Dirichlet Process, Related Priors and Posterior Asymptotics

Subhashis Ghoshal

Department of Statistics
North Carolina State University

http://www.stat.ncsu.edu/∼ghoshal/
Email: ghoshal@stat.ncsu.edu

Bayesian Nonparametric Regression, Cambridge, Aug 6-10

NC State University
DEPARTMENT OF STATISTICS

Table of Contents

- Introduction and goals
- Dirichlet Process
  - Construction
  - Basic properties
  - Dirichlet mixtures
- Posterior consistency
  - Examples of inconsistency
  - Kullback-Leibler property
  - Entropy and exponentially consistent tests
  - Applications to Dirichlet mixtures, Polya trees and Gaussian process prior

  - Semiparametric applications
  - Alternative approaches (Doob's theorem, tail-free, martingale)
- Rate of convergence
- Illustrations
- Adaptation and model selection
- Bernstein-von Mises theorems

Bayesian nonparametrics

Data $(X_1, \ldots, X_n) \sim P_\theta^{(n)}$ modeled by some infinite dimensional parameter θ.

Make Bayesian inference on cdf, density, regression function, spectral density etc.

- Construct "default" prior distribution Π on appropriate spaces
- Compute posterior distribution $\Pi(\theta|X_1, \ldots, X_n)$ (typically through special computing devices like MCMC)
- Establish good (or at least reasonable) frequentist properties, typically through asymptotics

Consistency: “What if” study — If the data are generated from a model with true parameter θ_0, then the posterior $\Pi(\theta|X_1, \ldots, X_n)$ should concentrate around θ_0 in the desired sense, that is, posterior probability of an ϵ-neighborhood should go to one.

Rate of convergence: The size ϵ_n of the smallest shrinking ball which still retains most of the posterior probability:
$\Pi(d(\theta, \theta_0) \geq M_n\epsilon_n|X_1, \ldots, X_n) \to 0$ a.s./probability for any $M_n \to \infty$.

The Dirichlet process

Problem of estimation cdf (probability law) governing i.i.d. data.

Grouping leads to finite dimensional multinomial.

Dirichlet distribution on the simplex is the natural conjugate prior.

Definition (Ferguson): A random measure P on $(\mathfrak{M}, \mathscr{M})$ is said to have a Dirichlet process prior $\mathcal{D}_\alpha$ with base measure α if for every finite measurable partition $A_1, \ldots, A_k$ of $\mathfrak{X}$,

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

where $\alpha(\cdot)$ is a finite positive Borel measure on $\mathfrak{X}$.

Why a measure? Allows unambiguous specification: If some sets in the partition are merged together, resulting probabilities follow lower dimensional Dirichlet with parameters group sums.

Basic Properties of a Dirichlet Process

1. $(P(A), P(A^c)) \sim \mathrm{Dir}(2; \alpha(A), \alpha(A^c))$, that is, $P(A) \sim \mathrm{Be}(\alpha(A), \alpha(A^c))$.

   - $\mathrm{E}(P(A)) = \frac{\alpha(A)}{\alpha(\mathbb{R})} =: \bar{\alpha}(A)$
     (Implication: If $X \sim P$ and $P \sim \mathcal{D}_\alpha$, then marginally $X \sim \bar{\alpha}$
   - $\mathrm{var}(P(A)) = \frac{\bar{\alpha}(A)(1-\bar{\alpha}(A))}{1+|\alpha|}$
   - $\mathrm{E}\left(\int \psi dP\right) = \int \psi d\bar{\alpha}$

   Comment: $\bar{\alpha}$ is called the center measure and $|\alpha|$ is called the precision.

2. $P \sim \mathcal{D}_\alpha$ on $\mathfrak{X}$ and $\psi : \mathfrak{X} \to \mathfrak{Y}$ implies $P\psi^{-1} \sim \mathcal{D}_{\alpha\psi^{-1}}$.

3. $P|X_1, \ldots, X_n \sim \mathcal{D}_{\alpha^*}$, where $\alpha^* = \alpha + \sum_{i=1}^n \delta_{X_i}$, gives a version of the posterior distribution.

   Sketch of Proof.

   - Can just do $n = 1$ and iterate.
   - Reduce to a finite measurable partition $\{A_1, \ldots, A_k\}$. To show its posterior is $\mathrm{Dir}(k; \alpha(A_1), \ldots, \alpha(A_{i-1}), \alpha(A_i) + 1, \alpha(A_{i+1}), \ldots, \alpha(A_k))$, when $X \in A_i$.
   - Clearly true if only that much were known.
   - Refine partition and corresponding information. The posterior does not change.
   - Apply the martingale convergence theorem to pass to the limit.

