EcoSta2018, City University of Hong Kong, 19 June 2018

What is asymptotically testable and what is not?

Bas Kleijn

KdV Institute for Mathematics

Universiteit van Amsterdam

Asymptotic symmetric testing

Observe *i.i.d.* data $X^n \sim P^n$, model $P \in \mathscr{P}$; for disjoint $B, V \subset \mathscr{P}$,

$$H_0 : P \in B, \quad \text{or} \quad H_1 : P \in V.$$

Look for test functions $\phi_n : \mathscr{X}^n \to [0, 1]$ s.t.

$$P^n \phi_n(X^n) \to 0, \quad \text{and} \quad Q^n(1 - \phi_n(X^n)) \to 0$$

for all $P \in B$ and all $Q \in V$.

Equivalently, we want,

> A testing procedure that chooses for B or V based on $X^n \sim P^n$ for every $n \geq 1$, has property (D) if it is wrong only a finite number of times with P^∞-probability one.

Property (D) is sometimes referred to as "discernibility".

Some examples and unexpected answers (I)

Consider non-parametric regression with $f : X \to \mathbb{R}$ and test for smoothness,

$$H_0 : f \in C^1(X \to \mathbb{R}), \quad H_1 : f \in C^2(X \to \mathbb{R}),$$

Consider a non-parametric density estimation with $p : \mathbb{R} \to [0, \infty)$ and test for square-integrability,

$$H_0 : \int x^2 \, p(x) \, dx < \infty, \quad H_1 : \int x^2 \, p(x) \, dx = \infty.$$

Practical problem we cannot use the data to determine with asymptotic certainty, if CLT applies with our data.

Some examples and unexpected answers (II)

Coin-flip $X^n \sim \text{Bernoulli}(p)^n$ with $p \in [0, 1]$.

Consider Cover's rational mean problem (1973):

$$H_0 : p \in [0, 1] \cap \mathbb{Q}, \quad H_1 : p \in [0, 1] \setminus \mathbb{Q}.$$

Consider also Dembo and Peres's irrational alternative (1995):

$$H_0 : p \in [0, 1] \cap \mathbb{Q}, \quad H_1 : p \in [0, 1] \cap \sqrt{2} + \mathbb{Q},$$

Consider ultimately fractal hypotheses, *e.g.* with Cantor set C,

$$H_0 : p \in C, \quad H_1 : p \in [0, 1] \setminus C.$$

Three forms of testability

Definition 5.1 (ϕ_n) *is a uniform test sequence for* B *vs* V*, if,*

$$\sup_{P\in B} P^n\phi_n \to 0, \quad \sup_{Q\in V} Q^n(1-\phi_n) \to 0. \tag{1}$$

Definition 5.2 (ϕ_n) *is a pointwise test sequence for* B *vs* V*, if,*

$$\phi_n(X^n) \xrightarrow{P} 0, \quad \phi_n(X^n) \xrightarrow{Q} 1, \tag{2}$$

for all $P \in B$ *and* $Q \in V$*.*

Definition 5.3 (ϕ_n) *is a Bayesian test sequence for* B *vs* V*, if,*

$$\phi_n(X^n) \xrightarrow{P} 0, \quad \phi_n(X^n) \xrightarrow{Q} 1, \tag{3}$$

for Π*-almost-all* $P \in B$ *and* $Q \in V$*.*

Posterior odds model selection for frequentists

Johnson & Rossell (JRSSB, 2010), Taylor & Tibshirani (PNAS, 2016)

Theorem 6.1 *Given measurable* $B, V \subset \Theta$ $(\Pi(B), \Pi(V) > 0)$ *and,*

i. there are Bayesian tests for B *vs* V *of power* $a_n \downarrow 0$,

$$\int_B P^n \phi_n \, d\Pi(P) + \int_V Q^n (1 - \phi_n) \, d\Pi(Q) = o(a_n),$$

ii. and, for all $P \in B$, $P^n \lhd a_n^{-1} P_n^{\Pi|B}$; *for all* $Q \in V$, $Q^n \lhd a_n^{-1} P_n^{\Pi|V}$,

then posterior odds give rise to a pointwise test for B *vs* V.

