Nonparametric Bayesian Methods - Lecture I

Harry van Zanten

Korteweg-de Vries Institute for Mathematics

University of Amsterdam

Finnish Summer school in Probability and Statistics
Lammi 30 May–3 June, 2016

Overview of the lectures

I Intro to nonparametric Bayesian statistics

II Consistency and contraction rates

III Contraction rates for Gaussian process priors

(IV Rate-adaptive BNP, Challenges, . . .)

Overview of Lecture I

- Bayesian statistics
- Nonparametric Bayesian statistics
- Nonparametric priors
  - Dirichlet processes
    $\rightarrow$ distribution function estimation
  - Gaussian processes
    $\rightarrow$ nonparametric regression
  - Conditionally Gaussian processes
  - Dirichlet mixtures
    $\rightarrow$ nonparametric density estimation
- Some more examples
- Concluding remarks

Bayesian statistics

Bayesian vs. frequentist statistics

Mathematical statistics:

Have data X, 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" $\theta_0 \in \Theta$. Assume $X \sim P_{\theta_0}$.
- Bayesian paradigm:
  Think of data as being generated in steps as follows:
  - Parameter is random: $\theta \sim \Pi$. Terminology Π: prior.
  - Data given parameter: $X \mid \theta \sim P_\theta$.
  - Can then consider $\theta \mid X$: posterior distribution.

Bayesian vs. frequentist statistics

Mathematical statistics:

Have data X, 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" $\theta_0 \in \Theta$. Assume $X \sim P_{\theta_0}$.
- Bayesian paradigm:
  Think of data as being generated in steps as follows:
  - Parameter is random: $\theta \sim \Pi$. Terminology Π: prior.
  - Data given parameter: $X \mid \theta \sim P_\theta$.
  - Can then consider $\theta \mid X$: posterior distribution.

Bayesian vs. frequentist statistics

Mathematical statistics:

Have data X, 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" $\theta_0 \in \Theta$. Assume $X \sim P_{\theta_0}$.
- Bayesian paradigm:
  Think of data as being generated in steps as follows:
  - Parameter is random: $\theta \sim \Pi$. Terminology Π: prior.
  - Data given parameter: $X \,|\, \theta \sim P_\theta$.
  - Can then consider $\theta \,|\, X$: posterior distribution.

Bayes' example - 1

[Bayes, Price (1763)]

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 $42/50 = 0.84$.

Bayes' example - 1

[Bayes, Price (1763)]

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?

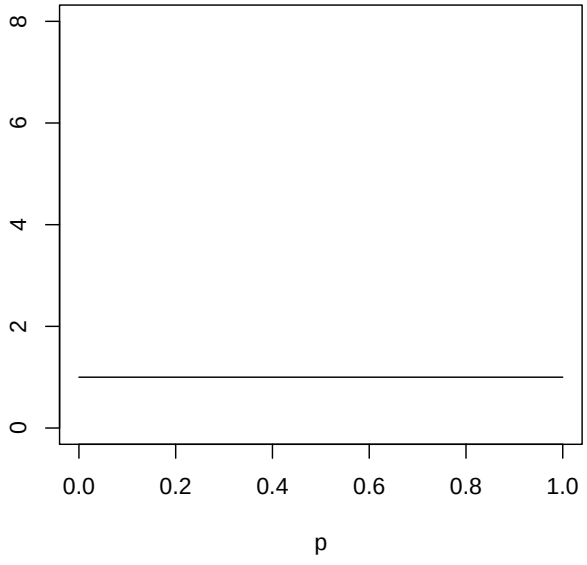
Here we have an observation (the number 42) from a binomial distribution with parameters 50 and p and want to estimate p.

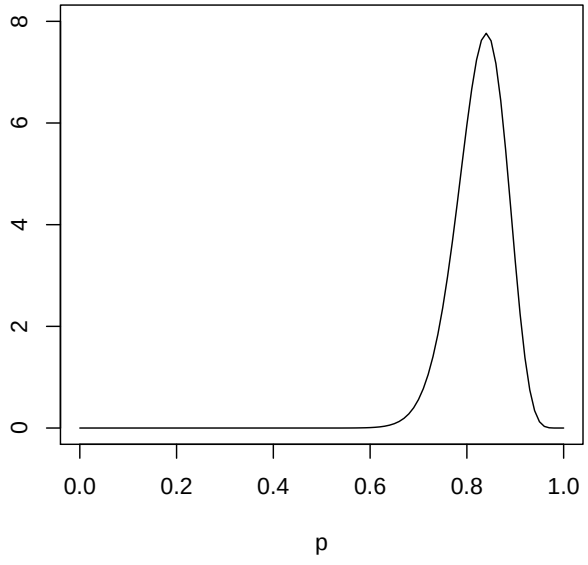
Standard frequentist solution: take the estimate $42/50 = 0.84$.

Bayes' Example - 2

Bayesian approach: choose a prior distribution on p, say uniform on $[0, 1]$. Compute the posterior: beta$(43, 9)$-distribution

(mode is at $42/50 = 0.84$).





prior $\xrightarrow{\text{data}}$ posterior

Bayes' rule

Observations X take values in sample space $\mathcal{X}$. Model $\{P_\theta : \theta \in \Theta\}$. All P_θ dominated: $P_\theta \ll \mu$, density $p_\theta = dP_\theta/d\mu$. Prior distribution Π on the parameter θ.

For the Bayesian: $\theta \sim \Pi$ and $X \,|\, \theta \sim P_\theta$. Hence, the pair (θ, X) has density $(\theta, x) \mapsto p_\theta(x)$ relative to $\Pi \times \mu$. Then X has marginal density

$$x \mapsto \int_\Theta p_\theta(x)\, \Pi(d\theta),$$

and hence the conditional distribution of θ given $X = x$, i.e. the posterior, has density

$$\theta \mapsto \frac{p_\theta(x)}{\int_\Theta p_\theta(x)\, \Pi(d\theta)}$$

relative to the prior Π.

Bayes' example again

Have $X \sim \text{Bin}(n, \theta)$, $\theta \in (0, 1)$. Likelihood:

$$p_\theta(X) = \binom{n}{X} \theta^X (1-\theta)^{n-X}.$$

Prior: uniform distribution on $(0, 1)$. By Bayes' rule, posterior density proportional to

$$\theta \mapsto \theta^X (1-\theta)^{n-X}.$$

Hence, posterior is $\text{Beta}(X+1, n-X+1)$.

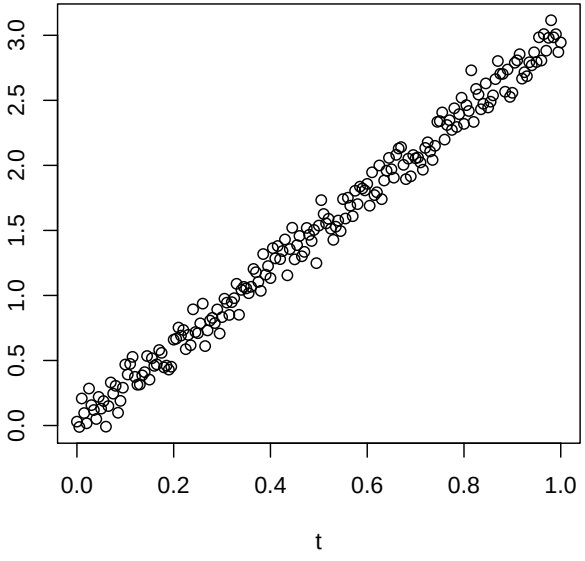
Bayesian nonparametrics

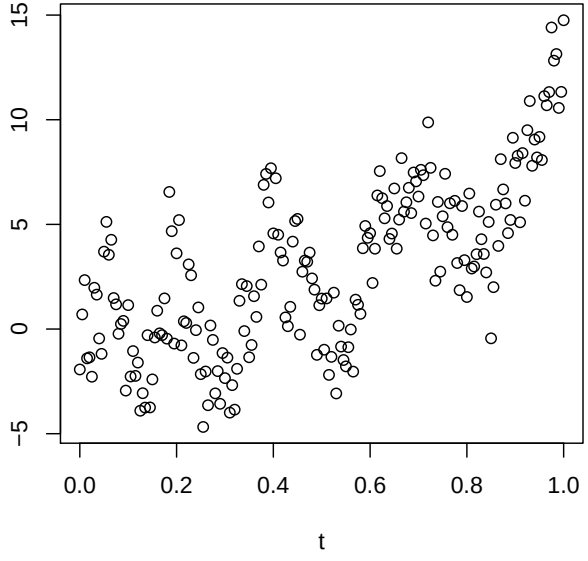
Bayesian nonparametrics

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

Illustration 1: parametric vs. nonparametric regression

$$Y_i = f(t_i) + \text{error}_i$$

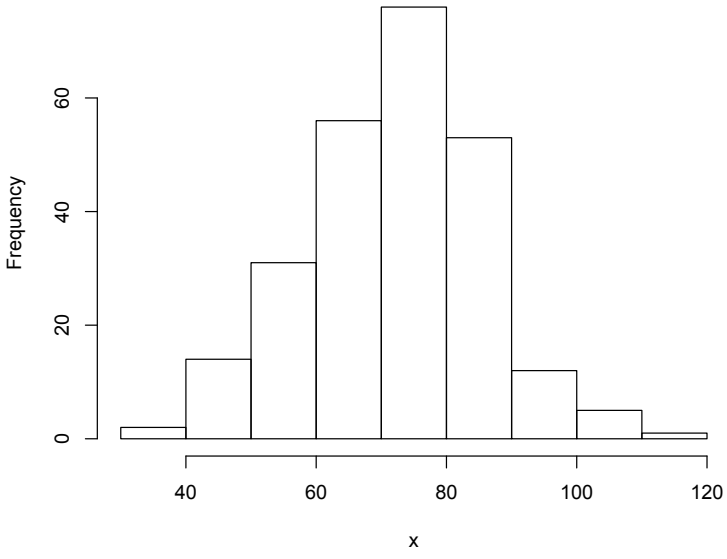


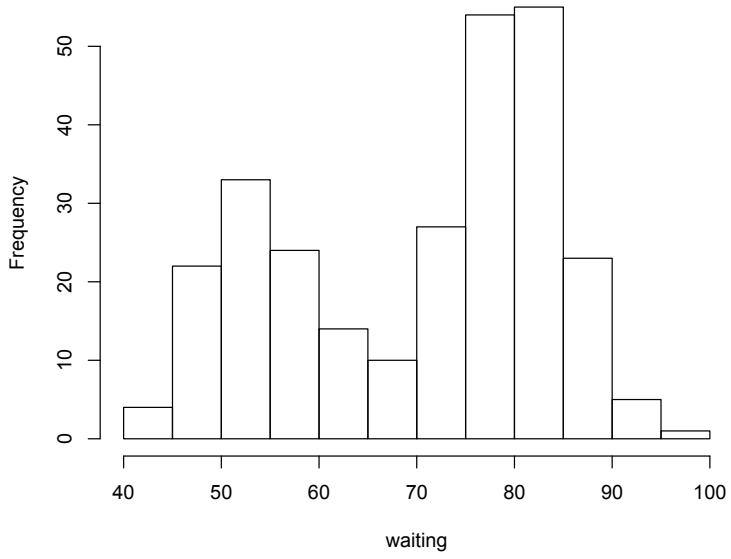


Bayesian nonparametrics

Illustration 2: parametric vs. nonparametric density estimation

$$X_1, \ldots, X_n \sim f$$





Bayesian nonparametrics

In nonparametric problems, the parameter of interest is typically a function: e.g. a density, regression function, distribution function, hazard rate, . . . , or some other infinite-dimensional object.

Bayesian approach is not at all fundamentally restricted to the parametric case, but:

- How do we construct priors on infinite-dimensional (function) spaces?
- How do we compute posteriors, or generate draws?
- What is the fundamental performance of procedures?

Nonparametric priors

Nonparametric priors - first remarks

- Often enough to describe how realizations are generated
- Possible ways to construct priors on an infinite-dimensional space Θ:
  - Discrete priors: Consider (random) points $\theta_1, \theta_2, \ldots,$ in Θ and (random) probability weights $w_1, w_2, \ldots$ and define $\Pi = \sum w_j \delta_{\theta_j}$.
  - Stochastic Process approach: If Θ is a function space, use machinery for constructing stochastic processes
  - Random series approach: If Θ is a function space, consider series expansions, put priors on coefficients
  - ...

Nonparametric priors

- Dirichlet process

Dirichlet process - 1

Step 1: prior on simplex of probability vectors of length k:

$$\Delta^{k-1} = \{(y_1, \ldots, y_k) \in \mathbb{R}^k : y_1 \geq, \ldots, y_k \geq 0, \sum y_i = 1\}.$$

For $\alpha = (\alpha_1, \ldots, \alpha_k) \in (0, \infty)^k$, define

$$f_\alpha(y_1, \ldots, y_{k-1}) = C_\alpha \prod_{i=0}^{k} y_i^{\alpha_i - 1} 1_{(y_1, \ldots, y_k) \in \Delta^{k-1}}$$

on $\mathbb{R}^{k-1}$, where $y_k = 1 - y_1 - \cdots - y_{k-1}$ and C_α is the appropriate normalizing constant.

A random vector $(Y_1, \ldots, Y_k)$ in $\mathbb{R}^k$ is said to have a Dirichlet distribution with parameter $\alpha = (\alpha_1, \ldots, \alpha_k)$ if $(Y_1, \ldots, Y_{k-1})$ has density f_α and $Y_k = 1 - Y_1 + \cdots + Y_{k-1}$.

Dirichlet process - 2

Step 2: definition of DP:

Let α be a finite measure on $\mathbb{R}$. A random probability measure P on $\mathbb{R}$ is called a Dirichlet Process with parameter α if for every partition $A_1, \ldots, A_k$ of $\mathbb{R}$, the vector $(P(A_1), \ldots, P(A_k))$ has a Dirichlet distribution with parameter $(\alpha(A_1), \ldots, \alpha(A_k))$.

Notation: $P \sim DP(\alpha)$.

Dirichlet process - 3

Step 3: Prove that DP exists!

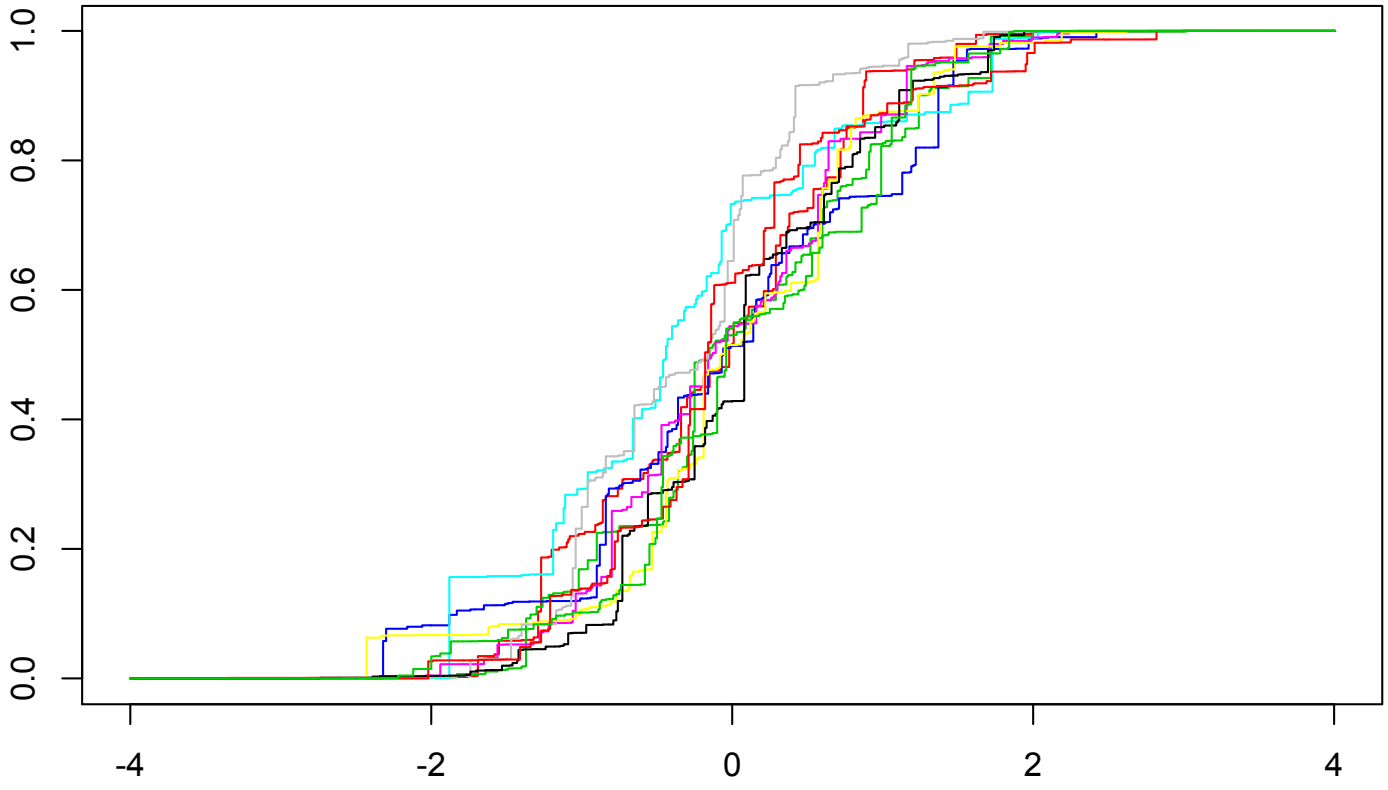
Theorem.
For any finite measure α on $\mathbb{R}$, the Dirichlet process with parameter α exists.

Proof.
For instance:

- Use Kolmogorov's consistency theorem to show $\exists$ a process $P = (P(A) : A \in \mathcal{B}(\mathbb{R}))$ with the right fdd's.
- Prove there exists a version of P such that every realization is a measure.

