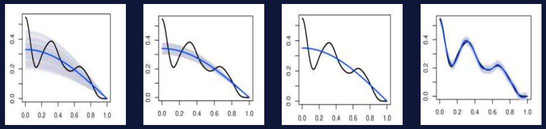
Bayesian Statistics in High Dimensions

Lecture 1: Curve and surface estimation

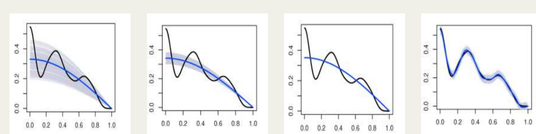
Aad van der Vaart

Universiteit Leiden, Netherlands



47th John H. Barrett Memorial Lectures, Knoxville, Tenessee, May 2017

Contents



Introduction

The Bayesian paradigm

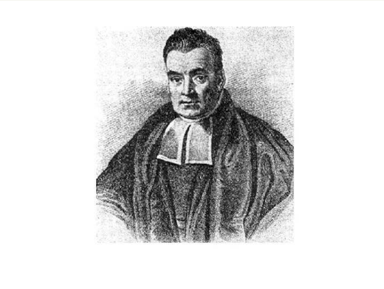
- A parameter θ is generated according to a prior distribution Π.
- Given θ the data X is generated according to a measure P_θ.

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

- Given observed data X the statistician computes the conditional distribution of θ given X, the posterior distribution:

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

The Bayesian paradigm



- A parameter θ is generated according to a prior distribution Π.
- Given θ the data X is generated according to a measure P_θ.

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

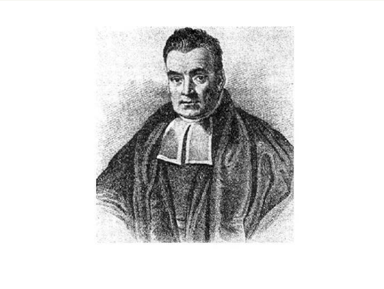
- Given observed data X the statistician computes the conditional distribution of θ given X, the posterior distribution:

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

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 θ *binomial* (n, θ).

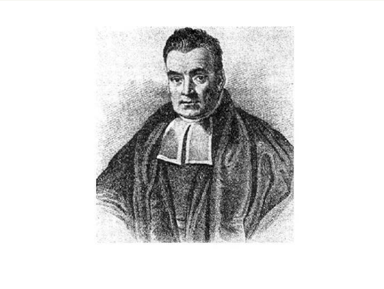
The posterior distribution is then *Beta*$(X+1, n-X+1)$.

$$d\Pi(\theta) = 1, \qquad 0 < \theta < 1,$$

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

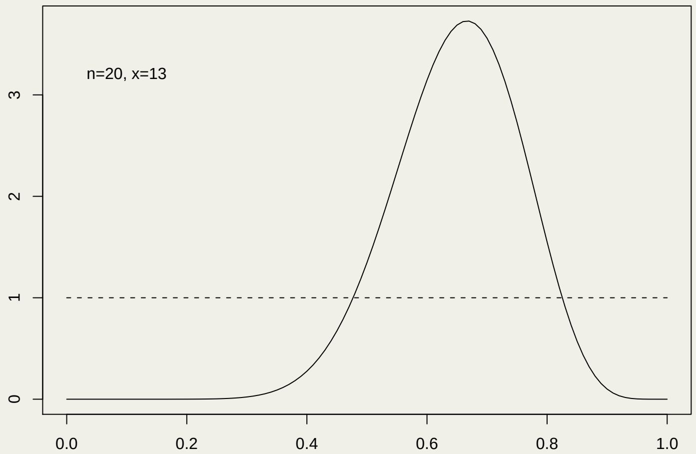
$$d\Pi(\theta|\,X) = \theta^X(1-\theta)^{n-X}\cdot 1.$$

Reverend Thomas

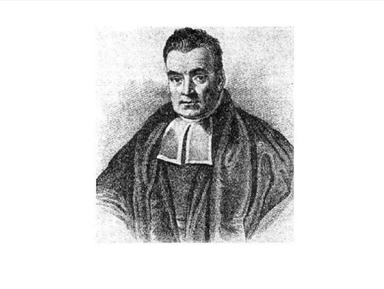


Thomas Bayes (1702–1761, 1763) followed this argument with θ possessing the *uniform* distribution and X given θ *binomial* (n, θ).

The posterior distribution is then *Beta*$(X + 1, n - X + 1)$.

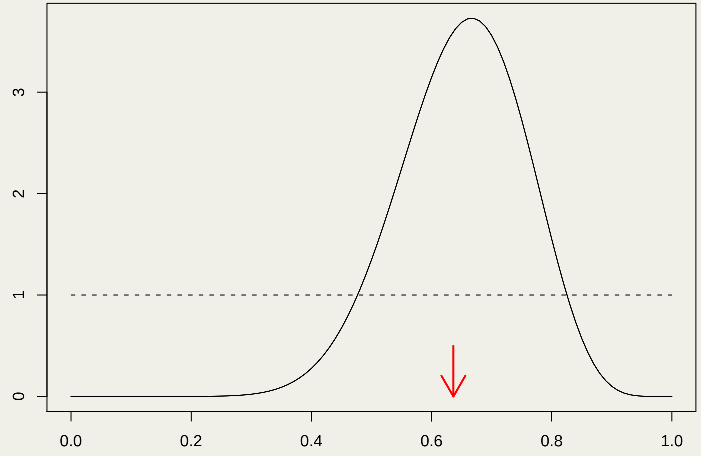


Reverend Thomas



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

The posterior distribution is then *Beta*$(X + 1, n - X + 1)$.

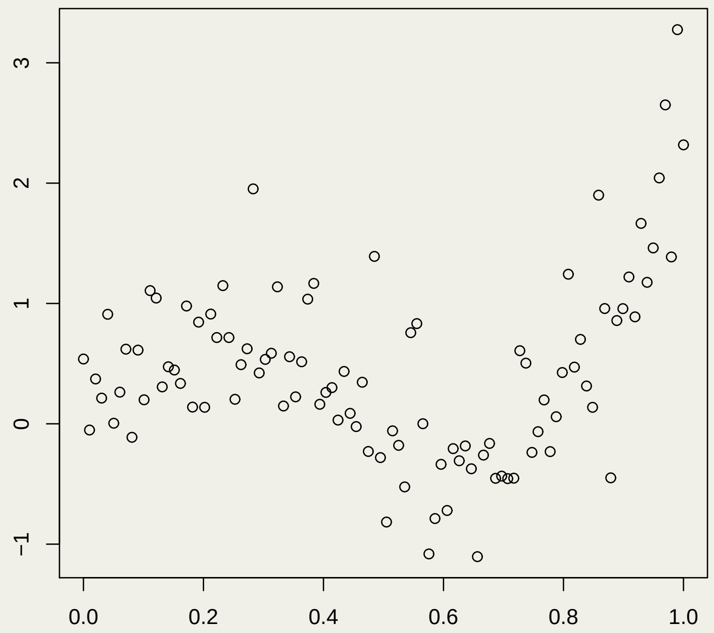


Nonparametric Bayes

If θ is a function, then the prior is a probability distribution on a function space. So is the posterior, given the data.
Bayes's formula does not change:

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

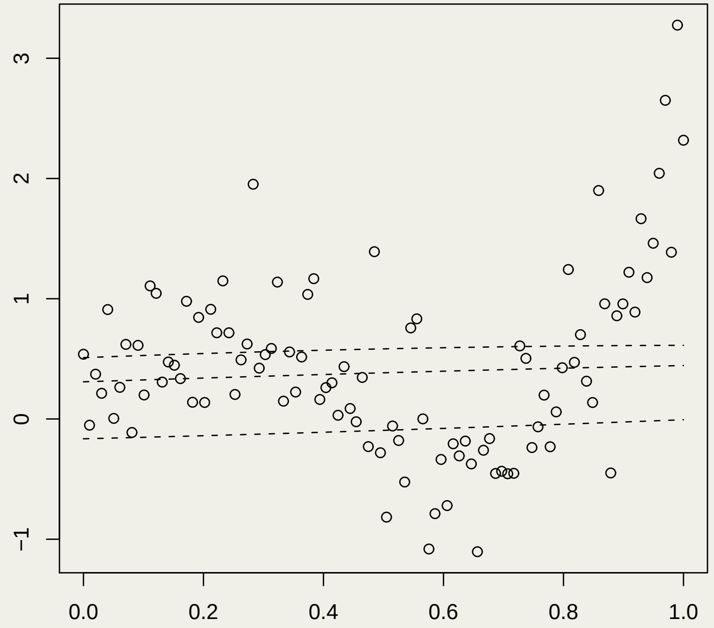
Prior and posterior can be visualized by plotting functions that are simulated from these distributions.



