Bayesian inference in infinite dimensions

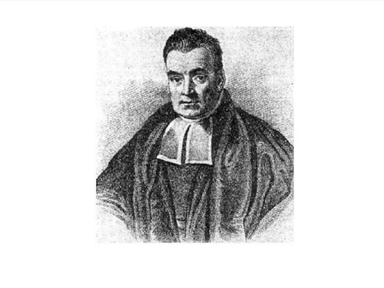
Aad van der Vaart
TU Delft, Netherlands

Mordell Lecture, Cambridge, May 4, 2023

Contents

Bayes

The Bayesian paradigm



- The unknown 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): $\mathrm{P}(X \in A, \theta \in B) = \int_B P_\theta(A)\, d\Pi(\theta)$.

- The scientist updates Π to the conditional distribution of θ given X, the posterior distribution:

$$\Pi(\theta \in B \mid X)$$

The Bayesian paradigm

- The unknown 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): $\mathrm{P}(X \in A, \theta \in B) = \int_B P_\theta(A)\, d\Pi(\theta)$.

- The scientist updates Π to the conditional distribution of θ given X, the posterior distribution:

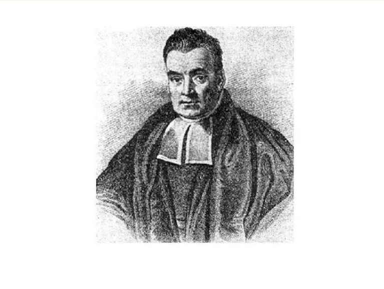
$$\Pi(\theta \in B |\, X)$$

If $P_\theta(A) = \int_A p_\theta(x)\, d\mu(x)$, then **Bayes's rule** gives

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

In general *disintegration*: $\mathrm{P}(X \in A, \theta \in B) = \int_A \Pi(\theta \in B |\, x)\, dP(x)$.

Thomas Bayes



Thomas Bayes (1702–1761) studied logic and theology in Edinburgh and was a Presbyterian minister in London and Turnbridge Wells.

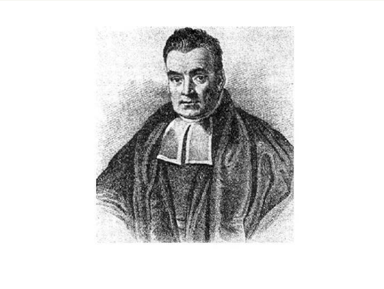
In his paper read to the Royal Society in 1763, he followed this argument with $\theta \in [0,1]$ *uniformly distributed* and X given θ *binomial* (n,θ).

$$d\Pi(\theta) = 1 \cdot d\theta, \qquad 0 < \theta < 1,$$

$$p_\theta(x) = \mathrm{P}(X = x|\,\theta) = \binom{n}{x}\theta^x(1-\theta)^{n-x}, \qquad x = 0, 1, \dots, n,$$

$$d\Pi(\theta|\,X) \propto \theta^X(1-\theta)^{n-X} \cdot 1 \cdot d\theta.$$

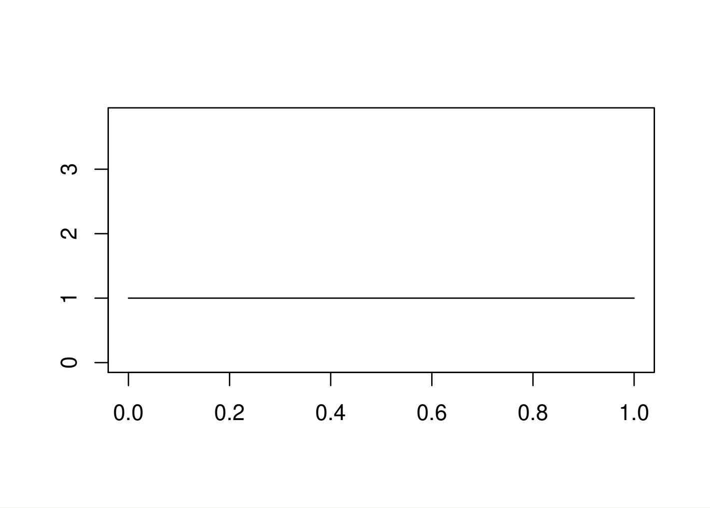
Thomas Bayes



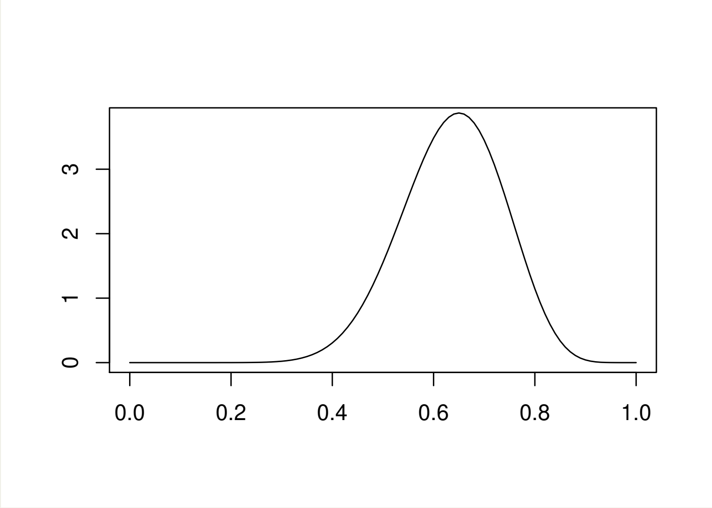
Thomas Bayes (1702–1761) studied logic and theology in Edinburgh and was a Presbyterian minister in London and Turnbridge Wells.

In his paper read to the Royal Society in 1763, he followed this argument with $\theta \in [0, 1]$ *uniformly distributed* and X given θ *binomial* (n, θ).

prior



posterior $(x = 13, n = 20)$



Parametric Bayes

Pierre-Simon Laplace (1749–1827) and Carl Friedrich Gauss (1777–1855) rediscovered Bayes's argument and applied it to general parametric models: models smoothly indexed by a Euclidean vector θ.

Ronald Aylmer Fisher (1890–1962) did not buy into it, but advocated maximum likelihood, with much success.

Parametric Bayes

Pierre-Simon Laplace (1749–1827) and Carl Friedrich Gauss (1777–1855) rediscovered Bayes's argument and applied it to general parametric models: models smoothly indexed by a Euclidean vector θ.

Ronald Aylmer Fisher (1890–1962) did not buy into it, but advocated maximum likelihood, with much success.

The Bayesian method regained popularity following the development of MCMC methods in the 1980/90s.

It is a method of choice for many applied scientists.

Nonparametric Bayes

In nonparametric statistics, the unknown θ is infinite-dimensional.

In nonparametric Bayesian statistics, prior and posterior are probability distributions on an infinite-dimensional space.

Bayes's formalism does not change, and his rule remains:

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

Nonparametric Bayes

In nonparametric statistics, the unknown θ is infinite-dimensional.

In nonparametric Bayesian statistics, prior and posterior are probability distributions on an infinite-dimensional space.

Bayes's formalism does not change, and his rule remains:

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

Nonparametric Bayesian statistics set off in the 1970/80/90s, but before 2000 it was thought not to work, except in very special cases.

Bayesian distribution estimation

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

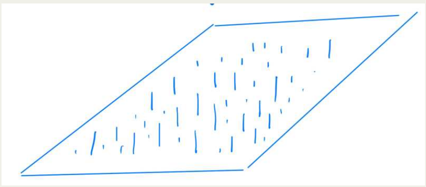
An obvious nonparametric estimator is the empirical distribution

$$\mathbb{P}_n = \frac{1}{n}\sum_{i=1}^{n}\delta_{X_i}$$

Bayesian approach starts with a prior over the set of distributions.

