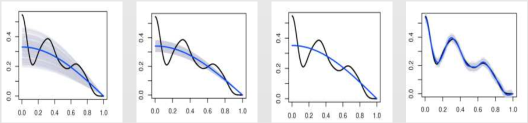
# Confidence in Nonparametric Bayes?

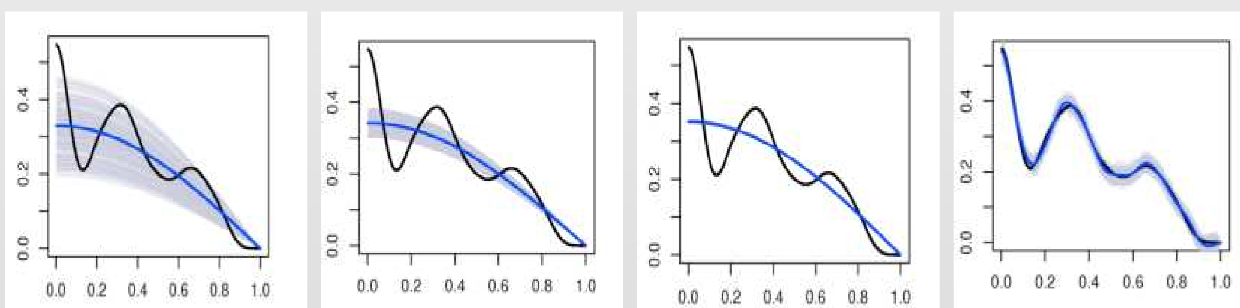
Aad van der Vaart
Universiteit Leiden, Netherlands



Bayes Lectures, Edinburgh, August 2012

# Contents





0.0 0.2 0.4
0.0 0.2 0.4 0.6 0.8 1.0
0.0 0.2 0.4
0.0 0.2 0.4 0.6 0.8 1.0
0.0 0.2 0.4
0.0 0.2 0.4 0.6 0.8 1.0
0.0 0.2 0.4
0.0 0.2 0.4 0.6 0.8 1.0

## Co-authors

Bartek Knapik
(Amsterdam)

Suzanne Sniekers
(Leiden/Amsterdam)

Botond Szabo
(Eindhoven)

Harry van Zanten
(Eindhoven)

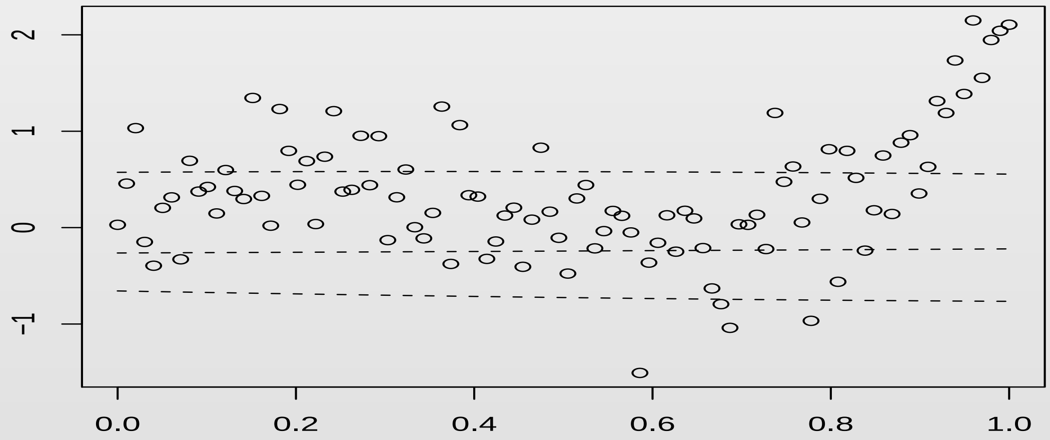
## Disclaimer

*For the sake of presentation some of the contents were edited to fit the talk (without asking my co-authors)*

# 1. Nonparametric Bayes

## Bayesian nonparametric inference

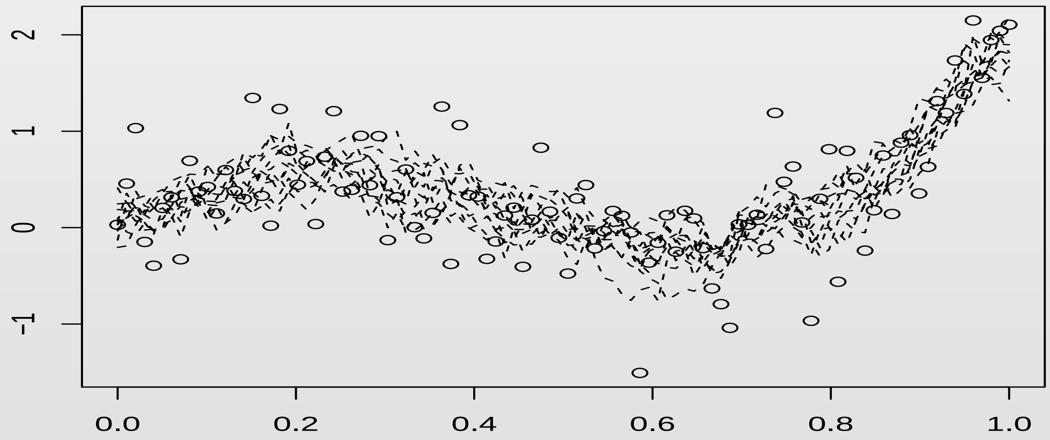
We model a function or surface by a prior on a function space. We visualize this by some draws.



## Bayesian nonparametric inference

We model a function or surface by a prior on a function space. We visualize this by some draws.

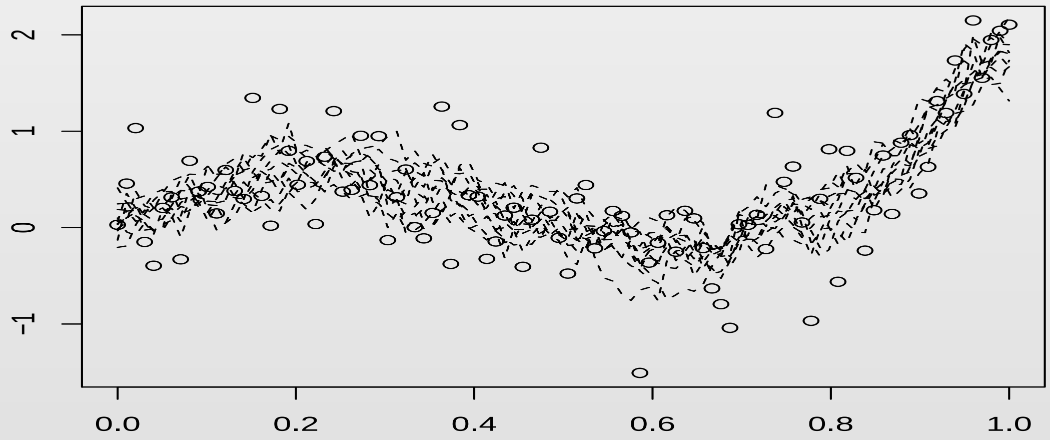
By the usual Bayesian machine we combine this with the likelihood to produce a posterior distribution for the function given the data. We visualize this by some draws.



## Bayesian nonparametric inference

We model a function or surface by a prior on a function space. We visualize this by some draws.

By the usual Bayesian machine we combine this with the likelihood to produce a posterior distribution for the function given the data. We visualize this by some draws.



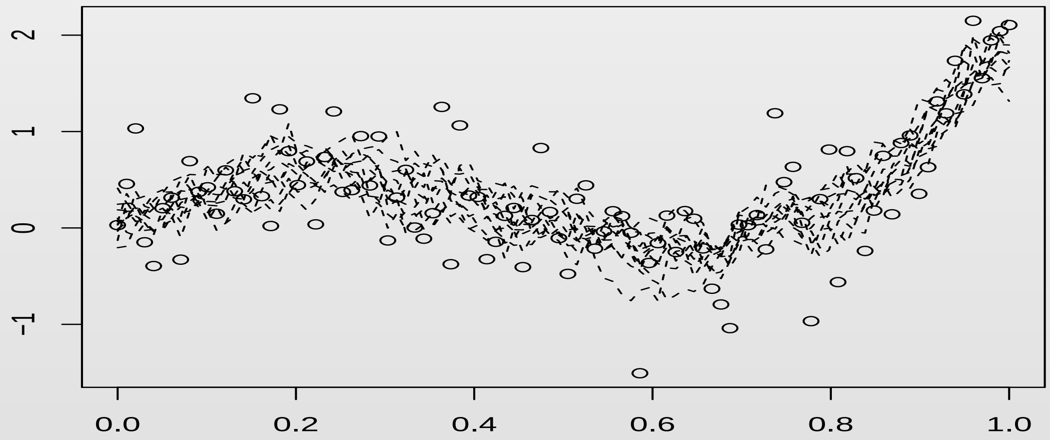

*Does this give good reconstructions?*

*Does the posterior distribution give a correct sense of remaining uncertainty?*

## Gaussian priors

We model a function or surface a-priori as the sample path of a Gaussian process. We visualize this by some draws.

By the usual Bayesian machine we combine this with the likelihood to produce a posterior distribution for the function given the data. We visualize this by some draws.




*Does this give good reconstructions?*

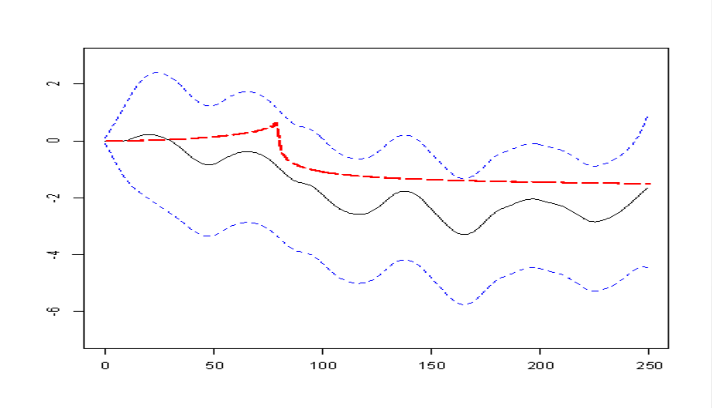
*Does the posterior distribution give a correct sense of remaining uncertainty?*

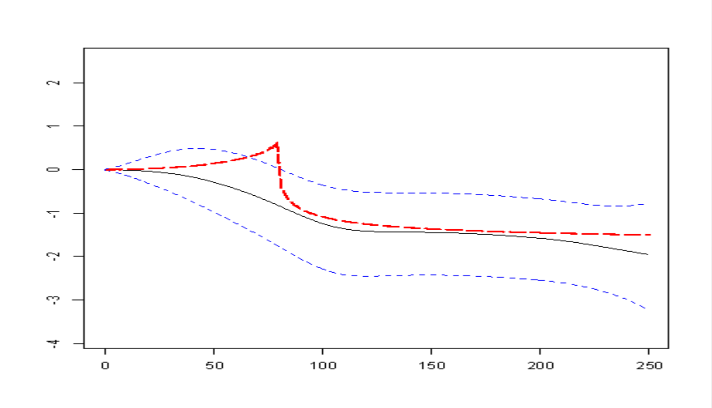
## Example: Logistic regression

Bayesian model:

$$\begin{cases} \theta \sim \text{ scaled integrated Brownian motion,} \\ (X_1,Y_1),\ldots,(X_n,Y_n)|\,\theta \sim \text{ i.i.d.: } \mathrm{P}(Y_i=1|\,X_i=x)=1/(1+e^{-\theta(x)}). \end{cases}$$

The posterior distribution is the law of $\theta$ given $(X_1,Y_1),\ldots,(X_n,Y_n)$.




Simulation experiment ($n = 250$). Two realisations of the posterior mode (black, solid) and 95 % posterior credible bands (blue, dotted), overlaid with true curve $\theta_0$ (red, dashed). Two different scalings of IBM. Computations by the INLA package.

## Example: heat equation

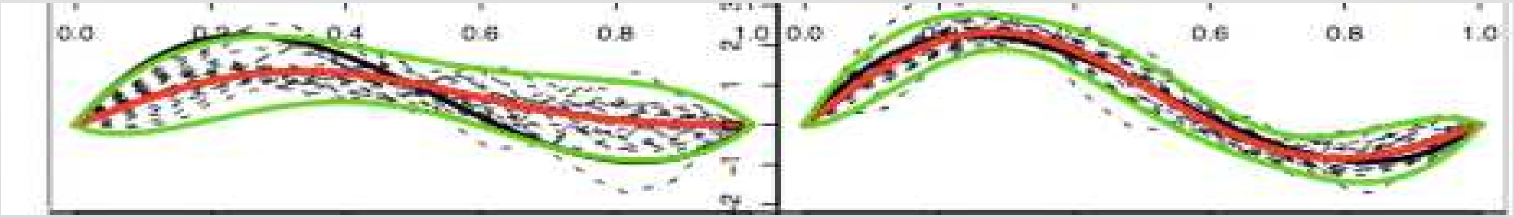
For given $\theta$: $[0,1] \to \mathbb{R}$ let $K\theta = u(\cdot, 1)$ for $u$: $[0,1] \times [0,1] \to \mathbb{R}$ solving

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

Bayesian model: for $(e_i)$ eigenbasis of $K^T K$.