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

   - Convex combination, relative weights $|\alpha|$ and n
   - $|\alpha|$ can be interpreted as prior sample size, $|\alpha| \to 0$ noninformative limit (the Bayesian bootstrap)
   - Asymptotically equivalent to sample mean upto $O(n^{-1})$, converges to true P_0 at $n^{-1/2}$ rate.

5. $\mathrm{var}(P(A)|X_1,\ldots,X_n) \leq 1/(4n) \to 0$

6. $\Pi(P : |P(A) - P_0(A)| \geq M_n n^{-1/2}|X_1,\ldots,X_n)$, by Chebyshev, bounded by $M_n^{-2} n\mathrm{E}((P(A) - P_0(A))^2|X_1,\ldots,X_n) \to 0$ a.s. under P_0

Construction of a Dirichlet process

1. Naive construction

   - Embed $\mathfrak{M}$ in $[0,1]^{\mathscr{B}}$. Finite dimensional joint distributions are consistently specified, so a unique infinite dimensional joint distribution exists on $[0,1]^{\mathscr{B}}$.
   - (If the constructed measure sits in $\mathfrak{M}$) Restrict to $\mathfrak{M}$.
   - BUT $\mathfrak{M}$ is not a measurable subset of $[0,1]^{\mathscr{B}}$.
   - Is outer measure one? That is, is countable additivity holds except on a null set?
   - For any fixed sequence of measurable sets $B_n \downarrow \varnothing$, $P(B_n) \downarrow 0$ a.s.
   - BUT too many sequences, too many null sets. No way to form a grand null set.

2. Construction through a countable generator

   - Assume space is $\mathbb{R}$ (wlog) and approach through cdf at points in $\mathbb{Q}$.
   - Joint law of these sits in $[0,1]^{\aleph_0}$, much more manageable.
   - Thus $\mathfrak{M}$ can be identified as a measurable subset
   - Problem of null sets disappear

3. Construction through a gamma process

- Assume space is $(0, \infty)$ (wlog) and let F be the cdf.
- Consider J a gamma process, the unique independent increment process with $J(t) \sim \mathrm{Ga}(\alpha(0, t], 1)$.
- Define $F(t) = J(t)/J(\infty)$ and induce a prior through the gamma process.
- Intimate relation between independent gamma variables and Dirichlet distributions give the desired finite dimensional laws.

Further Properties of a Dirichlet Process

1. Discreteness:

   $\mathcal{D}_\alpha(\mathfrak{D}) := \mathcal{D}_\alpha(P : P \text{ is discrete}) = 1.$

   Sketch of the proof.

   - $P \in \mathfrak{D}$ iff $P\{x : P(\{x\}) > 0\} = 1.$
   - $\mathfrak{D}$ is a measurable subset of $\mathfrak{M}$.
   - Assertion holds iff $(\mathcal{D}_\alpha \times P)((P, X) : P\{X\} > 0) = 1.$
   - Equivalent to $(\bar{\alpha} \times \mathcal{D}_{\alpha+\delta_X})((X, P) : P\{X\} > 0) = 1.$
   - True by Fubini since P has positive mass at X.

   Note: Null sets do matter.

2. Self-similarity:
   $P(A) \perp P|_A \perp P|_{A^c}$ and $P|_A \sim \mathcal{D}_{\alpha(A)\bar{\alpha}|_A}$.
   Can be shown using the relations between finite dimensional Dirichlet and independent gamma variables.

   If the Dirichlet process is localized to A and A^c, then both are Dirichlet and are independent of each other, and also of the "macro level" variable $P(A)$. Thus at any locality, mass is distributed according to a Dirichlet process, independent of what happens to the "outside world", that is locally a Dirichlet process is like itself, or is "self similar" like a fractal.

3. Weak Support:

   $\text{supp}(\mathcal{D}_\alpha) = \{P^* : \text{supp}(P^*) \subset \text{supp}(\alpha)\}$

   Sketch of the Proof.

   - No P^* which supports points outside $\text{supp}(\alpha)$ can be in the support since the corresponding first beta parameter would be zero.
   - Any compliant P^* would be in the weak support by fine partitioning and nonsingularity of corresponding finite dimensional Dirichlet distribution.
   - Further, if P^* in the weak support is continuous, then assertion automatically upgrades to Kolmorov-Smirnov support by Polya's theorem.

