Lectures on Nonparametric Bayesian Statistics

Aad van der Vaart

Universiteit Leiden, Netherlands

Bad Belzig, March 2013

Contents

Introduction

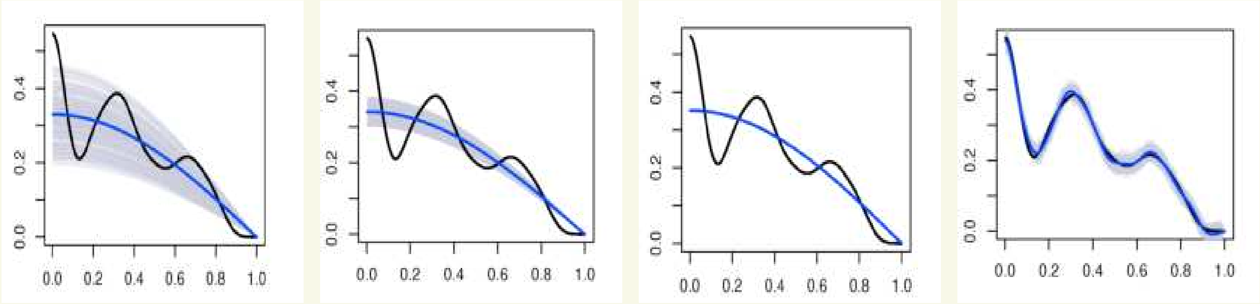
Dirichlet process

Consistency and rates

Gaussian process priors

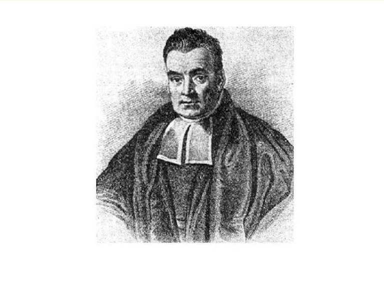
Dirichlet mixtures

All the rest



Introduction

The Bayesian paradigm



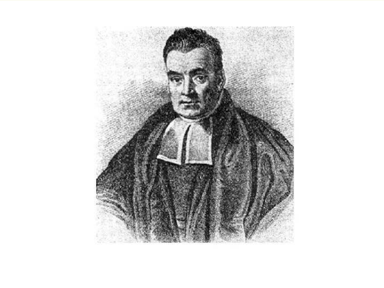
- A parameter Θ is generated according to a prior distribution Π.
- Given $\Theta = \theta$ the data X is generated according to a measure P_θ.

This gives a joint distribution of (X, Θ).

- Given observed data x the statistician computes the conditional distribution of Θ given $X = x$, the posterior distribution:

$$\Pi(\theta \in B |\, X).$$

The Bayesian paradigm



- A parameter Θ is generated according to a prior distribution Π.
- Given $\Theta = \theta$ the data X is generated according to a measure P_θ.

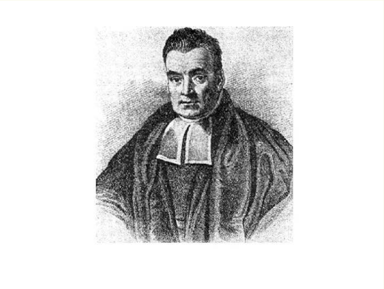
This gives a joint distribution of (X, Θ).

- Given observed data x the statistician computes the conditional distribution of Θ given $X = x$, the posterior distribution:

$$\Pi(\theta \in B|\, X).$$

We assume whatever needed (e.g. Θ Polish and Π a probability distribution on its Borel σ-field; Polish sample space) to make this well defined.

Bayes's rule



- A parameter Θ is generated according to a prior distribution Π.
- Given $\Theta = \theta$ the data X is generated according to a measure P_θ.

This gives a joint distribution of (X, Θ).

- Given observed data x the statistician computes the conditional distribution of Θ given $X = x$, the posterior distribution:

$$\Pi(\theta \in B |\, X).$$

If P_θ is given by a density $x \mapsto p_\theta(x)$, then **Bayes's rule** gives

$$\Pi(\Theta \in B |\, X) = \frac{\int_B p_\theta(X)\, d\Pi(\theta)}{\int_\Theta p_\theta(X)\, d\Pi(\theta)}$$

Bayes's rule

- A parameter Θ is generated according to a prior distribution Π.
- Given $\Theta = \theta$ the data X is generated according to a measure P_θ.

This gives a joint distribution of (X, Θ).

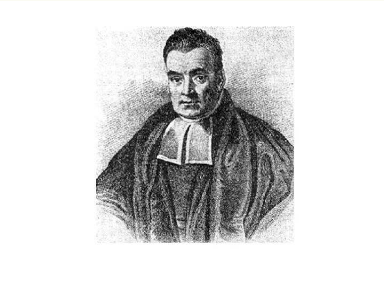
- Given observed data x the statistician computes the conditional distribution of Θ given $X = x$, the posterior distribution:

$$\Pi(\theta \in B|\, X).$$

If P_θ is given by a density $x \mapsto p_\theta(x)$, then **Bayes's rule** gives

$$d\Pi(\theta|\, X) \propto p_\theta(X)\, d\Pi(\theta)$$

Reverend Thomas



Thomas Bayes (1702–1761, 1763) followed this argument with Θ possessing the *uniform* distribution and X given $\Theta = \theta$ *binomial* (n, θ).

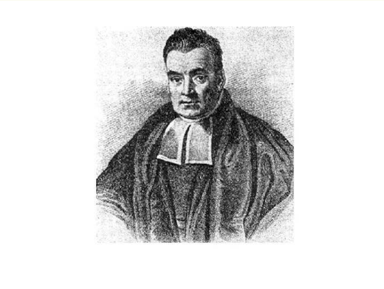
Using his famous rule he computed that the posterior distribution is then *Beta*$(X+1, n-X+1)$.

$$\Pr(a \le \Theta \le b) = b - a, \qquad 0 < a < b < 1,$$

$$\Pr(X = x |\, \Theta = \theta) = \binom{n}{x} \theta^x (1-\theta)^{n-x}, \qquad x = 0, 1, \ldots, n,$$

$$\Pr(a \le \Theta \le b |\, X = x) = \int_a^b \theta^x (1-\theta)^{n-x} \, d\theta / B(x+1, n-x+1).$$

Reverend Thomas



Thomas Bayes (1702–1761, 1763) followed this argument with Θ possessing the *uniform* distribution and X given $\Theta = \theta$ *binomial* (n, θ).

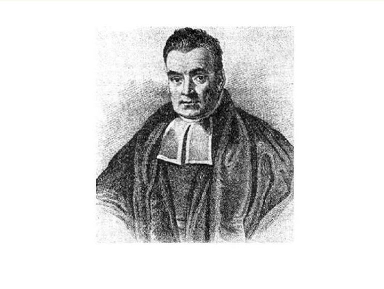
Using his famous rule he computed that the posterior distribution is then *Beta*$(X+1, n-X+1)$.

$$\Pr(a \leq \Theta \leq b) = b - a, \qquad 0 < a < b < 1,$$

$$\Pr(X = x |\, \Theta = \theta) = \binom{n}{x} \theta^x (1-\theta)^{n-x}, \qquad x = 0, 1, \ldots, n,$$

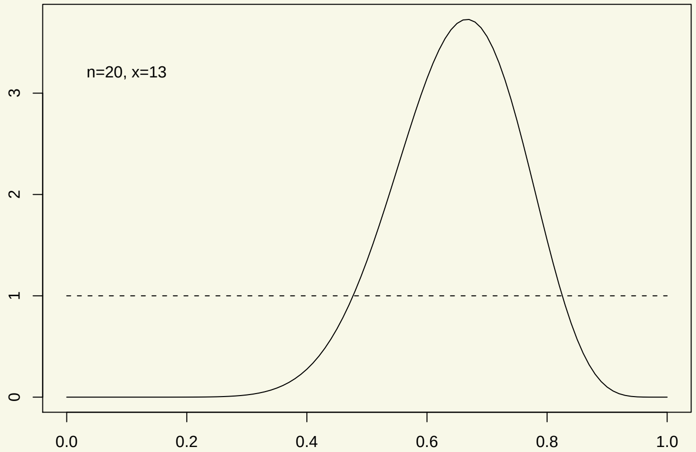
$$d\Pi(\theta |\, X) = \theta^X (1-\theta)^{n-X} \cdot 1.$$

Reverend Thomas

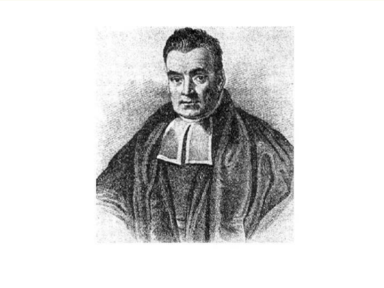


Thomas Bayes (1702–1761, 1763) followed this argument with Θ possessing the *uniform* distribution and X given $\Theta = \theta$ *binomial* (n, θ).

Using his famous rule he computed that the posterior distribution is then *Beta*$(X+1, n-X+1)$.

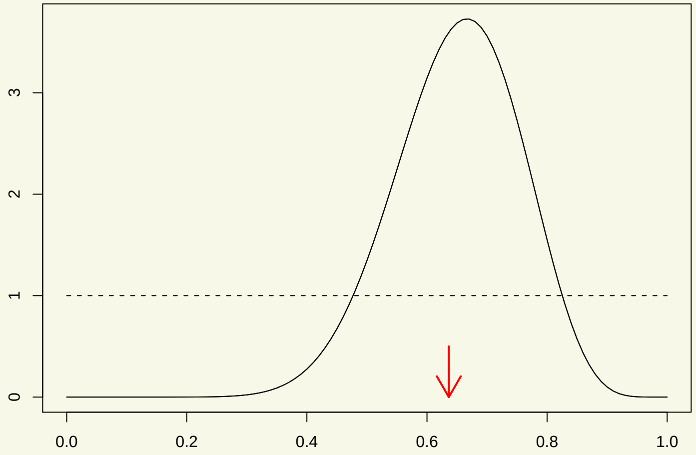


Reverend Thomas

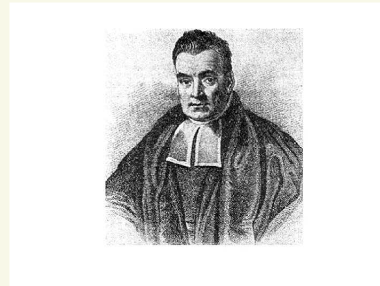


Thomas Bayes (1702–1761, 1763) followed this argument with Θ possessing the *uniform* distribution and X given $\Theta = \theta$ *binomial* (n, θ).

Using his famous rule he computed that the posterior distribution is then *Beta*$(X + 1, n - X + 1)$.

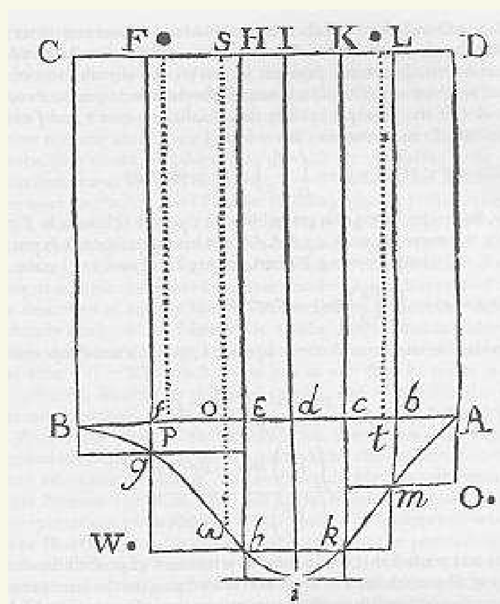


Reverend Thomas



Thomas Bayes (1702–1761, 1763) followed this argument with Θ possessing the *uniform* distribution and X given $\Theta = \theta$ *binomial* (n, θ).

Using his famous rule he computed that the posterior distribution is then *Beta*$(X + 1, n - X + 1)$.



Parametric Bayes

Pierre-Simon Laplace (1749-1827) rediscovered Bayes' argument and applied it to general parametric models: models smoothly indexed by a Euclidean parameter θ.

For instance, the linear regression model, where one observes $(x_1, Y_n), \ldots, (x_n, Y_n)$ following

$$Y_i = \theta_0 + \theta_1 x_i + e_i,$$

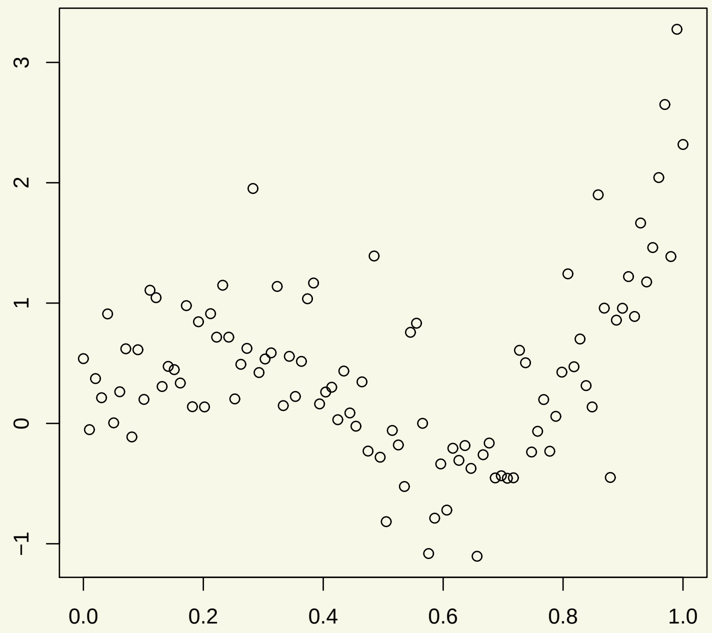
for $e_1, \ldots, e_n$ independent normal errors with zero mean.

Nonparametric Bayes

If the parameter θ is a function, then the prior is a probability distribution on an function space. So is the posterior, given the data. Bayes' formula does not change:

$$d\Pi(\theta|\, X) \propto p_\theta(X)\, d\Pi(\theta).$$

Prior and posterior can be visualized by plotting functions that are simulated from these distributions.

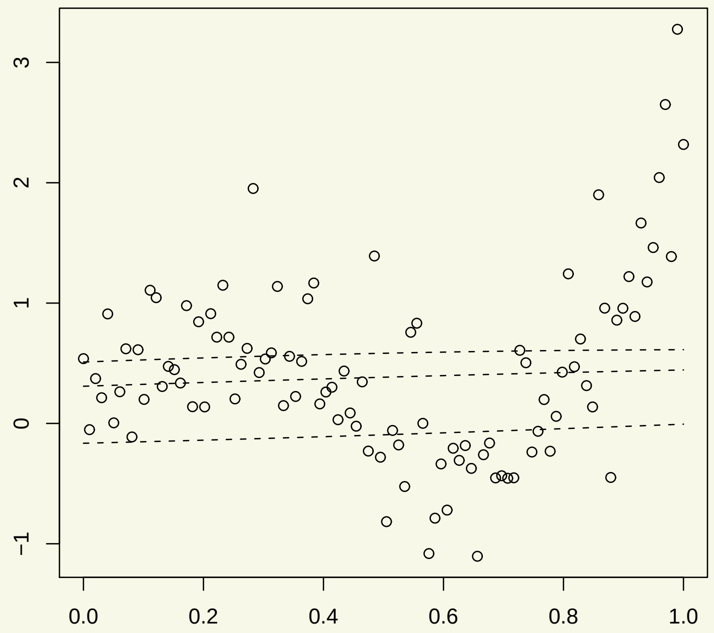


Nonparametric Bayes

If the parameter θ is a function, then the prior is a probability distribution on an function space. So is the posterior, given the data. Bayes' formula does not change:

$$d\Pi(\theta|\, X) \propto p_\theta(X)\, d\Pi(\theta).$$

Prior and posterior can be visualized by plotting functions that are simulated from these distributions.

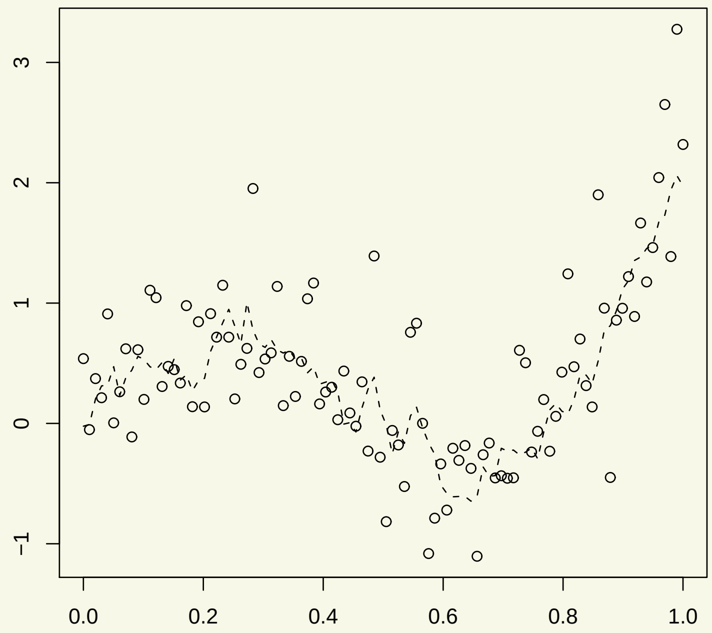


Nonparametric Bayes

If the parameter θ is a function, then the prior is a probability distribution on an function space. So is the posterior, given the data. Bayes' formula does not change:

$$d\Pi(\theta|\, X) \propto p_\theta(X)\, d\Pi(\theta).$$

Prior and posterior can be visualized by plotting functions that are simulated from these distributions.

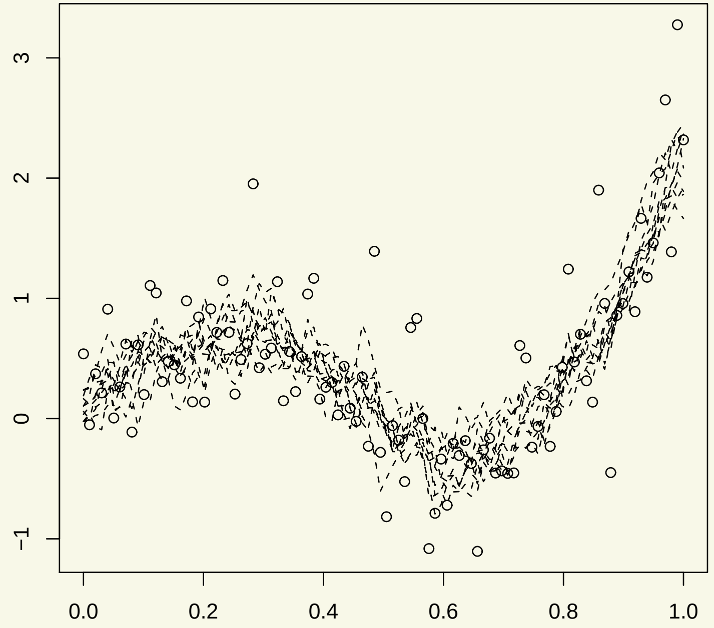


Nonparametric Bayes

If the parameter θ is a function, then the prior is a probability distribution on an function space. So is the posterior, given the data. Bayes' formula does not change:

$$d\Pi(\theta|\, X) \propto p_\theta(X)\, d\Pi(\theta).$$

Prior and posterior can be visualized by plotting functions that are simulated from these distributions.

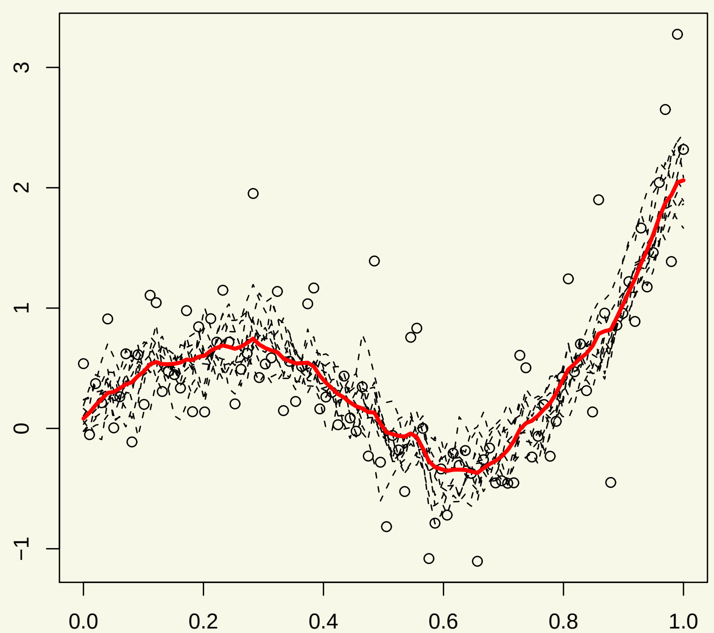


Nonparametric Bayes

If the parameter θ is a function, then the prior is a probability distribution on an function space. So is the posterior, given the data. Bayes' formula does not change:

$$d\Pi(\theta|\, X) \propto p_\theta(X)\, d\Pi(\theta).$$

Prior and posterior can be visualized by plotting functions that are simulated from these distributions.

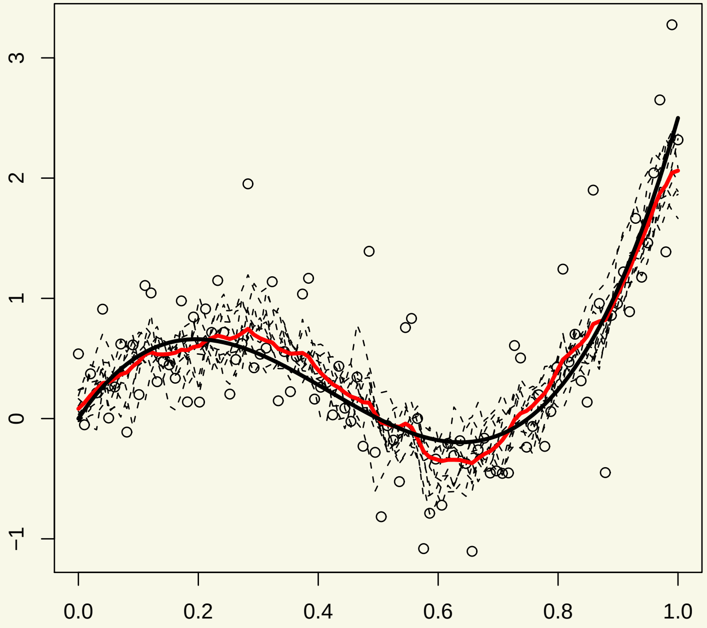


Nonparametric Bayes

If the parameter θ is a function, then the prior is a probability distribution on an function space. So is the posterior, given the data. Bayes' formula does not change:

$$d\Pi(\theta|\, X) \propto p_\theta(X)\, d\Pi(\theta).$$

Prior and posterior can be visualized by plotting functions that are simulated from these distributions.



Subjectivism

A philosophical Bayesian statistician views the prior distribution as an expression of his personal beliefs on the state of the world, before gathering the data.

After seeing the data he updates his beliefs into the posterior distribution.

Most scientists do not like dependence on subjective priors.

- One can opt for objective or noninformative priors.
- One can also mathematically study the role of the prior, and hope to find that it is small.

Frequentist Bayesian

Assume that the data X is generated according to a given parameter θ_0 and consider the posterior $\Pi(\theta \in \cdot \mid X)$ as a random measure on the parameter set dependent on X.

