Minimax lower bounds for estimation

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 $\int p_0 \psi_1 \leq \alpha$. 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 $\psi_0$ and $\psi_1$ for which $\psi_0(x) + \psi_1(x) = 1$ for all $x$. This problem also has a simple solution because

$$p_1(x) \psi_0(x) + p_0(x) \psi_1(x) \geq p_0(x) \wedge p_1(x) := \min(p_0(x), p_1(x))$$

with equality when $\psi_1(x) = 1\{x : p_0(x) < p_1(x)\}$. 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 $\tau$ maps $\Theta$ into some metric space $(\mathcal{T}, d)$.

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

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

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

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

$$\hat{\theta}_T(x) = \underset{\theta \in \Theta_0}{\text{argmin}} d(T(x), \tau(\theta)),$$

with any convenient rule for breaking ties. If we choose the finite subset $\Theta_0$ so that $d(\tau(\theta_0), \tau(\theta_1)) \geq 2\eta$ for distinct $\theta_0$ and $\theta_1$ in $\Theta_0$, which implies $\hat{\theta}_T(x) = \theta$. Put another way

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

Equivalently,

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

so that

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

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

Remark. Effectively the simplification replaces the loss function $L_\eta(\theta, t) = 1\{d(t, \tau(\theta)) \geq \eta\}$, for $(\theta, t) \in \Theta \times T$ by a loss function $1\{\theta \neq t\}$ for $(\theta, t) \in \Theta \times \Theta$. The $\hat{\theta}$ then corresponds to a nonrandomized test between $\theta_0$ against $\theta_1$.

0.3 Two point comparisons

The easiest case occurs when $\Theta_0$ is a set of two points, $\theta_0$ and $\theta_1$, chosen so that $d(\tau(\theta_0), \tau(\theta_1)) \geq 2\eta$. The $\hat{\theta}$ then corresponds to a nonrandomized test between $\theta_0$ against $\theta_1$. 
**Theorem.** For every estimator $T$ for $\tau(\theta)$,

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

**Proof** Consider $\Theta_0 = \{\theta_0, \theta_1\}$ for a pair with $d(\tau(\theta_0), \tau(\theta_1)) \geq 2\eta$. Abbreviate $P_{\theta_i}$ to $P_i$ and $p_{\theta_i}$ to $p_i$. By inequality $<2>,$

$$2M(\Theta, \eta) \geq 2 \max (P_0\{\theta_T \neq \theta_0\} + P_1\{\theta_T \neq \theta_1\})$$

$$\geq \int p_0(x) \mathbf{1}\{\theta_T \neq \theta_0\} + p_1(x) \mathbf{1}\{\theta_T \neq \theta_1\}$$

$$\geq \int p_0 \wedge p_1 \mathbf{1}\{\theta_T \neq \theta_0\} + p_0 \wedge p_1 \mathbf{1}\{\theta_T \neq \theta_1\} = \int p_0 \wedge p_1.$$ 

We have equality at the start of the last line if $p_0 \leq p_1$ whenever $\theta_T = \theta_1$ and $p_1 \leq p_0$ whenever $\theta_T = \theta_0$.

Complete the proof by taking a supremum over all such $\theta_0$ and $\theta_1$ pairs.

□

**Example.** For $\theta > 0$ write $P_{\theta}$ 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 $\max_i x_i$, the maximum likelihood estimator. For each $r > 0$,

$$P_{\theta}\{M_n(x) \leq \theta - r/n\} = P_{\theta}\{x_i \leq \theta - r/n \text{ for all } i \leq n\}$$

$$= (1 - r/(n\theta))^n$$

$$\to \exp(-r/\theta) \quad \text{as } n \to \infty$$

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

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

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

$$M(\Theta, \alpha, T_n) = \sup_{0 < \theta \leq C} P_{\theta}\{|T_n - \theta| \geq a\} \leq \epsilon.$$
How small could \( \alpha \) be? Consider \( \Theta_0 = \{1, 1 + 2\alpha\} \). Then
\[
2\epsilon \geq \int p_1 \wedge p_{1+2\alpha} \\
= \int (1 + 2\alpha)^{-n} 1\{0 \leq \min_i x_i \leq \max_i x_i \leq 1\} \, dx_1 \ldots dx_n \\
= (1 + 2\alpha)^{-n},
\]
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.

\[ \square \]

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

**Proof.** For each \( A \subseteq \mathcal{X} \),
\[
\mathbb{P}_0 A - \mathbb{P}_1 A = \int 1\{x \in A\} (p_0(x) - p_1(x)).
\]
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
\[
\int (p_0 - p_1)^+ - \int (p_0 - p_1)^- = \int (p_0 - p_1) = 0 \\
\int (p_0 - p_1)^+ + \int (p_0 - p_1)^- = \int |p_0 - p_1|
\]
Finally, note that
\[
1 - \int p_0 \land p_1 = \int p_0 - p_0 \land p_1 = \int (p_0 - p_1)^+
\]
because \( a - a \land b = \max(a - b, 0) \) for all \( a, b \in \mathbb{R} \).

\[\square\]

**Remark.** The quantity \( \int |p_0 - p_1| \) is often denoted by \( \|P_0 - P_1\|_1 \) and is called the \( L^1 \)-distance between \( P_0 \) and \( P_1 \).

### 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 distribution from Example <4> is a rare exception.) Instead one usually works with other measures of affinity or distance, such as the so-called \( f \)-divergences.

