physical sciences forum

MDPI

Proceeding Paper

Kangaroos in Cambridge †

Romke Bontekoe [1,*] and Barrie J. Stokes [2]

[1] Bontekoe Research, 1052 WJ Amsterdam, The Netherlands
[2] New Lambton Heights, Newcastle, NSW 2305, Australia
* Correspondence: romke@bontekoe.nl
† Presented at the 41st International Workshop on Bayesian Inference and Maximum Entropy Methods in Science and Engineering, Paris, France, 18–22 July 2022.

Abstract: In this tutorial paper the Gull–Skilling kangaroo problem is revisited. The problem is used as an example of solving an under-determined system by variational principles, the maximum entropy principle (MEP), and Information Geometry. The relationship between correlation and information is demonstrated. The Kullback–Leibler divergence of two discrete probability distributions is shown to fail as a distance measure. However, an analogy with rigid body rotations in classical mechanics is motivated. A table of proper "geodesic" distances between probability distributions is presented. With this paper the authors pay tribute to their late friend David Blower.

Keywords: kangaroo problem; variational principle; maximum entropy principle; information geometry; Kullback–Leibler divergence; metric tensor; Bhattacharyya angle; Wolfram Mathematica

1. Introduction

On my (RB) first meeting with Dr. John Skilling and Dr. Steve Gull in Cambridge in 1987, I was posed the following problem [1–3]:

> In Australia, 3/4 of the kangaroos are right-handed and 1/3 have blue eyes. Can you construct the 2 × 2 probability table?

Having no clue about the use of their shorter forelegs, let alone any handedness, nor of the colour of their eyes, I assumed that:

1. a kangaroo is right-handed or left-handed; and
2. a kangaroo has blue eyes or green eyes.

This means that there are four distinct possibilities: right-handed with blue eyes, right-handed with green eyes, left-handed with blue eyes, and left-handed with green eyes. The statement space is of dimension 2 × 2 and has 4 cells, and a bare probability table looks like Table 1, showing the two given marginal values and the sum.

Table 1. Probability table: version 1.

Probability Table			
	Blue eyes	Green eyes	
Right-handed			3/4
Left-handed			
	1/3		1

The two other marginal values result from normalizing the sum of the joint probabilities. Filling in the table a little more, we obtain Table 2. The notation Q_i for probabilities originates with David Blower, who avoids the overused P-symbol. In this paper we follow Blower's notation closely [4].

check for updates

Citation: Bontekoe, R.; Stokes, B.J. Kangaroos in Cambridge. *Phys. Sci. Forum* **2022**, *1*, 22. https://doi.org/10.3390/psf2022005022

Academic Editors: Frédéric Barbaresco, Ali Mohammad-Djafari, Frank Nielsen and Martino Trassinelli

Published: 8 December 2022

Publisher's Note: MDPI stays neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Table 2. Probability table: version 2.

Probability Table			
	Blue eyes	Green eyes	
Right-handed	Q_1	Q_2	$\boxed{3/4}$
Left-handed	Q_3	Q_4	$1/4$
	$\boxed{1/3}$	$2/3$	$\boxed{1}$

I thought about this problem for a short while and filled in the table by multiplying the row and column marginal values, as in Table 3.

Table 3. Probability table, version 3.

Probability Table			
	Blue eyes	Green eyes	
Right-handed	$1/4$	$1/2$	$\boxed{3/4}$
Left-handed	$1/12$	$1/6$	$1/4$
	$\boxed{1/3}$	$2/3$	$\boxed{1}$

However, then I was presented with the following set of equations

$$\begin{aligned} Q_1 + Q_2 &= 3/4 \\ Q_1 + Q_3 &= 1/3 \\ Q_1 + Q_2 + Q_3 + Q_4 &= 1. \end{aligned} \tag{1}$$

There are only three equations in four unknowns, leaving any other (consistent) equations relating to the Q_i redundant. This is an under-determined system. In my proposed solution, I must have used a fourth equation. So, where did this fourth equation come from? My answer was that I assumed that handedness and eye colour are independent, and thus the marginal probabilities could be multiplied. "*Aah*", they said, "you have applied the Maximum Entropy Principle!"

Jaynes discussed and extended the kangaroo problem in the Fourth Maximum Entropy Workshop in 1984 [3].

This under-determined system has one free variable. Choosing Q_1 as the free variable, the equations reduce to

$$\begin{aligned} Q_2 &= 3/4 - Q_1 \\ Q_3 &= 1/3 - Q_1 \\ Q_4 &= -1/12 + Q_1. \end{aligned} \tag{2}$$

A symbolic solution can be obtained by using Wolfram Mathematica's `Reduce[]` function [5] as shown in Figure 1.

```
Reduce[{q1 + q2 == 3/4, q1 + q3 == 1/3, q1 + q2 + q3 + q4 == 1,
        0 <= q1 <= 1, 0 <= q2 <= 1, 0 <= q3 <= 1, 0 <= q4 <= 1},
  {q1, q2, q3, q4} ]
```

Figure 1. Wolfram Mathematica code for solving the under-determined problem (2).

In this code snippet, the three equations can be recognized as well as the positivity condition. The solution is

$$1/12 \leq Q_1 \leq 1/3. \tag{3}$$

With this solution the probability table can be filled in as in Table 4.

Table 4. Probability table, version 4.

Probability Table			
	Blue eyes	Green eyes	
Right-handed	$1/12 \leq Q_1 \leq 1/3$	$3/4 - Q_1$	$\boxed{3/4}$
Left-handed	$1/3 - Q_1$	$-1/12 + Q_1$	$1/4$
	$\boxed{1/3}$	$2/3$	$\boxed{1}$

Figure 2 shows a range of solutions to this problem. This figure illustrates the correlation and anti-correlation between the various Q_i-s. Since $Q_1$and Q_2 have to maintain their sum as $3/4$, they must be anti-correlated. Therefore the coloured lines cross each other between Q_1 and Q_3. Similarly, Q_1 is anti-correlated with Q_3. Therefore, Q_2 and Q_3 have to be correlated, and the coloured lines between them do not cross. Finally, Q_1 is correlated with Q_4, which can be seen from the repeated Q_1-axis at the right.

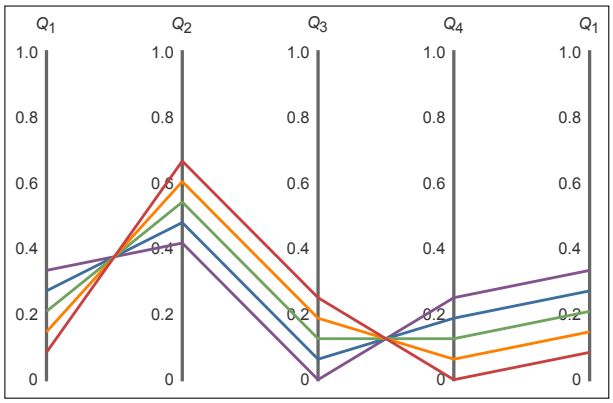

Figure 2. Parallel-axis plot of the Q_1, Q_2, Q_3, and Q_4, for Q_1 between $1/12$ (red) and $1/3$ (purple) in equidistant steps of $1/16$. For clarity, the Q_1 axis is repeated on the right.

2. Variational Principles

A possible solution for an under-determined problem can be found by adopting a *variational principle*. This is a *function* of the joint probabilities to be optimized (maximized or minimized) under some constraints, whose free parameters correspond to the missing equations. Sivia considers four variational functions: the entropy, the sum of squares, the sum of logarithms, and the sum of square roots, as shown in Table 5 [1].

Table 5. Sivia's four variational functions: entropy, sum of squares, sum of logarithms, sum of square roots.

Variational Functions	
	Function
Maximum entropy	$-\sum_{i=1}^{n} Q_i \log Q_i$
Least squares	$\sum_{i=1}^{n} Q_i^2$
Maximum logarithms	$\sum_{i=1}^{n} \log Q_i$
Maximum square roots	$\sum_{i=1}^{n} \sqrt{Q_i}$

In the case of the Least Squares variational function we have

$$\begin{aligned} f(Q) &= Q_1^2 + Q_2^2 + Q_3^2 + Q_4^2 \\ &= Q_1^2 + (3/4 - Q_1)^2 + (1/3 - Q_1)^2 + (-1/12 + Q_1)^2 \\ &= 4\,Q_1^2 - 7/3\,Q_1 + 49/72. \end{aligned} \tag{4}$$

This is a quadratic function and has a unique minimum at

$$Q_1 = 7/24, \tag{5}$$

which yields the exact solution of $M_{\text{VP,LeastSq}}$

$$Q_i = (7/24, 11/24, 1/24, 5/24). \tag{6}$$

For the Maximum Entropy, the variational function

$$f(Q) = -Q_1 \log Q_1 - Q_2 \log Q_2 - Q_3 \log Q_3 - Q_4 \log Q_4 \tag{7}$$

has to be maximized, subject to the constraints. This function has a unique maximum at

$$Q_1 = 1/4, \tag{8}$$

which yields the exact solution of $M_{\text{VP,MaxEnt}}$

$$Q_i = (1/4, 1/2, 1/12, 1/6). \tag{9}$$

The solutions for Q_1 for the Maximum logarithms and Maximum square roots variational equations only can be obtained via numerical optimization. For each solution Q_1, the other three Q_i values follow directly from (2). The Variational Principle solutions are tabulated in Table 6 and visualized in Figure 3.

