Some Results in Nonparametric Bayesian Inference

Aad van der Vaart

http://www.math.vu.nl/∼aad

Vrije Universiteit Amsterdam

based on joint work with
Subhashis Ghosal (North Carolina)
Harry van Zanten (Amsterdam)
Frank van der Meulen(Amsterdam)

general purpose

Study rates of contraction of posterior distributions
for priors on infinite-dimensional models
to the "true" distribution of the observations

general purpose

Study rates of contraction of posterior distributions
for priors on infinite-dimensional models
to the “true” distribution of the observations

OUTLINE
PART 1: generalities
PART 2: Gaussian process priors

GENERALITIES

setting (1)

$n = 1, 2, \ldots$

$(\mathcal{X}^{(n)}, \mathcal{A}^{(n)}, P_\theta^{(n)} : \theta \in \Theta_n)$ experiment

(Θ_n, d_n) metric space

$X^{(n)}$ observation, law $P_{\theta_0}^{(n)}$

Π_n prior

$$\Pi_n(B|X^{(n)}) = \frac{\int_B p_\theta^{(n)}(X^{(n)})\, d\Pi_n(\theta)}{\int_{\Theta_n} p_\theta^{(n)}(X^{(n)})\, d\Pi_n(\theta)} \qquad \text{posterior}$$

Rate of contraction is at least ϵ_n if $\forall M_n \to \infty$

$$P_{\theta_0}^{(n)} \Pi_n(\theta \in \Theta_n : d_n(\theta, \theta_0) \geq M_n \epsilon_n | X^{(n)}) \to 0$$

toy problem

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

prior:

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

Find rate if $\log p_0 \in C^\alpha[0,1]$

setting (1)

$\exists \xi > 0, K > 0$ such that $\forall n$
$\exists$ metric $e_n \geq d_n$ such that $\forall \epsilon > 0$

$\forall \theta_1 \in \Theta_n$ with $d_n(\theta_1, \theta_0) > \epsilon$ $\exists$ test ϕ_n with

$$P_{\theta_0}^{(n)} \phi_n \leq e^{-Kn\epsilon^2}, \qquad \sup_{\theta \in \Theta : e_n(\theta, \theta_1) < \epsilon\xi} P_\theta^{(n)}(1 - \phi_n) \leq e^{-Kn\epsilon^2}$$

Le Cam 73,75,86, Birgé 83:
$\exists$ estimators $\hat{\theta}_n$ with $d_n(\hat{\theta}_n, \theta_0) = O_P(\epsilon_n)$ if

$$\sup_{\epsilon > \epsilon_n} \log N(\epsilon\xi, \{\theta : d_n(\theta, \theta_0) \leq \epsilon\}, e_n) \leq n\epsilon_n^2$$

rate theorem (1)

$\epsilon_n \to 0$, $\epsilon_n \gg 1/\sqrt{n}$

$\log N(\epsilon_n, \Theta_n, e_n) \leq n\epsilon_n^2$
$\Pi_n(\Theta_n^c) = o(e^{-3n\epsilon_n^2})$
$\Pi_n(B_n(\theta_0, \epsilon_n; k)) \geq e^{-n\epsilon_n^2}$

Then $P_{\theta_0}^{(n)} \Pi_n(\theta \in \Theta_n : d_n(\theta, \theta_0) \geq M_n \epsilon_n | X^{(n)}) \to 0$

$B_n(\theta_0, \epsilon; k) =$
$\Big\{\theta \in \Theta : K(p_{\theta_0}^{(n)}, p_\theta^{(n)}) \leq n\epsilon^2, V_k(p_{\theta_0}^{(n)}, p_\theta^{(n)}) \leq n^{k/2}\epsilon^k\Big\}$

$K(p, q) = P \log(p/q)$
$V_k(p, q) = P\Big|\log(p/q) - K(p, q)\Big|^k$

rate theorem (2)

