0.1 Neyman-Pearson and the testing affinity

The Neyman-Pearson Lemma solves a problem for testing a $\mathbb{P}_0$, with density $p_0(x)$, against a $\mathbb{P}_1$, with density $p_1(x)$. It finds a (randomized) test $\Psi = (\psi_0, \psi_1)$ for which $\int p_1\psi_1$ is maximized subject to $\int p_0\psi_1 \le \alpha$. Equivalently, it minimizes $\int p_1\psi_0$ subject to the same constraint.

There are other plausible quantities to optimize. For example, we could try to minimize

$$\int p_1(x)\psi_0(x) + p_0(x)\psi_1(x)$$

over all nonnegative ψ_0 and ψ_1 for which $\psi_0(x) + \psi_1(x) = 1$ for all x. This problem also has a simple solution because

$$p_1(x)\psi_0(x) + p_0(x)\psi_1(x) \ge p_0(x) \wedge p_1(x) := \min\left(p_0(x), p_1(x)\right)$$

with equality when $\psi_1(x) = \mathbf{1}\{x : p_0(x) < p_1(x)\}$. That is,

$$\min_{\Psi} \int p_1(x)\psi_0(x) + p_0(x)\psi_1(x) = \int p_0 \wedge p_1.$$

The quantity $\int p_0 \wedge p_1$ is called the ***testing affinity*** between $\mathbb{P}_0$ and $\mathbb{P}_1$. It is sometimes denoted by $\|\mathbb{P}_0 \wedge \mathbb{P}_1\|_1$.

0.2 Estimators defining tests

Suppose we have a model $\mathcal{P} = \{\mathbb{P}_\theta : \theta \in \Theta\}$ where each $\mathbb{P}_\theta$ is a probability corresponding to some density $p_\theta(x)$ on a set $\mathcal{X}$. We are interested in estimating some function $\tau(\theta)$, where τ maps Θ into some metric space $(\mathcal{T}, d)$.

For a minimax approach for each $\eta > 0$, we judge each estimator T by the value

$$\mathcal{M}(\Theta, \eta, T) := \sup_{\theta\in\Theta} \mathbb{P}_\theta\{d(T, \tau(\theta)) \ge \eta\}$$

We seek a lower bound,

$$\mathcal{M}(\Theta, \eta) := \inf_T \mathcal{M}(\Theta, \eta, T),$$

the infimum running over all estimators $T : \mathcal{X} \to \mathcal{T}$. (Also it would be satisfying to find some T that achieves the lower bound, but that is sometimes more than we can manage.)

Remark. The quantity $\mathcal{M}(\Theta, \eta)$ is the minimax lower bound for the loss function $L_\eta(\theta, t) = \mathbf{1}\{d(t, \tau(\theta)) \geq \eta\}$, for $(\theta, t) \in \Theta \times \mathcal{T}$. The story can also be told with other loss functions.

The search for a lower bound $\mathcal{M}(\Theta, \eta)$ can be turned into a multiple hypothesis testing problem by focusing on some finite subset Θ_0 of Θ. For each estimator T, define $\widehat{\theta}_T : \mathcal{X} \to \Theta_0$ by

<1> $$\widehat{\theta}_T(x) = \operatorname*{argmin}_{\theta \in \Theta_0} d(T(x), \tau(\theta)),$$

with any convenient rule for breaking ties. If we choose the finite subset Θ_0 so that $d(\tau(\theta), \tau(\theta')) \geq 2\eta$ for distinct θ and θ' is Θ_0 then

$$d(T(x), \tau(\theta)) + d(T(x), \tau(\theta')) \geq 2\eta \qquad \text{for all } \theta \neq \theta'.$$

In particular, if $d(T(x), \tau(\theta)) < \eta$ then $d(T(x), \tau(\theta')) > \eta$ for all other θ' in Θ_0, which implies $\widehat{\theta}_T(x) = \theta$. Put another way

$$\{x : d(T(x), \tau(\theta)) < \eta\} \subseteq \{x : \widehat{\theta}_T(x) = \theta\} \qquad \text{for each } \theta \in \Theta_0.$$

