invariance_priors

September 12, 2022

1 Invariance Priors

This document provides a summary of the idea of obtaining a prior probability density function via the specification of invariance properties. The approach was championed by Jaynes, which he called *transformation group invariance* [1].

Jaynes' basic proposition was that if there is a continous transformation between the viewpoints of two observers, and further that if knowledge of the transformation does not make the prior knowledge of the observers differ, then they must both, for consistency, assign *exactly* the same prior distribution.

However, Jaynes himself offered an example that was an exception to this rule, namely the case of a contraction mapping from a circle to a smaller, concentric circle. In this case, Jaynes rightly states that the distribution conditional on the inner circle is not equal to, but instead *proportional* to, the distribution conditional on the outer circle, with the constant of proportionality being the ratio of the factors needed to normalise the respective proper distributions.

In fact, the need for proportionality was also demonstrated by Milne [2] for the improper distributions resulting from nonlinear scaling (as discussed in this document in detail in the applied section on nonlinear scale invariance).

Consequently, in this document we derive the general invariance relation, including proportionality, from first principles. We should note, however, that all of the examples for which Jaynes did *not* include proportionality actually result in a proportionality constant of unity, and hence remain correct.

1.1 Conservation of probability

We start with a continuous space $\mathcal{X} \subseteq \mathbb{R}^n$, with some as-yet undefined probability distribution $f : \mathcal{X} \mapsto \mathbb{R}^+$ that satisfies $\int_{\mathcal{X}} f(\mathbf{x}) \, |d\mathbf{x}| = 1$, where $|d\mathbf{x}| \doteq dx_1 dx_2 \dots dx_n$ is taken to be some infinitesimal volume element surrounding each point $\mathbf{x} \in \mathcal{X}$.

Next, we consider a continuous mapping $\mathbf{h} : \mathcal{X} \mapsto \mathcal{Y}$ into some transformed space $\mathcal{Y} \subseteq \mathbb{R}^m$. This mapping induces another probability distribution $g : \mathcal{Y} \mapsto \mathbb{R}^+$ that obeys $\int_{\mathcal{Y}} g(\mathbf{y}) \, |d\mathbf{y}| = 1$, where $|d\mathbf{y}| \doteq |dy_1 dy_2 \dots dy_m|$ is an infinitesimal volume element surrounding the point $\mathbf{y} = \mathbf{h}(\mathbf{x}) \in \mathcal{Y}$.

The first invariance is that of conservation of probability mass. Since the transformation $\mathbf{h}$ is continuous, we suppose that the volume element $|d\mathbf{x}|$ is continuously transformed into the volume element $|d\mathbf{y}|$. Hence, the probability mass of the untransformed element must equal that of the transformed element, giving rise to the invariance

$$f(\mathbf{x}) \, |d\mathbf{x}| \quad = \quad g(\mathbf{y}) \, |d\mathbf{y}| \, . \tag{1}$$

For common dimensionality $m = n$, we have $|d\mathbf{y}| = |J|\,|d\mathbf{x}|$, with Jacobian $J(\mathbf{x}) \doteq \mathtt{det}\left(\frac{\partial \mathbf{h}}{\partial \mathbf{x}}\right)$. The invariance therefore becomes

$$f(\mathbf{x}) \quad = \quad |J(\mathbf{x})|\; g(\mathbf{y})\,. \tag{2}$$

1.2 Distributional invariance

The next invariance is the key to the entire method. Consider two observers, X and Y, who both observe some event, but who take measurements $\mathbf{x} \in \mathcal{X}$ and $\mathbf{y} \in \mathcal{Y}$ of the event, respectively. We suppose that before the event both observers had exactly the same background knowledge. The invariance is to suppose that after the event, both observers still have the same knowledge, since they observed the same event. Clearly, we are excluding imperfect measurements where observation was partially obscured for either observer.

As noted in the intoduction, Jaynes [1] opined that if the knowledge of the two observers remains the same, then the observations are invariant to nature of the transformation, and furthermore observers X and Y must (for the sake of consistency) assign the same prior distribution to the event. Hence, Jaynes supposes that $g = f$.

Milne [2] disagreed, and demonstrated that although this invariance works for the linear scaling $y = \alpha x$, it fails to work for the nonlinear scaling $y = \alpha x^{\beta}$. Hence, Milne proposed that the invariance instead holds up to a constant factor, and that, in general, $g = \gamma f$ (although Milne explicitly stated it as $f = \gamma g$, his further working out in fact only makes sense if $g = \gamma f$ instead).

