Bayesian curve estimation using Gaussian process priors

Aad van der Vaart
Vrije Universiteit Amsterdam

Utrecht, May 2009

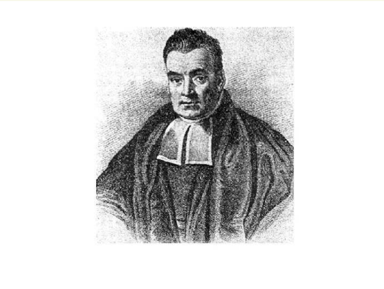
Co-author

Harry van Zanten

Contents

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 measure 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

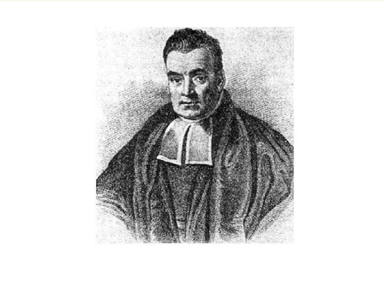
- A parameter Θ is generated according to a prior distribution Π.
- Given $\Theta = \theta$ the data X is generated according to a measure P_θ.

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



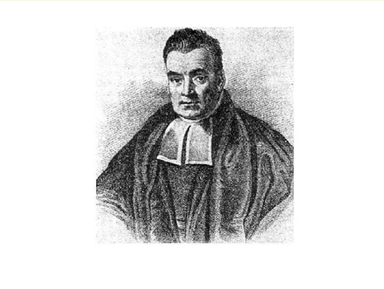
- A parameter Θ is generated according to a prior distribution Π.
- Given $\Theta = \theta$ the data X is generated according to a measure P_θ.

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*$(X + 1, n - X + 1)$.

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

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

$$\mathrm{P}(a \leq \Theta \leq b | \, X = x) = \int_a^b \theta^x (1 - \theta)^{n-x} \, d\theta / B(x + 1, n - x + 1).$$

Reverend Thomas

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

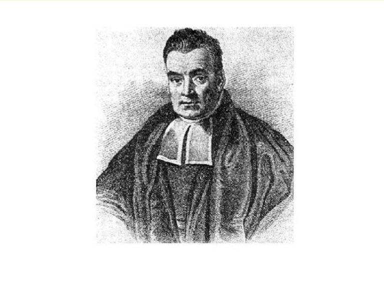
Using his famous rule he computed that the posterior distribution is then *Beta*$(X+1, n-X+1)$.

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

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

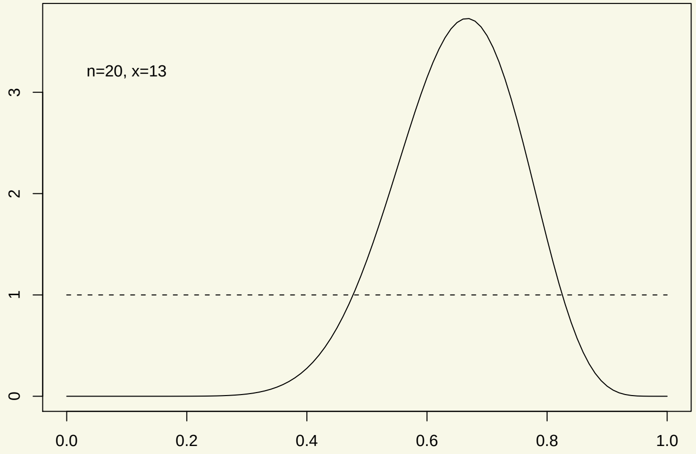
$$d\Pi(\theta | X) = \theta^X (1-\theta)^{n-X} \cdot 1.$$

Reverend Thomas

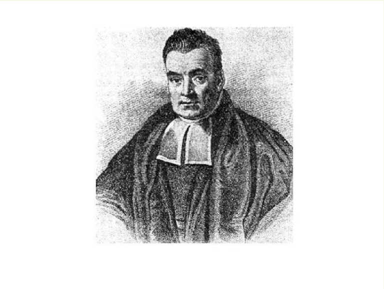


Thomas Bayes (1702–1761, 1763) followed this argument with Θ possessing the *uniform* distribution and X given $\Theta = \theta$ *binomial* (n, θ).

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

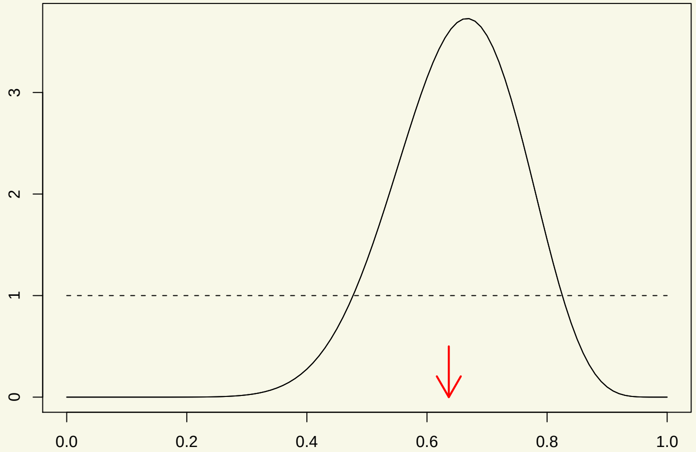


Reverend Thomas

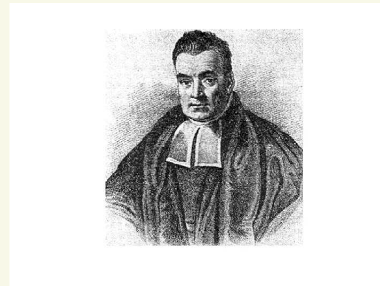


Thomas Bayes (1702–1761, 1763) followed this argument with Θ possessing the *uniform* distribution and X given $\Theta = \theta$ *binomial* (n, θ).

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

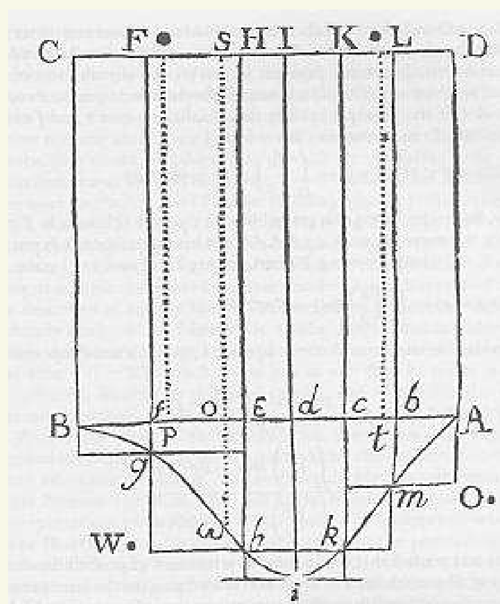


Reverend Thomas



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*$(X + 1, n - X + 1)$.



