

Asymptotic theory for Gaussian process priors

III: Adaptation using conditionally Gaussian priors

Harry van Zanten (TU Eindhoven)

YES IV Workshop
8 – 10 November 2010

Overview

- Rates for rescaled Gaussian process priors
- Adaptation using random scaling
- Gaussian kernel mixtures

Rescaling Gaussian process priors

Rescaled Gaussian process priors

Idea: instead of a different Gaussian process prior for every smoothness level, use a single Gaussian process and **rescale** it appropriately.

Instead of

$$t \mapsto W_t$$

use

$$t \mapsto W_{t/c_n}$$

for scaling constants c_n : **roughening or smoothing**.

Rescaled Gaussian process priors

Base process: e.g. the centered Gaussian process W with covariance

$$\mathbb{E}W_s W_t = e^{-(t-s)^2}$$

(squared exponential process).

Intuition: too smooth as prior on α -smooth functions, should use rescaling constants $c_n \rightarrow 0$.

Illustration: rescaled squared exponential process

W a squared exponential process. Consider rescaled process $(W_{at})_{t \in [0,1]}$ for different values of a :

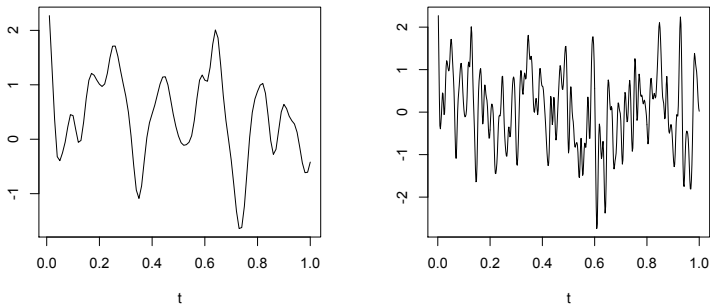


Figure: $a = 1$ versus $a = 5$

Properties of rescaled stationary processes

Let W be a centered stationary GP with spectral measure μ , i.e.

$$\mathbb{E}W_sW_t = \int e^{i\lambda(s-t)} \mu(d\lambda).$$

Set $W_t^c = W_{t/c}$. Suppose for some $\delta > 0$,

$$\int e^{\delta|\lambda|} \mu(d\lambda) < \infty,$$

and μ has Lebesgue density that is $\gg 0$ near 0.

Rescaled process W^c has spectral measure $\mu_c(B) = \mu(cB)$.

RKHS: functions $h_\psi(t) = \int e^{i\lambda t} \psi(\lambda) \mu_c(d\lambda)$, $\|h_\psi\|_{\mathbb{H}} = \|\psi\|_{L^2(\mu_c)}$

Properties of rescaled stationary processes

Lemma.

If $w \in C^\beta[0, 1]$, then

$$\inf_{h \in \mathbb{H}^c: \|h-w\|_\infty < C_w c^\beta} \|h\|_{\mathbb{H}^c}^2 \leq D_w \frac{1}{c}$$

Proof.

Use convolutions. □

Properties of rescaled stationary processes

RKHS ball \mathbb{H}_1^c contained in space of functions **analytic and bounded on a strip** in \mathbb{C} . Metric entropy: Kolmogorov and Tihomirov (1961).

Lemma.

$$\log N(\varepsilon, \mathbb{H}_1^c, \|\cdot\|_\infty) \lesssim \frac{1}{c} \left(\log \frac{1}{\varepsilon} \right)^2$$

Using entropy of \mathbb{H}_1 - small ball connection:

Lemma.

$$-\log \mathbb{P}(\|W^c\|_\infty < 2\varepsilon) \lesssim \frac{1}{c} \left(\log \frac{1}{c\varepsilon^2} \right)^2$$

Rates for rescaled Gaussian process priors

X_1, X_2, \dots, X_n : sample from a positive, continuous density p on $[0, 1]$.