$\square$

Dirichlet process - 4



Ten realizations from Dirichlet process with parameter $25 \times N(0, 1)$

Dirichlet process - 5

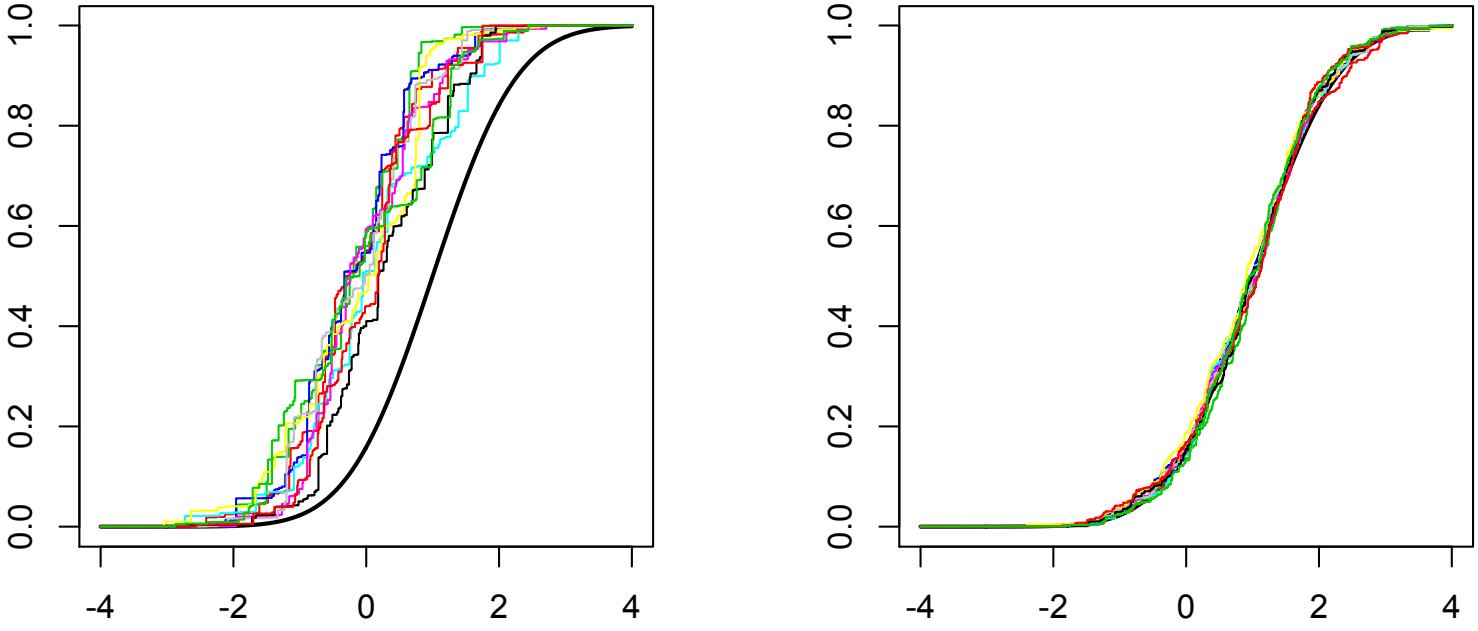
Draws from the DP are discrete measures on $\mathbb{R}$:

Theorem.
Let α be a finite measure, define $M = \alpha(\mathbb{R})$ and $\bar{\alpha} = \alpha/M$. If we have independent $\theta_1, \theta_2, \ldots \sim \bar{\alpha}$ and $Y_1, Y_2, \ldots \sim Beta(1, M)$ and $V_j = Y_j \prod_{l=1}^{j-1}(1 - Y_l)$, then $\sum_{j=1}^{\infty} V_j \delta_{\theta_j} \sim DP(\alpha)$.

This is the stick-breaking representation.

Distribution function estimation

The DP is a conjugate prior for full distribution estimation: if $P \sim DP(\alpha)$ and $X_1, \ldots, X_n \,|\, P \sim P$, then $P \,|\, X_1, \ldots, X_n \sim DP(\alpha + \sum_{i=1}^{n} \delta_{X_i})$.



Simulated data: 500 draws from a $N(1, 1)$-distribution, prior: Dirichlet process with parameter $25 \times N(0, 1)$.

Left: 10 draws from the prior. Right: 10 draws from the posterior.

Nonparametric priors

- Gaussian processes

Gaussian process priors - 1

A stochastic process $W = (W_t : t \in T)$ is called Gaussian if for all $n \in \mathbb{N}$ and $t_1, \ldots, t_n \in T$, the vector $(W_{t_1}, \ldots, W_{t_n})$ has an n-dimensional Gaussian distribution.

Associated functions:

- mean function: $m(t) = \mathbb{E}W_t$,
- covariance function: $r(s,t) = \mathbb{C}\mathrm{ov}(W_s, W_t)$.

The GP is called centered, or zero-mean if $m(t) = 0$ for all $t \in T$.

Gaussian process priors - 2

For $a_1, \ldots, a_n \in \mathbb{R}$ and $t_1, \ldots, t_n \in T$,

$$\sum_i \sum_j a_i a_j r(t_i, t_j) = \mathbb{V}\mathrm{ar}\Big(\sum a_i W_{t_i} \Big) \geq 0,$$

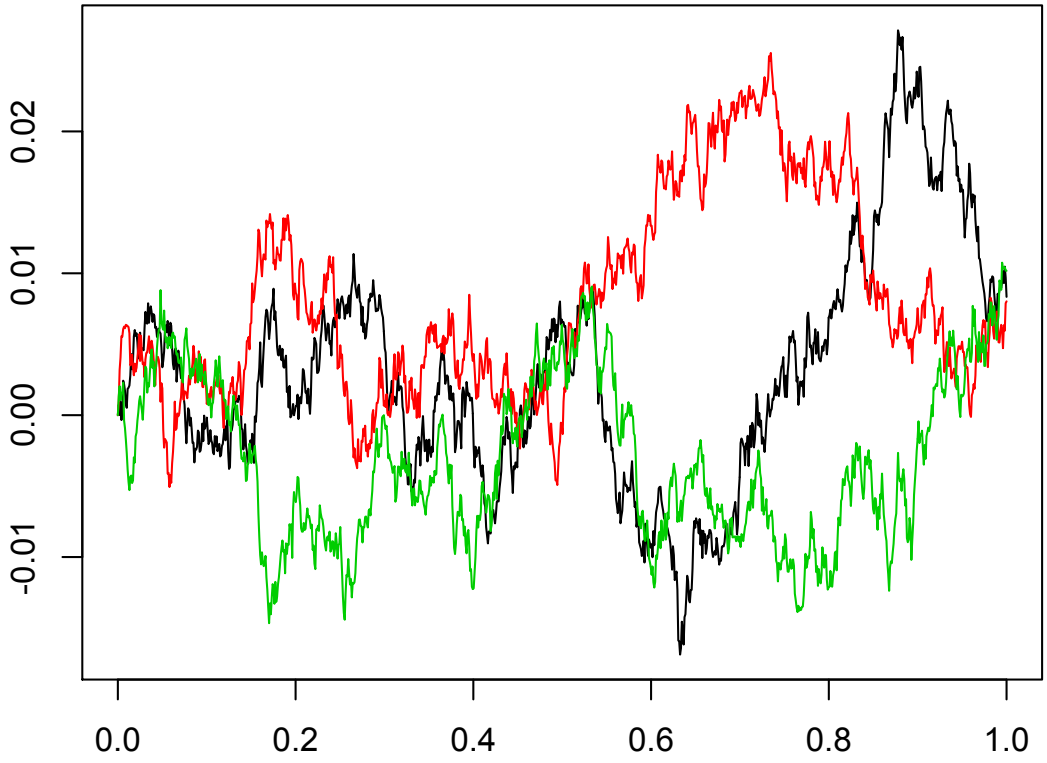
hence r is a positive definite, symmetric function on $T \times T$.

Theorem.
Let T be a set, $m : T \to \mathbb{R}$ a function and $r : T \times T \to \mathbb{R}$ a positive definite, symmetric function. Then there exists a Gaussian process with mean function m and covariance function r.

Proof.
Kolmogorov's consistency theorem. □

Gaussian process priors: examples - 1

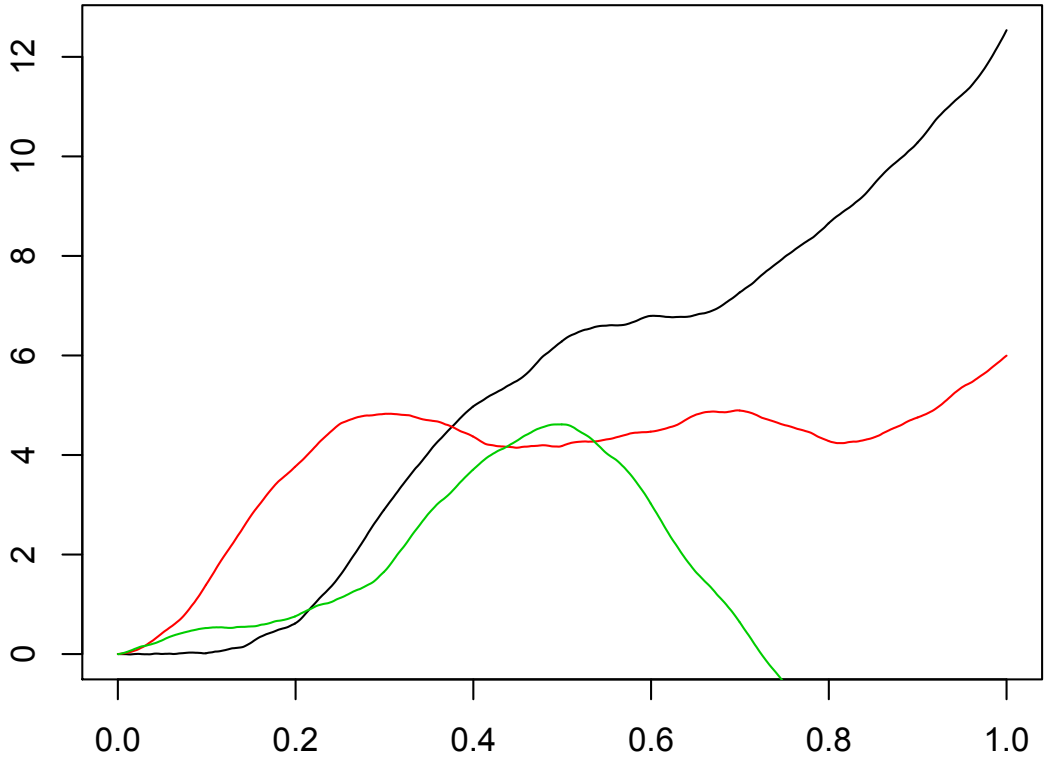
Brownian motion: $m(t) = 0$, $r(s,t) = s \wedge t$.



Regularity: $1/2$.

Gaussian process priors: examples - 2

Integrated Brownian motion: $\int_0^t W_s \, ds$, for W a Brownian motion.
$m(t) = 0$, $r(s,t) = s^2 t/2 - t^3/6$.



Regularity: $3/2$.

Gaussian process priors: examples - 3

By Fubini and integration by parts,

$$\int_0^t \int_0^{t_n} \cdots \int_0^{t_2} W_{t_1}\, dt_1 dt_2 \cdots dt_n = \frac{1}{(n-1)!} \int_0^t (t-s)^{n-1} W_s\, ds$$
$$= \frac{1}{n!} \int_0^t (t-s)^n\, dW_s.$$

The Riemann-Liouville process with parameter $\alpha > 0$:

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

Process has regularity α.

Gaussian process priors: examples - 4

Consider a centered Gaussian process $W = (W_t : t \in T)$, with $T \subseteq \mathbb{R}^d$, such that

$$\mathbb{E} W_s W_t = r(t - s), \qquad s, t \in T,$$

for a continuous $r : \mathbb{R}^d \to \mathbb{R}$. Such a process is called stationary, or homogenous.

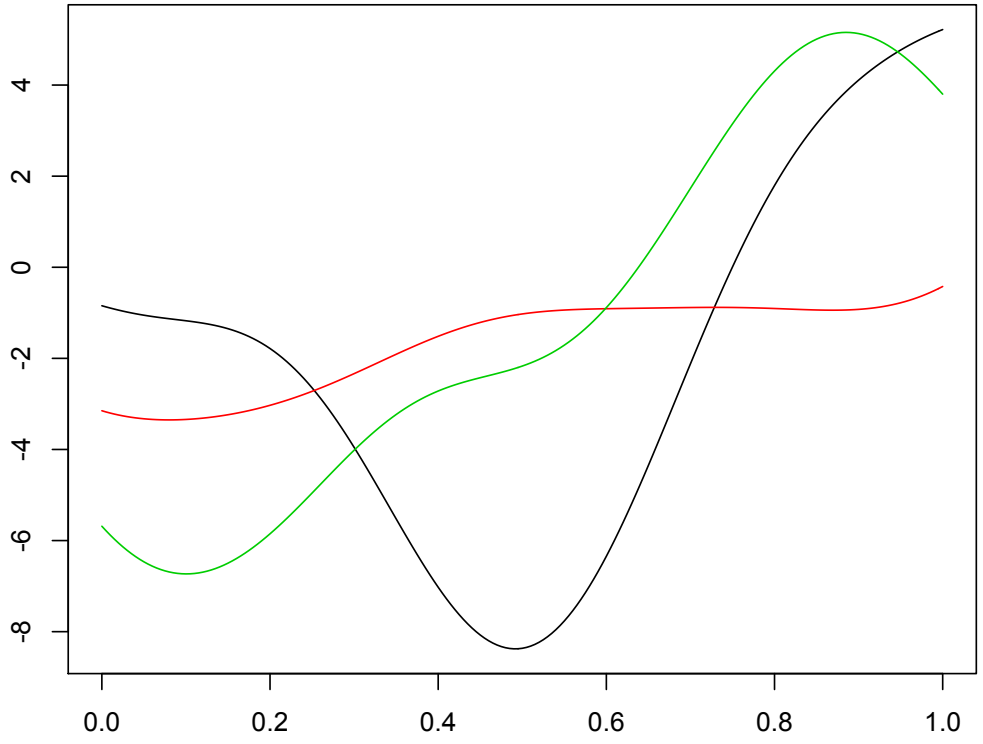
By Bochner's theorem:

$$r(t) = \int_{\mathbb{R}^d} e^{-i\langle \lambda, t \rangle} \, \mu(d\lambda),$$

for a finite Borel measure μ, called the spectral measure of the process.

Gaussian process priors: examples - 5

The squared exponential process: $r(s,t) = \exp(-\|t-s\|^2)$
Spectral measure: $2^{-d}\pi^{-d/2}\exp(-\|\lambda\|^2/4)\,d\lambda$.



Regularity: ∞.

Gaussian process priors: examples - 6

The Matérn process: $\mu(d\lambda) \propto (1 + \|\lambda\|^2)^{-(\alpha+d/2)}\, d\lambda$, $\alpha > 0$.

Covariance function:

$$r(s,t) = \frac{2^{1-\alpha}}{\Gamma(\alpha)} \|t-s\|^{\alpha} K_{\alpha}(\|t-s\|),$$

where K_α is the modified Bessel function of the second kind of order α.

Regularity: α.

For $d = 1$, $\alpha = 1/2$, get the Ornstein-Uhlenbeck process.

Gaussian process priors: examples - 6

The Matérn process: $\mu(d\lambda) \propto (1 + \|\lambda\|^2)^{-(\alpha+d/2)}\, d\lambda$, $\alpha > 0$.

Covariance function:

$$r(s,t) = \frac{2^{1-\alpha}}{\Gamma(\alpha)} \|t-s\|^{\alpha} K_{\alpha}(\|t-s\|),$$

where K_α is the modified Bessel function of the second kind of order α.

Regularity: α.

For $d = 1$, $\alpha = 1/2$, get the Ornstein-Uhlenbeck process.

Gaussian process regression - 1

Observations:

$$X_i = f(t_i) + \varepsilon_i,$$

$t_i \in [0,1]$ fixed ε_i independent $N(0,1)$.

Prior on f: law of a centered GP with covariance function r.

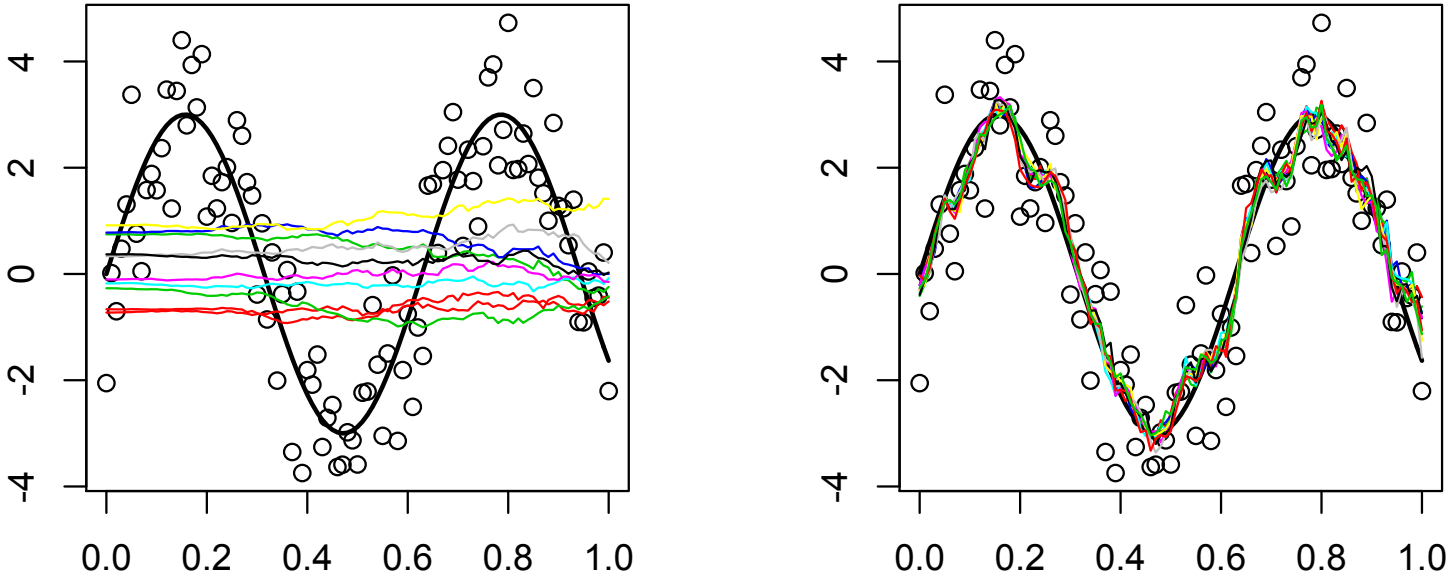
Posterior: this prior is conjugate for this model:

$$(f(t_1), \ldots, f(t_n)) \,|\, X_1, \ldots, X_n \sim N_n((I + \Sigma^{-1})^{-1} X, (I + \Sigma^{-1})^{-1}),$$

where Σ the is matrix with $\Sigma_{ij} = r(t_i, t_j)$.

