A new way of achieving Bayesian nonparametric adaptation

Sergios Agapiou

Verona, 7th November 2024

Department of Mathematics and Statistics, University of Cyprus

Table of Contents

References and collaborators

- *p*-exponential priors
  - S. Agapiou, M. Dashti and T. Helin, *Rates of contraction of posterior distributions based on p-exponential priors*, Bernoulli, 2021
  - S. Agapiou and S. Wang, *Laplace priors and spatial inhomogeneity in Bayesian inverse problems*, Bernoulli, 2024
  - S. Agapiou and A. Savva, *Adaptive inference over Besov spaces in the white noise model using p-exponential priors*, Bernoulli, 2024




- Oversmoothed heavy-tailed priors
  - S. Agapiou and I. Castillo, *Heavy-tailed Bayesian nonparametric adaptation*, The Annals of Statistics, 2024

- Works in progress with Ismaël Castillo and Paul Egels

Start point - rates of contraction with Gaussian priors

Some nonparametric models

- Gaussian white noise model

$$dX(t) = f(t)dt + \frac{1}{\sqrt{n}}dB(t), \; t \in [0,1]^d$$

- Gaussian nonparametric regression, design points $t_i \in [0,1]^d$

$$X_i = f(t_i) + \epsilon_i, \quad 1 \leq i \leq n$$

- Inverse problems, observe $\mathcal{G}(f)$ subject to noise

- Density estimation, $X_i \overset{iid}{\sim} f, \; 1 \leq i \leq n$, for f pdf on $[0,1]^d$

- Nonparametric classification, independent observations $X_i|Z_i, \; 1 \leq i \leq n$, predictor $Z \in [0,1]^d$, response $X \in \{0,1\}$, $f(z) = P(X = 1|Z = z)$

Some nonparametric models

- Gaussian white noise model

$$dX(t) = f(t)dt + \frac{1}{\sqrt{n}}dB(t),\ t \in [0,1]^d$$

- Gaussian nonparametric regression, design points $t_i \in [0,1]^d$

$$X_i = f(t_i) + \epsilon_i, \quad 1 \leq i \leq n$$

- Inverse problems, observe $\mathcal{G}(f)$ subject to noise

- Density estimation, $X_i \overset{iid}{\sim} f,\ 1 \leq i \leq n$, for f pdf on $[0,1]^d$

- Nonparametric classification, independent observations $X_i | Z_i,\ 1 \leq i \leq n$, predictor $Z \in [0,1]^d$, response $X \in \{0,1\}$, $f(z) = P(X = 1 | Z = z)$

Interested in inferring unknown function f, as $n \to \infty$

Classical adaptation problem

Typical minimax estimation rate for 'β-smooth' function f

$$n^{-\frac{\beta}{d+2\beta}} \quad \left(n^{-\frac{\beta}{d+2\beta+2\nu}},\ \nu \text{ ill-posedness}\right)$$

- Find estimator T of f converging at minimax rate without knowledge of β
- Some methods: Lepski's method (90s-), wavelet thresholding (95s-), model selection (98s-), Bayesian nonparametrics (2000s-)

Classical adaptation problem

Typical minimax estimation rate for 'β-smooth' function f

$$n^{-\frac{\beta}{d+2\beta}} \quad \left(n^{-\frac{\beta}{d+2\beta+2\nu}},\ \nu \text{ ill-posedness}\right)$$

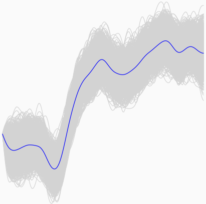
- Find estimator T of f converging at minimax rate without knowledge of β
- Some methods: Lepski's method (90s-), wavelet thresholding (95s-), model selection (98s-), Bayesian nonparametrics (2000s-)
- Restrict presentation to $d = 1$

Bayesian nonparametric framework

- $f \sim \Pi$ prior, distribution on parameter space $\mathcal{F}$ (say L_2)
- $X^{(n)}|f \sim P_f^{(n)}$ likelihood (suppress n, write $X|f \sim P_f$)
- $f|X \sim \Pi(\cdot|X)$ posterior, given by Bayes' rule

$$\Pi(B|X) = \frac{\int_B P_f(X) d\Pi(f)}{\int_{\mathcal{F}} P_f(X) d\Pi(f)}$$

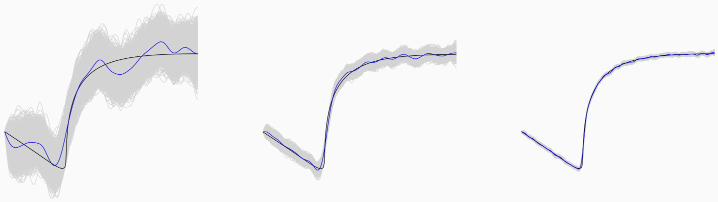
- Result is a data-dependent distribution $\Pi(\cdot|X)$
- Appealing because of uncertainty quantification and flexibility in prior's choice



Frequentist performance of Bayesian posteriors

- Assume there exists fixed true f_0 such that $X \sim P_{f_0}$ (recall suppressed n)
- Study the behaviour of $\Pi(\cdot|X)$ under P_{f_0} as $n \to \infty$:
  - convergence to f_0
  - rate of convergence
- ε_n is a posterior contraction rate at f_0 wrt loss ℓ, if as $n \to \infty$

$$E_{f_0}\Pi(f : \ell(f, f_0) > \varepsilon_n|X) \to 0$$



Frequentist performance of Bayesian posteriors

Why?

