Nonparametric Bayesian Methods - Lecture I

Harry van Zanten

Korteweg-de Vries Institute for Mathematics

University of Amsterdam

CRiSM Masterclass, April 4-6, 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?

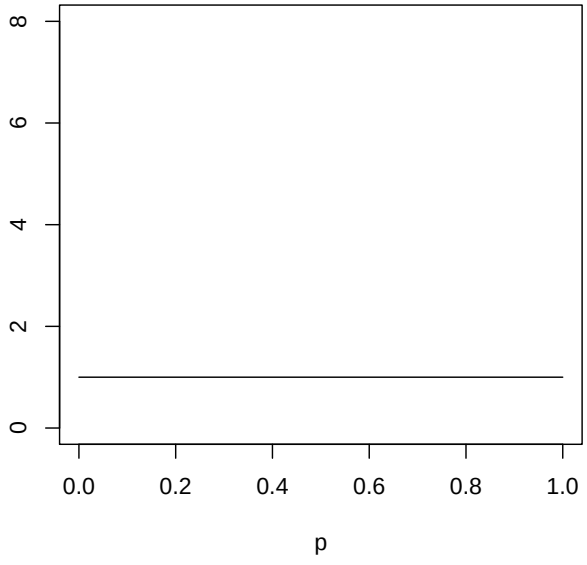
Here we have an observation (the number 42) from a binomial distribution with parameters 50 and p and want to estimate p.

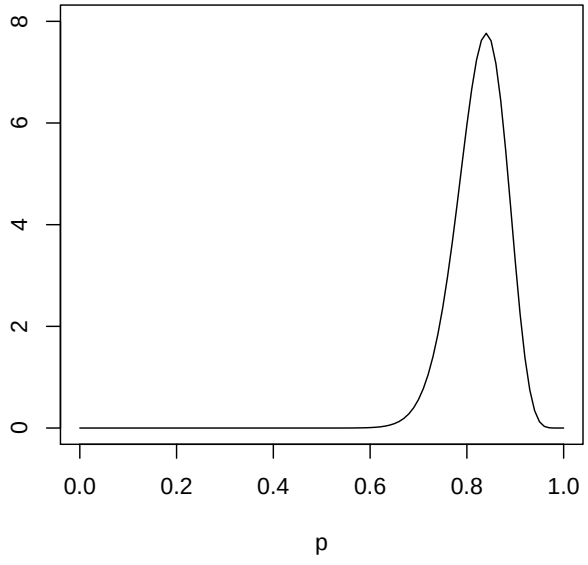
Standard frequentist solution: take the estimate $42/50 = 0.84$.