Parametric Bayes

Pierre-Simon Laplace (1749-1827) rediscovered Bayes' argument and applied it to general parametric models: models smoothly indexed by a Euclidean parameter θ.

For instance, the linear regression model, where one observes $(x_1, Y_n), \ldots, (x_n, Y_n)$ following

$$Y_i = \theta_0 + \theta_1 x_i + e_i,$$

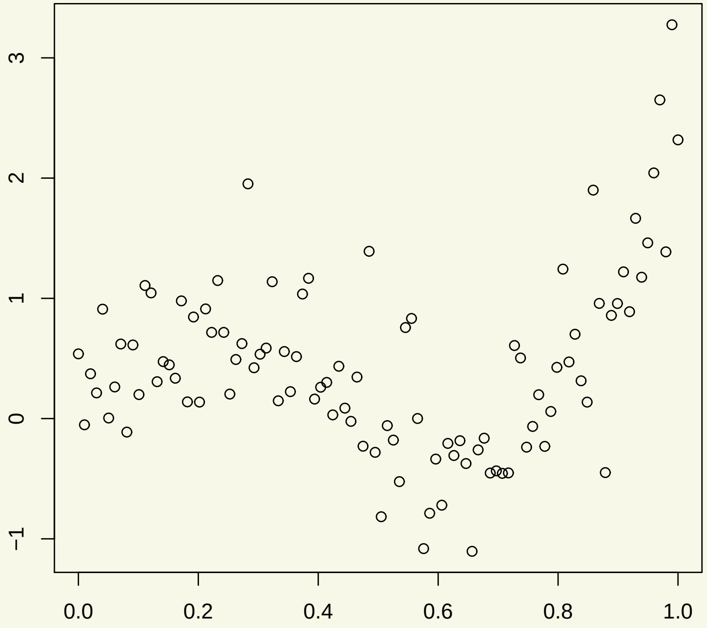
for $e_1, \ldots, e_n$ independent normal errors with zero mean.

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.

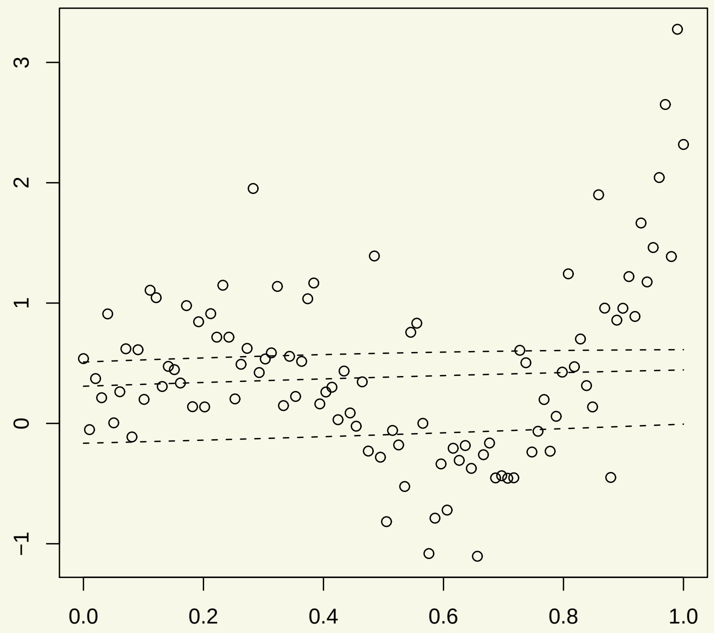


Nonparametric Bayes

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.

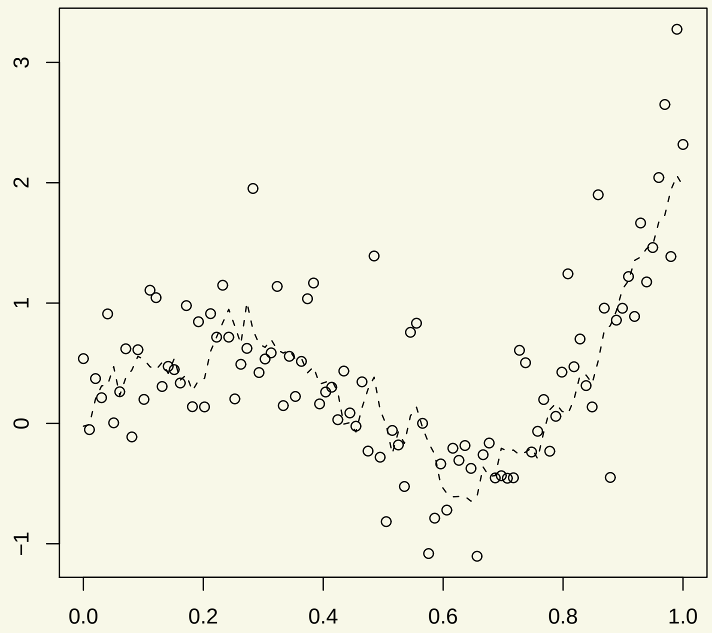


Nonparametric Bayes

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.

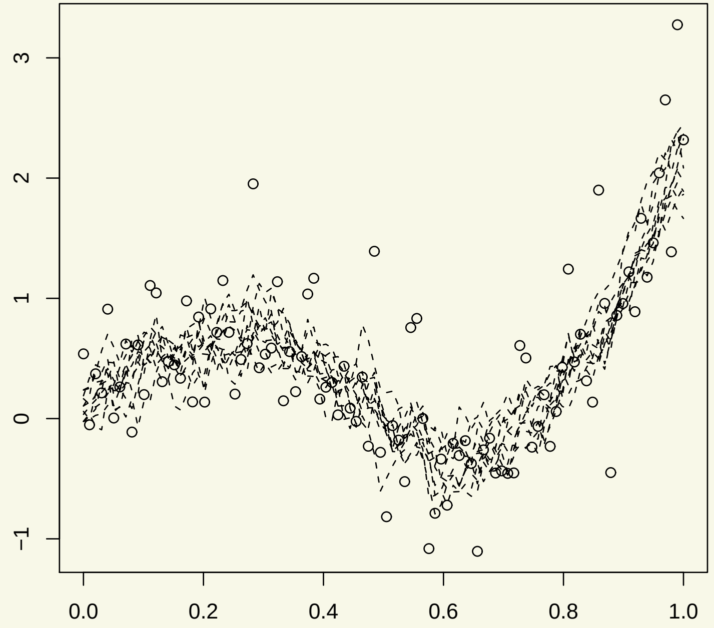


Nonparametric Bayes

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.

