Unsorted notes

Gustavo Pereira

October 15, 2019

Here I store some random notes that I may or may note talk about during recitations.

1 Lectures 1 & 2

- Finite additivity

  Let's define some notation. I can define the following for any indexed collection of sets A_i:

  $$A_1 + A_2 := A_1 \cup A_2$$

  or, more generally

  $$\sum_i A_i := \bigcup_i A_i$$

  whenever the collection A_i is pairwise disjoint.

  The idea of assuming additivity – without any further qualification – is that set-function $\mathbf{P}$ satisfies some form of linearity, that is

  $$\mathbf{P}\left(\sum A_i\right) = \sum_i \mathbf{P}\left(A_i\right)$$

  It turns out that the set of indices over which this assumption is made is consequential.

  We call $\mathbf{P}$ *finitely additive* if the above is required to hold for all finite sets of indices. Similarly, if the relationship holds for countably many indices, $\mathbf{P}$ is called *countably additive.*

  Let's investigate an example of finitely, but not countably, additive measure. Here, we are working with a triple $(X, \mathcal{A}, \mathbf{P})$. $\mathcal{A}$ is an *algebra* of sets. Very similar to the usual $\sigma -$ algebra couterpart, but we don't require the assumptions of closedness under unions and intersections to hold for infinitely many set, only finitely many.

  We will work with the following algebra, which is not a σ-algebra. Let X be the set of all natural numbers, $\mathbf{N}$. Define also

$$\mathcal{A} = \{A \subset \mathbf{N} : A \text{ is finite or } A^c \text{ is finite}\}$$

Example of sets in $\mathcal{A}$: $\{1, 2, 3\}$ and $\{5001, 5002, \ldots\}$. Example of a sets *not* in $\mathcal{A}$: the set of all odd/even/prime numbers.[1]

It's not hard to see that this is satisfies: $\emptyset \in \mathcal{A}$ (since $\emptyset$ is finite) and closedness under intersections/unions. The reason why $\mathcal{A}$ is not a σ-algebra is that each $A_i = \{1, 3, \ldots, 2i + 1\}$ is in $\mathcal{A}$, but its infinite union, the set of all odd numbers, is not.

Now consider the probability measure: $\mathbf{P} : \mathcal{A} \to [0, 1]$:

$$\mathbf{P}(A) = \begin{cases} 1 & \text{if } A \text{ is infinite} \\ 0 & \text{ otherwise} \end{cases}$$

Thus, for example, $\mathbf{P}(1, 2, 3) = 0$ and $P(\{1023, 1024, \ldots\}) = 1$.

Such $\mathbf{P}$ trivially satisfies $\mathbf{P}(A + A') = \mathbf{P}(A) + \mathbf{P}(A')$ because the finite union of finite sets is finite.

This probability measure is interesting because it provides a counter-example to continuity when $\mathbf{P}$ is only finitely, but not countably, additive.

For example, it holds that $\{1, 2, \ldots, n\} \uparrow \mathbf{N}$, but

$$1 = \mathbf{P}(\mathbf{N}) = \mathbf{P}\left(\bigcup_n \{1, 2, \ldots, n\}\right) \neq \lim_n \mathbf{P}\left(\{1, 2, \ldots, n\}\right) = 0$$

Moreover, $\{n + 1, n + 2, \ldots\} \downarrow \emptyset$, but

$$0 = \mathbf{P}(\emptyset) = \mathbf{P}\left(\bigcap_n \{n + 1, n + 2, \ldots\}\right) \neq \lim_n \mathbf{P}\left(\{n + 1, n + 2, \ldots\}\right) = 1$$

The CDF of the random variable $X : \mathbf{N} \to \mathbf{N}$, $X(n) = n$ according to $\mathbf{P}$ will satisfy:

$$F_X(k) = \mathbf{P}\{n : X(n) \leq k\} = 0$$

for all n, so $\lim F_X(k) = 0$ for $k \to \infty$.