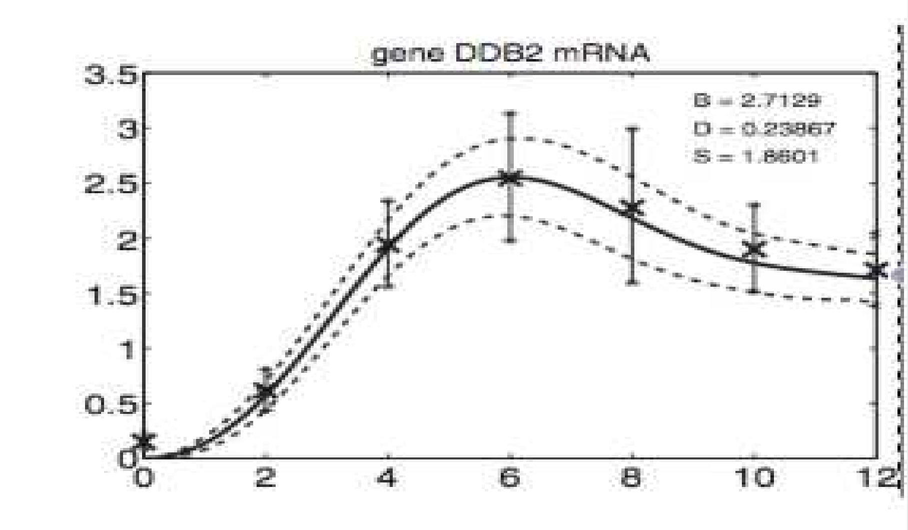
$$\begin{cases} \theta = \sum_i \theta_i e_i, \quad \theta_i \sim N(0, \tau^2 i^{-\alpha-1/2}) \text{ and independent,} \\ Z \sim \text{Gaussian white noise, independent of } \theta, \\ \text{data } Y = K\theta + n^{-1/2} Z. \end{cases}$$

The posterior distribution is the law of $\theta = \sum_i \theta_i e_i$ given $Y$.



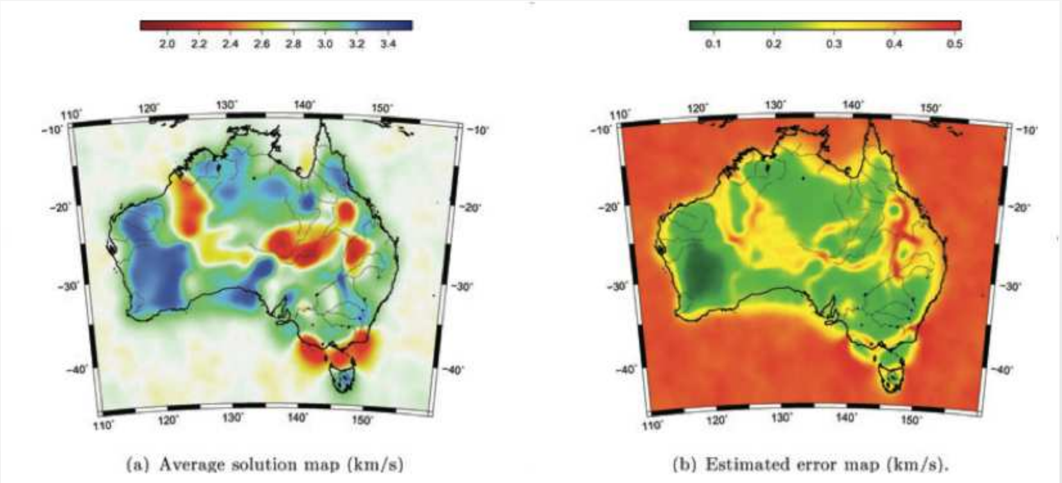

True $\theta_0$ (black), posterior mean (red), 20 realizations from the posterior (dashed black), and posterior credible bands (green). Left: $n = 10^4$ and right: $n = 10^8$.

# Example: genomics



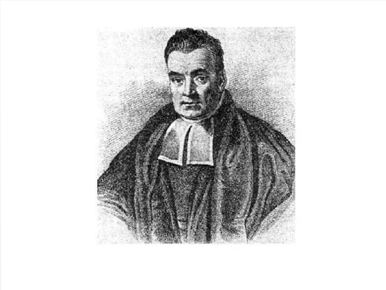

Nonparametric Bayesian analysis in *genomics*. Estimated abundance of a transcription factor as function of time: posterior mean curve and 95% credible bands. From Gao et al. *Bioinformatics*, 2008, 70–75.

# Example: earth science




Travel times of surfaces waves: nonparametric Bayesian analysis in *earth science*. Left: posterior mean (a two-dimensional surface shown by colour coding); right: uncertainty quantification by the posterior spread. From Bodin and Sambridge, *Geophs. J. Int.* 178, 2009, 1411–1436.

## Notation: the Bayesian machine



Given a prior model $\theta \sim \Pi$ and a likelihood $Y|\,\theta \sim p(y|\,\theta)$, the posterior distribution $\theta|\,Y$ is given by

$$d\Pi(\theta|\,Y) \propto p(Y|\,\theta)\,d\Pi(\theta).$$

Two uses:

- recovery, e.g. by mode, or mean.
- expression of uncertainty, e.g. by a credible set: a set $C(Y)$ with

$$\Pi\big(C(Y)|\,Y\big) = 0.95.$$

## Notation: the Bayesian machine — asymptotics in $n$

Given a prior model $\theta \sim \Pi_n$ and a likelihood $Y_n |\, \theta \sim p_n(y |\, \theta)$, the posterior distribution $\theta |\, Y_n$ is given by

$$d\Pi_n(\theta |\, Y_n) \propto p_n(Y_n |\, \theta)\, d\Pi_n(\theta).$$

Two uses:

- recovery, e.g. by mode, or mean.
- expression of uncertainty, e.g. by a credible set: a set $C_n(Y_n)$ with

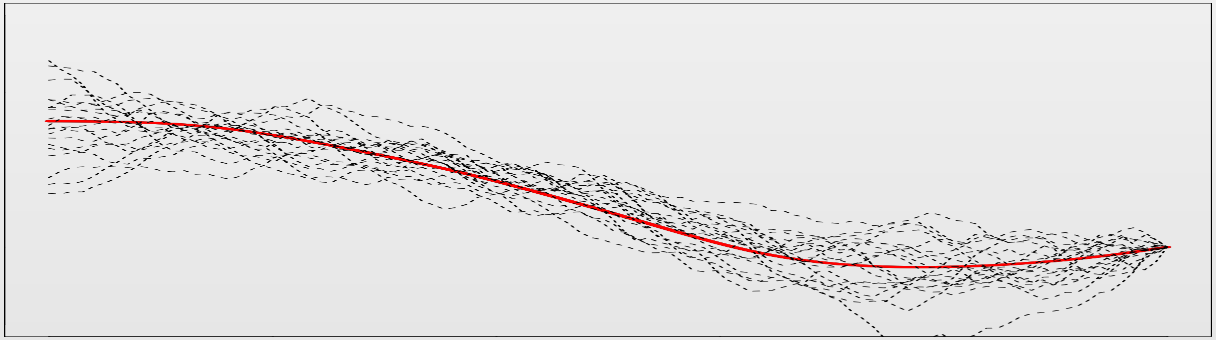
$$\Pi_n\big(C_n(Y_n) |\, Y_n\big) = 0.95.$$

## Frequentist Bayes

Assume that data $Y_n$ is generated according to $\theta_0$ ('truth').

The rate of contraction is (at least) $\varepsilon_n = \varepsilon_n(\theta_0)$ if

$$\mathrm{E}_{\theta_0}\Pi_n\big(d(\theta,\theta_0) > \varepsilon_n \big| Y_n\big) \to 0.$$

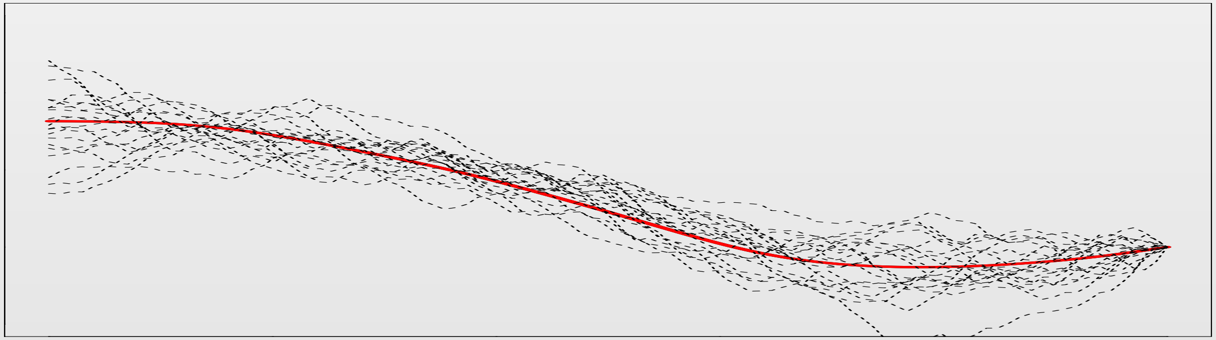


## Frequentist Bayes

Assume that data $Y_n$ is generated according to $\theta_0$ ('truth').

The rate of contraction is (at least) $\varepsilon_n = \varepsilon_n(\theta_0)$ if

$$\mathrm{E}_{\theta_0}\Pi_n\big(d(\theta,\theta_0) > \varepsilon_n \big| Y_n\big) \to 0.$$



The coverage of the credible region $C_n(Y_n)$ is

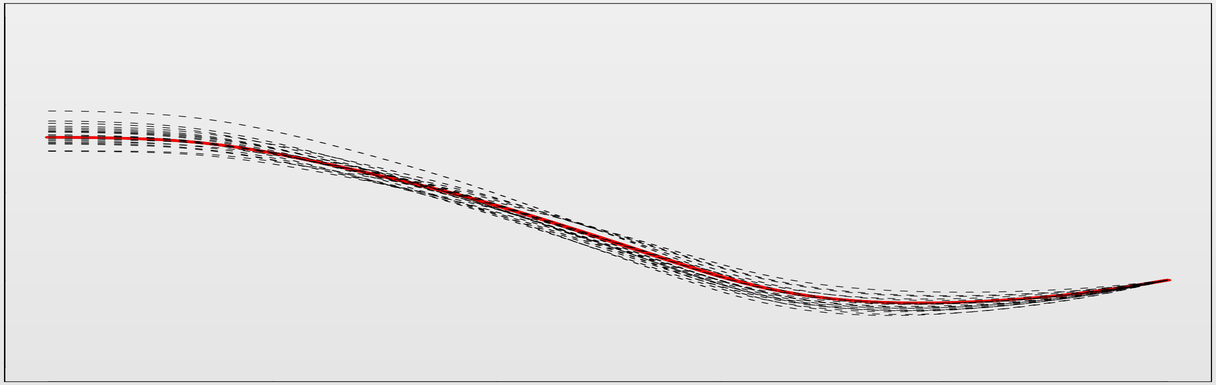
$$\mathrm{P}_{\theta_0}\big(C_n(Y_n) \ni \theta_0\big).$$

## Frequentist Bayes

Assume that data $Y_n$ is generated according to $\theta_0$ ('truth').

The rate of contraction is (at least) $\varepsilon_n = \varepsilon_n(\theta_0)$ if

$$\mathrm{E}_{\theta_0}\Pi_n\big(d(\theta,\theta_0) > \varepsilon_n \,|\, Y_n\big) \to 0.$$



The coverage of the credible region $C_n(Y_n)$ is

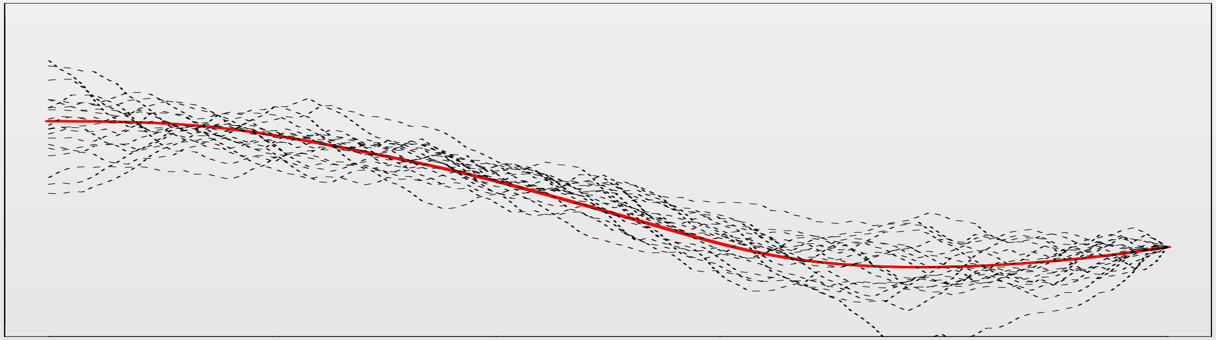
$$\mathrm{P}_{\theta_0}\big(C_n(Y_n) \ni \theta_0\big).$$

## Frequentist Bayes

Assume that data $Y_n$ is generated according to $\theta_0$ ('truth').

The rate of contraction is (at least) $\varepsilon_n = \varepsilon_n(\theta_0)$ if

$$\mathrm{E}_{\theta_0} \Pi_n \big( d(\theta, \theta_0) > \varepsilon_n \,|\, Y_n \big) \to 0.$$



The coverage of the credible region $C_n(Y_n)$ is

$$\mathrm{P}_{\theta_0} \big( C_n(Y_n) \ni \theta_0 \big).$$

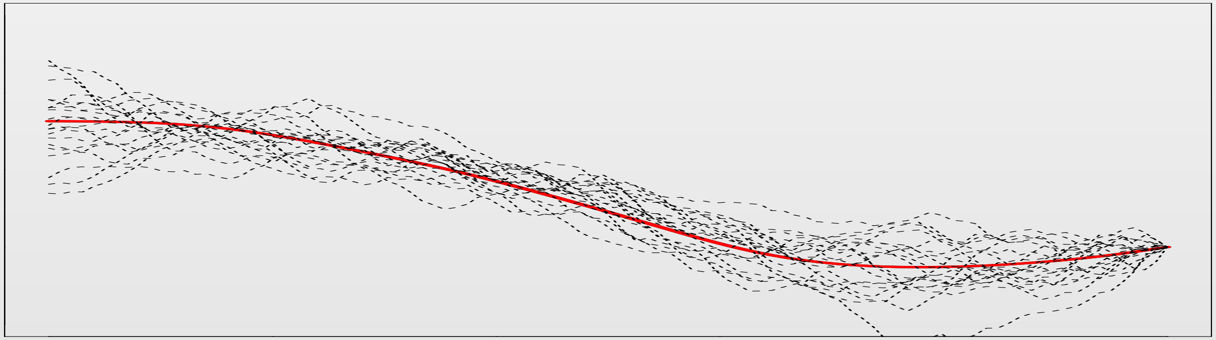
*Does it tend to 95 %?*

## Frequentist Bayes

Assume that data $Y_n$ is generated according to $\theta_0$ ('truth').

