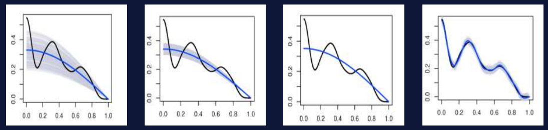
Nonparametric Bayes: review and challenges

Aad van der Vaart

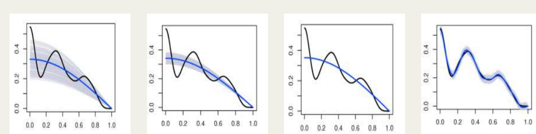
Universiteit Leiden, Netherlands



European Meeting of Statisticians

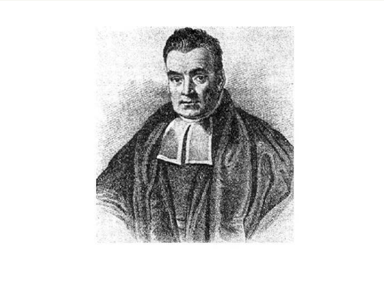
Palermo, Italy, July 2019

Contents



Introduction

The Bayesian paradigm



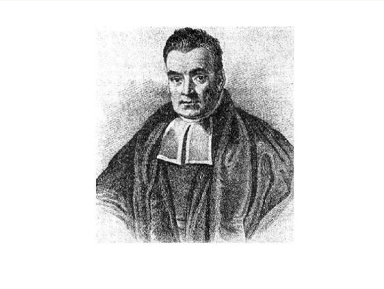
- A parameter θ is generated according to a prior distribution Π
- Given θ the data X is generated according to a measure P_θ

This gives a joint distribution of (X, θ)

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

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

The Bayesian paradigm



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

This gives a joint distribution of (X, θ)

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

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

If P_θ is given by a density $x \mapsto p_\theta(x)$, then **Bayes's rule** gives

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

Frequentist Bayes

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

Frequentist Bayes

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

Recovery

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

Frequentist Bayes

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

Recovery

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

Uncertainty quantification

We like the "spread" of $\Pi(\theta \in \cdot \mid X)$ to indicate remaining uncertainty

Frequentist Bayes

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

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

Uncertainty quantification
We like the "spread" of $\Pi(\theta \in \cdot \mid X)$ to indicate remaining uncertainty

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

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

Parametric Bayes

Laplace 1812, Bernstein 1910, von Mises, Le Cam 1989

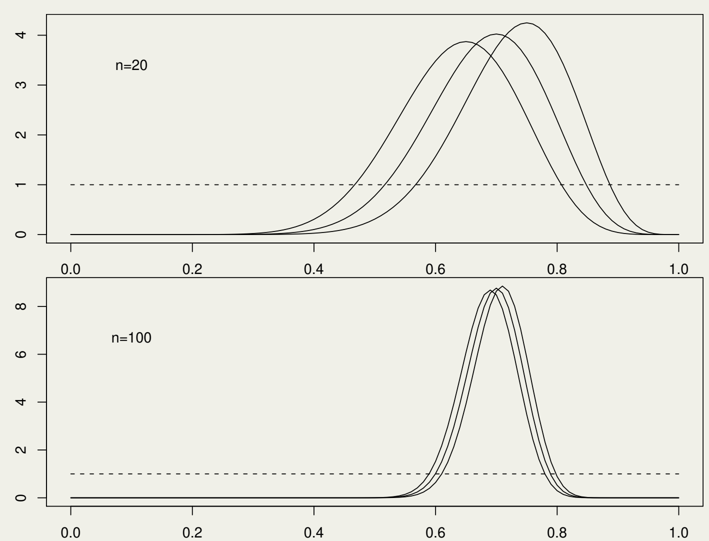
Data: $X_1, \ldots, X_n \overset{\text{iid}}{\sim} p_\theta$

$\mathbb{R}^d \ni \theta \mapsto p_\theta$ smooth and **identifiable**

Thm Under θ_0, for any prior with positive density,

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

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



Parametric Bayes

Laplace 1812, Bernstein 1910, von Mises, Le Cam 1989

Data: $X_1, \ldots, X_n \overset{\text{iid}}{\sim} p_\theta$
$\mathbb{R}^d \ni \theta \mapsto p_\theta$ smooth and **identifiable**

Thm Under θ_0, for any prior with positive density,

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

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

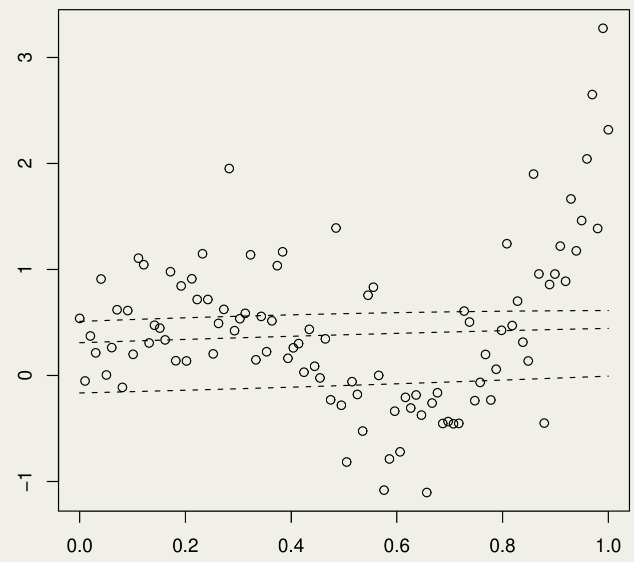
Recovery:
The posterior distribution concentrates most of its mass on balls of radius $O(1/\sqrt{n})$ around θ_0

Uncertainty quantification:
A central set of posterior probability 95 % is equivalent to the usual Wald confidence set $\{\theta\colon n(\theta - \tilde\theta_n)^T I_{\tilde\theta_n} (\theta - \tilde\theta_n) \le \chi^2_{d,1-\alpha}\}$

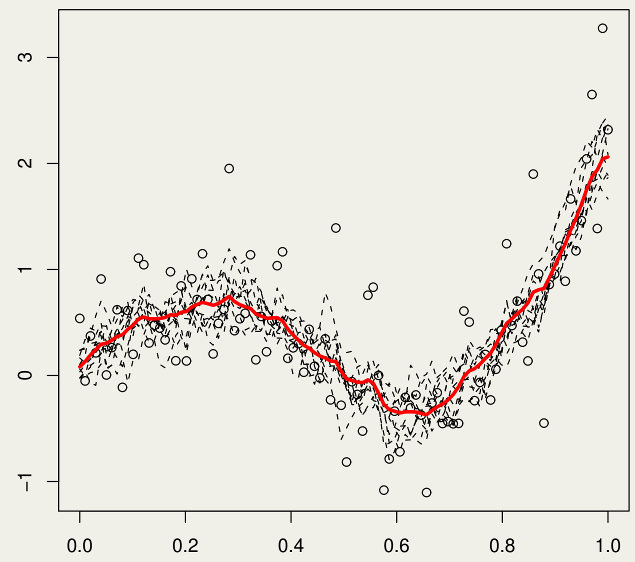
Bayesian curve estimation

A prior and posterior of a function can be visualized by plotting functions that are simulated from the prior and posterior distributions

prior



posterior

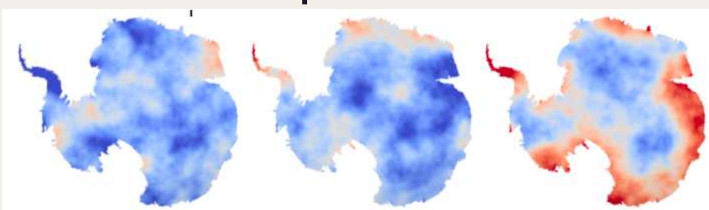


Many examples of priors: Dirichlet, Gaussian, random series,...
Recovery well understood

