Entropy Methods in Statistics

Aad van der Vaart
Vrije Universiteit Amsterdam

18–20 May 2009

Overview

1: ENTROPY
2: EMPIRICAL PROCESSES AND MAXIMAL INEQUALITIES
3: M-ESTIMATORS AND MODEL SELECTION
4: TESTS AND LECAM-BIRGÉ ESTIMATORS
5: POSTERIOR DISTRIBUTIONS

Overview

1: ENTROPY
2: EMPIRICAL PROCESSES AND MAXIMAL INEQUALITIES
3: M-ESTIMATORS AND MODEL SELECTION
4: TESTS AND LECAM-BIRGÉ ESTIMATORS
5: POSTERIOR DISTRIBUTIONS

REFERENCES:
1: Kolmogorov, Dudley, . . .—1960s–
2: Dudley, Pollard, Koltchinkii, Giné, Talagrand—1970/80/90s
3: Wong, Van de Geer, Birgé & Massart—1990s
4: Le Cam, Birgé—1970/80s, 2003
5: Ghosal & Van der Vaart—2000s
General: Van der Vaart & Wellner (1996, 2010)

1: Entropy

Covering numbers

Let $(\mathcal{P}, d)$ be a metric space.

DEFINITION The covering number $N\big(\varepsilon, \mathcal{P}, d\big)$ is the minimal number of balls of radius ε needed to cover the set $\mathcal{P}$.

DEFINITION The packing number $D(\varepsilon, \mathcal{P}, d)$ of $\mathcal{P}$ is the maximal number of points in $\mathcal{P}$ such that the distance between every pair is at least ε.

DEFINITION The entropy is the logarithm of these numbers

Covering numbers

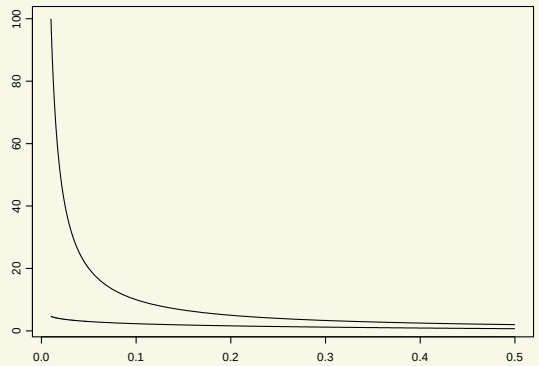
DEFINITION The covering number $N(\varepsilon, \mathcal{P}, d)$ is the minimal number of balls of radius ε needed to cover the set $\mathcal{P}$.

DEFINITION The packing number $D(\varepsilon, \mathcal{P}, d)$ of $\mathcal{P}$ is the maximal number of points in $\mathcal{P}$ such that the distance between every pair is at least ε.

DEFINITION The entropy is the logarithm of these numbers.

LEMMA $N(\varepsilon, \mathcal{P}, d) \leq D(\varepsilon, \mathcal{P}, d) \leq N(\varepsilon/2, \mathcal{P}, d)$.

LEMMA $N(\varepsilon, \mathcal{P}, d) < \infty$ for all $\varepsilon > 0$ if and only if the completion of $\mathcal{P}$ is compact.



Le Cam dimension

DEFINITION The Le Cam dimension or local covering number at P_0 is

$$L(\varepsilon, \mathcal{P}, d; P_0) = \sup_{\eta \geq \varepsilon} N\big(\eta, B(P_0, 2\eta, d), d\big),$$

where $B(P_0, \varepsilon, d)$ is the ball of radius ε around P_0.

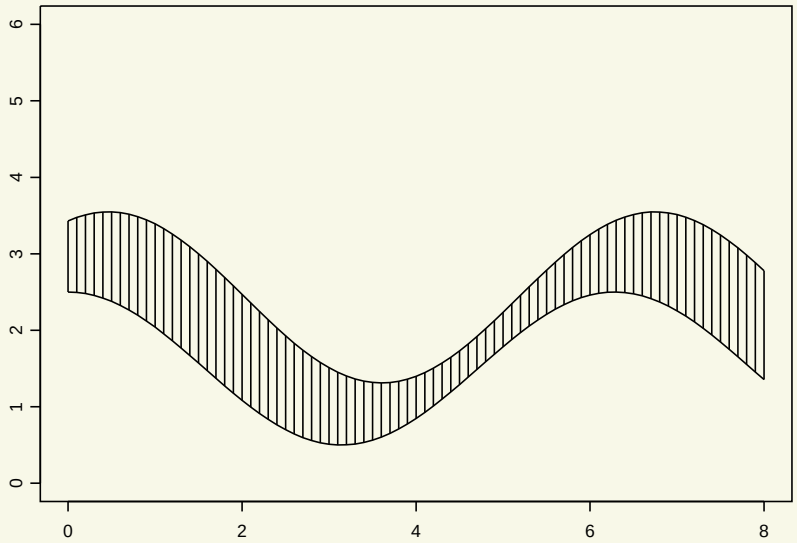
LEMMA $L(\varepsilon, \mathcal{P}, d; P_0) \leq \log N(\varepsilon, \mathcal{P}, d)$.

Bracketing numbers

DEFINITION Given $l, u: \mathcal{X} \to \mathbb{R}$ the bracket $[l, u]$ is the set of all functions f with $l \leq f \leq u$.
An ε-bracket relative to a metric d is a bracket $[l, u]$ with $d(u, l) < \varepsilon$.

DEFINITION The bracketing number $N_{[]}\big(\varepsilon, \mathcal{F}, d\big)$ is the minimum number of ε-brackets needed to cover $\mathcal{F}$.

DEFINITION The entropy is the logarithm of the bracketing number.



Bracketing versus covering

Suppose $|f| \leq |g|$ implies $\|f\| \leq \|g\|$

LEMMA $N(\varepsilon, \mathcal{F}, \|\cdot\|) \leq N_{[]}\big(2\varepsilon, \mathcal{F}, \|\cdot\|\big)$.

PROOF If f is in the 2ε-bracket $[l, u]$, then it is in the ball of radius ε around $(l+u)/2$. ∎

Three statistical results

Distances

Hellinger distance

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

L_r-distance

$$\|p - q\|_{\mu,r} = \left(\int |p - q|^r \, d\mu \right)^{1/r}.$$

Uniform norm of $f \colon \mathcal{X} \to \mathbb{R}$

$$\|f\|_\infty = \sup_{x \in \mathcal{X}} |f(x)|.$$

Le Cam-Birgé

Let the observations be a random sample $X_1, \ldots, X_n$ from some P_0 that is known to belong to a model $\mathcal{P}$. Let p denote the density of P.

THEOREM If

$$\log N(\varepsilon_n, \mathcal{P}, h) \leq n\varepsilon_n^2,$$

then there exist estimators $\hat{p}_n$ such that

$$h(\hat{p}_n, p_0) = O_{P_0}(\varepsilon_n).$$

MLE

Let the observations be a random sample $X_1, \ldots, X_n$ from some P_0 that is known to belong to a model $\mathcal{P}$. Let p denote the density of P.