[1] The sets whose complement is finite are called co-finite sets.

2 Best linear predictor, matrix version

Let $M(n,k)$ denote the linear space of all matrix of dimension $n \times k$.

Suppose we have random vectors $(\mathbf{y}(\omega), \mathbf{z}(\omega))'$. We know additionally that $\mathbf{y} \in M(n,1)$ and $\mathbf{z} \in M(k,1)$ and these vectors have finite mean and variance. Denote their mean by

$$\begin{bmatrix} \mu_y \\ \mu_z \end{bmatrix}$$

and their variance matrix by

$$\begin{bmatrix} \Sigma_{yy} & \Sigma_{yz} \\ \Sigma_{zy} & \Sigma_{zz} \end{bmatrix}$$

We define the **best linear predictor** of $\mathbf{y}$ given $\mathbf{z}$ as the random variable $\mathbf{w}$ such that

$$\mathbf{w}^* = \alpha^* + \beta^*(\mathbf{z} - \mu_z)$$

where $\alpha^* \in M(n,1)$ and $\beta^* \in M(n,k)$ solve the minimzation problem

$$\min_{\alpha,\beta} \mathbf{E}\left[\|\mathbf{y} - \alpha - \beta(\mathbf{z} - \mu_z)\|^2\right]$$

You can solve it either by using calculus – which can be cumbersome if you're not used to matrix derivatives – or by noting that the minimand is a squared norm generated by the inner product

$$\langle \mathbf{y}, \mathbf{w} \rangle := \mathbf{E}[\mathbf{w}'\mathbf{y}]$$

of all vectors of the type $\mathbf{y} - \mathbf{w}$ where $\mathbf{w} = \alpha + \beta(\mathbf{z} - \mu_z)$ for some α, β.

Let $\epsilon := \mathbf{y} - \mathbf{w}^*$ denote the residual of the minimization problem. Then ϵ must be orthogonal (by Hilbert's projection theorem) to every $\mathbf{w} = \alpha + \beta(\mathbf{z} - \mu_z)$.

Taking $\beta = 0$, we see that $\mathbf{w}^*$ must satisfy

$$0 = \langle \mathbf{y} - \mathbf{w}^*, \alpha \rangle = \mathbf{E}\left[\alpha'\mathbf{y}\right] - \mathbf{E}\left[\alpha'\alpha^*\right]$$

for all vectors $\alpha \in M(n,1)$. Taking these to be the elements of the canonical basis, we conclude that

$$\alpha^* = \mu_y$$

Now take $\alpha = 0$. The orthogonality condition now implies that for any $\beta \in M(n,k)$,

$$0 = \langle \mathbf{y} - \beta^*(\mathbf{z} - \mu_z), \beta(\mathbf{z} - \mu_z) \rangle = \mathbf{E}\left[(\mathbf{z} - \mu_z)'\beta' y\right] - \mathbf{E}\left[(\mathbf{z} - \mu_z)'\beta'\beta^*(\mathbf{z} - \mu_z)\right]$$

Use the properties of the trace – namely, that it's linear and that matrix multiplication commutes inside it – and of the expectation operator to conclude that

$$\mathrm{tr}\left(\beta'\mathbf{E}\left[\mathbf{y}(\mathbf{z}-\mu_z)'\right]\right) = \mathrm{tr}\left(\beta'\beta^*\mathbf{E}\left[(\mathbf{z}-\mu_z)(\mathbf{z}-\mu_z)'\right]\right)$$

note that $\mathbf{E}[\mathbf{y}(\mathbf{z}-\mu_z)'] = \Sigma_{yz}$ and $\mathbf{E}[(\mathbf{z}-\mu_z)(\mathbf{z}-\mu_z)'] = \Sigma_{zz}$. The equation above then implies that

$$\mathrm{tr}(\beta'\Sigma_{yz}) = \mathrm{tr}(\beta'\beta^*\Sigma_{zz})$$

should hold for all matrices $\beta \in M(n,k)$. That implies,[2]