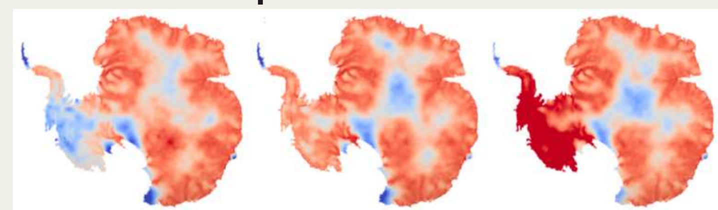
Bayesian inverse problems — data assimilation

A prior and posterior of a surface can be visualized by plotting surfaces that are simulated from the prior and posterior distributions.

prior



posterior

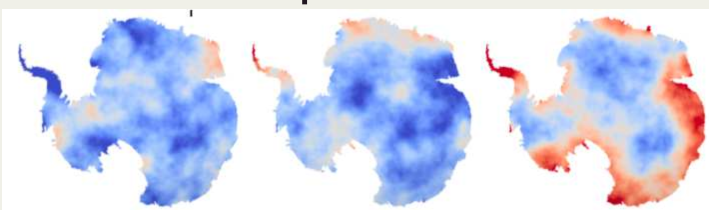


From Stadler et al., 2017

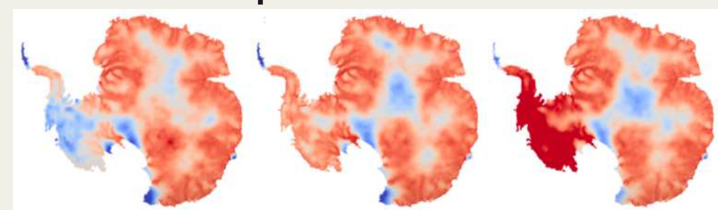
Bayesian inverse problems — data assimilation

A prior and posterior of a surface can be visualized by plotting surfaces that are simulated from the prior and posterior distributions.

prior



posterior



From Stadler et al., 2017

$\boxed{\text{Data: } = B_\theta u + \text{noise}, \quad \text{for } D_\theta u = 0, \quad u \in G_\theta}$

$$\begin{aligned} -\boldsymbol{\nabla}\cdot[2\eta(\boldsymbol{u},n)\,\dot{\boldsymbol{\varepsilon}}_{\boldsymbol{u}} - \boldsymbol{I}p] &= \rho\boldsymbol{g} && \text{in } \Omega \\ \boldsymbol{\nabla}\cdot\boldsymbol{u} &= 0 && \text{in } \Omega \\ \boldsymbol{\sigma}_{\boldsymbol{u}}\boldsymbol{n} &= \boldsymbol{0} && \text{on } \Gamma_t \\ \boldsymbol{u}\cdot\boldsymbol{n} = 0,\ \boldsymbol{T}\boldsymbol{\sigma}_{\boldsymbol{u}}\boldsymbol{n} + \exp(\beta)\boldsymbol{T}\boldsymbol{u} &= \boldsymbol{0} && \text{on } \Gamma_b \end{aligned}$$

$$\hat\theta = \operatorname*{argmin}_\theta \Big[L(\theta, \text{Data}) + \text{penalty}(\theta) \Big]$$

Thm Dashti et al 2013

If $\text{penalty}(\theta) = \|\theta\|^2_{\mathbb{H}}$ for RKHS norm, $\hat\theta$ is posterior mode for Gaussian prior

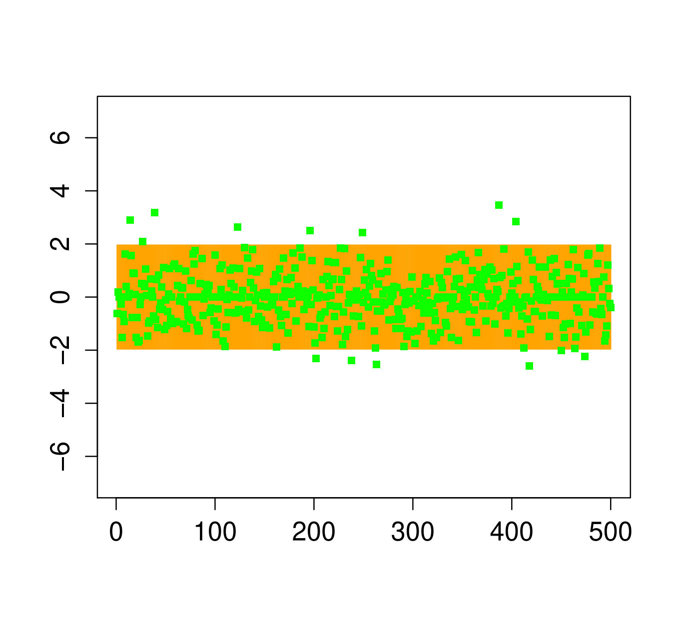
Connects to applied analysis

Bayesian understanding starting to develop *Stuart, Agapiou, Nickl,...*

High-dimensional Bayes

A high-dimensional parameter vector (or matrix) may be visualized through a plot of marginal distributions versus an index

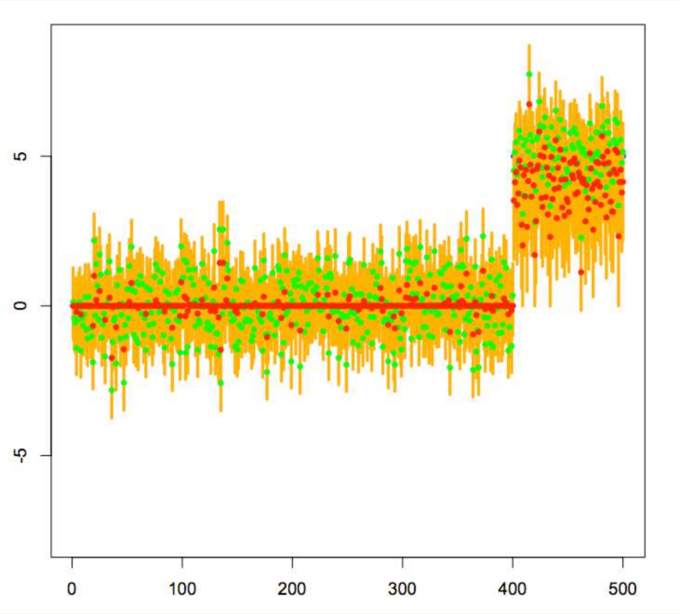
prior



Parameters $\theta_1, \ldots, \theta_{500}$ versus index $1, \ldots, 500$
Orange: marginal prior predictive intervals

