gaussian prior is reasonable in some way. We shall study performance with “smoothness” classes as test case.

Example: Brownian density estimation

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

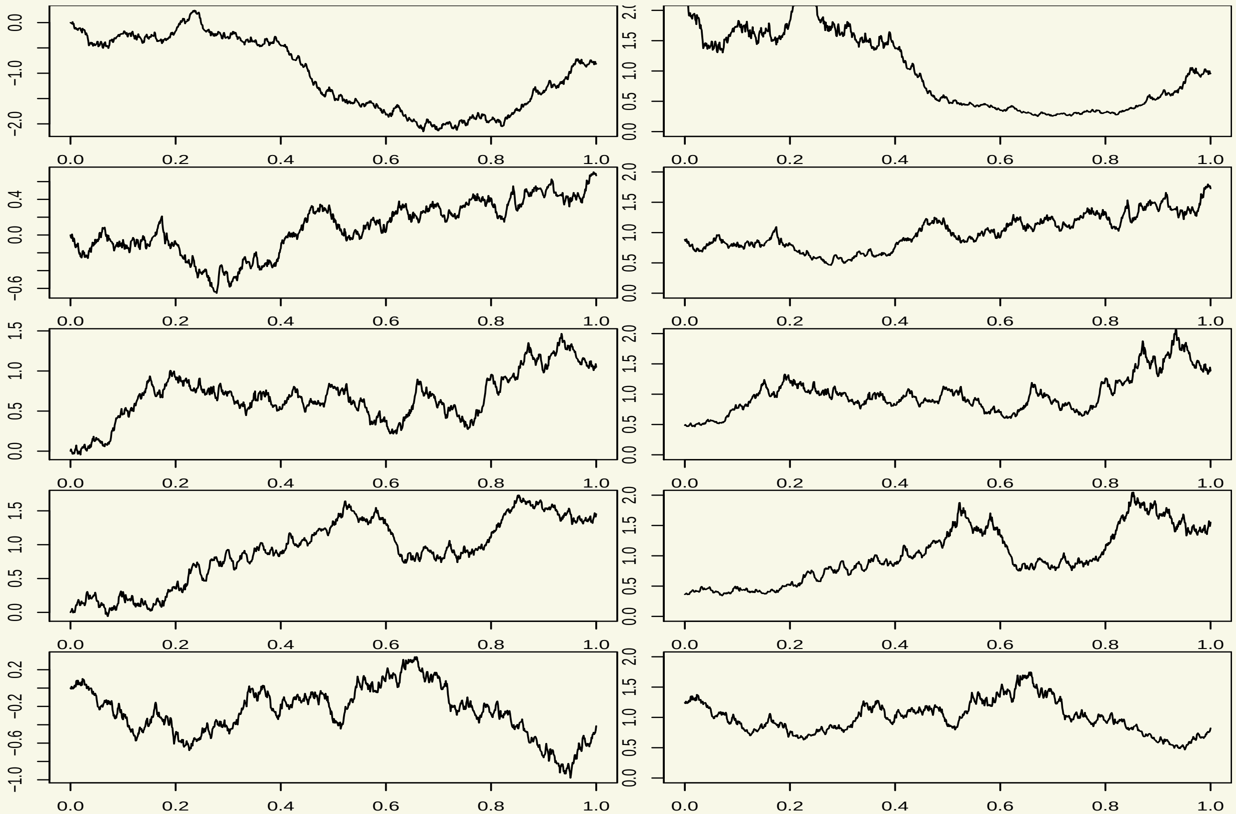
$$x \mapsto \frac{e^{W_x}}{\int_0^1 e^{W_y}\,dy}.$$

[Leonard, Lenk, Tokdar & Ghosh]

