Confidence in Credible Sets?

Aad van der Vaart
Universiteit Leiden, Netherlands

Oberwolfach, March 2014
Contents

Credible sets
Bernstein-von Mises theorem
Nonparametric credible sets
Linear Gaussian inverse problems
Fixed priors
Adapting priors
Extensions (if time)
Conclusions
Co-authors

Botond Szabo
(Eindhoven)

Harry van Zanten
(Amsterdam)
Credible sets
Credible sets

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

Gives posterior distribution $\theta \mid Y_n \sim \Pi_n(\cdot \mid Y_n)$ as usual.
Credible sets

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

Gives posterior distribution $\theta \mid Y_n \sim \Pi_n(\cdot \mid Y_n)$ as usual.

Two uses:
- recovery, e.g. by mode, or mean.
- expression of uncertainty, e.g. by a credible set.
Credible sets

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

Gives **posterior distribution** $\theta|Y_n \sim \Pi_n(\cdot|Y_n)$ as usual.

Two uses:
- **recovery**, e.g. by mode, or mean.
- **expression of uncertainty**, e.g. by a credible set.

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

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

[Plotting method for credible balls: simulate 1000 realizations from the posterior, plot the 95% that are closest to the posterior mean.]
Example: geology

From presentation of Sambridge, Lorentz Centre Leiden, 2011.
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, *Geophys. J. Int.* 178, 2009, 1411–1436.
Example: logistic regression

Bayesian model:

\[
\begin{align*}
\theta & \sim \text{scaled integrated Brownian motion}, \\
(X_1, Y_1), \ldots, (X_n, Y_n) \mid \theta & \sim \text{i.i.d.: } P(Y_i = 1 \mid X_i = x) = 1 / (1 + e^{-\theta(x)}).
\end{align*}
\]

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.
Do credible sets correctly quantify *remaining uncertainty*?
Do credible sets correctly quantify *remaining uncertainty*?

Does

\[ \Pi_n(\theta \in C_n(Y_n) | Y_n) = 0.95. \]

imply

\[ P_{\theta_0}(\theta_0 \in C_n(Y_n)) = 0.95? \]
Do credible sets correctly quantify remaining uncertainty?

Does the spread in the posterior give the correct order of the discrepancy between $\theta_0$ and the posterior mean?
Bernstein-von Mises theorem
THEOREM

If $X_1, \ldots, X_n (iid) \sim p_\theta$, for $\{p_\theta : \theta \in \mathbb{R}^k\}$ a smoothly parametrized model, then, under mild conditions,

$$\left\| \Pi_n(\theta \in \cdot | X_1, \ldots, X_n) - N(\hat{\theta}_n, (nI_\theta)^{-1}) \right\|_{TV} \to 0,$$

for $\hat{\theta}_n$ random variables with $\sqrt{n}(\hat{\theta}_n - \theta) \rightsquigarrow N(0, I_\theta^{-1})$.

Consequently credible sets are asymptotically equivalent to sets of the form

$$\{ \theta \in \mathbb{R}^k : \sqrt{n}I_\theta^{1/2}(\theta - \hat{\theta}_n) \in C \},$$

which are confidence sets.
THEOREM

If \( X_1, \ldots, X_n \) iid \( \sim p_\theta \), for \( \{p_\theta : \theta \in \mathbb{R}^k\} \) a smoothly parametrized model, then, under mild conditions,

\[
\left\| \Pi_n(\theta \in \cdot \mid X_1, \ldots, X_n) - N(\hat{\theta}_n, (nI_\theta)^{-1}) \right\|_{TV} \to 0,
\]

for \( \hat{\theta}_n \) random variables with \( \sqrt{n}(\hat{\theta}_n - \theta) \rightsquigarrow N(0, I_\theta^{-1}) \).

Consequently credible sets are asymptotically equivalent to sets of the form

\[
\left\{ \theta \in \mathbb{R}^k : \sqrt{n}I_\theta^{1/2}(\theta - \hat{\theta}_n) \in C \right\},
\]

which are confidence sets.