Table 6. The Variational Principle solutions.

Variational Functions		
Model	Function	Q_i
$M_{\text{VP,MaxEnt}}$	$-\sum_{i=1}^{n} Q_i \log Q_i$	$(0.25, 0.50, 0.08, 0.17)$
$M_{\text{VP,LeastSq}}$	$\sum_{i=1}^{n} Q_i^2$	$(0.29, 0.46, 0.04, 0.21)$
$M_{\text{VP,MaxLog}}$	$\sum_{i=1}^{n} \log Q_i$	$(0.23, 0.52, 0.11, 0.14)$
$M_{\text{VP,MaxSqrt}}$	$\sum_{i=1}^{n} \sqrt{Q_i}$	$(0.24, 0.51, 0.10, 0.15)$

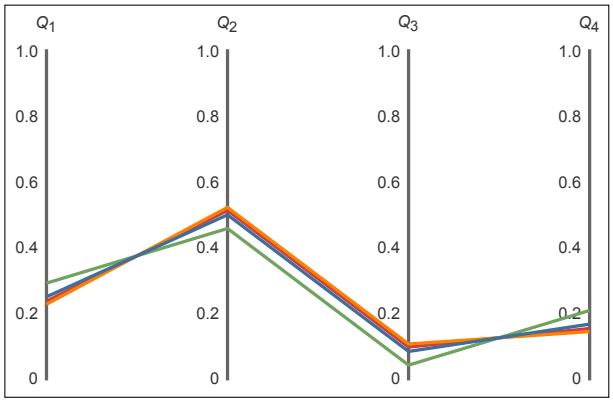

Figure 3. Parallel axis plot of the Variational Principle solutions: $M_{\text{VP,MaxEnt}}$ is blue, $M_{\text{VP,LeastSq}}$ is green, $M_{\text{VP,MaxLogs}}$ is orange, and $M_{\text{VP,MaxSqrt}}$ is red.

However, given these four different solutions to the kangaroo problem, we need a rationale for choosing one of them. Which one is 'best'? Sivia states that barring some evidence about a gene-linkage between handedness and eye colour for kangaroos, the MaxEnt model is preferred because this model provides the only uncorrelated assignment of the Q_i. This is shown in Section 4.

3. State Space and Constraint Functions

In the kangaroo problem, we have two traits: handedness and eye colour. Each trait has a set of features; for the handedness they are "right-handed" and "left-handed"; for the eye colour "blue" and "green". Mixtures of features are not allowed. Therefore, for every trait, one, and only one, feature applies; the features are mutually exclusive.

More abstractly, the features can be represented as *statements*. The combined features from different traits form *joint statements*. The joint statements define a *state space* of dimension $n = 4$. The n cells uniquely number the joint statements. Table 7 shows the general setup.

Table 7. The n cells of the state space uniquely number the joint statements.

State Space Table			
	Blue eyes	Green eyes	
Right-handed	$(X = x_1)$	$(X = x_2)$	$\boxed{3/4}$
Left-handed	$(X = x_3)$	$(X = x_4)$	$1/4$
	$\boxed{1/3}$	$2/3$	$\boxed{1}$

Any joint statement about a kangaroo can be placed in one and only one cell of the state space. For example, a left-handed and blue-eyed kangaroo is uniquely defined by the joint statement $(X = x_3)$. In this notation, the X denotes the two traits, and the x_3 specifies the features in cell 3. The state space is congruent to the probability table of Table 1, but it has a different role. The joint statements, $(X = x_i)$, are logical statements which can be either True or False.

A constraint function is defined over the state space, as shown in Table 8. The function F assigns a Boolean value to each joint statement and returns a vector of values ([4], Ch. 21)

$$(F(X = x_1), F(X = x_2), F(X = x_3), F(X = x_4)). \tag{10}$$

The constraint function vector specifies the operation of a constraint.

Table 8. The constraint function $F(X = x_i)$ is a function defined on the space of joint statements.

State Space Table			
	Blue eyes	Green eyes	
Right-handed	$F(X = x_1)$	$F(X = x_2)$	$\boxed{3/4}$
Left-handed	$F(X = x_3)$	$F(X = x_4)$	$1/4$
	$\boxed{1/3}$	$2/3$	$\boxed{1}$

The constraint function F_1 for our first constraint, "In Australia $3/4$ of the kangaroos are right-handed ...," is shown in Table 9. Writing out the constraint function vector for F_1, we have

$$(F_1(X = x_1), F_1(X = x_2), F_1(X = x_3), F_1(X = x_4)) = (1, 1, 0, 0). \tag{11}$$

The corresponding constraint function vector for the left-handed kangaroos is its complement, $(0, 0, 1, 1)$.

Table 9. The constraint function F_1 for the constraint "$3/4$ of the kangaroos are right-handed."

State Space Table			
	Blue eyes	Green eyes	
Right-handed	$F_1(X = x_1) = 1$	$F_1(X = x_2) = 1$	$\boxed{3/4}$
Left-handed	$F_1(X = x_3) = 0$	$F_1(X = x_4) = 0$	$1/4$
	$\boxed{1/3}$	$2/3$	$\boxed{1}$

The constraint function F_2 for the second constraint, "... and $1/3$ have blue eyes," is shown in Table 10. Writing out the constraint function vector F_2, we obtain

$$(F_2(X = x_1), F_2(X = x_2), F_2(X = x_3), F_2(X = x_4)) = (1, 0, 1, 0). \tag{12}$$

The constraint function vector for the blue-eyed kangaroos is $(1, 0, 1, 0)$, and $(0, 1, 0, 1)$ for the green-eyed 'roos.

Table 10. The constraint function vector for the second constraint ($1/3$ of the kangaroos have blue eyes).

State Space Table			
	Blue eyes	Green eyes	
Right-handed	$F_2(X = x_1) = 1$	$F_2(X = x_2) = 0$	$\boxed{3/4}$
Left-handed	$F_2(X = x_3) = 1$	$F_2(X = x_4) = 0$	$1/4$
	$\boxed{1/3}$	$2/3$	$\boxed{1}$

The probability distribution is normalized, which means that the sum of all joint probabilities is unity. This is also a constraint. The overall normalization is a universal constraint function vector

$$(F_0(X = x_1), F_0(X = x_2), F_0(X = x_3), F_0(X = x_4)) = (1, 1, 1, 1). \tag{13}$$

This whole business of creating constraint function vectors for assigning probabilities may seem overly elaborate but conceptually, and operationally, we need a way to connect a *statement* $(X = x_i)$ with a *numerical value*. Technically, F is an *operator* that accepts a joint statement as its variable and returns a Boolean value. Furthermore, the constraint function vectors F_j become the *basis vectors* $\mathbf{e}_j$ in the vector space of the information geometry in Section 6.

4. Correlation, Covariance, and Entropy

What do correlation and covariance actually mean, and what is the difference? Sometimes the two terms are used interchangeably.

We all have an intuitive interpretation. For instance, people's heights and weights are correlated, which means that generally, tall persons weigh more than short ones. The two variables vary together; they are *co-varying*. However, this does not necessarily reflect a *causal* relationship. Gaining weight does not automatically imply becoming taller, as we all know.

4.1. Expectation

Suppose that a function $V(X = x_i)$ is defined over the state space and returns a numerical value for each joint statement. The *expectation* of V is

$$\langle V \rangle = \sum_{i=1}^{n} V(X = x_i)\, Q_i. \tag{14}$$

The sum is over all $V(X = x_i)$ values in the state space, whereas the Q_i are from the probability table. The expectation value, $\langle V \rangle$, is a numerical quantity.

With this definition, let's compute the expectation for "right-handedness". The constraint function vector for right-handedness, $F_1 = (1, 1, 0, 0)$, acts as the quantity V

$$\begin{aligned}\langle F_1 \rangle &= \sum_{i=1}^{n} F_1(X = x_i)\, Q_i \\ &= F_1(X = x_1)\, Q_1 + F_1(X = x_2)\, Q_1 + F_1(X = x_3)\, Q_1 + F_1(X = x_4)\, Q_1 \\ &= 1\, Q_1 + 1\, Q_2 + 0\, Q_3 + 0\, Q_4 \\ &= Q_1 + Q_2 \\ &= 3/4.\end{aligned} \tag{15}$$

In the last step, we have used the information given in Table 2. The expectation for right-handedness thus equals its marginal value.

Similarly for "blue eyes", with $F_2 = (1, 0, 1, 0)$

$$\begin{aligned}\langle F_2 \rangle &= \sum_{i=1}^{n} F_2(X = x_i)\, Q_i \\ &= 1\, Q_1 + 0\, Q_2 + 1\, Q_3 + 0\, Q_4 \\ &= Q_1 + Q_3 \\ &= 1/3.\end{aligned} \tag{16}$$

Furthermore, the expectation value for blue eyes again equals its marginal value.

4.2. Variance

The variance of the $V(X = x_i)$ values is defined as

$$\begin{aligned}\mathrm{var}(V) &= \sum_{i=1}^{n} (V(X = x_i) - \langle V \rangle)^2\, Q_i \\ &= \left\langle (V(X = x_i) - \langle V \rangle)^2 \right\rangle.\end{aligned} \tag{17}$$

Notice that there are two nested sets of brackets $\langle . \rangle$ involved. The $\langle V \rangle$ is defined by (14).

