Supplementary Information for

Power laws derived from a Bayesian decision-making model in non-stationary environments

Shuji Shinohara[1*], Nobuhito Manome[12], Yoshihiro Nakajima[3], Yukio Pegio Gunji[4], Toru Moriyama[5], Hiroshi Okamoto[1], Shunji Mitsuyoshi[1], Ung-il Chung[1]

1. Department of Bioengineering, Graduate School of Engineering, The University of Tokyo, Tokyo, 113-8656, Japan
2. Department of Research and Development, SoftBank Robotics Group Corp., Tokyo, 105-0021, Japan
3. Graduate School of Economics, Osaka City University, Osaka, 558-8585, Japan
4. Department of Intermedia Art and Science, School of Fundamental Science and Technology, Waseda University, Tokyo, 169-8555, Japan
5. Faculty of Textile Science, Shinshu University, Ueda, 386-8567, Japan

Extended Bayesian Inference

In the field of cognitive psychology, many causal induction experiments have been conducted to determine how humans evaluate the strength of the causal relationship between two events [1-5]. In the case of the usual conditional statement 'if *p*, then *q*', one would think that the confidence in this statement would be proportional to the probability $P(q \mid p)$ of *q* occurring after *p* occurs [6].

In contrast, it has been experimentally demonstrated that humans have a strong sense of causal relation between a cause *p* and an effect *q* when $P(p \mid q)$ is high as well as when $P(q \mid p)$ is high. Specifically, the causal intensity that people feel between *p* and *q* can be approximated by the geometric mean of $P(q \mid p)$ and $P(p \mid q)$. This is called the dual-factor heuristics (DFH) model [1]. If the causal intensity between *p* and *q* is denoted as $DFH(q \mid p)$, then $DFH(q \mid p) = \sqrt{P(q \mid p) P(p \mid q)}$. Here, note that $DFH(q \mid p) = DFH(p \mid q)$ is valid. Such an inference is called 'symmetry inference'.

More generally, Shinohara et al. proposed the following model to express the causal strength $C(q \mid p)$ between *p* and *q* as the generalised weighted average of $P(q \mid p)$ and $P(p \mid q)$ [7].

$$C(q \mid p) = \left[\alpha P(q \mid p)^m + (1-\alpha) P(p \mid q)^m \right]^{1/m} \quad \text{(S1)}$$

The generalised weighted average of the variables x and y is expressed by the following equation using the parameters α and m.

$$\mu(\alpha, m) = \left[(1-\alpha) x^m + \alpha y^m \right]^{1/m} \quad \text{(S2)}$$

Here, α takes a value of the range $0.0 \le \alpha \le 1.0$ and represents the weighting of x and y. m takes a value of the range $-\infty \le m \le \infty$ and represents the way the average is taken. For example, if $\alpha = 0.5$ and $m = 1.0$, then $\mu(0.5, 1.0) = 0.5x + 0.5y$, which represents the arithmetic mean. If $\alpha = 0.5$ and $m = -1.0$, then $\mu(0.5, -1.0) = 2xy/(x+y)$, which represents the harmonic mean. Although equation (S2) cannot be defined when $m = 0.0$, if we denote the mean value in the limit of $m \to 0.0$ as $\mu(\alpha, 0.0)$, then we have $\mu(\alpha, 0.0) = x^{1-\alpha} y^{\alpha}$ and the geometric mean $\mu(0.5, 0.0) = \sqrt{xy}$ when $\alpha = 0.5$. If we set $\alpha = 0.0$ here, we obtain $C(q \mid p) = P(q \mid p)$ regardless of the value of *m*, and *C* corresponds to the conditional probability *P*.