We like $\Pi(\theta \in \cdot \mid X)$ to put “most” of its mass near θ_0 for “most” X.

Frequentist Bayesian

Assume that the data X is generated according to a given parameter θ_0 and consider the posterior $\Pi(\theta \in \cdot \mid X)$ as a random measure on the parameter set dependent on X.

We like $\Pi(\theta \in \cdot \mid X)$ to put “most” of its mass near θ_0 for “most” X.

Asymptotic setting: data X^n where the information increases as $n \to \infty$. We like the posterior $\Pi_n(\cdot \mid X^n)$ to contract to $\{\theta_0\}$, at a good rate.

Frequentist Bayesian

Assume that the data X is generated according to a given parameter θ_0 and consider the posterior $\Pi(\theta \in \cdot \mid X)$ as a random measure on the parameter set dependent on X.

We like $\Pi(\theta \in \cdot \mid X)$ to put "most" of its mass near θ_0 for "most" X.

Asymptotic setting: data X^n where the information increases as $n \to \infty$.
We like the posterior $\Pi_n(\cdot \mid X^n)$ to contract to $\{\theta_0\}$, at a good rate.

Desirable properties:

- Consistency + rate
- Adaptation
- Distributional approximations
- Uncertainty quantification

Frequentist Bayesian

Assume that the data X is generated according to a given parameter θ_0 and consider the posterior $\Pi(\theta \in \cdot \,|\, X)$ as a random measure on the parameter set dependent on X.

We like $\Pi(\theta \in \cdot \,|\, X)$ to put “most” of its mass near θ_0 for “most” X.

Asymptotic setting: data X^n where the information increases as $n \to \infty$. We like the posterior $\Pi_n(\cdot \,|\, X^n)$ to contract to $\{\theta_0\}$, at a good rate.

Desirable properties:

- Consistency + rate
- Adaptation
- Distributional approximations
- Uncertainty quantification

We assume that $P_{\theta_0} \ll \int P_\theta \, d\Pi(\theta)$ to make these questions well posed.

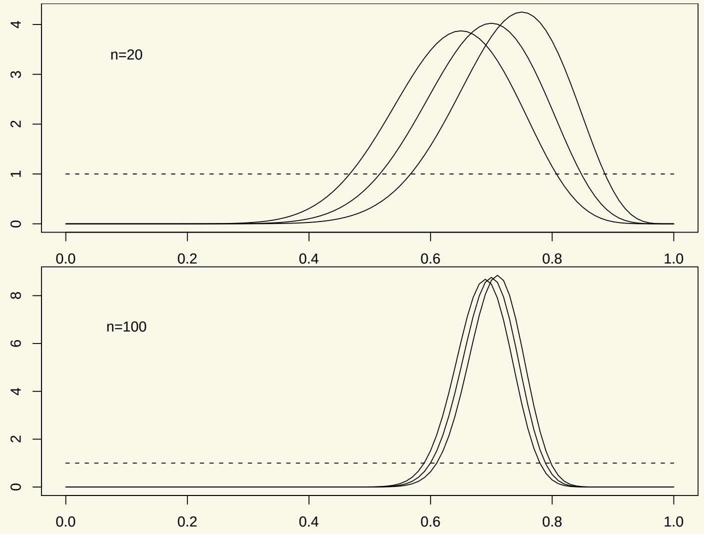
Parametric models

Suppose the data are a random sample $X_1, \ldots, X_n$ from a density $x \mapsto p_\theta(x)$ that is smoothly and **identifiably** parametrized by a vector $\theta \in \mathbb{R}^d$ (e.g. $\theta \mapsto \sqrt{p_\theta}$ continuously differentiable as map in $L_2(\mu)$).

Theorem (Laplace, Bernstein, von Mises, LeCam 1989)**.** *Under $P^n_{\theta_0}$-probability, for any prior with density that is positive around θ_0,*

$$\left\| \Pi(\cdot \,|\, X_1, \ldots, X_n) - N_d\Big(\tilde\theta_n, \frac{1}{n} I_{\theta_0}^{-1}\Big)(\cdot) \right\| \to 0.$$

Here $\tilde\theta_n$ is any efficient estimator of θ.



Parametric models

Suppose the data are a random sample $X_1, \ldots, X_n$ from a density $x \mapsto p_\theta(x)$ that is smoothly and **identifiably** parametrized by a vector $\theta \in \mathbb{R}^d$ (e.g. $\theta \mapsto \sqrt{p_\theta}$ continuously differentiable as map in $L_2(\mu)$).

Theorem (Laplace, Bernstein, von Mises, LeCam 1989)**.** *Under $P^n_{\theta_0}$-probability, for any prior with density that is positive around θ_0,*

$$\Big\| \Pi(\cdot \,|\, X_1, \ldots, X_n) - N_d\big(\tilde\theta_n, \frac{1}{n} I_{\theta_0}^{-1}\big)(\cdot)\Big\| \to 0.$$

Here $\tilde\theta_n$ is any efficient estimator of θ.

In particular, the posterior distribution concentrates most of its mass on balls of radius $O(1/\sqrt{n})$ around θ_0, and a central set of posterior probability 95 % is equivalent to the usual Wald confidence set.

The prior washes out completely.

Support

Definition. The support of a prior Π is the smallest closed set F with $\Pi(F) = 1$.

In nonparametrics we like priors with big (or even full) support, equal to a infinite-dimensional set.

Full support means that every open set has positive (prior) probability.

Support

Definition. The support of a prior Π is the smallest closed set F with $\Pi(F) = 1$.

In nonparametrics we like priors with big (or even full) support, equal to a infinite-dimensional set.

Full support means that every open set has positive (prior) probability.

Support depends on topology. It is well defined, e.g. if the parameter space is Polish.

Dirichlet process

Random measures

- $\mathfrak{M}$: all probability measures on (Polish) sample space $(\mathfrak{X}, \mathscr{X})$.
- $\mathscr{M}$: σ-field generated by all maps $M \mapsto M(A)$, $A \in \mathscr{X}$.

Lemma. *$\mathscr{M}$ is also the Borel σ-field on $\mathfrak{M}$ equipped with the weak topology ("of convergence in distribution").*

Random measures

- $\mathfrak{M}$: all probability measures on (Polish) sample space $(\mathfrak{X}, \mathscr{X})$.
- $\mathscr{M}$: σ-field generated by all maps $M \mapsto M(A)$, $A \in \mathscr{X}$.

Lemma. *$\mathscr{M}$ is also the Borel σ-field on $\mathfrak{M}$ equipped with the weak topology ("of convergence in distribution").*

Definition. A *random probability measure* on $(\mathfrak{X}, \mathscr{X})$ is a map $P\colon (\Omega, \mathscr{U}, \Pr) \to \mathfrak{M}$ such that $P(A)$ is a random variable for every $A \in \mathscr{X}$. (Equivalently, a Borel measurable map in $\mathfrak{M}$.)

Random measures

- $\mathfrak{M}$: all probability measures on (Polish) sample space $(\mathfrak{X}, \mathscr{X})$.
- $\mathscr{M}$: σ-field generated by all maps $M \mapsto M(A)$, $A \in \mathscr{X}$.

Lemma. *$\mathscr{M}$ is also the Borel σ-field on $\mathfrak{M}$ equipped with the weak topology ("of convergence in distribution").*

Definition. A *random probability measure* on $(\mathfrak{X}, \mathscr{X})$ is a map $P\colon (\Omega, \mathscr{U}, \Pr) \to \mathfrak{M}$ such that $P(A)$ is a random variable for every $A \in \mathscr{X}$. (Equivalently, a Borel measurable map in $\mathfrak{M}$.)

The law of P is a prior on $(\mathfrak{M}, \mathscr{M})$.

Random measures

- $\mathfrak{M}$: all probability measures on (Polish) sample space $(\mathfrak{X}, \mathscr{X})$.
- $\mathscr{M}$: σ-field generated by all maps $M \mapsto M(A)$, $A \in \mathscr{X}$.

Lemma. *$\mathscr{M}$ is also the Borel σ-field on $\mathfrak{M}$ equipped with the weak topology ("of convergence in distribution").*

Definition. A *random probability measure* on $(\mathfrak{X}, \mathscr{X})$ is a map $P\colon (\Omega, \mathscr{U}, \Pr) \to \mathfrak{M}$ such that $P(A)$ is a random variable for every $A \in \mathscr{X}$. (Equivalently, a Borel measurable map in $\mathfrak{M}$.)

The law of P is a prior on $(\mathfrak{M}, \mathscr{M})$.

Definition. The *mean measure* of P is the measure $A \mapsto \mathrm{E}P(A)$.

Discrete random measures

- $W_1, W_2, \ldots$ nonnegative variables with $\sum_{i=1}^{\infty} W_i = 1$, independent of
- $\theta_1, \theta_2, \ldots \overset{\text{iid}}{\sim} G$, random variables with values in $\mathfrak{X}$.

$$P = \sum_{i=1}^{\infty} W_i \delta_{\theta_i}.$$

Discrete random measures

- $W_1, W_2, \ldots$ nonnegative variables with $\sum_{i=1}^{\infty} W_i = 1$, independent of
- $\theta_1, \theta_2, \ldots \stackrel{\text{iid}}{\sim} G$, random variables with values in $\mathfrak{X}$.

$$P = \sum_{i=1}^{\infty} W_i \delta_{\theta_i}.$$

Lemma. *If $(W_1, \ldots, W_n)$ has (full) support the unit simplex for every n and the law of θ_1 has (full) support $\mathfrak{X}$, then P has full support $\mathfrak{M}$ relative to the weak topology.*

Discrete random measures

- $W_1, W_2, \ldots$ nonnegative variables with $\sum_{i=1}^{\infty} W_i = 1$, independent of
- $\theta_1, \theta_2, \ldots \overset{\text{iid}}{\sim} G$, random variables with values in $\mathfrak{X}$.

$$P = \sum_{i=1}^{\infty} W_i \delta_{\theta_i}.$$

Lemma. *If $(W_1, \ldots, W_n)$ has (full) support the unit simplex for every n and the law of θ_1 has (full) support $\mathfrak{X}$, then P has full support $\mathfrak{M}$ relative to the weak topology.*

Proof.

- Finitely discrete distributions are weakly dense in $\mathfrak{M}$.
- It suffices to show that P gives positive probability to any weak neighbourhood of a distribution of the form $P^* = \sum_{i=1}^{k} w_i^* \delta_{\theta_i^*}$.
- $\left\{\sum_{i>k} W_i < \epsilon, \max_{i\leq k} |W_i - w_i^*| \vee |\theta_i - \theta_i^*| < \epsilon\right\}$ is open and hence has positive probability.

$\square$

Discrete random measures

- $W_1, W_2, \ldots$ nonnegative variables with $\sum_{i=1}^{\infty} W_i = 1$, independent of
- $\theta_1, \theta_2, \ldots \stackrel{\text{iid}}{\sim} G$, random variables with values in $\mathfrak{X}$.

$$P = \sum_{i=1}^{\infty} W_i \delta_{\theta_i}.$$

Lemma. *If $(W_1, \ldots, W_n)$ has (full) support the unit simplex for every n and the law of θ_1 has (full) support $\mathfrak{X}$, then P has full support $\mathfrak{M}$ relative to the weak topology.*

Proof.

- Finitely discrete distributions are weakly dense in $\mathfrak{M}$.
- It suffices to show that P gives positive probability to any weak neighbourhood of a distribution of the form $P^* = \sum_{i=1}^{k} w_i^* \delta_{\theta_i^*}$.
- $\left\{\sum_{i>k} W_i < \epsilon, \max_{i \le k} |W_i - w_i^*| \vee |\theta_i - \theta_i^*| < \epsilon\right\}$ is open and hence has positive probability.

$\square$

Stick breaking

Given i.i.d. $Y_1, Y_2, \ldots$ in $[0, 1]$,

$$W_1 = Y_1, W_2 = (1 - Y_1)Y_2, W_3 = (1 - Y_1)(1 - Y_2)Y_3, \ldots$$

Stick breaking

Given i.i.d. $Y_1, Y_2, \ldots$ in $[0,1]$,

$$W_1 = Y_1, W_2 = (1 - Y_1)Y_2, W_3 = (1 - Y_1)(1 - Y_2)Y_3, \ldots$$

Lemma. $(W_1, W_2, \ldots)$ *is a random probability measure on* $\mathbb{N}$ *iff* $\mathrm{P}(Y_1 > 0) > 0$*, and has full support if* Y_1 *has support* $[0,1]$*.*

Stick breaking

Given i.i.d. $Y_1, Y_2, \ldots$ in $[0,1]$,

$$W_1 = Y_1, W_2 = (1-Y_1)Y_2, W_3 = (1-Y_1)(1-Y_2)Y_3, \ldots$$

Lemma. $(W_1, W_2, \ldots)$ *is a random probability measure on* $\mathbb{N}$ *iff* $\mathrm{P}(Y_1 > 0) > 0$*, and has full support if* Y_1 *has support* $[0,1]$*.*

Proof.

- The remaining length of the stick at stage j is $1 - \sum_{l=1}^{j} W_l = \prod_{l=1}^{j}(1-Y_l)$, and tends to zero a.s. iff $\prod_{l=1}^{j}(1-\mathrm{E}Y_l) \to 0$.
- $(W_1, \ldots, W_n)$ is a continuous function of $(Y_1, \ldots, Y_n)$ with full range, for every k.

$\square$

Stick breaking

Given i.i.d. $Y_1, Y_2, \ldots$ in $[0,1]$,

$$W_1 = Y_1, W_2 = (1 - Y_1)Y_2, W_3 = (1 - Y_1)(1 - Y_2)Y_3, \ldots$$

Lemma. $(W_1, W_2, \ldots)$ *is a random probability measure on* $\mathbb{N}$ *iff* $\mathrm{P}(Y_1 > 0) > 0$*, and has full support if* Y_1 *has support* $[0,1]$*.*

Proof.

- The remaining length of the stick at stage j is $1 - \sum_{l=1}^{j} W_l = \prod_{l=1}^{j}(1 - Y_l)$, and tends to zero a.s. iff $\prod_{l=1}^{j}(1 - \mathrm{E}Y_l) \to 0$.
- $(W_1, \ldots, W_n)$ is a continuous function of $(Y_1, \ldots, Y_n)$ with full range, for every k.

$\square$

EXAMPLE OF PARTICULAR INTEREST: $Y_1, Y_2, \overset{\text{iid}}{\sim} \mathrm{Be}(1, M)$.

Random measures as stochastic processes

A random measure P induces the distributions on $\mathbb{R}^k$ of the random vectors

$$\big(P(A_1), \ldots, P(A_k)\big), \qquad A_1, \ldots, A_k \in \mathscr{X}.$$

Random measures as stochastic processes

A random measure P induces the distributions on $\mathbb{R}^k$ of the random vectors

$$\big(P(A_1), \ldots, P(A_k)\big), \qquad A_1, \ldots, A_k \in \mathscr{X}.$$

Conversely suppose we want a measure with particular distributions, and can construct a stochastic process $\big(P(A)\colon A \in \mathscr{X}\big)$ with these distributions (e.g. by Kolmogorov's consistency theorem).

Random measures as stochastic processes

A random measure P induces the distributions on $\mathbb{R}^k$ of the random vectors

$$\big(P(A_1), \ldots, P(A_k)\big), \qquad A_1, \ldots, A_k \in \mathscr{X}.$$

Conversely suppose we want a measure with particular distributions, and can construct a stochastic process $\big(P(A)\colon A \in \mathscr{X}\big)$ with these distributions (e.g. by Kolmogorov's consistency theorem).

It will be true that

(i). $P(\emptyset) = 0$, $P(\mathscr{X}) = 1$, a.s.

(ii). $P(A_1 \cup A_1) = P(A_1) + P(A_2)$, a.s., for any disjoint A_1, A_2.

Random measures as stochastic processes

A random measure P induces the distributions on $\mathbb{R}^k$ of the random vectors

$$\big(P(A_1), \ldots, P(A_k)\big), \qquad A_1, \ldots, A_k \in \mathscr{X}.$$

Conversely suppose we want a measure with particular distributions, and can construct a stochastic process $\big(P(A)\colon A \in \mathscr{X}\big)$ with these distributions (e.g. by Kolmogorov's consistency theorem).

It will be true that

(i). $P(\emptyset) = 0$, $P(\mathscr{X}) = 1$, a.s.
(ii). $P(A_1 \cup A_1) = P(A_1) + P(A_2)$, a.s., for any disjoint A_1, A_2.

However, we do not automatically have that P is a random measure.

Theorem. *If $\big(P(A)\colon A \in \mathscr{X}\big)$ is a stochastic process that satisfies (i) and (ii) and whose mean $A \mapsto \mathrm{E}P(A)$ is a Borel measure on $\mathfrak{X}$, then there exists a version of P that is a random measure on $(\mathfrak{X}, \mathscr{X})$.*

Finite-dimensonal Dirichlet distribution

Definition. $(X_1, \dots, X_k)$ possesses a *Dirichlet* $(k, \alpha_1, \dots, \alpha_k)$ *distribution* for $\alpha_i > 0$ it has (Lebesgue) density on the unit simplex proportional to

$$x \mapsto x_1^{\alpha_1 - 1} \cdots x_k^{\alpha_k - 1}.$$

Finite-dimensonal Dirichlet distribution

Definition. $(X_1, \ldots, X_k)$ possesses a *Dirichlet* $(k, \alpha_1, \ldots, \alpha_k)$ *distribution* for $\alpha_i > 0$ it has (Lebesgue) density on the unit simplex proportional to

$$x \mapsto x_1^{\alpha_1 - 1} \cdots x_k^{\alpha_k - 1}.$$

We extend to $\alpha_i = 0$ for one or more i on the understanding that $X_i = 0$.

Finite-dimensonal Dirichlet distribution

Definition. $(X_1, \dots, X_k)$ possesses a *Dirichlet* $(k, \alpha_1, \dots, \alpha_k)$ *distribution* for $\alpha_i > 0$ it has (Lebesgue) density on the unit simplex proportional to

$$x \mapsto x_1^{\alpha_1 - 1} \cdots x_k^{\alpha_k - 1}.$$

We extend to $\alpha_i = 0$ for one or more i on the understanding that $X_i = 0$.

EXAMPLES

- For $k = 2$ we have $X_1 \sim \mathrm{Be}(\alpha_1, \alpha_2)$ and $X_2 = 1 - X_1 \sim \mathrm{Be}(\alpha_2, \alpha_1)$.
- The $\mathrm{Dir}(k; 1, \dots, 1)$-distribution is the uniform distribution on the simplex.

Dirichlet distribution — properties

Proposition (Gamma representation). *If* $Y_i \overset{\text{ind}}{\sim} \mathrm{Ga}(\alpha_i, 1)$*, then* $(Y_1/Y, \ldots, Y_k/Y) \sim \mathrm{Dir}(k; \alpha_1, \ldots, \alpha_k)$*, and is independent of and* $Y := \sum_{i=1}^k Y_i$.

Proposition (Aggregation). *If* $X \sim \mathrm{Dir}(k; \alpha_1, \ldots, \alpha_k)$ *and* $Z_j = \sum_{i \in I_j} X_i$ *for a given partition* $I_1, \ldots, I_m$ *of* $\{1, \ldots, k\}$*, then*

(i). $(Z_1, \ldots, Z_m) \sim \mathrm{Dir}(m; \beta_1, \ldots, \beta_m)$*, where* $\beta_j = \sum_{i \in I_j} \alpha_i$.

(ii). $(X_i/Z_j : i \in I_j) \overset{\text{ind}}{\sim} \mathrm{Dir}(\#I_j; \alpha_i, i \in I_j)$*, for* $j = 1, \ldots, m$.

(iii). $(Z_1, \ldots, Z_m)$ *and* $(X_i/Z_j : i \in I_j, j = 1, \ldots, m)$ *are independent.*

Conversely, if X *is a random vector such that (i)–(iii) hold, for a given partition* $I_1, \ldots, I_m$ *and* $Z_j = \sum_{i \in I_j} X_i$*, then* $X \sim \mathrm{Dir}(k; \alpha_1, \ldots, \alpha_k)$.

Proposition. $\mathrm{E}(X_i) = \alpha_i/|\alpha|$ *and* $\mathrm{var}(X_i) = \alpha_i(|\alpha| - \alpha_i)/(|\alpha|^2(|\alpha|+1))$*, for* $|\alpha| = \sum_{i=1}^k \alpha_i$.

Proposition (Conjugacy). *If* $p \sim \mathrm{Dir}(k; \alpha)$ *and* $N|\, p \sim \mathrm{MN}(n, k; p)$*, then* $p|\, N \sim \mathrm{Dir}(k; \alpha + N)$.

Dirichlet process

Definition. A random measure P on $(\mathfrak{X}, \mathscr{X})$ is a *Dirichlet process* with *base measure* α, if for every partition $A_1, \ldots, A_k$ of $\mathfrak{X}$,

$$\big(P(A_1), \ldots, P(A_k)\big) \sim \mathrm{Dir}\big(k; \alpha(A_1), \ldots, \alpha(A_k)\big).$$

We write $P \sim \mathrm{DP}(\alpha)$, $|\alpha| := \alpha(\mathfrak{X})$ and $\bar{\alpha} := \alpha/|\alpha|$.