By expanding the square, this can be rewritten as

$$\begin{aligned}\mathrm{var}(V) &= \left\langle (V(X = x_i) - \langle V \rangle)^2 \right\rangle \\ &= \left\langle V(X = x_i)^2 - 2\, V(X = x_i)\, \langle V \rangle + \langle V \rangle^2 \right\rangle \\ &= \left\langle V(X = x_i)^2 \right\rangle - 2\, \langle V(X = x_i)\, \langle V \rangle \rangle + \left\langle \langle V \rangle^2 \right\rangle \\ &= \left\langle V(X = x_i)^2 \right\rangle - 2\, \langle V(X = x_i) \rangle\, \langle V \rangle + \langle V \rangle^2 \\ &= \left\langle V(X = x_i)^2 \right\rangle - 2\, \langle V \rangle\, \langle V \rangle + \langle V \rangle^2 \\ &= \sum_{i=1}^{n} V(X = x_i)^2\, Q_i - \langle V \rangle^2.\end{aligned} \tag{18}$$

We have used the properties $\langle \langle V \rangle \rangle = \langle V \rangle$ and $\left\langle \langle V \rangle^2 \right\rangle = \langle V \rangle^2$ in the above derivation, because $\langle V \rangle$ is a constant.

So what is the variance of "right-handedness"? Taking $V = F_1$, we obtain

$$\begin{aligned} \mathrm{var}(F_1) &= \sum_{i=1}^{n} F_1(X = x_i)^2 \, Q_i - \langle F_1 \rangle^2 \\ &= 1^2 \, Q_1 + 1^2 \, Q_2 + 0^2 \, Q_3 + 0^2 \, Q_4 - \langle F_1 \rangle^2 \\ &= Q_1 + Q_2 - \langle F_1 \rangle^2 \\ &= 3/4 - (3/4)^2 \\ &= 3/16. \end{aligned} \tag{19}$$

The variance of “blue eyes” is

$$\begin{aligned} \mathrm{var}(F_2) &= \sum_{i=1}^{n} F_2(X = x_i)^2 \, Q_i - \langle F_2 \rangle^2 \\ &= 1^2 \, Q_1 + 0^2 \, Q_2 + 1^2 \, Q_3 + 0^2 \, Q_4 - \langle F_2 \rangle^2 \\ &= Q_1 + Q_3 - \langle F_2 \rangle^2 \\ &= 1/3 - (1/3)^2 \\ &= 2/9. \end{aligned} \tag{20}$$

We conclude that both variances are independent of Q_1.

4.3. Covariance

The covariance between two variables $V(X = x_i)$ and $W(X = x_i)$ is defined by

$$\mathrm{cov}(\mathrm{V}, \mathrm{W}) = \langle (\mathrm{V}(\mathrm{X} = \mathrm{x_i}) - \langle \mathrm{V} \rangle) \, (\mathrm{W}(\mathrm{X} = \mathrm{x_i}) - \langle \mathrm{W} \rangle) \rangle. \tag{21}$$

By a similar expansion as above, the product can be written as

$$\begin{aligned} \mathrm{cov}(V, W) &= \langle V(X = x_i) \, W(X = x_i) - V(X = x_i) \, \langle W \rangle - W(X = x_i) \, \langle V \rangle + \langle V \rangle \, \langle W \rangle \rangle \\ &= \langle V(X = x_i) \, W(X = x_i) \rangle - \langle V(X = x_i) \rangle \, \langle W \rangle - \langle W(X = x_i) \rangle \, \langle V \rangle + \langle V \rangle \, \langle W \rangle \\ &= \langle V(X = x_i) \, W(X = x_i) \rangle - \langle V \rangle \, \langle W \rangle \\ &= \sum_{i=1}^{n} V(X = x_i) \, W(X = x_i) \, Q_i - \langle V \rangle \, \langle W \rangle. \end{aligned} \tag{22}$$

What does this give for the $\mathrm{cov}(F_1, F_2)$? Expanding the sum and substituting the constraint function vectors $F_1 = (1, 1, 0, 0)$ and $F_2 = (1, 0, 1, 0)$, we obtain

$$\begin{aligned} \mathrm{cov}(F_1, F_2) &= 1 * 1 \, Q_1 + 1 * 0 \, Q_2 + 0 * 1 \, Q_3 + 0 * 0 \, Q_4 - \langle F_1 \rangle \, \langle F_2 \rangle \\ &= Q_1 - 3/4 * 1/3 \\ &= Q_1 - 1/4. \end{aligned} \tag{23}$$

We find that $\mathrm{cov}(F_1, F_2)$ *does* depend on Q_1.

The variances and covariances can be combined in the *variance-covariance* matrix, which is defined by

$$\begin{aligned} \Sigma(F_1, F_2) &= \begin{pmatrix} \mathrm{var}(F_1) & \mathrm{cov}(F_1, F_2) \\ \mathrm{cov}(F_1, F_2) & \mathrm{var}(F_2) \end{pmatrix} \\ &= \begin{pmatrix} 3/16 & Q_1 - 1/4 \\ Q_1 - 1/4 & 2/9 \end{pmatrix}. \end{aligned} \tag{24}$$

The variance-covariance matrix is related to the metric tensor g from information geometry in Section 6.

4.4. Correlation

The correlation coefficient is a single value derived from the variance and covariance values. It is defined as

$$\rho(V,W) = \frac{\operatorname{cov}(V,W)}{\sqrt{\operatorname{var}(V)\operatorname{var}(W)}}. \tag{25}$$

Therefore the correlation between the eye colour and the handedness of the kangaroos is

$$\begin{aligned} \rho(F_1,F_2) &= \frac{\operatorname{cov}(F_1,F_2)}{\sqrt{\operatorname{var}(F_1)\operatorname{var}(F_2)}} \\ &= \frac{Q_1 - 1/4}{\sqrt{3/16 * 2/9}} \\ &= 2\sqrt{6}\,(Q_1 - 1/4). \end{aligned} \tag{26}$$

This finally confirms that indeed, the MaxEnt solution, with $Q_1 = 1/4$, has zero correlation. We agree with Sivia that the other variational functions yield a positive or negative correlation between handedness and eye colour. (Notice that our correlation coefficients have the opposite sign, because Sivia correlates the left-handedness with blue eyes [1].) Table 11 shows the model solutions Q_i and the corresponding correlation values.

Table 11. The numerical details for the variational principle solutions.

Variational Functions				
Model	Function	Q_i	$\rho(F_1,F_2)$	$H(Q)$ (bits)
$M_{\text{VP,MaxEnt}}$	$-\sum_{i=1}^{n} Q_i \log Q_i$	$(0.25, 0.50, 0.08, 0.17)$	0.00	1.730
$M_{\text{VP,LeastSq}}$	$\sum_{i=1}^{n} Q_i^2$	$(0.29, 0.46, 0.04, 0.21)$	0.20	1.697
$M_{\text{VP,MaxLog}}$	$\sum_{i=1}^{n} \log Q_i$	$(0.23, 0.52, 0.11, 0.14)$	-0.11	1.721
$M_{\text{VP,MaxSqrt}}$	$\sum_{i=1}^{n} \sqrt{Q_i}$	$(0.24, 0.51, 0.10, 0.15)$	-0.07	1.727

One may have gotten the impression that the constraint function values are always 0 or 1, but these are specific for the problem treated in this paper. In general, a constraint function may yield any numerical value. The construction of a constraint function can be intricate; see, for example, Blower ([4], p. 63).

4.5. Entropy

The *information entropy* is a measure of the amount of *missing information* in a probability distribution. The information entropy $H(Q)$ of a discrete probability distribution is

$$H(Q) = -\sum_{i=1}^{n} Q_i \log Q_i. \tag{27}$$

Of all possible probability distributions, the discrete uniform distribution has the maximum missing information. Thus for $n = 4$, we have $q = (1/4, 1/4, 1/4, 1/4)$ with

$$\begin{aligned} H(q) &= -\sum_{i=1}^{n} \frac{1}{4} \log \frac{1}{4} \\ &= \log 4 \\ &\approx 1.39. \end{aligned} \tag{28}$$

When the natural logarithm $\log_e$ is used, the units of entropy are *nats*. However, the entropy can also be defined in terms of the more familiar *bits* when $\log_2$ is used. The conversion of $H(Q)$ to bits by multiplying by $\log_2 e \approx 1.44$ gives

$$\begin{aligned} H(q) &= \log 4 * \log_2 e \\ &= 1.39 * 1.44 \\ &= 2 \text{ bits.} \end{aligned} \tag{29}$$

Maximum missing information of two bits exactly describes our minimum state of knowledge in a 2×2 state space with four equally probable states. We need one bit to choose a column and another bit to choose a row. Combined, we have fully specified one of four equally probable states or cells in the state space.

Absolute certainty is described by zero bits of missing information. This is attained when one $Q_i = 1$ and all other $Q_{j \neq i} = 0$. Then our state of knowledge is fully specified and there is no missing information. For example, a "certain distribution" is $p = (0, 0, 1, 0)$, for which the entropy is

$$H(p) = 0. \tag{30}$$

Here we have used

$$\lim_{x \to 0^+} \; x \log x = 0, \tag{31}$$

and $\log 1 = 0$.