THEOREM If

$$\int_0^{\varepsilon_n} \sqrt{\log N_{[]}(\eta, \mathcal{P}, h)}\, d\eta \leq \sqrt{n}\varepsilon_n^2.$$

then maximizer $\hat{p}_n$ of $p \mapsto \prod_{i=1}^n p(X_i)$ over $\mathcal{P}$ satisfies

$$h(\hat{p}_n, p_0) = O_{P_0}(\varepsilon_n).$$

Le Cam-Birge versus MLE

$$J_{[]}(\varepsilon, \mathcal{P}, h) := \int_0^{\varepsilon} \sqrt{\log N_{[]}\left(\eta, \mathcal{P}, h\right)}\, d\eta$$

LE CAM-BIRGÉ: $\log N(\varepsilon_n, \mathcal{P}, h) \leq n\varepsilon_n^2$

MLE: $J_{[]}(\varepsilon_n, \mathcal{P}, h) \leq \sqrt{n}\varepsilon_n^2$

Always $J_{[]}(\varepsilon, \mathcal{P}, h) \geq \varepsilon\sqrt{\log N_{[]}\left(\varepsilon, \mathcal{P}, h\right)}$.

If two sides of inequality are equivalent up to constants, then

MLE: $\log N_{[]}\left(\varepsilon_n, \mathcal{P}, h\right) \leq n\varepsilon_n^2$.

Bayes

Let the observations be a random sample $X_1, \ldots, X_n$ from some P_0 that is known to belong to a model $\mathcal{P}$. Let p denote the density of P.

prior: Π_n

posterior: $$\Pi_n(B|\, X_1, \ldots, X_n) = \frac{\int_B \prod_{i=1}^n p(X_i)\, d\Pi_n(P)}{\int \prod_{i=1}^n p(X_i) d\Pi_n(P)}$$

THEOREM If $n\varepsilon_n^2 \to \infty$ and

$$\log N(\varepsilon_n, \mathcal{P}, h) \le n\varepsilon_n^2,$$
$$\Pi_n\big(B(P_0, \varepsilon_n)\big) \ge e^{-n\varepsilon_n^2},$$

then $\mathrm{E}_{P_0}\Pi_n(P\colon h(P, P_0) \ge M\varepsilon_n|\, X_1, \ldots, X_n) \to 0$.

$$B(P_0, \varepsilon_n) = \left\{P\colon -P_0\left(\log \frac{p}{p_0}\right) \le \varepsilon_n^2, P_0\left(\log \frac{p}{p_0}\right)^2 \le \varepsilon_n^2\right\}$$

Examples of entropy

Finite-dimensional sets

$U \subset \mathbb{R}^d$ unit ball, arbitrary norm $\|\cdot\|$.

THEOREM For any $\varepsilon \in (0,1)$,

$$\left(\frac{1}{\varepsilon}\right)^d \leq N\big(\varepsilon, U, \|\cdot\|\big) \leq \left(\frac{3}{\varepsilon}\right)^d.$$

Furthermore,

$$2^d \leq N\big(\varepsilon, 2\varepsilon U, \|\cdot\|\big) \leq 5^d.$$

Hölder classes

$\mathcal{X} \subset \mathbb{R}^d$ bounded, convex, nonempty interior.
$C_M^\alpha(\mathcal{X})$ all continuous functions $f\colon \mathcal{X} \to \mathbb{R}$ with $\|f\|_\alpha \le M$.

$$\|f\|_\alpha = \max_{k_\cdot \le \underline{\alpha}} \sup_x \left|D^k f(x)\right| + \max_{k_\cdot = \underline{\alpha}} \sup_{x,y} \frac{\left|D^k f(x) - D^k f(y)\right|}{\|x-y\|^{\alpha-\underline{\alpha}}},$$

$$D^k = \frac{\partial^{\sum k_i}}{\partial x_1^{k_1} \cdots \partial x_d^{k_d}}$$

THEOREM For every $\varepsilon > 0$

$$\log N\left(\varepsilon, C_1^\alpha(\mathcal{X}), \|\cdot\|_\infty\right) \le K_{\alpha,d,\mathcal{X}} \left(\frac{1}{\varepsilon}\right)^{d/\alpha}.$$

Monotone functions

$\mathcal{F}$ all monotone functions $f\colon \mathbb{R} \to [0,1]$.

THEOREM For every probability measure Q and every $r \geq 1$,

$$\log N_{[]}\big(\varepsilon, \mathcal{F}, L_r(Q)\big) \leq K_r\Big(\frac{1}{\varepsilon}\Big).$$

Lipschitz dependence on a parameter

$\mathcal{F} = \{f_t : t \in T\}$ for metric space (T, d)

$$\big|f_s(x) - f_t(x)\big| \le d(s,t)\, F(x), \qquad \text{every } x$$

THEOREM For any norm $\|\cdot\|$,

$$N_{[]}\big(2\varepsilon\|F\|, \mathcal{F}, \|\cdot\|\big) \le N(\varepsilon, T, d).$$

Gaussian mixtures

ϕ_σ density of $N(0, \sigma^2)$-distribution

$$p_{F,\sigma}(x) = \int \phi_\sigma(x - z)\, dF(z).$$

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

THEOREM For d the uniform, L_1 or the Hellinger distance,

$$\log N_{[\,]}(\varepsilon, \mathcal{P}_{a,\tau}, d) \lesssim \frac{a}{\tau}\left(\log \frac{a}{\varepsilon\tau}\right)^2.$$

Proof — key lemma

LEMMA Let $0 < \varepsilon < 1/4$ and $0 < \sigma \lesssim a$ be given. For any F on $[-a, a]$ there exists a discrete F' on $[-a - \sigma, a + \sigma]$ with at most $N \lesssim (a/\sigma) \log(1/\varepsilon)$ support points such that

$$\|p_{F,\sigma} - p_{F',\sigma}\|_1 \lesssim \varepsilon \left(\log \frac{1}{\varepsilon} \right)^{1/2}.$$

Within level of precision ε Gaussian mixtures are an N-dimensional model for $N \lesssim (a/\sigma) \log(1/\varepsilon)$. This suggests that the covering numbers are of the order $(1/\varepsilon)^N$.

Unbounded mixtures

LEMMA For $a > 0$ and F_a the renormalized restriction of the probability measure F to $[-a, a]$: $\|p_{F,\sigma} - p_{F_a,\sigma}\|_1 \leq 2F[-a, a]^c$.

Unbounded mixtures

LEMMA For $a > 0$ and F_a the renormalized restriction of the probability measure F to $[-a, a]$: $\|p_{F,\sigma} - p_{F_a,\sigma}\|_1 \leq 2F[-a, a]^c$.

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

THEOREM For positive numbers $b_1 < b_2$ and $\tau \leq 1/4$ with $b_2\tau \leq 1/4$, and $0 < \varepsilon < 1/4$

$$\log N(3\varepsilon, \mathcal{P}_{\Phi,\tau}, d) \lesssim \frac{1}{\tau}\left(\log \frac{1}{\varepsilon\tau}\right)^{5/2},$$

where the constant in $\lesssim$ depends on b_1, b_2 only.

VC-Classes

A collection of subsets $\mathcal{C}$ of $\mathcal{X}$ shatters a set of points $\{x_1, \ldots, x_n\}$ if each of its 2^n subsets can be obtained as $C \cap \{x_1, \ldots, x_n\}$ for some $C \in \mathcal{C}$.

DEFINITION The VC-dimension $V(\mathcal{C})$ of $\mathcal{C}$ is the largest n for which some set of size n is shattered by $\mathcal{C}$.

DEFINITION The VC-dimension $V(\mathcal{F})$ of a collection $\mathcal{F}$ of functions $f: \mathcal{X} \to \mathbb{R}$ is VC-dimension of its set of subgraphs in $\mathcal{X} \times \mathbb{R}$.