[Result fails under model misspecification.]
Semiparametric Bernstein-von Mises theorem

For the marginal posterior distribution of a smooth functional $\psi(\theta) \in \mathbb{R}$ of an infinite-dimensional $\theta$, there may be a Bernstein-von Mises theorem:

$$\left\| \Pi_n (\psi(\theta) \in \cdot | Y_n) - N (\hat{\psi}_n, s_n^2) \right\|_{TV} \to 0.$$ 

Credible sets for $\psi(\theta)$ are confidence sets if $(\hat{\psi}_n - \psi(\theta))/s_n \rightsquigarrow N(0, 1)$. 
Semiparametric Bernstein-von Mises theorem

For the marginal posterior distribution of a smooth functional $\psi(\theta) \in \mathbb{R}$ of an infinite-dimensional $\theta$, there may be a Bernstein-von Mises theorem:

$$\left\| \Pi_n(\psi(\theta) \in \cdot | Y_n) - N(\hat{\psi}_n, s_n^2) \right\|_{TV} \to 0.$$

Credible sets for $\psi(\theta)$ are confidence sets if $(\hat{\psi}_n - \psi(\theta))/s_n \sim N(0, 1)$.

The essence is that estimation of $\psi(\theta)$ should not involve a bias-variance trade-off: the bias should be negligible.

[Lo, Kim, Castillo, Kleijn, Rousseau+Rivoirard, Knapik et al.,...]

Nonparametric Bernstein-von Mises theorem

A full parameter $\theta$ may be identified through a set $\Psi$ of smooth functionals $\psi(\theta)$.

If there is a joint Bernstein-von Mises theorem

$$\sup_{B} \left| \Pi_n(\theta: \sqrt{n}(\psi(\theta) - \hat{\psi}_n)_{\psi \in \Psi} \in B \mid Y_n) - P((Z(\psi)_{\psi \in \Psi} \in B) \right| \to 0,$$

then this may be used to get valid credible sets.
Weak nonparametric Bernstein-von Mises theorem
[Castillo+Nickl, 2013]

View $\theta$ and observation $Y_n$ in the white noise model $dY_n(t) = \theta(t) \, dt + dW_t$ as elements of logarithmic Sobolev space

$$H = \left\{ (\theta_1, \theta_2, \ldots) : \sum_{i=1}^{\infty} \frac{1}{i(\log i)^{1+\delta}} \theta_i^2 < \infty \right\}.$$  

**THEOREM**

Under mild conditions on the prior,

$$d_{BL(H)} \left( \Pi_n \left( \sqrt{n}(\theta - Y_n) \in \cdot \mid Y_n \right), \mathcal{N} \right) \to 0.$$  

This can be used to define valid credible sets for $\theta$.

To control their diameter these can be intersected under suitable strong a-priori assumptions.
Nonparametric Bernstein-von Mises theorem [Leahu, 2012]

**THEOREM**

For parameter $\theta(x) = \sum_i \theta_i e_i(x)$ the Bernstein-von Mises theorem holds in $\ell_2$, but the prior on $(\theta_1, \theta_2, \ldots)$ must not concentrate on $\ell_2$.

E.g. $\theta_i \sim N(0, v_i)$ with $v_i \to \infty$ fast.
Nonparametric credible sets
History: credible sets are a disaster! [Cox, 1993]