Furthermore, Shinohara et al. proposed an extended Bayesian inference that incorporates such a causal inference element into Bayesian inference [7, 8].

$$C^{t+1}(h_k) \leftarrow \left[\alpha C^t(d^t \mid h_k)^m + (1-\alpha) C^t(h_k \mid d^t)^m \right]^{1/m} = \frac{C^t(h_k) C^t(d^t \mid h_k)}{\left[(1-\alpha) C^t(d^t)^{-m} + \alpha C^t(h_k)^{-m} \right]^{1/m}} \quad \text{(S3)}$$

$$C^{t+1}(d^t \mid h_k) \leftarrow \left[(1-\alpha) C^t(d^t \mid h_k)^m + \alpha C^t(h_k \mid d^t)^m \right]^{1/m} = \frac{C^t(h_k) C^t(d^t \mid h_k)}{\left[(1-\alpha) C^t(h_k)^{-m} + \alpha C^t(d^t)^{-m} \right]^{1/m}} \quad \text{(S4)}$$

Here,

$$C^t(d^t) = \sum_k C^t(h_k) C^t(d^t \mid h_k) \quad \text{(S5)}$$

In equation (S3), we omit the description of the normalisation process to make the confidence a probability. If we set $\alpha = 0$ in equation (S3), we obtain the same form as the case of Bayesian inference.

$$C^{t+1}(h_k) \leftarrow \frac{C^t(h_k) C^t(d^t \mid h_k)}{C^t(d^t)} \quad \text{(S6)}$$

When $\alpha = 0$, equation (S4) is expressed as follows, and the model of the hypothesis is invariant.

$$C^{t+1}(d^t \mid h_k) \leftarrow C^t(d^t \mid h_k) \quad \text{(S7)}$$

That is, equation (S4) is greatly reduced and the extended Bayesian inference agrees with the Bayesian inference.

On the other hand, when $\alpha > 0$, the model is deformed by equation (S4). In this paper, we do not update the models of all the hypotheses, but only the model of the hypothesis $h^t_{\max}$ that has the highest confidence at that time.

$$C^{t+1}\left(d^t \mid h^t_{\max}\right) \leftarrow \frac{C^t\left(h^t_{\max}\right) C^t\left(d^t \mid h^t_{\max}\right)}{\left[(1-\alpha) C^t\left(h^t_{\max}\right)^{-m} + \alpha C^t\left(d^t\right)^{-m}\right]^{-1/m}} \quad \text{(S8)}$$

If there are multiple hypotheses with an equally high maximum degree of confidence, one of them is selected at random.

For simplicity, we have fixed $m = 0$ in this paper. When $m = 0$, equation (S3) can be transformed as follows:

$$C^{t+1}\left(h_k\right) \leftarrow \frac{C^t\left(h_k\right) C^t\left(d^t \mid h_k\right)}{C^t\left(d^t\right)^{1-\alpha} C^t\left(h_k\right)^{\alpha}} = \left[\frac{C^t\left(h_k\right)}{C^t\left(d^t\right)}\right]^{1-\alpha} C^t\left(d^t \mid h_k\right) \quad \text{(S9)}$$

Noting the recurrent nature of $C^t\left(h_k\right)$, equation (S9) can be further transformed as follows:

$$C^{t+1}\left(h_k\right) \leftarrow \left[C^1\left(h_k\right)\right]^{(1-\alpha)^t} \prod_{i=1}^{t} \frac{\left[C^i\left(d^i \mid h_k\right)\right]^{(1-\alpha)^{t-i}}}{\left[C^i\left(d^i\right)\right]^{(1-\alpha)^{t+1-i}}} \quad \text{(S10)}$$

In equation (S10), the denominator $C^t\left(d^t\right)$ of the right-hand side is common in each hypothesis and can be considered as a constant, so if the normalisation process is omitted, it can be written as follows:

$$C^{t+1}\left(h_k\right) \leftarrow \left[C^1\left(h_k\right)\right]^{(1-\alpha)^t} \prod_{i=1}^{t} \left[C^i\left(d^i \mid h_k\right)\right]^{(1-\alpha)^{t-i}} \quad \text{(S11)}$$

