

Small deviations and Bayesian nonparametrics

Harry van Zanten

Eindhoven University of Technology

SPA 2009, Berlin

Some people involved

The good reverend, 1702(?)–1761:



Co-workers: F. Aurzada, I.A. Ibragimov, R. de Jonge, M.A. Lifshits, A.W. van der Vaart, ...

Outline

- Bayesian statistics in a nutshell
- Illustration: nonparametric regression
- Frequentist asymptotics for Bayes procedures
- Rates of contractions for Gaussian priors
- Noncentered small ball probabilities: RL-processes
- Smooth stationary processes
- Sequences of processes
- Concluding remarks

Bayesian inference in a nutshell

Have data X and a model $\{P_\theta : \theta \in \Theta\}$.

Classical/frequentist point of view:

Data are distributed according to P_{θ_0} , for some fixed, but unknown **true parameter** θ_0 .

Bayesian point of view:

Data are generated by first drawing θ from a **prior** distribution Π on Θ and then given θ , drawing X from P_θ . The pair (θ, X) then has a joint distribution, and hence we can consider the conditional distribution of θ given X : the **posterior** distribution. Further inference is based on this posterior.

Illustration: nonparametric regression - 1

Suppose we have observations

$$Y_i = f(t_i) + \varepsilon_i, \quad i = 1, \dots, n,$$

where $t_i = i/n$, f is an unknown, continuous function, ε_i are independent $N(0, \sigma^2)$ for some unknown σ .

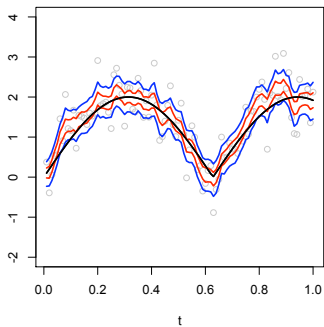
Aim: reconstruct “signal” f .

Approach:

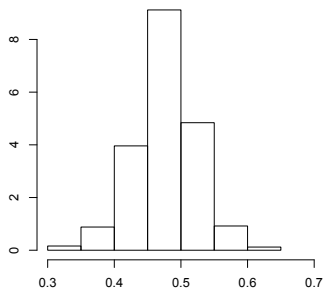
- put priors on f and σ (for f : $\Gamma^{-1} \times$ BM, for σ : Γ^{-1}),
- numerically compute posteriors using Gibbs sampler.

Illustration: nonparametric regression - 2

posterior for signal (red: 50%, blue: 90%)



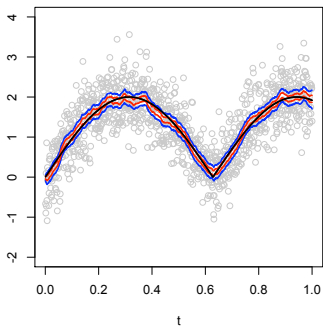
posterior for noise stdev



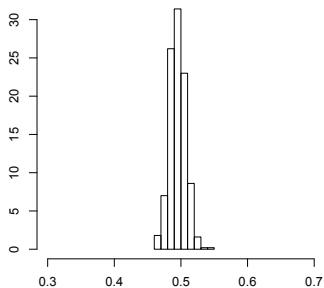
100 observations

Illustration: nonparametric regression - 3

posterior for signal (red: 50%, blue: 90%)



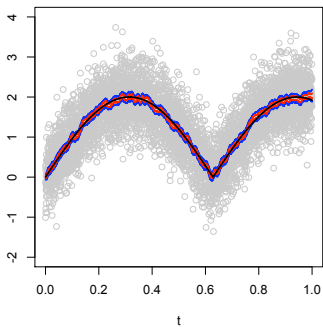
posterior for noise stdev



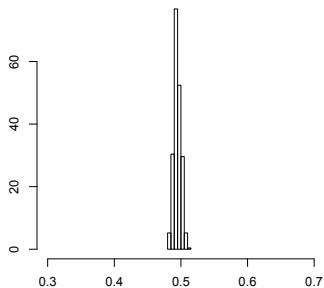
1000 observations

Illustration: nonparametric regression - 4

posterior for signal (red: 50%, blue: 90%)



posterior for noise stdev



5000 observations

Illustration: nonparametric regression - 5

Some questions for this concrete problem:

- The posterior seems to contract around the true f . Why?
- How fast is the convergence to the unknown signal?

