Nonparametric Bayesian Uncertainty Quantification

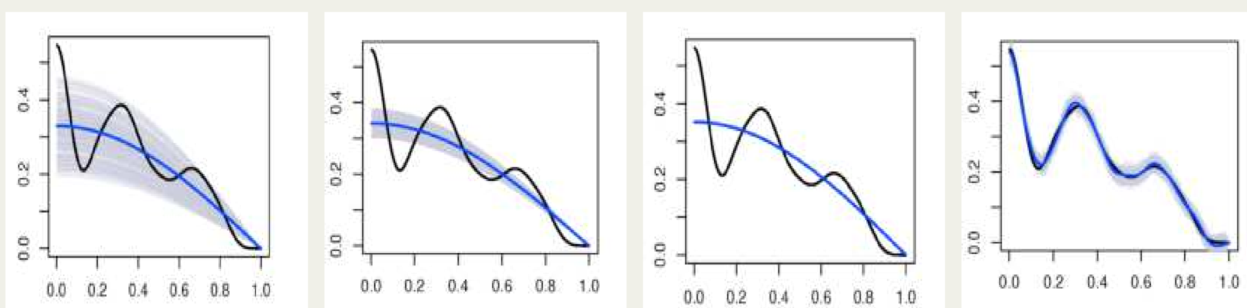
Lecture 1: Introduction to Nonparametric Bayes

Aad van der Vaart

Universiteit Leiden, Netherlands

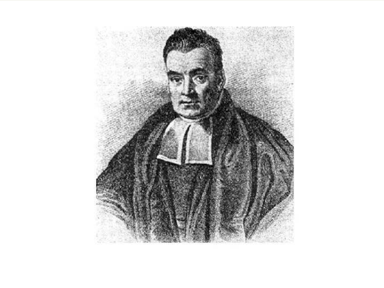
YES, Eindhoven, January 2017

Contents



Introduction

The Bayesian paradigm



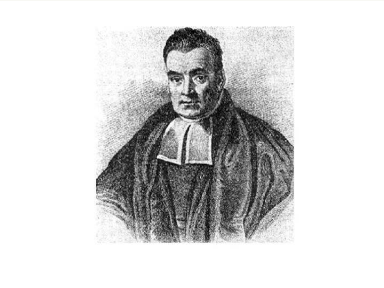
- A parameter Θ is generated according to a prior distribution Π.
- Given $\Theta = \theta$ the data X is generated according to a measure P_θ.

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:

$$\Pi(\theta \in B|\, X).$$

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.

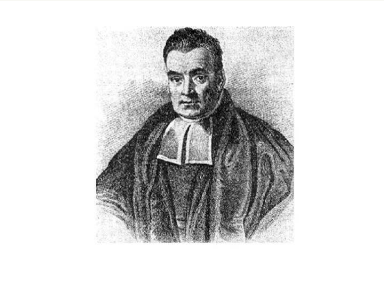
This gives a joint distribution of (X, Θ).

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

$$\Pi(\theta \in B |\, X).$$

We assume whatever needed (e.g. Θ Polish and Π a probability distribution on its Borel σ-field; Polish sample space) to make this well defined.

Bayes's rule



- 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).

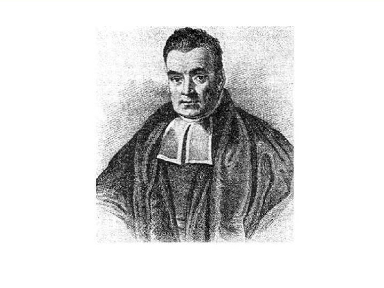
- Given observed data x the statistician computes the conditional distribution of Θ given $X = x$, the posterior distribution:

$$\Pi(\theta \in B |\, X).$$

If P_θ is given by a density $x \mapsto p_\theta(x)$, then **Bayes's rule** gives

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

Bayes's rule



- 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).

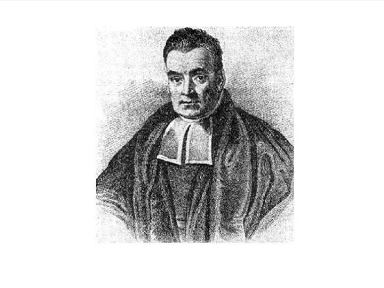
- Given observed data x the statistician computes the conditional distribution of Θ given $X = x$, the posterior distribution:

$$\Pi(\theta \in B | X).$$

If P_θ is given by a density $x \mapsto p_\theta(x)$, then **Bayes's rule** gives

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

Reverend Thomas



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

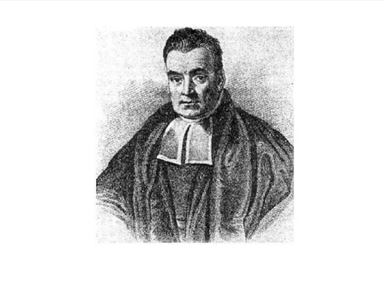
Using his famous rule he computed that the posterior distribution is then *Beta*$(X+1, n-X+1)$.

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

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

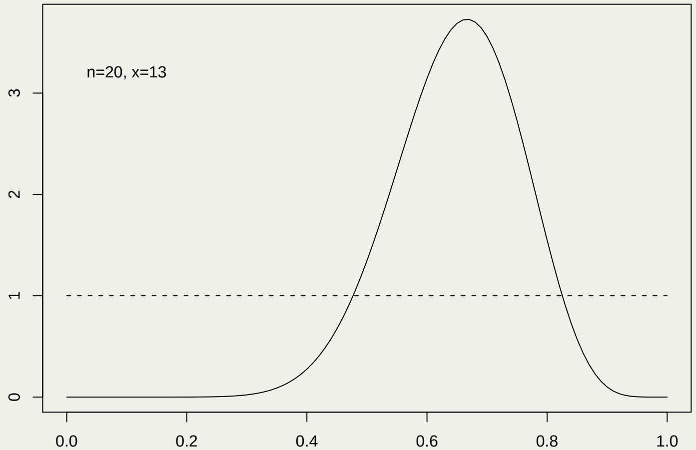
$$d\Pi(\theta|\, X) = \theta^X (1-\theta)^{n-X} \cdot 1.$$

Reverend Thomas

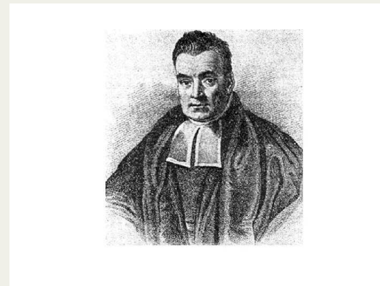


Thomas Bayes (1702–1761, 1763) followed this argument with Θ possessing the *uniform* distribution and X given $\Theta = \theta$ *binomial* (n, θ).

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

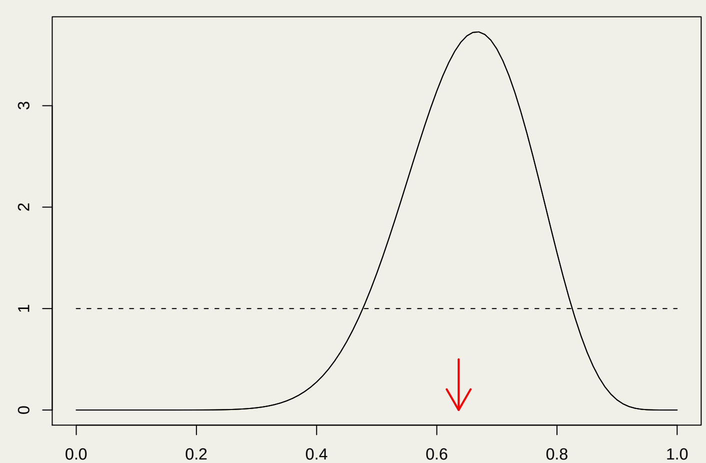


Reverend Thomas



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



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.

He had many followers, but Ronald Aylmer Fisher (1890–1962) did not buy into it.

The Bayesian method regained popularity following the development of MCMC methods in the 1980/90s.