4. Convergence: Let α_m are such that $\bar{\alpha}_m \to_w \bar{\alpha}$, then

   (i) If $|\alpha_m| \to M$, $0 < M < \infty$, then $\mathcal{D}_{\alpha_m} \to_w \mathcal{D}_{M\bar{\alpha}}$;

   (ii) If $|\alpha_m| \to 0$, then $\mathcal{D}_{\alpha_m} \to_w \mathcal{D}^*_{\bar{\alpha}} := \mathcal{L}(\delta_X : X \sim \bar{\alpha})$;

   (iii) If $|\alpha_m| \to \infty$, then $\mathcal{D}_{\alpha_m} \to_w \delta_{\bar{\alpha}}$.

   Sketch of the Proof.

   - A random measure is tight iff its expectation measure is tight, so tighness holds here. To check finite dimensional convergence.
   - Work with a finite partition.
   - For (i), Dirichlet goes to Dirichlet by Scheffe.
   - For (ii) and (iii), check convergence of all mixed moments.

   Corollary. If $X_i \overset{\text{iid}}{\sim} P_0$, then $\mathcal{D}_{\alpha+\sum_{i=1}^n \delta_{X_i}}$ weakly converges to

   (i) δ_{P_0} if $n \to \infty$ (consistency of posterior)

   (ii) $\mathcal{D}_{\sum_{i=1}^n \delta_{X_i}}$ (the Bayesian bootstrap) if $|\alpha| \to 0$.

Generalized Polya Urn Scheme

$X_1 \sim \bar{\alpha}$.

$X_2|P, X_1 \sim P$ and $P|X_1 \sim \mathcal{D}_{\alpha+\delta_{X_1}}$, so

$$X_2|X_1 \sim \frac{|\alpha|}{|\alpha|+1}\bar{\alpha} + \frac{1}{|\alpha|+1}\delta_{X_1}$$

that is,

$$X_2|X_1 \sim \begin{cases} = X_1, & \text{w.p. } \frac{1}{|\alpha|+1}, \\ \text{new draw } \sim \bar{\alpha}, & \text{w.p. } \frac{|\alpha|}{|\alpha|+1}. \end{cases}$$

More generally, $P|X_1,\ldots,X_{i-1} \sim \mathcal{D}_{\alpha+\sum_{j=1}^{i-1}\delta_{X_j}}$, so that

$$X_i|X_1,\ldots,X_{i-1} \sim \begin{cases} = X_j, & \text{w.p. } \frac{1}{|\alpha|+i-1},\ j=1,\ldots,i-1, \\ \bar{\alpha}, & \text{w.p. } \frac{|\alpha|}{|\alpha|+i-1}, \end{cases}$$

By exchangeability of $(X_1,\ldots,X_n)$,

$$X_i|X_{[-i]} \sim \begin{cases} = X_j, & \text{w.p. } \frac{1}{|\alpha|+n-1},\ j \neq i \\ \bar{\alpha}, & \text{w.p. } \frac{|\alpha|}{|\alpha|+i-1}, \end{cases}$$

Note: A lot of ties produced by Dirichlet samples, and hence induce random partitions on $\{1,\ldots,n\}$. Expected number of distinct elements grows like $|\alpha|\log(n/|\alpha|)$.

Sethuraman Representation

Let $\theta_1, \theta_2, \ldots \sim \bar{\alpha}$ and $Y_1, Y_2, \ldots \sim \mathrm{Be}(1, |\alpha|)$, all mutually independent.

Put $V_i = Y_i \prod_{j=1}^{i-1}(1 - Y_j)$, Stick breaking

$P \sim \mathcal{D}_\alpha$ can be represented as $P = \sum_{i=1}^\infty V_i \delta_{\theta_i}$.

Allows to actually simulate DP, at least approximately.

Distributional equation

$$\begin{aligned} P &= V_1 \delta_{\theta_1} + \sum_{i=2}^\infty V_i \delta_{\theta_i} \\ &= Y \delta_\theta + (1 - Y) \sum_{i=1}^\infty V_i' \delta_{\theta_i'} \\ &=_d Y \delta_\theta + (1 - Y) P \end{aligned}$$

Good for “functional MCMC”

Intuition

Since prior is posterior averaged wrt marginal distribution of the observation, $\mathcal{D}_\alpha$ is also $\int \mathcal{D}_{\alpha+\delta_x} d\bar{\alpha}(x)$.