For instance, a random discrete distribution

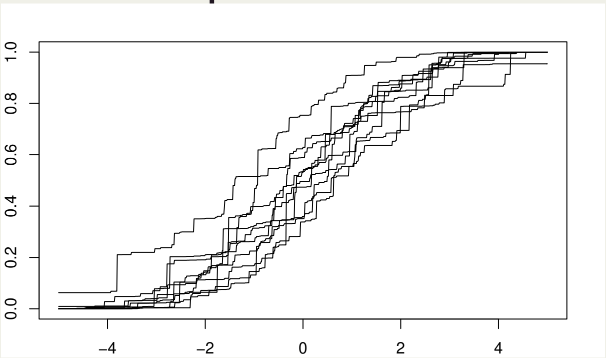
$$P = \sum_{i=1}^{\infty} W_i \delta_{\theta_i}$$



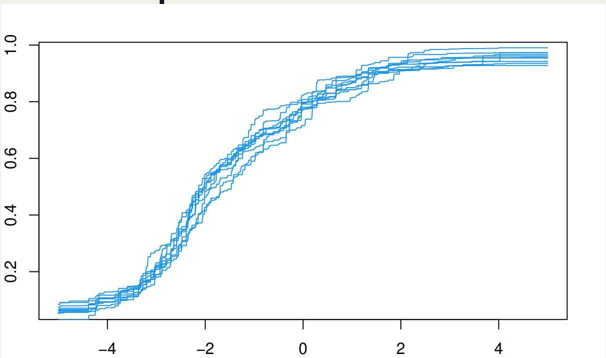
Bayesian distribution estimation – Dirichlet prior [Ferguson 1973]

Def $P = \sum_{i=1}^{\infty} W_i \delta_{\theta_i}$ is *Dirichlet process* if $W_i = V_i \prod_{j<i}(1 - V_j)$,
where $V_j \overset{\text{iid}}{\sim} \mathrm{Be}(1, M)$, $\theta_i \overset{\text{iid}}{\sim} G$

prior cdf



posterior cdf

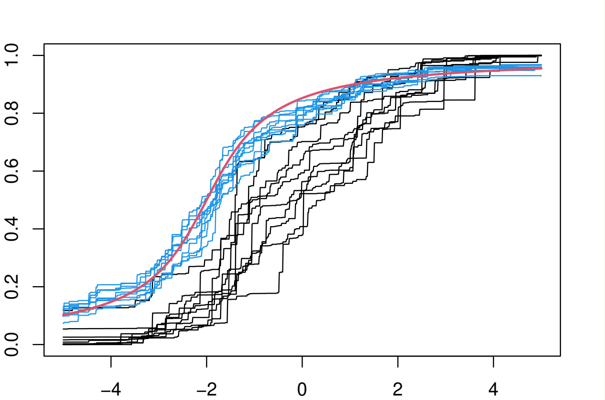


Sample of 100 from Cauchy (-2,1)

Bayesian distribution estimation – Dirichlet prior [Ferguson 1973]

Def $P = \sum_{i=1}^{\infty} W_i \delta_{\theta_i}$ is *Dirichlet process* if $W_i = V_i \prod_{j<i}(1 - V_j)$, where $V_j \overset{\text{iid}}{\sim} \mathrm{Be}(1, M)$, $\theta_i \overset{\text{iid}}{\sim} G$

prior and posterior cdf



Sample of 100 from Cauchy (-2,1).

Thm [Ferguson, Lo 1983-86]

If $P \sim DP(MG)$ and $X_1, \ldots, X_n |\, P \overset{\text{iid}}{\sim} P$, then

$$\mathrm{E}(P|\, X_1 \ldots, X_n) - \mathbb{P}_n = O(1/n)$$

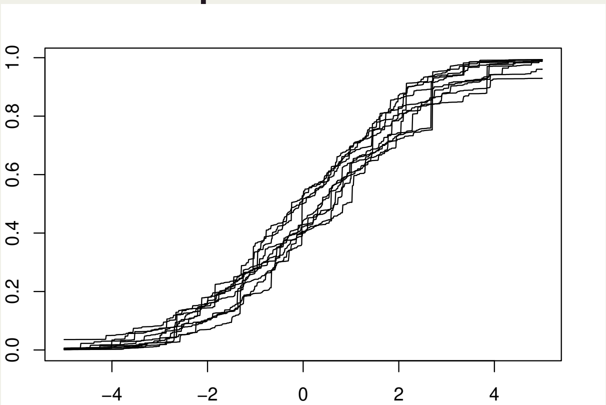
$$\sqrt{n}(P - \mathbb{P}_n)|\, X_1, \ldots, X_n \rightsquigarrow \text{Brownian bridge}$$

"Bayesian bootstrap"

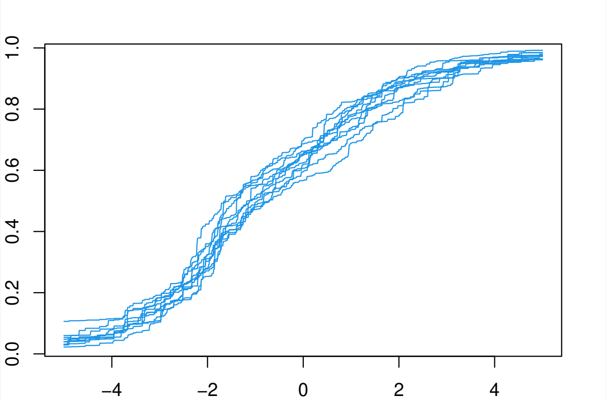
Bayesian distribution estimation – Pitman-Yor prior

Def $P = \sum_{i=1}^{\infty} W_i \delta_{\theta_i}$ is *Pitman-Yor process* if $W_i = V_i \prod_{j<i}(1 - V_j)$, where $V_j \overset{\text{iid}}{\sim} \mathrm{Be}(1-\sigma, M+j\sigma)$, $\theta_i \overset{\text{iid}}{\sim} G$

prior cdf



posterior cdf

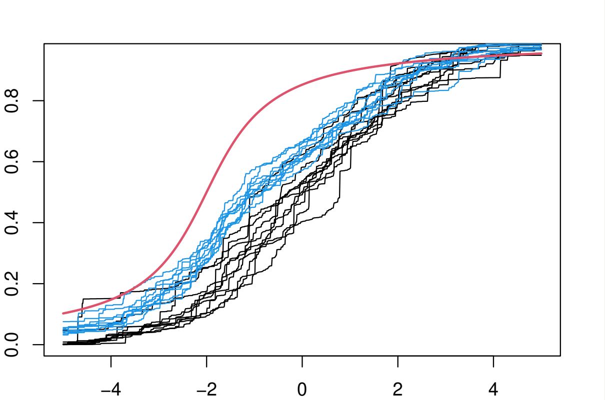


Sample of 100 from Cauchy (-2,1)

Bayesian distribution estimation – Pitman-Yor prior

Def $P=\sum_{i=1}^\infty W_i\delta_{\theta_i}$ is *Pitman-Yor process* if $W_i=V_i\prod_{j<i}(1-V_j)$, where $V_j\overset{\text{iid}}{\sim}\mathrm{Be}(1-\sigma, M+j\sigma)$, $\theta_i\overset{\text{iid}}{\sim}G$

prior and posterior



Sample of 100 from Cauchy (-2,1).

Thm [James 2006, Franssen vdV 2022]

If $P\sim PY(MG,\sigma)$ and $X_1,\ldots,X_n|\,P\overset{\text{iid}}{\sim}P$, then

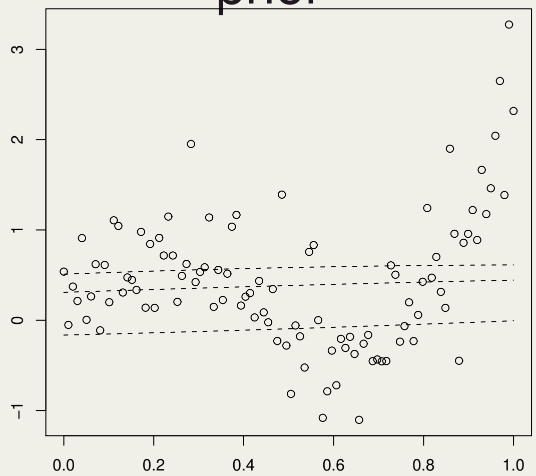
$$P|\,X_1,\ldots,X_n\rightsquigarrow\delta_{(1-\lambda)P_0^d+\lambda(1-\sigma)P_0^c+\sigma\lambda G}$$

$$\sqrt{n}\Big(P-\mathbb{P}_n-\frac{\sigma K_n}{n}(G-\tilde{\mathbb{P}}_n)\Big)\Big|\,X_1,\ldots,X_n\rightsquigarrow\text{Gaussian.}$$