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

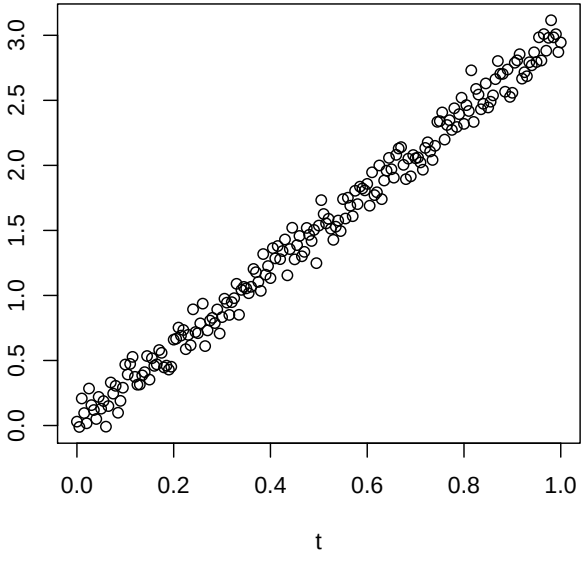
Bayesian nonparametrics

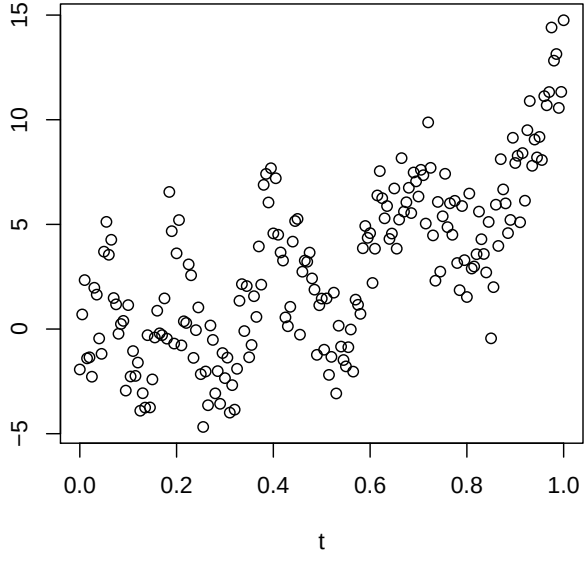
Bayesian nonparametrics

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

Illustration 1: parametric vs. nonparametric regression

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

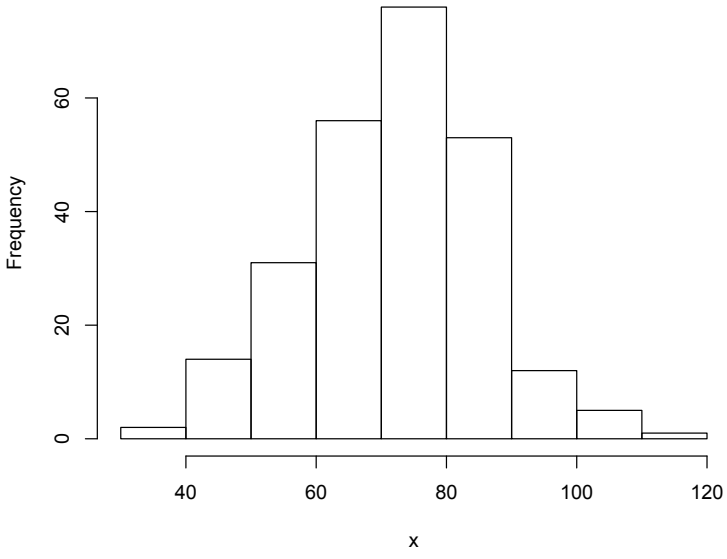


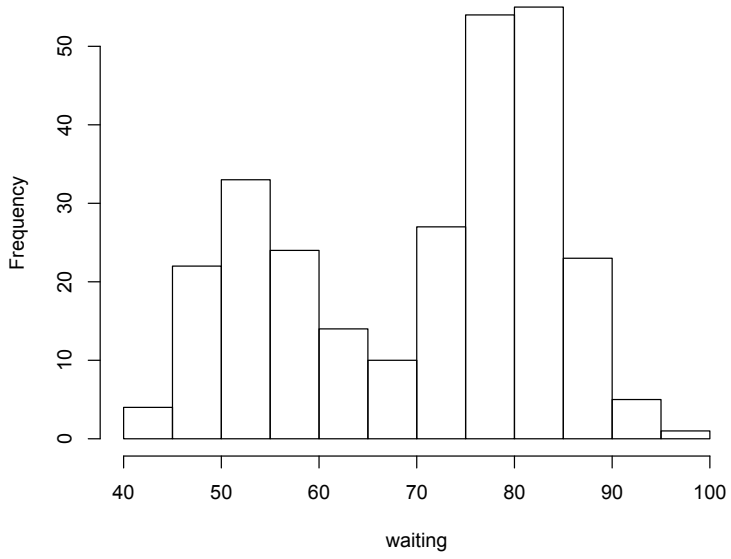


Bayesian nonparametrics

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!

