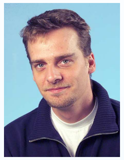
Some frequentist results about posterior distributions on infinite-dimensional parameter spaces
1+2

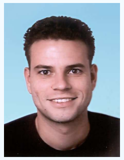
Aad van der Vaart
Vrije Universiteit Amsterdam

Göttingen, January 2009

Co-authors

 Bas Kleijn

Frank van der Meulen

 Harry van Zanten

 Ismael Castillo

Jyri Lember

 Subhashis Ghosal

 Willem Kruijer

Talk 1 — Contents

- Bayesian inference
- Examples of priors
- Frequentist Bayesian inference
- Some results

Talk 2 — Contents

- Rates — i.i.d. observations
- Rates — general
- Gaussian process priors — main result
- Gaussian process priors — settings
- Gaussian process priors — a proof

Bayesian inference

The Bayesian machine

- 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.

The Bayesian machine

- 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.

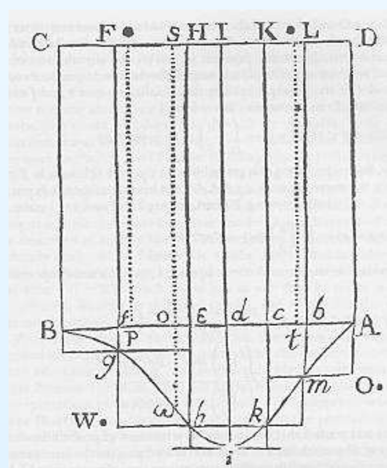
The prior expresses our uncertainty about the parameter.

The posterior expresses our remaining uncertainty after seeing the data.

The Reverend Thomas Bayes

Thomas Bayes followed this argument with Θ possessing the *Beta(1,1)* distribution and X given $\Theta = \theta$ *binomial* (n, θ).

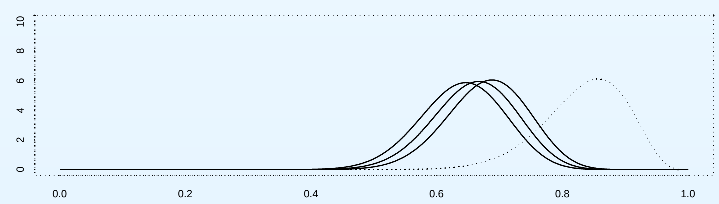
Using his famous rule he could compute that the posterior distribution is then *Beta*$(X + 1, n - X + 1)$.

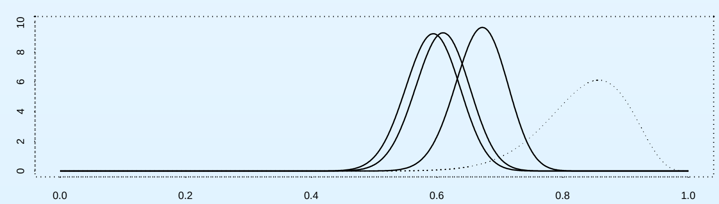


The Reverend Thomas Bayes

Thomas Bayes followed this argument with Θ possessing the *Beta(1,1)* distribution and X given $\Theta = \theta$ *binomial* (n, θ).

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

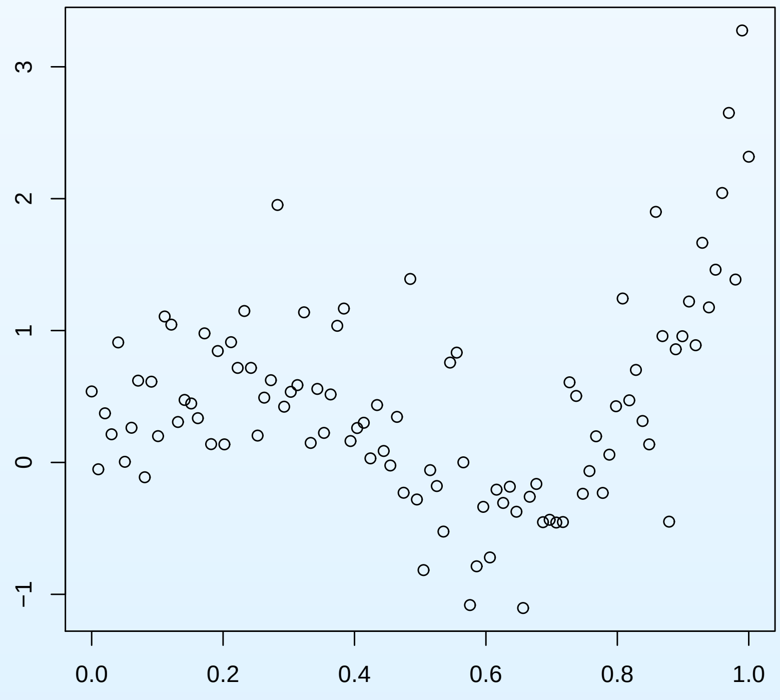




Nonparametric Bayes

If the parameter θ is a function, then the prior is a probability distribution on a function space.
So is the posterior, given the data.

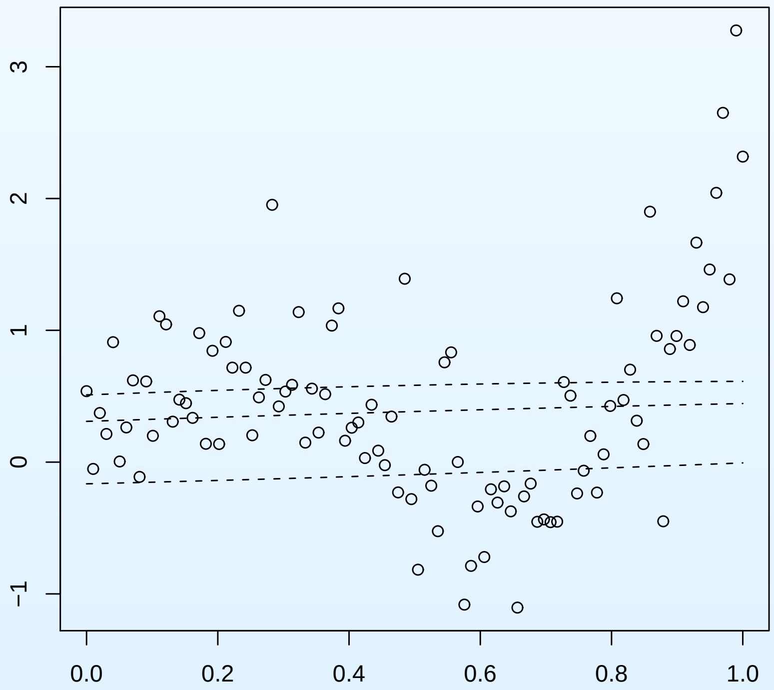
Prior and posterior are typically visualized by plotting functions that are simulated from these distributions.



Nonparametric Bayes

If the parameter θ is a function, then the prior is a probability distribution on a function space.
So is the posterior, given the data.

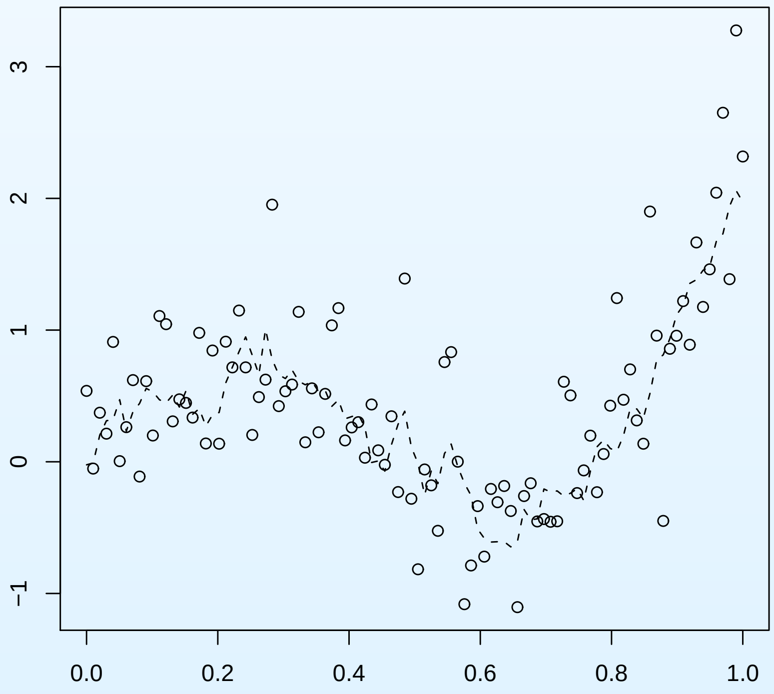
Prior and posterior are typically visualized by plotting functions that are simulated from these distributions.