$\epsilon_n \to 0$

$$\sup_{\epsilon > \epsilon_n} \log N(\epsilon\xi, \{\theta \in \Theta_n : d_n(\theta, \theta_0) < \epsilon\}, e_n) \le n\epsilon_n^2$$

$$\frac{\Pi_n(\Theta_n^c)}{\Pi_n(B_n(\theta_0, \epsilon_n; k))} = o(e^{-2n\epsilon_n^2})$$

$$\frac{\Pi_n(\theta \in \Theta_n : d_n(\theta, \theta_0) \le 2j\epsilon_n)}{\Pi_n(B_n(\theta_0, \epsilon_n; k))} \le e^{Kn\epsilon_n^2 j^2/2} \qquad \forall j$$

Then $P_{\theta_0}^{(n)} \Pi_n(\theta \in \Theta_n : d_n(\theta, \theta_0) \ge M_n \epsilon_n | X^{(n)}) \to 0$

$$B_n(\theta_0, \epsilon; k) =$$
$$\left\{\theta \in \Theta : K(p_{\theta_0}^{(n)}, p_\theta^{(n)}) \le n\epsilon^2, V_k(p_{\theta_0}^{(n)}, p_\theta^{(n)}) \le n^{k/2}\epsilon^k\right\}$$

examples

Entropy: from Birgé 83a,b

Prior mass: some work necessary

independent observations

$X^{(n)} = (X_1, X_2, \ldots, X_n)$
$X_1, \ldots, X_n$ independent, $X_i \sim p_{\theta,i}$

MAIN RESULT HOLDS WITH

$$d_n^2(\theta, \theta') = \frac{1}{n} \sum_{i=1}^n \int \left(\sqrt{p_{\theta,i}} - \sqrt{p_{\theta',i}} \right)^2 \, d\mu_i$$

$$B_n(\theta_0, \epsilon; 2) = \{\theta : \frac{1}{n} \sum_{i=1}^n K_i(\theta_0, \theta) \le \epsilon^2, \frac{1}{n} \sum_{i=1}^n V_i(\theta_0, \theta) \le \epsilon^2\}$$

markov chains

$X^{(n)} = (X_0, X_1, \ldots, X_n)$

$\cdots, X_0, X_1, X_2, \cdots$ stationary Markov chain $\quad p_\theta(y|x)$, $q_\theta(x)$

$r(y) \lesssim p_\theta(y|x) \lesssim r(y)$, $\qquad \int r\, d\mu < \infty$

α-mixing, $\sum_{h=0}^{\infty} \alpha_h^{1-1/k} < \infty$ for some $k > 2$

MAIN RESULT HOLDS WITH

$$d^2(\theta, \theta') = \iint \left[\sqrt{p_\theta(y|x)} - \sqrt{p_{\theta'}(y|x)}\right]^2 d\mu(y)\, r(x)\, d\mu(x)$$

$$B_n(\theta_0, \epsilon; k) =$$

$$\left\{\theta : P_{\theta_0} \log \frac{p_{\theta_0}}{p_\theta}(X_1|X_0) \le \epsilon^2, P_{\theta_0} \left|\log \frac{p_{\theta_0}}{p_\theta}(X_1|X_0)\right|^k \le \epsilon^k\right\}$$

gaussian white noise model

$X^{(n)} = (X_t^{(n)} : 0 \leq t \leq 1)$

$dX_t^{(n)} = \theta(t)\, dt + n^{-1/2} dB_t$ Brownian motion B

MAIN RESULT HOLDS WITH

d L_2-norm

$B(\theta_0, \epsilon; 2)$ L_2-ball

gaussian time series

$X^{(n)} = (X_0, X_1, \ldots, X_n)$
$\cdots, X_0, X_1, X_2, \cdots$ stationary mean zero Gaussian process
spectral density θ

$\sup_\theta \|\log\theta\|_\infty < \infty$
$\sup_\theta \sum_{h=-\infty}^{\infty} |h|(\mathrm{E}_\theta X_h X_0)^2 < \infty$

