Contraction rates for Gaussian process priors

Harry van Zanten

Vrije Universiteit Amsterdam

Workshop on Limit Theorems , January 14–16, 2008

Nonparametric Bayesian inference I

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?

Nonparametric Bayesian inference II

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.

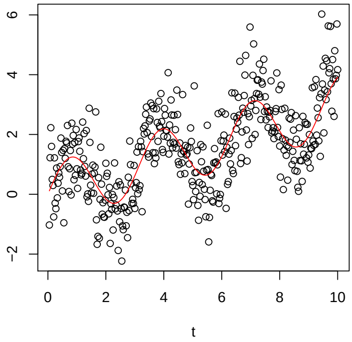
Attractive stochastic process priors: use Gaussian processes as building blocks.

- flexible class
- relatively tractable mathematically

Example: fixed design regression I

Simple example: data (t_i, Y_i) satisfying $Y_i = f(t_i) + \varepsilon_i$ for an unknown, continuous regression function f, ε_i independent $N(0, 1)$.

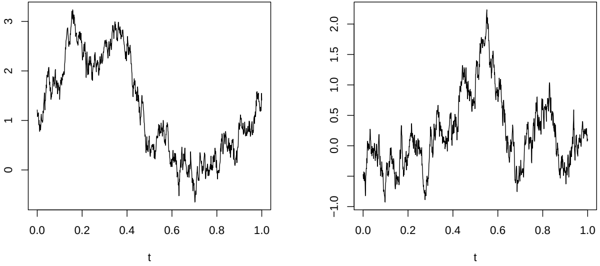
Simulated data:



Example: fixed design regression II

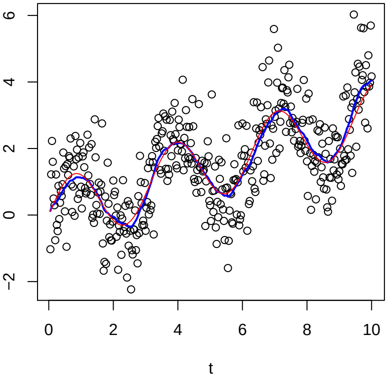
- Prior on $C[0,10]$: $f \sim$ Brownian motion (started in a random point).

  Example realizations of prior:



Example: fixed design regression III

- Compute posterior: $f \sim$ "some Gaussian random process"
- Compute posterior mean:



Example: Density estimation I

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 (2007), . . .)

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

Example: Density estimation II

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

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

Example: Density estimation III

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.

Example: Density estimation III

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.

Example: Density estimation IV

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)

Here $w_0 = \log \theta_0$.

Example: Density estimation V

(Bibliography on small ball probabilities: Lifshits (2007), $>$ 200 papers.)

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 do we achieve the rate $n^{-1/4}$?

Example: Density estimation VI

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 (Cameron-Martin):

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

Prior mass condition:

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

where

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

(concentration function)

Example: Density estimation VII

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 $\varphi_{w_0}(\varepsilon_n) \leq n\varepsilon_n^2$ 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?

Example: Density estimation VII

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 $\varphi_{w_0}(\varepsilon_n) \leq n\varepsilon_n^2$ 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!

Example: Density estimation VIII

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 I

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 $b^* \in \mathbb{B}^*$).

Reproducing kernel Hilbert space (RKHS) $\mathbb{H}$ associated with W: closure of

$$\{\mathbb{E}Wb^*(W) : b^* \in \mathbb{B}^*\}$$

with respect to the inner product

$$\langle \mathbb{E}Wb_1^*(W), \mathbb{E}Wb_2^*(W)\rangle_{\mathbb{H}} = \mathbb{E}b_1^*(W)b_2^*(W).$$

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

Concentration of Gaussian measures II

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.

Concentration of Gaussian measures III

$\mathbb{B}_1$, $\mathbb{H}_1$: unit balls in $\mathbb{B}$, $\mathbb{H}$. For $w_0 \in \mathbb{B}$,

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

Borell (1975):

$$\mathbb{P}(W \notin \varepsilon\mathbb{B}_1 + M\mathbb{H}_1) \leq 1 - \Phi(\Phi^{-1}(e^{-\varphi_0(\varepsilon)}) + M).$$

Kuelbs and Li (1973):

> $\mathbb{H}_1$ is compact in $\mathbb{B}$, metric entropy related to small ball probability $\varphi_0(\varepsilon)$.

Concentration of Gaussian measures IV

Theorem.
Let w_0 be in the support of W and $\varepsilon_n > 0$ such that $n\varepsilon_n^2 \to \infty$ and

$$\varphi_{w_0}(\varepsilon_n) \leq n\varepsilon_n^2.$$

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

$$\begin{aligned} \log N(3\varepsilon_n, B_n, \|\cdot\|) &\leq 6Cn\varepsilon_n^2, \\ \mathbb{P}(W \notin B_n) &\leq e^{-Cn\varepsilon_n^2}, \\ \mathbb{P}(\|W - w_0\| < 2\varepsilon_n) &\geq e^{-n\varepsilon_n^2}. \end{aligned}$$

Consequences of the general result

- Can deal with several statistical settings: density estimation, regression, signal in white noise, classification, . . .
- Can exhibit optimal priors for smoothness classes. Basic idea: if the true function is α-smooth, the sample paths of the Gaussian prior should be α-smooth as well.
- Sheds some light on how we might treat more general priors, e.g. rescaled Gaussian process priors or conditionally Gaussian priors.

Optimal priors for smoothness classes I

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 $n^{-\alpha/(1+2\alpha)}$?

Optimal priors for smoothness classes II

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.

Optimal priors for smoothness classes III

Known results for the RL-process:

Li and Linde (1998):

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

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

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

Optimal priors for smoothness classes IV

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 $\varphi_w(\varepsilon) = O(\varepsilon^{-1/\alpha})$ as $\varepsilon \to 0$.

Optimal priors for smoothness classes V

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 $n^{-\alpha/(1+2\alpha)}$.

Rescaled Gaussian process priors I

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 II

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 $c_n \to 0$.

Rescaled Gaussian process priors III

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

Prior distribution on θ: law of

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

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

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

Adaptive density estimation

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

Prior distribution Π on θ:

- Let W be a centered Gaussian process with $\mathbb{E} W_s W_t = e^{-(t-s)^2}$.
- Let A be $[1,\infty)$-valued, independent of W, density $g(a) \sim C_1 e^{C_2 a \log^2 a}$ for $a \to \infty$.
- Define Π to be the law of the random density

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

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

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

Thanks!

Based on joint work with Aad van der Vaart:

- Rates of contraction of posterior distributions based on Gaussian process priors. To appear in *Annals of Statistics*.
- Reproducing kernel Hilbert spaces of Gaussian priors. To appear in IMS volume in honour of J.K. Ghosh.
- Bayesian inference with rescaled Gaussian process priors. *Electronic Journal of Statistics*, 2007.
- Adaptive Bayesian estimation with rescaled Gaussian process priors. In preparation.

See: www.math.vu.nl/~harry