Example: Brownian density estimation

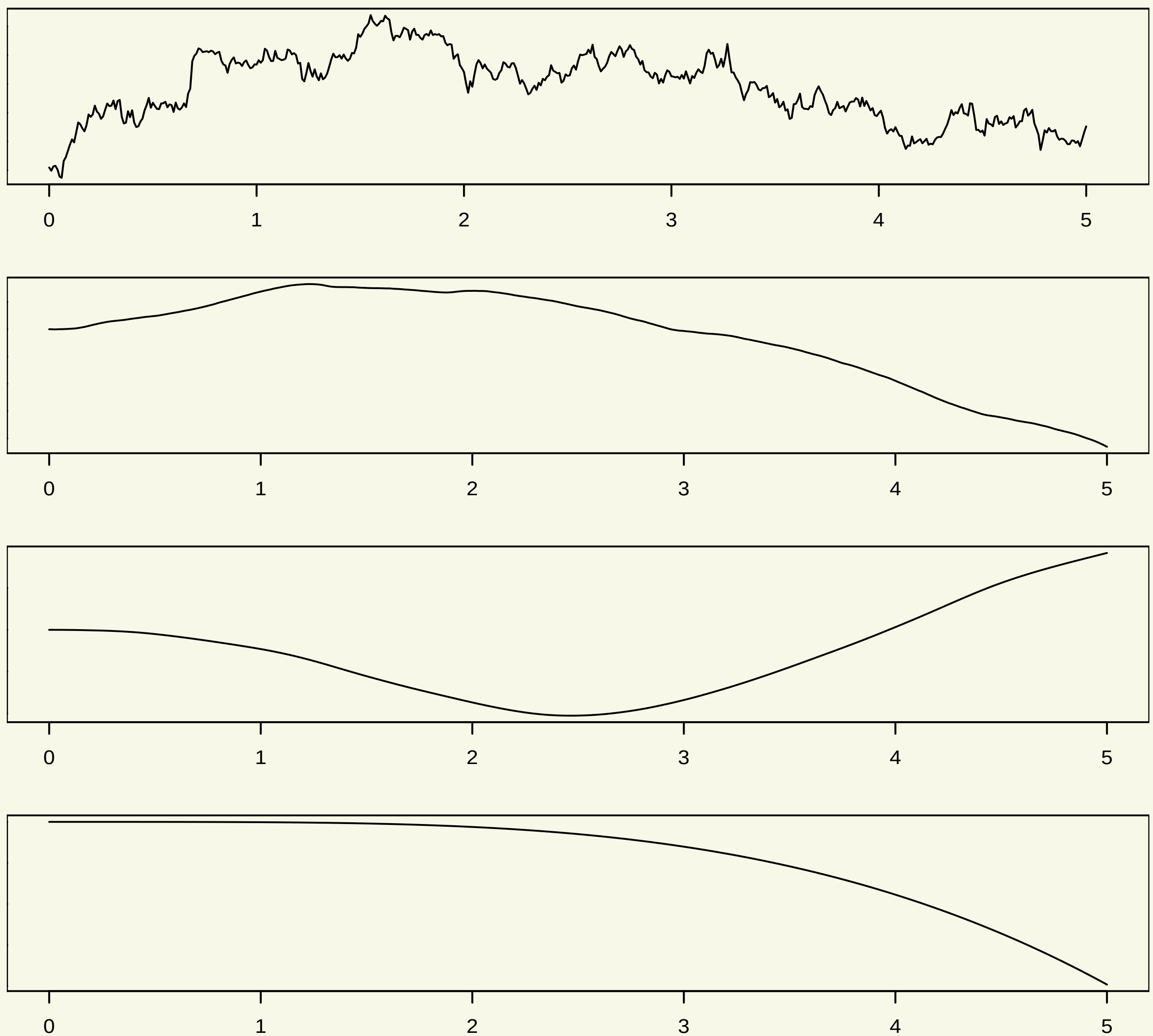
For W Brownian motion use as prior on a density p on $[0,1]$:

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



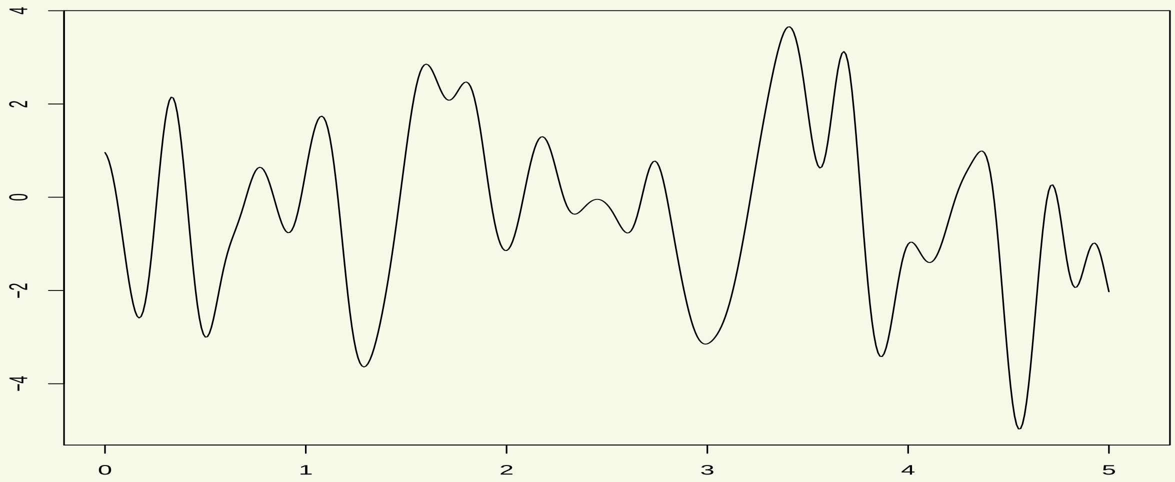
Brownian motion $t \mapsto W_t$ — Prior density $t \mapsto c\exp(W_t)$

Integrated Brownian motion



0, 1, 2 and 3 times integrated Brownian motion

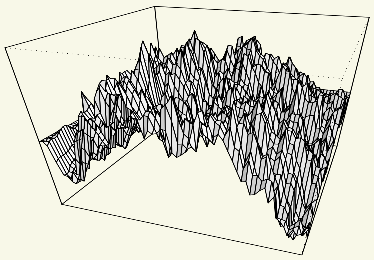
Stationary processes



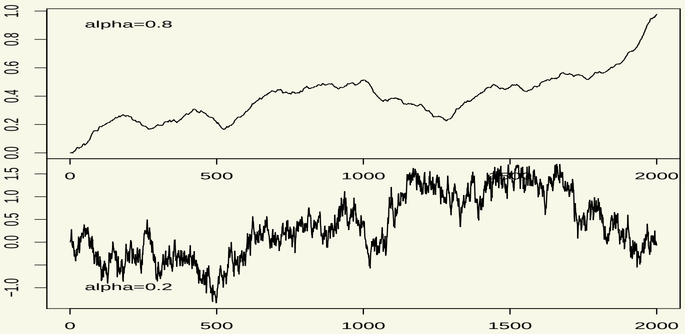
Gaussian spectral measure

Matérn spectral measure (3/2)

Other Gaussian processes



Brownian sheet



Fractional Brownian motion

$$w = \sum_i w_i e_i, \qquad w_i \sim_{ind} N(0, \sigma_i^2)$$

Series prior

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

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

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 for $\varepsilon_n \gtrsim 1/\sqrt{n}$.

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 for $\varepsilon_n \gtrsim 1/\sqrt{n}$.

Settings

Density estimation
$X_1, \ldots, X_n$ iid in $[0,1]$,

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

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

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

$$\mathrm{P}_w(Y = 1|X = x) = \frac{1}{1 + e^{-w_x}}.$$

- Distance on parameter: $L_2(G)$ on P_w. (G marginal of X_i.)
- Norm on W: $L_2(G)$.

