Nonparametric Bayesian Methods - Lecture IV

Harry van Zanten

Korteweg-de Vries Institute for Mathematics


University of Amsterdam

CRiSM Masterclass, April 4-6, 2016

Overview of Lecture IV

- Recall: contraction rates for GP priors
- Deterministic rescaling of Gaussian process priors
- Adaptation using a prior on the length scale

- Other examples of rate-adaptive nonparametric Bayes

- Challenges in theory for BNP

Contraction rates for Gaussian process priors

Recall: general theorem for GP priors - 1

Let $W = (W_t)_{t\in[0,1]}$ be a centered, continuous GP, with RKHS $\mathbb{H}$.

Define, for a function f_0,

$$\varphi_{f_0}(\varepsilon) = \inf_{h\in\mathbb{H}:\|h-f_0\|_\infty<\varepsilon} \|h\|_{\mathbb{H}}^2 - \log \mathbb{P}(\|W\|_\infty < \varepsilon).$$

Theorem.
If $\varepsilon_n > 0$ is such that $n\varepsilon_n^2 \to \infty$ and $\varphi_{f_0}(\varepsilon_n) \le n\varepsilon_n^2$, then $\forall C > 1$, there exist $\mathcal{F}_n \subset C[0,1]$ s.t.

$$\mathbb{P}(\|W - f_0\|_\infty < 2\varepsilon_n) \ge e^{-n\varepsilon_n^2},$$
$$\mathbb{P}(W \notin \mathcal{F}_n) \le e^{-Cn\varepsilon_n^2}$$
$$\log N(3\varepsilon_n, \mathcal{F}_n, \|\cdot\|_\infty) \le 6Cn\varepsilon_n^2.$$

Recall: general theorem for GP priors - 1

Let $W = (W_t)_{t\in[0,1]}$ be a centered, continuous GP, with RKHS $\mathbb{H}$.

Define, for a function f_0,

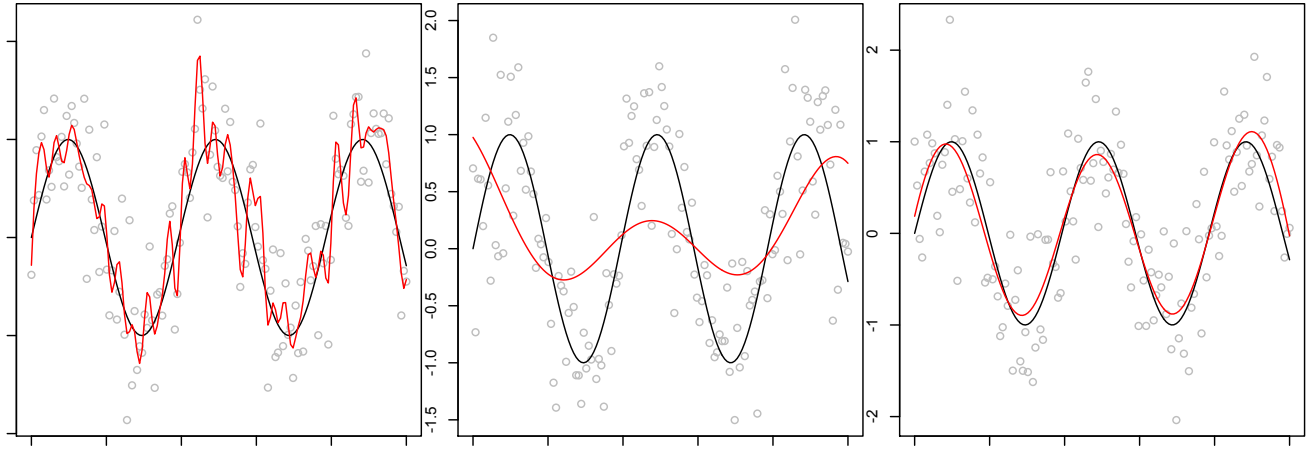
$$\varphi_{f_0}(\varepsilon) = \inf_{h\in\mathbb{H}:\|h-f_0\|_\infty<\varepsilon} \|h\|_{\mathbb{H}}^2 - \log \mathbb{P}(\|W\|_\infty < \varepsilon).$$

Theorem.
If $\varepsilon_n > 0$ is such that $n\varepsilon_n^2 \to \infty$ and $\varphi_{f_0}(\varepsilon_n) \leq n\varepsilon_n^2$, then $\forall C > 1$, there exist $\mathcal{F}_n \subset C[0,1]$ s.t.

$$\mathbb{P}(\|W - f_0\|_\infty < 2\varepsilon_n) \geq e^{-n\varepsilon_n^2},$$
$$\mathbb{P}(W \notin \mathcal{F}_n) \leq e^{-Cn\varepsilon_n^2}$$
$$\log N(3\varepsilon_n, \mathcal{F}_n, \|\cdot\|_\infty) \leq 6Cn\varepsilon_n^2.$$

Recall: general theorem for GP priors - 2

- Get optimal rates if regularity of true function equals regularity of the prior
- If there is a mismatch, get over- or undersmoothing (that is, under- or over fitting)
- Have to be extremely lucky to guess the correct hyperparameters



Q: can we get optimal rates without knowing the true regularity?