The rate of contraction is (at least) $\varepsilon_n = \varepsilon_n(\theta_0)$ if

$$\mathrm{E}_{\theta_0}\Pi_n\big(d(\theta,\theta_0) > \varepsilon_n \,|\, Y_n\big) \to 0.$$



The coverage of the credible region $C_n(Y_n)$ is

$$\mathrm{P}_{\theta_0}\big(C_n(Y_n) \ni \theta_0\big).$$

*Does it tend to 95 %?*

*Does at least the posterior spread express remaining uncertainty?*

## What do the frequentists say? — rates

Nonparametric theory is often concerned with smooth functions.

A typical nonparametric rate of estimation has the form

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

This is the optimal rate for the root mean square error of an estimator of a function $\theta_0\colon [0,1]^d \to \mathbb{R}$ that is known to be $\beta$ times differentiable:

$$\inf_T \sup_{\theta_0 \in C_1^\beta} \mathrm{E}_{\theta_0} d^2\big(T(Y_n), \theta_0\big) = O\big(n^{-2\beta/(2\beta+d)}\big).$$

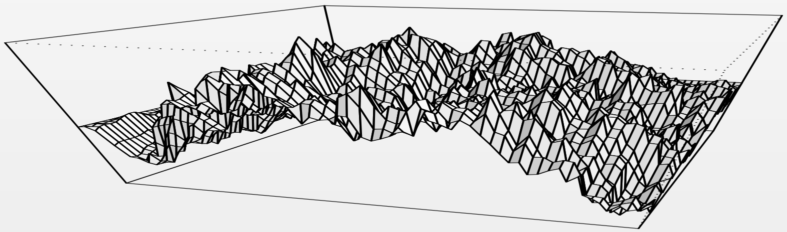
As $\beta \uparrow \infty$ the rate improves, to $n^{-1/2}$ at $\beta = \infty$.

[Adaptive estimators can attain this rate for any $\beta$ without knowing it.]

# 2. Gaussian Process Priors

## Gaussian process

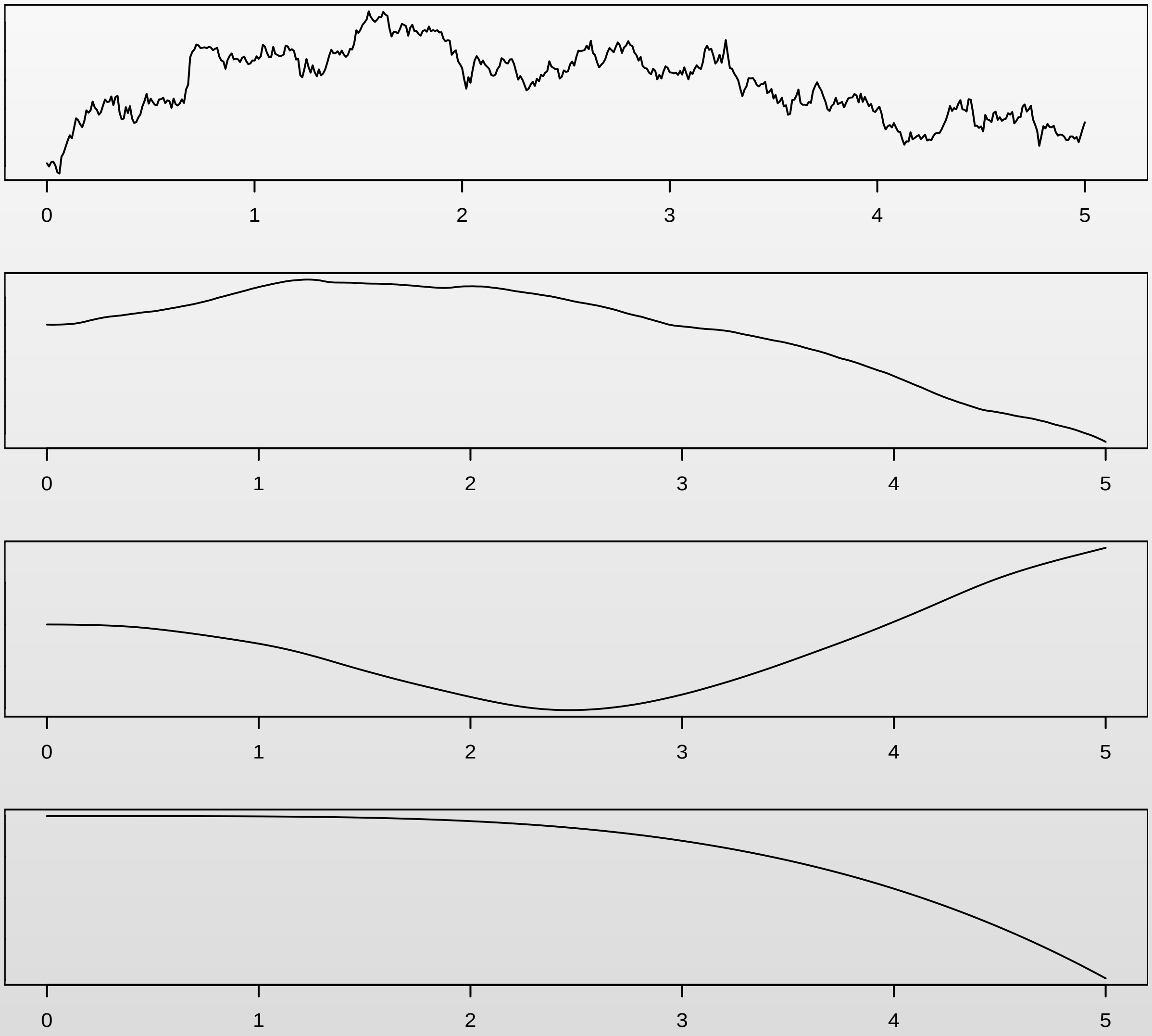
The law of a stochastic process $W = (W_t\colon t \in T)$ is a prior distribution on the space of functions $\theta\colon T \to \mathbb{R}$.



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

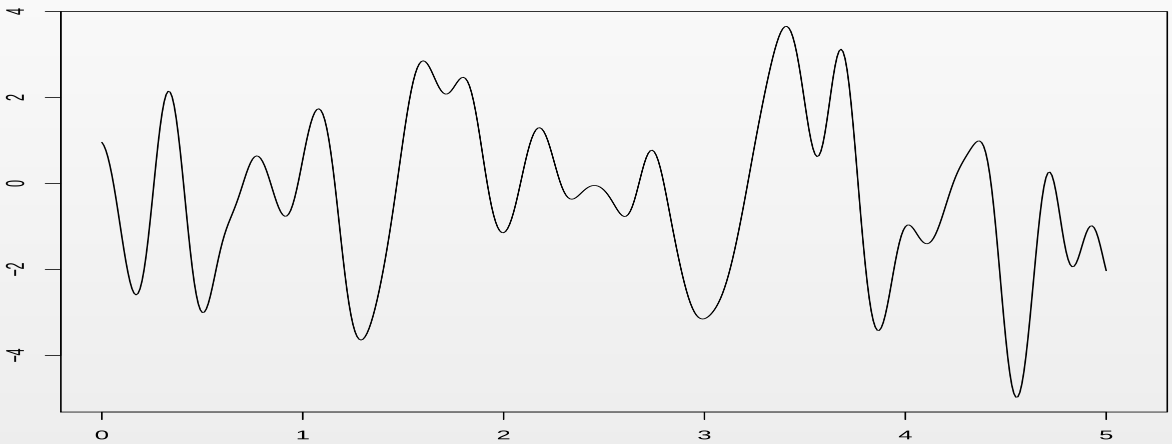
- Every Gaussian prior is reasonable in some way.
- Tuning by (random) hyperparameter is often desirable.

# Integrated Brownian motion

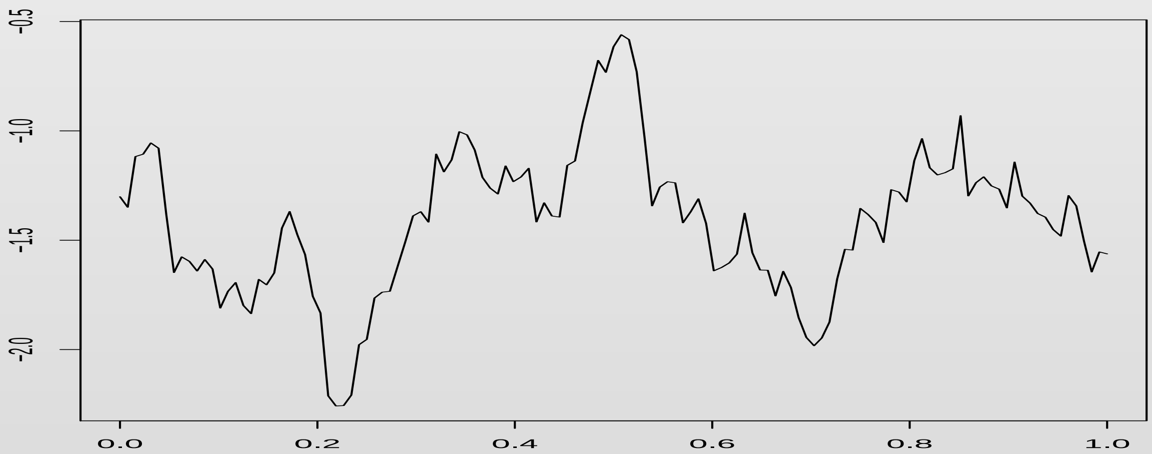



0, 1, 2 and 3 times integrated Brownian motion

## Stationary processes

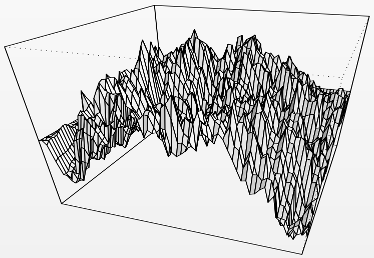



Gaussian spectral measure

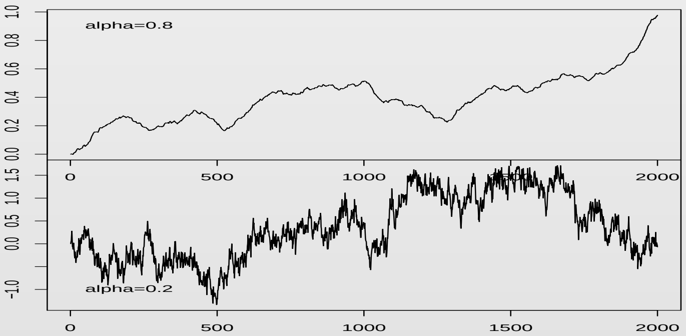



Matérn spectral measure (3/2)

## Other Gaussian processes



Brownian sheet




Fractional Brownian motion

$$\theta(x) = \sum_i \theta_i e_i(x), \quad \theta_i \sim_{indep} N(0, \lambda_i)$$

Series prior

## Posterior contraction rates for Gaussian priors

Prior $W$ is centered Gaussian map in Banach space $(\mathbb{B}, \|\cdot\|)$.
$\theta_0 \in \mathbb{B}$ true parameter.

THEOREM
If statistical distances on the model combine appropriately with the norm $\|\cdot\|$ of $\mathbb{B}$, then the posterior rate of contraction is $\varepsilon_n$ if

$$\Pi(\|W - \theta_0\| < \varepsilon_n) \geq e^{-n\varepsilon_n{}^2}.$$

$\theta_0$-centered small ball probability determines the rate.

## Posterior contraction rates for Gaussian priors

Prior $W$ is centered Gaussian map in Banach space $(\mathbb{B}, \|\cdot\|)$.
$\theta_0 \in \mathbb{B}$ true parameter.

THEOREM
If statistical distances on the model combine appropriately with the norm $\|\cdot\|$ of $\mathbb{B}$, then the posterior rate of contraction is $\varepsilon_n$ if

$$\Pi(\|W - \theta_0\| < \varepsilon_n) \geq e^{-n\varepsilon_n{}^2}.$$

$\theta_0$-centered small ball probability determines the rate.

# 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 $p_\theta$.
- 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 $\mathrm{P}_\theta$. ($G$ 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 $\theta$.
- Norm on $W$: empirical $L_2$-distance.

Ergodic diffusions
$(X_t: t \in [0,n])$, 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 $W$: $L_2(\mu_0)$.
($\mu_0$ stationary measure.)

## Settings (2)

For inverse problems the rate equation is different.

(Only special cases understood.)

## Posterior contraction rates for Gaussian priors (2)

Prior $W$ is centered Gaussian map in Banach space $(\mathbb{B}, \|\cdot\|)$ with RKHS $(\mathbb{H}, \|\cdot\|_{\mathbb{H}})$ and small ball exponent

$$\phi_0(\varepsilon) = -\log \Pi(\|W\| < \varepsilon).$$

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

$$\phi_0(\varepsilon_n) \leq n{\varepsilon_n}^2 \qquad \text{AND} \qquad \inf_{h \in \mathbb{H}: \|h-\theta_0\| < \varepsilon_n} \|h\|_{\mathbb{H}}^2 \leq n{\varepsilon_n}^2.$$

## Posterior contraction rates for Gaussian priors (2)

Prior $W$ is centered Gaussian map in Banach space $(\mathbb{B}, \|\cdot\|)$ with RKHS $(\mathbb{H}, \|\cdot\|_{\mathbb{H}})$ and small ball exponent

$$\phi_0(\varepsilon) = -\log \Pi(\|W\| < \varepsilon).$$

THEOREM

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

$$\phi_0(\varepsilon_n) \leq n{\varepsilon_n}^2 \qquad \text{AND} \qquad \inf_{h \in \mathbb{H}: \|h - \theta_0\| < \varepsilon_n} \|h\|_{\mathbb{H}}^2 \leq n{\varepsilon_n}^2.$$

Both inequalities give lower bound on $\varepsilon_n$.
The first depends on $W$ and not on $\theta_0$.
If $\theta_0 \in \mathbb{H}$, then second inequality is satisfied for $\varepsilon_n \gtrsim 1/\sqrt{n}$.