AN ANALYSIS OF BAYESIAN INFERENCE FOR NONPARAMETRIC REGRESSION

BY DENNIS D. COX

Rice University

The observation model, $y_i = \beta(t_{i}) + \epsilon_i$, $1 \leq i \leq n$, is considered, where the $\epsilon_i$’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 for the prior distribution of $\beta$ by the solution of a high-order stochastic differential equation. The univariate normal $N(\beta, \sigma^2)$ is approximated, where $\beta$ is the posterior expectation of $\beta$. Asymptotic posterior and sampling distribution approximation are given for $|\beta|$ when $|\beta|$ is one of a class of smooth functions in 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 $\alpha > 0$ as $\alpha \to 0$ with a prior probability of 1. A related nonparametric 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

$$Y_i = \beta(t_{i}) + \epsilon_i, \quad 1 \leq i \leq n,$$

where $t_{i} = i/n$, $\beta : [0, 1] \rightarrow R$ is an unknown smooth function, and $\epsilon_1, \epsilon_2, \ldots$ are i.i.d. random errors with mean 0 and known variance $\sigma^2 < \infty$. The $t_i$’s are modeled as $N(0, \sigma^2)$. A Gaussian prior for $\beta$ will now be specified. Let $m \geq 2$ and for some constants $a_0, \ldots, a_m$ with $a_m \neq 0$ let

$$L = \sum_{i=0}^{m} a_i \beta^{i}$$
"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.,"

\[
\lim \inf_{n \to \infty} P_{\theta_0} \left( C_n(Y_n) \ni \theta_0 \right) = 0, \quad \text{for } \Pi\text{-a.e. } \theta_0.
\]

Cox's credible set \( C_n(Y_n) \) is an \( L_2 \)-ball of posterior mass 95 % around the posterior mean. The prior is multiply integrated Brownian motion.
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 \( Y(t_i) = g(t_i) + \epsilon_i, \quad i = 1, 2, \ldots, n \), where \( g(t_i) \) is a smooth function and the \( \{\epsilon_i\} \) are independent \( N(0, \sigma^2) \) errors with \( \sigma^2 \) unknown. The cross-validated smoothing spline can be used to estimate \( g \) non-parametrically from observations on \( Y(t_i), \quad i = 1, 2, \ldots, n \), 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 \( g(t_i), \quad i = 1, 2, \ldots, n \). 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

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

where \( \epsilon = (\epsilon_1, \ldots, \epsilon_n)^T \sim N(0, \sigma^2 I_n) \), \( \sigma^2 \) is unknown and \( g(\cdot) \) is a fixed but unknown function with \( p-1 \) continuous derivatives and \( \int_0^1 (g^{(p)}(t))^2 dt < \infty \). The smoothing spline estimate of \( g \) given \( Y(t_i) = y_i, \quad i = 1, 2, \ldots, n \), which we will call \( g_{n, \lambda} \), is the miniser of

\[ n^{-1} \sum_{i=1}^n (g(t_i) - y_i)^2 + \lambda \int_0^1 (g^{(p)}(t))^2 dt \]
Wahba's credible set $C_{T_1} (Y_{T_1})$ is a pointwise band of marginal posterior masses 95% around the posterior mean. The prior is multiply integrated Brownian motion. The numbers are the percentages of $x$-values at which the band contains the truth.
Example: heat equation

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

for \( u: [0, 1] \times [0, 1] \rightarrow \mathbb{R} \),

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

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

\[
Y_n = K\theta + \frac{1}{\sqrt{n}} Z.
\]

We put a Gaussian proces prior on \( \theta \).
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).

Left: $n = 10^4$; right: $n = 10^8$. Top to bottom: prior of increasing smoothness.
What is happening?

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

**Recovery:** the prior influences the rate of recovery by the posterior, but “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.
What is happening?

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

Recovery: the prior influences the rate of recovery by the posterior, but “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).

Cox oversmoothes, Wahba undersmoothes
where $1 - \alpha$ depends on $m$ as indicated in Table 1 (see Corollary 3.4).

The procedure analyzed here differs from that advocated by Wahba (1983) and Wecker and Ansley (1983) in the following ways:

1. These authors use a model wherein $\beta = \beta_0 + \beta_1$ and $\beta_1$ is given a proper prior of the type above but $\beta_0$ is finite dimensional and given a Lebesgue prior.
2. We have assumed $\sigma^2$ is known.
3. These authors assume $\beta = b\beta_0$ where $\beta_0$ has a given prior and $b > 0$ is an unknown scale factor which is estimated.
4. Wahba (1983) considers the true function $\beta$ to be fixed and smoother than one generated by the Gaussian prior.