**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

\[
D_f(P, Q) = D_f(p, q) := \int qf(p/q),
\]

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 \( L^1 \) 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 \( f(t) = \sup_i (a_i t + b_i) \). The function \( f^* \) defined on \((0, \infty)\) by

\[
f^*(t) = tf(1/t) = \sup_i (a_i t + b_i)
\]
is also convex and \( f^*(1) = f(1) = 0 \). It also defines a divergence,

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

$$D_f(P,Q) = \int qf\left(\frac{\alpha_1 p_1 + \alpha_2 p_2}{q}\right)$$

$$\leq \int \alpha_1 qf(p_1/q) + \alpha_2 qf(p_2/q)$$

$$= \alpha_1 D_f(P_1,Q) + \alpha_2 D_f(P_2,Q)$$

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 $H(P,Q) = \left(\int (\sqrt{p} - \sqrt{q})^2\right)^{1/2}$ 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 **$\chi^2$ distance**, sometimes denoted by $\chi^2(P,Q)$. 

Draft: 5 Nov 2014

Statistics 610 © David Pollard
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 \( q \land p = 0 \). See Liese and Miescke (2008, page 35).

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

### 0.6 Fano’s inequality

Suppose \( \Theta_0 \) is a finite subset of \( \Theta \) with \( \#\Theta_0 = N \). One version of Fano’s inequality asserts that, for each \( \hat{\theta} : X \rightarrow \Theta_0 \),

\[
\max_{\theta \in \Theta_0} \mathbb{P}_\theta \{ \hat{\theta}(x) \neq \theta \} \geq \frac{\log N - \log 2 - N^{-1} \sum_{\theta \in \Theta_0} KL(\mathbb{P}_\theta, Q))}{\log(N - 1)}
\]

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

\[
N^{-1} \sum_{\theta \in \Theta_0} KL(\mathbb{P}_\theta, Q)) \leq N^{-2} \sum_{\theta, \ell} KL(\mathbb{P}_\theta, \mathbb{P}_\ell) \leq \max_{\theta, \ell \in \Theta_0} KL(\mathbb{P}_\theta, \mathbb{P}_\ell).
\]

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

\[
\max_{\theta \in \Theta_0} \mathbb{P}_\theta \{ \hat{\theta}(x) \neq \theta \} \geq 1 - \frac{\log 2 + \max_{\theta, \ell} KL(\mathbb{P}_\theta, \mathbb{P}_\ell)}{\log N}
\]

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

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

\[
\mathbb{E}_{\mathbb{P}} g(x, \theta) = \sum_{\theta} \pi_\theta \int p_\theta(x) g(x, \theta).
\]

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,

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

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

\[
\tau := \mathbb{P}\{\tau(x) \neq \theta\} = 1 - \int \max_\theta (\pi_\theta p_\theta(x)).
\]

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

\[
\mathbb{E}_Q g(x, \theta) = \sum_\theta \pi_\theta \int q(x)g(x, \theta).
\]

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

\[
1 - \tau = \int \sum_\theta \pi_\theta q(x)\tilde{p}(x, \theta)1\{x : \tau(x) = \theta\} = \mathbb{E}_Q \tilde{p}(x, \theta)1\{(x, \theta) \in A\}
\]

and \( \tau = \mathbb{E}_Q \tilde{p}(x, \theta)1\{(x, \theta) \in A^c\} \).

Define

\[
\alpha := \mathbb{Q}A = \int q(x)\sum_\theta \pi_\theta 1\{\tau(x) = \theta\} = \int q(x)\pi_{\tau(x)}.
\]

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

\[
1 - \tau = \alpha \mathbb{E}_A \tilde{p}(x, \theta) \quad \text{AND} \quad \tau = (1 - \alpha) \mathbb{E}_{A^c} \tilde{p}(x, \theta).
\]

The conditioning idea also works well with the average \( f \)-divergence between \( P_\theta \) and \( Q \):

\[
\Delta := \sum_\theta \pi_\theta D_f(P_\theta, Q) = \sum_\theta \pi_\theta \int q(x)f(p_\theta(x)/q(x)) = \mathbb{E}_Q f(\tilde{p}(x, \theta)) = \alpha \mathbb{E}_A f(\tilde{p}(x, \theta)) + (1 - \alpha) \mathbb{E}_{A^c} f(\tilde{p}(x, \theta)) \geq \alpha f(\mathbb{E}_A \tilde{p}(x, \theta)) + (1 - \alpha) f(\mathbb{E}_{A^c} \tilde{p}(x, \theta)) \quad \text{by Jensen’s inequality}
\]

\[
= \alpha f\left(\frac{1 - \tau}{\alpha}\right) + (1 - \alpha) f\left(\frac{\tau}{1 - \alpha}\right).
\]
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
\[
\Delta \geq \Psi_\alpha(\tau)
\]
could be inverted (or approximately inverted), for various choices of $f$, to deduce various lower bounds for $\tau$.

The Fano inequality $<8>$ comes from the choice $f(t) = t \log t$ and $\pi$ the uniform distribution on $\Theta_0$ (so that $\alpha = 1/N$). For that case
\[
\Psi_\alpha(t) = t \log t + (1-t) \log (1-t) - t \log(1-\alpha) - (1-t) \log(\alpha)
\geq -\log 2 + \log N - t \log(1-t).
\]
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,
\[
\Delta \geq -\log 2 + \log N - \tau \log(N-1),
\]
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