When $m = 0$, equation (S8) can be transformed with respect to $h^t_{\max}$ as follows:

$$C^{t+1}\left(d^t \mid h^t_{\max}\right) \leftarrow \frac{C^t\left(h^t_{\max}\right) C^t\left(d^t \mid h^t_{\max}\right)}{C^t\left(h^t_{\max}\right)^{1-\alpha} C^t\left(d^t\right)^{\alpha}} = \left[\frac{C^t\left(h^t_{\max}\right)}{C^t\left(d^t\right)}\right]^{\alpha} C^t\left(d^t \mid h^t_{\max}\right) \quad \text{(S12)}$$

Through the processes described above, the confidence values for each hypothesis and the model for the hypothesis with maximum confidence are corrected whenever the data are observed. We refer to the latter process of modifying the model for $h^t_{\max}$ as inverse Bayesian inference [9]. If the former process of updating the confidence values for hypotheses is referred to as inference, inverse Bayesian inference can be called ‘learning’ because it forms a model for a hypothetical instead of an inference. In this sense, although α in equation (S9) and α in equation (S12) are the same parameter, they can be considered as forgetting and learning rates, respectively, and can be kept separate. In this paper, we introduce a forgetting rate β and a learning rate γ, and transform equations (S9) and (S12) as follows:

$$C^{t+1}\left(h_k\right) \leftarrow \frac{C^t\left(h_k\right)C^t\left(d^t \mid h_k\right)}{C^t\left(d^t\right)^{1-\beta}C^t\left(h_k\right)^{\beta}} = \left[\frac{C^t\left(h_k\right)}{C^t\left(d^t\right)}\right]^{1-\beta} C^t\left(d^t \mid h_k\right) \qquad \text{(S13)}$$

$$C^{t+1}\left(d^t \mid h^t_{\max}\right) \leftarrow \frac{C^t\left(h^t_{\max}\right)C^t\left(d^t \mid h^t_{\max}\right)}{C^t\left(h^t_{\max}\right)^{1-\gamma}C^t\left(d^t\right)^{\gamma}} = \left[\frac{C^t\left(h^t_{\max}\right)}{C^t\left(d^t\right)}\right]^{\gamma} C^t\left(d^t \mid h^t_{\max}\right) \qquad \text{(S14)}$$

See the next section for methods that apply the normal distribution as a specific generative model in the extended Bayesian inference.

Applying a normal distribution

In our model, the confidences $C^t\left(h_k\right)$ for each hypothesis h_k and the model $C^t\left(d^t \mid h^t_{\max}\right)$ for the hypothesis $h^t_{\max}$ with maximum confidence are corrected whenever the data d^t are observed at time *t* using equations (S13) and (S14).

In this paper, we consider the following one-dimensional normal distribution as a concrete model of the hypothesis. For simplicity, we assume that the variance Σ is the same at all times for all hypotheses, and we consider only the difference in the mean μ^t_k. See [8] for a method to estimate the mean and variance simultaneously.

$$C^t\left(d \mid h_k\right) = N\left(d \mid \mu^t_k, \Sigma\right) = \frac{1}{\sqrt{2\pi\Sigma}}\exp\left[-\frac{\left(d-\mu^t_k\right)^2}{2\Sigma}\right] \qquad \text{(S15)}$$

When adopting a normal distribution as a model, if the number of hypotheses is discrete and finite, $C^t\left(h_k\right)$ is a probability and $C^t\left(d \mid h_k\right)$ or $C^t\left(d\right)$ is a probability density. For this reason, we introduce a positive number Δ when computing equations (S13) and (S14) as follows:

$$C^{t+1}\left(h_k\right) \leftarrow \left[C^t\left(h_k\right)\right]^{1-\beta} \Delta N\left(d^t \mid \mu^t_k, \Sigma\right) \qquad \text{(S16)}$$

$$\begin{aligned} C^{t+1}\left(d^t \mid h^t_{\max}\right) &\leftarrow \frac{1}{\Delta}\left[\frac{C^t\left(h^t_{\max}\right)}{\sum_k C^t\left(h_k\right)\Delta N\left(d^t \mid \mu^t_k, \Sigma\right)}\right]^{\gamma} \Delta N\left(d^t \mid \mu^t_{\max}, \Sigma\right) \\ &= \frac{1}{\Delta^{\gamma}}\left[\frac{C^t\left(h^t_{\max}\right)}{\sum_k C^t\left(h_k\right) N\left(d^t \mid \mu^t_k, \Sigma\right)}\right]^{\gamma} N\left(d^t \mid \mu^t_{\max}, \Sigma\right) \end{aligned} \qquad \text{(S17)}$$

Here, $\mu^t_{\max}$ is the mean of the model of hypothesis $h^t_{\max}$.