Equivalently,

$$\{x : d(T(x), \tau(\theta)) \geq \eta\} \supseteq \{x : \widehat{\theta}_T(x) \neq \theta\} \qquad \text{for each } \theta \in \Theta_0$$

so that

<2> $$\mathcal{M}(\Theta, \eta, T) = \sup_{\theta \in \Theta} \mathbb{P}_\theta\{d(T, \tau(\theta)) \geq \eta\} \geq \max_{\theta \in \Theta_0} \mathbb{P}_\theta\{\widehat{\theta}_T(x) \neq \theta\}.$$

If we find a lower bound for $\max_{\theta \in \Theta_0} \mathbb{P}_\theta\{\widehat{\theta}(x) \neq \theta\}$ that is valid for all maps $\widehat{\theta} : \mathcal{X} \to \Theta_0$ then it also provides a lower bound for every $\mathcal{M}(\Theta, \eta, T)$.

Remark. Effectively the simplification replaces the loss function $L_\eta(\theta, t) = \mathbf{1}\{d(t, \tau(\theta)) \geq \eta\}$, for $(\theta, t) \in \Theta \times \mathcal{T}$ by a loss function $\mathbf{1}\{\theta \neq t\}$ for $(\theta, t) \in \Theta_0 \times \Theta_0$.

0.3 Two point comparisons

The easiest case occurs when Θ_0 is a set of two points, θ_0 and θ_1, chosen so that $d(\tau(\theta_0), \tau(\theta_1)) \geq 2\eta$. The $\widehat{\theta}$ then corresponds to a nonrandomized test between θ_0 against θ_1.

<3> **Theorem.** *For every estimator T for $\tau(\theta)$,*

$$2\mathcal{M}(\Theta,\eta,T) \geq \sup\{\|\mathbb{P}_{\theta_0} \wedge \mathbb{P}_{\theta_1}\|_1 : \theta_i \in \Theta \text{ and } d(\tau(\theta_0),\tau(\theta_1)) \geq 2\eta\}.$$

PROOF Consider $\Theta_0 = \{\theta_0, \theta_1\}$ for a pair with $d(\tau(\theta_0),\tau(\theta_1)) \geq 2\eta$. Abbreviate $\mathbb{P}_{\theta_i}$ to $\mathbb{P}_i$ and p_{θ_i} to p_i. By inequality <2>,

$$\begin{aligned}
2\mathcal{M}(\Theta,\eta) &\geq 2\max\left(\mathbb{P}_0\{\widehat{\theta}_T \neq \theta_0\}, \mathbb{P}_1\{\widehat{\theta}_T \neq \theta_1\}\right) \\
&\geq \mathbb{P}_0\{\widehat{\theta}_T \neq \theta_0\} + \mathbb{P}_1\{\widehat{\theta}_T \neq \theta_1\} \\
&= \int p_0(x)\mathbf{1}\{\widehat{\theta}_T \neq \theta_0\} + p_1(x)\mathbf{1}\{\widehat{\theta}_T \neq \theta_1\} \\
&\geq \int p_0 \wedge p_1 \mathbf{1}\{\widehat{\theta}_T \neq \theta_0\} + p_0 \wedge p_1 \mathbf{1}\{\widehat{\theta}_T \neq \theta_1\} = \int p_0 \wedge p_1.
\end{aligned}$$

We have equality at the start of the last line if $p_0 \leq p_1$ whenever $\widehat{\theta}_T = \theta_1$ and $p_1 \leq p_0$ whenever $\widehat{\theta}_T = \theta_0$.

Complete the proof by taking a supremum over all such θ_0 and θ_1 pairs.

