invariance_priors

September 24, 2022

Author: G. Jarrad

This document describes and summarises various methodologies for obtaining continuous probability distributions based on specified invariance properties. Typically, such an invariance distribution might be used as an *ignorance* prior, when little is known in advance about the distributional properties.

1 Transformation Group Invariance

The approach of *transformation group invariance* was championed by Jaynes [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 improper distributions resulting from nonlinear transformation (as discussed later in the example of nonlinear scale invariance).

Consequently, in this chapter we derive the general invariance relations, 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 Transformation geometry

We begin with two continuous spaces, $\mathcal{U}\subseteq\mathbb{R}^n$ and $\mathcal{V}\subseteq\mathbb{R}^m$, and a continuous transformation, $\mathbf{h}:\mathcal{U}\mapsto\mathcal{V}$, such that the point $\mathbf{u}=(u_1,\ldots,u_n)\in\mathcal{U}$ is mapped to its counterpart $\mathbf{v}=(v_1,\ldots,v_m)\in\mathcal{V}$ via $\mathbf{v}=\mathbf{h}(\mathbf{u})$.

We further suppose that $\mathbf{h}$ is differentiable, such that an infinitesimal displacement from $\mathbf{u}$, denoted by $d\mathbf{u}\doteq(du_1,\ldots,du_n)$, is mapped to its counterpart displacement $d\mathbf{v}\doteq(dv_1,\ldots,dv_n)$ from $\mathbf{v}$ via

$$d\mathbf{v} \;=\; \frac{\partial\mathbf{h}}{\partial\mathbf{u}}(\mathbf{u})\,d\mathbf{u}\,. \tag{1}$$

Similarly, we suppose that the infinitesimal volume element $dU = |d\mathbf{u}| \doteq du_1 \cdots du_n$ about $\mathbf{u}$ is mapped to its counterpart volume element $dV = |d\mathbf{v}| \doteq dv_1 \cdots dv_m$ about $\mathbf{v}$. At this juncture, we further require $\mathbf{h}$ to be invertible, such that no information is lost in the transformation. Consequently, spaces $\mathcal{U}$ and $\mathcal{V}$ must share common dimensionality, i.e. $n = m$. By geometrical reasoning, the two volume elements are now related via

$$|d\mathbf{v}| \;=\; |J(\mathbf{u})|\;|d\mathbf{u}|\,, \qquad J \;\doteq\; \mathtt{det}\left[\frac{\partial\mathbf{h}}{\partial\mathbf{u}}\right]\,, \tag{2}$$

where J is known as the *Jacobian* of the transformation.

1.2 Conservation of probability

In addition to the transformation geometry of the previous section, we now suppose that spaces $\mathcal{U}$ and $\mathcal{V}$ have associated (but as yet unspecified) probability distributions denoted by $p : \mathcal{U} \mapsto \mathbb{R}^+$ and $q : \mathcal{V} \mapsto \mathbb{R}^+$, respectively. We initially assume that p and q are *proper* distributions, such that they obey $\int_{\mathcal{U}} p(\mathbf{u})\,|d\mathbf{u}| = 1$ and $\int_{\mathcal{V}} q(\mathbf{v})\,|d\mathbf{v}| = 1$, respectively. It then follows that however we partition $\mathcal{U}$ and $\mathcal{V}$, the total probability mass must always be unity. This is the key to the first invariance principle of conservation of probability mass.

Note, however, that we shall later relax this constraint (without justification), allowing p and q to be *improper* distributions that do not normalise to unity (in fact, their normalisation factors will be infinite).

Suppose we notionally partition $\mathcal{U}$ into a finite number N of finite volume elements, $(\delta U_1, \ldots, \delta U_N)$, such that $\mathcal{U} = \bigcup_{i=1}^N \delta U_i$. Then the total probability mass must remain constant:

$$\int_{\mathcal{U}} p(\mathbf{u})\,|d\mathbf{u}| \;=\; \sum_{i=1}^N \int_{\delta U_i} p(\mathbf{u})\,|d\mathbf{u}| \;=\; 1\,. \tag{3}$$

Next, we apply transformation $\mathbf{h}$ to continuously map each volume element δU_i into its counterpart δV_i. Since we assumed that $\mathbf{h}$ is invertible, then the volume elements in $\mathcal{V}$ cannot overlap. Additionally, since $\mathcal{U}$ is mapped entirely into $\mathcal{V}$, there cannot be any space left over. Consequently, we must have partitioned $\mathcal{V}$ as $\mathcal{V} = \bigcup_{i=1}^N \delta V_i$, such that

$$\int_{\mathcal{V}} q(\mathbf{v})\,|d\mathbf{v}| \;=\; \sum_{i=1}^N \int_{\delta V_i} q(\mathbf{v})\,|d\mathbf{v}| \;=\; 1\,. \tag{4}$$

It therefore follows that

$$\sum_{i=1}^N \left(\int_{\delta U_i} p(\mathbf{u})\,|d\mathbf{u}| - \int_{\delta V_i} q(\mathbf{v})\,|d\mathbf{v}|\right) \;=\; 0\,. \tag{5}$$

Now, since this relation holds for arbitrary N, it must hold for all N, such that

$$\int_{\delta U} p(\mathbf{u})\,|d\mathbf{u}| \;=\; \int_{\delta V} q(\mathbf{v})\,|d\mathbf{v}|\,. \tag{6}$$

In particular, in the limit as $N \to \infty$ we have $\delta U \to dU$ and $\delta V \to dV$, such that the relation reduces to

$$p(\mathbf{u})\,|d\mathbf{u}| \;=\; q(\mathbf{v})\,|d\mathbf{v}|\,. \tag{7}$$

From the transformation geometry of the previous section, we therefore obtain the (standard) probability conservation relation

$$p(\mathbf{u}) \quad = \quad |J(\mathbf{u})|\; q(\mathbf{h}(\mathbf{u}))\,, \tag{8}$$

for all $\mathbf{u}\in\mathcal{U}$.