Green dots: prior draws

posterior



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

Green dots: data points

Connects to Empirical Bayes and Large Scale Inference
Recent progress

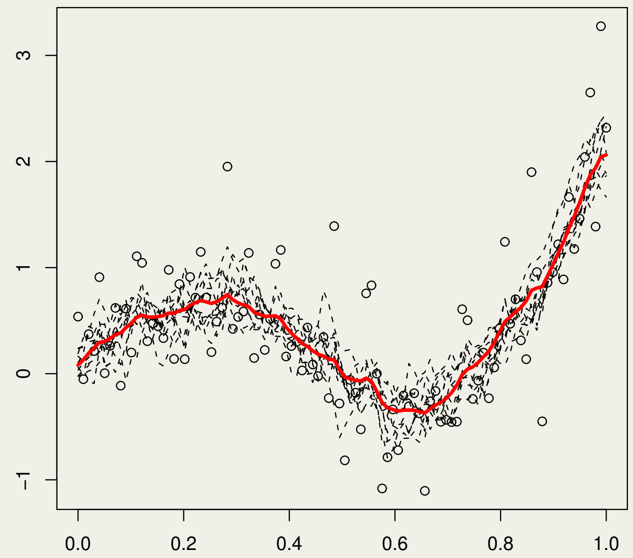
Recovery

Rate of contraction

Data: $X^{(n)} \sim P_\theta^{(n)} \qquad \theta \in (\Theta, d)$

Def *contraction rate* at θ_0 is ϵ_n if, for large M,

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



Rate of contraction

Data: $X^{(n)} \sim P_\theta^{(n)} \qquad \theta \in (\Theta, d)$

Def *contraction rate* at θ_0 is ϵ_n if, for large M,

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

Benchmark rate for (inverse) curve fitting:
A function θ of d variables with bounded derivatives of order β is estimable based on n observations at rate

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

Benchmark rate for sparse estimation:
A vector θ in $\mathbb{R}^n$ of which $s \ll n$ coordinates are nonzero is estimable based on 1 observation per parameter at rate

$$\sqrt{s\log(n/s)}.$$

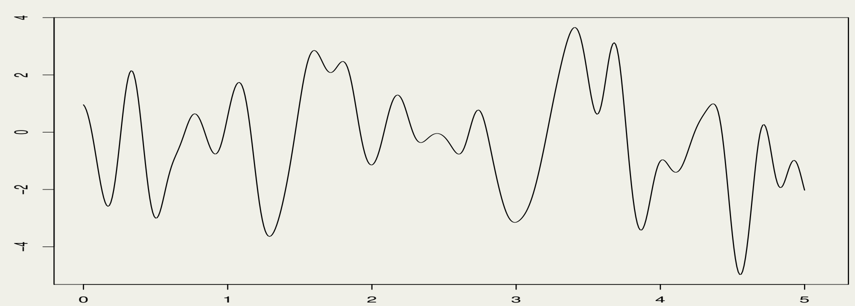
Curve fitting with square exponential prior

vdV, van Zanten

Data: Sample of size n in regression model or from density

Prior on regression function or log density θ: centered Gaussian process with

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



Thm Rate of contraction is $(\log n)^\gamma/\sqrt{n}$ if θ_0 is analytic, but is $(1/\log n)^k$ if θ_0 is only ordinary smooth

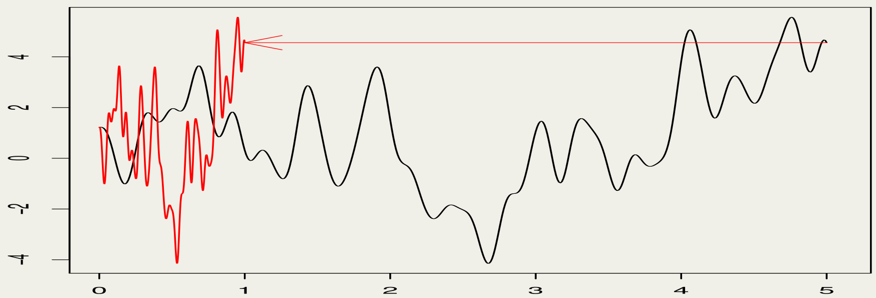
$$\mathrm{P}\big(\|\theta\|_\infty < \varepsilon\big) \gtrsim e^{-C(\log \varepsilon^{-1})^{1+d/2}}$$

Curve fitting with square exponential prior — adaptation

Data: Sample of size n in regression model or from density

Prior on regression function or log density θ:

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



Thm Rate of contraction is:

- if $\theta_0 \in C^\beta[0, 1]^d$, then nearly $n^{-\beta/(2\beta+d)}$
- if θ_0 is analytic, then nearly $n^{-1/2}$

Linear Gaussian inverse problems

Knapik, Szabó, van Zanten et al. 2011, 2015

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

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

Linear Gaussian inverse problems

Knapik, Szabó, van Zanten et al. 2011, 2015

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

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

Thm If $\sum_{i=1}^{\infty} i^{2\beta}\theta_{i,0}^2 < \infty$, then rate:

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

Linear Gaussian inverse problems

Knapik, Szabó, van Zanten et al. 2011, 2015

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

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

Thm If $\sum_{i=1}^\infty i^{2\beta}\theta_{i,0}^2 < \infty$, then rate:

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

- Prior on α

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

- $n^{-\beta/(2\beta+2p+1)}$, any $\beta > 0$

Gaussian process priors

When using a Gaussian process as prior for a function:

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

(Generalizes to non-Gaussian priors Ray, Yan, Agapiou,...)

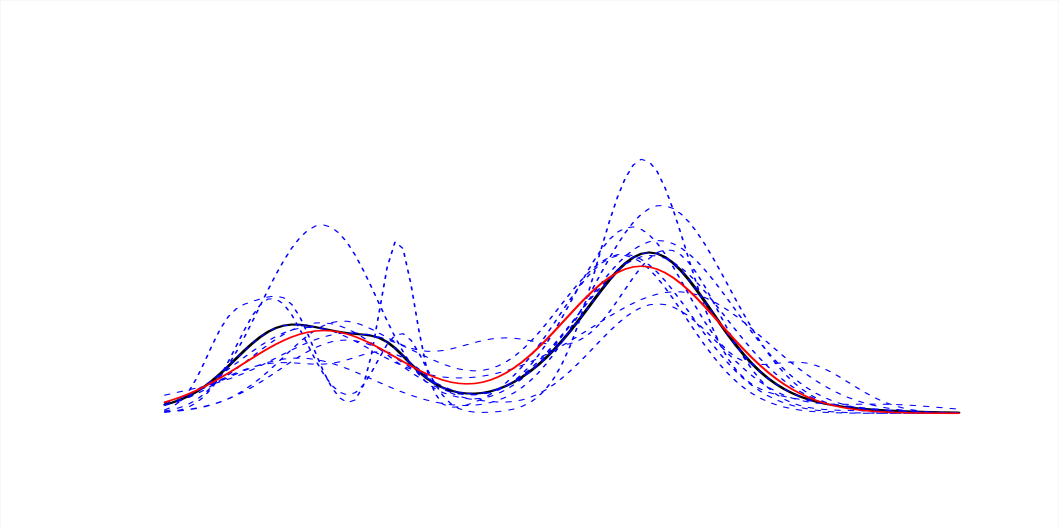
Dirichlet normal mixtures

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

Data: $X_1, \ldots, X_n \overset{\text{iid}}{\sim} p$

- $F \sim$ Dirichlet process
- $1/c \sim \Gamma(a,b)$, independent of F

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



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

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

[Plot by DPpackage, Jara et al., 2011]