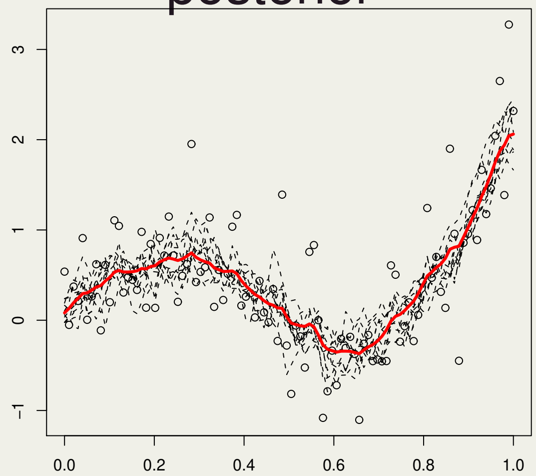
Bayesian regression

Data: $Y_i = \theta(x_i) + \varepsilon_i$, for $i = 1, \ldots, n$.

prior



posterior

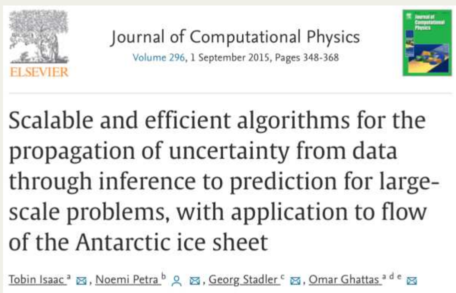


Bayesian data assimilation [Stuart 2010, ...]

Data: $Y_i = u_\theta(x_i) + \varepsilon_i$, for u_θ solution of a PDE with unknown θ

Bayesian data assimilation [Stuart 2010, ...]

Data: $Y_i = u_\theta(x_i) + \varepsilon_i$, for u_θ solution of a PDE with unknown θ



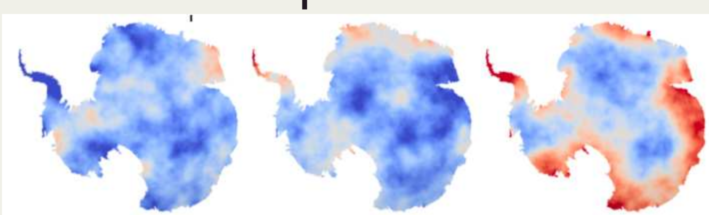
ELSEVIER
Journal of Computational Physics
Volume 296, 1 September 2015, Pages 348-368

Scalable and efficient algorithms for the propagation of uncertainty from data through inference to prediction for large-scale problems, with application to flow of the Antarctic ice sheet

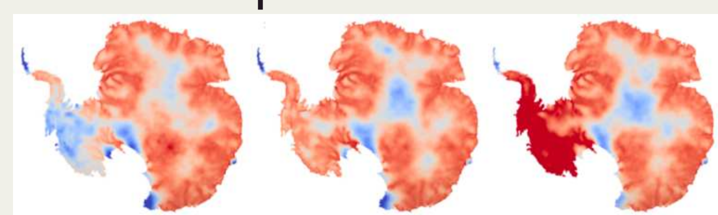
Tobin Isaac [a], Noemi Petra [b], Georg Stadler [c], Omar Ghattas [a d e]

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

prior



posterior



Bayesian data assimilation [Stuart 2010, ...]

Data: $Y_i = u_\theta(x_i) + \varepsilon_i$, for u_θ solution of a PDE with unknown θ

If $\varepsilon_1, \ldots, \varepsilon_n$ are i.i.d. Gaussian, then Bayes's rule gives

$$d\Pi(\theta \mid Y_1, \ldots, Y_n) \propto \prod_{i=1}^n e^{-(Y_i - u_\theta(x_i))^2/2\sigma^2} \, d\Pi(\theta).$$

Thm [posterior mode] [Wahba 1978, Dashti et al 2013]

For Gaussian prior on θ with RKHS norm $\|\cdot\|$

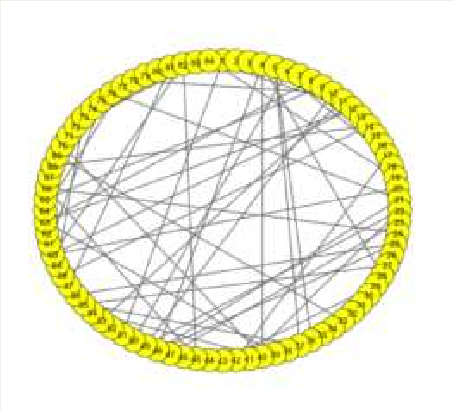
$$\lim_{\varepsilon \downarrow 0} \operatorname*{argmax}_\theta \Pi\big(B(\theta, \varepsilon) \mid Y^{(n)}\big) = \operatorname*{argmin}_\theta \sum_{i=1}^n \big(Y_i - u_\theta(x_i)\big)^2 + \sigma^2 \|\theta\|^2$$

Full posterior distribution quantifies uncertainty

Other settings

- density estimation
- high-dimensional inference
- networks
- deep learning
- diffusion processes
- hierarchical models
- ...

Prior used to model sparsity or network structure or “to borrow strength”.



network of genes involved in lung cancer

[Kpogbezan et al. 2016]

Does it work?

Frequentist Bayes

Assume 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

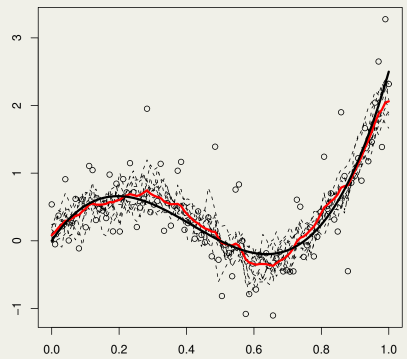
Assume data X is generated according to a given parameter θ_0.
Consider the posterior $\Pi(\theta \in \cdot \mid X)$ as a given random measure.

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 $X^{(n)}$, with information increasing as $n \to \infty$.

Parametric Bayes

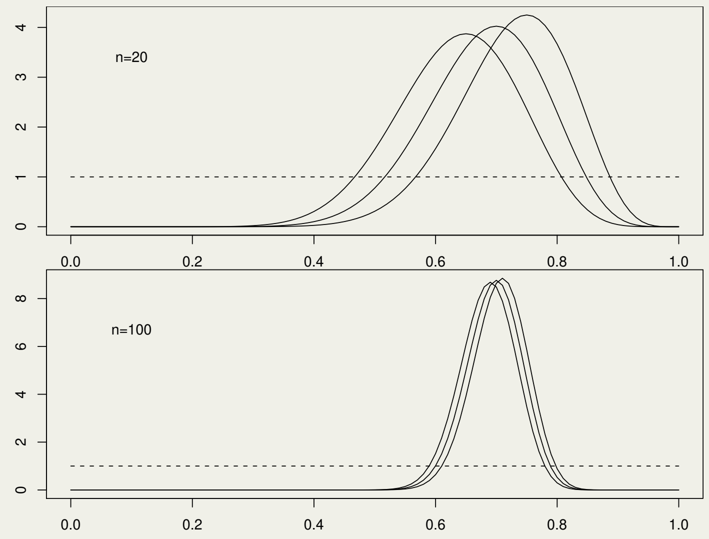
[Laplace, Bernstein, von Mises, Le Cam 1986]

Data: $X_1, \ldots, X_n$ i.i.d. sample from density $x \mapsto p_\theta(x)$ that is **smoothly** and **identifiably** parametrized by $\theta \in \mathbb{R}^d$.