Dirichlet process

Definition. A random measure P on $(\mathfrak{X}, \mathscr{X})$ is a *Dirichlet process* with *base measure* α, if for every partition $A_1, \ldots, A_k$ of $\mathfrak{X}$,

$$\big(P(A_1), \ldots, P(A_k)\big) \sim \mathrm{Dir}\big(k; \alpha(A_1), \ldots, \alpha(A_k)\big).$$

We write $P \sim \mathrm{DP}(\alpha)$, $|\alpha|\colon= \alpha(\mathfrak{X})$ and $\bar{\alpha}\colon= \alpha/|\alpha|$.

$$\mathrm{E}P(A) = \bar{\alpha}(A), \qquad \mathrm{var}\, P(A) = \frac{\bar{\alpha}(A)\bar{\alpha}(A^c)}{1+|\alpha|}.$$

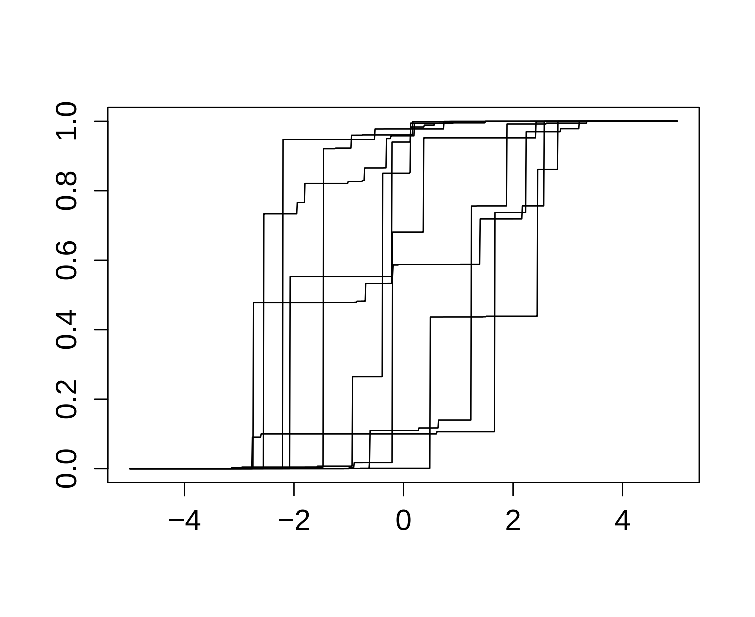
Dirichlet process

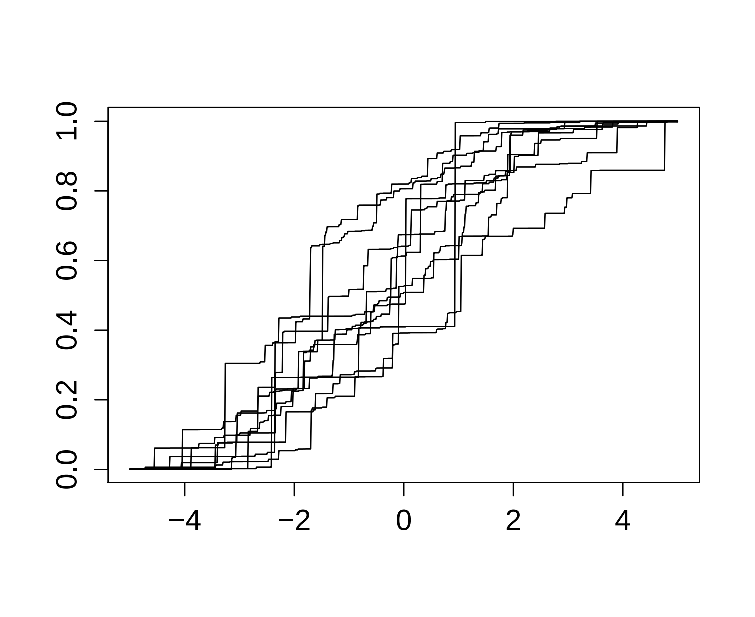
Definition. A random measure P on $(\mathfrak{X}, \mathscr{X})$ is a *Dirichlet process* with *base measure* α, if for every partition $A_1, \ldots, A_k$ of $\mathfrak{X}$,

$$\big(P(A_1), \ldots, P(A_k)\big) \sim \mathrm{Dir}\big(k; \alpha(A_1), \ldots, \alpha(A_k)\big).$$

We write $P \sim \mathrm{DP}(\alpha)$, $|\alpha| := \alpha(\mathfrak{X})$ and $\bar{\alpha} := \alpha/|\alpha|$.

$$\mathrm{E}P(A) = \bar{\alpha}(A), \qquad \operatorname{var} P(A) = \frac{\bar{\alpha}(A)\bar{\alpha}(A^c)}{1 + |\alpha|}.$$





Dirichlet process — existence

Theorem. *For any Borel measure α the Dirichlet process exists as a Borel measure on $\mathfrak{M}$.*

Dirichlet process — existence

Theorem. *For any Borel measure α the Dirichlet process exists as a Borel measure on $\mathfrak{M}$.*

Proof.

- An arbitrary collection of sets $A_1, \ldots, A_k$ can be partitioned in 2^k atoms $B_j = A_1^* \cap A_2^* \cap \cdots \cap A_k^*$, where A^* stands for A or A^c.
- The distribution of $\big(P(B_j)\colon j = 1, \ldots, 2^k\big)$ must be Dirichlet.
- Define the distribution of $\big(P(A_1), \ldots, P(A_k)\big)$ corresponding to the fact that each $P(A_i)$ must be a sum of some set of $P(B_j)$.
- Using properties of finite-dimensional Dirichlets, check that this is consistent in the sense of Kolmogorov, so that a version of the stochastic process $\big(P(A)\colon A \in \mathscr{X}\big)$ exists.
- Apply the general theorem on existence of random measures.

$\square$

Sethuraman representation

Theorem. *If* $\theta_1, \theta_2, \ldots \overset{iid}{\sim} \bar{\alpha}$ *and* $Y_1, Y_2, \ldots \overset{iid}{\sim} \mathrm{Be}(1, M)$ *are independent random variables and* $W_j = Y_j \prod_{l=1}^{j-1}(1 - Y_l)$*, then* $\sum_{j=1}^{\infty} W_j \delta_{\theta_j} \sim \mathrm{DP}(M\bar{\alpha})$*.*

Sethuraman representation

Theorem. *If* $\theta_1, \theta_2, \ldots \overset{iid}{\sim} \bar{\alpha}$ *and* $Y_1, Y_2, \ldots \overset{iid}{\sim} \mathrm{Be}(1, M)$ *are independent random variables and* $W_j = Y_j \prod_{l=1}^{j-1}(1 - Y_l)$, *then* $\sum_{j=1}^{\infty} W_j \delta_{\theta_j} \sim \mathrm{DP}(M\bar{\alpha})$.

Proof.

$$P := W_1\delta_{\theta_1} + \sum_{j=2}^{\infty} W_j \delta_{\theta_j} = Y_1\delta_{\theta_1} + (1 - Y_1)P', \qquad P' = \sum_{j=2}^{\infty}(Y_j \prod_{l=2}^{j-1}(1 - Y_l))\delta_{\theta_j}.$$

Hence $Q = \big(P(A_1), \ldots, P(A_k)\big)$ and $N = \big(\delta_{\theta_1}(A_1), \ldots, \delta_{\theta_1}(A_k)\big)$ satisfy

$$Q =_d YN + (1 - Y)Q.$$

Now

- For given $Y \sim \mathrm{Be}(1, M)$ and independent $\theta \sim G$ there is at most one solution in distribution Q.
- A Dirichlet vector Q is a solution.

Second follows by properties of Dirichlet (not obvious!).
First: see next slide. $\square$

Sethuraman representation

Proof. (Continued)

$$Q =_d YN + (1-Y)Q.$$

Given i.i.d. copies (Y_n, N_n) and given independent solutions Q and Q':

$$\begin{aligned} Q_0 &= Q, & Q'_0 &= Q', \\ Q_n &= Y_n N_n + (1-Y_n)Q_{n-1}, & Q'_n &= Y_n N_n + (1-Y_n)Q'_{n-1}. \end{aligned}$$

Then $Q_n =_d Q$ and $Q'_n =_d Q'$ for every n, and

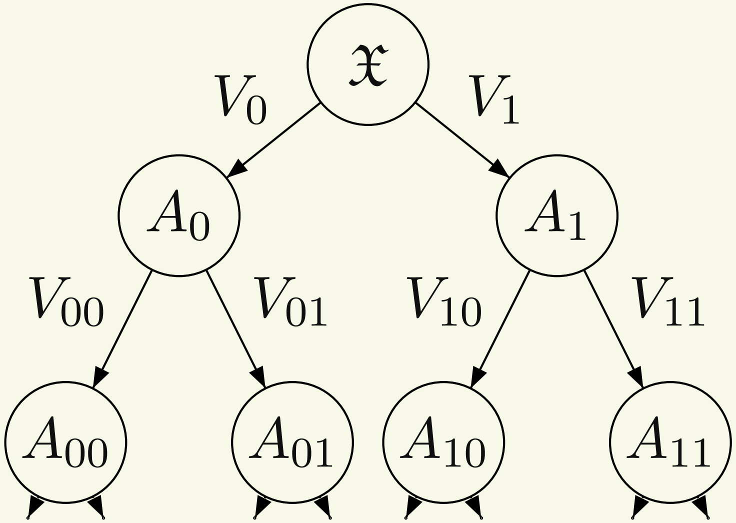
$$\|Q_n - Q'_n\| = |1-Y_n|\,\|Q_{n-1} - Q'_{n-1}\| = \prod_{i=1}^{n} |1-Y_i|\,\|Q - Q'\| \to 0$$

Hence $Q =_d Q'$.

$\square$

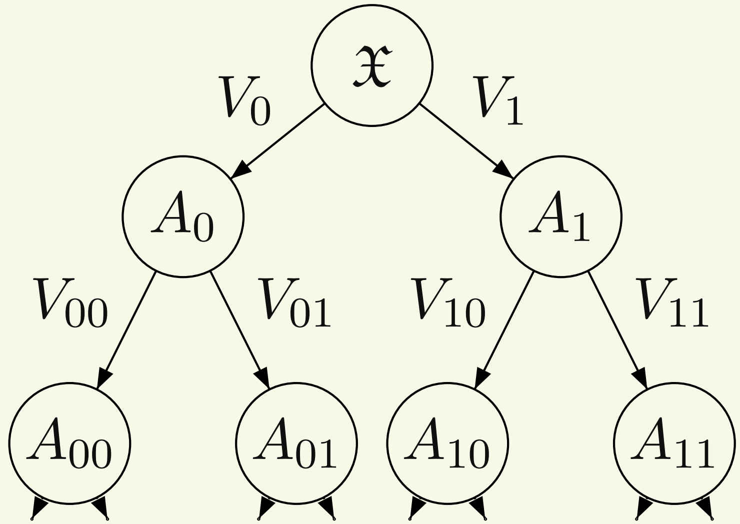
Tail-free processes

Let $\mathfrak{X} = A_0 \cup A_1 = (A_{00} \cup A_{01}) \cup (A_{10} \cup A_{11}) = \cdots$ be nested partitions, rich enough that they generates the Borel σ-field.



Tail-free processes

Let $\mathfrak{X} = A_0 \cup A_1 = (A_{00} \cup A_{01}) \cup (A_{10} \cup A_{11}) = \cdots$ be nested partitions, rich enough that they generates the Borel σ-field.



Splitting variables:

$$V_{\varepsilon 0} = P(A_{\varepsilon 0} | \, A_\varepsilon), \qquad \text{and} \qquad V_{\varepsilon 1} = P(A_{\varepsilon 1} | \, A_\varepsilon).$$

$$P(A_{\varepsilon_1 \cdots \varepsilon_m}) = V_{\varepsilon_1} V_{\varepsilon_1 \varepsilon_2} \cdots V_{\varepsilon_1 \cdots \varepsilon_m}, \qquad \varepsilon = \varepsilon_1 \cdots \varepsilon_m \in \{0, 1\}^m.$$

Tail-free processes (2)

$$P(A_{\varepsilon_1\cdots\varepsilon_m}) = V_{\varepsilon_1}V_{\varepsilon_1\varepsilon_2}\cdots V_{\varepsilon_1\cdots\varepsilon_m}, \qquad \varepsilon = \varepsilon_1\cdots\varepsilon_m \in \{0,1\}^m. \tag{1}$$

Theorem. *Suppose*

- $A_\varepsilon = \cup\{A_{\varepsilon\delta}\colon \overline{A}_{\varepsilon\delta}\text{compact}, \overline{A}_{\varepsilon\delta} \subset A_\varepsilon\}$
- $(V_\varepsilon\colon \varepsilon \in \mathcal{E}^*)$ *stochastic process with* $0 \le V_\varepsilon \le 1$ *and* $V_{\varepsilon 0} + V_{\varepsilon 1} = 1$.
- *There is a Borel measure with* $\mu(A_\varepsilon)\colon= \mathrm{E}V_{\varepsilon_1}V_{\varepsilon_1\varepsilon_2}\cdots V_{\varepsilon_1\cdots\varepsilon_m}$.

Then there exists a random Borel measure P *satisfying (1)*

Tail-free processes (2)

$$P(A_{\varepsilon_1\cdots\varepsilon_m}) = V_{\varepsilon_1}V_{\varepsilon_1\varepsilon_2}\cdots V_{\varepsilon_1\cdots\varepsilon_m}, \qquad \varepsilon=\varepsilon_1\cdots\varepsilon_m\in\{0,1\}^m. \qquad (1)$$

Theorem. *Suppose*

- $A_\varepsilon = \cup\{A_{\varepsilon\delta}: \overline{A}_{\varepsilon\delta}$*compact*$, \overline{A}_{\varepsilon\delta}\subset A_\varepsilon\}$
- $(V_\varepsilon: \varepsilon\in\mathcal{E}^*)$ *stochastic process with* $0\le V_\varepsilon\le 1$ *and* $V_{\varepsilon 0}+V_{\varepsilon 1}=1$*.*
- *There is a Borel measure with* $\mu(A_\varepsilon):=\mathrm{E}V_{\varepsilon_1}V_{\varepsilon_1\varepsilon_2}\cdots V_{\varepsilon_1\cdots\varepsilon_m}$*.*

Then there exists a random Borel measure P *satisfying (1)*

SPECIAL CASE: *Polya tree prior*: all V_ε independent Beta variables.

Tail-free processes (3)

$$V_{\varepsilon 0} = P(A_{\varepsilon 0} |\, A_\varepsilon), \qquad \text{and} \qquad V_{\varepsilon 1} = P(A_{\varepsilon 1} |\, A_\varepsilon).$$

$$P(A_{\varepsilon_1 \cdots \varepsilon_m}) = V_{\varepsilon_1} V_{\varepsilon_1 \varepsilon_2} \cdots V_{\varepsilon_1 \cdots \varepsilon_m}, \qquad \varepsilon = \varepsilon_1 \cdots \varepsilon_m \in \{0,1\}^m.$$

Tail-free processes (3)

$$V_{\varepsilon 0} = P(A_{\varepsilon 0} |\, A_\varepsilon), \qquad \text{and} \qquad V_{\varepsilon 1} = P(A_{\varepsilon 1} |\, A_\varepsilon).$$

$$P(A_{\varepsilon_1 \cdots \varepsilon_m}) = V_{\varepsilon_1} V_{\varepsilon_1 \varepsilon_2} \cdots V_{\varepsilon_1 \cdots \varepsilon_m}, \qquad \varepsilon = \varepsilon_1 \cdots \varepsilon_m \in \{0,1\}^m.$$

Notation:

$U \perp V$ means “U and V are independent”

$U \perp V |\, Z$ means “U and V are conditionally independent given Z”.

Tail-free processes (3)

$$V_{\varepsilon 0} = P(A_{\varepsilon 0} \mid A_\varepsilon), \qquad \text{and} \qquad V_{\varepsilon 1} = P(A_{\varepsilon 1} \mid A_\varepsilon).$$

$$P(A_{\varepsilon_1 \cdots \varepsilon_m}) = V_{\varepsilon_1} V_{\varepsilon_1 \varepsilon_2} \cdots V_{\varepsilon_1 \cdots \varepsilon_m}, \qquad \varepsilon = \varepsilon_1 \cdots \varepsilon_m \in \{0,1\}^m.$$

Notation:
$U \perp V$ means "U and V are independent"
$U \perp V \mid Z$ means "U and V are conditionally independent given Z".

Definition (Tail-free)**.** The random measure P is a *tail-free process* with respect to the sequence of partitions if
$\{V_0\} \perp \{V_{00}, V_{10}\} \perp \cdots \perp \{V_{\varepsilon 0} : \varepsilon \in \mathcal{E}^m\} \perp \cdots$.

Tail-free processes (3)

$$V_{\varepsilon 0} = P(A_{\varepsilon 0} |\, A_\varepsilon), \qquad \text{and} \qquad V_{\varepsilon 1} = P(A_{\varepsilon 1} |\, A_\varepsilon).$$

$$P(A_{\varepsilon_1 \cdots \varepsilon_m}) = V_{\varepsilon_1} V_{\varepsilon_1 \varepsilon_2} \cdots V_{\varepsilon_1 \cdots \varepsilon_m}, \qquad \varepsilon = \varepsilon_1 \cdots \varepsilon_m \in \{0,1\}^m.$$

Notation:
$U \perp V$ means "U and V are independent"
$U \perp V |\, Z$ means "U and V are conditionally independent given Z".

Definition (Tail-free)**.** The random measure P is a *tail-free process* with respect to the sequence of partitions if
$\{V_0\} \perp \{V_{00}, V_{10}\} \perp \cdots \perp \{V_{\varepsilon 0} : \varepsilon \in \mathcal{E}^m\} \perp \cdots$.

Theorem. *The* $\mathrm{DP}(\alpha)$ *prior is tail free. All splitting variables* $V_{\varepsilon 0}$ *are independent and* $V_{\varepsilon 0} \sim \mathrm{Be}\big(\alpha(A_{\varepsilon 0}), \alpha(A_{\varepsilon 1})\big)$.

Proof. This follows from properties of the finite-dimensional Dirichlet. ∎

Posterior distribution

For $X_1, \ldots, X_n |\, P \overset{\text{iid}}{\sim} P$ define count variables:

$$N_\varepsilon := \#\{1 \leq i \leq n \colon X_i \in A_\varepsilon\}.$$

Posterior distribution

For $X_1, \ldots, X_n | \, P \stackrel{\text{iid}}{\sim} P$ define count variables:

$$N_\varepsilon \colon = \#\{1 \le i \le n \colon X_i \in A_\varepsilon\}.$$

Theorem. *If P is tail-free, then for every m and n the posterior distribution of $\big(P(A_\varepsilon) \colon \varepsilon \in \mathcal{E}^m\big)$ given $X_1, \ldots, X_n$ depends only on $(N_\varepsilon \colon \varepsilon \in \mathcal{E}^m)$.*

Posterior distribution

For $X_1,\ldots,X_n|\,P \overset{\text{iid}}{\sim} P$ define count variables:

$$N_\varepsilon{:}= \#\{1\le i\le n{:}\, X_i\in A_\varepsilon\}.$$

Theorem. *If P is tail-free, then for every m and n the posterior distribution of $\bigl(P(A_\varepsilon){:}\,\varepsilon\in\mathcal{E}^m\bigr)$ given $X_1,\ldots,X_n$ depends only on $(N_\varepsilon{:}\,\varepsilon\in\mathcal{E}^m)$.*

Proof. We may generate the variables $P, X_1,\ldots,X_n$ in four steps:

(a) Generate $\theta{:}= \bigl(P(A_\varepsilon){:}\,\varepsilon\in\mathcal{E}^m\bigr)$ from its prior.
(b) Given θ generate $N=(N_\varepsilon{:}\,\varepsilon\in\mathcal{E}^m)$ multinomial (n,θ).
(c) Generate $\eta{:}= \bigl(P(A|\,A_\varepsilon){:}\,A\in\mathscr{X},\varepsilon\in\mathcal{E}^m\bigr)$.
(d) Given (N,η) generate for every $\varepsilon\in\mathcal{E}^m$ a random sample of size N_ε from $P(\cdot|\,A_\varepsilon)$, independently across $\varepsilon\in\mathcal{E}^m$; let $X_1,\ldots,X_n$ be the n values in a random order.

Then $\eta\perp\theta$ and $N\perp\eta|\,\theta$ and $X\perp\theta|\,(N,\eta)$.
Thus $\theta\perp X|\,N$. $\square$

Posterior distribution (continued)

Theorem. *If P is tail-free, then the posterior $P|\, X_1, \ldots, X_n$ is tail-free.*

Posterior distribution (continued)

Theorem. *If P is tail-free, then the posterior $P|\,X_1,\ldots,X_n$ is tail-free.*

Proof. Suffices to show, for every level:

$$(V_{\varepsilon 0}\colon \varepsilon\in\mathcal{E}^m) \perp \bigl(P(A_\varepsilon)\colon \varepsilon\in\mathcal{E}^m\bigr)|\,X_1,\ldots,X_n.$$

In view of preceding theorem, suffices:

$$(V_{\varepsilon 0}\colon \varepsilon\in\mathcal{E}^m) \perp \bigl(P(A_\varepsilon)\colon \varepsilon\in\mathcal{E}^m\bigr)|\,(N_{\varepsilon\delta}\colon \varepsilon\in\mathcal{E}^m,\delta\in\mathcal{E}).$$

The likelihood for (V,θ,N), where $\theta_\varepsilon=P(A_\varepsilon)$, takes the form

$$\binom{n}{N}\prod_{\varepsilon\in\mathcal{E}^m,\delta\in\mathcal{E}}(\theta_\varepsilon V_{\varepsilon\delta})^{N_{\varepsilon\delta}}\,d\Pi_1(V)\,d\Pi_2(\theta).$$

This factorizes in parts involving (V,N) and involving (θ,N). $\square$

Conjugacy of Dirichlet process

$$P \sim \mathrm{DP}(\alpha), \qquad X_1, X_2, \ldots \mid P \overset{\text{iid}}{\sim} P.$$

Conjugacy of Dirichlet process

$$P \sim \mathrm{DP}(\alpha), \qquad\qquad X_1, X_2, \ldots \,|\, P \overset{\text{iid}}{\sim} P.$$

Theorem. $P|\, X_1, \ldots, X_n \sim \mathrm{DP}(\alpha + n\mathbb{P}_n)$, *for* $\mathbb{P}_n = n^{-1} \sum_{i=1}^n \delta_{X_i}$.

Conjugacy of Dirichlet process

$$P \sim \mathrm{DP}(\alpha), \qquad\qquad X_1, X_2, \ldots \,|\, P \stackrel{\text{iid}}{\sim} P.$$

Theorem. $P|\,X_1, \ldots, X_n \sim \mathrm{DP}(\alpha + n\mathbb{P}_n)$, *for* $\mathbb{P}_n = n^{-1}\sum_{i=1}^n \delta_{X_i}$.

Proof. $\big(P(A_1), \ldots, P(A_k)\big)|\,X_1, \ldots, X_n \sim \big(P(A_1), \ldots, P(A_k)\big)|\,N$.
Apply result for finite-dimensional Dirichlet. $\square$

Conjugacy of Dirichlet process

$$P \sim \mathrm{DP}(\alpha), \qquad\qquad X_1, X_2, \ldots \,|\, P \overset{\text{iid}}{\sim} P.$$

Theorem. $P|\,X_1,\ldots,X_n \sim \mathrm{DP}(\alpha + n\mathbb{P}_n)$, *for* $\mathbb{P}_n = n^{-1}\sum_{i=1}^n \delta_{X_i}$.

Proof. $\big(P(A_1),\ldots,P(A_k)\big)|\,X_1,\ldots,X_n \sim \big(P(A_1),\ldots,P(A_k)\big)|\,N$.
Apply result for finite-dimensional Dirichlet. $\square$

$$\mathrm{E}\big(P(A)|\,X_1,\ldots,X_n\big) = \frac{|\alpha|}{|\alpha|+n}\bar{\alpha}(A) + \frac{n}{|\alpha|+n}\mathbb{P}_n(A),$$

$$\mathrm{var}\big(P(A)|\,X_1,\ldots,X_n\big) = \frac{\tilde{\mathbb{P}}_n(A)\tilde{\mathbb{P}}_n(A^c)}{1+|\alpha|+n} \leq \frac{1}{4(1+|\alpha|+n)}.$$

Corollary. $P(A)|\,X_1,\ldots,X_n \rightarrow_d \delta_{P_0(A)}$ *as* $n \to \infty$, *a.s.* $[P_0^\infty]$.

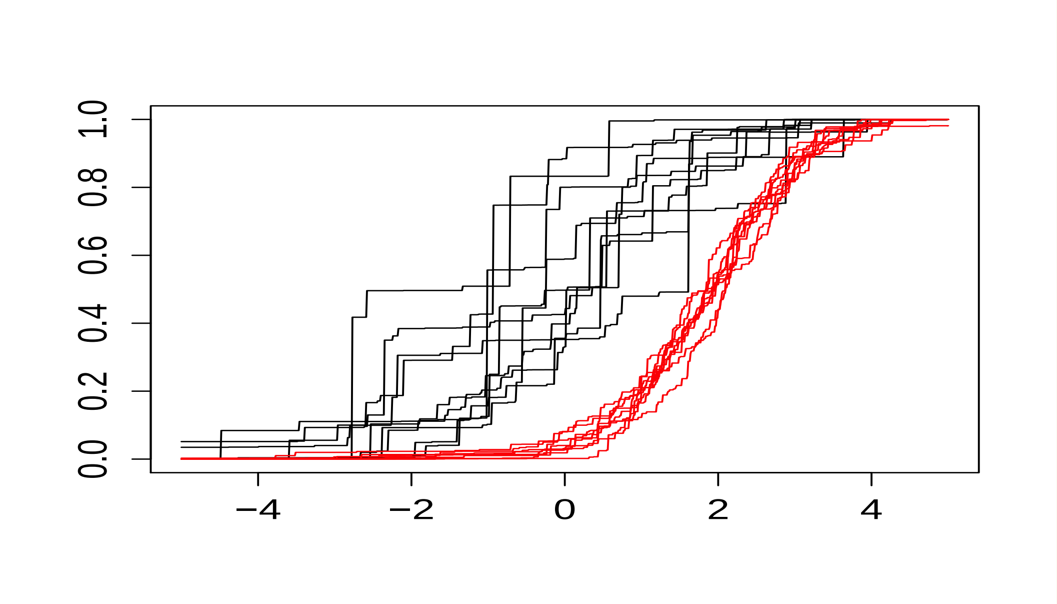
Conjugacy of Dirichlet process

$$P \sim \mathrm{DP}(\alpha), \qquad\qquad X_1, X_2, \ldots \,|\, P \overset{\text{iid}}{\sim} P.$$

Theorem. $P|\, X_1, \ldots, X_n \sim \mathrm{DP}(\alpha + n\mathbb{P}_n)$, *for* $\mathbb{P}_n = n^{-1} \sum_{i=1}^n \delta_{X_i}$.