$\rightarrow$ adaptation

Deterministic rescaling of GP 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

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

for scaling constants ℓ: roughening or smoothing.

Rescaled Gaussian process priors

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

$$r(s,t) = e^{-(t-s)^2}$$

(squared exponential process).

Rescaled process has covariance

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

Hyperparameter ℓ: length scale parameter.

Intuition: W itself "too smooth" as prior on β-smooth functions, should use length scale $\ell \to 0$.

Illustration: rescaled squared exponential process

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

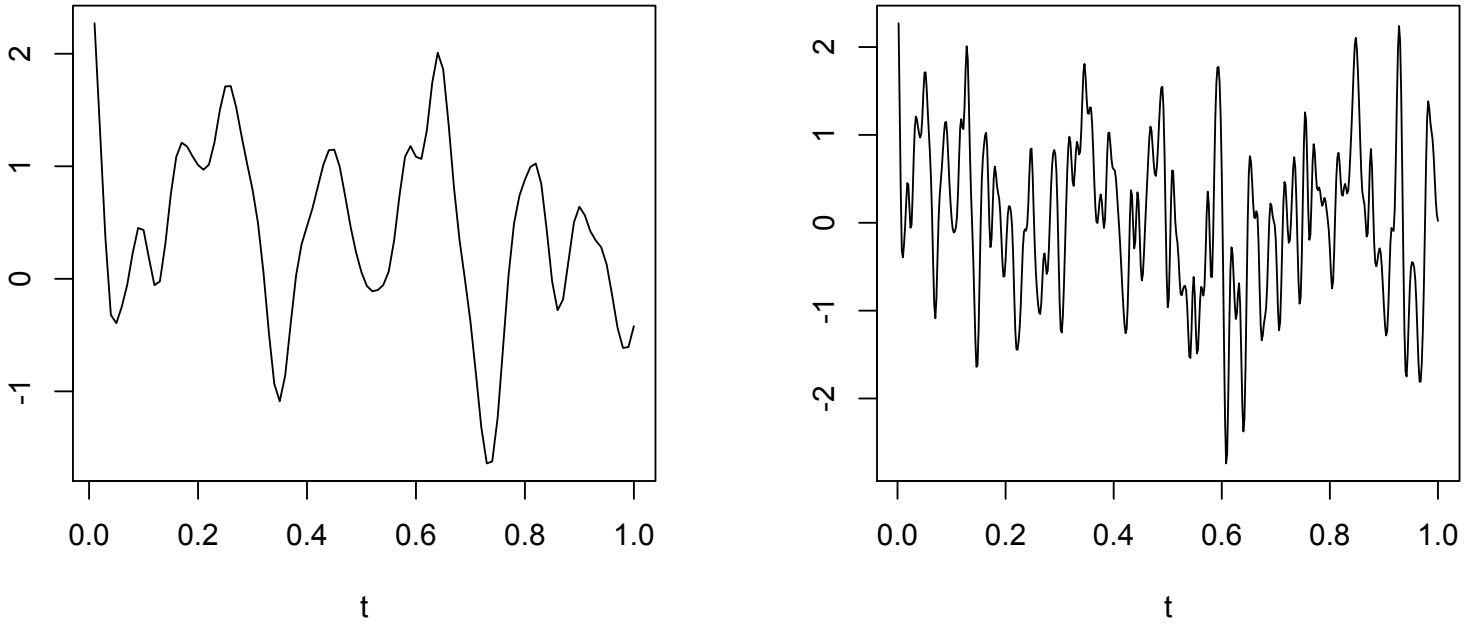


Figure: $\ell = 1$ versus $\ell = 0.2$

RKHS of rescaled stationary GP's

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

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

Set $W_t^\ell = W_{t/\ell}$. 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^ℓ has spectral measure $\mu_\ell(B) = \mu(\ell B)$.

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

Approximating smooth functions by RKHS elements

Lemma.
If $f_0 \in C^\beta[0,1]$, then

$$\inf_{h \in \mathbb{H}^\ell : \|h - f_0\|_\infty < C_{f_0} \ell^\beta} \|h\|^2_{\mathbb{H}^\ell} \leq D_{f_0} \frac{1}{\ell}$$

Proof.
Use convolutions. □

Centered small ball probability

RKHS ball $\mathbb{H}_1^\ell$ contained in space of functions analytic and bounded on a strip in $\mathbb{C}$.

Lemma. [Kolmogorov and Tihomirov (1961)]

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

Using "entropy of $\mathbb{H}_1$" – "small ball" connection:

Lemma.

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

Centered small ball probability

RKHS ball $\mathbb{H}_1^\ell$ contained in space of functions analytic and bounded on a strip in $\mathbb{C}$.

Lemma. [Kolmogorov and Tihomirov (1961)]

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

Using "entropy of $\mathbb{H}_1$" – "small ball" connection:

Lemma.

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

Rates for rescaled Gaussian process priors

Observations: $Y_1, \ldots, Y_n$ satisfying

$$Y_i = f_0(i/n) + e_i,$$

with e_i i.i.d. $N(0, \sigma^2)$.