1.3 Distributional invariance

The second invariance is the key to the entire method of transformation groups. Consider two observers, X and Y, who both observe some event, but who take measurements $\mathbf{u}\in\mathcal{U}$ and $\mathbf{v}\in\mathcal{V}$ 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 $q=p$.

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. [Aside: In fact, taking $p=f$ and $q=g$, Milne explicitly stated the constraint as $f=\gamma g$, altough his further working out only makes sense if $g=\gamma f$ instead.]

Indeed, as noted in the introduction, Jaynes himself saw the occasional need for $q=\gamma p$, 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 define *distributional invariance* as

$$q(\mathbf{v}) \quad = \quad \gamma p(\mathbf{v})\,, \tag{9}$$

for all $\mathbf{v}\in\mathcal{V}$, and for some constant $\gamma>0$. Note that this now presupposes that $\mathcal{V}\subseteq\mathcal{U}$, since $p:\mathcal{U}\mapsto\mathbb{R}^+$.

It now follows that if p and q are *proper* distributions, as assumed in the previous section, then γ is determined uniquely by

$$\gamma \quad = \quad \frac{\int_{\mathcal{V}} q(\mathbf{v})\,|d\mathbf{v}|}{\int_{\mathcal{V}} p(\mathbf{v})\,|d\mathbf{v}|} \quad = \quad \frac{1}{\int_{\mathcal{V}} p(\mathbf{u})\,|d\mathbf{u}|}\,. \tag{10}$$

Clearly, we must then have $\gamma=1$ for $\mathcal{V}=\mathcal{U}$, and $\gamma>1$ for $\mathcal{V}\subset\mathcal{U}$.

However, if the distributions are *improper*, then γ must be determined from the above invariance relation, once the forms of p and q have been found.

1.4 Unconditional probability invariance

The third invariance follows from the distributional invariance of the previous section. Stated briefly, 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 example).

Let the transformation $\mathbf{h}$ now be redefined explicitly as a function, $\mathbf{h} : \mathcal{U} \times \Psi \mapsto \mathcal{V}$, of parameters $\psi \in \Psi$, such that $\mathbf{v} = \mathbf{h}(\mathbf{u}; \psi)$, and likewise $J(\mathbf{u}; \psi) = \mathtt{det}\left(\frac{\partial \mathbf{h}}{\partial \mathbf{u}}\right)$. Then the combination of conservation of probability and distributional invariance requires that the (unconditional) probability distribution $p(\mathbf{u})$ obeys the unified invariance relation

$$p(\mathbf{u}) \quad = \quad \gamma \, |J(\mathbf{u}; \psi)| \, p(\mathbf{h}(\mathbf{u}; \psi)) \,, \tag{11}$$

for every point $\mathbf{u} \in \mathcal{U}$, and *typically* (see the caveat further below) for all parameter values $\psi \in \Psi$. We shall call this the *unconditional (probability) invariance* relation.

Since this relation is invariant to changes in the transformation parameter values ψ, we may take derivatives with respect to ψ, giving rise to the parameter invariance relation(s)

$$\frac{\partial |J|}{\partial \psi}(\mathbf{u}; \psi) \, p(\mathbf{h}(\mathbf{u}; \psi)) + |J(\mathbf{u}; \psi)| \, \frac{\partial \mathbf{h}}{\partial \psi}(\mathbf{u}; \psi) \, p'(\mathbf{h}(\mathbf{u}; \psi)) \quad = \quad \mathbf{0} \,. \tag{12}$$

These invariances again hold for all $\mathbf{u} \in \mathcal{U}$, and typically for all $\psi \in \Psi$ (as per the caveat below).

Furthermore, if these derivatives are invariant to the values of the parameters, then the relations hold for all parameter values, including the special value ψ_0 which leads to the identity mapping $\mathbf{h}(\mathbf{u}; \psi_0) = \mathbf{u}$. Consequently, we may simplify the derivative invariances at $\psi = \psi_0$, and then solve the simplified equations for $p(\mathbf{u})$.

Caveat: Note that we stated above that the invariance relation in $\mathbf{u}$ is also typically satisfied for all transformation parameters ψ. However, in some circumstances, as we shall see in the example on nonlinear rescaling, the invariance relation might be satisfied only for variation in some transformation parameters but not others. When this occurs, we have a *family* of invariance relations dependent on (the constant values of) the non-invariant parameters. This includes the normalisation factor γ, which then more generally becomes $\gamma(\psi)$. Consequently, the more general invariance relation is in fact

$$p(\mathbf{u}) \quad = \quad \gamma(\psi) \, |J(\mathbf{u}; \psi)| \, p(\mathbf{h}(\mathbf{u}; \psi)) \,, \tag{13}$$

for all $\mathbf{u} \in \mathcal{U}$, and for all $\psi \in \Psi$.

1.5 Conditional probability invariance

So far in this chapter, we have considered only the unconditional probability distributions $p(\mathbf{u})$ and $q(\mathbf{v})$. As Jaynes repeatedly mentions, all distributions are conditional in the sense that they depend (at the minimum) upon prior information about the problem to be solved (including, but not limited to, the respective domains $\mathcal{U}$ and $\mathcal{V}$). Hence, by *unconditional* we mean not conditioned on any distributional parameters.