MAIN RESULT HOLDS WITH

d L_2-norm, e supremum-norm
$B(\theta_0, \epsilon; 2)$ L_2-ball

ergodic diffusions

$X^{(n)} = (X_t : 0 \leq t \leq n)$
$dX_t = \theta(X_t)\,dt + \sigma(X_t)\,dB_t$ Brownian motion B
stationary ergodic, state space I, stationary measure μ_{θ_0}

MAIN RESULT HOLDS WITH

$d(\theta, \theta') = \|(\theta - \theta')1_J/\sigma\|_{\mu_{\theta_0},2}$ $J \subset I$
$e(\theta, \theta') = \|(\theta - \theta')/\sigma\|_{\mu_{\theta_0},2}$
$B(\theta_0, \epsilon; 2)$ $\| \cdot /\sigma\|_{\mu_{\theta_0},2}$-ball

ergodic diffusions (2)

$X^{(n)} = (X_t : 0 \leq t \leq n)$
$dX_t = \theta(X_t)\,dt + \sigma(X_t)\,dB_t$ Brownian motion B
stationary, ergodic solution, state space $I = (0,1)$, μ_{θ_0}

$\sigma(x) \sim x^{1+p}$ as $x \downarrow 0$; $\sigma(x) \sim (1-x)^{1+q}$ as $x \uparrow 1$
$\theta : [0,1] \to \mathbb{R}$, $\downarrow$, $\theta(0) > 0$, $\theta(1) < 0$

$\big(D(0,1/L), D(1/L,2/L), \ldots, D(1-1/L,1)\big) \sim$ Dirichlet(α)
+linear interpolation
$L \sim n^{1/3} \log n$
$U \sim U[0,1]$, $V \sim \exp(1)$ independent
$\theta \sim V(U - D)$
gives rate $n^{-1/3} \log n$

GAUSSIAN PROCESS PRIORS

setting (1)

$W = (W_x : x \in \mathcal{X})$ zero-mean Gaussian process

sample path is a-priori model for "unknown function"

Examples

1: density estimation $p_W(x) = e^{W_x} / \int e^{W_y}\, d\nu(y)$

2: logistic regression $\Pr(Y = 1 | X = x) = 1/(1 + e^{-W_x})$

3: regression $Y = W_x + \mathrm{error}$

4: white noise $dX_t = W_t + n^{-1/2}\, dB_t$

reproducing kernel hilbert space (1)

W zero-mean Gaussian map in Banach space $(\mathbb{B}, \|\cdot\|)$
$S: \mathbb{B}^* \to \mathbb{B}$, $\quad Sb^* = \mathrm{E} W b^*(W)$ [Pettis integral]

RKHS $\mathbb{H}$ is completion of $S\mathbb{B}^*$ under
$\langle Sb_1^*, Sb_2^* \rangle_{\mathbb{H}} = \mathrm{E} b_1^*(W) b_2^*(W)$

FACT support of W is closure of $\mathbb{H}$ in $\mathbb{B}$

CONSEQUENCE posterior inconsistent if $\|w_0 - \mathbb{H}\| > 0$
[smoothness is not enough]

reproducing kernel hilbert space (2)

$W = (W_x : x \in \mathcal{X})$ map in $\ell^\infty(\mathcal{X})$
covariance function $K(x, y) = \mathrm{E} W_x W_y$

RKHS is completion of the set of functions
$x \mapsto \sum_i \alpha_i K(y_i, x)$
for
$\left\langle \sum_i \alpha_i K(y_i, \cdot), \sum_j \beta_j K(z_j, \cdot) \right\rangle_{\mathbb{H}} = \sum_i \sum_j \alpha_i \beta_j K(y_i, z_j)$

small ball probability

W map in $(\mathbb{B}, \|\cdot\|)$, RKHS $(\mathbb{H}, \|\cdot\|_{\mathbb{H}})$, unit balls $\mathbb{B}_1$, $\mathbb{H}_1$

$\Pr(\|W\| < \epsilon) = e^{-\phi_0(\epsilon)}$

FACT (Borel, 1975)

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

FACT (Kuelbs and Li, 1993)