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

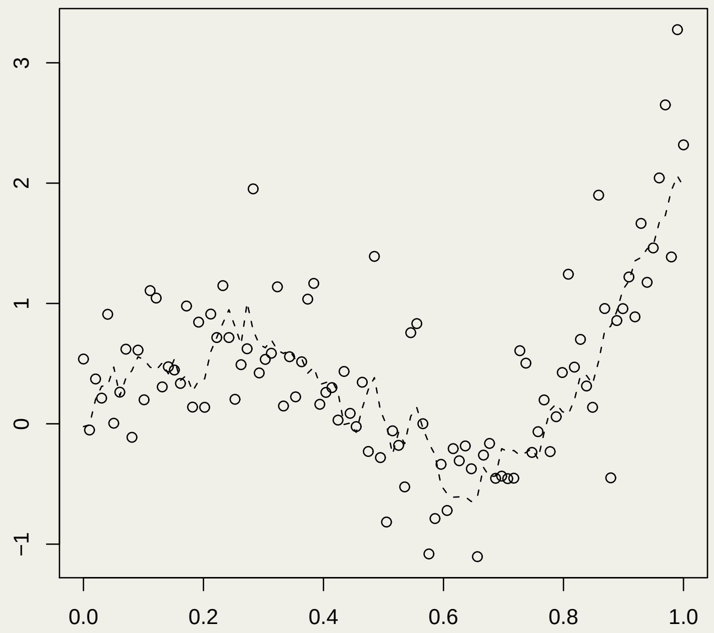


Nonparametric Bayes

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

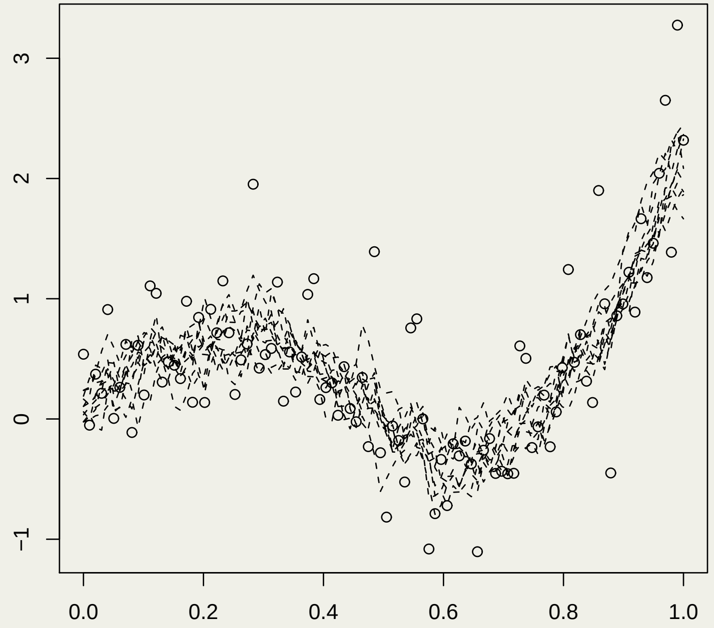


Nonparametric Bayes

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

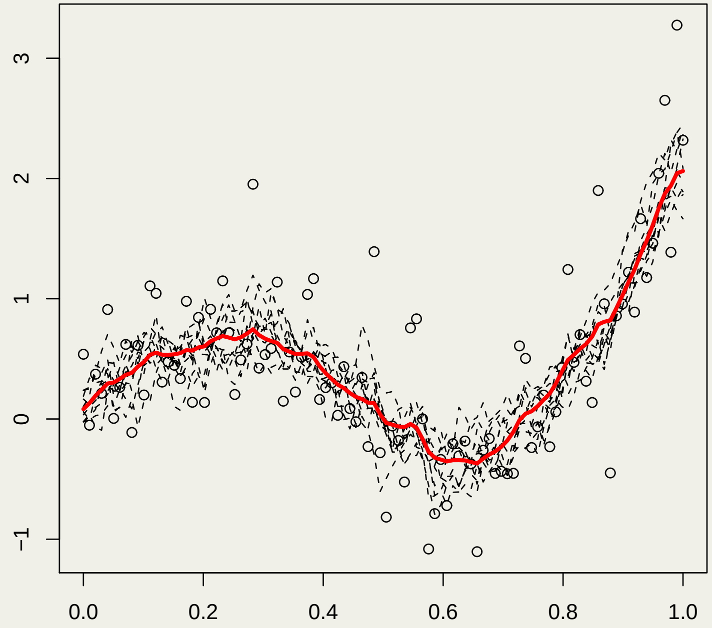


Nonparametric Bayes

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

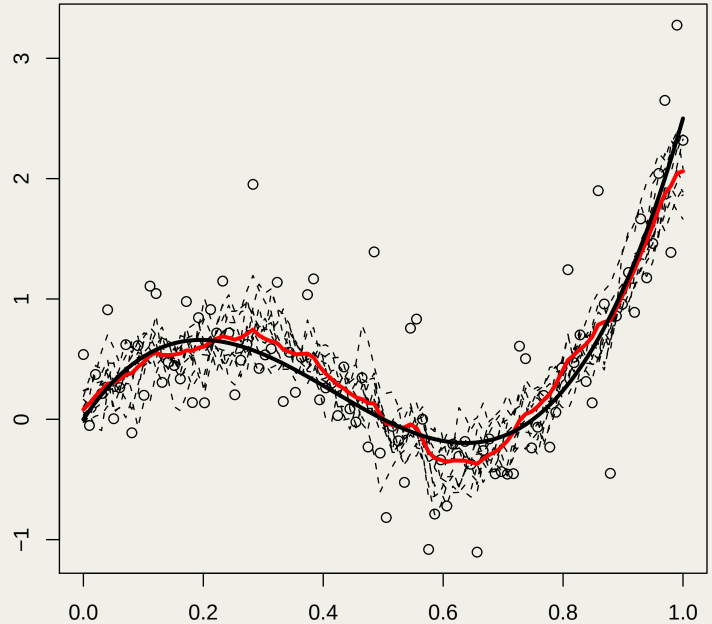


Nonparametric Bayes

If θ is a function, then the prior is a probability distribution on a function space. So is the posterior, given the data.
Bayes's formula does not change:

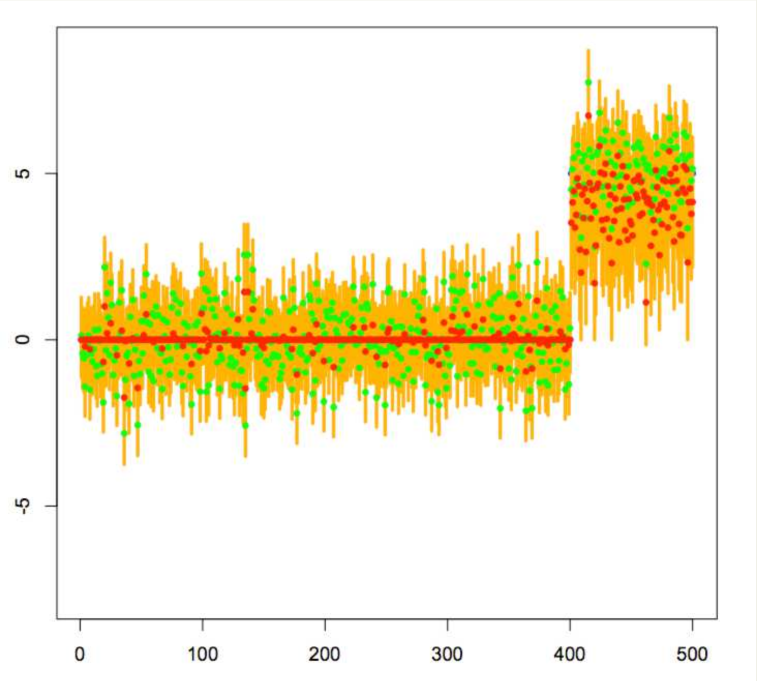
$$d\Pi(\theta|\,X) \propto p_\theta(X)\,d\Pi(\theta).$$

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



High-dimensional Bayes

A high-dimensional parameter vector (or matrix) is similar to a function.
Visualization may be through a plot versus an index.



Parameters $\theta_1, \ldots, \theta_{500}$ (vertical) versus index $1, \ldots, 500$.
Red dots: marginal posterior medians
Orange: marginal credible intervals

Green dots: data points.

Frequentist Bayes

Assume the data X is generated according to a given parameter θ_0.
Consider the posterior $\Pi(\theta \in \cdot \mid X)$ as a given random measure.

Frequentist Bayes

Assume the data X is generated according to a given parameter θ_0.
Consider the posterior $\Pi(\theta \in \cdot \mid X)$ as a given random measure.

Recovery

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

Frequentist Bayes

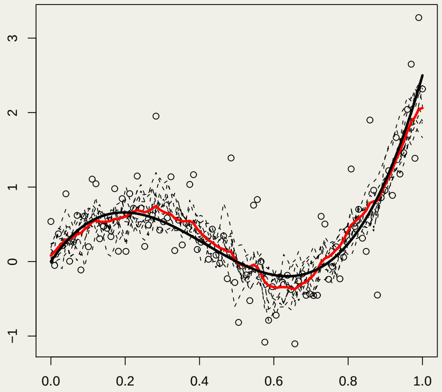
Assume the data X is generated according to a given parameter θ_0.
Consider the posterior $\Pi(\theta \in \cdot \mid X)$ as a given random measure.

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 Bayes

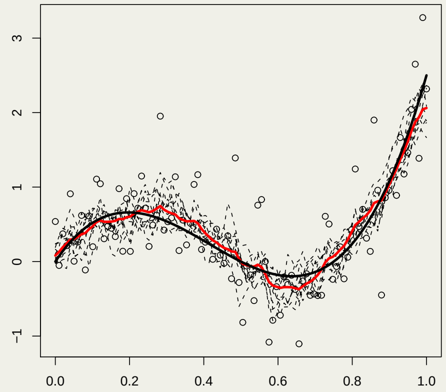
Assume the data X is generated according to a given parameter θ_0.
Consider the posterior $\Pi(\theta \in \cdot \mid X)$ as a given random measure.

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.



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

- We want $\Pi_n(\cdot \mid X^{(n)}) \rightsquigarrow \delta_{\theta_0}$, at a good rate.
- We like the *coverage* of a set of large posterior mass to be large.

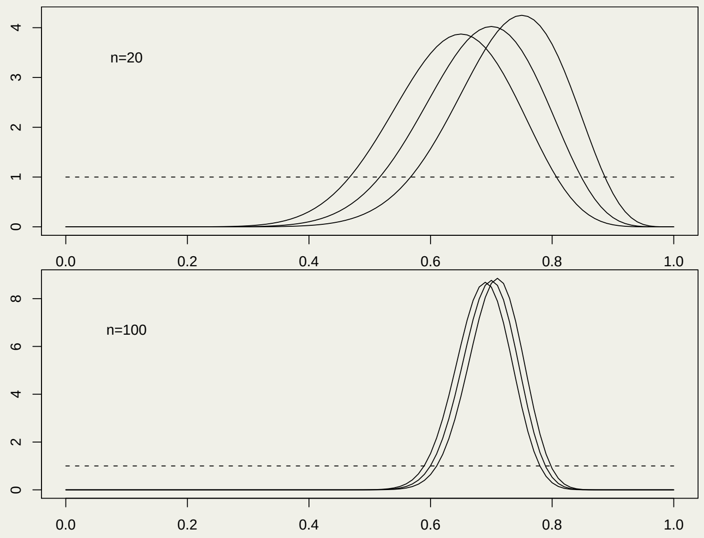
Parametric models Laplace, Bernstein, von Mises, Le Cam 1989

Suppose the data are a random sample $X_1, \ldots, X_n$ from a density $x \mapsto p_\theta(x)$ that is smoothly and **identifiably** parametrized by $\theta \in \mathbb{R}^d$.