## Brownian Motion

THEOREM

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

$\boxed{\text{The rate is minimax iff } \beta = 1/2.}$

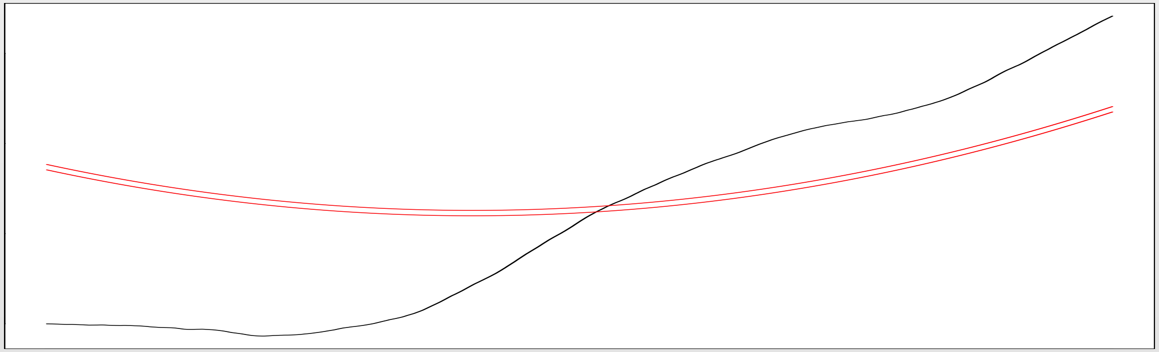
The small ball exponent of Brownian motion is $\phi_0(\varepsilon) \asymp (1/\varepsilon)^2$ as $\varepsilon \downarrow 0$. This gives the $n^{-1/4}$-rate, even for very smooth truths.

## Integrated Brownian Motion

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 minimax iff $\beta = \alpha$.



## Stationary processes

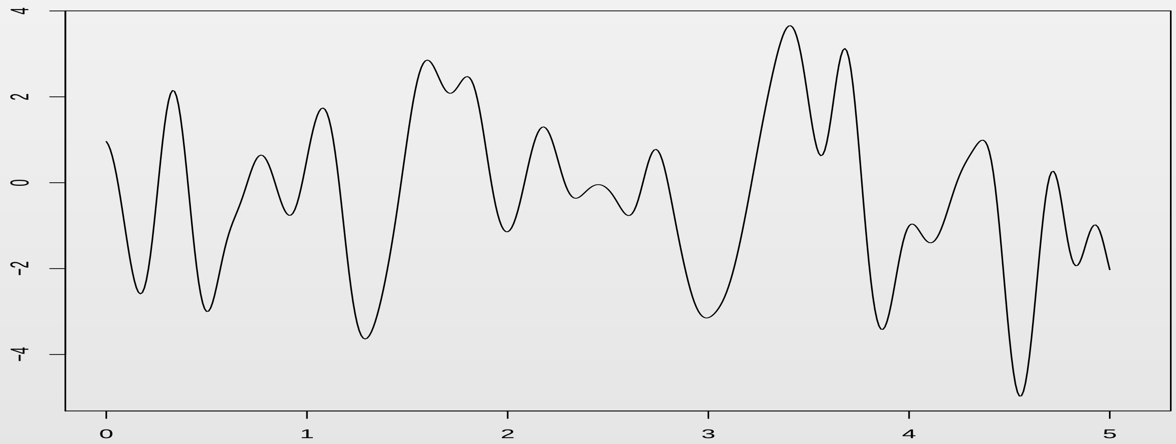
A stationary Gaussian field $(W_t : t \in \mathbb{R}^d)$ is characterized through a spectral measure $\mu$, by

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

## Stationary processes — radial basis

Stationary Gaussian field $(W_t: t \in \mathbb{R}^d)$ characterized through

$$\mathrm{cov}(W_s, W_t) = \int e^{i\lambda^T(s-t)}\, e^{-\lambda^2}\, d\lambda.$$




THEOREM

Let $\hat{\theta}_0$ be the Fourier transform of the true parameter $\theta_0: [0,1]^d \to \mathbb{R}$.

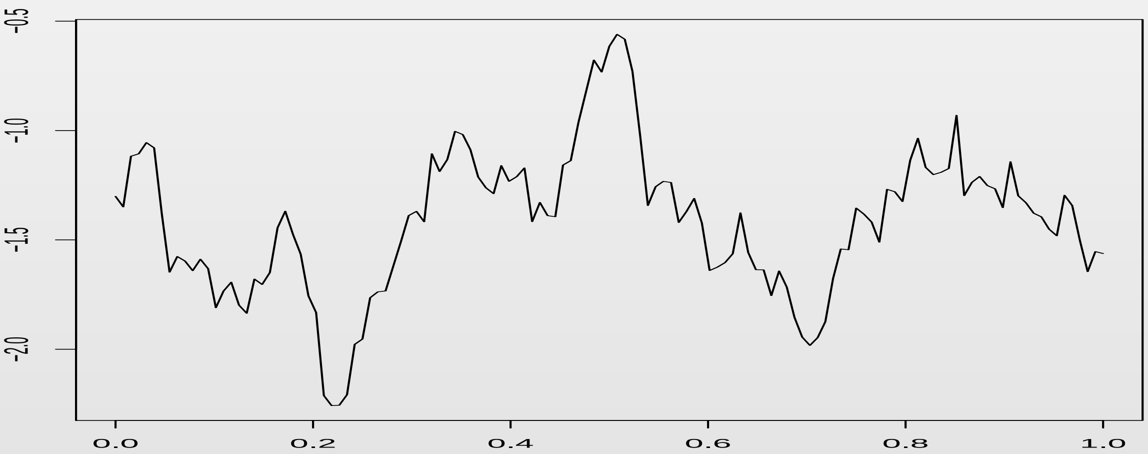
- If $\int e^{\|\lambda\|} |\hat{\theta}_0(\lambda)|^2\, d\lambda < \infty$, then rate of contraction is near $1/\sqrt{n}$.
- If $|\hat{\theta}_0(\lambda)| \gtrsim (1 + \|\lambda\|^2)^{-\beta}$, then rate is power of $1/\log n$.

Excellent if truth is supersmooth; disastrous otherwise.

## Stationary processes — Matérn

Stationary Gaussian field $(W_t\colon t \in \mathbb{R}^d)$ characterized through

$$\mathrm{cov}(W_s, W_t) = \int e^{i\lambda^T(s-t)} \frac{1}{(1+\|\lambda\|^2)^{(\alpha+d/2)}}\, d\lambda.$$



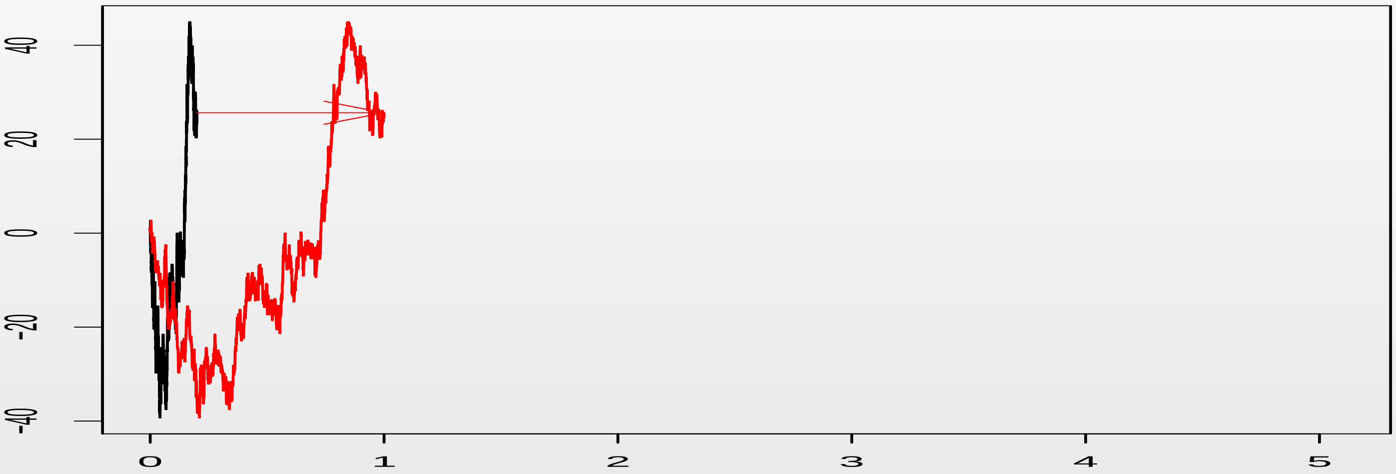

THEOREM

- If $\theta_0 \in C^\beta[0,1]^d$, then rate of contraction is $n^{-(\alpha\wedge\beta)/(2\alpha+d)}$.

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

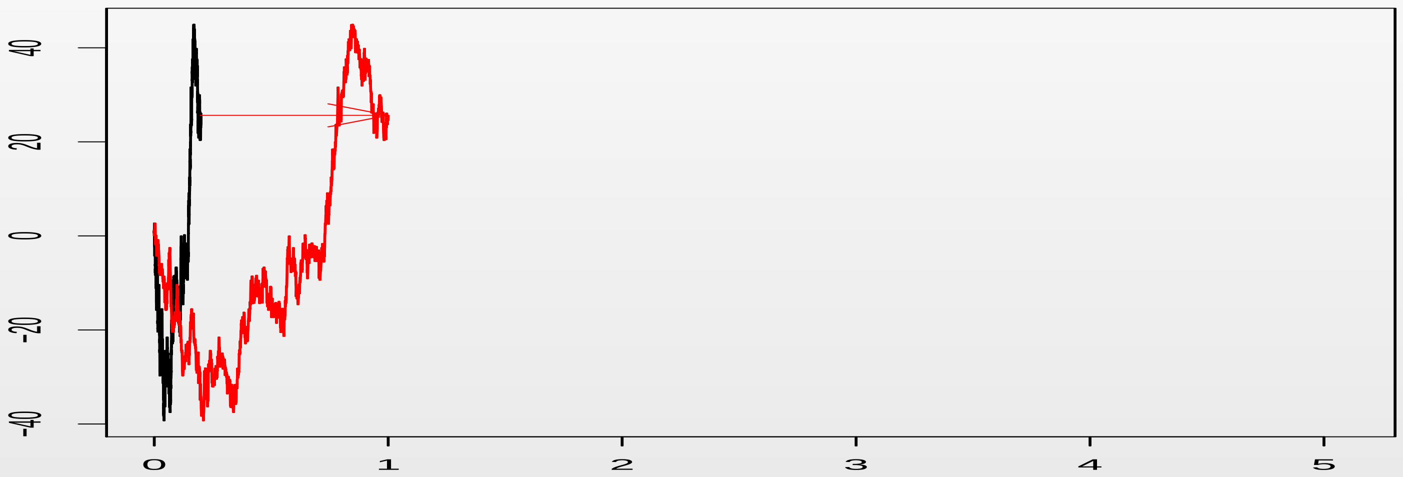
## Time-scaling Gaussian processes

Sample paths can be smoothed by stretching

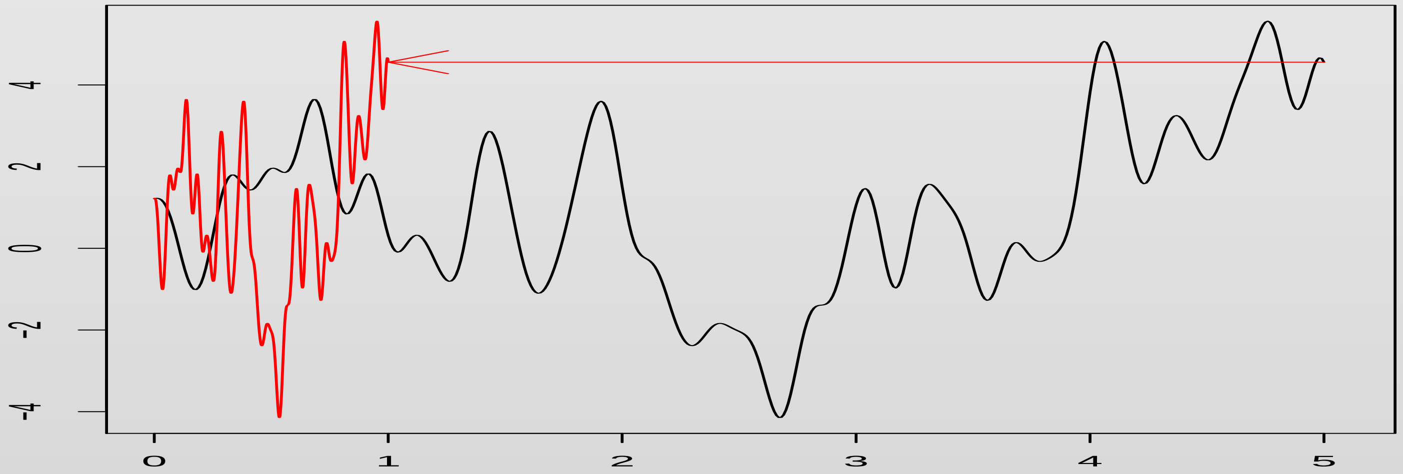


## Time-scaling Gaussian processes

Sample paths can be smoothed by stretching




or roughened by shrinking



## Time-scaling integrated Brownian motion

$G = (G_t\colon t > 0)$ the $k$-fold integral of Brownian motion “released at zero” and

$$c_n \sim n^{(\beta-k-1/2)/(2\beta+1)(k+1/2)}.$$

THEOREM

The prior $W = (G_{t/c_n}\colon 0 \le t \le 1)$ gives optimal rate for $\theta_0 \in C^\beta[0,1]$, $\beta \in (0, k+1]$.