Gaussian process regression - 2

Data: 200 simulated data points.
Prior: multiple of integrated Brownian motion.



Left: 10 draws from the prior. Right: 10 draws from the posterior.

Nonparametric priors

- Conditionally Gaussian processes

CGP's - 1

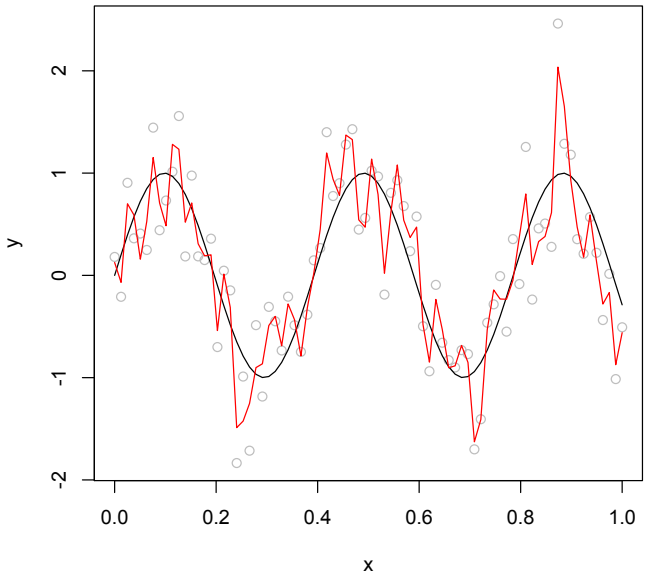
Observation about GP's:

- Families of GP's typically depend on auxiliary parameters: hyper parameters.
- Performance can heavily depend on tuning of parameters.
- How to choose values of hyper parameters?

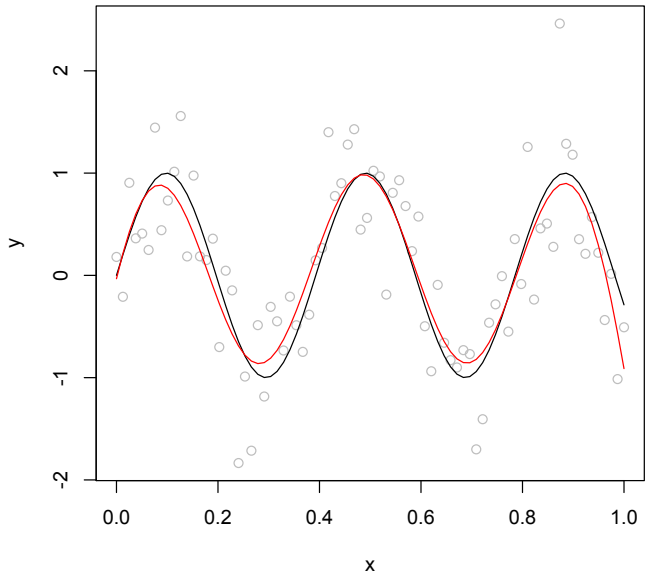
CGP's - 2

Regression with a squared exponential GP with covariance $(x, y) \mapsto \exp(-(x-y)^2/\ell^2)$, for different length scale hyper parameters ℓ.

ℓ too small:



ℓ correct:



CGP's - 3

- Q: How to choose the best values of hyper parameters?
- A: Let the data decide!

Possible approaches:

- Put a prior on the hyper parameters as well: full Bayes
- Estimate hyper parameters : empirical Bayes

CGP's - 3

- Q: How to choose the best values of hyper parameters?
- A: Let the data decide!

Possible approaches:

- Put a prior on the hyper parameters as well: full Bayes
- Estimate hyper parameters : empirical Bayes

CGP's - 4

Squared exponential GP with gamma length scale:

$$\ell \sim \Gamma(a, b)$$
$$f \mid \ell \sim GP \quad \text{with cov } (x, y) \mapsto \exp(-(x-y)^2/\ell^2)$$

- Example of a hierarchical prior
- Prior is only conditionally Gaussian

Q: does this solve the bias-variance issue?

CGP's - 4

Squared exponential GP with gamma length scale:

$$\ell \sim \Gamma(a, b)$$
$$f \mid \ell \sim GP \quad \text{with cov } (x, y) \mapsto \exp(-(x-y)^2/\ell^2)$$

- Example of a hierarchical prior
- Prior is only conditionally Gaussian

Q: does this solve the bias-variance issue?

Nonparametric priors

- Dirichlet mixtures

DP mixture priors - 1

Idea:

- Consider location/scale mixtures of Gaussians of the form

$$p_G(x) = \int\int \varphi_\sigma(x-\mu)\, G(d\mu, d\sigma),$$

where

$\varphi_\sigma(\cdot - \mu)$ is the $N(\mu, \sigma^2)$-density

G is a probability measure (mixing measure).

- Construct a prior on densities by making G random.

DP mixture priors - 2

Draw g from a Gaussian DP mixture prior:

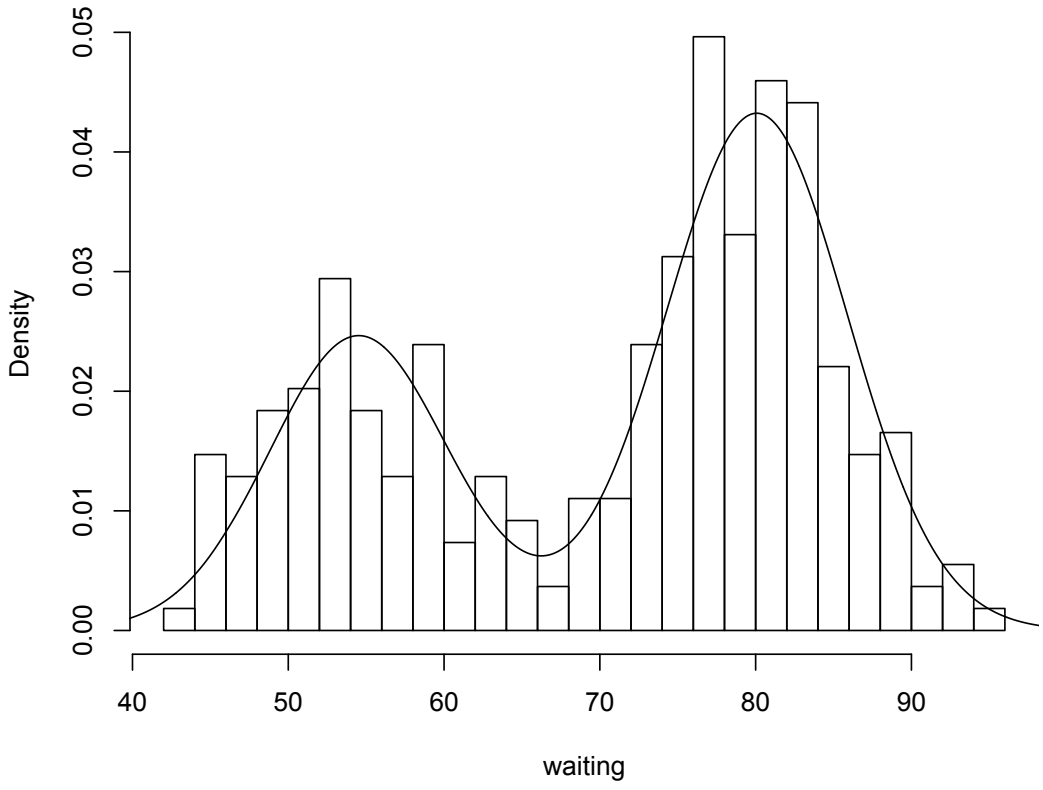
$$G \sim DP(G_0) \qquad (G_0 \text{ often } N \times IW)$$
$$p \,|\, G \sim p_G$$

Another example of a hierarchical prior

DP mixture density estimation

Data: 272 waiting times between geyser eruptions
Prior: DP mixture of normals

Posterior mean:



Some more examples

Estimating the drift of a diffusion

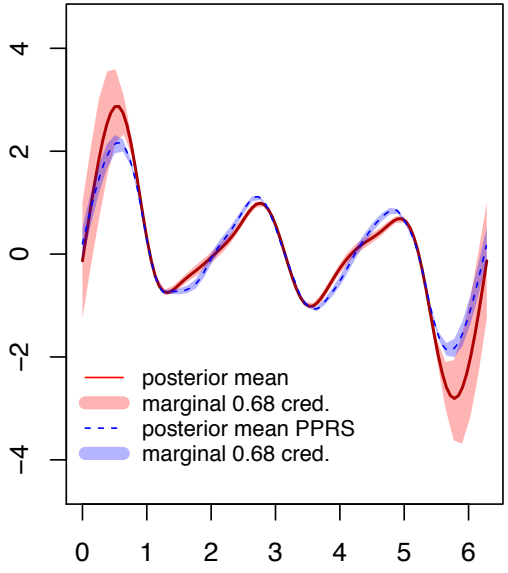
Observation model: $dX_t = b(X_t)\,dt + dW_t$. Goal: estimate b.

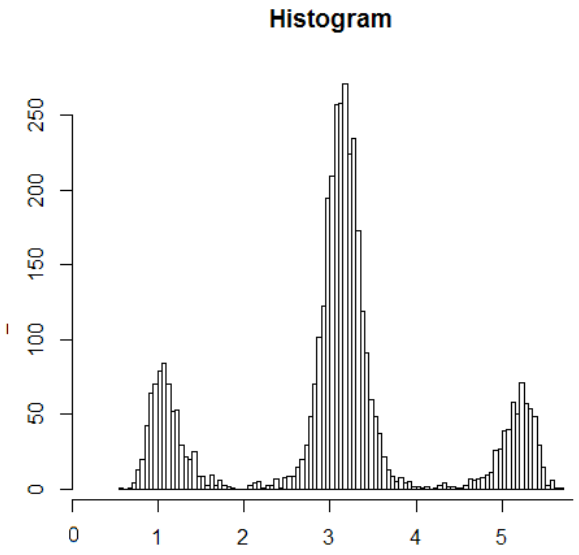
Prior:

$$s \sim IG(a, b)$$

$$J \sim Ps(\lambda)$$

$$b \,|\, s, J \sim s\sum_{j=1}^{J} j^{-2} Z_j e_j \quad e_j\text{: Fourier basis, } Z_j \sim N(0,1)$$



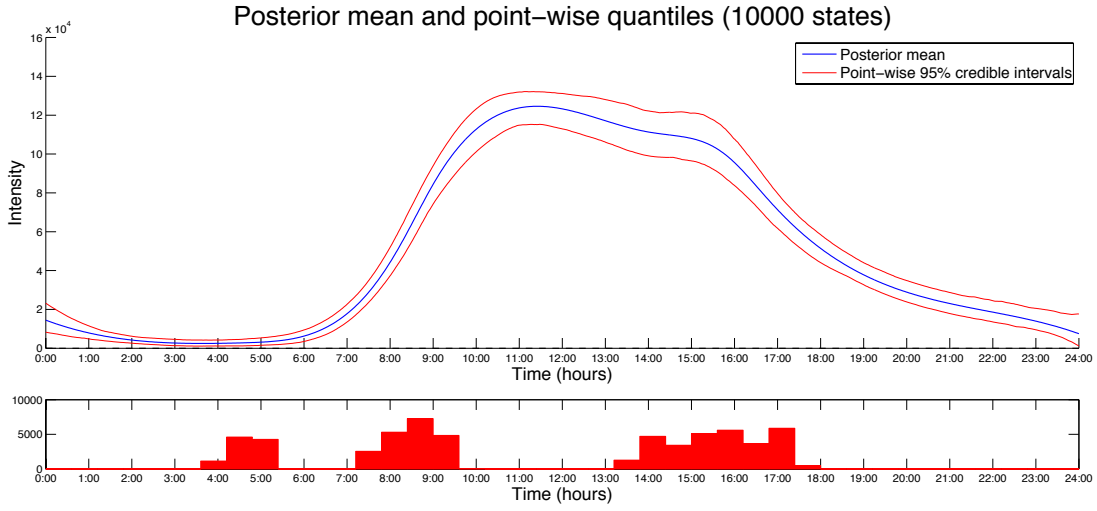


[Van der Meulen, Schauer, vZ. (2014)]

Nonparametric estimation of a Poisson intensity

Observation model: counts from an inhomogenous Poisson process with periodic intensity λ. Goal: estimate λ.

Prior: B-spline expansion with priors on knots and coefficients

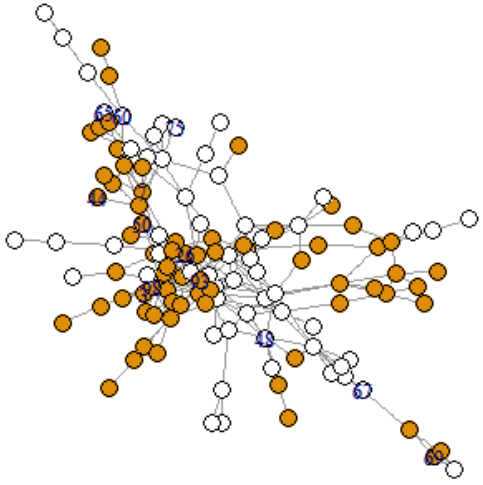


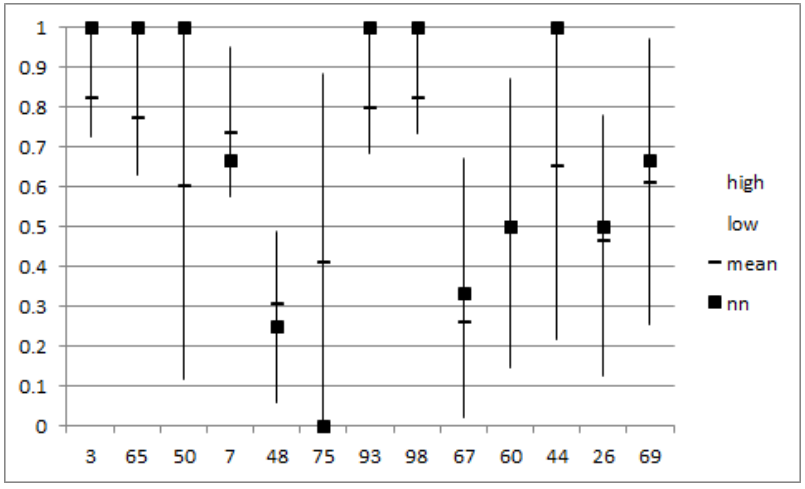
[Belitser, Serra, vZ. (2015)]

Binary prediction on a graph

Observation model: $\mathbb{P}(Y_i = 1) = \Psi(f(i))$, for $f : G \to \mathbb{R}$.

Prior: Conditionally Gaussian with precision L^p, L: graph Laplacian





[Hartog, vZ. (in prep.)]

Concluding remarks

Take home from Lecture I

- Within the Bayesian paradigm it is perfectly possible and natural to deal with nonparametric statistical problems.
- Many nonparametric priors have been proposed and studied: DP's, GP's, DP mixtures, series expansion, ...
- Numerical techniques have been developed to sample from the corresponding posteriors
- In a variety of statistical settings, the results can be quite satisfactory.

Some (theoretical) questions:

- So do these procedures do what we expect them to do?
- Why/why not?
- Do they have desirable properties like consistency?
- Can we say something more about performance, e.g. about (optimal) convergence rates?

Take home from Lecture I

- Within the Bayesian paradigm it is perfectly possible and natural to deal with nonparametric statistical problems.
- Many nonparametric priors have been proposed and studied: DP's, GP's, DP mixtures, series expansion, ...
- Numerical techniques have been developed to sample from the corresponding posteriors
- In a variety of statistical settings, the results can be quite satisfactory.

Some (theoretical) questions:

- So do these procedures do what we expect them to do?
- Why/why not?
- Do they have desirable properties like consistency?
- Can we say something more about performance, e.g. about (optimal) convergence rates?

Some references for Lecture I - 1

DP:

- Ferguson, T. S. (1973). A Bayesian analysis of some nonparametric problems. Annals of Statistics, 1, 209–30.

DP mixtures:

- Ferguson, T. S. (1983). Bayesian density estimation by mixtures of normal distributions. In Recent Advances in Statistics, ed. M. Rizvi et al., 287–302.
- Escobar, M. and West, M. (1995). Bayesian density estimation and inference using mixtures. Journal of the American Statistical Association, 90, 577–88.
- MacEachern, S. N. and Muller, P. (1998) Estimating mixture of Dirichlet Process Models. Journal of Computational and Graphical Statistics, 7 (2), 223–338.
- Neal, R. M. (2000). Markov Chain sampling methods for Dirichlet process mixture models. Journal of Computational and Graphical Statistics, 9, 249–265.

Some references for Lecture I - 2

GP priors:

- Lenk, P. J. (1988). The logistic normal distribution for Bayesian, nonparametric, predictive densities. J. Amer. Statist. Assoc. 83 509–516.
- Lenk, P. J. (1991). Towards a practicable Bayesian nonparametric density estimator. Biometrika 78 531–543.
- Rasmussen, C. E. and Williams, C. K. (2006). Gaussian Processes for Machine Learning. MIT Press, Cambridge, MA.