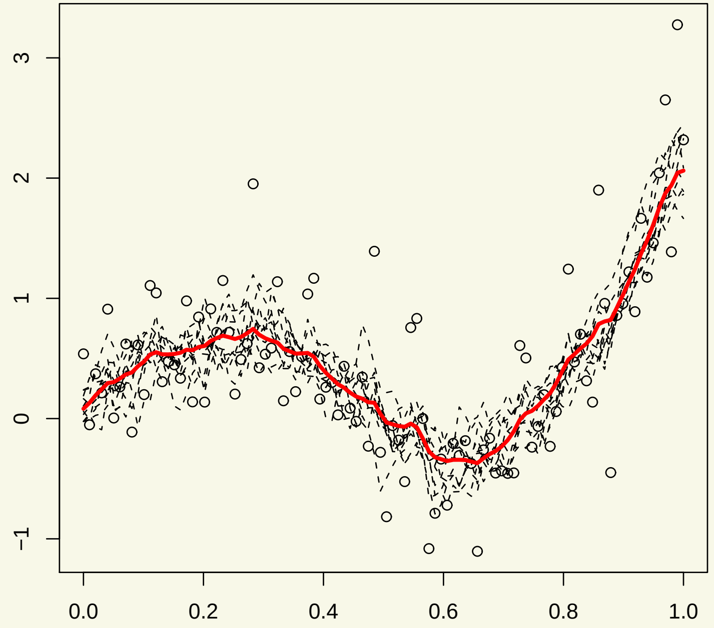


Nonparametric Bayes

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.

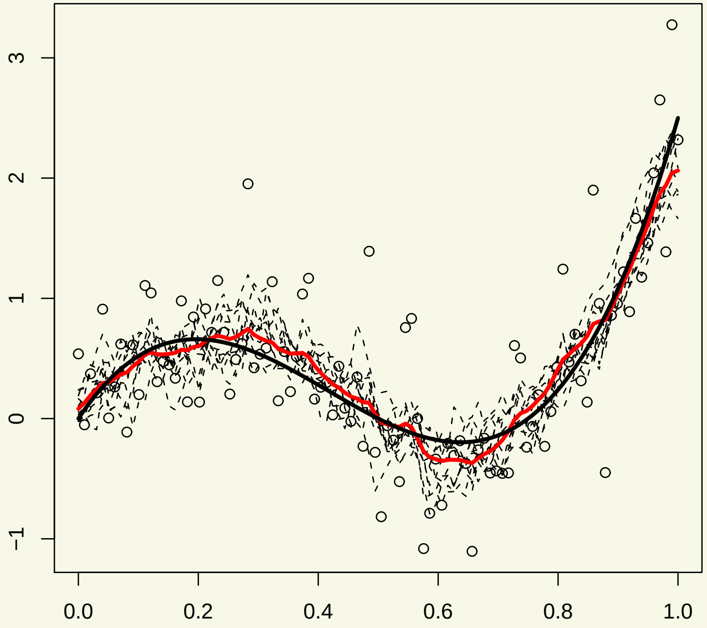


Nonparametric Bayes

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:

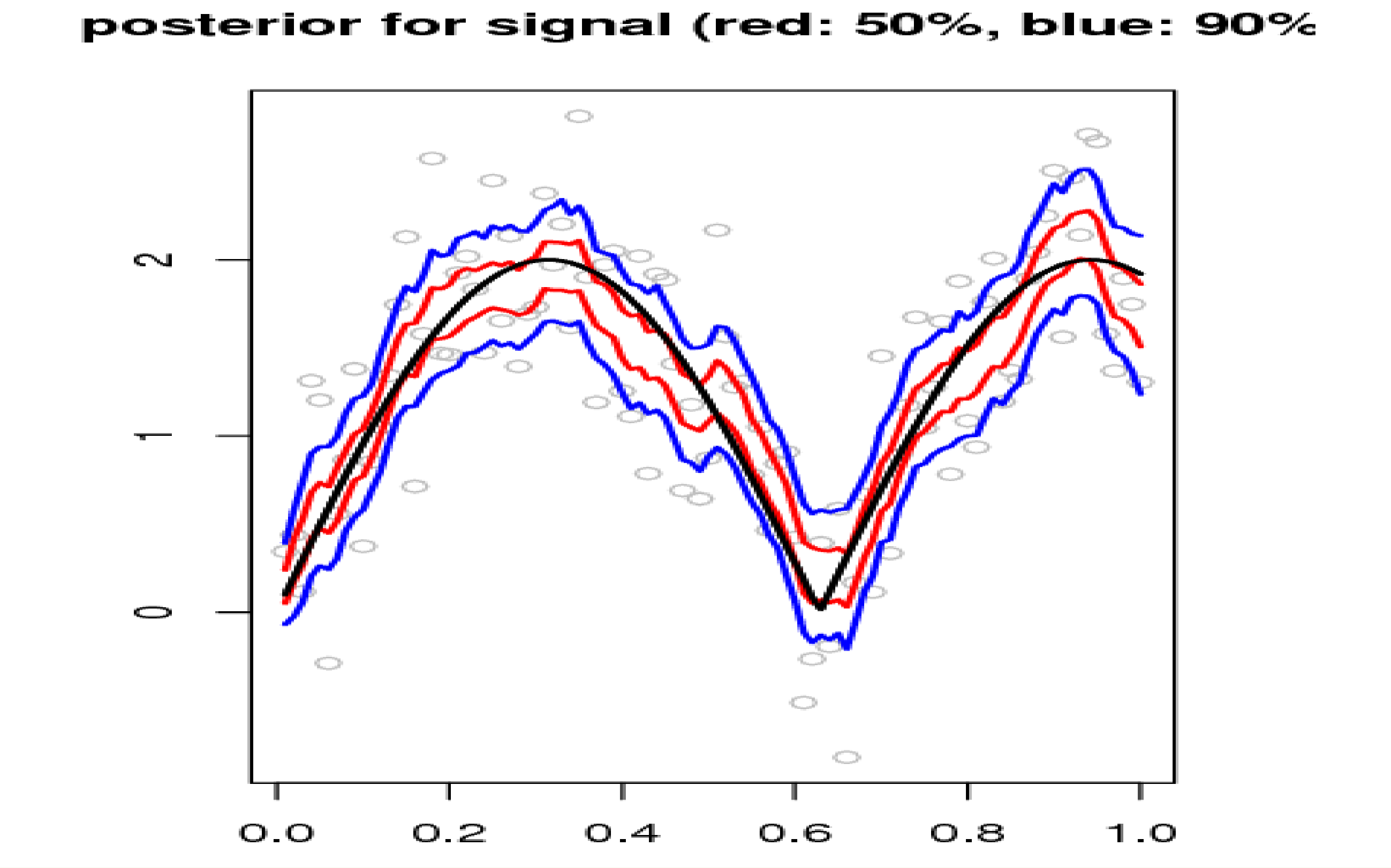
$$d\Pi(\theta|\, X) \propto p_\theta(X)\, d\Pi(\theta).$$