Nonparametric Bayes

If the parameter θ is a function, then the prior is a probability distribution on a function space.
So is the posterior, given the data.

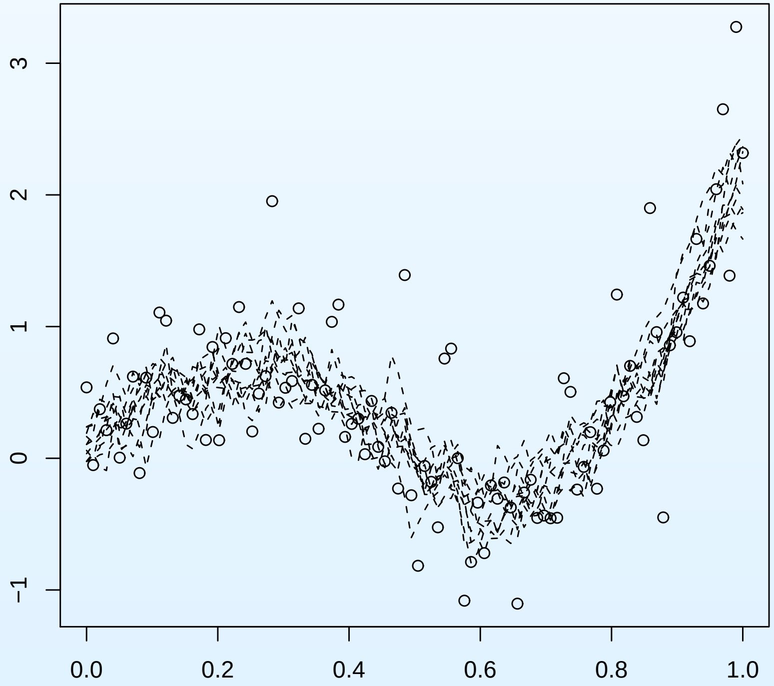
Prior and posterior are typically visualized by plotting functions that are simulated from these distributions.



Nonparametric Bayes

If the parameter θ is a function, then the prior is a probability distribution on a function space.
So is the posterior, given the data.

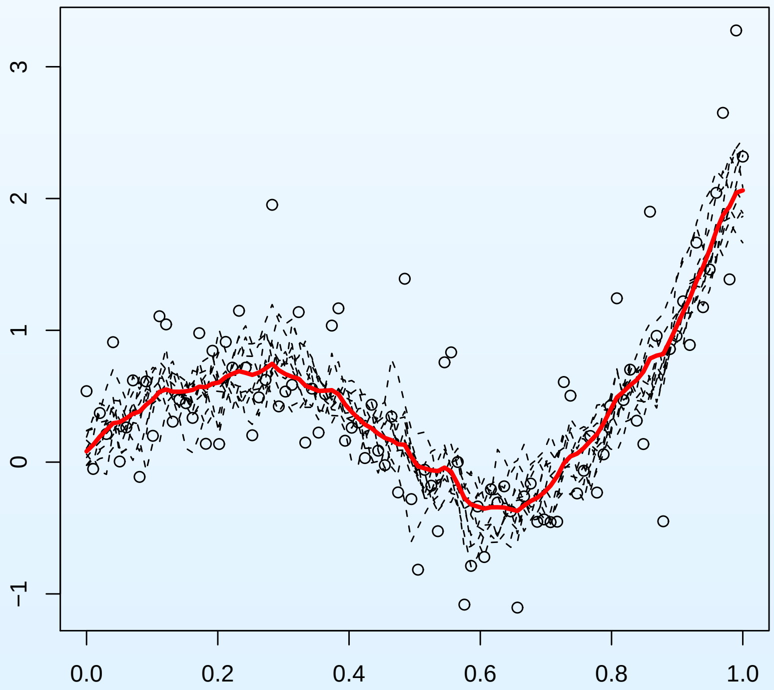
Prior and posterior are typically visualized by plotting functions that are simulated from these distributions.



Nonparametric Bayes

If the parameter θ is a function, then the prior is a probability distribution on a function space.
So is the posterior, given the data.

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

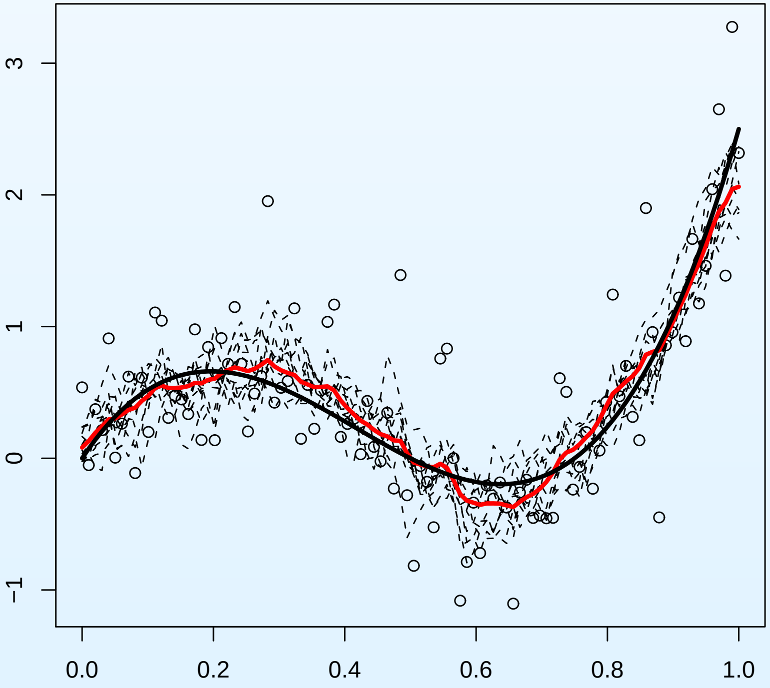


Nonparametric Bayes

If the parameter θ is a function, then the prior is a probability distribution on a function space.

So is the posterior, given the data.

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



Why Bayesian?

If you are a Bayesian, then you find this a stupid question.

If you are an ordinary person, then you might like Bayesian methods, because:

- they work better
- they are more elegant
- they allow to incorporate prior information better
- they are easier to implement
- they are computationally efficient

Why Bayesian?

If you are a Bayesian, then you find this a stupid question.

If you are an ordinary person, then you might like Bayesian methods, because:

- they work better [NO]
- they are more elegant [YES]
- they allow to incorporate prior information better [YES]
- they are easier to implement [SOMETIMES]
- they are computationally efficient [NO]

Computation

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

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

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

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

Computation

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

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

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

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

Time-consuming, must be tuned properly, many short-cuts suggested.

Computation (2) — MCMC

A Markov chain $\theta_1, \theta_2, \ldots$ is a sequence of random variables such that the distribution of θ_{k+1} given $\theta_1, \ldots, \theta_k$ depends only on θ_k. A distribution Π is stationary if every θ_i is marginally distributed according to Π.

Two important MCMC algorithms

- Metropolis-Hastings: given θ_k generate $\tilde{\theta}_{k+1}$ from some $Q(\cdot \,|\, \theta_k)$ and set $\theta_{k+1} = \tilde{\theta}_{k+1}$ with probability $\alpha_{Q,\Pi}(\theta_k, \tilde{\theta}_{k+1})$ and $\theta_{k+1} = \theta_k$ otherwise.

- Gibbs: for multivariate $\theta_{k+1} = (\theta_{k+1,1}, \ldots, \theta_{k+1,d})$ simulate one coordinate $\theta_{k+1,i}$ at a time from its conditional distribution given the other current coordinates.

Typically only approximately stationary, as it is impossible to simulate θ_1 correctly, whence burn-in is necessary.

Computation (3) — Hierarchical priors

Many priors are defined by a hierarchy of the type:

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

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

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

Regularization

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

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

The posterior mode maximizes

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

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

Bayesian thinking suggests penalties.

Bayesian inference gives a full posterior distribution.

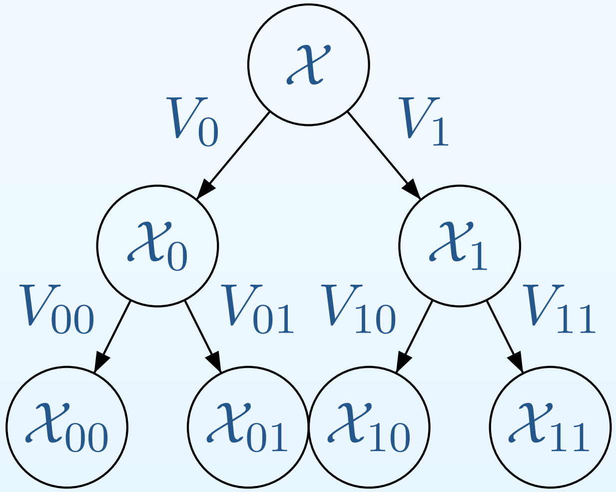
Examples of priors

Polya trees and Dirichlet process

Given a sequence of binary partitions:

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

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



The Dirichlet process prior is the special case that the parameters of V_ε are $\big(\alpha(\mathcal{X}_{\varepsilon 0}), \alpha(\mathcal{X}_{\varepsilon 1})\big)$ for a fixed measure α, the mean measure. It puts mass on discrete measures only.