Frequentist performance of Bayesian posteriors

Why?

- Mathematical foundation of Bayesian procedures

Frequentist performance of Bayesian posteriors

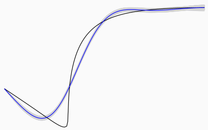
Why?

- Mathematical foundation of Bayesian procedures
- Implies existence of point estimator converging at this rate (meaningful to compare to minimax rate)

Frequentist performance of Bayesian posteriors

Why?

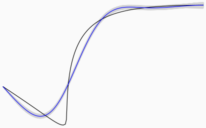
- Mathematical foundation of Bayesian procedures
- Implies existence of point estimator converging at this rate (meaningful to compare to minimax rate)
- Insight on choice of prior



Frequentist performance of Bayesian posteriors

Why?

- Mathematical foundation of Bayesian procedures
- Implies existence of point estimator converging at this rate (meaningful to compare to minimax rate)
- Insight on choice of prior



Trade offs:

- Ability to optimally capture complex unknown functions
- Prior's complexity
- Computability

How?

- Sometimes can do explicit or semi-explicit calculations

Frequentist performance of Bayesian posteriors

How?

- Sometimes can do explicit or semi-explicit calculations

- 'GGV' general theory [Ghosal, Ghosh and van der Vaart 00], [Ghosal and van der Vaart 07]
  - Prior mass condition

    'The prior should put enough mass around the truth'

  - Testing/entropy condition on sieve sets
  - Sieve sets need to capture 'bulk' of prior mass

Frequentist performance of Bayesian posteriors

How?

- Sometimes can do explicit or semi-explicit calculations

- 'GGV' general theory [Ghosal, Ghosh and van der Vaart 00], [Ghosal and van der Vaart 07]
  - Prior mass condition

    'The prior should put enough mass around the truth'
  - Testing/entropy condition on sieve sets
  - Sieve sets need to capture 'bulk' of prior mass

- Prior mass condition alone suffices for contraction of ρ-posteriors

$$\Pi_\rho(B|X) = \frac{\int_B (P_f(X))^\rho d\Pi(f)}{\int_{\mathcal{F}} (P_f(X))^\rho d\Pi(f)}, \qquad 0 < \rho < 1$$

[T. Zhang 06, Bhattacharya et al. 19, L'Huillier et al. 24]

Priors on functions - Gaussian process priors

- [A. van der Vaart and H. van Zanten 08] showed that posterior contraction rates for GP priors can be studied via their concentration function at f_0

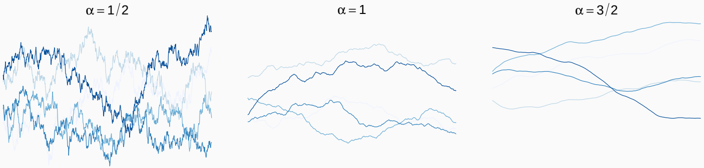
$$\phi_{f_0}(\varepsilon) = \inf_{h \in \mathbb{H}: \|h - f_0\|_{\mathcal{F}} \leq \varepsilon} \|h\|_{\mathbb{H}}^2 - \log \Pi(\varepsilon B_{\mathcal{F}})$$

- α-smooth Gaussian priors by random expansions in orthonormal bases, e.g.

$$f(\cdot) = \sum_{k \geq 1} \sigma_k \zeta_k \varphi_k(\cdot)$$

with

$$\sigma_k = k^{-1/2-\alpha}, \qquad \zeta_k \overset{iid}{\sim} \mathcal{N}(0,1)$$



Posterior contraction rates for Gaussian priors

- Contraction rate for α-smooth GP prior for β-Sobolev smooth truth f_0

$$\varepsilon_n \lesssim \begin{cases} n^{-\beta/(1+2\alpha)}, & \text{if } \alpha \geq \beta, \\ n^{-\alpha/(1+2\alpha)}, & \text{if } \alpha \leq \beta \end{cases}$$

- Rate cannot be improved, [Castillo 08]

- GPs are not adaptive to smoothness

- Adaptation can be achieved by making the prior more complex, e.g. by
  - making α random, [Belitser and Ghosal 03], [Knapik et al. 16]
  - introducing random rescaling, [van der Vaart and van Zanten 09]
  - randomly truncating the series expansion, e.g. [Arbel et al 13]

Spatially inhomogeneous unknowns

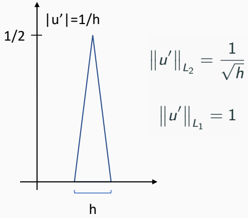
In many applications, the unknown function has 'edges' and 'blocky structure'