Thm For any prior with positive density,

$$\mathrm{E}_{\theta_0} \Big\| \Pi_n(\cdot \,|\, X_1, \ldots, X_n) - N_d\big(\tilde\theta_n, \frac{1}{n} i_{\theta_0}^{-1}\big)(\cdot) \Big\|_{TV} \to 0.$$

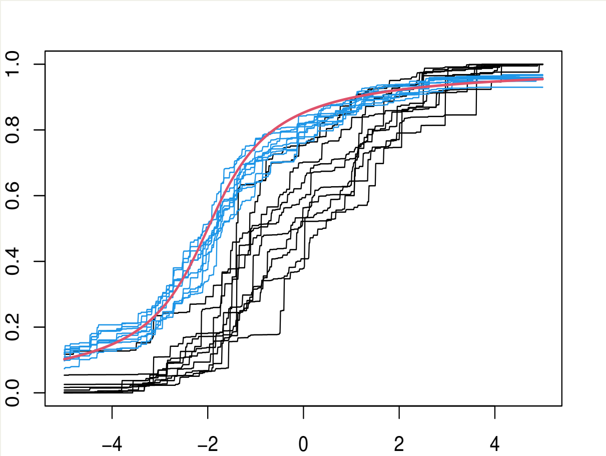
Here $\tilde\theta_n = \tilde\theta_n(X_1, \ldots, X_n)$ satisfy $\sqrt{n}(\tilde\theta_n - \theta_0) \rightsquigarrow N_d(0, i_{\theta_0}^{-1})$.

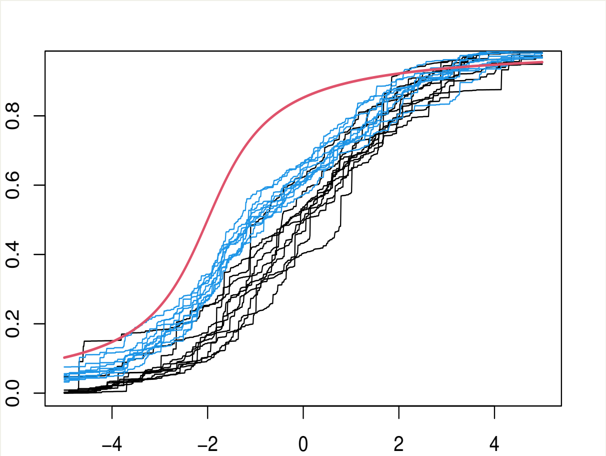


The prior washes out.

Nonparametric Bayes

For infinite-dimensional θ, the prior matters!





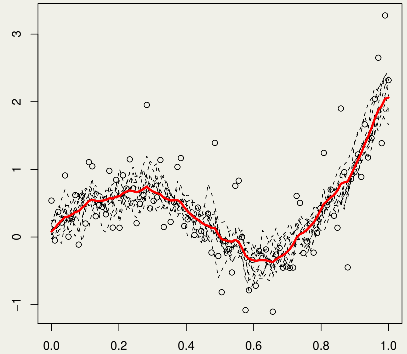
Recovery

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 enough M,

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

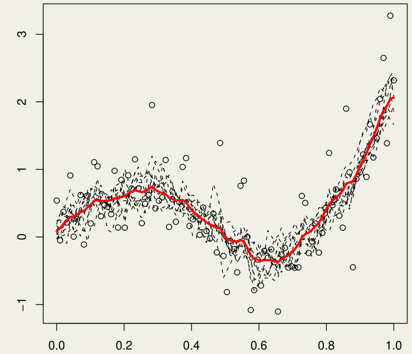


Rate of contraction

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

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

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



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

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

Minimaxity and adaptation

Minimax rate

To model Θ_β is attached an optimal rate of recovery defined by the minimax criterion

$$\varepsilon_{n,\beta} = \inf_T \sup_{\theta \in \Theta_\beta} \mathrm{E}_\theta d\big(T(X), \theta\big).$$

Adaptation

Given models $(\Theta_\beta : \beta \in B)$, there often exists a single T that attains the minimax rate for every β.

Minimaxity and adaptation

Minimax rate

To model Θ_β is attached an optimal rate of recovery defined by the minimax criterion

$$\varepsilon_{n,\beta} = \inf_T \sup_{\theta\in\Theta_\beta} \mathrm{E}_\theta d\big(T(X),\theta\big).$$

Adaptation

Given models $(\Theta_\beta\colon \beta \in B)$, there often exists a single T that attains the minimax rate for every β.

A good prior gives a posterior such that

$$\forall\beta\colon\quad \forall\theta_0\in\Theta_\beta\colon\quad \mathrm{E}_{\theta_0}\Pi_n\big(\theta\colon d(\theta,\theta_0) > M\epsilon_{n,\beta}|\, X^{(n)}\big)\to 0.$$

Basic contraction theorem – direct problem

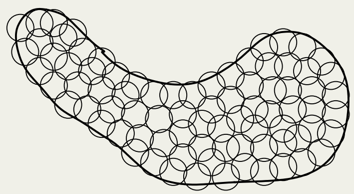
[Ghosal, Ghosh, vdV, 2000]

Data: Sample of size n from density p.

Prior Π, on set of densities $\mathcal{P}$, with convex metric d, $d \leq$ Hellinger.

Kolmogorov entropy

$N(\epsilon, \mathcal{P}, d)$ is the minimal number of d-balls of radius ϵ needed to cover $\mathcal{P}$.



Thm If

$$\Pi\Big(p\colon P_0\Big(\log\frac{p_0}{p}\Big) < \varepsilon_n^2\Big) \geq e^{-n\epsilon_n^2}, \qquad \text{(prior mass)}$$

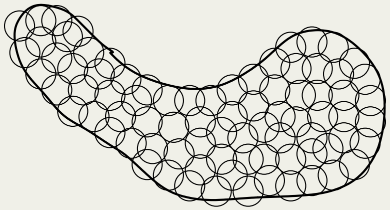
$$\exists \mathcal{P}_n\colon \quad \log N\big(\epsilon_n, \mathcal{P}_n, d\big) \leq n\epsilon_n^2 \quad \text{and} \quad \Pi(\mathcal{P}_n^c) \leq e^{-4n\epsilon_n^2}, \qquad \text{(complexity)}$$

then posterior rate of contraction at p_0 is ϵ_n.

[Hellinger metric: $h(p,q) = \|\sqrt{p} - \sqrt{q}\|_2$.]

*Interpretation

Let $p_1, \ldots, p_N \in \mathcal{P}$ be maximal set with $d(p_i, p_j) \geq \epsilon_n, \quad N \asymp N(\epsilon_n, \mathcal{P}, d)$



The complexity bound says

$$N \asymp N(\epsilon_n, \mathcal{P}, d) \leq e^{n\epsilon_n^2}.$$

A "uniform prior" would give each ball of radius ε_n mass

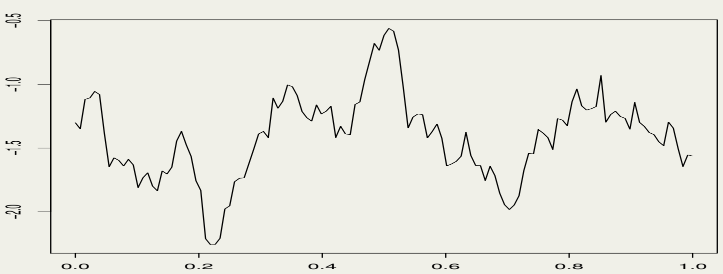
$$\Pi\big(B(p_j, \varepsilon_n)\big) \asymp \frac{1}{N} \geq e^{-n\epsilon_n^2}.$$

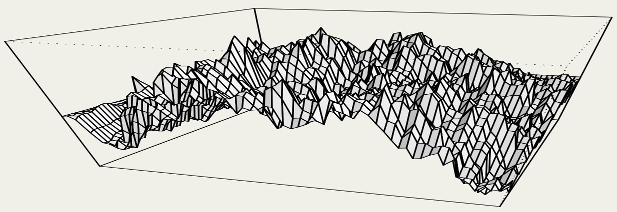
This is the prior mass bound.

Suggestion:
Contraction rate is ε_n at every $p_0 \in \mathcal{P}$ for priors that
"distribute mass uniformly over $\mathcal{P}$, at discretization level ϵ_n".

Gaussian process prior

Stochastic process $W = (W_t: t \in T)$ gives prior on functions $\theta: T \to \mathbb{R}$.