Prior on f: law of $(W_{t/\ell_n} : t \in [0,1])$, with W the squared exponential process and, for $\beta > 0$,

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

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

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

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

Rates for rescaled Gaussian process priors

- Using a squared exponential GP, can get optimal rates for any smoothness level (up to a log-factor), by appropriate choice of the length scale hyperparameter.
- 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. Not adaptive!

Adaptation using a prior on the length scale

Scaling when the true regularity is unknown

How to choose the right scaling parameter ℓ? "Correct" choice will depend on the unknown function of interest.

Statsticians solution: let the data choose the parameter ℓ.

Full Bayesian approach: view scaling constant ℓ as a hyperparameter and endow it with a prior distribution as well. Use the hierarchical prior model

$$\ell \sim p(\ell)$$
$$W \mid \ell \sim \text{squared exp GP with length scale } \ell$$

Popular choice for prior on the length scale: inverse gamma distribution.

Natural question: is this a good idea?

Scaling when the true regularity is unknown

How to choose the right scaling parameter ℓ? "Correct" choice will depend on the unknown function of interest.

Statsticians solution: let the data choose the parameter ℓ.

Full Bayesian approach: view scaling constant ℓ as a hyperparameter and endow it with a prior distribution as well. Use the hierarchical prior model

$$\ell \sim p(\ell)$$
$$W \mid \ell \sim \text{squared exp GP with length scale } \ell$$

Popular choice for prior on the length scale: inverse gamma distribution.

Natural question: is this a good idea?

Scaling when the true regularity is unknown

How to choose the right scaling parameter ℓ? "Correct" choice will depend on the unknown function of interest.

Statsticians solution: let the data choose the parameter ℓ.

Full Bayesian approach: view scaling constant ℓ as a hyperparameter and endow it with a prior distribution as well. Use the hierarchical prior model

$$\begin{aligned} \ell &\sim p(\ell) \\ W \mid \ell &\sim \text{squared exp GP with length scale } \ell \end{aligned}$$

Popular choice for prior on the length scale: inverse gamma distribution.

Natural question: is this a good idea?

Scaling when the true regularity is unknown

How to choose the right scaling parameter ℓ? "Correct" choice will depend on the unknown function of interest.

Statsticians solution: let the data choose the parameter ℓ.

Full Bayesian approach: view scaling constant ℓ as a hyperparameter and endow it with a prior distribution as well. Use the hierarchical prior model

$$\begin{aligned} \ell &\sim p(\ell) \\ W \mid \ell &\sim \text{squared exp GP with length scale } \ell \end{aligned}$$

Popular choice for prior on the length scale: inverse gamma distribution.

Natural question: is this a good idea?

Scaling when the true regularity is unknown

How to choose the right scaling parameter ℓ? "Correct" choice will depend on the unknown function of interest.

Statsticians solution: let the data choose the parameter ℓ.

Full Bayesian approach: view scaling constant ℓ as a hyperparameter and endow it with a prior distribution as well. Use the hierarchical prior model

$$\begin{aligned} \ell &\sim p(\ell) \\ W \mid \ell &\sim \text{squared exp GP with length scale } \ell \end{aligned}$$

Popular choice for prior on the length scale: inverse gamma distribution.

Natural question: is this a good idea?

Rates for the randomly rescaled SEQ prior - 1

Data: $Y_1, \ldots, Y_n$, with $Y_i = f_0(i/n) + e_i$, for e_i i.i.d. $N(0, \sigma^2)$, $f_0 : [0,1] \to \mathbb{R}$.

Prior on f:

$$\ell \sim \text{inverse gamma}$$
$$f \,|\, \ell \sim \text{squared exp GP with length scale } \ell$$

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

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

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

Rates for the randomly rescaled SEQ prior - 2

Some remarks regarding this result:

- Up to a log-factor, the rate of contraction is the optimal minimax rate for estimating β-regular functions.
- The prior does not depend on the unknown smoothness level β: 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 and priors on length scale for which the result is valid is slightly larger.

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

Rates for the randomly rescaled SEQ prior - 2

Some remarks regarding this result:

- Up to a log-factor, the rate of contraction is the optimal minimax rate for estimating β-regular functions.
- The prior does not depend on the unknown smoothness level β: 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 and priors on length scale for which the result is valid is slightly larger.

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

Other examples of rate-adaptive BNP

Priors that provably yield adaptive, rate-optimal priors

An incomplete list:

- Squared exponential GP, with prior on the length scale [Van der Vaart, vZ. (2009), Bhattacharya et al. (2014)]
- Dirichlet process mixtures of Gaussians [Ghosal et al. (2013)]
- Mixtures of Beta's [Rousseau (2010)]
- Discrete location-scale mixtures [De Jonge, vZ (2010), Kruijer et al. (2010)]
- Spline-based priors [Huang (2004), De Jonge, vZ. (2012)]
- ...

Challenges in theory for BNP

Topics of current/future interest

- Empirical Bayes
- Inverse problems
- Uncertainty quantification
- Models with implicit likelihoods
- Distributed methods
- Statistical efficiency / computational efficiency

...