See BK, "The frequentist validity of Bayesian limits", arXiv:1611.08444 [math.ST]

Example: KL-neighbourhoods

Definition 7.1 *Given (P_n), (Q_n) and a $a_n \downarrow 0$, Q_n is a_n-remotely contiguous w.r.t. P_n ($Q_n \lhd a_n^{-1} P_n$), if for any msb $\psi_n : \mathscr{X}^n \to [0,1]$*

$$P_n\psi_n = o(a_n) \quad \Rightarrow \quad Q_n\psi_n = o(1)$$

Example 7.2 *Let $\mathscr{P}$ be a model for i.i.d. data X^n. Let P_0, P and $\epsilon > 0$ be such that $-P_0 \log(dP/dP_0) < \epsilon^2$. Then, for large enough n,*

$$\frac{dP^n}{dP_0^n}(X^n) \geq e^{-\frac{n}{2}\epsilon^2}, \tag{4}$$

with P_0^n-probability one. So for any tests ψ_n,

$$P^n\psi_n \geq e^{-\frac{1}{2}n\epsilon^2} P_0^n\psi_n. \tag{5}$$

So if $P^n\phi_n = o(\exp(-\frac{1}{2}n\epsilon^2))$ then $P_0^n\phi_n = o(1)$: $P_0^n \lhd a_n^{-1} P^n$ with $a_n = \exp(-\frac{1}{2}n\epsilon^2)$.

Example: select the DAG (I)

Observe an *i.i.d.* X^n of vectors of discrete random variables $X_i = (X_{1,i}, \ldots, X_{k,i}) \in \mathbb{Z}^k$, $1 \le i \le n$.

Define a family $\mathscr{F}$ of kernels $p_\theta(\cdot|\cdot) : \mathbb{Z} \times \mathbb{Z}^l \to [0,1]$, for $\theta \in \Theta$, $1 \le l \le k$. Assume that Θ is compact and,

$$\theta \mapsto \sum_{x \in \mathbb{Z}} f(x) P_\theta(x | z_1, \ldots, z_l)$$

is continuous, for every bounded $f : \mathbb{Z} \to \mathbb{R}$ and all $z_1, \ldots, z_l \in \mathbb{Z}$.

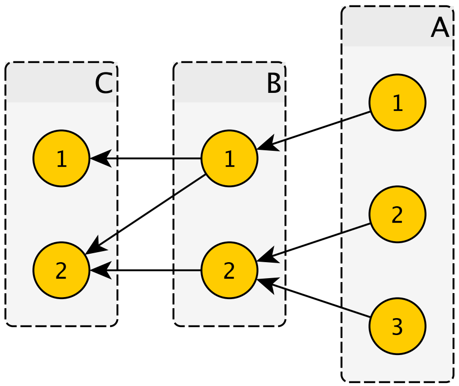
$X \sim P$ follows a graphical model,

$$P_{\mathscr{A},\theta}(X_1 \in B_1, \ldots, X_k \in B_k) = \prod_{i=1}^{k} P_{\theta_i}(X_i \in B_i | \mathscr{A}_i)$$

where $\mathscr{A}_i \subset \{1, \ldots, k\}$ denotes the *parents* of X_i (and $\mathscr{A}_{ij} = \mathscr{A}_i \cup \mathscr{A}_j$). Together, the $\mathscr{A}_i$ describe a directed, a-cyclical graph (DAG).

Example: select the DAG (II)

The DAG $\mathscr{A} = (\mathscr{A}_i : 1 \leq i \leq k)$ represents a number of conditional independence statements concerning the components $X_1, \ldots, X_k$.



$$P_{\mathscr{A},\theta}(C_1 \in \cdot, \ldots, A_3 \in \cdot)$$
$$= P_{\theta_{C,1}}(\cdot|B_1) \times P_{\theta_{C,2}}(\cdot|B_1, B_2)$$
$$\times P_{\theta_{B,1}}(\cdot|A_1) \times P_{\theta_{B_2}}(\cdot|A_2, A_3)$$
$$\times P_{\theta_{A,1}}(\cdot) \times P_{\theta_{A,2}}(\cdot) \times P_{\theta_{A,3}}(\cdot)$$

Fig 1. An small example DAG: No arrow means $X_i \perp X_j | \mathscr{A}_{ij}$. $\mathscr{A}_{C_1} = \{B_1\}$, $\mathscr{A}_{B_2} = \{A_2, A_3\}$, so given B_1, A_2 and A_3, C_1 is independent of B_2.