Theorem. *Under $P^n_{\theta_0}$, for any prior with positive density,*

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

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



Parametric models Laplace, Bernstein, von Mises, Le Cam 1989

Suppose the data are a random sample $X_1, \ldots, X_n$ from a density $x \mapsto p_\theta(x)$ that is smoothly and **identifiably** parametrized by $\theta \in \mathbb{R}^d$.

Theorem. *Under $P^n_{\theta_0}$, for any prior with positive density,*

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

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 $\{\theta: n(\theta - \tilde\theta_n)^T I_{\tilde\theta_n}(\theta - \tilde\theta_n) \le \chi^2_{d,1-\alpha}\}$.

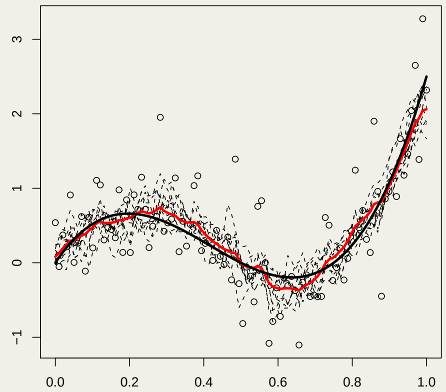
These lectures

Recovery and uncertainty quantification for high-dimensional models.

LECTURE 1: Curve and surface fitting.

LECTURE 2: Sparsity.

> Interest
> Reliability of the posterior distribution for natural priors,
> in particular for priors that adapt to complexity in the data.



These lectures

In these lectures no attention for computing or simulating the posterior.

For small datasets: Markov Chain Monte Carlo.

For bigger datasets: iterative methods and approximations, e.g.:

- ABC
- expectation propagation
- variational Bayes

Interest in scalable methods for very big datasets is recent.
E.g. variational methods, distributed computations, stochastic descent.

Recovery

Rate of contraction

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

Definition. *Posterior* contraction rate *at* θ_0 *is* ϵ_n *if, for large* M*,*

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

Rate of contraction

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

Definition. *Posterior* contraction rate *at* θ_0 *is* ϵ_n *if, for large* M*,*

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

Benchmark rate for curve fitting:

A function θ of d variables with bounded derivatives of order β is estimable based on n observations at rate

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

Proposition. *If the contraction rate at* θ_0 *is* ϵ_n*, then the center* $\hat\theta_n$ *of a (nearly) smallest ball of posterior mass* $\geq 1/2$ *satisfies* $d(\hat\theta_n, \theta_0) = O_P(\epsilon_n)$*.*

Basic contraction theorem (Ghosal, Ghosh, vdV 2000)

- $p \sim \Pi$, prior on set of densities $\mathcal{P}$.
- $X_1, \ldots, X_n | p \overset{\text{iid}}{\sim} p$.

$$B(p_0, \varepsilon) = \Big\{p \colon P_0 \log \frac{p_0}{p} < \varepsilon^2, P_0\Big(\log \frac{p_0}{p}\Big)^2 < \varepsilon^2\Big\}.$$

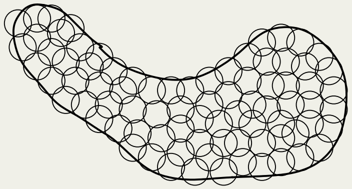
Theorem. *Let d convex metric bounded above by Hellinger metric such that that there exist $\mathcal{P}_n \subset \mathcal{P}$ and $C > 0$ with*

$$\Pi_n\big(B(p_0, \varepsilon_n)\big) \geq e^{-Cn\epsilon_n^2} \qquad \textit{(prior mass)}$$

$$\log N\big(\epsilon_n, \mathcal{P}_n, d\big) \leq n\epsilon_n^2 \quad \textit{and} \quad \Pi_n(\mathcal{P}_n^c) \leq e^{-(C+4)n\epsilon_n^2} \qquad \textit{(complexity)}.$$

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

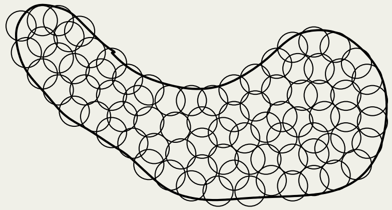
$N(\epsilon, \mathcal{P}, d)$ is the minimal number of d-balls of radius ϵ needed to cover $\mathcal{P}$.



[Hellinger distance: $h(p, q) = \|\sqrt{p} - \sqrt{q}\|_2$.]

Interpretation

Let $p_1, \ldots, p_N$ in $\mathcal{P}$ be a maximal set with $d(p_i, p_j) \geq \epsilon_n$.



Under the complexity bound,

$$N \asymp N(\epsilon_n, \mathcal{P}, d) \geq e^{n\epsilon_n^2}.$$

If prior mass were evenly distributed, then each ball of radius $\varepsilon_n/2$ would have mass of order

$$\frac{1}{N} \leq e^{-n\epsilon_n^2}.$$

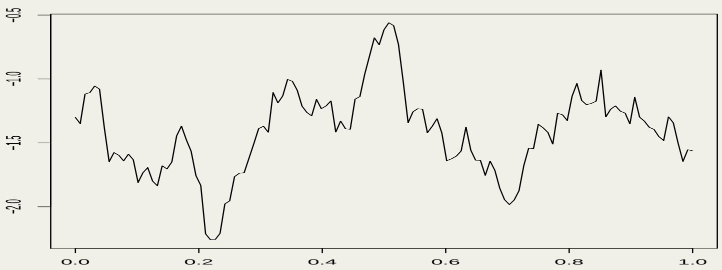
This is the order of the prior mass bound.

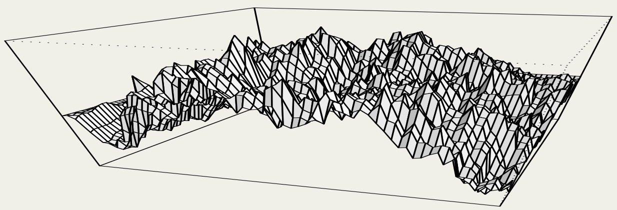
> Suggestion:
> The conditions can be satisfied for every $p_0 \in \mathcal{P}$ if the prior *"distributes its mass uniformly over $\mathcal{P}$, at discretization level ϵ_n"*.

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 $\theta : T \to \mathbb{R}$.



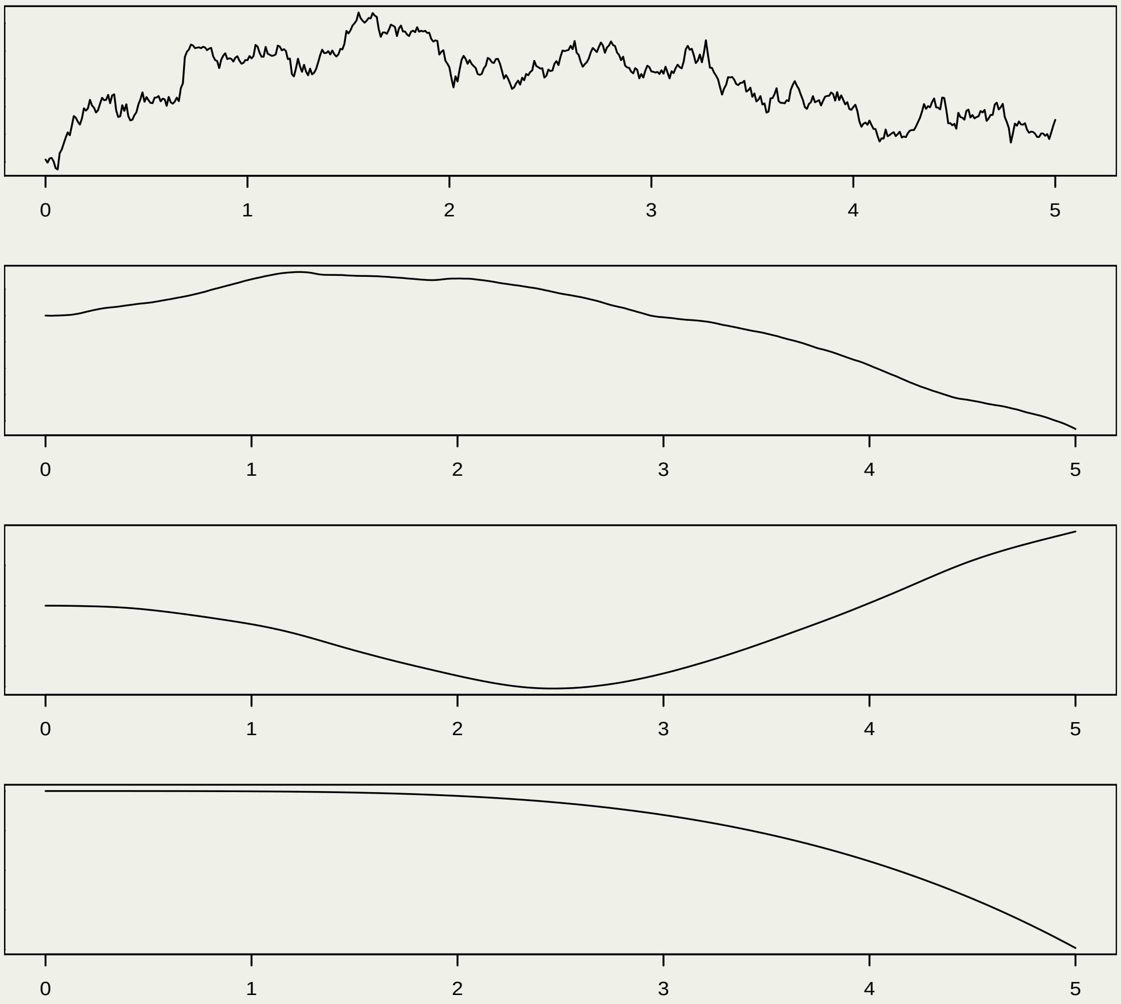


W is a Gaussian process if
$(W_{t_1}, \ldots, W_{t_k})$ is multivariate Gaussian, for every $t_1, \ldots, t_k$.

Mean and covariance function:

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

Example: Brownian motion and its primitives



0, 1, 2 and 3 times integrated Brownian motion

Posterior contraction rates for Gaussian priors vdV+van Zanten, 2007-2011

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

Theorem. *If statistical distances combine appropriately with $\|\cdot\|$, 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 vdV+van Zanten, 2007-2011

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

Theorem. *If statistical distances combine appropriately with $\|\cdot\|$, then the posterior rate is ε_n if*

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

Proof.