This now leads us to consider conditional, i.e., parameterised, distributions of the form $p(\mathbf{u} \mid \theta)$ and $q(\mathbf{v} \mid \phi)$. In addition, we also have unconditional, i.e. prior, distributions of the form $p(\theta)$ and $q(\phi)$.

The prior distributions are handled by the unconditional invariance relation of the previous section. In particular, we consider the transformation $\mathbf{g}:\Theta\times\Psi\mapsto\Phi$ that maps distributional parameters $\theta\in\Theta$ to $\phi=\mathbf{g}(\theta;\psi)\in\Phi\subseteq\Theta$. The corresponding invariance relation is then

$$p(\theta) \;=\; \gamma_{\mathbf{g}}\,|J_{\mathbf{g}}(\theta;\psi)|\,p(\mathbf{g}(\theta;\psi))\,, \qquad J_{\mathbf{g}} \;\doteq\; \mathtt{det}\left[\frac{\partial\mathbf{g}}{\partial\theta}\right]\,, \tag{14}$$

for all $\theta\in\Theta$ and all $\psi\in\Psi$, with constant $\gamma_{\mathbf{g}}>0$. Note for later use that the infinitesimal volumes $|d\theta|$ and $|d\phi|$ are related by $|d\phi| \;=\; |J_{\mathbf{g}}(\theta;\psi)|\,|d\theta|$, from the transformation geometry.

Next, we consider the joint distributions, namely $p(\mathbf{u},\theta)=p(\mathbf{u}\mid\theta)\,p(\theta)$ and $q(\mathbf{v},\phi)=q(\mathbf{v}\mid\phi)\,q(\phi)$, which in this form are also to be considered unconditional. To the existing transformation $\mathbf{g}$, we thus add the additional transformation $\mathbf{h}:\mathcal{U}\times\Theta\times\Psi\mapsto\mathcal{V}$, which maps $\mathbf{u}\in\mathcal{U}$ to $\mathbf{v}=\mathbf{h}(\mathbf{u};\theta,\psi)\in\mathcal{V}\subseteq\mathcal{U}$.

From the transformation geometry, we deduce that the infinitesimal joint displacement $d(\mathbf{u},\theta)$ is related to the corresponding joint displacement $d(\mathbf{v},\phi)$ via

$$\left[\begin{array}{c} d\mathbf{v} \\ d\phi \end{array}\right] \;=\; \left[\begin{array}{cc} \frac{\partial\mathbf{h}}{\partial\mathbf{u}} & \frac{\partial\mathbf{h}}{\partial\theta} \\ \mathbf{0} & \frac{\partial\mathbf{g}}{\partial\theta} \end{array}\right]\left[\begin{array}{c} d\mathbf{u} \\ d\theta \end{array}\right]\,. \tag{15}$$

The Jacobian of the joint transformation is therefore

$$J_{\mathbf{h},\mathbf{g}} \;\doteq\; \mathtt{det}\left[\begin{array}{cc} \frac{\partial\mathbf{h}}{\partial\mathbf{u}} & \frac{\partial\mathbf{h}}{\partial\theta} \\ \mathbf{0} & \frac{\partial\mathbf{g}}{\partial\theta} \end{array}\right] \;=\; \mathtt{det}\left[\frac{\partial\mathbf{h}}{\partial\mathbf{u}}\right]\,\mathtt{det}\left[\frac{\partial\mathbf{g}}{\partial\theta}\right]\,, \tag{16}$$

$$\Rightarrow J_{\mathbf{h},\mathbf{g}}(\mathbf{u},\theta;\psi) \;\equiv\; J_{\mathbf{h}}(\mathbf{u};\theta,\psi)\,J_{\mathbf{g}}(\theta;\psi)\,. \tag{17}$$

Consequently, the infinitesimal joint volumes $|d(\mathbf{u},\theta)|$ and $|d(\mathbf{v},\phi)|$ are related via

$$|d(\mathbf{v},\phi)| \;=\; |J_{\mathbf{h},\mathbf{g}}(\mathbf{u},\theta;\psi)|\,|d(\mathbf{u},\theta)| \tag{18}$$

$$\Rightarrow |d\mathbf{v}|\,|d\phi| \;=\; |J_{\mathbf{h}}(\mathbf{u};\theta,\psi)|\,|d\mathbf{u}|\,|J_{\mathbf{g}}(\theta;\psi)|\,|d\theta|\,, \tag{19}$$

$$\Rightarrow |d\mathbf{v}| \;=\; |J_{\mathbf{h}}(\mathbf{u};\theta,\psi)|\,|d\mathbf{u}|\,, \tag{20}$$

since we noted earlier that $|d\phi|=|J_{\mathbf{g}}(\theta;\psi)|\,|d\theta|$ for the prior distributions. Hence, the relation above holds for the conditional distributions, such that the joint invariance must factor into a prior invariance and a conditional invariance.

