Gaussian process priors in Bayesian nonparametrics

Harry van Zanten

joint work with Aad van der Vaart

Vrije Universiteit Amsterdam

2007 DYNSTOCH+ Meeting

Nonparametric Bayesian inference

Observations X^n taking values in sample space $\mathcal{X}^n$. Model $\{\mathbb{P}_\theta^n : \theta \in \Theta^n\}$. All $\mathbb{P}_\theta^n$ dominated, density p_θ^n.
Put a prior distribution Π^n on the parameter θ and base the inference on the posterior distribution

$$\Pi^n(B \mid X^n) = \frac{\int_B p_\theta^n(X^n)\,\Pi^n(d\theta)}{\int_{\Theta^n} p_\theta^n(X^n)\,\Pi^n(d\theta)}.$$

Frequentist questions:

- Does the posterior contract around the true parameter θ_0 as $n \to \infty$?
- What is the rate of contraction?

Infinite-dimensional models: parameter θ is a function (density, regression function, drift function, ...), parameter space Θ is a function space.

View prior Π^n as the law of a stochastic process with sample paths in Θ.

Attractive stochastic process priors: use Gaussian processes as building blocks.

- flexible class
- relatively tractable mathematically

Example: Density estimation

Let $X_1, X_2, \ldots, X_n$ be a sample from a distribution with positive, continuous density θ on $[0, 1]$.

Prior distribution on θ: take a Brownian motion W_t and let Π be the law of the random density

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

(Leonard (1978), Lenk (1988), Tokdar and Ghosh (2005), ...)

At what rate does the posterior based on this prior converge to the true density θ_0?

Ghosal, Ghosh and Van der Vaart (2000):

If there exist $\Theta_n \subset \Theta$ and positive numbers ε_n such that $n\varepsilon_n^2 \to \infty$ and, for some $c > 0$,

$$\sup_{\varepsilon > \varepsilon_n} \log N(\varepsilon, \Theta_n, h) \leq n\varepsilon_n^2, \qquad \text{(entropy)}$$

$$\Pi(\Theta \backslash \Theta_n) \leq e^{-(c+4)n\varepsilon_n^2}, \qquad \text{(remaining mass)}$$

$$\Pi(B_n(\theta_0, \varepsilon_n)) \geq e^{-cn\varepsilon_n^2}, \qquad \text{(prior mass)}$$

then for M large enough

$$\mathbb{E}_{\theta_0} \Pi(\theta : h(\theta, \theta_0) > M\varepsilon_n \,|\, X_1, \ldots, X_n) \to 0.$$

Step 1: Relate the relevant metrics (Hellinger, Kullback-Leibler, ...) on the densities

$$p_w(t) = \frac{e^{w_t}}{\int_0^1 e^{w_t}\, dt}$$

to the uniform distance on the functions w.

Step 1: Relate the relevant metrics (Hellinger, Kullback-Leibler, ...) on the densities

$$p_w(t) = \frac{e^{w_t}}{\int_0^1 e^{w_t}\, dt}$$

to the uniform distance on the functions w.

Step 2: Solve the corresponding problem for Brownian motion.

To get a rate ε_n (with $n\varepsilon_n^2 \to \infty$), need to show that there exist $C_n \subset C[0,1]$ such that, for some $c > 0$,

$$\sup_{\varepsilon > \varepsilon_n} \log N(\varepsilon, C_n, \|\cdot\|_\infty) \leq n\varepsilon_n^2,$$

$$\mathbb{P}(W \notin C_n) \leq e^{-(c+4)n\varepsilon_n^2},$$

$$\mathbb{P}(\|W - w_0\|_\infty < \varepsilon_n) \geq e^{-cn\varepsilon_n^2},$$

(small ball probability)

where $w_0 = \log \theta_0$.

(Bibliography: Lifshits (2006))

Brownian motion:

$$\mathbb{P}(\|W - w_0\|_\infty < \varepsilon) \leq \mathbb{P}(\|W\|_\infty < \varepsilon) \sim e^{-(1/\varepsilon)^2}.$$

Hence, can not do better than $\varepsilon_n \sim Cn^{-1/4}$.

Question: under which conditions on $w_0 = \log \theta_0$ do we achieve the rate $n^{-1/4}$?

Reproducing kernel Hilbert space (RKHS):

$$\mathbb{H} = \{h = \int h' : h' \in L^2\}, \qquad \|h\|_{\mathbb{H}} = \|h'\|_{L^2}.$$

Non-centered vs. centered small ball probability:

$$\mathbb{P}(\|W - h\|_\infty < \varepsilon) \geq e^{-\frac{1}{2}\|h\|_{\mathbb{H}}^2} \mathbb{P}(\|W\|_\infty < \varepsilon).$$

Prior mass condition:

$$\phi_{w_0}(\varepsilon_n) \leq n\varepsilon_n^2,$$

where

$$\phi_{w_0}(\varepsilon) = \inf_{h \in \mathbb{H}: \|h - w_0\|_\infty < \varepsilon} \|h\|_{\mathbb{H}}^2 - \log \mathbb{P}(\|W\|_\infty < \varepsilon).$$

(concentration function)

Lemma.