The table in Table 11 shows the values for $H(Q)$, in bits, in the last column. Although all models have an entropy that is smaller than two bits, the numerical values of the entropy are not easily assessed intuitively. Jaynes gives an excellent explanation to guide one's intuition ([6], Ch. 11.3).

Suppose we were first told about the kangaroos' handedness, namely $p_1 = 3/4$ versus $p_2 = 1/4$. The information entropy of this binary case is

$$\begin{aligned} H_2(p_1, p_2) &= -\frac{3}{4} \log_2 \frac{3}{4} - \frac{1}{4} \log_2 \frac{1}{4} \\ &= 0.81. \end{aligned} \tag{32}$$

Next, we learn that the first alternative consists of two possibilities, namely blue and green eyes, with $p_1 = q_1 + q_2$, where $q_1 = 1/4$ and $q_2 = 1/2$. The information entropy for the ternary case becomes

$$\begin{aligned} H_3(q_1, q_2, p_2) &= H_2(p_1, p_2) + p_1 * H_2\left(\frac{q_1}{p_1}, \frac{q_2}{p_1}\right) \\ &= H_2(p_1, p_2) + \frac{3}{4}\left(-\frac{1}{3} \log_2 \frac{1}{3} - \frac{2}{3} \log_2 \frac{2}{3}\right) \\ &= 0.81 + 0.69 \\ &= 1.50. \end{aligned} \tag{33}$$

Finally, the second alternative also consists of two possibilities, namely $p_2 = q_3 + q_4$, with $q_3 = 1/12$ and $q_4 = 1/6$. The information entropy becomes

$$\begin{aligned} H_4(q_1, q_2, q_3, q_4) &= H_3(q_1, q_2, p_2) + p_2 * H_2\left(\frac{q_3}{p_2}, \frac{q_4}{p_2}\right) \\ &= H_3(q_1, q_2, p_2) + \frac{1}{4}\left(-\frac{1}{3} \log_2 \frac{1}{3} - \frac{2}{3} \log_2 \frac{2}{3}\right) \\ &= 1.50 + 0.23 \\ &= 1.73. \end{aligned} \tag{34}$$

We recognize the same value as for $M_{\mathrm{VP,MaxEnt}}$ in Table 11. In this example, the state space is gradually expanded and, as the number of cells increases one's ambivalence also increases, which is reflected in an increase in the entropy. The example also shows that the subsequent H_n are additive. Notice that the above partitioning of the p_1 and p_2 is proportional to the blue- and green-eyed kangaroos ratio.

For a given set of constraints, of all possible models, the maximum entropy solution has the highest information entropy ([4], Ch. 24.2), which is confirmed in Table 11. This means that the $M_{\mathrm{VP,MaxEnt}}$ solution has the most missing information. Consequently, in one way or another, some extra information was introduced by the other variational functions. From the example above, one may surmise that the additional information originates from a different partitioning of the p_1 and p_2 into the q_i-s.

This extra information also shows up as non-zero correlations; the higher the absolute value of the correlation, the lower the information entropy. Therefore, *correlation induces information*, reducing the amount of *missing* information.

5. Maximum Entropy Principle

Although we have already obtained several solutions to the kangaroo problem by the optimization of various variational functions, the procedure may be seen as ad hoc. The Maximum Entropy Principle (MEP) is a versatile problem-solving method based on the work of Shannon and Jaynes ([6], Ch. 11; [3,7]). The MEP is a method with highly desirable features for making numerical assignments, and, most importantly, all conceivable legitimate numerical assignments may be made, and are made, via the MEP. The book by Blower [4] is entirely devoted to the MEP.

5.1. Interactions

Blower defines the *interaction* between two (or more) constraints as the product of their constraint function vectors. Here we have two constraints, which can have only one interaction, namely between "right-handed" and "blue eyes". In problems with more dimensions, higher-dimensional interactions can be defined by the product of three or more constraint function vectors.

The interaction vector is the element-wise product of the relevant constraint function vectors

$$\begin{aligned} F_3(X = x_i) &= F_1(X = x_i) * F_2(X = x_i) \\ &= (1,1,0,0) * (1,0,1,0) \\ &= (1,0,0,0). \end{aligned} \tag{35}$$

From Table 12, we see how $F_3(X = x_i)$ selects the interaction between "right-handed" and "blue eyes". This interaction singles out the $(X = x_1)$ statement in the state space and, consequently, the Q_1 joint probability. Keeping our terminology simple, this interaction vector is also called a constraint function vector.

Table 12. $F_3(X = x_i)$ selects the interaction between "right-handed" and "blue eyes".

State Space Table			
	Blue eyes	Green eyes	
Right-handed	$F_3(X = x_1) = 1$	$F_3(X = x_2) = 0$	$\boxed{3/4}$
Left-handed	$F_3(X = x_3) = 0$	$F_3(X = x_4) = 0$	$1/4$
	$\boxed{1/3}$	$2/3$	$\boxed{1}$

There are now three constraint function vectors

$$\begin{aligned} F_1(X = x_i) &= (1,1,0,0) \\ F_2(X = x_i) &= (1,0,1,0) \\ F_3(X = x_i) &= (1,0,0,0), \end{aligned} \tag{36}$$

which can be combined to form the *constraint function matrix*

$$M = \begin{pmatrix} 1 & 1 & 0 & 0 \\ 1 & 0 & 1 & 0 \\ 1 & 0 & 0 & 0 \end{pmatrix}. \tag{37}$$

The constraint function matrix has dimensions $m \times n$. As in Section 4, the expectation value of the interaction $\langle F_3 \rangle$ is

$$\begin{aligned} \langle F_3 \rangle &= \sum_{i=1}^{n} F_3(X = x_i)\, Q_i \\ &= 1\, Q_1 + 0\, Q_2 + 0\, Q_3 + 0\, Q_4 \\ &= Q_1. \end{aligned} \tag{38}$$

The three expectation values are combined to form the *constraint function average vector*

$$\begin{pmatrix} \langle F_1 \rangle \\ \langle F_2 \rangle \\ \langle F_3 \rangle \end{pmatrix} = \begin{pmatrix} 3/4 \\ 1/3 \\ Q_1 \end{pmatrix}. \tag{39}$$

The constraint function average vector $(\langle F_1 \rangle, \langle F_2 \rangle, \langle F_3 \rangle)$ is related to the contravariant coordinates (η_1, η_2, η_3) in information geometry in Section 6.

In an under-determined problem, the number of constraints (primary and interaction) is $m < n - 1$. In our case $m = 3$, therefore combined with the normalization of the probability distribution, we have a linear system of four equations with four unknowns. However, in this paper, we take a general approach as if we had an under-determined system with $m < n - 1$.

Returning to our kangaroo problem, from the MEP perspective, we will obtain four models $M_{\mathrm{MEP},k}$ defined by their constraint function averages $(\langle F_1 \rangle, \langle F_2 \rangle, \langle F_3 \rangle)$. The set-up of the problem fixes values of $\langle F_1 \rangle$ and $\langle F_2 \rangle$, whereas the third value, $\langle F_3 \rangle$, is taken as the Q_1-s from the $M_{\mathrm{VP},k}$ model solutions, as shown in Table 11.

5.2. The Maximum Entropy Principle

The MEP involves a constrained optimization problem utilizing the method of Lagrange multipliers. According to Jaynes, the MEP provides the most conservative, non-committal distribution where the missing information is as 'spread-out' as possible, yet which accords with no other constraints than those explicitly taken into account.

The MEP solution in its canonical form is ([4], p. 50)

$$Q_i = \frac{\exp\left(\sum_{j=1}^{m} \lambda_j\, F_j(X = x_i)\right)}{Z(\lambda)}. \tag{40}$$

Here Q_i is the probability for the joint statement $(X = x_i)$. The $F_j(X = x_i)$ is the j-th constraint function operator acting on the i-th joint statement. The λ_j are the Lagrange multipliers, each corresponding to a constraint function. The summation is over all m constraints. The $Z(\lambda)$ in the denominator normalizes the joint probabilities and is called the *partition function*

$$Z(\lambda) = \sum_{i=1}^{n} \exp\left(\sum_{j=1}^{m} \lambda_j\, F_j(X = x_i)\right). \tag{41}$$

For our kangaroo problem the MEP solution can be written as

$$Q_i = \frac{\exp(\lambda_1\, F_1(X = x_i) + \lambda_2\, F_2(X = x_i) + \lambda_3\, F_3(X = x_i))}{Z(\lambda_1, \lambda_2, \lambda_3)}, \tag{42}$$

with

$$Z(\lambda_1, \lambda_2, \lambda_3) = \sum_{i=1}^{n} \exp(\lambda_1 F_1(X = x_i) + \lambda_2 F_2(X = x_i) + \lambda_3 F_3(X = x_i)). \tag{43}$$

The arguments of the exponents can be written in vector-matrix notation, using the constraint function matrix (37)

$$(\lambda_1, \lambda_2, \lambda_3) \cdot \begin{pmatrix} 1 & 1 & 0 & 0 \\ 1 & 0 & 1 & 0 \\ 1 & 0 & 0 & 0 \end{pmatrix} = (\lambda_1 + \lambda_2 + \lambda_3, \lambda_1, \lambda_2, 0). \tag{44}$$