## Time-scaling integrated Brownian motion

$G = (G_t \colon t > 0)$ the $k$-fold integral of Brownian motion “released at zero” and

$$c_n \sim n^{(\beta-k-1/2)/(2\beta+1)(k+1/2)}.$$

THEOREM
The prior $W = (G_{t/c_n} \colon 0 \le t \le 1)$ gives optimal rate for $\theta_0 \in C^\beta[0,1]$, $\beta \in (0, k+1]$.

- $\beta < k + 1/2$: $c_n \to 0$ (shrink).
- $\beta \in (k + 1/2, k+1]$: $c_n \to \infty$ (stretch).

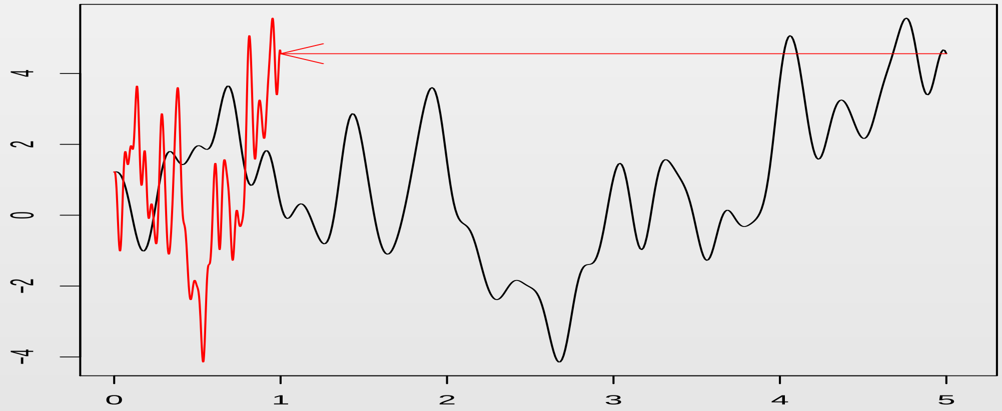
Stretching helps a little, shrinking helps a lot.

[By self-similarity *time-scaling* $G_{t/c}$ is equivalent to *space-scaling* $c^{k+1/2} G_t$.]

## Time-scaling smooth stationary process

$G = (G_t\colon t \in \mathbb{R}^d)$ the stationary Gaussian field with Gaussian spectral measure and

$$c_n \sim n^{-1/(2\beta+d)}.$$




THEOREM
The prior $W_t = G_{t/c_n}$ gives nearly optimal rate for $\theta_0 \in C^\beta[0,1]$, any $\beta > 0$.

Shrinking can adapt supersmooth prior to everything.

## Adaptation

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

This can be alleviated by adapting the prior to the data by

- *hierarchical Bayes:* putting a prior on the regularity, or on a scaling.
- *empirical Bayes:* using a regularity or scaling determined by maximum likelihood on the marginal distribution of the data.

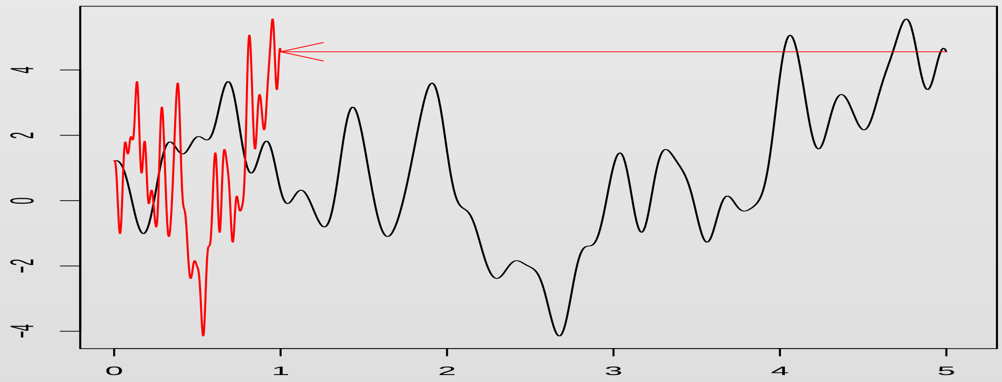
The first is known to work in some generality.
For the second there are some, but not many results.

## Adaptation by random scaling — example

- Choose $A^d$ from a Gamma distribution.
- Choose $(G_t\colon t > 0)$ centered stationary Gaussian with Gaussian spectral measure.
- Set $W_t \sim G_{At}$.

THEOREM

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




Full Bayes solves the bandwidth problem.

## Recovery: summary

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

# 3. Credible Sets

## Notation: the Bayesian machine

Given a prior model $\theta \sim \Pi_n$ and a likelihood $Y_n |\, \theta \sim p_n(y |\, \theta)$, the posterior distribution $\theta |\, Y_n$ is given by

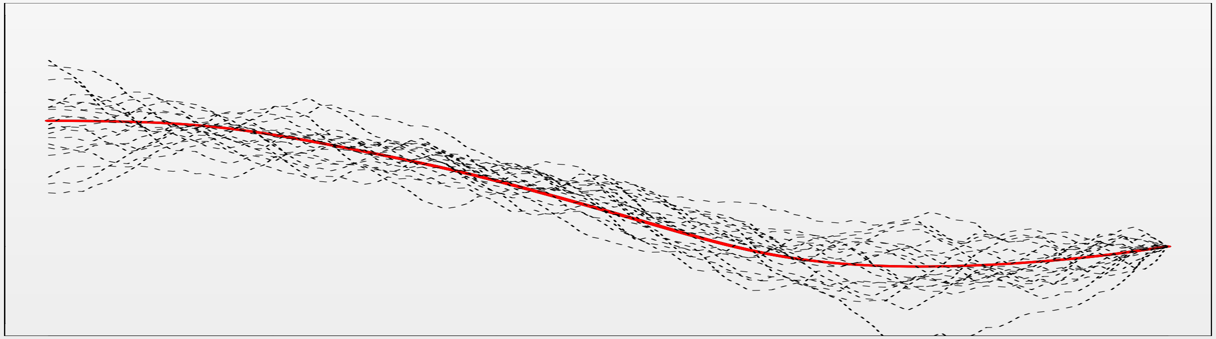
$$d\Pi_n(\theta |\, Y_n) \propto p_n(Y_n |\, \theta)\, d\Pi_n(\theta).$$

Two uses:

- recovery, e.g. by mode, or mean.
- expression of uncertainty, e.g. by a credible set: a set $C_n(Y_n)$ with $\Pi_n\big(C_n(Y_n) |\, Y_n\big) = 0.95$.

## Frequentist Bayes

Assume that data $Y_n$ is generated according to $\theta_0$.



The coverage of the credible region $C_n(Y_n)$ is

$$\mathrm{P}_{\theta_0}\big(C_n(Y_n) \ni \theta_0\big).$$

*Does it tend to 95 %?*

*Does at least the posterior spread express remaining uncertainty?*

# Uncertainty quantification: an early answer

*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 $\varepsilon$'s are i.i.d. with mean zero and variance $\sigma^2$ and $\beta$ is an unknown smooth function. A Gaussian prior distribution is specified by assuming $\beta$ 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 $\beta$. 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 $t_{ni} = i/n$, $\beta\colon [0,1] \to \mathbb{R}$ is an unknown smooth function, and $\varepsilon_1, \varepsilon_2, \dots$ are i.i.d. random errors with mean 0 and known variance $\sigma^2 < \infty$. The $\varepsilon_i$ are modeled as $N(0,\sigma^2)$. A Gaussian prior for $\beta$ will now be specified. Let $m \ge 2$ and for some constants $a_0, \dots, a_m$ with $a_m \ne 0$ let

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

*"Non-Bayesians often find such Bayesian procedures attractive because as $n \to \infty$, the frequentist coverage probability of the Bayesian regions tends to the posterior coverage probability in "typical" cases. It was my hope that this would also hold in the nonparametric setting $[\cdots]$ Unfortunately, the hoped for result is false in about the worst possible way, viz.,"*

$$\liminf_{n\to\infty} \mathrm{P}_{\theta_0}\Big(C_n(Y_n) \ni \theta_0\Big) = 0, \qquad \text{for } \Pi\text{-a.e. } \theta_0.$$

## Linear Gaussian inverse problems

The model of Cox (1993) can be cast in sequence form by representing functions $\theta$ on a suitable basis $e_1, e_2, \ldots$ as

$$\theta(x) = \sum_{i=1}^{\infty} \theta_i e_i(x).$$

DATA: independent $Y_{n,1}, Y_{n,2}, \ldots$ with $Y_{n,i} |\, \theta_i \sim N(\kappa_i \theta_i, n^{-1})$ for known $\kappa_i$.

PRIOR: independent $\theta_i \sim N(0, \lambda_i)$.

## Linear Gaussian inverse problems

The model of Cox (1993) can be cast in sequence form by representing functions $\theta$ on a suitable basis $e_1, e_2, \ldots$ as

$$\theta(x) = \sum_{i=1}^{\infty} \theta_i e_i(x).$$

DATA: $Y_n \,|\, \theta \sim N_\infty(K\theta, n^{-1}I)$ for known $K$.

PRIOR: $\theta \sim N_\infty(0, \Lambda)$.

## Linear Gaussian inverse problems

The model of Cox (1993) can be cast in sequence form by representing functions $\theta$ on a suitable basis $e_1, e_2, \ldots$ as

$$\theta(x) = \sum_{i=1}^{\infty} \theta_i e_i(x).$$

DATA: $Y_n |\, \theta \sim N_\infty(K\theta, n^{-1}I)$ for known $K$.

PRIOR: $\theta \sim N_\infty(0, \Lambda)$.

POSTERIOR: $\theta |\, Y_n \sim N_\infty\big(AY_n, S\big)$, for

$$A = \Lambda K^T \Big(\frac{1}{n} I + K\Lambda K^T\Big)^{-1}, \qquad S = \Lambda - A(n^{-1}I + K\Lambda K^T)A^T.$$

CREDIBLE SET: $\mathrm{ball}(AY_n, r)$, for $r$ with $N_\infty(0, S)(\mathrm{ball}(0, r)) = 0.95$.

## Sobolev models and priors

TRUTH: $\theta_0 \in S^\beta$, for

$$S^\beta = \Big\{ \sum_i \theta_i e_i : \sum_i i^{2\beta} \theta_i^2 < \infty \Big\}.$$

PRIOR: $\theta_1, \theta_2, \ldots$ independent with $\theta_i \sim N(0, \lambda_i)$, for

$$\lambda_i \asymp \frac{1}{i^{2\alpha+1}}.$$

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

# Sobolev models and priors

TRUTH: $\theta_0 \in S^\beta$, for

$$S^\beta = \Big\{ \sum_i \theta_i e_i : \sum_i i^{2\beta} \theta_i^2 < \infty \Big\}.$$

PRIOR: $\theta_1, \theta_2, \ldots$ independent with $\theta_i \sim N(0, \lambda_i)$, for

$$\lambda_i \asymp \frac{1}{i^{2\alpha+1}}.$$

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

[Alternative definition $S^\beta$: use $\sup_i |i^{2\beta}\theta_i^2|$ instead of $\sum_i i^{2\beta}\theta_i^2$.]

## Linear Gaussian inverse problem — rate of contraction

DATA: $Y_n|\,\theta \sim N_\infty(K\theta, n^{-1}I)$ for $\kappa_i \sim i^{-p}$.

PRIOR: $\theta \sim N_\infty(0, \Lambda)$.

THEOREM
For an $\alpha$-smooth prior and $\beta$-smooth truth, the posterior rate of contraction is

$$\left(\frac{1}{n}\right)^{\frac{\alpha\wedge\beta}{2\alpha+2p+1}}.$$

This is as usual:

- contraction for any combination of truth and prior ($\beta$ and $\alpha$).
- minimax rate of contraction iff prior and truth match ($\alpha = \beta$).

## Example: reconstruct derivative

The Volterra operator $K\colon L_2[0,1] \to L_2[0,1]$ is given by

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

The observation is $\big(Y_n(x)\colon x \in [0,1]\big)$, for $Z$ Gaussian white noise,

$$\dot{Y}_n(x) = \int_0^x \theta(s)\,ds + \frac{1}{\sqrt{n}} Z(x), \qquad x \in [0,1].$$

## Example: reconstruct derivative

The Volterra operator $K: L_2[0,1] \to L_2[0,1]$ is given by

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

The observation is $\big(Y_n(x): x \in [0,1]\big)$, for $Z$ Gaussian white noise,