General related remarks:

- In nonparametric Bayes problems, consistency is not automatic. [Freedman, Diaconis]
- In nonparametric Bayes problems, convergence rates depend on the choice of the prior. [Castillo]

Frequentist asymptotics for Bayes procedures - 1

Suppose there is a **true parameter** θ_0 . When does the posterior corresponding to the prior Π contract around θ_0 as the number of observations grows indefinitely?

- 40's: **Doob's consistency theorem**: identifiability implies consistency for Π -almost all θ .
- 60's: **negative consistency examples** of David Freedman
- '65: **Schwartz consistency theorem**: prior mass condition, testing condition
- 80's: **more negative consistency examples** by Diaconis and Freedman
- '99: **negative Bernstein-Von Mises examples** by Freedman
- '00/'01: **Rate of contraction results** by Ghosal, Ghosh & Van der Vaart and Shen & Wasserman.

Frequentist asymptotics for Bayes procedures - 2

Typical list of sufficient conditions for having a certain rate of contraction ε_n relative to the metric d on the parameter space Θ :

- **Prior mass condition:**

$$\Pi(\theta : B(\theta_0, \varepsilon_n)) \geq c_1 e^{-n\varepsilon_n^2}$$

There should exist *sieves* $\Theta_n \subset \Theta$ such that

- **Entropy condition:**

$$\log N(\varepsilon_n, \Theta_n, d) \leq c_2 n \varepsilon_n^2$$

- **Remaining mass condition:**

$$\Pi(\Theta \setminus \Theta_n) \leq c_3 e^{-c_4 n \varepsilon_n^2}$$

Frequentist asymptotics for Bayes procedures - 3

Results of this type exist for

- i.i.d. **density estimation** [Ghosal, Ghosh & Van der Vaart (2000)],
- nonparametric **regression** [Ghosal & Van der Vaart (2007)],
- nonparametric **classification** [GvdV '07],
- nonparametric **signal denoising** [GvdV '07],
- nonparametric **spectrum estimation** for time series [GvdV '07],
- nonparametric estimation for **diffusion processes** [Van der Meulen, Van der Vaart & vZ (2006), Panzar & vZ (2009)],
- ...

Rates of contraction for Gaussian priors - 1

Data:

Independent draws X_1, \dots, X_n from a positive, continuous density ρ on $[0, 1]$.

Prior:

Take a centered, continuous Gaussian process $W = (W_t : t \in [0, 1])$. Define the prior Π as the law of the random density $t \mapsto p_W(t)$, where

$$p_w(t) = \frac{e^{w(t)}}{\int_0^1 e^{w(s)} ds}.$$

[Leonard (1978), Lenk (1988)]

Rates of contraction for Gaussian priors - 2

Concentration function:

Let \mathbb{H} be the RKHS associated to W . For $w \in C[0, 1]$, define

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

Rates of contraction for Gaussian priors - 2

Concentration function:

Let \mathbb{H} be the RKHS associated to W . For $w \in C[0, 1]$, define

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

Theorem. [Van der Vaart & vZ. (2008)]

Let $\varepsilon_n \rightarrow 0$ be such that $n\varepsilon_n^2 \rightarrow \infty$ and, for $w_0 = \log p_0$,
 $\varphi_{w_0}(\varepsilon_n) \leq n\varepsilon_n^2$. Then for all $M > 0$ large enough,

$$\Pi(p : h(p, p_0) \geq M\varepsilon_n \mid X_1, \dots, X_n) \xrightarrow{P_0} 0$$

as $n \rightarrow \infty$.

Rates of contraction for Gaussian priors - 3

Proof.

Step 1:

Relate the relevant statistical metrics on the densities $p_w = e^w / \int e^w$ (Hellinger, Kullback-Leibler, ...) to the uniform distance on w .

After that it suffices to show that

$$\mathbb{P}(\|W - \log p_0\|_\infty \leq \varepsilon_n) \geq e^{-n\varepsilon_n^2}$$

and that for every $K > 0$ there exist $C_n \subset C[0, 1]$ such that

$$\mathbb{P}(W \notin C_n) \leq e^{-Kn\varepsilon_n^2},$$

$$\log N(\varepsilon_n, C_n, \|\cdot\|_\infty) \leq n\varepsilon_n^2.$$

Rates of contraction for Gaussian priors - 4

Step 2:

The prior mass condition is equivalent to the condition $\varphi_{w_0}(\varepsilon_n) \leq n\varepsilon_n^2$ [Kuelbs & Linde (1994)].