$\forall h \in \mathbb{H}\ \Pr(\|W - h\| < \epsilon) \geq e^{-\|h\|_{\mathbb{H}}^2/2} \Pr(\|W\| < \epsilon)$

FACT (Kuelbs and Li, 1993)

$\phi_0(\epsilon) \sim \log N(\epsilon/\sqrt{\phi_0(\epsilon)}, \mathbb{H}_1, \|\cdot\|)$

gaussian rate conditions

W map in $(\mathbb{B}, \|\cdot\|)$, RKHS $(\mathbb{H}, \|\cdot\|_{\mathbb{H}})$
$\Pr(\|W\| < \epsilon) = e^{-\phi_0(\epsilon)}$

THEOREM $\forall\, w_0 \in \overline{\mathbb{H}}$ and $\epsilon_n > 0$ with
$\inf_{h\in\mathbb{H}:\|h-w_0\|<\epsilon_n} \|h\|_{\mathbb{H}}^2 + \phi_0(\epsilon_n) \leq n\epsilon_n^2$

$\exists\, B_n \subset \mathbb{B}$ with
$\log N\big(\epsilon_n, B_n, \|\cdot\|\big) \lesssim n\epsilon_n^2$
$\Pr(W \notin B_n) \lesssim e^{-4n\epsilon_n^2}$
$\Pr(\|W - w_0\| < \epsilon_n) \gtrsim e^{-n\epsilon_n^2}$

CONSEQUENCE
If distances on model combine well with $\|\cdot\|$, rate is ϵ_n if
$\inf_{h\in\mathbb{H}:\|h-w_0\|<\epsilon_n} \|h\|_{\mathbb{H}}^2 \leq n\epsilon_n^2$ AND $\phi_0(\epsilon_n) \leq n\epsilon_n^2$

rate theorem (1)

$\epsilon_n \to 0$, $\epsilon_n \gg 1/\sqrt{n}$

$\log N(\epsilon_n, \Theta_n, e_n) \le n\epsilon_n^2$
$\Pi_n(\Theta_n^c) = o(e^{-3n\epsilon_n^2})$
$\Pi_n(B_n(\theta_0, \epsilon_n; k)) \ge e^{-n\epsilon_n^2}$

Then $P_{\theta_0}^{(n)} \Pi_n(\theta \in \Theta_n : d_n(\theta, \theta_0) \ge M_n \epsilon_n | X^{(n)}) \to 0$

$B_n(\theta_0, \epsilon; k) =$
$\Big\{\theta \in \Theta : K(p_{\theta_0}^{(n)}, p_\theta^{(n)}) \le n\epsilon^2, V_k(p_{\theta_0}^{(n)}, p_\theta^{(n)}) \le n^{k/2}\epsilon^k\Big\}$

$K(p,q) = P \log(p/q)$
$V_k(p,q) = P\big|\log(p/q) - K(p,q)\big|^k$

log gaussian densities

$p_W(x) = e^{W_x} / \int e^{W_y}\, d\nu(y)$

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

CONSEQUENCE
If W is map in $\ell^\infty(\mathcal{X})$ rate is ϵ_n if
$\phi_0(\epsilon_n) \leq n\epsilon_n^2$ AND $\inf_{h\in\mathbb{H}:\|h-w_0\|_\infty<\epsilon_n} \|h\|_{\mathbb{H}}^2 \leq n\epsilon_n^2$

logistic regression

$\Pr(Y = 1|X = x) = 1/(1 + e^{-W_x})$
p_w density of (X, Y) given $W = w$

LEMMA $\forall v, w$
$\|p_v - p_w\|_2 = 2^{1/2}\|\Psi(v) - \Psi(w)\|_{2,G} \lesssim \|v - w\|_{2,G}$
$K(p_w, p_{w_0}) \leq \|w - w_0\|_{2,G}^2$
$V(p_w, p_{w_0}) \leq \|w - w_0\|_{2,G}^2$