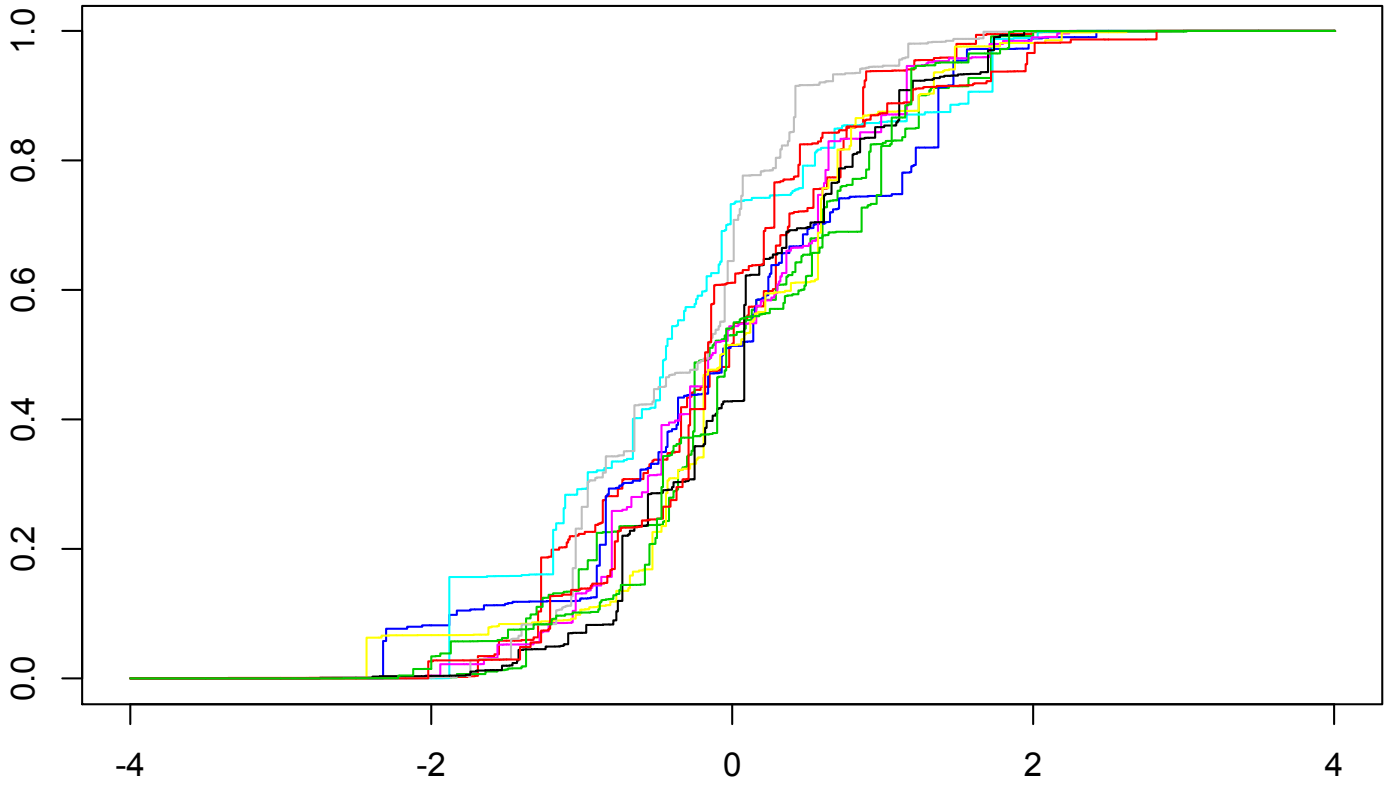
Theorem.
For any finite measure α on $\mathbb{R}$, the Dirichlet process with parameter α exists.

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

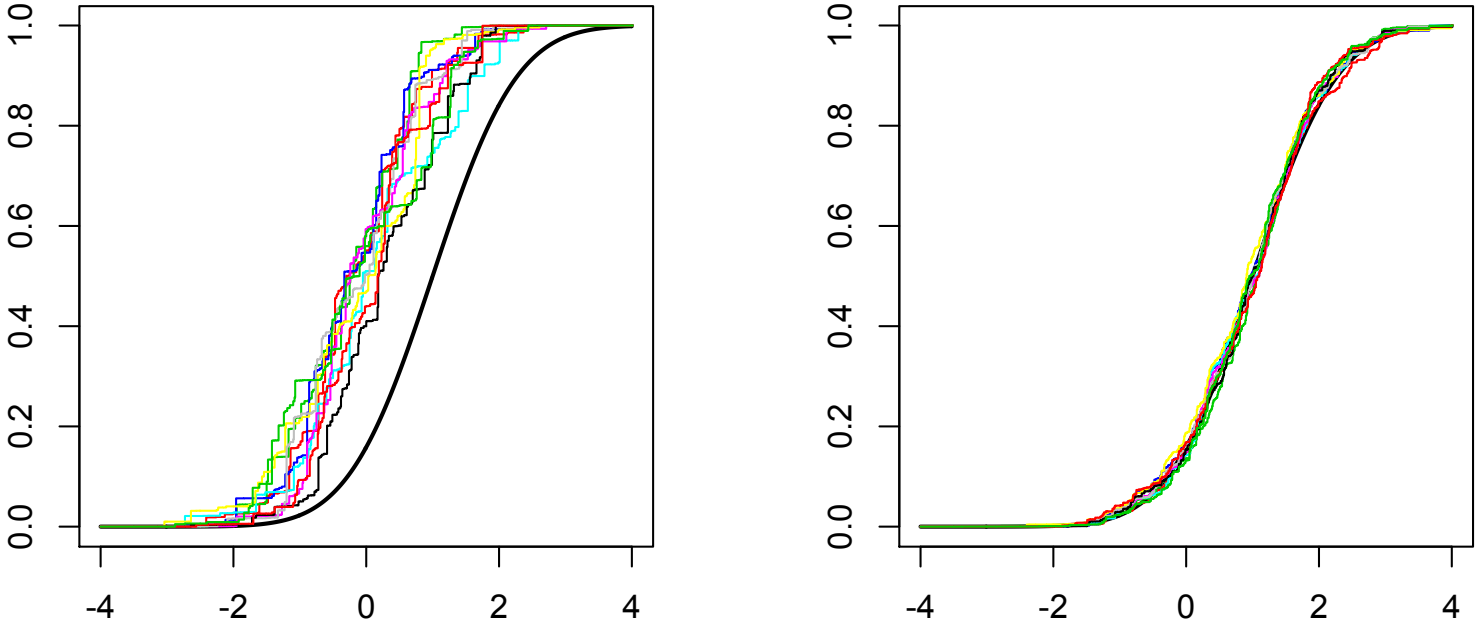
Draws from the DP are discrete measures on $\mathbb{R}$:

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 $t_1, \ldots, t_n \in T$,

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

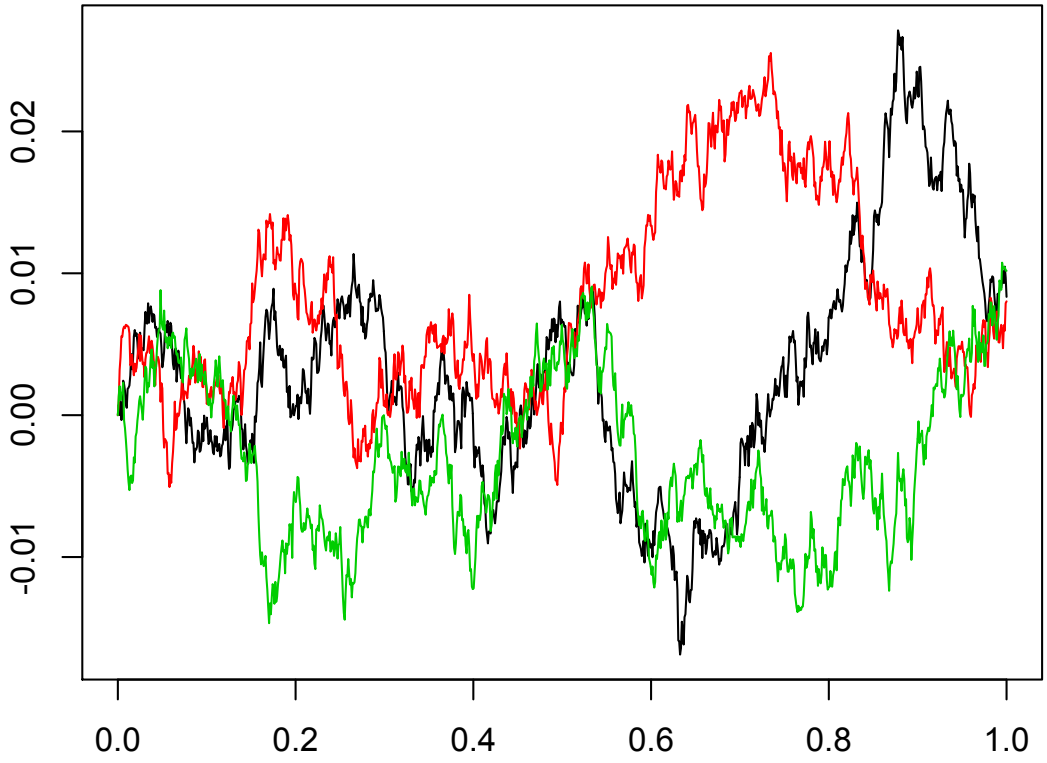
hence r is a positive definite, symmetric function on $T \times T$.