THEOREM For a VC-class of functions with envelope function F and $r \geq 1$, and any probability measure Q, for $0 < \varepsilon < 1$,

$$N\big(\varepsilon \|F\|_{Q,r}, \mathcal{F}, L_r(Q)\big) \leq KV(\mathcal{F})(16e)^{V(\mathcal{F})} \left(\frac{1}{\varepsilon}\right)^{r(V(\mathcal{F})-1)},$$

VC-Classes — Sauer's lemma

Let $\Delta_n(\mathcal{C}, x_1, \ldots, x_n)$ be the number of subsets of $x_1, \ldots, x_n$ picked out by $\mathcal{C}$.
Thus $\Delta_n(\mathcal{C}, x_1, \ldots, x_n) < 2^n$ for $n > V(\mathcal{C})$ and any $x_1, \ldots, x_n$.

LEMMA For any VC-class $\mathcal{C}$ of sets,

$$\max_{x_1,\ldots,x_n} \Delta_n(\mathcal{C}, x_1, \ldots, x_n) \leq \sum_{j=0}^{V(\mathcal{C})} \binom{n}{j} \leq \left(\frac{ne}{V(\mathcal{C})}\right)^{V(\mathcal{C})}.$$

Empirical processes and maximal inequalities

Overview

1: ENTROPY
2: EMPIRICAL PROCESSES AND MAXIMAL INEQUALITIES
3: M-ESTIMATORS AND MODEL SELECTION
4: TESTS AND LECAM-BIRGÉ ESTIMATORS
5: POSTERIOR DISTRIBUTIONS

Entropy

DEFINITION The covering number $N(\varepsilon, \mathcal{P}, d)$ is the minimal number of balls of radius ε needed to cover the set $\mathcal{P}$.

DEFINITION The bracketing number $N_{[]}(\varepsilon, \mathcal{F}, d)$ is the minimum number of ε-brackets needed to cover $\mathcal{F}$.

DEFINITION The entropy is the logarithm of these numbers.

Empirical process

$X_1, \ldots, X_n$ be i.i.d. random elements in $(\mathcal{X}, \mathcal{A})$.
$\mathcal{F} \subset \mathcal{L}_1(\mathcal{X}, \mathcal{A}, P)$.
$\|z\|_{\mathcal{F}} = \sup_{f \in \mathcal{F}} \big|z(f)\big|$; $\qquad Pf = \int f \, dP$.

DEFINITION The empirical measure is the map $\mathbb{P}_n \colon \mathcal{F} \to \mathbb{R}$ given by

$$\mathbb{P}_n f = \frac{1}{n} \sum_{i=1}^{n} f(X_i).$$

The empirical process is the map $\mathbb{G}_n \colon \mathcal{F} \to \mathbb{R}$ given by

$$\mathbb{G}_n f = \sqrt{n}(\mathbb{P}_n - P)f = \frac{1}{\sqrt{n}} \sum_{i=1}^{n} \big(f(X_i) - Pf\big).$$

Empirical process

$X_1, \ldots, X_n$ be i.i.d. random elements in $(\mathcal{X}, \mathcal{A})$.
$\mathcal{F} \subset \mathcal{L}_1(\mathcal{X}, \mathcal{A}, P)$.
$\|z\|_{\mathcal{F}} = \sup_{f \in \mathcal{F}} \big|z(f)\big|$; $\qquad Pf = \int f \, dP$.

DEFINITION The empirical measure is the map $\mathbb{P}_n \colon \mathcal{F} \to \mathbb{R}$ given by

$$\mathbb{P}_n f = \frac{1}{n} \sum_{i=1}^{n} f(X_i).$$

The empirical process is the map $\mathbb{G}_n \colon \mathcal{F} \to \mathbb{R}$ given by

$$\mathbb{G}_n f = \sqrt{n}(\mathbb{P}_n - P)f = \frac{1}{\sqrt{n}} \sum_{i=1}^{n} \big(f(X_i) - Pf\big).$$

EXAMPLE
$\mathcal{X} = \mathbb{R}$, $\mathcal{F} = \{1_{(-\infty, t]} \colon t \in \mathbb{R}\}$.
$\mathbb{P}_n$ is the empirical distribution function $\mathbb{F}_n(t) = \#(X_i \leq t)/n$.
$\mathbb{G}_n = \sqrt{n}(\mathbb{F}_n - F)$ the classical empirical process.
$\|\mathbb{G}_n\|_{\mathcal{F}}$ is the Kolmogorov-Smirnov statistic.

Limit theorems

The LLN gives that $\mathbb{P}_n f \to Pf$ almost surely and in mean if $P|f| < \infty$.

DEFINITION A class $\mathcal{F}$ is called Glivenko-Cantelli if $\sup_f \left|\mathbb{P}_n f - Pf\right| \to 0$ almost surely.

The CLT gives that $\left(\mathbb{G}_n f_1, \ldots, \mathbb{G}_n f_k\right) \rightsquigarrow N_k(0, \Sigma)$ whenever $f_i \in L_2(P)$ for $i = 1, \ldots, k$, where

$$\Sigma_{i,j} = Pf_i f_j - Pf_i Pf_j,$$

DEFINITION A class $\mathcal{F}$ is called Donsker if $\mathbb{G}_n$ converges in distribution to a Gaussian random element (called Brownian bridge) in the space $\ell^\infty(\mathcal{F})$ of all bounded functions $z\colon \mathcal{F} \to \mathbb{R}$, normed with the uniform norm.

Entropy integrals

An envelope function of a class $\mathcal{F}$ of functions $f\colon \mathcal{X} \to \mathbb{R}$ is a function $F\colon \mathcal{X} \to \mathbb{R}$ such that

$$|f(x)| \leq F(x), \text{ every } x \in \mathcal{X}, f \in \mathcal{F}.$$

DEFINITION Uniform entropy integral relative to envelope function F:

$$J(\delta, \mathcal{F}) = \sup_Q \int_0^\delta \sqrt{1 + \log N\big(\varepsilon \|F\|_{Q,2}, \mathcal{F}, \|\cdot\|_{Q,2}\big)}\, d\varepsilon.$$

DEFINITION Bracketing integral:

$$J_{[]}\big(\delta, \mathcal{F}, \|\cdot\|\big) = \int_0^\delta \sqrt{1 + \log N_{[]}\big(\varepsilon, \mathcal{F}, \|\cdot\|\big)}\, d\varepsilon.$$

Maximal inequalities driven by envelope functions

DEFINITION Uniform entropy integral relative to envelope function F:

$$J(\delta, \mathcal{F}) = \sup_Q \int_0^\delta \sqrt{1 + \log N\big(\varepsilon \|F\|_{Q,2}, \mathcal{F}, \|\cdot\|_{Q,2}\big)}\, d\varepsilon.$$

DEFINITION Bracketing integral:

$$\bar{J}_{[]}\big(\delta, \mathcal{F}, \|\cdot\|\big) = \int_0^\delta \sqrt{1 + \log N_{[]}\big(\varepsilon, \mathcal{F}, \|\cdot\|\big)}\, d\varepsilon.$$

THEOREM If either $J_{[]}(1, \mathcal{F}, \|\cdot\|_{P_0,2}) < \infty$ or $J(1, \mathcal{F}) < \infty$, then

$$\mathrm{E}_{P_0} \|\mathbb{G}_n\|_{\mathcal{F}} \leq C \|F\|_{P_0,2},$$

for C a multiple of the entropy integral at 1.