If $w_0 \in C^\alpha[0,1]$, $\alpha > 0$, then

$$\inf_{h \in \mathbb{H}: \|h - w_0\|_\infty < \varepsilon} \|h\|_{\mathbb{H}}^2 \lesssim \varepsilon^{-(2-2\alpha)/\alpha}.$$

Hence for $w_0 \in C^\alpha[0,1]$ the prior mass condition $\phi_{w_0}(\varepsilon_n) \leq n\varepsilon_n^2$ holds for

$$\varepsilon_n \sim \begin{cases} n^{-1/4} & \text{if } \alpha \geq 1/2 \\ n^{-\alpha/2} & \text{if } \alpha \leq 1/2. \end{cases}$$

How about the entropy and remaining mass conditions?

Lemma.
If $w_0 \in C^\alpha[0,1]$, $\alpha > 0$, then

$$\inf_{h \in \mathbb{H}: \|h - w_0\|_\infty < \varepsilon} \|h\|_{\mathbb{H}}^2 \lesssim \varepsilon^{-(2-2\alpha)/\alpha}.$$

Hence for $w_0 \in C^\alpha[0,1]$ the prior mass condition $\phi_{w_0}(\varepsilon_n) \leq n\varepsilon_n^2$ holds for

$$\varepsilon_n \sim \begin{cases} n^{-1/4} & \text{if } \alpha \geq 1/2 \\ n^{-\alpha/2} & \text{if } \alpha \leq 1/2. \end{cases}$$

How about the entropy and remaining mass conditions?

They are automatically fulfilled!

Let $X_1, X_2, \ldots, X_n$ be a sample from a density θ on $[0,1]$.

Prior distribution on θ: law of

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

with W a Brownian motion

Theorem.
Suppose $\log \theta_0 \in C^\alpha[0,1]$. Then the posterior contracts around θ_0 at the rate

$$\varepsilon_n \sim \begin{cases} n^{-1/4} & \text{if } \alpha \geq 1/2 \\ n^{-\alpha/2} & \text{if } \alpha \leq 1/2. \end{cases}$$

Concentration of Gaussian measures

Abstract formulation:

Let $\mathbb{B}$ be a separable Banach space with norm $\|\cdot\|$. Let W be a Borel measurable random element in $\mathbb{B}$, centered and Gaussian (i.e. $b^*(W)$ is Gaussian and centered for $b^* \in \mathbb{B}^*$).

Consider $S : \mathbb{B}^* \to \mathbb{B}, \quad Sb^* = \mathbb{E}Wb^*(W)$.
Reproducing kernel Hilbert space (RKHS) $\mathbb{H}$ associated with W: closure of $S\mathbb{B}^*$ with respect to the inner product

$$\langle Sb_1^*, Sb_2^* \rangle_{\mathbb{H}} = \mathbb{E}b_1^*(W)b_2^*(W).$$

Always $\mathbb{H} \subset \mathbb{B}$.

Example:

Let W be a Borel measurable centered and Gaussian random element in $\mathbb{B} = C[0,1]$ with norm $\|\cdot\|_\infty$.

$\mathbb{B}^* = \{\text{finite signed measures}\}$ and for $\nu \in \mathbb{B}^*$

$$(S\nu)_t = \mathbb{E}W_t \int_0^1 W_s \, \nu(ds).$$

It follows that $\mathbb{H} = \{t \mapsto \mathbb{E}W_t H : H \in \text{first chaos of } W\}$,

$$\langle t \mapsto \mathbb{E}W_t G, t \mapsto \mathbb{E}W_t H \rangle_{\mathbb{H}} = \mathbb{E}GH.$$

Support of W: smallest closed subset $\mathbb{B}_0$ of $\mathbb{B}$ such that $\mathbb{P}(W \in \mathbb{B}_0) = 1$.

Fact:

The support of W is the closure of $\mathbb{H}$ in $\mathbb{B}$.

(Consequence of Hahn-Banach.)

Much more precise: Borell's inequality.

$\mathbb{B}_1$, $\mathbb{H}_1$: unit balls in $\mathbb{B}$, $\mathbb{H}$.

$$\phi_w(\varepsilon) = \inf_{h \in \mathbb{H}: \|h - w_0\| < \varepsilon} \|h\|_{\mathbb{H}}^2 - \log \mathbb{P}(\|W\| < \varepsilon).$$

Borell (1975):

$$\mathbb{P}(W \notin \varepsilon \mathbb{B}_1 + M\mathbb{H}_1) \leq 1 - \Phi(\Phi^{-1}(e^{-\phi_0(\varepsilon)}) + M).$$

Kuelbs and Li (1973):

> $\mathbb{H}_1$ is compact in $\mathbb{B}$, metric entropy related to small ball probability $\phi_0(\varepsilon)$.