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



Credibility Bands

Posterior gives measure of uncertainty.



Subjectivism

A philosophical Bayesian statistician views the prior distribution as an expression of his personal beliefs on the state of the world, before gathering the data.

After seeing the data he updates his beliefs into the posterior distribution.

Most scientists do not like dependence on subjective priors.

- One can opt for objective or noninformative priors.
- One can also mathematically study the role of the prior, and hope to find that it is small.

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 this random measure 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.

Two desirable properties:

- Consistency + rate
- Adaptation

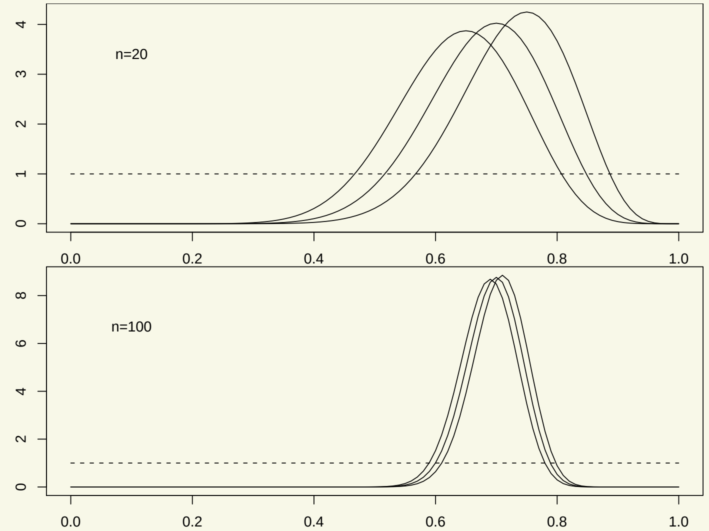
Parametric models

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 a vector $\theta \in \mathbb{R}^d$ (e.g. $\theta \mapsto \sqrt{p_\theta}$ continuously differentiable as map in $L_2(\mu)$).

THEOREM [Laplace, Bernstein, von Mises, LeCam 1989]
Under $P^n_{\theta_0}$-probability, for any prior with density that is positive around θ_0,

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

Here $\tilde\theta_n$ 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 a vector $\theta \in \mathbb{R}^d$ (e.g. $\theta \mapsto \sqrt{p_\theta}$ continuously differentiable as map in $L_2(\mu)$).

THEOREM [Laplace, Bernstein, von Mises, LeCam 1989]
Under $P^n_{\theta_0}$-probability, for any prior with density that is positive around θ_0,

$$\left\| \Pi(\cdot \mid X_1, \ldots, X_n) - N_d\Big(\tilde\theta_n, \frac{1}{n} I_{\theta_0}^{-1}\Big)(\cdot) \right\| \to 0.$$

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

In particular, the posterior distribution concentrates most of its mass on balls of radius $O(1/\sqrt{n})$ around θ_0.

The prior washes out completely.

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

We like $\varepsilon_n = \varepsilon_n(\theta_0)$ to tend to 0 fast, for every θ_0 in some model Θ.

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. For a good prior 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 $\theta\colon [0,1] \to \mathbb{R}$ based on data $(x_1, Y_1), \ldots, (x_n, Y_n)$, with

$$Y_i = \theta(x_i) + e_i, \qquad i = 1, \ldots, n,$$

for $e_1, \ldots, e_n$ independent "errors" drawn from a normal distribution with mean zero.

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

$$\|\theta\|_{C_\alpha} = \sup_x \bigl|\theta(x)\bigr| + \sup_{x \neq y} \frac{\bigl|\theta^{(\alpha - \underline{\alpha})}(x) - \theta^{(\alpha - \underline{\alpha})}(y)\bigr|}{|x-y|^\alpha}.$$

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

$$\varepsilon_{n,\alpha}^2 = \inf_T \sup_{\|\theta\|_{C_\alpha} \leq 1} \mathrm{E}_\theta \int_0^1 \bigl|T(x_1, Y_1, \ldots, x_n, Y_n)(s) - \theta(s)\bigr|^2 \, ds \asymp \Bigl(\frac{1}{n}\Bigr)^{2\alpha/(2\alpha+1)}.$$

Minimaxity and adaptation: other models

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

Gaussian process priors

Gaussian priors

A Gaussian random variable W with values in a (separable) Banach space $\mathbb{B}$ is a Borel measurable map from some probability space into $\mathbb{B}$ such that b^*W is normally distributed for every b^* in the dual space $\mathbb{B}^*$.