In fact, as again noted in the introduction, Jaynes himself saw the occasional need for $g = \gamma f$, as explicitly used in his solution [3] to the Bertrand paradox (or supposed 'solution', since Drory [4] counters Jaynes' claim to have a unique solution).

To consolidate these two viewpoints, we may suppose that invariance means here that the transformation does not alter the shape of the probability distribution, but only moves points along it, and possibly changes its constant of normalisation. Hence, we take

$$g(\mathbf{y}) \quad = \quad \gamma f(\mathbf{y})\,, \tag{3}$$

which now presupposes that $\mathcal{Y} \subseteq \mathcal{X}$.

It now follows that if f and g are *proper* distributions, such that they can be normalised, then γ is determined uniquely by

$$\gamma \quad = \quad \frac{\int_{\mathcal{Y}} g(\mathbf{y})\,|d\mathbf{y}|}{\int_{\mathcal{Y}} f(\mathbf{y})\,|d\mathbf{y}|} \quad = \quad \frac{1}{\int_{\mathcal{Y}} f(\mathbf{x})\,|d\mathbf{x}|}\,. \tag{4}$$

Clearly, we must have $\gamma = 1$ for $\mathcal{Y} = \mathcal{X}$, and $\gamma > 1$ for $\mathcal{Y} \subset \mathcal{X}$.

However, if the distributions are *improper*, then $\gamma > 0$ must be determined from the above invariance relation, once the forms of f and g have been found.

1.3 Parameter invariance

The third invariance follows from the second. If the distribution is invariant to the transformation (up to a normalising constant), then it is invariant to the specific values of the transformation parameters (or at least a subset of the parameters, as we shall see in a later section). Let the

mapping $\mathbf{h}$ now explicitly be a function of parameters θ, such that $\mathbf{y} = \mathbf{h}(\mathbf{x};\theta)$, and likewise $J(\mathbf{x};\theta) = \texttt{det}\left(\frac{\partial \mathbf{h}}{\partial \mathbf{x}}\right)$. Then the combination of conservation of probability and distributional invariance leads to the unified invariance relation

$$f(\mathbf{x}) \;=\; \gamma\,|J(\mathbf{x};\theta)|\,f(\mathbf{h}(\mathbf{x};\theta))\,. \tag{5}$$

This relation is now invariant to changes in the parameter values, and so we may take derivatives with respect to the parameters. This gives rise to the parameter invariance relation

$$\frac{\partial |J|}{\partial \theta}(\mathbf{x};\theta)\,f(\mathbf{h}(\mathbf{x};\theta)) + |J(\mathbf{x};\theta)|\;f'(\mathbf{h}(\mathbf{x};\theta))\,\frac{\partial \mathbf{h}}{\partial \theta}(\mathbf{x};\theta) \;=\; \mathbf{0}\,. \tag{6}$$

Furthermore, if these derivatives are invariant to the values of the parameters, then the relations hold for all parameter values, including the special value θ_0 which leads to the identity mapping $\mathbf{h}(\mathbf{x};\theta_0) = \mathbf{x}$. Consequently, we may simplify the derivative invariances at $\theta = \theta_0$, and then solve the simplified equations for $f(\mathbf{x})$.

2 Examples of Application

2.1 Nonlinear scale invariance

We now examine the nonlinear scale invariance discussed by Milne [2], namely $y = h(x;\alpha,\beta) = \alpha x^\beta$ over the domain(s) $\mathcal{Y} = \mathcal{X} = (0,\infty)$ for $\alpha > 0$ and $\beta \neq 0$. We observe that

$$\frac{\partial h}{\partial \alpha} \;=\; x^\beta\,, \qquad \frac{\partial h}{\partial \beta} \;=\; \alpha x^\beta \ln x\,, \tag{7}$$

such that

$$|J| \;=\; \left|\frac{\partial h}{\partial x}\right| \;=\; \alpha|\beta|x^{\beta-1} \;=\; \frac{|\beta|h}{x}\,, \tag{8}$$

$$\Rightarrow \frac{\partial |J|}{\partial \alpha} \;=\; |\beta|x^{\beta-1}\,, \quad \frac{\partial |J|}{\partial \beta} \;=\; \alpha x^{\beta-1}(\texttt{sign}(\beta) + |\beta|\ln x)\,. \tag{9}$$

The derivatives of the unified invariance relation are therefore

$$|\beta|x^{\beta-1}\,f(\alpha x^\beta) + \alpha|\beta|x^{\beta-1}\,f'(\alpha x^\beta)\,x^\beta \;=\; 0\,, \tag{10}$$

$$\alpha x^{\beta-1}(\texttt{sign}(\beta) + |\beta|\ln x)\,f(\alpha x^\beta) + \alpha|\beta|x^{\beta-1}\,f'(\alpha x^\beta)\,\alpha x^\beta \ln x \;=\; 0\,. \tag{11}$$

