Nonparametric Credible Sets

Aad van der Vaart
Universiteit Leiden, Netherlands

Padova, March 2013
Contents

0. Credible sets

1. Examples

2. Linear Gaussian inverse problems

3. Theory: fixed priors

4. Theory: adapting priors

5. Concluding remarks
<table>
<thead>
<tr>
<th>Co-authors</th>
<th>Image</th>
<th>Image</th>
<th>Image</th>
</tr>
</thead>
<tbody>
<tr>
<td>Bartek Knapik (Paris)</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Suzanne Sniekers (Leiden)</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Botond Szabo (Eindhoven)</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Harry van Zanten (Amsterdam)</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
0. Credible sets
Nonparametric credible sets

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

Form the posterior distribution $\theta | Y_n$ 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

$$\Pi_n(\theta \in C_n(Y_n) | Y_n) = 0.95.$$
1. Examples
Example: Logistic regression

Bayesian model:

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

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.
Example: genomics

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: reconstruct derivative

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

$$K \theta(x) = \int_0^x \theta(s) \, ds.$$
Example: reconstruct derivative

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

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

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

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

The Volterra operator \( K : L^2[0, 1] \to L^2[0, 1] \) is given by

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

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

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

We put a Gaussian process prior on \(\theta\).
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).
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 $u: [0, 1] \times [0, 1] \rightarrow \mathbb{R}$,

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

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

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

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

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

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

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

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

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

We put a Gaussian proces prior on $\theta$. 
Example: heat equation (n=10 000)

True $\theta_0$ (black), posterior mean (red), 20 realizations from the posterior (dashed black), and posterior credible bands (green).

In all ten panels $n = 10^4$. Left: rough prior. Right: smooth prior.
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.
History: credible sets are a disaster!

AN ANALYSIS OF BAYESIAN INFERENCE FOR NONPARAMETRIC REGRESSION

By Dennis D. Cox

Rice University

The observation model $y_i = g(x_i) + \epsilon_i, 1 \leq i \leq m$, is considered, where the $\epsilon$'s are i.i.d. with mean zero and variance $\sigma^2$, and $g$ 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 $\hat{g} - g$ is analyzed, where $\hat{g}$ is the posterior expectation of $\beta$. Asymptotic posterior and sampling distribution approximations are given for $(\hat{g})^a$ when $a$ is 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 $\alpha > 0$ as $\alpha \to 0$ with prior probability. 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

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

where $t_{i\alpha} = i/n, \beta: [0, 1] \to R$ 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 $\epsilon_i$ 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

$$L = \sum_{i=0}^{m} a_i |D^i|$$
“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 [· · · ]

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 \( Y(t_i) = g(t_i) + \varepsilon_i \), \( i = 1, 2, \ldots, n \), where \( g(t_i), t_i \in [0, 1] \) is a smooth function and the \( \{\varepsilon_i\} \) 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 \( Y(t_i), i = 1, 2, \ldots, n \), 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
\[
Y(t_i) = g(t_i) + \varepsilon_i, \quad i = 1, 2, \ldots, n, \quad t_i \in [0, 1],
\]
where \( \varepsilon = (\varepsilon_1, \ldots, \varepsilon_n)^T \sim N(0, \sigma^2 I_{n \times n}) \), \( \sigma^2 \) 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 \( \hat{g}_{m, n} \), is the minimizer of
\[
n^{-1} \sum_{i=1}^n (g(t_i) - y_i)^2 + \lambda \int_0^1 (g^{(m)}(t))^2 dt
\]
Wahba’s credible set $C_{Y \tau} (Y_{\tau})$ 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 $\alpha$-values at which the band contains the truth.
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*).
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
Example: more bad news?

What about adapting the prior smoothness by hierarchical or empirical Bayes?

Empirical Bayes:

- Take a family of priors $\Pi_{\alpha}$ of varying smoothness $\alpha > 0$.
- Determine $\hat{\alpha}_n$ as the MLE in the model $\theta | \alpha \sim \Pi_{\alpha}, Y_n | (\theta, \alpha) \sim \Pi_{\alpha}$.
- Use the plug-in posterior $\theta | (Y_n, \alpha)$, with $\alpha = \hat{\alpha}_n$. 
Example: more bad news?