Example: select the DAG (III)

Define the submodels $\mathscr{P}_{\mathscr{A}} = \{P_{\mathscr{A},\theta} : \theta \in \Theta^k\}$, for all $\mathscr{A}$. Given any $\mathscr{A}' \neq \mathscr{A}$, there is a pair $X_i \perp X_j | \mathscr{A}_{ij}$ but $X_i \not\perp X_j | \mathscr{A}'_{ij}$.

Require that, for all θ, all $A, B \subset \mathbb{Z}$,

$$\left| P_{\mathscr{A}',\theta}(X_i \in A, X_j \in B | \mathscr{A}_{ij}) - P_{\mathscr{A}',\theta}(X_i \in A | \mathscr{A}_{ij})\, P_{\mathscr{A}',\theta}(X_j \in B | \mathscr{A}_{ij}) \right| > \epsilon,$$

for some $\epsilon > 0$ that depends only on $\mathscr{A}$ and $\mathscr{A}'$.

With a KL-prior posterior odds for $\mathscr{P}_{\mathscr{A}}$ select the correct DAG $\mathscr{A}$.

Uniform testability: equivalent formulations

Proposition 11.1 *Let $\mathscr{P}$ be a model for i.i.d. data with disjoint B and V. The following are equivalent:*

i. there exists a uniform test sequence (ϕ_n),

$$\sup_{P\in B} P^n\phi_n \to 0, \qquad \sup_{Q\in V} Q^n(1-\phi_n) \to 0,$$

ii. there is a exponentially powerful uniform test sequence (ψ_n),

$$\sup_{P\in B} P^n\psi_n \leq e^{-nD}, \qquad \sup_{Q\in V} Q^n(1-\psi_n) \leq e^{-nD}.$$