General text:

- Hjort, N.L., et al., eds. Bayesian nonparametrics. Vol. 28. Cambridge University Press, 2010.

Nonparametric Bayesian Methods - Lecture II

Harry van Zanten

Korteweg-de Vries Institute for Mathematics

University of Amsterdam

Finnish Summer school in Probability and Statistics
Lammi 30 May–3 June, 2016

Overview of Lecture II

- Frequentist asymptotics
- Parametric models: Bernstein - Von Mises
- Consistency: Doob and Schwartz
- General rate of contraction results
- Concluding remarks

Frequentist asymptotics

Illustration: nonparametric regression

Suppose we have observations

$$Y_i = f(t_i) + e_i, \qquad i = 1, \ldots, n,$$

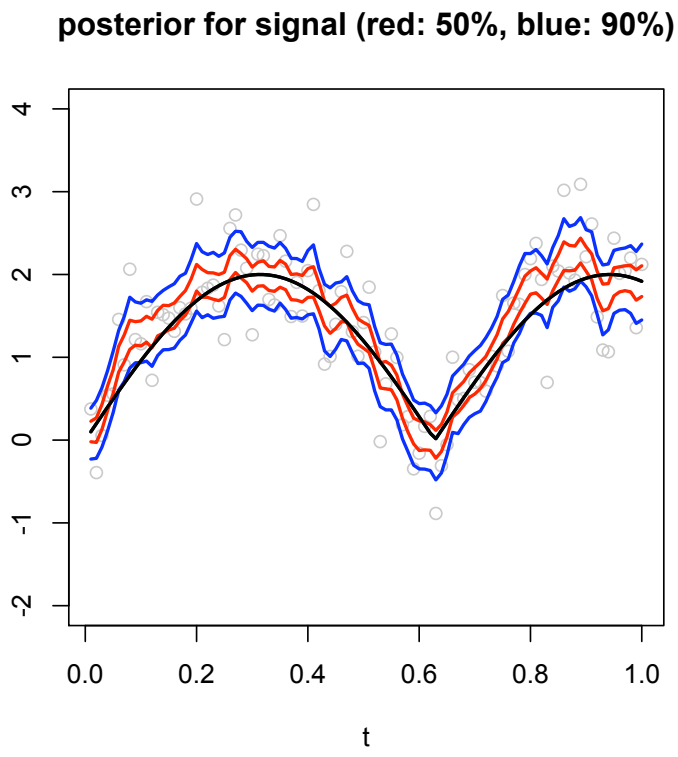
where $t_i = i/n$, f is an unknown, continuous function, e_i are independent $N(0, \sigma^2)$ for some unknown σ.

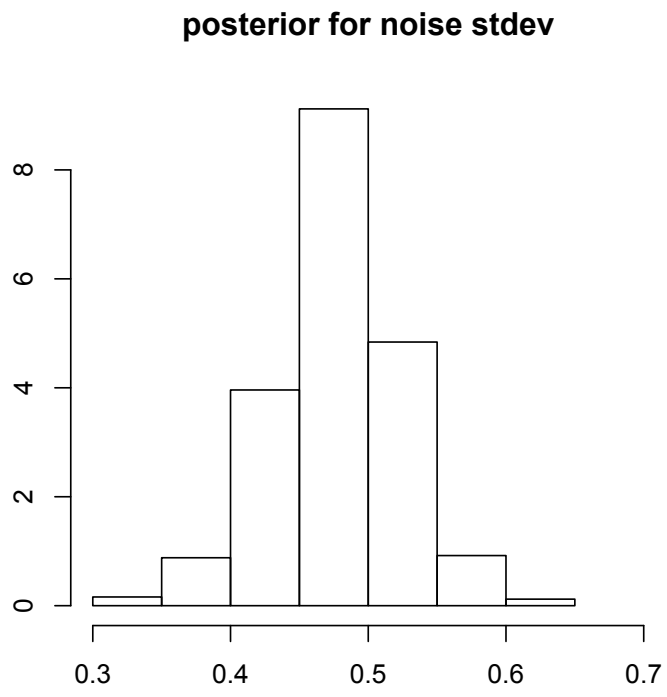
Aim: reconstruct "signal" f.

Approach:

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

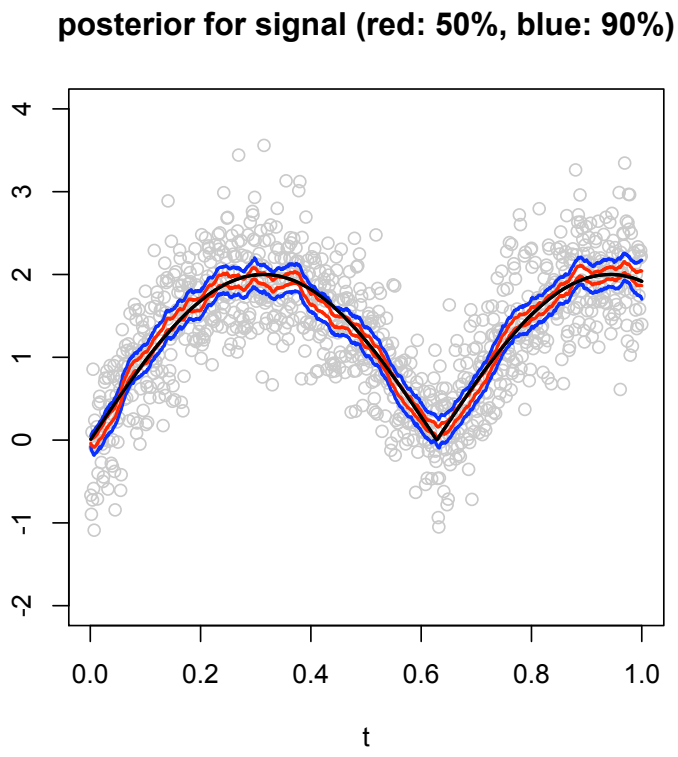
Illustration: nonparametric regression

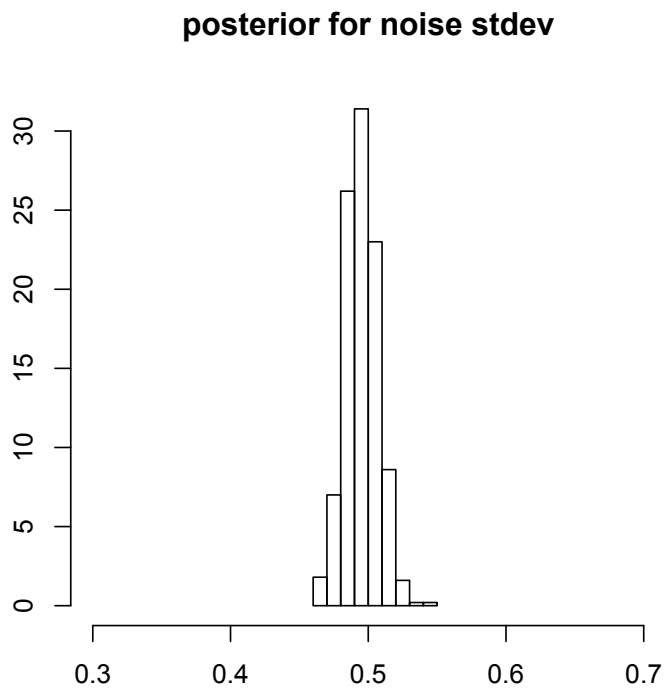




100 observations

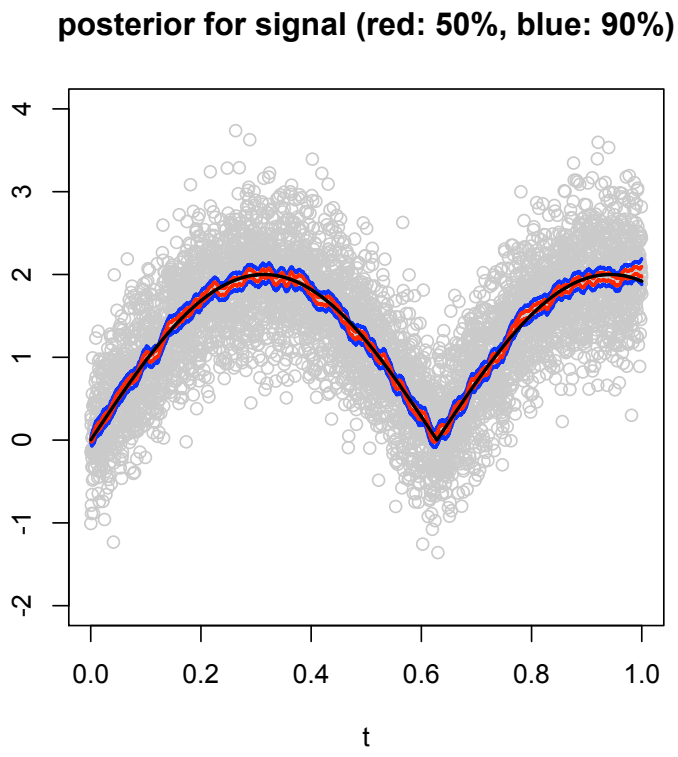
Illustration: nonparametric regression

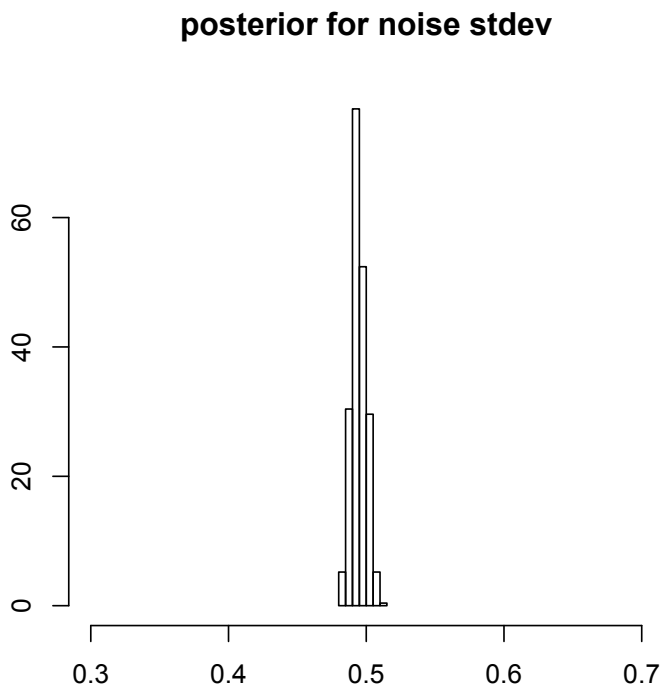




1000 observations

Illustration: nonparametric regression





5000 observations

Illustration: nonparametric regression

Questions we are interested in:

- Why does this work?
- How fast is the convergence to the unknown signal?
- Is this procedure optimal, or can we do better?
- Does this work in other statistical settings as well?
- ...

Frequentist asymptotics - 1

Suppose there is a true parameter θ_0 $(= (f_0, \sigma_0))$ generating the data. Say $\theta_0 \in \Theta$, for a parameter space Θ.

Consider a prior Π on the space Θ, compute the corresponding posterior $\Pi(\cdot \mid Y_1, \ldots, Y_n)$. (Usually, Θ is defined indirectly, as the support of the prior Π).

Say there is some natural distance d on the space Θ. Want to understand if for "large n", "most" of the posterior mass is concentrated "close to" the true parameter θ_0.

Frequentist asymptotics - 2

Main questions, more precisely:

- Consistency: does the posterior contract around θ_0 as $n \to \infty$? I.e., for all $\varepsilon > 0$, is it true that

$$\Pi(\theta \in \Theta : d(\theta, \theta_0) > \varepsilon \,|\, Y_1, \ldots, Y_n) \overset{P_{\theta_0}}{\to} 0?$$

- Contraction rate: how fast can we let $\varepsilon_n \downarrow 0$ such that still

$$\Pi(\theta \in \Theta : d(\theta, \theta_0) > M\varepsilon_n \,|\, Y_1, \ldots, Y_n) \overset{P_{\theta_0}}{\to} 0$$

for all $M > 0$ large enough?

Frequentist asymptotics: a short history

Parametric models: Bernstein-Von Mises theorem (Laplace (1800), Von Mises (30's), Le Cam (80's))

Nonparametrics:

- 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
- '98/'99: Important extensions of Schwartz by Barron, Schervish & Wasserman, and Ghosal, Ghosh & Ramamoorthi
- '00/'01: Rate of contraction results by Ghosal, Ghosh & Van der Vaart and Shen & Wasserman.

Asymptotics for parametric models:

Bernstein - Von Mises

BvM - 1

Let $X_1, \ldots, X_n$ be a sample from a density p_{θ_0}, $\theta_0 \in \Theta \subset \mathbb{R}$. Suppose $\theta \mapsto p_\theta$ is "smooth". Consider MLE and Fisher info:

$$\hat{\theta}_n = \operatorname*{argmax}_{\theta \in \Theta} \prod_{i=1}^{n} p_\theta(X_i),$$

$$i_\theta = \mathbb{V}\mathrm{ar}_\theta \frac{\partial \log p_\theta(X_1)}{\partial \theta}$$

Then under "regularity conditions",

$$\sqrt{n}(\hat{\theta}_n - \theta_0) \Rightarrow N(0, i_{\theta_0}^{-1})$$

under P_{θ_0} as $n \to \infty$.

BvM - 2

Consider a prior Π on Θ with Lebesgue density π. Posterior:

$$\Pi(B \mid X_1, \ldots, X_n) = \frac{\int_B \prod p_\theta(X_i)\pi(\theta)\, d\theta}{\int_\Theta \prod p_\theta(X_i)\pi(\theta)\, d\theta}.$$

BvM: If π is positive and continuous at θ_0, then, under "regularity conditions"

$$\left\| \Pi(\cdot \mid X_1, \ldots, X_n) - N(\hat{\theta}_n, (n i_{\theta_0})^{-1}) \right\|_{TV} \overset{P_{\theta_0}}{\to} 0$$

as $n \to \infty$.

In particular: rate of contraction in the parametric case is $n^{-1/2}$ under mild conditions.

BvM - "proof" - 1

Set $h = \sqrt{n}(\theta - \theta_0)$. Have

$$\Pi(h \in B \,|\, X_1, \ldots, X_n) = \frac{\int_{\theta:\sqrt{n}(\theta-\theta_0)\in B} e^{\sum \ell_\theta(X_i)} \pi(\theta)\, d\theta}{\int_{\mathbb{R}} e^{\sum \ell_\theta(X_i)} \pi(\theta)\, d\theta},$$

with $\ell_\theta(x) = \log p_\theta(x)$. By Taylor,

$$\ell_\theta(x) - \ell_{\theta_0}(x) \approx (\theta - \theta_0)\dot{\ell}_{\theta_0}(x) + \frac{1}{2}(\theta - \theta_0)^2 \ddot{\ell}_{\theta_0}(x).$$

By the LLN, we have $\mathbb{P}_{\theta_0}$-a.s.

$$-\frac{1}{n}\sum_{i=1}^{n} \ddot{\ell}_{\theta_0}(X_i) \to -\mathbb{E}_{\theta_0}\ddot{\ell}_{\theta_0}(X_1) = \mathbb{V}\mathrm{ar}_{\theta_0}\dot{\ell}_{\theta_0}(X_1) = i_{\theta_0}.$$

Hence,

$$\begin{aligned} e^{\sum(\ell_\theta - \ell_{\theta_0})(X_i)} &\approx e^{-\frac{1}{2} i_{\theta_0}(n(\theta-\theta_0)^2 - 2\sqrt{n}(\theta-\theta_0)\Delta_n)} \\ &= e^{-\frac{1}{2} i_{\theta_0}(h-\Delta_n)^2} e^{\frac{1}{2} i_{\theta_0}\Delta_n^2}, \end{aligned}$$

BvM - "proof" - 2

where

$$\Delta_n = \frac{1}{i_{\theta_0}\sqrt{n}} \sum_{i=1}^{n} \dot{\ell}_{\theta_0}(X_i).$$

We get

$$\Pi(\sqrt{n}(\theta - \theta_0) \in B \,|\, X_1, \ldots, X_n) \approx \frac{\int_B e^{-\frac{1}{2} i_{\theta_0}(h-\Delta_n)^2} \pi(\theta_0 + h/\sqrt{n})\, dh}{\int e^{-\frac{1}{2} i_{\theta_0}(h-\Delta_n)^2} \pi(\theta_0 + h/\sqrt{n})\, dh}.$$

Now let $n \to \infty$ and conclude that

$$\Pi(B \,|\, X_1, \ldots, X_n) \approx N\Big(\theta_0 + \frac{\Delta_n}{\sqrt{n}}, \frac{1}{n i_{\theta_0}}\Big)(B).$$

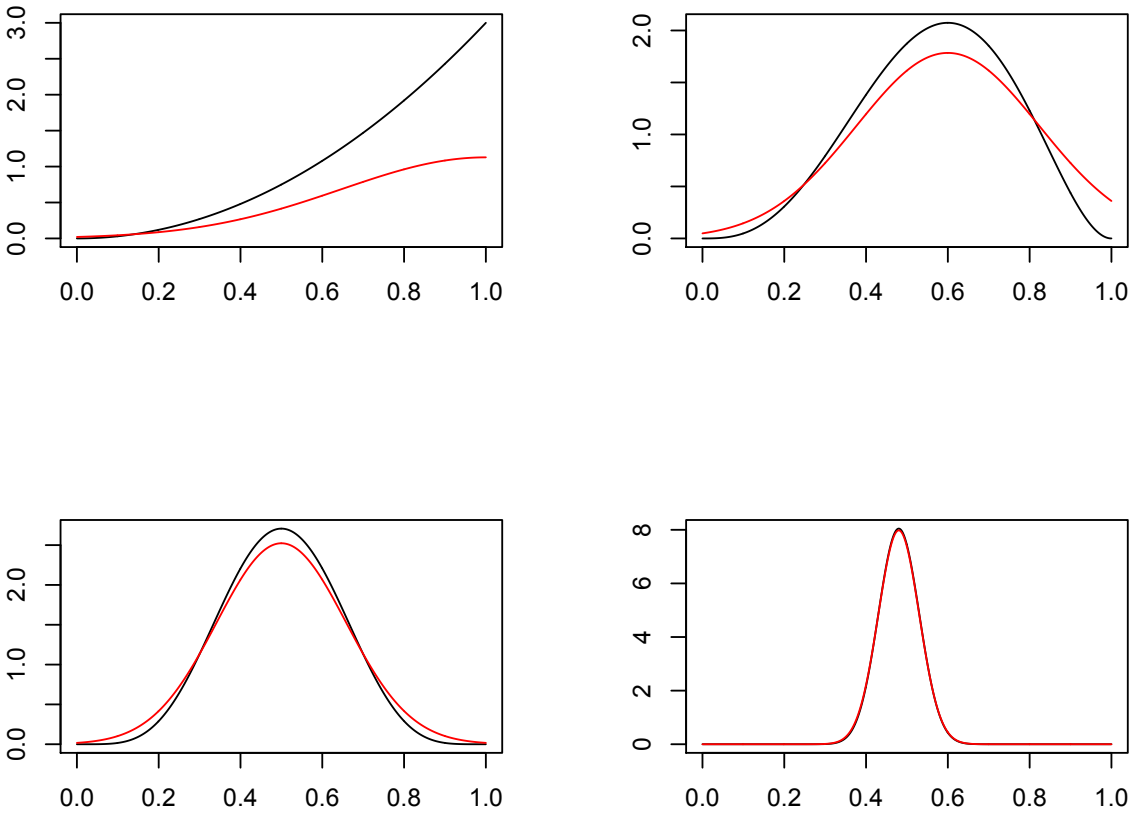
The LAN expansion also implies that

$$\hat{\theta}_n \approx \theta_0 + \frac{\Delta_n}{\sqrt{n}}.$$

Combining the last two displays completes the "proof". □

BvM - numerical illustration

Consider $X \sim \text{Bin}(n, p)$, uniform prior on p. Here $\hat{p} = X/n$, $I_p = n/(p(1-p))$. Posterior is $\text{Beta}(X+1, n-X+1)$.



n = 2, 5, 10, 100

Consistency:

Doob and Schwartz

Doob's theorem

Consider i.i.d. $X_1, \ldots, X_n \sim P_\theta$, $\theta \in \Theta$, for a "nice" metric space (Θ, d). Assume that $\theta \mapsto P_\theta$ is appropriately measurable. Let Π be a prior on (the Borel sets of) Θ.