To see this in more detail, note that the invariance relation for the joint distribution(s) is

$$p(\mathbf{u},\theta) \;=\; \gamma_{\mathbf{h},\mathbf{g}}\,|J_{\mathbf{h},\mathbf{g}}(\mathbf{u};\theta,\psi)|\,p(\mathbf{h}(\mathbf{u};\theta,\psi),\mathbf{g}(\theta;\psi)) \tag{21}$$

$$\Rightarrow p(\mathbf{u}\mid\theta)\,p(\theta) \;=\; \gamma_{\mathbf{h},\mathbf{g}}\,|J_{\mathbf{h}}(\mathbf{u};\theta,\psi)|\,p(\mathbf{h}(\mathbf{u};\theta,\psi)\mid\mathbf{g}(\theta;\psi))\,|J_{\mathbf{g}}(\theta;\psi)|\,p(\mathbf{g}(\theta;\psi))\,, \tag{22}$$

$$\Rightarrow \gamma_{\mathbf{g}}\,p(\mathbf{u}\mid\theta) \;=\; \gamma_{\mathbf{h},\mathbf{g}}\,|J_{\mathbf{h}}(\mathbf{u};\theta,\psi)|\,p(\mathbf{h}(\mathbf{u};\theta,\psi)\mid\mathbf{g}(\theta;\psi))\,, \tag{23}$$

where the last line follows directly from the prior invariance relation, $p(\theta)=\gamma_{\mathbf{g}}|J_{\mathbf{g}}(\theta;\psi)|\,p(\mathbf{g}(\theta;\psi))$, given earlier. Hence, letting $\gamma_{\mathbf{h},\mathbf{g}}=\gamma_{\mathbf{h}}\gamma_{\mathbf{g}}$, we finally obtain the *conditional (probability) invariance* relation

$$p(\mathbf{u}\mid\theta) \;=\; \gamma_{\mathbf{h}}\,|J_{\mathbf{h}}(\mathbf{u};\theta,\psi)|\,p(\mathbf{h}(\mathbf{u};\theta,\psi)\mid\mathbf{g}(\theta;\psi))\,, \tag{24}$$

which holds for all $\mathbf{u}\in\mathcal{U}$, all $\theta\in\Theta$, and *typically* for all $\psi\in\Psi$.

However, from the caveat given in the previous section, if invariance does **not** in fact hold for all subsets of the transformation parameter ψ, then we require the more general invariance relation

$$p(\mathbf{u}\mid\theta) \;=\; \gamma_{\mathbf{h}}(\psi)\,|J_{\mathbf{h}}(\mathbf{u};\theta,\psi)|\,p(\mathbf{h}(\mathbf{u};\theta,\psi)\mid\mathbf{g}(\theta;\psi))\,, \tag{25}$$

which **does** now hold for all $\psi\in\Psi$.

1.6 Examples

1.6.1 Nonlinear rescaling plus relocation

Consider a scaling parameter $\sigma \in (0, \infty)$ and a location parameter $\mu \in \mathbb{R}$, such that the prior distriubtion $p(\sigma, \mu)$ is invariant both to further rescaling and to translation.

Jaynes [1], solved this problem for simple translation $\mu' = \mu + \nu$ and linear rescaling $\sigma' = \alpha\sigma$. However, Milne [2], solved the rescaling invariance problem (without translation) for nonlinear scaling $\sigma' = \alpha\sigma^\beta$. Here we combine both approaches.

Let the prior distribution $p(\mu)$ be invariant to the transformation

$$f(\mu; \nu) \quad = \quad \mu + \nu \,, \tag{26}$$

which has Jacobian

$$J_f(\mu; \nu) \quad = \quad \frac{\partial f}{\partial \mu} \quad = \quad 1 \,. \tag{27}$$

Note that the transformation becomes the identity for $\nu = 0$.

The unconditional invariance relation is thus

$$p(\mu) \quad = \quad \gamma_f \, p(\mu + \nu) \,. \tag{28}$$

Taking the derivative with respect to the transformation parameter ν, and setting $\nu = 1$, gives

$$0 \;=\; \gamma_f \, p'(\mu) \quad \Rightarrow \quad p(\mu) \;=\; k_f \,, \tag{29}$$

for abitrary constant $k_f > 0$. Note that this is an improper prior distribution. Substitution back into the invariance relation then gives $\gamma_f = 1$.

Next, suppose that the prior distribution $p(\sigma)$ is invariant to the transformation

$$g(\sigma; \alpha, \beta) \quad = \quad \alpha\sigma^\beta \,, \tag{30}$$

with $\alpha > 0$ and $\beta \neq 0$. Note that the transformation becomes the identity for $\alpha = 1$ and $\beta = 1$. The Jacobian of the transformation is then

$$J_g(\mu; \alpha, \beta) \quad = \quad \frac{\partial g}{\partial \sigma} \quad = \quad \alpha\beta\sigma^{\beta-1} \,, \tag{31}$$

such that the unconditional invariance relation is given by

$$p(\sigma) \quad = \quad \gamma_g \, \alpha |\beta| \sigma^{\beta-1} \, p(\alpha\sigma^\beta) \,. \tag{32}$$

Taking derivatives with respect to the transformation parameters α and β therefore give

$$0 \quad = \quad \gamma_g \, |\beta| \sigma^{\beta-1} \left[p(\alpha\sigma^\beta) + \alpha \, p'(\alpha\sigma^\beta) \, \sigma^\beta \right] \,, \tag{33}$$

$$0 \quad = \quad \gamma_g \, \alpha \, \sigma^{\beta-1} \left[\mathtt{sign}(\beta) \, p(\alpha\sigma^\beta) + |\beta| \, \ln\sigma \, p(\alpha\sigma^\beta) + |\beta| \, p'(\alpha\sigma^\beta) \, \alpha\sigma^\beta \, \ln\sigma \right] \,, \tag{34}$$

respectively.