The model as a uniform space

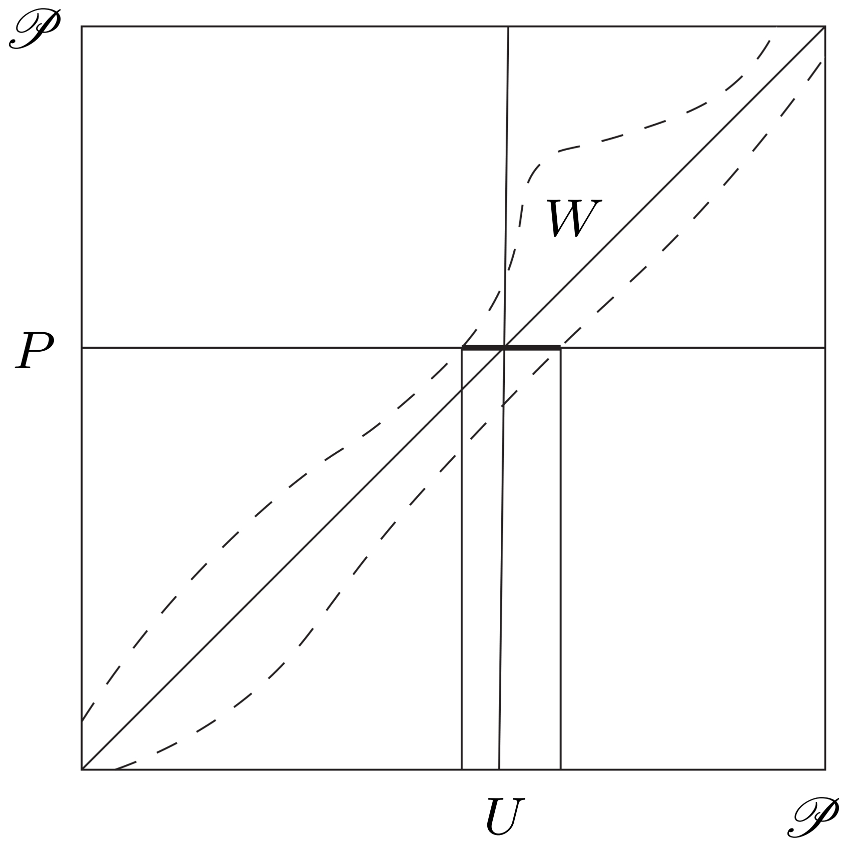


Fig 2. Let $P \in \mathscr{P}$ and entourage $W \in \mathscr{U}_\infty$ be given. Define neighbourhood $U \in \mathscr{T}_\infty$ as $U = \{Q \in \mathscr{P} : (Q, P) \in W\}$

Uniform separation (II)

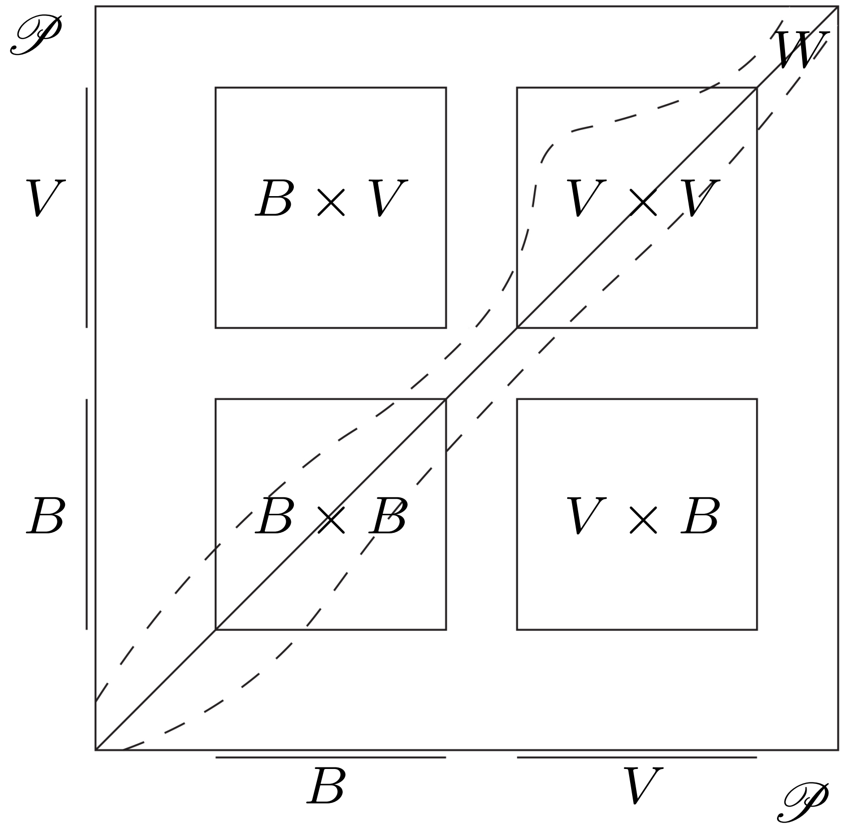


Fig 3. B and V are uniformly separated by $\mathscr{U}_\infty$ if there is a $W \in \mathscr{U}_\infty$ that does not meet $B \times V$ and $V \times B$.

Characterisation of uniform testability

Theorem 14.1 *Let $\mathscr{P}$ be a model for i.i.d. data with disjoint B and V. The following are equivalent:*

(i.) there are uniform tests ϕ_n for B versus V,

(ii.) B and V are uniformly separated by $\mathscr{U}_\infty$.