Conditionally on observed x, $P\{x\} \sim \text{Be}(1, |\alpha|)$ and $P|_{\{x\}^c} \sim \mathcal{D}_\alpha$ independently by the self similarity property.

Steps in the formal proof

(a) DP is a solution of the distributional equation,

(b) the solution is unique.

Dirichlet Mixtures

Ferguson;

Lo

$p_F(x) = \int \psi(x, \theta) dF(\theta)$.

Eg: location mixture of the form $\sigma^{-1} k((x - \mu)/\sigma)$ form rich class

$$\int \frac{1}{\sigma} k\left(\frac{x-\mu}{\sigma}\right) f(\mu) d\mu \to f(x) \quad \text{as } \sigma \to 0$$

i.i.d. samples from p_F equivalent to the model

$$X_i | \theta_i, \varphi \overset{\text{ind}}{\sim} \psi(\cdot; \theta_i, \varphi), \quad \theta_i | F \overset{\text{iid}}{\sim} F, \quad F \sim \mathcal{D}_\alpha, \quad \varphi \sim \pi(\cdot).$$

Posterior Consistency

For any neighborhood U of the true parameter θ_0, the posterior probability $\Pi(U^c|X^{(n)}) \to 0$.

Clearly, the prior Π must support the true θ_0; otherwise very little chance of consistency (Dirichlet is an exception).

We tend to think (based on experience with the parametric situation) that if the prior puts positive probability in the neighborhood of θ_0, we must have consistency, at least when the data are i.i.d.

Not quite true in infinite dimension.

Examples of Inconsistency

1. (Freedman): Infinite multinomial; p unknown p.m.f.; true p.m.f. p_0 is $\mathrm{Geo}(\frac{1}{4})$. We can construct a prior μ which gives positive mass to every neighborhood of p_0 but the posterior concentrates at $p^* := \mathrm{Geo}(\frac{3}{4})$.

   Example is actually generic: The collection of (p, Π) which leads to consistency is topologically very small.

2. (Diaconis & Freedman): To estimate the point of symmetry θ of a symmetric density. Put normal prior on θ and symmetrized DP with Cauchy base measure on the rest of the density. Then there is a symmetric true density under which the posterior concentrates at two wrong values of θ.

3. (Kim & Lee): Consider estimating hazard function H. There are two priors Π_1, Π_2, both having prior mean equal to the true hazard H_0, Π_1 with a uniformly smaller prior variance that Π_2, such that the posterior for Π_2 is consistent but the posterior for Π_1 is inconsistent.

Schwartz Theory of Consistency

Schwartz;

Barron, Schervish & Wasserman;

Ghosal, Ghosh & Ramamoorthi

Work with the dominated case only — measures admit densities.

$$\Pi(\theta \in U^c | X^{(n)}) = \frac{\int_{U^c} \frac{p^{(n)}(X^{(n)};\theta)}{p^{(n)}(X^{(n)};\theta_0)} d\Pi(\theta)}{\int_{\Theta} \frac{p^{(n)}(X^{(n)};\theta)}{p^{(n)}(X^{(n)};\theta_0)} d\Pi(\theta)}.$$

Typical strategy: Show that the numerator is exponentially small (i.e., goes to 0 even after multiplying by some suitable e^{nc}) and the denominator is not exponentially small (i.e., goes to infinity if multiplied by any e^{nc}.)

Note: The domain of integration Θ can be replaced by any suitable subset for our purpose.

Outside a neighborhood, likelihood ratios drop exponentially, so numerator will be exponentially small. But uniformity is the real challenge. Conditions for uniform exponential decay are too strong and unrealistic.

We shall use a classical device of tests. Existence of exponentially powerful tests, a condition much weaker than uniform exponential decay of likelihood, will be seen to bound the numerator suitably.

Kullback-Leibler Property

For the denominator, restrict to i.i.d. observations and note that

$$\frac{p^{(n)}(X^{(n)})}{p_0^{(n)}(X^{(n)})} > e^{-2n\epsilon}$$

iff

$$n^{-1}\sum_{i=1}^{n}\log\frac{p(X_i)}{p_0(X_i)} > -2\epsilon,$$

which will hold ultimately under P_0 by SLLN if

$$\int p_0(x)\log\frac{p_0(x)}{p(x)}d\nu(x) < \epsilon.$$

Recall: The above quantity is the Kullback-Leibler divergence $K(p_0; p)$.