Theorem.
Suppose that if $\theta_1 \neq \theta_2$, then $P_{\theta_1} \neq P_{\theta_2}$ (identifiability). Then for Π-almost all $\theta_0 \in \Theta$ and all $\varepsilon > 0$:

$$\Pi(\theta : d(\theta, \theta_0) > \varepsilon \,|\, X_1, \ldots X_n) \to 0,$$

$\mathbb{P}_{\theta_0}$-a.s..

This is called (strong) posterior consistency, or consistency at θ_0.

Doob's theorem

Consider i.i.d. $X_1, \ldots, X_n \sim P_\theta$, $\theta \in \Theta$, for a "nice" metric space (Θ, d). Assume that $\theta \mapsto P_\theta$ is appropriately measurable. Let Π be a prior on (the Borel sets of) Θ.

Theorem.
Suppose that if $\theta_1 \neq \theta_2$, then $P_{\theta_1} \neq P_{\theta_2}$ (identifiability). Then for Π-almost all $\theta_0 \in \Theta$ and all $\varepsilon > 0$:

$$\Pi(\theta : d(\theta, \theta_0) > \varepsilon \,|\, X_1, \ldots X_n) \to 0,$$

$\mathbb{P}_{\theta_0}$-a.s..

This is called (strong) posterior consistency, or consistency at θ_0.

Side remark: relation to consistency of estimators

Proposition.

Suppose we have posterior consistency at θ_0, relative to the metric d. Define the estimator $\hat{\theta}_n$ as the center of a ball of minimal radius that has posterior mass at least $1/2$. Then $\hat{\theta}_n$ is consistent at θ_0, i.e.

$$d(\hat{\theta}_n, \theta_0) \to 0,$$

$\mathbb{P}_{\theta_0}$-a.s..

Proof.

Let $B(\hat{\theta}_n, \hat{r})$ be a ball of minimal radius that has posterior mass at least $1/2$. For every $\varepsilon > 0$, $B(\theta_0, \varepsilon)$ asymptotically contains posterior mass 1. Hence, $\hat{r} \leq \varepsilon$. Moreover, the balls can not be disjoint. By the triangle inequality, it follows that, asymptotically, $d(\hat{\theta}_n, \theta_0) \leq \hat{r} + \varepsilon \leq 2\varepsilon$. □

Doob's theorem - "proof" - 1

Let Q be the joint distribution of θ and $X_1, X_2, \ldots$ in the Bayesian framework, i.e. under Q have $\theta \sim \Pi$ and $X_1, X_2, \ldots \mid \theta$ are i.i.d. P_θ.

The posterior is the Q-conditional distribution of $\theta \mid X_1, \ldots, X_n$.

Note:

- LLN implies that $\forall \theta$, for P_θ-almost all $X_1, X_2, \ldots$, can identify P_θ from $X_1, X_2, \ldots$:

$$\frac{1}{n} \sum_{i=1}^{n} 1_A(X_i) \to P_\theta(A), \qquad P_\theta\text{-a.s..}$$

- Identifiability assumption implies we can identify θ from P_θ.

Doob's theorem - "proof" - 2

Using some measure theory, it follows $\exists$ measurable $h : \mathbb{R}^\infty \to \Theta$:

$$h(x_1, x_2, \ldots) = \theta, \qquad \text{for } Q\text{-almost all } (\theta, x_1, x_2, \ldots).$$

Using Doob's martingale convergence theorem, we get, Q-a.s.

$$\Pi(\theta : d(\theta, \theta_0) > \varepsilon \,|\, X_1, \ldots, X_n) = \mathbb{E}_Q(1_{d(\theta,\theta_0)>\varepsilon} \,|\, X_1, \ldots, X_n)$$
$$\to \mathbb{E}_Q(1_{d(\theta,\theta_0)>\varepsilon} \,|\, X_1, X_2, \ldots) = 1_{d(\theta,\theta_0)>\varepsilon} = 1_{d(h(X_1,X_2,\ldots),\theta_0)>\varepsilon}.$$

From this, can derive that for Π-almost all θ_0, P_{θ_0}-a.s.

$$\Pi(\theta : d(\theta, \theta_0) > \varepsilon \,|\, X_1, \ldots, X_n) \to 0.$$

□

Limitations of Doob's theorem

Main issues:

- In infinite-dimensional spaces, null sets can be very large.
  $\rightarrow \exists$ examples of inconsistent procedures (take $\Pi = \delta_\theta$).
- Result is very pessimistic!
  $\rightarrow$ in many cases of interest, consistency actually holds for many more θ_0 than Doob says.

Need a different approach to obtain less pessimistic results...

Limitations of Doob's theorem

Main issues:

- In infinite-dimensional spaces, null sets can be very large.
  $\rightarrow \exists$ examples of inconsistent procedures (take $\Pi = \delta_\theta$).
- Result is very pessimistic!
  $\rightarrow$ in many cases of interest, consistency actually holds for many more θ_0 than Doob says.

Need a different approach to obtain less pessimistic results...

Schwartz' theorem - setting

Observations: sample $X_1, \ldots, X_n$ from a density $p_0 \in \mathcal{P}$, for $\mathcal{P}$ the collection of densities on the unit interval.

Prior: measure Π on $\mathcal{P}$

Posterior:

$$\Pi(B \mid X_1, \ldots, X_n) = \frac{\int_B \prod_{i=1}^n p(X_i)\, \Pi(dp)}{\int_{\mathcal{P}} \prod_{i=1}^n p(X_i)\, \Pi(dp)}.$$

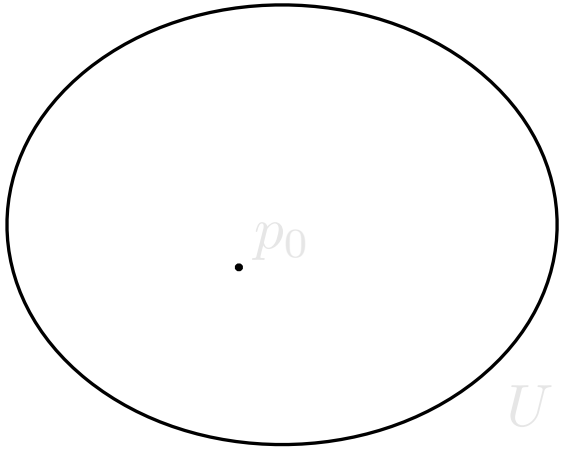
Schwartz' theorem - 1

Idea: replace identifiability by a stronger condition on testability.

Assume that for a neighborhood U of p_0: can consistently test $H_0 : p = p_0$ against $H_1 : p \in U^c$.

More precisely, assume there exist $[0, 1]$-valued tests $\varphi_n = \varphi_n(X_1, \ldots, X_n)$ such that

$$\mathbb{E}_{p_0} \varphi_n \to 0 \qquad \sup_{p \in U^c} \mathbb{E}_p (1 - \varphi_n) \to 0.$$



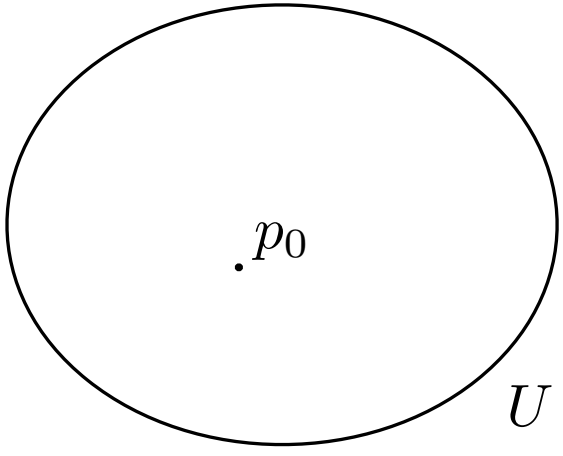
Interpretation: φ_n = probab. of rejecting $H_0 : p = p_0$.

Schwartz' theorem - 1

Idea: replace identifiability by a stronger condition on testability.
Assume that for a neighborhood U of p_0: can consistently test $H_0 : p = p_0$ against $H_1 : p \in U^c$.

More precisely, assume there exist $[0, 1]$-valued tests $\varphi_n = \varphi_n(X_1, \ldots, X_n)$ such that

$$\mathbb{E}_{p_0} \varphi_n \to 0 \qquad \sup_{p \in U^c} \mathbb{E}_p (1 - \varphi_n) \to 0.$$



Interpretation: φ_n = probab. of rejecting $H_0 : p = p_0$.

Schwartz' theorem - 2

Consistency for which p_0?

Define the Kulback-Leibler divergence

$$K(p,q) = \int p(x) \log \frac{p(x)}{q(x)}\, dx.$$

Consistency for Π-almost all p_0 will be replaced by consistency for all p_0 in the KL-support of Π: say that $p_0 \in KL(\Pi)$ if for all $\varepsilon > 0$,

$$\Pi(p : K(p_0, p) < \varepsilon) > 0.$$

Hence, get consistency if prior puts mass "arbitrarily close" to p_0 (in KL-sense) $\rightarrow$ enter approximation theory.

Schwartz' theorem - 2

Consistency for which p_0?

Define the Kulback-Leibler divergence

$$K(p,q) = \int p(x) \log \frac{p(x)}{q(x)}\, dx.$$

Consistency for Π-almost all p_0 will be replaced by consistency for all p_0 in the KL-support of Π: say that $p_0 \in KL(\Pi)$ if for all $\varepsilon > 0$,

$$\Pi(p : K(p_0, p) < \varepsilon) > 0.$$

Hence, get consistency if prior puts mass "arbitrarily close" to p_0 (in KL-sense) $\rightarrow$ enter approximation theory.

Schwartz' theorem - 2

Consistency for which p_0?

Define the Kulback-Leibler divergence

$$K(p, q) = \int p(x) \log \frac{p(x)}{q(x)} \, dx.$$

Consistency for Π-almost all p_0 will be replaced by consistency for all p_0 in the KL-support of Π: say that $p_0 \in KL(\Pi)$ if for all $\varepsilon > 0$,

$$\Pi(p : K(p_0, p) < \varepsilon) > 0.$$

Hence, get consistency if prior puts mass "arbitrarily close" to p_0 (in KL-sense) $\rightarrow$ enter approximation theory.

Schwartz' theorem - 2

Consistency for which p_0?

Define the Kulback-Leibler divergence

$$K(p, q) = \int p(x) \log \frac{p(x)}{q(x)} \, dx.$$

Consistency for Π-almost all p_0 will be replaced by consistency for all p_0 in the KL-support of Π: say that $p_0 \in KL(\Pi)$ if for all $\varepsilon > 0$,

$$\Pi(p : K(p_0, p) < \varepsilon) > 0.$$

Hence, get consistency if prior puts mass "arbitrarily close" to p_0 (in KL-sense) $\rightarrow$ enter approximation theory.

Schwartz' theorem - 1

Theorem.
Suppose that p_0 is in the KL-support of the prior and that for a neighborhood $U \subset \mathcal{P}$ of p_0, there exist tests such that $\mathbb{E}_{p_0}\varphi_n \to 0$ and $\sup_{p \in U^c} \mathbb{E}_p(1-\varphi_n) \to 0$. Then

$$\Pi(U^c \mid X_1, \ldots, X_n) \to 0$$

$\mathbb{P}_0$-a.s..

Hence we have (strong) consistency at p_0 if (i) tests exist for *every* neighborhood U of p_0 and (ii) p_0 is in the KL-support of the prior.

Schwartz' theorem - 1

Theorem.
Suppose that p_0 is in the KL-support of the prior and that for a neighborhood $U \subset \mathcal{P}$ of p_0, there exist tests such that $\mathbb{E}_{p_0}\varphi_n \to 0$ and $\sup_{p \in U^c} \mathbb{E}_p(1-\varphi_n) \to 0$. Then

$$\Pi(U^c \mid X_1, \ldots, X_n) \to 0$$

$\mathbb{P}_0$-a.s..

Hence we have (strong) consistency at p_0 if (i) tests exist for *every* neighborhood U of p_0 and (ii) p_0 is in the KL-support of the prior.

Schwartz' theorem - "proof"

Write

$$\Pi(U^c \mid X_1, \ldots, X_n) \leq \varphi_n + \frac{\int_{U^c}(1-\varphi_n)\frac{d\mathbb{P}^n}{d\mathbb{P}_0^n}\,\Pi(dp)}{\int_{\mathcal{P}}\frac{d\mathbb{P}^n}{d\mathbb{P}_0^n}\,\Pi(dp)}.$$

Denominator: restrict integral to KL ball + Jensen+ LLN, get

$$\text{denominator} \geq e^{-n\varepsilon}\Pi(p_0 : KL(p_0, p) < \varepsilon).$$

Combine with testing assumptions. Use that by Hoeffding, can assume tests have exponential power. □

Schwartz' theorem - weak topology

Note: the testing condition depends on the topology.

Example.

If U is a weak neighborhood of the form

$$U = \left\{ p : \mathbb{E}_p \psi(X_1) < \mathbb{E}_{p_0} \psi(X_1) + \varepsilon \right\}$$

for a bounded, continuous ψ and $\varepsilon > 0$, then required tests always exist, by Hoeffding.

Hence, always have consistency relative to the weak topology for every p_0 in the KL-support of the prior.

For consistency in stronger topologies (e.g. Hellinger, L^1), more is needed.

Extended Schwartz theorem

More useful version:

Theorem.
Suppose that p_0 is in the KL-support of the prior and that for a neighborhood $U \subset \mathcal{P}$ of p_0, there exist tests φ_n and $C > 0$ and $\mathcal{P}_n \subset \mathcal{P}$ such that

$$\mathbb{E}_{p_0}\varphi_n \leq e^{-Cn}, \qquad \sup_{p \in U^c \cap \mathcal{P}_n} \mathbb{E}_p(1-\varphi_n) \leq e^{-Cn}, \quad \Pi(\mathcal{P}_n^c) \leq e^{-Cn}.$$

Then

$$\Pi(U^c \mid X_1, \ldots, X_n) \to 0$$

$\mathbb{P}_0$-a.s..

$\mathcal{P}_n$: sieves. Idea: sets with very little prior mass (like $\mathcal{P}_n^c$) get no posterior mass, asymptotically.