$$\Sigma_{yz} = \beta^*\Sigma_{zz}$$

which in turn yields $\beta^* = \Sigma_{yz}\Sigma_{zz}^{-1}$ whenever Σ_{zz} has an inverse. In that case, the BLP is

$$\mathbf{w}^* = \mu_y + \Sigma_{yz}\Sigma_{zz}^{-1}(\mathbf{z}-\mu_z) \tag{1}$$

2.1 Appendix: the Trace operator

- let $A(i,j)$ denote the entry (i,j) of any matrix
- Let A be a $m \times n$ matrix. The trace is defined as

$$\mathrm{tr}A = \sum_{i=1}^{\min\{m,n\}} A(i,i)$$

  in other words, it's just the sum of elements in the main diagonal.

- Some properties of the trace:

  1. $\mathrm{tr}(A+B) = \mathrm{tr}(A) + \mathrm{tr}(B)$ whenever A and B have similar dimensions
  2. $\mathrm{tr}(kA) = k\,\mathrm{tr}(A)$ for all scalars k
  3. $\mathrm{tr}(AB) = \mathrm{tr}(BA)$ whenever dimensions are such that both multiplications make sense

  Curiosity: any operation $\tilde{\mathrm{tr}}$ that satisfies the properties above is equal to tr (modulo multiplication by a constant)

[2] See the appendix on the trace operator for details.

- The trace and expectation operators commute:

$$\mathrm{tr}(\mathbf{E}A) = \mathbf{E}(\mathrm{tr}A)$$

- Suppose $A \in M(m,n)$ and you want to select element (i,j) from it. Note that

$$A(i,j) = e_i' A \varepsilon_j = tr(e_i' A \varepsilon_j) = tr(\varepsilon_j e_i' A)$$

  where e_i is the i-th element in the canonical basis of R^m and ε_j is the j-th element of the canonical basis of R^n.

  Hence for any (i,j), letting $B = \varepsilon_j e_i' \in M(n,m)$ we have

$$A(i,j) = \mathrm{tr}(BA)$$

- This implies that if A and $\tilde{A}$ are fixed $m \times n$ matrices, and

$$\mathrm{tr}(BA) = tr(B\tilde{A})$$

  holds for every $B \in M(n,m)$, then

$$A = \tilde{A}$$

3 Admissible tests and maximization of power subject to size

In lecture notes 9-10, Proposition 1 characterizes admissible tests in terms of the solution of an problem of maximizing power subject to a size constraint. I reproduce the statement of that proposition below.

Proposition 1. *Suppose that for any set* $A \subseteq \mathbf{X}$

$$\int_A f(x, \theta_0)dx > 0 \implies \int_A f(x, \theta_1)dx > 0.$$

A randomized test ϕ *is admissible if and only if there exists* $\alpha \in [0, 1]$ *such that* ϕ *maximizes power subject to having size at most* α*; that is*

$$\phi \in \arg\max_{\phi} \left(1 - R(\phi, \theta_1)\right) \tag{2}$$

s.t.

$$R(\phi, \theta_0) \leq \alpha \tag{3}$$

That proposition is actually really nice. In standard statistics courses, we sometimes take this maximization problem as the starting point, as if it's somehow self-evident that we should seek tests that *maximize power subject to size.* With the decision theoretic framework we built in the first few lectures, we can actually understand why tests that solve this maximization problem are of any interest to us. The reason is that this procedure yields tests that aren't dominated.

Another way of framing the proposition is the following. For a fixed $\alpha \in [0, 1]$, let $\Phi^*(\alpha)$ denote the set of all tests ϕ^* that maximize (2) subject to (3).

The correspondence $\Phi^*(\alpha)$ depends on a single parameter $\alpha \in [0, 1]$. What proposition 1 says is that, as we vary α, we cover all possible admissible tests. In other words,

$$\mathcal{A} = \bigcup_{\alpha \in [0,1]} \Phi^*(\alpha)$$

is *exactly* the set of all admissible tests.