Regression
$Y_1, \ldots, Y_n$ independent $N(w(x_i), \sigma^2)$, for fixed design points $x_1, \ldots, x_n$.

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

Ergodic diffusions
$(X_t\colon t \in [0,n])$, ergodic, recurrent:

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

- Distance on parameter: random Hellinger h_n ($\approx \|\cdot/\sigma\|_{\mu_0,2}$).
- Norm on W: $L_2(\mu_0)$.
  (μ_0 stationary measure.)

Reproducing kernel Hilbert space

To every Gaussian random element with values in a Banach space $(\mathbb{B}, \|\cdot\|)$ 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

To every Gaussian random element with values in a Banach space $(\mathbb{B}, \|\cdot\|)$ 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

To every Gaussian random element with values in a Banach space $(\mathbb{B}, \|\cdot\|)$ 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

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

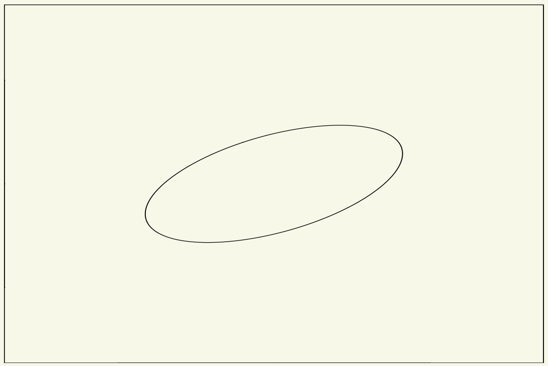
Reproducing kernel Hilbert space

To every Gaussian random element with values in a Banach space $(\mathbb{B}, \|\cdot\|)$ 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

To every Gaussian random element with values in a Banach space $(\mathbb{B}, \|\cdot\|)$ 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

To every Gaussian random element with values in a Banach space $(\mathbb{B}, \|\cdot\|)$ 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

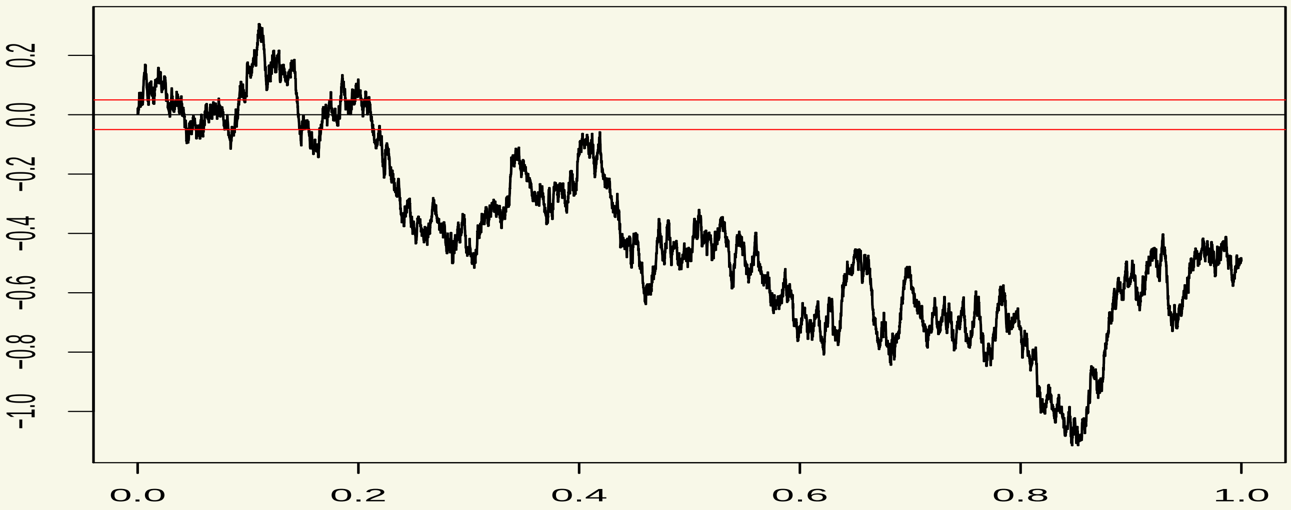
Brownian motion is a random element in $C[0,1]$.
Its RKHS is $\mathbb{H} = \{h\colon \int h'(t)^2\,dt < \infty\}$ with norm $\|h\|_{\mathbb{H}} = \|h'\|_2$.