In equation (S16), the term Δ is common to all hypotheses and can be cancelled by normalisation. Therefore, if we omit the normalisation process, we can express (S16) as follows:

$$C^{t+1}(h_k) \leftarrow \left[C^t(h_k)\right]^{1-\beta} N\left(d^t \mid \mu_k^t, \Sigma\right) \qquad \text{(S18)}$$

In equation (S18), once the confidence $C^t(h_k)$ of each hypothesis becomes 0, it is always 0 thereafter. To prevent this, a normalisation process (smoothing) is performed by adding a small positive constant ε to the confidence of each hypothesis obtained by Equation (S18).

$$C^{t+1}(h_k) \leftarrow \frac{C^{t+1}(h_k)+\varepsilon}{\sum_{j=1}^{K}\left[C^{t+1}(h_j)+\varepsilon\right]} = \frac{C^{t+1}(h_k)+\varepsilon}{K\varepsilon+\sum_{j=1}^{K}C^{t+1}(h_j)} \text{(S19)}$$

In this paper, we set $\varepsilon = 10^{-8}$. K represents the number of hypotheses.

When observing the data d^t at time t, the likelihood is changed to $C^{t+1}\left(d^t \mid h_{\max}^t\right)$ by equation (S17). Accordingly, we modify the mean of the model of the hypothesis $h_{\max}^t$ from $\mu_{\max}^t$ to $\mu_{\max}^{t+1}$ so that the following equation is satisfied:

$$C^{t+1}\left(d^t \mid h_{\max}^t\right) = N\left(d^t \mid \mu_{\max}^{t+1}, \Sigma\right) = \frac{1}{\sqrt{2\pi\Sigma}}\exp\left[-\frac{\left(d^t-\mu_{\max}^{t+1}\right)^2}{2\Sigma}\right] \qquad \text{(S20)}$$

Solving equation (S20) for $\mu_{\max}^{t+1}$ yields the following two solutions.

$$\begin{aligned}\mu_1 &= d^t + \sqrt{-2\Sigma\log\left[C^{t+1}\left(d^t \mid h_{\max}^t\right)\sqrt{2\pi\Sigma}\right]}\\ \mu_2 &= d^t - \sqrt{-2\Sigma\log\left[C^{t+1}\left(d^t \mid h_{\max}^t\right)\sqrt{2\pi\Sigma}\right]}\end{aligned} \qquad \text{(S21)}$$

We define $\mu_{\max}^{t+1}$ as the solution that is closer to $\mu_{\max}^t$. Specifically,

$$\mu_{\max}^{t+1} = \begin{cases}\mu_1 & if \quad \left|\mu_1-\mu_{\max}^t\right| \le \left|\mu_2-\mu_{\max}^t\right| \\ \mu_2 & otherwise\end{cases} \qquad \text{(S22)}$$

However, in order to solve equation (S20) as an equation of $\mu_{\max}^{t+1}$, $C^{t+1}\left(d^t \mid h_{\max}^t\right)$ must be in the range of $0 < C^{t+1}\left(d^t \mid h_{\max}^t\right) \le \frac{1}{\sqrt{2\pi\Sigma}}$.

For this reason, we set the following constraint when calculating $C^{t+1}\left(d^t \mid h_{\max}^t\right)$ using equation (S17):

$$C^{t+1}\left(d^t \mid h_{\max}^t\right) \leftarrow \min\left(\max\left(C^{t+1}\left(d^t \mid h_{\max}^t\right), \varepsilon\right), \frac{1}{\sqrt{2\pi\Sigma}}\right) \text{(S23)}$$

We set $\varepsilon = 10^{-8}$.