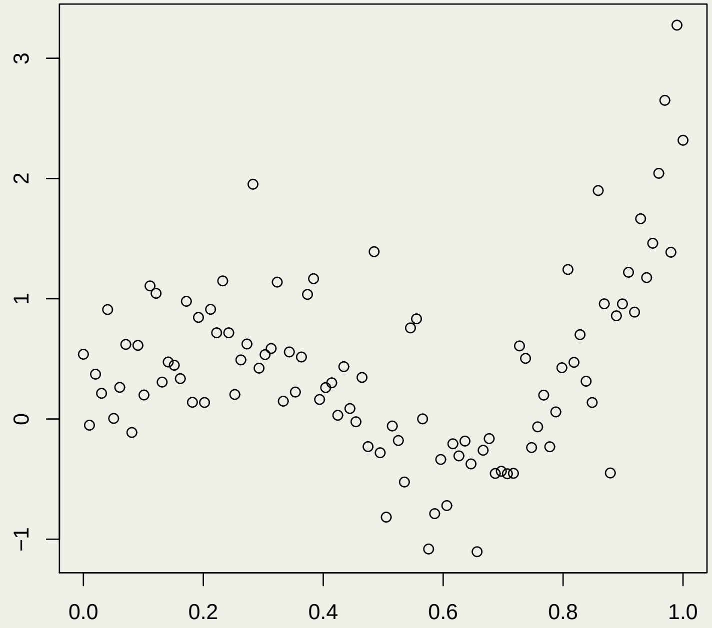
Nonparametric Bayesian statistics set off in the 1970/80/90s, although it was long thought not too work.

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's 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.

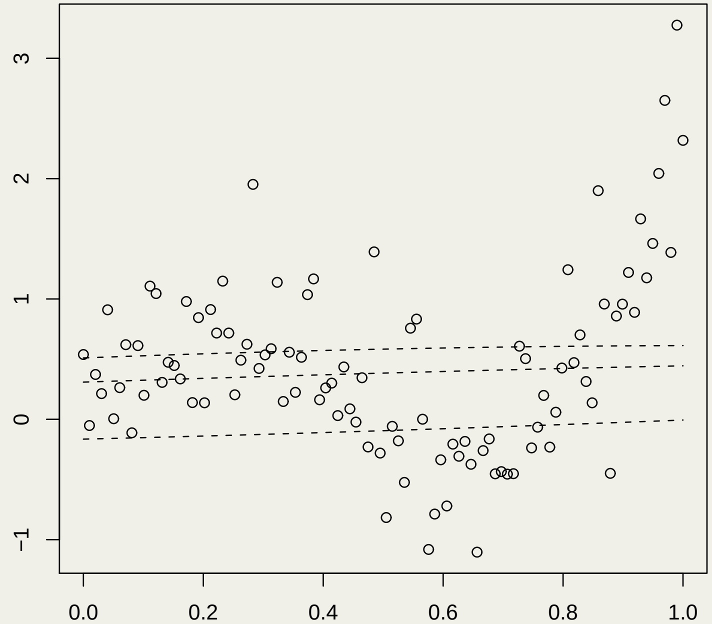


Nonparametric Bayes

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's 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.

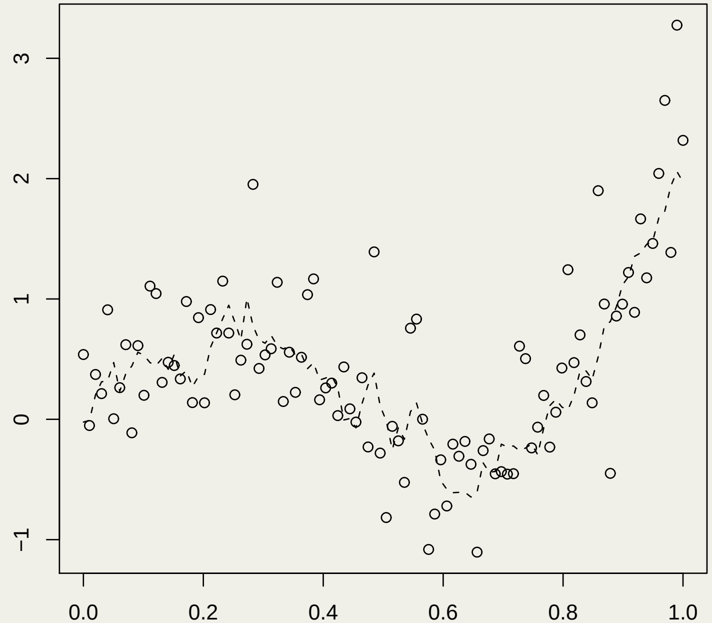


Nonparametric Bayes

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's 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.

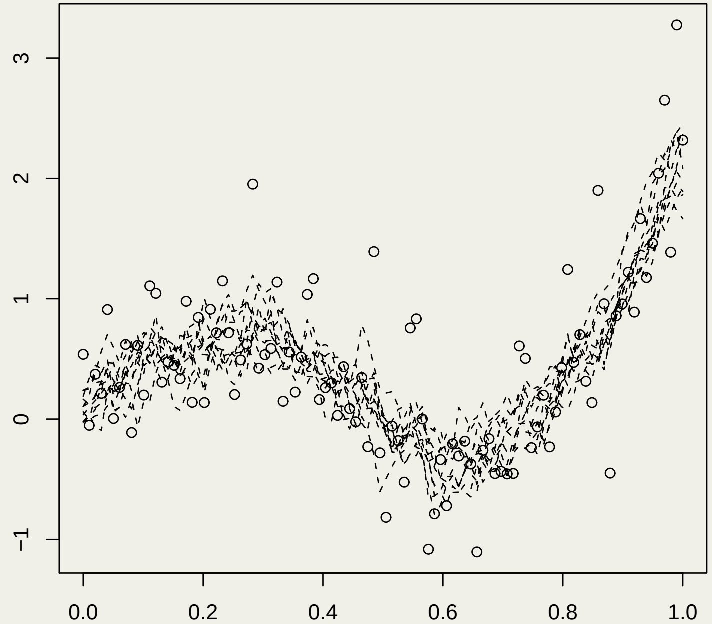


Nonparametric Bayes

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's 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.

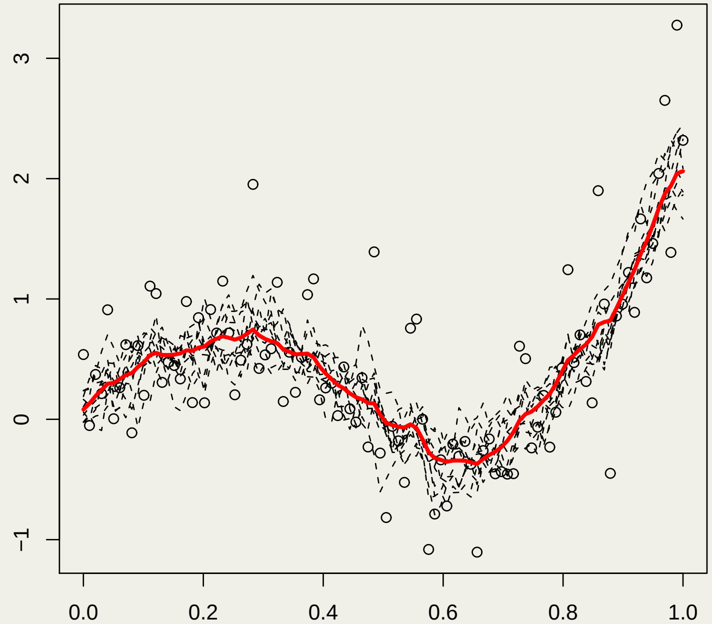


Nonparametric Bayes

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's 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.