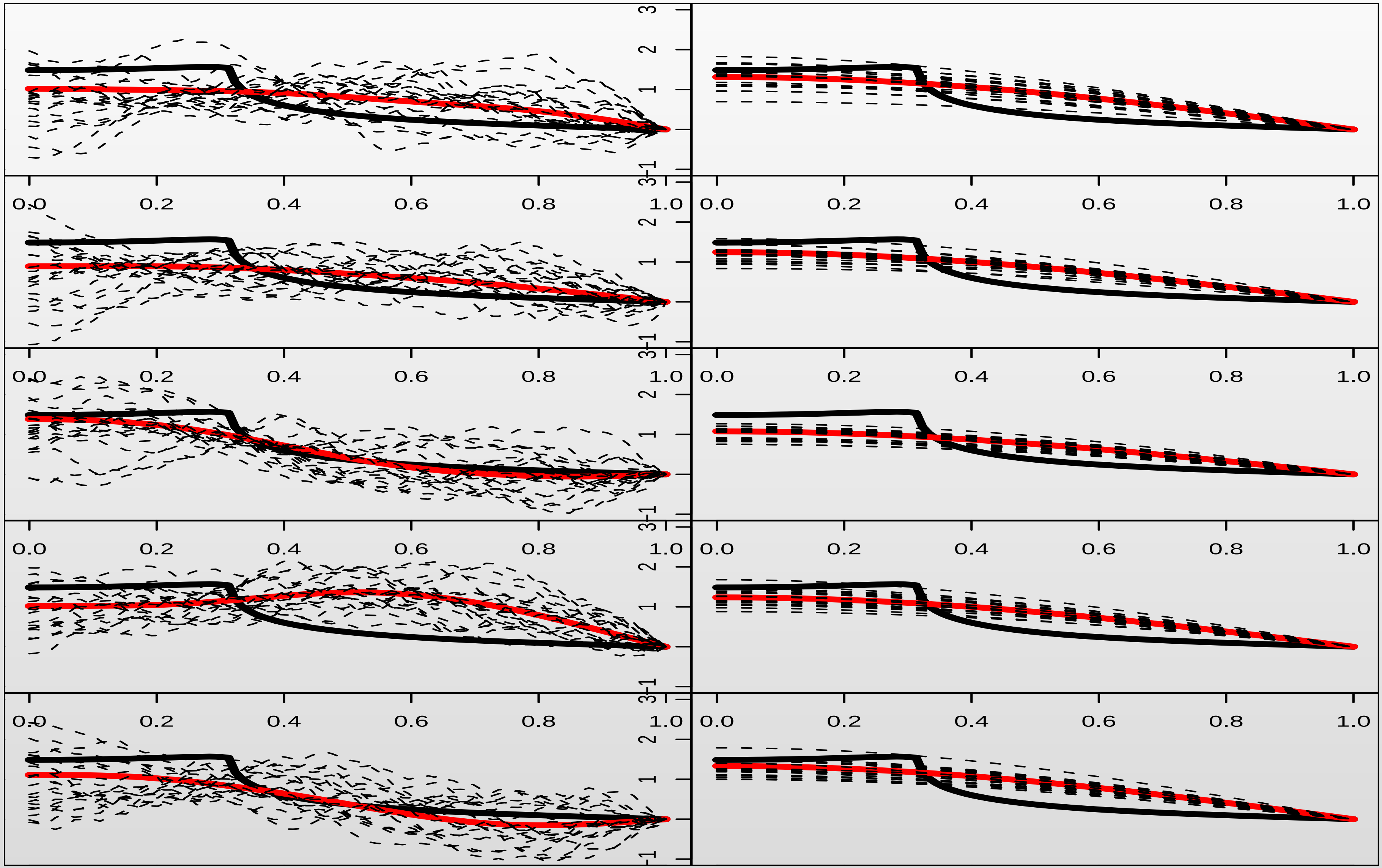
$$\dot{Y}_n(x) = \int_0^x \theta(s)\,ds + \frac{1}{\sqrt{n}} Z(x), \qquad x \in [0,1].$$

*mildly inverse problem:* $Y_{n,i} \,|\, \theta_i \sim N(\kappa_i \theta_i, n^{-1})$ for

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

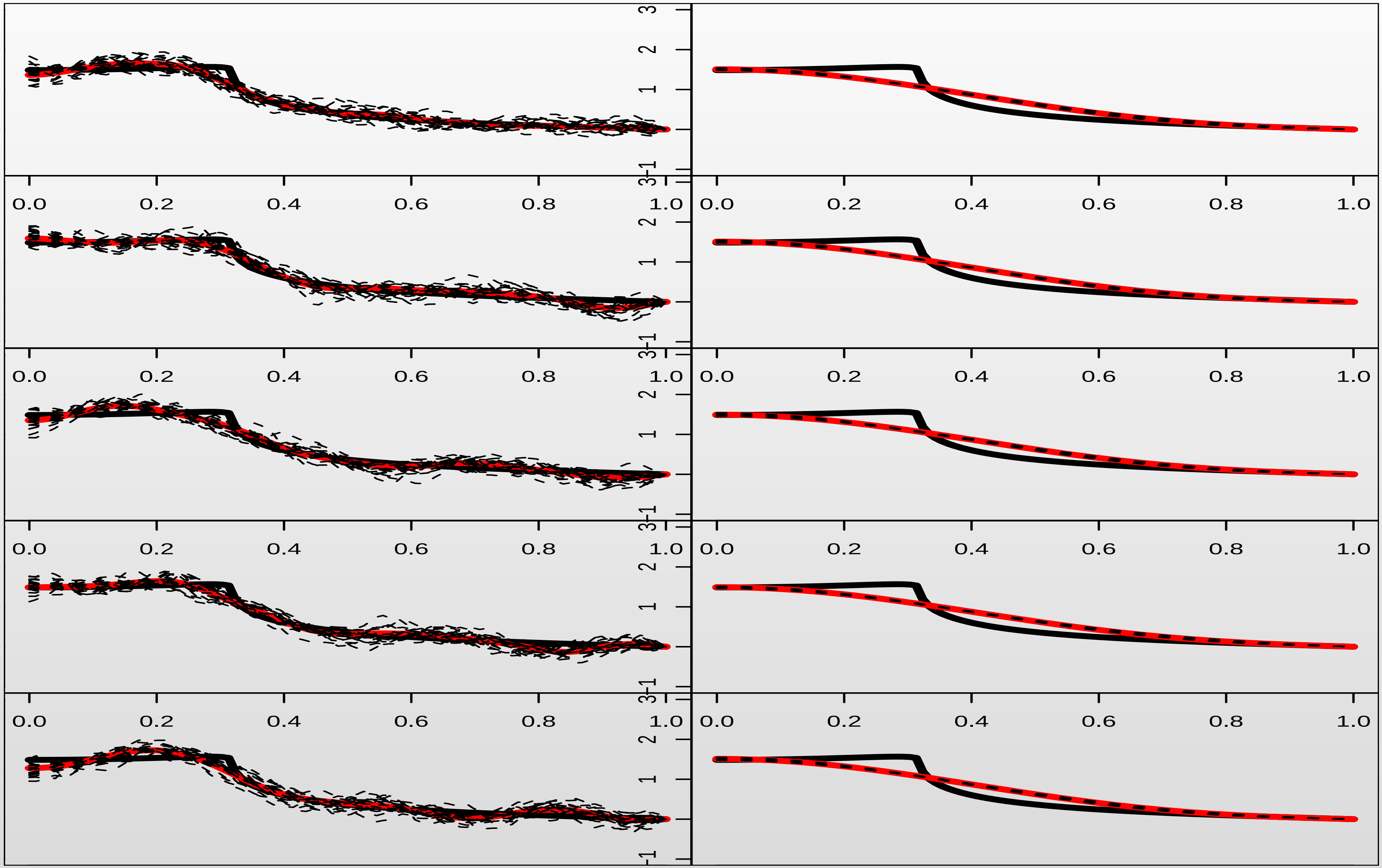
$$(i = 0, 1, 2, \ldots).$$

# Example: reconstruct derivative (n=100)



True $\theta_0$ (black), posterior mean (red), and 20 realizations from the posterior, repeated 5 times for a rough prior (left) and a smooth prior (right).

# Example: reconstruct derivative (n=100 000)




True $\theta_0$ (black), posterior mean (red), and 20 realizations from the posterior, repeated 5 times for a rough prior (left) and a smooth prior (right).

## Linear Gaussian inverse problem — credible balls

POSTERIOR: $\theta | Y_n \sim N_\infty\big(AY_n, S\big)$.

CREDIBLE SET: $\mathrm{ball}(AY_n, r)$, for $r$ with $N_\infty(0,S)(\mathrm{ball}(0,r)) = 0.95$.

THEOREM

For $\alpha$-smooth prior and $\beta$-smooth truth:

- If $\alpha < \beta$, then asymptotic coverage is 1 (uniformly).
- If $\alpha = \beta$, then any asymptotic coverage $c \in (0,1)$ occurs along some sequence in $S^\beta$.
- If $\alpha > \beta$, then for some $\theta \in S^\beta$ asymptotic coverage is 0.

The credible ball has the correct order of magnitude iff $\alpha \leq \beta$.

If $\alpha > \beta$, then the prior oversmoothes and creates bias.
If $\alpha < \beta$, then credible balls are conservative, but OK as a rough indication of statistical uncertainty.

## Linear Gaussian inverse problem — credible balls

POSTERIOR: $\theta|\, Y_n \sim N_\infty\big(AY_n, S\big)$.

CREDIBLE SET: $\mathrm{ball}(AY_n, r)$, for $r$ with $N_\infty(0,S)(\mathrm{ball}(0,r)) = 0.95$.

THEOREM

For $\alpha$-smooth prior and $\beta$-smooth truth:

- If $\alpha < \beta$, then asymptotic coverage is 1 (uniformly).
- If $\alpha = \beta$, then any asymptotic coverage $c \in (0,1)$ occurs along some sequence in $S^\beta$.
- If $\alpha > \beta$, then for some $\theta \in S^\beta$ asymptotic coverage is 0.

The credible ball has the correct order of magnitude iff $\alpha \leq \beta$.

If $\alpha > \beta$, then the prior oversmoothes and creates bias.
If $\alpha < \beta$, then credible balls are conservative, but OK as a rough indication of statistical uncertainty.

*Cox's result: truths $\theta_0$ generated from an $\alpha$-smooth prior belong with probability one to $S^\beta$ for any $\beta < \alpha$, but not to $S^\alpha$. Their coverage is 0.*

## Linear Gaussian inverse problem — scaling the prior

DATA: $Y_n \,|\, \theta \sim N_\infty(K\theta, n^{-1}I)$ for $\kappa_i \sim i^{-p}$.

PRIOR: $\theta \sim N_\infty(0, \tau_n^2 \Lambda)$ for $\lambda_i = i^{-1-2\alpha}$.

THEOREM

For $\theta_0 \in S^\beta$ the *best rescaling rate* is $\tilde{\tau}_n = n^{(\alpha-\tilde{\beta})/(2\tilde{\beta}+2p+1)}$, for $\tilde{\beta} = \beta \wedge (1 + 2\alpha + 2p)$ .

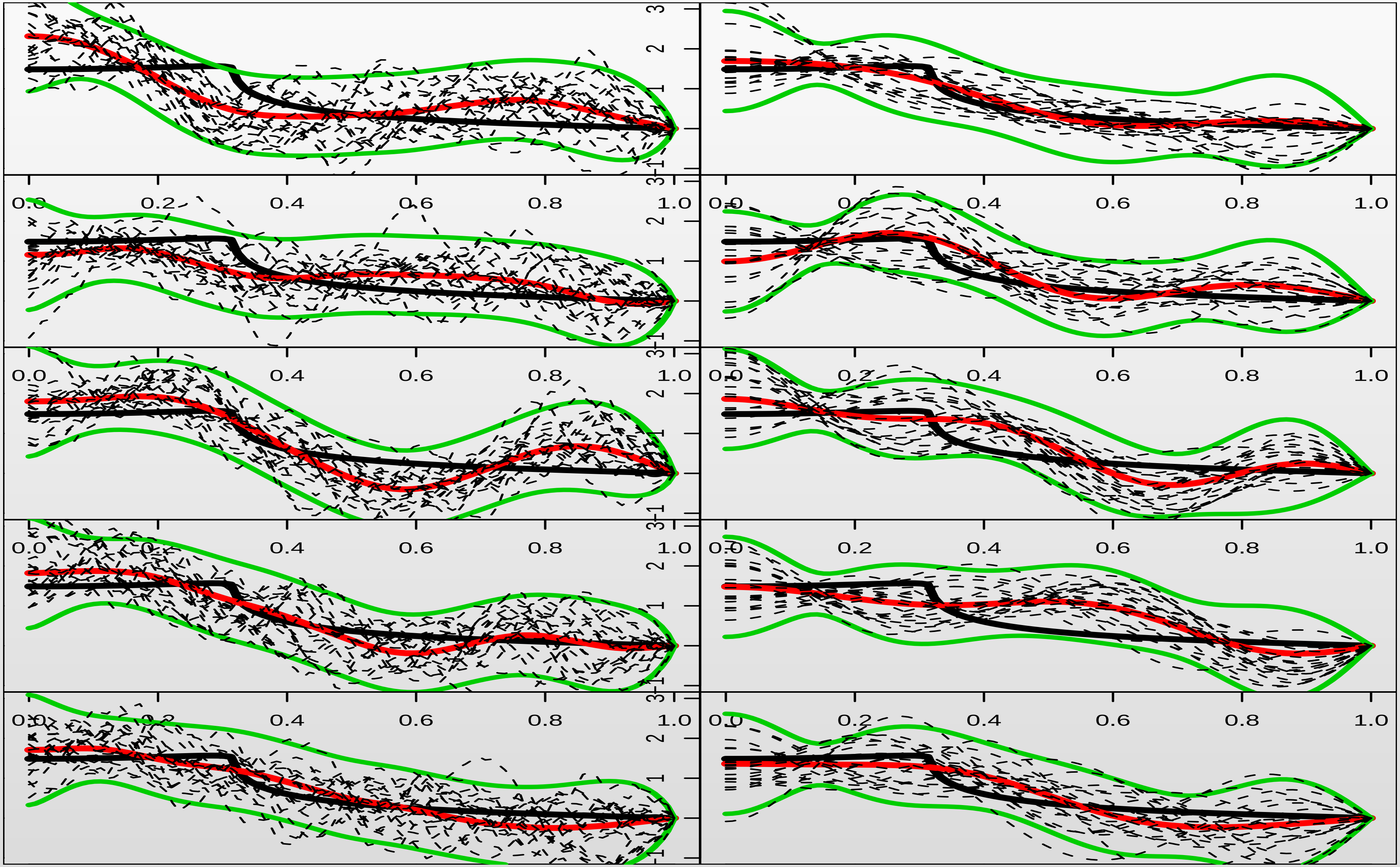
- If $\tau_n \gg \tilde{\tau}_n$, then the asympotic coverage is 1.
- If $\tau_n \asymp \tilde{\tau}_n$, then any asymptotic coverage occurs.
- If $\tau_n \ll \tilde{\tau}_n$, then the asymptotic coverage is 0.

In the first two cases the size of the credible sets has the correct order.

| Appropriate scaling solves the problem. |
|---|

[The contraction rate is minimax iff $\beta \leq 2\alpha + 2p + 1$. Can scale a smooth prior to become rougher, but not conversely.]

# Example: reconstruct derivative (n=1000)




True $\theta_0$ (black), posterior mean (red), and 20 realizations from the posterior, repeated 5 times for a rescaled rough prior (left) and an optimally rescaled smooth prior (right).

## Credible sets: first summary

In a nonparametric set-up the prior is not washed out by the data.