Dirichlet mixtures

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

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

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

This is often formulated in a Bayesian hierarchy:

- μ and τ are chosen from priors.
- P is chosen from a Dirichlet with mean measure $N(\mu,\tau)$.
- $Z_1,\ldots,Z_n$ are chosen i.i.d. from P.
- σ is chosen from an inverse Gamma.
- $\varepsilon_1,\ldots,\varepsilon_n$ are i.i.d. from $N(0,1)$.
- Observations $X_i = Z_i + \sigma\varepsilon_i$.

Dirichlet mixtures — computation

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

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

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

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

Gaussian priors

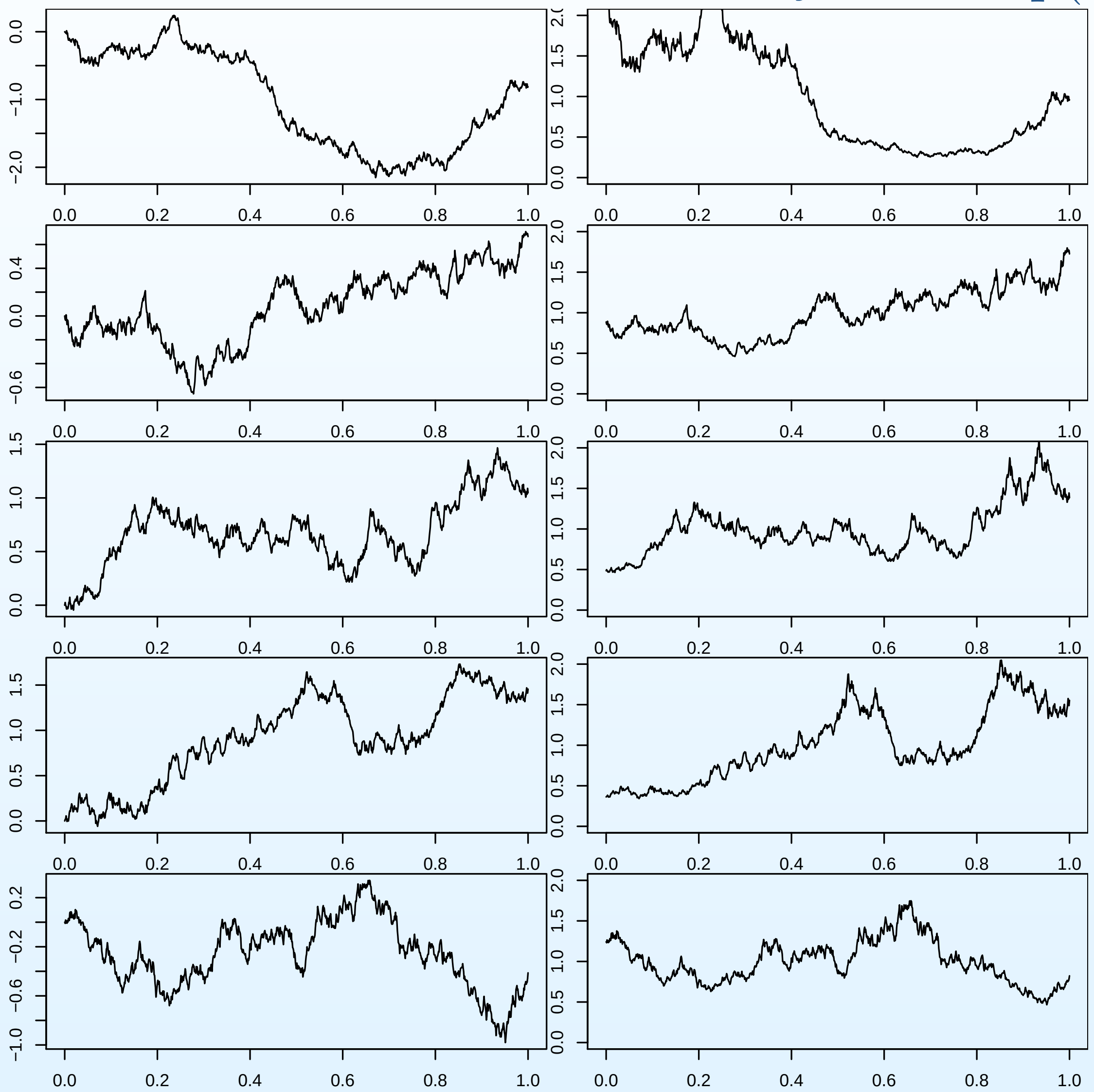
The law of a stochastic process $(W_t : t \in T)$ is a prior distribution on the space of functions $w : T \to \mathbb{R}$