Shepp-Logan phantom [Shepp and Logan 74], road cut in Chimborazo volcano [www.geologyin.com], and NMR signal [Donoho et al. 95]

- Empirically, Gaussian priors are known to perform poorly for such spatially inhomogeneous unknowns

Waypoint - p-exponential priors

Besov-space priors



- (Hilbert-)Sobolev spaces measure differentiability in L^2-sense, functions with spikes get high norms
- $\mathcal{B}_{11}^{\beta}$-Besov spaces, 'measure differentiability in L^1-sense'
- $\mathcal{B}_{pp}^{\beta}$-Besov spaces with $1 \leq p < 2$, permit spatially inhomogeneous functions with small norm ($p = 2$ gives Sobolev spaces)
- Motivation for introduction of $\mathcal{B}_{pp}^{s}$-Besov priors in [Lassas et al. 2009], 'penalizing B_{pp}^{s} norms', $p \in [1, 2]$
  - for $p = 2$ Gaussian priors
  - for $p = 1$ Laplace priors, permitting SI functions with non-trivial probability

p-exponential priors

- [Agapiou et al 21] consider p-exponential priors, $p \in [1,2]$

$$f(\cdot) = \sum_{k \geq 1} \sigma_k \zeta_k \varphi_k(\cdot)$$

with

$$(\sigma_k) \in \ell_2, \qquad \zeta_k \overset{iid}{\sim} c_p \exp(-|x|^p/p)$$

- $p = 1$ Laplace, $p = 2$ Gaussian, for appropriate σ_k get Besov priors
- Developed abstract concentration theory, strongly relying on log-concavity

- [Agapiou et al 21] consider p-exponential priors, $p \in [1, 2]$

$$f(\cdot) = \sum_{k \geq 1} \sigma_k \zeta_k \varphi_k(\cdot)$$

with

$$(\sigma_k) \in \ell_2, \qquad \zeta_k \overset{iid}{\sim} c_p \exp(-|x|^p/p)$$

- $p = 1$ Laplace, $p = 2$ Gaussian, for appropriate σ_k get Besov priors
- Developed abstract concentration theory, strongly relying on log-concavity
- Let $\mathcal{Z}$ be the Banach space with norm $\|h\|_{\mathcal{Z}} = \left(\sum_{k=1}^{\infty} |h_k/\sigma_k|^p\right)^{1/p}$.

Theorem (A., Dashti, Helin 21)

Can study rates of contraction under p-exponential priors via concentration function

$$\phi_{f_0}(\varepsilon) = \inf_{h \in \mathcal{Z}: \|h - f_0\|_{\mathcal{F}} \leq \varepsilon} \|h\|_{\mathcal{Z}}^p - \log \Pi(\varepsilon B_{\mathcal{F}})$$

Rates of contraction under Besov smoothness in the WNM

- In WNM, [Donoho and Johnstone 98]
  - minimax rate over $B^{\beta}_{rq}, r \in [1,2]$ in L_2-loss is $n^{-\frac{\beta}{1+2\beta}}$
  - for $r \in [1,2)$ linear estimators limited by slower rate $n^{-\frac{\beta-\gamma/2}{1+2\beta-\gamma}}$, $\gamma = \frac{2-r}{r}$
    (for $r = 2$ linear estimators achieve minimax rate)

'Linear estimators not flexible enough to fit both smooth and spiky part'

- α-smooth p-exponential priors, $p \in [1,2]$,

$$f(\cdot) = \sum_{k \geq 1} \sigma_k \zeta_k \varphi_k(\cdot), \quad \sigma_k = k^{-1/2-\alpha}, \quad \zeta_k \overset{iid}{\sim} c_p \exp(-|x|^p/p)$$

or wavelet version

$$f(\cdot) = \sum_{l \geq 0} \sum_{k=0}^{2^l-1} \sigma_l \zeta_{lk} \psi_{lk}(\cdot), \quad \sigma_l = 2^{-(1/2+\alpha)l}, \quad \zeta_{lk} \overset{iid}{\sim} c_p \exp(-|x|^p/p)$$

Rates of contraction under Besov smoothness in the WNM

- [Agapiou et al 21], see also [Savva - PhD thesis 23], derived upper bounds
- Over Sobolev spaces, p-exponential priors with any $p \in [1,2]$ contract at the minimax rate only for $\alpha = \beta$
- Over $\mathcal{B}^{\beta}_{rq}$, $r \in [1,2)$

  - Rates for α-smooth Gaussian priors at best match the (suboptimal) linear minimax rate
  - Laplace priors can achieve the minimax rate for $\alpha = \beta - 1$ and appropriate rescaling

- In [Agapiou and Wang 24] established lower bound over $\mathcal{B}^{\beta}_{rq}$, for arbitrary sequences of Gaussian priors: GP priors limited by linear minimax rate!
- Open problem whether Laplace rates can be improved