What about adapting the prior smoothness by hierarchical or empirical Bayes?

Empirical Bayes:
- Take a family of priors $\Pi_\alpha$ of varying smoothness $\alpha > 0$.
- Determine $\hat{\alpha}_n$ as the MLE in the model $\theta | \alpha \sim \Pi_\alpha$, $Y_n | (\theta, \alpha) \sim \Pi_\alpha$.
- Use the plug-in posterior $\theta | (Y_n, \alpha)$, with $\alpha = \hat{\alpha}_n$.

$n = 10^3$  

$n = 10^4$  

$n = 10^6$  

$n = 10^8$

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.
Example: more bad news?

What about adapting the prior smoothness by **hierarchical** or **empirical Bayes**?

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

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

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.

**Hierarchical Bayes:**
- Full Bayes, with prior on $\alpha$. 
What is happening?

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

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

Empirical or hierarchical Bayes try to learn the smoothness of the true function from the data, but they can be tricked.
What is happening?

$n = 10^3$

$n = 10^4$

$n = 10^6$

$n = 10^8$

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

Empirical or hierarchical Bayes try to learn the smoothness of the true function from the data, but they can be tricked.

- Smoothness can be expressed in the speed of convergence $\theta_i \downarrow 0$ of the coefficients $\theta_1, \theta_2, \ldots$ of $\theta$ relative to an appropriate basis.
- Finite data allows inference only on $\theta_1, \ldots, \theta_N$ for an “effective dimension” $N = N_n$.
- Trouble arises if $\theta_1, \ldots, \theta_N$ does not resemble the full sequence $\theta_1, \theta_2, \ldots$. 
2. Linear Gaussian inverse problems
Theory

Theory so far available for:

- Regression with Gaussian errors.
- Linear Gaussian inverse problems.
The model of Wahba (1983) and 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 Wahba (1983) and 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 Wahba (1983) and 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

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

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

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

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

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

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

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: \( Y_{n,i} \mid \theta_i \sim N(\kappa_i \theta_i, n^{-1}) \) for

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

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

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

**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 $\sum_i i^{2\beta} \theta_i^2$.]
3. Theory: 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:** \( \theta | Y_n \sim N_\infty (A_\alpha Y_n, S_\alpha) \).

**THEOREM**

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

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

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

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

**PRIOR:** \( \theta \sim N_\infty(0, \Lambda_\alpha) \).

**POSTERIOR:** \( \theta \mid Y_n \sim N_\infty(A_\alpha Y_n, S_\alpha) \).
Linear Gaussian inverse problem — credible sets

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

**PRIOR:** $\theta \sim N_\infty \left( 0, \Lambda_\alpha \right)$.

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

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

$$\Pi_n(\theta \in C_n(Y_n) | Y_n) = 0.95.$$
**Linear Gaussian inverse problem — credible sets**

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

**PRIOR:** $\theta \sim N_\infty(0, \Lambda_\alpha)$.

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

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

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

The shape may depend on one’s “loss function”: e.g.

- a central ball
- a uniform band
- a central ellipsoid that takes account of intrinsic difficulties
- ... 

For function estimation one likes sets that can be plotted.

*Plotting method for credible balls: simulate 1000 realizations from the posterior, plot the 95% that are closest to the posterior mean.*
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:** 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 OK as a rough indication of statistical uncertainty.
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.
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.]
4. Theory: adapting priors
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

- **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 on a scaling.

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

- **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 on a scaling.

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

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

Empirical Bayes method: plug in the MLE $\hat{\alpha}$ of the
MARGINAL MODEL: $Y_n \sim N_\infty(0, K\Lambda_\alpha K^T + n^{-1}I)$:
**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:** $\theta | Y_n \sim N_\infty(A_\alpha Y_n, S_\alpha)$.

**Empirical Bayes method:** plug in the MLE $\hat{\alpha}$ of the

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

$$\hat{\alpha} = \arg\max_\alpha \sum_{i=1}^\infty \left( \frac{n^2}{i^{1+2\alpha+2p}} + n Y_{n,i}^2 - \log \left( 1 + \frac{n}{i^{1+2\alpha+2p}} \right) \right).$$
**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:** $\theta \mid Y_n \sim N_\infty(A_\alpha Y_n, S_\alpha)$.