Dirichlet normal mixtures

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

Data: $X_1, \ldots, X_n \overset{\text{iid}}{\sim} p$

- $F \sim$ Dirichlet process
- $1/c \sim \Gamma(a,b)$, independent of F

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

Thm Hellinger rate of contraction for $X_1, \ldots, X_n \overset{\text{iid}}{\sim} p_0$ is, any $\beta > 0$,

- nearly $n^{-1/2}$ if $p_0 = p_{F_0,c_0}$, some F_0, c_0
- nearly $n^{-\beta/(2\beta+d)}$ if $p_0 \in C^\beta(\mathbb{R}^d)$ with exponentially small tails

Dirichlet normal mixtures

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

Data: $X_1, \ldots, X_n \overset{\text{iid}}{\sim} p$

- $F \sim$ Dirichlet process
- $1/c \sim \Gamma(a, b)$, independent of F

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

Thm Hellinger rate of contraction for $X_1, \ldots, X_n \overset{\text{iid}}{\sim} p_0$ is, any $\beta > 0$,

- nearly $n^{-1/2}$ if $p_0 = p_{F_0,c_0}$, some F_0, c_0
- nearly $n^{-\beta/(2\beta+d)}$ if $p_0 \in C^\beta(\mathbb{R}^d)$ with exponentially small tails

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

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

Data: $Y^n \sim N_n(\theta, I)$, for $\theta = (\theta_1, \ldots, \theta_n) \in \mathbb{R}^n$

- $\tau \sim B(1, n+1)$
- $\theta_i \overset{\text{iid}}{\sim} (1-\tau)\delta_0 + \tau G$, e.g. $G =$ Laplace

Thm For $s_n \to \infty$ with $s_n \ll n$,

$$\sup_{\#(\theta_{0,i}\neq 0)\leq s_n} \mathrm{E}_{\theta_0}\, \Pi_n\big(\theta\colon \|\theta-\theta_0\|_2^2 \gtrsim s_n \log(n/s_n) \,|\, Y^n\big) \to 0.$$

Shrinkage controlled by sparsity parameter τ

(interpretation: $\tau \approx (s/n)\sqrt{\log n/s}$)

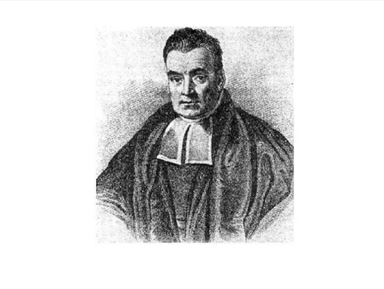
Uncertainty quantification

Credible sets

Def A credible set is a data-dependent set $C(X)$ with

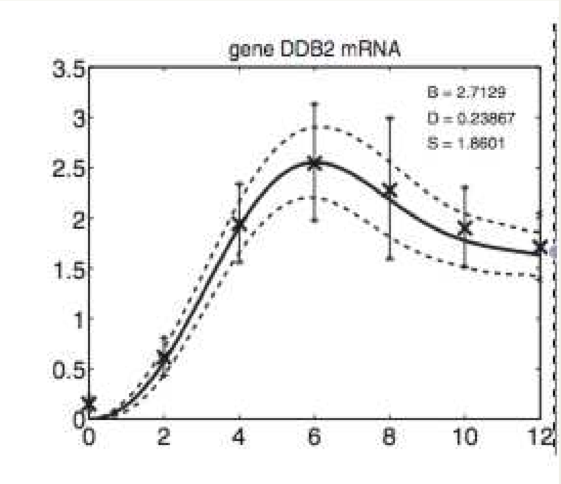
$$\Pi\big(\theta \in C(X) |\, X\big) = 0.95.$$

Credible sets

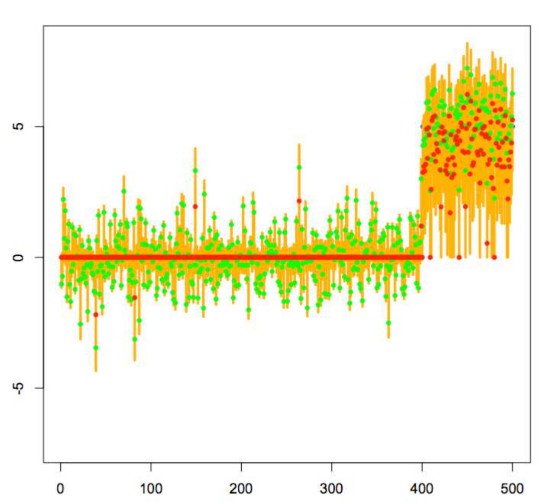


Def A credible set is a data-dependent set $C(X)$ with

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



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



Red dots: marginal posterior medians
Orange: marginal credible intervals

Green dots: data points

Do credible sets correctly quantify *remaining uncertainty*?

Is a credible set a confidence set?

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

Do credible sets correctly quantify *remaining uncertainty*?

Is a credible set a confidence set?

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

Meta Thm Cox 1993, Freedman 2000, Leahu 2012

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

Do credible sets correctly quantify *remaining uncertainty*?

Is a credible set a confidence set?

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

Meta Thm Cox 1993, Freedman 2000, Leahu 2012

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

Identify θ with a set Ψ of smooth functionals $\psi(\theta)$

A joint Bernstein-von Mises theorem

$$d\Big(\Pi_n\big(\theta\colon \sqrt{n}(\psi(\theta) - \hat{\psi}_n)_{\psi\in\Psi} \in \cdot \mid X^{(n)}\big), \mathrm{P}\big(Z(\psi)_{\psi\in\Psi} \in \cdot\big)\Big) \to 0$$

may be used to get valid credible sets

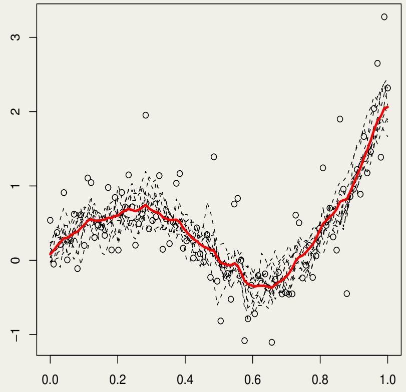
Castillo, Nickl, Ray .. make this operational using weak norms

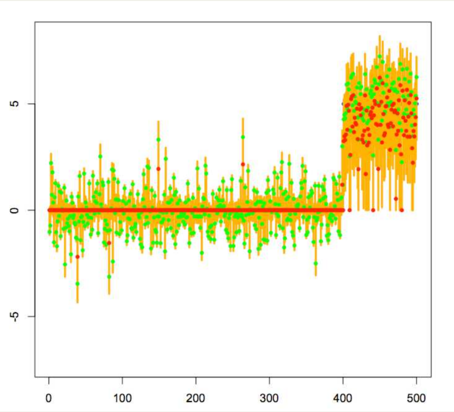
Do credible sets correctly quantify *remaining uncertainty*?

