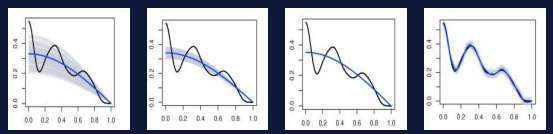
Nonparametric Bayesian Uncertainty Quantification

Lecture 2: Curve Estimation

Aad van der Vaart

Universiteit Leiden, Netherlands



YES, Eindhoven, January 2017

Co-authors

Bartek Knapik
(Amsterdam)

Suzanne Sniekers
(Leiden)

Botond Szabó
(Leiden)

Harry van Zanten
(Amsterdam)

Nonparametric credible sets

Credible sets

- prior model $\theta \sim \Pi_n$ for parameter θ,
- likelihood $Y_n | \, \theta \sim p_n(y | \, \theta)$ for the data,

Gives posterior distribution $\theta | \, Y_n \sim \Pi_n(\cdot | \, Y_n) \propto p_n(Y_n | \, \theta) \, d\Pi(\theta)$ as usual.

Two uses:

- recovery, e.g. by mode, or mean.
- expression of uncertainty.

Credible sets

- prior model $\theta \sim \Pi_n$ for parameter θ,
- likelihood $Y_n | \, \theta \sim p_n(y | \, \theta)$ for the data,

Gives posterior distribution $\theta | \, Y_n \sim \Pi_n(\cdot | \, Y_n) \propto p_n(Y_n | \, \theta) \, d\Pi(\theta)$ as usual.

Two uses:

- recovery, e.g. by mode, or mean.
- expression of uncertainty.

A credible set is a data-dependent set $C_n(Y_n)$ with

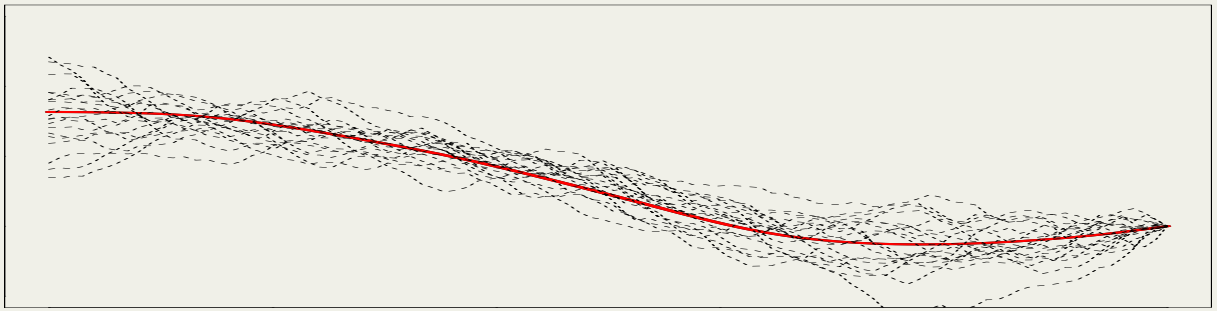
$$\Pi_n\big(\theta \in C_n(Y_n) | \, Y_n\big) = 0.95.$$

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?

Does

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

imply

$$\mathrm{P}_{\theta_0}\big(\theta_0 \in C_n(Y_n)\big) = 0.95?$$

Do credible sets correctly quantify *remaining uncertainty*?

Is a credible set a confidence set?

Does

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

imply

$$\mathrm{P}_{\theta_0}\big(\theta_0 \in C_n(Y_n)\big) = 0.95?$$

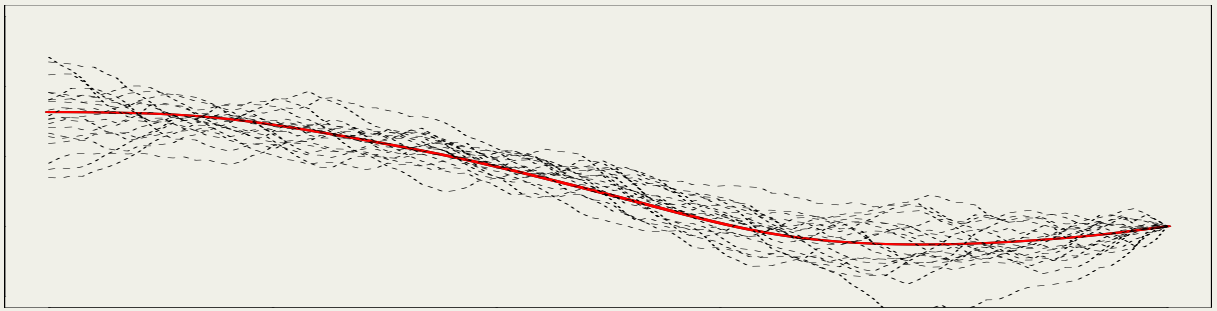
Rarely!

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*?

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?

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

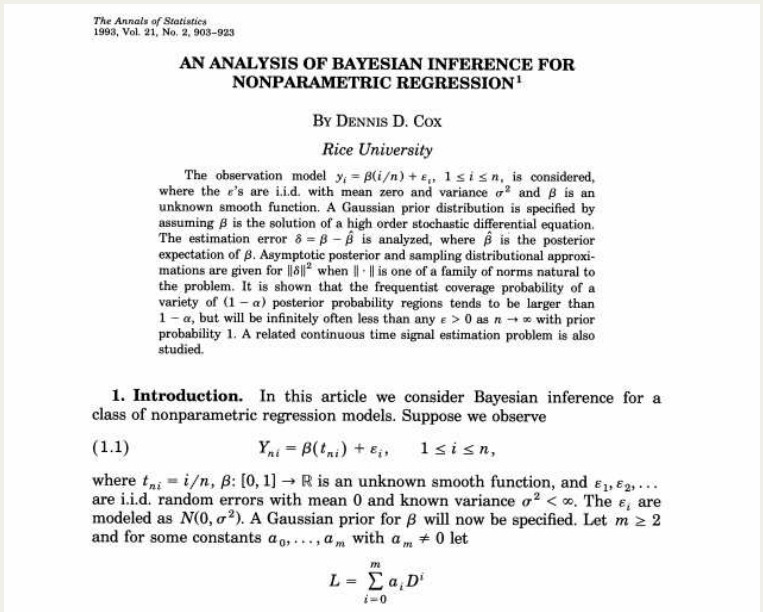
$$Y(t_i) = g(t_i) + \epsilon_i, \quad i = 1, 2, \ldots, n, \quad t_i \in [0, 1], \tag{1.1}$$

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.

