Ignorance priors and transformation groups

← Back to Chapters

Comments on 12.4.1

I find the discussion in 12.4 somewhat challenging. "Statistical decision theory and bayesian analysis" by James O. Berger was a better reference for me.

To put the main issue up first: The prior 12.27 is in fact "best" in this problem, but not because 12.18 (which has multiple issues) is better than 12.30, but because the deduction of 12.36 from 12.30 is flawed, and a better analysis of 12.30 (using the framework of invarinat decision rules) does indeed lead to 12.27!

Jaynes's paragraph (page 381) trying to find problem with 12.30 is itself flawed. There is in fact no prefered choice of the $x = 0$ point. However, this does not invalidate 12.30. The equations 12.30 take a particular form once the origin $x = 0$ is chosen, but even when the origin is not selected, the group of affine transformations encoded in 12.30 acts on the affine line $\mathbb{A}^1$. This action is abstract, and is only represented in coordinates via 12.30. This is in contrast to 12.18 which represents an action of the commutative group $\mathbb{R} \times \mathbb{R}_+$ on the 2D space of parameters, as well as the 3D space of "parameters together with coordinates", but not any action on the space of xs from which we draw the observed data. The idea of "rescaling from the current mean" encoded in 12.18 seems interesting, but hard to justify or generalize.

I flesh some of this out in the writeup below, which is largely based Berger's book, particularly sections 3.3.2 and 6.6.

The word "invariance" presupposes a group action. The most natural setting is that in which a group G acts on the space X in which we get data.

Example 1: (location-scale in 1D) Take $X = \mathbb{A}^1$ the affine line, and G the group of (orientation-preserving) affine transformations of X. As is common (for example in computer graphics), we represent elements of X as vectors $\vec{x} = \begin{pmatrix} x \\ 1 \end{pmatrix}$ (the line $y = 1$ inside $\mathbb{R}^2$) and then every element of G has unique representation as $g = \begin{pmatrix} a & b \\ 0 & 1 \end{pmatrix}$ with $a > 0$ (i.e. the set of linear transforms of $\mathbb{R}^2$ that take the line $y = 1$ to itself in orientation-preserving way) so that

$$g \cdot \vec{x} = \begin{pmatrix} a & b \\ 0 & 1 \end{pmatrix} \begin{pmatrix} x \\ 1 \end{pmatrix} = \begin{pmatrix} ax + b \\ 1 \end{pmatrix}$$

is indeed affine-linear.

The product in G is then:

$$\begin{pmatrix} a & b \\ 0 & 1 \end{pmatrix} \begin{pmatrix} a' & b' \\ 0 & 1 \end{pmatrix} = \begin{pmatrix} aa' & ab'+b \\ 0 & 1 \end{pmatrix}$$

This is a non-commutative group.

We have the following pair of group homomorphisms (known as a short exact sequence):

$$\mathbb{R} \hookrightarrow G \twoheadrightarrow \mathbb{R}_+$$

Here $\mathbb{R}$ is the real numbers under addition, and $\hookrightarrow$ sends a number to a translation by that amount; on the other hand $\mathbb{R}_+$ is positive reals under multiplication, and $\twoheadrightarrow$ sends an affine map to its stretching factor.

These maps do not depend on any particular way of representing X and G. On the other hand, there are "backward maps" $\mathbb{R}_+ \to G$ and $G \to \mathbb{R}$, but those are depend on additional choices, like the choice to represent X and G in terms of vectors and matrices as above. (In group theory one says that the sequence above is split, and thus G is "semi-direct product of $\mathbb{R}_+$ acting on $\mathbb{R}$, or"a split extension of $\mathbb{R}_+$ by $\mathbb{R}$"")

Example 2: (Location-scale in higher dimensions) When $X = \mathbb{A}^n$ the location-scale group G is the subgroup of those affine transforms of X whose linear part is a pure rescaling by a (positive) factor. We then have $\mathbb{R}^n \hookrightarrow G \twoheadrightarrow \mathbb{R}_+$.