Now, by substituting $\alpha = 1$ and $\beta = 1$, and simplifying the results, we obtain

$$
\begin{aligned}
p(\sigma) + \sigma\, p'(\sigma) &= 0\,, && (35)\\
[1 + \ln\sigma]\; p(\sigma) + \sigma \ln\sigma\, p'(\sigma) &= 0\,. && (36)
\end{aligned}
$$

It turns out that we cannot (usefully) satisfy both of these equations simultaneously. The former equation (due to variation in α) has the solution

$$
p(\sigma) = \frac{k_g}{\sigma}\,, \qquad (37)
$$

for arbitrary constant $k_g > 0$. However, the latter equation (due to variation in β) then reduces to $k_g = 0$.

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 unconditional invariance relation above, we have

$$
\begin{aligned}
p(\sigma) &= \gamma_g\, \alpha|\beta|\sigma^{\beta-1}\, p(\alpha\sigma^\beta)\,, && (38)\\
\Rightarrow \frac{k_g}{\sigma} &= \gamma_g\, \alpha|\beta|\sigma^{\beta-1}\, \frac{k_g}{\alpha\sigma^\beta}\,, && (39)\\
\Rightarrow \gamma_g &= \frac{1}{|\beta|}\,. && (40)
\end{aligned}
$$

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

In fact, we actually require the general form of the unconditional invariance, namely

$$
\begin{aligned}
p(\sigma) &= \gamma_g(\alpha,\beta)\, \alpha|\beta|\sigma^{\beta-1}\, p(\alpha\sigma^\beta)\,, && (41)\\
\Rightarrow 0 &= \left[\alpha\frac{\partial\gamma_g}{\partial\alpha} + \gamma_g\right]\, p(\sigma') + \gamma_g\alpha\sigma^\beta\, p'(\sigma')\,, && (42)\\
0 &= \frac{\partial\gamma_g}{\partial\beta}|\beta|\, p(\sigma') + \gamma_g\left[\mathtt{sign}(\beta)\, p(\sigma') + |\beta|\, \ln\sigma\, p(\sigma') + |\beta|\, p'(\sigma')\, \alpha\sigma^\beta\, \ln\sigma\right]\,, && (43)\\
\Rightarrow 0 &= \frac{\partial\gamma_g}{\partial\beta}|\beta| + \gamma_g\, \mathtt{sign}(\beta) - \alpha\frac{\partial\gamma_g}{\partial\alpha}\, |\beta|\ln\sigma\,. && (44)
\end{aligned}
$$

However, since $\gamma_g(\alpha,\beta)$ is not a function of σ, then

$$
\frac{\partial\gamma_g}{\partial\alpha} = 0 \;\Rightarrow\; \frac{\partial\gamma_g}{\partial\beta} = -\gamma_g\frac{\mathtt{sign}(\beta)}{|\beta|} \;\Rightarrow\; \gamma_g = \frac{1}{|\beta|}\,. \qquad (45)
$$

We see that taking $\beta = 1$ results in the linear scaling of Jaynes [1], for which $\gamma_g = 1$. Also note that now $\gamma_g \neq 1$ in general for $\beta \neq 1$ (except for $\beta = -1$), despite the fact that both $\sigma, \sigma' \in (0,\infty)$. This is due, as discussed in the section on distributional invariance, to the fact that $p(\sigma)$ is an *improper* distribution that cannot actually be properly normalised. Despite this, we maintain (without proof) that conservation of probability continues to hold.

Finally, we now turn to the subsequent problem posed by Jaynes [1], namely that of invariance of the conditional distribution $p(x \mid \sigma, \mu)$ to rescaling and relocation, specifically of the form $x' =$

$(\sigma'/\sigma)(x-\mu)+\mu'$. Allowing for nonlinear rescaling, we therefore suppose that $p(x \mid \sigma, \mu)$ is invariant to the transformation

$$h(x;\sigma,\mu;\alpha,\beta,\nu) \;=\; \alpha\sigma^{\beta-1}\,x - (\alpha\sigma^{\beta-1}-1)\,\mu + \nu\,. \tag{46}$$

Note that this transformation becomes the identity for $\alpha = 1$, $\beta = 1$ and $\nu = 0$. Also note that its Jacobian is

$$J_h(x;\sigma,\mu;\alpha,\beta,\nu) \;=\; \frac{\partial h}{\partial x} \;=\; \alpha\sigma^{\beta-1}\,. \tag{47}$$

Hence, via conditional invariance, we obtain the further invariance relation

$$p(x \mid \sigma,\mu) \;=\; \gamma_h\,|J_h|\,p(x' \mid \sigma',\mu')\,, \tag{48}$$

$$\Rightarrow p(x \mid \sigma,\mu) \;=\; \gamma_h\,\alpha\sigma^{\beta-1}\,p(\alpha\sigma^{\beta-1}\,x - (\alpha\sigma^{\beta-1}-1)\,\mu + \nu \mid \alpha\sigma^{\beta}, \mu+\nu)\,. \tag{49}$$