Corollary 14.2 (Parametrised models) *Suppose $\mathscr{P} = \{p_\theta : \theta \in \Theta\}$, with (Θ, d) compact, metric space and $\theta \to P_\theta$ identifiable and $\mathscr{T}_\infty$-continuous, (that is, for every $f \in \mathscr{F}_n$, $\theta \mapsto \int f \, dP_\theta^n$ is continuous). If $B_0, V_0 \subset \Theta$ with $d(B_0, V_0) > 0$, then the images $B = \{P_\theta : \theta \in B_0\}$, $V = \{P_\theta : \theta \in V_0\}$ are uniformly testable.*

Pointwise testability: equivalent formulations

Proposition 15.1 *Let* $\mathscr{P}$ *be a model for i.i.d. data and let* B, V *be disjoint model subsets. The following are equivalent:*

i. there are tests (ϕ_n) *such that, for all* $P \in B$ *and* $Q \in V$*,*

$$P^n \phi_n \to 0, \quad Q^n(1-\phi_n) \to 0,$$

ii. there are tests (ϕ_n) *such that, for all* $P \in B$ *and* $Q \in V$*,*

$$\phi_n(X^n) \xrightarrow{P} 0, \quad (1-\phi_n(X^n)) \xrightarrow{Q} 0,$$

iii. there are tests (ϕ_n) *such that, for all* $P \in B$ *and* $Q \in V$*,*

$$\phi_n(X^n) \xrightarrow{P\text{-a.s.}} 0, \quad (1-\phi_n(X^n)) \xrightarrow{Q\text{-a.s.}} 0.$$

Pointwise testability in dominated models

Definition 16.1 *The testing problem has a (uniform) representation on* X*, if there exists a* $\mathscr{T}_\infty$*-(uniformly-)continuous, surjective map* $f : B \cup V \to X$ *such that* $f(B) \cap f(V) = \varnothing$*.*

Definition 16.2 *The model is parametrised by* Θ*, if there exists a* $\mathscr{T}_\infty$*-continuous bijection* $P_\cdot : \Theta \to \mathscr{P}$ *(i.e. for every* $m \geq 1$ *and measurable* $f : \mathscr{X}^m \to [0,1]$*, the map* $\theta \mapsto \int f \, dP_\theta^m$ *is continuous).*

Characterisation of pointwise testability

Theorem 17.1 *Let $\mathscr{P}$ be a dominated model for i.i.d. data with disjoint B,V. The following are equivalent,*

i. there exists a pointwise test for B vs V,

ii. the problem has a representation $f : B \cup V \to X$ on a normal space X and there exist disjoint F_σ-sets $B', V' \subset X$ such that $f(B) \subset B', f(V) \subset V'$,

iii. the problem has a uniform representation $\psi : B \cup V \to X$ on a separable, metrizable space X with $\psi(B), \psi(V)$ both F_σ- and G_δ-sets.

Finite entropy and uniform integrability

Corollary 18.1 *Suppose that $\mathscr{P}$ is dominated and TV-totally-bounded. Then disjoint $B, V \subset \mathscr{P}$ are pointwise testable, if and only if, B, V are both F_σ- and G_δ-sets in $B \cup V$ (for $\mathscr{T}_{TV}$).*

Corollary 18.2 *Suppose that $\mathscr{P}$ is dominated by a probability measure, with a uniformly integrable family of densities. Then disjoint $B, V \subset \mathscr{P}$ are pointwise testable, if and only if, B, V are both F_σ- and G_δ-sets in $B \cup V$ (for $\mathscr{T}_C$).*

Bayesian testability: equivalent formulations

Theorem 19.1 *Let a model* $(\mathscr{P}, \mathscr{G}, \Pi)$ *with* $B, V \in \mathscr{G}$ *be given, with* $\Pi(B) > 0, \Pi(V) > 0$. *The following are equivalent,*

i. there exist Bayesian tests for B *vs* V,

ii. there are tests ϕ_n *such that for* Π*-almost-all* $P \in B, Q \in V$,

$$P^n\phi_n \to 0, \quad Q^n(1-\phi_n) \to 0,$$

iii. there are tests $\phi_n : \mathscr{X}^n \to [0,1]$ *such that,*

$$\int_B P^n\phi_n \, d\Pi(P) + \int_V Q^n(1-\phi_n)\, d\Pi(Q) \to 0,$$

iv. for Π*-almost-all* $P \in B$, $Q \in V$,

$$\Pi(V|X^n) \xrightarrow{P} 0, \quad \Pi(B|X^n) \xrightarrow{Q} 0.$$

Characterisation of Bayesian testability

Theorem 20.1 *Let $(\mathscr{P},\mathscr{G})$ be a measurable model with a prior* Π *that is a Radon measure and hypotheses B,V. There is a Bayesian test sequence for B vs V, if and only if, B,V are $\mathscr{G}$-measurable.*