3.1 Elaborating Proposition 1

I modify the proposition's exposition to make it a bit more digestible.

First, let's define the following.

Definition 1. Let $\{f_\theta(x)\}_{\theta\in\Theta}$ be a statistical model. We say that

$$f_{\theta_0} \ll f_{\theta_1}$$

(in plain English: f_{θ_0} is *dominated* by f_{θ_1}) if, for every measurable set A,

$$\mathbf{P}_{\theta_1}(A) = 0 \implies \mathbf{P}_{\theta_0}(A) = 0$$

Important remark. The relation $\ll$ has *nothing* to do with risk, loss, etc. It also has nothing to do with stochastic dominance.

Let's translate the definition above. What it means for f_{θ_0} to be dominated by f_{θ_1} is that, if the statistical model under θ_1 assigns zero probability to a set A – that is, there is a zero probability that we observe data in the set A under the alternative – then the probability that we observe data in the set A under the null must also be zero.

In other words, if that condition didn't hold, there would be a set of data realizations that are "impossible" under the alternative, but "possible" under the null.

Note that we can rewrite the definition in terms of integrals, since

$$\mathbf{P}_\theta(A) = \int_A f_\theta(x)dx$$

Hence, $f_{\theta_0} \ll f_{\theta_1}$ if and only if

$$\int_A f_{\theta_1}(x)dx = 0 \implies \int_A f_{\theta_0}(x)dx = 0$$

Or yet (by contraposition): $f_{\theta_0} \ll f_{\theta_1}$ iff

$$\int_A f_{\theta_0}(x)dx > 0 \implies \int_A f_{\theta_1}(x)dx > 0$$

All of these are restatements of the assumption that we can't observe under the null things that can't be observed under the alternative.

That assumption gives us an important result, that I state as a lemma.

Lemma 1. *Let $\{f_\theta\}_{\theta\in\Theta}$ be a statistical model with $\Theta = \{\theta_0, \theta_1\}$. Suppose $f_{\theta_0} \ll f_{\theta_1}$.*

Then any test ϕ achieving full power must have size equals one. Mathematically:

$$\mathbf{E}_{\theta_1}[\phi(X)] = 1 \implies \mathbf{E}_{\theta_0}[\phi(X)] = 1$$

Moreover, tests achieving zero size must have trivial power:

$$\mathbf{E}_{\theta_0}[\phi(X)] = 0 \implies \mathbf{E}_{\theta_1}[\phi(X)] = 0$$

Proof. Since $\phi(X) \leq 1$, full power – ie $\mathbf{E}_{\theta_1}\phi(X) = 1$ – implies that the set $A = \{x \in \mathcal{X} : \phi(x) < 1\}$ has zero probability under θ_1. Thus

$$\int_{\phi(x)<1} f_{\theta_1}(x)dx = 0$$

Since f_{θ_0} is dominated by f_{θ_1},

$$\mathbf{E}_{\theta_0}\phi(X) = \int_{\{\phi(x)=1\}} \phi(x) f_{\theta_0}(x)dx + \cancelto{0}{\int_{\{\phi(x)<1\}} \phi(x) f_{\theta_0}(x)dx} = 1$$

□

I'll now restate one directions of Proposition 1, for the particular case when $0 < \alpha < 1$.

Proposition 2. *Let $\{f_\theta\}_{\theta\in\Theta}$ be a statistical model with $\Theta = \{\theta_0, \theta_1\}$.*

Suppose $f_{\theta_0} \ll f_{\theta_1}$. Then any (randomized) test ϕ^* that solves the problem below is admissible in a decision problem with 0-1 loss, when $\alpha \in (0, 1)$.

$$\begin{aligned} \max_{\phi} \quad & E_{\theta_1}\phi(X) \\ s.t. \quad & E_{\theta_0}\phi(X) \leq \alpha \end{aligned} \tag{P}$$