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

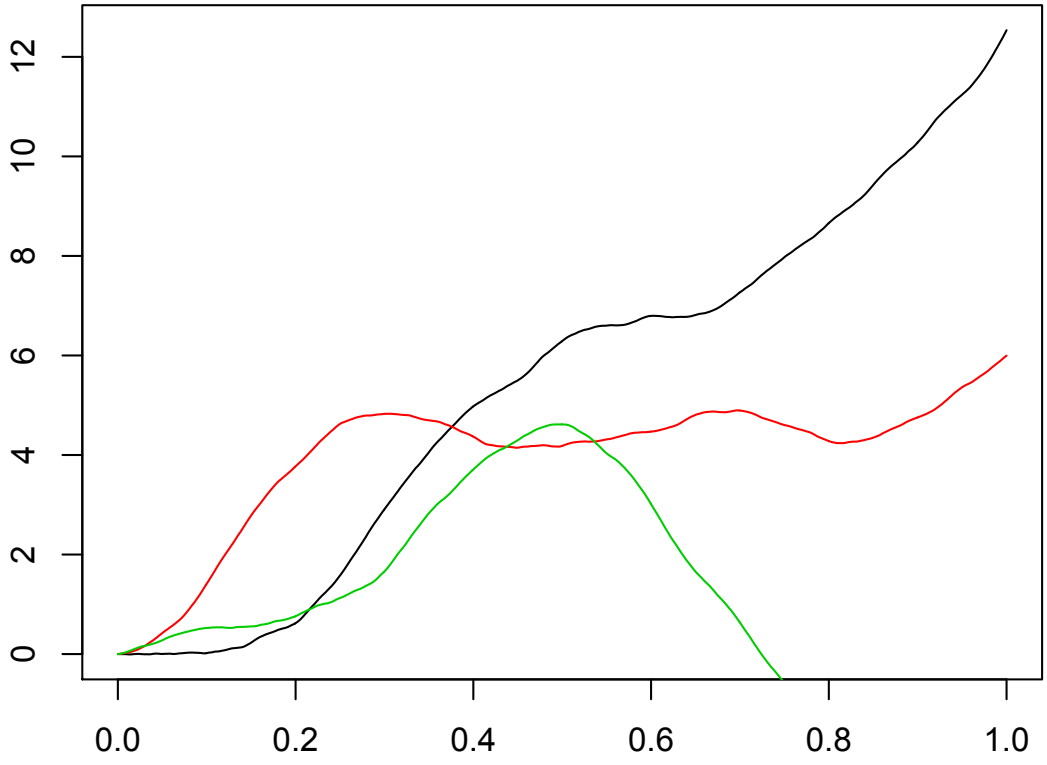
Brownian motion: $m(t) = 0$, $r(s,t) = s \wedge t$.



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.

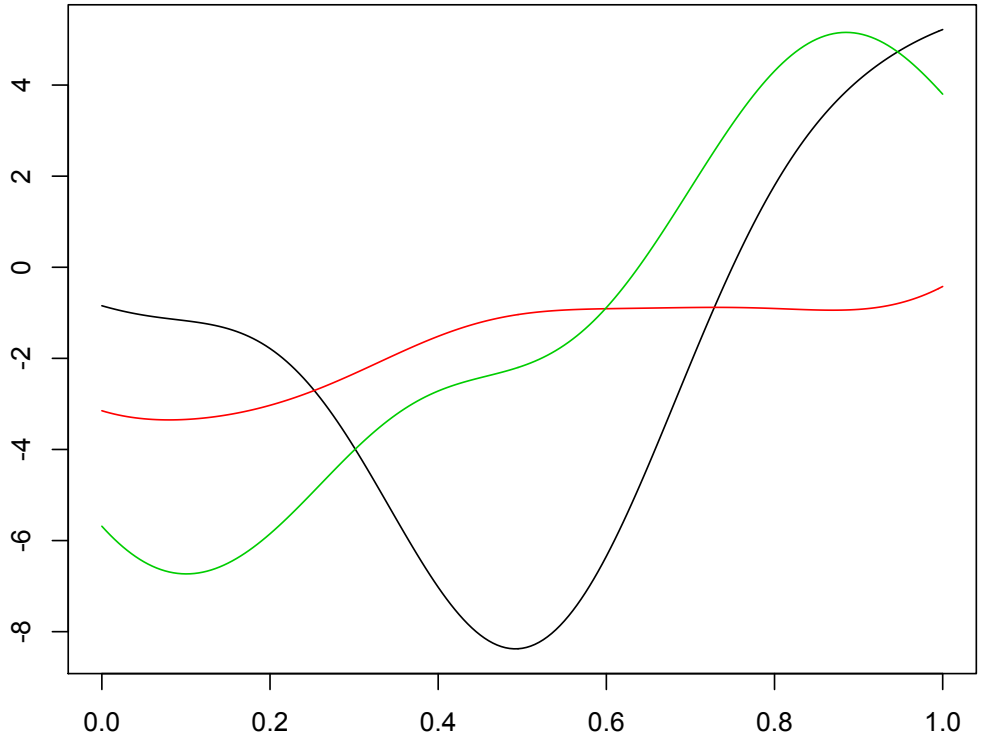
By Bochner's theorem:

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

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.