Gaussian processes have been found useful, because

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

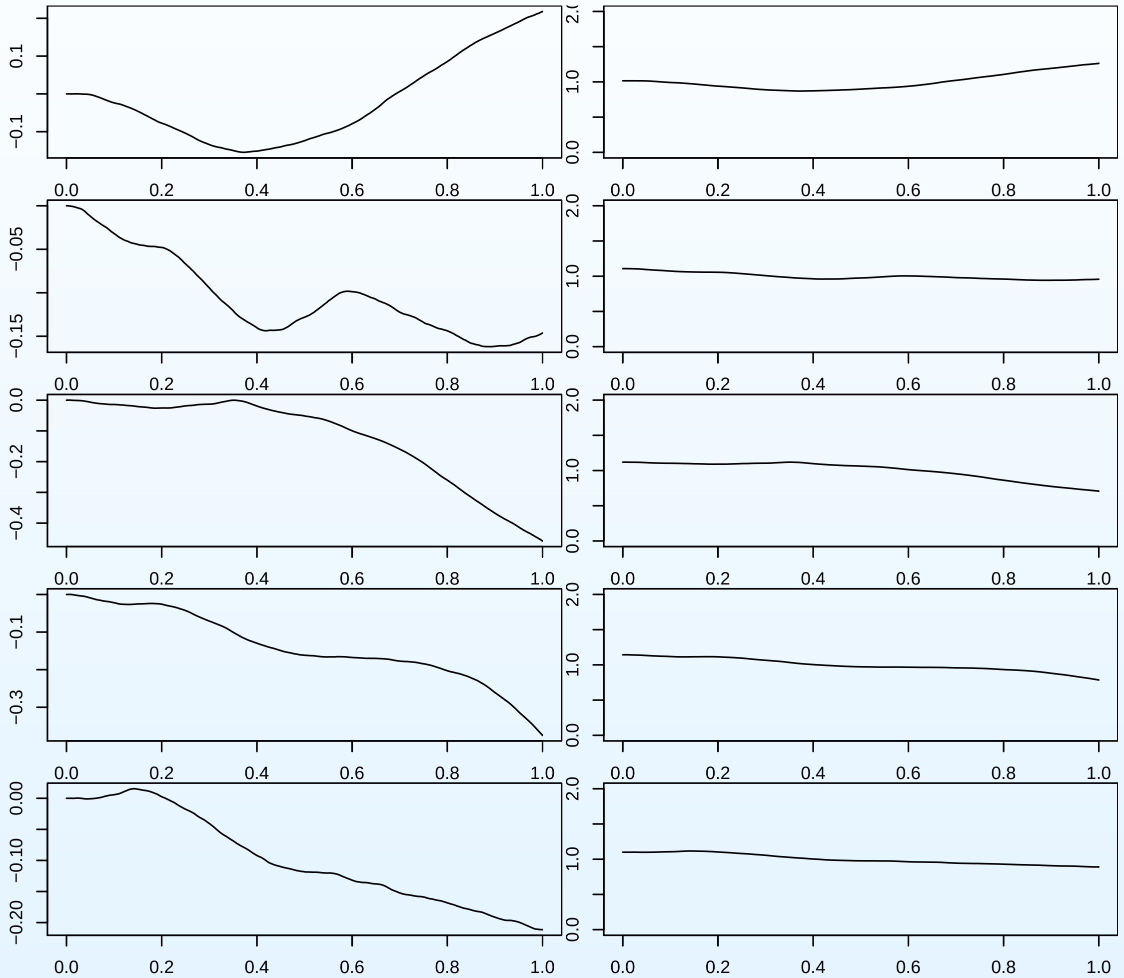
Gaussian processes

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



Gaussian processes

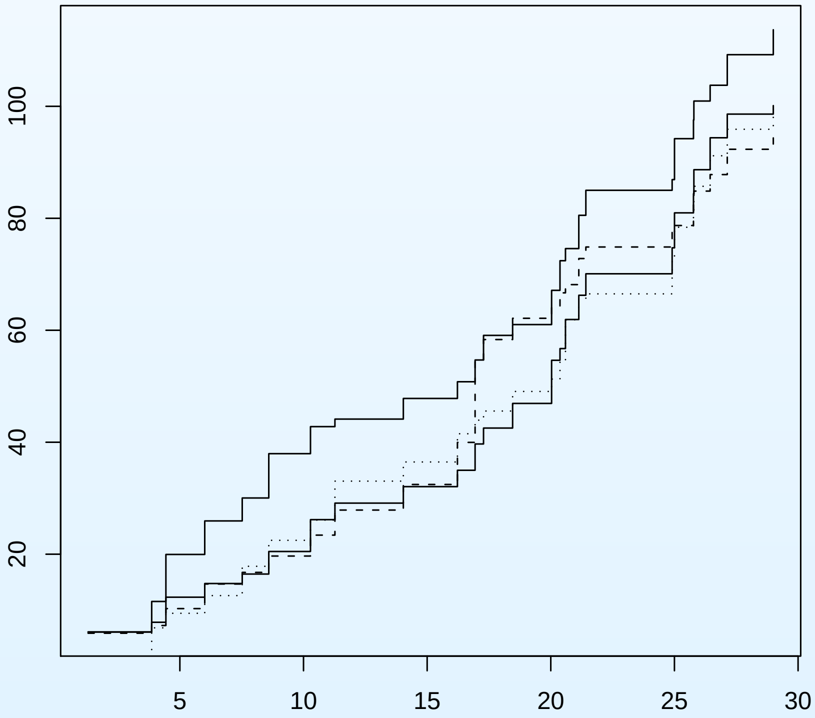
Integrated Brownian motion — Prior density



Independent increment processes

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

For better results we need more jumps, as in Lévy processes or general independent increment processes.



Sparsity (1)

Parameter $\theta = (\theta_1, \dots, \theta_n) \in \mathbb{R}^n$. We think only few θ_i are nonzero.

Prior on $\theta \in \mathbb{R}^n$:

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

We can build in more a-priori knowledge, e.g. to model genetic networks in micro-array analysis.

Sparsity (2)

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

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

Series priors

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

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

A practical approach is to choose $\theta_{k+1}, \theta_{k+2}, \ldots$ zero for some randomly chosen k.

Adaptation

Nonparametric estimation often works with scales of regularity classes. For instance, functions having $\alpha > 0$ derivatives (bounded by a given constant).

For a given α there are many appropriate priors Π_α.

Put prior w on α and next given α use Π_α, yielding the overall prior

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

This should solve the bandwidth problem.

Frequentist Bayesian theory

Frequentist Bayesian

If you are a Bayesian, then you worry

- about using the "right" prior
- about computation of the posterior.

If you are an ordinary person, then you worry about this too AND

- you can study the posterior as a random measure from a frequentist point of view:

You assume that the data X is generated according to a given parameter θ_0 and want the posterior $\Pi(\theta \in \cdot \mid X)$ to put "most" of its mass near θ_0 for "most" X.

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 $\theta \in \mathbb{R}^d$.

THEOREM [Bernstein, von Mises, LeCam,..]
Under $P^n_{\theta_0}$-probability, for any prior with density that is positive around θ_0, for $\tilde\theta_n = \theta_0 + n^{-1}\sum_{i=1}^n I_{\theta_0}^{-1}\dot\ell_{\theta_0}(X_i)$,

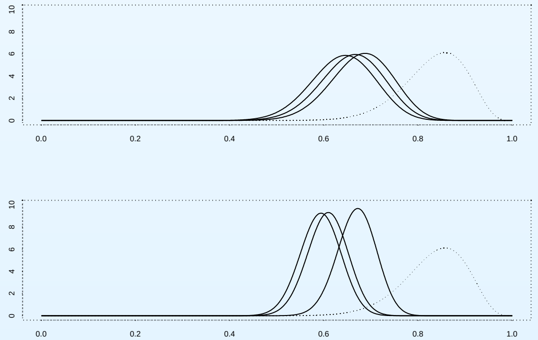
$$\Big\|\Pi(\cdot|\, X_1, \ldots, X_n) - N_d\big(\tilde\theta_n, \frac{1}{n} I_{\theta_0}^{-1}\big)(\cdot)\Big\| \to 0.$$

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 $\theta \in \mathbb{R}^d$.

THEOREM [Bernstein, von Mises, LeCam,..]
Under $P^n_{\theta_0}$-probability, for any prior with density that is positive around θ_0, for $\tilde\theta_n = \theta_0 + n^{-1}\sum_{i=1}^n I_{\theta_0}^{-1}\dot\ell_{\theta_0}(X_i)$,

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



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 $\theta \in \mathbb{R}^d$.

THEOREM [Bernstein, von Mises, LeCam,..]
Under $P^n_{\theta_0}$-probability, for any prior with density that is positive around θ_0, for $\tilde\theta_n = \theta_0 + n^{-1}\sum_{i=1}^n I_{\theta_0}^{-1}\dot\ell_{\theta_0}(X_i)$,

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

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

Semi- or nonparametric models

Does Bayes do a good job for infinite-dimensional models too?

Does the posterior contract to the truth at a good rate?

Does the posterior adapt to unknown regularity?

Does the posterior detect sparsity?

Complete class theorem

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

Complete class theorem

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

Complete class theorem

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

Which priors?

Complete class theorem

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

Which priors?

Most priors do not work! [Freedman and Diaconis, 1970s/80s]

Rate of contraction

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

DEFINITION

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

Distributional convergence

The posterior of a "parameter" $\phi(\theta)$ is obtained from the posterior for θ by marginalization.

For $\phi(\theta) \in \mathbb{R}$ we may hope to obtain distributional approximations, such as the Bernstein-von Mises theorem:

$$\Pi\Big(\phi(\theta) \in \cdot \,|\, X^{(n)}\Big) - N\Big(\Delta_n(X^{(n)}), \frac{\Sigma}{n}\Big)(\cdot) \xrightarrow{P} 0.$$

$\Delta_n(X^{(n)})$ and Σ defined from the efficient score function.

For nonregular parameters we expect a nonnormal distribution instead.

Minimaxity and adaptation

To a given regularity class is attached an optimal rate of convergence defined by the minimax criterion.

Minimaxity and adaptation

To a given regularity class is attached an optimal rate of convergence defined by the minimax criterion.

We like the posterior to contract at this rate.

Minimaxity and adaptation

To a given regularity class is attached an optimal rate of convergence defined by the minimax criterion.

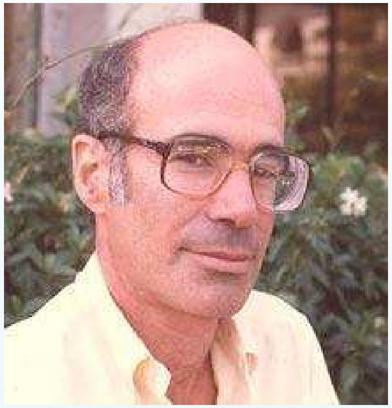
We like the posterior to contract at this rate.

Given a scale of regularity classes, indexed by a parameter α, we like the posterior to adapt: if the true parameter has regularity α, then we like the contraction rate to be the minimax rate for the α-class.

General findings

In infinite-dimensional situations the performance does depend on the prior.

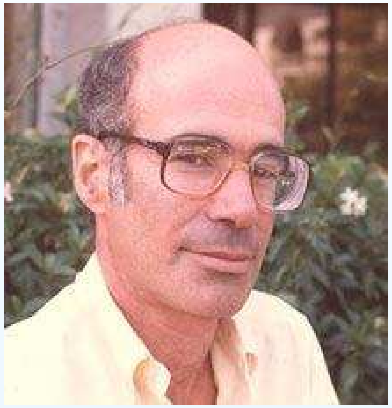
The prior does not wash out as $n \longrightarrow \infty$.



Bayesians, too, need to proceed with caution in the infinite-dimensional case, unless they are convinced of the fine details of their priors. Indeed, the consistency of their estimates and the coverage probability of their confidence sets depend on the details of their priors. [DAVID FREEDMAN, 1999.]