The partition function then becomes

$$Z(\lambda_1, \lambda_2, \lambda_3) = \exp(\lambda_1 + \lambda_2 + \lambda_3) + \exp(\lambda_1) + \exp(\lambda_2) + 1. \tag{45}$$

The joint probabilities (42) are expressed in full as

$$\begin{aligned} Q_1 &= \frac{\exp(\lambda_1 + \lambda_2 + \lambda_3)}{\exp(\lambda_1 + \lambda_2 + \lambda_3) + \exp(\lambda_1) + \exp(\lambda_2) + 1} \\ Q_2 &= \frac{\exp(\lambda_1)}{\exp(\lambda_1 + \lambda_2 + \lambda_3) + \exp(\lambda_1) + \exp(\lambda_2) + 1} \\ Q_3 &= \frac{\exp(\lambda_2)}{\exp(\lambda_1 + \lambda_2 + \lambda_3) + \exp(\lambda_1) + \exp(\lambda_2) + 1} \\ Q_4 &= \frac{1}{\exp(\lambda_1 + \lambda_2 + \lambda_3) + \exp(\lambda_1) + \exp(\lambda_2) + 1}, \end{aligned} \tag{46}$$

and the three Lagrange parameters $(\lambda_1, \lambda_2, \lambda_3)$ are the solutions of the three constraint equations

$$\begin{aligned} Q_1 + Q_2 &= \frac{\exp(\lambda_1 + \lambda_2 + \lambda_3) + \exp(\lambda_1)}{\exp(\lambda_1 + \lambda_2 + \lambda_3) + \exp(\lambda_1) + \exp(\lambda_2) + 1} &&= \langle F_1 \rangle \\ Q_1 + Q_3 &= \frac{\exp(\lambda_1 + \lambda_2 + \lambda_3) + \exp(\lambda_2)}{\exp(\lambda_1 + \lambda_2 + \lambda_3) + \exp(\lambda_1) + \exp(\lambda_2) + 1} &&= \langle F_2 \rangle \\ Q_1 &= \frac{\exp(\lambda_1 + \lambda_2 + \lambda_3)}{\exp(\lambda_1 + \lambda_2 + \lambda_3) + \exp(\lambda_1) + \exp(\lambda_2) + 1} &&= \langle F_3 \rangle. \end{aligned} \tag{47}$$

This is a non-linear problem in three unknowns. Solving the Lagrange parameters usually requires an advanced numerical approximation technique. The Legendre transform provides such a method, which is described in detail by Blower ([4], Ch. 24), and demonstrated in the code example in Figure 4. In some cases, the λ_j can be obtained exactly, as we will see below.

```
cfm = {{1, 1, 0, 0}, {1, 0, 1, 0}, {1, 0, 0, 0}};
cfa = {3 / 4, 1 / 3, Q₁};
lsymbolic = {λ₁, λ₂, λ₃};
zsymbolic = Total[Exp[Dot[lsymbolic, cfm]]];
{entropy, solution} = NMinimize[Log[zsymbolic] - Dot[lsymbolic, cfa], lsymbolic]
lnumeric = lsymbolic /. solution
```

Figure 4. Wolfram Mathematica code for finding the Lagrange parameters (47) using the Legendre transform as a function of Q_1.

Our four models are distinguished only by their constraint function average, $\langle F_3 \rangle = Q_1$, in (39). The details are shown in Table 13.

Table 13. MEP-solution of the kangaroo problem.

	$\langle F_j \rangle$	λ_j	Q_i
$M_{\text{MEP,MaxEnt}}$	$(0.75, 0.33, 0.25)$	$(1.10, -0.69, 0.00)$	$(0.25, 0.50, 0.08, 0.17)$
$M_{\text{MEP,LeastSq}}$	$(0.75, 0.33, 0.29)$	$(0.79, -1.61, 1.16)$	$(0.29, 0.46, 0.04, 0.21)$
$M_{\text{MEP,MaxLog}}$	$(0.75, 0.33, 0.23)$	$(1.29, -0.31, -0.53)$	$(0.23, 0.52, 0.11, 0.14)$
$M_{\text{MEP,MaxSqrt}}$	$(0.75, 0.33, 0.24)$	$(1.21, -0.46, -0.31)$	$(0.24, 0.51, 0.10, 0.15)$

The constraint function vectors are shown in the second column. The three Lagrange parameters are shown in the third column. From this column one can learn that all three Lagrange parameters λ_j vary, even when only the value of $\langle F_3 \rangle$ is varied. Substituting these $(\lambda_1, \lambda_2, \lambda_3)$ in (46), the probability distributions Q_i of the last column are obtained. In our case, these MEP solutions are the same as those obtained by the variational principle methods in table in Table 6, but this need not be so in general. The Lagrange parameters $(\lambda_1, \lambda_2, \lambda_3)$ are related to the covariant coordinates $\left(\theta^1, \theta^2, \theta^3\right)$ of information geometry in Section 6.

Close inspection of the table in Table 13 reveals that the Lagrange multiplier $\lambda_3 = 0$ for $M_{\text{MEP,MaxEnt}}$ solution. This is an important observation because it signals that the $F_3(X = x_i)$ constraint function is redundant and, consequently, can be removed. The solution for the joint probabilities Q_i using only $(\langle F_1 \rangle, \langle F_2 \rangle)$ is identical to the one with $(\langle F_1 \rangle, \langle F_2 \rangle, \langle F_3 \rangle)$. Actually, we knew this already, as this was the basis of the solution in Table 3, but the MEP provides a systematic method for detecting redundancies ([6], p. 369, [8], p. 108).

The Lagrange parameters can be solved algebraically for the $M_{\text{MEP,MaxEnt}}$ and the $M_{\text{MEP,LeastSq}}$ models. Recall that the $M_{\text{VP,MaxEnt}}$ and $M_{\text{VP,LeastSq}}$ models gave exact solutions for the Q_i, namely from substituting (8) and (5) in (2). From (46), we see that $Z = 1/Q_4$, therefore the value of the partition function is exactly known. Subsequently, the $\exp \lambda_j$ can be solved algebraically from (46).

Since the $M_{\text{VP,k}}$ and the $M_{\text{MEP,k}}$ model results turn out to be identical, the distinction based on their solution method can now be dropped. For consistency, we keep the redundant $F_3(X = x_i)$ constraint function in the M_{MaxEnt} model.

6. Information Geometry

6.1. Coordinate Systems

In Information Geometry (IG), a discrete probability distribution Q_i is represented by a point in a *manifold* S. A manifold of dimension n is denoted by S^n; in our case $n = 4$. The probability distribution is parameterized by two dual coordinate systems, namely a covariant system denoted by superscripts $\left(\theta^0, \theta^1, \theta^2, \theta^3\right)$ and a contravariant system denoted by subscripts $(\eta_0, \eta_1, \eta_2, \eta_3)$. This notation corresponds to the work of Amari [9]. The book by Blower [8] is entirely devoted to IG, and in this section we follow his notation.

The contravariant coordinate system corresponds to the constraint function averages

$$(\eta_0, \eta_1, \eta_2, \eta_3) = (\langle F_0 \rangle, \langle F_1 \rangle, \langle F_2 \rangle, \langle F_3 \rangle), \tag{48}$$

whereas the covariant coordinates are the Lagrange multipliers

$$\left(\theta^0, \theta^1, \theta^2, \theta^3\right) = (\lambda_0, \lambda_1, \lambda_2, \lambda_3). \tag{49}$$

The normalization of the probability distribution is given by

$$\langle F_0 \rangle = \eta_0 = 1.$$

This definition yields for the first covariant coordinate

$$\lambda_0 = \theta^0 = 1 - \log Z,$$

where Z is the partition function (41). For example, the uniform distribution q in the covariant coordinate system is

$$(\lambda_0, \lambda_1, \lambda_2, \lambda_3) = (1 - \log 4, 0, 0, 0), \tag{50}$$

and in the contravariant coordinate system

$$(\langle F_0 \rangle, \langle F_1 \rangle, \langle F_2 \rangle, \langle F_3 \rangle) = (1, 1/2, 1/2, 1/4). \tag{51}$$

In IG, the normalization is always implicitly assumed; therefore the coordinates η_0 and θ^0 are never shown explicitly. In the remainder of this paper, only three coordinates are used, namely

$$(\langle F_1 \rangle, \langle F_2 \rangle, \langle F_3 \rangle) = (\eta_1, \eta_2, \eta_3), \tag{52}$$

and

$$(\lambda_1, \lambda_2, \lambda_3) = \left(\theta^1, \theta^2, \theta^3\right). \tag{53}$$

6.2. Tangent Space

All modeling takes place in a sub-manifold S^m, which is tangent to the manifold S^n. This is illustrated in Figure 5. In our kangaroo problem $m = 3$.

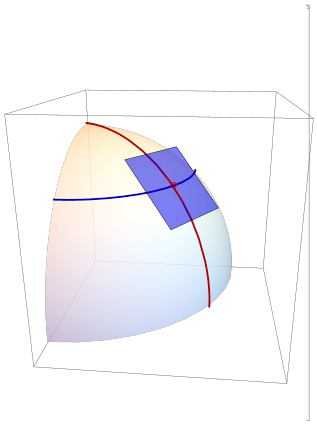

Figure 5. The sub-manifold S^m (blue) is tangent to the manifold S^n. The red line is a meridian of longitude and the blue line is a parallel of latitude through the point of tangency.