Prior distribution on p : law of

$$t \mapsto \frac{e^{W_t/c_n}}{\int_0^1 e^{W_t/c_n} dt},$$

with W the squared exponential process and, for $\alpha > 0$,

$$c_n = \left(\frac{\log^2 n}{n} \right)^{\frac{1}{1+2\alpha}}.$$

Theorem. [Van der Vaart and vZ. (2007)]

Suppose $\log p_0 \in C^\alpha[0, 1]$. Then the posterior contracts around p_0 at the rate

$$\varepsilon_n \sim \left(\frac{n}{\log^2 n} \right)^{-\frac{\alpha}{1+2\alpha}}.$$

Rates for rescaled Gaussian process priors

- Using a **single** GP, can get optimal rate for any smoothness level (up to a log-factor), by appropriate scaling.
- Have similar results for multiply integrated BM priors.
- Have similar results for different statistical settings.
- Still need to know the regularity of the truth to get the optimal rate. . .

Adaptation using random scaling

Scaling when the true regularity is unknown

How to choose the right scaling constant a ? “Correct” choice will depend on the unknown function of interest.

Scaling when the true regularity is unknown

How to choose the right scaling constant a ? “Correct” choice will depend on the unknown function of interest.

Standard Bayesian solution: let the **data** choose the parameter a .

View scaling constant a as a **hyperparameter** and endow it with a prior distribution as well. Use the **hierarchical prior** model $(W_{At})_{t \in [0,1]^d}$, where A is some random variable independent of W .

Scaling when the true regularity is unknown

How to choose the right scaling constant a ? “Correct” choice will depend on the unknown function of interest.

Standard Bayesian solution: let the **data** choose the parameter a .

View scaling constant a as a **hyperparameter** and endow it with a prior distribution as well. Use the **hierarchical prior** model $(W_{At})_{t \in [0,1]^d}$, where A is some random variable independent of W .

Popular choice for prior on the scaling constant: **gamma distribution**.

Scaling when the true regularity is unknown

How to choose the right scaling constant a ? “Correct” choice will depend on the unknown function of interest.

Standard Bayesian solution: let the **data** choose the parameter a .

View scaling constant a as a **hyperparameter** and endow it with a prior distribution as well. Use the **hierarchical prior** model $(W_{At})_{t \in [0,1]^d}$, where A is some random variable independent of W .

Popular choice for prior on the scaling constant: **gamma distribution**.

Natural question: **is this a good idea?**

Rates for the randomly rescaled GP prior

Data Y_1, \dots, Y_n , with $Y_i = f_0(t_i) + \varepsilon_i$, for ε_i i.i.d. $N(0, \sigma^2)$, $t_i \in [0, 1]$, $f_0 : [0, 1] \rightarrow \mathbb{R}$.

Prior on f : law of $(W_{At})_{t \in [0,1]}$ for W the squared exponential process, A Gamma-distributed, A and W independent.

Theorem. [Van der Vaart and vZ. (2009)]

Suppose $f_0 \in C^\alpha[0, 1]$ for $\alpha > 0$. Then the posterior contracts around f_0 at the rate

$$\varepsilon_n = \left(\frac{\log^2 n}{n} \right)^{\frac{\alpha}{1+2\alpha}}.$$

Rates for the randomly rescaled GP prior

Some remarks regarding this result:

- Up to a log-factor, the rate of contraction is the **optimal minimax rate** for estimating α -regular functions.
- The prior does not depend on the unknown smoothness level α : the procedure is fully **rate-adaptive**.
- Similar result is true in the **multivariate** case ($d > 1$).
- Similar results are true for different statistical settings: **density estimation, classification, . . .**
- Class of Gaussian processes W and scaling distributions A for which the result is valid is slightly larger.

Rates for the randomly rescaled GP prior

Some remarks regarding this result:

- Up to a log-factor, the rate of contraction is the **optimal minimax rate** for estimating α -regular functions.
- The prior does not depend on the unknown smoothness level α : the procedure is fully **rate-adaptive**.
- Similar result is true in the **multivariate** case ($d > 1$).
- Similar results are true for different statistical settings: **density estimation, classification, . . .**
- Class of Gaussian processes W and scaling distributions A for which the result is valid is slightly larger.

So in many ways: **yes**, it is a good idea to use such priors!