Let us consider a situation $C^t\left(h_{\max}^t\right) \approx 1$ where the confidence of the hypothesis with the highest confidence is almost 1. Because the confidence of any other hypothesis other than $h_{\max}^t$ is almost zero by the constraint of $\sum_k C^t(h_k) = 1$, we obtain

$C^t\left(d^t\right)=\sum_k C^t\left(h_k\right)C^t\left(d^t \mid h_k\right)\approx C^t\left(d^t \mid h^t_{\max}\right)$ from $C^t\left(d^t\right)=\sum_k C^t\left(h_k\right)C^t\left(d^t \mid h_k\right)$. Therefore, equation (S17) can be transformed as follows:

$$C^{t+1}\left(d^t \mid h^t_{\max}\right)\leftarrow\frac{1}{\Delta^\gamma}\left[\frac{C^t\left(h^t_{\max}\right)}{\sum_k C^t\left(h_k\right)N\left(d^t \mid \mu^t_k,\Sigma\right)}\right]^\gamma N\left(d^t \mid \mu^t_{\max},\Sigma\right)\approx\left(\frac{1}{\Delta}\right)^\gamma\left[C^t\left(d^t \mid h^t_{\max}\right)\right]^{1-\gamma} \quad \text{(S24)}$$

If equation (S24) is denoted by $x^{t+1}=f\left(x^t\right)=\left(\frac{1}{\Delta}\right)^\gamma\left(x^t\right)^{1-\gamma}$, then $f\left(x^t\right)$ becomes a concave function.

Solving $x^t=f\left(x^t\right)$ results in $x^t=0,\frac{1}{\Delta}$.

The fixed point $\left(x^t,f\left(x^t\right)\right)=\left(\frac{1}{\Delta},\frac{1}{\Delta}\right)$ is a stable point because $x^t\geq f\left(x^t\right)$ when $x^t>\frac{1}{\Delta}$ and $x^t\leq f\left(x^t\right)$ when $x^t<\frac{1}{\Delta}$. In this study, we set $\Delta=\sqrt{2\pi\Sigma}$. In this case, $C^{t+1}\left(d^t \mid h^t_{\max}\right)$ approaches the vertex $\frac{1}{\Delta}=\frac{1}{\sqrt{2\pi\Sigma}}$ of the normal distribution whenever data d^t are observed.

As shown in formula (S20), $\mu^{t+1}_{\max}$ is determined to satisfy the condition $C^{t+1}\left(d^t \mid h^t_{\max}\right)=N\left(d^t \mid \mu^{t+1}_{\max},\Sigma\right)$. This means that $\mu^{t+1}_{\max}$ approaches the observation data d^t.

To summarize the above ideas:

1. Set values for parameters $\beta,\gamma,\varepsilon,K$.
2. Establish initial values for $\Sigma,\mu^1_k,C^1\left(h_k\right)\ (k=1,2,\cdots K)$.
3. Repeat the following whenever data d^t are observed.
   - Find the hypothesis $h^t_{\max}$ with the maximum confidence.
   - Update the confidence $C^{t+1}\left(h_k\right)$ of each hypothesis using formulas (S18) and (S19).
   - Update the likelihood $C^{t+1}\left(d^t \mid h^t_{\max}\right)$ of the hypothesis $h^t_{\max}$ for the observed data d^t using formula (S17).
   - Correct the mean $\mu^{t+1}_{\max}$ of the model for the hypothesis $h^t_{\max}$ using formulas (S21) and (S22) to match the new likelihood $C^{t+1}\left(d^t \mid h^t_{\max}\right)$.