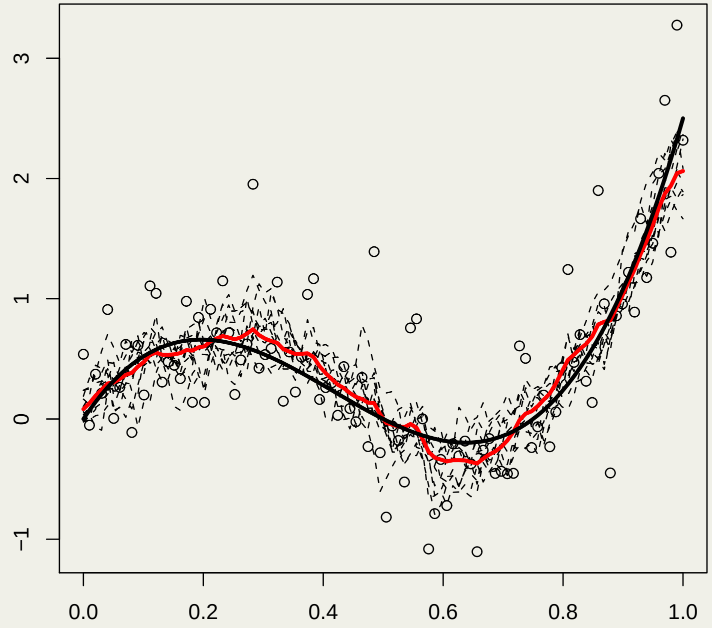


Nonparametric Bayes

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's 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 $\Pi(\theta \in \cdot \,|\, X)$ as a random measure on the parameter set dependent on X.

RECOVERY

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

Frequentist Bayesian

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

RECOVERY

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

UNCERTAINTY QUANTIFICATION

We like the "spread" of $\Pi(\theta \in \cdot \mid X)$ to indicate remaining uncertainty.

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.

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

UNCERTAINTY QUANTIFICATION
We like the "spread" of $\Pi(\theta \in \cdot \,|\, X)$ to indicate remaining uncertainty.

A set with C_X with $\Pi(\theta \in C_X \,|\, X) = 0.95$ is called a credible set.

Is it a confidence set? Is $\mathrm{P}_{\theta_0}(C_X \ni \theta_0) = 0.95$? Is its order of magnitude correct?

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.

RECOVERY

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

UNCERTAINTY QUANTIFICATION

We like the “spread” of $\Pi(\theta \in \cdot \,|\, X)$ to indicate remaining uncertainty.

A set with C_X with $\Pi(\theta \in C_X \,|\, X) = 0.95$ is called a credible set.

Is it a confidence set? Is $\mathrm{P}_{\theta_0}(C_X \ni \theta_0) = 0.95$? Is its order of magnitude correct?

Asymptotic setting: data $X^{(n)}$ where the information increases as $n \to \infty$.

- We want $\Pi_n(\cdot \,|\, X^{(n)}) \to \delta_{\theta_0}$, at a good rate.
- We like $\mathrm{P}_{\theta_0}(C_{X^{(n)}} \ni \theta_0) \to 0.95$.

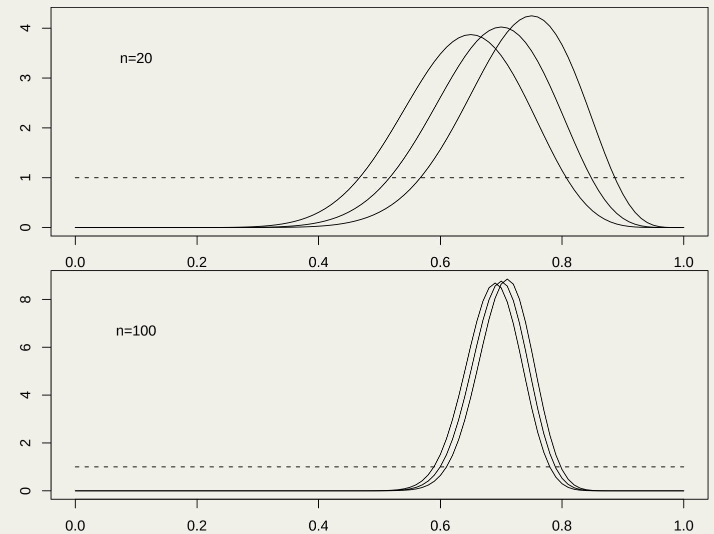
Parametric models

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,*

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

Here $\tilde\theta_n$ are estimators with $\sqrt{n}(\tilde\theta_n - \theta_0) \rightsquigarrow N(0, I_{\theta_0}^{-1})$.



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 $\tilde\theta_n$ are estimators with $\sqrt{n}(\tilde\theta_n - \theta_0) \rightsquigarrow N(0, I^{-1}_{\theta_0})$.

RECOVERY:
The posterior distribution concentrates most of its mass on balls of radius $O(1/\sqrt{n})$ around θ_0.

UNCERTAINTY QUANTIFICATION:
A central set of posterior probability 95 % is equivalent to the usual Wald confidence set $\big\{\theta\colon n(\theta - \tilde\theta_n)^T I_{\tilde\theta_n}(\theta - \tilde\theta_n) \le \chi^2_{d,1-\alpha}\big\}$.

These lectures

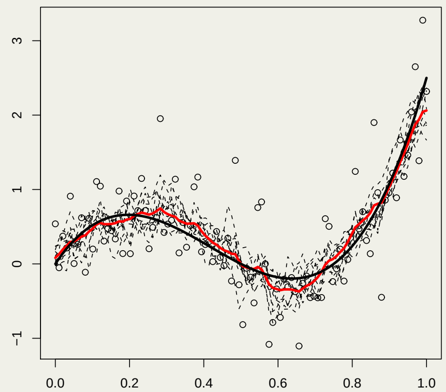
Recovery and uncertainty quantification for nonparametric models.

LECTURE 1: Introduction to recovery.

LECTURE 2: Uncertainty quantification for curve fitting/smoothing.

LECTURE 3: Uncertainty quantification in high dimensions under sparsity.

> Point of view:
> How does the posterior distribution for natural priors behave, in particular for priors that adapt to complexity in the data.



Recovery

Consistency

- $X^{(n)}$ observation in sample space $(\mathfrak{X}^{(n)}, \mathcal{X}^{(n)})$ with distribution $P_\theta^{(n)}$.
- θ belongs to metric space (Θ, d).

Definition. *The posterior distribution is* consistent *at* $\theta_0 \in \Theta$ *if for every* $\epsilon > 0$*.*

$$\mathrm{E}_{\theta_0} \Pi_n\big(\theta\colon d(\theta, \theta_0) > \epsilon \,|\, X^{(n)}\big) \to 0, \qquad n \to \infty.$$

Schwartz's theorem (1965)

Parameter p: ν-density on sample space $(\mathfrak{X}, \mathcal{X})$. True value p_0.

$$K(p_0; p) = \int p_0 \log(p_0/p)\, d\nu.$$

Schwartz's theorem (1965)

Parameter p: ν-density on sample space $(\mathfrak{X}, \mathcal{X})$. True value p_0.

$$K(p_0; p) = \int p_0 \log(p_0/p)\, d\nu.$$

Definition. p_0 *is said to possess the* Kullback-Leibler property *relative to* Π *if* $\Pi\big(p\colon K(p_0; p) < \epsilon\big) > 0$ *for every* $\epsilon > 0$.

Bayesian model:

$$X_1, \ldots, X_n | \, p \overset{\text{iid}}{\sim} p, \qquad p \sim \Pi.$$

Schwartz's theorem (1965)

Parameter p: ν-density on sample space $(\mathfrak{X}, \mathscr{X})$. True value p_0.

$$K(p_0; p) = \int p_0 \log(p_0/p)\, d\nu.$$

Definition. p_0 *is said to possess the* Kullback-Leibler property *relative to* Π *if* $\Pi\big(p\colon K(p_0; p) < \epsilon\big) > 0$ *for every* $\epsilon > 0$.

Bayesian model:

$$X_1, \ldots, X_n |\, p \overset{\text{iid}}{\sim} p, \qquad p \sim \Pi.$$

Theorem. *If* p_0 *has KL-property, and for every neighbourhood* $\mathcal{U}$ *of* p_0 *there exist tests* ϕ_n *such that*

$$P_0^n \phi_n \to 0, \qquad \sup_{p \in \mathcal{U}^c} P^n(1 - \phi_n) \to 0,$$

then $\Pi_n(\cdot|\, X_1, \ldots, X_n)$ *is consistent at* p_0*.*

Extended Schwartz's theorem

Bayesian model:

$$X_1, \ldots, X_n | \, p \overset{\text{iid}}{\sim} p, \qquad p \sim \Pi.$$

Theorem. *If If p_0 has KL-property and for every neighbourhood $\mathcal{U}$ of p_0 there exist $C > 0$, sets $\mathcal{P}_n \subset \mathcal{P}$ and tests ϕ_n such that*

$$\Pi(\mathcal{P} \setminus \mathcal{P}_n) < e^{-Cn}, \qquad P_0^n \phi_n \le e^{-Cn}, \qquad \sup_{p \in \mathcal{P}_n \cap \mathcal{U}^c} P^n(1 - \phi_n) \le e^{-Cn},$$

then $\Pi_n(\cdot | \, X_1, \ldots, X_n)$ is consistent at p_0.