□

<4> **Example.** For $\theta > 0$ write $\mathbb{P}_\theta$ for the uniform distribution on $[0,\theta]^n$. Consider estimation of $\tau(\theta) = \theta$. For $x \in \mathbb{R}^n_+$ write $M_n(x)$ for $\max_{i\leq n} x_i$, the maximum likelihood estimator. For each $r > 0$,

$$\begin{aligned}
\mathbb{P}_\theta\{M_n(x) \leq \theta - r/n\} &= \mathbb{P}_\theta\{x_i \leq \theta - r/n \text{ for all } i \leq n\} \\
&= (1 - r/(n\theta))^n \\
&\to \exp(-r/\theta) \qquad \text{as } n \to \infty
\end{aligned}$$

More precisely, for each $\epsilon > 0$ and each $C > 0$ we can find an r, depending on both ϵ and C, for which

$$\sup_{0<\theta\leq C} \mathbb{P}_\theta\{|M_n - \theta| \geq r/n\} \leq \epsilon.$$

We have an estimator that achieves the n^{-1} rate, at least for $\Theta = (0, C]$.

To prove that n^{-1} is the best rate possible, suppose T_n is another function of $x_1, \ldots, x_n$ for which

$$\mathcal{M}(\Theta,\alpha,T_n) = \sup_{0<\theta\leq C} \mathbb{P}_\theta\{|T_n - \theta| \geq a\} \leq \epsilon.$$

How small could α be? Consider $\Theta_0 = \{1, 1+2\alpha\}$. Then

$$\begin{aligned} 2\epsilon &\geq \int p_1 \wedge p_{1+2\alpha} \\ &= \int (1+2\alpha)^{-n} \mathbf{1}\{0 \leq \min_i x_i \leq \max_i x_i \leq 1\}\, dx_1 \ldots dx_n \\ &= (1+2\alpha)^{-n}, \end{aligned}$$

which forces

$$2\alpha \geq \log(1+2\alpha) \geq n^{-1} \log(1/2\epsilon).$$

We can't do better than an n^{-1} rate. □

0.4 Total variation

The testing affinity is closely related to the ***total variation distance***,

$$d_{TV}(\mathbb{P}_0, \mathbb{P}_1) := \sup_A |\mathbb{P}_0 A - \mathbb{P}_1 A|$$

between $\mathbb{P}_0$ and $\mathbb{P}_1$.

For a real valued function f on $\mathcal{X}$ remember that $f^+(x) := \max(f(x), 0)$ and $f^-(x) := \max(-f(x), 0)$, which ensures that $f = f^+ - f^-$ and $|f| = f^+ + f^-$.

<5> **Lemma.** *For probabilities $\mathbb{P}_0$ and $\mathbb{P}_1$ with densities p_0 and p_1,*

$$d_{TV}(\mathbb{P}_0, \mathbb{P}_1) = 1 - \int p_0 \wedge p_1 = \int (p_0 - p_1)^+ = \int (p_0 - p_1)^- = \tfrac{1}{2} \int |p_0 - p_1|.$$

Proof For each $A \subseteq \mathcal{X}$,

$$\mathbb{P}_0 A - \mathbb{P}_1 A = \int \mathbf{1}\{x \in A\}(p_0(x) - p_1(x)).$$

The integral takes its maximum value, $\int (p_0 - p_1)^+$, when A picks out only the nonnegative values for $p_0(x) - p_1(x)$, that is, when $A = \{x : p_0(x) \geq p_1(x)\}$. It takes its minimum value (most negative), $-\int (p_0 - p_1)^-$, when A picks out values where $p_0(x) - p_1(x) < 0$, that is, $A = \{x : p_0(x) < p_1(x)\}$.

The integrals $\int (p_0 - p_1)^+$ and $\int (p_0 - p_1)^-$ are both equal to $\frac{1}{2} \int |p_0 - p_1|$ because

$$\begin{aligned} \int (p_0 - p_1)^+ - \int (p_0 - p_1)^- &= \int (p_0 - p_1) = 0 \\ \int (p_0 - p_1)^+ + \int (p_0 - p_1)^- &= \int |p_0 - p_1| \end{aligned}$$

Finally, note that

$$1 - \int p_0 \wedge p_1 = \int p_0 - p_0 \wedge p_1 = \int (p_0 - p_1)^+$$

because $a - a \wedge b = \max(a - b, 0)$ for all $a, b \in \mathbb{R}$.

□

Remark. The quantity $\int |p_0 - p_1|$ is often denoted by $\|\mathbb{P}_0 - \mathbb{P}_1\|_1$ and is called the ***$\mathcal{L}^1$-distance between $\mathbb{P}_0$ and $\mathbb{P}_1$***.