We observed above that the transformation g for σ is not invariant to variation in the transformation parameter β. Hence, taking only derivatives with respect to the other transformation parameters, α and ν, we obtain

$$0 \;=\; \gamma_h\,\sigma^{\beta-1}\left[p(x' \mid \sigma',\mu') + \alpha\,\frac{\partial p}{\partial x}(x' \mid \sigma',\mu')\,\sigma^{\beta-1}(x-\mu) + \alpha\,\frac{\partial p}{\partial \sigma}(x' \mid \sigma',\mu')\,\sigma^{\beta}\right], \tag{50}$$

$$0 \;=\; \gamma_h\,\alpha\sigma^{\beta-1}\left[\frac{\partial p}{\partial x}(x' \mid \sigma',\mu') + \frac{\partial p}{\partial \mu}(x' \mid \sigma',\mu')\right], \tag{51}$$

respectively. Thus, at $\alpha = 1$, $\beta = 1$ and $\nu = 0$, we have

$$p + (x-\mu)\frac{\partial p}{\partial x} + \sigma\frac{\partial p}{\partial \sigma} \;=\; 0\,, \tag{52}$$

$$\frac{\partial p}{\partial x} + \frac{\partial p}{\partial \mu} \;=\; 0\,. \tag{53}$$

The latter equation indicates that $p(x \mid \sigma, \mu)$ is a function of $x - \mu$, and the former equation is then satisfied in general for

$$p(x \mid \sigma,\mu) \;=\; \frac{k_h}{\sigma}\varphi\left(\frac{x-\mu}{\sigma}\right), \tag{54}$$

where $\varphi(\cdot)$ is known as a *radial basis function*. Observe that this transformation invariance therefore holds for the Gaussian distribution with basis function $\varphi(z) = e^{-\frac{1}{2}z^2}$.

Finally, substituting this back into the conditional invariance relation, we obtain

$$\frac{k_h}{\sigma}\varphi\left(\frac{x-\mu}{\sigma}\right) \;=\; \gamma_h\,\alpha\sigma^{\beta-1}\,\frac{k_h}{\alpha\sigma^{\beta}}\varphi\left(\frac{x-\mu}{\sigma}\right) \quad\Rightarrow\quad \gamma_h \;=\; 1\,. \tag{55}$$

Lastly, just for the sake of consistency, we check the derivative of the invariance relation with respect to β, namely

$$0 \;=\; \gamma_h\,\alpha\left[\sigma^{\beta-1}\ln\sigma\,p(x' \mid \sigma',\mu') + \sigma^{\beta-1}\frac{\partial p}{\partial x}(x' \mid \sigma',\mu')\,\alpha\sigma^{\beta-1}(x-\mu)\,\ln\sigma\right. \tag{56}$$

$$\left. + \,\sigma^{\beta-1}\frac{\partial p}{\partial \sigma}(x' \mid \sigma',\mu')\,\alpha\sigma^{\beta}\ln\sigma\right]. \tag{57}$$

Evaluating this at $\alpha = 1$, $\beta = 1$ and $\nu = 0$ gives

$$\ln\sigma\, p + (x-\mu)\, \ln\sigma \frac{\partial p}{\partial x} + \sigma \ln\sigma \frac{\partial p}{\partial \sigma} \quad = \quad 0\,. \tag{58}$$

It turns out that this is just the invariance relation for α multiplied by $\ln\sigma$, and hence is entirely consistent. In fact, this consistency means that $p(x \mid \sigma, \mu)$ is also invariant to changes in β, and is thus the reason that $\gamma_h = 1$ is not a function of β.

1.6.2 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

$$f(\rho) \;=\; |J(\rho)|\, g(\nu) \quad\Rightarrow\quad g(\nu) \;=\; \frac{1}{\nu^2} f\left(\frac{1}{\nu}\right) \;=\; \frac{1}{(b-a)\nu^2}\,. \tag{59}$$

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

$$f(\rho) \;=\; \frac{k}{\rho}\,, \qquad g(\nu) \;=\; \frac{1}{\nu^2} f\left(\frac{1}{\nu}\right) \;=\; \frac{k}{\nu}\,, \tag{60}$$

respectively. Since we assumed a finite domain $\rho \in [a,b]$, we can 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.

1.6.3 Linear regression coefficients

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 $p(\beta)$, of the regression coefficients.

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 $\alpha' = \alpha + \nu$. The unconditional invariance relation is thus

$$p(\alpha) \;=\; \gamma_1\, p(\alpha+\nu)\,. \tag{61}$$

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

$$0 \;=\; \gamma_1\, p'(\alpha) \;\;\Rightarrow\;\; p(\alpha) \;=\; k_1\,,\;\; \gamma_1 \;=\; 1\,. \tag{62}$$

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:

$$\begin{aligned} \beta &= \tan\theta \;=\; \frac{\sin\theta}{\cos\theta} \;=\; \frac{\mathtt{sign}(\theta)\sqrt{1-\cos^2\theta}}{\cos\theta} & (63)\\ \Rightarrow \cos\theta &= \frac{1}{\sqrt{1+\beta^2}}\,. & (64) \end{aligned}$$

Next, we see that

$$\begin{aligned} \frac{d\tan\theta}{d\theta} &= \frac{d}{d\theta}\left(\frac{\sin\theta}{\cos\theta}\right) \;=\; \frac{\cos^2\theta+\sin^2\theta}{\cos^2\theta} & (65)\\ &= 1+\tan^2\theta \;=\; \frac{1}{\cos^2\theta} \;=\; \sec^2\theta\,, & (66)\\ \Rightarrow \frac{d\beta}{d\theta} &= 1+\beta^2\,. & (67) \end{aligned}$$

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 unconditional invariance relation is then

$$p(\beta)\left|\frac{d\beta}{d\theta}\right| \;=\; \gamma_2\, p(\beta')\left|\frac{\partial\beta'}{\partial\theta}\right|\,, \tag{68}$$

where

$$\frac{\partial\beta'}{\partial\theta} \;=\; \frac{\partial\beta'}{\partial\psi} \;=\; 1+\beta'^2\,. \tag{69}$$