I conjecture that 1, 2 and 4 do not change the negative result (1.3). See Cox (1989). Concerning point 3, (1.3) depends on the law of the iterated logarithm fluctuations of the bias $E[\hat{\beta}_n - \beta\beta]$ about its mean (Lemma 3.2). Such fluctuations undoubtedly impact the smoothing parameter estimation procedure of Wahba (1983), known as generalized cross validation [see also Craven and Wahba (1979) and Speckman (1983)], so (1.3) may not hold when $b$ is estimated.

In Section 3 we consider a more general Gaussian prior on an abstract space. One

From Cox, 1993
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.,

\[
\lim \inf_{n \to \infty} P_{\theta_0} \left( \text{ball} \left( A_\alpha Y_n, r_\alpha \right) \ni \theta_0 \right) = 0, \quad \text{for } \Pi\text{-a.e. } \theta_0.
\]
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.,

\[
\liminf_{n \to \infty} P_{\theta_0}\left( \text{ball}\left( A_\alpha Y_n, (\log n) r_\alpha \right) \ni \theta_0 \right) = 1, \quad \text{for } \Pi\text{-a.e. } \theta_0.
\]

Slightly enlarging the ball gives full coverage.
In practice the prior smoothness is determined by the data.

Empirical Bayes:

- Take a family of priors $\Pi_\alpha$ of varying smoothness $\alpha > 0$.
- Let $\hat{\alpha}_n$ be the MLE in the model $Y_n \mid (\theta, \alpha) \sim p_n(\cdot \mid \theta)$ and $\theta \mid \alpha \sim \Pi_\alpha$.
- Use the plug-in posterior $\theta \mid (Y_n, \alpha)$, with $\alpha = \hat{\alpha}_n$.

$$\hat{\alpha}_n = \arg\max_{\alpha} \int p_n(Y_n \mid \theta) d\Pi_\alpha(\theta).$$
Adaptation

In practice the prior smoothness is determined by the data.

**Empirical Bayes:**
- Take a family of priors $\Pi_\alpha$ of varying smoothness $\alpha > 0$.
- Let $\hat{\alpha}_n$ be the MLE in the model $Y_n \mid (\theta, \alpha) \sim p_n(\cdot \mid \theta)$ and $\theta \mid \alpha \sim \Pi_\alpha$.
- Use the plug-in posterior $\theta \mid (Y_n, \alpha)$, with $\alpha = \hat{\alpha}_n$.

\[
\hat{\alpha}_n = \arg\max_\alpha \int p_n(Y_n \mid \theta) d\Pi_\alpha(\theta).
\]

**Hierarchical Bayes:**
- Full Bayes, with prior $\pi$ on $\alpha$.

\[
\theta \sim \int \Pi_\alpha d\pi(\alpha).
\]
Adaptive credible sets

By general principles honest, adaptive confidence sets do not exist.

Thus adaptive credible sets must be misleading for some truths.
Adaptive credible sets

By general principles honest, adaptive confidence sets do not exist.

Thus adaptive credible sets must be misleading for some truths.

\[ n = 10^3 \quad n = 10^4 \quad n = 10^6 \quad n = 10^8 \]

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


The pictures show an “inconvenient” truth. For some (most?) truths the results are good.

[Not “asymptotical”: for still bigger \( n \) it can become bad again!]
Confidence in Credible Sets?

Aad van der Vaart
Universiteit Leiden, Netherlands

Oberwolfach, March 2014
Credible sets

• prior model $\theta \sim \Pi_n$ for parameter $\theta$,
• likelihood $Y_n \mid \theta \sim p_n(y \mid \theta)$ for the data,

Gives posterior distribution $\theta \mid Y_n \sim \Pi_n(\cdot \mid Y_n)$ as usual.

Two uses:
• recovery, e.g. by mode, or mean.
• expression of uncertainty, e.g. by a credible set.