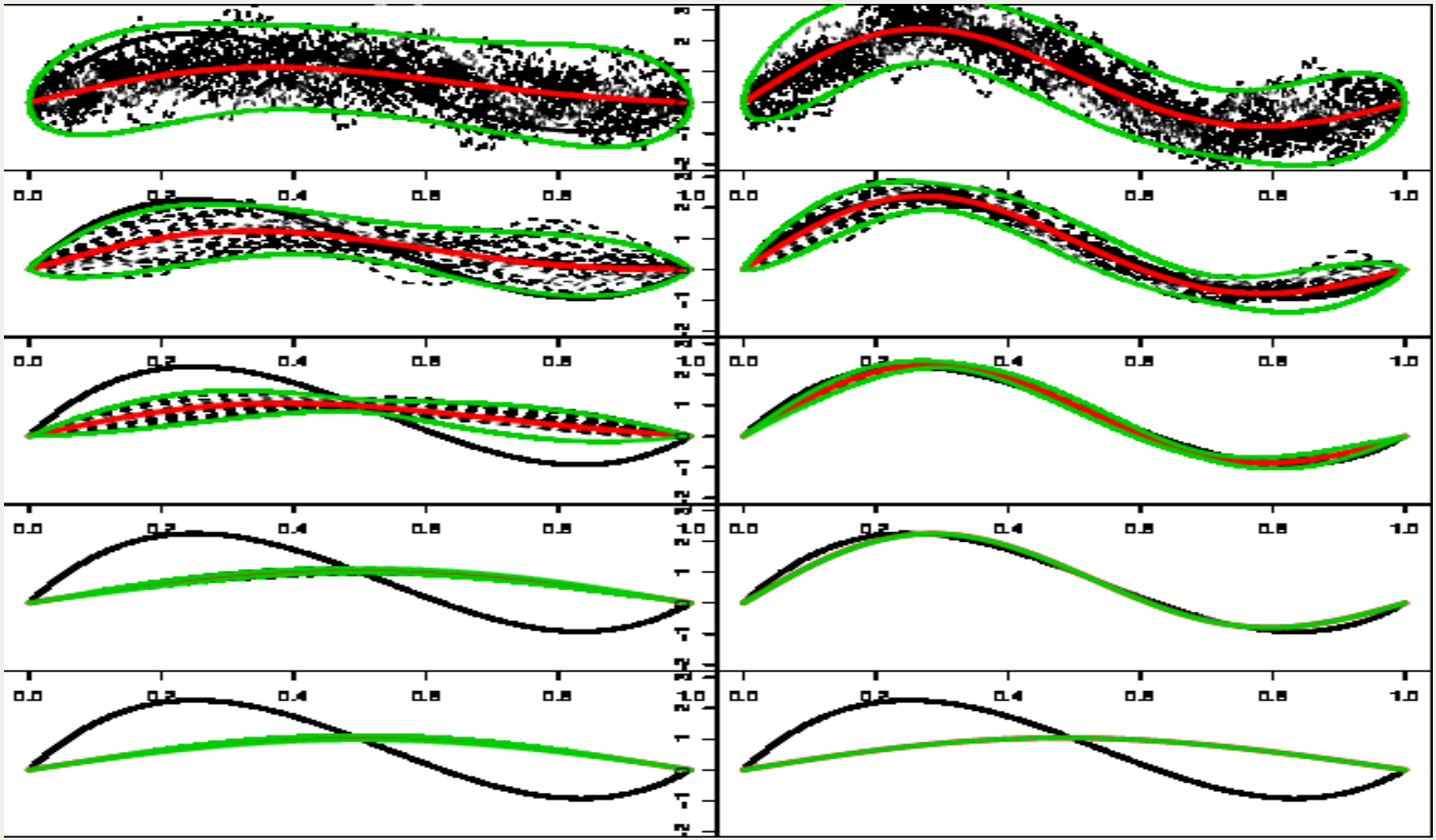
EXAMPLE

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

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

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

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



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_c of varying smoothness; posteriors $\Pi_{n,c}(\cdot \,|\, Y_n)$.

Empirical Bayes:

- $\hat{c}_n$ some "estimator".
- Plug-in posterior $\Pi_{n,\hat{c}_n}(\cdot \,|\, Y_n)$.

Bayesian adaptation

Family of priors Π_c of varying smoothness; posteriors $\Pi_{n,c}(\cdot|\, Y_n)$.

Empirical Bayes:

- $\hat{c}_n$ some "estimator".
- Plug-in posterior $\Pi_{n,\hat{c}_n}(\cdot|\, Y_n)$.

Hierarchical Bayes:

- Full Bayes, with prior π on c.
- Posterior $\int \Pi_{n,c}(\cdot|\, Y_n)\, \pi_n(c|\, Y_n)\, dc$.

Both methods (in particular Hierarchical Bayes) are known to give adaptive reconstructions in some generality:
if the true function is smoother, then the reconstruction is better.

Bayesian adaptation

Family of priors Π_c of varying smoothness; posteriors $\Pi_{n,c}(\cdot \,|\, Y_n)$.

Empirical Bayes:

- $\hat{c}_n$ some "estimator".
- Plug-in posterior $\Pi_{n,\hat{c}_n}(\cdot \,|\, Y_n)$.

Hierarchical Bayes:

- Full Bayes, with prior π on c.
- Posterior $\int \Pi_{n,c}(\cdot \,|\, Y_n)\, \pi_n(c \,|\, Y_n)\, dc$.

Both methods (in particular Hierarchical Bayes) are known to give adaptive reconstructions in some generality:
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

$C_n(Y_n)$ is an honest confidence set over a model Θ if

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

Honesty and impossibility of adaptation

$C_n(Y_n)$ is an honest confidence set over a model Θ if

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

THEOREM [Low, Cai & Low, Lepski, Juditzky et al., Robins&vdV, Bull& Nickl]

For any $\Theta_1 \subset \Theta$ the diameter of honest $C_n(Y_n)$ cannot have smaller order, uniformly over Θ_1, than:

(a) any $\varepsilon_n \to 0$ such that, for any T_n and some $\beta > 0.05$,

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

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

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

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.