- The stated condition is prior mass.
- Complexity is automatic due to concentration of Gaussian processes.

Posterior contraction rates for Gaussian priors vdV+van Zanten, 2007-2011

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

Theorem. *If statistical distances combine appropriately with $\|\cdot\|$, 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, for $(\mathbb{H}, \|\cdot\|_{\mathbb{H}})$ the RKHS,

$$\mathrm{P}\big(\|W\| < \varepsilon_n\big) \geq e^{-n\varepsilon_n^2} \qquad \text{AND} \qquad \inf_{h \in \mathbb{H}: \|h - w_0\| < \varepsilon_n} \|h\|_{\mathbb{H}}^2 \leq n{\varepsilon_n}^2.$$

- *Both inequalities give lower bound on ε_n.*
- *The first does not depend on w_0.*

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.)

Brownian Motion prior

Theorem. *If $\theta_0 \in C^\beta[0,1]$, then rate for Brownian motion is*

- $n^{-\beta/2}$ *if* $\beta \leq 1/2$,
- $n^{-1/4}$ *for every* $\beta \geq 1/2$.

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

Brownian Motion prior

Theorem. *If $\theta_0 \in C^\beta[0,1]$, then rate for Brownian motion is*

- $n^{-\beta/2}$ *if* $\beta \leq 1/2$,
- $n^{-1/4}$ *for every* $\beta \geq 1/2$.

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

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

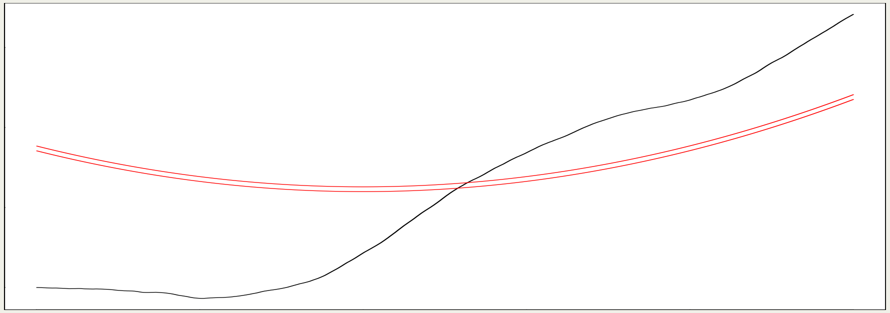
Small ball probability causes $n^{-1/4}$-rate even for smooth truths.

Example: integrated Brownian Motion prior

Theorem. *If* $\theta_0 \in C^\beta[0,1]$*, then rate for* $(\alpha - 1/2)$*-times integrated Brownian motion is*

- $n^{-\beta/(2\alpha+1)}$, *if* $\beta \leq \alpha$,
- $n^{-\alpha/(2\alpha+1)}$, *if* $\beta \geq \alpha$.

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



$$\mathrm{P}\big(\|W\|_\infty < \varepsilon\big) \sim e^{-(1/\varepsilon)^{1/\alpha}}.$$

Integrated Brownian motion prior — adaptation by random scaling

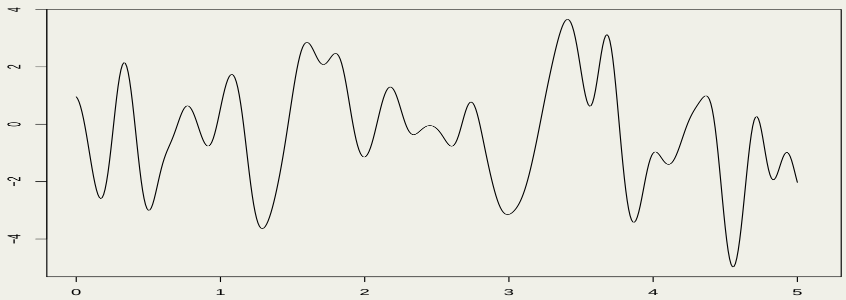
- $1/c \sim \Gamma(a, b)$.
- $(G_t\colon t > 0)$ k-times integrated Brownian motion "released at zero",
- $W_t \sim \sqrt{c}\, G_t$.

Theorem. *If $\theta_0 \in C^\beta[0,1]$ rate for prior W is $n^{-\beta/(2\beta+1)}$, for any $\beta \in (0, k+1]$.*

Bayes solves the bandwidth problem.

Example: square exponential prior

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



$$\mathrm{P}\big(\|W\|_\infty < \varepsilon\big) \gtrsim e^{-C(\log \varepsilon^{-1})^{1+d/2}}.$$

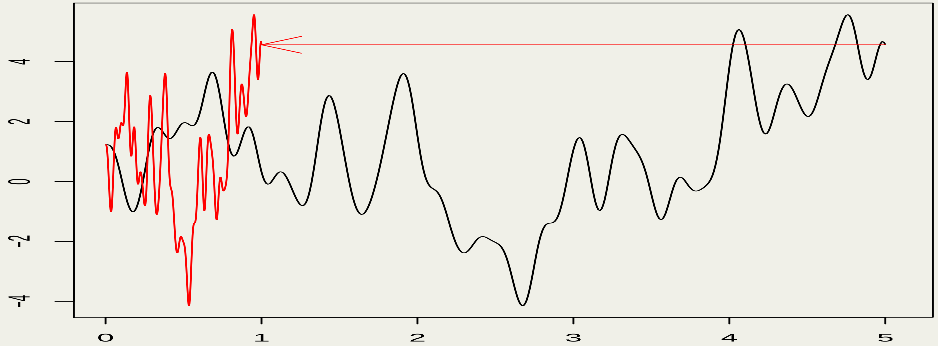
Theorem. *For prior G a is $(\log n)^\gamma/\sqrt{n}$ if θ_0 is analytic, but may be $(\log n)^{-\gamma'}$ if θ_0 is only ordinary smooth.*

Square exponential prior — adaptation by random time scaling

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

Theorem. *For prior $(W_t: t \in [0,1]^d)$:*

- *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 analytic, then the rate is nearly $n^{-1/2}$.*



Gaussian processes: summary

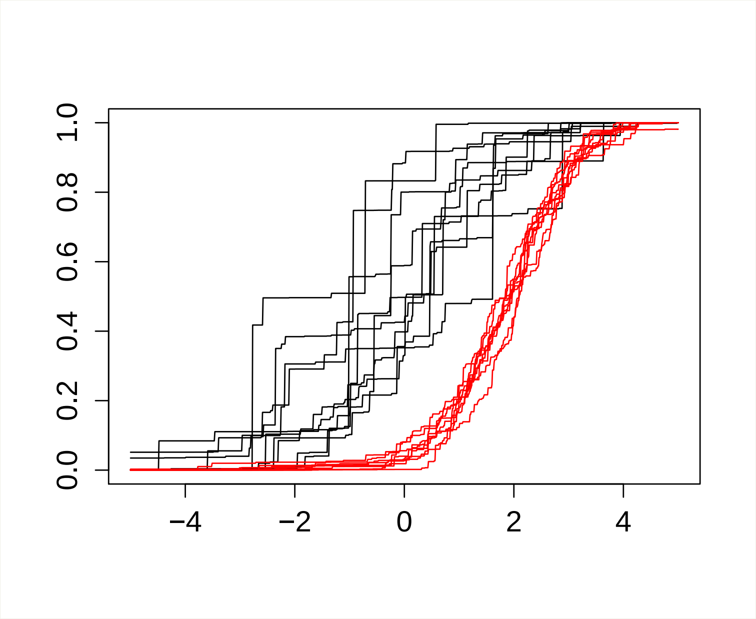
Recovery is best if prior ‘matches’ truth.
Mismatch slows down, but does not prevent, recovery.
Mismatch can be prevented by using hyperparameters.

Dirichlet process mixtures

Dirichlet process [Ferguson 1973]

Definition. *A* Dirichlet process *is a random measure P on $(\mathfrak{X}, \mathscr{X})$ such that for every partition $A_1, \ldots, A_k$ of $\mathfrak{X}$,*

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

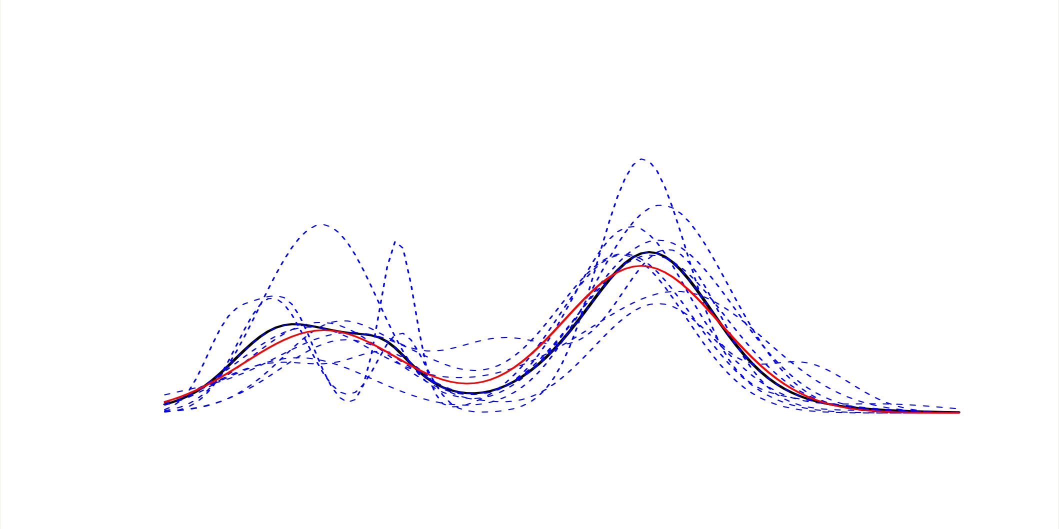


Draws from Dirichlet prior (black) and posterior based on random sample from P (red).