Small ball probability

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

$$P(\|W\| < \varepsilon),$$

and the small ball exponent $\phi_0(\varepsilon)$ is minus the logarithm of this.



small ball for uniform norm

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 $\phi_0(\varepsilon)$ is minus the logarithm of this.

It can be computed either by probabilistic arguments, or analytically from the RKHS.

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. (In July 2009 243 entries in database maintained by Michael Lifshits.)

Rates for Gaussian priors — proof

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

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

Rates for Gaussian priors — proof

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

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

PROOF
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 $\mathrm{P}(W \in \mathbb{B}_n) = 1 - o(e^{-3n\varepsilon_n^2})$. entropy.

(2) $\mathrm{P}\big(\|W - w_0\| < \varepsilon_n\big) \geq e^{-n\varepsilon_n^2}$. prior mass.

Rates for Gaussian priors — proof

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

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

PROOF

The posterior rate is ε_n if there exist sets $\mathbb{B}_n$ such that

(1) $\log N(\varepsilon_n, \mathbb{B}_n, d) \le n\varepsilon_n^2$ and $\mathrm{P}(W \in \mathbb{B}_n) = 1 - o(e^{-3n\varepsilon_n^2})$. entropy.

(2) $\mathrm{P}\big(\|W - w_0\| < \varepsilon_n\big) \ge e^{-n\varepsilon_n^2}$. prior mass.

Take $\mathbb{B}_n = M_n\mathbb{H}_1 + \varepsilon_n\mathbb{B}_1$ for large M_n ($\mathbb{H}_1$, $\mathbb{B}_1$ the unit balls of $\mathbb{H}$, $\mathbb{B}$).

Proof (2) — key results

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

THEOREM [Kuelbs & Li (93)]

Concentration function measures concentration around w_0 (up to factors 2):

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

THEOREM [Borell (75)]

For $\mathbb{H}_1$ and $\mathbb{B}_1$ the unit balls of RKHS and $\mathbb{B}$

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

(Integrated) Brownian Motion

THEOREM

If $w_0 \in C^\beta[0,1]$, then the rate for Brownian motion is: $n^{-1/4}$ if $\beta \geq 1/2$;
$n^{-\beta/2}$ if $\beta \leq 1/2$.

The small ball exponent of Brownian motion is $\phi_0(\varepsilon) \asymp (1/\varepsilon)^2$ as $\varepsilon \downarrow 0$.
This gives the $n^{-1/4}$-rate, even for very smooth truths.

Truths with $\beta \leq 1/2$ are "far from" the RKHS, giving the rate $n^{-\beta/2}$.

The minimax rate is attained iff $\beta = 1/2$.

(Integrated) Brownian Motion

THEOREM

If $w_0 \in C^\beta[0,1]$, then the rate for Brownian motion is: $n^{-1/4}$ if $\beta \geq 1/2$;
$n^{-\beta/2}$ if $\beta \leq 1/2$.

The small ball exponent of Brownian motion is $\phi_0(\varepsilon) \asymp (1/\varepsilon)^2$ as $\varepsilon \downarrow 0$.
This gives the $n^{-1/4}$-rate, even for very smooth truths.

Truths with $\beta \leq 1/2$ are “far from” the RKHS, giving the rate $n^{-\beta/2}$.

The minimax rate is attained iff $\beta = 1/2$.