General findings

In infinite-dimensional situations the performance does depend on the prior.
The prior does not wash out as $n \to \infty$.



Bayesians, too, need to proceed with caution in the infinite-dimensional case, unless they are convinced of the fine details of their priors. Indeed, the consistency of their estimates and the coverage probability of their confidence sets depend on the details of their priors. [DAVID FREEDMAN, 1999.]

The good news: with a correct prior a Bayesian method works as well as the best nonBayesian method, it does adapt, and it does detect sparsity.

Some results

Dirichlet mixtures

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

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

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

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

Dirichlet mixtures

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

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

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

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

Conjecture: if $p_0 \in C^\alpha$ and the prior on σ is fixed with sufficient mass near 0, then rate is $(\log n)^\lambda/n^{-\alpha/(2\alpha+1)}$.

Adaptation — general

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

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

THEOREM [Lember&vdV 07]
Under general conditions the posterior rate is at least $\varepsilon_{n,\beta}$ if the true parameter belongs to model β.

Under more complicated conditions similar results hold for more general weights $w_{n,\alpha}$. There are also elegant special constructions. [See later.]

Misspecification

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

THEOREM Kleijn & vdV, 2006
Under general conditions the posterior contracts to the parameter “in the support” at minimal Kullback-Leibler divergence to the true parameter, at a rate as if it were “in the support”.

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

Brownian density estimation (Toy example)

- $X_1, \ldots, X_n$ i.i.d. from density p_0 on $[0,1]$
- $(W_x \colon x \in [0,1])$ Brownian motion

Prior on p:

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

THEOREM [vdV & van Zanten 07, Castillo 08]
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 (Toy example)

- $X_1, \ldots, X_n$ i.i.d. from density p_0 on $[0,1]$
- $(W_x : x \in [0,1])$ Brownian motion

Prior on p:

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

THEOREM [vdV & van Zanten 07, Castillo 08]
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 minimax optimal if and only if $\alpha = 1/2$.
- Rate does not improve if α increases from $1/2$.
- Consistency for any $\alpha > 0$.

Brownian density estimation (Toy example)

- $X_1, \ldots, X_n$ i.i.d. from density p_0 on $[0, 1]$
- $(W_x : x \in [0, 1])$ Brownian motion

Prior on p:

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

THEOREM [vdV & van Zanten 07, Castillo 08]
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 minimax optimal if and only if $\alpha = 1/2$.
- Rate does not improve if α increases from $1/2$.
- Consistency for any $\alpha > 0$.

Similar results hold for Gaussian regression, with w_0 the true regression function.

Other Gaussian priors

Integrated Brownian motion (released at zero) is an optimal prior if $w_0 \in C^\alpha[0,1]$ for $\alpha = 3/2$.

More generally $(\alpha - 1/2)$ times (fractionally) integrated Brownian motion (released at zero) is an optimal prior if $w_0 \in C^\alpha[0,1]$.

Alternative optimal priors can be constructed from fractional Brownian motion or by using series expansions.

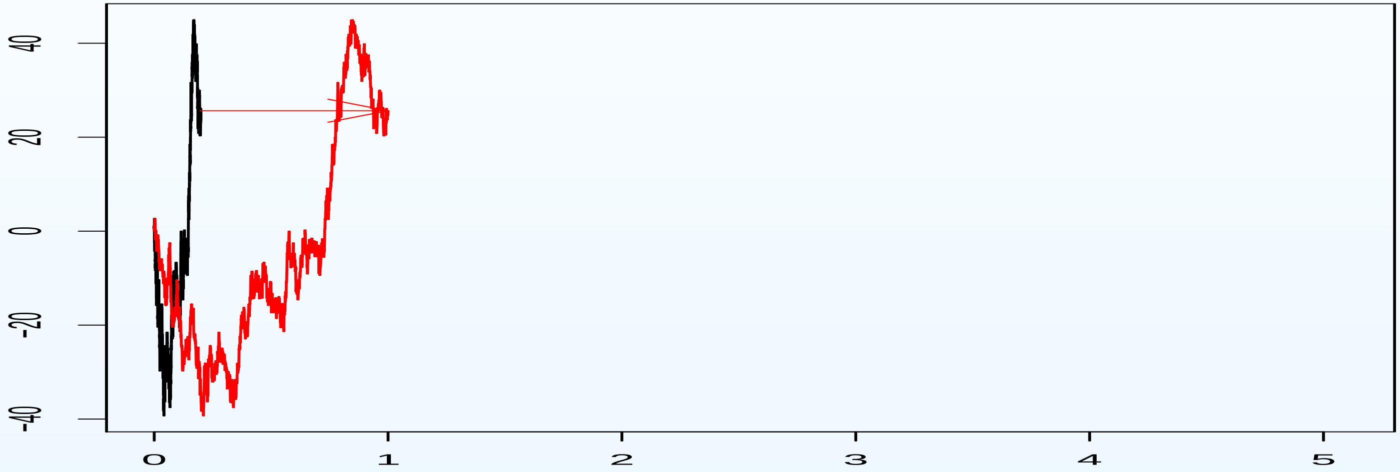
Stationary priors correspond to centered Gaussian processes G with

$$\mathrm{E} G_s G_t = \psi(s-t).$$

Appropriate smoothness obtained by consideration of the tail of $\hat{\psi}$.

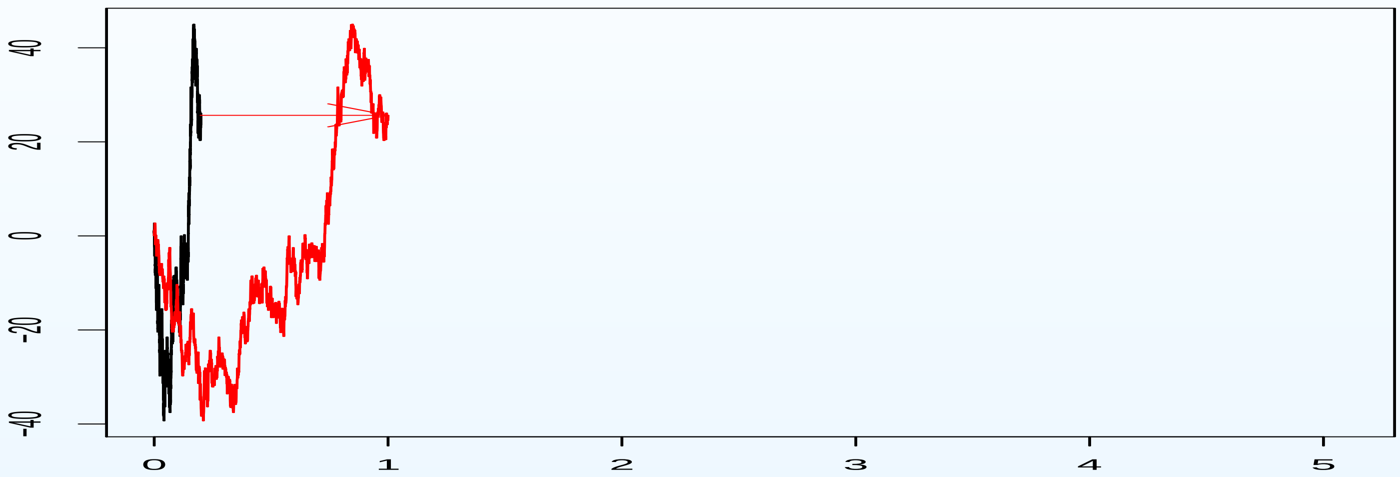
Rescaling

Sample paths can be smoothed by stretching

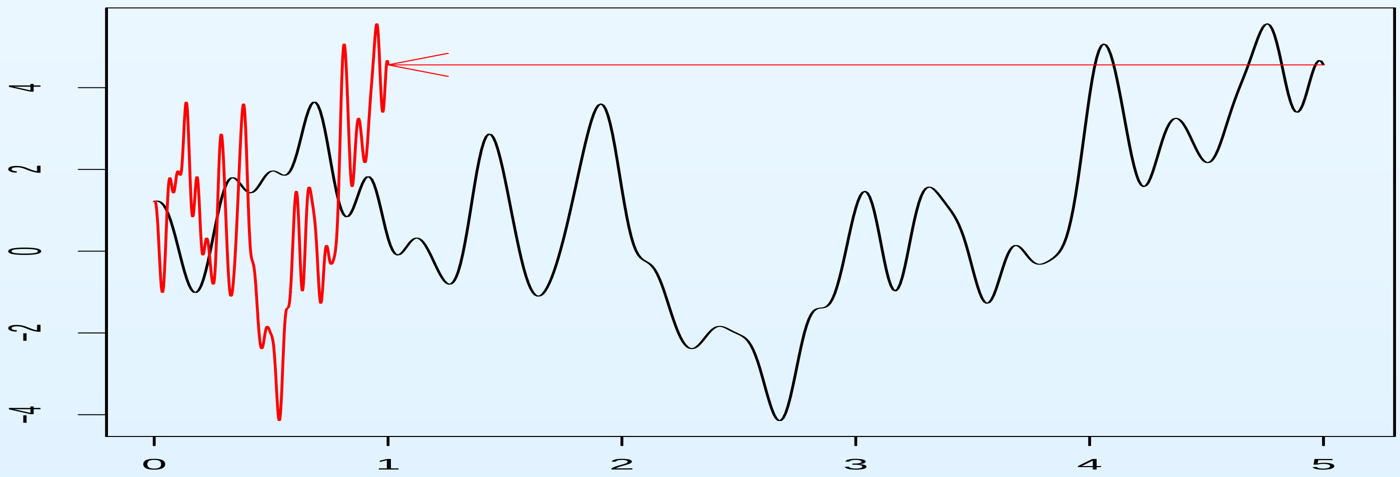


Rescaling

Sample paths can be smoothed by stretching



and roughened by shrinking



Rescaled Brownian motion (Toy example)