Theorem.
Let w be in the support of W and $\varepsilon_n > 0$ such that $n\varepsilon_n^2 \to \infty$ and

$$\phi_w(\varepsilon_n) \leq n\varepsilon_n^2.$$

Then for all $C > 1$ there exist measurable $B_n \subset \mathbb{B}$ such that

$$\begin{aligned} \log N(3\varepsilon_n, B_n, \|\cdot\|) &\leq 6Cn\varepsilon_n^2, \\ \mathbb{P}(W \notin B_n) &\leq e^{-Cn\varepsilon_n^2}, \\ \mathbb{P}(\|W - w\| < 2\varepsilon_n) &\geq e^{-n\varepsilon_n^2}. \end{aligned}$$

Rates of convergence in various settings

$X_1, X_2, \ldots, X_n$: i.i.d. from density θ w.r.t. a measure ν on $\mathcal{X}$.

Prior distribution on θ: law of $x \mapsto e^{W_x} / \int_{\mathcal{X}} e^{W_x} \, \nu(dx)$, with W a centered, Borel measurable Gaussian process on $\mathcal{X}$ with uniformly bounded sample paths.

Theorem.
Suppose $w_0 = \log \theta_0$ is in the support of W. Let $\varepsilon_n > 0$ be such that $n\varepsilon_n^2 \to \infty$ and $\phi_{w_0}(\varepsilon_n) \leq n\varepsilon_n^2$ (with the uniform norm). Then, relative to the Hellinger metric, the posterior concentrates around θ_0 at the rate ε_n.

$X^{(n)}$: sample path of the process

$$X_t^{(n)} = \int_0^t \theta(s)\, ds + \frac{1}{\sqrt{n}} B_t, \quad t \in [0,1],$$

with B a Brownian motion.

Prior on θ: law of a centered Gaussian process W with sample paths in $L^2[0,1]$.

Theorem.

Suppose θ_0 is in the support of W. Let $\varepsilon_n > 0$ be such that $n\varepsilon_n^2 \to \infty$ and $\phi_{w_0}(\varepsilon_n) \leq n\varepsilon_n^2$ (with the L^2-norm). Then, relative to the L^2-norm, the posterior concentrates around θ_0 at the rate ε_n.

Similar results can be derived for

- classification
- regression
- ergodic diffusion
- ...

Gaussian process priors for smoothness classes

Let $X_1, X_2, \ldots, X_n$ be a sample from a distribution with a positive, continuous density θ on $[0,1]$.

Prior distribution on θ: take a centered Gaussian process W_t and let Π be the law of the random density

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

Suppose that $\log \theta_0 \in C^\alpha[0,1]$ for $\alpha > 0$.

Which Gaussian process W leads to the optimal rate $n^{-\alpha/(1+2\alpha)}$?

Candidate: Riemann-Liouville process

$$W_t = \int_0^t (t-s)^{\alpha-1/2}\, dB_s.$$

For $\alpha - 1/2$ integer: W is $(\alpha - 1/2)$-fold repeated integral of B.
For other α: use fractional calculus.

Intuition: good model for α-smooth functions.

Known results for the RL-process:

Li and Linde (1998):

$$-\log \mathbb{P}(\|W\|_\infty < \varepsilon) \sim \varepsilon^{-1/\alpha}$$

RKHS is $I_{0+}^{\alpha+1/2}(L^2)$,

$$\|I_{0+}^{\alpha+1/2} f\|_{\mathbb{H}} = \frac{\|f\|_{L^2}}{\Gamma(\alpha + 1/2)}.$$

Modified RL-process with parameter $\alpha > 0$:

$$W_t = \sum_{k=0}^{\underline{\alpha}+1} Z_k t^k + \int_0^t (t-s)^{\alpha-1/2}\, dB_s.$$

Theorem.

The support of the process W is $C[0,1]$. For $w \in C^\alpha[0,1]$ we have $\phi_w(\varepsilon) = O(\varepsilon^{-1/\alpha})$ as $\varepsilon \to 0$.

Let $X_1, X_2, \ldots, X_n$ be a sample from a distribution with a positive, continuous density θ on $[0, 1]$.

Prior distribution on θ: take a modified RL-process W_t with parameter $\alpha > 0$ and let Π be the law of the random density

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

Theorem.

Suppose $\log \theta_0 \in C^\alpha[0, 1]$. Then, relative to the Hellinger metric, the posterior concentrates around θ_0 at the rate $n^{-\alpha/(1+2\alpha)}$.

Alternative Gaussian process priors:

- fractional Brownian motion
- truncated series expansions
- wavelet expansions
- ...

Concluding remarks

Additional questions:

- Concrete priors: priors on functions of several variables, ...
- Scaling: W and cW lead to the same rate. How to choose c?
- Adapting to smoothness: When does mixing over the smoothness parameter work?
- Non-Gaussian processes: Lévy processes, ...