If the Banach space is a space of functions $w: T \to \mathbb{R}$, then W is usually written $W = (W_t: t \in T)$ and the map is often determined by the distributions of all vectors $(W_{t_1}, \ldots, W_{t_k})$, for $t_1, \ldots, t_k \in T$. These are determined by their mean vectors and the covariance function

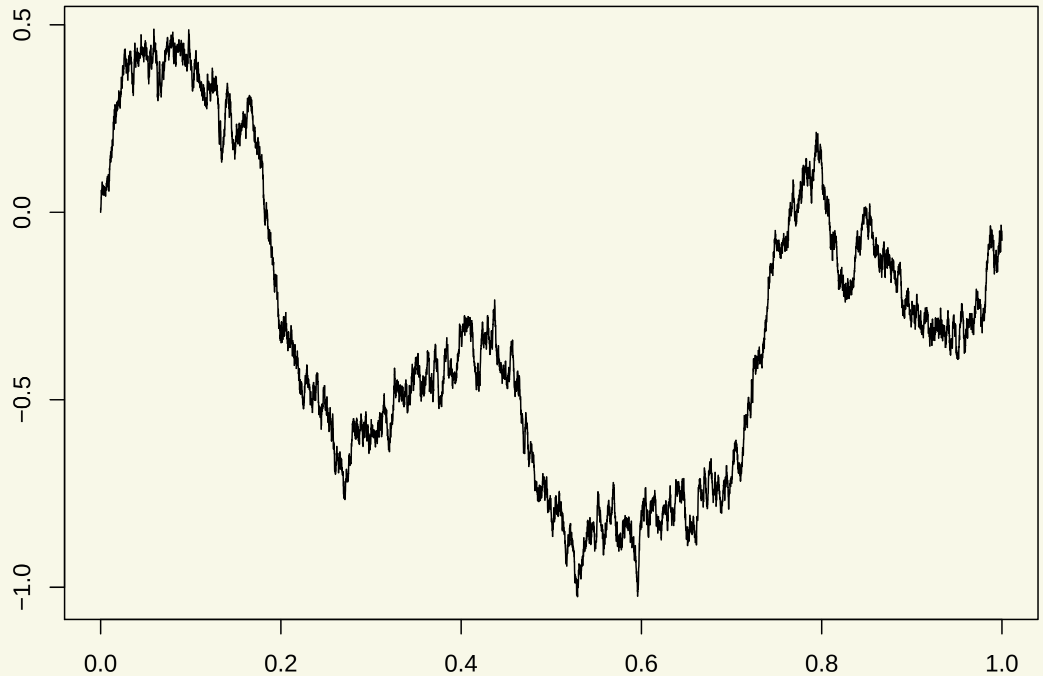
$$K(s,t) = \mathrm{E} W_s W_t, \qquad s, t \in T.$$

Gaussian priors have been found useful, because

- they offer great variety
- they are easy (?) to understand through their covariance function
- they can be computationally attractive (e.g. `www.gaussianprocess.org`)

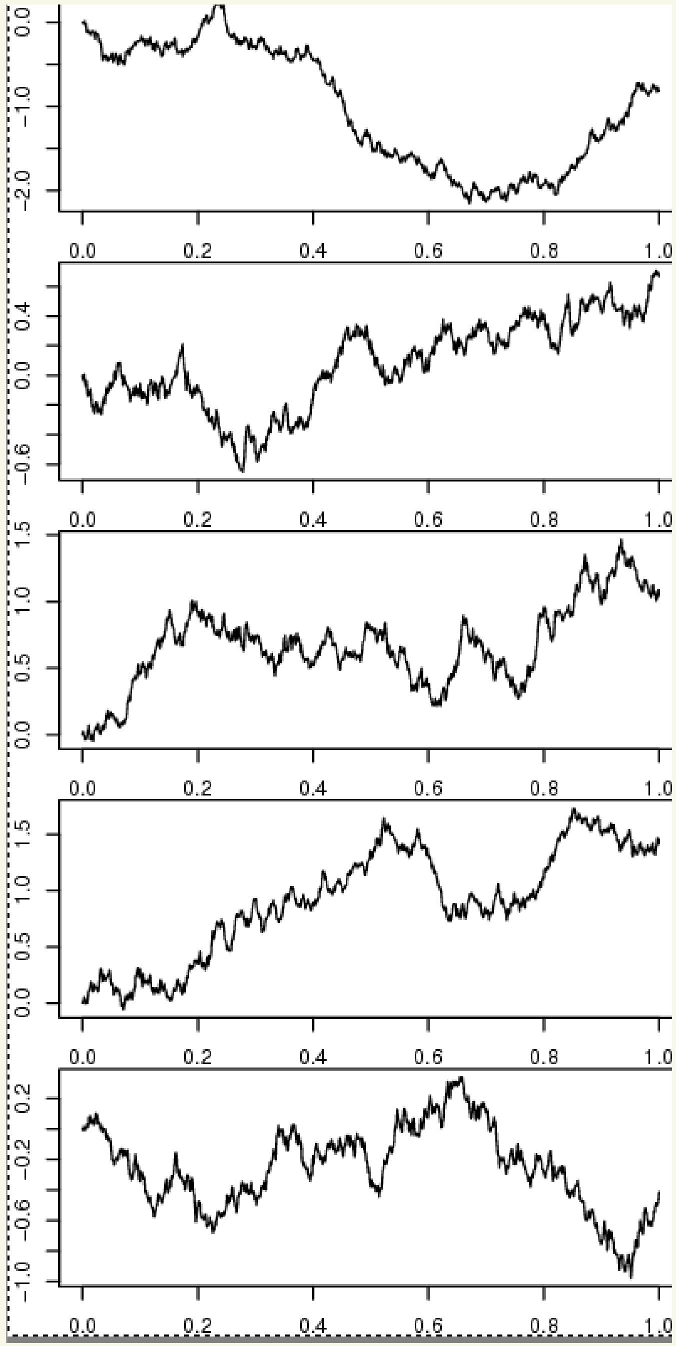
Example: Brownian motion

$$\mathrm{E}W_sW_t = s \wedge t, \qquad 0 \le s, t \le 1.$$



Brownian motion is usually viewed as map in $C[0,1]$.
It can be constructed so that it takes values in $C^\alpha[0,1]$ for every $\alpha < 1/2$ and also in $B^{1/2}_{1,\infty}[0,1]$.

Brownian motion—5 realizations



Brownian regression

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

$$Y_i = w_0(x_i) + e_i, \qquad i = 1, \ldots, n,$$

for $e_1, \ldots, e_n$ independent "errors" drawn from a normal distribution with mean zero.