Maximal Inequalities for small functions

THEOREM If $Pf^2 < \delta^2$ and $\|f\|_\infty \leq 1$ for every f in $\mathcal{F}$, then

$$\mathrm{E}_P^* \|\mathbb{G}_n\|_{\mathcal{F}} \lesssim J_{[]}\big(\delta, \mathcal{F}, L_2(P)\big) \left(1 + \frac{J_{[]}\big(\delta, \mathcal{F}, L_2(P)\big)}{\delta^2 \sqrt{n}}\right).$$

THEOREM If $Pf^2 < \delta^2 PF^2$ and $F \leq 1$ for every $f \in \mathcal{F}$, then

$$\mathrm{E}_P^* \|\mathbb{G}_n\|_{\mathcal{F}} \lesssim J\big(\delta, \mathcal{F}, L_2\big) \left(1 + \frac{J(\delta, \mathcal{F}, L_2)}{\delta^2 \sqrt{n} \|F\|_{P,2}}\right) \|F\|_{P,2}.$$

Maximal Inequalities for small functions

THEOREM If $Pf^2 < \delta^2$ and $\|f\|_\infty \leq 1$ for every f in $\mathcal{F}$, then

$$\mathrm{E}_P^* \|\mathbb{G}_n\|_{\mathcal{F}} \lesssim J_{[]}\big(\delta, \mathcal{F}, L_2(P)\big) \left(1 + \frac{J_{[]}\big(\delta, \mathcal{F}, L_2(P)\big)}{\delta^2 \sqrt{n}}\right).$$

THEOREM If $Pf^2 < \delta^2 PF^2$ and $F \leq 1$ for every $f \in \mathcal{F}$, then

$$\mathrm{E}_P^* \|\mathbb{G}_n\|_{\mathcal{F}} \lesssim J\big(\delta, \mathcal{F}, L_2\big) \left(1 + \frac{J(\delta, \mathcal{F}, L_2)}{\delta^2 \sqrt{n} \|F\|_{P,2}}\right) \|F\|_{P,2}.$$

$$\|f\|_{P,B} = \Big(2P\big(e^{|f|} - 1 - |f|\big)\Big)^{1/2}$$ Bernstein "norm"

LEMMA If $\|f\|_{P,B} \leq \delta$ for every $f \in \mathcal{F}$, then

$$\mathrm{E}_P^* \|\mathbb{G}_n\|_{\mathcal{F}} \lesssim J_{[]}\big(\delta, \mathcal{F}, \|\cdot\|_{P,B}\big) \left(1 + \frac{J_{[]}\big(\delta, \mathcal{F}, \|\cdot\|_{P,B}\big)}{\delta^2 \sqrt{n}}\right).$$

Deviations from the mean

THEOREM [Bounded Difference]
If $\left|f(x) - f(y)\right| \leq 1$ for every $f \in \mathcal{F}$ and every $x, y \in \mathcal{X}$, then, for all $t \geq 0$,

$$\mathrm{P}^*\Big(\big|\|\mathbb{G}_n\|_{\mathcal{F}} - \mathrm{E}^*\|\mathbb{G}_n\|_{\mathcal{F}}\big| \geq t\Big) \leq 2\exp(-2t^2).$$

THEOREM [Talagrand]
If $f(x) - Pf \leq 1$ for every $x \in \mathcal{X}$ and
$w_n := \sup_f P(f - Pf)^2 + 2n^{-1/2}\mathrm{E}\sup_f \mathbb{G}_n f$, for every $t \geq 0$,

$$\mathrm{P}\Big(\sup_f \mathbb{G}_n f - \mathrm{E}\sup_f \mathbb{G}_n f \geq t\Big) \leq \exp\Big(-\tfrac{1}{2}\frac{t^2}{w_n + \frac{1}{3}t/\sqrt{n}}\Big).$$

Orlicz norm — maximal inequality

For $\psi\colon [0,\infty) \to [0,\infty)$ convex, nondecreasing with $\psi(0) = 0$

$$\|X\|_\psi = \inf\left\{C\colon \mathrm{E}\psi\Big(\frac{|X|}{C}\Big) \le 1\right\}.$$

LEMMA For arbitrary random variables $X_1, \ldots, X_N$,

$$\mathrm{E} \max_{1\le i\le N} |X_i| \le \psi^{-1}(N) \max_{1\le i\le N} \|X_i\|_\psi.$$

EXAMPLES

$\psi(x) = x^p$, for $p \ge 1$. $\quad \psi^{-1}(N) = N^{1/p}$.

$\psi_p(x) = e^{x^p} - 1$, for $p \ge 1$. $\quad \psi^{-1}(N) = \big(\log(1+N)\big)^{1/p}$.

LEMMA $\|X\|_{\psi_p} < \infty$ if and only if $\mathrm{P}\big(|X| > x\big) \le Ke^{-C|x|^p}$ for all (large) x and some C, K.

Maximal inequality — chaining

(T, d) a metric space.
$\{X_t : t \in T\}$ a separable stochastic process.

THEOREM If for every s, t,

$$\|X_s - X_t\|_\psi \le d(s,t),$$

then

$$\mathrm{E} \sup_{s,t \in T} |X_s - X_t| \le K_{\psi,C} \int_0^{\operatorname{diam} T} \psi^{-1}\big(D(\varepsilon, T, d)\big)\, d\varepsilon.$$

Symmetrization

For $\varepsilon_1, \ldots, \varepsilon_n$ i.i.d. random variables, each equal to -1 or 1 with probability 1/2 each (Rademacher variables), independent of $X_1, \ldots, X_n$, set

$$\mathbb{G}_n^o f = \frac{1}{\sqrt{n}} \sum_{i=1}^{n} \varepsilon_i f(X_i).$$

LEMMA

$$\mathrm{E}^* \|\mathbb{G}_n\|_{\mathcal{F}} \leq 2\mathrm{E}^* \|\mathbb{G}_n^o\|_{\mathcal{F}}.$$

Bounded difference inequality

Suppose for every $x_1, \ldots, x_n, y_i \in \mathcal{X}$,

$$\bigl|T(x_1, \ldots, x_n) - T(x_1, \ldots, x_{i-1}, y_i, x_{i+1}, \ldots, x_n)\bigr| \le c_i.$$

THEOREM For all $t > 0$,

$$\mathrm{P}\bigl(T - \mathrm{E}T \ge t\bigr) \le \exp\Bigl(-\frac{2t^2}{\sum_{i=1}^n c_i^2}\Bigr).$$

EXAMPLES

[Hoeffding] $T = \sum_{i=1}^n Y_i$, $\quad Y_1, \ldots, Y_n$ independent with $c_i = \|Y_i\|_\infty$.

[McDiarmid] $T = \|\mathbb{G}_n\|_{\mathcal{F}}$ or $T = \sup_{f \in \mathcal{F}} \mathbb{G}_n f$, with

$$c_i = \bigl(\sup_f \sup_{x,y} f(x) - f(y)\bigr)/\sqrt{n}.$$

[SEP] $T = \mathbb{G}_n^o f$ as function of $\varepsilon_1, \ldots, \varepsilon_n$ for fixed $X_1, \ldots, X_n$, with $\sum_i c_i^2 = \mathbb{P}_n f^2$.

Proof of UE maximal inequality

By the bounded difference inequality, for $\|\cdot\|_n$ the $L_2(\mathbb{P}_n)$-norm,

$$\|\mathbb{G}_n^o f - \mathbb{G}_n^o g\|_{\psi_2|X_1,\ldots,X_n} \leq \|f-g\|_n.$$

Therefore, by the subGaussian maximal inequality

$$\begin{aligned}
\mathrm{E}_\varepsilon \|\mathbb{G}_n^o\|_{\mathcal{F}} &\lesssim \int_0^\infty \sqrt{1+\log N\big(\varepsilon, \mathcal{F}, L_2(\mathbb{P}_n)\big)}\, d\varepsilon \\
&\leq \int_0^1 \sqrt{1+\log N\big(\varepsilon\|F\|_n, \mathcal{F}, L_2(\mathbb{P}_n)\big)}\, d\varepsilon \, \|F\|_n \\
&\leq J(1,\mathcal{F})\, \|F\|_n.
\end{aligned}$$

Finally take the expectation over $X_1, \ldots, X_n$.