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

$$\Pi_n(\theta \in C_n(Y_n) \mid Y_n) = 0.95.$$
Linear Gaussian inverse problems
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 \( \theta_i \sim N(0, \lambda_i) \).
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: $Y_n \mid \theta \sim N_{\infty}(K\theta, n^{-1}I)$ 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:** $\theta | Y_n \sim N_{\infty}(AY_n, S)$, for some $A$ and $S$.

$$A = \Lambda K^T \left( \frac{1}{n} I + K \Lambda K^T \right)^{-1} \quad \text{and} \quad S = \Lambda - A(n^{-1}I + K \Lambda K^T)A^T.$$
Example: heat equation

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

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

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

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

$$(i = 1, 2, \ldots).$$
Example: reconstruct derivative

The Volterra operator $K: L^2[0, 1] \rightarrow L^2[0, 1]$ is given by

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

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

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

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

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

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

\[
\lambda_i \sim \frac{1}{i^{2\alpha+1}}.
\]

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

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

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

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

**POSTERIOR:** \( \theta \mid Y_n \sim N_\infty(A_\alpha Y_n, S_\alpha). \)

**THEOREM** [Zhao, Knapik et al.]

For an \( \alpha \)-smooth prior and \( \beta \)-smooth truth \( \theta_0 \), and

\[
r_{n,\alpha,\beta} = n^{-(\alpha \wedge \beta)/(2\alpha+2p+1)},
\]

\[
\Pi_n(\theta: \|\theta - \theta_0\|_2 \gtrsim r_{n,\alpha,\beta} \mid Y_n) \rightarrow 0, \quad \text{a.s.} \quad [Y_n \sim N_\infty(K\theta_0, n^{-1}I)].
\]

In other words, the **posterior rate of contraction** is \( r_{n,\alpha,\beta}. \)
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:** $\theta | Y_n \sim N_\infty (A_\alpha Y_n, S_\alpha)$.

**THEOREM** [Zhao, Knapik et al.] For an $\alpha$-smooth prior and $\beta$-smooth truth $\theta_0$, and

\[ r_{n,\alpha,\beta} = n^{-(\alpha \wedge \beta)/(2\alpha + 2p + 1)}, \]

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

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 ($\beta$ and $\alpha$).
- 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: $\theta|Y_n \sim N_\infty(A_\alpha Y_n, S_\alpha)$.

CREDIBLE SET: $\text{ball}(A_\alpha Y_n, r_\alpha)$ 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:** \( \theta | Y_n \sim N_\infty(A_\alpha Y_n, S_\alpha) \).

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

**THEOREM**

For \( \alpha \)-smooth prior and \( \beta \)-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.

**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.
**Posterior credible intervals**

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

**THEOREM**

For $\alpha$-smooth prior and $\beta$-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. The rate of contraction is then $L(n)/\sqrt{n}$ 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 $\theta_0$ (black), posterior mean (red), 20 realizations from the posterior (dashed black), and posterior credible bands (green). Left: $n = 10^4$ and right: $n = 10^8$. Top to bottom: increasing prior smoothness.
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.]
Adapting priors
Bayesian adaptation

A given prior may be good for some regularity class, but very bad for another.

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

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

A given prior may be good for some regularity class, but very bad for another.

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

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

For recovery the second is known to work in some generality. The first is thought to be equivalent.
Bayesian adaptation

A given prior may be good for some regularity class, but very bad for another.

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

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

**EXAMPLES:**

- Gaussian process $x \mapsto W_{hx}$ with random time-scaling $h$.
- Series $x \mapsto \sum_{i=1}^{N} Z_i e_i(x)$ with random $Z_i$ and $N$.
- Mixture $x \mapsto \int \phi_h(x - z) dF(z)$ with random $F$ and bandwidth $h$.
- Series $x \mapsto \sum_{i=1}^{\infty} \tau Z_i e_i(x)$ with random $Z_i$ and random $\tau$. 
Bayesian adaptation

A given prior may be good for some regularity class, but very bad for another.