Hence, the completed relation is

$$(1+\beta^2)\,p(\beta) \;=\; \gamma_2\,(1+\beta'^2)\,p(\beta')\,. \tag{70}$$

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

$$0 \;=\; \gamma_2\left[2\beta'(1+\beta'^2)\,p(\beta') + (1+\beta'^2)^2\,p'(\beta')\right]\,. \tag{71}$$

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

$$\begin{aligned} 2\beta\, p(\beta) + (1+\beta^2)\, p'(\beta) &= 0\,, & (72)\\ \Rightarrow \frac{p'(\beta)}{p(\beta)} &= -\frac{2\beta}{1+\beta^2}\,, & (73)\\ \Rightarrow \ln p(\beta) &= \ln k_2 - \ln(1+\beta^2)\,, & (74)\\ \Rightarrow p(\beta) &= \frac{k_2}{1+\beta^2}\,. & (75) \end{aligned}$$

Substitution back into the unified invariance relation then gives $\gamma_2 = 1$. Furthermore, we recognise that $p(\beta)$ is just the Cauchy distribution, which properly normalises with $k_2 = \frac{1}{\pi}$.

1.6.4 Time-scale invariance

Consider the conditional distribution $p(t \mid \lambda)$ for a temporal variate $t \in [0, \infty)$ and a frequency parameter $\lambda \in (0, \infty)$. The proper scale of time is thus given by $\tau \doteq \frac{1}{\lambda}$.

Next, as per Jaynes' suggestion, consider two observers, X and Y, whose watches run at different rates. Hence, observer X has time-scale τ, but observer Y has time-scale $\tau' = \alpha\tau$, for $\alpha > 0$. Assuming the prior $p(\tau)$ is invariant to this transformation, we obtain, from the examples on rescaling and the Von Kries paradox, that

$$p(\tau) = \frac{k}{\tau}\,, \qquad p(\lambda) = \frac{k}{\lambda}\,. \tag{76}$$

Similarly, we see that if observer X measures time as t, then observer Y measures times as $t' = \frac{\tau'}{\tau} t = \alpha t$. Assuming that $p(t \mid \lambda)$ is invariant to this transformation, we obtain the conditional invariance relation

$$p(t \mid \lambda) = \gamma\alpha\, p\left(\alpha t \mid \frac{1}{\alpha}\lambda\right)\,, \tag{77}$$

for some $\gamma > 0$, where we define $\lambda' \doteq \frac{1}{\tau'} = \frac{\lambda}{\alpha}$.

Hence, taking the derivative with respect to the transformation parameter α gives

$$0 = \gamma\left[p(t' \mid \lambda') + \alpha\, t\frac{\partial p}{\partial t}(t' \mid \lambda') - \frac{\lambda}{\alpha}\frac{\partial p}{\partial \lambda}(t' \mid \lambda')\right]\,. \tag{78}$$

Lastly, we substitute $\alpha = 1$, at which point the transformation is the identity, and simplify the result, giving

$$p(t \mid \lambda) + t\frac{\partial p}{\partial t}(t \mid \lambda) - \lambda\frac{\partial p}{\partial \lambda}(t \mid \lambda) = 0\,. \tag{79}$$

This has the solution

$$p(t \mid \lambda) = \lambda e^{-\lambda t}\,, \tag{80}$$

which we recognise as the *exponential distribution.*

2 Minimum Information Priors

Apart from the transformation group invariance discussed in the previous chapter, there are other methodologies for choosing prior distributions on the grounds on *noninformativeness*. One such method, as discussed by Zellner [5], is that of choosing the prior to minimise a formal notion of entropic information.

2.1 Minimum information principle

To set the scene, consider the conditional distribution $p(\mathbf{u} \mid \theta)$, for $\mathbf{u} \in \mathcal{U}$ and $\theta \in \Theta$. The information-theoretic entropy of this distribution is defined as

$$H(\mathcal{U} \mid \theta) \doteq -\int_{\mathcal{U}} p(\mathbf{u} \mid \theta) \, \log p(\mathbf{u} \mid \theta) \, |d\mathbf{u}| \, , \tag{81}$$

which is conditional on some arbitrary but fixed value of the distributional parameters θ. Hence, for some prior distribution $p(\theta)$, the average entropy is just the conditional entropy, defined as

$$H(\mathcal{U} \mid \Theta) \doteq \int_{\Theta} H(\mathcal{U} \mid \theta) \, p(\theta) \, |d\theta| \, . \tag{82}$$

The prior distribution $p(\theta)$ itself has the entropy

$$H(\Theta) \doteq -\int_{\Theta} p(\theta) \, \log p(\theta) \, |d\theta| \, , \tag{83}$$

such that the joint entropy is given by

$$\begin{aligned} H(\mathcal{U}, \Theta) &\doteq -\int_{\Theta} \int_{\mathcal{U}} p(\mathbf{u}, \theta) \, \log p(\mathbf{u}, \theta) \, |d\mathbf{u}| \, |d\theta| && (84) \\ &= -\int_{\Theta} \int_{\mathcal{U}} p(\mathbf{u} \mid \theta) \, p(\theta) \, \log \left[p(\mathbf{u} \mid \theta) \, p(\theta) \right] \, |d\mathbf{u}| \, |d\theta| && (85) \\ &= H(\mathcal{U} \mid \Theta) + H(\Theta) \, . && (86) \end{aligned}$$

The maximum entropy principle now asserts that the distribution with the greatest entropy embodies the least prior information, such that a gain in information corresponds to a loss in entropy. In fact, Zellner [5] would appear to define information as the negative of entropy.