Step 3:

For $C_n = M\mathbb{H}_1 + \varepsilon\mathbb{B}_1$, we have [Borell-Sudakov]

$$\mathbb{P}(W \notin C_n) \leq 1 - \Phi(\Phi^{-1}(\mathbb{P}(\|W\|_\infty < \varepsilon)) + M)$$

Choosing $\varepsilon = \varepsilon_n$ and $M^2 \sim n\varepsilon_n^2$ gives the remaining mass condition.

Step 4: Entropy condition is fulfilled by the small deviations assumption and the entropy-small deviations connection [Kuelbs & Li (1993), Li & Linde (1999)]



Noncentered small ball probabilities: RL-process - 1

Riemann-Liouville process:

$$R_t^\alpha = \int_0^t (t-s)^{\alpha-1/2} dW_s.$$

[Kimeldorf, Wahba 1970s]

Modified Riemann-Liouville process:

$$X_t^\alpha = \sum_{k=0}^{[\alpha]+1} Z_k t^k + \int_0^t (t-s)^{\alpha-1/2} dW_s.$$

Noncentered small ball probabilities: RL-process - 2

Li & Linde (1998):

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

Noncentered small ball probabilities: RL-process - 2

Li & Linde (1998):

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

Lemma. [Van der Vaart & vZ (2008)]

For $f \in C^\alpha[0, 1]$,

$$\inf_{h \in \mathbb{H}^\alpha: \|h-f\|_\infty < \varepsilon} \|h\|_{\mathbb{H}^\alpha}^2 \lesssim \varepsilon^{-1/\alpha}.$$

So for $f \in C^\alpha[0, 1]$,

$$-\log \mathbb{P}(\|X^\alpha - f\|_\infty < \varepsilon) \lesssim \varepsilon^{-1/\alpha}.$$

Noncentered small ball probabilities: RL-process - 3

Statistical implications

Note that

$$\varepsilon_n^{-1/\alpha} \leq n\varepsilon_n^2 \quad \text{iff} \quad \varepsilon_n \geq n^{-\frac{\alpha}{1+2\alpha}}.$$

Hence if the true log-density belongs to $C^\alpha[0, 1]$, the prior that is build from the modified RL-process X^α yields the contraction rate $\varepsilon_n = n^{-\alpha/(1+2\alpha)}$.

Noncentered small ball probabilities: RL-process - 3

Statistical implications

Note that

$$\varepsilon_n^{-1/\alpha} \leq n\varepsilon_n^2 \quad \text{iff} \quad \varepsilon_n \geq n^{-\frac{\alpha}{1+2\alpha}}.$$

Hence if the true log-density belongs to $C^\alpha[0, 1]$, the prior that is build from the modified RL-process X^α yields the contraction rate $\varepsilon_n = n^{-\alpha/(1+2\alpha)}$.

This is the **optimal** rate of convergence!

Noncentered small ball probabilities: RL-process - 4

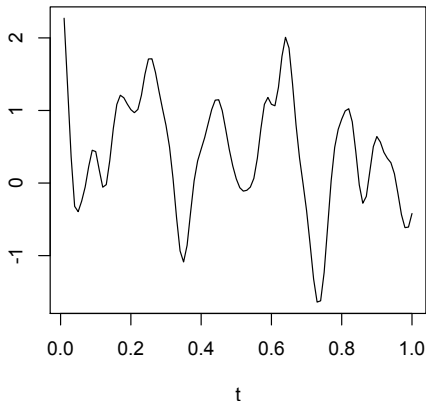
Message for the statisticians: Gaussian process priors (only) yield good contraction rates if the regularity of the prior matches the regularity of the unknown function.

Message for the probabilists: Inequalities for **noncentered** small ball probabilities are extremely useful for the asymptotic analysis of nonparametric Bayes procedures.