Tests exist if:

- Weak topology: always
- L_1-distance: if $\log N\big(\epsilon, \mathcal{P}_n, \| \cdot \|_1\big) \le n\epsilon^2/3$, for all $\varepsilon > 0$.

Definition. *The covering number $N(\epsilon, \mathcal{P}, d)$ is the minimal number of d-balls of radius ϵ needed to cover $\mathcal{P}$.*

Rate of contraction

Bayesian model:

$$X_1, \ldots, X_n | \, p \overset{\text{iid}}{\sim} p, \qquad p \sim \Pi.$$

Definition. *The posterior distribution* $\Pi_n(\cdot | \, X^{(n)})$ contracts at rate $\epsilon_n \to 0$ *at* $\theta_0 \in \Theta$ *if, for every* $M_n \to \infty$*,*

$$\mathrm{E}_{\theta_0} \Pi_n\bigl(\theta\colon d(\theta, \theta_0) > M_n \epsilon_n | \, X^{(n)}\bigr) \to 0, \qquad n \to \infty.$$

Rate of contraction

Bayesian model:

$$X_1, \ldots, X_n | \, p \overset{\text{iid}}{\sim} p, \qquad p \sim \Pi.$$

Definition. *The posterior distribution* $\Pi_n(\cdot | \, X^{(n)})$ contracts at rate $\epsilon_n \to 0$ *at* $\theta_0 \in \Theta$ *if, for every* $M_n \to \infty$,

$$\mathrm{E}_{\theta_0} \Pi_n\bigl(\theta \colon d(\theta, \theta_0) > M_n \epsilon_n | \, X^{(n)}\bigr) \to 0, \qquad n \to \infty.$$

Benchmark rate for curve fitting: A function θ of d variables that has bounded derivatives of order β is estimable based on n observations at rate

$$n^{-\beta/(2\beta+d)}.$$

Proposition. *If the posterior distribution contracts at rate* ϵ_n *at* θ_0*, then* $\hat{\theta}_n$ *defined as the center of a (nearly) smallest ball that contains posterior mass at least* $1/2$ *satisfies* $d(\hat{\theta}_n, \theta_0) = O_P(\epsilon_n)$ *under* $P_{\theta_0}^{(n)}$.

Basic contraction theorem

Bayesian model:

$$X_1, \ldots, X_n | \, p \overset{\text{iid}}{\sim} p, \qquad p \sim \Pi.$$

$$K(p_0; p) = P_0 \log \frac{p_0}{p}, \quad V(p_0; p) = P_0 \Big(\log \frac{p_0}{p}\Big)^2, \quad h^2(p_0, p) = \int (\sqrt{p_0} - \sqrt{p})^2 \, d\nu.$$

Theorem. *Given metric $d \leq h$ whose balls are convex suppose that there exist $\mathcal{P}_n \subset \mathcal{P}$ and $C > 0$, such that,*

(i) $\Pi_n\big(p: K(p_0; p) < \epsilon_n^2, V(p_0; p) < \epsilon_n^2\big) \geq e^{-Cn\epsilon_n^2}$, *(prior mass)*

(ii) $\log N\big(\epsilon_n, \mathcal{P}_n, d\big) \leq n\epsilon_n^2$. *(complexity)*

(iii) $\Pi_n(\mathcal{P}_n^c) \leq e^{-(C+4)n\epsilon_n^2}$.

Then the posterior rate of convergence for d is $\epsilon_n \vee n^{-1/2}$.

Basic contraction theorem — proof

Proof.

- There exist tests ϕ_n with

$$P_0^n\phi_n \le e^{n\epsilon_n^2}\frac{e^{-nM^2\epsilon_n^2/8}}{1-e^{-nM^2\epsilon_n^2/8}}, \qquad \sup_{p\in\mathcal{P}_n:d(p,p_0)>M\epsilon_n} P^n(1-\phi_n) \le e^{-nM^2\epsilon_n^2/8}.$$

- For $A_n = \bigl\{\int \prod_{i=1}^n (p/p_0)(X_i)\,d\Pi_n(p) \ge e^{-(2+C)n\epsilon_n^2}\bigr\}$

$$\Pi_n\bigl(p\colon d(p,p_0) > M\epsilon_n \bigm| X_1,\ldots,X_n\bigr)$$
$$\le \phi_n + 1\!\!1\{A_n^c\} + e^{(2+C)n\epsilon_n^2}\int_{d(p,p_0)>M\epsilon_n}\prod_{i=1}^n \frac{p}{p_0}(X_i)\,d\Pi_n(p)(1-\phi_n).$$

- $P_0^n(A_n^c)\to 0$. See further on.

Basic contraction theorem — proof continued

Proof. (Continued)

- By Fubini

$$P_0^n \int_{p\in\mathcal{P}_n: d(p,p_0)>M\epsilon_n} \prod_{i=1}^n \frac{p}{p_0}(X_i)\, d\Pi_n(p)$$
$$\leq \int_{p\in\mathcal{P}_n: d(p,p_0)>M\epsilon_n} P^n(1-\phi_n)\, d\Pi_n(p) \leq e^{-nM^2\epsilon_n^2/8}$$

- By Fubini

$$P_0^n \int_{\mathcal{P}\setminus\mathcal{P}_n} \prod_{i=1}^n \frac{p}{p_0}(X_i)\, d\Pi_n(p) \leq \Pi_n(\mathcal{P}\setminus\mathcal{P}_n).$$

Bounding the denominator

Lemma. *For any probability measure* Π *on* $\mathcal{P}$*, and positive constant* ϵ*, with* P_0^n*-probability at least* $1-(n\epsilon^2)^{-1}$*,*

$$\int \prod_{i=1}^{n} \frac{p}{p_0}(X_i)\, d\Pi(p) \geq \Pi\Big(p\colon K(p_0;p) < \epsilon^2, V(p_0;p) < \epsilon^2\Big) e^{-2n\epsilon^2}.$$

Bounding the denominator

Lemma. *For any probability measure Π on $\mathcal{P}$, and positive constant ϵ, with P_0^n-probability at least $1-(n\epsilon^2)^{-1}$,*

$$\int \prod_{i=1}^n \frac{p}{p_0}(X_i)\, d\Pi(p) \geq \Pi\big(p\colon K(p_0;p) < \epsilon^2, V(p_0;p) < \epsilon^2\big) e^{-2n\epsilon^2}.$$

Proof. $B\colon = \big\{p\colon K(p_0;p) < \epsilon_n^2, V(p_0;p) < \epsilon_n^2\big\}$.

$$\log \int \prod_{i=1}^n \frac{p}{p_0}(X_i)\, d\Pi(P) \geq \sum_{i=1}^n \int \log \frac{p}{p_0}(X_i)\, d\Pi(P) =: Z.$$

$$\mathrm{E} Z = -n \int K(p_0;p)\, d\Pi(p) > -n\epsilon^2,$$

$$\mathrm{var}\, Z \leq nP_0 \Big(\int \log \frac{p_0}{p}\, d\Pi(p)\Big)^2 \leq nP_0 \int \Big(\log \frac{p_0}{p}\Big)^2 d\Pi(p) \leq n\epsilon^2,$$

Apply Chebyshev's inequality.

Interpretation

Consider a maximal set of points $p_1, \ldots, p_N$ in $\mathcal{P}_n$ with $d(p_i, p_j) \geq \epsilon_n$.

Maximality implies $N \geq N(\epsilon_n, \mathcal{P}_n, d) \geq e^{c_1 n \epsilon_n^2}$, under the entropy bound.

The balls of radius $\epsilon_n/2$ around the points are disjoint and hence the sum of their prior masses will be less than 1.