THEOREM

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

Brownian regression

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

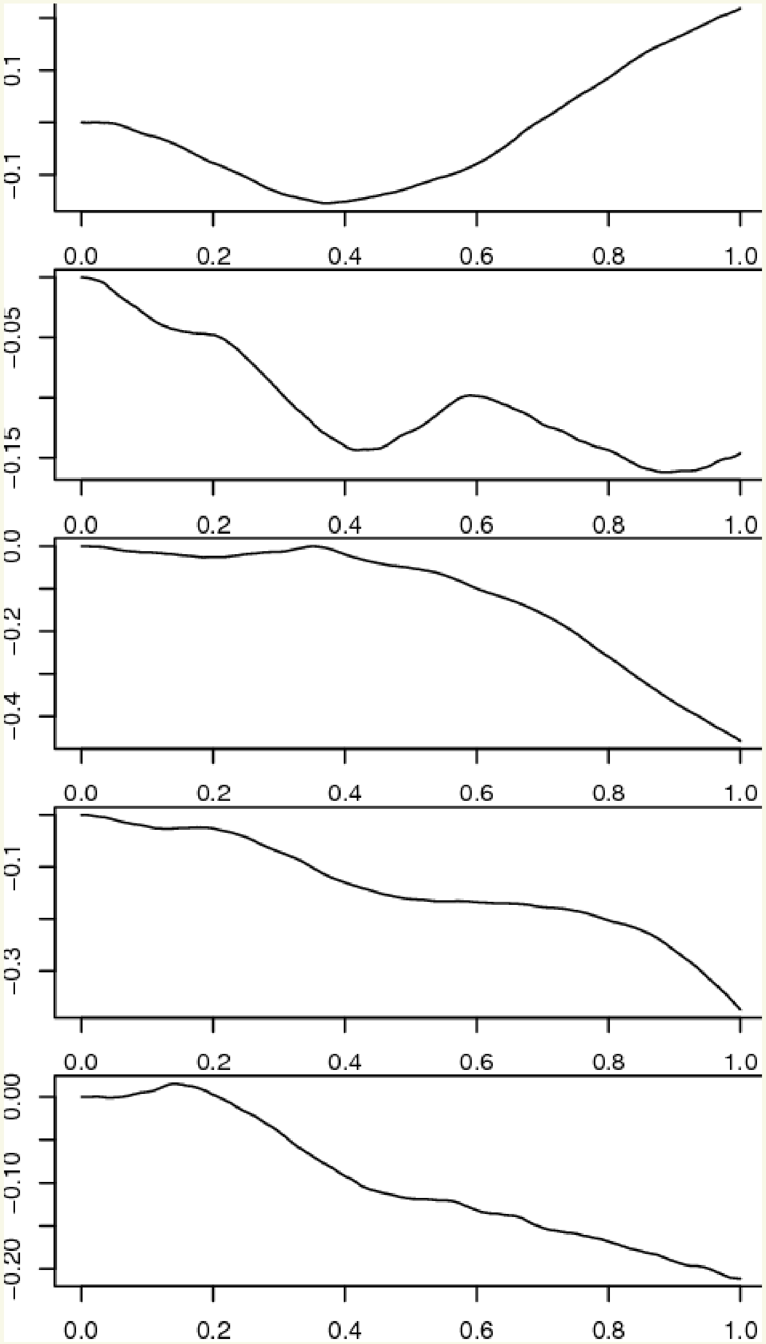
$$Y_i = w_0(x_i) + e_i, \qquad i = 1, \ldots, n,$$

for $e_1, \ldots, e_n$ independent "errors" drawn from a normal distribution with mean zero.

THEOREM
If $w_0 \in C^\alpha[0,1]$, 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 $\alpha > 0$.

Integrated Brownian motion — 5 realizations



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

If $w_0 \in C^\beta[0,1]$, then L_2-rate is: $n^{-\alpha/(2\alpha+1)}$ if $\beta \geq \alpha$;
$n^{-\beta/(2\alpha+1)}$ if $\beta \leq \alpha$.

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

Other priors

Fractional Brownian motion [Hurst index $0 < \alpha < 1$]:

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

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

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

Stationary processes: For a given "spectral measure" μ

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

Smoothness of $t \mapsto W_t$ can be controlled by the tails of μ. For instance, exponentially small tails give analytic sample paths.

Adaptation

Two methods for adaptation

The Gaussian priors considered so far possess itself a certain regularity, and are optimal iff this matches the regularity of the true regression function.

To obtain a prior that is suitable for estimating a function of unknown regularity $\alpha > 0$, there are two methods:

- Hierarchical prior
- Rescaling

Hierarchical priors

For each $\alpha > 0$ there are several good priors Π_α (Riemann-Liouville, Fractional, Series,...).

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

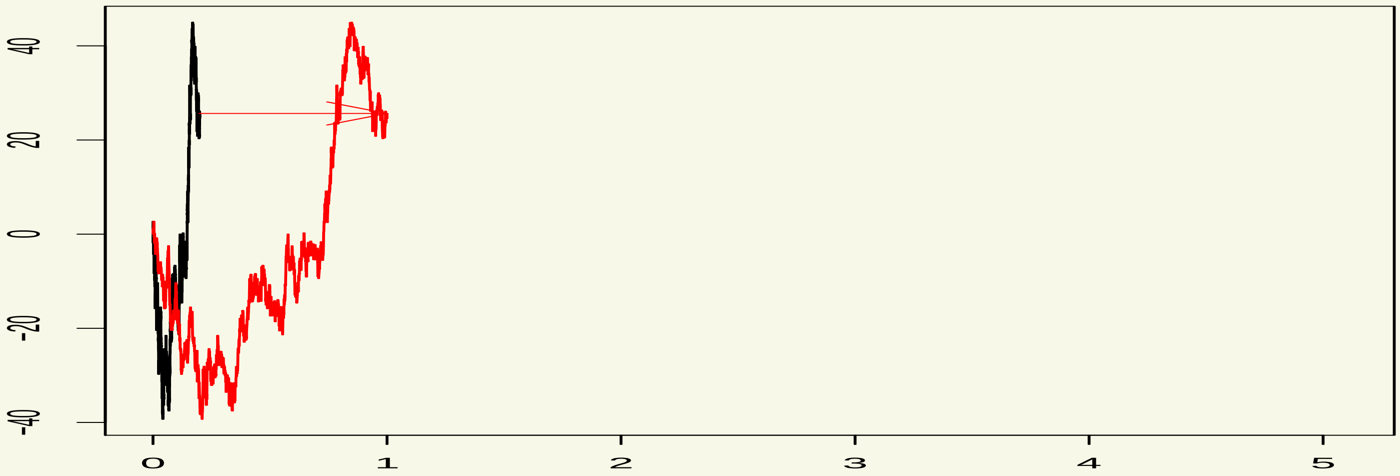
This gives a mixture prior

$$\Pi = \int \Pi_\alpha \, dw(\alpha).$$

Disadvantage: computations are expensive.