Smooth stationary Gaussian processes - 1

Squared exponential process:

$$\mathbb{E}W_s W_t = c_1 e^{-c_2(s-t)^2}$$



Smooth stationary Gaussian processes - 2

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

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

Smooth stationary Gaussian processes - 2

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

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

Tsirelson's bound: [Lifshits & Tsirelson (1986)]

$$-\log \mathbb{P}(\|W\|_\infty < \varepsilon) \gtrsim \log^{3/2} \frac{1}{\varepsilon}$$

Smooth stationary Gaussian processes - 2

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

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

Tsirelson's bound: [Lifshits & Tsirelson (1986)]

$$-\log \mathbb{P}(\|W\|_\infty < \varepsilon) \gtrsim \log^{3/2} \frac{1}{\varepsilon}$$

What is the true small ball behaviour ??????

Smooth stationary Gaussian processes - 3

Let W_ν be the centered, stationary Gaussian process with spectral measure

$$\mu_\nu(d\lambda) = e^{-|\lambda|^\nu} d\lambda.$$

Smooth stationary Gaussian processes - 3

Let W_ν be the centered, stationary Gaussian process with spectral measure

$$\mu_\nu(d\lambda) = e^{-|\lambda|^\nu} d\lambda.$$

Theorem. [Aurzada, Ibragimov, Lifshits, vZ (2009?)]

For $0 < \nu \leq 1$,

$$-\log \mathbb{P}(\|W_\nu\|_\infty < \varepsilon) \asymp \log^{1+\frac{1}{\nu}} \frac{1}{\varepsilon}$$

Smooth stationary Gaussian processes - 3

Let W_ν be the centered, stationary Gaussian process with spectral measure

$$\mu_\nu(d\lambda) = e^{-|\lambda|^\nu} d\lambda.$$

Theorem. [Aurzada, Ibragimov, Lifshits, vZ (2009?)]

For $0 < \nu \leq 1$,

$$-\log \mathbb{P}(\|W_\nu\|_\infty < \varepsilon) \asymp \log^{1+\frac{1}{\nu}} \frac{1}{\varepsilon}$$

For $\nu > 1$

$$-\log \mathbb{P}(\|W_\nu\|_\infty < \varepsilon) \asymp \frac{\log^2 \frac{1}{\varepsilon}}{\log \log \frac{1}{\varepsilon}}.$$

Smooth stationary Gaussian processes - 4

Let \tilde{W}_ν be the centered, stationary Gaussian process with spectral measure

$$\tilde{\mu}_\nu = \sum_{k \in \mathbb{Z}} e^{-|k|^\nu} \delta_{2\pi k}.$$

Smooth stationary Gaussian processes - 4

Let \tilde{W}_ν be the centered, stationary Gaussian process with spectral measure

$$\tilde{\mu}_\nu = \sum_{k \in \mathbb{Z}} e^{-|k|^\nu} \delta_{2\pi k}.$$

Theorem. [Aurzada, Ibragimov, Lifshits, vZ (2009?)]

For all $\nu > 0$,

$$-\log \mathbb{P}(\|\tilde{W}_\nu\|_\infty < \varepsilon) \asymp \log^{1+\frac{1}{\nu}} \frac{1}{\varepsilon}.$$

Sequences of simple processes - 1

Let $p : \mathbb{R}^d \rightarrow [0, \infty)$ standard Gaussian probability density, $m \in \mathbb{N}$, $\sigma > 0$.

$$W_t^{m,\sigma} = \sum_{k \in \{1, \dots, m\}^d} Z_k \frac{1}{m^{d/2}} \frac{1}{\sigma^d} p\left(\frac{t - k/m}{\sigma}\right).$$

Sequences of simple processes - 1

Let $p : \mathbb{R}^d \rightarrow [0, \infty)$ standard Gaussian probability density, $m \in \mathbb{N}$, $\sigma > 0$.

$$W_t^{m,\sigma} = \sum_{k \in \{1, \dots, m\}^d} Z_k \frac{1}{m^{d/2}} \frac{1}{\sigma^d} p\left(\frac{t - k/m}{\sigma}\right).$$

Theorem. [De Jonge & vZ (2009)]

$$-\log \mathbb{P}(\|W^{m,\sigma}\|_\infty < \varepsilon) \lesssim \frac{1}{\sigma^d} \log^{1+d} \frac{1}{\varepsilon \sigma}.$$

Concluding remarks

- Small deviations theory extremely useful for Bayesian asymptotics.
- In particular, inequality for **noncentered** balls are important.
- The statistics connection has motivated new small deviations results for “unusual” classes of processes.
- Often interest in small deviations bounds for **collections** of (relatively simple) processes.
- How about **non-Gaussian** processes, e.g. Lévy processes, processes with independent increments?