We therefore consider the change in entropy specified via

$$G(\mathcal{U}, \Theta) \doteq H(\mathcal{U} \mid \Theta) - H(\Theta) \, . \tag{87}$$

Observe that as $p(\theta)$ becomes more informative, then $H(\Theta)$ decreases, and (roughly speaking) $G(\mathcal{U}, \Theta)$ increases. Thus, $G(\mathcal{U}, \Theta)$ apparently measures the additional information contained in the prior distribution $p(\theta)$ that is not also present in the sampling distribution $p(\mathbf{u} \mid \theta)$. Consequently, minimising G should correspond to the choosing the least informative prior $p(\theta)$, relative to a fixed distribution $p(\mathbf{u} \mid \theta)$. This is the *minimum information principle*.

Subject to the constraint that

$$\int_{\Theta} p(\theta) \, |d\theta| = 1 \, , \tag{88}$$

we therefore introduce the Lagrangian multiplier λ, and minimise the functional

$$F[p] \quad \doteq \quad \lambda + \int_{\Theta} \left\{ H(\mathcal{U} \mid \theta) + \log p(\theta) - \lambda \right\} \, p(\theta) \, |d\theta| \, . \tag{89}$$

By the calculus of variations, F is minimised when

$$H(\mathcal{U} \mid \theta) - \lambda + \frac{1}{\ln b} + \log p(\theta) \quad = \quad 0 \, , \tag{90}$$

where b is the base of the logarithm. Consequently, we obtain

$$p(\theta) \quad = \quad k \, e^{-H(\mathcal{U}|\theta) \, \ln b} \, , \tag{91}$$

as the *minimum information prior.*

2.2 Examples

2.2.1 Gaussian distribution prior

Following Zellner [5], we consider the univariate Gaussian distribution

$$p(x \mid \mu, \sigma) \quad = \quad \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(x-\mu)^2}{2\sigma^2}} \, , \tag{92}$$

such that

$$\begin{aligned}
\ln p(x \mid \mu, \sigma) \quad &= \quad -\frac{1}{2} \ln(2\pi\sigma^2) - \frac{(x-\mu)^2}{2\sigma^2} \, , && (93)\\
\Rightarrow H(X \mid \mu, \sigma) \quad &= \quad \int_{-\infty}^{\infty} \left[\frac{1}{2} \ln(2\pi\sigma^2) + \frac{(x-\mu)^2}{2\sigma^2} \right] p(x \mid \mu, \sigma), dx && (94)\\
&= \quad \frac{1}{2} \ln(2\pi\sigma^2) + \frac{1}{2\sigma^2} \mathtt{Var}(X) && (95)\\
&= \quad \frac{1}{2} \left[1 + \ln(2\pi\sigma^2) \right] \quad = \quad \frac{1 + \ln(2\pi)}{2} + \ln \sigma \, , && (96)
\end{aligned}$$

since $\mathtt{Var}[X] = \sigma^2$. The minimum information prior is therefore (with logarithmic base $b = e$) given by

$$p(\mu, \sigma) \quad = \quad k \, e^{-H(X|\theta)} \quad = \quad \frac{k'}{\sigma} \, . \tag{97}$$

This agrees with the joint prior found using transformation group invariance.

2.2.2 Exponential distribution prior

We turn now to the exponential distribution

$$\begin{aligned}
p(t \mid \lambda) \quad &= \quad \lambda e^{-\lambda t} \, , && (98)\\
\Rightarrow \ln p(t \mid \lambda) \quad &= \quad \ln \lambda - \lambda t \, , && (99)\\
\Rightarrow H(T \mid \lambda) \quad &= \quad \int_0^{\infty} \left[\lambda t - \ln \lambda \right] p(t \mid \lambda) \, dt && (100)\\
&= \quad \lambda E[T \mid \lambda] - \ln \lambda \quad = \quad 1 - \ln \lambda \, , && (101)
\end{aligned}$$

since we know that $E[T \mid \lambda] = \frac{1}{\lambda}$. The minimum information prior is therefore

$$p(\lambda) \;=\; k\, e^{-H(T|\lambda)} \;=\; k'\lambda\,. \tag{102}$$

However, this form of prior is unexpected, since we saw that invariance to scaling led to the prior $p(\lambda) \propto \frac{1}{\lambda}$.

In fact, if we reparameterise using the time-scale $\tau = \frac{1}{\lambda}$, then we instead obtain

$$p(t \mid \tau) \;=\; \frac{1}{\tau} e^{-\frac{t}{\tau}}\,, \tag{103}$$

$$\Rightarrow \ln p(t \mid \tau) \;=\; -\ln\tau - \frac{t}{\tau}\,, \tag{104}$$

$$\Rightarrow H(T \mid \tau) \;=\; \int_0^\infty \left[\frac{t}{\tau} + \ln\tau\right] p(t \mid \tau)\, dt \tag{105}$$

$$= \;\frac{1}{\tau} E[T \mid \tau] + \ln\tau \;=\; 1 + \ln\tau\,, \tag{106}$$

since $E[T \mid \tau] = \tau$. The minimum information prior is now

$$p(\tau) \;=\; k\, e^{-H(T|\tau)} \;=\; \frac{k'}{\tau}\,. \tag{107}$$

This has the form of prior expected from invariance to scaling. Hence, we conclude that the minimum information prior is **not** generally invariant to transformation.

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)

[5] A. Zellner (1996): "*An introduction to Bayesian inference in econometrics*" (ref)