We are interested in distribution of the data, i.e. in probability distributions over X. Thus we consider a collection of $\mathcal{P}$ of distributions over X.

Defintion: The collection $\mathcal{P}$ is said to be invariant under the action of G if for any $p \in P$ and any $g \in G$ the pushforward distribution g_*p is also in $\mathcal{P}$.

Since $(g \cdot h)_* p = g_*(h_* p)$, this means that G acts on $\mathcal{P}$ as well. If $\mathcal{P}$ is a parametric family, parametrized by space Θ then weconclude that G acts on Θ.

Example 1 continued:

a) Let $\mathcal{P}$ be the collection of all Gaussian distributions on $\mathbb{A}^1$. Note that in order to talk about the collection $\mathcal{P}$ specifying the origin is not necessary. This collection is a invariant under the action of G. If we pick an origin, then we can use mean μ and standard deviation σ as parameters for $\mathcal{P}$. We can also use representation of G by matrices that we have discussed above. Then $\Theta = \mathbb{R} \times \mathbb{R}_+$ is the parameter space and G acts on by sending (μ, σ) to $(a\mu + b, a\sigma)$.

(Together with x being sent to $ax + b$ this appears as 12.30 in Jaynes.)

b) Let $\mathcal{P}$ be the collection of all Cauchy distributions on $\mathbb{A}^1$; after choice of the origin this is the family with pdfs $\frac{1}{\pi\gamma\left[1+\left(\frac{(x-x_0)}{\gamma}\right)^2\right]}$. We no longer have mean or standard deviation available as a parameters, but we do have location x_0 and scale γ. They transform under G by the same formulas as (μ, σ) did before.

c) Let $\mathcal{P}$ be the collection of all mixures of normal distributions on $\mathbb{A}^1$. After picking the origin, this is the collection of distributions which can be written as $p = \sum_i^m w_i \mathcal{N}(\mu_i, \sigma_i^2)$ for some $m \in \mathbb{N}$ and $w_i > 0$ with $\sum w_i = 1$. This collection is invariant under G. When m is fixed the supbfamily $\mathcal{P}_m$ it is parametric, with parameters being $3m$ dimensional vectors (μ_i, σ_i, w_i). If m is not fixed, however, the collection $\mathcal{P}$ is not parametric in the usual sense of the word.

Example 3: Consider the family $\mathcal{P}$ of all Gamma distributions on $X = \mathbb{R}_+$, with pdfs $\Gamma_{(k,\theta)}(x) = \frac{1}{\Gamma(k)\theta}(\frac{x}{\theta})^{k-1}\exp(-\frac{x}{\theta})$. Here $G = \mathbb{R}+$ acts on X by multiplication, $\mathcal{P}$ is invariant, and the action of $a \in G$ sends (k, θ) to $(k, a\theta)$.

Now a **prior** is a probability distribution π over $\mathcal{P}$. This is easiest to understand when the collection $\mathcal{P}$ is parametric, so that we have $\Theta \subset \mathbb{R}^d$. When Θ is open, we may, as usual, describe the prior by its pdf, which we by abuse of notation will denote by $\pi(\theta)$. Since G acts on Θ, this action will transform the prior; namely, given $g \in G$, we obtain new distribution $g_*\pi$. When the action is differentiable, we have

$$[g_*\pi](\theta) = \pi(g^{-1}(\theta))|J_{g^{-1}}(\theta)|$$

(Recall: If $g(\psi) = \theta$ then $p(\psi)d\psi = p(g^{-1}(\theta))\frac{d\psi}{d\theta}d\theta = p(g^{-1}(\theta))|J_{g^{-1}}(\theta)|d\theta$.)

Example 1 continued: For the location-scale $g = \begin{pmatrix} a & b \\ 0 & 1 \end{pmatrix}$ sends $\theta = (m, s)$ to $g(\theta) = (am + b, as)$.