M-Estimators and model selection

Overview

1: ENTROPY
2: EMPIRICAL PROCESSES AND MAXIMAL INEQUALITIES
3: M-ESTIMATORS AND MODEL SELECTION
4: TESTS AND LECAM-BIRGÉ ESTIMATORS
5: POSTERIOR DISTRIBUTIONS

M-Estimators

(Θ, d) metric space
$\hat{\theta}_n$ maximizes stochastic process $\theta \mapsto \mathbb{M}_n(\theta)$ over Θ
θ_0 maximizes deterministic function $\theta \mapsto \mathbb{M}(\theta)$

THEOREM Assume that for some ϕ_n such that $\delta \mapsto \phi_n(\delta)/\delta^\alpha$ is decreasing for some $\alpha < 2$,

$$\mathbb{M}(\theta) - \mathbb{M}(\theta_0) \lesssim -d^2(\theta, \theta_0), \qquad \text{every } \theta \in \Theta,$$

$$\mathrm{E}_{P_0} \sup_{d(\theta,\theta_0)<\delta} \left|\sqrt{n}(\mathbb{M}_n - \mathbb{M})(\theta) - \sqrt{n}(\mathbb{M}_n - \mathbb{M})(\theta_0)\right| \lesssim \phi_n(\delta).$$

If $\phi_n(\varepsilon_n) \leq \sqrt{n}\varepsilon_n^2$, then $d(\hat{\theta}_n, \theta_0) = O_P^*(\varepsilon_n)$

Empirical process — recall notation

$X_1, \ldots, X_n$ be i.i.d. random elements in $(\mathcal{X}, \mathcal{A})$.
$\mathcal{F} \subset \mathcal{L}_1(\mathcal{X}, \mathcal{A}, P)$.
$\|z\|_{\mathcal{F}} = \sup_{f \in \mathcal{F}} \big|z(f)\big|$; $\qquad Pf = \int f \, dP$.

DEFINITION The empirical measure is the map $\mathbb{P}_n \colon \mathcal{F} \to \mathbb{R}$ given by

$$\mathbb{P}_n f = \frac{1}{n} \sum_{i=1}^{n} f(X_i).$$

The empirical process is the map $\mathbb{G}_n \colon \mathcal{F} \to \mathbb{R}$ given by

$$\mathbb{G}_n f = \sqrt{n}(\mathbb{P}_n - P)f = \frac{1}{\sqrt{n}} \sum_{i=1}^{n} \big(f(X_i) - Pf\big).$$

Sieved M-Estimators—iid+Refinement

(Θ_n, d_n) metric spaces, $\theta_n \in \Theta_n$.

$$\hat{\theta}_n \in \Theta_n, \qquad \text{and} \qquad \mathbb{P}_n m_{n,\hat{\theta}_n} \geq \mathbb{P}_n m_{n,\theta_n}.$$

$\mathcal{M}_{n,\delta} := \{m_{n,\theta} - m_{n,\theta_n} : d_n(\theta, \theta_n) < \delta, \theta \in \Theta_n\}$.

THEOREM Assume for $\delta > \delta_n$ and ϕ_n such that $\delta \to \phi_n(\delta)/\delta^\alpha$ is decreasing for some $\alpha < 2$,

$$P_0(m_{n,\theta} - m_{n,\theta_n}) \leq -d^2(\theta, \theta_n),$$

$$\mathrm{E}_{P_0} \sup_{m \in \mathcal{M}_{n,\delta}} |\mathbb{G}_n m| \lesssim \phi_n(\delta).$$

If $\varepsilon_n \gtrsim \delta_n$ and $\phi_n(\varepsilon_n) \leq \sqrt{n}\varepsilon_n^2$, then $d_n(\hat{\theta}_n, \theta_n) = O_P^*(\varepsilon_n)$.

Bias-Variance trade-off

If $\Theta_n \uparrow \Theta$ and $\theta_0 \in \Theta$

$$d(\hat{\theta}_n, \theta_0) = O_P^*(\varepsilon_n) + d(\theta_n, \theta_0).$$

Rate ε_n determined by $\phi_n(\delta) =: \mathrm{E}_{P_0} \sup_{m \in \mathcal{M}_{n,\delta}} |\mathbb{G}_n m|$

Small sieves Θ_n give small $\mathcal{M}_{n,\delta}$, hence fast rate ε_n.

Large sieves give slow $d(\theta_n, \theta_0)$.

Example: Lipschitz criteria

$\hat{\theta}_n$ maximizes $\theta \mapsto \mathbb{P}_n m_\theta$ over $\Theta \subset \mathbb{R}^d$
$\left|m_{\theta_1}(x) - m_{\theta_2}(x)\right| \leq \dot{m}(x)\|\theta_1 - \theta_2\|$ for all θ_1, θ_2
$P_0 \dot{m}^2 < \infty$

Then $M_\delta = \delta \dot{m}$
$N_{[]}\left(\varepsilon \|M_\delta\|_{P_0,2}, \mathcal{M}_\delta, \|\cdot\|_{P_0,2}\right) \leq N\left(\varepsilon, \Theta, \|\cdot\|\right)$

So can apply theorem with $\phi(\delta) = \delta$.

If $\theta \subset \mathbb{R}^d$ and $\theta \mapsto P_0 m_\theta$ is $2\times$ differentiable at point of maximum θ_0 with nonsingular second-derivative, then rate $\varepsilon_n = n^{-1/2}$

$\mathcal{M}_\delta := \left\{m_\theta - m_{\theta_0} : d(\theta, \theta_0) < \delta, \theta \in \Theta\right\}$

Example: interval mode

$\hat{\theta}_n$ center of an interval of length 2 that contains the largest possible fraction of the observations
$\hat{\theta}_n$ maximizes $\theta \mapsto \mathbb{P}_n[\theta - 1, \theta + 1]$

$\mathcal{M}_\delta = \{1_{[\theta-1,\theta+1]} : |\theta - \theta_0| < \delta\}$ is VC-class with envelope
$M_\delta = \sup_{|\theta-\theta_0|<\delta} \left|1_{[\theta-1,\theta+1]} - 1_{[\theta_0-1,\theta_0+1]}\right|$
$\|M_\delta\|_{P_0,2} \leq C\sqrt{\delta}$

So we can apply theorem with $\phi(\delta) = \sqrt{\delta}$

If second derivative of $\theta \mapsto P_0[\theta - 1, \theta + 1]$ *is strictly negative at* θ_0*, then rate* $\varepsilon_n = n^{1/3}$

Maximal inequalities for small functions

THEOREM If $Pf^2 < \delta^2$ and $\|f\|_\infty \leq 1$ for every f in $\mathcal{F}$, then

$$\mathrm{E}_P^* \|\mathbb{G}_n\|_{\mathcal{F}} \lesssim J_{[]}\bigl(\delta, \mathcal{F}, L_2(P)\bigr) \left(1 + \frac{J_{[]}\bigl(\delta, \mathcal{F}, L_2(P)\bigr)}{\delta^2 \sqrt{n}}\right).$$

THEOREM If $Pf^2 < \delta^2 PF^2$ and $F \leq 1$ for every $f \in \mathcal{F}$, then

$$\mathrm{E}_P^* \|\mathbb{G}_n\|_{\mathcal{F}} \lesssim J\bigl(\delta, \mathcal{F}, L_2\bigr) \left(1 + \frac{J(\delta, \mathcal{F}, L_2)}{\delta^2 \sqrt{n} \|F\|_{P,2}}\right) \|F\|_{P,2}.$$