Proof. $\big(P(A_1), \ldots, P(A_k)\big)|\, X_1, \ldots, X_n \sim \big(P(A_1), \ldots, P(A_k)\big)|\, N$.
Apply result for finite-dimensional Dirichlet. $\square$

Corollary. $P(A)|\, X_1, \ldots, X_n \rightarrow_d \delta_{P_0(A)}$ *as* $n \to \infty$, *a.s.* $[P_0^\infty]$.



Predictive distribution

$$P \sim \mathrm{DP}(\alpha), \qquad\qquad X_1, X_2, \ldots \,|\, P \overset{\text{iid}}{\sim} P.$$

Theorem.

$$X_i | \, X_1, \ldots, X_{i-1} \sim \begin{cases} \delta_{X_1}, & \textit{with probability } \frac{1}{|\alpha|+i-1}, \\ \vdots & \qquad\qquad \vdots \\ \delta_{X_{i-1}}, & \textit{with probability } \frac{1}{|\alpha|+i-1}, \\ \bar{\alpha}, & \textit{with probability } \frac{|\alpha|}{|\alpha|+i-1}. \end{cases}$$

Predictive distribution

$$P \sim \mathrm{DP}(\alpha), \qquad\qquad X_1, X_2, \ldots \,|\, P \overset{\text{iid}}{\sim} P.$$

Theorem.

$$X_i \,|\, X_1, \ldots, X_{i-1} \sim \begin{cases} \delta_{X_1}, & \textit{with probability } \frac{1}{|\alpha|+i-1}, \\ \vdots & \qquad\qquad \vdots \\ \delta_{X_{i-1}}, & \textit{with probability } \frac{1}{|\alpha|+i-1}, \\ \bar{\alpha}, & \textit{with probability } \frac{|\alpha|}{|\alpha|+i-1}. \end{cases}$$

Proof.

(i). $\Pr(X_1 \in A) = \mathrm{E}\Pr(X_1 \in A|\, P) = \mathrm{E}P(A) = \bar{\alpha}(A)$.

(ii). Preceding step means: $X_1|\, P \sim P$ and $P - \mathrm{DP}(\alpha)$ imply $X_1 \sim \bar{\alpha}$. Hence $X_2|\,(P, X_1) \sim P$ and $P|\, X_1 \sim \mathrm{DP}(\alpha + \delta_{X_1})$ imply $X_2|\, X_1 \sim (\alpha + \delta_{X_1})/(|\alpha| + 1)$.

(iii). etc.

$\square$

Dirichlet process mixtures

Given a probability density $x \mapsto \psi(x;\theta)$ consider data

$$X_1, \ldots, X_n | \, F \overset{\text{iid}}{\sim} p_F(x) := \int \psi(x;\theta)\, dF(\theta).$$

Dirichlet process mixtures

Given a probability density $x \mapsto \psi(x;\theta)$ consider data

$$X_1,\ldots,X_n|\,F \overset{\text{iid}}{\sim} p_F(x) := \int \psi(x;\theta)\,dF(\theta).$$

For $F \sim \mathrm{DP}(\alpha)$, this gives Bayesian model:

$$X_1,\ldots,X_n|\,\theta_1,\ldots,\theta_n, F \overset{\text{ind}}{\sim} \psi(\cdot;\theta_i), \qquad \theta_1,\ldots,\theta_n|\,F \overset{\text{iid}}{\sim} F, \qquad F \sim \mathrm{DP}(\alpha).$$

Dirichlet process mixtures

Given a probability density $x \mapsto \psi(x;\theta)$ consider data

$$X_1,\ldots,X_n|\,F \overset{\text{iid}}{\sim} p_F(x) := \int \psi(x;\theta)\,dF(\theta).$$

For $F \sim \mathrm{DP}(\alpha)$, this gives Bayesian model:

$$X_1,\ldots,X_n|\,\theta_1,\ldots,\theta_n, F \overset{\text{ind}}{\sim} \psi(\cdot;\theta_i), \qquad \theta_1,\ldots,\theta_n|\,F \overset{\text{iid}}{\sim} F, \qquad F \sim \mathrm{DP}(\alpha).$$

Lemma. *For any $\theta \mapsto \psi(\theta)$ (e.g. $\psi(x,\cdot)$),*

$$\mathrm{E}\Big(\int \psi\,dF|\,\theta_1,\ldots,\theta_n, X_1,\ldots,X_n\Big) = \frac{1}{|\alpha|+n}\Big[\int \psi\,d\alpha + \sum_{j=1}^n \psi(\theta_j)\Big].$$

Dirichlet process mixtures

Given a probability density $x \mapsto \psi(x;\theta)$ consider data

$$X_1,\ldots,X_n|\,F \overset{\text{iid}}{\sim} p_F(x) := \int \psi(x;\theta)\,dF(\theta).$$

For $F \sim \mathrm{DP}(\alpha)$, this gives Bayesian model:

$$X_1,\ldots,X_n|\,\theta_1,\ldots,\theta_n,F \overset{\text{ind}}{\sim} \psi(\cdot;\theta_i), \qquad \theta_1,\ldots,\theta_n|\,F \overset{\text{iid}}{\sim} F, \qquad F \sim \mathrm{DP}(\alpha).$$

Lemma. *For any* $\theta \mapsto \psi(\theta)$ *(e.g.* $\psi(x,\cdot)$*),*

$$\mathrm{E}\Big(\int \psi\,dF|\,\theta_1,\ldots,\theta_n,X_1,\ldots,X_n\Big) = \frac{1}{|\alpha|+n}\Big[\int \psi\,d\alpha + \sum_{j=1}^n \psi(\theta_j)\Big].$$

Proof. $F \perp X_1,\ldots,X_n|\,\theta_1,\ldots,\theta_n; \quad F|\,\theta_1,\ldots,\theta_n \sim \mathrm{DP}(\alpha + \sum_{i=1}^n \delta_{\theta_i})$. $\square$

Dirichlet process mixtures

Given a probability density $x \mapsto \psi(x;\theta)$ consider data

$$X_1,\ldots,X_n|\,F \overset{\text{iid}}{\sim} p_F(x){:=}\int \psi(x;\theta)\,dF(\theta).$$

For $F \sim \mathrm{DP}(\alpha)$, this gives Bayesian model:

$$X_1,\ldots,X_n|\,\theta_1,\ldots,\theta_n, F \overset{\text{ind}}{\sim} \psi(\cdot;\theta_i), \qquad \theta_1,\ldots,\theta_n|\,F \overset{\text{iid}}{\sim} F, \qquad F \sim \mathrm{DP}(\alpha).$$

Lemma. *For any $\theta \mapsto \psi(\theta)$ (e.g. $\psi(x,\cdot)$),*

$$\mathrm{E}\Big(\int \psi\,dF|\,\theta_1,\ldots,\theta_n, X_1,\ldots,X_n\Big) = \frac{1}{|\alpha|+n}\Big[\int \psi\,d\alpha + \sum_{j=1}^n \psi(\theta_j)\Big].$$

Proof. $F \perp X_1,\ldots,X_n|\,\theta_1,\ldots,\theta_n; \quad F|\,\theta_1,\ldots,\theta_n \sim \mathrm{DP}(\alpha+\sum_{i=1}^n \delta_{\theta_i})$. $\quad\square$

Compute conditional expectation given $X_1,\ldots,X_n$ by generating samples $\theta_1,\ldots,\theta_n$ from $\theta_1,\ldots,\theta_n|\,X_1,\ldots,X_n$, and averaging.

Gibbs sampler

$$X_i|\,\theta_i, F \overset{\text{ind}}{\sim} \psi(\cdot;\theta_i), \qquad \theta_i|\,F \overset{\text{iid}}{\sim} F, \qquad F \sim \mathrm{DP}(\alpha).$$

Theorem (Gibbs sampler)**.**

$$\theta_i|\,\theta_{-i} X_1, \ldots, X_n \sim \sum_{j\neq i} q_{i,j}\delta_{\theta_j} + q_{i,0} G_{b,i},$$

where $(q_{i,j}\colon j \in \{0,1,\ldots,n\} - \{i\})$ *is the probability vector satisfying*

$$q_{i,j} \propto \begin{cases} \psi(X_i;\theta_j), & j \neq i, j \geq 1, \\ \int \psi(X_i;\theta)\, d\alpha(\theta), & j = 0, \end{cases}$$

and $G_{b,i}$ *is the “baseline posterior measure” given by*

$$dG_{b,i}(\theta|\,X_i) \propto \psi(X_i;\theta)\, d\alpha(\theta).$$

Gibbs sampler — proof

Proof.

$$
\begin{aligned}
&\mathrm{E}\big(1\!\!1_A(X_i)1\!\!1_B(\theta_i)|\,\theta_{-i},X_{-i}\big)\\
&\qquad= \mathrm{E}\Big(\mathrm{E}\big(1\!\!1_A(X_i)1\!\!1_B(\theta_i)|\,F,\theta_{-i},X_{-i}\big)|\,\theta_{-i},X_{-i}\Big)\\
&\qquad= \mathrm{E}\Big(\int\!\!\int 1\!\!1_A(x)1\!\!1_B(\theta)\psi(x;\theta)\,d\mu(x)dF(\theta)|\,\theta_{-i}\Big)\\
&\qquad= \frac{1}{|\alpha|+n}\int\!\!\int 1\!\!1_A(x)1\!\!1_B(\theta)\psi(x;\theta)\,d\mu(x)\,d\Big(\alpha+\sum_{j\neq i}\delta_{\theta_j}\Big)(\theta).
\end{aligned}
$$

By Bayes's rule (applied conditionally given (θ_{-i},X_{-i}))

$$
\Pr\big(\theta_i\in B|\,X_i,\theta_{-i},X_{-i}\big)=\frac{\int_B \psi(X_i;\theta)\,d(\alpha+\sum_{j\neq i}\delta_{\theta_j})(\theta)}{\int \psi(X_i;\theta)\,d(\alpha+\sum_{j\neq i}\delta_{\theta_j})(\theta)}.
$$

$\square$

Further properties

- The number of distinct values in $(X_1, \ldots, X_n)$ is $O_P(\log n)$.
- The pattern of equal values induces the same random partition of the set $\{1, 2, \ldots, n\}$ as the *Kingman coalescent*.
- The Dirichlet distribution has full support relative to the weak topology.
- $\mathrm{DP}(\alpha_1) \perp \mathrm{DP}(\alpha_2)$ as soon as $\alpha_1^c \neq \alpha_2^c$ or α_1^d and α_1^d have different supports.
- In particular prior $\mathrm{DP}(\alpha)$ and posterior $\mathrm{DP}(\alpha + n\mathbb{P}_n)$ are typically orthogonal.
- The cdf of $P \sim DP(\alpha)$ is a normalized Gamma process.
- The tails of $P \sim DP(\alpha)$ are much thinner than the tails of α.
- The Dirichlet is the only prior that is tail-free relative to *any* partition.
- The splitting variables of a Polya tree can be defined so that the prior is absolutely continuous.

Consistency and rates

Consistency

$X^{(n)}$ observation in sample space $(\mathfrak{X}^{(n)}, \mathscr{X}^{(n)})$ with distribution $P_\theta^{(n)}$.
θ belongs to metric space (Θ, d).

Definition. The posterior distribution is *consistent* at $\theta_0 \in \Theta$ if

$$\Pi_n\big(\theta: d(\theta, \theta_0) > \epsilon |\, X^{(n)}\big) \to 0$$

in $P_{\theta_0}^{(n)}$-probability, as $n \to \infty$, for every $\epsilon > 0$.

Point estimator

Proposition. *If the posterior distribution is consistent at* θ_0 *then* $\hat{\theta}_n$ *defined as the center of a (nearly) smallest ball that contains posterior mass at least* $1/2$ *satisfies* $d(\hat{\theta}_n, \theta_0) \to 0$ *in* $P_{\theta_0}^{(n)}$*-probability.*

Point estimator

Proposition. *If the posterior distribution is consistent at* θ_0 *then* $\hat\theta_n$ *defined as the center of a (nearly) smallest ball that contains posterior mass at least* $1/2$ *satisfies* $d(\hat\theta_n, \theta_0) \to 0$ *in* $P_{\theta_0}^{(n)}$*-probability.*

Proof. For $B(\theta, r) = \{s \in \Theta \colon d(s, \theta) \le r\}$ let

$$\hat r_n(\theta) = \inf\{r \colon \Pi_n\big(B(\theta, r) |\, X^{(n)}\big) \ge 1/2\}.$$

Then $\hat r_n(\hat\theta_n) \le \inf_\theta \hat r_n(\theta)$.

- $\Pi_n\big(B(\theta_0, \epsilon) |\, X^{(n)}\big) \to 1$ in probability.
- $\hat r_n(\theta_0) \le \epsilon$ with probability tending to 1, whence $\hat r_n(\hat\theta_n) \le \hat r_n(\theta_0) \le \epsilon$.
- $B(\theta_0, \epsilon)$ and $B\big(\hat\theta_n, \hat r_n(\hat\theta_n)\big)$ cannot be disjoint.
- $d(\theta_0, \hat\theta_n) \le \epsilon + \hat r_n(\hat\theta_n) \le 2\epsilon$.

$\square$

Point estimator

Proposition. *If the posterior distribution is consistent at* θ_0 *then* $\hat\theta_n$ *defined as the center of a (nearly) smallest ball that contains posterior mass at least* $1/2$ *satisfies* $d(\hat\theta_n,\theta_0)\to 0$ *in* $P_{\theta_0}^{(n)}$*-probability.*

Proof. For $B(\theta,r)=\{s\in\Theta\colon d(s,\theta)\le r\}$ let

$$\hat r_n(\theta)=\inf\{r\colon \Pi_n\bigl(B(\theta,r)|\,X^{(n)}\bigr)\ge 1/2\}.$$

Then $\hat r_n(\hat\theta_n)\le \inf_\theta \hat r_n(\theta)$.

- $\Pi_n\bigl(B(\theta_0,\epsilon)|\,X^{(n)}\bigr)\to 1$ in probability.
- $\hat r_n(\theta_0)\le\epsilon$ with probability tending to 1, whence $\hat r_n(\hat\theta_n)\le \hat r_n(\theta_0)\le\epsilon$.
- $B(\theta_0,\epsilon)$ and $B\bigl(\hat\theta_n,\hat r_n(\hat\theta_n)\bigr)$ cannot be disjoint.
- $d(\theta_0,\hat\theta_n)\le\epsilon+\hat r_n(\hat\theta_n)\le 2\epsilon$.

$\square$

Alternative: posterior mean $\int\theta\,d\Pi_n(\theta|\,X^{(n)})$.

Doob's theorem

Theorem (Doob). *Let $(\mathfrak{X}, \mathscr{X}, P_\theta\colon \theta \in \Theta)$ be experiments with $(\mathfrak{X}, \mathscr{X})$ a standard Borel space and Θ a Borel subset of a Polish space such that $\theta \mapsto P_\theta(A)$ is Borel measurable for every $A \in \mathscr{X}$ and the map $\theta \mapsto P_\theta$ is one-to-one. Then for any prior Π on the Borel sets of Θ the posterior $\Pi_n(\cdot \,|\, X_1, \ldots, X_n)$ in the model $X_1, \ldots, X_n \,|\, \theta \overset{iid}{\sim} p_\theta$ and $\theta \sim \Pi$ is consistent at θ, for Π-almost every θ.*

Kullback-Leibler property

Parameter p: ν-density on sample space $(\mathfrak{X}, \mathscr{X})$. True value p_0.
Kullback-Leibler divergence:

$$K(p_0; p) = \int p_0 \log(p_0/p)\, d\nu, \qquad K(p_0; \mathcal{P}_0) = \inf_{p \in \mathcal{P}_0} K(p_0; p).$$

Kullback-Leibler property

Parameter p: ν-density on sample space $(\mathfrak{X}, \mathscr{X})$. True value p_0.
Kullback-Leibler divergence:

$$K(p_0; p) = \int p_0 \log(p_0/p)\, d\nu, \qquad K(p_0; \mathcal{P}_0) = \inf_{p \in \mathcal{P}_0} K(p_0; p).$$

Definition. p_0 is said to possess the *Kullback-Leibler property* relative to Π if $\Pi\big(p\colon K(p_0; p) < \epsilon\big) > 0$ for every $\epsilon > 0$.

Kullback-Leibler property

Parameter p: ν-density on sample space $(\mathfrak{X}, \mathscr{X})$. True value p_0.
Kullback-Leibler divergence:

$$K(p_0; p) = \int p_0 \log(p_0/p)\, d\nu, \qquad K(p_0; \mathcal{P}_0) = \inf_{p \in \mathcal{P}_0} K(p_0; p).$$

Definition. p_0 is said to possess the *Kullback-Leibler property* relative to Π if $\Pi\big(p\colon K(p_0; p) < \epsilon\big) > 0$ for every $\epsilon > 0$.

EXAMPLES

- Polya tree prior with dyadic partition and splitting variables $V_{\varepsilon 0} \sim \mathrm{Be}(a_{|\varepsilon|}, a_{|\varepsilon|})$ for $\sum_m a_m^{-1} < \infty$ and $K(p_0, \lambda) < \infty$.
- Dirichlet mixtures $\int \psi(\cdot, \theta)\, dF(\theta)$ with $F \sim \mathrm{DP}(\alpha)$, under some regularity conditions.

Schwartz's theorem

Bayesian model:

$$X_1, \ldots, X_n | \, p \overset{\text{iid}}{\sim} p, \qquad p \sim \Pi.$$

Schwartz's theorem

Bayesian model:

$$X_1, \ldots, X_n | \, p \overset{\text{iid}}{\sim} p, \qquad p \sim \Pi.$$

Theorem. *If p_0 has KL-property, and for every neighbourhood $\mathcal{U}$ of p_0 there exist tests ϕ_n such that*

$$P_0^n \phi_n \to 0, \qquad \sup_{p \in \mathcal{U}^c} P^n(1 - \phi_n) \to 0,$$

then $\Pi_n(\cdot | \, X_1, \ldots, X_n)$ is consistent at p_0.

Schwartz's theorem

Bayesian model:

$$X_1,\ldots,X_n|\,p \stackrel{\text{iid}}{\sim} p, \qquad p\sim\Pi.$$

Theorem. *If p_0 has KL-property, and for every neighbourhood $\mathcal{U}$ of p_0 there exist tests ϕ_n such that*

$$P_0^n\phi_n\to 0, \qquad \sup_{p\in\mathcal{U}^c} P^n(1-\phi_n)\to 0,$$

then $\Pi_n(\cdot|\,X_1,\ldots,X_n)$ is consistent at p_0.

Proof. By grouping the observations and using Hoeffding's inequality we can find tests ψ_n with

$$P_0^n\psi_n\le e^{-Cn}, \qquad \sup_{p\in\mathcal{U}^c} P^n(1-\psi_n)\le e^{-Cn}.$$

Then apply the theorem later on. $\square$

Weak consistency

Consider the topology induced on p by the weak topology on the probability measures P.

Weak consistency

Consider the topology induced on p by the weak topology on the probability measures P.

Theorem. *The posterior distribution is consistent for the weak topology at any p_0 with the Kullback-Leibler property.*

Weak consistency

Consider the topology induced on p by the weak topology on the probability measures P.

Theorem. *The posterior distribution is consistent for the weak topology at any p_0 with the Kullback-Leibler property.*

Proof. Consistent tests always exist:

- Subbasis for the weak neighbourhoods are sets of the type $\mathcal{U} = \bigl\{p: P\psi < P_0\psi + \epsilon\bigr\}$, for $\psi: \mathfrak{X} \to [0, 1]$ continuous and $\epsilon > 0$.
- Given a test for each neighbourhood the maximum of the tests works for a finite intersection.
- Use Hoeffding's inequality to bound the error probabilities of the test

$$\phi_n = \mathbb{1}\Bigl\{\frac{1}{n}\sum_{i=1}^{n}\psi(X_i) > P_0\psi + \epsilon/2\Bigr\}.$$

$\square$

Extended Schwartz's theorem

Bayesian model:

$$X_1, \ldots, X_n \,|\, p \overset{\text{iid}}{\sim} p, \qquad p \sim \Pi.$$

Theorem. *If If p_0 has KL-property and for every neighbourhood $\mathcal{U}$ of p_0 there exist $C > 0$, sets $\mathcal{P}_n \subset \mathcal{P}$ and tests ϕ_n such that*

$$\Pi(\mathcal{P} - \mathcal{P}_n) < e^{-Cn}, \qquad P_0^n \phi_n \leq e^{-Cn}, \qquad \sup_{p \in \mathcal{P}_n \cap \mathcal{U}^c} P^n(1 - \phi_n) \leq e^{-Cn},$$

then the posterior distribution $\Pi_n(\cdot \,|\, X_1, \ldots, X_n)$ is consistent at p_0.

Extended Schwartz's theorem

Bayesian model:

$$X_1, \ldots, X_n |\, p \overset{\text{iid}}{\sim} p, \qquad p \sim \Pi.$$

Theorem. *If If p_0 has KL-property and for every neighbourhood $\mathcal{U}$ of p_0 there exist $C > 0$, sets $\mathcal{P}_n \subset \mathcal{P}$ and tests ϕ_n such that*

$$\Pi(\mathcal{P} - \mathcal{P}_n) < e^{-Cn}, \qquad P_0^n \phi_n \le e^{-Cn}, \qquad \sup_{p \in \mathcal{P}_n \cap \mathcal{U}^c} P^n(1 - \phi_n) \le e^{-Cn},$$

then the posterior distribution $\Pi_n(\cdot|\, X_1, \ldots, X_n)$ is consistent at p_0.

Proof.

$$\Pi_n(\mathcal{U}^c) = \frac{\int_{\mathcal{U}^c} \prod_{i=1}^n (p/p_0)(X_i)\, d\Pi(p)}{\int \prod_{i=1}^n (p/p_0)(X_i)\, d\Pi(p)}.$$

Follow steps 1–4. $\square$

Extended Schwartz's theorem — proof

Proof. continued.

- Step 1: for any $\epsilon > 0$ eventually a.s. $[P_0^\infty]$:

$$\int \prod_{i=1}^{n} \frac{p}{p_0}(X_i)\, d\Pi(p) \geq \Pi\big(p\colon K(p_0; p) < \epsilon\big) e^{-n\epsilon}. \tag{2}$$

Extended Schwartz's theorem — proof

Proof. continued.