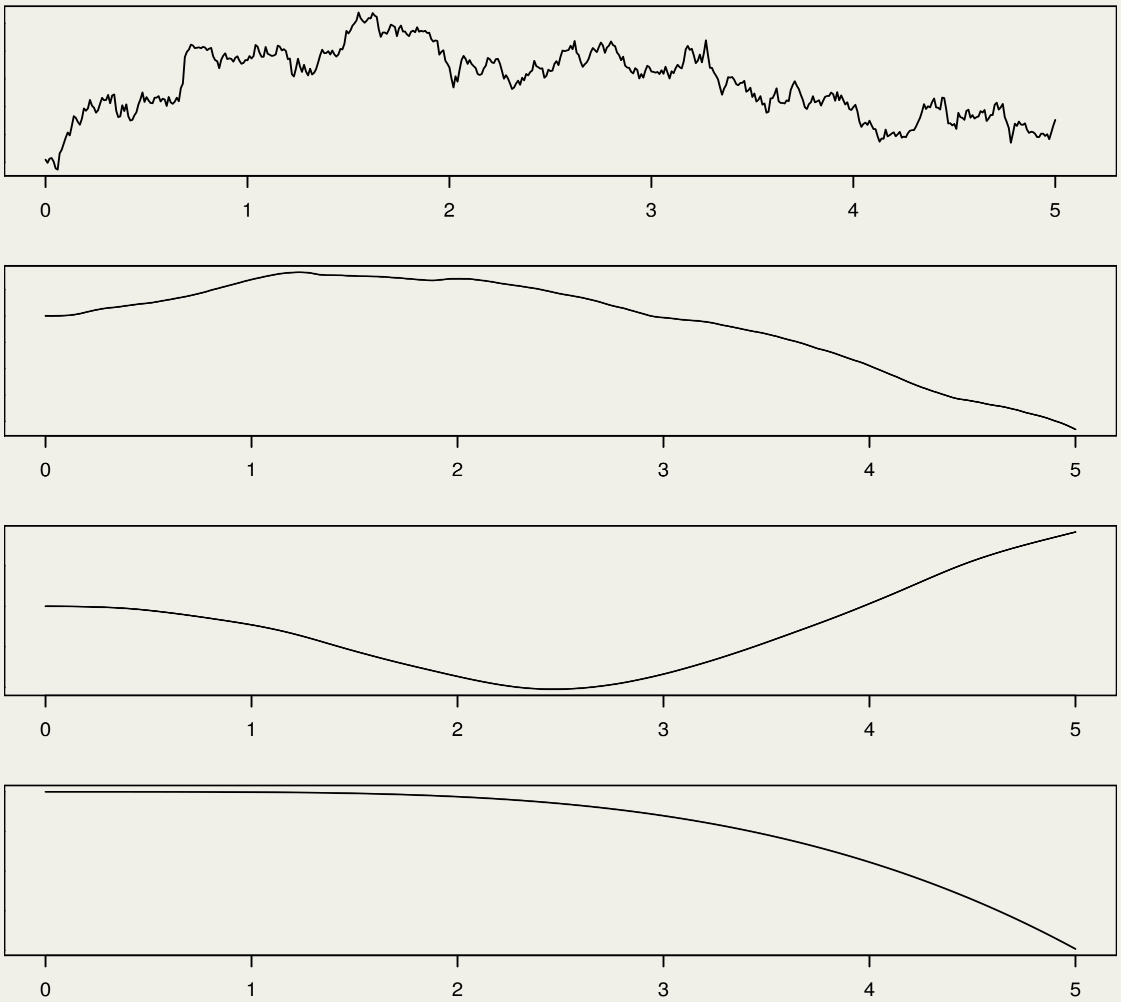


W is a **Gaussian process** if
$\sum_{i=1}^k \alpha_i W_{t_i}$ is Gaussian, for every $\alpha_1, \ldots, \alpha_k, t_1, \ldots, t_k$.

For every positive-definite $c: T \times T \to \mathbb{R}$, there exists W with

$$c(s,t) = \mathrm{E} W_s W_t, \qquad s, t \in T.$$

Example: Brownian motion and its primitives



0, 1, 2 and 3 times integrated Brownian motion

Posterior contraction rates for Gaussian priors [vdV+van Zanten, 2007-2011]

View Gaussian process W as map into Banach space $(\mathbb{B}, \|\cdot\|)$.

It comes with a RKHS $\mathbb{H}$.

Thm If statistical distances combine appropriately with $\|\cdot\|$, then contraction rate is ε_n if both

$$\mathrm{P}\big(\|W\| < \varepsilon_n\big) \geq e^{-n\varepsilon_n^2} \quad \text{and} \quad \inf_{h\in\mathbb{H}:\|h-\theta_0\|<\varepsilon_n} \|h\|_{\mathbb{H}}^2 \leq n\varepsilon_n^2$$

Posterior contraction rates for Gaussian priors

[vdV+van Zanten, 2007-2011]

View Gaussian process W as map into Banach space $(\mathbb{B}, \|\cdot\|)$.

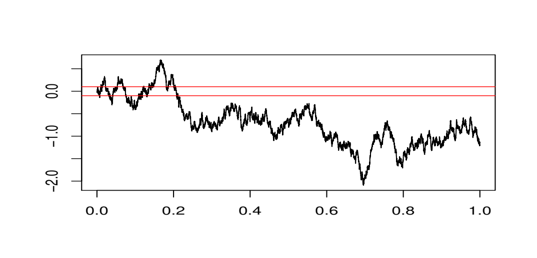
It comes with a RKHS $\mathbb{H}$.

Thm If statistical distances combine appropriately with $\|\cdot\|$, then contraction rate is ε_n if both

$$\mathrm{P}\big(\|W\| < \varepsilon_n\big) \geq e^{-n\varepsilon_n^2} \quad \text{and} \quad \inf_{h\in\mathbb{H}:\|h-\theta_0\|<\varepsilon_n} \|h\|_{\mathbb{H}}^2 \leq n\varepsilon_n^2$$

Example Integrated Brownian motion viewed as map in $C[0,1]$ has

$$\mathbb{H} = H^{k+1} = \big\{h: \|h\|_{\mathbb{H}} := \|h^{(k+1)}\|_2 < \infty\big\}$$



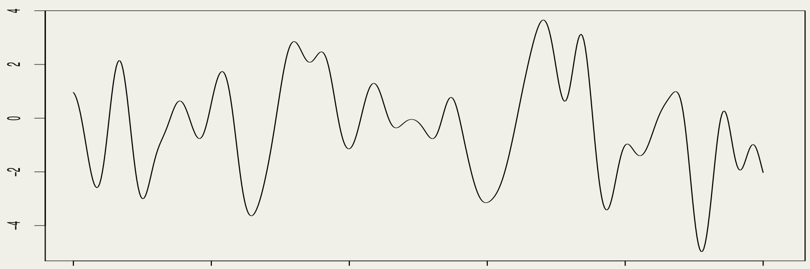
$$-\log \mathrm{P}\big(\|W\|_\infty < \varepsilon\big) \asymp (1/\varepsilon)^{2/(2k+1)}$$

Contraction rate $n^{-(\beta\wedge(k+1/2))/(2k+2)}$ if $\theta_0 \in C^\beta$. Optimal if $k+1/2=\beta$.

Regression with square exponential prior

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

Prior Gaussian with $\mathrm{cov}(\theta_{\tau x}, \theta_{\tau x'}) = e^{-\|x-x'\|^2 \tau^2}$.



$\tau = 1$

$$\mathrm{P}\Big(\sup_{0<x<1} |\theta(x)| < \varepsilon\Big) \gtrsim e^{-C(\log \varepsilon^{-1})^{1+d/2}}.$$

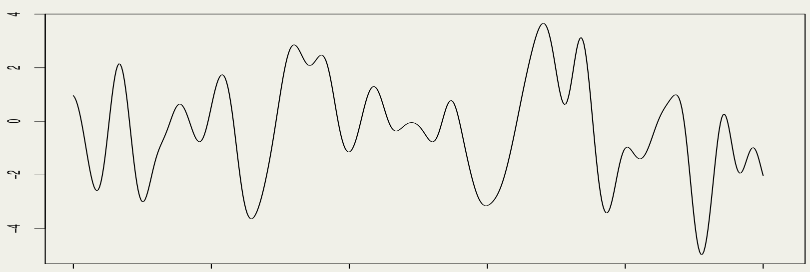
Thm If τ fixed,

- if θ_0 analytic, then contraction rate nearly $n^{-1/2}$.
- if θ_0 only ordinary smooth, then contraction rate $(\log n)^{-k}$.