Maximal inequalities for small functions

THEOREM If $Pf^2 < \delta^2$ and $\|f\|_\infty \leq 1$ for every f in $\mathcal{F}$, then

$$\mathrm{E}_P^* \|\mathbb{G}_n\|_{\mathcal{F}} \lesssim J_{[]}\bigl(\delta, \mathcal{F}, L_2(P)\bigr) \left(1 + \frac{J_{[]}\bigl(\delta, \mathcal{F}, L_2(P)\bigr)}{\delta^2 \sqrt{n}}\right).$$

THEOREM If $Pf^2 < \delta^2 PF^2$ and $F \leq 1$ for every $f \in \mathcal{F}$, then

$$\mathrm{E}_P^* \|\mathbb{G}_n\|_{\mathcal{F}} \lesssim J\bigl(\delta, \mathcal{F}, L_2\bigr) \left(1 + \frac{J(\delta, \mathcal{F}, L_2)}{\delta^2 \sqrt{n} \|F\|_{P,2}}\right) \|F\|_{P,2}.$$

$\|f\|_{P,B} = \Bigl(2P\bigl(e^{|f|} - 1 - |f|\bigr)\Bigr)^{1/2}$ Bernstein “norm”

LEMMA If $\|f\|_{P,B} \leq \delta$ for every $f \in \mathcal{F}$, then

$$\mathrm{E}_P^* \|\mathbb{G}_n\|_{\mathcal{F}} \lesssim J_{[]}\bigl(\delta, \mathcal{F}, \|\cdot\|_{P,B}\bigr) \left(1 + \frac{J_{[]}\bigl(\delta, \mathcal{F}, \|\cdot\|_{P,B}\bigr)}{\delta^2 \sqrt{n}}\right).$$

MLE

$\hat{p}_n$ maximizes $p \mapsto \prod_{i=1}^n p(X_i)$ over $\mathcal{P}_n$.
This implies $\mathbb{P}_n m_{n,\hat{p}_n} \geq \mathbb{P}_n m_{n,p_n}$ for any $p_n \in \mathcal{P}_n$ and

$$m_{n,p} = \log \frac{p + p_n}{2p_n}.$$

LEMMA For $p_0/p_n \leq M$: $P_0(m_{n,p} - m_{n,p_n}) \lesssim -h^2(p, p_n)$ if $h(p, p_n) \geq 32M\, h(p_n, p_0)$

LEMMA For $p_0/p_n \leq M$: $\|m_{n,p} - m_{n,p_n}\|_{P_0,B} \lesssim h(p, p_n)$
For $p \leq q$: $\|m_{n,p} - m_{n,q}\|_{P_0,B} \lesssim h(p, q)$

COROLLARY $J_{[]}\big(\delta, \mathcal{M}_{n,\delta}, \|\cdot\|_{P,B}\big) \leq J_{[]}\big(c\delta, \mathcal{P}_n, h\big)$

$\mathcal{M}_{n,\delta} = \big\{m_{n,p} - m_{n,p_n} : p \in \mathcal{P}_n, h(p, p_n) < \delta\big\}$

MLE (2)

$$\phi_n(\delta) = J_{[]}\big(\delta, \mathcal{P}_n, h\big)\left(1 + \frac{J_{[]}\big(\delta, \mathcal{P}_n, h\big)}{\delta^2\sqrt{n}}\right).$$

The condition $\phi_n(\varepsilon_n) \lesssim \sqrt{n}\varepsilon_n^2$ is equivalent to

$$J_{[]}(\varepsilon_n, \mathcal{P}_n, h) \lesssim \sqrt{n}\varepsilon_n^2.$$

THEOREM

Unsieved MLE ($\mathcal{P}_n = \mathcal{P}$): $h(\hat{p}_n, p_0) = O_{P_0}(\varepsilon_n)$.

Sieved MLE: $h(\hat{p}_n, p_0) = O_P^*(\varepsilon_n) + O(h(p_n, p_0))$ for sequence p_n such that p_0/p_n is uniformly bounded.

Classification

$(X_1, Y_1), \ldots, (X_n, Y_n)$ i.i.d. in $\mathcal{X} \times \{0, 1\}$.
Θ a collection of measurable functions $\theta\colon \mathcal{X} \to \{0, 1\}$.
$m_\theta(x, y) = 1_{y \neq \theta(x)}$

LEMMA The risk $Pm_\theta = \mathrm{P}\big(Y \neq \theta(X)\big)$ is minimized over *all* measurable maps $\theta\colon \mathcal{X} \to \{0, 1\}$ by the Bayes classifier

$$\theta_0(x) = 1_{\eta(x) > 1/2}, \qquad \eta(x) = \mathrm{P}(Y = 1 | \, X = x).$$

LEMMA [Tsybakov's condition] If $P1_{|\eta - 1/2| \leq t} \leq Ct^\kappa$ for every $t > 0$ and some $\kappa > 0$, then $P(m_\theta - m_{\theta_0}) \gtrsim \bar{d}^2(\theta, \theta_0)$ for

$$\bar{d}(\theta, \theta_0) = \|m_\theta - m_{\theta_0}\|_1^{(1+1/\kappa)/2} = \|m_\theta - m_{\theta_0}\|_2^{1+1/\kappa}.$$

THEOREM If $\theta_0 \in \Theta$ and Tsybakov's condition holds, then $P(m_{\hat{\theta}} - m_{\theta_0}) = O_P(\varepsilon_n^2)$ for ε_n satisfying $J(\varepsilon_n^{\kappa/(\kappa+1)}) \lesssim \sqrt{n}\varepsilon_n^2$, where J is $J(\cdot, \Theta, L_2)$ or $J_{[]}\big(\cdot, \Theta, L_2(P)\big)$.

Model selection

Θ_k subset of a metric space (Θ, d), for $k \in \mathcal{K}$, countable.
$\theta^* = \operatorname{argmin}_{\theta\in\Theta} Pm_\theta$.
$\theta_k^* = \operatorname{argmin}_{\theta\in\Theta_k} Pm_\theta$.

$$\hat\theta_k = \operatorname*{argmin}_{\theta\in\Theta_k} \mathbb{P}_n m_\theta, \qquad \hat k = \operatorname*{argmin}_{k\in\mathcal{K}} \big(\mathbb{P}_n m_{\hat\theta_k} + J(k)\big).$$

THEOREM If
$P(m_\theta - m_{\theta'})^2 \le d^2(\theta,\theta')$.
$\|m_\theta\|_\infty \le 1, \qquad \theta \in \Theta_k$.
$P(m_\theta - m_{\theta^*}) \ge d^2(\theta,\theta^*), \qquad \theta \in \Theta_k$.
$\mathrm{E}^* \sup_{\theta\in\Theta_k : d(\theta,\bar\theta_k)<\delta} \mathbb{G}_n(m_{\bar\theta_k} - m_\theta) \le \phi_{n,k}(\delta), \qquad \bar\theta_k = \operatorname{argmin}_{\theta\in\Theta_k} d(\theta,\theta^*)$.
$\phi_{n,k}(\varepsilon_{n,k}) \le \sqrt{n}\varepsilon_{n,k}^2$.
$J(k) \gtrsim \varepsilon_{n,k}^2 + (x_k+1)/n, \qquad \sum_k e^{-x_k} \le 1$.
then $\hat\theta = \hat\theta_{\hat k}$ satisfies

$$\mathrm{E}^* P(m_{\hat\theta} - m_{\theta^*}) \lesssim \inf_{k\in\mathcal{K}} \left(P(m_{\theta_k^*} - m_{\theta^*}) + J(k) + \frac{1}{n} \right).$$

Tests and Le Cam-Birgé estimators

Overview

1: ENTROPY
2: EMPIRICAL PROCESSES AND MAXIMAL INEQUALITIES
3: M-ESTIMATORS AND MODEL SELECTION
4: TESTS AND LECAM-BIRGÉ ESTIMATORS
5: POSTERIOR DISTRIBUTIONS

Minimax risk for testing

$\mathcal{P}$ and $\mathcal{Q}$ collections of probability measures

DEFINITION *Minimax risk for testing* $\mathcal{P}$ versus $\mathcal{Q}$:

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

The infimum is taken over all measurable functions $\phi \colon \mathcal{X} \to [0,1]$.