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

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

*How does this work for credible sets?*
Linear Gaussian inverse problem — empirical Bayes

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

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

Empirical Bayes method: plug in the MLE $\hat{\alpha}$ of marginal model:
Linear Gaussian inverse problem — empirical Bayes

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

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

**POSTERIOR:** \( \theta \mid Y_n \sim N_{\infty}(A_\alpha Y_n, S_\alpha) \).

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

**Empirical Bayes method:** plug in the MLE \( \hat{\alpha} \) of marginal model:

\[
\hat{\alpha} = \arg\max_{\alpha} \sum_{i=1}^{\infty} \left( \frac{n^2}{i^{1+2\alpha+2p}} + nY_{n,i}^2 - \log\left(1 + \frac{n}{i^{1+2\alpha+2p}}\right) \right).
\]
Linear Gaussian inverse problem — 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 $\lambda_i = i^{-1-2\alpha}.$

**PRIOR:** $\alpha \sim \pi$ (with “correct” tails, e.g. gamma).

**POSTERIOR:** $\theta | Y_n \sim \int N_\infty ( A_\alpha Y_n, S_\alpha ) \pi_n ( \alpha | Y_n ) d\alpha.$
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 $\theta_0 \in S^\infty = \{ \theta : \sum_i e^{i\theta_i^2} < \infty \}$.

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

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

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

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

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

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

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

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

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

- 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

\[ n = 10^3 \quad n = 10^4 \quad n = 10^6 \quad n = 10^8 \]

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.

The pictures show an “inconvenient” truth. For some (most?) truths the results are good.

[Not “asymptotical”: for still bigger \( n \) it can become bad again!]
Honesty and impossibility of adaptation

A set $C_n(Y_n)$ is an honest confidence set over a model $\Theta$ if

$$P_{\theta_0}(C_n(Y_n) \ni \theta_0) \geq 0.95, \quad \text{for all } \theta_0 \in \Theta.$$
Honesty and impossibility of adaptation

A set $C_n(Y_n)$ is an honest confidence set over a model $\Theta$ if

$$P_{\theta_0}(C_n(Y_n) \ni \theta_0) \geq 0.95, \quad \text{for all } \theta_0 \in \Theta.$$  

THEOREM

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

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

$$\liminf_{n \to \infty} \sup_{\theta \in \Theta_1} P_{\theta}(\|T_n - \theta\| \geq \varepsilon_n) > \beta.$$

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

(a) typically gives the minimax rate of estimation for the model $\Theta_1$.
(b) is determined by the biggest model $\Theta$ rather than $\Theta_1$. 


A set $C_n(Y_n)$ is an honest confidence set over a model $\Theta_0$ if

$$P_{\theta_0}(C_n(Y_n) \ni \theta_0) \geq 0.95,$$

for all $\theta_0 \in \Theta_0$. 

**Honesty**
**Honesty**

A set $C_n(Y_n)$ is an honest confidence set over a model $\Theta_0$ if

$$P_{\theta_0}(C_n(Y_n) \ni \theta_0) \geq 0.95,$$

for all $\theta_0 \in \Theta_0$.

**THEOREM**

If $C_n(Y_n)$ is honest over $\bigcup_{\beta \geq \beta_0} S^\beta_1$, then its diameter is of the uniform order $O_P(n^{-\beta_0/(1/2 + 2\beta_0)})$ over $S^\beta$ for $\beta \geq 2\beta_0$. 

Honesty

A set $C_n(Y_n)$ is an honest confidence set over a model $\Theta_0$ if

$$P_{\theta_0} \left( C_n(Y_n) \ni \theta_0 \right) \geq 0.95, \quad \text{for all } \theta_0 \in \Theta_0.$$

**THEOREM**

If $C_n(Y_n)$ is honest over $\bigcup_{\beta \geq \beta_0} S_1^\beta$, then its diameter is of the uniform order $O_P \left( n^{-\beta_0/(1/2+2\beta_0)} \right)$ 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 (and many functionals) this is even more true.