Consistent model selection

Let $\mathscr{P}$ be a model for *i.i.d.* data $X^n \sim P^n$, $(n \geq 1)$, and suppose that $(\mathscr{P}, \mathscr{G}, \Pi)$ has finite, measurable partition,

$$P \in \mathscr{P} = \mathscr{P}_1 \cup \ldots \cup \mathscr{P}_M.$$

Model-selection Which $1 \leq i \leq M$? (such that $P \in \mathscr{P}_i$)

Theorem 21.1 *Assume that for all $1 \leq i < j \leq M$,*

$\mathscr{P}_i$ and $\mathscr{P}_j$ are $\mathscr{U}_\infty$-uniformly separated.

Let $1 \leq i \leq M$ be such that $P \in \mathscr{P}_i$. If Π is a KL-prior, then indicators for posterior odds,

$$\phi_n(X^n) = 1\Big\{X^n : \Pi(\mathscr{P}_i | X^n) \geq \sum_{j \neq i} \Pi(\mathscr{P}_j | X^n)\Big\},$$

are a pointwise test for $\mathscr{P}_i$ vs $\cup_{j \neq i} \mathscr{P}_j$.

Thank you for your attention

BK, ”The frequentist validity of Bayesian limits”
arXiv:1611.08444 [math.ST]

Remote contiguity

Definition 23.1 *Given* (P_n), (Q_n) *and a* $a_n \downarrow 0$, Q_n *is* a_n*-remotely contiguous w.r.t.* P_n ($Q_n \lhd a_n^{-1} P_n$), *if for any msb* $\psi_n : \mathscr{X}^n \to [0, 1]$

$$P_n \psi_n = o(a_n) \quad \Rightarrow \quad Q_n \psi_n = o(1)$$

Lemma 23.2 $Q_n \lhd a_n^{-1} P_n$ *if any of the following holds:*

(i) For any bnd msb $T_n : \mathscr{X}^n \to \mathbb{R}$, $a_n^{-1} T_n \xrightarrow{P_n} 0$, *implies* $T_n \xrightarrow{Q_n} 0$

(ii) Given $\epsilon > 0$, *there is a* $\delta > 0$ *s.t.* $Q_n(dP_n/dQ_n < \delta\, a_n) < \epsilon$ *f.l.e.n.*

(iii) There is a $b > 0$ *s.t.* $\liminf_{n\to\infty} b\, a_n^{-1}\, P_n(dQ_n/dP_n > b\, a_n^{-1}) = 1$

(iv) Given $\epsilon > 0$, *there is a* $c > 0$ *such that* $\|Q_n - Q_n \wedge c\, a_n^{-1} P_n\| < \epsilon$

(v) Under Q_n, *every subsequence of* $(a_n (dP_n/dQ_n)^{-1})$ *has a further subsequence that converges in* $\mathscr{T}_C$.

The model as a uniform space

Take $\mathscr{X}$ a separable metrizable space, with Borel σ-algebra $\mathscr{B}$.

The class $\mathscr{F}_n$ contains all bounded, $\mathscr{B}^n$-measurable $f : \mathscr{X}^n \to \mathbb{R}$.