Adaptive rates of contraction in the WNM

- [Agapiou and Savva 24], see also [Savva - PhD thesis 23], studied adaptation in WNM
- No conjugacy to exploit, used general theory of [Rousseau and Szabo 17]

- Adaptation over Sobolev spaces with p-exponential priors for any $p \in [1, 2]$, by making α random or introducing random rescaling
- Adaptation over Besov spaces $\mathcal{B}^{\beta}_{rq}$, $r \in [1, 2]$, $q \in [1, \infty]$ with Laplace priors, by simultaneously randomizing α and introducing random rescaling

- MMLE empirical Bayes choice of hyper-parameters leads to same rates

Other models

- [Agapiou and Wang 24] rates of contraction with Laplace priors in (nonlinear) PDE inverse problems, for Besov truths
- [Giordano and Ray 22] rates of contraction with p-exponential priors over Sobolev spaces, in drift estimation of multidimensional diffusions
- [Giordano 23] adaptation over Besov spaces with Laplace priors in density estimation

Can we do better?

- Sampling hyper-parameters [Agapiou et al 14] or maximizing the marginal likelihood can be computationally hard

- Rates for α-smooth p-exponential prior in WNM for β-Sobolev truth f_0

$$\varepsilon_n \lesssim \begin{cases} n^{-\beta/(1+2\beta+p(\alpha-\beta))}, & \text{if } \alpha \geq \beta, \\ n^{-\alpha/(1+2\alpha)}, & \text{if } \alpha \leq \beta \end{cases}$$

'Oversmoothing' rate slightly improves when p goes from 2 to 1

Can we do better?

- Sampling hyper-parameters [Agapiou et al 14] or maximizing the marginal likelihood can be computationally hard

- 'Heavy' tails correspond to $p \to 0$ and would give

$$\varepsilon_n \text{ (??) } \begin{cases} n^{-\beta/(1+2\beta)}, & \text{if } \alpha \geq \beta, \\ n^{-\alpha/(1+2\alpha)}, & \text{if } \alpha \leq \beta \end{cases}$$

  'Oversmoothing' rate is minimax!

Can we do better?

- Sampling hyper-parameters [Agapiou et al 14] or maximizing the marginal likelihood can be computationally hard

- 'Heavy' tails correspond to $p \to 0$ and would give

$$\varepsilon_n \text{ (??) } \begin{cases} n^{-\beta/(1+2\beta)}, & \text{if } \alpha \geq \beta, \\ n^{-\alpha/(1+2\alpha)}, & \text{if } \alpha \leq \beta \end{cases}$$

  'Oversmoothing' rate is minimax!

- If heuristic correct
  - adaptation 'for free' if $\alpha \geq \beta$ (prior oversmoothing)
  - rate still limited by prior's smoothness, try '$\alpha \to \infty$'

Promised land? - oversmoothed heavy-tailed priors

Nonparametric regression

- Model: project WNM on given orthonormal basis (φ_k) of $L^2[0,1]$

$$X_k \overset{ind}{\sim} \mathcal{N}(f_k, 1/n)$$

Observation is sequence $X = (X_1, X_2, \ldots)$ and unknown $f = (f_1, f_2, \ldots)$

- Truth: suppose f_0 is β-smooth in Sobolev sense

$$f_0 \in \mathcal{S}^\beta(L) = \left\{ f = (f_k),\ \sum_{k \geq 1} k^{2\beta} f_k^2 \leq L^2 \right\}$$

- Prior on f: for ζ_k iid of heavy-tailed density h, $h(x) \asymp |x|^{-m}$

$$f_k \overset{ind}{\sim} \sigma_k \zeta_k$$

$$\sigma_k = k^{-1/2-\alpha}, \quad \text{HT}(\alpha)\text{-prior}$$

$$\sigma_k = e^{-(\log k)^2}, \quad \text{OT-prior} \quad (\text{`}\alpha \to \infty\text{'})$$

Nonparametric regression - main result

Theorem (A. and Castillo 24+)

If h has two moments ($m > 3$), then for the OT-prior and any $\beta > 0$

$$E_{f_0}\Pi\left[\left\{f : \|f - f_0\|_2 > \varepsilon_n\right\}|X\right] \to 0$$

$$\varepsilon_n = \mathcal{L}_n n^{-\beta/(1+2\beta)} \qquad (\mathcal{L}_n = (\log n)^{\omega})$$

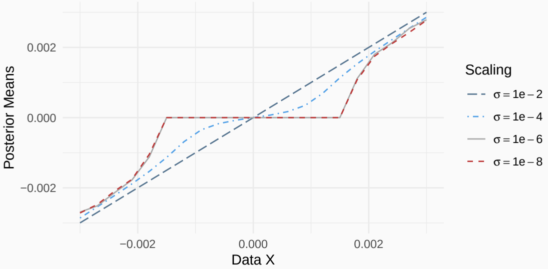
For HT(α)-prior the same holds provided $\beta \leq \alpha$.