Low, Cai+Low, Hoffmann+Lepski, Juditzky+Lambert-Lacroix, Robins+vdV.]
Estimation versus uncertainty quantification

Adaptive estimation:

- A more regular true function is easier to estimate.
- Estimators can be simultaneously optimal for multiple regularities.
- Bayesian methods can achieve this by a prior on the “bandwidth”.

Estimation versus uncertainty quantification

Adaptive estimation:
- A more regular true function is easier to estimate.
- Estimators can be simultaneously optimal for multiple regularities.
- Bayesian methods can achieve this by a prior on the “bandwidth”.

Uncertainty quantification:
- 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.
Estimation versus uncertainty quantification

Adaptive estimation:
- A more regular true function is easier to estimate.
- Estimators can be simultaneously optimal for multiple regularities.
- Bayesian methods can achieve this by a prior on the “bandwidth”.

Uncertainty quantification:
- 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.

SOLUTION 1: be honest; make conditional confidence statements only.
Estimation versus uncertainty quantification

Adaptive estimation:
- A more regular true function is easier to estimate.
- Estimators can be simultaneously optimal for multiple regularities.
- Bayesian methods can achieve this by a prior on the “bandwidth”.

Uncertainty quantification:
- 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.

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

SOLUTION 2: determine which $\theta$ cause the trouble; argue that these are implausible.
**Self-similarity** [after Giné+Nickl, Hoffmann+Nickl, Bull, 2010-12]

**DEFINITION**
A parameter $\theta \in \Theta^\beta (M)$ is *self-similar* if, for fixed $\varepsilon > 0$, $N_0$, and $\rho \geq 2$,

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

Interpretation:
$\theta$ hits $\varepsilon$ times the 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 $\rho \geq 2$,

$$\sum_{i=N}^{\infty} \theta_i^2 \leq L_0 \sum_{i=N}^{\rho N} \theta_i^2, \quad \forall N \geq N_0.$$
Credible sets are honest over polished tail sequences

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

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

**POSTERIOR**: $\theta \mid Y_n \sim N_{\infty}(A_\alpha Y_n, S_\alpha)$.

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

**THEOREM**

For large enough $L$ the empirical Bayes credible ball $\text{ball}(A_{\hat{\alpha}} Y_n, L \hat{r}_{\hat{\alpha}})$ with $\hat{\alpha}$ restricted to a compact is honest over the set of all polished tail sequences (for given $(L_0, N_0, \rho)$).

[For unrestricted $\alpha$ this is true with a slowly varying sequence $L_n \to \infty$.]

**Conjecture**: hierarchical Bayes works equivalently.
“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
\[
\Pi_\alpha\left(\bigcup_{N_0}\{\theta: \theta \in \text{polished tail}(2^{2+2\alpha}, N_0, 2)\}\right) = 1.
\]
Credible sets have optimal diameter

THEOREM
For all $\beta$ 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 $\beta$ 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 $[0, K \sqrt{\log n}]$,

$$\inf_{\theta_0 \in S^{\infty}(M)} P_{\theta_0} \left( r_{\hat{\alpha}_n}^2 \leq e^{(1/2+p)\sqrt{\log n} \log \log n} n^{-1} \right) \to 1.$$
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).
Extensions (if time)
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:** \( \theta \sim N_\infty (0, \tau \Lambda_{\gamma}) \) for \( \lambda_i = i^{-1-2\gamma} \) and \( \tau > 0 \); fixed \( \gamma \).

An empirical or hierarchical Bayes approach on \( \tau \), for fixed \( \gamma \), works as before if \( \theta_0 \in S^\beta \) and \( \beta \leq 2\gamma + 2p + 1 \).
Pointwise estimation

For estimating \( \theta_0(x) \), the function \( \theta_0 \) at the point \( x \), the “apparent smoothness” of \( \theta_0 \in S^\beta \) is \( \beta - 1/2 \), not \( \beta \).