Dirichlet normal mixtures [Ghosal, vdV, Rousseau, Kruijer, Tokdar, Shen, 2001–2013]

- $F \sim$ Dirichlet process, independent of $1/c \sim \Gamma(a,b)$.
- Data: $X_1, \ldots, X_n |\, F, c \stackrel{\text{iid}}{\sim} p_{F,c}$, for

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



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 [Ghosal, vdV, Rousseau, Kruijer, Tokdar, Shen, 2001–2013]

- $F \sim$ Dirichlet process, independent of $1/c \sim \Gamma(a, b)$.
- Data: $X_1, \ldots, X_n | F, c \overset{\text{iid}}{\sim} p_{F,c}$, for

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

Theorem. *Hellinger rate of contraction for* $X_1, \ldots, X_n \overset{\text{iid}}{\sim} p_0$ *is*

- *nearly* $n^{-1/2}$ *if* $p_0 = p_{F_0, c_0}$*, some* F_0, c_0.
- *nearly* $n^{-\beta/(2\beta+1)}$ *if* p_0 *has* β *derivatives and exponentially small tails.*

Dirichlet normal mixtures [Ghosal, vdV, Rousseau, Kruijer, Tokdar, Shen, 2001–2013]

- $F \sim$ Dirichlet process, independent of $1/c \sim \Gamma(a,b)$.
- Data: $X_1,\ldots,X_n |\, F, c \overset{\text{iid}}{\sim} p_{F,c}$, for

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

Theorem. *Hellinger rate of contraction for $X_1,\ldots,X_n \overset{\text{iid}}{\sim} p_0$ is*

- *nearly $n^{-1/2}$ if $p_0 = p_{F_0,c_0}$, some F_0, c_0.*
- *nearly $n^{-\beta/(2\beta+1)}$ if p_0 has β derivatives and exponentially small tails.*

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

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

Linear Gaussian inverse problems

Linear Gaussian inverse problems

Data: $X^{(n)} = K\theta + n^{-1/2}\dot{W}$, for white noise $\dot{W}$.

- K compact operator with eigen basis (e_i).
- Prior: $\theta = \sum_{i=1}^{\infty} \theta_i e_i$, with $\theta_i | \, \alpha \overset{\text{ind}}{\sim} N(0, i^{-2\alpha-1})$.

Linear Gaussian inverse problems

Data: $X^{(n)} = K\theta + n^{-1/2}\dot{W}$, for white noise $\dot{W}$.

- K compact operator with eigen basis (e_i).
- Prior: $\theta = \sum_{i=1}^{\infty} \theta_i e_i$, with $\theta_i | \alpha \overset{\text{ind}}{\sim} N(0, i^{-2\alpha-1})$.

Theorem. *If $\sum_{i=1}^{\infty} i^{2\beta}\theta_{i,0}^2 < \infty$ and eigenvalues $\kappa_i \asymp i^{-p}$, then rate:*

- $n^{-\beta/(2\alpha+2p+1)}$, *if $\beta \leq \alpha$,*
- $n^{-\alpha/(2\alpha+2p+1)}$, *if $\beta \geq \alpha$.*

Optimal rate if and only if truth and prior "match".

Linear Gaussian inverse problems — adaptation

Data: $X^{(n)} = K\theta + n^{-1/2}\dot{W}$, for white noise $\dot{W}$.

- K compact operator with eigen basis (e_i).
- Prior: $\theta = \sum_{i=1}^{\infty} \theta_i e_i$, with $\theta_i |\, \alpha \overset{\text{ind}}{\sim} N(0, i^{-2\alpha-1})$.
- Prior on α.

Theorem. *If* $\sum_{i=1}^{\infty} i^{2\beta}\theta_{0,i}^2 < \infty$ *and eigenvalues* $\kappa_i \asymp i^{-p}$*, then rate* $n^{-\beta/(2\beta+2p+1)}$*, any* $\beta > 0$*.*

Example: reconstructing a derivative

Volterra operator $K\colon L_2[0,1] \to L_2[0,1]$

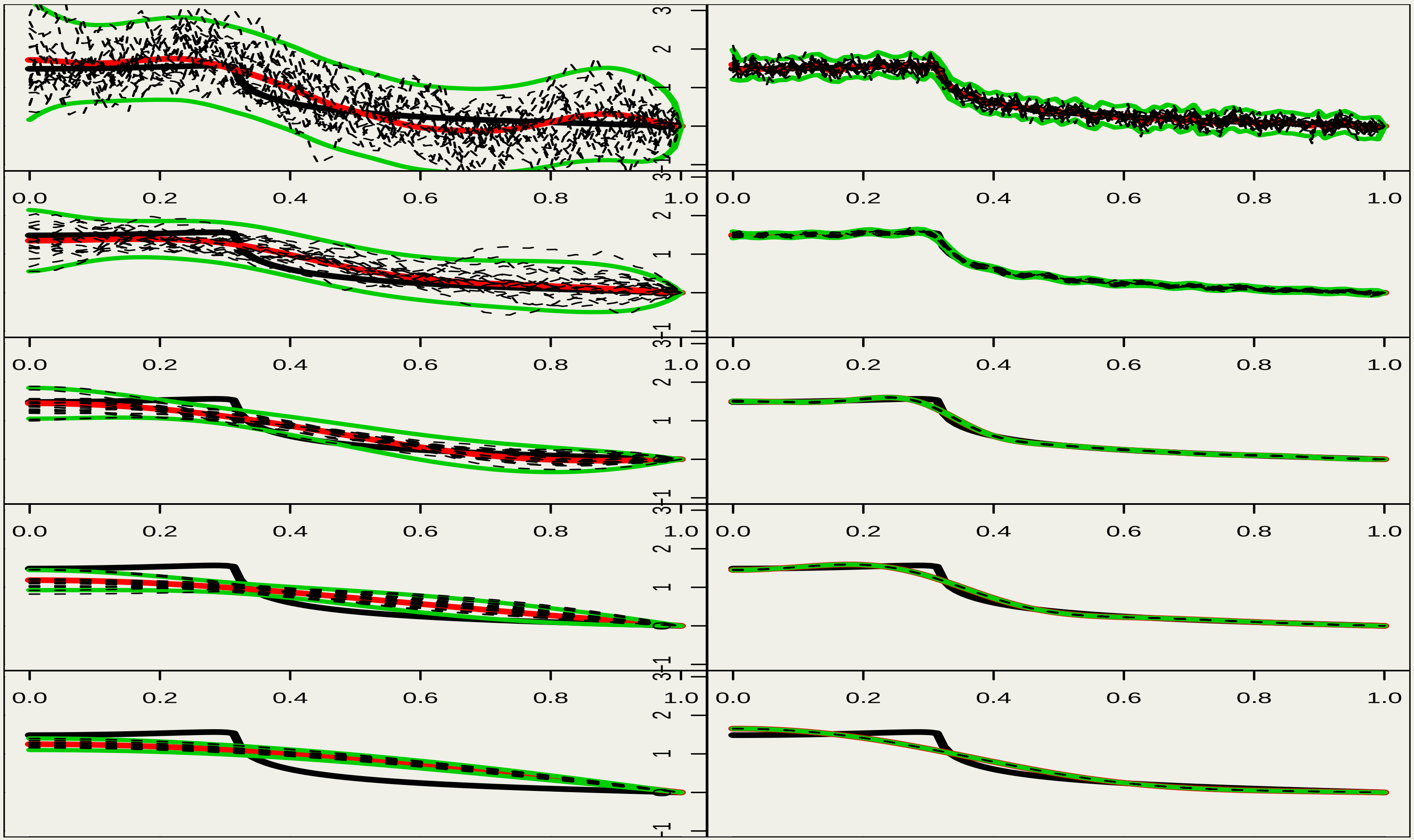
$$K\theta(t) = \int_0^t \theta(s)\, ds.$$

mildly ill-posed inverse problem with eigenvalues and functions:

$$\kappa_i = \frac{1}{(i-1/2)\pi} \qquad e_i(t) = \sqrt{2}\cos\big((i-1/2)\pi t\big),$$

$$(i = 0, 1, 2, \ldots).$$

Example: reconstructing derivative

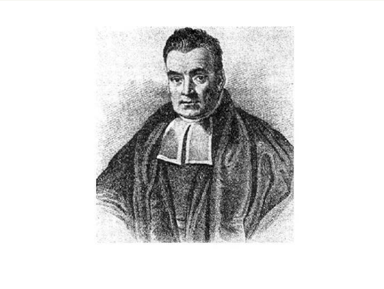


True θ_0 (black), posterior mean (red), and 20 realizations from the posterior,

for $\alpha = 0.5, 1, 2, 3, 5$ (top to bottom) and $n = 1000, 10^8$ (left and right).

Uncertainty quantification

Credible sets



- A parameter Θ is generated according to a prior distribution Π.
- Given θ 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, the posterior distribution:

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

Definition. *A credible set is a data-dependent set $C(X)$ with*

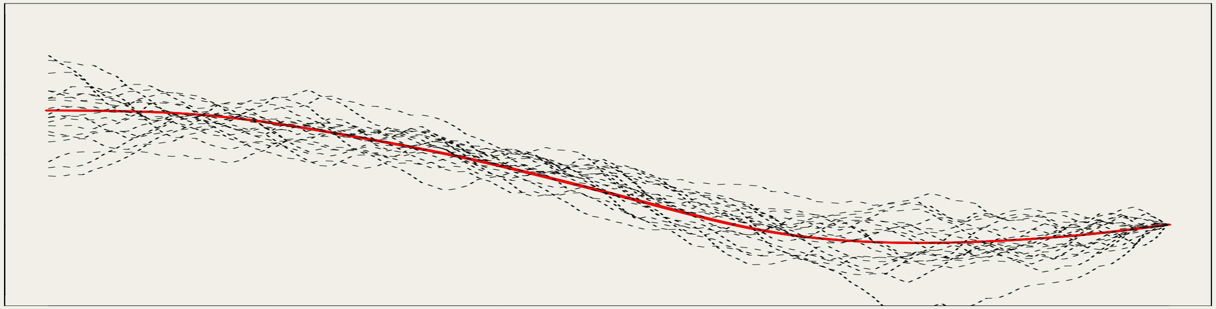
$$\Pi\big(\theta \in C(X) |\, X\big) = 0.95.$$

Nonparametric credible sets

Nonparametric credible sets are sets in function space.
They can take many forms:

- Plots of realizations from the posterior distribution.
- Credible bands.
- Credible balls.

They are routinely produced from MCMC output.



20 realizations from the posterior.

Do credible sets correctly quantify *remaining uncertainty*?