Fitting to simulation data

Fitting to truncated power law distribution (TP)

Here, we describe a method to fit the frequency distribution of duration T observed by simulation to the truncated power law distribution (TP). The method was based on references [10-14]. Specifically, we want to find the minimum value $\hat{T}_{\min}$ and maximum value $\hat{T}_{\max}$ of the data to be fitted to the TP and the exponent $\hat{\eta}$ of the TP model that best fits the data in the range of $\hat{T}_{\min} \le T \le \hat{T}_{\max}$. First, $\hat{T}_{\max}$ is the longest step length in the observation data. Next, we describe the method of calculating $\hat{T}_{\min}$. In the case of a discrete distribution, the TP in the range of $T_{\min} \le T \le T_{\max}$ is expressed by the following formula:

$$p(T;\eta,T_{\min},T_{\max}) = \frac{T^{-\eta}}{\zeta(\eta,T_{\min},T_{\max})}, \quad \zeta(\eta,T_{\min},T_{\max}) = \sum_{i=T_{\min}}^{T_{\max}} i^{-\eta} \qquad \text{(S25)}$$

The CDF of $p(l;\eta,l_{\min},l_{\max})$ is expressed in the following equation.

$$P(T;\eta,T_{\min},T_{\max}) = \frac{\zeta(\mu,T,T_{\max})}{\zeta(\mu,T_{\min},T_{\max})} \qquad \text{(S26)}$$

If the observed data in the range of $T_{\min} \le T \le T_{\max}$ are $\{T_1,T_2,\cdots,T_n\}$, then the log-likelihood of these data for TP is calculated using equation (S25) as follows:

$$L(\eta;T_{\min},T_{\max}) = \sum_{i=1}^{n} \ln p(T_i;T_{\min},T_{\max},\eta) = -n\ln\zeta(\eta,T_{\min},T_{\max}) - \eta\sum_{i=1}^{n}\ln T_i \qquad \text{(S27)}$$

The exponent $\hat{\eta}(T_{\min},T_{\max})$ of the TP model that best fits the data in the range of $T_{\min} \le T \le T_{\max}$ is η, which maximizes $L(\eta;T_{\min},T_{\max})$. Specifically, we varied η from 0.5 to 3.5 in increments of 0.01 to obtain $\hat{\eta}(T_{\min},T_{\max})$, which numerically maximizes equation (S27).

We introduce the Kolmogorov-Smirnov static $D(T_{\min},T_{\max})$ to measure the closeness of the cumulative frequency distribution $S(T;T_{\min},T_{\max})$ created from the data in the range of $T_{\min} \le T \le T_{\max}$ and the theoretical cumulative frequency distribution $P\left(T;\hat{\eta}(T_{\min},T_{\max}),T_{\min},T_{\max}\right)$ represented by equation (S26).

$$D(T_{\min},T_{\max}) = \max_{T_{\min}\le T\le T_{\max}} \left| S(T;T_{\min},T_{\max}) - P\left(T;\hat{\eta}(T_{\min},T_{\max}),T_{\min},T_{\max}\right)\right| \qquad \text{(S28)}$$

If we fix $T_{\max} = \hat{T}_{\max}$, then $D\left(T_{\min},\hat{T}_{\max}\right)$ is a function of $T_{\min}$. We numerically choose $T_{\min}$ out of the observed data, which minimizes $D\left(T_{\min},\hat{T}_{\max}\right)$. That is, $\hat{T}_{\min} = \arg\min_{T_{\min}} D\left(T_{\min},\hat{T}_{\max}\right)$. In the above, $\hat{T}_{\min}$ and $\hat{T}_{\min}$ were obtained. Finally, we find the exponent $\hat{\mu} = \hat{\mu}\left(\hat{T}_{\min},\hat{T}_{\max}\right)$ of the TP model that best fits the data in the range of $\hat{T}_{\min} \le T \le \hat{T}_{\max}$ using the formula (S27).

Fitting to exponential distribution (EP)

In this section, our goal is to find the minimum value $\hat{T}_{\min}$ of the observed data to be fitted to the exponential distribution (EP) model and the exponent $\hat{\lambda}$ of the EP model that best fits the data in the range of $\hat{T}_{\min} \le T$. In the discrete case, the EP in the range of $T_{\min} \le T$ is expressed in the following equation:

$$p\left(T;T_{\min},\lambda\right)=\left(1-e^{-\lambda}\right)e^{-\lambda\left(T-T_{\min}\right)} \qquad \text{(S29)}$$