The relation defines a "neighborhood" of p_0.

Thus if $\Pi(p : K(p_0; p) < \epsilon) > 0$ for all $\epsilon > 0$, Fatou's lemma and SLLN implies that the denominator is not exponentially small.

The prior positivity condition is a fundamental requirement, known as the Kullback-Leibler Property.

It is prior positivity of Kullback-Leibler (KL) neighborhoods, rather than standard neighborhoods that matter in consistency studies.

Generally, KL neighborhoods are not easily characterized, so we typically find a subset contained in the KL neighborhood which has positive prior probability. Such a set is to be found by bounding the KL divergence at a given "good" p_0.

Also, since SLLN applies to a much wider class, the approach can be extended beyond the i.i.d. set up.

Exponentially Consistent Tests

Suppose the KL property holds at true p_0.

Suppose there is a test $\phi_n(X^{(n)})$ for testing $H_0 : p = p_0$ vs $H : p \in U^c$ such that $P_0\phi_n$ and $\sup\{P(1-\phi_n) : p \in U^c\}$ are exponentially small.

$$\Pi(U^c|X^{(n)}) \leq \phi_n + \frac{\int_{U^c}(1-\phi_n)\prod_{i=1}^n \frac{p(X_i)}{p_0(X_i)}d\Pi(p)}{\int \prod_{i=1}^n \frac{p(X_i)}{p_0(X_i)}d\Pi(p)}$$

Expected values of the first term and that of the numerator of the second term are exponentially small by testing condition.

How to find an exponentially consistent test?

[Diagram session]

What about the posterior probability of the complement $\mathcal{P}_n^c$ of the sieve?

We can choose it to have small prior probability.

Does it mean small posterior probability?

No, but if the prior probability is exponentially small, then so is the posterior probability under the KL property. Show by taking expectations (straightforward).

Thus to establish consistency, (i) verify KL property, (ii) find a sieve $\mathcal{P}_n$ which has entropy smaller than cn (iii) show that the complement of the sieve has exponentially small prior probability.

ϵ-entropy: Logarithm of the least number of ϵ-balls you need to cover a set.

Illustrations

1. Dirichlet mixture of normal

- Approximate true p_0 by $p_m = P_0|_{[-m,m]} * \phi_{h_m}$, a genuine normal mixture
- KL neighborhood around p_m includes $Q * \phi_h$, Q in a weak neighborhood of $P_0|_{[-m,m]}$, h in a neighborhood of h_m, and hence has positive prior probability.
- Argument extends to a variety of kernels, even outside location scale.
- For consistency in strong (Hellinger, L_1) topologies, consider sieve $\{Q * \phi_h : Q[-a_n, a_n] > 1 - \delta, h > \sigma_n\}$, $a_n/\sigma_n < cn$, $\bar{\alpha}[-a_n, a_n]^c < e^{-bn}$, $\Pi(h < h_n) < e^{-bn}$.

2. Polya tree process

- Definition [diagram session]
- For convenience, partitions generated by quantiles of a probability density λ
- Parameters depend only at the level: $\alpha_{\varepsilon_1\cdots\varepsilon_m} = a_m$
- If $\sum_{m=1}^{\infty} a_m^{-1} < \infty$, Polya trees admit densities
- If $\sum_{m=1}^{\infty} a_m^{-1/2} < \infty$ and $\int p_0 \log(p_0/\lambda) < \infty$, Polya tree has KL property at p_0.
- Polya tree sample paths lack smoothness, so the condition for strong consistency is somewhat stringent. But consistency of posterior mean holds only under KL property in this case.

3. Gaussian process prior

- For density estimation on a compact set I, consider a Gaussian process $\xi(t)$ and a prior on p induced by $p(x) = \exp(\xi(x))/\int_I \exp[\xi(t)]dt$
- KL neighborhood for p includes uniform neighborhood for ξ
- Gaussian process supports any function in its Reproducing Kernel Hilbert Space (RKHS)
- Two effective ways to construct a Gaussian process with large RKHS: (a) random polynomial plus integrated Wiener process; (b) allow random scaling in the covariance kernel of a stationary Gaussian process
- Smoothness of Gaussian paths is controlled giving appropriate entropy bounds
- Small probability of complement of sieves obtained by maximal inequalities for Gaussian processes

Semiparametric Applications

KL property plays a vital role. If nonparametric part is nuisance, in many cases, even a sieve is not required.