Linear Gaussian inverse problems

Linear Gaussian inverse problems

Represent functions θ 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 κ_i.

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

Equivalent to observing Y for

$$dY_t = (K\theta)(t)\, dt + n^{-1/2} dW_t.$$

Linear Gaussian inverse problems

Represent functions θ 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

Represent functions θ 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 some A and S.

$$A = \Lambda K^T \left(\tfrac{1}{n} I + K \Lambda K^T\right)^{-1} \text{ and } \qquad S = \Lambda - A(n^{-1} I + K \Lambda K^T) A^T.$$

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: for $u\colon [0,1] \times [0,1] \to \mathbb{R}$,

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

We observe a noisy version of the final curve: for Z white noise:

$$Y_n = K\theta + n^{-1/2} Z.$$

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

$$\kappa_i = e^{-i^2 \pi^2} \qquad e_i = \sqrt{2} \sin(i \pi x),$$

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

Example: reconstruct derivative

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

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

We observe $\big(Y_n(x)\colon x \in [0,1]\big)$, for W Brownian motion,

$$dY_n(x) = K\theta(x)\, dx + \frac{1}{\sqrt{n}} dW(x), \qquad x \in [0,1].$$

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

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

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

Sobolev models and priors — smooth functions

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

$$S^\beta = \left\{(\theta_1, \theta_2, \ldots): \sum_i i^{2\beta}\theta_i^2 < \infty\right\}.$$

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.

Hyperrectangles — smooth functions

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

$$\Theta^\beta = \left\{(\theta_1, \theta_2, \ldots): \sup_i i^{2\beta+1}\theta_i^2 < \infty\right\}.$$

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.

– Fixed priors

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

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

THEOREM [Zhao, Knapik et al.]

For an α-smooth prior and β-smooth truth θ_0, and $r_{n,\alpha,\beta} = n^{-(\alpha\wedge\beta)/(2\alpha+2p+1)}$,

$$\Pi_n\big(\theta\colon \|\theta-\theta_0\|_2 \gtrsim r_{n,\alpha,\beta}|\,Y_n\big) \to 0, \qquad \text{a.s. } \big[Y_n \sim N_\infty(K\theta_0, n^{-1}I)\big].$$

In other words, the posterior rate of contraction is $r_{n,\alpha,\beta}$.

This is as usual:

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

Linear Gaussian inverse problem — credible balls

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 $\lambda_i \sim i^{-2\alpha-1}$.

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

CREDIBLE SET: $\mathrm{ball}(A_\alpha Y_n, r_\alpha)$ of posterior mass 0.95.

Linear Gaussian inverse problem — credible balls

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 $\lambda_i \sim i^{-2\alpha-1}$.

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

CREDIBLE SET: $\mathrm{ball}(A_\alpha Y_n, r_\alpha)$ of posterior mass 0.95.

THEOREM

For α-smooth prior and β-smooth truth:

- If $\alpha < \beta$, then asymptotic coverage is 1 (uniformly).
- If $\alpha = \beta$, then asymptotic coverage is $c \in (0, 1)$ for some $\theta_0 \in S^\beta$.
- 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 of correct size.

Linear Gaussian inverse problem — credible intervals

A 95% credible interval for a functional $\psi(\theta) = \sum_i l_i \theta_i$ is a central interval in the marginal posterior of posterior mass 0.95.