This difference between local smoothness and global smoothness gives trouble for global bandwidth selection methods. E.g. hierarchical or (standard) empirical Bayes cannot differentiate between estimating the full function or a function at a point.

Consequence: empirical or hierarchical credible intervals for functions at a point will be off.
Pointwise estimation (2)

**DATA:** \( Y_i \mid \theta \sim N_n(\theta(x_i), n^{-1}) \), for \( i = 1, 2 \ldots, n \).

**PRIOR:** \( \theta \) Gaussian process, \( E\theta(s)\theta(t) = n^{(1-2\alpha)/(1+2\alpha)}s \land t, \alpha \in (0, 1] \).

**THEOREM**

The coverage \( C_n \) of a marginal credible interval for \( \theta_0(x) \) satisfies

- \( C_n \to P(|Z| < 1.96 \times 2) \) if \( \theta_0 \in C^\beta [0, 1] \) for \( \beta > \alpha \).
- \( C_n \to p < P(|Z| < 1.96 \times 2) \) for some \( \theta_0 \in C^\beta [0, 1] \) if \( \beta = \alpha \).
- \( C_n \to 0 \) if \( \theta_0 \in C^\beta [0, 1] \) for \( \beta < \alpha \).

**CONJECTURE:** Appropriate empirical Bayes (following Wahba) can make prior independent of \( \alpha \) and keep coverage for polished tail sequences.
Adaptation of radius of credible ball

**DATA:** $Y_n \mid \theta \sim N_{\infty}(\theta, n^{-1}I)$.

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

**POSTERIOR:** $\theta \mid Y_n \sim N_{\infty}(A_\alpha Y_n, S_\alpha)$.

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

$$\tilde{\alpha}_n = \sup\left\{\alpha \leq 2\beta_0: \sum_{i=1}^{n^{(1/2+2\beta_0)}} \frac{i^{2+4\alpha}}{(i^{1+2\alpha} + n)^2} \left(Y_i^2 - \frac{1}{n}\right) \leq n^{-2\alpha/(1+2\alpha)} \right\}.$$ 

**THEOREM**

Credible sets based on $\tilde{\alpha}_n$ are honest over polished tail sequences and of radius $O_P(n^{-\beta/(1+2\beta)})$ under $\theta_0 \in S^{\beta}$ for any $\beta \leq 2\beta_0$. 
Sparse regression

**DATA:** $Y \sim N_n(X\theta, \sigma^2 I)$, for $X$ known, $n \times p$.

**PRIOR:**
- Number of nonzero coordinates $s \sim (e^{-cs \log(cp)})$
- Random set $S \subset \{1, \ldots, p\}$ of size $s$.
- $\theta_i \sim $ Laplace for $i \in S$; $\theta_i = 0$ for $i \notin S$.

Single data with $\theta_0 = (0, \ldots, 0, 5, \ldots, 5)$ and $n = 500$ and $\|\theta_0\|_0 = 100$.

Red dots: marginal posterior medians
Orange: marginal credible intervals
Green dots: data points.
Conclusions
Conclusions and Conjectures

Nonparametric credible regions are never “correct” confidence regions.
Conclusions and Conjectures

Nonparametric credible regions are never “correct” confidence regions.

Priors that undersmooth give posteriors with a correct idea of their accuracy.
Conclusions and Conjectures

Nonparametric credible regions are never “correct” confidence regions.

Priors that *undersmooth* give posteriors with a correct idea of their accuracy.

Priors that *oversmooth* give misguided posteriors that wrongly believe they “know”.

Conclusions and Conjectures

Nonparametric credible regions are never “correct” confidence regions.

Priors that undersmooth give posteriors with a correct idea of their accuracy.

Priors that oversmooth give misguided posteriors that wrongly believe they “know”.

This effect disappears if the prior is adapted, by an hierarchical or empirical Bayesian method, but only for convenient truths.
Final conclusion

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”.
Final conclusion

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.
Final conclusion

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.