Is a credible set a confidence set?

credible set	confidence set
$\Pi\big(\theta \in C(X) \,\big\vert\, X\big) = 0.95.$	$\mathrm{P}_{\theta_0}\big(\theta_0 \in C_n(X)\big) = 0.95, \ \forall \theta_0.$

Do credible sets correctly quantify *remaining uncertainty*?

Is a credible set a confidence set?

credible set	confidence set
$\Pi\big(\theta \in C(X) \mid X\big) = 0.95.$	$\mathrm{P}_{\theta_0}\big(\theta_0 \in C_n(X)\big) = 0.95, \ \forall \theta_0.$

Rarely!

Only if some version of the Bernstein-von Mises theorem holds.

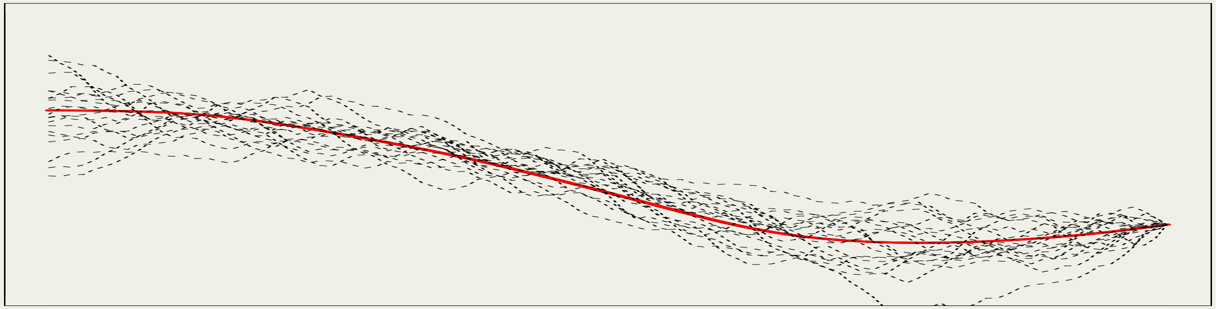
[Cox (1993), Freedman (2000), Leahu (2012), Castillo & Nickl (2013), Ray (2014).]

Do credible sets correctly quantify *remaining uncertainty*?

Is a credible set a confidence set?

credible set	confidence set
$\Pi\big(\theta \in C(X) \mid X\big) = 0.95.$	$\mathrm{P}_{\theta_0}\big(\theta_0 \in C_n(X)\big) = 0.95, \ \forall \theta_0.$

Does the spread in the posterior give the correct order of the discrepancy between θ_0 and the posterior mean?



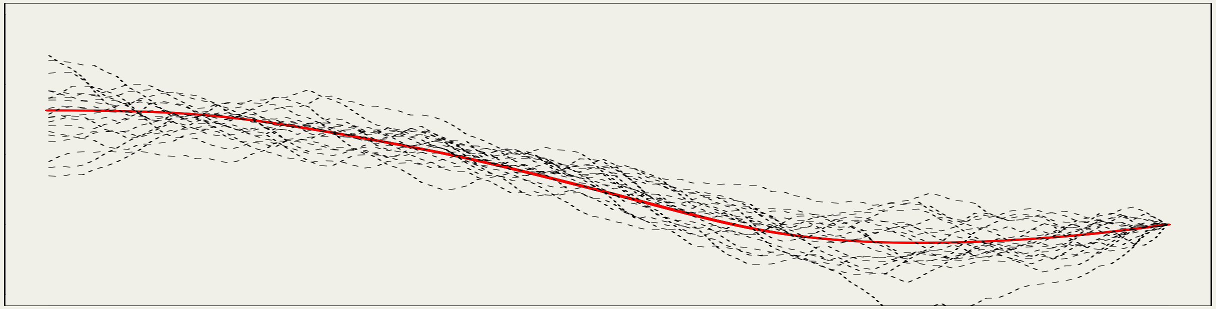
20 realizations from the posterior.

Do credible sets correctly quantify *remaining uncertainty*?

Is a credible set a confidence set?

credible set	confidence set
$\Pi\big(\theta \in C(X) \mid X\big) = 0.95.$	$P_{\theta_0}\big(\theta_0 \in C_n(X)\big) = 0.95, \ \forall \theta_0.$

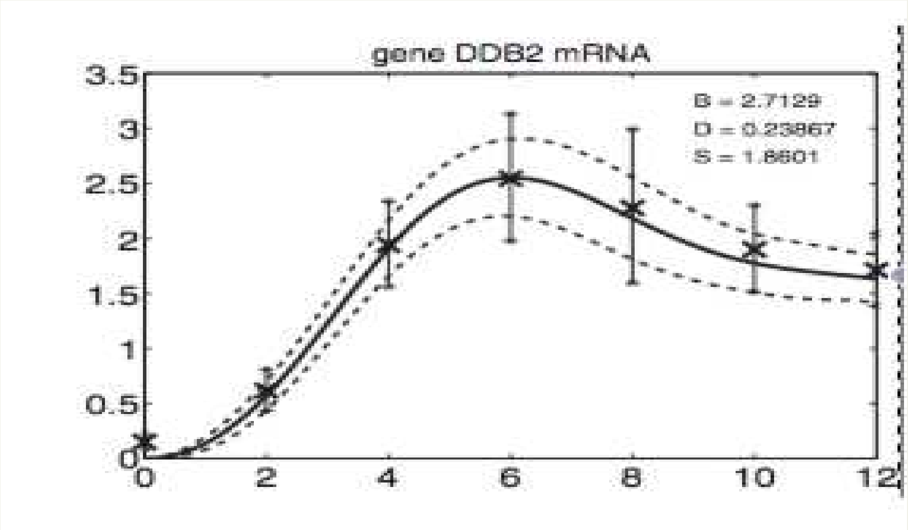
Does the spread in the posterior give the correct order of the discrepancy between θ_0 and the posterior mean?



20 realizations from the posterior.

Is this picture interesting?

Example: genomics



Estimated abundance of a transcription factor as function of time:
posterior mean curve and 95% credible bands.
From Gao et al. *Bioinformatics*, 2008, 70–75.

History

Wahba, 1975

J. R. Statist. Soc. B (1983),
45, *No.* 1, *pp.* 133–150

Bayesian "Confidence Intervals" for the Cross-validated Smoothing Spline

By GRACE WAHBA

University of Wisconsin, USA

[Received August 1981. Revised August 1982]

SUMMARY

We consider the model $Y(t_i) = g(t_i) + \epsilon_i$, $i = 1, 2, \ldots, n$, where $g(t)$, $t \in [0, 1]$ is a smooth function and the $\{\epsilon_i\}$ are independent $N(0, \sigma^2)$ errors with σ^2 unknown. The cross-validated smoothing spline can be used to estimate g non-parametrically from observations on $Y(t_i)$, $i = 1, 2, \ldots, n$, and the purpose of this paper is to study confidence intervals for this estimate. Properties of smoothing splines as Bayes estimates are used to derive confidence intervals based on the posterior covariance function of the estimate. A small Monte Carlo study with the cubic smoothing spline is carried out to suggest by example to what extent the resulting 95 per cent confidence intervals can be expected to cover about 95 per cent of the true (but in practice unknown) values of $g(t_i)$, $i = 1, 2, \ldots, n$. The method was also applied to one example of a two-dimensional thin plate smoothing spline. An asymptotic theoretical argument is presented to explain why the method can be expected to work on fixed smooth functions (like those tried), which are "smoother" than the sample functions from the prior distributions on which the confidence interval theory is based.

Keywords: SPLINE SMOOTHING; CROSS-VALIDATION; CONFIDENCE INTERVALS

1. INTRODUCTION

Consider the model

$$Y(t_i) = g(t_i) + \epsilon_i, \quad i = 1, 2, \ldots, n, \quad t_i \in [0, 1], \tag{1.1}$$

where $\epsilon = (\epsilon_1, \ldots, \epsilon_n)' \sim N(0, \sigma^2 I_{n \times n})$, σ^2 is unknown and $g(\cdot)$ is a fixed but unknown function with $m - 1$ continuous derivatives and $\int_0^1 (g^{(m)}(t))^2 dt < \infty$. The smoothing spline estimate of g given $Y(t_i) = y_i$, $i = 1, 2, \ldots, n$, which we will call $g_{n,\lambda}$, is the minimizer of

$$n^{-1} \sum_{i=1}^{n} (g(t_i) - y_i)^2 + \lambda \int_0^1 (g^{(m)}(t))^2 dt$$

Works great!

Cox, 1993

The Annals of Statistics
1993, Vol. 21, No. 2, 903–923

AN ANALYSIS OF BAYESIAN INFERENCE FOR NONPARAMETRIC REGRESSION[1]

BY DENNIS D. COX

Rice University

The observation model $y_i = \beta(i/n) + \varepsilon_i$, $1 \le i \le n$, is considered, where the ε's are i.i.d. with mean zero and variance σ^2 and β is an unknown smooth function. A Gaussian prior distribution is specified by assuming β is the solution of a high order stochastic differential equation. The estimation error $\delta = \beta - \hat{\beta}$ is analyzed, where $\hat{\beta}$ is the posterior expectation of β. Asymptotic posterior and sampling distributional approximations are given for $\|\delta\|^2$ when $\|\cdot\|$ is one of a family of norms natural to the problem. It is shown that the frequentist coverage probability of a variety of $(1 - \alpha)$ posterior probability regions tends to be larger than $1 - \alpha$, but will be infinitely often less than any $\varepsilon > 0$ as $n \to \infty$ with prior probability 1. A related continuous time signal estimation problem is also studied.

1. Introduction. In this article we consider Bayesian inference for a class of nonparametric regression models. Suppose we observe

$$Y_{ni} = \beta(t_{ni}) + \varepsilon_i, \qquad 1 \le i \le n, \tag{1.1}$$

where $t_{ni} = i/n$, β: $[0, 1] \to \mathbb{R}$ is an unknown smooth function, and $\varepsilon_1, \varepsilon_2, \ldots$ are i.i.d. random errors with mean 0 and known variance $\sigma^2 < \infty$. The ε_i are modeled as $N(0, \sigma^2)$. A Gaussian prior for β will now be specified. Let $m \ge 2$ and for some constants $a_0, \ldots, a_m$ with $a_m \ne 0$ let

$$L = \sum_{i=0}^{m} a_i D^i$$

Fails miserably!

Priors of fixed regularity

Coverage requires undersmoothing

In *nonparametric statistics*:
oversmoothing gives big bias and small variance and hence no coverage.