- OT-prior leads to fully adaptive posterior (up to logs) over Sobolev smoothness!
- Conjecture in limit $p \downarrow 0$ of p-exponential priors, holds true for HT(α)-prior
- Moment assumption not necessary (ongoing with I. Castillo and P. Egels, cover, e.g. Cauchy and horseshoe priors)

Idea underlying proof

Consider univariate model $X \sim \mathcal{N}(\mu, 1/n)$, $\mu \in \mathbb{R}$ unknown, prior $\mu \sim \sigma\Pi$

- For Π standard Gaussian: $E[\mu|X] = n\sigma^2 X/(1 + n\sigma^2)$
  - shrinking of data X determined by $n\sigma^2$
- For Π standard Student ($n = 10^7$)



  - for large σ posterior mean preserves the data X
  - for small σ posterior mean resembles thresholding estimator
  - good recovery independently of σ, for $|X| \gg 1/\sqrt{n}$

Idea used in semi-explicit bounds

Linear inverse problems

- For ν 'degree of ill-posedness', one observes

$$X_k \overset{ind}{\sim} \mathcal{N}(\kappa_k f_k, 1/n), \qquad \kappa_k \asymp k^{-\nu}$$

- Example (Volterra equation, $\nu = 1$)

$$X(t) = \int_0^t \int_0^s f(u)duds + \frac{1}{\sqrt{n}} B(t), \quad t \in [0, 1]$$

Theorem (A. and Castillo 24+)

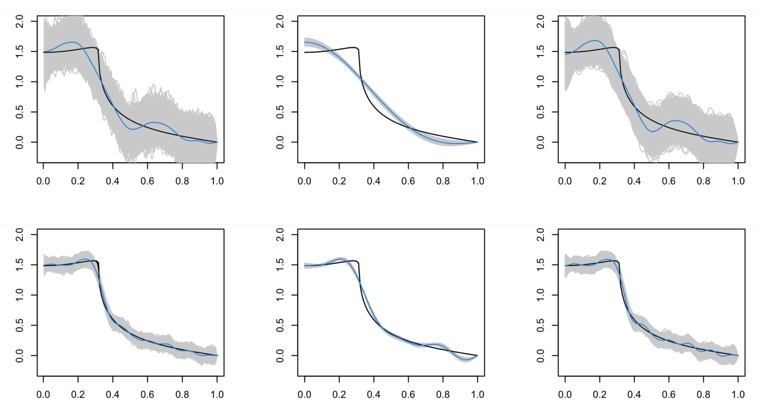
If h has two moments ($m > 3$), then for the OT-prior and any $\beta > 0$

$$E_{f_0} \Pi \Big[\{f : \ \|f - f_0\|_2 > \varepsilon_n\} \, | \, X \Big] \to 0$$

$$\varepsilon_n = \mathcal{L}_n n^{-\beta/(1+2\beta+2\nu)} \qquad (\mathcal{L}_n = (\log n)^{\omega})$$

For HT(α)-prior the same holds provided $\beta \leq \alpha$.

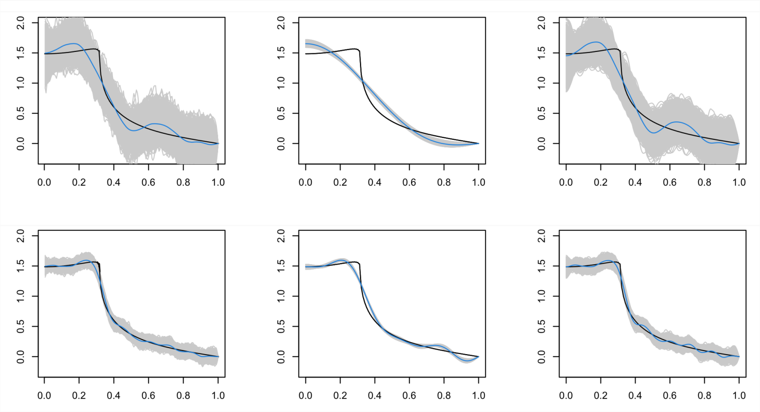
Simulations I: Volterra operator with homogeneously smooth truth



Left: GP+random regularity; Middle: HT(α)–prior; Right: OT-prior

True $\beta = 1$, here $\alpha = 5$

Simulations I: Volterra operator with homogeneously smooth truth



Left: GP+random regularity; Middle: HT(α)–prior; Right: OT-prior

True $\beta = 1$, here $\alpha = 5$

Comments on computation

Multiscale OT prior

- Consider $(\psi_{lk}, l \geq 0, k \in \mathcal{K}_l)$ appropriate wavelet basis. Adapt scaling of OT-prior accordingly