Extended Schwartz theorem - Hellinger topology - 1

Define the Hellinger distance by

$$h^2(p,q) = \int (\sqrt{p(x)} - \sqrt{q(x)})^2 \, dx.$$

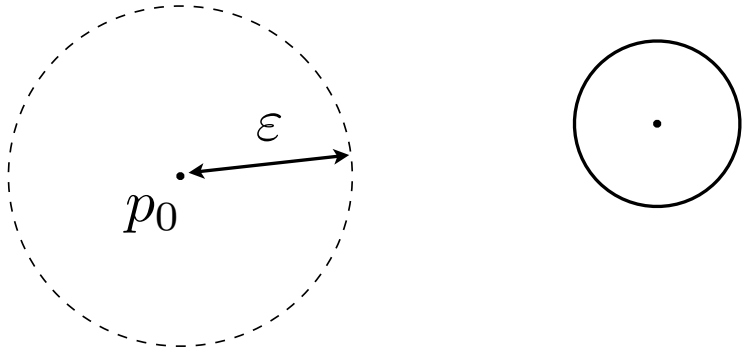
Theorem.
For every *convex* $\mathcal{Q} \subset \mathcal{P}$ such that $h(p_0, p) > \varepsilon$ for all $p \in \mathcal{Q}$, there exists a test φ_n s.t.

$$\mathbb{E}_{p_0} \varphi_n \leq e^{-n\varepsilon^2/2}, \qquad \sup_{p \in \mathcal{Q}} \mathbb{E}_p (1 - \varphi_n) \leq e^{-n\varepsilon^2/2}.$$

"Proof."
Minimize $\varphi_n \mapsto \mathbb{E}_{p_0} \varphi_n + \sup_{p \in \mathcal{Q}} \mathbb{E}_p (1 - \varphi_n)$ over all tests φ_n using the minimax theorem. □

Extended Schwartz theorem - Hellinger topology - 2

Theorem gives tests for $H_0 : p = p_0$ against the hypothesis H_1 that p is in a ball at Hellinger distance at least ε from p_0:

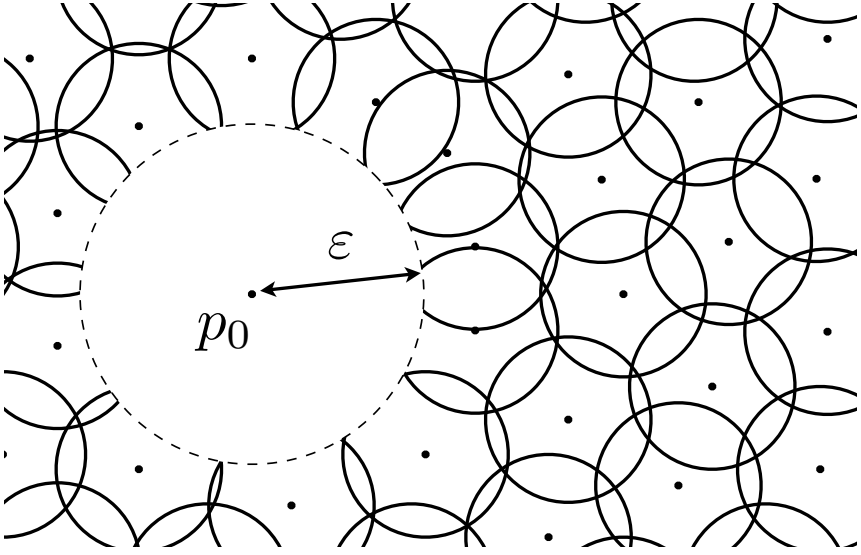


For consistency relative to the Hellinger distance, need test for p_0 against the complement of the ε-ball around p_0 (intersected with a sieve $\mathcal{P}_n$).

Extended Schwartz theorem - Hellinger topology - 3

Idea:

- Cover the complement of the ε-ball with small balls.
- For every such small ball, have a local test for p_0 against that small ball.
- Make a new global test that rejects $H_0 : p = p_0$ if any of the local tests rejects it.



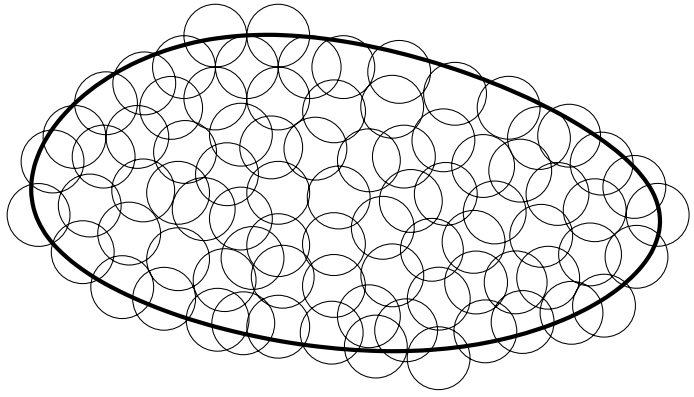
Extended Schwartz theorem - Hellinger topology - 4

Power of the new global test depends on the number of small balls needed.

Covering number: for $\mathcal{Q} \subset \mathcal{P}$ and $\varepsilon > 0$, define

$N(\varepsilon, \mathcal{Q}, h) =$

minimum number of h-balls of radius ε needed to cover $\mathcal{Q}$.



We call $\log N(\varepsilon, \mathcal{Q}, h)$ the metric entropy of $\mathcal{Q}$ w.r.t. h.

Extended Schwartz theorem - Hellinger topology - 5

Theorem.
Suppose that $p_0 \in KL(\Pi)$ and that for every $\varepsilon > 0$, there exist $\mathcal{P}_n \subset \mathcal{P}$ and constants $C < 6$ and $D > 0$ such that $N(\varepsilon, \mathcal{P}_n, h) \leq \exp(Cn\varepsilon^2)$ and $\Pi(\mathcal{P}_n^c) \leq \exp(-Dn)$. Then we have posterior consistency w.r.t. the Hellinger distance.

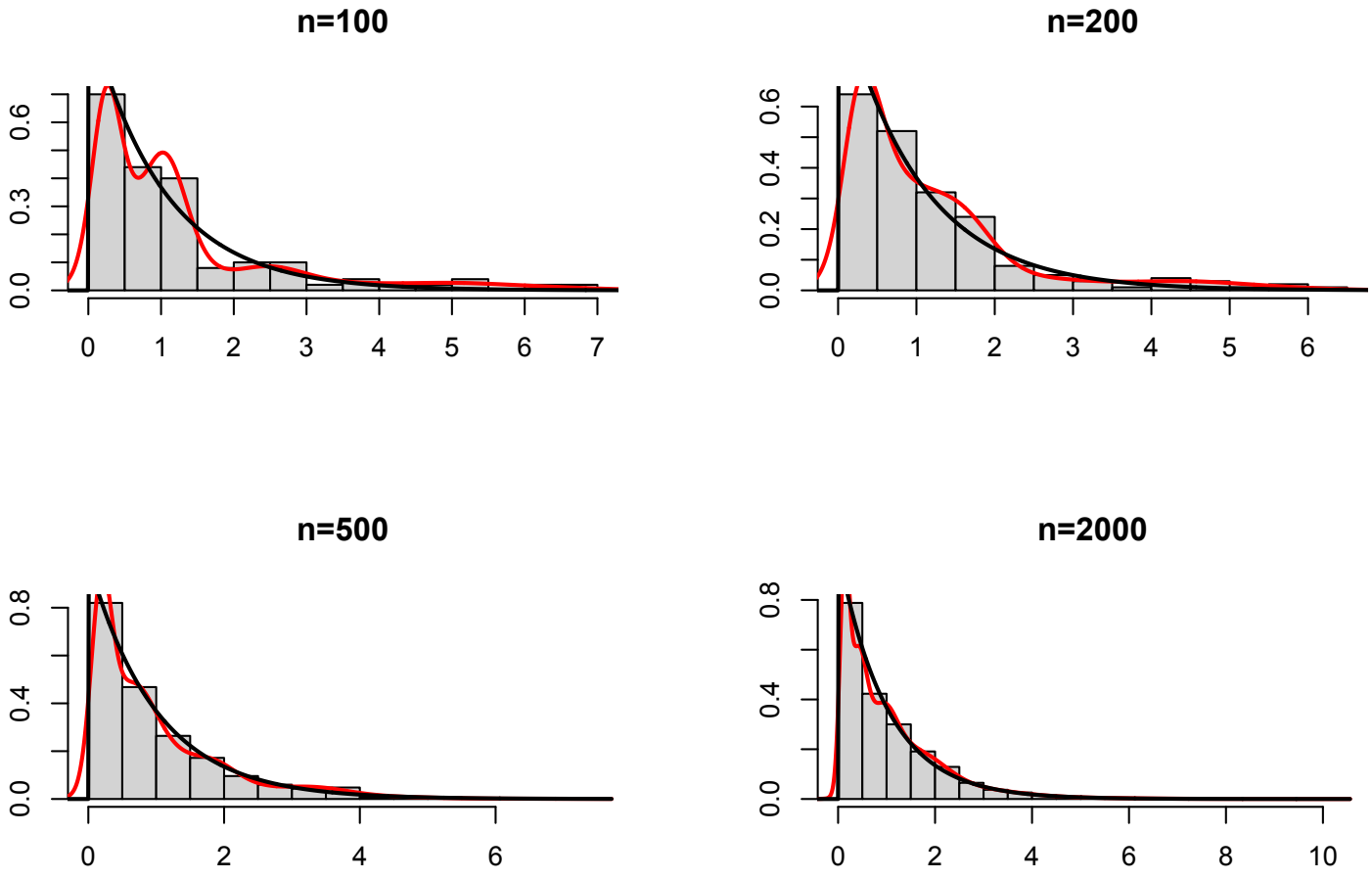
Conditions essentially:

- Should have true density in KL-support of the prior.
- All but a negligible amount of prior mass should be should be concentrated on a set whose "size", or "complexity" is not too large.

Concrete examples: many, but case-by-case analysis...

Consistency - example

Truth: exponential, prior: Dirichlet mixture of normals



[Ghosal, Ghosh, Ramamoorthi (1999)]

General rate of contraction results

Posterior contraction

Consistency: all posterior mass is ultimately contained in arbitrarily small neighborhoods of the true parameter.

Posterior contraction: how fast can we let these neighborhoods shrink, while still capturing all the posterior mass in the limit?

In other words: find the fastest converging $\varepsilon_n \downarrow 0$ such that asymptotically, all posterior mass in located in balls around θ_0 with radius of the order ε_n.

Definition: we say the posterior contracts around θ_0 at the rate ε_n if

$$\Pi(\theta \in \Theta : d(\theta, \theta_0) > M\varepsilon_n \,|\, X_1, \ldots, X_n) \xrightarrow{P_{\theta_0}} 0$$

for all $M > 0$ large enough.

Posterior contraction

Consistency: all posterior mass is ultimately contained in arbitrarily small neighborhoods of the true parameter.

Posterior contraction: how fast can we let these neighborhoods shrink, while still capturing all the posterior mass in the limit?

In other words: find the fastest converging $\varepsilon_n \downarrow 0$ such that asymptotically, all posterior mass in located in balls around θ_0 with radius of the order ε_n.

Definition: we say the posterior contracts around θ_0 at the rate ε_n if

$$\Pi(\theta \in \Theta : d(\theta, \theta_0) > M\varepsilon_n \,|\, X_1, \ldots, X_n) \overset{P_{\theta_0}}{\to} 0$$

for all $M > 0$ large enough.

Posterior contraction

Consistency: all posterior mass is ultimately contained in arbitrarily small neighborhoods of the true parameter.

Posterior contraction: how fast can we let these neighborhoods shrink, while still capturing all the posterior mass in the limit?

In other words: find the fastest converging $\varepsilon_n \downarrow 0$ such that asymptotically, all posterior mass in located in balls around θ_0 with radius of the order ε_n.

Definition: we say the posterior contracts around θ_0 at the rate ε_n if

$$\Pi(\theta \in \Theta : d(\theta, \theta_0) > M\varepsilon_n \,|\, X_1, \ldots, X_n) \xrightarrow{P_{\theta_0}} 0$$

for all $M > 0$ large enough.

Posterior contraction

Consistency: all posterior mass is ultimately contained in arbitrarily small neighborhoods of the true parameter.

Posterior contraction: how fast can we let these neighborhoods shrink, while still capturing all the posterior mass in the limit?

In other words: find the fastest converging $\varepsilon_n \downarrow 0$ such that asymptotically, all posterior mass in located in balls around θ_0 with radius of the order ε_n.

Definition: we say the posterior contracts around θ_0 at the rate ε_n if

$$\Pi(\theta \in \Theta : d(\theta, \theta_0) > M\varepsilon_n \,|\, X_1, \ldots, X_n) \overset{P_{\theta_0}}{\to} 0$$

for all $M > 0$ large enough.

Relation to convergence rates of estimators - 1

Proposition.

Suppose we have posterior contraction around θ_0 at the rate ε_n, relative to the metric d. Define the estimator $\hat{\theta}_n$ as the center of a ball of minimal radius that has posterior mass at least $1/2$. Then $d(\hat{\theta}_n, \theta_0) = O_{P_{\theta_0}}(\varepsilon_n)$, i.e. for all $\varepsilon > 0$, there exists $M > 0$ s.t.

$$\mathbb{P}_{\theta_0}(d(\hat{\theta}_n, \theta_0) > M\varepsilon_n) \leq \varepsilon.$$

Proof.

Let $B(\hat{\theta}_n, \hat{r})$ a ball of minimal radius that has posterior mass at least $1/2$. For every $\varepsilon > 0$, there exists an $M > 0$ s.t. $B(\theta_0, M\varepsilon_n)$ asymptotically contains posterior mass $1 - \varepsilon$ w.p. $1 - \varepsilon$. Hence on that event, $\hat{r} \leq M\varepsilon_n$. Moreover, the balls can not be disjoint. By the triangle inequality, it follows that, asymptotically w.p. $1 - \varepsilon$, $d(\hat{\theta}_n, \theta_0) \leq \hat{r} + M\varepsilon_n \leq 2M_n\varepsilon_n$. □

Relation to convergence rates of estimators - 2

So if we have posterior contraction at rate ε_n, then there exists an estimator with convergence rate ε_n.

Important consequence: posterior contraction rates are limited by the best possible convergence rates of frequentist estimators.

For many statistical problems, lower bounds (e.g. of minimax type) for convergence rates of estimators are known. Posteriors can never contract faster.

Under regularity conditions, best possible rate at which we can estimate a β-smooth function of d variables is typically $n^{-\beta/(d+2\beta)}$.

Q: which priors produce posteriors that contract at this optimal rate?

Relation to convergence rates of estimators - 2

So if we have posterior contraction at rate ε_n, then there exists an estimator with convergence rate ε_n.

Important consequence: posterior contraction rates are limited by the best possible convergence rates of frequentist estimators.

For many statistical problems, lower bounds (e.g. of minimax type) for convergence rates of estimators are known. Posteriors can never contract faster.

Under regularity conditions, best possible rate at which we can estimate a β-smooth function of d variables is typically $n^{-\beta/(d+2\beta)}$.

Q: which priors produce posteriors that contract at this optimal rate?

Recall: Extended Schwartz theorem

Observations: sample $X_1, \ldots, X_n$ from a density $p_0 \in \mathcal{P}$, for $\mathcal{P}$ the collection of densities on the unit interval. Prior: measure Π on $\mathcal{P}$

Theorem.
If there exist $\mathcal{P}_n \subset \mathcal{P}$ and $C < 6$, $D > 0$ such that for every $\varepsilon > 0$

$$\begin{aligned} \Pi(p : K(p_0, p) < \varepsilon) &> 0, \\ \Pi(\mathcal{P}_n^c) &\leq e^{-Dn}, \\ \log N(\varepsilon, \mathcal{P}_n, h) &\leq Cn\varepsilon^2. \end{aligned}$$

Then

$$\Pi(p : h(p, p_0) > \varepsilon \,|\, X_1, \ldots, X_n) \xrightarrow{P_0} 0$$

as $n \to \infty$.

General contraction rate theorem - 1

Distances:

$$h^2(p,q) = \int (\sqrt{p(x)} - \sqrt{q(x)})^2 \, dx \qquad \text{(Hellinger)}$$

$$K(p,q) = \int p(x) \log \frac{p(x)}{q(x)} \, dx \qquad \text{(Kulback-Leibler)}$$

$$V(p,q) = \int p(x) \Big(\log \frac{p(x)}{q(x)} \Big)^2 \, dx$$

KL-type ball:

$$B_n(p_0, \varepsilon) = \{p \in \mathcal{P} : K(p_0, p) \leq \varepsilon^2, V(p_0, p) \leq \varepsilon^2\}.$$

General contraction rate theorem - 2

Theorem.
If there exist $\mathcal{P}_n \subset \mathcal{P}$ and positive numbers ε_n such that $n\varepsilon_n^2 \to \infty$ and, for some $c > 0$,

$$\begin{aligned} \Pi(B_n(p_0, \varepsilon_n)) &\geq e^{-cn\varepsilon_n^2}, \\ \Pi(\mathcal{P}_n^c) &\leq e^{-(c+4)n\varepsilon_n^2}, \\ \log N(\varepsilon_n, \mathcal{P}_n, h) &\leq n{\varepsilon_n}^2, \end{aligned}$$

then for $M > 0$ large enough

$$\Pi(p : h(p, p_0) > M\varepsilon_n \,|\, X_1, \ldots, X_n) \xrightarrow{P_0} 0.$$

[Ghosal, Ghosh and Van der Vaart (2000)]

General contraction rate theorem - 3

- Proof: refinement of Schwartz.
- Theorem gives the "right" rates for many priors.
- Verifying the three conditions for a particular prior can be hard! [Shen, Tokdar, Ghosal (2013)]
- Versions of this theorem now exist for many nonparametric statistical settings: regression, classification, signal-in-white-noise, drift estimation for diffusions, Markov chains, time series, ...

The general theorem only becomes useful when combined with techniques tailored to specific (classes of) priors!

General contraction rate theorem - 3

- Proof: refinement of Schwartz.
- Theorem gives the "right" rates for many priors.
- Verifying the three conditions for a particular prior can be hard! [Shen, Tokdar, Ghosal (2013)]
- Versions of this theorem now exist for many nonparametric statistical settings: regression, classification, signal-in-white-noise, drift estimation for diffusions, Markov chains, time series, ...

The general theorem only becomes useful when combined with techniques tailored to specific (classes of) priors!

General contraction rate theorem - 3

- Proof: refinement of Schwartz.
- Theorem gives the "right" rates for many priors.
- Verifying the three conditions for a particular prior can be hard! [Shen, Tokdar, Ghosal (2013)]
- Versions of this theorem now exist for many nonparametric statistical settings: regression, classification, signal-in-white-noise, drift estimation for diffusions, Markov chains, time series, . . .

The general theorem only becomes useful when combined with techniques tailored to specific (classes of) priors!

General contraction rate theorem - 3

- Proof: refinement of Schwartz.
- Theorem gives the "right" rates for many priors.
- Verifying the three conditions for a particular prior can be hard! [Shen, Tokdar, Ghosal (2013)]
- Versions of this theorem now exist for many nonparametric statistical settings: regression, classification, signal-in-white-noise, drift estimation for diffusions, Markov chains, time series, ...

