

Bayesian nonparametrics, Gaussian random elements in Banach space, and associated approximation problems

Harry van Zanten (TU/e)

Delft Analysis Colloquium
February 16, 2010

Some people involved

The good reverend, 1702(?)–1761:



Co-workers: Aad van der Vaart, Frank van der Meulen, Laura Panzar, René de Jonge, Bartek Knapik, Botond Szabó, Haralambie Leahu, Frank Aurzada, Michael Lifshits, Ildar Ibragimov, ...

Outline

- What is Bayesian statistics?/ Why use it?
- Nonparametric Bayesian statistics / Questions that arise
- Frequentist asymptotics for Bayes procedures

- General asymptotic results for Gaussian priors
- Concrete example: BM as a prior in nonparametric regression
- Other Gaussian priors and related approximation problems

- Concluding remarks

What is Bayesian statistics?

Mathematical statistics:

Have data X and a collection of possible distributions $\{P_\theta : \theta \in \Theta\}$. Want to make inference about θ on the basis of X .

Paradigms in mathematical statistics:

- **Classical/frequentist paradigm:**

There is a “true value” $\theta_0 \in \Theta$. Data have distribution P_{θ_0} .

- **Bayesian paradigm:**

Data is generated as follows: first θ is sampled from a *prior distribution* Π . Given the value of θ , the data X is sampled from the distribution P_θ .

Can then define the *posterior distribution*: the conditional distribution of θ given X . Further inference is based on this posterior.

What is Bayesian statistics?

Example:

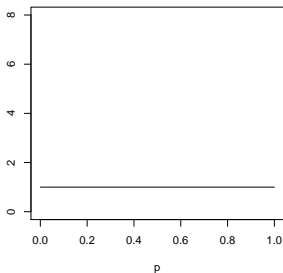
Suppose we have a coin that has probability p of turning up heads. We do 50 independent tosses and observe 42 heads. What can we say about p ?

Here we have an observation (the number 42) from a binomial distribution with parameters 50 and p and want to estimate p .

Standard **frequentist** solution: take the estimate $42/50 = 0.84$.

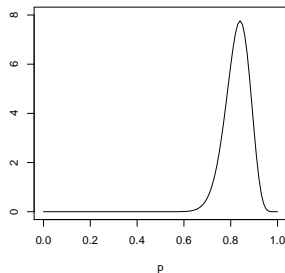
What is Bayesian statistics?

Bayesian approach: choose a prior distribution on p , say uniform on $[0, 1]$. Compute the posterior: beta(43, 8)-distribution (mean is $43/51 \approx 0.843$).



prior

data
→



posterior

Why Bayesian inference?

Philosophical issues, but many practical advantages:

- **Bayesian machine** does the work automatically.
- Flexible modelling for **high-dimensional, complex systems**.
- Use prior to incorporate **expert knowledge**.
- Methods for numerical implementation: **MCMC** (**Metropolis-Hastings sampling, Gibbs sampling, ...**).
- Natural ways to deal with **model selection** and **adaptation**.
- ...

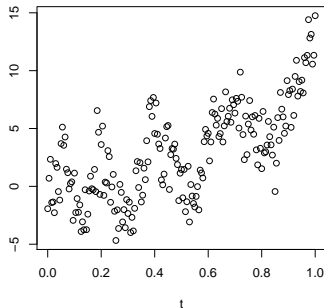
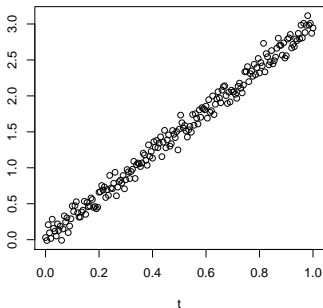
Both practitioners and theoretical statisticians tend to take a pragmatic position nowadays.

Bayesian nonparametrics

Challenges lie in particular in the area of **high-dimensional** or **nonparametric** models.

Illustration: parametric vs. nonparametric regression

$$X_i = \theta(t_i) + \text{error}_i$$



Bayesian nonparametrics

- **When does it work?** There are consistency problems in infinite-dimensional problems (Freedman (1963, 1965), Diaconis and Freedman (1986a,b)).
- **If it works, how well does it work?** At what rate does the posterior corresponding to a given prior contract? Find priors yielding optimal convergence rates.
- **Computational issues:** numerical computation of posterior distributions.

Bayesian practice is currently ahead of the mathematics!

Illustration: nonparametric regression

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 .

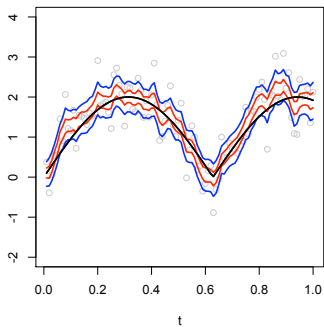
Aim: reconstruct “signal” f .