If the prior mass were evenly distributed over these balls, then each would have no more mass than $e^{-c_1 n \epsilon_n^2}$.

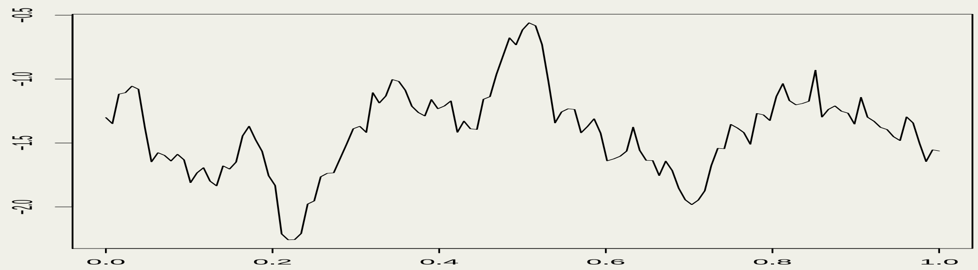
This is of the same order as the prior mass bound.

This argument suggests that the conditions can only be satisfied for every p_0 in the model if the prior "distributes its mass uniformly, at discretization level ϵ_n".

Gaussian process priors

Gaussian process prior

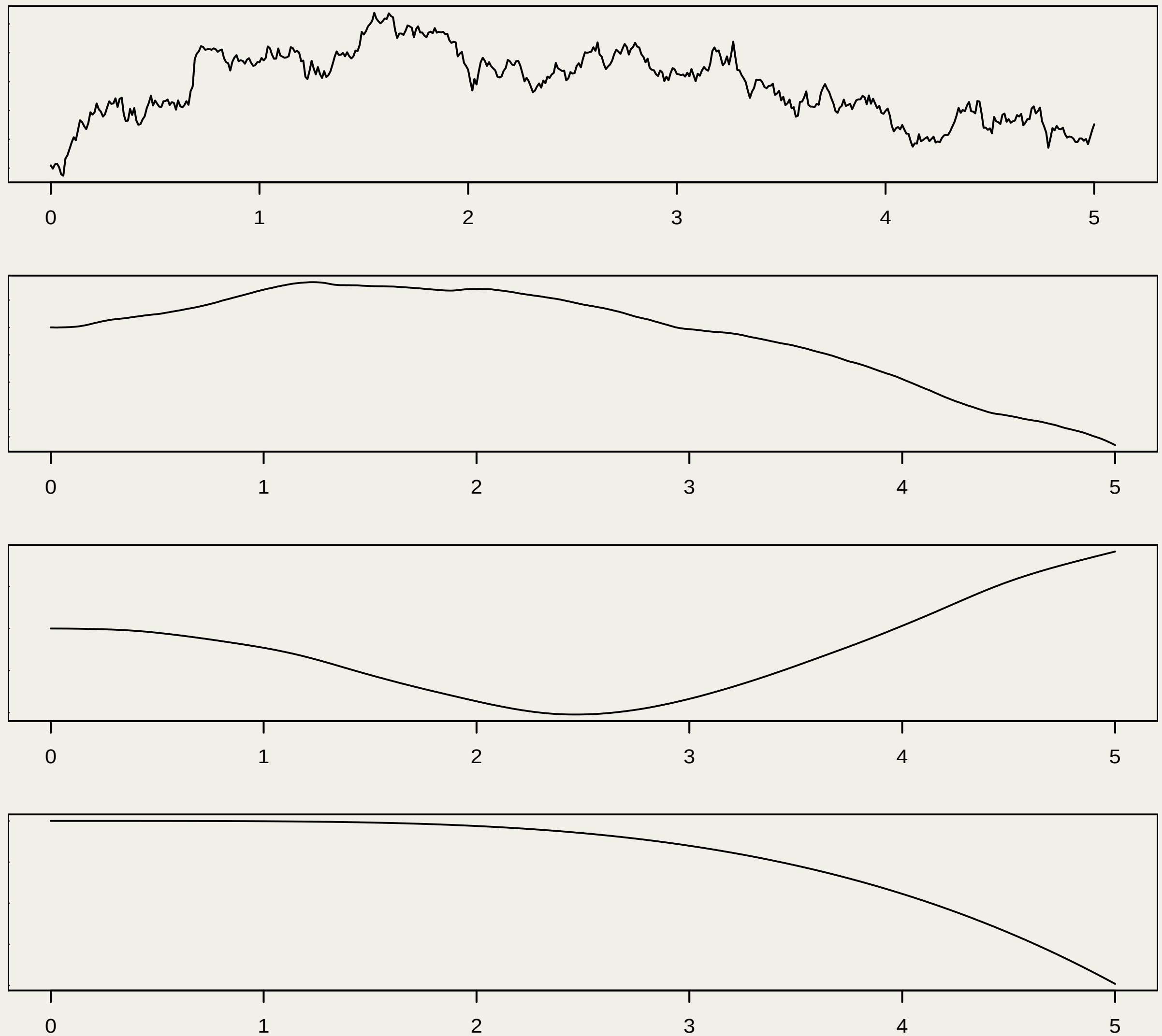
The law of a stochastic process $W = (W_t: t \in T)$ is a prior distribution on the space of functions $\theta: T \to \mathbb{R}$.



W is a Gaussian process if $(W_{t_1}, \ldots, W_{t_k})$ is multivariate Gaussian, for every $t_1, \ldots, t_k$.
Mean and covariance function:

$$t \mapsto \mathrm{E}W_t, \qquad \text{and} \qquad (s,t) \mapsto \mathrm{cov}(W_s, W_t), \qquad s,t \in T.$$

Brownian motion and its primitives



0, 1, 2 and 3 times integrated Brownian motion

Settings

Density estimation

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

$$p_\theta(x) = \frac{e^{\theta(x)}}{\int_0^1 e^{\theta(t)}\,dt}.$$

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

Classification

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

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

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

Regression

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

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

Ergodic diffusions

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

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

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

Posterior contraction rates for Gaussian priors

View Gaussian process W as measurable map into Banach space $(\mathbb{B}, \|\cdot\|)$.

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

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

Posterior contraction rates for Gaussian priors

View Gaussian process W as measurable map into Banach space $(\mathbb{B}, \|\cdot\|)$.

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

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

Proof. Stated condition is prior mass.
Complexity is automatic due to concentration of Gaussian processes.

Posterior contraction rates for Gaussian priors

View Gaussian process W as measurable map into Banach space $(\mathbb{B}, \|\cdot\|)$.

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

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

An equivalent condition is

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

where $\phi_0(\varepsilon) = -\log \Pi(\|W\| < \varepsilon)$ is the small ball exponent and $(\mathbb{H}, \|\cdot\|_{\mathbb{H}})$ is the RKHS.

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

Brownian Motion

Theorem.
If $\theta_0 \in C^\beta[0,1]$, then the rate for Brownian motion is $n^{-\beta/2}$ if $\beta \leq 1/2$ and $n^{-1/4}$ for every $\beta \geq 1/2$.

$$\boxed{\text{The rate is } n^{-\beta/(2\beta+1)} \text{ iff } \beta = 1/2.}$$

Brownian Motion

Theorem.
If $\theta_0 \in C^\beta[0,1]$, then the rate for Brownian motion is $n^{-\beta/2}$ if $\beta \leq 1/2$ and $n^{-1/4}$ for every $\beta \geq 1/2$.

$$\boxed{\text{The rate is } n^{-\beta/(2\beta+1)} \text{ iff } \beta = 1/2.}$$

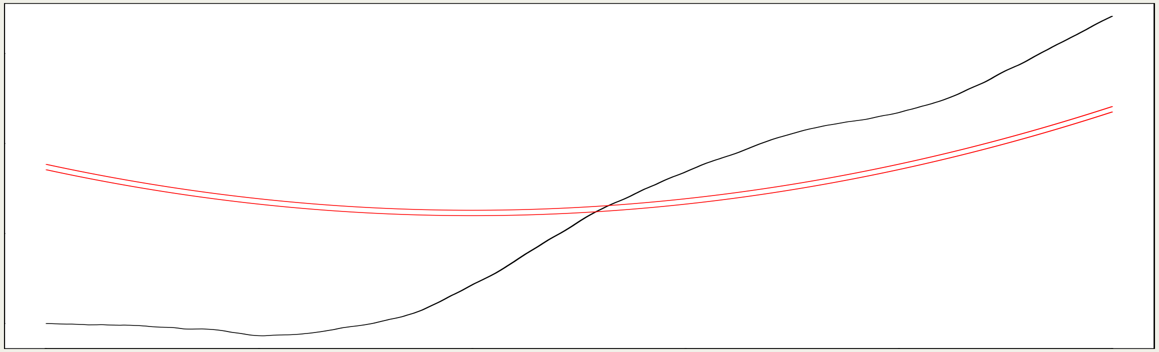
The small ball probability *of Brownian motion is*

$$\mathrm{P}\big(\|W\|_\infty < \varepsilon\big) \sim e^{-(1/\varepsilon)^2}, \qquad \varepsilon \downarrow 0.$$

This causes a $n^{-1/4}$-rate even for very smooth truths.