Rescaling

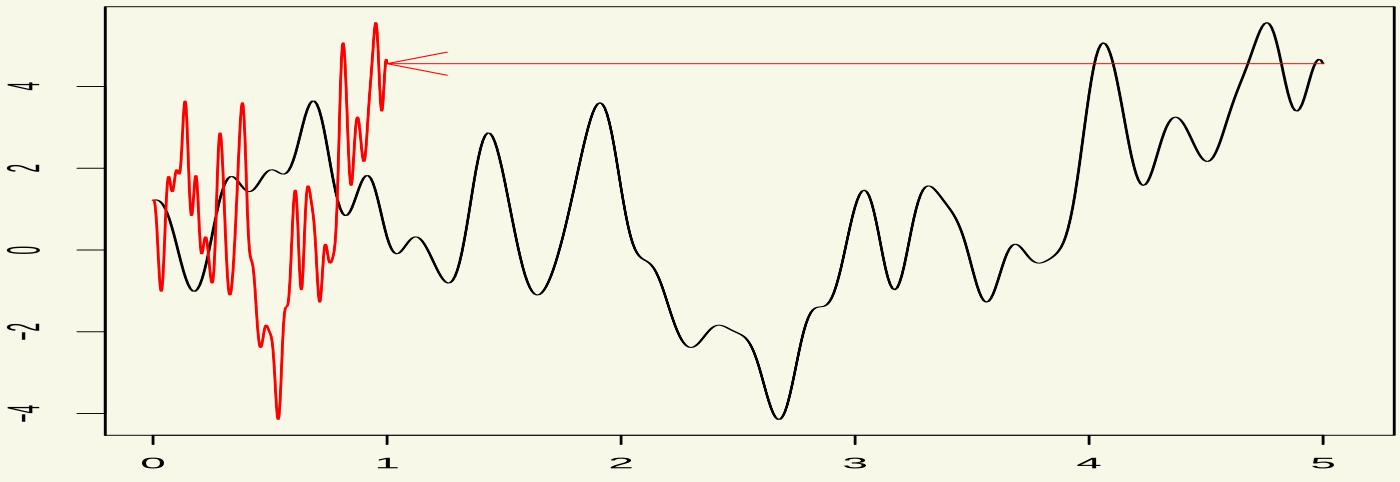
Sample paths can be smoothed by stretching



Rescaling

Sample paths can be smoothed by stretching

and roughened by shrinking



Rescaling (2)

It turns out that one can rescale k times integrated Brownian motion in such a way that it gives an appropriate prior for α-smooth functions, for any $\alpha \in (0, k + 1/2]$.

Similarly one any rescale (shrink) an analytic stationary Gaussian process, so that it becomes appropriate for α-smooth functions, for any $\alpha > 0$.

Unfortunately, the rescaling rate depends on α.

Adaptation by rescaling

- Choose c 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 $W_t \sim G_{t/c}$

THEOREM

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

Reverend Thomas solved the bandwidth problem!?

Determination of Rates

Two ingredients

Two ingredients:

- RKHS
- Small ball exponent

Reproducing kernel Hilbert space

W zero-mean Gaussian in $(\mathbb{B}, \|\cdot\|)$.

$S\colon \mathbb{B}^* \to \mathbb{B}$, $\quad Sb^* = \mathrm{E} W b^*(W)$.

DEFINITION RKHS $(\mathbb{H}, \|\cdot\|_{\mathbb{H}})$ is the completion of $S\mathbb{B}^*$ under

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

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

Reproducing kernel Hilbert space (2)

Any Gaussian random element in a separable Banach space can be represented as

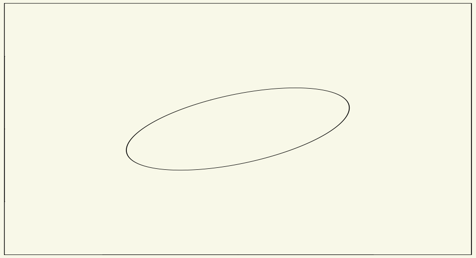
$$W = \sum_{i=1}^{\infty} \mu_i Z_i e_i$$

for

- $\mu_i \downarrow 0$
- $Z_1, Z_2, \ldots$ i.i.d. $N(0,1)$
- $\|e_1\| = \|e_2\| = \cdots = 1$

The RKHS consists of all elements $h := \sum_i h_i e_i$ with

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



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

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

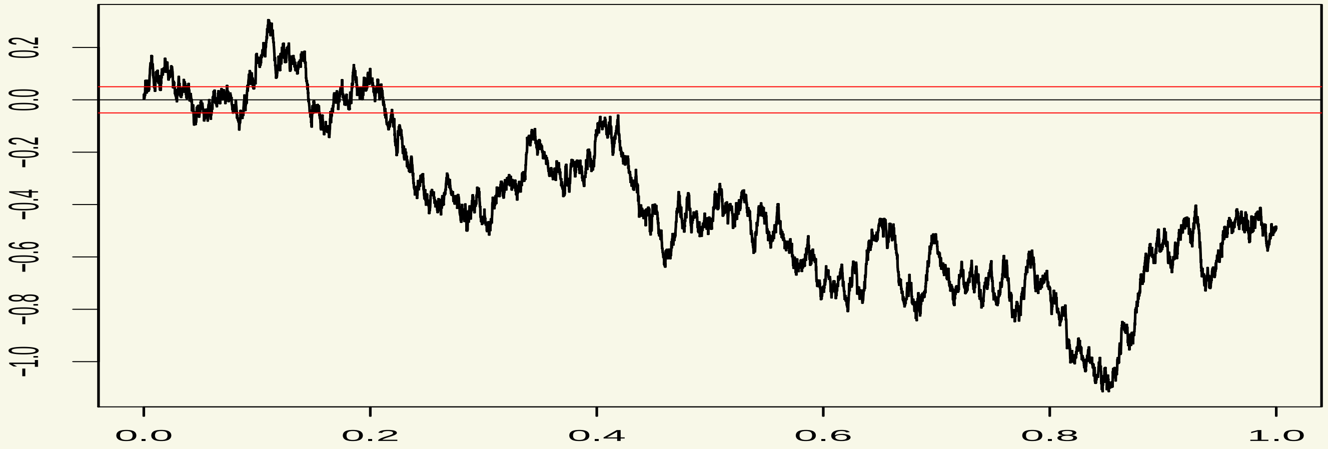
Small ball probability

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

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



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

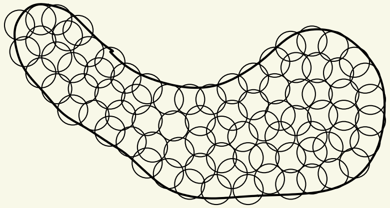
$$\phi_0(\varepsilon) = -\log \mathrm{P}(\|W\| < \varepsilon).$$

Computable for many examples, by probabilistic arguments, or using:

THEOREM [Kuelbs & Li 93]

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

$N\big(\varepsilon, B, \|\cdot\|\big)$ is the minimal number of balls of radius ε needed to cover B.