The CDF of $p\left(T;T_{\min},\lambda\right)$ is expressed as follows:

$$P\left(T;T_{\min},\lambda\right)=e^{-\lambda\left(T-T_{\min}\right)} \qquad \text{(S30)}$$

If the data in the range of $T_{\min} \le T$ is $\left\{T_1,T_2,\cdots,T_m\right\}$, then the log likelihood for these data is expressed as:

$$L\left(\lambda;T_{\min}\right)=\sum_{i=1}^{m}\ln p\left(T_i;T_{\min},\lambda\right)=m\ln\left(1-e^{-\lambda}\right)-\lambda\sum_{i=1}^{m}\left(T_i-T_{\min}\right) \text{ (S31)}$$

The exponent $\hat{\lambda}\left(T_{\min}\right)$ that maximizes $L\left(\lambda;T_{\min}\right)$ is found as a solution to $\frac{\partial L\left(\lambda;T_{\min}\right)}{\partial\lambda}=0$ by the following formula:

$$\hat{\lambda}\left(T_{\min}\right)=\ln\left(\frac{m}{\sum_{i=1}^{m}\left(T_i-T_{\min}\right)}+1\right) \qquad \text{(S32)}$$

$\hat{T}_{\min}$ is calculated from the simulation data and D_{adj} obtained from equation (S30), as in the case of TP. The final value is $\hat{\lambda}=\hat{\lambda}\left(\hat{T}_{\min}\right)$.

Comparison of truncated power law distribution (TP) and exponential distribution (EP)

In this section, we describe a method to determine which of the two distribution models, TP or EP, is more suitable for the simulation data. We use Akaike Information Criteria weights (AICw) for comparison [14]. First, the Akaike Information Criterion (AIC) for data in the range of $T_{\min} \le T \le T_{\max}$ is defined as follows:

$$\begin{aligned} AIC_{TP} &= -2\ln\left(L\left(\hat{\eta};T_{\min},T_{\max}\right)\right)+2 \\ AIC_{EP} &= -2\ln\left(L\left(\hat{\lambda};T_{\min}\right)\right)+2 \end{aligned} \qquad \text{(S33)}$$

The AIC difference Δ is then calculated as follows:

$$
\begin{aligned}
AIC_{\min} &= \min\left(AIC_{TP}, AIC_{EP}\right) \\
\Delta_{TP} &= AIC_{TP} - AIC_{\min} \\
\Delta_{EP} &= AIC_{EP} - AIC_{\min}
\end{aligned}
\tag{S34}
$$

Finally, AICw are calculated as follows:

$$
\begin{aligned}
w_{TP} &= \frac{e^{-\Delta_{TP}/2}}{e^{-\Delta_{TP}/2} + e^{-\Delta_{EP}/2}} \\
w_{EP} &= \frac{e^{-\Delta_{EP}/2}}{e^{-\Delta_{TP}/2} + e^{-\Delta_{EP}/2}}
\end{aligned}
\tag{S35}
$$

First, using the data in the range of $\hat{T}_{\min} \le T \le \hat{T}_{\max}$ calculated during the fitting of the TP, we find the most appropriate exponents $\hat{\eta}$ and $\hat{\lambda}$ for each model.

Next, these exponents are used to calculate and compare AICw. Then, we change $\hat{T}_{\min}$ to the one calculated during the fitting of the EP and compare. If $w_{TP} > w_{EP}$ for both data, the TP is considered to fit the simulated data better. On the other hand, if $w_{TP} < w_{EP}$ for both data, the EP is considered to fit the simulated data better. In case of discrepancies between the results in both data, the following indicators were defined and judged according to reference [10].

$$
\begin{aligned}
D_{adj,TP} &= \frac{\ln N}{\ln n_{TP}} D_{TP} \\
D_{adj,EP} &= \frac{\ln N}{\ln n_{EP}} D_{EP}
\end{aligned}
\tag{S36}
$$

where D_{TP} and D_{EP} are the Kolmogorov-Smirnov static calculated during the model fitting of the TP and EP, respectively. N is the total number of observed data points, and n_{TP} and n_{EP} are the number of observed data points used in each model fitting. In other words, the index considers a model that can fit more observational data to be a better model. In the case of $D_{adj,TP} < D_{adj,EP}$, the TP is considered to fit the simulation data better. Conversely, when $D_{adj,TP} > D_{adj,EP}$, EP is considered to be a better fit to the simulation data. If the optimal model was judged to be a TP, it is considered to be a Lévy walk if $1 < \hat{\eta} \le 3$ was satisfied.