- Prior on $f = \sum_{l=0}^{\infty} \sum_{k \in \mathcal{K}_l} f_{lk} \psi_{lk}$: for ζ_{lk} iid of heavy–tailed density h

$$f_{lk} \overset{\text{ind}}{\sim} \sigma_l \zeta_{lk}$$

$$\sigma_l = 2^{-l^2}, \qquad h(x) \asymp x^{-m} \qquad \text{OT–prior}$$

- For $1 \leq r \leq 2$ set

$$\mathcal{B}_{rr}^{\beta}(L) = \left\{ f = (f_{lk}), \quad \sum_{l \geq 0} 2^{rl(\beta + 1/2 - 1/r)} \sum_{k \in \mathcal{K}_l} |f_{lk}|^r < L^r \right\}$$

Theorem (A. and Castillo 24+)

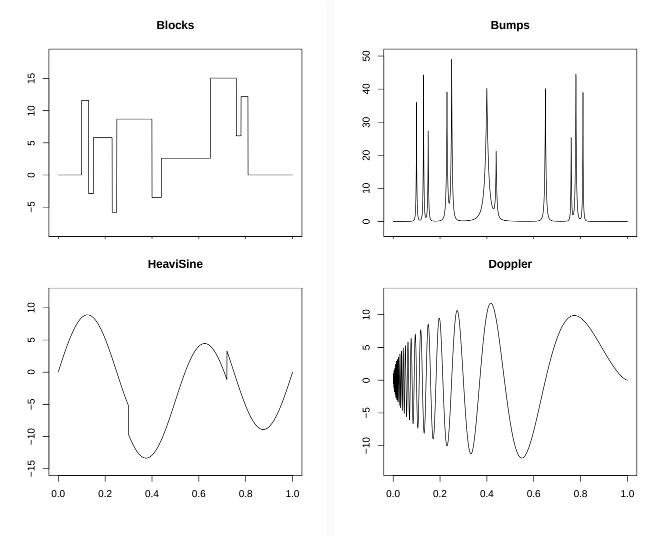
If h has two moments ($m > 3$), then for the multiscale OT-prior, any $1 \leq r \leq 2$ and $\beta > 1/r - 1/2$, and any $f_0 \in \mathcal{B}^{\beta}_{rr}(L)$,

$$E_{f_0} \Pi\Big[\{f : \|f - f_0\|_2 > \varepsilon_n\} \,|\, X\Big] \to 0$$

$$\varepsilon_n = \mathcal{L}_n n^{-\beta/(1+2\beta)} \qquad (\mathcal{L}_n = (\log n)^{\omega})$$

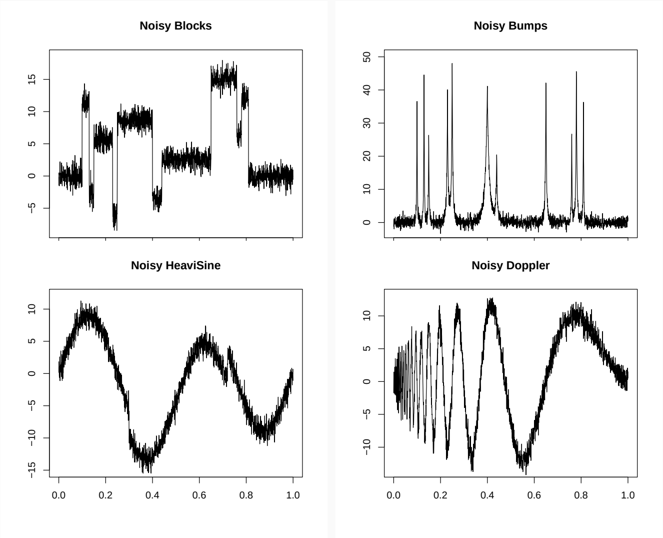
OT-prior adaptive (up to logs) on spatially inhomogeneous Besov spaces without the need of randomizing hyperparameters

Simulations II: direct regression with spatially inhomogeneous truth



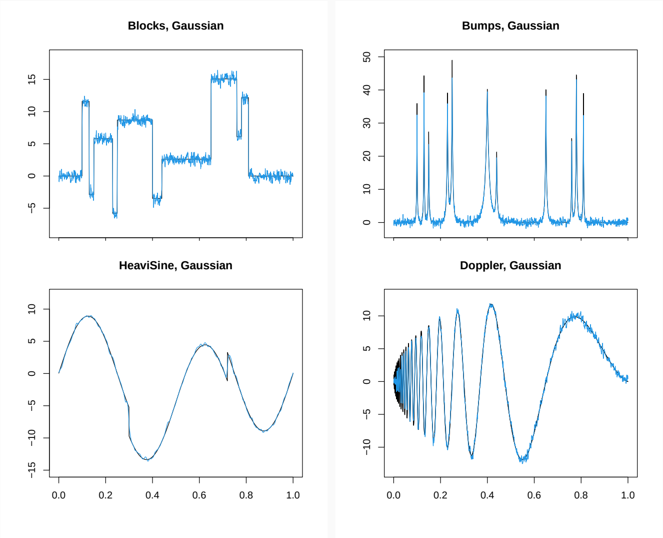
Model truths from [Donoho and Johnstone 94]