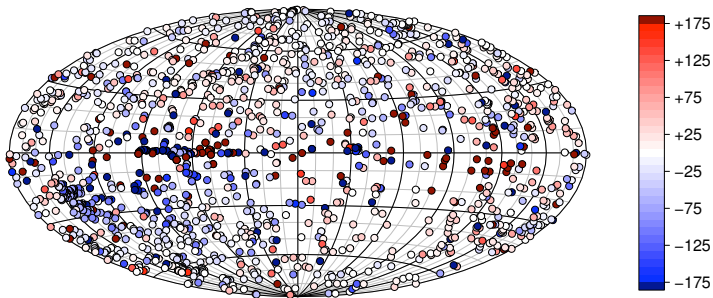
Gaussian kernel mixtures

Motivation: rotation measure data

Short, Higdon and Kronberg (2007), Bayesian Analysis:

- **Goal:** estimate/visualize strength and structure of magnetic fields generated by our galaxy and its neighborhood.
- **Data:** Faraday RM data for radio sources outside our galaxy.

RM data: measurements



1566 observations
(source: Short et al. (2007))

RM data: statistical model

Model:

- Smooth function $f : S^2 \rightarrow \mathbb{R}$ describes magnetic field
- We observe f at 1566 locations, corrupted with noise.
- Essentially a bivariate, nonparametric, fixed design regression problem.

RM data: prior distribution

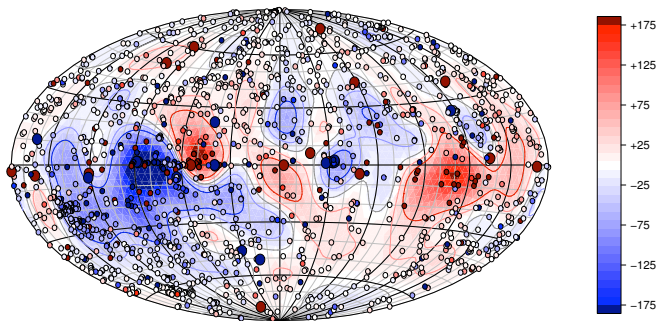
Proposed prior for f :

- Choose a (fine) **grid** x_1, \dots, x_m on S^2 .
- Choose a centered, **smooth kernel** p (e.g. Gaussian density with mean 0).
- Prior for f : law of the random **kernel mixture**

$$x \mapsto \sum_{i=1}^m w_i p(x - x_i),$$

where w_1, \dots, w_m are independent, normally distributed.

RM data: posterior



posterior mean
(source: Short et al. (2007))

RM data: some questions that arise

- How to choose (the number of) **grid points** x_j ?
- How to choose the **shape** of the kernel p ?
- How to choose the **bandwidth** of the kernels?
- How to choose the **scale** of the weights w_j ?
- Can we make the choices in such a way that we get **consistent** estimates for all possible smooth f 's?
- Can we make the choices in such a way that the procedure becomes **optimal** in any way?
-

Construction of the priors I

Suppose $0 < a < b < 1$ and $d \in \mathbb{N}$. Want to construct a prior on functions on $\mathcal{X} = [a, b]^d$.

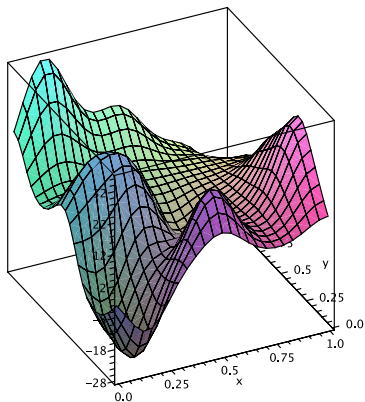
Basic construction:

- Choose a **kernel function** $p : \mathbb{R}^d \rightarrow \mathbb{R}$.
- Draw a random number m from a distribution (p_m) on \mathbb{N} (**gridsize**) and form the gridpoints $x_i = i/m$, $i \in \{1, \dots, m\}^d$.
- Draw a random number σ from a density g on $(0, \infty)$ (**bandwidth**).
- Draw independent standard Gaussians w_i , $i \in \{1, \dots, m\}^d$ (**weights**).

Define

$$W(x) = \sum_{i \in \{1, \dots, m\}^d} w_i \frac{1}{m^{d/2}} \frac{1}{\sigma^d} p\left(\frac{x - x_i}{\sigma}\right).$$

Construction of the priors II



Realization of W ($m = 10$, $\sigma = .1$).

Construction of the priors III

Can now construct priors for various statistical settings:

- **Regression**: use law of random function W as prior.
- **Density estimation**: use law of

$$x \mapsto \frac{e^{W(x)}}{\int_{[a,b]^d} e^{W(y)} dy}.$$

- **Binary regression**: use law of $x \mapsto \Psi(W(x))$, for $\Psi : \mathbb{R} \rightarrow [0, 1]$ a link function.
- ⋮

Concrete questions

In many statistical settings, the **optimal rate** for estimating a function of d variables of regularity α is $n^{-\alpha/(d+2\alpha)}$.