Theorem.
If $\theta_0 \in C^\beta[0,1]$, then the rate for $(\alpha - 1/2)$-times integrated Brownian is $n^{-(\alpha\wedge\beta)/(2\alpha+d)}$.

The rate is $n^{-\beta/(2\beta+1)}$ iff $\beta = \alpha$.



Integrated Brownian motion

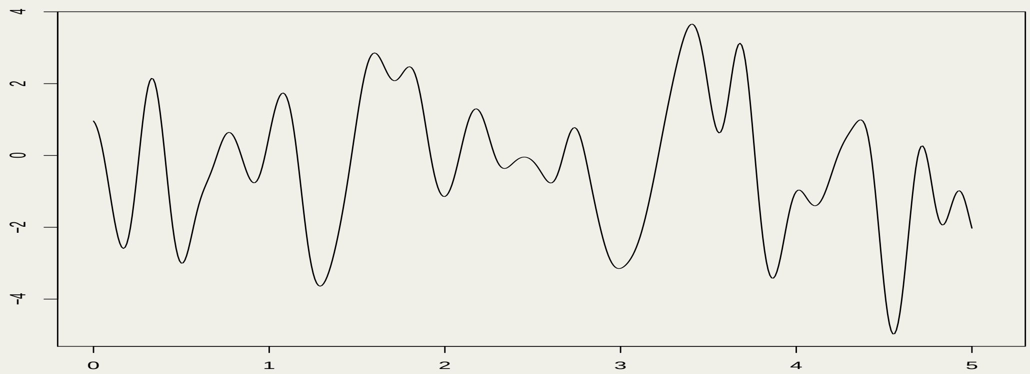
- $1/c \sim \Gamma(a, b)$.
- $(G_t : t > 0)$ the k-fold integral of Brownian motion “released at zero”,
- $W_t \sim \sqrt{c}\, G_t$.

Theorem.
The prior $W = (\sqrt{c}\, G_t : 0 \leq t \leq 1)$ *gives contraction rate* $n^{-\beta/(2\beta+1)}$ *for* $\theta_0 \in C^\beta[0, 1]$*, for any* $\beta \in (0, k+1]$.

Bayes solves the bandwidth problem.

Square exponential process

$$\mathrm{cov}(G_s, G_t) = e^{-\|s-t\|^2}, \qquad s, t \in \mathbb{R}^d.$$



Theorem.

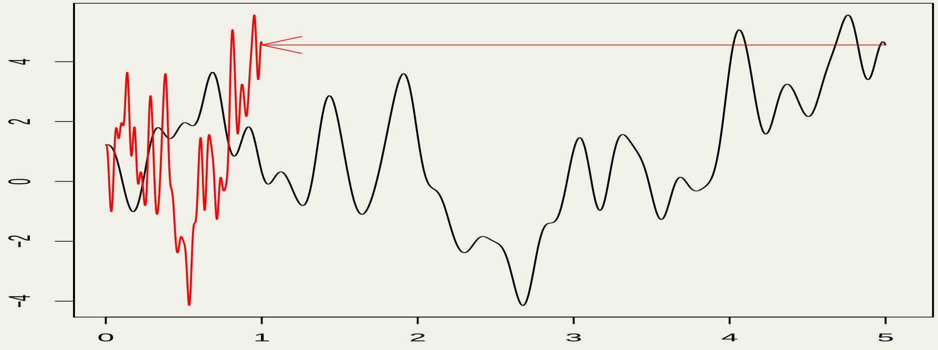
The prior G gives a rate $(\log n)^\gamma/\sqrt{n}$ if θ_0 is analytic, but may give a rate $(\log n)^{-\gamma'}$ if θ_0 is only ordinary smooth.

Square exponential process — adaptation by random time scaling

- $c^d \sim \Gamma$.
- $(G_t\colon t > 0)$ square exponential process.
- $W_t \sim G_{ct}$.

Theorem.

- *if $\theta_0 \in C^\beta[0,1]^d$, then the rate of contraction is nearly $n^{-\beta/(2\beta+d)}$.*
- *if θ_0 is supersmooth, then the rate is nearly $n^{-1/2}$.*



Gaussian processes: summary

- Recovery is best if prior ‘matches’ truth.
- Mismatch slows down, but does not prevent, recovery.
- Mismatch can be prevented by using hyperparameters.

Dirichlet process

Finite-dimensonal Dirichlet distribution

Definition. $(\theta_1, \ldots, \theta_k)$ *possesses a* $\mathrm{Dir}(k, \alpha_1, \ldots, \alpha_k)$*-distribution for* $\alpha_i > 0$ *it has (Lebesgue) density on the unit simplex proportional to*

$$\theta \mapsto \theta_1^{\alpha_1 - 1} \cdots \theta_k^{\alpha_k - 1}.$$

Finite-dimensonal Dirichlet distribution

Definition. $(\theta_1, \ldots, \theta_k)$ *possesses a* $\mathrm{Dir}(k, \alpha_1, \ldots, \alpha_k)$*-distribution for* $\alpha_i > 0$ *it has (Lebesgue) density on the unit simplex proportional to*

$$\theta \mapsto \theta_1^{\alpha_1 - 1} \cdots \theta_k^{\alpha_k - 1}.$$

We extend to $\alpha_i = 0$ for one or more i on the understanding that $\theta_i = 0$.

Theorem. *If* $\theta \sim \mathrm{Dir}(k, \alpha)$ *and* $N |\, \theta \sim \mathrm{Multinom}(n, \alpha)$*, then* $\theta |\, N \sim \mathrm{Dir}(k, \alpha + N)$*.*

Proof. $\Pi(d\theta |\, N) \propto \binom{n}{N} \prod_{i=1}^k \theta_i^{N_i} \prod_{i=1}^k \theta_i^{\alpha_i - 1}$.

Dirichlet process

Definition. *A random measure P on a measurable space $(\mathfrak{X}, \mathscr{X})$ is a* Dirichlet process *with* base measure α, *if for every partition $A_1, \ldots, A_k$ of $\mathfrak{X}$,*

$$\big(P(A_1), \ldots, P(A_k)\big) \sim \mathrm{Dir}\big(k; \alpha(A_1), \ldots, \alpha(A_k)\big).$$

Dirichlet process

Definition. *A random measure* P *on a measurable space* $(\mathfrak{X}, \mathcal{X})$ *is a* Dirichlet process *with* base measure α, *if for every partition* $A_1, \ldots, A_k$ *of* $\mathfrak{X}$,

$$\big(P(A_1), \ldots, P(A_k)\big) \sim \mathrm{Dir}\big(k; \alpha(A_1), \ldots, \alpha(A_k)\big).$$