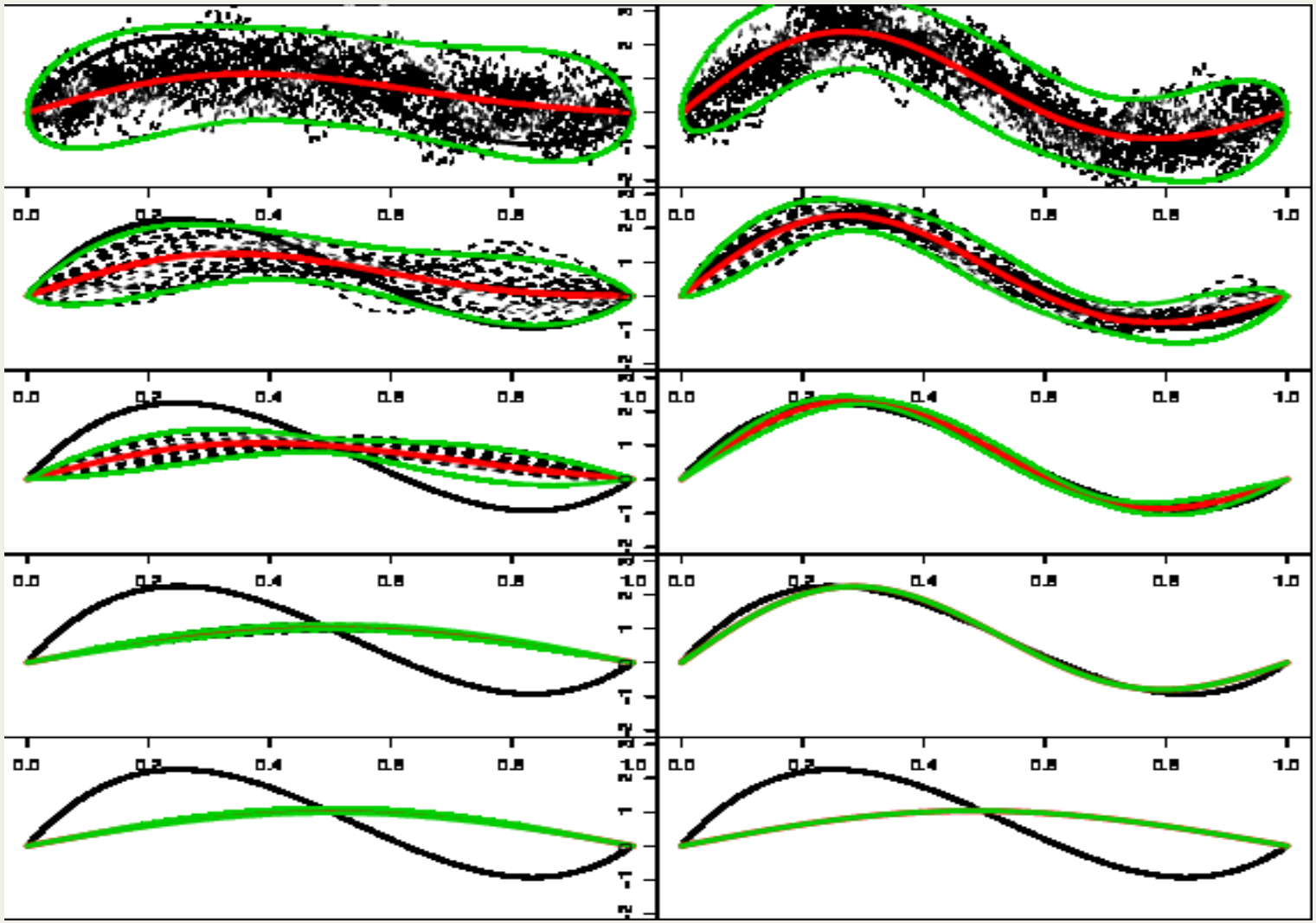
THEOREM

For α-smooth prior and β-smooth truth, and l regularly varying of order $-q$.

- If $q < p$ and $\alpha < \beta - 1/2$, then asymptotic coverage is in $(0.95, 1)$.
- If $q \geq p$ and $\alpha < \beta - 1/2 + (q - p)$, then asymptotic coverage is correct.
- If $\alpha > \beta - 1/2 + (q - p)^+$, then for some $\theta \in S^\beta$ asymptotic coverage is 0.

Correct coverage iff the Bernstein-von Mises theorem holds. Then rate of contraction is $L(n)/\sqrt{n}$ for a slowly varying function L.
If the prior undersmoothes, then credible interval is OK.

Example: heat equation (n=10000 and n=100 000 000)



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

– Adapting priors

Linear Gaussian inverse problem — empirical and hierarchical Bayes

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

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

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

MARGINAL MODEL: $Y_n|\,\alpha \sim N_\infty(0, K\Lambda_\alpha K^T + n^{-1}I)$.

Empirical Bayes method: plug in the MLE $\hat\alpha$ of marginal model:

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

Linear Gaussian inverse problem — empirical and hierarchical Bayes

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

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

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

MARGINAL MODEL: $Y_n|\,\alpha \sim N_\infty(0, K\Lambda_\alpha K^T + n^{-1}I)$.

Empirical Bayes method: plug in the MLE $\hat\alpha$ of marginal model:

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

Hierarchical Bayes method:

PRIOR: $\alpha \sim \pi$ (with "correct" tails, e.g. inverse gamma).

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

Linear Gaussian inverse problem — contraction rates

THEOREM [Knapik et al., 2012]

For any $\beta > 0$ the (plugged-in) empirical or hierarchical posterior distribution $\theta|\, Y_n$ contracts

- at (nearly) the rate $n^{-\beta/(2\beta+2p+1)}$ if $\theta_0 \in S^\beta$ or if $\theta_0 \in \Theta^\beta$.
- at (nearly) the rate $n^{-1/2}$ if $\theta_0 \in S^\infty = \{\theta\colon \sum_i e^i \theta_i^2 < \infty\}$.

Difficulty of proof: $\hat{\alpha}_n$ does not necessarily settle down.

Credible balls — counter example

THEOREM

For $n_1 \geq 2$ and $n_j \geq n_{j-1}^4$ for every j, $\beta > 0$ and suitable $M > 0$, define $\theta = (\theta_1, \theta_2, ...)$ by

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

Then $\theta \in S^\beta$, but, for every $L_n \ll n^\delta$,

$$\liminf P_\theta\big(\theta \in \text{ball}(A_{\hat{\alpha}} Y_n, L_n r_{\hat{\alpha}})\big) = 0.$$

Credible balls — counter example

THEOREM

For $n_1 \geq 2$ and $n_j \geq n_{j-1}^4$ for every j, $\beta > 0$ and suitable $M > 0$, define $\theta = (\theta_1, \theta_2, ...)$ by

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

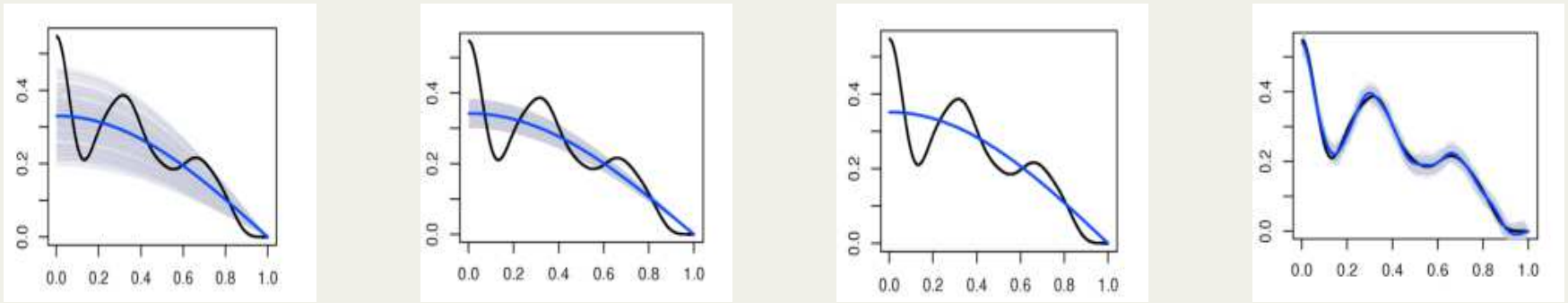
Then $\theta \in S^\beta$, but, for every $L_n \ll n^\delta$,

$$\liminf P_\theta\big(\theta \in \text{ball}(A_{\hat{\alpha}} Y_n, L_n r_{\hat{\alpha}})\big) = 0.$$

- Data allows inference on $\theta_1, \ldots, \theta_N$ for an *effective dimension* $N = N_n$.
- Trouble if $\theta_1, \ldots, \theta_N$ does not resemble $\theta_1, \theta_2, \ldots$.

Example: reconstructing a derivative

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



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*. For some (most?) truths the results are good.

[Szabo, vdV, van Zanten, 2016.]

[Not "asymptotical": for still bigger n it can become good and bad again!]