Perhaps it is tempting to think of a probability distribution Q as a vector in S^n, with a coordinate system along the axes as in Figure 6. However, this notion is conceptually wrong because the probability distribution is normalized by

$$Q_1 + Q_2 + Q_3 + Q_4 = 1, \tag{54}$$

and not as

$$\sqrt{Q_1^2 + Q_2^2 + Q_3^2 + Q_4^2} = 1. \tag{55}$$

We will return to the issue of normalization in Section 6.6.

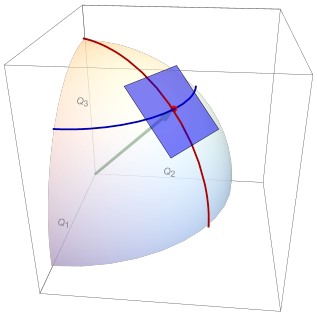

Figure 6. Incorrect view of the probability distribution as a vector (green) to the point of tangency in S^n, with a coordinate system along the axes.

The manifold has no familiar extrinsic set of coordinate axes by which all points can be referenced. All we have is this austere representation of points mapped to a coordinate system ([8], p.46). The tangent space is spanned by a set of m basis vectors. The natural basis vectors of the tangent space are

$$\mathbf{e}_r = F_r(X = x_i) - \langle F_r \rangle, \tag{56}$$

where we recognize the constraint function vector $F_r(X = x_i)$ and the corresponding constraint function average $\langle F_r \rangle$. Notice that the constraint function average $\langle F_r \rangle$ is subtracted from every element of the constraint function vector $F_r(X = x_i)$. For the least squares model solution M_{LeastSq}, the basis vectors are

$$(\mathbf{e}_1, \mathbf{e}_2, \mathbf{e}_3) = \begin{pmatrix} 1/4 & 2/3 & 17/24 \\ 1/4 & -1/3 & -7/24 \\ -3/4 & 2/3 & -7/24 \\ -3/4 & -1/3 & -7/24 \end{pmatrix}, \tag{57}$$

where we have used (36) and (39), and substituted the Q_i using (5).

These basis vectors are not orthogonal. The angle ϕ between two vectors $\mathbf{v}$ and $\mathbf{w}$ is given by

$$\cos(\phi) = \frac{\mathbf{v} \cdot \mathbf{w}}{\|\mathbf{v}\| \|\mathbf{w}\|}. \tag{58}$$

This gives for the angles in degrees between $\mathbf{e}_1$ and $\mathbf{e}_2$, $\mathbf{e}_1$ and $\mathbf{e}_3$, and $\mathbf{e}_2$ and $\mathbf{e}_3$: 98.1°, 56.1°, and 59.0°, respectively. The basis vectors are also not normalized; their lengths are defined as $\|\mathbf{e}_r\|$ and found to be 1.12, 1.05, and 0.87, respectively. However, the $\mathbf{e}_r$ of (57) are perpendicular to the probability distribution (6) from the model $M_{\text{VP,LeastSq}}$

$$Q_i = (7/24, 11/24, 1/24, 5/24); \tag{59}$$

all mutual angles ϕ are 90.0°.

Since the basis vectors $\mathbf{e}_r$ do not form an orthogonal coordinate system, for an arbitrary vector there are two possible projections. Covariant coordinates are obtained by a projection parallel to the basis vectors, while contravariant coordinates are obtained by a perpendicular projection onto the basis vectors.

6.3. Metric Tensor

Each probability distribution p in the manifold S^n has an associated metric tensor $\mathbf{G}(p)$. The metric tensor is an additional structure that allows the definition of distances and angles in the manifold.

The metric tensor is a symmetric matrix, and it comes in covariant and contravariant forms which are each other's matrix inverse. The contravariant metric tensor turns out to be the same as the variance-covariance matrix ([8], p. 50). In our notation the covariant form is g^{rc}, and the contravariant form is g_{rc}, where the superscripts and subscripts r and c refer to the matrix *row* and *column* index.

The elements of the contravariant metric tensor are defined as inner products

$$\begin{aligned} g_{rc} &= \langle (F_r(X = x_i) - \langle F_r \rangle), (F_c(X = x_i) - \langle F_c \rangle) \rangle \\ &= \sum_{i=1}^{n} (F_r(X = x_i) - \langle F_r \rangle) \, (F_c(X = x_i) - \langle F_c \rangle) \, Q_i \\ &= \sum_{i=1}^{n} F_r(X = x_i) \, F_c(X = x_i) \, Q_i - \langle F_r \rangle \, \langle F_c \rangle . \end{aligned} \tag{60}$$

The sum is over all state space cells, whereas the r and c are fixed. Notice that this is the same computation as (22) for the covariance between two vectors.

In the locally flat tangent space S^m, the two coordinate systems are non-orthogonal, and the metric tensor forms the local transformation between the two coordinate systems,

$$\frac{\partial \langle F_c \rangle}{\partial \lambda_r} = \frac{\partial \eta_c}{\partial \theta^r} = g_{rc}, \tag{61}$$

and its inverse

$$\frac{\partial \lambda_r}{\partial \langle F_c \rangle} = \frac{\partial \theta^r}{\partial \eta_c} = g^{rc}. \tag{62}$$

In Blower's notation the contravariant $\langle F_j \rangle$ and covariant λ_j vector indices do not follow the common Einstein convention.

The metric tensor can be computed by

$$g_{rc} = \frac{\partial^2 \log Z}{\partial \lambda_r \, \partial \lambda_c}, \tag{63}$$

with Z the partition function (43)

$$Z(\boldsymbol{\lambda}) = e^{(\lambda_1 + \lambda_2 + \lambda_3)} + e^{\lambda_1} + e^{\lambda_2} + 1. \tag{64}$$

The *contravariant* metric tensor for our kangaroo problem is most easily expressed in the *covariant* coordinates $(\lambda_1, \lambda_2, \lambda_3)$

$$\mathbf{G}(\boldsymbol{\lambda}) = \frac{1}{Z^2} \begin{pmatrix} e^{\lambda_1}\left(e^{\lambda_2}+1\right)\left(e^{\lambda_2+\lambda_3}+1\right) & e^{\lambda_1+\lambda_2}\left(e^{\lambda_3}-1\right) & e^{\lambda_1+\lambda_2+\lambda_3}\left(e^{\lambda_2}+1\right) \\ e^{\lambda_1+\lambda_2}\left(e^{\lambda_3}-1\right) & e^{\lambda_2}\left(e^{\lambda_1}+1\right)\left(e^{\lambda_1+\lambda_3}+1\right) & e^{\lambda_1+\lambda_2+\lambda_3}\left(e^{\lambda_1}+1\right) \\ e^{\lambda_1+\lambda_2+\lambda_3}\left(e^{\lambda_2}+1\right) & e^{\lambda_1+\lambda_2+\lambda_3}\left(e^{\lambda_1}+1\right) & e^{\lambda_1+\lambda_2+\lambda_3}\left(e^{\lambda_1}+e^{\lambda_2}+1\right) \end{pmatrix}. \tag{65}$$

The Wolfram Mathematica [5] code which yields this symbolic expression is surprisingly compact, as shown in Figure 7. This short piece of code demonstrates the indispensability of a good symbolic tool when doing IG.

```
cfm = {{1, 1, 0, 0}, {1, 0, 1, 0}, {1, 0, 0, 0}};
lsymbolic = {λ1, λ2, λ3};
z = Total[Exp[Dot[lsymbolic, cfm] ]];
Outer [D[Log[z], #1, #2] &, lsymbolic, lsymbolic]
```

Figure 7. Wolfram Mathematica code for calculating the metric tensor (65).

Substituting the appropriate Lagrange parameters from Table 13, the metric tensor for the least squares model solution M_{LeastSq} is

$$\mathbf{G}_{\text{LeastSq}} = \begin{pmatrix} 3/16 & 1/24 & 7/96 \\ 1/24 & 2/9 & 7/36 \\ 7/96 & 7/36 & 119/576 \end{pmatrix}, \tag{66}$$

and for the maximum entropy model M_{MaxEnt}, we obtain

$$\mathbf{G}_{\text{MaxEnt}} = \begin{pmatrix} 3/16 & 0 & 1/16 \\ 0 & 2/9 & 1/6 \\ 1/16 & 1/6 & 3/16 \end{pmatrix}. \tag{67}$$

Here we can see that the upper-left 2×2 sub-matrices are identical to the variance-covariance matrix of (24). The extension to the 3×3 matrices is due to the added interactions $F_3(X = x_i)$.

6.4. Kullback–Leibler Divergence

The Kullback–Leibler divergence allows for the determination of the differences in information content between two probability distributions. The Kullback–Leibler *divergence* between two discrete probability distributions p and q is defined as

$$KL(p \parallel q) = \sum_{i=1}^{n} p_i \log\left(\frac{p_i}{q_i}\right). \tag{68}$$

The divergence is not a distance because the expression is not symmetric in p and q. A common way to refer to Kullback–Leibler divergence (KL) is as the relative entropy of p with respect to q or the information gained from p over q.