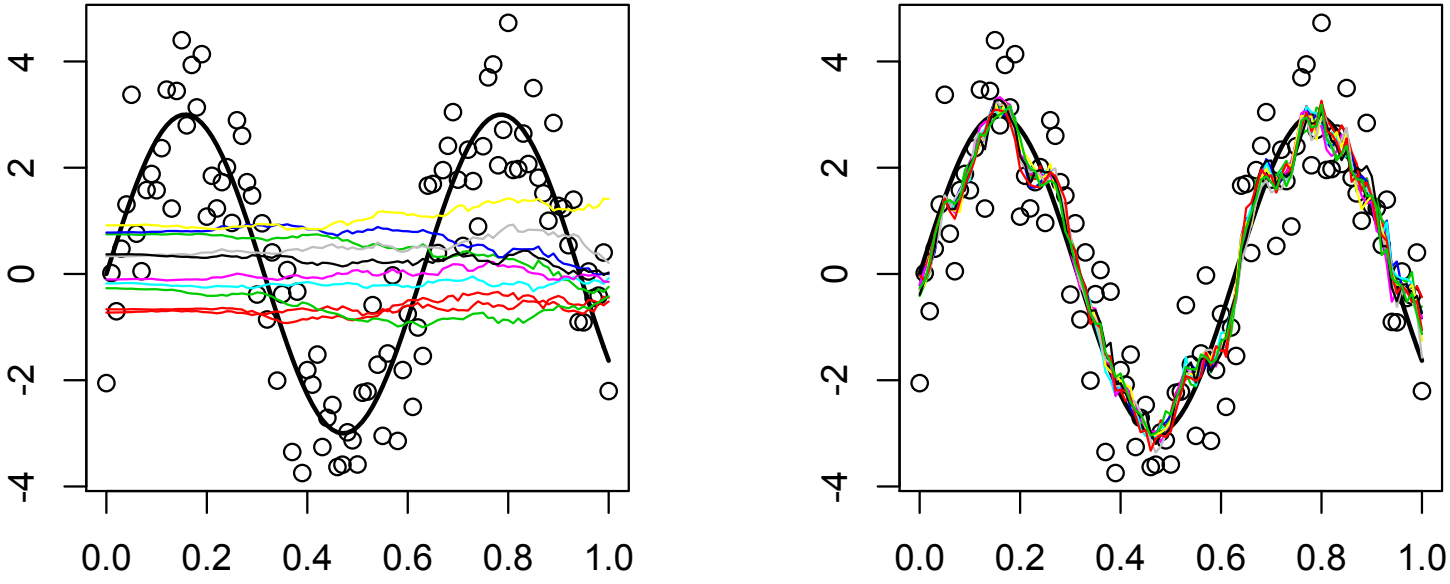
Posterior: this prior is conjugate for this model:

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

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

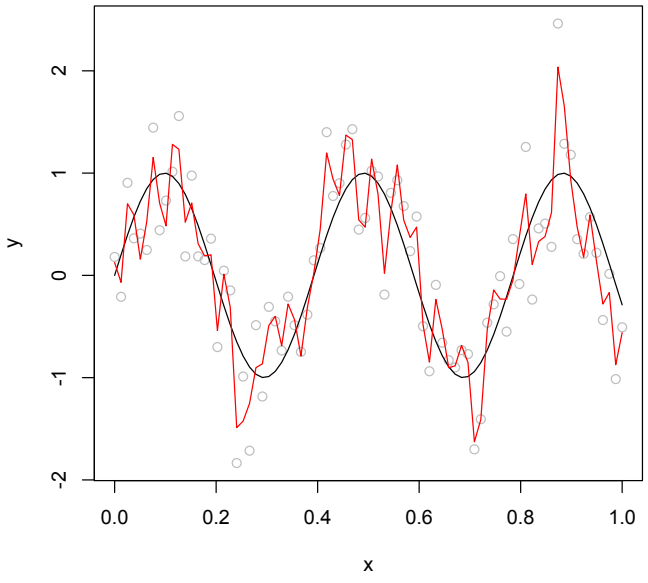
Observation about GP's:

- 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?

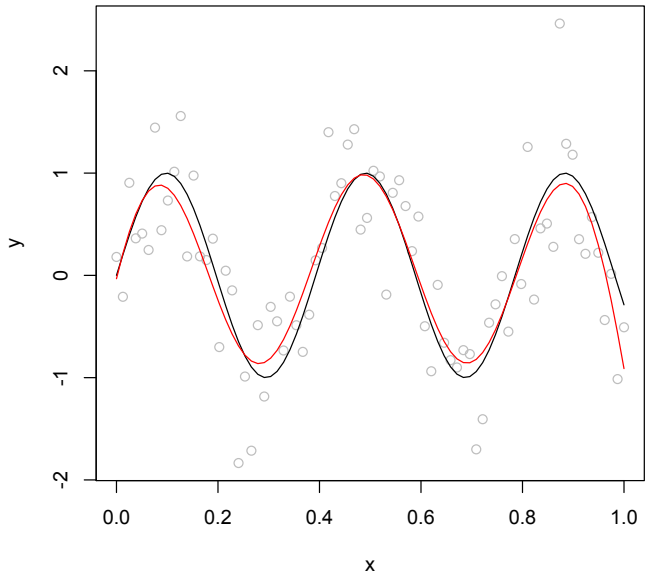
CGP's - 2

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:

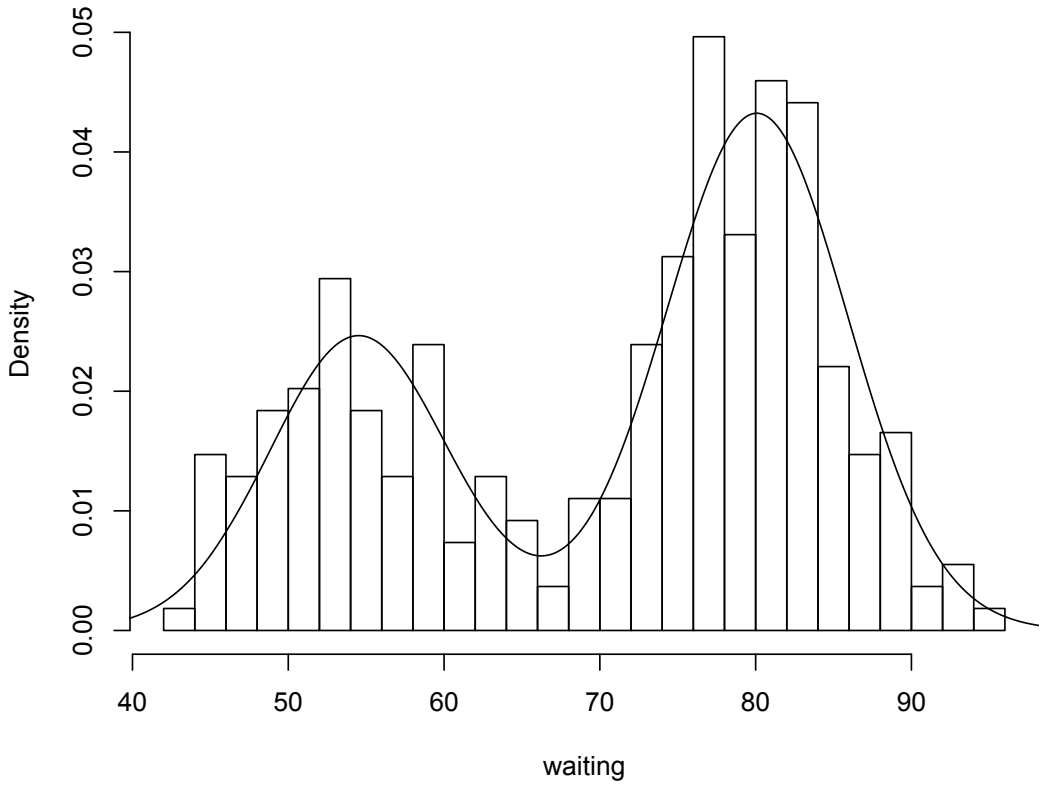
$$G \sim DP(G_0) \qquad (G_0 \text{ often } N \times IW)$$
$$p \,|\, G \sim p_G$$

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

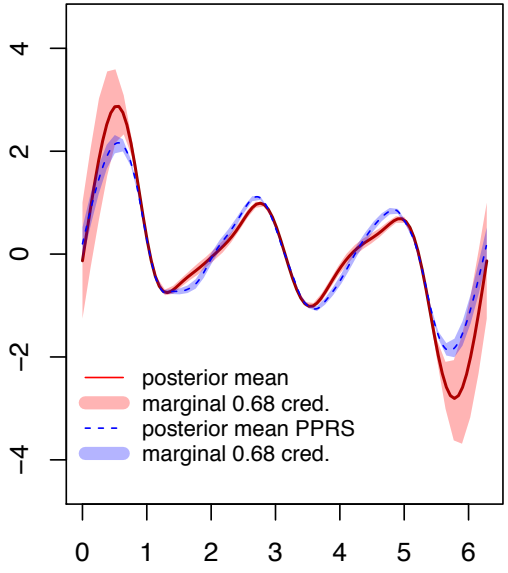
Observation model: $dX_t = b(X_t)\,dt + dW_t$. Goal: estimate b.