Coverage requires undersmoothing

In *nonparametric statistics*:
oversmoothing gives big bias and small variance and hence no coverage.

In *nonparametric Bayesian statistics*:
this occurs if the prior produces too smooth functions.

Coverage requires undersmoothing

In *nonparametric statistics*:
oversmoothing gives big bias and small variance and hence no coverage.

In *nonparametric Bayesian statistics*:
this occurs if the prior produces too smooth functions.

EXAMPLE

Truth: $$\theta_0(t) = \sum_{i=1}^{\infty} \theta_{0,i} e_i(t), \qquad \theta_{0,i} \asymp i^{-1-2\beta}.$$

Prior: $$x \mapsto \sum_{i=1}^{\infty} \theta_i e_i(t), \qquad \theta_i \overset{\text{ind}}{\sim} N(0, i^{-1-2\alpha}).$$

Interpretation:
$\alpha = \beta$: prior and truth match.
$\alpha > \beta$: prior oversmoothes.
$\alpha < \beta$: prior undersmoothes.

Example: heat equation

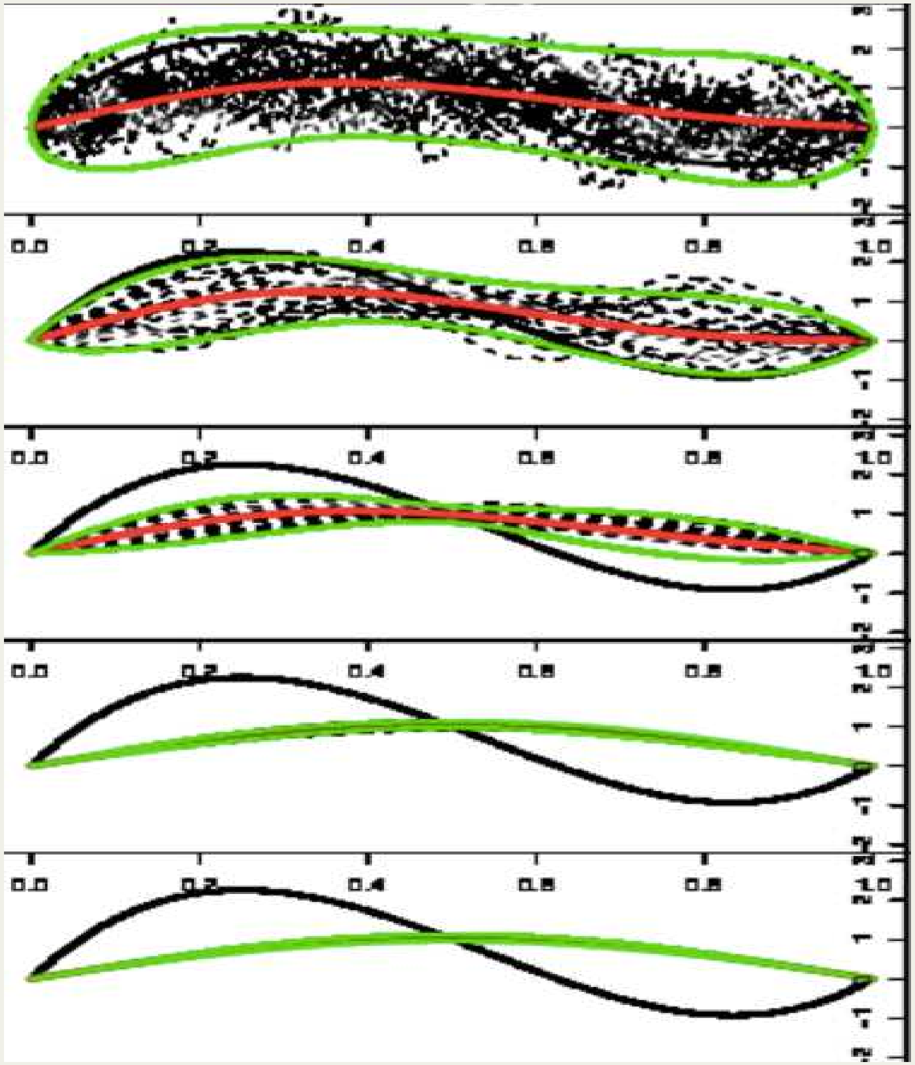
For given initial heat curve $\theta\colon [0,1] \to \mathbb{R}$ let $K\theta = u(\cdot, 1)$ be the final curve:

$$\frac{\partial}{\partial t}u(x,t) = \frac{\partial^2}{\partial x^2}u(x,t), \quad u(\cdot, 0) = \theta, \quad u(0,t) = u(1,t) = 0.$$

Observe noisy version $(X_t^{(n)}\colon 0 \le t \le 1)$ of final curve: for $\dot{W}$ white noise:

$$X^{(n)} = K\theta + n^{-1/2}\dot{W}.$$

Example: heat equation (n=10 000)



True θ_0 (black), posterior mean (red), 20 realizations from the posterior (dashed black), and posterior credible bands (green). Left: $n = 10^4$; right: $n = 10^8$. Top to bottom: prior of increasing smoothness.

[Knapik, vdV and Van Zanten, 2013.]

Priors of flexible regularity

Bayesian adaptation

Family of priors Π_α of varying smoothness; posteriors $\Pi_\alpha(\cdot \mid X)$.

Examples

- $t \mapsto \sum_{i=1}^\infty \theta_i e_i(t)$, for $\theta_i \overset{\text{ind}}{\sim} N(0, i^{-1-2\alpha})$.
- $t \mapsto G_{\alpha t}$, for Gaussian process G.
- $t \mapsto \int \alpha^{-1}\phi(\alpha^{-1}(t-z))\, dF(z)$, with $F \sim$ Dirichlet process.

Bayesian adaptation

Family of priors Π_α of varying smoothness; posteriors $\Pi_\alpha(\cdot|\, X)$.

Hierarchical Bayes:

- Prior on α.
- Ordinary posterior.

Empirical Bayes:

- $\hat\alpha =$ marginal MLE.
- Plug-in posterior $\Pi_{\hat\alpha}(\cdot|\, X)$.

$$\hat\alpha = \operatorname*{argmax}_\alpha \int p(X|\,\theta)\, d\Pi_\alpha(\theta).$$

Both methods give adaptive reconstructions:
if the true function is smoother, then the reconstruction is better.

Bayesian adaptation

Family of priors Π_α of varying smoothness; posteriors $\Pi_\alpha(\cdot|\,X)$.

Hierarchical Bayes:

- Prior on α.
- Ordinary posterior.

Empirical Bayes:

- $\hat\alpha =$ marginal MLE.
- Plug-in posterior $\Pi_{\hat\alpha}(\cdot|\,X)$.

$$\hat\alpha = \underset{\alpha}{\operatorname{argmax}} \int p(X|\,\theta)\, d\Pi_\alpha(\theta).$$

Both methods give adaptive reconstructions:
if the true function is smoother, then the reconstruction is better.

This implies that they cannot give honest confidence sets.

Honesty and impossibility of adaptation [Low, Cai & Low, Lepski, Juditzky et al., Robins&vdV, Bull& Nickl]

Definition. $C_n(X^{(n)})$ *is a (honest) confidence set over a model* Θ *if*

$$\mathrm{P}_{\theta_0}\bigl(C_n(X^{(n)}) \ni \theta_0\bigr) \geq 0.95, \qquad \textit{for all } \theta_0 \in \Theta.$$

Honesty and impossibility of adaptation [Low, Cai & Low, Lepski, Juditzky et al., Robins&vdV, Bull& Nickl]

Definition. $C_n(X^{(n)})$ *is a (honest) confidence set over a model* Θ *if*

$$\mathrm{P}_{\theta_0}\big(C_n(X^{(n)}) \ni \theta_0\big) \geq 0.95, \qquad \text{for all } \theta_0 \in \Theta.$$

Theorem. *For* $\Theta_1 \subset \Theta$ *the diameter of* $C_n(X^{(n)})$ *cannot be smaller, uniformly in* $\theta \in \Theta_1$*, than:*

(a) ε_n *such that, for any* T_n*,*

$$\liminf_{n\to\infty} \sup_{\theta\in\Theta_1} \mathrm{P}_\theta\big(d(T_n,\theta) \geq \varepsilon_n\big) > 0.501.$$

(b) rate ε_n *of minimax testing, for any given* $\Theta_1' \subset \Theta_1$ *of* $H_0\colon \theta \in \Theta_1'$ *versus* $H_1\colon \theta \in \Theta, d(\theta, \Theta_1') > \varepsilon_n$.

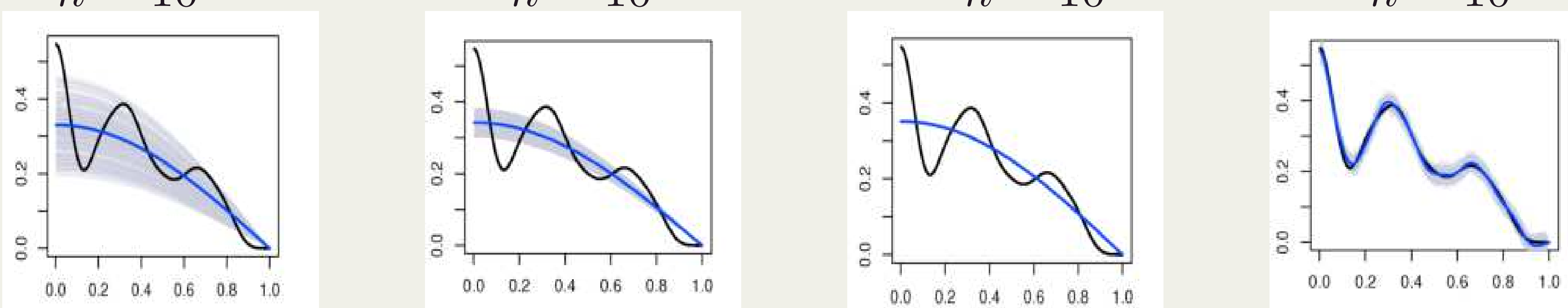
(a) typically gives minimax rate of estimation for model Θ_1.
(b) is determined by biggest model Θ rather than Θ_1.

Credible balls — counter example — reconstructing a derivative

Data: $X^{(n)} = K\theta + n^{-1/2}\dot{W}$, for white noise $\dot{W}$.