Basic questions:

- Can we attain these rates with our kernel mixture priors?
- Can we do so without knowledge of the regularity of the function? (**adaptation**)

Concretely:

- How to choose the gridsize m ?
- How to choose the bandwidth σ ?
- How to choose the shape of the kernel p ?

Preparations: class of kernels

Class of kernels \mathcal{P}_γ :

For $\gamma \in (0, \infty]$, $p \in L^1(\mathbb{R}^d)$ belongs to \mathcal{P}_γ if $\int_{\mathbb{R}^d} p(x) dx = 1$, it is uniformly Lipschitz, it has finite moments of every order, and it satisfies one of the following conditions:

- For $\gamma < \infty$: p belongs to $C^\gamma(\mathbb{R}^d)$.
- For $\gamma = \infty$: p is the restriction to \mathbb{R}^d of a function that is defined on the set $S = \{(z_1, \dots, z_d) \in \mathbb{C}^d : |\Im z_j| \leq 1 \forall j\}$, and that is bounded and analytic on S .

Preparations: assumptions

(Am) For the prior on the gridsize:

$$p_m \geq Cm^{-s}$$

for some $C > 0$, $s > 1$.

(A σ) For the prior on the bandwidth:

$$D_1\sigma^{-q}e^{-D_2(\frac{1}{\sigma})^{d_\gamma}(\log \frac{1}{\sigma})^r} \leq g(\sigma) \leq D_3\sigma^{-q}e^{-D_4(\frac{1}{\sigma})^{d_\gamma}(\log \frac{1}{\sigma})^r}$$

for some $D_1, D_2, D_3, D_4 > 0$ and $q, r \geq 0$, for all σ in a neighborhood of 0. Here

$$d_\gamma = \frac{2d(d + \gamma)}{2\gamma - d}, \quad \delta_\gamma = \frac{d}{2\gamma - d}.$$

Main abstract result

Theorem. [De Jonge & vZ (2010)]

Assume $p \in \mathcal{P}_\gamma$ for $\gamma \in (d/2, \infty]$ and (Am) and $(A\sigma)$. If $w_0 \in C^\alpha(\mathcal{X})$ for $\alpha > 0$, there exist for every $C > 1$ measurable $B_n \subset C([0, 1]^d)$ and $D > 0$ such that, for n large enough,

$$\begin{aligned}\log N(\bar{\varepsilon}_n, B_n, \|\cdot\|_\infty) &\leq Dn\bar{\varepsilon}_n^2, \\ \mathbb{P}(W \notin B_n) &\leq e^{-Cn\varepsilon_n^2}, \\ \mathbb{P}\left(\sup_{x \in \mathcal{X}} |W(x) - w_0(x)| \leq \varepsilon_n\right) &\geq e^{-n\varepsilon_n^2}.\end{aligned}$$

Here if $\gamma < \infty$:

$$\varepsilon_n = n^{-\frac{\alpha}{d\gamma + 2\alpha(1+\delta\gamma)}}, \quad \bar{\varepsilon}_n = n^{-\frac{\alpha(1-(d\delta\gamma)/(2\gamma))}{(d\gamma + 2\alpha(1+\delta\gamma))(1+d/(2\gamma))}.$$

If $\gamma = \infty$:

$$\varepsilon_n = n^{-\frac{\alpha}{d+2\alpha}} \log^{\frac{r\vee(1+d)}{2+d/\alpha}} n, \quad \bar{\varepsilon}_n = n^{-\frac{\alpha}{d+2\alpha}} \log^{\frac{r\vee(1+d)}{2+d/\alpha} + \left(\frac{1+d-r}{2}\right)_+} n.$$

First remarks

- This result will lead to contraction rates of the order $\varepsilon_n \vee \overline{\varepsilon}_n$ for various statistical models.
- For $\gamma < \infty$, only get a rate for $\gamma > (1 + \sqrt{5})d/4 \approx (0.81)d$.
- For $\gamma < \infty$ the rate is worse than minimax, but it approaches the minimax rate as $\gamma \rightarrow \infty$.
- For $\gamma = \infty$ we get the minimax rate, up to a logarithmic factor.
- The prior does not depend on the unknown regularity α .
- We don't need higher order kernels to get good rates for smooth truths.