The general theorem only becomes useful when combined with techniques tailored to specific (classes of) priors!

General contraction rate theorem - 3

- Proof: refinement of Schwartz.
- Theorem gives the "right" rates for many priors.
- Verifying the three conditions for a particular prior can be hard! [Shen, Tokdar, Ghosal (2013)]
- Versions of this theorem now exist for many nonparametric statistical settings: regression, classification, signal-in-white-noise, drift estimation for diffusions, Markov chains, time series, ...

The general theorem only becomes useful when combined with techniques tailored to specific (classes of) priors!

Concluding remarks

Take home from Lecture II

- Frequentist notions like consistency and convergence rates can be useful to assess performance of nonparametric Bayes procedures.
- Contrary to the parametric case, performance depends crucially on the fine properties of the prior!
- We have general theorems that give conditions for consistency or contraction rates in terms of (i) the amount of mass that the prior gives to neighborhoods of the truth, (ii) the "size", or "complexity" of the sets where the prior puts all but a negligible amount of mass.

Q: how do we verify these conditions for interesting priors?

Take home from Lecture II

- Frequentist notions like consistency and convergence rates can be useful to assess performance of nonparametric Bayes procedures.
- Contrary to the parametric case, performance depends crucially on the fine properties of the prior!
- We have general theorems that give conditions for consistency or contraction rates in terms of (i) the amount of mass that the prior gives to neighborhoods of the truth, (ii) the "size", or "complexity" of the sets where the prior puts all but a negligible amount of mass.

Q: how do we verify these conditions for interesting priors?

Some references for Lecture II - 1

BvM and Doob's theorem:

- Van der Vaart, A. W. (2000). Asymptotic statistics. Cambridge university press.

Consistency:

- Doob, J. L. (1948). Application of the theory of martingales. Le calcul des probabilites et ses applications, 23-27.
- Schwartz, L. (1965). On Bayes procedures. Probability Theory and Related Fields, 4(1), 10-26.

Freedman/Diaconis:

- Diaconis, P., and Freedman, D. (1986). On the consistency of Bayes estimates. The Annals of Statistics, 1-26.
- Freedman, D. (1999). On the Bernstein-von Mises theorem with infinite-dimensional parameters. Annals of Statistics, 1119-1140.

Some references for Lecture II - 2

Extended consistency theorems:

- Barron, A., Schervish, M. J., and Wasserman, L. (1999). The consistency of posterior distributions in nonparametric problems. The Annals of Statistics, 27(2), 536-561.
- Ghosal, S., Ghosh, J. K., and Ramamoorthi, R. V. (1999). Posterior consistency of Dirichlet mixtures in density estimation. Ann. Statist, 27(1), 143-158.

Contraction rate theorems:

- Ghosal, S., Ghosh, J. K., and Van Der Vaart, A. W. (2000). Convergence rates of posterior distributions. Annals of Statistics, 28(2), 500-531.
- Shen, X., and Wasserman, L. (2001). Rates of convergence of posterior distributions. Annals of Statistics, 687-714.

Nonparametric Bayesian Methods - Lecture III

Harry van Zanten

Korteweg-de Vries Institute for Mathematics

University of Amsterdam

Finnish Summer school in Probability and Statistics
Lammi 30 May–3 June, 2016

Overview of Lecture III

- Recall: general rate of contraction results
- Gaussian process priors
  - reproducing kernel Hilbert space
  - small ball probabilities
  - concentration of measure
  - general theorem
- Rate results for concrete GP priors

General rate of contraction results

General contraction rate theorem - regression case - 1

Observations: pairs $(x_1, Y_1), \ldots, (x_n, Y_n)$, where

$$Y_i = f(x_i) + e_i,$$

with fixed $x_1, \ldots, x_n \in \mathcal{X}$ and e_i i.i.d. $N(0, \sigma^2)$.

Object of interest: regression function $f : \mathcal{X} \to \mathbb{R}$. Element of class $\mathcal{F}$ of all such functions.

Norm on regression functions:

$$\|f\|_n^2 = \frac{1}{n} \sum_{i=1}^{n} f^2(x_i).$$

General contraction rate theorem - regression case - 2

Theorem.
If there exist $\mathcal{F}_n \subset \mathcal{F}$ and positive numbers ε_n such that $n\varepsilon_n^2 \to \infty$ and, for some $c > 0$,

$$\begin{aligned} \Pi(f : \|f - f_0\|_n \leq \varepsilon_n)) &\geq e^{-cn\varepsilon_n^2}, \\ \Pi(\mathcal{F}_n^c) &\leq e^{-(c+8)n\varepsilon_n^2}, \\ \log N(\varepsilon_n, \mathcal{F}_n, \|\cdot\|_n) &\leq n{\varepsilon_n}^2, \end{aligned}$$

then for $M > 0$ large enough

$$\Pi(f : \|f - f_0\|_n > M\varepsilon_n \mid Y_1, \ldots, Y_n) \overset{P_{f_0}}{\to} 0.$$

Rates for Gaussian process priors

If $W = (W_x)_{x \in \mathcal{X}}$ is a stochastic process indexed by $\mathcal{X}$, we can use the law, or distribution of W as prior Π: for a set of functions $B \subset \mathcal{F}$,

$$\Pi(B) = \mathbb{P}(W \in B).$$

Convenient/popular choice: take W a Gaussian process (GP): Gaussian process prior (GP prior).

Examples:
If $\mathcal{X} = [0, 1]$: Brownian motion, (multiply) integrated Brownian motion, ...
If $\mathcal{X} = [0, 1]^d$: Matérn process, squared exponential GP, ...

Q: how do we verify the three conditions for GP's?

Gaussian process priors

What do we want to know about the GP?

For simplicity: take $\mathcal{X} = [0,1]$, bound $\|f\|_n$ by $\|f\|_\infty$,

$$\|f\|_\infty = \sup_{x \in [0,1]} |f(x)|.$$

For a GP $W = (W_t : t \in [0,1])$, want to

- lower bound probabilities of the form $\mathbb{P}(\|W - f_0\|_\infty < \varepsilon)$,
- find sets of functions $\mathcal{F}$ such that
  - $\mathbb{P}(W \in \mathcal{F}^c)$ is exponentially small
  - $N(\varepsilon, \mathcal{F}, \|\cdot\|_\infty)$ is "small"

Gaussian processes

$\rightarrow$ reproducing kernel Hilbert space

GP's - linear functionals

Let $W = (W_t : t \in [0,1])$ be a *centered, continuous* Gaussian process with covariance function $r(s,t) = \mathbb{E} W_s W_t$. Say defined on $(\Omega, \mathcal{F}, \mathbb{P})$.

(Can view W as a random element in $\mathbb{B} = C[0,1]$.)

Space $\mathcal{L}$ of linear functionals of W:

- $\mathcal{L}_0$: all finite linear combinations of the form $\sum c_i W_{t_i}$
- $\mathcal{L}$: closure of $\mathcal{L}_0$ in $L^2(\Omega, \mathcal{F}, \mathbb{P})$.

$\mathcal{L}$ is a separable Hilbert space with norm $\|L\| = \sqrt{\mathbb{E} L^2}$.

GP's - RKHS - 1

The reproducing kernel Hilbert space (RKHS) of W:

$$\mathbb{H} = \{t \mapsto \mathbb{E} W_t L : L \in \mathcal{L}\}.$$

Inner product:

$$\langle t \mapsto \mathbb{E} W_t L_1, t \mapsto \mathbb{E} W_t L_2 \rangle_{\mathbb{H}} = \mathbb{E} L_1 L_2.$$

We have isometry

$$\mathcal{L} \ni L \leftrightarrow (t \mapsto \mathbb{E} W_t L) \in \mathbb{H}.$$

In particular, $\mathbb{H}$ is a separable Hilbert space of functions.

GP's - RKHS - 1

The reproducing kernel Hilbert space (RKHS) of W:

$$\mathbb{H} = \{t \mapsto \mathbb{E} W_t L : L \in \mathcal{L}\}.$$

Inner product:

$$\langle t \mapsto \mathbb{E} W_t L_1, t \mapsto \mathbb{E} W_t L_2 \rangle_{\mathbb{H}} = \mathbb{E} L_1 L_2.$$

We have isometry

$$\mathcal{L} \ni L \leftrightarrow (t \mapsto \mathbb{E} W_t L) \in \mathbb{H}.$$

In particular, $\mathbb{H}$ is a separable Hilbert space of functions.

GP's - RKHS - 2

For fixed s, have that $t \mapsto r(s,t) = \mathbb{E}W_t W_s \in \mathbb{H}$ and for $h(t) = \mathbb{E}W_t L$, $L \in \mathcal{L}$,

$$\langle h, r(s,\cdot)\rangle_{\mathbb{H}} = \mathbb{E}LW_s = h(s).$$

This is the reproducing property.

We have $\mathbb{H} \subset \mathbb{B}$ and for $L \in \mathcal{L}$ and $h(t) = \mathbb{E}W_t L$,

$$\|h\|_\infty \leq \sup_{t\in[0,1]} \sqrt{\mathbb{E}W_t^2}\sqrt{\mathbb{E}L^2} = \sigma(W)\|h\|_{\mathbb{H}},$$

for $\sigma(W) = \sup_{t\in[0,1]} \sqrt{\mathbb{E}W_t^2}$.

Example: RKHS of Brownian motion

Let W be Brownian motion, with $r(s,t) = s \wedge t$. Then for all $s \in [0,1]$,

$$t \mapsto \int_0^t 1_{[0,s]}(u)\, du = t \wedge s = \mathbb{E} W_t W_s \in \mathbb{H}$$

and

$$\left\langle \int_0^{\cdot} 1_{[0,s_1]}(u)\, du, \int_0^{\cdot} 1_{[0,s_2]}(u)\, du \right\rangle_{\mathbb{H}} = s_1 \wedge s_2 = \left\langle 1_{[0,s_1]}, 1_{[0,s_2]} \right\rangle_{L^2}.$$

Hence

$$\mathbb{H} = \left\{ \int_0^{\cdot} f(u)\, du : f \in L^2 \right\},$$

$$\left\langle \int_0^{\cdot} f(u)\, du, \int_0^{\cdot} g(u)\, du \right\rangle_{\mathbb{H}} = \langle f, g \rangle_{L^2}.$$

(Cameron-Martin space)

Gaussian processes

$\rightarrow$ small ball probabilities

Important tool: Cameron-Martin formula

Let $U : \mathbb{H} \to \mathcal{L}$ be the isometry defined by $U(t \mapsto \mathbb{E}W_t L) = L$.

Let P^W be the law of W on $\mathbb{B} = C[0,1]$, i.e. $P^W(B) = \mathbb{P}(W \in B)$.

Theorem. [Cameron-Martin (1944)]
If $h \in \mathbb{H}$, then P^W and P^{W+h} are equivalent and

$$\frac{dP^{W+h}}{dP^W}(W) = e^{Uh - \frac{1}{2}\|h\|_{\mathbb{H}}^2}.$$

(If $h \notin \mathbb{H}$, then P^W and P^{W+h} are orthogonal.)

(For BM, compare with Girsanov)

Support of a GP

Support of W:

- smallest closed subset $\mathbb{B}_0 \subseteq \mathbb{B}$ such that $\mathbb{P}(W \in \mathbb{B}_0) = 1$.
- f_0 in support iff $\forall \varepsilon > 0$, $\mathbb{P}(\|W - f_0\|_\infty < \varepsilon) > 0$.

Theorem. [Kallianpur (1971)]

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

Proof.

- Elementary arguments: $\mathbb{P}(\|W\|_\infty < \varepsilon) > 0$ for all $\varepsilon > 0$.
- Cameron-Martin: $\mathbb{H} \subseteq \mathbb{B}_0$.
- Closing off: $\overline{\mathbb{H}} \subseteq \mathbb{B}_0$.
- Hahn-Banach: $\overline{\mathbb{H}} = \mathbb{B}_0$.

□

Example: support of BM is $\{f \in C[0,1] : f(0) = 0\}$.

Non-centered small ball probabilities - 1

For $h \in \mathbb{H}$, by Cameron-Martin,

$$\begin{aligned}\mathbb{P}(\|W-h\|_\infty < \varepsilon) &= \mathbb{E}\frac{dP^{W-h}}{dP^W}(W)1_{\|W\|_\infty<\varepsilon} \\ &= \mathbb{E}e^{-Uh-\frac{1}{2}\|h\|_{\mathbb{H}}^2}1_{\|W\|_\infty<\varepsilon}.\end{aligned}$$

Since $W \stackrel{\mathrm{d}}{=} -W$, also

$$\begin{aligned}\mathbb{P}(\|W-h\|_\infty < \varepsilon) &= \mathbb{E}\frac{dP^{W+h}}{dP^W}(W)1_{\|W\|_\infty<\varepsilon} \\ &= \mathbb{E}e^{Uh-\frac{1}{2}\|h\|_{\mathbb{H}}^2}1_{\|W\|_\infty<\varepsilon}.\end{aligned}$$

Together,

$$\begin{aligned}\mathbb{P}(\|W-h\|_\infty < \varepsilon) &= e^{-\frac{1}{2}\|h\|_{\mathbb{H}}^2}\mathbb{E}1_{\|W\|_\infty<\varepsilon}\frac{1}{2}(e^{Uh}+e^{-Uh}) \\ &\geq e^{-\frac{1}{2}\|h\|_{\mathbb{H}}^2}\mathbb{P}(\|W\|_\infty < \varepsilon).\end{aligned}$$

Non-centered small ball probabilities - 2

Concentration function: for $f_0 \in \mathbb{B}$ and $\varepsilon > 0$:

$$\varphi_{f_0}(\varepsilon) = \inf_{h \in \mathbb{H}: \|h - f_0\| < \varepsilon} \tfrac{1}{2}\|h\|_{\mathbb{H}}^2 - \log \mathbb{P}(\|W\| < \varepsilon).$$

Theorem.
for $f_0 \in \mathbb{B}$ and $\varepsilon > 0$

$$\varphi_{f_0}(\varepsilon) \leq -\log \mathbb{P}(\|W - f_0\| < \varepsilon) \leq \varphi_{f_0}(\varepsilon/2).$$

Centered small ball probabilities - 1

Let $h_1, \ldots, h_N$ be elements of $\mathbb{H}_1$ that are 2ε-separated in $\mathbb{B}$. Then

$$\begin{aligned}\sqrt{e} &\geq \sqrt{e} \sum_{j=1}^{N} \mathbb{P}(\|W - h_j\|_\infty < \varepsilon) \\ &\geq \sqrt{e} \sum_{j=1}^{N} e^{-\frac{1}{2}\|h_j\|_{\mathbb{H}}^2} \mathbb{P}(\|W\|_\infty < \varepsilon) \\ &\geq N\mathbb{P}(\|W\|_\infty < \varepsilon).\end{aligned}$$

Hence:

large metric entropy of $\mathbb{H}_1 \sim$ small probability $\mathbb{P}(\|W\|_\infty < \varepsilon)$.

Centered small ball probabilities - 2

More careful (much more) analysis gives (Kuelbs and Li (1993), Li and Linde (1999)):

$$\log N(\varepsilon, \mathbb{H}_1, \|\cdot\|_\infty) \asymp \varepsilon^{-\frac{2\alpha}{2+\alpha}} \iff -\log \mathbb{P}(\|W\|_\infty < \varepsilon) \asymp \varepsilon^{-\alpha},$$

$$\log N(\varepsilon, \mathbb{H}_1, \|\cdot\|_\infty) \asymp \log^\gamma \frac{1}{\varepsilon} \iff -\log \mathbb{P}(\|W\|_\infty < \varepsilon) \asymp \log^\gamma \frac{1}{\varepsilon}.$$

Gaussian processes

$\longrightarrow$ concentration of measure

Concentration of measure - 1

Finite-dimensional situation:

Let $X \sim N_d(0, \Sigma)$, Σ invertible. Then $\mathbb{H} = \mathbb{R}^d$ and

$$\langle x, y \rangle_{\mathbb{H}} = x^T \Sigma^{-1} y.$$

Hence, the balls in $\mathbb{H}$ (ellipsoids in the ordinary metric) are precisely the level sets of the density of X.

In other words: the RKHS unit ball describes the "geometry" of the support of the distribution of X.

Concentration of measure - 2

$\mathbb{B}_1$: unit ball of $\mathbb{B}$, $\mathbb{H}_1$: unit ball of $\mathbb{H}$.

Theorem. [Borell (1975), Sudakov-Tsirelson (1974)]
For $\varepsilon > 0$ and $M \geq 0$,

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

Let $M(W)$ be the median of $\|W\|_\infty$. Take $\varepsilon = M(W)$ and $M = x/\sigma(W)$, use $\mathbb{H}_1 \subset \sigma(W)\mathbb{B}_1$.

Corollary.
For $x > 0$,

$$\mathbb{P}(\|W\|_\infty - M(W) > x) \leq 1 - \Phi(x/\sigma(W)).$$

Concentration of measure - 2

$\mathbb{B}_1$: unit ball of $\mathbb{B}$, $\mathbb{H}_1$: unit ball of $\mathbb{H}$.

Theorem. [Borell (1975), Sudakov-Tsirelson (1974)]

For $\varepsilon > 0$ and $M \geq 0$,

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

Let $M(W)$ be the median of $\|W\|_\infty$. Take $\varepsilon = M(W)$ and $M = x/\sigma(W)$, use $\mathbb{H}_1 \subset \sigma(W)\mathbb{B}_1$.

Corollary.