For example, with $p = (0, 0, 1, 0)$ and $q = (1/4, 1/4, 1/4, 1/4)$ we have

$$\begin{aligned} KL(p \parallel q) &= 0 \log\left(\frac{0}{1/4}\right) + 0 \log\left(\frac{0}{1/4}\right) + 1 \log\left(\frac{1}{1/4}\right) + 0 \log\left(\frac{0}{1/4}\right) \\ &= \log(4), \end{aligned} \tag{69}$$

where we have used the limit expression (31) again. However, when we interchange p and q we obtain

$$\begin{aligned} KL(q \parallel p) &= 1/4 \log\left(\frac{1/4}{0}\right) + 1/4 \log\left(\frac{1/4}{0}\right) + 1/4 \log\left(\frac{1/4}{1}\right) + 1/4 \log\left(\frac{1/4}{0}\right) \\ &= \infty. \end{aligned} \tag{70}$$

Therefore, figuratively speaking, we have gained a finite amount of information when learning that we are certain, but we have lost an "infinite" amount when we lose our certainty. Learning and forgetting are asymmetric.

Therefore, the notion of the KL-divergence as a distance measure between distinct probability distributions is flawed. Rewriting (68) we obtain

$$\begin{aligned} KL(p \parallel q) &= \sum_{i=1}^{n} p_i \log\left(\frac{p_i}{q_i}\right) \\ &= \sum_{i=1}^{n} p_i \log p_i - \sum_{i=1}^{n} p_i \log q_i \\ &= -\langle \log q \rangle_p - H(p) \end{aligned} \tag{71}$$

where $H(p)$ is the entropy of p. The first term on the right is the expectation of $\log q$ with respect to p. When $q \neq p$, $KL(p \parallel q)$ and $-\langle \log q \rangle_p$ are strictly positive quantities.

The KL-divergence can be expressed in bits when (68) is multiplied by $\log_2 e \approx 1.44$. Table 14 shows the values for our four models. As expected, the table is not symmetric.

Table 14. The Kullback–Leibler divergence $KL(p \parallel q)$ (bits) between the models M_k, where p and q are the models in the rows and columns, respectively.

	$M_{\mathbf{MaxEnt}}$	$M_{\mathbf{LeastSq}}$	$M_{\mathbf{MaxLog}}$	$M_{\mathbf{MaxSqrt}}$
M_{MaxEnt}	0	0.0368	0.0085	0.0030
M_{LeastSq}	0.0327	0	0.0729	0.0547
M_{MaxLog}	0.0087	0.0834	0	0.0014
M_{MaxSqrt}	0.0031	0.0623	0.0014	0

When the distributions p and $q = p + dp$ are infinitesimally close, writing

$$q_i = p_i + dp_i, \tag{72}$$

we have

$$\sum_{i=1}^{n} dp_i = 0. \tag{73}$$

Expanding the KL-divergence for small dp

$$\begin{aligned} KL(p \parallel q) &= \sum_{i=1}^{n} p_i \log\left(\frac{p_i}{q_i}\right) \\ &= -\sum_{i=1}^{n} p_i \log\left(\frac{q_i}{p_i}\right) \\ &= -\sum_{i=1}^{n} p_i \log\left(1 + \frac{dp_i}{p_i}\right) \\ &= -\sum_{i=1}^{n} dp_i + \sum_{i=1}^{n} \frac{1}{2} \frac{dp_i^2}{p_i} - O\left(dp^3\right) \\ &\approx \frac{1}{2} \sum_{i=1}^{n} \frac{dp_i^2}{p_i}. \end{aligned} \tag{74}$$

This expansion is a sum of squares, which is symmetric. Therefore, the KL-divergence is *commutative* for infinitesimal separations between p and q.

This property of the Kullback–Leibler divergence has an analogy in classical mechanics, namely that two infinitesimal rotations of a rigid body along different principal axes are commutative, while finite rotations are not.

6.5. Distances

What is the distance between two discrete probability distributions p and q in the manifold S^n? This is at the heart of Information Geometry. For a distance we need a curve connecting the two points. There are many possibilities. What would be the length of such curves? Which one is the shortest? The shortest of all possible curves is called a *geodesic*. Suppose that s is a curve connecting p and q, then any point t on the curve s is a probability distribution. Therefore, we have a continuum of probability distributions along s in the manifold S^n.

For two close-by points p and $q = p + dp$, their distance is a function of the KL-divergence, namely ([8], pp. 77–78)

$$ds = \sqrt{2\, KL(p \parallel q)}. \tag{75}$$

The same distance is given by

$$ds = \sqrt{\sum_{r=1}^{m}\sum_{c=1}^{m} g_{rc}(p)\, d\lambda_r\, d\lambda_c}, \tag{76}$$

where the covariant coordinates $\boldsymbol{\lambda}$ and $\boldsymbol{\lambda} + d\boldsymbol{\lambda}$ of p and q are used, and the metric tensor $g_{rc}(p)$ is evaluated as in (65). However, there is a subtle difference here, namely the KL-divergence in (75) is computed in the full manifold S^n, whereas ds in (76) is computed in the tangent space S^m, with $m < n$.

When the two distributions are finitely separated, as is the case for our models M_k, the length of the curve $s(t)$ is the integral from p to q of

$$L(s) = \int_p^q \left|s'(t)\right| dt, \tag{77}$$

where $s(t)$ is the curve in S^n parameterized by the probability distribution t, and $s'(t)$ is its first derivative. The tangent sub-manifold $S^m(t)$ follows t along $s(t)$ from p to q, and the Lagrange parameters $\boldsymbol{\lambda}(t)$ and the metric tensor $g_{rc}(t)$ vary with t. However, finding the distance $D = \min L(s)$ is an Euler–Legendre variational problem beyond the scope of this paper [10].

6.6. Angular Distances

The distance between two probability distributions can also be found as the arc length of a great circle on a sphere in S^n. This is known as the *Bhattacharyya angle*.

Substituting (74) in (75) we can write

$$\begin{aligned}(ds)^2 &= \sum_{i=1}^{n} \frac{(dp_i)^2}{p_i} \\ &= \sum_{r=1}^{m}\sum_{c=1}^{m} g_{rc}(p)\, dp_r\, dp_c,\end{aligned} \tag{78}$$

with a metric tensor

$$g_{rc}(p) = \begin{cases} 1/p_r & \text{for } r = c \\ 0 & \text{otherwise.} \end{cases} \tag{79}$$

Using the transformation

$$\psi_i = \sqrt{p_i} \tag{80}$$

we define ψ as a point on the *positive orthant* of the unit sphere with

$$\sum_{i=1}^{n} \psi_i^2 = \sum_{i=1}^{n} p_i = 1. \tag{81}$$

This effectively restricts ψ to a sub-manifold of dimension S^{n-1}. The geometry is illustrated by Figure 8. In the ψ-coordinate system, the infinitesimal distance becomes

$$\begin{aligned}(ds)^2 &= \sum_{i=1}^{n} \frac{(dp_i)^2}{p_i} \\ &= \sum_{i=1}^{n} \frac{(2\psi\, d\psi_i)^2}{\psi_i^2} \\ &= 4\sum_{i=1}^{n} (d\psi_i)^2,\end{aligned} \tag{82}$$

or

$$ds = 2\,d\psi. \tag{83}$$

Notice that in this coordinate system the metric tensor is the Euclidean metric tensor

$$g_{rc}(\psi) = \begin{cases} 1 & \text{for } r = c \\ 0 & \text{otherwise.} \end{cases} \tag{84}$$

With this transformation the probability distributions become points on a hypersphere with a unit radius in $(n-1)$ dimensions. However, it is well known that geodesics on a sphere are great circles. Therefore, the distance can be obtained by the path integral (77) along a great circle connecting the two points. The arc length between two points is the subtended angle θ between two points ψ_1 and ψ_2 on the unit hypersphere

$$\begin{aligned} \theta &= \arccos \psi_1 \cdot \psi_2 \\ &= \arccos \sum_{i=1}^{n} \psi_{1,i}\, \psi_{2,i} \\ &= \arccos \sum_{i=1}^{n} \sqrt{p_i}\, \sqrt{q_i}. \end{aligned} \tag{85}$$

This remarkable result is the Bhattacharyya angle between two probability distributions [11]. The distance D between p and q is twice the arc length from (83)

$$\begin{aligned} D(p,q) &= 2\,\theta \\ &= 2 \arccos \sum_{i=1}^{n} \sqrt{p_i}\, \sqrt{q_i}. \end{aligned} \tag{86}$$

The units of D are radians. The maximum distance of π radians is achieved between two orthogonal distributions.

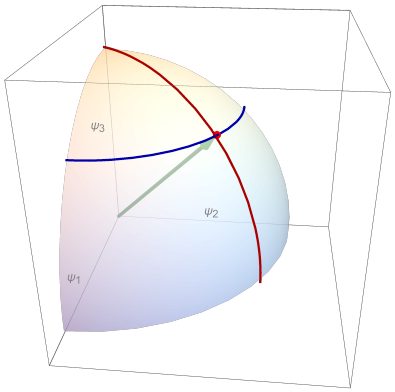

Figure 8. Positive orthant S^{n-1}. In the ψ-coordinate system, the ψ_i are orthonormal coordinates.