**Empirical Bayes method:** plug in the MLE $\hat{\alpha}$ of the

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

$$
\hat{\alpha} = \arg\max_\alpha \sum_{i=1}^\infty \left( \frac{n^2}{i^{1+2\alpha+2p} + n} Y_{n,i}^2 - \log \left( 1 + \frac{n}{i^{1+2\alpha+2p}} \right) \right).
$$

**THEOREM**

The posterior distribution $\theta \mid Y_n$ with plugged in $\hat{\alpha}_n$ contracts nearly at the optimal rate $n^{-\beta/(2\beta+2p+1)}$ if $\theta_0 \in S^\beta$, for any $\beta > 0$. 
**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:** $\theta | Y_n \sim N_\infty(A_\alpha Y_n, S_\alpha)$.

**Empirical Bayes method:** plug in the MLE $\hat{\alpha}$ of the

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

$$\hat{\alpha} = \arg\max_\alpha \sum_{i=1}^{\infty} \left( \frac{n^2}{i^{1+2\alpha+2p}} + \frac{Y_{n,i}^2}{n} - \log \left( 1 + \frac{n}{i^{1+2\alpha+2p}} \right) \right).$$

**THEOREM**

The posterior distribution $\theta | Y_n$ with plugged in $\hat{\alpha}_n$ contracts nearly at the optimal rate $n^{-\beta/(2\beta+2p+1)}$ if $\theta_0 \in S^\beta$, for any $\beta > 0$.

**CREDIBLE SET:** $\text{ball}(A_\alpha Y_n, r_\alpha)$ of posterior mass 0.95, with $\alpha = \hat{\alpha}_n$ plugged in.
**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).

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

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

THEOREM
The posterior distribution $\theta | Y_n$ contracts nearly at the optimal rate $n^{-\beta/(2\beta+2p+1)}$ if $\theta_0 \in S^\beta$, for any $\beta > 0$. 
Linear Gaussian inverse problem — hierarchical Bayes

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

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

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

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

**THEOREM**

The posterior distribution $\theta \mid Y_n$ contracts nearly at the optimal rate $n^{-\beta/(2\beta+2p+1)}$ if $\theta_0 \in S^{\beta}$, for any $\beta > 0$.

**CREDIBLE SET:** $C_n(Y_n)$ with posterior mass 0.95.
Given a family of priors $\Pi_\alpha$ of varying smoothness the parameter $\alpha$ can be chosen by empirical or hierarchical Bayes to give optimal recovery of the true parameter.

Does the posterior distribution also give correct uncertainty quantification?
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.

The pictures show an “inconvenient” truth. For some (most?) truths the results are good.
A set $C_n(Y_n)$ is an honest confidence set if

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

for all $\theta_0 \in \Theta_0$.

$\Theta_0$ contains ‘all possible truths’, e.g. $\Theta_0 = S_1^\beta$, Sobolev ball of regularity $\beta$. 
What do the frequentists say? — honesty

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

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

$\Theta_0$ 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 $S_{1}^{\beta}$. 
A set $C_n(Y_n)$ is an honest confidence set if

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

$\Theta_0$ 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 $\bigcup_{\beta \geq \beta_0} S_1^\beta$, then its diameter is of the uniform order $O_P(n^{-\beta_0/(1/2+2\beta_0)})$ over $S_1^\beta$ 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:
- A more regular true function is easier to estimate.
- Estimators can be simultaneously optimal for multiple regularities.
- Bayesian estimators achieve this by a prior on the “bandwidth”.
What do the frequentists say? — discrepancy between estimation and uncertainty quantification

Adaptive estimation:
- A more regular true function is easier to estimate.
- Estimators can be simultaneously optimal for multiple regularities.
- Bayesian estimators 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.
What do the frequentists say? — discrepancy between estimation and uncertainty quantification

Adaptive estimation:
- A more regular true function is easier to estimate.
- Estimators can be simultaneously optimal for multiple regularities.
- Bayesian estimators 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.