$W_t = B_{t/c_n}$ for B Brownian motion, $t \in [0,1]$ 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 (Toy example)

$W_t = B_{t/c_n}$ for B Brownian motion, $t \in [0,1]$ 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!)

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

Rescaled smooth stationary process

A Gaussian process with infinitely-smooth sample paths is obtained with

$$\mathrm{E} G_s G_t = \psi(s-t), \qquad \int e^{|\lambda|} \hat{\psi}(\lambda)\, d\lambda < \infty.$$

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

Adaptation by rescaling (1)

- Choose c from a Gamma distribution
- Choose $(G_t\colon t > 0)$ centered Gaussian with $\mathrm{E} G_s G_t = \exp\bigl(-(s-t)^2\bigr)$
- Set $W_t \sim G_{t/c}$

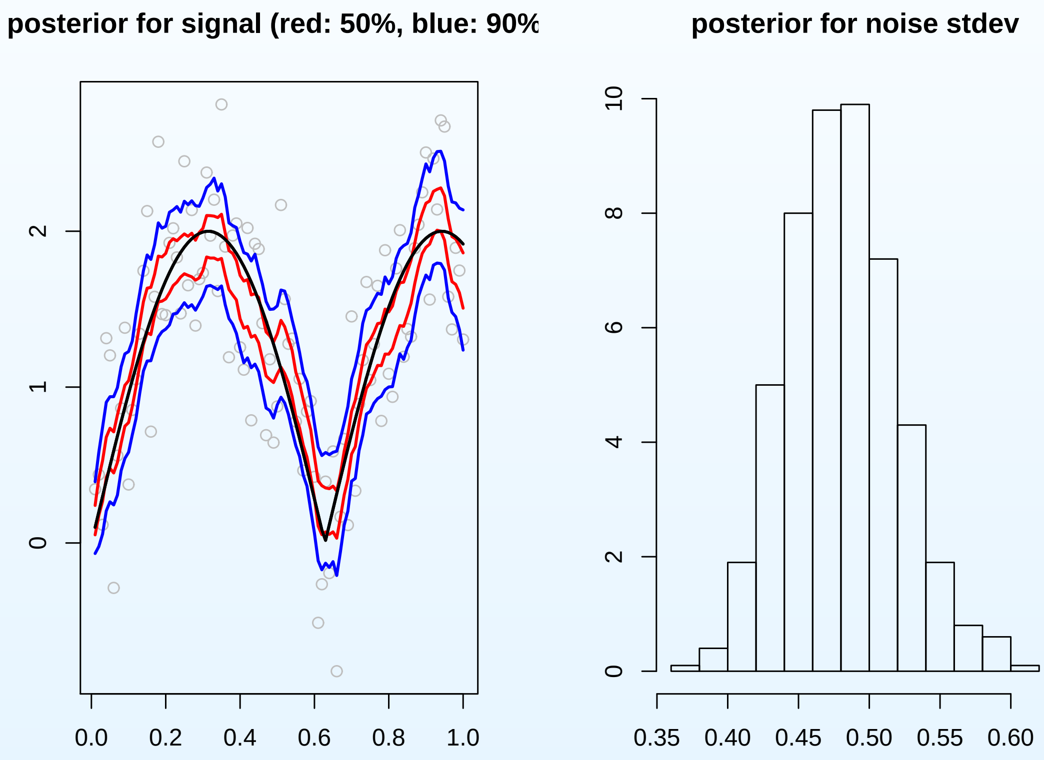
THEOREM [vdV & van Zanten 09]

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

Sir Thomas solved the bandwidth problem!?

Adaptation by rescaling (2)

Gaussian regression with Brownian motion rescaled by an inverse Gamma variable.



Conjecture: this (nearly) gives the optimal rate $n^{-\alpha/(2\alpha+1)}$ if true regression function is in $C^\alpha[0,1]$ for $\alpha \in (0,1]$. Integrated BM extends this to higher α.

Sparsity

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

$$p_n{:}= \#(1 \le i \le n{:}\, \theta_i \ne 0).$$

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

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

THEOREM (?)
If $\pi_n(p) \propto e^{-p \log(n/p)}$ and g_S has heavy tails (e.g. Cauchy or Laplace), then rate of “contraction” for Euclidean norm is $p_n \log n$.

CONCLUSION

Correctly chosen priors yield

fully adaptive nonparametrically optimal procedures

.

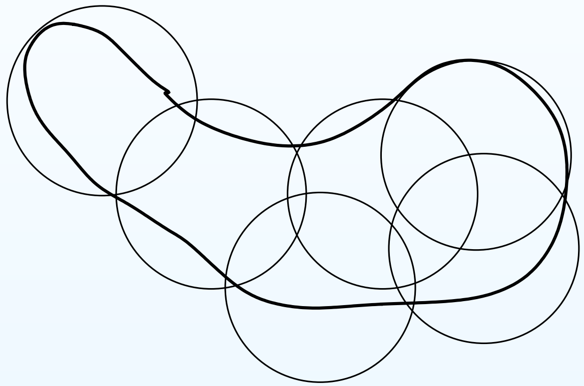
Talk 2 — Contents

- Rates — i.i.d.
- Rates — general
- Gaussian process priors — main result
- Gaussian process priors — settings
- Gaussian process priors — a proof

Rates — i.i.d.

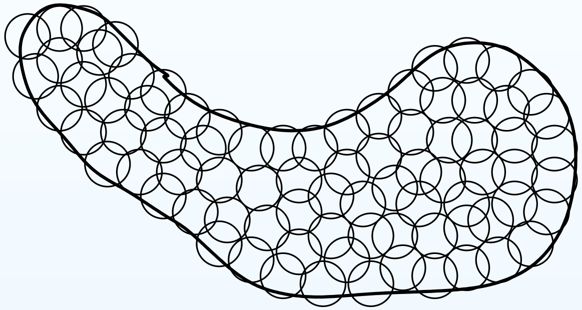
Entropy

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



Entropy

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



Entropy

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

Covering numbers characterize the minimax rate of convergence by the equation [Le Cam 73 75 86, Birgé 83 06]

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

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

Covering numbers characterize the minimax rate of convergence by the equation [Le Cam 73 75 86, Birgé 83 06]

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

For instance, for estimating a density based on a random sample of n observations with d the Hellinger distance

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

Rate — iid observations

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

$$\Pi_n(B|X_1, \ldots, X_n) := \frac{\int_B \prod_{i=1}^n p(X_i)\, d\Pi(p)}{\int_{\mathcal{P}} \prod_{i=1}^n p(X_i)\, d\Pi(p)}$$

THEOREM [Ghosal & vdV 00]
If there exist $\mathcal{P}_n \subset \mathcal{P}$ such that

- $\log N(\varepsilon_n, \mathcal{P}_n, h) \leq n\varepsilon_n^2$ entropy
- $\Pi(\mathcal{P}_n) = 1 - o(e^{-3n\varepsilon_n^2})$
- $\Pi(B_{KL}(p_0, \varepsilon_n)) \geq e^{-n\varepsilon_n^2}$ prior mass

then the Hellinger contraction rate is at least ε_n.

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

Dirichlet mixtures of normal

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

Put Dirichlet prior on F, and positive prior on $\sigma \in (a,b) \subset (0,\infty)$.

KEY LEMMA
Given ε and (F,σ) there exists F_ε with at most $d_\varepsilon := \sigma^{-1}\log(1/\varepsilon)$ support points and $d(p_{F,\sigma}, p_{F_\varepsilon,\sigma}) < \varepsilon$.