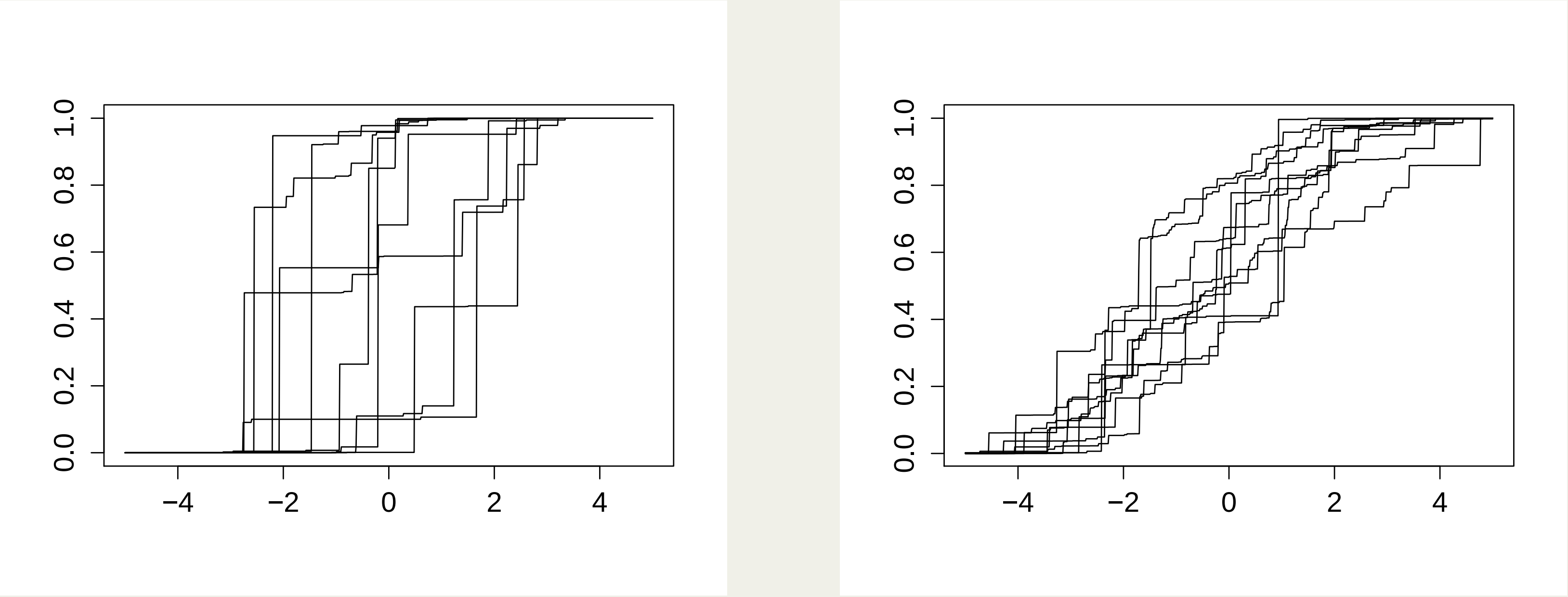
Lemma. $\mathrm{E}P(A) = \frac{\alpha(A)}{\|\alpha\|}$.

Dirichlet process

Definition. *A random measure P on a measurable space $(\mathfrak{X}, \mathcal{X})$ is a* **Dirichlet process** *with* **base measure** *α, if for every partition $A_1, \ldots, A_k$ of $\mathfrak{X}$,*

$$\big(P(A_1), \ldots, P(A_k)\big) \sim \mathrm{Dir}\big(k; \alpha(A_1), \ldots, \alpha(A_k)\big).$$

Lemma. $\mathrm{E}P(A) = \frac{\alpha(A)}{\|\alpha\|}$.



Cumulative distribution functions of 10 draws from the Dirichlet process with base measure $N(0, 2)$ (left) and $10N(0, 2)$ (right).

Dirichlet process

Definition. *A random measure P on a measurable space $(\mathfrak{X}, \mathcal{X})$ is a* Dirichlet process *with* base measure α*, if for every partition $A_1, \ldots, A_k$ of $\mathfrak{X}$,*

$$\big(P(A_1), \ldots, P(A_k)\big) \sim \mathrm{Dir}\big(k; \alpha(A_1), \ldots, \alpha(A_k)\big).$$

Lemma. $\mathrm{E}P(A) = \frac{\alpha(A)}{\|\alpha\|}$.

Theorem. *If $\theta_1, \theta_2, \ldots \overset{iid}{\sim} \bar{\alpha}$ and $Y_1, Y_2, \ldots \overset{iid}{\sim} \mathrm{Be}(1, M)$ are independent, and $W_j = (1 - Y_1) \cdots (1 - Y_{j-1}) Y_j$, then*

$$P := \sum_{j=1}^{\infty} W_j \delta_{\theta_j} \sim \mathrm{DP}(M\bar{\alpha}).$$

Conjugacy of Dirichlet process

$$P \sim \mathrm{DP}(\alpha), \qquad\qquad X_1, X_2, \ldots \,|\, P \overset{\text{iid}}{\sim} P.$$

Theorem. $P | \, X_1, \ldots, X_n \sim \mathrm{DP}(\alpha + n\mathbb{P}_n)$, *for* $\mathbb{P}_n = n^{-1} \sum_{i=1}^n \delta_{X_i}$.

Conjugacy of Dirichlet process

$$P \sim \mathrm{DP}(\alpha), \qquad\qquad X_1, X_2, \ldots \,|\, P \overset{\text{iid}}{\sim} P.$$

Theorem. $P|\, X_1, \ldots, X_n \sim \mathrm{DP}(\alpha + n\mathbb{P}_n)$, *for* $\mathbb{P}_n = n^{-1}\sum_{i=1}^n \delta_{X_i}$.

Proof.
$\big(P(A_1), \ldots, P(A_k)\big)|\, X_1, \ldots, X_n \sim \big(P(A_1), \ldots, P(A_k)\big)|\, (N_1, \ldots, N_k)$,
where $N_i = n\mathbb{P}_n(A_i)$.
Apply result for finite-dimensional Dirichlet.

Conjugacy of Dirichlet process

$$P \sim \mathrm{DP}(\alpha), \qquad\qquad X_1, X_2, \ldots \,|\, P \overset{\text{iid}}{\sim} P.$$

Theorem. $P|\, X_1, \ldots, X_n \sim \mathrm{DP}(\alpha + n\mathbb{P}_n)$, *for* $\mathbb{P}_n = n^{-1}\sum_{i=1}^n \delta_{X_i}$.

Corollary. $\mathrm{E}\big(P(A)|\, X_1, \ldots, X_n\big) = \frac{\alpha(A)}{\|\alpha\|+n} + \frac{n}{\|\alpha\|+n}\mathbb{P}_n(A)$.

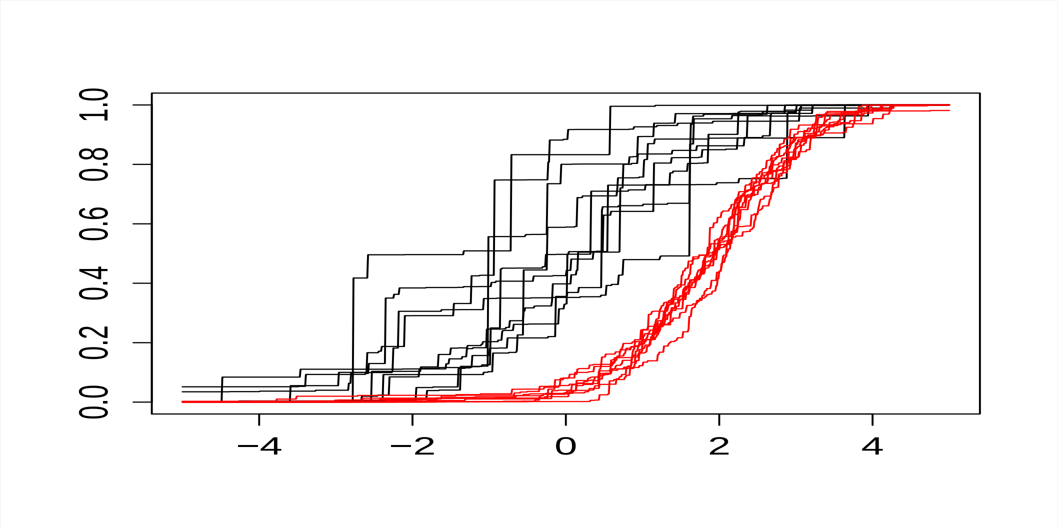
Conjugacy of Dirichlet process

$$P \sim \mathrm{DP}(\alpha), \qquad\qquad X_1, X_2, \ldots \,|\, P \overset{\text{iid}}{\sim} P.$$

Theorem. $P|\, X_1, \ldots, X_n \sim \mathrm{DP}(\alpha + n\mathbb{P}_n)$, *for* $\mathbb{P}_n = n^{-1}\sum_{i=1}^n \delta_{X_i}$.