- Step 1: for any $\epsilon > 0$ eventually a.s. $[P_0^\infty]$:

$$\int \prod_{i=1}^n \frac{p}{p_0}(X_i)\, d\Pi(p) \geq \Pi\big(p\colon K(p_0;p) < \epsilon\big) e^{-n\epsilon}. \qquad (2)$$

Proof: for $\Pi_\epsilon(\cdot) = \Pi(\cdot \cap \mathcal{P}_\epsilon)/\Pi(\mathcal{P}_\epsilon)$, and $\mathcal{P}_\epsilon = \{p\colon K(p_0;p) < \epsilon\}$,

$$\begin{aligned}
&\log \int_{\mathcal{P}_\epsilon} \prod_{i=1}^n \frac{p}{p_0}(X_i)\, d\Pi(p) - \log \Pi(\mathcal{P}_\epsilon) \\
&= \log \int \prod_{i=1}^n \frac{p}{p_0}(X_i)\, d\Pi_\epsilon(p) \geq \int \log \prod_{i=1}^n \frac{p}{p_0}(X_i)\, d\Pi_\epsilon(p), \\
&= \sum_{i=1}^n \int \log \frac{p}{p_0}(X_i)\, d\Pi_\epsilon(p) = -n \int K(p_0;p)\, d\Pi_\epsilon(p) + o(n), \qquad a.s.
\end{aligned}$$

$\square$

Extended Schwartz's theorem — proof (2)

Proof. continued.

- Step 2:

$$\Pi_n(\mathcal{U}^c |\, X_1, \ldots, X_n) \le \phi_n + (1-\phi_n) \frac{\int_{\mathcal{U}^c} \prod_{i=1}^n (p/p_0)(X_i)\, d\Pi(p)}{\int \prod_{i=1}^n (p/p_0)(X_i)\, d\Pi(p)}$$

$$\le \phi_n + \Pi\big(p \colon K(p_0; p) < \epsilon\big) e^{n\epsilon} (1-\phi_n) \int_{\mathcal{U}^c} \prod_{i=1}^n (p/p_0)(X_i)\, d\Pi(p)$$

Extended Schwartz's theorem — proof (2)

Proof. continued.

- Step 2:

$$\Pi_n(\mathcal{U}^c|\, X_1,\ldots,X_n) \le \phi_n + (1-\phi_n)\frac{\int_{\mathcal{U}^c}\prod_{i=1}^n (p/p_0)(X_i)\,d\Pi(p)}{\int\prod_{i=1}^n (p/p_0)(X_i)\,d\Pi(p)}$$
$$\le \phi_n + \Pi\big(p\colon K(p_0;p)<\epsilon\big)e^{n\epsilon}(1-\phi_n)\int_{\mathcal{U}^c}\prod_{i=1}^n (p/p_0)(X_i)\,d\Pi(p)$$

- Step 3:

$$P_0^n\Big((1-\phi_n)\int_{\mathcal{U}^c}\prod_{i=1}^n \frac{p}{p_0}(X_i)\,d\Pi(p)\Big) = \int_{\mathcal{U}^c} P_0^n\Big[(1-\phi_n)\prod_{i=1}^n \frac{p}{p_0}(X_i)\Big]\,d\Pi(p)$$
$$\le \int_{\mathcal{U}^c} P^n(1-\phi_n)\,d\Pi(p).$$

Extended Schwartz's theorem — proof (2)

Proof. continued.

- Step 2:

$$\Pi_n(\mathcal{U}^c|\, X_1,\ldots,X_n) \le \phi_n + (1-\phi_n)\frac{\int_{\mathcal{U}^c}\prod_{i=1}^n (p/p_0)(X_i)\,d\Pi(p)}{\int \prod_{i=1}^n (p/p_0)(X_i)\,d\Pi(p)}$$
$$\le \phi_n + \Pi\big(p\colon K(p_0;p)<\epsilon\big)e^{n\epsilon}(1-\phi_n)\int_{\mathcal{U}^c}\prod_{i=1}^n (p/p_0)(X_i)\,d\Pi(p)$$

- Step 3:

$$P_0^n\left((1-\phi_n)\int_{\mathcal{U}^c}\prod_{i=1}^n \frac{p}{p_0}(X_i)\,d\Pi(p)\right) = \int_{\mathcal{U}^c} P_0^n\left[(1-\phi_n)\prod_{i=1}^n \frac{p}{p_0}(X_i)\right]d\Pi(p)$$
$$\le \int_{\mathcal{U}^c} P^n(1-\phi_n)\,d\Pi(p).$$

- Step 4: Split $\mathcal{U}^c$ in $\mathcal{U}^c\cap\mathcal{P}_n$ and $\mathcal{U}^c\cap\mathcal{P}_n^c$ and use that $P^n(1-\phi_n)\le e^{-Cn}$ on first set, while $\Pi(\mathcal{U}^c\cap\mathcal{P}_n^c)\le e^{-Cn}$.

∎

Strong consistency and entropy

Definition (Covering number). $N(\epsilon, \mathcal{P}, d)$ is the minimal number of d-balls of radius ϵ needed to cover $\mathcal{P}$.

Theorem. *The posterior distribution is consistent relative to the L_1-distance at every p_0 with the KL-property if for every $\epsilon > 0$ there exist a partition $\mathcal{P} = \mathcal{P}_{n,1} \cup \mathcal{P}_{n,2}$ (which may depend on ϵ) such that, for $C > 0$,*

(i) $\Pi(\mathcal{P}_{n,2}) \leq e^{-Cn}$.

(ii) $\log N\big(\epsilon, \mathcal{P}_{n,1}, \|\cdot\|_1\big) \leq n\epsilon^2/3$.

Strong consistency and entropy

Definition (Covering number). $N(\epsilon, \mathcal{P}, d)$ is the minimal number of d-balls of radius ϵ needed to cover $\mathcal{P}$.

Theorem. *The posterior distribution is consistent relative to the L_1-distance at every p_0 with the KL-property if for every $\epsilon > 0$ there exist a partition $\mathcal{P} = \mathcal{P}_{n,1} \cup \mathcal{P}_{n,2}$ (which may depend on ϵ) such that, for $C > 0$,*

(i) $\Pi(\mathcal{P}_{n,2}) \leq e^{-Cn}$.
(ii) $\log N\big(\epsilon, \mathcal{P}_{n,1}, \|\cdot\|_1\big) \leq n\epsilon^2/3$.

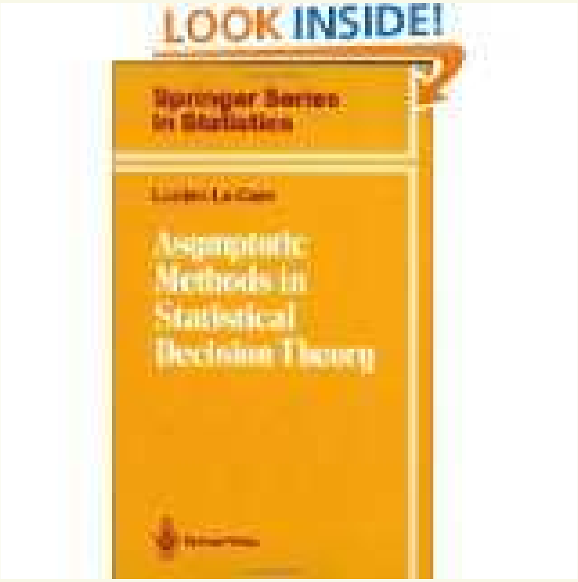
Proof.

- Entropy gives tests. See below.
- Apply Extended Schwartz's theorem.

$\square$

Tests

Tests


LOOK INSIDE!
Springer Series in Statistics
Asymptotic Methods in Statistical Decision Theory

Tests — the two Luciens

Lucien le Cam

Lucien Birgé

Tests — minimax theorem

minimax risk for testing P versus $\mathcal{Q}$:

$$\pi(P, \mathcal{Q}) = \inf_{\phi}\Bigl(P\phi + \sup_{Q\in\mathcal{Q}} Q(1-\phi)\Bigr).$$

Tests — minimax theorem

minimax risk for testing P versus $\mathcal{Q}$:

$$\pi(P, \mathcal{Q}) = \inf_{\phi} \Big(P\phi + \sup_{Q \in \mathcal{Q}} Q(1-\phi) \Big).$$

Hellinger affinity:

$$\rho_{1/2}(p, q) = \int \sqrt{p}\sqrt{q}\, d\mu = 1 - h^2(p, q)/2,$$

for $h^2(p, q) = \int \big(\sqrt{p} - \sqrt{q}\big)^2\, d\mu$ square *Hellinger distance*

Tests — minimax theorem

minimax risk for testing P versus $\mathcal{Q}$:

$$\pi(P,\mathcal{Q}) = \inf_{\phi}\Big(P\phi + \sup_{Q\in\mathcal{Q}} Q(1-\phi)\Big).$$

Hellinger affinity:

$$\rho_{1/2}(p,q) = \int \sqrt{p}\sqrt{q}\,d\mu = 1 - h^2(p,q)/2,$$

for $h^2(p,q) = \int\big(\sqrt{p}-\sqrt{q}\big)^2\,d\mu$ square *Hellinger distance*

Proposition. *For dominated probability measures* P *and* $\mathcal{Q}$

$$\pi(P,\mathcal{Q}) = 1 - \tfrac{1}{2}\|P - \mathrm{conv}(\mathcal{Q})\|_1 \le \sup_{Q\in\mathrm{conv}(\mathcal{Q})} \rho_{1/2}(p,q).$$

Tests — minimax risk

Proof.

- $$\begin{aligned} \pi(P,\mathcal{Q}) &= \inf_{\phi} \sup_{Q\in\mathrm{conv}(\mathcal{Q})} \Big(P\phi + Q(1-\phi)\Big) \\ &= \sup_{Q\in\mathrm{conv}(\mathcal{Q})} \inf_{\phi} \Big(P\phi + Q(1-\phi)\Big) \\ &= \sup_{Q\in\mathrm{conv}(\mathcal{Q})} \Big(P\mathbb{1}\{p<q\} + Q\mathbb{1}\{p\geq q\}\Big) \\ &= \sup_{Q\in\mathrm{conv}(\mathcal{Q})} \big(1-\tfrac{1}{2}\|p-q\|_1\big). \end{aligned}$$

- $$P\mathbb{1}\{p<q\} + Q\mathbb{1}\{p\geq q\} = \int_{p<q} p\,d\mu + \int_{p\geq q} q\,d\mu \leq \int \sqrt{p}\sqrt{q}\,d\mu.$$

$\square$

Tests — product measures

$$\rho_{1/2}(p_1 \times p_2, q_1 \times q_2) = \rho_{1/2}(p_1, q_1)\rho_{1/2}(p_2, q_2).$$

Tests — product measures

$$\rho_{1/2}(p_1 \times p_2, q_1 \times q_2) = \rho_{1/2}(p_1, q_1)\rho_{1/2}(p_2, q_2).$$

Lemma. *For any probability measures P_i and Q_i*

$$\rho_{1/2}\big(\otimes_i P_i, \mathrm{conv}(\otimes_i \mathcal{Q}_i)\big) \leq \prod_i \rho_{1/2}\big(P_i, \mathrm{conv}(\mathcal{Q}_i)\big).$$

Tests — product measures

$$\rho_{1/2}(p_1 \times p_2, q_1 \times q_2) = \rho_{1/2}(p_1, q_1)\rho_{1/2}(p_2, q_2).$$

Lemma. *For any probability measures P_i and $\mathcal{Q}_i$*

$$\rho_{1/2}\big(\otimes_i P_i, \mathrm{conv}(\otimes_i \mathcal{Q}_i)\big) \leq \prod_i \rho_{1/2}\big(P_i, \mathrm{conv}(\mathcal{Q}_i)\big).$$

Proof. Suffices to consider products of 2.
If $q(x,y) = \sum_j \kappa_j q_{1j}(x) q_{2j}(y)$, then $\rho_{1/2}(p_1 \times p_2, q) =$

$$\int p_1(x)^{1/2}\Big(\sum_j \kappa_j q_{1j}(x)\Big)^{1/2}\Big[\int p_2(y)^{1/2}\Big(\frac{\sum_j \kappa_j q_{1j}(x) q_{2j}(y)}{\sum_j \kappa_j q_{1j}(x)}\Big)^{1/2} d\mu_2(y)\Big]\, d\mu_1(x).$$

$\square$

Tests — product measures (2)

Corollary.

$$\pi(P^n, \mathcal{Q}^n) \leq \rho_{1/2}(P^n, \mathrm{conv}(\mathcal{Q}^n)) \leq \rho_{1/2}(P, \mathrm{conv}(\mathcal{Q}))^n.$$

Tests — product measures (2)

Corollary.

$$\pi(P^n, \mathcal{Q}^n) \leq \rho_{1/2}(P^n, \mathrm{conv}(\mathcal{Q}^n)) \leq \rho_{1/2}(P, \mathrm{conv}(\mathcal{Q}))^n.$$

Theorem. *For any probability measure P and convex set of dominated probability measures $\mathcal{Q}$ with $h(p,q) > \epsilon$ for every $q \in \mathcal{Q}$ and any $n \in \mathbb{N}$, there exists a test ϕ such that*

$$P^n\phi \leq e^{-n\epsilon^2/2}, \qquad \sup_{Q\in\mathcal{Q}} Q^n(1-\phi) \leq e^{-n\epsilon^2/2}.$$

Tests — product measures (2)

Corollary.

$$\pi(P^n, \mathcal{Q}^n) \le \rho_{1/2}(P^n, \mathrm{conv}(\mathcal{Q}^n)) \le \rho_{1/2}(P, \mathrm{conv}(\mathcal{Q}))^n.$$

Theorem. *For any probability measure P and convex set of dominated probability measures $\mathcal{Q}$ with $h(p,q) > \epsilon$ for every $q \in \mathcal{Q}$ and any $n \in \mathbb{N}$, there exists a test ϕ such that*

$$P^n\phi \le e^{-n\epsilon^2/2}, \qquad \sup_{Q \in \mathcal{Q}} Q^n(1-\phi) \le e^{-n\epsilon^2/2}.$$

Proof.

- $\rho_{1/2}(P, \mathcal{Q}) = 1 - \frac{1}{2}h^2(P, \mathcal{Q}) \le 1 - \epsilon^2/2$.
- $\pi(P^n, \mathcal{Q}^n) \le (1-\epsilon^2/2)^n \le e^{-n\epsilon^2/2}$.

$\square$

Tests — nonconvex alternatives

Definition (Covering number). $N(\epsilon, \mathcal{Q}, d)$ is the minimal number of d-balls of radius ϵ needed to cover $\mathcal{Q}$.

Proposition. *Let $d \leq h$ be a metric whose balls are convex. If $N(\epsilon/4, \mathcal{Q}, d) \leq N(\epsilon)$ for every $\epsilon > \epsilon_n > 0$ and some nonincreasing function $N\colon (0,\infty) \to (0,\infty)$, then for every $\epsilon > \epsilon_n$ and n there exists a test ϕ such that, for all $j \in \mathbb{N}$,*

$$P^n\phi \leq N(\epsilon)\frac{e^{-n\epsilon^2/2}}{1-e^{-n\epsilon^2/8}}, \qquad \sup_{Q\in\mathcal{Q}: d(P,Q)>j\epsilon} Q^n(1-\phi) \leq e^{-n\epsilon^2 j^2/8}.$$

Tests — nonconvex alternatives

Proof.

- For $j \in \mathbb{N}$, choose a maximal set of $j\epsilon/2$-separated points $Q_{j,1}, \ldots, Q_{j,N_j}$ in $\mathcal{Q}_j := \bigl\{Q \in \mathcal{Q}: j\epsilon < d(P,Q) < 2j\epsilon\bigr\}$.
  (i). $N_j \leq N(j\epsilon/4, \mathcal{Q}_j, d)$.
  (ii). The N_j balls $B_{j,l}$ of radius $j\epsilon/2$ around the $Q_{j,l}$ cover $\mathcal{Q}_j$.
  (iii). $h(P, B_{j,l}) \geq d(P, B_{j,l}) > j\epsilon/2$ for every ball $B_{j,l}$.
- For every ball take a test $\phi_{j,l}$ of P versus $B_{j,l}$. Let ϕ be their supremum.

$$P^n\phi \leq \sum_{j=1}^{\infty}\sum_{l=1}^{N_j} e^{-nj^2\epsilon^2/8} \leq \sum_{j=1}^{\infty} N(j\epsilon/4, \mathcal{Q}_j, d)e^{-nj^2\epsilon^2/8} \leq N(\epsilon)\frac{e^{-n\epsilon^2/8}}{1-e^{-n\epsilon^2/8}}$$

and, for every $j \in \mathbb{N}$,

$$\sup_{Q \in \cup_{l>j}\mathcal{Q}_l} Q^n(1-\phi) \leq \sup_{l>j} e^{-nl^2\epsilon^2/8} \leq e^{-nj^2\epsilon^2/8}.$$

$\square$

Rate of contraction

Definition. The posterior distribution $\Pi_n(\cdot|\, X^{(n)})$ *contracts at rate* $\epsilon_n \to 0$ at $\theta_0 \in \Theta$ if $\Pi_n\big(\theta\colon d(\theta,\theta_0) > M_n\epsilon_n|\, X^{(n)}\big) \to 0$ in $P_{\theta_0}^{(n)}$-probability, for every $M_n \to \infty$ as $n \to \infty$.

Rate of contraction

Definition. The posterior distribution $\Pi_n(\cdot|\, X^{(n)})$ *contracts at rate* $\epsilon_n \to 0$ at $\theta_0 \in \Theta$ if $\Pi_n\big(\theta\colon d(\theta,\theta_0) > M_n\epsilon_n|\, X^{(n)}\big) \to 0$ in $P_{\theta_0}^{(n)}$-probability, for every $M_n \to \infty$ as $n \to \infty$.

Proposition (Point estimator)**.** *If the posterior distribution contracts at rate* ϵ_n *at* θ_0*, then* $\hat\theta_n$ *defined as the center of a (nearly) smallest ball that contains posterior mass at least* $1/2$ *satisfies* $d(\hat\theta_n,\theta_0) = O_P(\epsilon_n)$ *under* $P_{\theta_0}^{(n)}$.

Basic contraction theorem

$$K(p_0; p) = P_0 \log \frac{p_0}{p}, \qquad V(p_0; p) = P_0 \Big(\log \frac{p_0}{p}\Big)^2.$$

Theorem. *Given* $d \leq h$ *whose balls are convex suppose that there exist* $\mathcal{P}_n \subset \mathcal{P}$ *and* $C > 0$*, such that,*

(i) $\Pi_n\big(p\colon K(p_0; p) < \epsilon_n^2, V(p_0; p) < \epsilon_n^2\big) \geq e^{-Cn\epsilon_n^2}$,

(ii) $\log N\big(\epsilon_n, \mathcal{P}_n, d\big) \leq n\epsilon_n^2$.

(iii) $\Pi_n(\mathcal{P}_n^c) \leq e^{-(C+4)n\epsilon_n^2}$.

Then the posterior rate of convergence for d *is* $\epsilon_n \vee n^{-1/2}$*.*

Basic contraction theorem — proof

Proof.

- There exist tests ϕ_n with

$$P_0^n\phi_n \le e^{n\epsilon_n^2}\frac{e^{-nM^2\epsilon_n^2/8}}{1-e^{-nM^2\epsilon_n^2/8}}, \qquad \sup_{p\in\mathcal{P}_n: d(p,p_0)>M\epsilon_n} P^n(1-\phi_n) \le e^{-nM^2\epsilon_n^2/8}.$$

- For $A_n = \{\int\prod_{i=1}^n (p/p_0)(X_i)\,d\Pi_n(p) \ge e^{-(2+C)n\epsilon_n^2}\}$

$$\Pi_n\bigl(p\colon d(p,p_0)>M\epsilon_n|\,X_1,\ldots,X_n\bigr) \le \phi_n + 1\!\!1\{A_n^c\} + e^{(2+C)n\epsilon_n^2}\int_{d(p,p_0)>M\epsilon_n}\prod_{i=1}^n\frac{p}{p_0}(X_i)\,d\Pi_n(p)(1-\phi_n).$$

- $P_0^n(A_n^c)\to 0$. See further on.

$\square$

Basic contraction theorem — proof continued

Proof. (Continued)

- $$P_0^n \int_{p\in\mathcal{P}_n : d(p,p_0)>M\epsilon_n} \prod_{i=1}^n \frac{p}{p_0}(X_i)\, d\Pi_n(p) \le \int_{p\in\mathcal{P}_n : d(p,p_0)>M\epsilon_n} P^n(1-\phi_n)\, d\Pi_n(p) \le e^{-nM^2\epsilon_n^2/8}$$

- $$P_0^n \int_{\mathcal{P}-\mathcal{P}_n} \prod_{i=1}^n \frac{p}{p_0}(X_i)\, d\Pi_n(p) \le \Pi_n(\mathcal{P}-\mathcal{P}_n).$$

$\square$

Bounding the denominator

Lemma. *For any probability measure* Π *on* $\mathcal{P}$*, and positive constant* ϵ*, with* P_0^n*-probability at least* $1-(n\epsilon^2)^{-1}$*,*

$$\int \prod_{i=1}^n \frac{p}{p_0}(X_i)\, d\Pi(p) \geq \Pi\Big(p\colon K(p_0;p) < \epsilon^2, V(p_0;p) < \epsilon^2\Big) e^{-2n\epsilon^2}.$$

Bounding the denominator

Lemma. *For any probability measure* Π *on* $\mathcal{P}$*, and positive constant* ϵ*, with* P_0^n*-probability at least* $1-(n\epsilon^2)^{-1}$*,*

$$\int \prod_{i=1}^n \frac{p}{p_0}(X_i)\, d\Pi(p) \geq \Pi\big(p\colon K(p_0;p) < \epsilon^2, V(p_0;p) < \epsilon^2\big) e^{-2n\epsilon^2}.$$

Proof. $B\colon= \big\{p\colon K(p_0;p) < \epsilon_n^2, V(p_0;p) < \epsilon_n^2\big\}$.

$$\log \int \prod_{i=1}^n \frac{p}{p_0}(X_i)\, d\Pi(P) \geq \sum_{i=1}^n \int \log \frac{p}{p_0}(X_i)\, d\Pi(P) =\colon Z.$$

$$\mathrm{E} Z = -n \int K(p_0;p)\, d\Pi(p) > -n\epsilon^2,$$

$$\operatorname{var} Z \leq nP_0\Big(\int \log\frac{p_0}{p}\, d\Pi(p)\Big)^2 \leq nP_0 \int \Big(\log\frac{p_0}{p}\Big)^2 d\Pi(p) \leq n\epsilon^2,$$

Apply Chebyshev's inequality. $\square$

Interpretation

Consider a maximal set of points $p_1, \ldots, p_N$ in $\mathcal{P}_n$ with $d(p_i, p_j) \geq \epsilon_n$.

Maximality implies $N \geq N(\epsilon_n, \mathcal{P}_n, d) \geq e^{c_1 n \epsilon_n^2}$, under the entropy bound.

The balls of radius $\epsilon_n/2$ around the points are disjoint and hence the sum of their prior masses will be less than 1.

If the prior mass were evenly distributed over these balls, then each would have no more mass than $e^{-c_1 n \epsilon_n^2}$.

This is of the same order as the prior mass bound.

This argument suggests that the conditions can only be satisfied for every p_0 in the model if the prior "distributes its mass uniformly, at discretization level ϵ_n".