- $K\theta(t) = \int_0^t \theta(s)\,ds$, for $0 < t < 1$.
- **Prior**: $\theta = \sum_{i=1}^{\infty} \theta_i e_i$, with $\theta_i |\, \alpha \overset{\text{ind}}{\sim} N(0, i^{-2\alpha-1})$.
- **Prior** on α or empirical Bayes $\hat{\alpha}$.

$n = 10^3$ $\quad$ $n = 10^4$ $\quad$ $n = 10^6$ $\quad$ $n = 10^8$



Gaussian prior in white noise model of smoothness determined by empirical Bayes.

Black: true curve. Blue: posterior mean. Grey: draws from posterior.

The pictures show an *inconvenient truth*.

[Szabo, vdV, van Zanten, 2016.]

Credible balls — counter example — reconstructing a derivative

Theorem. *For $n_j \geq n_{j-1}^4$ for every j, define $\theta = (\theta_1, \theta_2, ...)$ by*

$$\theta_i^2 = \begin{cases} n_j^{-\frac{1+2\beta}{1+2\beta+2p}}, & \textit{if } n_j^{\frac{1}{1+2\beta+2p}} \leq i < 2n_j^{\frac{1}{1+2\beta+2p}}, \qquad j = 1, 2, \ldots, \\ 0, & \textit{otherwise.} \end{cases}$$

Then $\sum_j j^{2\beta}\theta_j^2 \leq 1$, but the central 95%-credible ball $\hat{C}_n$, blown up by $L_n \ll n^\delta$, satisfies

$$\liminf P_\theta\big(\theta \in \hat{C}_n\big) = 0.$$

Credible balls — counter example — reconstructing a derivative

Theorem. *For* $n_j \geq n_{j-1}^4$ *for every* j*, define* $\theta = (\theta_1, \theta_2, ...)$ *by*

$$\theta_i^2 = \begin{cases} n_j^{-\frac{1+2\beta}{1+2\beta+2p}}, & \text{if } n_j^{\frac{1}{1+2\beta+2p}} \leq i < 2n_j^{\frac{1}{1+2\beta+2p}}, \qquad j = 1, 2, \ldots, \\ 0, & \text{otherwise.} \end{cases}$$

Then $\sum_j j^{2\beta}\theta_j^2 \leq 1$*, but the central 95%-credible ball* $\hat{C}_n$*, blown up by* $L_n \ll n^\delta$*, satisfies*

$$\liminf P_\theta\big(\theta \in \hat{C}_n\big) = 0.$$

- Data allows inference only on $\theta_1, \ldots, \theta_{N_n}$.
- Trouble if $\theta_1, \ldots, \theta_{N_n}$ does not resemble $\theta_1, \theta_2, \ldots$.
- Example θ has repeated runs of 0s of increasing lengths.

Estimation versus uncertainty quantification

Adaptive estimation:

- Estimators can be simultaneously optimal for multiple regularities.
- (Bayesian procedures are natural.)

Uncertainty quantification:

- The size of an honest confidence set is determined by the smallest possible regularity level.
- (Bayesian constructions can be misleading.)

Estimation versus uncertainty quantification

Adaptive estimation:

- Estimators can be simultaneously optimal for multiple regularities.
- (Bayesian procedures are natural.)

Uncertainty quantification:

- The size of an honest confidence set is determined by the smallest possible regularity level.
- (Bayesian constructions can be misleading.)

SOLUTION 1: *be honest*; only make conditional confidence statements.

Estimation versus uncertainty quantification

Adaptive estimation:

- Estimators can be simultaneously optimal for multiple regularities.
- (Bayesian procedures are natural.)

Uncertainty quantification:

- The size of an honest confidence set is determined by the smallest possible regularity level.
- (Bayesian constructions can be misleading.)

SOLUTION 1: *be honest*; only make conditional confidence statements.
SOLUTION 2: determine which θ cause the trouble; argue that these are implausible.

Polished tail sequences

Definition. $\theta \in \ell^2$ *satisfies the* polished tail condition *if*

$$\sum_{i=N}^{1000N} \theta_i^2 \geq 0.001 \sum_{i=N}^{\infty} \theta_i^2, \qquad \forall \text{ large } N.$$

Interpretation:
every block of frequencies $(N, 1000N)$
contains a fraction of the total energy above frequency N.

Polished tail sequences

Definition. $\theta \in \ell^2$ *satisfies the* polished tail condition *if*

$$\sum_{i=N}^{1000N} \theta_i^2 \geq 0.001 \sum_{i=N}^{\infty} \theta_i^2, \qquad \forall \textit{ large } N.$$

"Everything" is polished tail...:

Polished tail sequences

Definition. $\theta \in \ell^2$ *satisfies the* polished tail condition *if*

$$\sum_{i=N}^{1000N} \theta_i^2 \geq 0.001 \sum_{i=N}^{\infty} \theta_i^2, \qquad \forall \textit{ large } N.$$

"Everything" is polished tail...:

- For the *topologist* [Giné+Nickl, 2010]:
  Non polished tail sequences are meagre in a natural topology.

Polished tail sequences

Definition. *$\theta \in \ell^2$ satisfies the* polished tail condition *if*

$$\sum_{i=N}^{1000N} \theta_i^2 \geq 0.001 \sum_{i=N}^{\infty} \theta_i^2, \qquad \forall \text{ large } N.$$

“Everything” is polished tail...:

- For the *topologist* [Giné+Nickl, 2010]:
  Non polished tail sequences are meagre in a natural topology.
- For the *minimax expert*:
  Intersecting the usual models with polished tail sequences decreases the minimax risk by at most a logarithmic factor.

Polished tail sequences

Definition. $\theta \in \ell^2$ *satisfies the* polished tail condition *if*

$$\sum_{i=N}^{1000N} \theta_i^2 \geq 0.001 \sum_{i=N}^{\infty} \theta_i^2, \qquad \forall \textit{ large } N.$$

"Everything" is polished tail...:

- For the *topologist* [Giné+Nickl, 2010]:
  Non polished tail sequences are meagre in a natural topology.
- For the *minimax expert*:
  Intersecting the usual models with polished tail sequences decreases the minimax risk by at most a logarithmic factor.
- For the *Bayesian*:
  Almost every parameter generated from a prior $\theta_i \overset{\text{ind}}{\sim} N(0, ci^{-\alpha-1/2})$ is polished tail.

Linear Gaussian inverse problems

Data: $X^{(n)} = K\theta + n^{-1/2}\dot{W}$, for white noise $\dot{W}$.

- K compact operator with eigenvalues $\kappa_i \asymp i^{-p}$ and eigen basis (e_i).
- Prior: $\theta = \sum_{i=1}^{\infty} \theta_i e_i$, with $\theta_i \,|\, \alpha \overset{\text{ind}}{\sim} N(0, i^{-2\alpha-1})$.
- Prior on α.

Linear Gaussian inverse problems

Data: $X^{(n)} = K\theta + n^{-1/2}\dot{W}$, for white noise $\dot{W}$.

- K compact operator with eigenvalues $\kappa_i \asymp i^{-p}$ and eigen basis (e_i).
- Prior: $\theta = \sum_{i=1}^\infty \theta_i e_i$, with $\theta_i |\, \alpha \stackrel{\text{ind}}{\sim} N(0, i^{-2\alpha-1})$.
- Prior on α.

Credible ball:

$$\hat{C}_n(M) := \{\theta : \|\theta - \hat{\theta}_n\| < Mr\}$$

$$\hat{\theta}_n = \mathrm{E}(\theta |\, X^{(n)})$$
$$\Pi\big(\theta : \|\theta - \hat{\theta}_n\| < r |\, X^{(n)}\big) = 0.95$$

Theorem. *For not too small M, uniformly in polished tail functions θ,*

$$\mathrm{P}_\theta\big(\theta \in \hat{C}_n(M)\big) \to 1.$$

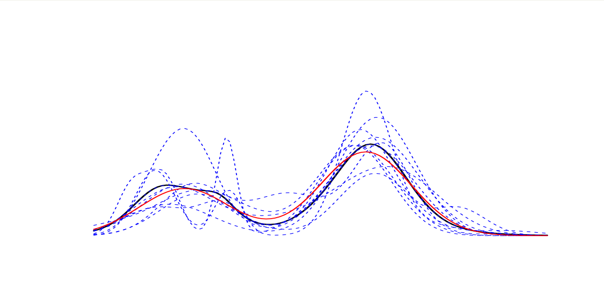
Similar results for empirical Bayes.

Closing remarks

Work in progress

Story on uncertainty quantification appears to be generic,
but conditions for good behaviour depend on prior and model.

There is further work [e.g. by Szabó et al.], but much is unknown.



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

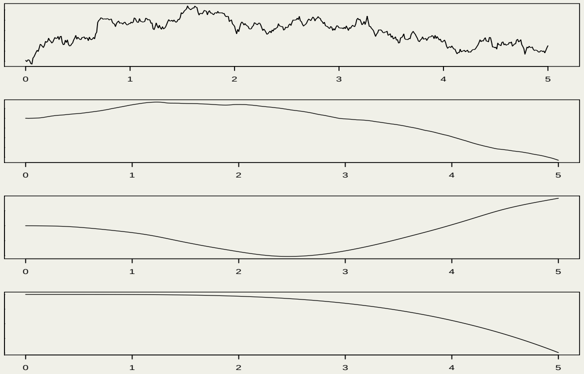
for a sample of size 50 from a mixture of two normals (red).

Summary

In nonparametric statistics uncertainty quantification is problematic for both Bayesian and non-Bayesian methods.

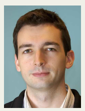
It necessarily extrapolates into features of the world that cannot be seen in the data.

Bayesians are perhaps more easily misled as they trust their priors. In nonparametrics they should not, as the fine details of a prior are not obvious.



Co-authors

 Subhashis Ghoshal

 Ismael Castillo

 Stéphanie van der Pas

 Willem Kruijer

 Bartek Knapik

 Suzanne Sniekers

 Gwenael Leday

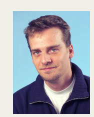
 Mark van de Wiel

 Harry van Zanten

 Johannes Schmidt-Hieber

 Botond Szabo

 Judith Rousseau

 Bas Kleijn

 Fengnan Gao

 Gino Kpogbezan

 Wessel van Wieringen