For $x > 0$,

$$\mathbb{P}(\|W\|_\infty - M(W) > x) \leq 1 - \Phi(x/\sigma(W)).$$

Gaussian processes

$\rightarrow$ general theorem

Recall what we want

For a continuous GP $W = (W_t)_{t\in[0,1]}$ and function f_0, want to find smallest possible $\varepsilon_n \downarrow 0$ for which there exist $\mathcal{F}_n \subset C[0,1]$ and $c > 0$ such that

1. $$\mathbb{P}(\|W - f_0\|_\infty < \varepsilon_n) \geq e^{-cn\varepsilon_n^2},$$
2. $$\mathbb{P}(W \notin \mathcal{F}_n) \leq e^{-(c+8)n\varepsilon_n^2},$$
3. $$\log N(\varepsilon_n, \mathcal{F}_n, \|\cdot\|_\infty) \leq n\varepsilon_n^2.$$

Putting things together - 1

Prior mass condition 1. is fulfilled if $\varphi_{f_0}(\varepsilon_n) \leq n\varepsilon_n^2$, where

$$\varphi_{f_0}(\varepsilon) = \inf_{h \in \mathbb{H}: \|h - f_0\|_\infty < \varepsilon} \frac{1}{2}\|h\|_{\mathbb{H}}^2 - \log \mathbb{P}(\|W\|_\infty < \varepsilon).$$

Borell-Sudakov suggests to take sieve $\mathcal{F}_n$ of the form $\mathcal{F}_n = M_n \mathbb{H}_1 + \varepsilon_n \mathbb{B}_1$. Have

$$\mathbb{P}(W \notin \mathcal{F}_n) \leq 1 - \Phi(\Phi^{-1}(\mathbb{P}(\|W\|_\infty \leq \varepsilon_n)) + M_n).$$

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

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

Putting things together - 1

Prior mass condition 1. is fulfilled if $\varphi_{f_0}(\varepsilon_n) \leq n\varepsilon_n^2$, where

$$\varphi_{f_0}(\varepsilon) = \inf_{h \in \mathbb{H}: \|h - f_0\|_\infty < \varepsilon} \tfrac{1}{2}\|h\|_{\mathbb{H}}^2 - \log \mathbb{P}(\|W\|_\infty < \varepsilon).$$

Borell-Sudakov suggests to take sieve $\mathcal{F}_n$ of the form $\mathcal{F}_n = M_n \mathbb{H}_1 + \varepsilon_n \mathbb{B}_1$. Have

$$\mathbb{P}(W \notin \mathcal{F}_n) \leq 1 - \Phi(\Phi^{-1}(\mathbb{P}(\|W\|_\infty \leq \varepsilon_n)) + M_n).$$

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

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

Putting things together - 2

so

$$\mathbb{P}(W \notin \mathcal{F}_n) \leq 1 - \Phi(\Phi^{-1}(e^{-n\varepsilon_n^2})) + M_n).$$

Now take $M_n = -2\Phi^{-1}(e^{-Cn\varepsilon_n^2})$ for some $C > 1$. Then

$$\begin{aligned}\mathbb{P}(W \notin \mathcal{F}_n) &\leq 1 - \Phi(\Phi^{-1}(e^{-n\varepsilon_n^2})) + M_n) \\ &\leq 1 - \Phi(-\Phi^{-1}(e^{-Cn\varepsilon_n^2})) \\ &= e^{-Cn\varepsilon_n^2}.\end{aligned}$$

So for this sieve $\mathcal{F}_n$, the remaining mass condition 2. holds, provided we choose C large enough.

How about the entropy condition?

Putting things together - 2

so

$$\mathbb{P}(W \notin \mathcal{F}_n) \leq 1 - \Phi(\Phi^{-1}(e^{-n\varepsilon_n^2})) + M_n).$$

Now take $M_n = -2\Phi^{-1}(e^{-Cn\varepsilon_n^2})$ for some $C > 1$. Then

$$\begin{aligned}\mathbb{P}(W \notin \mathcal{F}_n) &\leq 1 - \Phi(\Phi^{-1}(e^{-n\varepsilon_n^2})) + M_n) \\ &\leq 1 - \Phi(-\Phi^{-1}(e^{-Cn\varepsilon_n^2})) \\ &= e^{-Cn\varepsilon_n^2}.\end{aligned}$$

So for this sieve $\mathcal{F}_n$, the remaining mass condition 2. holds, provided we choose C large enough.

How about the entropy condition?

Putting things together - 3

The condition $\varphi(\varepsilon_n) \leq n\varepsilon_n^2$ implies a lower bound for the centered small ball probability $\mathbb{P}(\|W\|_\infty < \varepsilon_n)$.

This gives an upper bound on $N(2\varepsilon_n, \mathbb{H}_1, \|\cdot\|_\infty)$, hence also on $N(2\varepsilon_n, M_n\mathbb{H}_1, \|\cdot\|_\infty)$!

Since $N(3\varepsilon_n, M_n\mathbb{H}_1 + \varepsilon_n\mathbb{B}_1, \|\cdot\|_\infty) \leq N(2\varepsilon_n, M_n\mathbb{H}_1, \|\cdot\|_\infty)$, we obtain a bound for $N(3\varepsilon_n, \mathcal{F}_n, \|\cdot\|_\infty)$.

It turns out that with our choice of M_n, we get

$$\log N(3\varepsilon_n, \mathcal{F}_n, \|\cdot\|_\infty) \leq 6Cn\varepsilon_n^2.$$

This takes care of the entropy condition 3!

General theorem for Gaussian process priors

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. [Van der Vaart and vZ. (2008)]

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.

$$\mathbb{P}(\|W - f_0\|_\infty < 2\varepsilon_n) \geq e^{-n\varepsilon_n^2},$$

$$\mathbb{P}(W \notin \mathcal{F}_n) \leq e^{-Cn\varepsilon_n^2}$$

$$\log N(3\varepsilon_n, \mathcal{F}_n, \|\cdot\|_\infty) \leq 6Cn\varepsilon_n^2.$$

General theorem for Gaussian process priors

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. [Van der Vaart and vZ. (2008)]
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.

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

Rate results for concrete GP priors

Splitting up the problem

To get a rate ε_n solving $\varphi_{f_0}(\varepsilon_n) \leq n\varepsilon_n^2$ we need two things:

1. $$-\log \mathbb{P}(\|W\|_\infty < \varepsilon_n) \leq n\varepsilon_n^2,$$

2. $$\inf_{h \in \mathbb{H}: \|h - f_0\|_\infty < \varepsilon_n} \|h\|_{\mathbb{H}}^2 \leq n\varepsilon_n^2.$$

Problem 1. only depends on the GP. Widely studied in the "small deviations" or "small balls" community (Lifshits (2015): 329 refs). Result only depends on the prior.

For problem 2. we have to study the approximation of the function f_0 by elements of the RKHS of the GP. Result depends on the relation between f_0 and the prior.

Splitting up the problem

To get a rate ε_n solving $\varphi_{f_0}(\varepsilon_n) \leq n\varepsilon_n^2$ we need two things:

1.
$$-\log \mathbb{P}(\|W\|_\infty < \varepsilon_n) \leq n\varepsilon_n^2,$$

2.
$$\inf_{h \in \mathbb{H}: \|h - f_0\|_\infty < \varepsilon_n} \|h\|_{\mathbb{H}}^2 \leq n\varepsilon_n^2.$$

Problem 1. only depends on the GP. Widely studied in the "small deviations" or "small balls" community (Lifshits (2015): 329 refs). Result only depends on the prior.

For problem 2. we have to study the approximation of the function f_0 by elements of the RKHS of the GP. Result depends on the relation between f_0 and the prior.

Example: Brownian motion

Brownian motion W:

$$\mathbb{P}(\|W - f_0\|_\infty < \varepsilon) \leq \mathbb{P}(\|W\|_\infty < \varepsilon) \asymp e^{-c(1/\varepsilon)^2}.$$

Can only have $\varphi_{f_0}(\varepsilon_n) \leq n\varepsilon_n^2$ for $\varepsilon_n \geq n^{-1/4}$.

Hence, can not get faster rate than $\varepsilon_n \asymp n^{-1/4}$ with BM prior. (See Castillo (2008)).

Question: under which conditions on f_0 do we achieve the rate $n^{-1/4}$?

Example: Brownian motion

Brownian motion W:

$$\mathbb{P}(\|W - f_0\|_\infty < \varepsilon) \leq \mathbb{P}(\|W\|_\infty < \varepsilon) \asymp e^{-c(1/\varepsilon)^2}.$$

Can only have $\varphi_{f_0}(\varepsilon_n) \leq n\varepsilon_n^2$ for $\varepsilon_n \geq n^{-1/4}$.

Hence, can not get faster rate than $\varepsilon_n \asymp n^{-1/4}$ with BM prior. (See Castillo (2008)).

Question: under which conditions on f_0 do we achieve the rate $n^{-1/4}$?

Example: Brownian motion

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

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

Proof.
Approximate f_0 by convolutions. □

Hence for $f_0 \in C^\beta[0,1]$ the prior mass condition $\varphi_{f_0}(\varepsilon_n) \leq n\varepsilon_n^2$ holds for

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

Example: Brownian motion

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

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

Proof.
Approximate f_0 by convolutions. □

Hence for $f_0 \in C^\beta[0,1]$ the prior mass condition $\varphi_{f_0}(\varepsilon_n) \leq n\varepsilon_n^2$ holds for

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

Univariate nonparametric regression with a BM prior - 1

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 Π of Brownian motion (with standard normal initial distribution).

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

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

Univariate nonparametric regression with a BM prior - 2

Note: rate equals $n^{-\beta/(1+2\beta)}$ only for $\beta = 1/2$.

In other words:

The BM prior is rate-optimal if the smoothness of the true regression function equals the smoothness of the BM paths.

Suppose that $f_0 \in C^\beta[0,1]$ for $\beta > 0$.

Which GP prior leads to the optimal rate $n^{-\beta/(1+2\beta)}$?

Univariate nonparametric regression with a BM prior - 2

Note: rate equals $n^{-\beta/(1+2\beta)}$ only for $\beta = 1/2$.

In other words:

The BM prior is rate-optimal if the smoothness of the true regression function equals the smoothness of the BM paths.

Suppose that $f_0 \in C^{\beta}[0,1]$ for $\beta > 0$.

Which GP prior leads to the optimal rate $n^{-\beta/(1+2\beta)}$?

Priors for other smoothness levels - 1

Candidate: Riemann-Liouville process

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

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

Idea: should be good prior model for β-smooth functions.

Priors for other smoothness levels - 2

Known results for the RL-process:

Li and Linde (1998):

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

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

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

Priors for other smoothness levels - 3

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

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

Theorem.
The support of the process W is $C[0,1]$. For $f_0 \in C^\beta[0,1]$ we have $\varphi_{f_0}(\varepsilon) = O(\varepsilon^{-1/\beta})$ as $\varepsilon \to 0$.

Univariate nonparametric regression with a RL prior

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 Π of a modified RL-process with parameter $\beta > 0$.

Theorem.
Suppose $f_0 \in C^\beta[0,1]$ for $\beta > 0$. Then the posterior contracts around f_0 at the rate $\varepsilon_n \asymp n^{-\beta/(1+2\beta)}$.

Possible extensions

Other Gaussian priors that can be handled:

- (multiply integrated) fractional BM
- series priors
- Matérn process
- squared exponential process
- ...

Can also handle other statistical settings: density estimation, extracting a signal in white noise, nonparametric classification, drift estimation for diffusions, ...

Concluding remarks

Take home from Lecture III

- Gaussian process theory provides powerful tools for studying asymptotic behaviour of procedures using Gaussian priors.
- Have a powerful general theorem for GP's that matches with general contraction rate theorems.
- For many important nonparametric problems, rate-optimal GP priors can be exhibited.
- Performance depends very much on the fine properties of the GP that is used.
- To get rate-optimal performance, a Gaussian prior has to be carefully tuned to the true parameter to avoid over- or undersmoothing.

Q: How can we get optimal rates without knowledge of the true regularity?

Take home from Lecture III

- Gaussian process theory provides powerful tools for studying asymptotic behaviour of procedures using Gaussian priors.
- Have a powerful general theorem for GP's that matches with general contraction rate theorems.
- For many important nonparametric problems, rate-optimal GP priors can be exhibited.
- Performance depends very much on the fine properties of the GP that is used.
- To get rate-optimal performance, a Gaussian prior has to be carefully tuned to the true parameter to avoid over- or undersmoothing.

Q: How can we get optimal rates without knowledge of the true regularity?

Main references for Lecture III

Review of Gaussian process theory useful for Bayesian nonparametrics:

- van der Vaart, A.W., van Zanten, J.H. (2008). Reproducing kernel Hilbert spaces of Gaussian priors. In: Pushing the limits of contemporary statistics: contributions in honor of Jayanta K. Ghosh, Inst. Math. Stat. Collect., 3, pp. 200–222, Beachwood, OH.

General theorem for GP priors, and RL and other examples:

- van der Vaart, A.W., van Zanten, J.H. (2008). Rates of contraction of posterior distributions based on Gaussian process priors. Ann. Statist. 36, no. 3, 1435–1463.

Nonparametric Bayesian Methods - Lecture IV

Harry van Zanten

Korteweg-de Vries Institute for Mathematics

University of Amsterdam

Finnish Summer school in Probability and Statistics
Lammi 30 May–3 June, 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,

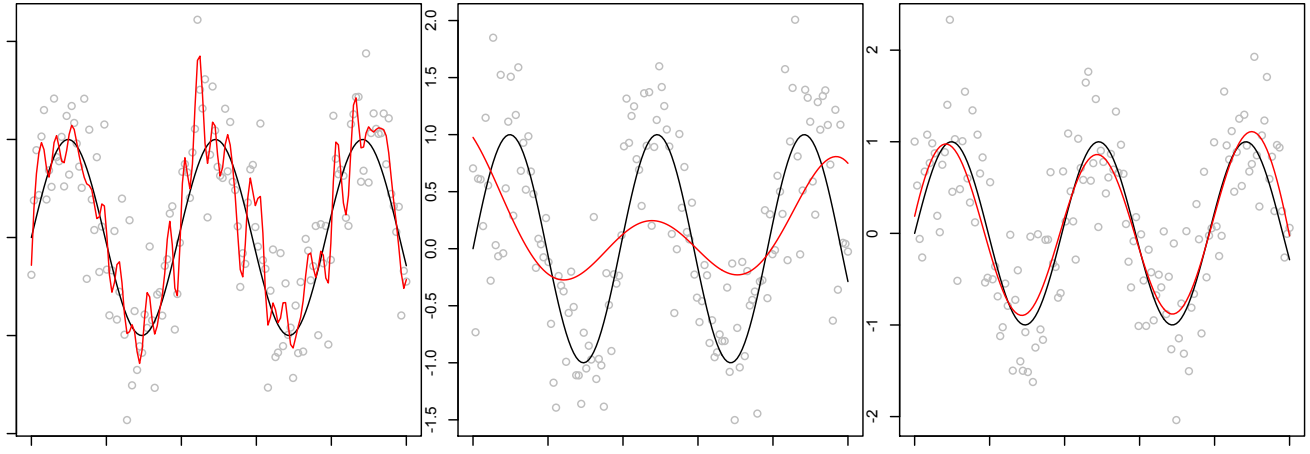
$$\varphi_{f_0}(\varepsilon) = \inf_{h\in\mathbb{H}:\|h-f_0\|_\infty<\varepsilon} \|h\|_{\mathbb{H}}^2 - \log \mathbb{P}(\|W\|_\infty < \varepsilon).$$

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:

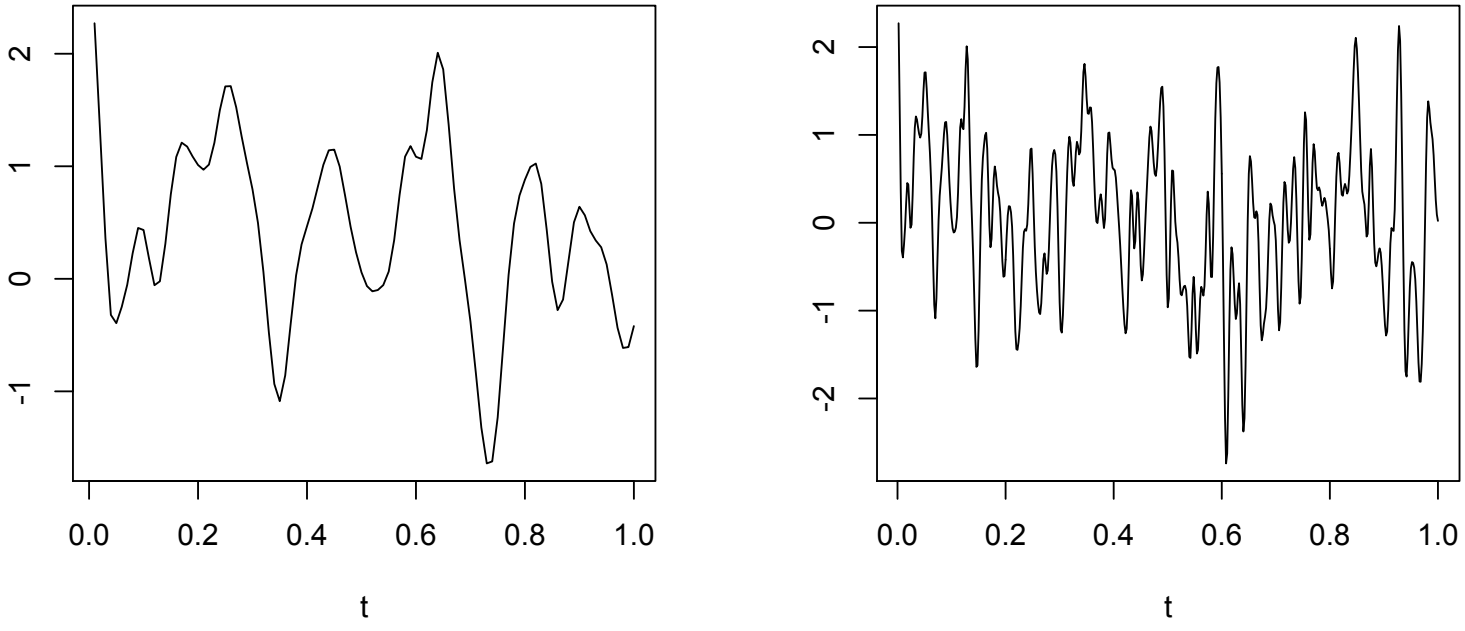


Figure: $\ell = 1$ versus $\ell = 0.2$

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

...