General observations

Experiments $(\mathfrak{X}^{(n)}, \mathscr{X}^{(n)}, P_\theta^{(n)}\colon \theta \in \Theta_n)$, with observations $X^{(n)}$, and true parameters $\theta_{n,0} \in \Theta_n$.

d_n and e_n semi-metrics on Θ_n such that: there exist $\xi, K > 0$ such that for every $\epsilon > 0$ and every $\theta_{n,1} \in \Theta_n$ with $d_n(\theta_1, \theta_{n,0}) > \epsilon$, there exists a test ϕ_n such that

$$P_{\theta_{n,0}}^{(n)} \phi_n \leq e^{-Kn\epsilon^2}, \qquad \sup_{\theta \in \Theta_n : e_n(\theta, \theta_{n,1}) < \xi\epsilon} P_\theta^{(n)} (1 - \phi_n) n \leq e^{-Kn\epsilon^2}.$$

General observations — rate of contraction

$$B_{n,k}(\theta_{n,0},\epsilon)=\Big\{\theta\in\Theta_n\colon K(p^{(n)}_{\theta_{n,0}};p^{(n)}_\theta)\le n\epsilon^2, V_{k,0}(p^{(n)}_{\theta_{n,0}};p^{(n)}_\theta)\le n^{k/2}\epsilon^k\Big\}.$$

Theorem. *If for arbitrary $\Theta_{n,1}\subset\Theta_n$ and $k>1$, $n\epsilon_n^2\ge 1$, and every $j\in\mathbb{N}$,*

(i) $$\frac{\Pi_n\big(\theta\in\Theta_{n,1}\colon j\epsilon_n<d_n(\theta,\theta_0)\le 2j\epsilon_n\big)}{\Pi_n\big(B_{n,k}(\theta_0,\epsilon_n)\big)}\le e^{Kn\epsilon_n^2 j^2/2},$$

(ii) $$\sup_{\epsilon>\epsilon_n}\log N\big(\xi\epsilon,\{\theta\in\Theta_{n,1}\colon d_n(\theta,\theta_{n,0})<2\epsilon\},e_n\big)\le n\epsilon_n^2,$$

then $\Pi_n\big(\theta\in\Theta_{n,1}\colon d_n(\theta,\theta_{n,0})\ge M_n\epsilon_n|\,X^{(n)}\big)\to 0$, in $P^{(n)}_{\theta_{n,0}}$-probability, for every $M_n\to\infty$.

Theorem. *If for arbitrary $\Theta_{n,2}\subset\Theta_n$, some $k>1$,*

(iii) $$\frac{\Pi_n(\Theta_{n,2})}{\Pi_n\big(B_{n,k}(\theta_{n,0},\epsilon_n)\big)}=o\left(e^{-2n\epsilon_n^2}\right). \qquad (3)$$

then $\Pi_n(\Theta_{n,2}|\,X^{(n)})\to 0$, in $P^{(n)}_{\theta_{n,0}}$-probability if,

Gaussian process priors

Gaussian processes

Definition. A Gaussian process is a set of random variables (or vectors) $W = (W_t\colon t \in T)$ such that $(W_{t_1}, \ldots, W_{t_k})$ is multivariate normal, for every $t_1, \ldots, t_k \in T$.

The finite-dimensional distributions are determined by the mean function and the covariance function

$$\mu(t) = \mathrm{E}W_t, \qquad K(s,t) = \mathrm{E}W_s W_t, \qquad s, t \in T.$$

Gaussian processes

Definition. A Gaussian process is a set of random variables (or vectors) $W = (W_t: t \in T)$ such that $(W_{t_1}, \ldots, W_{t_k})$ is multivariate normal, for every $t_1, \ldots, t_k \in T$.

The finite-dimensional distributions are determined by the mean function and the covariance function

$$\mu(t) = \mathrm{E}W_t, \qquad K(s,t) = \mathrm{E}W_s W_t, \qquad s, t \in T.$$

The law of a Gaussian process is a prior for a function.

Gaussian processes

Definition. A Gaussian process is a set of random variables (or vectors) $W = (W_t : t \in T)$ such that $(W_{t_1}, \ldots, W_{t_k})$ is multivariate normal, for every $t_1, \ldots, t_k \in T$.

The finite-dimensional distributions are determined by the mean function and the covariance function

$$\mu(t) = \mathrm{E}W_t, \qquad K(s,t) = \mathrm{E}W_s W_t, \qquad s,t \in T.$$

The law of a Gaussian process is a prior for a function.

Gaussian process priors have been found useful, because

- they offer great variety
- they are easy (?) to understand through their covariance function
- they can be computationally attractive (e.g. `www.gaussianprocess.org`)

Brownian density estimation

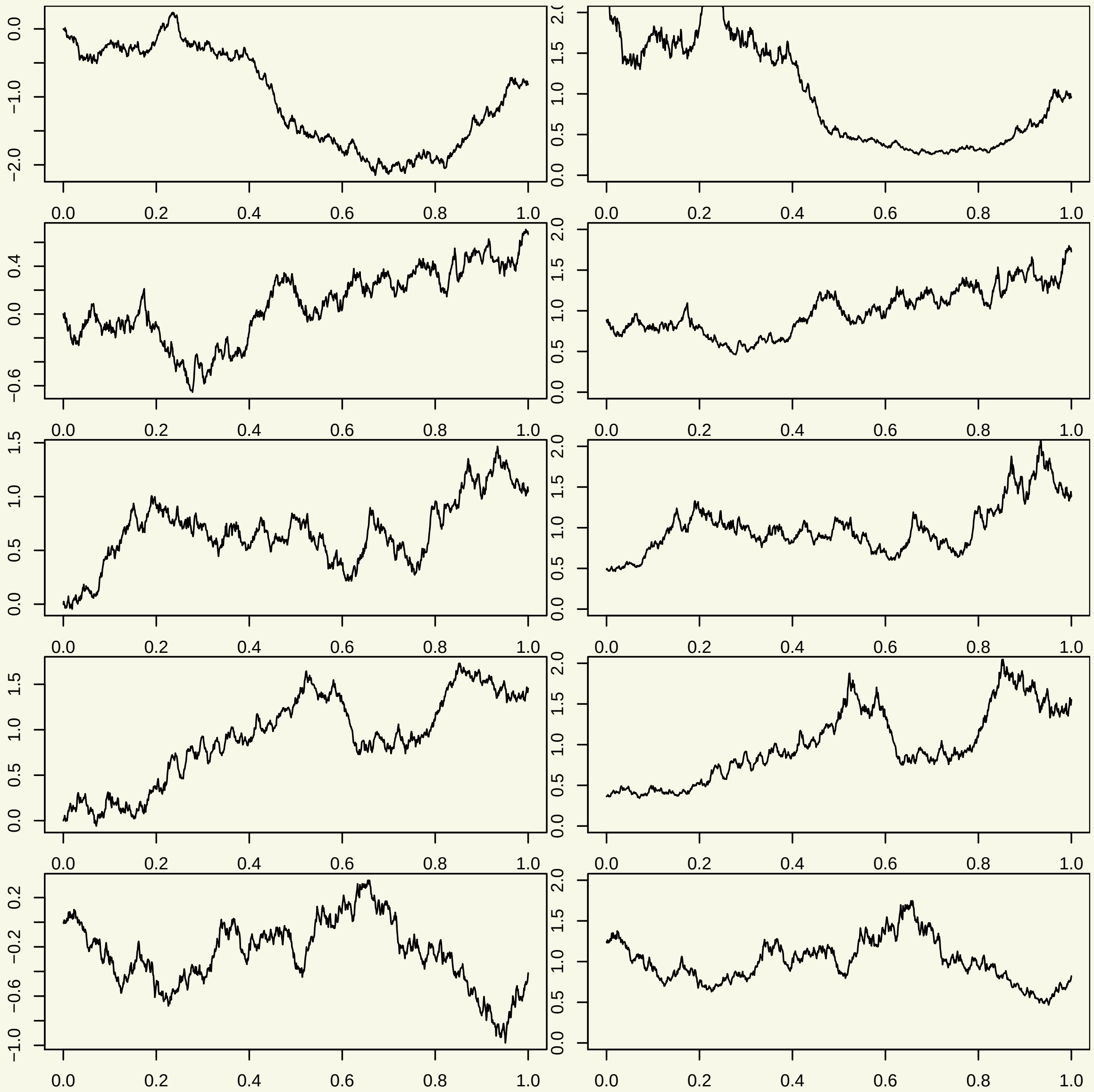
- $X_1, \ldots, X_n$ i.i.d. from density p_0 on $[0,1]$
- $(W_x: x \in [0,1])$ Brownian motion

As prior on p use:

$$x \mapsto \frac{e^{W_x}}{\int_0^1 e^{W_y}\,dy}$$

Brownian density estimation

Brownian motion $t \mapsto W_t$ — Prior density $t \mapsto c\exp(W_t)$



Brownian density estimation

- $X_1, \ldots, X_n$ i.i.d. from density p_0 on $[0,1]$
- $(W_x\colon x \in [0,1])$ Brownian motion

As prior on p use:

$$x \mapsto \frac{e^{W_x}}{\int_0^1 e^{W_y}\,dy}$$

Brownian density estimation

- $X_1, \ldots, X_n$ i.i.d. from density p_0 on $[0,1]$
- $(W_x \colon x \in [0,1])$ Brownian motion

As prior on p use:

$$x \mapsto \frac{e^{W_x}}{\int_0^1 e^{W_y}\,dy}$$

Theorem. *If $w_0 \colon = \log p_0 \in C^\alpha[0,1]$, then L_2-rate is:*

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

Brownian density estimation

- $X_1, \ldots, X_n$ i.i.d. from density p_0 on $[0,1]$
- $(W_x : x \in [0,1])$ Brownian motion

As prior on p use:

$$x \mapsto \frac{e^{W_x}}{\int_0^1 e^{W_y}\,dy}$$

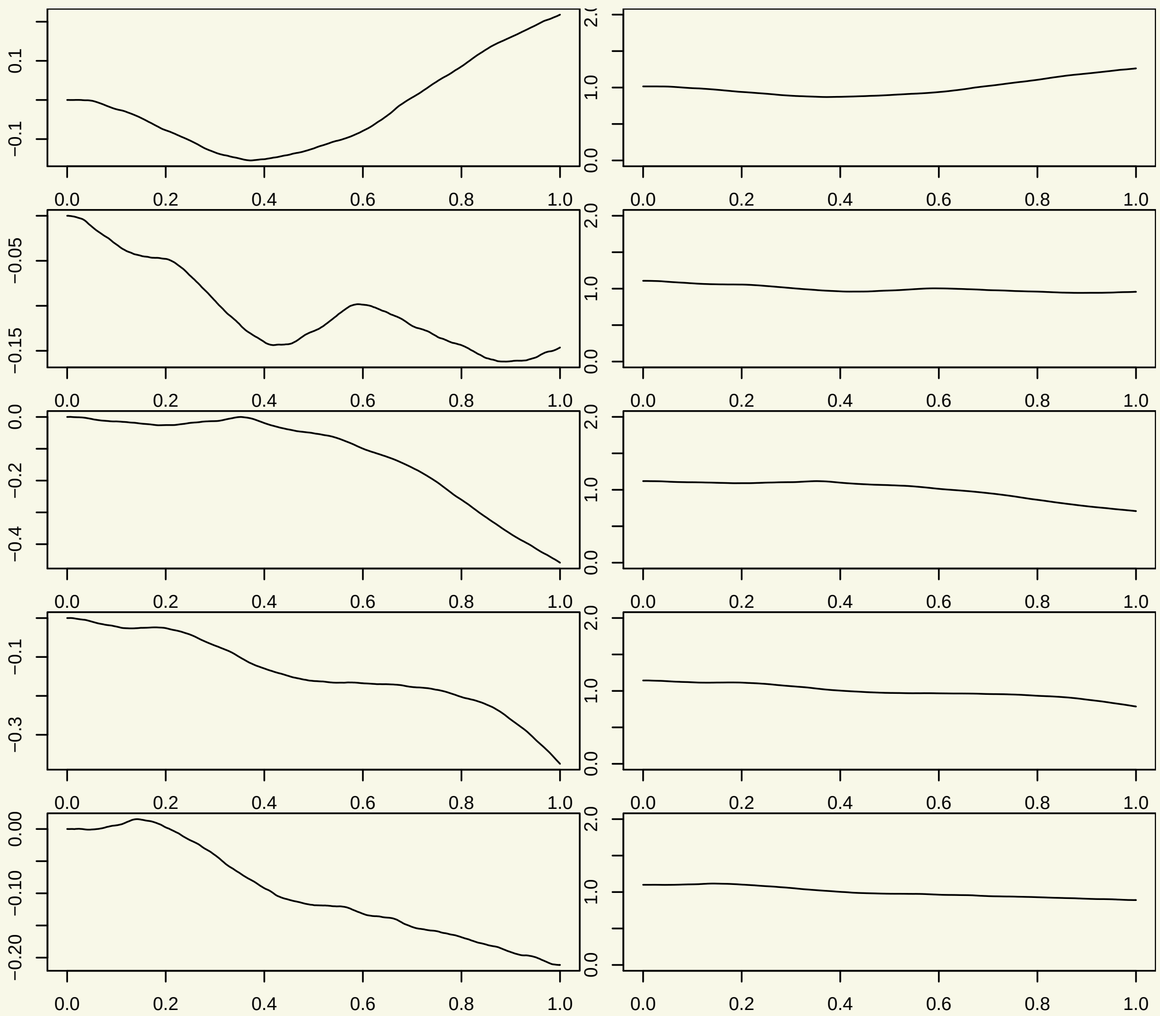
Theorem. *If $w_0 := \log p_0 \in C^\alpha[0,1]$, then L_2-rate is:*

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

- *This is optimal if and only if $\alpha = 1/2$.*
- *Rate does not improve if α increases from $1/2$.*
- *Consistency for any $\alpha > 0$.*

Integrated Brownian density estimation

Integrated Brownian motion — Prior density



Integrated Brownian motion: Riemann-Liouville process

$\alpha - 1/2$ times integrated Brownian motion, released at 0

$$W_t = \int_0^t (t-s)^{\alpha-1/2}\, dB_s + \sum_{k=0}^{[\alpha]+1} Z_k t^k$$

[B Brownian motion, $\alpha > 0$, (Z_k) iid $N(0,1)$, “fractional integral”]

Theorem. *IBM gives appropriate model for α-smooth functions: consistency if $w_0 \in C^\beta[0,1]$ for any $\beta > 0$, but the optimal $n^{-\beta/(2\beta+1)}$ if and only if $\alpha = \beta$.*

Settings

Density estimation

$X_1, \ldots, X_n$ iid in $[0,1]$,

$$p_\theta(x) = \frac{e^{\theta(x)}}{\int_0^1 e^{\theta(t)}\,dt}.$$

- Distance on parameter: Hellinger on p_θ.
- Norm on W: uniform.

Classification

$(X_1, Y_1), \ldots, (X_n, Y_n)$ iid in $[0,1] \times \{0,1\}$

$$\Pr_\theta(Y = 1 |\, X = x) = \frac{1}{1 + e^{-\theta(x)}}.$$

- Distance on parameter: $L_2(G)$ on Pr_θ. (G marginal of X_i.)
- Norm on W: $L_2(G)$.

Regression

$Y_1, \ldots, Y_n$ independent $N(\theta(x_i), \sigma^2)$, for fixed design points $x_1, \ldots, x_n$.

- Distance on parameter: empirical L_2-distance on θ.
- Norm on W: empirical L_2-distance.

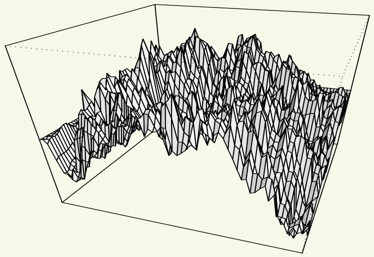
Ergodic diffusions

$(X_t \colon t \in [0,n])$, ergodic, recurrent:

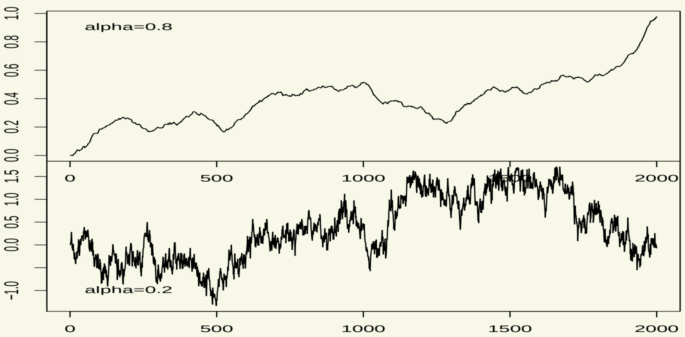
$$dX_t = \theta(X_t)\,dt + \sigma(X_t)\,dB_t.$$

- Distance on parameter: random Hellinger h_n ($\approx \|\cdot/\sigma\|_{\mu_0,2}$).
- Norm on W: $L_2(\mu_0)$. (μ_0 stationary measure.)

Other Gaussian processes



Brownian sheet



Fractional Brownian motion

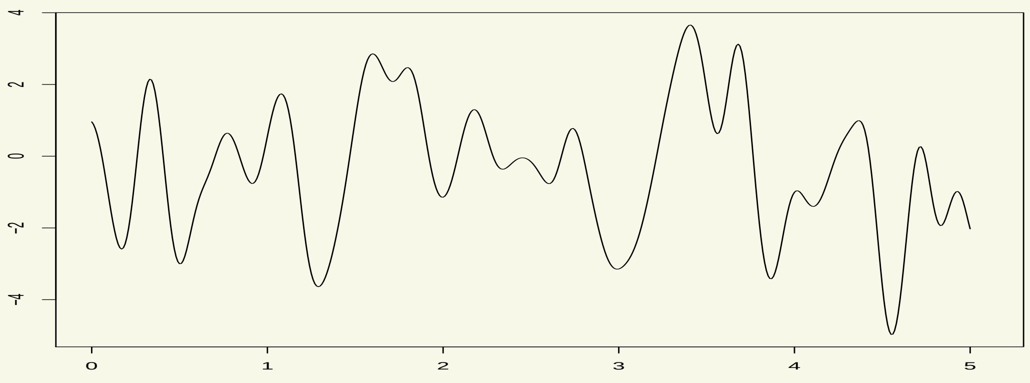
$$\theta(x) = \sum_i \theta_i e_i(x), \qquad \theta_i \sim_{indep} N(0, \lambda_i)$$

Series prior

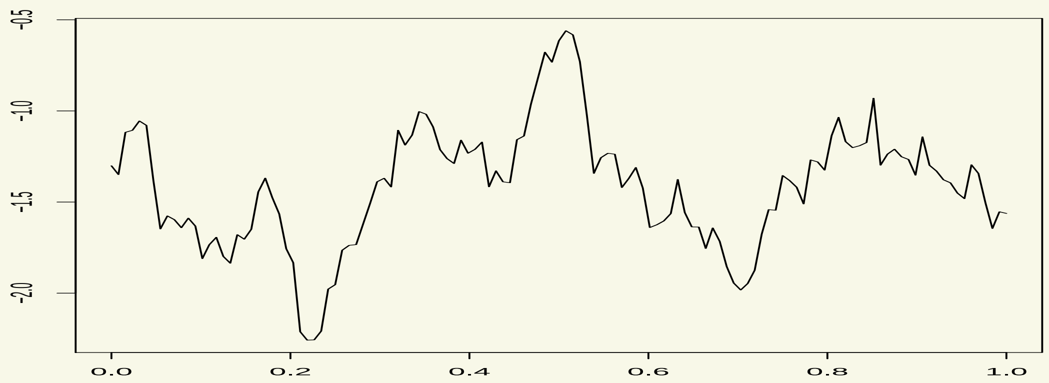
Stationary processes

A stationary Gaussian field $(W_t \colon t \in \mathbb{R}^d)$ is characterized through a spectral measure μ, by

$$\operatorname{cov}(W_s, W_t) = \int e^{i\lambda^T (s-t)}\, d\mu(\lambda).$$



Gaussian spectral measure; "radial basis"

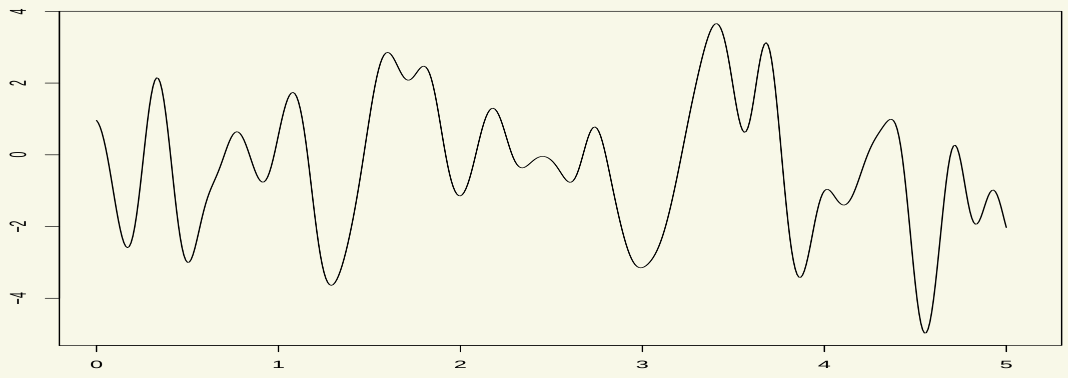


Matérn spectral measure (3/2)

Stationary processes — radial basis

Stationary Gaussian field $(W_t\colon t \in \mathbb{R}^d)$ characterized through

$$\operatorname{cov}(W_s, W_t) = \int e^{i\lambda^T(s-t)}\, e^{-\lambda^2}\, d\lambda.$$



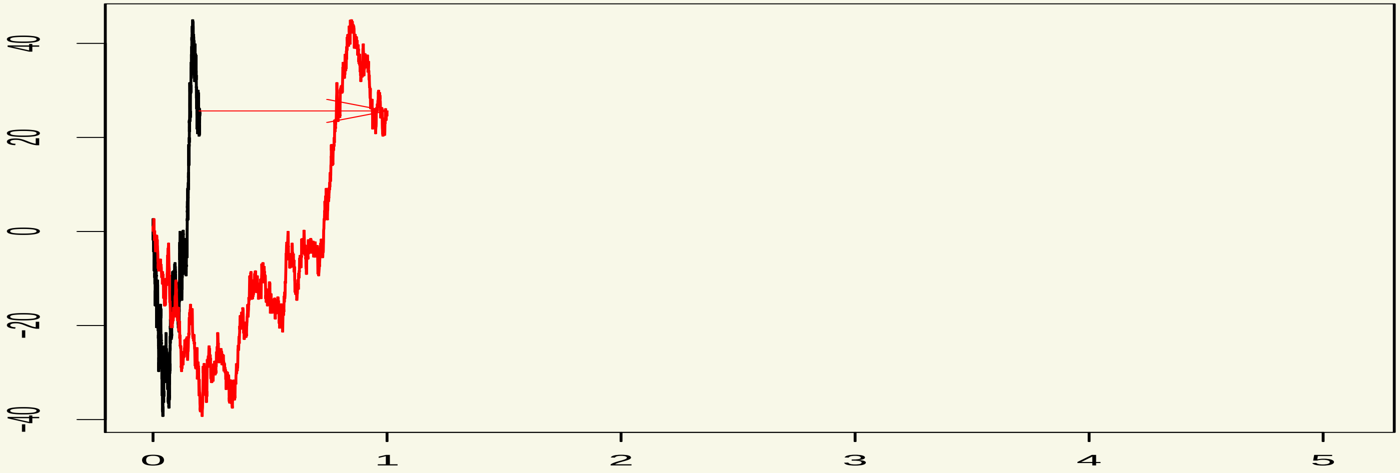
Theorem. *Let $\hat{w}_0$ be the Fourier transform of the true parameter $w_0\colon [0,1]^d \to \mathbb{R}$.*

- *If $\int e^{\|\lambda\|}|\hat{w}_0(\lambda)|^2\, d\lambda < \infty$, then rate of contraction is near $1/\sqrt{n}$.*
- *If $|\hat{w}_0(\lambda)| \gtrsim (1+\|\lambda\|^2)^{-\beta}$, then rate is power of $1/\log n$.*

Excellent if truth is supersmooth; disastrous otherwise.