Then $J_g(\theta) = \begin{pmatrix} a & b \\ 0 & a \end{pmatrix}$, and so $|J_g| = a^2$, $|J_{g^{-1}}| = \frac{1}{a^2}$

$$[g_*\pi](\theta) = \frac{1}{a^2}\pi(g^{-1}(\theta)).$$

Now, one may argue that a transformation induced by g is simply a "change of coordinates" and the problems of forming a prior about θ and about $g(\theta)$ are equivalent, and so we should posit

$$[g_*\pi](\theta) = \pi(\theta)$$

A prior satisfying this is called **left invariant**. For such a prior we have in the differentiable case

$$\pi(\theta) = \pi(g^{-1}(\theta))|J_{g^{-1}}(\theta)|$$

In the cases where G acts transitively on Θ this specifies π uniquely, up to a constant since if we set $\pi(\theta_0) = C$ then for any θ there exists g such that $\theta = g(\theta_0)$ and then

$$\pi(\theta) = \pi(g^{-1}(\theta))|J_{g^{-1}}(\theta)| = C|J_{g^{-1}}(\theta)|$$

Example 1 continued: If π is left-invariant then taking $\theta_0 = (m_0 = 0, s_0 = 1)$ and given $\theta = (m, s)$ we have $g = \begin{pmatrix} m & s \\ 0 & 1 \end{pmatrix}$ and

$$\pi(m, s) = \frac{C}{s^2}.$$

(This agrees with 12.36 in Jaynes.)

However, as pointed out by Berger (p.86) there is a logical flaw in requiring $[g_*\pi](\theta) = \pi(\theta)$. In fact, often the prior π in question is improper, and so is only determined up to a constant. Thus we can only require a weaker equality

$$[g_*\pi](\theta) = K(g)\pi(\theta)$$

for some g-dependent scaling function $K(g)$.

We will call priors satisfying this "weakly" invariant, and the ones with $K(g) = 1$ "strictly" invariant.

In the differentiable case this leads to

$$K(g)\pi(\theta) = \pi(g^{-1}(\theta))|J_{g^{-1}}(\theta)|,$$

which can have many solutions.

Example 1 continued: Take $\pi(\theta) = Cs^\alpha$, $K(\begin{pmatrix} a & b \\ 0 & 1 \end{pmatrix}) = a^{-(\alpha+2)}$. All of these solve the above equation.

The question of how to choose "invariant prior" (even when the transformation group is known) is hence not settled. One method is to use the framework of "invariant decision problems", which suggests, that when the action of G on Θ is "simply transitive", one should use the **right invariant** prior. The story goes s follows:

Suppose that the group G acts on Θ ins such a way that picking some starting θ_0 we have for any $\theta \in \Theta$ a unique $g \in G$ with $g_\theta \theta_0 = \theta$. Thus after making this choice of θ_0 we can identify this θ with that unique $g = f(\theta)$. (Note that this f is a "map of G-sets: $f(g\theta) = g \cdot f(\theta)$, where $\cdot$ is the multiplication in G.)

Now a distributions/measure over Θ is a measure over G. "Strictly" invariant measures on Θ correspond to what's known as "left-invariant" measures on G (they are invariant under all left multiplication maps $L_g : G \to G$ sending $g' \in G$ to $L_g(g') = gg'$). It is a theorem that such a measure is uniques up to scaling. There are, however, more "weakly" invariant measures (as we have seen), those that are "invarinat up to scaling function $K(g)$". Among those, there is unique up to scaling measure which is what is called **right-invariant**: it is invariant under all maps $R_g : G \to G$ sending g' to $R_g(g') = g'g$ (the uniqueness up to scale and the fact that right invariant measures are "weakly" left invariant are basic facts of the theory of Haar measures on groups).

We illustrate the result on the location-scale example.

(We postpone more the discussion of **why** the right-invarinat prior is the best; see section 6.6 in Berger).

Example 1 continued: Using $\theta_0 = (m_0 = 0, s_0 = 1)$ the map $f : \Theta \to G$ sends (m, s) to $\begin{pmatrix} s & m \\ 0 & 1 \end{pmatrix}$.