Simulations II: noisy observations



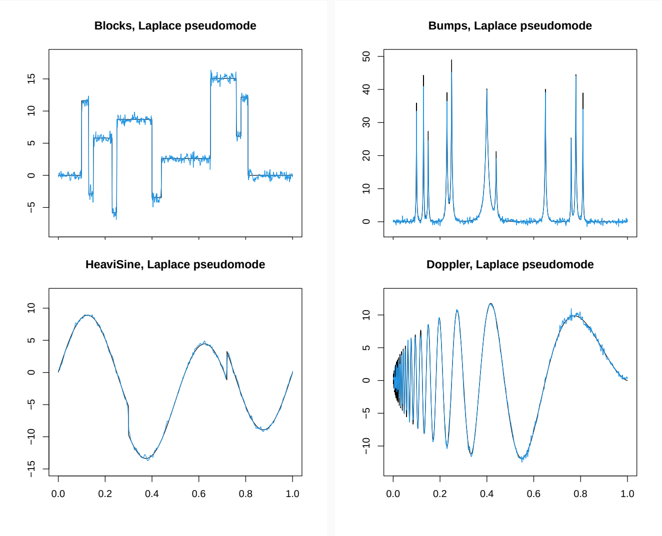
Signal-to-noise ratio ≈7

Simulations II: Gaussian prior with random smoothness and scaling



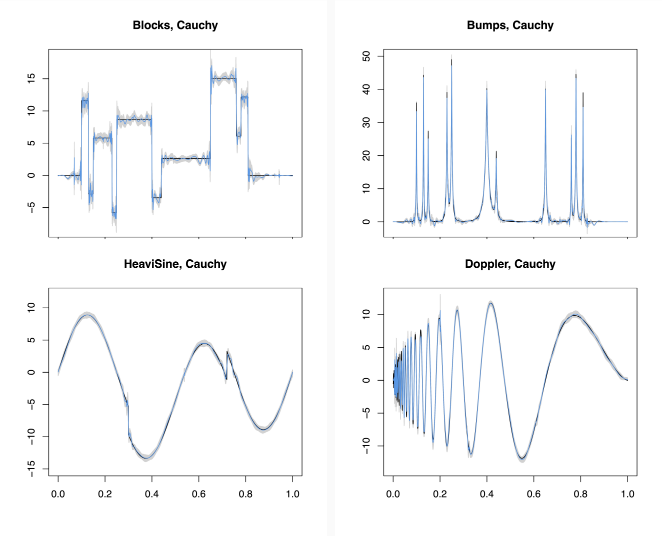
non-centered Gibbs sampler

Simulations II: Laplace prior with random smoothness and scaling



wpCN-within NC-GS (200 draws of f per hyperparameter update) [Chen et al 18]

Simulations II: OT-prior



no GS, wMALA [Chen et al 18], similar results with coordinate-wise sampling in Stan

Nonparametric Regression - Supremum norm contraction

$$\mathcal{H}^\beta(L) = \left\{ f = (f_{lk}), \quad \max_{k \in \mathcal{K}_l} |f_{lk}| \leq 2^{-l(1/2+\beta)} L \;\text{ for all }\; l \geq 0 \right\}$$

Theorem (A. and Castillo 24+)

If h has two moments ($m > 3$), then for the multiscale OT-prior, any $\beta > 0$ and $f_0 \in \mathcal{H}^\beta(L)$

$$E_{f_0} \Pi \left[\{ f : \; \|f - f_0\|_\infty > \varepsilon_n \} \,|\, X \right] \to 0$$

$$\varepsilon_n = \mathcal{L}_n (\log n / n)^{\beta/(1+2\beta)} \qquad (\mathcal{L}_n = (\log n)^\omega)$$

- Adaptation also holds in supremum norm (up to logs)
- So far existing results for priors with spikes (spike–and–slab, BCART) only
- Can also derive adaptive nonparametric Bernstein–von Mises theorem in multiscale space for the OT-prior

Posterior contraction in general models

- To go to more general models, the standard path is via the theory of [Ghosal, Ghosh and van der Vaart 00]
    - prior mass condition
    - testing/entropy condition on sieve set
    - sieve set needs to contain 'bulk' of prior mass

- For heavy-tailed priors sets containing 'bulk' of prior mass are too big to be used as sieve sets

- Use ρ-posteriors, $\rho \in (0, 1)$, for which the prior mass condition

$$\Pi(B_n(f_0, \varepsilon_n)) \geq \exp(-n\varepsilon_n^2)$$

suffices for contraction with rate ε_n in Rényi divergence

Generic prior mass for OT priors

$$\varepsilon_n := (\log n)^{\omega} n^{-\beta/(1+2\beta)}$$

where ω may vary along lines below

Theorem (A. and Castillo 24+)

Consider OT-prior (no moment condition). Given $\beta, L > 0$,

- if $f_0 \in \mathcal{S}^{\beta}(L)$, for any $d_2 > 0$ there exists $d_1 > 0$ sufficiently large s.t.

$$\Pi\left[\|f - f_0\|_2 < d_1 \varepsilon_n\right] \geq e^{-d_2 n \varepsilon_n^2}$$