CONSEQUENCE
If W is map in $L_2(G)$, rate is ϵ_n if
$\phi_0(\epsilon_n) \leq n\epsilon_n^2$ AND $\inf_{h\in\mathbb{H}:\|h-w_0\|_{G,2}<\epsilon_n} \|h\|_{\mathbb{H}}^2 \leq n\epsilon_n^2$

example: brownian motion

W one-dimensional Brownian motion on $[0,1]$

$\phi_0(\epsilon) \sim (1/\epsilon)^2$
$\mathbb{H} = \{h : h(0) = 0, \int h'(t)^2\, dt < \infty\}$
start at $N(0,1)$ to replace $\mathbb{H}$ by $\mathbb{H} \oplus 1$

LEMMA If $w_0 \in C^\alpha[0,1]$ for $0 < \alpha < 1$, then
$\inf_{h \in \mathbb{H}: \|h - w_0\|_\infty < \epsilon_n} \|h\|_{\mathbb{H}}^2 \sim (1/\epsilon)^{(2-2\alpha)/\alpha}$

PROOF: $w_0 * \psi_\sigma \in \mathbb{H}$
$\|w_0 * \psi_\sigma - w_0\|_\infty \sim \sigma^\alpha$
$\int (w_0 * \psi_\sigma)'(t)^2\, dt \sim (1/\sigma)^{2-2\alpha}$

example: brownian motion

W one-dimensional Brownian motion on $[0,1]$

$\phi_0(\epsilon) \sim (1/\epsilon)^2$
$\mathbb{H} = \{h : h(0) = 0, \int h'(t)^2\, dt < \infty\}$
start at $N(0,1)$ to replace $\mathbb{H}$ by $\mathbb{H} \oplus 1$

LEMMA If $w_0 \in C^\alpha[0,1]$ for $0 < \alpha < 1$, then
$\inf_{h \in \mathbb{H}: \|h - w_0\|_\infty < \epsilon_n} \|h\|_{\mathbb{H}}^2 \sim (1/\epsilon)^{(2-2\alpha)/\alpha}$

CONSEQUENCE
rate is $n^{-1/4}$ if $\alpha \geq 1/2$; $n^{-\alpha/2}$ if $\alpha \leq 1/2$
Brownian motion: $(1/\epsilon_n)^2 \leq n\epsilon_n^2 \implies \epsilon_n \geq n^{-1/4}$
$\log p_0 \in C^\alpha[0,1]$: $(1/\epsilon_n)^{(2-2\alpha)/\alpha} \leq n\epsilon_n^2 \implies \epsilon_n \geq n^{-\alpha/2}$

alternatives

RIEMANN-LIOUVILLE process

$W_t = \int_0^t (t-s)^{\alpha-1/2}\, dB_s + W_0$

[B Brownian motion, $\alpha > 0$]

FRACTIONAL BROWNIAN MOTION

$\mathrm{E} W_s W_t = s^{2\alpha} + t^{2\alpha} - |t-s|^{2\alpha}$

[Hurst index $0 < \alpha < 1$]

Give proper models for α-smooth functions

SERIES EXPANSIONS

$W_t = \sum_{i=1}^{\infty} \lambda_i Z_i e_i(t)$

[(e_i) basis, (Z_i) i.i.d. $N(0,1)$, $\lambda_i \to 0$]

adaptation

To obtain “correct” rate for any $\alpha > 0$
can use mixture of Gausian priors with prior weights

$$w_{n,\alpha} \sim e^{-Cn^{1/(2\alpha+1)}}$$

[Other weights too?]

example: brownian sheet

W d-dimensional Brownian sheet on $[0,1]^d$

$\phi_0(\epsilon) \sim (1/\epsilon)^2 (\log(1/\epsilon))^{2d-2}$
$\mathbb{H} = \{h : h(0) = 0, \int_{[0,1]^d} h'(t)^2 \, dt < \infty\}$

can approximate “Bounded variation” functions only

Some Results on Nonparametric Bayesian Inference

Aad van der Vaart

http://www.math.vu.nl/∼aad

Vrije Universiteit Amsterdam