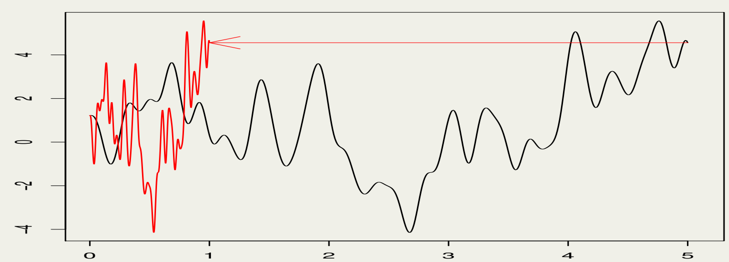
Regression with square exponential prior

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

Prior Gaussian with $\mathrm{cov}(\theta_{\tau x}, \theta_{\tau x'}) = e^{-\|x-x'\|^2 \tau^2}$.



$\tau = 1$



$\tau \gg 1$

Thm If τ fixed,

- if θ_0 analytic, then contraction rate nearly $n^{-1/2}$.
- if θ_0 only ordinary smooth, then contraction rate $(\log n)^{-k}$.

Thm If $\tau^d \sim \Gamma(a, b)$,

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

*Rates in sequence space

Data: $X^{(n)} = \theta + n^{-1/2}\dot{\mathbb{W}}$

Prior $\theta_i \overset{\text{ind}}{\sim} N(0, \tau^2 i^{-2\alpha-1})$ on coefficients on orthonormal basis

Lem For all $s < \alpha$, prior concentrates on

$$G^s = \{\theta \in \ell_2: \sum_{i=1}^{\infty} i^{2s}\theta_i^2 < \infty\}.$$

Thm τ fixed
If $\theta_0 \in G^\beta$, then contraction rate $n^{-(\alpha\wedge\beta)/(1+2\alpha)}$.

Thm $\tau^{-1} \sim \Gamma(c, d)$
If $\theta_0 \in G^\beta$ and $\beta < \alpha + 1/2$, then contraction rate $n^{-\beta/(1+2\beta)}$.

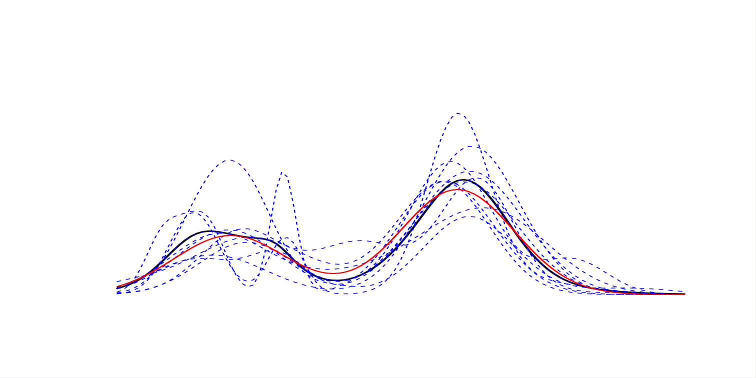
*Dirichlet normal mixtures

[Ghosal vdV Rousseau Kruijer Tokdar Shen 2001–2013]

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

Prior on p

- $F \sim$ Dirichlet process.
- $1/\tau \sim \Gamma(c, d)$, independent of F.
- $p_{F,\tau}(x) = \int \frac{1}{\tau}\phi\left(\frac{x-z}{\tau}\right) dF(z)$.



Thm Hellinger contraction rate is

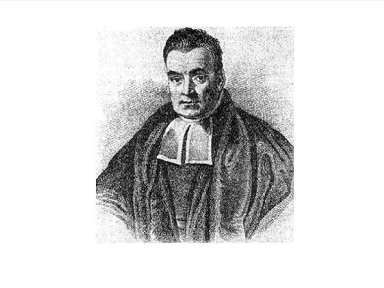
- nearly $n^{-1/2}$ if $p_0 = p_{F_0,\tau_0}$, some F_0, τ_0.
- nearly $n^{-\beta/(2\beta+d)}$ if p_0 has β derivatives and small tails.

Recovery: summary

Recovery is best if prior matches truth.
Mismatch slows down, but does not prevent, recovery.
Mismatch can be prevented by using a hyperparameter.

Uncertainty quantification

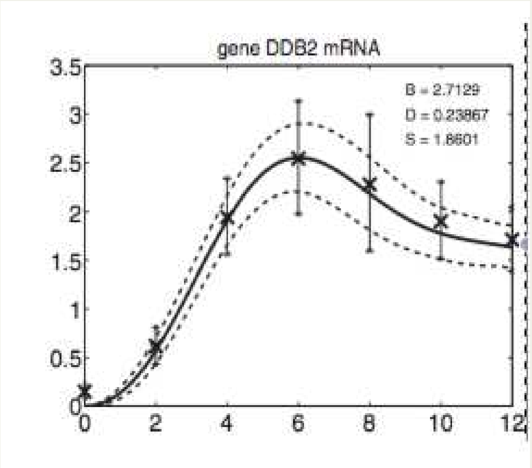
Credible sets



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

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

credible bands $C(X)$ are natural

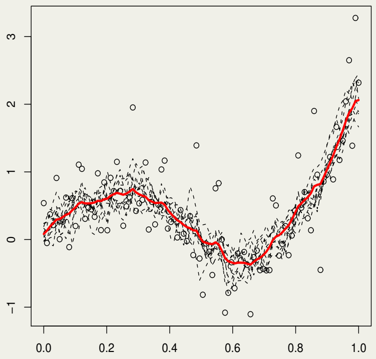


Estimated abundance of a transcription factor as function of time:
posterior mean curve and 95% credible bands
(Gao et al. *Bioinformatics* 2008)

Are credible sets confidence sets?

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

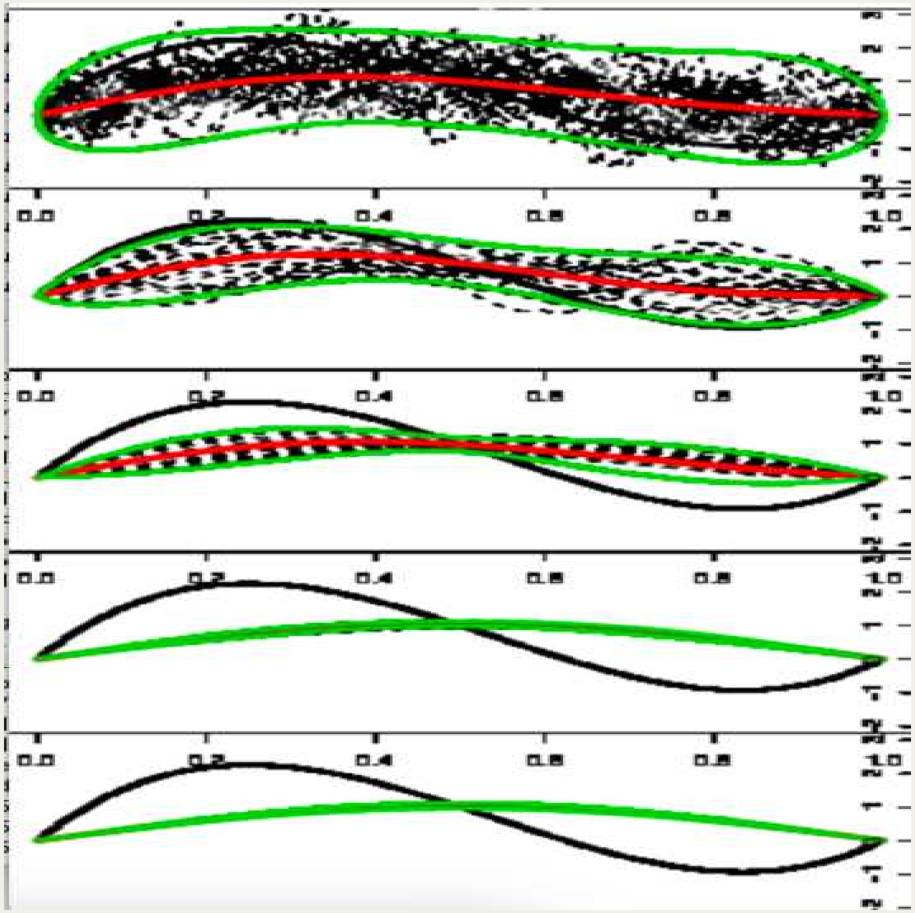
- Finite-dimensional θ: yes *(by Bernstein-von Mises)*
- Smooth projections of infinite-dimensional θ: yes
- Truly nonparametric θ: no



Does spread of posterior give correct order of uncertainty?