Interpretation: within accuracy ε the model is of dimension d_ε. Therefore prior mass is of order

$$\varepsilon^{d_\varepsilon}$$

and entropy is

$$\log(1/\varepsilon)^{d_\varepsilon} \approx \frac{1}{\sigma}\Big(\log\frac{1}{\varepsilon}\Big)^2.$$

Interpretation — flat prior

THEOREM [Ghosal & vdV 00]
If there exist $\mathcal{P}_n \subset \mathcal{P}$ such that

- $\log N(\varepsilon_n, \mathcal{P}_n, h) \leq n\varepsilon_n^2$ entropy
- $\Pi(\mathcal{P}_n) = 1 - o(e^{-3n\varepsilon_n^2})$
- $\Pi(B_{KL}(p_0, \varepsilon_n)) \geq e^{-n\varepsilon_n^2}$ prior mass

then the Hellinger contraction rate is at least ε_n.

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

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

Rates — general

Setting

For $n = 1, 2, \ldots$

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

Given prior Π_n on Θ_n form posterior

$$\Pi_n(B|X^{(n)}) = \frac{\int_B p_\theta^{(n)}(X^{(n)})\, d\Pi_n(\theta)}{\int_{\Theta_n} p_\theta^{(n)}(X^{(n)})\, d\Pi_n(\theta)}$$

Setting

For $n = 1, 2, \ldots$

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

Given prior Π_n on Θ_n form posterior

$$\Pi_n(B|X^{(n)}) = \frac{\int_B p_\theta^{(n)}(X^{(n)})\, d\Pi_n(\theta)}{\int_{\Theta_n} p_\theta^{(n)}(X^{(n)})\, d\Pi_n(\theta)}$$

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

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

Setting — Le Cam's testing criterion

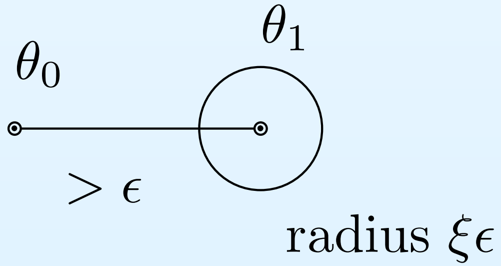
For $n = 1, 2, \dots$

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

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

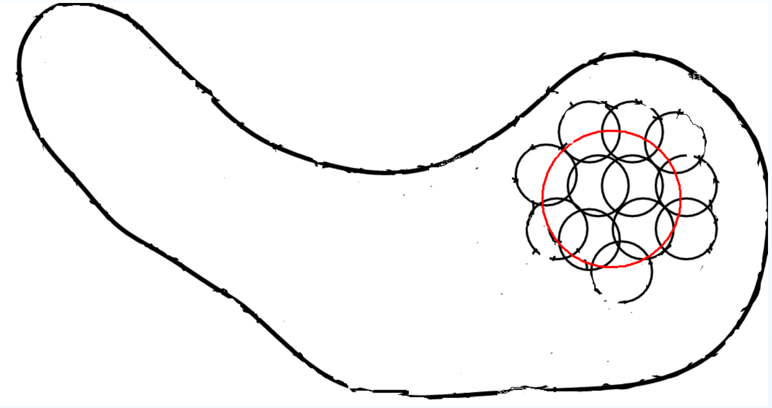
$\forall \theta_1 \in \Theta_n$ with $d_n(\theta_1, \theta_0) > \varepsilon$ $\exists$ test ϕ_n with

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



Le Cam dimension

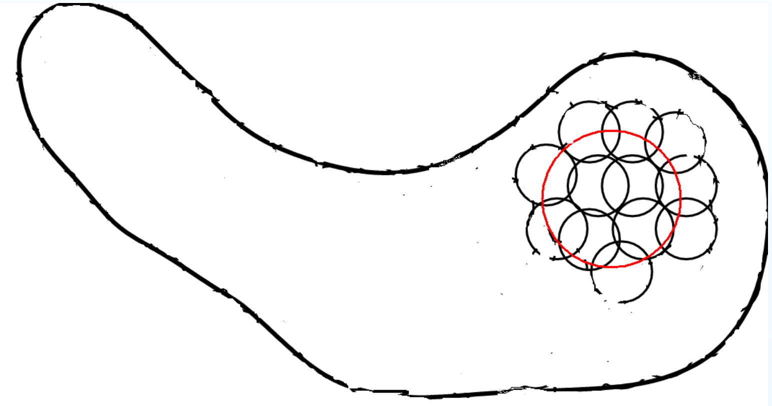
$N(\varepsilon, \Theta, d)$ = smallest number of balls of radius ε needed to cover Θ



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

Le Cam dimension

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



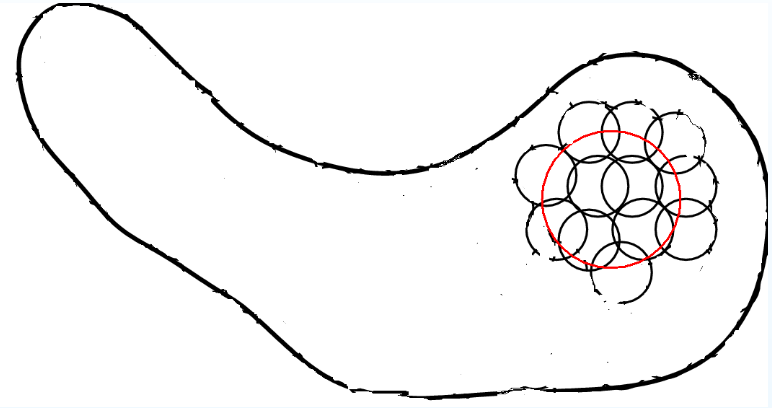
$$D_n(\varepsilon, \Theta, d_n, \bar{d}_n) = \sup_{\eta > \varepsilon} \log N\big(\varepsilon\xi, \{\theta \in \Theta_n : d_n(\theta, \theta_0) \leq \eta\}, \bar{d}_n\big).$$

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

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

Le Cam dimension

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



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

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

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

Rate theorem

THEOREM [Ghosal & vdV, 2006]
For $\varepsilon_n \to 0$, $\varepsilon_n \gg 1/\sqrt{n}$, assume $\exists \tilde\Theta_n \subset \Theta_n$:

- $D_n(\varepsilon_n, \tilde\Theta_n, d_n, \bar d_n) \le n\varepsilon_n^2$ entropy
- $\Pi_n(\tilde\Theta_n - \Theta_n) = o(e^{-3n\varepsilon_n^2})$
- $\Pi_n(B_n(\theta_0, \varepsilon_n; k)) \ge e^{-n\varepsilon_n^2}$ prior mass

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

$$B_n(\theta_0, \varepsilon; k) = \Big\{\theta \in \Theta_n\colon K(p_{\theta_0}^{(n)}, p_\theta^{(n)}) \le n\varepsilon^2, V_k(p_{\theta_0}^{(n)}, p_\theta^{(n)}) \le n^{k/2}\varepsilon^k\Big\}$$

(Kullback-Leibler neighborhood)

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

Rate theorem — refined

THEOREM [Ghosal & vdV, 2006]

For $\varepsilon_n \to 0$, assume $\exists \tilde\Theta_n \subset \Theta_n$:

- $D_n(\varepsilon_n, \tilde\Theta_n, d_n, \bar d_n) \le n\varepsilon_n^2$
- $\dfrac{\Pi_n(\tilde\Theta_n - \Theta_n)}{\Pi_n(B_n(\theta_0, \varepsilon_n; k))} = o(e^{-2n\varepsilon_n^2})$
- $\dfrac{\Pi_n(\theta \in \Theta_n \colon d_n(\theta, \theta_0) \le 2j\varepsilon_n)}{\Pi_n(B_n(\theta_0, \varepsilon_n; k))} \le e^{Kn\varepsilon_n^2 j^2/2} \qquad \forall j$

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

Further trade-off between complexity and prior mass possible.