Theorem. *The posterior distribution of $P(A)$ satisfies a Bernstein-von Mises theorem:*

$$\Big\|\Pi\big(P(A) \in \cdot \,|\, X_1, \ldots, X_n\big) - N\Big(\mathbb{P}_n(A), \frac{P_0(A)(1 - P_0)(A)}{n}\Big)\Big\| \to 0.$$

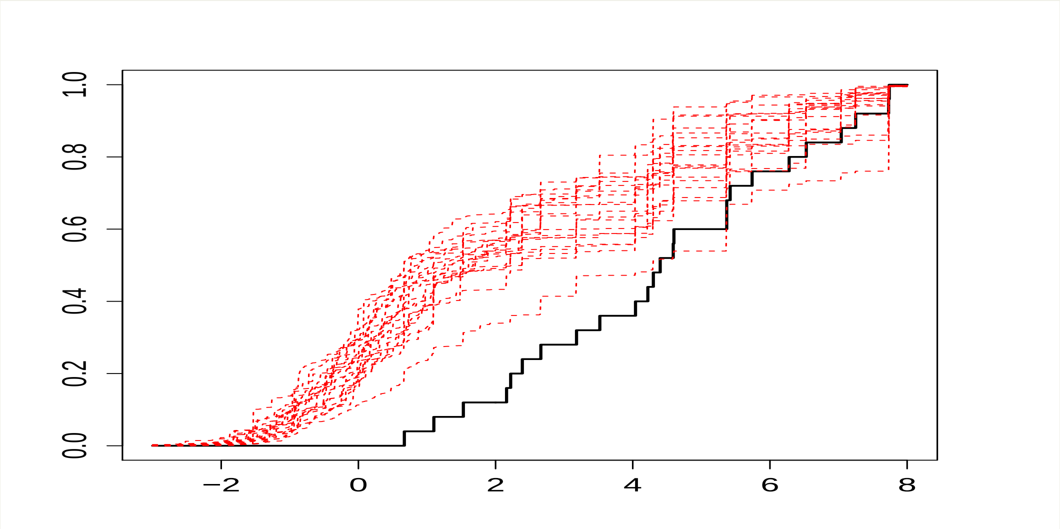


Pitman-Yor process

$\theta_1, \theta_2, \ldots \overset{\text{iid}}{\sim} \bar{\alpha}$ independent of $Y_j \overset{\text{ind}}{\sim} \mathrm{Be}(1-\sigma, M+j\sigma)$.
$W_j = (1-Y_1)\cdots(1-Y_{j-1})Y_j$.

$$P := \sum_{j=1}^{\infty} W_j \delta_{\theta_j}.$$

Theorem. *The posterior distribution based on $X_1, \ldots, X_n | P \sim P$ is inconsistent unless $\sigma = 0$ or $P_0 = \alpha$ or P_0 is discrete.*

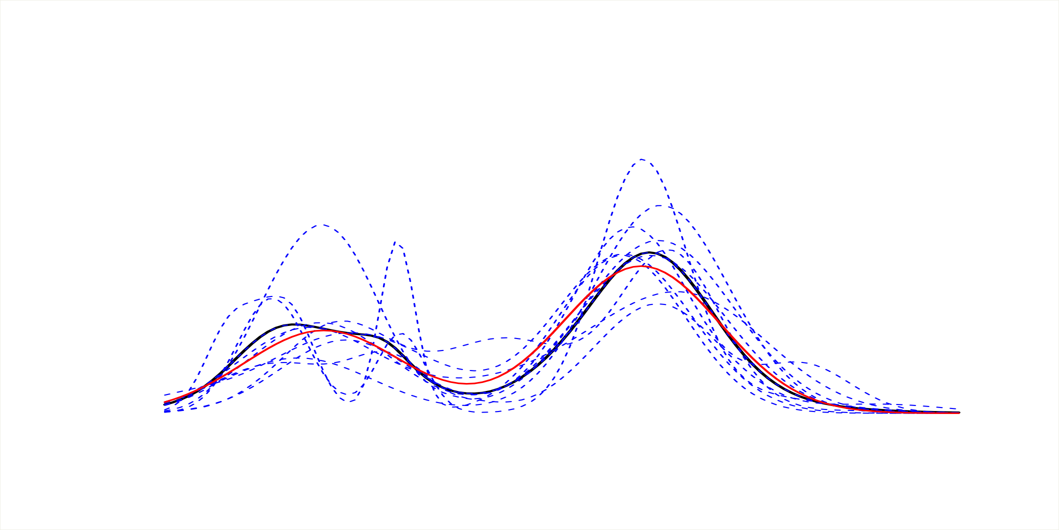


Dirichlet process mixtures

Dirichlet normal mixtures

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

$$X_1,\ldots,X_n|\,F,\sigma \stackrel{\text{iid}}{\sim} p_{F,\sigma}, \qquad F\sim \mathrm{DP}(\alpha) \quad \perp \quad \sigma\sim\pi.$$



Posterior mean (solid black) and 10 draws of the posterior distribution

for a sample of size 50 from a mixture of two normals (red).

Dirichlet normal mixtures

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

$$X_1, \ldots, X_n | \, F, \sigma \overset{\text{iid}}{\sim} p_{F,\sigma}, \qquad F \sim \mathrm{DP}(\alpha) \quad \perp \quad \sigma \sim \pi.$$

Two cases for the true density p_0:

- Supersmooth: $p_0 = p_{F_0,\sigma_0}$, for compactly supported F_0.
  Take prior for σ with continuous positive density on $(a, b) \ni \sigma_0$.
- Ordinary smooth: p_0 has β derivatives and exponentially small tails.
  Take $1/\sigma$ a priori Gamma distributed.

Dirichlet normal mixtures

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

$$X_1,\ldots,X_n|\, F,\sigma \overset{\text{iid}}{\sim} p_{F,\sigma}, \qquad F \sim \mathrm{DP}(\alpha) \quad \perp \quad \sigma\sim\pi.$$

Two cases for the true density p_0:

- Supersmooth: $p_0 = p_{F_0,\sigma_0}$, for compactly supported F_0.
  Take prior for σ with continuous positive density on $(a,b)\ni\sigma_0$.
- Ordinary smooth: p_0 has β derivatives and exponentially small tails.
  Take $1/\sigma$ a priori Gamma distributed.

Theorem. *Hellinger rate of contraction is*

- *nearly $n^{-1/2}$ in the supersmooth case.*
- *nearly $n^{-\beta/(2\beta+1)}$ in the ordinary smooth case.*

Dirichlet normal mixtures

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

$$X_1,\ldots,X_n|\,F,\sigma \overset{\text{iid}}{\sim} p_{F,\sigma}, \qquad F \sim \mathrm{DP}(\alpha) \quad \perp \quad \sigma \sim \pi.$$

Two cases for the true density p_0:

- Supersmooth: $p_0 = p_{F_0,\sigma_0}$, for compactly supported F_0.
  Take prior for σ with continuous positive density on $(a,b) \ni \sigma_0$.
- Ordinary smooth: p_0 has β derivatives and exponentially small tails.
  Take $1/\sigma$ a priori Gamma distributed.

Adaptation to any smoothness with a **Gaussian** kernel!
Kernel density estimation needs higher order kernels.

$$\frac{1}{n\sigma}\sum_{i=1}^n \phi\Big(\frac{x-X_i}{\sigma}\Big) = p_{\mathbb{F}_n,\sigma}(x).$$

Key lemma: finite approximation

Lemma. *For any probability measure F on the interval $[0,1]$ there exists a discrete probability measure F' on with at most*

$$N \lesssim \log \frac{1}{\varepsilon}$$

support points, such that

$$\|p_{F,1} - p_{F',1}\|_\infty \lesssim \varepsilon, \qquad \|p_{F,1} - p_{F',1}\|_1 \lesssim \varepsilon \Big(\log \frac{1}{\varepsilon}\Big)^{1/2}.$$

Proof.

- Match moments of F and F' up to order $\log(1/\epsilon)$.
- Taylor expand the kernel $z \mapsto \phi(x-z)$.