Special case I: regression

Observe

$$Y_j = f_0(t_j) + \varepsilon_j, \quad j = 1, \dots, n,$$

with fixed $t_j \in \mathcal{X} = [a, b]^d$ and ε_j indep. $N(0, \tau^2)$. Use law Π of W as prior on f .

Special case I: regression

Observe

$$Y_j = f_0(t_j) + \varepsilon_j, \quad j = 1, \dots, n,$$

with fixed $t_j \in \mathcal{X} = [a, b]^d$ and ε_j indep. $N(0, \tau^2)$. Use law Π of W as prior on f .

Theorem. [De Jonge & vZ (2010)]

Suppose $p \in \mathcal{P}_\infty$ and (Am) and (A σ). If $f_0 \in C^\alpha(\mathcal{X})$ for $\alpha > 0$, then for $L > 0$ sufficiently large,

$$\Pi\left(f : \frac{1}{n} \sum_{j=1}^n (f(t_j) - f_0(t_j))^2 > L^2 \varepsilon_n^2 \mid Y_1, \dots, Y_n\right) \xrightarrow{P_{f_0}} 0,$$

for

$$\varepsilon_n = n^{-\frac{\alpha}{d+2\alpha}} \log \frac{r\sqrt{(1+d)}}{2+d/\alpha} + \left(\frac{1+d-r}{2}\right)_+ n.$$

Special case I: regression

Corollary. [De Jonge & vZ (2010)]

Suppose p is the standard Gaussian kernel, (Am) holds and σ^d is inverse gamma. Then if $f_0 \in C^\alpha(\mathcal{X})$ for $\alpha > 0$, then for $L > 0$ sufficiently large,

$$\Pi\left(f : \frac{1}{n} \sum_{j=1}^n (f(t_j) - f_0(t_j))^2 > L^2 \varepsilon_n^2 \mid Y_1, \dots, Y_n\right) \xrightarrow{P_{f_0}} 0,$$

for

$$\varepsilon_n = n^{-\frac{\alpha}{d+2\alpha}} \log \frac{4\alpha+4\alpha d+d+d^2}{4\alpha+2d} n.$$

Special case II: density estimation

Observe sample X_1, \dots, X_n from a positive density f_0 on $\mathcal{X} = [a, b]^d$. Use law Π of

$$x \mapsto \frac{e^{W(x)}}{\int_{\mathcal{X}} e^{W(y)} dy}$$

as prior on f .

Special case II: density estimation

Observe sample X_1, \dots, X_n from a positive density f_0 on $\mathcal{X} = [a, b]^d$. Use law Π of

$$x \mapsto \frac{e^{W(x)}}{\int_{\mathcal{X}} e^{W(y)} dy}$$

as prior on f .

Theorem. [De Jonge & vZ (2010)]

Suppose $p \in \mathcal{P}_\infty$ and (Am) and $(A\sigma)$. If $\log f_0 \in C^\alpha(\mathcal{X})$ for $\alpha > 0$, then for $L > 0$ sufficiently large,

$$\Pi\left(f : h(f, f_0) > L\varepsilon_n \mid X_1, \dots, X_n\right) \xrightarrow{P_{f_0}} 0,$$

for

$$\varepsilon_n = n^{-\frac{\alpha}{d+2\alpha}} \log \frac{r\sqrt{(1+d)}}{2+d/\alpha} + \left(\frac{1+d-r}{2}\right)_+ n.$$

Other special cases

- Nonparametric classification
- Signal in white noise
- Drift estimation for an ergodic diffusion
- ⋮

Concluding remarks I

- Kernel mixture priors can yield **optimal contraction rates** for estimating smooth functions of several variables.
- They can do so in a wide **variety of statistical settings**, not just in density estimation.
- They can do so **adaptively**, i.e. without using knowledge of the true regularity.
- It is not necessary to use higher order kernels in order to estimate very smooth functions.
- Careful construction of the prior is necessary to avoid sub-optimal rates.

Concluding remarks

Concluding remarks

- In combination with conditioning, GP tools allow us to analyze conditionally GP priors.
- When carefully constructed, conditionally GP priors can lead to **rate-adaptive** procedures.
- Results provide theoretical underpinning and guidelines for procedures used in practice.