Is a credible set a confidence set?

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

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





Deterministic bandwidth: coverage requires undersmoothing

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

Deterministic bandwidth: coverage requires undersmoothing

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

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

** Example: heat equation

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

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

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

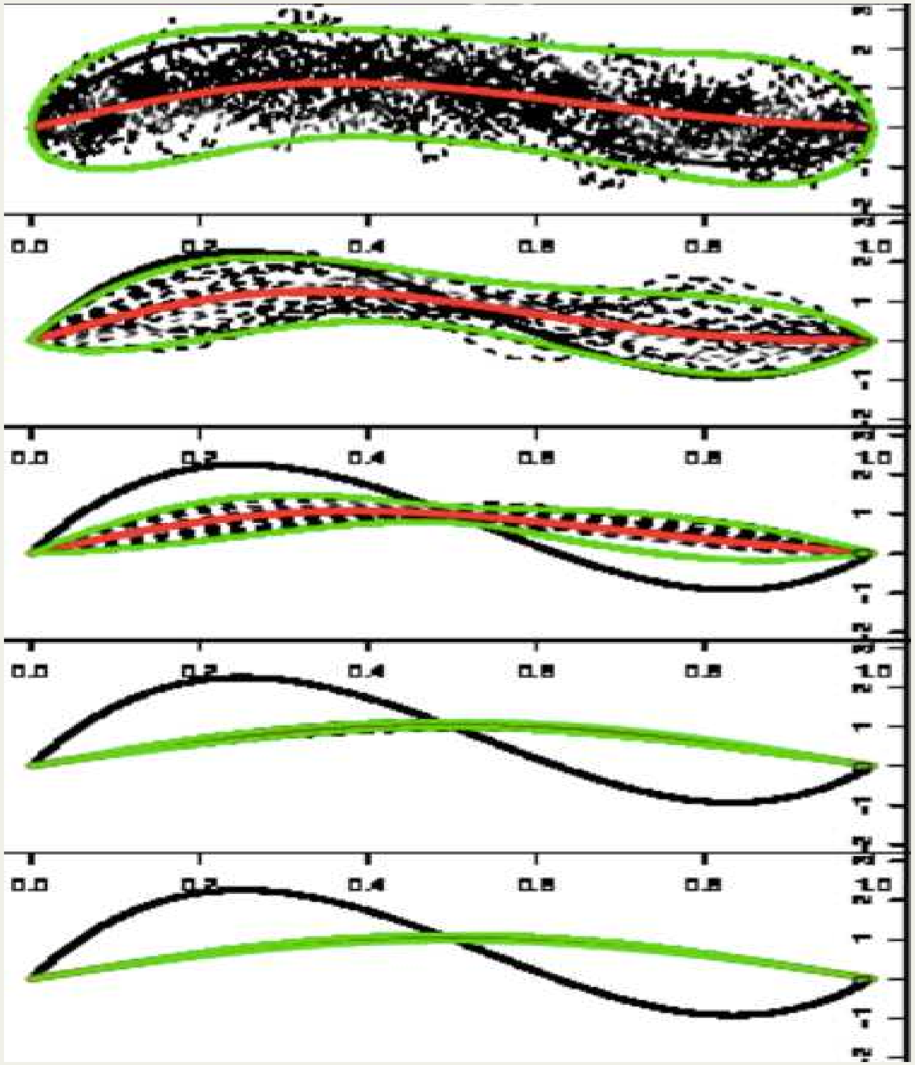
Observe noisy version $(X_x^{(n)}: 0 \leq x \leq 1)$ of final curve

- $\theta = \sum_{i=1}^{\infty} \theta_i e_i$, for e_i eigenbasis of K
- Truth: $\theta_{0,i} \asymp i^{-1-2\beta}$
- Prior: $\theta_i \overset{\text{ind}}{\sim} N(0, i^{-1-2\alpha})$

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

Example: heat equation (n=10 000)

Knapik et al, 2013



True θ_0 (black), posterior mean (red)

20 realizations from the posterior (dashed black)

posterior credible bands (green)

- $\theta = \sum_{i=1}^{\infty} \theta_i e_i$
- Truth:
  $\theta_{0,i} \asymp i^{-1-2\beta}$
- Prior:
  $\theta_i \overset{\text{ind}}{\sim} N(0, i^{-1-2\alpha})$

Top to bottom: increasing α

Bayesian adaptation

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

Examples

- $t \mapsto \sum_{i=1}^\infty \theta_i e_i(t)$, for $\theta_i \overset{\text{ind}}{\sim} N(0, i^{-1-2\alpha})$
- $t \mapsto G_{\alpha t}$, for Gaussian process G
- $t \mapsto \int \alpha^{-1}\phi(\alpha^{-1}(t-z))\, dF(z)$, with $F \sim$ Dirichlet process
- $i \mapsto \theta_i \sim \alpha\delta_0 + (1-\alpha)$Laplace

Bayesian adaptation

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

Examples

- $t \mapsto \sum_{i=1}^\infty \theta_i e_i(t)$, for $\theta_i \overset{\text{ind}}{\sim} N(0, i^{-1-2\alpha})$
- $t \mapsto G_{\alpha t}$, for Gaussian process G
- $t \mapsto \int \alpha^{-1}\phi(\alpha^{-1}(t-z))\, dF(z)$, with $F \sim$ Dirichlet process
- $i \mapsto \theta_i \sim \alpha\delta_0 + (1-\alpha)$Laplace

Hierarchical Bayes:

- Prior on α
- Ordinary posterior

Empirical Bayes:

- $\hat\alpha = \operatorname{argmax}_\alpha \int p(X|\,\theta)\, d\Pi_\alpha(\theta)$
- Plug-in posterior $\Pi_{\hat\alpha}(\cdot|\, X)$

Both methods give adaptive reconstructions:
for smoother true function better reconstruction

Bayesian adaptation

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

Examples

- $t \mapsto \sum_{i=1}^\infty \theta_i e_i(t)$, for $\theta_i \overset{\text{ind}}{\sim} N(0, i^{-1-2\alpha})$
- $t \mapsto G_{\alpha t}$, for Gaussian process G
- $t \mapsto \int \alpha^{-1}\phi(\alpha^{-1}(t-z))\, dF(z)$, with $F \sim$ Dirichlet process
- $i \mapsto \theta_i \sim \alpha\delta_0 + (1-\alpha)$Laplace

Hierarchical Bayes:

- Prior on α
- Ordinary posterior

Empirical Bayes:

- $\hat\alpha = \operatorname{argmax}_\alpha \int p(X|\,\theta)\, d\Pi_\alpha(\theta)$
- Plug-in posterior $\Pi_{\hat\alpha}(\cdot|\, X)$

Both methods give adaptive reconstructions:
for smoother true function better reconstruction

This implies that they *cannot* give honest confidence sets

Honesty and impossibility of adaptation

Low, Cai, Lepski, Hoffmann, Juditzky, Lambert-Lacroix, Robins, vdV, Bull, Nickl

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

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

Honesty and impossibility of adaptation

Low, Cai, Lepski, Hoffmann, Juditzky, Lambert-Lacroix, Robins, vdV, Bull, Nickl

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

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

Thm The diameter of $C_n(X^{(n)})$ cannot be smaller, uniformly in $\theta \in \Theta_1 \subset \Theta$, than:

(a) ε_n such that, for any T_n,

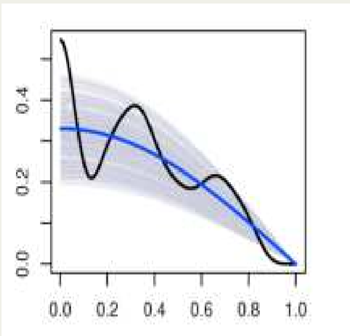
$$\liminf_{n\to\infty} \sup_{\theta\in\Theta_1} \mathrm{P}_\theta\bigl(d(T_n,\theta) \geq \varepsilon_n\bigr) > 0.501$$

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

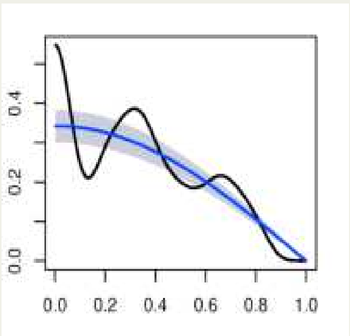
(a) typically gives minimax rate of estimation for model Θ_1
(b) is determined by biggest model Θ rather than Θ_1

* Credible balls — counter example — reconstructing a derivative

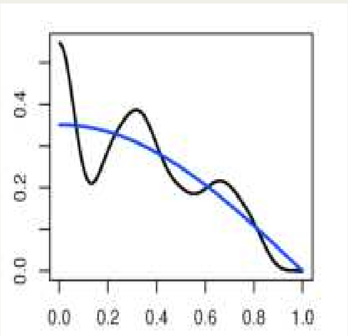
$n = 10^3$



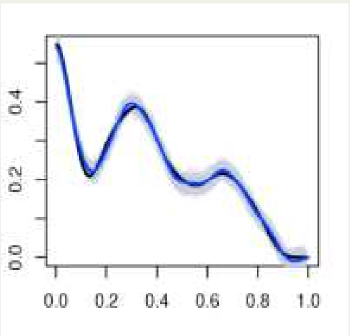
$n = 10^4$



$n = 10^6$



$n = 10^8$



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

** Credible balls — counter example — reconstructing a derivative

Data: $X^{(n)} = \int_0^\cdot \theta(t)\,dt + n^{-1/2}\dot W$, for white noise $\dot W$

- Prior: $\theta = \sum_{i=1}^\infty \theta_i e_i$, with $\theta_i |\, \alpha \overset{\text{ind}}{\sim} N(0, i^{-2\alpha-1})$
- Prior on α or empirical Bayes $\hat\alpha$

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

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

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

$$\liminf P_\theta\big(\theta \in \hat C_n\big) = 0$$

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

Estimation versus uncertainty quantification

Adaptive estimation:

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

Uncertainty quantification:

- Size of honest confidence set is determined by smallest considered regularity
- (Data-driven constructions can be misleading)

Estimation versus uncertainty quantification

Adaptive estimation:

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

Uncertainty quantification:

- Size of honest confidence set is determined by smallest considered regularity
- (Data-driven constructions can be misleading)

SOLUTION 1: *be honest*
make conditional confidence statements

Estimation versus uncertainty quantification

Adaptive estimation:

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

Uncertainty quantification:

- Size of honest confidence set is determined by smallest considered regularity
- (Data-driven constructions can be misleading)

SOLUTION 1: *be honest*
make conditional confidence statements

SOLUTION 2: believe your prior 

Estimation versus uncertainty quantification

Adaptive estimation:

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

Uncertainty quantification:

- Size of honest confidence set is determined by smallest considered regularity
- (Data-driven constructions can be misleading)

SOLUTION 1: *be honest*
make conditional confidence statements

SOLUTION 2: believe your prior 

SOLUTION 3: determine which θ cause the trouble
argue that these are implausible

Uncertainty quantification for curve estimation

Polished tail sequences

Szabó et al, 2015

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

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

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

Polished tail sequences

Szabó et al, 2015

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

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

"Everything" is polished tail...:

Polished tail sequences

Szabó et al, 2015

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

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

“Everything” is polished tail...:

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

Polished tail sequences
Szabó et al, 2015

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

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

"Everything" is polished tail...:

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

Polished tail sequences

Szabó et al, 2015

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

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

"Everything" is polished tail...:

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

Credible balls in linear Gaussian inverse problems

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

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

Credible balls in linear Gaussian inverse problems

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

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

Credible ball:

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

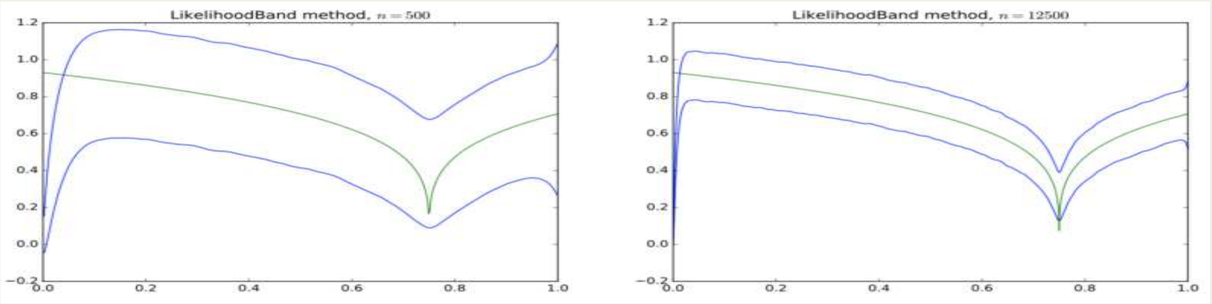
$$\hat{\theta}_n = \mathrm{E}(\theta|\, X^{(n)})$$
$$\Pi\big(\theta{:}\, \|\theta - \hat{\theta}_n\| < r|\, X^{(n)}\big) = 0.95$$

Thm For not too small M, uniformly in polished tail functions θ,

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

Similar results for empirical Bayes

Credible bands and other models



- Rousseau & Szabó, 2017-20: empirical Bayes and credible balls for general models
- Yoo 2017: ‘Bayesian Lepski method’ for adaptive credible bands; spline and wavelet priors
- Sniekers & vdV 2017, 19: bands for scaled Gaussian prior under ‘self-similarity’, ’good bias’ and ’discrete polished tail’.
- Belitser & Nurushev 2015-19: general projection estimators; ’excessive-bias restriction’
- Ray 2017: intersect credible set from weak and strong norms
- Hadji& Szabó, 2019: supersmooth priors
- ...