Recovery: the prior influences the posterior contraction rate (although "consistency" occurs for most priors).

Uncertainty quantification: the prior makes it felt strongly: if it mistakes the truth for being more regular than it is, the posterior will:

- be too concentrated (*leave too little uncertainty*).
- centre far away from the truth (*oversmooth*).

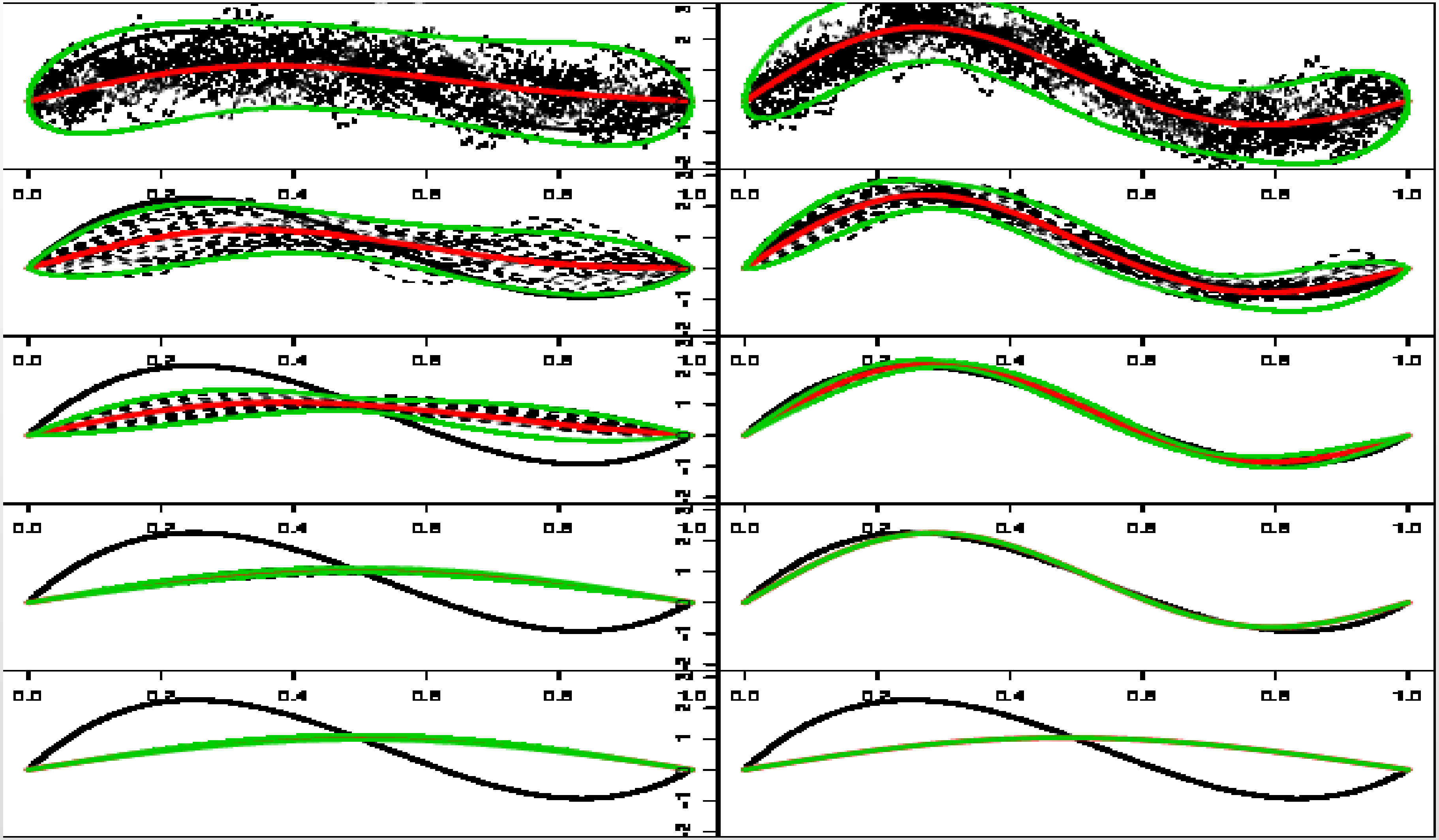
Together these may make for disastrous credible sets.

A solution:

- Undersmooth! Make the prior at least as rough as the truth *(Undersmoothing gives coverage)*.
- but not too much! *(Undersmoothing deteriorates recovery)*.

[Much work to be done. Results available only for the linear Gaussian inverse problem and Gaussian regression.]

# Example: heat equation (n=10 000, n=100 000 000)



True $\theta_0$ (black), posterior mean (red), 20 realizations from the posterior (dashed black), and posterior credible bands (green).
In all ten panels $\beta = 2.5$. Left: $n = 10^4$ and $\alpha = 0.5, 1, 2, 5, 10$ (top to bottom); right: $n = 10^8$ and $\alpha = 0.5, 1, 2, 5, 10$ (top to bottom).

# 4. Adaptive Credible Sets

## Adaptation

For recovery it can be useful to make a prior depend on a hyperparameter, in a hierarchical or empirical Bayes set-up.

*How does this work for credible sets?*

## Linear Gaussian inverse problem — random smoothness

DATA: $Y_n |\, \theta \sim N_\infty(K\theta, n^{-1}I)$ for $\kappa_i \sim i^{-p}$.

PRIOR: $\theta \sim N_\infty(0, \Lambda_\alpha)$ for $\lambda_i = i^{-1-2\alpha}$.

POSTERIOR: $\theta |\, Y_n \sim N_\infty\big(A_\alpha Y_n, S_\alpha\big)$.

CREDIBLE SET: $\mathrm{ball}(A_\alpha Y_n, r_\alpha)$, for $r_\alpha$ with $N_\infty(0, S_\alpha)(\mathrm{ball}(0, r_\alpha)) = 0.95$.

The empirical Bayes method uses the MLE $\hat\alpha$ for the marginal model $Y_n \sim N_\infty(0, K\Lambda_\alpha K^T + n^{-1}I)$:

$$\hat\alpha = \operatorname*{argmax}_\alpha \sum_{i=1}^\infty \left( \frac{n^2}{i^{1+2\alpha+2p} + n} Y_{n,i}^2 - \log\Big(1 + \frac{n}{i^{1+2\alpha+2p}}\Big)\right).$$

## Linear Gaussian inverse problem — random smoothness

DATA: $Y_n|\,\theta \sim N_\infty(K\theta, n^{-1}I)$ for $\kappa_i \sim i^{-p}$.

PRIOR: $\theta \sim N_\infty(0, \Lambda_\alpha)$ for $\lambda_i = i^{-1-2\alpha}$.

POSTERIOR: $\theta|\,Y_n \sim N_\infty\bigl(A_\alpha Y_n, S_\alpha\bigr)$.

CREDIBLE SET: $\mathrm{ball}(A_\alpha Y_n, r_\alpha)$, for $r_\alpha$ with $N_\infty(0, S_\alpha)(\mathrm{ball}(0, r_\alpha)) = 0.95$.

The empirical Bayes method uses the MLE $\hat\alpha$ for the marginal model $Y_n \sim N_\infty(0, K\Lambda_\alpha K^T + n^{-1}I)$:

$$\hat\alpha = \operatorname*{argmax}_\alpha \sum_{i=1}^\infty \biggl(\frac{n^2}{i^{1+2\alpha+2p}+n} Y_{n,i}^2 - \log\Bigl(1+\frac{n}{i^{1+2\alpha+2p}}\Bigr)\biggr).$$

This works for recovery.

*Does it also work for uncertainty quantification?*

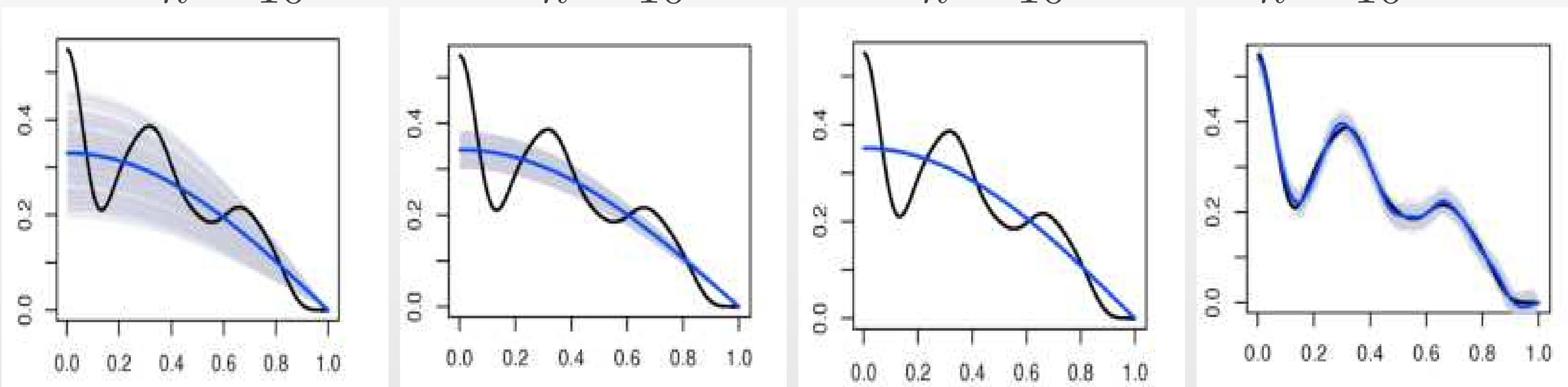
*Does* $\mathrm{ball}(A_{\hat\alpha} Y_n, r_{\hat\alpha})$ *cover?*

[Hierarchical Bayes, with prior on $\alpha$, probably works similarly.]

## Example: reconstructing a derivative

*Credible sets determined by empirical Bayes can be terribly wrong.*

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




True $\theta_0$ (black), posterior mean (blue) and 95 % realizations (out of 2000) that are closest to the posterior mean.
Same truth, different $n$, prior smoothness determined by empirical Bayes.

This is a *counterexample of a truth*. For some truths the results are good.

## What do the frequentists say? — Honesty

A set $C_n(Y_n)$ is an honest confidence set if

$$\mathrm{P}_{\theta_0}\big(C_n(Y_n) \ni \theta_0\big) \geq 0.95, \qquad \text{for all } \theta_0 \in \Theta_0.$$

$\Theta_0$ contains *'all possible truths'*, e.g. $\Theta_0 = S_1^\beta$, Sobolev ball of regularity $\beta$.

THEOREM

For given $\beta$ there exist $C_n(Y_n)$ of diameter of the order $O_P\big(n^{-\beta/(1+2\beta)}\big)$ that are honest over $S_1^\beta$.

## What do the frequentists say? — Honesty

A set $C_n(Y_n)$ is an honest confidence set if

$$\mathrm{P}_{\theta_0}\big(C_n(Y_n) \ni \theta_0\big) \geq 0.95, \qquad \text{for all } \theta_0 \in \Theta_0.$$

$\Theta_0$ contains *'all possible truths'*, e.g. $\Theta_0 = S_1^\beta$, Sobolev ball of regularity $\beta$.

THEOREM
For given $\beta$ there exist $C_n(Y_n)$ of diameter of the order $O_P(n^{-\beta/(1+2\beta)})$ that are honest over $S_1^\beta$.

THEOREM [Low, Robins+vdV, Juditzky+Lacroix.]
If $C_n(Y_n)$ is honest over $\cup_{\beta \geq \beta_0} S_1^\beta$, then its diameter is of the uniform order $O_P(n^{-\beta_0/(1/2+2\beta_0)})$ over $S^\beta$ for $\beta \geq 2\beta_0$.

The diameter is determined by the biggest model (smallest $\beta$).

## What do the frequentists say? — Honesty

A set $C_n(Y_n)$ is an honest confidence set if

$$\mathrm{P}_{\theta_0}\big(C_n(Y_n) \ni \theta_0\big) \geq 0.95, \qquad \text{for all } \theta_0 \in \Theta_0.$$

$\Theta_0$ contains *'all possible truths'*, e.g. $\Theta_0 = S_1^\beta$, Sobolev ball of regularity $\beta$.

THEOREM
For given $\beta$ there exist $C_n(Y_n)$ of diameter of the order $O_P(n^{-\beta/(1+2\beta)})$ that are honest over $S_1^\beta$.

THEOREM [Low, Robins+vdV, Juditzky+Lacroix.]
If $C_n(Y_n)$ is honest over $\cup_{\beta \geq \beta_0} S_1^\beta$, then its diameter is of the uniform order $O_P(n^{-\beta_0/(1/2+2\beta_0)})$ over $S^\beta$ for $\beta \geq 2\beta_0$.

The diameter is determined by the biggest model (smallest $\beta$).

[One should also consider adaptation to the *radius* of the Sobolev balls.
For credible bands the diameter is of the order $n^{-\beta_0/(1+2\beta_0)}$ for $\beta \geq \beta_0$.]

## What do the frequentists say? — Discrepancy between estimation and uncertainty quantification

Adaptive estimation: [1990s]

- A more regular true function is easier to estimate.
- Estimators can be simultaneously optimal for multiple regularities (e.g. *wavelet shrinkage*).
- Bayesian estimators can achieve this by a prior on a ‘bandwidth parameter’.

Uncertainty quantification: [2000s]

- Honest uncertainty quantification must argue from the worst case scenario: the smallest possible regularity level.
- The size of an honest confidence set cannot adapt (much) to unknown regularity.

## What do the frequentists say? — Discrepancy between estimation and uncertainty quantification

Adaptive estimation: [1990s]

- A more regular true function is easier to estimate.
- Estimators can be simultaneously optimal for multiple regularities (e.g. *wavelet shrinkage*).
- Bayesian estimators can achieve this by a prior on a ‘bandwidth parameter’.

Uncertainty quantification: [2000s]

- Honest uncertainty quantification must argue from the worst case scenario: the smallest possible regularity level.
- The size of an honest confidence set cannot adapt (much) to unknown regularity.