The map $R_g = R_{(a,b)}$ sends $\begin{pmatrix} s & m \\ 0 & 1 \end{pmatrix}$ to

$\begin{pmatrix} s & m \\ 0 & 1 \end{pmatrix} \begin{pmatrix} a & b \\ 0 & 1 \end{pmatrix} = \begin{pmatrix} as & bs + m \\ 0 & 1 \end{pmatrix}$, and thus has Jacobian matrix in (s, m) coordinates equal to $\begin{pmatrix} a & 0 \\ b & 1 \end{pmatrix}$, and determinant $|J_{R_g}| = a$, and $|J_{R_{g^{-1}}}| = \frac{1}{a}$.

The invariance condition $R_g * \pi(g') = \pi(g')$ is then

$$\pi(g') = \pi(g^{-1}g')\frac{1}{a}$$

and, setting $\pi(e) = C$ and $g' = g$ we get

$\pi(g) = \frac{C}{a}$.

This is in fact the "invariant prior" 12.27.

So, to repeat what we said in the beginning, 12.27 is in fact "best" in this problem, but not because 12.18 (which has multiple issues) is better than 12.30, but because the deduction of 12.36 from 12.30 is flawed, and a better analysis of 12.30 (using the framework of invariant decision rules) does indeed lead to 12.27.

So, why right invariant priors as opposed to left invariant ones? I don't feel that I understand the answer completely. One thing I can say is that, roughly speaking, this may be considered to arize as follows: Suppose that you have a function of parameter and data $f(x, \theta)$ invariant under action $(x, \theta) \to f(gx, g\theta)$ (say, a loss function of some decision procedure). Then integrating "over X" $\int f(x, \theta_0)dx$ can be rewritten as integrating "over G" via $\int f(gx_0, \theta)dg$, then as $\int f(x_0, g^{-1}\theta)dg$ and then, finally, as integration "over Θ"" i.e. $\int f(x_0, g^{-1}\theta)d\theta$. In the X to G transition "preserves left-invariance" but G to X transition has a g^{-1} and "moves left-invariance to right-invariance".

Comments on 12.4.3

The derivation in 12.4.3 is suspect. Apart from dubious justification for invariance through "total confusion", should we not ask why is it that $a = \frac{p(E|Sx)}{p(E|FX)}$ is the same for every member of this imaginary population of individuals?

Would it not be better to say that the group that is acting is the translation group acting on log likeliehoods (aka "evidence")? I.e. that in terms of the odds ratio parameter $l = log\frac{\theta}{1-\theta}$ the (left and right) invariant prior is uniform? That, is the invariant prior is Cdl, which is $Cdl = C\frac{dl}{d\theta}d\theta = C\frac{1-\theta}{\theta}\frac{1}{(1-\theta)^2} = \frac{C}{\theta(1-\theta)}$.

(See Example 8 in Section 3.4.3 in Berger and the Kevin Van Horn's page for further options and discussion.)

12.50 from 12.48

Plug in $\theta = 1/2$ to get $af(\frac{a}{1+a}) = \frac{(1+a)^2}{4}f(1/2)$

If $\theta = \frac{a}{1+a}$ then $(1 + a)(1 - \theta) = 1$ so $a = \frac{\theta}{1-\theta}$ and $(1 + a) = \frac{1}{1-\theta}$; plugging this in we have

$$\frac{\theta}{1-\theta}f(\theta) = \frac{C}{(1-\theta)^2}$$

$$f(\theta) = \frac{C}{\theta(1-\theta)}.$$

Comments on 12.4.4

It certainly **does** violence to Bertrand's paradox to rephrase it in terms of throwing strawes. The point of this paradox from the probability theory point of view os that saying "at random" is meaningless – one has to provide a probability distribution; replacing the words "at random" by the "throwing straws" procedure makes this point moot – the issue then becomes not the fact that the probability is unspecified, but, rather, that it is specified via a "physical" procedure, and one has to deduce the probability distribution from this procedure. This is what Jaynes proceed to do. But this is an entirely different matter!