Stretching or shrinking: “length scale”

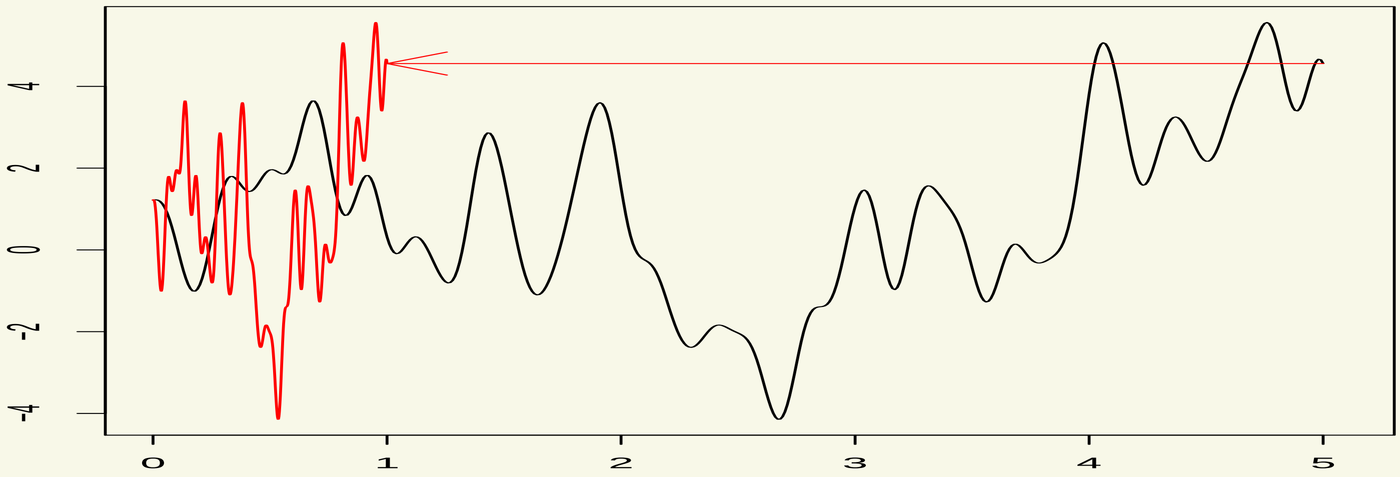
Sample paths can be smoothed by stretching



Stretching or shrinking: “length scale”

Sample paths can be smoothed by stretching

and roughened by shrinking



Rescaled Brownian motion

$W_t = B_{t/c_n}$ for B Brownian motion, and $c_n \sim n^{(2\alpha-1)/(2\alpha+1)}$

- $\alpha < 1/2$: $c_n \to 0$ (shrink)
- $\alpha \in (1/2, 1]$: $c_n \to \infty$ (stretch)

Theorem. *The prior $W_t = B_{t/c_n}$ gives optimal rate for $w_0 \in C^\alpha[0,1]$, $\alpha \in (0,1]$.*

Rescaled Brownian motion

$W_t = B_{t/c_n}$ for B Brownian motion, and $c_n \sim n^{(2\alpha-1)/(2\alpha+1)}$

- $\alpha < 1/2$: $c_n \to 0$ (shrink)
- $\alpha \in (1/2, 1]$: $c_n \to \infty$ (stretch)

Theorem. *The prior $W_t = B_{t/c_n}$ gives optimal rate for $w_0 \in C^\alpha[0,1]$, $\alpha \in (0,1]$.*

Surprising? (Brownian motion is self-similar!)

Rescaled Brownian motion

$W_t = B_{t/c_n}$ for B Brownian motion, and $c_n \sim n^{(2\alpha-1)/(2\alpha+1)}$

- $\alpha < 1/2$: $c_n \to 0$ (shrink)
- $\alpha \in (1/2, 1]$: $c_n \to \infty$ (stretch)

Theorem. *The prior* $W_t = B_{t/c_n}$ *gives optimal rate for* $w_0 \in C^\alpha[0,1]$*,* $\alpha \in (0,1]$*.*

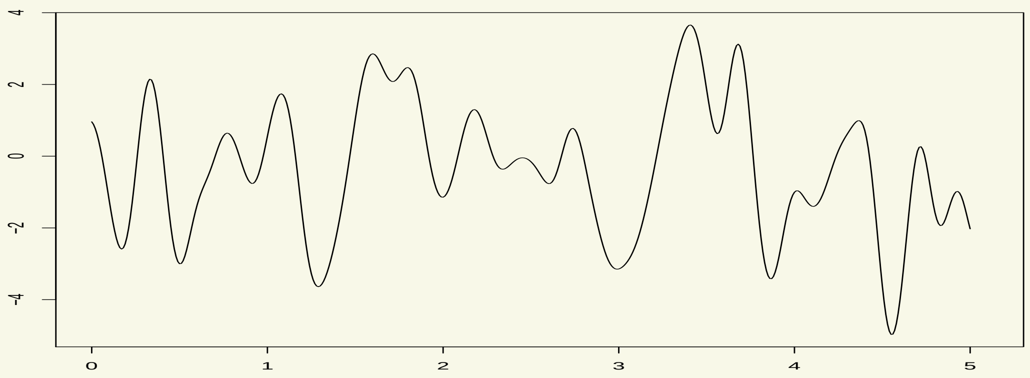
Surprising? (Brownian motion is self-similar!)

Appropriate rescaling of k *times integrated Brownian motion gives optimal prior for every* $\alpha \in (0, k+1]$*.*

Rescaled smooth stationary process

A Gaussian field with infinitely-smooth sample paths is obtained with

$$\mathrm{E} G_s G_t = \psi(s-t), \qquad \int e^{\|\lambda\|} \hat{\psi}(\lambda)\, d\lambda < \infty.$$



Gaussian spectral measure; "radial basis"

Theorem. *The prior $W_t = G_{t/c_n}$ for $c_n \sim n^{-1/(2\alpha+d)}$ gives nearly optimal rate for $w_0 \in C^\alpha[0,1]$, any $\alpha > 0$.*

Gaussian elements in a Banach space

Definition. A Gaussian random variable in a (separable) Banach space $\mathbb{B}$ is a Borel measurable map $W: (\Omega, \mathscr{U}, \Pr) \to \mathbb{B}$ such that $b^* W$ is normally distributed for every b^* in the dual space $\mathbb{B}^*$.

Many Gaussian processes $(W_t: t \in T)$ can be viewed as a Gaussian variable in a space of functions $w: T \to \mathbb{R}^d$.

EXAMPLES

- Brownian motion can be viewed as a map in $C[0,1]$, equipped with the uniform norm $\|w\| = \sup_{t\in[0,1]} |w(t)|$.

Gaussian elements in a Banach space

Definition. A Gaussian random variable in a (separable) Banach space $\mathbb{B}$ is a Borel measurable map $W\colon (\Omega, \mathscr{U}, \Pr) \to \mathbb{B}$ such that b^*W is normally distributed for every b^* in the dual space $\mathbb{B}^*$.

Many Gaussian processes $(W_t\colon t \in T)$ can be viewed as a Gaussian variable in a space of functions $w\colon T \to \mathbb{R}^d$.

EXAMPLES

- Brownian motion can be viewed as a map in $C[0,1]$, equipped with the uniform norm $\|w\| = \sup_{t\in[0,1]} |w(t)|$.
- Brownian motion is also a map in $L_2[0,1]$, or $C^{1/4}[0,1]$, or some Besov space.

RKHS — definition

W zero-mean Gaussian in Banach space $(\mathbb{B}, \|\cdot\|)$.

$S\colon \mathbb{B}^* \to \mathbb{B}$, $\quad Sb^* = \mathrm{E}W b^*(W)$.

Definition. The *reproducing kernel Hilbert space* $(\mathbb{H}, \|\cdot\|_{\mathbb{H}})$ of W is the completion of $S\mathbb{B}^*$ under

$$\langle Sb_1^*, Sb_2^* \rangle_{\mathbb{H}} = \mathrm{E} b_1^*(W) b_2^*(W)$$

.

RKHS — definition (2)

$W=(W_t\colon t\in T)$ Gaussian process that can be seen as tight, Borel measurable map in $\ell^\infty(T)=\{f\colon T\to\mathbb{R}\colon \|f\|\colon=\sup_t|f(t)|<\infty\}$. with covariance function $K(s,t)=\mathrm{E}W_sW_t$.

Theorem. *Then RKHS is completion of the set of functions*

$$t\mapsto\sum_i\alpha_iK(s_i,t)$$

relative to inner product

$$\Big\langle\sum_i\alpha_iK(r_i,\cdot),\sum_j\beta_jK(s_j,\cdot)\Big\rangle_{\mathbb{H}}=\sum_i\sum_j\alpha_i\beta_jK(r_i,t_j).$$

RKHS — definition (2)'

$W=(W_t\colon t\in T)$ Gaussian process that can be seen as tight, Borel measurable map in $\ell^\infty(T)=\{f\colon T\to\mathbb{R}\colon \|f\|\colon=\sup_t|f(t)|<\infty\}$, with covariance function $K(s,t)=\mathrm{E}W_sW_t$.

Theorem. *Then RKHS is completion of the set of functions*

$$t\mapsto\sum_i\alpha_iK(s_i,t)=\mathrm{E}\Big(\sum_i\alpha_iW_{s_i}\Big)W_t$$

relative to inner product

$$\Big\langle\sum_i\alpha_iK(r_i,\cdot),\sum_j\beta_jK(s_j,\cdot)\Big\rangle_{\mathbb{H}}=\sum_i\sum_j\alpha_i\beta_jK(r_i,t_j)$$

i.e. all functions $t\mapsto h_L(t)\colon=\mathrm{E}LW_t$, *where* $L\in L_2(W)$, *with inner product*

$$\langle h_{L_1},h_{L_2}\rangle_{\mathbb{H}}=\mathrm{E}L_1L_2.$$

RKHS — definition (3)

Any Gaussian random element in a separable Banach space can be represented (in many ways, e.g. spectral decomposition) as

$$W = \sum_{i=1}^{\infty} \mu_i Z_i e_i$$

for

- $\mu_i \downarrow 0$
- $Z_1, Z_2, \ldots$ i.i.d. $N(0,1)$
- $\|e_1\| = \|e_2\| = \cdots = 1$

RKHS — definition (3)

Any Gaussian random element in a separable Banach space can be represented (in many ways, e.g. spectral decomposition) as

$$W = \sum_{i=1}^{\infty} \mu_i Z_i e_i$$

for

- $\mu_i \downarrow 0$
- $Z_1, Z_2, \ldots$ i.i.d. $N(0,1)$
- $\|e_1\| = \|e_2\| = \cdots = 1$

Theorem. *The RKHS consists of all elements* $h := \sum_i h_i e_i$ *with*

$$\|h\|_{\mathbb{H}}^2 := \sum_i \frac{h_i^2}{\mu_i^2} < \infty$$

EXAMPLE — Brownian motion

Theorem. *The RKHS of k times IBM is*

$$\{f \colon f^{(k+1)} \in L_2[0,1], f(0) = \cdots = f^{(k)}(0) = 0\}, \qquad \|f\|_{\mathbb{H}} = \|f^{(k+1)}\|_2.$$

EXAMPLE — Brownian motion

Theorem. *The RKHS of k times IBM is*

$$\{f\colon f^{(k+1)} \in L_2[0,1], f(0) = \cdots = f^{(k)}(0) = 0\}, \quad \|f\|_{\mathbb{H}} = \|f^{(k+1)}\|_2.$$

Proof.

- For $k = 0$: $\mathrm{E}W_sW_t = s \wedge t = \int_0^t 1\!\!1_{[0,s]}\, d\lambda$. The set of all linear combinations $\sum_i \alpha_i 1\!\!1_{[0,s_i]}$ is dense in $L_2[0,1]$.
- For $k > 0$: use the general result that the RKHS is “equivariant” under continous linear transformations, like integration.

$\square$

EXAMPLE — Brownian motion

Theorem. *The RKHS of k times IBM is*

$$\{f\colon f^{(k+1)} \in L_2[0,1], f(0) = \cdots = f^{(k)}(0) = 0\}, \quad \|f\|_{\mathbb{H}} = \|f^{(k+1)}\|_2.$$

Proof.

- For $k = 0$: $\mathrm{E} W_s W_t = s \wedge t = \int_0^t 1\!\!1_{[0,s]}\, d\lambda$. The set of all linear combinations $\sum_i \alpha_i 1\!\!1_{[0,s_i]}$ is dense in $L_2[0,1]$.
- For $k > 0$: use the general result that the RKHS is “equivariant” under continous linear transformations, like integration.

$\square$

Theorem. *The RKHS of the sum of k times IBM and* $t \mapsto \sum_{i=0}^k Z_i t^i$ *is*

$$\{f\colon f^{(k+1)} \in L_2[0,1]\}, \qquad \|f\|_{\mathbb{H}}^2 = \|f^{(k+1)}\|_2^2 + \sum_{i=0}^k f^{(i)}(0)^2.$$

Example — stationary processes

A stationary Gaussian process is characterized through a spectral measure μ, by

$$\operatorname{cov}(W_s, W_t) = \int e^{i\lambda^T(s-t)}\, d\mu(\lambda).$$

Theorem. *The RKHS of $(W_t : t \in T)$ is the set of real parts of the functions*

$$t \mapsto \int e^{i\lambda^T t}\psi(\lambda)\, d\mu(\lambda), \qquad \psi \in L_2(\mu),$$

with RKHS-norm

$$\|h\|_{\mathbb{H}} = \inf\{\|\psi\|_2 : h_\psi = h\}.$$

If the interior of T is nonempty and $\int e^{\|\lambda\|}\, \mu(d\lambda) < \infty$, then ψ is unique and $\|h\|_{\mathbb{H}} = \|\psi\|_2$.

Proof.

$$\mathrm{E} W_s W_t = \langle e_s, e_t\rangle_{2,\mu}, \qquad e_s(\lambda) = e^{i\lambda^T s}.$$

$\square$

Small ball probability

Definition. The small ball probability of a Gaussian random element W in $(\mathbb{B}, \|\cdot\|)$ is $\Pr(\|W\| < \epsilon)$, and the small ball exponent is

$$\phi_0(\epsilon) = -\log \Pr(\|W\| < \epsilon).$$

Small ball probability

Definition. The small ball probability of a Gaussian random element W in $(\mathbb{B}, \|\cdot\|)$ is $\Pr(\|W\| < \epsilon)$, and the small ball exponent is

$$\phi_0(\epsilon) = -\log \Pr(\|W\| < \epsilon).$$

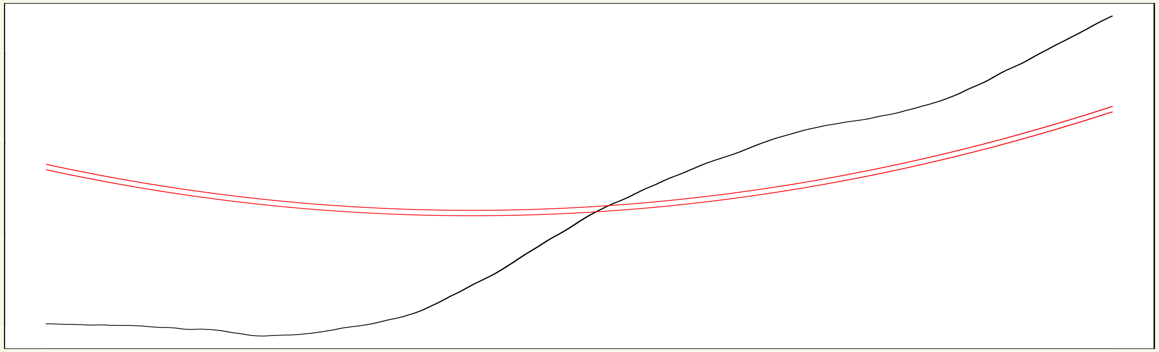
EXAMPLES

- Brownian motion: $\phi_0(\epsilon) \asymp (1/\epsilon)^2$.

Small ball probability

Definition. The small ball probability of a Gaussian random element W in $(\mathbb{B}, \|\cdot\|)$ is $\Pr(\|W\| < \epsilon)$, and the small ball exponent is

$$\phi_0(\epsilon) = -\log \Pr(\|W\| < \epsilon).$$



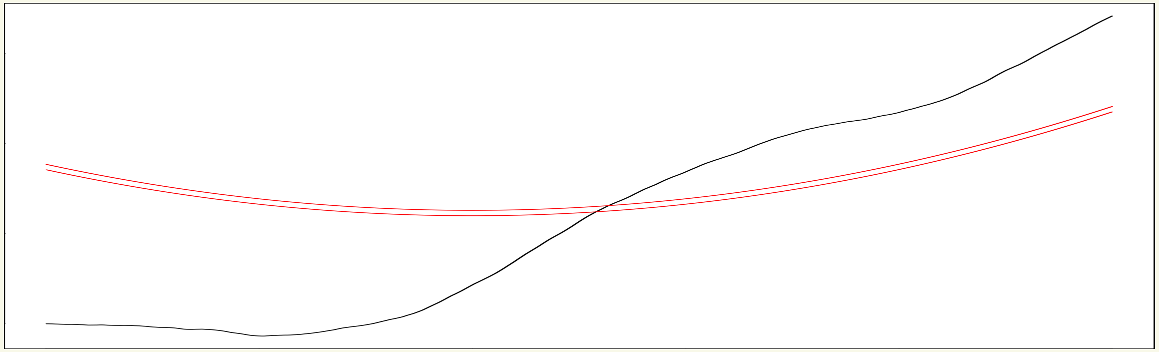
EXAMPLES

- Brownian motion: $\phi_0(\epsilon) \asymp (1/\epsilon)^2$.
- $\alpha - 1/2$ times integrated BM: $\phi_0(\epsilon) \asymp (1/\epsilon)^{1/\alpha}$.

Small ball probability

Definition. The small ball probability of a Gaussian random element W in $(\mathbb{B}, \|\cdot\|)$ is $\Pr(\|W\| < \epsilon)$, and the small ball exponent is

$$\phi_0(\epsilon) = -\log \Pr(\|W\| < \epsilon).$$



EXAMPLES

- Brownian motion: $\phi_0(\epsilon) \asymp (1/\epsilon)^2$.
- $\alpha - 1/2$ times integrated BM: $\phi_0(\epsilon) \asymp (1/\epsilon)^{1/\alpha}$.
- Radial basis: $\phi_0(\epsilon) \lesssim \bigl(\log(1/\epsilon)\bigr)^{1+d}$.

Small ball probability

Definition. The small ball probability of a Gaussian random element W in $(\mathbb{B}, \|\cdot\|)$ is $\Pr(\|W\| < \epsilon)$, and the small ball exponent is

$$\phi_0(\epsilon) = -\log \Pr(\|W\| < \epsilon).$$

Small ball probabilities can be computed either by probabilistic arguments, or analytically from the RKHS.

Theorem.

$$\phi_0(\epsilon) \asymp \log N\Big(\frac{\epsilon}{\sqrt{\phi_0(\epsilon)}}, \mathbb{H}_1, \|\cdot\|\Big)$$

EXAMPLE
RKHS of Brownian motion is Sobolev space of first order.
Unit ball has entropy $1/\epsilon$ for uniform norm.

$$\frac{1}{\epsilon^2} \asymp \log N\Big(\frac{\epsilon}{\sqrt{(1/\epsilon)^2}}, \mathbb{H}_1, \|\cdot\|\Big)$$

.

Posterior contraction rates for Gaussian priors

Prior W is centered Gaussian map in Banach space $(\mathbb{B}, \|\cdot\|)$ with RKHS $(\mathbb{H}, \|\cdot\|_{\mathbb{H}})$ and small ball exponent

$$\phi_0(\epsilon) = -\log \Pi(\|W\| < \epsilon).$$

Theorem. *If statistical distances on the model combine appropriately with the norm $\|\cdot\|$ of $\mathbb{B}$, then the posterior rate is ϵ_n if*

$$\phi_0(\epsilon_n) \le n{\epsilon_n}^2 \qquad \textit{AND} \qquad \inf_{h \in \mathbb{H}: \|h - w_0\| < \epsilon_n} \|h\|_{\mathbb{H}}^2 \le n{\epsilon_n}^2.$$

Posterior contraction rates for Gaussian priors

Prior W is centered Gaussian map in Banach space $(\mathbb{B}, \|\cdot\|)$ with RKHS $(\mathbb{H}, \|\cdot\|_{\mathbb{H}})$ and small ball exponent

$$\phi_0(\epsilon) = -\log \Pi(\|W\| < \epsilon).$$

Theorem. *If statistical distances on the model combine appropriately with the norm $\|\cdot\|$ of $\mathbb{B}$, then the posterior rate is ϵ_n if*

$$\phi_0(\epsilon_n) \le n{\epsilon_n}^2 \qquad \textit{AND} \qquad \inf_{h \in \mathbb{H}: \|h - w_0\| < \epsilon_n} \|h\|_{\mathbb{H}}^2 \le n{\epsilon_n}^2.$$

- *Both inequalities give lower bound on ϵ_n.*
- *The first depends on W and not on w_0.*
- *If $w_0 \in \mathbb{H}$, then second inequality is satisfied for $\epsilon_n \gtrsim 1/\sqrt{n}$.*

Density estimation

As prior on density p use p_W for:

$$p_w(x) = \frac{e^{w_x}}{\int_0^1 e^{w_t}\,dt}.$$

Density estimation

As prior on density p use p_W for:

$$p_w(x) = \frac{e^{w_x}}{\int_0^1 e^{w_t}\,dt}.$$

Lemma. $\forall v, w$

- $h(p_v, p_w) \leq \|v - w\|_\infty \, e^{\|v-w\|_\infty/2}$
- $K(p_v, p_w) \lesssim \|v - w\|_\infty^2 \, e^{\|v-w\|_\infty}(1 + \|v - w\|_\infty)$
- $V(p_v, p_w) \lesssim \|v - w\|_\infty^2 \, e^{\|v-w\|_\infty}(1 + \|v - w\|_\infty)^2$

Settings

Density estimation
$X_1, \ldots, X_n$ iid in $[0,1]$,

$$p_\theta(x) = \frac{e^{\theta(x)}}{\int_0^1 e^{\theta(t)}\,dt}.$$

- Distance on parameter: Hellinger on p_θ.
- Norm on W: uniform.

Classification
$(X_1,Y_1),\ldots,(X_n,Y_n)$ iid in $[0,1]\times\{0,1\}$

$$\Pr_\theta(Y=1|\,X=x) = \frac{1}{1+e^{-\theta(x)}}.$$

- Distance on parameter: $L_2(G)$ on $\Pr_\theta$.
  (G marginal of X_i.)
- Norm on W: $L_2(G)$.

Regression
$Y_1,\ldots,Y_n$ independent $N(\theta(x_i),\sigma^2)$, for fixed design points $x_1,\ldots,x_n$.