Basic result

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
The posterior rate is ε_n if

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

- Both inequalities give lower bound on ε_n.
- The first depends on W and not on w_0.

Example — Brownian motion

One-dimensional Brownian motion is a map in $C[0,1]$.

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

Example — Brownian motion

One-dimensional Brownian motion is a map in $C[0,1]$.

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

CONSEQUENCE:
The rate is never faster than the solution of

$$(1/\varepsilon_n)^2 \leq n\varepsilon_n^2$$

It also depends on the approximation of w_0 in uniform norm by functions from the first order Sobolev space, through

$$\inf_{h:\|w_0-h\|_\infty<\varepsilon_n} \|h'\|_2^2 \leq n\varepsilon_n^2.$$

Proof

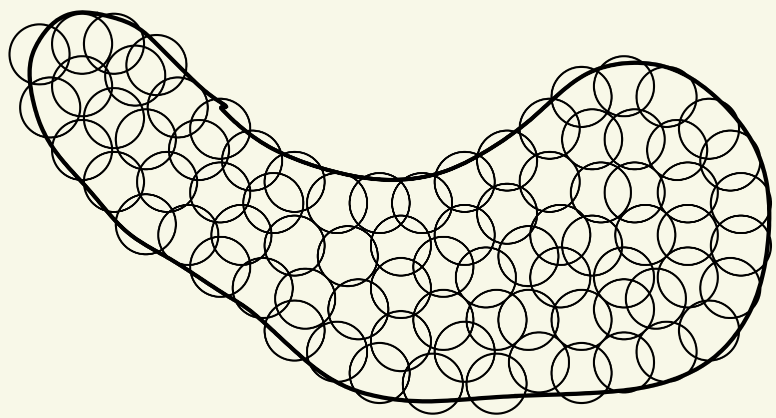
General results by Ghosal and vdV (2000, 2006) show that the rate of posterior contraction is ε_n if there exist sets $\mathbb{B}_n$ such that

(1) $\log N\big(\varepsilon_n, \mathbb{B}_n, \|\cdot\|\big) \leq n\varepsilon_n^2$ entropy

(2) $\Pi_n(\mathbb{B}_n) = 1 - o(e^{-3n\varepsilon_n^2})$

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

$N\big(\varepsilon, B, \|\cdot\|\big)$ is the minimal number of balls of radius ε needed to cover B.



Proof

General results by Ghosal and vdV (2000, 2006) show that the rate of posterior contraction is ε_n if there exist sets $\mathbb{B}_n$ such that

(1) $\log N\bigl(\varepsilon_n, \mathbb{B}_n, \|\cdot\|\bigr) \leq n\varepsilon_n^2$ entropy

(2) $\Pi_n(\mathbb{B}_n) = 1 - o(e^{-3n\varepsilon_n^2})$

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

$N\bigl(\varepsilon, B, \|\cdot\|\bigr)$ is the minimal number of balls of radius ε needed to cover B.

The interpretation of these conditions is that the prior should be "flat". By (1) we need $N\bigl(\varepsilon_n, \mathbb{B}_n, \|\cdot\|\bigr) \approx e^{n\varepsilon_n^2}$ balls to cover the model. If the mass is "uniformly spread" then every ball has mass as required by (3):

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

Proof

General results by Ghosal and vdV (2000, 2006) show that the rate of posterior contraction is ε_n if there exist sets $\mathbb{B}_n$ such that

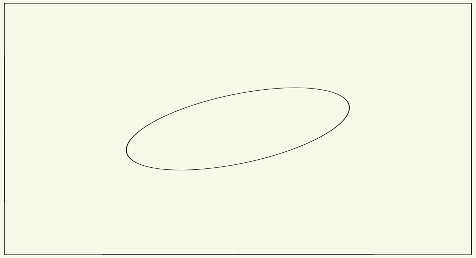
(1) $\log N\big(\varepsilon_n, \mathbb{B}_n, \|\cdot\|\big) \leq n\varepsilon_n^2$ entropy

(2) $\Pi_n(\mathbb{B}_n) = 1 - o(e^{-3n\varepsilon_n^2})$

(3) $\Pi_n\big(w: \|w - w_0\| < \varepsilon_n\big) \geq e^{-n\varepsilon_n^2}$ prior mass

$N\big(\varepsilon, B, \|\cdot\|\big)$ is the minimal number of balls of radius ε needed to cover B.

Existence of sets $\mathbb{B}_n$ can be verified using characterizations of the geometry of Gaussian measures.



Geometry

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

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

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

Geometry

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

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

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

EXAMPLE: One-dimensional Brownian motion is a map in $C[0,1]$.

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

$$\mathrm{P}\big(\|W - \{h: \|h'\|_2 \leq M\}\|_\infty > \varepsilon\big) \leq 1 - \Phi\big(\Phi^{-1}(e^{-1/\varepsilon^2}) + M\big).$$

Geometry

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

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

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

THEOREM [Kuelbs & Li 93]

$$\log \mathrm{P}\big(\|W - w_0\| < \varepsilon\big) \asymp \log \mathrm{P}\big(\|W\| < \varepsilon\big) - \inf_{h \in \mathbb{H}: \|h - w_0\| < \varepsilon} \|h\|_{\mathbb{H}}^2.$$

(up to factors 2)

Geometry

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

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

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

THEOREM [Kuelbs & Li 93]

$$\log \mathrm{P}\bigl(\|W - w_0\| < \varepsilon\bigr) \asymp \log \mathrm{P}\bigl(\|W\| < \varepsilon\bigr) - \inf_{h \in \mathbb{H}: \|h - w_0\| < \varepsilon} \|h\|_{\mathbb{H}}^2.$$

(up to factors 2)
For Brownian motion this is a consequence of Girsanov's formula

$$\frac{dP^{W+h}}{dP^W}(W) = e^{\int h'\, dW - \|h'\|_2^2/2}.$$

Bayesian curve estimation using Gaussian process priors

Aad van der Vaart
Vrije Universiteit Amsterdam

Utrecht, May 2009