Location problem: $X \sim p_\theta = p(\cdot - \theta)$. Dirichlet does not admit densities; use Polya tree, Dirichlet mixtures or any other prior with KL property.

Interesting observation: $(\theta, p) \mapsto p_\theta$ is bi-continuous, so just need to get weak consistency for the density p_θ based on i.i.d. observations.

It can be shown that KL property is essentially preserved under location change.

Many other applications possible: Linear regression (using a non-i.i.d. generalization of Schwartz theory), exponential frailty model, Cox proportional hazard model, etc.

Alternative Approaches

1. Doob's result: Consider i.i.d. observations $X_1, X_2, \ldots$ from a model (or more generally, when there is a consistent estimator of the parameter θ). Let θ be given a prior Π.

   Consider $(\theta, X_1, X_2, \ldots)$ as random, distributed as $\Pi \times P_\theta^\infty$. Then $\mathrm{E}(f(\theta)|X_1, \ldots, X_n) \to \mathrm{E}(f(\theta)|X_1, X_2, \ldots) = f(\theta)$ a.s., and so by Fubini, consistency holds at θ a.e. $[\Pi]$.

   Null set could be large, so generally not very useful. Exception: Countable parameter space or when Π is dominated by a well known measure like the Lebesgue.

2. Tail free process (Freedman): Like the Polya tree but more general from 3 aspects: (a) partitions need not be binary, (b) independence within the same level is not required; only independence across different levels is needed, (c) distributions need not be beta.

   Assume partitions are fine enough to generate Borel sigma-field.

   If we are studying consistency for the weak topology, we can stop at a finite stage.

   Tail free (independence across different levels) imply that the posterior distribution depends only at the counts at the required level. Thus the problem reduces to finite dimensional multinomial where consistency holds.

3. Martingale method (Walker): Alternative method to obtain exponential bound for the numerator $L_n = \int_{U^c} \prod_{i=1}^{n} \frac{p(X_i)}{p_0(X_i)} d\Pi(p)$.

   Key observation: $\mathrm{E}(\sqrt{L_{n+1}/L_n} | X_1, \ldots, X_n) = 1 - d_H^2(p_0, p_n)$, where p_n is the posterior expectation when Π is restricted to U^c.

   Allows to construct a martingale and hence assess convergence properties, if p_n can be made stay away from p_0.

   Conclusion:

   For weak topology, consistency can be obtained under only KL property.

   For Hellinger neighborhoods, need the existence of an infinite partition $\{A_j : \mathrm{diam}(A_j) < \delta\}$ such that $\sum_{j=1}^{\infty} \sqrt{\Pi(A_j)} < \infty$. Condition is related to entropy and sieve condition.

Convergence rates

Conditions quantitative analogue of those in Schwartz theory for consistency.

Ghosal, Ghosh & van der Vaart

If ϵ_n is the targeted rate, verify

(i) $\Pi(K_{\epsilon_n}(p_0)) \geq e^{-n\epsilon_n^2}$

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

(iii) $\Pi(\mathcal{P}_n^c) \leq e^{-cn\epsilon_n^2}$

Applications

1. Optimal rate by bracketing:

   Cover a space of densities by $N_{[\,]}(\epsilon_j, \mathcal{P})$ many brackets.

   Normalize upper brackets to obtain a discrete approximation and let Π_j be uniform on the collection. Take a convex combination of these as the final prior. Then the convergence rate is given by $\epsilon_n : \log N_{[\,]}(\epsilon, \mathcal{P}) \leq n\epsilon^2$.

   Often bracketing numbers are essentially equal to usual covering numbers, so the device will produce optimal rates, for instance for Hölder's α smooth class of densities $[n^{-\alpha/(2\alpha+1)}]$ or monotone densities $[n^{-1/3}]$.

2. Optimal rate via splines

   Density estimation on $[0,1]$. Split $[0,1]$ into K_n equal intervals.

   Form an exponential family using the corresponding J_n many B-splines and put uniform (say) prior on the coefficients. If p_0 is C^α, spline density approximates p_0 upto $J_n^{-2\alpha}$ in KL.

   Hellinger and Euclidean is comparable, so calculations reduce to Euclidean. (local) Entropy grows like J_n while prior concentrates like $e^{-J_n(c+\log(\epsilon_n\sqrt{J_n}))}$.