Self-similarity [after Giné+Nickl, Hoffmann+Nickl, Bull, 2010-12]

DEFINITION

A parameter *self-similar* of order β if

$$\sup_i i^{2\beta+1}\theta_i^2 \leq M,$$

and for fixed $\varepsilon > 0$, N_0, and $\rho \geq 2$,

$$\sum_{i=N}^{\rho N} \theta_i^2 \geq \varepsilon M N^{-2\beta}, \quad \forall N \geq N_0.$$

Interpretation:
θ hits ε times maximal possible *energy level* at *any frequency* N.

Polished tail sequences

DEFINITION

A parameter $\theta \in \ell^2$ satisfies the *polished tail condition* if, for fixed L_0, N_0 and $\rho \geq 2$,

$$\sum_{i=N}^{\infty} \theta_i^2 \leq L_0 \sum_{i=N}^{\rho N} \theta_i^2, \qquad \forall N \geq N_0.$$

Same interpretation, but self-referencing.

Credible sets are honest over polished tail sequences

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

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

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

CREDIBLE SET: $\mathrm{ball}(A_\alpha Y_n, r_\alpha)$ of posterior mass 0.95.

EB: replace α by $\hat\alpha_n$.
HE: integrate out over $\alpha|\,Y_n$.

THEOREM

For large enough L the EB and HB balls are honest over the set of all polished tail sequences (for given (L_0, N_0, ρ)).

[We must restrict α to a compact. For unrestricted $\hat\alpha$ this is true with a slowly varying sequence $L_n \to \infty$.]

“Everything” is polished tail..

For the *topologist*:

THEOREM [Giné+Nickl, 2010]
Non self-similar sequences are meagre relative to a natural topology.

“Everything” is polished tail..

For the *topologist*:

THEOREM [Giné+Nickl, 2010]
Non self-similar sequences are meagre relative to a natural topology.

For the *minimax expert*:

THEOREM
By intersecting a model with the polished tail sequences the minimax risk decreases by at most

- a constant if the model is a hyperrectangle.
- a logarithmic factor if the model is a Sobolev ball.

“Everything” is polished tail..

For the *topologist*:

THEOREM [Giné+Nickl, 2010]
Non self-similar sequences are meagre relative to a natural topology.

For the *minimax expert*:

THEOREM
By intersecting a model with the polished tail sequences the minimax risk decreases by at most

- a constant if the model is a hyperrectangle.
- a logarithmic factor if the model is a Sobolev ball.

For the *Bayesian*:

THEOREM
For every $\alpha > 0$ the prior $\Pi_\alpha = N_\infty(0, \Lambda)$ with $\lambda_i \sim i^{-1-2\alpha}$ satisfies

$$\Pi_\alpha\big(\cup_{N_0}\{\theta\colon \theta \in \text{polished tail}(2^{2+2\alpha}, N_0, 2)\}\big) = 1.$$

Credible sets have optimal diameter

THEOREM

For all β in a compact, and $M > 0$,

$$\inf_{\theta_0 \in \Theta^\beta(M)} P_{\theta_0}\left(r_{\hat{\alpha}_n} \lesssim M^{\frac{1/2+p}{1+2\beta+2p}} n^{-\frac{\beta}{1+2\beta+2p}}\right) \to 1.$$

THEOREM

For all β in a compact, and $M > 0$,

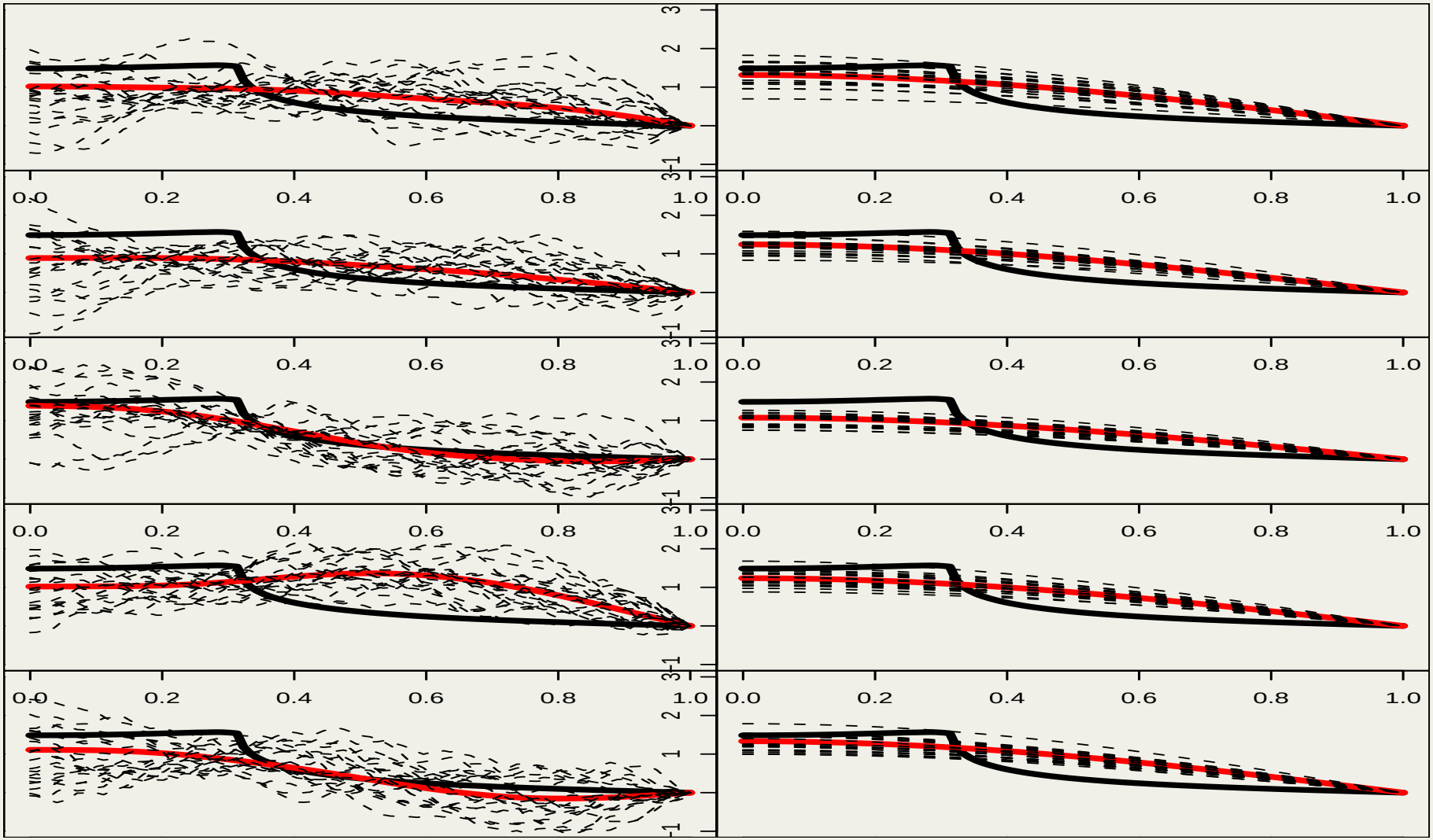
$$\inf_{\theta_0 \in S^\beta(M)} P_{\theta_0}\left(r_{\hat{\alpha}_n} \lesssim M^{\frac{1/2+p}{1+2\beta+2p}} n^{-\frac{\beta}{1+2\beta+2p}}\right) \to 1.$$