Different answers for deterministic bandwidth and data-driven bandwidth

Deterministic bandwidth: coverage requires undersmoothing



True θ_0 (black), posterior mean (red)

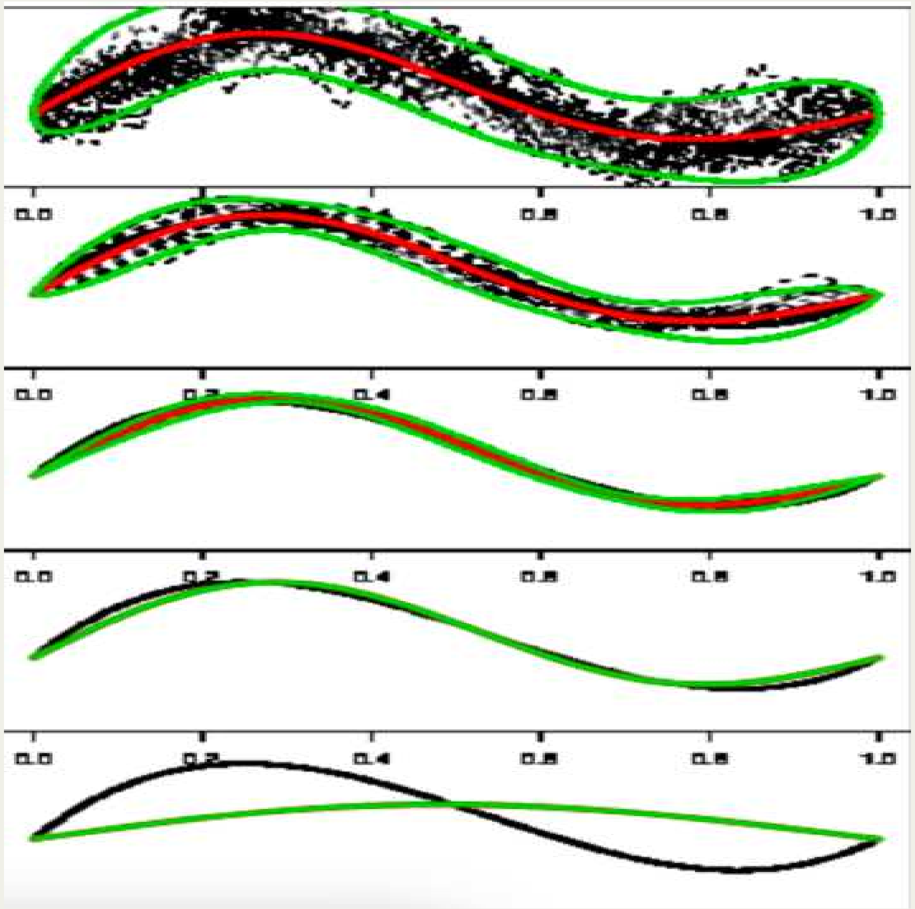
- $\theta = \sum_{i=1}^{\infty} \theta_i e_i$
- Truth: $\theta_{0,i} \asymp i^{-1-2\beta}$
- Prior: $\theta_i \overset{\text{ind}}{\sim} N(0, i^{-1-2\alpha})$

Top to bottom: increasing α

Black: truth

Green: bands

Deterministic bandwidth: coverage requires undersmoothing



True θ_0 (black), posterior mean (red)

- $\theta = \sum_{i=1}^{\infty} \theta_i e_i$
- 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

**Deterministic bandwidth: coverage requires undersmoothing

Data: $X^{(n)} = \theta + n^{-1/2}\dot{\mathbb{W}}$

Prior $\theta_i \overset{\text{ind}}{\sim} N(0, i^{-2\alpha-1})$ for coefficients on orthonormal basis

$$\|\theta\|_{G^s}^2 = \sum_{i=1}^{\infty} i^{2s}\theta_i^2$$

Thm $\hat{\theta}_n = \mathrm{E}(\theta|\, X^{(n)})$

- If $\alpha < \beta$, then $\mathrm{E}\big(\|\theta - \hat{\theta}_n\|_{\ell_2}^2 |\, X^{(n)}\big) \gg \|\mathrm{E}\hat{\theta}_n - \theta_0\|_{\ell_2}^2$, all $\theta_0 \in G^\beta$.
- If $\alpha > \beta$, then $\mathrm{E}\big(\|\theta - \hat{\theta}_n\|_{\ell_2}^2 |\, X^{(n)}\big) \ll \|\mathrm{E}\hat{\theta}_n - \theta_0\|_{\ell_2}^2$, some $\theta_0 \in G^\beta$.

Cor For $C_n = \{\theta\colon \|\theta - \hat{\theta}_n\|_{\ell_2} < R_n\}$, for $\Pi(C_n|\, X^{(n)}) = 0.95$.

- If $\alpha < \beta$, then $\mathrm{P}_{\theta_0}(\theta_0 \in C_n) \to 1$, all $\theta_0 \in G^\beta$.
- If $\alpha > \beta$, then $\mathrm{P}_{\theta_0}(\theta_0 \in C_n) \to 0$, some $\theta_0 \in G^\beta$.

Data-driven bandwidth

Family of priors Π_τ of varying smoothness τ.

Examples

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

Data-driven bandwidth

Family of priors Π_τ of varying smoothness τ.

Prior on bandwidth τ gives adaptive recovery:
for smoother true function better reconstruction

Data-driven bandwidth

Family of priors Π_τ of varying smoothness τ.

Prior on bandwidth τ gives adaptive recovery:
for smoother true function better reconstruction

This implies that data-driven posteriors *must* be tricked by some inconvenient truths and sometimes be misleading in their uncertainty quantification

- Estimation: $\forall\beta\colon\ \forall\theta\in\Theta_\beta\colon$ rate $\varepsilon_{n,\beta}$.
- Uncertainty: $\forall\theta\in\cup_\beta\Theta_\beta\colon \mathrm{P}_\theta\big(\theta\in C(X)\big)\geq 0.95.$

"We may know that a given statistical procedures is optimal in many settings simultaneously, but we cannot know how good it is" *[Lucien Birgé]*

Inconvenient truths

$$\theta_1, \theta_2, \ldots, \theta_{N_1}, 0, 0, \ldots, 0, \theta_{n_2}, \theta_{n_2+1}, \ldots, \theta_{N_2}, 0, 0, \ldots, 0, \theta_{n_3}, \ldots, \theta_{N_3}, 0, \ldots$$

Length of zero runs increasing.

Convenient truths [Szabó et al 2015]

Def $\theta \in \ell_2$ satisfies the *polished tail condition* if

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

Convenient truths [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 θ from a prior $\theta_i \overset{\text{ind}}{\sim} N(0, ci^{-\alpha-1/2})$ is polished tail

Data-driven bandwidth – credible balls

[Szabó et al 2015, Sniekers et al 2020]

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

- 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

$$\hat{C}_{n,M} := \{\theta : \|\theta - \hat{\theta}_n\| < MR\}$$

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

Thm For not too small M, uniformly in polished tail functions θ_0,

$$\mathrm{P}_{\theta_0}\big(\theta_0 \in \hat{C}_{n,M}\big) \to 1$$

Inverse problems

Inverse problems

Data: $Y^{(n)} = u_\theta + n^{-1/2}\dot{\mathbb{W}}$, for u_θ solution to a PDE

Estimation of u_θ is ordinary regression problem

However, contraction rate for u_θ does not imply rate for θ

The prior must regularize both u_θ and the inverse $u_\theta \mapsto \theta$.