For every $n \geq 1$ and $f \in \mathscr{F}_n$, define the entourage,

$$W_{n,f} = \{(P,Q) \in \mathscr{P} \times \mathscr{P} : |P^n f - Q^n f| < 1\}.$$

Defines uniformity $\mathscr{U}_n$ (with topology $\mathscr{T}_n$). Take $\mathscr{U}_\infty = \cup_{n\geq 1}\mathscr{U}_n$.

$$P \to Q \ \text{ in } \mathscr{T}_\infty \quad \Leftrightarrow \quad \int f\, dP^n \to \int f\, dQ^n,$$

for all $n \geq 1$ and all $f \in \mathscr{F}_n$. Note also,

$$\mathscr{U}_C \subset \mathscr{U}_1 \subset \cdots \subset \mathscr{U}_\infty \subset \mathscr{U}_{TV}.$$

The Dunford-Pettis theorem

Theorem 25.1 (Dunford-Pettis) *Assume $\mathscr{P}$ is dominated by a probability measure Q with densities in $\mathscr{P}_Q \subset L^1(Q)$; $\mathscr{P}_Q$ is relatively weakly compact, if and only if, for every $\epsilon > 0$ there is an $M > 0$ such that,*

$$\sup_{P \in \mathscr{P}} \int_{\{dP/dQ > M\}} \frac{dP}{dQ} \, dQ < \epsilon,$$

that is, $\mathscr{P}_Q$ is uniformly Q-integrable.

Uniform separation

Definition 26.1 *Subsets $B, V \subset \mathscr{P}$ are uniformly separated by $\mathscr{U}_\infty$, if there exists an entourage $W \in \mathscr{U}_\infty$ such that,*

$$(B \times V \cup V \times B) \cap W = \varnothing.$$

In other words, there are $J, m \geq 1$, $\epsilon > 0$ and bounded, measurable functions $f_1, \ldots, f_J : \mathscr{X}^m \to [0, 1]$ such that, for any $P, Q \in B \cup V$, if,

$$\max_{1 \leq j \leq J} \left| P^m f_j - Q^m f_j \right| < \epsilon,$$

then either $P, Q \in B$, or $P, Q \in V$. (If the model is $\mathscr{T}_\infty$-compact, $m = 1$ suffices).

The Le Cam-Schwartz theorem

Theorem 27.1 (Le Cam-Schwartz, 1960) *Let $\mathscr{P}$ be a model for i.i.d. data X^n with disjoint subsets B, V. The folllowing are equivalent:*

i. there exist (uniformly) consistent tests for B vs V,

ii. there is a sequence of $\mathscr{U}_\infty$-uniformly continuous $\psi_n : \mathscr{P} \to [0,1]$,

$$\psi_n(P) \to 1_V(P), \qquad (6)$$

(uniformly) for all $P \in B \cup V$.

Example: how many clusters? (I)

Observe *i.i.d.* $X^n \sim P^n$, where P dominated with density p.

Clusters Family $\mathscr{F}$ of kernels $\varphi_\theta : \mathbb{R} \to [0, \infty)$, with parameter $\theta \in \Theta$. Assume Θ compact and,

$$\theta \mapsto \int f(x) \varphi_\theta(x)\, dx,$$

is continuous, for every bounded, measurable $f : \mathbb{R} \to \mathbb{R}$. Define $\Theta'_M = \Theta^M / \sim$.

Model Assume that there is an $M > 0$ such that p can be written as,

$$p_{\lambda,\theta}(x) = \sum_{m=1}^{M} \lambda_m p_{\theta_m}(x),$$

for some $M \geq 1$, with $\lambda \in S_M = \{\lambda \in [0,1]^M : \sum_m \lambda_m = 1\}$, $\theta \in \Theta'_M$.

Example: how many clusters? (II)

Assume M less than some known M'. Choose prior $\Pi_{\lambda,M}$ for $\lambda \in S_M$ such that, for some $\epsilon > 0$,

$$\Pi_{\lambda,M}\Big(\lambda \in S_M : \epsilon < \min\{\lambda_m\}, \max\{\lambda_m\} < 1-\epsilon\Big) = 1.$$

For $\theta \in \Theta'_M$ also choose a prior $\Pi_{\theta,M}$ that 'stays away from the edges'. Define,

$$\Pi = \sum_{M=1}^{M'} \mu_M \, \Pi_{\lambda,M} \times \Pi_{\theta,M}.$$

(for $\sum_M \mu_M = 1$).

If Π is a KL-prior, posterior odds select the correct number of clusters M. If there are no M' and ϵ known, there are sequences $M'_n \to \infty$ and $\epsilon_n \downarrow 0$ with priors Π_n that finds the correct number of clusters.