THEOREM

If $w_0 \in C^\beta[0,1]$, then the rate for $(\alpha - 1/2)$-times integrated Brownian is $n^{-(\beta\wedge\alpha)/(2\alpha+d)}$.

The minimax rate is attained iff $\beta = \alpha$.

Stationary processes

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

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

Stationary processes

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

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

THEOREM

Suppose that μ is Gaussian. Let $\hat{w}_0$ be the Fourier transform of $w_0: [0,1]^d \to \mathbb{R}$.

- If $\int e^{\|\lambda\|}|\hat{w}_0(\lambda)|^2\, d\lambda < \infty$, then rate of contraction is near $1/\sqrt{n}$.
- If $\int (1+\|\lambda\|^2)^\beta |\hat{w}_0(\lambda)|^2\, d\lambda < \infty$, then rate is $(1/\log n)^{\kappa_\beta}$.

Stationary processes

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

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

THEOREM
Suppose that μ is Gaussian. Let $\hat{w}_0$ be the Fourier transform of $w_0: [0,1]^d \to \mathbb{R}$.

- If $\int e^{\|\lambda\|}|\hat{w}_0(\lambda)|^2\, d\lambda < \infty$, then rate of contraction is near $1/\sqrt{n}$.
- If $\int (1+\|\lambda\|^2)^\beta |\hat{w}_0(\lambda)|^2\, d\lambda < \infty$, then rate is $(1/\log n)^{\kappa_\beta}$.

THEOREM
Suppose that $d\mu(\lambda) = (1+\|\lambda\|^2)^{-(\alpha-d/2)}\, d\lambda$.

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

Adaptation

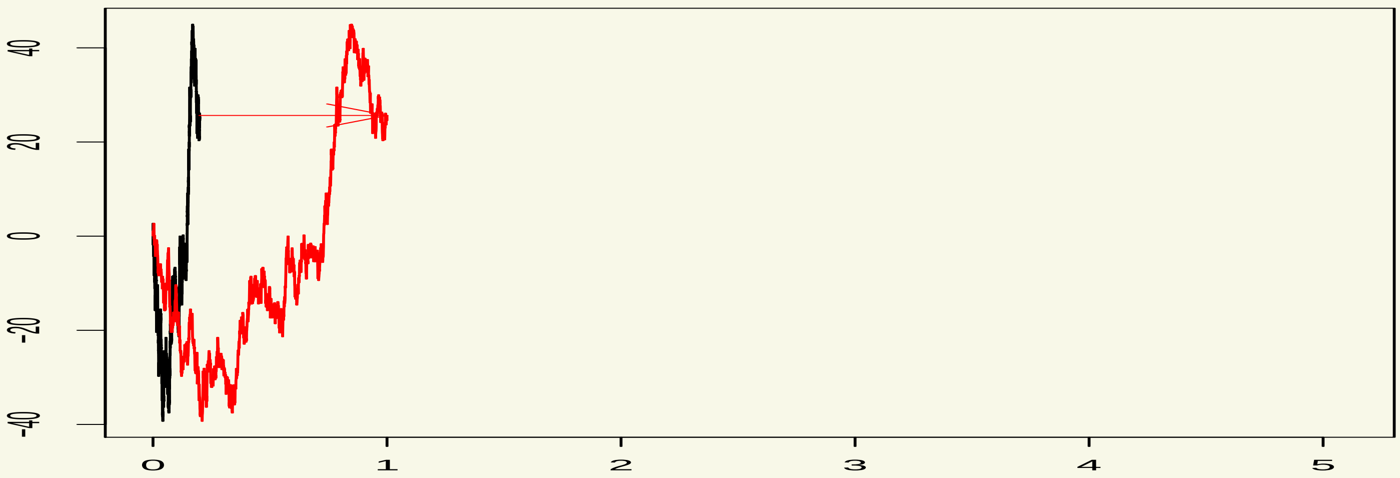
Every Gaussian prior is good for some regularity class, but may be very bad for another.