Uncertainty quantification for sparse high-dimensional parameters

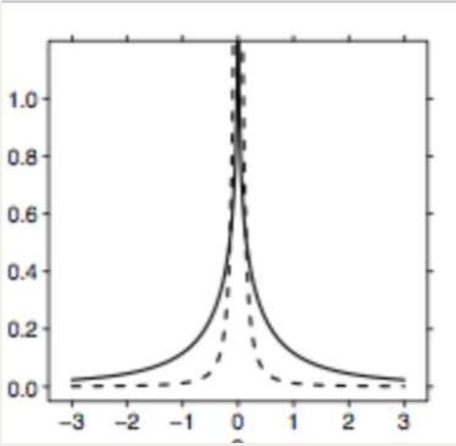
Horseshoe prior

Carvalho & Polson & Scott, 2010, van der Pas et al.

Data: $Y^n \sim N_n(\theta, I)$, for $\theta = (\theta_1, \ldots, \theta_n) \in \mathbb{R}^n$

Constructive definition of prior Π for $\theta \in \mathbb{R}^p$:

(1) Choose "sparsity level" τ from prior or by empirical Bayes
(2) Generate $\sqrt{\psi_1}, \ldots, \sqrt{\psi_n}$ iid from $\mathrm{Cauchy}^+(0, \tau)$
(3) Generate independent $\theta_i \sim N(0, \psi_i)$



prior density of θ_i

This prior gives optimal recovery of sparse vectors θ

Credible intervals

van der Pas, Szabó et al. 2018

Credible interval:

$$\hat{C}_{ni}(L) = \Big\{\theta_i \colon \big|\theta_i - \hat{\theta}_i\big| \le L\hat{r}_i\Big\}$$

$$\boxed{\begin{gathered}\hat{\theta} = \mathrm{E}(\theta \mid Y^n) \\ \Pi\big(\theta_i \colon |\theta_i - \hat{\theta}_i| \le \hat{r}_i \mid Y^n\big) = 0.95\end{gathered}}$$

Credible intervals

van der Pas, Szabó et al. 2018

Credible interval:

$$\hat{C}_{ni}(L) = \left\{\theta_i \colon \left|\theta_i - \hat{\theta}_i\right| \leq L\hat{r}_i\right\}$$

$$\boxed{\begin{gathered}\hat{\theta} = \mathrm{E}(\theta \mid Y^n) \\ \Pi\big(\theta_i \colon |\theta_i - \hat{\theta}_i| \leq \hat{r}_i \mid Y^n\big) = 0.95\end{gathered}}$$

$$\mathbf{S}_a := \{1 \leq i \leq n \colon |\theta_{0,i}| \leq 1/n\}$$

$$\mathbf{M}_a := \{1 \leq i \leq n \colon (s_n/n)\sqrt{\log(n/s_n)} \ll |\theta_{0,i}| \leq 0.99\sqrt{2\log(n/s_n)}\}$$

$$\mathbf{L}_a := \{1 \leq i \leq n \colon 1.001\sqrt{2\log n} \leq |\theta_{0,i}|\}$$

Credible intervals

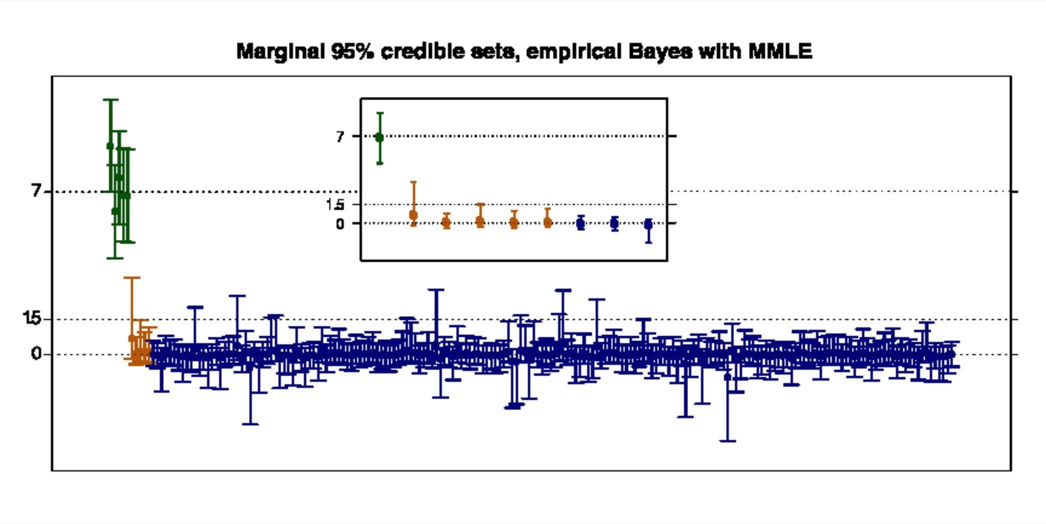
van der Pas, Szabó et al. 2018

Credible interval:

$$\hat{C}_{ni}(L) = \left\{\theta_i : \left|\theta_i - \hat{\theta}_i\right| \leq L\hat{r}_i\right\}$$

$$\hat{\theta} = \mathrm{E}(\theta \mid Y^n)$$
$$\Pi\left(\theta_i : |\theta_i - \hat{\theta}_i| \leq \hat{r}_i \mid Y^n\right) = 0.95$$

$$\mathbf{S}_a := \{1 \leq i \leq n : |\theta_{0,i}| \leq 1/n\}$$
$$\mathbf{M}_a := \{1 \leq i \leq n : (s_n/n)\sqrt{\log(n/s_n)} \ll |\theta_{0,i}| \leq 0.99\sqrt{2\log(n/s_n)}\}$$
$$\mathbf{L}_a := \{1 \leq i \leq n : 1.001\sqrt{2\log n} \leq |\theta_{0,i}|\}$$



marginal credible intervals for a single Y^n with $n = 200$ and $s_n = 10$

$\theta_1 = \cdots = \theta_5 = 7$, $\theta_6 = \cdots = \theta_{10} = 1.5$. Insert: credible sets 5 to 13

Credible intervals

van der Pas, Szabó et al. 2018

Credible interval:

$$\hat{C}_{ni}(L) = \Big\{\theta_i\colon \big|\theta_i - \hat{\theta}_i\big| \le L\hat{r}_i\Big\}$$

$$\hat{\theta} = \mathrm{E}(\theta |\, Y^n)$$
$$\Pi\big(\theta_i\colon |\theta_i - \hat{\theta}_i| \le \hat{r}_i |\, Y^n\big) = 0.95$$

$$\mathbf{S}_a\colon= \{1 \le i \le n\colon |\theta_{0,i}| \le 1/n\}$$
$$\mathbf{M}_a\colon= \{1 \le i \le n\colon (s_n/n)\sqrt{\log(n/s_n)} \ll |\theta_{0,i}| \le 0.99\sqrt{2\log(n/s_n)}\}$$
$$\mathbf{L}_a\colon= \{1 \le i \le n\colon 1.001\sqrt{2\log n} \le |\theta_{0,i}|\}$$