0.5 Distances between probabilities

The testing affinity and the total variation distance for two probability distributions are seldom easy to calculate directly. (The uniform distribuion from Example <4> is a rare exception.) Instead one usually works with other measures of affinity ordistance, such as the so-called ***f-divergences***.

<6> **Definition.** *Let $f : (0, \infty) \to \mathbb{R}$ be convex, with $f(1) = 0$. For probabilities P and Q (on the same set) with densities p and q define*

<7> $$D_f(P, Q) = D_f(p, q) := \int q f(p/q),$$

the f-divergence "distance" between P and Q.

I put "distance" in quotes because D_f is usually not a metric on the set of probabilities. (The $\mathcal{L}^1$ and Hellinger metrics are notable exceptions.) However, Jensen's inequality does show that $D_f(P, Q) \geq 0$ with inequality when $P = Q$.

The divergences come in pairs defined by an operation that preserves convexity. Remember that each convex f mapping $(0, \infty)$ into $\mathbb{R}$ can be written as a countable supremum of linear functions $f(t) = \sup_i (a_i + b_i t)$. The function f^* defined on $(0, \infty)$ by

$$f^*(t) = t f(1/t) = \sup_i (a_i t + b_i)$$

is also convex and $f^*(1) = f(1) = 0$. It also defines a divergence,

$$D_{f^*}(P, Q) = \int q f^*(p/q) = \int q(p/q) f(q/p) = D_f(Q, P).$$

The convexity of f ensures that the map $P \mapsto D_f(P,Q)$ is convex. For if P is a convex combination of P_1 and P_2, that is, $P = \alpha_1 P_1 + \alpha_2 P_2$, with density $p(x) = \alpha_1 p_1(x) + \alpha_2 p_2(x)$ then

$$\begin{aligned} D_f(P,Q) &= \int q f\left(\frac{\alpha_1 p_1 + \alpha_2 p_2}{q}\right) \\ &\le \int \alpha_1 q f(p_1/q) + \alpha_2 q f(p_2/q) \\ &= \alpha_1 D_f(P_1,Q) + \alpha_2 D_f(P_2,Q) \end{aligned}$$

Convexity of $Q \mapsto D_{f^*}(Q,P)$ then ensures that $D_f(P,Q)$ is separately convex in each argument.

Some examples

(i) for $f(t) = |t-1| = f^*(t)$,

$$D_f(P,Q) = \int |p-q| = \|P-Q\|_1 .$$

(ii) for $f(t) = (1-\sqrt{t}\,)^2 = f^*(t)$,

$$D_f(P,Q) = \int q(1-\sqrt{p/q}\,)^2 = \int (\sqrt{p}-\sqrt{q}\,)^2 .$$

The quantity $H(P,Q) = \left(\int \left(\sqrt{p}-\sqrt{q}\right)^2\right)^{1/2}$ is called the ***Hellinger distance*** between P and Q.

(iii) For $f(t) = t\log t$,

$$D_f(P,Q) = \int q(p/q)\log(p/q) = \int p\log(p/q),$$

which is called the ***Kullback-Leibler*** distance between P and Q. I denote it by $KL(P,Q)$. Note $f^*(t) = -\log t$.

(iv) for $f(t) = t^2 - 1$,

$$D_f(P,Q) = \int \frac{p^2}{q} - 1 = \int \frac{(p-q)^2}{q},$$

which is called the χ^2 ***distance***, sometimes denoted by $\chi^2(P,Q)$.

Remark. In all cases I have ignored possible 0/0 difficulties. A more precise treatment would pay more attention to contributions from the set where $q \wedge p = 0$. See Liese and Miescke (2008, page 35).

The KL and Hellinger distances are particularly convenient for dealing with independent observations. If $p(x) = \prod_{i \le n} g_i(x_i)$ and $q(x) = \prod_{i \le n} h_i(x_i)$ then $KL(p,q) = \sum_{i \le n} KL(g_i, h_i)$ and $H^2(p,q) \le \sum_{i \le n} H^2(g_i, h_i)$. See the homework for details.