We note that the nonlinear scaling becomes the identity mapping for $\alpha = 1$ and $\beta = 1$. Hence, at these values, the derivatives reduce to

$$f(x) + x f'(x) \;=\; 0\,, \tag{12}$$

$$(1+\ln x)\,f(x) + x\ln x\,f'(x) \;=\; 0\,. \tag{13}$$

The α-derivative is clearly satisfied for $f(x) = \frac{k}{x}$ with arbitrary constant $k > 0$. However, the β-derivative is then only satisfied for $k = 0$. Consequently, we see that the nonlinear scaling is

invariant to $\alpha > 0$, but is **not** invariant to β, which must therefore be held constant to a value fixed in advance. In fact, from the unified invariance relation, we have

$$\begin{aligned} f(x) &= \gamma\alpha|\beta|x^{\beta-1} f(\alpha x^{\beta}) \,, && (14) \\ \Rightarrow \frac{k}{x} &= \gamma\alpha|\beta|x^{\beta-1}\frac{k}{\alpha x^{\beta}} \,, && (15) \\ \Rightarrow \gamma &= \frac{1}{|\beta|} \,. && (16) \end{aligned}$$

This dependence of the renormalisation factor γ on β reinforces the fact that for each fixed value of β we obtain a family of transformations that are distributionally invariant to the scaling factor α.

We see that taking $\beta = 1$ results in the linear scaling of Jaynes [1], for which $\gamma = 1$. Also note that $\gamma \neq 1$ for $\beta \neq \pm 1$, despite the fact that $\mathcal{Y} = \mathcal{X}$. This is due, as discussed in the section on distributional invariance, to the fact that f is an *improper* distribution that cannot actually be properly normalised.

2.2 Von Kries paradox

We now turn to the resolution of the Von Kries paradox, as also discussed (briefly) by Milne [2].

Suppose we know the specific density ρ (my notation) of a fluid is restricted to the range $\rho \in [a, b]$ for some $0 < a < b < \infty$. If we know nothing else, then the principle of insufficient reason suggests all values are equally likely, and so we might take the uniform prior $f(\rho) = \frac{1}{b-a}$.

However, the specific volume ν (again, my notation) is inversely related to the specific density via $\nu = \frac{1}{\rho}$, giving rise to the Jacobian $J(\rho) = \frac{d\nu}{d\rho} = -\frac{1}{\rho^2}$. The corresponding prior $g(\nu)$ therefore satisfies the conservation of probability relation

$$f(\rho) \;=\; |J(\rho)|\, g(\nu) \quad\Rightarrow\quad g(\nu) \;=\; \frac{1}{\nu^2} f\left(\frac{1}{\nu}\right) \;=\; \frac{1}{(b-a)\nu^2} \,. \tag{17}$$

The paradox is therefore that the specific volume is deterministically known from the specific density, but uniform ignorance of the specific density does not translate to uniform ignorance of the specific volume. In other words, we seem to have acquired some knowledge of the specific volume for free.

The resolution of the paradox, of course, is that we must already have added this knowledge by the *unwarranted* assumption of a uniform prior for the specific density.

Instead we take ignorance to mean distributional invariance. From the previous section, we see that $\nu = \frac{1}{\rho}$ is a specific form of the nonlinear transformation $\nu = \alpha\rho^{\beta}$ for $\alpha = 1$ and $\beta = -1$. Consequently, we already know that $\gamma = \frac{1}{|\beta|} = 1$, such that the two prior distributions are given by

$$f(\rho) \;=\; \frac{k}{\rho} \,, \qquad g(\nu) \;=\; \frac{1}{\nu^2} f\left(\frac{1}{\nu}\right) \;=\; \frac{k}{\nu} \,, \tag{18}$$

respectively. Since we are assuming the finite domain $\rho \in [a, b]$, we easily obtain the normalisation constant as $k = \frac{1}{\ln b - \ln a}$.

We observe that now ignorance of the specific density gives rise to the same form as ignorance of the specific volume, which resolves the paradox.

2.3 Linear regression

Let us consider (briefly) the problem of fitting the straight-line model $y = \alpha + \beta x$, for $x, y \in \mathbb{R}$. Typically, in practice, we observe pairs of values (x, y), and from these data D we wish to infer the posterior distribution $p(\alpha, \beta \mid D)$. However, here we are not concerned with the data but with the prior distributions, $p(\alpha)$ and $p(\beta)$.