REFERENCES

1. M. Hattori and M. Oaksford. Adaptive non-interventional heuristics for covariation detection in causal induction: model comparison and rational analysis. Cogn Sci. 2007;31: 765–814. pmid:21635317

2. S. Waxman and T. Kosowski. Nouns mark category relations: Toddlers' and preschoolers' word-learning biases. Child Dev. 1990;61: 1461–1473. https://doi.org/10.1111/j.1467-8624.1990.tb02875.x. pmid:2245738

3. M. J. Buehner, P. W. Cheng, D. Clifford. From covariation to causation: A test of the assumption of causal power. J Exper Psychol Learn Mem Cogn. 2003;29: 1119–1140. http://dx.doi.org/10.1037/0278-7393.29.6.1119.

4. K. Lober, D. R. Shanks. Is causal induction based on causal power? Critique of Cheng (1997). Psychol Rev. 2000;107: 95–212. http://dx.doi.org/10.1037/0033-295X.107.1.195.

5. P. A. White. Making causal judgments from the proportion of confirming instances: the pCI rule. J Exper Psychol Learn Mem Cogn. 2003;29: 710–727. http://dx.doi.org/10.1037/0278-7393.29.4.710.

6. J. Evans, S. Handley, D. Over. Conditionals and conditional probability. J Exp Psychol Learn Mem Cogn. 2003;29: 321–335. pmid:12696819

7. S. Shinohara, N. Manome, K. Suzuki, U. I. Chung, T. Takahashi, P. Y. Gunji, et al. Extended Bayesian inference incorporating symmetry bias. Biosystems. 2020;190: 104104. https://doi.org/10.1016/j.biosystems.2020.104104.

8. S. Shinohara, N. Manome, K. Suzuki, U. I. Chung, T. Takahashi, H. Okamoto, et al. A new method of Bayesian causal inference in non-stationary environments (2020). PLoS ONE 15(5): e0233559. https://doi.org/10.1371/journal.pone.0233559

9. Y. P. Gunji, S. Shinohara, T. Haruna, and V. Basios. Inverse Bayesian inference as a key of consciousness featuring a macroscopic quantum logical structure. Biosystems. 2016;152: 44–65. pmid:28041845

10. N. E. Humphries et al. Foraging success of biological Lévy flights recorded in situ. Proceedings of the National Academy of Sciences of the United States of America vol. 109,19 (2012): 7169-74. doi:10.1073/pnas.1121201109

11. V. A. Jansen, A. Mashanova, and S. Petrovskii. Comment on "Lévy walks evolve through interaction between movement and environmental complexity". Science. 2012 Feb 24;335(6071):918; author reply 918. doi: 10.1126/science.1215747. PMID: 22362991.

12. E. P. White, B. J. Enquist, J. L. Green. On estimating the exponent of power-law frequency distributions. Ecology. 2008 Apr;89(4):905-12. doi: 10.1890/07-1288.1. Erratum in: Ecology. 2008 Oct;89(10):2971. PMID: 18481513.

13. A. Clauset, C. R. Shalizi, and M. E. J. Newman, Power-Law Distributions in Empirical Data SIAM Rev., 51(4), 661–703

14. A. M. Edwards, R. A. Phillips, N. W. Watkins, M. P. Freeman, E. J. Murphy, V. Afanasyev, S. V. Buldyrev, M. G. da Luz, E. P. Raposo, H. E. Stanley, and G. M. Viswanathan. Revisiting Lévy flight search patterns of wandering albatrosses, bumblebees and deer. Nature. 2007 Oct 25;449(7165):1044-8. doi: 10.1038/nature06199. PMID: 17960243.