THEOREM

For $\hat{\alpha}_n$ restricted to $[0, K\sqrt{\log n}]$,

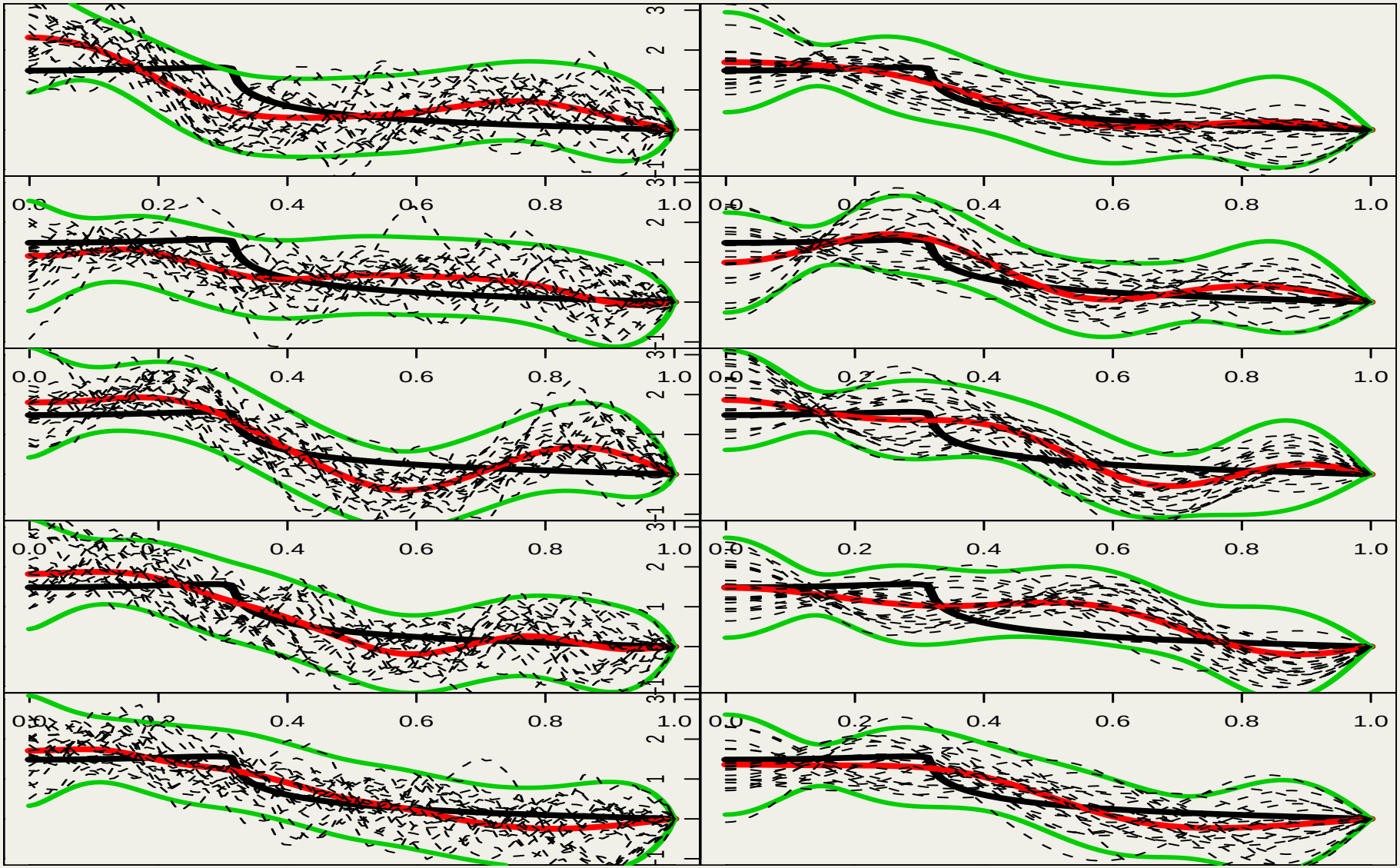
$$\inf_{\theta_0 \in S^\infty(M)} P_{\theta_0}\left(r^2_{\hat{\alpha}_n} \leq e^{(1/2+p)\sqrt{\log n}} \log\log n\, n^{-1}\right) \to 1.$$

Example: reconstruct derivative (n=100)



True θ_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=1000)



True θ_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).

Nonparametric regression

Nonparametric regression

Parameter $\theta\colon \mathcal{X} \to \mathbb{R}$; *design points* $x_{1,n}, \ldots, x_{n,n}$.

DATA: $Y_n | \, \theta \sim N_n(\vec{\theta}_n, I)$, for $\vec{\theta}_n := \big(\theta(x_{1,n}), \ldots, \theta(x_{n,n})\big)^T$.

PRIOR: $\theta \sim \sqrt{c}\, W$, for a Gaussian process W.

POSTERIOR: $\vec{\theta}_n | \, Y_n \sim N_n\big(\hat{\theta}_{n,c}, I - \Sigma_{n,c}^{-1}\big)$.

$$\hat{\theta}_{n,c} = (I - \Sigma_{n,c}^{-1})\vec{Y}_n,$$
$$\Sigma_{n,c} = I + cU_n,$$
$$U_n = \mathrm{cov}(\vec{W}_n).$$

EXAMPLE

- scaled Brownian motion
- discrete Laplacian $(n^2 L)^{-\alpha}\vec{W}_n \sim N_n(0, I)$, for $Lf(i) = \sum_{j:j\sim i}\big[f(j) - f(i)\big]$. [Kirichenko & van Zanten, 2015.]
- Brownian sheet
- eigenfunctions as Brownian sheet but “Sobolev eigenvalues”.