Minimax risk for testing

$\mathcal{P}$ and $\mathcal{Q}$ collections of probability measures

DEFINITION *Minimax risk for testing* $\mathcal{P}$ versus $\mathcal{Q}$:

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

Infimum taken over all measurable functions $\phi \colon \mathcal{X} \to [0, 1]$.

LEMMA If $\mathcal{P}$ and $\mathcal{Q}$ are dominated, then

$$\pi(\mathcal{P}, \mathcal{Q}) = \sup_{P \in \mathrm{conv}\ (\mathcal{P})} \sup_{Q \in \mathrm{conv}\ (\mathcal{Q})} \big(P(p < q) + Q(p \geq q) \big)$$

Testing affinity

DEFINITION

$$\rho_\alpha(\mathcal{P},\mathcal{Q}) = \sup_{P\in\mathcal{P}}\sup_{Q\in\mathcal{Q}} \int p^\alpha q^{1-\alpha}\, d\mu$$

LEMMA

$$\pi(\mathcal{P},\mathcal{Q}) \le \rho_\alpha\big(\text{conv }(\mathcal{P}),\text{conv }(\mathcal{Q})\big)$$

PROOF

$$\pi(\mathcal{P},\mathcal{Q}) = \sup_{P\in\text{conv }(\mathcal{P})}\sup_{Q\in\text{conv }(\mathcal{Q})} \big(P(p<q)+Q(p\ge q)\big)$$

$$\begin{aligned} P(p<q)+Q(p\ge q) &= \int_{p<q} p\, d\mu + \int_{q\le p} q\, d\mu \\ &\le \int_{p<q} p^\alpha q^{1-\alpha}\, d\mu + \int_{q\le p} p^\alpha q^{1-\alpha}\, d\mu \end{aligned}$$

Testing affinity

DEFINITION

$$\rho_\alpha(\mathcal{P}, \mathcal{Q}) = \sup_{P \in \mathcal{P}} \sup_{Q \in \mathcal{Q}} \int p^\alpha q^{1-\alpha} \, d\mu$$

LEMMA

$$\pi(\mathcal{P}, \mathcal{Q}) \leq \rho_\alpha\big(\text{conv }(\mathcal{P}), \text{conv }(\mathcal{Q})\big)$$

LEMMA

$$\rho_\alpha\big(\text{conv }(\mathcal{P}^n), \text{conv }(\mathcal{Q}^n)\big) \leq \rho_\alpha\big(\text{conv }(\mathcal{P}), \text{conv }(\mathcal{Q})\big)^n$$

$\mathcal{P}^n = \{P^n : P \in \mathcal{P}\}$

Testing product measures

THEOREM For convex, dominated $\mathcal{P}$, $\mathcal{Q}$ there exists a test ϕ_n

$$\sup_{P\in\mathcal{P}} P^n\phi_n \le e^{-\frac{1}{2}nh^2(\mathcal{P},\mathcal{Q})}$$

$$\sup_{Q\in\mathcal{Q}} Q^n(1-\phi_n) \le e^{-\frac{1}{2}nh^2(\mathcal{P},\mathcal{Q})}$$

PROOF

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

$$\rho_{1/2}(P,Q)^n \le e^{-\frac{1}{2}nh^2(P,Q)}$$

Testing product measures

THEOREM For convex, dominated $\mathcal{P}$, $\mathcal{Q}$ there exists a test ϕ_n

$$\sup_{P \in \mathcal{P}} P^n \phi_n \leq e^{-\frac{1}{2} n h^2(\mathcal{P}, \mathcal{Q})}$$

$$\sup_{Q \in \mathcal{Q}} Q^n (1 - \phi_n) \leq e^{-\frac{1}{2} n h^2(\mathcal{P}, \mathcal{Q})}$$

Consider metric d with: $d \leq h$, convex balls

COROLLARY For every p_1, p_2 there exists a test ϕ_n with

$$\sup_{p : d(p,p_1) < d(p_1,p_2)/3} P^n \phi_n \leq e^{-K n d^2(p_1,p_2)}$$

$$\sup_{p : d(p,p_2) < d(p_1,p_2)/3} P^n (1 - \phi_n) \leq e^{-K n d^2(p_1,p_2)}$$

Le Cam-Birgé estimators

Le Cam-Birgé

Let the observations be a random sample $X_1, \ldots, X_n$ from some P_0 that is known to belong to a model $\mathcal{P}$. Let p denote the density of P.

THEOREM If

$$\log N(\varepsilon_n, \mathcal{P}, h) \leq n\varepsilon_n^2,$$

then there exist estimators $\hat{p}_n$ such that

$$h(\hat{p}_n, p_0) = O_{P_0}(\varepsilon_n).$$

Le Cam-Birgé versus MLE

$$J_{[]}(\varepsilon, \mathcal{P}, h) := \int_0^{\varepsilon} \sqrt{\log N_{[]}\left(\eta, \mathcal{P}, h\right)}\, d\eta$$

LE CAM-BIRGÉ: $\log N(\varepsilon_n, \mathcal{P}, h) \leq n\varepsilon_n^2$

MLE: $J_{[]}(\varepsilon_n, \mathcal{P}, h) \leq \sqrt{n}\varepsilon_n^2$

Always $J_{[]}(\varepsilon, \mathcal{P}, h) \geq \varepsilon\sqrt{\log N_{[]}\left(\varepsilon, \mathcal{P}, h\right)}$.

If two sides of inequality are equivalent up to constants, then

MLE: $\log N_{[]}\left(\varepsilon_n, \mathcal{P}, h\right) \leq n\varepsilon_n^2$.

Le Cam-Birgé estimators

Take any metric $d \leq h$.
Fix a minimal ε_n-net $\mathcal{P}_n$ over $\mathcal{P}$ for d.
Fix a test ϕ_{n,p_1,p_2} of $B(p_1, \varepsilon_n)$ versus $B(p_2, \varepsilon_n)$ as before, for every $p_1, p_2 \in \mathcal{P}_n$ with $d(p_1, p_2) > 3\varepsilon_n$.
Let $\delta_{n,p_1} = \max\bigl\{d(p_1, p_2) \colon \phi_{n,p_1,p_2} = 1\bigr\}$.
Let $\hat{p}_n$ minimize $p \mapsto \delta_{n,p}$ over $\mathcal{P}_n$.

THEOREM If $\log N\bigl(\varepsilon_n, \mathcal{P}, d\bigr) \leq n\varepsilon_n^2$, then $d(\hat{p}_n, p_0) = O_{P_0}(\varepsilon_n)$.