“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 \(I = 1, 2, \ldots\),

\[
\sum_{i=I}^{I+1000} i^{2\beta} \theta_i^2 \geq \frac{1}{100000} \sup_i i^{2\beta} \theta_i^2.
\]

**THEOREM**  [after Bull and Nickl, 2012]

There exist \(C_n(Y_n)\) that

- are honest over the set of all self-similar \(\theta_0 \in \bigcup_\beta S^\beta_1\).
- have diameter of order \(O_P(n^{-\beta/(1+2\beta)})\) whenever \(\theta_0 \in S^\beta\), for any \(\beta\).

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 \((\hat{\theta}_1, \ldots, \hat{\theta}_{\hat{N}})\) at the ‘effective’ dimension \(\hat{N}\).
Credible sets are honest over self-similar functions

**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**

If $\hat{\alpha}$ is the MLE for the marginal law of $Y_n$, then credible ball $\text{ball}(A_{\hat{\alpha}} Y_n, \log \log n \hat{\alpha})$

- is honest over the set of all self-similar $\theta_0 \in \bigcup_{\beta} S_{1}^{\beta}$.
- has diameter nearly of the order $O_P(n^{-\beta/(1+2\beta)})$ whenever $\theta_0 \in S^{\beta}$, for any $\beta$.

Empirical Bayes works for self-similar truths.

Conjecture: hierarchical Bayes works equivalently.
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).
5. Concluding remarks
Linear Gaussian inverse problem — scaling the prior

**DATA:** $Y_n | \theta \sim N_\infty(K\theta, \kappa_1^{-1}I)$ for $\kappa_i \sim i^{-p}$. 
Linear Gaussian inverse problem — scaling the prior

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

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

An empirical or hierarchical Bayes approach on $\tau$, for fixed $\alpha$, works as before if $\theta_0 \in S^\beta$ and $\beta \leq 2\alpha + 2p + 1$. 
“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.,”

\[
\liminf_{n \to \infty} P_{\theta_0}\left(\text{ball}(A_{\alpha} Y_n, r_{\alpha}) \ni \theta_0\right) = 0, \quad \text{for } \Pi\text{-a.e. } \theta_0.
\]
Credible sets are honest over prior sets?

"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 [...]."

Unfortunately, the hoped for result is false in about the worst possible way, viz.,

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

Slightly enlarging the ball gives full coverage.
CONJECTURE
For $\hat{\alpha}$ determined by empirical Bayes in the linear inverse problem:

$$\liminf_{n \to \infty} P_{\theta_0} \left( \text{ball}(\hat{A}_{\hat{\alpha}} Y_n, (\log n) \hat{r}_{\hat{\alpha}}) \ni \theta_0 \right) = 1, \quad \text{for } \Pi_\alpha \text{-a.e. } \theta_0, \text{ for every } \alpha.$$
Functionals

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

$$\left\| \Pi_n(\psi(\theta) \in \cdot \mid Y_n) - N(\mu_n(Y_n), s_n) \right\| \to 0.$$  

Credible sets for $\psi(\theta)$ are then valid confidence sets.
Functionals

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

$$\left\| \Pi_n(\psi(\theta) \in \cdot | Y_n) - N(\mu_n(Y_n), s_n) \right\| \to 0.$$ 

Credible sets for $\psi(\theta)$ are then valid confidence sets.

The essence is that estimation of $\psi(\theta)$ does not involve a bias-variance trade-off: the bias should be negligible.
Functionals (2)

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.

The set of posterior distributions, and their marginals, for varying smoothing parameters $\alpha$ are “sufficient”, but the smoothing parameter may have to be chosen outside the Bayesian framework.
Example: other versions of empirical Bayes

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

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

**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.

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

**THEOREM**

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

This is optimal in a frequentist sense.
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”.
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 may disappear if the prior is adapted, for instance by an hierarchical or empirical Bayesian method,
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 may disappear if the prior is adapted, for instance by an hierarchical or empirical Bayesian method, *but only for truths that resemble the prior.*
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.

The fine details of a prior in nonparametrics are not obvious.