Thm For any $\gamma > 0$ and $\|\theta_0\|_0 \le s_n$,

$$P_{\theta_0}\Big(\frac{1}{\#\mathbf{S}_a}\#\{i \in \mathbf{S}_a\colon \theta_{0,i} \in \hat{C}_{ni}(L_{S,\gamma})\} \ge 1-\gamma\Big) \to 1,$$

$P_{\theta_0}\big(\theta_{0,i} \notin \hat{C}_{ni}(L)\big) \to 1,$ for any $i \in \mathbf{M}_a$ and any L

$$P_{\theta_0}\Big(\frac{1}{\#\mathbf{L_a}}\#\{i \in \mathbf{L}_a\colon \theta_{0,i} \in \hat{C}_{ni}(L_{L,\gamma})\} \ge 1-\gamma\Big) \to 1$$

Few false discoveries. Most easy discoveries made
Intermediate discoveries not made

False discovery rate

Castillo and Roquain, 2019

Data: $Y^n \sim N_n(\theta, I)$, for $\theta = (\theta_1, \ldots, \theta_n) \in \mathbb{R}^n$

- $\theta_i \overset{\text{iid}}{\sim} (1-\tau)\delta_0 + \tau G$, with $G =$ Laplace or Cauchy
- $\hat{\tau}$ determined by marginal empirical Bayes

$$\ell_\tau(x) = \Pi_\tau(\theta_i = 0 |\, X_i = x),$$
$$q_\tau(x) = \Pi_\tau(\theta_i = 0 |\, |X_i| \geq |x|)$$

Tests: Reject $H_0\colon \theta_{0,i} = 0$ if $\ell_{\hat{\tau}}(X_i) \leq t$ or $q_{\hat{\tau}}(X_i) \leq t$

Thm For $s_n \to \infty$ with $s_n \ll n^\nu$,

$$\sup_{\#(\theta_{0,i}\neq 0)\leq s_n} \mathrm{E}_{\theta_0} \frac{\#(i\colon \theta_{0,i} = 0, \text{rejected})}{\#(i\colon \text{rejected}) \vee 1} \lesssim t \log \frac{1}{t}$$

* Credible balls

A confidence ball

$$C_n(Y^n) = \{\theta \in \mathbb{R}^n \colon \|\theta - \hat{\theta}\| \leq \hat{r}\}$$

cannot have both:

- radius $\hat{r}$ of order the adaptive benchmark $\sqrt{s \log(n/s)}$ for sparsity,
- uniform coverage over multiple sparsity levels s

Meta Thm
A credible ball will cover "self-similar" parameters

** Simultaneous credible balls — impossibility of adaptation

General principle:
size of honest confidence set is determined by biggest model

Thm [Li, 1987]
If $P_{\theta_0}(C_n(Y^n) \ni \theta_0) \geq 0.95$, all $\theta_0 \in \mathbb{R}^n$, then $\mathrm{diam}(C_n(Y^n)) \gtrsim n^{-1/4}$, some θ_0

Thm [Nickl, van de Geer, 2013]
If $s_{1,n} \ll s_{2,n}$ and
$\mathrm{diam}(C_n(Y^n))$ is of optimal size, uniformly in $\|\theta_0\|_0 \leq s_{i,n}$ for $i = 1, 2$,
then $C_n(Y^n)$ cannot have uniform coverage over $\{\theta_0 \colon \|\theta_0\|_0 \leq s_{2,n}\}$.

Since the Bayesian procedure adapts to sparsity,
its credible balls *cannot* be honest confidence sets

[Optimal size is $\big((s_{i,n}/n)\log(n/s_{i,n})\big)^{1/2}$]

** Simultaneous credible balls — impossibility of adaptation — restricting the parameter

Coverage only when θ_0 *does not cause too much shrinkage*

Def [self-similarity]
For $s = \|\theta_0\|_0$ at least $0.001s$ coordinates of θ_0 satisfy

$$|\theta_{0,i}| \geq 1.001\sqrt{2\log(n/s)}.$$

** Simultaneous credible balls — impossibility of adaptation — restricting the parameter

Coverage only when θ_0 *does not cause too much shrinkage*

Def [self-similarity]
For $s = \|\theta_0\|_0$ at least $0.001s$ coordinates of θ_0 satisfy

$$|\theta_{0,i}| \geq 1.001\sqrt{2\log(n/s)}.$$

Def [excessive-bias restriction, Belitser & Nurushev, 2015]
$\|\theta\|_0 \leq s$ and $\exists \tilde{s}$ with $\tilde{s} \asymp \#\big(i\colon |\theta_{0,i}| \geq 1.001\sqrt{2\log(n/\tilde{s})}\big)$ and

$$\sum_{i:|\theta_{0,i}|\leq 1.001\sqrt{2\log(n/\tilde{s})}} \theta_{0,i}^2 \lesssim \tilde{s}\log(n/\tilde{s})$$

Excessive-bias restriction weaker than self-similarity
(Self-similarity allows to tighten up the sets $\mathbf{S}, \mathbf{M}, \mathbf{L}$)

** Simultaneous credible balls

Credible ball:

$$\hat{C}_n(L) = \{\theta\colon \|\theta - \hat{\theta}\| \le L\hat{r}\}$$

$$\hat{\theta} = \mathrm{E}(\theta |\, Y^n)$$
$$\Pi\big(\theta\colon \|\theta - \hat{\theta}\| \le \hat{r} |\, Y^n\big) = 0.95$$

Thm If $s_n/n \to 0$, for sufficiently large L,

$$\liminf_{n\to\infty} \inf_{\theta_0 \in \text{EBR}[s_n]} P_{\theta_0}\Big(\theta_0 \in \hat{C}_n(L)\Big) \ge 1 - \alpha$$

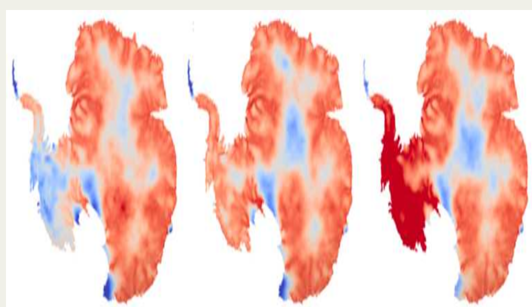
EBR$[s]$: vectors θ_0 that satisfy excessive bias restriction

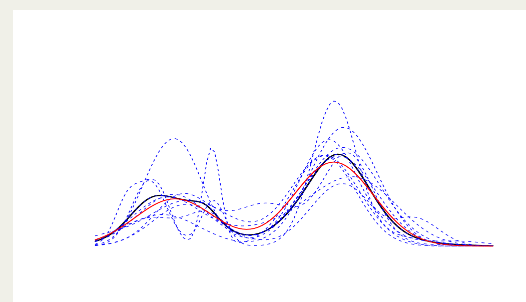
Closing remarks

Closing remarks

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*





Adaptive methods seem reasonable, even though
their confidence sets are dishonest