0.6 Fano's inequality

Suppose Θ_0 is a finite subset of Θ with $\#\Theta_0 = N$. One version of Fano's inequality asserts that, for each $\widehat{\theta} : \mathcal{X} \to \Theta_0$,

<8> $$\max_{\theta \in \Theta_0} \mathbb{P}_\theta\{\widehat{\theta}(x) \ne \theta\} \ge \frac{\log N - \log 2 - N^{-1} \sum_{\theta \in \Theta_0} KL(\mathbb{P}_\theta, Q))}{\log(N-1)}$$

where $Q = N^{-1} \sum_{\theta \in \Theta_0} \mathbb{P}_\theta$. To simplify the average of KL-dstances it is customary to use convexity of $Q \mapsto KL(\mathbb{P}_\theta, Q)$ to show that

$$N^{-1} \sum\nolimits_{\theta \in \Theta_0} KL(\mathbb{P}_\theta, Q)) \le N^{-2} \sum\nolimits_{\theta,t} KL(\mathbb{P}_\theta, \mathbb{P}_t) \le \max_{\theta,t \in \Theta_0} KL(\mathbb{P}_\theta, \mathbb{P}_t).$$

With an increase of $\log(N-1)$ to $\log N$ one then has the simpler form of Fano's inequality,

<9> $$\max_{\theta \in \Theta_0} \mathbb{P}_\theta\{\widehat{\theta}(x) \ne \theta\} \ge 1 - \frac{\log 2 + \max_{\theta,t} KL(\mathbb{P}_\theta, \mathbb{P}_t)}{\log N}$$

To derive inequality <8> I use (a minor modification) of an elegant method due to Aditya Guntuboyina (2011).

Put a prior π on Θ_0. (For inequality <8> it will turn out to be the uniform prior, which puts mass N^{-1} at each point of Θ_0.) The prior defines a joint distribution $\mathbb{P}$ for x and θ under which $\theta \sim \pi$ and $x \mid \theta \sim P_\theta$. More formally, for each real g on $\mathcal{X} \times \Theta_0$,

$$\mathbb{E}_{\mathbb{P}} g(x,\theta) = \sum\nolimits_\theta \pi_\theta \int p_\theta(x) g(x,\theta).$$

Under $\mathbb{P}$ the x-coordinate has marginal distribution $Q = \sum_\theta \pi_\theta P_\theta$ with density $q(x) = \sum_\theta \pi_\theta p_\theta(x)$.

The Bayes estimator $\tau(x)$ is chosen to minimize the Bayes risk,

$$\mathbb{P}\{\tau(x) \neq \theta\} = 1 - \sum_\theta \pi_\theta \mathbb{P}_\theta\{x : \tau(x) = \theta\} = 1 - \int \sum_\theta \pi_\theta p_\theta(x) \mathbf{1}\{x : \tau(x) = \theta\}.$$

That is, $\tau(x) = \text{argmax}_\theta\, \pi_\theta p_\theta(x)$, so that the minimum Bayes risk is

<10> $$\overline{r} := \mathbb{P}\{\tau(x) \neq \theta\} = 1 - \int \max_\theta \left(\pi_\theta p_\theta(x)\right).$$

It turns out that to be cleaner to write expectations in terms of another probability distribution $\mathbb{Q}$ on $\mathcal{X} \times \Theta_0$ under which $x \sim Q$ and $\theta \sim \pi$ independently. More formally,

$$\mathbb{E}_\mathbb{Q} g(x, \theta) = \sum\nolimits_\theta \pi_\theta \int q(x) g(x, \theta).$$

Define $\widetilde{p}(x, \theta) := p(x, \theta)/q(x)$ and $A = \{(x, \theta) : \tau(x) = \theta\}$ then

$$1 - \overline{r} = \int \sum_\theta \pi_\theta q(x) \widetilde{p}(x, \theta) \mathbf{1}\{x : \tau(x) = \theta\} = \mathbb{E}_\mathbb{Q} \widetilde{p}(x, \theta) \mathbf{1}\{(x, \theta) \in A\}$$

and $\overline{r} = \mathbb{E}_\mathbb{Q} \widetilde{p}(x, \theta) \mathbf{1}\{(x, \theta) \in A^c\}$.