Empirical Bayes and hierarchical Bayes

DATA: $Y_n | \, \theta, c \sim N_n(\vec{\theta}_n, I)$.

PRIOR: $\theta | \, c \sim \sqrt{c}\, W$, for a Gaussian process W.

POSTERIOR: $\vec{\theta}_n | \, Y_n, c \sim N_n\big(\hat{\theta}_{n,c}, I - \Sigma_{n,c}^{-1}\big)$.

- Risk-based Empirical Bayes method [Wahba, 1975]: plug in:

$$\hat{c}_n = \operatorname*{argmin}_c \underbrace{\Big[\operatorname{tr}\big((I - \Sigma_{n,c}^{-1})^2\big) - \operatorname{tr}(\Sigma_{n,c}^{-2}) + \vec{Y}_n^T \Sigma_{n,c}^{-2} \vec{Y}_n\Big]}_{\text{unbiased estimate of } \mathrm{E}_\theta \|\hat{\theta}_{n,c} - \vec{\theta}_n\|^2}.$$

Empirical Bayes and hierarchical Bayes

DATA: $Y_n | \, \theta, c \sim N_n(\vec{\theta}_n, I)$.

PRIOR: $\theta | \, c \sim \sqrt{c}\, W$, for a Gaussian process W.

POSTERIOR: $\vec{\theta}_n | \, Y_n, c \sim N_n\big(\hat{\theta}_{n,c}, I - \Sigma_{n,c}^{-1}\big)$.

- Risk-based Empirical Bayes method [Wahba, 1975]: plug in:

$$\hat{c}_n = \operatorname*{argmin}_c \underbrace{\Big[\operatorname{tr}\big((I - \Sigma_{n,c}^{-1})^2\big) - \operatorname{tr}(\Sigma_{n,c}^{-2}) + \vec{Y}_n^T \Sigma_{n,c}^{-2} \vec{Y}_n\Big]}_{\text{unbiased estimate of } \mathrm{E}_\theta \|\hat{\theta}_{n,c} - \vec{\theta}_n\|^2}.$$

MARGINAL MODEL: $Y_n | \, c \sim N_n(0, \Sigma_{n,c}), \qquad \Sigma_{n,c} = I + cU_n$.

- Likelihood-based Empirical Bayes method: plug in MLE:

$$\hat{c}_n = \operatorname*{argmin}_c \Big[\log\det \Sigma_{n,c} + \vec{Y}_n^T \Sigma_{n,c}^{-1} \vec{Y}_n\Big].$$

Empirical Bayes and hierarchical Bayes

DATA: $Y_n|\,\theta, c \sim N_n(\vec{\theta}_n, I)$.

PRIOR: $\theta|\,c \sim \sqrt{c}\,W$, for a Gaussian process W.

POSTERIOR: $\vec{\theta}_n|\,Y_n, c \sim N_n\big(\hat{\theta}_{n,c}, I - \Sigma_{n,c}^{-1}\big)$.

- Risk-based Empirical Bayes method [Wahba, 1975]: plug in:

$$\hat{c}_n = \operatorname*{argmin}_c \underbrace{\Big[\operatorname{tr}\big((I - \Sigma_{n,c}^{-1})^2\big) - \operatorname{tr}(\Sigma_{n,c}^{-2}) + \vec{Y}_n^T \Sigma_{n,c}^{-2} \vec{Y}_n\Big]}_{\text{unbiased estimate of } \mathrm{E}_\theta \|\hat{\theta}_{n,c} - \vec{\theta}_n\|^2}.$$

MARGINAL MODEL: $Y_n|\,c \sim N_n(0, \Sigma_{n,c}), \qquad \Sigma_{n,c} = I + cU_n$.

- Likelihood-based Empirical Bayes method: plug in MLE:

$$\hat{c}_n = \operatorname*{argmin}_c \Big[\log\det \Sigma_{n,c} + \vec{Y}_n^T \Sigma_{n,c}^{-1} \vec{Y}_n\Big].$$

- Hierarchical Bayes method:

PRIOR: $c^{-1} \sim \Gamma(a, b)$.

POSTERIOR: $\vec{\theta}_n|\,Y_n \sim \int N_n\big(\hat{\theta}_{n,c}, I - \Sigma_{n,c}^{-1}\big)\,\pi_n(c|\,Y_n)\,dc$.

Credible balls are honest over polished tail functions

DATA: $Y_n | \, \theta, c \sim N_n(\vec{\theta}_n, I)$.

PRIOR: $\theta | \, c \sim \sqrt{c}\, W$, for a Gaussian process W.

POSTERIOR: $\vec{\theta}_n | \, Y_n, c \sim N_n\big(\hat{\theta}_{n,c}, I - \Sigma_{n,c}^{-1}\big)$.

CREDIBLE BALL: for $s_n^2(c) = \mathrm{E}\big[\|\vec{\theta}_n - \hat{\theta}_{n,c}\|^2 | \, Y_n, c\big]$ and $\Pi_n(\hat{c}_{1,n} < c < \hat{c}_{2,n} | \, Y_n) = 1 - \eta$,

$$\hat{C}_{n,\eta,M} = \big\{\theta \colon \|\vec{\theta}_n - \hat{\theta}_{n,\hat{c}_n}\| < M s_n(\hat{c}_n)\big\}, \qquad (EB),$$

or

$$\hat{C}_{n,\eta,M} = \bigcup_{\hat{c}_{1,n} < c < \hat{c}_{2,n}} \big\{\theta \colon \|\vec{\theta}_n - \hat{\theta}_{n,c}\| < M s_n(c)\big\}, \qquad (HB).$$

THEOREM

For large M, uniformly in polished tail functions θ,

$$P_\theta(\theta \in \hat{C}_{n,\eta,M}) \to 1.$$

Credible intervals are honest over polished tail functions

DATA: $Y_n|\,\theta, c \sim N_n(\vec{\theta}_n, I)$.