*“Adaptive estimators [..] do the best that is possible in view of the properties (smoothness or complexity) of the underlying function to be estimated. [...] This is quite satisfactory but [..] the estimator does not tell you how well it does [...] you have no idea about the order of magnitude of the distance between your estimator and the truth [...].”*

*[Lucien Birgé, 2002, discussion of a paper by Hoffmann+Lepski.]*

## What do the frequentists say? — Self-similarity

A sequence $(\theta_1, \theta_2, \ldots) \in S^\beta$ is self-similar if, for all $I = 1, 2, \ldots$,

$$\sum_{i=I}^{1000I} i^{2\beta}\theta_i^2 \geq \frac{1}{1000} \sup_i i^{2\beta}\theta_i^2.$$

THEOREM [Bull and Nickl, 2012]
There exist $C_n(Y_n)$ that are honest over the set of all self-similar $\theta_0 \in \cup_\beta S_1^\beta$ such that the radius is of the order $O_P(n^{-\beta/(1+2\beta)})$ whenever $\theta_0 \in S^\beta$.

Interpretation of self-similarity: $(\theta_1, \theta_2, \ldots)$ has the *same character at any resolution level* ($i \to \infty$).
A noisy data set $Y_n$ can infer this character from the estimated sequence $(\hat\theta_1, \ldots, \hat\theta_{\hat N})$ for $\hat N$ the ‘effective’ dimension.

## Linear Gaussian inverse problems — Credible sets are honest over self-similar functions

DATA: $Y_n |\, \theta \sim N_\infty(K\theta, n^{-1}I)$ for $\kappa_i \sim i^{-p}$

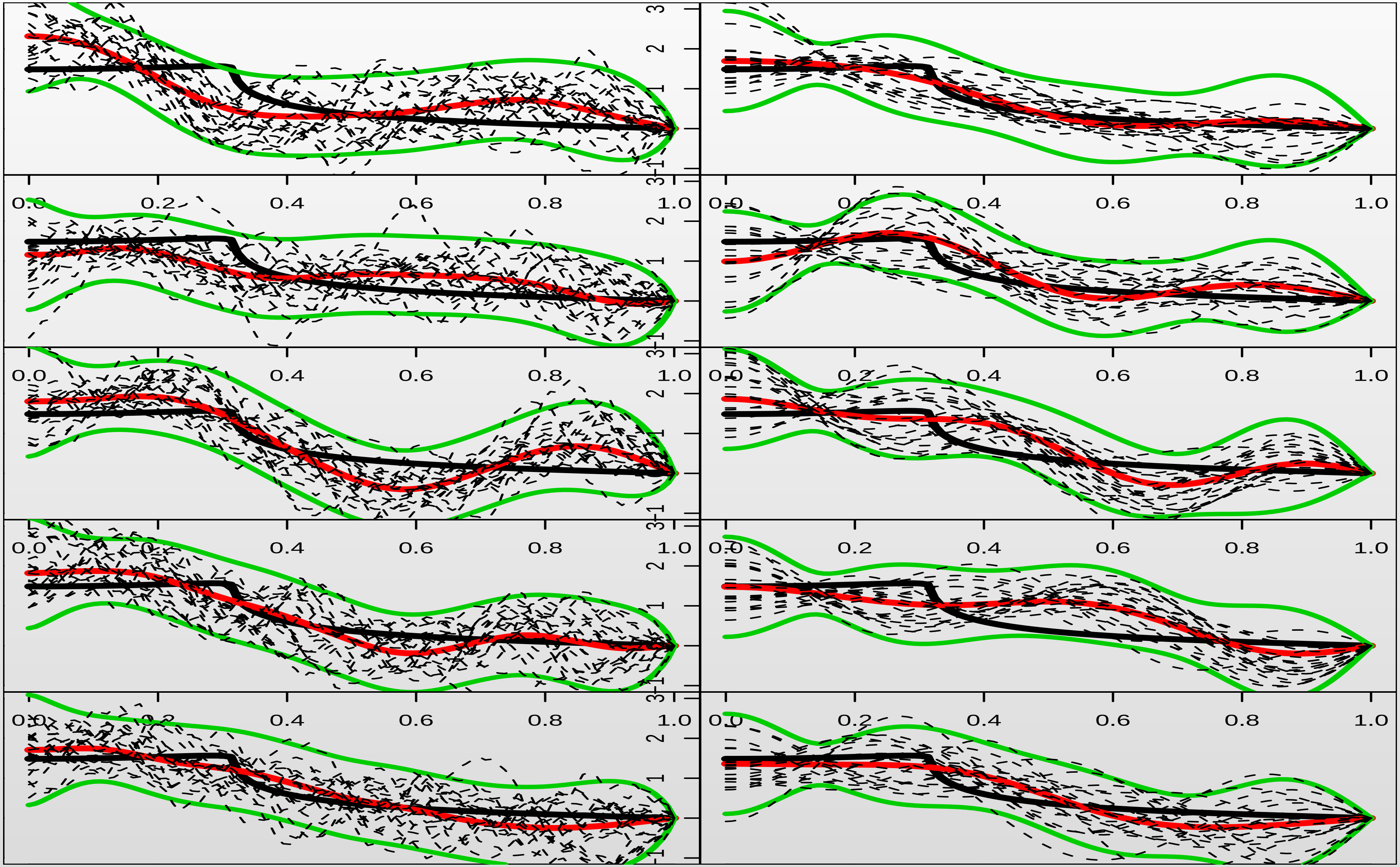
PRIOR: $\theta \sim N_\infty(0, \Lambda)$ for $\lambda_i = i^{-1-2\alpha}$.

CREDIBLE SET: $\mathrm{ball}(A_\alpha Y_n, r_\alpha)$, for $r_\alpha$ with $N_\infty(0, S_\alpha)(\mathrm{ball}(0, r_\alpha)) = 0.95$.

THEOREM
If $\hat\alpha$ is the MLE for the marginal law of $Y_n$, then credible ball $\mathrm{ball}(A_{\hat\alpha} Y_n, \hat r_{\hat\alpha})$ is nearly honest over the set of all self-similar $\theta_0 \in \cup_\beta S_1^\beta$, and has radius nearly of the order $O_P(n^{-\beta/(1+2\beta)})$ whenever $\theta_0 \in S^\beta$.

Empirical Bayes works for self-similar truths.

# Example: reconstruct derivative (n=1000)




True $\theta_0$ (black), posterior mean (red), and 20 realizations from the posterior, repeated 5 times for a rescaled rough prior (left) and a rescaled smooth prior (right).

# Credible sets are honest over prior sets?

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

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

*"Non-Bayesians often find such Bayesian procedures attractive because as $n \to \infty$, the frequentist coverage probability of the Bayesian regions tends to the posterior coverage probability in "typical" cases. It was my hope that this would also hold in the nonparametric setting [$\cdots$] Unfortunately, the hoped for result is false in about the worst possible way, viz.,"*

$$\liminf_{n\to\infty} \mathrm{P}_{\theta_0}\Big(\mathrm{ball}(A_\alpha Y_n, r_\alpha) \ni \theta_0\Big) = 0, \qquad \text{for } \Pi\text{-a.e. } \theta_0.$$

# Credible sets are honest over prior sets?

*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 $\varepsilon$'s are i.i.d. with mean zero and variance $\sigma^2$ and $\beta$ is an unknown smooth function. A Gaussian prior distribution is specified by assuming $\beta$ 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 $\beta$. 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 $t_{ni} = i/n$, $\beta\colon [0,1] \to \mathbb{R}$ is an unknown smooth function, and $\varepsilon_1, \varepsilon_2, \dots$ are i.i.d. random errors with mean 0 and known variance $\sigma^2 < \infty$. The $\varepsilon_i$ are modeled as $N(0, \sigma^2)$. A Gaussian prior for $\beta$ will now be specified. Let $m \ge 2$ and for some constants $a_0, \dots, a_m$ with $a_m \ne 0$ let

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

*"Non-Bayesians often find such Bayesian procedures attractive because as $n \to \infty$, the frequentist coverage probability of the Bayesian regions tends to the posterior coverage probability in "typical" cases. It was my hope that this would also hold in the nonparametric setting [$\cdots$] Unfortunately, the hoped for result is false in about the worst possible way, viz.,"*

$$\liminf_{n\to\infty} \mathrm{P}_{\theta_0}\Big(\mathrm{ball}\big(A_\alpha Y_n, (\log n) r_\alpha\big) \ni \theta_0\Big) = 1, \qquad \text{for } \Pi\text{-a.e. } \theta_0.$$

## Credible sets are honest over prior sets? (2)

CONJECTURE

For $\hat\alpha$ determined by empirical Bayes in the linear inverse problem:

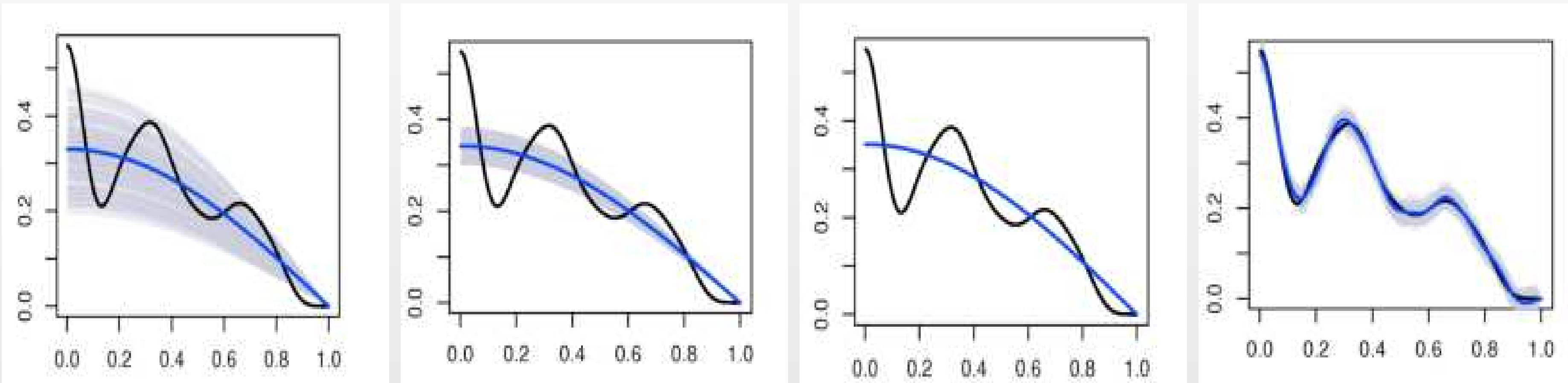
$$\liminf_{n\to\infty} \mathrm{P}_{\theta_0}\Bigl(\mathrm{ball}\bigl(\hat A_{\hat\alpha} Y_n, (\log n)\hat r_{\hat\alpha}\bigr) \ni \theta_0\Bigr) = 1, \qquad \text{for } \Pi_\alpha\text{-a.e. } \theta_0\text{, for every } \alpha.$$

[Honesty is questionable.]

## Example: reconstructing a derivative

*Credible sets determined by empirical Bayes can be terribly wrong.*

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




True $\theta_0$ (black), posterior mean (blue) and 95 % realizations (out of 2000) that are closest to the posterior mean.
Same truth, different $n$, prior smoothness determined by empirical Bayes.

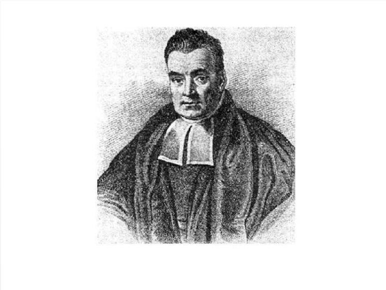
Same truth, different $n$, prior smoothness determined by empirical Bayes.

*WHAT CAUSES THIS BAD BEHAVIOUR?*

In this example the truth is very smooth, unlike any function that is generated from a prior.

[More fancy counterexamples exist.]

## Conclusions and Conjectures



Nonparametric credible regions are never “correct” frequentist confidence regions.

## Conclusions and Conjectures

Nonparametric credible regions are never "correct" frequentist confidence regions.

Priors that undersmooth the truth give a reasonable idea of the uncertainty in the posterior mean.

## Conclusions and Conjectures

Nonparametric credible regions are never “correct” frequentist confidence regions.

Priors that undersmooth the truth give a reasonable idea of the uncertainty in the posterior mean.

If the prior oversmoothes the truth, then the spread in the posterior is very misleading about the remaining uncertainty.

## Conclusions and Conjectures

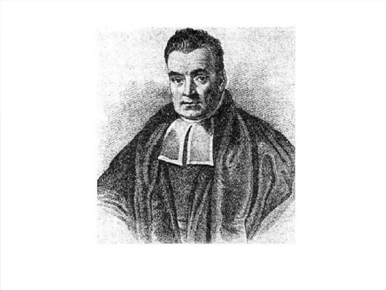
Nonparametric credible regions are never "correct" frequentist confidence regions.

Priors that undersmooth the truth give a reasonable idea of the uncertainty in the posterior mean.

If the prior oversmoothes the truth, then the spread in the posterior is very misleading about the remaining uncertainty.

This effect may disappear if the prior is scaled, for instance by an hierarchical or empirical Bayesian method,

## Conclusions and Conjectures



Nonparametric credible regions are never “correct” frequentist confidence regions.

Priors that undersmooth the truth give a reasonable idea of the uncertainty in the posterior mean.

If the prior oversmoothes the truth, then the spread in the posterior is very misleading about the remaining uncertainty.

This effect may disappear if the prior is scaled, for instance by an hierarchical or empirical Bayesian method, *but only for truths that resemble the prior*

## Conclusions and Conjectures

Nonparametric credible regions are never “correct” frequentist confidence regions.

Priors that undersmooth the truth give a reasonable idea of the uncertainty in the posterior mean.

If the prior oversmoothes the truth, then the spread in the posterior is very misleading about the remaining uncertainty.

This effect may disappear if the prior is scaled, for instance by an hierarchical or empirical Bayesian method, *but only for truths that resemble the prior* ???

## Conclusions and Conjectures

It seems we must either undersmooth or believe the fine details of our prior.

## Conclusions and Conjectures

It seems we must either undersmooth or believe the fine details of our prior.

Is that possible in nonparametrics?