Firstly, we observe that $\alpha \in \mathbb{R}$ uniquely specifies the intersection of the model line with the y-axis, and hence α is a location parameter. Consequently, if we know nothing else about α, then we might suppose that the prior distribution, $p(\alpha)$, is invariant to the translation $\alpha' = \alpha + \nu$. From the section on parameter invariance, we therefore obtain the invariance relation

$$p(\alpha) \;=\; \gamma\, p(\alpha + \nu)\,. \tag{19}$$

As usual, we take the derivative with respect to the transformation parameter ν, and then evaluate the result at the point $\nu = 0$, for which the transformation becomes an identity. This gives

$$0 \;=\; \gamma\, p'(\alpha) \;\;\Rightarrow\;\; p(\alpha) \;=\; k\,,\;\; \gamma \;=\; 1\,. \tag{20}$$

Thus, α has an improper, uniform invariance prior.

Next, we observe that the slope $\beta \in \mathbb{R}$ is the tangent of the angle $\theta \in (-\frac{\pi}{2}, \frac{\pi}{2})$ that the model line is rotated counter-clockwise from the x-axis.

Some trigonometry gives us, for example:

$$\beta \;=\; \tan\theta \;=\; \frac{\sin\theta}{\cos\theta} \;=\; \frac{\mathtt{sign}(\theta)\,\sqrt{1-\cos^2\theta}}{\cos\theta} \tag{21}$$

$$\Rightarrow \cos\theta \;=\; \frac{1}{\sqrt{1+\beta^2}}\,. \tag{22}$$

Next, we see that

$$\frac{d\tan\theta}{d\theta} \;=\; \frac{d}{d\theta}\left(\frac{\sin\theta}{\cos\theta}\right) \;=\; \frac{\cos\theta}{\cos\theta} + \frac{\sin^2\theta}{\cos^2\theta} \;=\; 1 + \tan^2\theta \tag{23}$$

$$\;=\; \frac{1}{\cos^2\theta} \;=\; \sec^2\theta\,, \tag{24}$$

$$\Rightarrow \frac{d\beta}{d\theta} \;=\; 1 + \beta^2\,. \tag{25}$$

If we are ignorant about any further properties of β, then we may suppose that the prior distribution, $p(\beta)$, is invariant to the rotational transformation $\beta' = \tan(\theta + \psi)$. The unified invariance relation is then

$$p(\beta)\left|\frac{d\beta}{d\theta}\right| \;=\; \gamma\, p(\beta')\left|\frac{\partial\beta'}{\partial\theta}\right|\,, \tag{26}$$

where

$$\frac{\partial\beta'}{\partial\theta} \;=\; \frac{\partial\beta'}{\partial\psi} \;=\; 1 + \beta'^2\,. \tag{27}$$

Hence, the completed relation is

$$(1+\beta^2)\,p(\beta) \;=\; \gamma\,(1+\beta'^2)\,p(\beta')\,. \tag{28}$$

Once again, we take the derivative with respect to the transformation parameter ψ. This gives

$$0 \;=\; \gamma\,\left[2\beta'(1+\beta'^2)\,p(\beta') + (1+\beta'^2)^2\,p'(\beta')\right]\,. \tag{29}$$

Next, we evaluate the derivative at the point $\psi = 0$ at which the transformation becomes the identity. With simplification, this gives

$$\begin{aligned}
2\beta\,p(\beta) + (1+\beta^2)\,p'(\beta) &= 0\,, &(30)\\
\Rightarrow \frac{p'(\beta)}{p(\beta)} &= -\frac{2\beta}{1+\beta^2}\,, &(31)\\
\Rightarrow \ln p(\beta) &= \ln k - \ln(1+\beta^2)\,, &(32)\\
\Rightarrow p(\beta) &= \frac{k}{1+\beta^2}\,. &(33)
\end{aligned}$$

Substitution back into the unified invariance relation then gives $\gamma = 1$. Furthermore, we recognise that $p(\beta)$ is just the Cauchy distribution, which properly normalises with $k = \frac{1}{\pi}$.

3 References

[1] E.T. Jaynes (1964): "*Prior probabilities and transformation groups*" (pdf)

[2] P. Milne (1983): "*A note on scale invariance*" (ref)

[3] E.T. Jaynes (1973): "*The well-posed problem*" (pdf)

[4] A. Drory (2015): "*Failure and uses of Jaynes' principle of transformation groups*" (ref to pdf)