Define

$$\alpha := \mathbb{Q}A = \int q(x) \sum\nolimits_\theta \pi_\theta \mathbf{1}\{\tau(x) = \theta\} = \int q(x) \pi_{\tau(x)}.$$

Note well: For the special case where $\pi_\theta = 1/N$ for all N we have $\alpha = 1/N$.

Aditya's wonderful idea was to write $\mathbb{Q}$ as a weighted average of two conditional distributions, $\mathbb{Q} = \alpha \mathbb{Q}(\cdot \mid A) + (1 - \alpha)\mathbb{Q}(\cdot \mid A^c)$. Abbreviating the expected values with respect to the conditional distributions to $\mathbb{E}_A$ and $\mathbb{E}_{A^c}$, we then have

$$1 - \overline{r} = \alpha \mathbb{E}_A \widetilde{p}(x, \theta) \quad \text{AND} \quad \overline{r} = (1 - \alpha) \mathbb{E}_{A^c} \widetilde{p}(x, \theta).$$

The conditioning idea also works well with the average f-divergence between $\mathbb{P}_\theta$ and Q:

$$\begin{aligned}
\Delta &:= \sum\nolimits_\theta \pi_\theta D_f(\mathbb{P}_\theta, Q) \\
&= \sum\nolimits_\theta \pi_\theta \int q(x) f(p_\theta(x)/q(x)) \\
&= \mathbb{E}_\mathbb{Q} f(\widetilde{p}(x, \theta)) \\
&= \alpha \mathbb{E}_A f(\widetilde{p}(x, \theta)) + (1 - \alpha) \mathbb{E}_{A^c} f(\widetilde{p}(x, \theta)) \\
&\geq \alpha f(\mathbb{E}_A \widetilde{p}(x, \theta)) + (1 - \alpha) f(\mathbb{E}_{A^c} \widetilde{p}(x, \theta)) \qquad \text{by Jensen's inequality} \\
&= \alpha f\left(\frac{1 - \overline{r}}{\alpha}\right) + (1 - \alpha) f\left(\frac{\overline{r}}{1 - \alpha}\right).
\end{aligned}$$

For each fixed $\alpha \in (0,1)$, the function

$$\Psi_\alpha(t) = \alpha f\left(\frac{1-t}{\alpha}\right) + (1-\alpha) f\left(\frac{t}{1-\alpha}\right)$$

is convex in t. Aditya noted that the inequality

<11> $$\Delta \geq \Psi_\alpha(\overline{r})$$

could be inverted (or approximately inverted), for various choices of f, to deduce various lower bounds for $\overline{r}$.

The Fano inequality <8> comes from the choice $f(t) = t \log t$ and π the uniform distribution on Θ_0 (so that $\alpha = 1/N$). For that case

$$\begin{aligned}\Psi_\alpha(t) &= t\log(t) + (1-t)\log(1-t) - t\log(1-\alpha) - (1-t)\log(\alpha) \\ &\geq -\log 2 + \log N - t\log(N-1).\end{aligned}$$

In the last line I have used the fact that the function $t\log t + (1-t)\log(1-t)$ achieves its minimum value of $-\log 2$ at $t = 1/2$. In particular,

$$\Delta \geq -\log 2 + \log N - \overline{r}\log(N-1),$$

which rearranges to give <8>.

See homework 9 for an application of Fano's inequality to the calculation of a nonparametric minimax lower bound.

0.7 Notes

The tutorial by Csiszár and Shields (2004) contains a chapter on f-divergences.

References

Csiszár, I. and P. C. Shields (2004). Information theory and statistics: a tutorial. *Foundations and Trends in Communications and Information Theory 1*(4), 417–528.

Guntuboyina, A. (2011). Lower bound for the minimax risk using f-divergences, and applications. *IEEE Transactions on Information Theory 57*(4), 2386–2399.

Liese, F. and K.-J. Miescke (2008). *Statistical Decision Theory: Estimation, Testing, and Selection*. Springer-Verlag.