PRIOR: $\theta \sim \sqrt{c}\,W$, for a Gaussian process W.

POSTERIOR: $\vec{\theta}_n|\,Y_n, c \sim N_n\big(\hat{\theta}_{n,c}, I - \Sigma_{n,c}^{-1}\big)$.

CREDIBLE INTERVALS: for $s_n^2(c,x) = \mathrm{E}\big[|\theta(x) - \hat{\theta}_{n,c}(x)|^2|\,\vec{Y}_n, c\big]$,

$$\hat{C}_{n,\eta,M}(x) = \big\{\theta\colon |\theta(x) - \hat{\theta}_{n,\hat{c}_n}(x)| < M s_n(\hat{c}_n, x)\big\}, \qquad (EB),$$

or

$$\hat{C}_{n,\eta,M}(x) = \bigcup_{\hat{c}_{1,n}<c<\hat{c}_{2,n}} \big\{\theta\colon |\theta(x) - \hat{\theta}_{n,c}(x)| < M s_n(c, x)\big\}, \qquad (HB).$$

THEOREM

If $x_{j,n}$ "uniformly spread relative to the prior", then for large M and all $\gamma < 1$, uniformly in polished tail functions θ

$$P_\theta\Big(\frac{1}{n}\sum_{i=1}^{n} 1\big\{\theta \in \hat{C}_{n,\eta,M}(x_{i,n})\big\} \geq \gamma\Big) \to 1.$$

Credible bands are honest for Bayesians

DATA: $Y_n | \, \theta, c \sim N_n(\vec{\theta}_n, I)$.

PRIOR: $\theta \sim \sqrt{c}\, W$, for W Brownian motion.

POSTERIOR: $\vec{\theta}_n | \, Y_n, c \sim N_n\big(\hat{\theta}_{n,c}, I - \Sigma_{n,c}^{-1}\big)$.

CREDIBLE BAND: for $s_n^2(c,x) = \mathrm{E}\big[|\theta(x) - \hat{\theta}_{n,c}(x)|^2 \,|\, \vec{Y}_n, c\big]$,

$$\tilde{C}_{n,\eta,M} = \bigcap_x \big\{\theta\colon |\theta(x) - \hat{\theta}_{n,\hat{c}_n}(x)| < M\sqrt{\log n}\, s_n(\hat{c}_n, x)\big\}, \qquad (EB),$$

or

$$\tilde{C}_{n,\eta,M} = \bigcap_x \bigcup_{\hat{c}_{1,n} < c < \hat{c}_{2,n}} \big\{\theta\colon |\theta(x) - \hat{\theta}_{n,c}(x)| < M\sqrt{\log n}\, s_n(c, x)\big\}, \qquad (HB).$$

THEOREM

For almost any realization θ from a Gaussian process prior

$$P_\theta\Big(\theta \in \tilde{C}_{n,\eta,M}\Big) \to 1.$$

Credible bands can be honest for non-Bayesians

DATA: $Y_n | \, \theta, c \sim N_n(\vec{\theta}_n, I)$.

PRIOR: $\theta \sim \sqrt{c}\, W$, for W Brownian motion.

POSTERIOR: $\vec{\theta}_n | \, Y_n, c \sim N_n\big(\hat{\theta}_{n,c}, I - \Sigma_{n,c}^{-1}\big)$.

CREDIBLE BAND: for $s_n^2(c,x) = \mathrm{E}\big[|\theta(x) - \hat{\theta}_{n,c}(x)|^2 \,|\, \vec{Y}_n, c\big]$,

$$\tilde{C}_{n,\eta,M} = \bigcap_x \big\{\theta\colon |\theta(x) - \hat{\theta}_{n,\hat{c}_n}(x)| < M\sqrt{\log n}\, s_n(\hat{c}_n, x)\big\}, \qquad (EB),$$

or

$$\tilde{C}_{n,\eta,M} = \bigcap_x \bigcup_{\hat{c}_{1,n} < c < \hat{c}_{2,n}} \big\{\theta\colon |\theta(x) - \hat{\theta}_{n,c}(x)| < M\sqrt{\log n}\, s_n(c, x)\big\}, \qquad (HB).$$

THEOREM

If θ is self-similar and Hölder of the same order, then

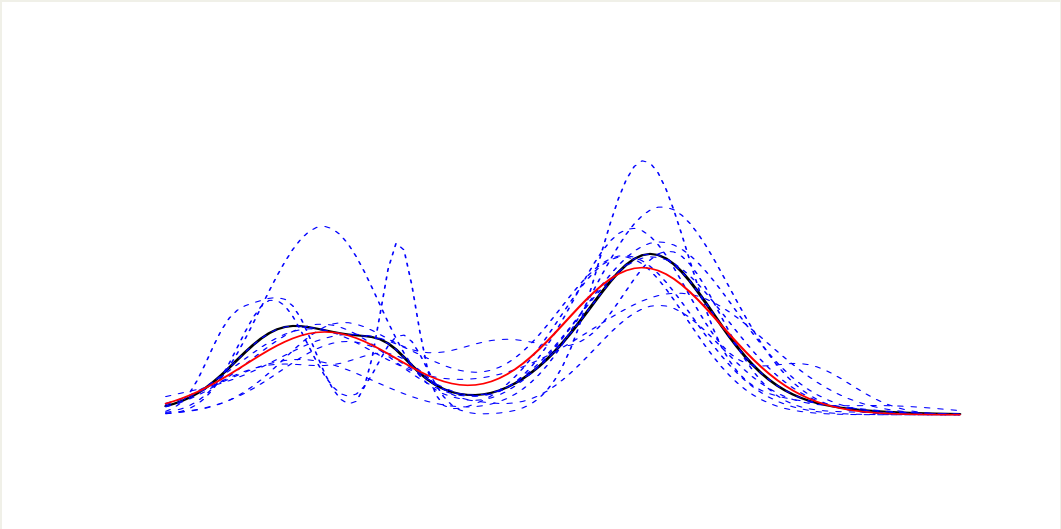
$$P_\theta\Big(\theta \in \tilde{C}_{n,\eta,M}\Big) \to 1.$$

Dirichlet process mixtures

Dirichlet normal mixtures

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

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



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

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

Valid uncertainty quantification???

Conclusion

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 mislead as they trust their priors. In nonparametrics they should not, as the fine details of a prior are not obvious.