Proof. Let's proceed by contradiction. Assume that ϕ^* solves the maximization problem but is not admissible. Then there exists some test ϕ that dominates ϕ^*, that is:

$$R(\phi, \theta_0) = \mathbf{E}_{\theta_0}[\phi(X)] \leq \mathbf{E}_{\theta_0}[\phi^*(X)] = R(\phi^*, \theta_0) \tag{4}$$

$$R(\phi, \theta_1) = 1 - \mathbf{E}_{\theta_1}[\phi(X)] \leq 1 - \mathbf{E}_{\theta_1}[\phi^*(X)] = R(\phi^*, \theta_1) \tag{5}$$

where one of the equalities holds strictly. We consider the two cases below.

1. Suppose 4 holds strictly, and 5 holds weakly. Since ϕ^* solves the maximization problem (P), the size constraint must be satisfied so

   $$\mathbf{E}_{\theta_0}[\phi(X)] < \mathbf{E}_{\theta_0}[\phi^*(X)] \leq \alpha < 1$$

   This first thing to note, which will only be used later on, is that since $\mathbf{E}_{\theta_0}[\phi(X)] < 1$, it must be that $\mathbf{E}_{\theta_1}[\phi(X)] < 1$ by the first part of Lemma 1.

   The idea of the proof is to construct yet another test that will use up the slack that ϕ has in the size constraint, $\mathbf{E}_{\theta_0}[\phi(X)] < \alpha$, to achieve higher power.

We can do that by mixing ϕ with the test that rejects the null for any realization,

$$\phi_R(X) \equiv 1$$

and by picking the right mix, we will increase power relative to ϕ, while still controlling for size. By 5, we will also improve relative to ϕ^*, a contradiction.

Now how do we find that combination? Consider, for arbitrary $\lambda \in [0, 1]$, the test

$$\phi_\lambda(X) \equiv \lambda \phi^R(X) + (1 - \lambda)\phi(X)$$

(Make sure you understand why we combine ϕ with ϕ^R, in particular why we don't combine ϕ^R with ϕ^*.) Its rate of type I error is given by

$$\mathbf{E}_{\theta_0}[\phi_\lambda(X)] = \lambda + (1 - \lambda)E_{\theta_0}[\phi(X)]$$

We pick $\bar{\lambda}$ that gives size exactly equal to α by setting

$$\bar{\lambda} = \frac{\alpha - \mathbf{E}_{\theta_0}[\phi(X)]}{1 - \mathbf{E}_{\theta_0}[\phi(X)]}$$

Since $0 \leq E_{\theta_0}[\phi(X)] < \alpha < 1$, we have $\bar{\lambda} \in (0, 1)$.

By construction, $\phi_{\bar{\lambda}}$ has a rate of type I error of exactly α. Its power on the other hand is given by

$$\mathbf{E}_{\theta_1}[\phi_{\bar{\lambda}}(X)] = \bar{\lambda} \cdot 1 + (1 - \bar{\lambda}) \cdot \mathbf{E}_{\theta_1}[\phi(X)]$$

Because $\mathbf{E}_{\theta_1}[\phi(X)] < 1$ and $\bar{\lambda} \in (0, 1)$, the above expression implies

$$\mathbf{E}_{\theta_1}[\phi_{\bar{\lambda}}(X)] > \mathbf{E}_{\theta_1}[\phi(X)] \geq \mathbf{E}_{\theta_1}[\phi^*(X)]$$

Where the last inequality comes from the assumption (5). That is a contradiction with the fact that ϕ^* is solves problem (P).

2. Suppose now that (5) holds strictly, while (4) holds weakly. Then (4) implies ϕ satisfies the size constraint, and

$$\mathbf{E}_{\theta_1}[\phi(X)] > \mathbf{E}_{\theta_1}[\phi^*(X)]$$

implies that ϕ achieves strictly higher power than ϕ^*, in direct contradiction with the fact that ϕ^* solves problem (P).

□