With this result we can compute the symmetric distance table for our four Kangaroo models, shown in Table 15; the numerical values are converted from radians to degrees. From this table we see that the largest distance is between the models M_{LeastSq} and M_{MaxLog}. This observation corresponds with these models having the biggest difference in their correlation coefficients $\rho(F_1, F_2)$ in Table 11.

Table 15. The distance D (in degree) between the models M_k.

$M_{\text{MaxEnt}} \leftrightarrow M_{\text{LeastSq}}$	12.5
$M_{\text{MaxEnt}} \leftrightarrow M_{\text{MaxLog}}$	6.3
$M_{\text{MaxEnt}} \leftrightarrow M_{\text{MaxSqrt}}$	3.7
$M_{\text{LeastSq}} \leftrightarrow M_{\text{MaxLog}}$	18.8
$M_{\text{LeastSq}} \leftrightarrow M_{\text{MaxSqrt}}$	16.3
$M_{\text{MaxLog}} \leftrightarrow M_{\text{MaxSqrt}}$	2.5

Interestingly, when we define lower and upper bounds

$$\begin{aligned} KL_{\min} &= \min(KL(p \parallel q), KL(q \parallel p)) \\ KL_{\max} &= \max(KL(p \parallel q), KL(q \parallel p)), \end{aligned} \tag{87}$$

all the distances from Table 15 have values

$$\sqrt{2\,KL_{\min}} < D(p,q) < \sqrt{2\,KL_{\max}}. \tag{88}$$

Although we have no proof, this observation suggests that the two forms of the KL-divergence may act as lower and upper limits for the true distance $D(p,q)$.

6.7. Geodesics

The arc of the great circle connecting the two points can be found as follows [12]. Let v_1 and v_2 be two points on the $(n-1)$ dimensional hypersphere, then

$$w = v_2 - (v_2 \cdot v_1)\, v_1 \tag{89}$$

$$u = \frac{1}{\|w\|}\, w. \tag{90}$$

Then

$$\alpha(\tau) = \cos(\tau)\, v_1 + \sin(\tau)\, u \tag{91}$$

traces out a great circle through v_1 and v_2. It starts at $\alpha(\tau) = v_1$ when $\tau = 0$, it reaches $\alpha(\tau) = v_2$ at $\tau = \arccos(v_2 \cdot v_1)$, and returns to v_1 when $\tau = 2\pi$. Here we recognize the Bhattacharyya angle again.

When $v_1 = \psi_1$ and $v_2 = \psi_2$ represent two probability distributions, they must remain on the positive orthant of the hypersphere. For $0 \le \tau \le \arccos(\psi_2 \cdot \psi_1)$,

$$t = \alpha^2(\tau) \tag{92}$$

is a probability distribution in S^n on the geodesic connecting ψ_1 and ψ_2.

Our under-determined problem is parametrized by a single variable, namely $1/12 \le Q_1 \le 1/3$ from (3), which implies that there is only one dimension involved. Therefore it seemed reasonable to surmise that varying Q_1 traces out probability distributions t along the shortest distance between the various models, but this turned out to be incorrect. The distributions t on the geodesic $s(t)$ connecting, for example, $M_{\text{LeastSq}}$to M_{MaxEnt}, do not comply with the constraint function average vector (39)

$$(\langle F_1 \rangle, \langle F_2 \rangle, \langle F_3 \rangle) = (3/4, 1/3, Q_1). \tag{93}$$

except for the endpoints.

6.8. Distances Revisited

Our knowledge of the geodesic $s(t)$ allows us to verify (75) with (76). The arc length of the geodesic between p and q is

$$\begin{aligned} s &= \int_p^q ds(t)\, dt \\ &= \int_p^q \sqrt{\sum_{r=1}^{m}\sum_{c=1}^{m} g_{rc}(t)\, d\lambda(t)_r\, d\lambda(t)_c}\, dt, \end{aligned} \tag{94}$$

where we have substituted (76). Notice that the metric tensor as well as the covariant coordinates depends on t. This integral can be approximated by a sum of many small steps in t ([8], p.78).

By taking K small segments, the distance s is approximated by

$$s = \sum_{k=0}^{K-1} \left(\sum_{r=1}^{m}\sum_{c=1}^{m} (\lambda(t_k) - \lambda(t_{k+1}))_r\, g_{rc}(\lambda(t_k))\, (\lambda(t_k) - \lambda(t_{k+1}))_c \right)^{\frac{1}{2}}. \tag{95}$$

Here $k = 0$ corresponds to the probability distribution $t_0 = p$ and $k = K$ is the distribution $t_K = q$. The intermediate points t_k are obtained by dividing the arc $0 \le \tau \le \arccos\left(\sqrt{p}\cdot\sqrt{q}\right)$ of the hypersphere into K equal angular segments. The corresponding probability distributions are $t_k = \alpha^2(\tau_k)$, using (92).

For each t_k in (95), the constraint function averages $(\langle F_1\rangle, \langle F_2\rangle, \langle F_3\rangle)_k$ are obtained through the multiplication by the constraint function vectors (36). The corresponding covariant coordinates $\lambda(t_k)$ are computed by solving the set of equations in (47), as illustrated by Figure 4. Finally, the metric tensor $g_{rc}(\lambda(t_k))$ is obtained through substitution of $\lambda(t_k)$ in (65). By taking $K = 128$ segments and performing the computation of (95) we have confirmed all the numerical values in Table 15. This confirms the equivalence of (75) and (76).

7. Conclusions

The Gull–Skilling kangaroo problem provides a useful setting for illustrating the solution of under-determined problems in probability. The Variational Principle—in conjunction a variational function—effectively creates enough missing information for a complete solution, but not necessarily the minimum amount. In this paper four different Variational Principle solutions are shown, only one of which introduces the minimum amount, when the variational function is the Shannon–Jaynes entropy function.

The Maximum Entropy Principle is an alternative method for solving under-determined problems, which however avoids any implicit introduction of extra information not in the original problem. This information manifests itself in our examples as added correlations in the solutions.

The Kullback–Leibler divergence allows for the determination of the differences in information content between two probability distributions, but it cannot be used as a distance measure. It is symmetrical for infinitesimal separations. We point out an analogy with infinitesimal rigid body rotations.

Through the lens of Information Geometry, the actual geometric distance between two probability distributions along a geodesic path, can also be expressed as twice the Bhattacharyya angle in a hypersphere. In this paper, we illustrate the equivalence of these two geometrical concepts.

We also find that the mutual differences in distance between any two models, are directly reflected in the difference of their correlation coefficients.

Our understanding of the kangaroo problem and its implications has been particularly facilitated by the symbolic programming capabilities of Wolfram Mathematica.

Author Contributions: The authors R.B. and B.J.S. contributed equally to the manuscript. All authors have read and agreed to the published version of the manuscript.

Funding: This research received no external funding.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: Not applicable.

Acknowledgments: This paper was written to honor our late friend David Blower. The reader may benefit from his books, as we have. We acknowledge the comments of John Skilling, who pointed out the Bhattacharyya angle to us. Further we thank Ann Stokes, Ali Mohammad-Djafari and two anonymous referees for supporting comments.

Conflicts of Interest: The authors declare no conflict of interest.

References

1. Sivia, D.S.; Skilling, J. *Data Analysis*, 2nd ed.; Oxford University Press: Oxford, UK, 2006; pp. 111–113.
2. Gull, S.F.; Skilling, J. Maximum entropy method in image processing. *IEE Proc.* **1984,** *131*, 646–659. [CrossRef]
3. Jaynes, E.T. Monkeys, kangaroos and N. In *Maximum Entropy and Bayesian Methods in Applied Statistics*; Justice, J. H., Ed.; Cambridge University Press: Calgary, AB, Canada, 1984; pp. 27–58.
4. Blower, D.J. *Information Processing, Volume II, The Maximum Entropy Principle*; Third Millennium Inferencing: Pensacola, FL, USA, 2013.
5. Wolfram Mathematica. Available online: www.wolfram.com (accessed on 7 December 2022).
6. Jaynes, E.T. *Probability Theory: The Logic of Science*; Bretthorst, G.L., Ed.; Cambridge University Press: New York, NY, USA, 2003.
7. Buck, B; Macaulay, V.A. *Maximum Entropy in Action*; Oxford University Press: Oxford, UK, 1991.
8. Blower, D.J. *Information Processing, Volume III, Introduction to Information Geometry*; Third Millennium Inferencing: Pensacola, FL, USA, 2016.
9. Amari, S.; Nagaoka, H. *Methods of Information Geometry*; Originally Published in Japanese by Iwanami Shoten Publishers, Tokyo; Translated by D. Harada; Oxford University Press: Oxford, UK, 1993.
10. Mathews, J.; Walker, R.L. *Mathematical Methods of Physics*, 2nd ed.; Addison-Wesley Publ.: Menlo Park, CA, USA, 1973; pp. 322–344.
11. Bhattacharyya, A. On a Measure of Divergence between Two Multinomial Populations. *Sankhyā* **1946**, *7*, 401–406.
12. Mathematics Stack Exchange. Available online: https://math.stackexchange.com/questions/1883904/a-time-parameterization-of-geodesics-on-the-sphere (accessed on 15 September 2022).