Le Cam-Birgé estimators

Take any metric $d \leq h$.
Fix a minimal ε_n-net $\mathcal{P}_n$ over $\mathcal{P}$ for d.
Fix a test ϕ_{n,p_1,p_2} of $B(p_1, \varepsilon_n)$ versus $B(p_2, \varepsilon_n)$ as before, for every $p_1, p_2 \in \mathcal{P}_n$ with $d(p_1, p_2) > 3\varepsilon_n$.
Let $\delta_{n,p_1} = \max\bigl\{d(p_1, p_2)\colon \phi_{n,p_1,p_2} = 1\bigr\}$.
Let $\hat{p}_n$ minimize $p \mapsto \delta_{n,p}$ over $\mathcal{P}_n$.

THEOREM If $\log N\bigl(\varepsilon_n, \mathcal{P}, d\bigr) \leq n\varepsilon_n^2$, then $d(\hat{p}_n, p_0) = O_{P_0}(\varepsilon_n)$.

Can replace global entropy by Le Cam dimension:

$$L(\varepsilon, \mathcal{P}, d; P_0) = \sup_{\eta \geq \varepsilon} N\bigl(\eta, \{p\colon d(p, p_0) < 2\eta, d\}, d\bigr).$$

THEOREM If $L(\varepsilon_n, \mathcal{P}, , d; P_0) \leq n\varepsilon_n^2$, then there exist estimators $\hat{p}_n$ such that $\mathrm{E}_{P_0} h^2(\hat{p}_n, p_0) \leq C\varepsilon_n^2$ for some C.

EXAMPLE If Le Cam dimension is finite, then $\varepsilon_n = n^{-1/2}$.

Testing complements of balls

Consider testing $\{P_0\}$ versus $\{p\colon d(p,p_0) \geq \varepsilon\}$.

THEOREM Suppose $d \leq h$ and for a nonincreasing $\varepsilon \mapsto D(\varepsilon)$

$$D\Big(\frac{\varepsilon}{2}, \big\{P\colon \varepsilon \leq d(P,P_0) \leq 2\varepsilon\big\}, d\Big) \leq D(\varepsilon), \qquad \text{every } \varepsilon > \varepsilon_n$$

Then for every $\varepsilon > \varepsilon_n$ there exist test ϕ_n

$$P_0^n \phi_n \leq D(\varepsilon) e^{-Kn\varepsilon^2} \frac{1}{1 - e^{-Kn\varepsilon^2}},$$

$$\sup_{d(P,P_0) > j\varepsilon} P^n(1-\phi_n) \leq e^{-Kn\varepsilon^2 j^2}, \qquad \text{every } j \in \mathbb{N}.$$

Posterior distributions

Overview

1: ENTROPY
2: EMPIRICAL PROCESSES AND MAXIMAL INEQUALITIES
3: M-ESTIMATORS AND MODEL SELECTION
4: TESTS AND LECAM-BIRGÉ ESTIMATORS
5: POSTERIOR DISTRIBUTIONS

Bayes

prior: Π_n

posterior: $$\Pi_n(B|\, X_1, \ldots, X_n) = \frac{\int_B \prod_{i=1}^n p(X_i)\, d\Pi_n(P)}{\int \prod_{i=1}^n p(X_i) d\Pi_n(P)}$$

THEOREM If $n\varepsilon_n^2 \to \infty$ and there exist $\mathcal{P}_n \subset \mathcal{P}$ with

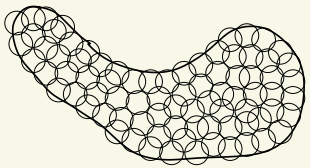
$$\log N(\varepsilon_n, \mathcal{P}_n, h) \leq n\varepsilon_n^2,$$
$$\Pi_n\big(\mathcal{P}_n^c\big) \leq e^{-4n\varepsilon_n^2},$$
$$\Pi_n\big(B(P_0, \varepsilon_n)\big) \geq e^{-n\varepsilon_n^2},$$

then $\mathrm{E}_{P_0}\Pi_n(P\colon h(P, P_0) \geq M\varepsilon_n|\, X_1, \ldots, X_n) \to 0$ for large M.

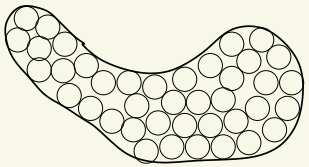
$$B(P_0, \varepsilon_n) = \left\{P\colon -P_0\left(\log \frac{p}{p_0}\right) \leq \varepsilon_n^2, P_0\left(\log \frac{p}{p_0}\right)^2 \leq \varepsilon_n^2\right\}$$

Motivation prior mass

$$\Pi_n\big(B(P_0, \varepsilon_n)\big) \geq e^{-n\varepsilon_n^2}$$



Need $N(\varepsilon, \mathcal{P}, d) \approx \exp(n\varepsilon_n^2)$ balls



Can place $\exp(Cn\varepsilon_n^2)$ balls

If Π_n “uniform”, then each ball receives mass $\exp(-Cn\varepsilon_n^2)$

Equivalence KL and Hellinger

$$B(P_0, \varepsilon_n) = \left\{ P \colon -P_0 \left(\log \frac{p}{p_0} \right) \leq \varepsilon_n^2, P_0 \left(\log \frac{p}{p_0} \right)^2 \leq \varepsilon_n^2 \right\}.$$

This is close to a Hellinger ball:

LEMMA For every p, q,

$$P \log \frac{p}{q} \lesssim h^2(p,q) \left(1 + \log \left\| \frac{p}{q} \right\|_\infty \right),$$
$$P \left(\log \frac{p}{q} \right)^2 \lesssim h^2(p,q) \left(1 + \log \left\| \frac{p}{q} \right\|_\infty \right)^2.$$

LEMMA For every $b > 0$ exists $\eta_b > 0$ with for every p, q with $0 < h^2(p,q) < \eta_b P(p/q)^b$,

$$P \log \frac{p}{q} \lesssim h^2(p,q) \left(1 + \frac{1}{b} \log_+ \frac{1}{h(p,q)} + \frac{1}{b} \log_+ P \left(\frac{p}{q} \right)^b \right),$$
$$P \left(\log \frac{p}{q} \right)^2 \lesssim h^2(p,q) \left(1 + \frac{1}{b} \log_+ \frac{1}{h(p,q)} + \frac{1}{b} \log_+ P \left(\frac{p}{q} \right)^b \right)^2.$$

Bayes — test version

prior: Π_n

posterior: $$\Pi_n(B|\, X_1, \ldots, X_n) = \frac{\int_B \prod_{i=1}^n p(X_i)\, d\Pi_n(P)}{\int \prod_{i=1}^n p(X_i) d\Pi_n(P)}.$$

THEOREM If $n\varepsilon_n^2 \to \infty$ and there exist $\mathcal{P}_n \subset \mathcal{P}$ and tests ϕ_n with for every j:

$$P_0^n \phi_n \to 0,$$

$$\sup_{P \in \mathcal{P}_n : d(P,P_0) > j\varepsilon_n} P^n(1-\phi_n) \le e^{-Kn\varepsilon_n^2 j^2},$$

and

$$\Pi_n\big(\mathcal{P}_n^c\big) \le e^{-4n\varepsilon_n^2},$$

$$\Pi_n\big(B(P_0, \varepsilon_n)\big) \ge e^{-n\varepsilon_n^2},$$

then $\mathrm{E}_{P_0} \Pi_n(P\colon h(P,P_0) \ge M\varepsilon_n |\, X_1, \ldots, X_n) \to 0$ for large M.

Technical lemma

LEMMA For every $\varepsilon > 0$

$$P_0^n\left(\int \prod_{i=1}^{n} \frac{p}{p_0}(X_i)\, d\Pi(P) \leq \Pi_n(B(P_0, \varepsilon_n))e^{-2n\varepsilon^2}\right) \leq \frac{1}{n\varepsilon^2}.$$