   Leads to rate equations $n\epsilon_n^2 \sim J_n$, $\epsilon_n \sim J_n^{-\alpha}$ giving optimal rate $\epsilon_n \sim n^{-\alpha/(1+2\alpha)}$.

3. Dirichlet mixture of normal (Ghosal & van der Vaart)

   Rate depends on the situation.

   (a) Super-smooth case: True p_0 a normal mixture $p_0 = \int \phi_{\sigma_0}(x-z)dF_0(z)$, $\underline{\sigma} \leq \sigma \leq \overline{\sigma}$.
   Basic technique in calculation of entropy and prior concentration is finding discrete mixture approximation with only $N = O(\log \frac{1}{\epsilon})$ support points, leading to calculation in N dimension. Entropy grows like $(\log \frac{1}{\epsilon})^2$ and prior concentration rate is $\epsilon^{-N} = e^{-(\log \frac{1}{\epsilon})^2}$, leading to rate $\epsilon_n \sim n^{-1/2}(\log n)$ for most favorable situation.

   (b) Smooth case: Take a prior and scale by a sequence σ_n like $n^{-1/5}$. Approximate p_0 by a normal mixture p_0^* with bandwidth σ_n up to σ_n^2, and work with p_0^* as target. Similar strategy as before but the number of support points increases to $N = \sigma_n^{-1} \log \frac{1}{\epsilon}$.
   Rate equations $n\epsilon_n^2 = \sigma_n^{-1}(\log \frac{1}{\epsilon})^2$ and $\epsilon_n = \sigma_n^2$ leading to $\epsilon_n \sim n^{-2/5}(\log n)^{4/5}$.

Adaptation

In many examples, we can get optimal based on brackets or splines if we know smoothness level. Can a single prior give optimal rate in all classes?

If so, the prior is called rate adaptive.

Natural approach: If Π_α gives the optimal rate for class $\mathcal{C}_\alpha$, then a mixture prior $\Pi = \int_\alpha \Pi_\alpha d\lambda(\alpha)$ may give the right rate for every class.

Strategy works in many cases.

Typically models are nested with different convergence rates, especially if they are indexed by a smoothness level.

1. Infinite dimensional normal

   Belitser & Ghosal

   $(X_1, X_2, \ldots) \sim N_\infty((\theta_1, \theta_2, \ldots), \boldsymbol{I}_\infty)$.

   Model q: $\sum_{i=1}^\infty i^{2q}\theta_i^2 < \infty$.

   Prior $\theta_i \overset{\text{ind}}{\sim} N(0, i^{-(2q+1)})$.

   Rate of convergence $n^{-2q/(2q+1)}$.

   Discrete countably many models. Posterior adapts to the right smoothness level in the sense that it has convergence rate $n^{-2q/(2q+1)}$ when q is the true smoothness value, for any q.

   Also, coarser models get asymptotically small posterior probability.

2. Spline basis adaptation

   Ghosal, Lember & van der Vaart

   Huang

   Finitely many smoothness classes of densities. Construct individual spline based prior for each class.

   Plain mixture prior with positive weights is rate adaptive up to a log factor.

   Smart choice of weights (depending on n) can remove the log factor.

   A general theorem in terms of entropy bounds and prior concentration rates possible.

Model selection

Can the Bayesian device automatically assign high probability to the correct model for large samples.

If one model is simple (singleton family) and KL property holds in the model, this happens by Doob's theorem, and Schwartz type arguments. [Dass & Lee]

One model with KL property, other not, then model with KL property is chosen [Walker *et al.*] under some conditions.

If models are with different prior concentration rate and order of entropy, then the correct model is chosen [Ghosal, Lember & van der Vaart]

Bernstein-von Mises theorem

The empirical $\mathbb{F}_n$ is the "MLE". $\sqrt{n}(\mathbb{F}_n - F)$ converges weakly to a Brownian Bridge (Donsker theorem).

When the prior is Dirichlet $\sqrt{n}(F - \mathbb{F}_n)$ also converges weakly to a Brownian Bridge a.s. given the sample. [Lo]

For estimating hazard (possibly with censoring) $\sqrt{n}(H - \mathbb{H}_n)$ converges weakly to a Brownian Bridge a.s. [Kim & Lee]

Does not seem to hold if convergence rate is slower than $n^{-1/2}$ [Cox; Freedman].