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

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



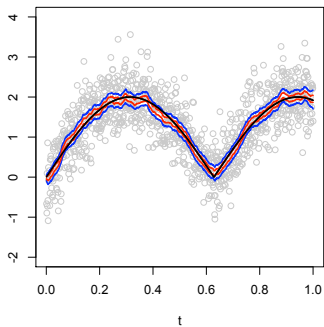
posterior for noise stdev



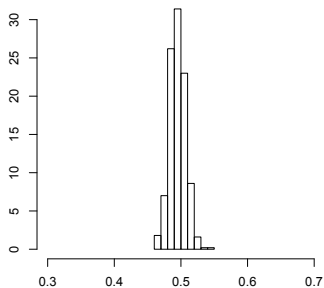
100 observations

Illustration: nonparametric regression

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



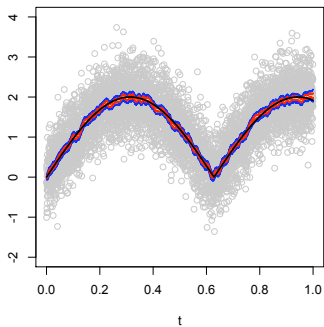
posterior for noise stdev



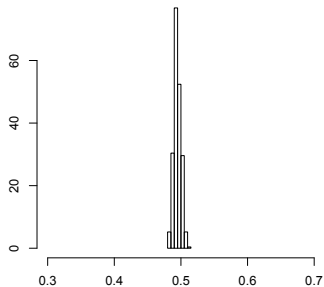
1000 observations

Illustration: nonparametric regression

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



posterior for noise stdev



5000 observations

Illustration: nonparametric regression

Some questions for this concrete problem:

- How did I do the numerical computations?
- Why does this work?
- How fast is the convergence to the unknown signal?
- Is this procedure optimal, or can we do better?
- What is the asymptotic shape of the posterior?
- Does this work in other statistical settings as well?

Frequentist asymptotics for Bayes procedures I

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 II

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:**

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



Gaussian priors I

Gaussian process priors: Π is the law of some (usually centered) Gaussian stochastic process W with continuous sample paths.

Some popular choices:

- **Multiply integrated Brownian motion:**

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

- **Matérn class:** stationary process with spectral measure


$$\mu(d\lambda) \sim \frac{1}{(1+\lambda^2)^{\alpha+1/2}} d\lambda$$

- **Squared exponential process:**

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

Gaussian priors II

Question:

How to verify the entropy, prior mass and remaining mass conditions  for these Gaussian priors?

Setting:

Let W now be a centered Gaussian random element in a separable Banach space $(\mathbb{B}, \|\cdot\|)$. (So: $W : (\Omega, \mathcal{F}, \mathbb{P}) \rightarrow \mathbb{B}$ is Borel measurable and $\forall b^* \in \mathbb{B}^*$: b^*W is a centered, real-valued Gaussian random variable.)

Consider $S : \mathbb{B}^* \rightarrow \mathbb{B}$, $Sb^* = \mathbb{E}Wb^*W$.

Reproducing kernel Hilbert space (RKHS) \mathbb{H} associated with W :
closure of $S\mathbb{B}^*$ with respect to the inner product

$$\langle Sb_1^*, Sb_2^* \rangle_{\mathbb{H}} = \mathbb{E}b_1^*Wb_2^*W.$$

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

Gaussian priors III

Prior mass condition: want to control $\mathbb{P}(\|W - w_0\| < \varepsilon)$.

Concentration function:

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

Lemma.

$$\varphi_{w_0}(\varepsilon) \leq -\log \mathbb{P}(\|W - w_0\| < \varepsilon) \leq \varphi_{w_0}(\varepsilon/2)$$

(main ingredient: Cameron-Martin formula)

Gaussian priors IV

So if ε_n solves

$$\varphi_{w_0}(\varepsilon_n) \leq n\varepsilon_n^2, \quad (1)$$

then

$$\mathbb{P}(\|W - w_0\| < 2\varepsilon_n) \geq e^{-n\varepsilon_n^2}.$$

To solve (1)

- find asymptotic behaviour of the **small ball probability** $\mathbb{P}(\|W\| < \varepsilon)$ for $\varepsilon \rightarrow 0$,
- find approximations of w_0 by elements of the RKHS \mathcal{H} .

Gaussian priors V

How about the entropy and remaining mass conditions?

Gaussian priors V

How about the entropy and remaining mass conditions?

They are **automatically fulfilled** if the prior mass condition (1) holds!

Gaussian priors V

How about the entropy and remaining mass conditions?

They are **automatically fulfilled** if the prior mass condition (1) holds!

Main reasons:

- Connection between small ball probabilities and the entropy of the RKHS unit ball (Kuelbs and Li (1993), Li and Linde (1999)).
- Concentration inequalities for Gaussian measures (Borell, Sudakov).