Linear inverse problems

[Ray 2017, Knapik Salomond 2018, vdV&Yan 2020]

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

Smoothness scale $\|\theta\|_{G^s}^2 = \sum_{i=1}^\infty i^{2s/d}\theta_i^2$ for $\theta = \sum_{i=1}^\infty \theta_i e_i$.

Smoothing property $K\colon G^0 \to L$, Hilbert space, with

$$\|K\theta\|_L \asymp \|\theta\|_{G^{-p}}.$$

Galerkin reconstruction $\theta^{(j)} = K^{-1}Q_jK\theta$, for $Q_j\colon L \to K\,\mathrm{lin}\,(e_1, ..., e_j)$.

Thm If $\exists j_n \lesssim n\varepsilon_n^2$ and $\eta_n \gtrsim \varepsilon_n j_n^p \vee j_n^{-\beta}$ with

$$\Pi\big(\theta\colon \|K\theta - K\theta_0\|_L < \varepsilon_n\big) \gtrsim e^{-n\varepsilon_n^2},$$
$$\Pi\big(\theta\colon \|\theta^{(j_n)} - \theta\|_{G^0} > \eta_n\big) \le e^{-4n\varepsilon_n^2},$$

then contraction rate for $\theta_0 \in G^0$ is η_n.

Linear Gaussian inverse problems [vdV&Yan 2020, Koers Szabó vdV 2023]

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

- Prior $\theta_i \overset{\text{ind}}{\sim} N(0, \tau^2 i^{-2\alpha/d-1})$
- Prior $\tau^{-1} \sim \Gamma(c, d)$

Smoothness scale $\|\theta\|_{G^s}^2 = \sum_{i=1}^\infty i^{2s/d}\theta_i^2$ for $\theta = \sum_{i=1}^\infty \theta_i e_i$.

Smoothing property $K\colon G^0 \to L$, Hilbert space, with

$$\|K\theta\|_L \asymp \|\theta\|_{G^{-p}}.$$

‘

Thm For $\delta < \beta < \alpha + 1/2$, contraction rate for $\theta_0 \in G^\delta$ is $n^{-(\beta-\delta)/(2\beta+2p+d)}$.

Inverse problems from PDEs [Stuart 2010, Nickl,...]

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

Forward map u_θ solves a PDE that depends on θ.

Examples

- Schrödinger [Nickl 2020] $\frac{1}{2}\Delta u = \theta u$.
- Heat with absorption [Kekkonen 2022] $\partial_t u - \frac{1}{2}\Delta u = \theta u$.
- Non-abelian X-ray transform [Monard Nickl Paternain 2019, 2021]
- Divergence/Darcy [Abraham Nickl 2019, Bohr 2022] $\nabla \cdot (\theta \nabla u) = g$.
- Navier-Stokes [Nickl Titi 2023]
- ...

Linear methods for nonlinear PDE inverse problems [Koers Szabó vdV 2023]

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

$$\begin{cases} \mathcal{L}u_\theta = c(\theta, u_\theta), & \text{on } \Omega, \\ u_\theta = g, & \text{on } \Gamma \subseteq \partial\Omega. \end{cases}$$

$\mathcal{L}$ linear and rich enough so that there exists (Lipschitz) e with

$$\theta = e(\mathcal{L}u_\theta).$$

Data: $Y^{(n)} = K\mathcal{L}u_\theta + n^{-1/2}\dot{\mathbb{W}}$, for $K = \mathcal{L}^{-1}$

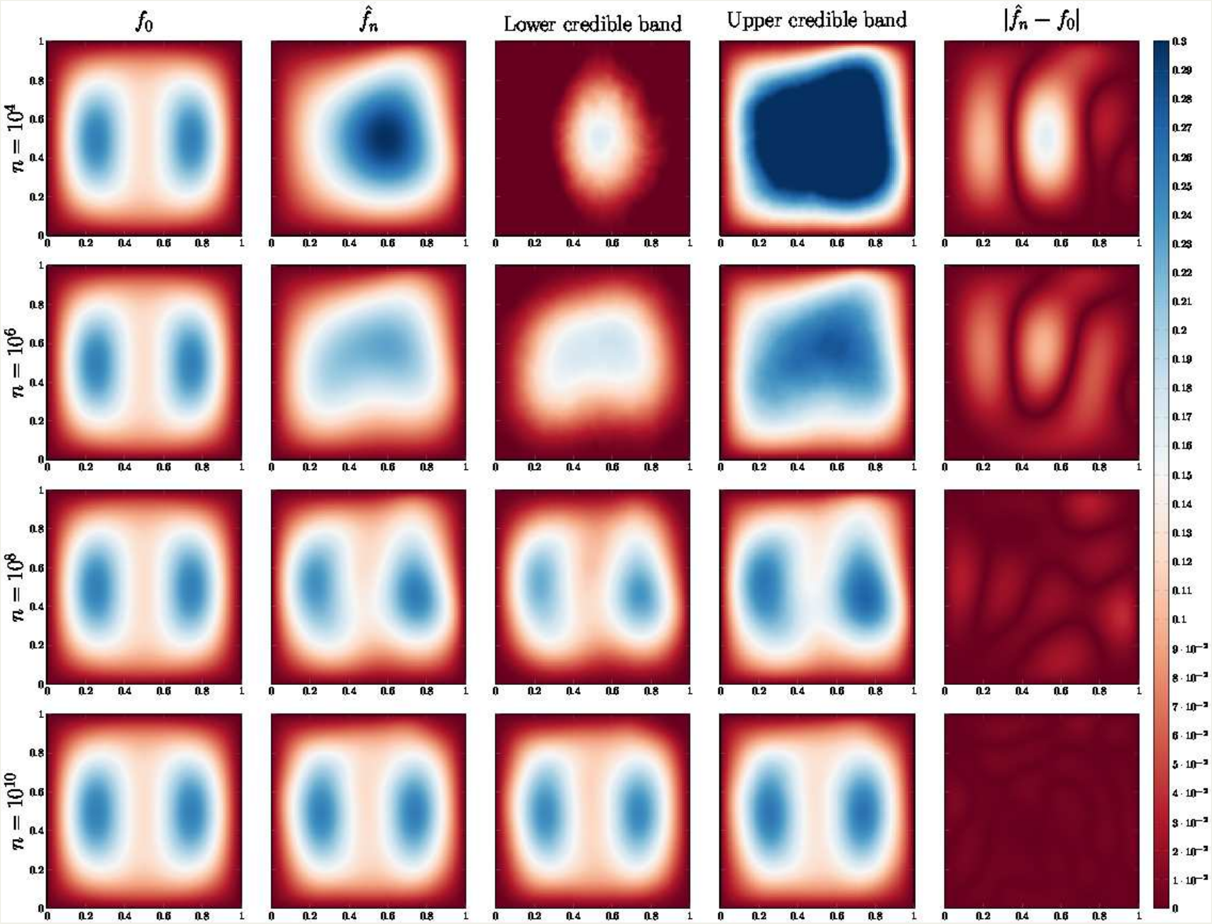
- Put prior on u_θ, equivalently on $v = \mathcal{L}u_\theta$
- Obtain posterior for v from $Y^{(n)} = Kv + n^{-1/2}\dot{\mathbb{W}}$
- Map to posterior of $\theta = e(v)$.

Thm If e Lipschitz on set of posterior mass tending to 1, then contraction rate and uncertainty quantification inherited from linear problem

Schroedinger equation

$$\begin{cases} \frac{1}{2}\Delta u_\theta = \theta u_\theta, & \text{on } \Omega, \\ u_\theta = g, & \text{on } \partial\Omega. \end{cases}$$

$$\theta = e(\Delta u_\theta) := \frac{\Delta u_\theta}{2\theta}.$$



Outlook

Contraction rates in direct problems understood

Uncertainty quantification much less understood

Growing insight in Bayesian methods for inverse problems