This can be alleviated by putting a prior on the regularity of the process.

An alternative, more attractive approach is scaling.

Stretching or shrinking

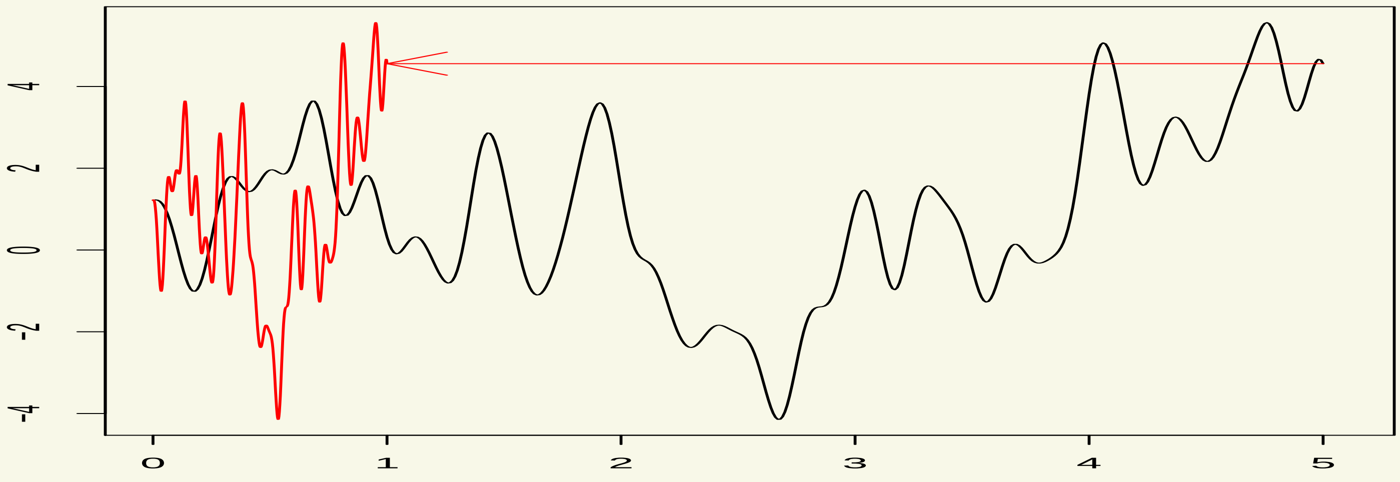
Sample paths can be smoothed by stretching



Stretching or shrinking

Sample paths can be smoothed by stretching

and roughened by shrinking



Scaled (integrated) 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 $w_0 \in C^\alpha[0,1]$, $\alpha \in (0,1]$.

Scaled (integrated) 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 $w_0 \in C^\alpha[0,1]$, $\alpha \in (0,1]$.

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

Scaled (integrated) 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 $w_0 \in C^\alpha[0,1]$, $\alpha \in (0,1]$.

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

Stretching helps a little, shrinking helps a lot.

Scaled 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 $\alpha > 0$.

Adaptation by random scaling

- Choose A^d from a Gamma distribution.
- Choose $(G_t: t > 0)$ centered Gaussian with $\mathrm{E} G_s G_t = \exp\bigl(-\|s-t\|^2\bigr)$.
- 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 $n^{-1/2}$.

Adaptation by random scaling

- 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 $n^{-1/2}$.

The first result is also true for randomly scaled k-times integrated Brownian motion and $\alpha \leq k+1$.

Conclusion

Conclusion and final remark

There exist natural (fixed) priors that yield fully automatic smoothing at the “correct” bandwidth. (For instance, randomly scaled Gaussian processes.)

Conclusion and final remark

There exist natural (fixed) priors that yield fully automatic smoothing at the "correct" bandwidth. (For instance, randomly scaled Gaussian processes.)

Similar statements are true for adaptation to the scale of models described by sparsity (*research in progress*).