Gaussian priors VI

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

Borell (1975):

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

Kuelbs and Li (1993), Li and Linde (1999): for

$H(\varepsilon) = \log N(\varepsilon, \mathbb{H}_1, \|\cdot\|)$,

$$H(\varepsilon) \asymp \varepsilon^{-\frac{2\alpha}{2+\alpha}} \iff -\log \mathbb{P}(\|W\| < \varepsilon) \asymp \varepsilon^{-\alpha},$$

$$H(\varepsilon) \asymp \log^\gamma \frac{1}{\varepsilon} \iff -\log \mathbb{P}(\|W\| < \varepsilon) \asymp \log^\gamma \frac{1}{\varepsilon}.$$

Gaussian priors VII

General result:

Theorem.

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

$$\log N(3\varepsilon_n, B_n, \|\cdot\|) \leq 6Cn\varepsilon_n^2,$$

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

$$\mathbb{P}(\|W - w_0\| < 2\varepsilon_n) \geq e^{-n\varepsilon_n^2}.$$

Back to the example I

Suppose we have observations

$$Y_i = f_0(t_i) + \varepsilon_i, \quad i = 1, \dots, n,$$

where $t_i = i/n$, f_0 is an unknown, continuous function, ε_i are independent $N(0, \sigma^2)$ for some *known* σ .

Prior: law Π of a Brownian motion $W = (W_t)_{t \in [0,1]}$ (with standard normal initial distribution).

Back to the example II

Here: $\mathbb{B} = C[0, 1]$ and \mathbb{H} is the Cameron-Martin space:

$$\mathbb{H} = \left\{ t \mapsto c + \int_0^t g(s) ds : c \in \mathbb{R}, g \in L^2[0, 1] \right\},$$

$$\left\| t \mapsto c + \int_0^t g(s) ds \right\|_{\mathbb{H}}^2 = c^2 + \|g\|_{L^2}^2.$$

Well known (consider for instance hitting times):

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

Back to the example III

Lemma.

If $f_0 \in C^\alpha[0, 1]$, $\alpha > 0$, then

$$\inf_{h \in \mathbb{H}: \|h - f_0\|_\infty < \varepsilon} \|h\|_{\mathbb{H}}^2 \lesssim \varepsilon^{-(2-2\alpha)/\alpha}.$$

(Use convolutions)

Back to the example IV

Theorem.

Suppose $f_0 \in C^\alpha[0, 1]$. Then the posterior contracts around θ_0 at the rate

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

This means that for M large enough,

$$\Pi\left(f : \frac{1}{n} \sum_{i=1}^n (f(i/n) - f_0(i/n))^2 > M^2 \varepsilon_n^2 \mid Y_1, \dots, Y_n\right) \xrightarrow{\mathbb{P}_{f_0}} 0$$

Results for other priors I

In Bayesian nonparametrics, we often encounter processes that are not well studied in the small deviations community.

Example: **squared exponential process** with covariance function $\mathbb{E}W_s W_t = ae^{-b(t-s)^2}$.

Small deviations results can then often be derived using the approximation theory connection:

- study/characterize the RKHS and its unit ball,
- derive entropy bounds for that space,
- use the Kuelbs-Li-Linde connection.

Results for other priors II

Example: let $W = (W_t)_{t \in [0,1]}$ be a **stationary** centered Gaussian process with covariance function

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

where

$$\mu(d\lambda) = e^{-|\lambda|^p}, \quad p > 0.$$

Theorem.

$$-\log \mathbb{P}(\|W\|_\infty < \varepsilon) \asymp \left(\log \frac{1}{\varepsilon}\right)^{1+1/p}, \quad 0 < p \leq 1,$$

$$-\log \mathbb{P}(\|W\|_\infty < \varepsilon) \asymp \frac{\left(\log \frac{1}{\varepsilon}\right)^2}{\log \log \frac{1}{\varepsilon}}, \quad 1 < p \leq \infty.$$

Elements of the proof

- For $p > 1$: a typical element

$$h(t) = \int e^{-i\lambda t} \psi(\lambda) \mu(d\lambda)$$

of the RKHS is an entire function on \mathbb{C} and for $h \in \mathbb{H}_1$,

$$|h(z)| \leq C_1 e^{C_2 |z|^{p/(p-1)}}, \quad z \in \mathbb{C}.$$

Kolmogorov and Tihomirov ('61):

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

Conversely, \mathbb{H}_1 contains a similar class of entire functions.

Elements of the proof

- For $p \leq 1$: write $h \in \mathbb{H}_1$ as

$$h(t) = \int_{|\lambda| \leq \nu} e^{-i\lambda t} \psi(\lambda) \mu(d\lambda) + \int_{|\lambda| > \nu} e^{-i\lambda t} \psi(\lambda) \mu(d\lambda).$$

For $\varepsilon > 0$, choose ν such that $|\text{second term}| < \varepsilon$. Then handle the first term as above to get the upper bound. Use Tsirelson's method to get the lower bound.

What we can do at the moment

- derive convergence rates for the most popular Gaussian process priors
- exhibit optimal priors for estimating functions belonging to classical smoothness classes
- understand how to combine Gaussian process priors in order to achieve adaptation

What we would like to understand much better

- asymptotic shape of posterior distributions
- behaviour of empirical Bayes procedures
- good priors for estimating sparse signals
- methods for non-Gaussian priors
-