- Distance on parameter: empirical L_2-distance on θ.
- Norm on W: empirical L_2-distance.

Ergodic diffusions
$(X_t\colon t\in[0,n])$, ergodic, recurrent:

$$dX_t = \theta(X_t)\,dt + \sigma(X_t)\,dB_t.$$

- Distance on parameter: random Hellinger h_n ($\approx \|\cdot/\sigma\|_{\mu_0,2}$).
- Norm on W: $L_2(\mu_0)$.
  (μ_0 stationary measure.)

Brownian Motion — rate calculation

- Small ball probability:

$$\phi_0(\epsilon) \asymp (1/\epsilon)^2 \leq n\epsilon^2 \text{ implies } \epsilon \geq n^{-1/4}.$$

- Approximation: if $w_0 \in C^\beta[0,1]$, $\beta \leq 1$,

$$\inf_{h \in \mathbb{H}: \|h - w_0\|_\infty < \epsilon} \|h'\|_2^2 \lesssim \epsilon^{-(2-2\beta)/\beta}$$

(Attained for $h = w_0 * \phi_\sigma$ with $\sigma \asymp \epsilon^{1/\beta}$.)

$$\epsilon^{-(2-2\beta)/\beta} \leq n\epsilon^2 \text{ implies } \epsilon \geq n^{-\beta/2}.$$

Contraction rate is the slowest of the two rates.

Example — radial basis stationary process

- Small ball pobabilility:

$$\phi_0(\epsilon) \asymp \big(\log(1/\epsilon)\big)^2 \le n\epsilon^2 \text{ implies } \epsilon \ge n^{-1/2}(\log n)^2.$$

- Approximation: since $\delta\mu(\lambda) = e^{-\lambda^2}\, d\lambda$:

$$w_0(t) = \int e^{it^T\lambda}\hat{w}_0(\lambda)\, d\lambda = \int e^{it^T\lambda}\hat{w}_0(\lambda)e^{\lambda^2}\, d\mu(\lambda).$$

  If the red function is in $L_2(\mu)$, then $w_0 \in \mathbb{H}$. Otherwise approximate it by $\psi(\lambda) = \hat{w}_0(\lambda)e^{\lambda^2}1\!\!1\{|\lambda| \le M\}$. Optimize over M.

Contraction rate is the slowest of the two rates, typically the second.

Posterior contraction rates for Gaussian priors

Prior W centered Gaussian map in Banach space $(\mathbb{B}, \|\cdot\|)$ with RKHS $(\mathbb{H}, \|\cdot\|_{\mathbb{H}})$ and small ball exponent

$$\phi_0(\epsilon) = -\log \Pi(\|W\| < \epsilon).$$

Theorem. *If statistical distances on the model combine appropriately with the norm $\|\cdot\|$ of $\mathbb{B}$, then the posterior rate is ϵ_n if*

$$\phi_0(\epsilon_n) \leq n{\epsilon_n}^2 \qquad \textit{AND} \qquad \inf_{h \in \mathbb{H}: \|h - w_0\| < \epsilon_n} \|h\|_{\mathbb{H}}^2 \leq n{\epsilon_n}^2.$$

Posterior contraction rates for Gaussian priors

Prior W centered Gaussian map in Banach space $(\mathbb{B}, \|\cdot\|)$ with RKHS $(\mathbb{H}, \|\cdot\|_{\mathbb{H}})$ and small ball exponent

$$\phi_0(\epsilon) = -\log \Pi(\|W\| < \epsilon).$$

Theorem. *If statistical distances on the model combine appropriately with the norm $\|\cdot\|$ of $\mathbb{B}$, then the posterior rate is ϵ_n if*

$$\phi_0(\epsilon_n) \leq n{\epsilon_n}^2 \qquad \textit{AND} \qquad \inf_{h\in\mathbb{H}:\|h-w_0\|<\epsilon_n} \|h\|_{\mathbb{H}}^2 \leq n{\epsilon_n}^2.$$

Proof. Suffices: existence of $\mathbb{B}_n \subset \mathbb{B}$ with

- $\log N\big(\epsilon_n, \mathbb{B}_n, \|\cdot\|\big) \leq n\epsilon_n^2$ complexity
- $\Pi_n(\mathbb{B}_n) = 1 - o(e^{-n\epsilon_n^2})$ remaining mass
- $\Pi_n\big(w: \|w - w_0\| < \epsilon_n\big) \geq e^{-n\epsilon_n^2}$ prior mass

Posterior contraction rates for Gaussian priors

Prior W centered Gaussian map in Banach space $(\mathbb{B}, \|\cdot\|)$ with RKHS $(\mathbb{H}, \|\cdot\|_{\mathbb{H}})$ and small ball exponent

$$\phi_0(\epsilon) = -\log \Pi(\|W\| < \epsilon).$$

Theorem. *If statistical distances on the model combine appropriately with the norm $\|\cdot\|$ of $\mathbb{B}$, then the posterior rate is ϵ_n if*

$$\phi_0(\epsilon_n) \le n{\epsilon_n}^2 \qquad \textit{AND} \qquad \inf_{h \in \mathbb{H}: \|h - w_0\| < \epsilon_n} \|h\|_{\mathbb{H}}^2 \le n{\epsilon_n}^2.$$

Proof. Suffices: existence of $\mathbb{B}_n \subset \mathbb{B}$ with

- $\log N\bigl(\epsilon_n, \mathbb{B}_n, \|\cdot\|\bigr) \le n\epsilon_n^2$ complexity
- $\Pi_n(\mathbb{B}_n) = 1 - o(e^{-n\epsilon_n^2})$ remaining mass
- $\Pi_n\bigl(w\colon \|w - w_0\| < \epsilon_n\bigr) \ge e^{-n\epsilon_n^2}$ prior mass

Take $\mathbb{B}_n = M_n \mathbb{H}_1 + \epsilon_n \mathbb{B}_1$ for appropriate M_n. $\square$

Prior mass — decentered small ball probability

W a centered Gaussian map in $(\mathbb{B}, \|\cdot\|)$ with RKHS $(\mathbb{H}, \|\cdot\|_{\mathbb{H}})$ and small ball exponent $\phi_0(\epsilon)$.

$$\phi_{w_0}(\epsilon) := \phi_0(\epsilon) + \frac{1}{2} \inf_{h \in \mathbb{H}: \|h - w_0\| < \epsilon} \|h\|_{\mathbb{H}}^2$$

Prior mass — decentered small ball probability

W a centered Gaussian map in $(\mathbb{B}, \|\cdot\|)$ with RKHS $(\mathbb{H}, \|\cdot\|_{\mathbb{H}})$ and small ball exponent $\phi_0(\epsilon)$.

$$\phi_{w_0}(\epsilon) := \phi_0(\epsilon) + \frac{1}{2} \inf_{h \in \mathbb{H}: \|h - w_0\| < \epsilon} \|h\|_{\mathbb{H}}^2$$

Theorem.

$$\Pr(\|W - w_0\| < 2\epsilon) \geq e^{-\phi_{w_0}(\epsilon)}$$

Prior mass — decentered small ball probability — proof

Proof. (Sketch)

- For $h \in \mathbb{H}$ the distribution of $W + h$ is absolute continuous relative to that of W and

$$\Pr\bigl(\|W - h\| < \epsilon\bigr) = \mathrm{E} e^{-Uh - \frac{1}{2}\|h\|_{\mathbb{H}}^2} 1\!\!1\{\|W\| < \epsilon\}.$$

The left side does not change if $-h$ replaces h. Take average:

$$\Pr\bigl(\|W - h\| < \epsilon\bigr) = \mathrm{E}\tfrac{1}{2}(e^{-Uh} + e^{Uh}) e^{-\frac{1}{2}\|h\|_{\mathbb{H}}^2} 1\!\!1\{\|W\| < \epsilon\}$$
$$\geq e^{-\frac{1}{2}\|h\|_{\mathbb{H}}^2} \Pr(\|W\| < \epsilon).$$

Prior mass — decentered small ball probability — proof

Proof. (Sketch)

- For $h \in \mathbb{H}$ the distribution of $W + h$ is absolute continuous relative to that of W and

$$\Pr\bigl(\|W - h\| < \epsilon\bigr) = \mathrm{E}e^{-Uh - \frac{1}{2}\|h\|_{\mathbb{H}}^2} 1\!\!1\{\|W\| < \epsilon\}.$$

  The left side does not change if $-h$ replaces h. Take average:

$$\begin{aligned}\Pr\bigl(\|W - h\| < \epsilon\bigr) &= \mathrm{E}\tfrac{1}{2}(e^{-Uh} + e^{Uh})e^{-\frac{1}{2}\|h\|_{\mathbb{H}}^2} 1\!\!1\{\|W\| < \epsilon\} \\ &\geq e^{-\frac{1}{2}\|h\|_{\mathbb{H}}^2} \Pr(\|W\| < \epsilon).\end{aligned}$$

- For general w_0: if $h \in \mathbb{H}$ with $\|w_0 - h\| < \epsilon$, then $\|W - h\| < \epsilon$ implies $\|W - w_0\| < 2\epsilon$.

$\square$

Complexity and remaining mass

Theorem. *The closure of $\mathbb{H}$ in $\mathbb{B}$ is support of the Gaussian measure (and hence posterior is inconsistent if $\|w_0 - \mathbb{H}\| > 0$).*

Complexity and remaining mass

Theorem. *The closure of $\mathbb{H}$ in $\mathbb{B}$ is support of the Gaussian measure (and hence posterior is inconsistent if $\|w_0 - \mathbb{H}\| > 0$).*

Theorem (Borell 75). *For $\mathbb{H}_1$ and $\mathbb{B}_1$ the unit balls of RKHS and $\mathbb{B}$,*

$$\Pr(W \notin M\mathbb{H}_1 + \epsilon\mathbb{B}_1) \leq 1 - \Phi\Big(\Phi^{-1}(e^{-\phi_0(\epsilon)}) + M\Big)$$

Complexity and remaining mass

Theorem. *The closure of $\mathbb{H}$ in $\mathbb{B}$ is support of the Gaussian measure (and hence posterior is inconsistent if $\|w_0 - \mathbb{H}\| > 0$).*

Theorem (Borell 75). *For $\mathbb{H}_1$ and $\mathbb{B}_1$ the unit balls of RKHS and $\mathbb{B}$,*

$$\Pr(W \notin M\mathbb{H}_1 + \epsilon\mathbb{B}_1) \le 1 - \Phi\big(\Phi^{-1}(e^{-\phi_0(\epsilon)}) + M\big)$$

Corollary. *For $M(W)$ a median of $\|W\|$ and $\sigma^2(W) = \sup_{\|b^*\|\le 1} \operatorname{var} b^* W$,*

$$\Pr(W - M(W) \ge x) \le 1 - \Phi(x/\sigma(W)) \le e^{-\frac{1}{2}x^2/\sigma^2(W)}$$

Adaptation

Every Gaussian prior is good for some regularity class, but may be very bad for another.

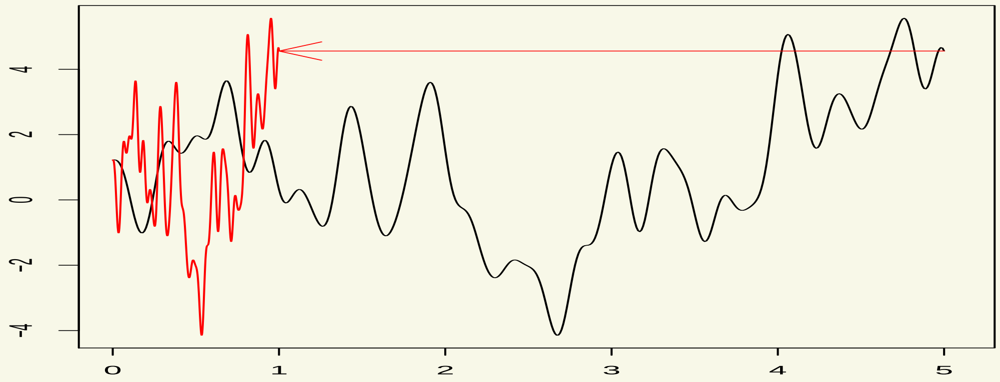
This can be alleviated by adapting the prior to the data by

- *hierarchical Bayes:* putting a prior on the regularity, or on a scaling.
- *empirical Bayes:* using a regularity or scaling determined by maximum likelihood on the marginal distribution of the data.

The first is known to work in some generality.
For the second there are some, but not many results.

Adaptation by random scaling — example

- Choose A^d from a Gamma distribution.
- Choose $(G_t : t \in \mathbb{R}_+^d)$ “radial basis” stationary Gaussian process.
- Set $W_t \sim G_{At}$.



Theorem. • *if $w_0 \in C^\beta[0,1]^d$, then the rate of contraction is nearly* $n^{-\beta/(2\beta+d)}$.

- *if w_0 is supersmooth, then the rate is nearly $n^{-1/2}$.*

Proof. Use the basic contraction theorem (and careful estimates). □

Acknowledgement

Harry van Zanten

Dirichlet mixtures

Acknowledgement

Subhashis Ghosal

Dirichlet mixtures

$$p_{F,\sigma}(x) = \int \sigma^{-1}\phi\big((x-z)/\sigma\big)\, dF(z).$$

$$X_1, \ldots, X_n |\, F, \sigma \overset{\text{iid}}{\sim} p_{F,\sigma}, \qquad F \sim \mathrm{DP}(\alpha) \perp \sigma \sim \pi.$$

Dirichlet mixtures

$$p_{F,\sigma}(x) = \int \sigma^{-1}\phi\big((x-z)/\sigma\big)\,dF(z).$$

$$X_1,\ldots,X_n|\,F,\sigma \overset{\text{iid}}{\sim} p_{F,\sigma}, \qquad F\sim \mathrm{DP}(\alpha)\perp \sigma\sim\pi.$$

Two cases for the true density p_0:

- Supersmooth: $p_0 = p_{F_0,\sigma_0}$, for some F_0, $\sigma_0>0$.
  Take prior for σ with continuous positive density on $(a,b)\ni\sigma_0$.

Dirichlet mixtures

$$p_{F,\sigma}(x) = \int \sigma^{-1}\phi\big((x-z)/\sigma\big)\,dF(z).$$

$$X_1,\ldots,X_n|\,F,\sigma \stackrel{\text{iid}}{\sim} p_{F,\sigma}, \qquad F \sim \mathrm{DP}(\alpha) \perp \sigma \sim \pi.$$

Two cases for the true density p_0:

- Supersmooth: $p_0 = p_{F_0,\sigma_0}$, for some F_0, $\sigma_0 > 0$.
  Take prior for σ with continuous positive density on $(a,b) \ni \sigma_0$.
- Ordinary smooth: p_0 has β derivatives.
  Take $1/\sigma$ a priori Gamma distributed.

Dirichlet mixtures

$$p_{F,\sigma}(x) = \int \sigma^{-1}\phi\big((x-z)/\sigma\big)\,dF(z).$$

$$X_1,\ldots,X_n|\,F,\sigma \overset{\text{iid}}{\sim} p_{F,\sigma}, \qquad F \sim \mathrm{DP}(\alpha) \perp \sigma \sim \pi.$$

Two cases for the true density p_0:

- Supersmooth: $p_0 = p_{F_0,\sigma_0}$, for some F_0, $\sigma_0 > 0$.
  Take prior for σ with continuous positive density on $(a,b) \ni \sigma_0$.
- Ordinary smooth: p_0 has β derivatives.
  Take $1/\sigma$ a priori Gamma distributed.

Compare to kernel density estimation

$$\frac{1}{n\sigma}\sum_{i=1}^{n}\phi\Big(\frac{x-X_i}{\sigma}\Big) = p_{\mathbb{F}_n,\sigma}(x).$$

Supersmooth truth

$$p_{F,\sigma}(x)=\int\sigma^{-1}\phi\bigl((x-z)/\sigma\bigr)\,dF(z).$$

$$X_1,\ldots,X_n|\,F,\sigma\overset{\text{iid}}{\sim}p_{F,\sigma},\qquad F\sim\mathrm{DP}(\alpha)\quad\perp\quad\sigma\sim\pi.$$

Theorem. *If* $p_0=p_{F_0,\sigma_0}$, *where*

- F_0 *has compact support* K,
- α *has a positive density on an open set* $G\supset K$,
- $\alpha(|z|>t)\lesssim e^{-C|t|^\delta}$ *for all* $t>0$, *some* $C>0,\delta>0$,
- π *has a continuous positive density on* $(a,b)\ni\sigma_0$,

then for some $M,\kappa>0$,

$$P_0^n\Pi\left(F,\sigma\colon h(p_{F,\sigma},p_0)>M\frac{(\log n)^\kappa}{\sqrt{n}}\,|\,X_1,\ldots,X_n\right)\to 0.$$

Ordinary smooth truth

$$p_{F,\sigma}(x) = \int \sigma^{-1}\phi\big((x-z)/\sigma\big)\, dF(z).$$

$$X_1,\ldots,X_n|\, F,\sigma \overset{\text{iid}}{\sim} p_{F,\sigma}, \qquad F \sim \mathrm{DP}(\alpha) \quad \perp \quad \sigma^{-1} \sim \Gamma(s,t).$$

Ordinary smooth truth

$$p_{F,\sigma}(x) = \int \sigma^{-1}\phi\big((x-z)/\sigma\big)\, dF(z).$$

$$X_1, \ldots, X_n |\, F, \sigma \overset{\text{iid}}{\sim} p_{F,\sigma}, \qquad F \sim \mathrm{DP}(\alpha) \quad \perp \quad \sigma^{-1} \sim \Gamma(s,t).$$

Let “β-smooth” mean:

$$\left|p^{(\underline{\beta})}(x) - p^{(\underline{\beta})}\right| \leq L(x)|y|^{\beta-\underline{\beta}},$$

for L satisfying, for $\beta' > \beta$,

$$P_0\Big(\frac{p^{(\underline{\beta})}}{p_0}\Big)^{2\beta'/\underline{\beta}} < \infty, \qquad P_0\Big(\frac{L}{p_0}\Big)^{2\beta'/\beta} < \infty, \qquad |p_0(x)| \lesssim e^{-C|x|^\tau}.$$

Ordinary smooth truth

$$p_{F,\sigma}(x) = \int \sigma^{-1}\phi\big((x-z)/\sigma\big)\,dF(z).$$

$$X_1,\ldots,X_n|\,F,\sigma \overset{\text{iid}}{\sim} p_{F,\sigma}, \qquad F \sim \mathrm{DP}(\alpha) \quad \perp \quad \sigma^{-1} \sim \Gamma(s,t).$$

Theorem. *If p_0 is β-smooth and*

- α *has a positive density on* $\mathbb{R}$*,*
- $\alpha(|z|>t) \lesssim e^{-C|t|^\delta}$ *for all* $t>0$*, some* $C>0,\delta>0$*,*

then for some $M,\kappa>0$*,*

$$P_0^n\Pi\left(F,\sigma\colon h(p_{F,\sigma},p_0) > Mn^{-\beta/(2\beta+1)}(\log n)^\kappa|\,X_1,\ldots,X_n\right) \to 0.$$

Adaptation to any smoothness with a Gaussian kernel.
Compare to kernel density estimation, which needs higher order kernels.

$$\frac{1}{n\sigma}\sum_{i=1}^n \phi\Big(\frac{x-X_i}{\sigma}\Big) = p_{\mathbb{F}_n,\sigma}(x).$$

Finite approximation

Lemma. *For any probability measure F on the interval $[0,1]$ there exists a discrete probability measure F' on with at most*

$$N \lesssim \log \frac{1}{\epsilon}$$

support points, such that

$$\|p_{F,1} - p_{F',1}\|_\infty \lesssim \epsilon, \qquad \|p_{F,1} - p_{F',1}\|_1 \lesssim \epsilon\Big(\log \frac{1}{\epsilon}\Big)^{1/2}.$$

Proof.

- Match moments of F and F' up to order $\log(1/\epsilon)$.
- Taylor expand the kernel $z \mapsto \phi(x-z)$.

$\square$

Prior mass

Lemma. *Let* $z_j \in U_j$ *for partition* $\mathbb{R} = \cup_{j=0}^N U_j$. *Then for* $F' = \sum_{j=1}^N p_j \delta_{z_j}$ *and any* F,

$$\|p_{F,\sigma} - p_{F',\sigma}\|_1 \quad \lesssim \quad \frac{1}{\sigma} \max_{1 \le j \le N} \lambda(U_j) + \sum_{j=1}^N |F(U_j) - p_j|.$$

By properties of finite-dimensional Dirichlet can bound prior probability that right side is smaller than ϵ

Entropy

For $b_1 < b_2$, $\tau < 1/4$ and $a \geq e$ let

$$\mathcal{P}_{a,\tau} = \left\{p_{F,\sigma}\colon F[-a,a] = 1,\ b_1\tau \leq \sigma \leq b_2\tau\right\}.$$

Theorem. *For $0 < \epsilon < 1/2$ and d the L_1-norm or Hellinger distance*

$$\log N(\epsilon, \mathcal{P}_{a,\tau}, d) \leq C_{b_1,b_2} \frac{a}{\tau}\Big(\log \frac{1}{\epsilon}\Big)\Big(\log \frac{a}{\epsilon\tau}\Big).$$

Proof.

- Partition $[-a, a]$ into $(1/\sigma)$ equal length intervals.
- On each interval approximate with discrete distribution with $\lesssim \log(1/\epsilon)$ support points.
- Use bounds on entropy in Euclidean space.

$\square$

Approximation

Under some regularity conditions on p_0, as $\sigma \to 0$.

$$d(p_{P_0,\sigma}, p_0) = d(\phi_\sigma * p_0, p_0) = O(\sigma^2).$$

Hence an ϵ-ball around $p_{P_0,\sigma}$ is contained in $\epsilon + \sigma$ ball around p_0, and prior mass condition can be verified.

Approximation

Under some regularity conditions on p_0, as $\sigma \to 0$.

$$d(p_{P_0,\sigma}, p_0) = d(\phi_\sigma * p_0, p_0) = O(\sigma^2).$$

Hence an ϵ-ball around $p_{P_0,\sigma}$ is contained in $\epsilon + \sigma$ ball around p_0, and prior mass condition can be verified.

This works, but only for smoothness up to 2.

Approximation

Under some regularity conditions on p_0, as $\sigma \to 0$.

$$d(p_{P_0,\sigma}, p_0) = d(\phi_\sigma * p_0, p_0) = O(\sigma^2).$$

Hence an ϵ-ball around $p_{P_0,\sigma}$ is contained in $\epsilon + \sigma$ ball around p_0, and prior mass condition can be verified.

This works, but only for smoothness up to 2.

For general result need to choose more clever approximations than $p_{P_0,\sigma}$.

All the rest

All the rest

- Adaptation
- Distributional approximation
- Survival analysis
- Credible sets
- Sparsity
- Inverse problems
- Structures

A few names names I should have mentioned..

- Dirichlet process: Ferguson, Lo, Antoniak, and many others.
- Consistency: Schwartz, Barron.
- Tests: Le Cam, Birgé.
- Frequentist Bayes: Ghosal, vdV.
- Gaussian variables in Banach spaces: Borell, Kuelbs, Li, Lifshitz.
- Gaussian process priors: van Zanten, vdV.
- Dirichlet mixtures: Ghosal, Kruijer, Rousseau, W. Shen, Tokdar, vdV.

Further reading:
Subhashis Ghosal, Aad van der Vaart:
Fundamentals of Nonparametric Bayesian Inference
Cambridge University Press, 2013(?)