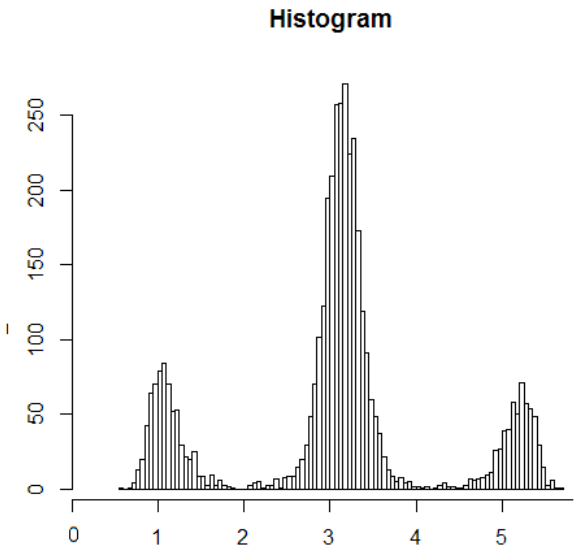
Prior:

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

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

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



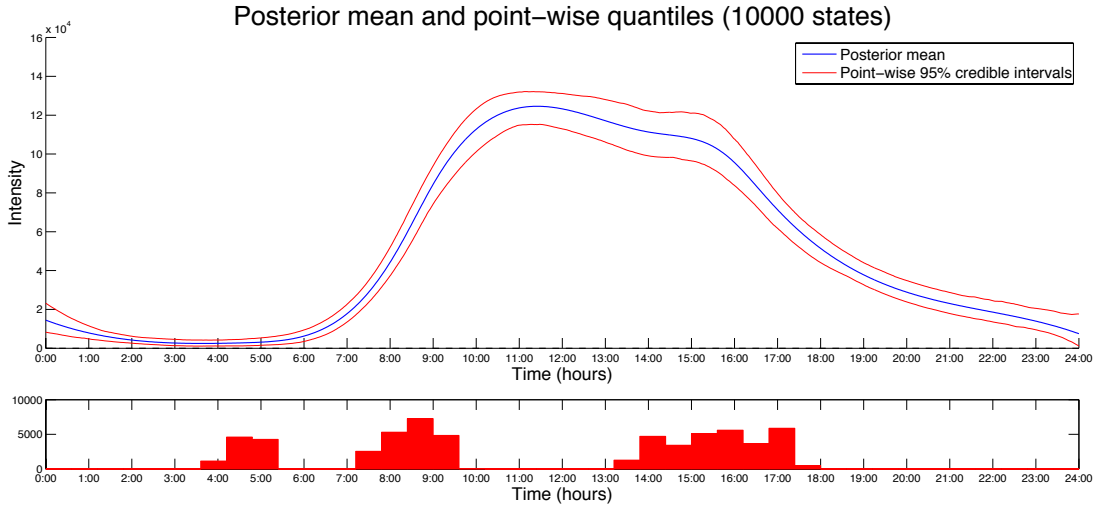


[Van der Meulen, Schauer, vZ. (2014)]

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

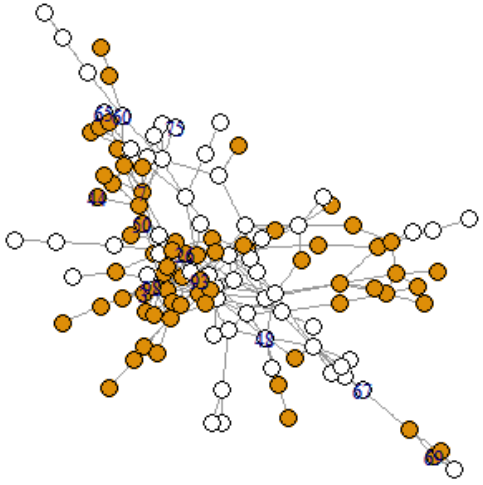


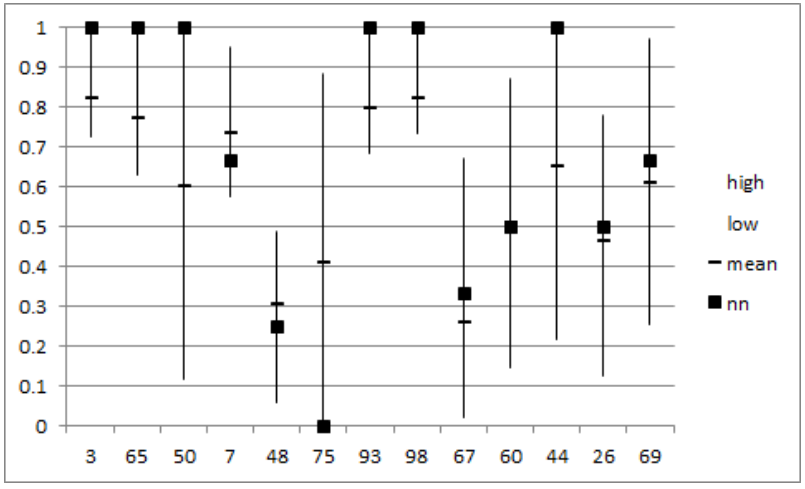
[Belitser, Serra, vZ. (2015)]

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.