- if $f_0 \in \mathcal{H}^{\beta}(L)$, for any $d_2 > 0$, for $d_1 > 0$ large enough

$$\Pi\left[\|f - f_0\|_\infty < d_1 \varepsilon_n\right] \geq e^{-d_2 n \varepsilon_n^2}$$

Similar prior mass control can be derived for HT(α)-prior

Application: density estimation

- $X = (X_1, \dots, X_n)$ where $X_j \overset{iid}{\sim} g_0(x),\ x \in [0,1]$, unknown pdf $g_0 \geq c > 0$
- Define prior on density $g : [0,1] \to R^+$ via prior on f and

$$g(x) = g_f(x) = \frac{e^{f(x)}}{\int e^{f(x)} dx}$$

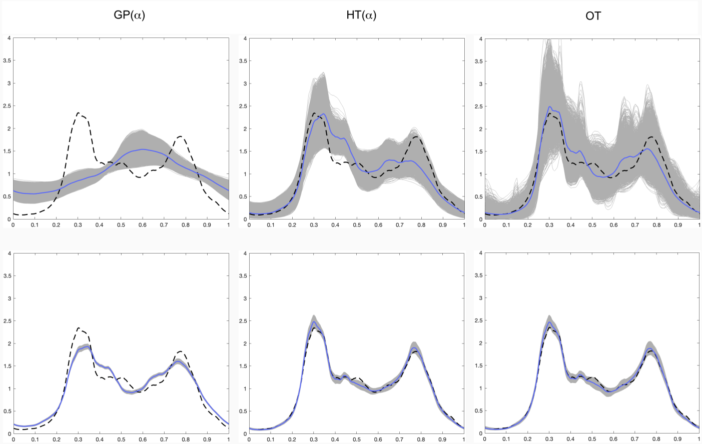
Theorem (A. and Castillo 24+)

Suppose $f_0 := \log g_0 \in \mathcal{H}^\beta(L)$ for some $\beta, L > 0$. Let Π be the prior induced on densities through g_f with f from OT-prior . Then for any $\rho < 1$, there exists $M > 0$ such that

$$E_{g_0} \Pi_\rho \left[\|g - g_0\|_1 > M\varepsilon_n \,|\, X \right] \to 0$$

- OT-prior leads to adaptation (up to logs) in density estimation
- Similar results in classification

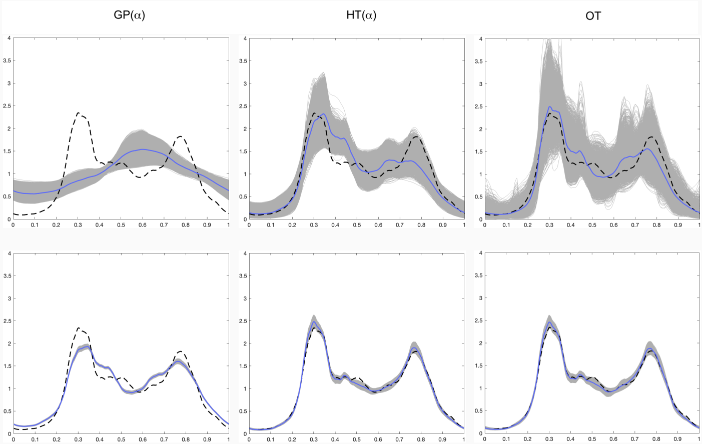
Simulations III: density estimation



Left: GP(α); Middle: HT(α)–prior; Right: OT-prior

True $\beta = 2$ (Hölder), here $\alpha = 5$ [Top $n = 10^2$, Bottom $n = 10^4$]

Simulations III: density estimation



Left: GP(α); Middle: HT(α)–prior; Right: OT-prior

True $\beta = 2$ (Hölder), here $\alpha = 5$ [Top $n = 10^2$, Bottom $n = 10^4$]

Comments on computation

Outlook

Outlook

Bayesian Adaptation in WNM	Sobolev (L^2)	Hölder (L^∞)	Besov (L^2)	Notes on Computation
Gaussian hierarchical	Knapik et al 2016	—	Agapiou and Wang 2024	Conditionally conjugate, GS required
Laplace hierarchical	Agapiou and Savva 2024	—	Agapiou and Savva 2024	Metropolis within Gibbs
OT	Agapiou and Castillo 2024+	Agapiou and Castillo 2024+	Agapiou and Castillo 2024+	Plain MCMC (no GS)
Spike and Slab	Hoffman et al 2015	Hoffman et al 2015	—	Combinatorial number of models to explore
Sieve	Ray 2013	Castillo and Rockova (2021)	—	Depends on base distribution, reversible jump required

- New approach to Bayesian adaptation to smoothness
- Main idea: combine heavy tails with oversmoothing deterministic scaling
- Computationally attractive algorithms (eg easy distributed learning)
- Applies for many models, also for broader adaptation, eg to compositional structure [Castillo and Egels 24+]

Thank you!