I.i.d. observations

Data $X_1, \ldots, X_n$, i.i.d. with density p_θ

MAIN RESULT HOLDS WITH

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

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

Independent observations

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

MAIN RESULT HOLDS WITH

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

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

Markov chains

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

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

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

MAIN RESULT HOLDS WITH

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

Gaussian white noise model

Data $(X_t^{(n)}: 0 \le t \le 1)$ for $dX_t^{(n)} = \theta(t)\,dt + n^{-1/2}\,dB_t$, where B is Brownian motion

MAIN RESULT HOLDS WITH

- d_n: L_2-norm
- $B_n(\theta_0, \varepsilon; 2)$: L_2-ball

Gaussian time series

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

Assume

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

MAIN RESULT HOLDS WITH

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

Ergodic diffusions

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

Assume

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

MAIN RESULT HOLDS WITH

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

Gaussian process priors — main result

Setting

Data $X^{(n)}$ follows density $p_{w_0}^{(n)}$ indexed by a function $w_0: T \to \mathbb{R}$

Prior Π for w is law of Gaussian process $(W_t: t \in T)$

Posterior:

$$\Pi_n(B|X^{(n)}) := \frac{\int_B p_w^{(n)}(X^{(n)})\, d\Pi(w)}{\int p_w^{(n)}(X^{(n)})\, d\Pi(w)}$$

Setting

Data $X^{(n)}$ follows density $p_{w_0}^{(n)}$ indexed by a function $w_0: T \to \mathbb{R}$

Prior Π for w is law of Gaussian process $(W_t: t \in T)$

Posterior:

$$\Pi_n(B|X^{(n)}) := \frac{\int_B p_w^{(n)}(X^{(n)})\, d\Pi(w)}{\int p_w^{(n)}(X^{(n)})\, d\Pi(w)}$$

Rate of contraction is defined to be at least ε_n if as $n, M \to \infty$,

$$P_{w_0}^{(n)} \Pi_n(w: d_n(w, w_0) \geq M\varepsilon_n | X^{(n)}) \to 0$$

Reproducing kernel Hilbert space

Think of the Gaussian process as a random element in a (complete) function space equipped with a norm: 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 (complete) function space equipped with a norm: 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 $\mathbb{H} = \{h: \int h'(t)^2\,dt < \infty\}$ with norm $\|h\|_{\mathbb{H}} = \|h'\|_2$

Small ball probability

W Gaussian map in $(\mathbb{B}, \|\cdot\|)$

Small ball probability $\mathrm{P}(\|W\| < \varepsilon)$

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

Small ball probability

W Gaussian map in $(\mathbb{B}, \|\cdot\|)$

Small ball probability $\quad \mathrm{P}(\|W\| < \varepsilon)$

Small ball exponent $\quad \phi_0(\varepsilon) = -\log \mathrm{P}(\|W\| < \varepsilon)$

EXAMPLE

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

Main result

Prior W is Gaussian map in $(\mathbb{B}, \|\cdot\|)$

RKHS $(\mathbb{H}, \|\cdot\|_{\mathbb{H}})$ $\qquad \mathrm{P}(\|W\| < \varepsilon) = e^{-\phi_0(\varepsilon)}$

THEOREM [vdV& van Zanten 07]

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

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

Both inequalities give lower bound on ε_n; first depends on W and not on w_0

Toy problem — Brownian motion

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

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

LEMMA

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

Toy problem — Brownian motion

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

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

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

CONSEQUENCE:
Rate is ε_n if
$(1/\varepsilon_n)^2 \leq n\varepsilon_n^2$ AND $(1/\varepsilon_n)^{(2-2\alpha)/\alpha} \leq n\varepsilon_n^2$

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

Gaussian process priors — settings

Main result

Prior W is Gaussian map in $(\mathbb{B}, \|\cdot\|)$

RKHS $(\mathbb{H}, \|\cdot\|_{\mathbb{H}})$ $\quad \mathrm{P}(\|W\| < \varepsilon) = e^{-\phi_0(\varepsilon)}$

THEOREM [vdV& van Zanten 07]

If statistical distances on the model combine appropriately with the norm $\|\cdot\|$ of $\mathbb{B}$ (see below), 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 $X_1, \ldots, X_n$ i.i.d. from density on $[0, 1]$

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

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

Density estimation

Data $X_1, \ldots, X_n$ i.i.d. from density on $[0,1]$

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

- Distance on parameter: Hellinger distance 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 $(X_1, Y_1), \ldots, (X_n, Y_n)$ i.i.d. in $[0, 1] \times \{0, 1\}$

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

E.g. Ψ 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 $Y_1, \ldots, Y_n$
$Y_i = w_0(x_i) + e_i$
$x_1, \ldots, x_n$ fixed design points
$e_1, \ldots, e_n$ i.i.d. Gaussian mean-zero errors

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

Can use posterior for Gaussian errors also if errors have only mean zero?
(Kleijn & vdV, 2006)

Gaussian white noise

Data $(X_t: t \in [0,1])$
$dX_t = w_t + n^{-1/2}\, dB_t$

- Distance on parameter: L_2
- Norm on W: L_2

Gaussian process priors — a proof

Reproducing kernel Hilbert space — definition

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

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

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

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

Reproducing kernel Hilbert space — definition (2)

$W = (W_x : x \in \mathcal{X})$ Gaussian stochastic process that can be seen as tight, Borel measurable map in $\ell^\infty(\mathcal{X}) = \{f : \mathcal{X} \to \mathbb{R} : \sup_x |f(x)| < \infty\}$

Covariance function $K(x, y) = \mathrm{E} W_x W_y$

Then RKHS is completion of the set of functions

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

relative to inner product

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

Reproducing kernel Hilbert space — definition (3)

Any Gaussian random element 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. $N(0,1)$
- $\|e_1\| = \|e_2\| = \cdots = 1$

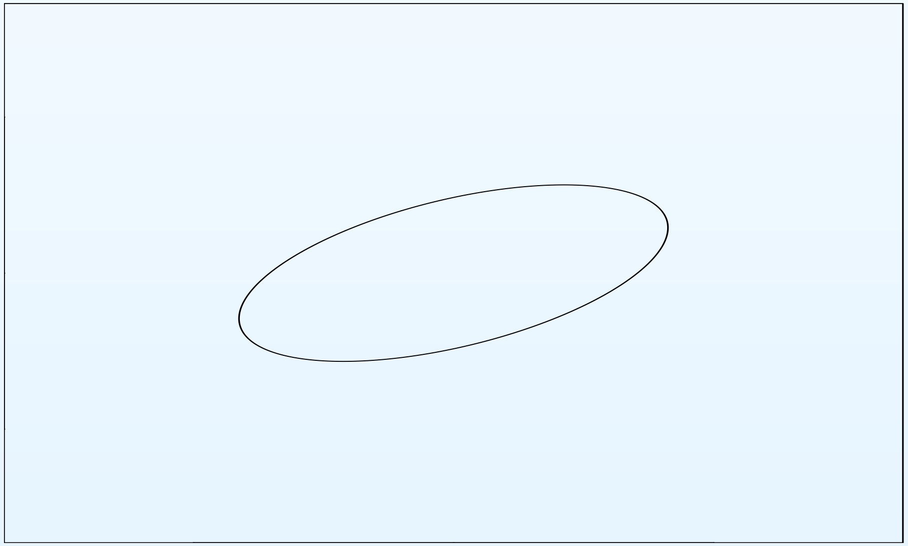
The RKHS consists of all elements $h := \sum_i h_i e_i$ with

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

Reproducing kernel Hilbert space — definition (4)

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

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



Geometry

RKHS gives the “geometry of the support of W”

Geometry

RKHS gives the "geometry of the support of W"

THEOREM
Norm closure of $\mathbb{H}$ in $\mathbb{B}$ is smallest closed set with probability one under Gaussian measure (and hence posterior inconsistent if $\|w_0 - \mathbb{H}\| > 0$)

THEOREM [Borell 75]

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

THEOREM [Kuelbs & Li 93]
For $\mathbb{H}_1$ the unit ball of RKHS

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

Decentered small ball probability

W Gaussian map in $(\mathbb{B}, \|\cdot\|)$

RKHS $(\mathbb{H}, \|\cdot\|_{\mathbb{H}})$ $\quad \mathrm{P}(\|W\| < \varepsilon) = e^{-\phi_0(\varepsilon)}$

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

Decentered small ball probability

W Gaussian map in $(\mathbb{B}, \|\cdot\|)$

RKHS $(\mathbb{H}, \|\cdot\|_{\mathbb{H}})$ $\quad \mathrm{P}(\|W\| < \varepsilon) = e^{-\phi_0(\varepsilon)}$

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

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

Proof

Sufficient for posterior rate of ε_n is existence of sets $\mathbb{B}_n$ with

- $\log N(\varepsilon_n, \mathbb{B}_n, d) \leq n\varepsilon_n^2$ entropy
- $\Pi_n(\mathbb{B}_n) = 1 - o(e^{-3n\varepsilon_n^2})$
- $\Pi_n(B_n(w_0, \varepsilon_n)) \geq e^{-n\varepsilon_n^2}$ prior mass

Take $\mathbb{B}_n = M_n\mathbb{H}_1 + \varepsilon_n\mathbb{B}_1$ for appropriate M_n.
Use Borell's inequality.

CONCLUSION

Correctly chosen priors yield

fully adaptive nonparametrically optimal procedures

.