PRIOR PROBABILITIES: AN INFORMATION-THEORETIC APPROACH

Philip Goyal

Cavendish Laboratory, Cambridge University.

Abstract. General theoretical principles that enable the derivation of prior probabilities are of interest both in practical data analysis and, more broadly, in the foundations of probability theory. In this paper, it is shown that the general rule for the assignment of priors proposed by Jeffreys can be obtained from, and is logically equivalent to, an intuitively reasonable information-theoretical invariance principle Some of the implications for the priors proposed by Hartigan [1], Jaynes [2], and Skilling [3], are also discussed.

Keywords: Bayesian Data Analysis, Prior probability, Jeffreys' prior, Entropic prior
PACS: 02.50.Cw, 02.50.Tt, 89.70.+c

1. INTRODUCTION

Any application of probability theory to a problem of data analysis requires the specification of a *prior probability* over the parameters being estimated on the basis of the given data. For example, suppose that we are given a coin, and suppose that we model the outcome of a toss of the coin as the outcome of a two-outcome probabilistic source with outcome probabilities P_1, P_2. After an experiment where the coin is tossed n times, we can estimate the value of P_1 on the basis of the data string $D_n = a_1 a_2 \ldots a_n$, where a_i is the outcome (1 or 2) of the ith toss, using Bayes rule,

$$\Pr(P_1|D_n, I) = \frac{\Pr(D_n|P_1, I)\Pr(P_1|I)}{\Pr(D_n|I)}. \tag{1}$$

The prior, $\Pr(P_1|I)$, is undetermined by the theory of probability, and represents our state of knowledge prior to performing the experiment.

Suppose that we have no knowledge about the origin or manufacture of the coin, so that, for example, the coin could be from general circulation or be a magician's coin. What prior over P_1 properly reflects this state of ignorance? Since the work of Bayes and Laplace, the question of how to assign $\Pr(P_1|I)$ and, more generally, the prior $\Pr(\mathbf{P}|I)$ where $\mathbf{P}$ is an N–dimensional probability vector, so as to reflect a state of ignorance, has received a succession of sometimes conflicting answers. For example, Bayes and Laplace assigned the uniform prior, $\Pr(P_1|I) = 1$, arguing on the ground of the general philosophical principle of uniformity that, lacking evidence to the contrary, one should assign equal weight to all possible values of P_1. However, if one were to parameterise P_1 by the parameter θ, then such an argument could, reasonably, also be applied to θ; but, if $P_1(\theta)$ were not a linear function of θ, it would then follow from $\Pr(P_1|I)|dP_1| = \Pr(\theta|I)|d\theta|$ that $\Pr(P_1|I)$ is *not* uniform. Although one could

CP803, *Bayesian Inference and Maximum Entropy Methods in Science and Engineering*
edited by Kevin H. Knuth, Ali E. Abbas, Robin D. Morris, and J. Patrick Castle

conceivably evade this objection by granting the parameter P_1 privileged status, it is difficult to find a compelling reason for doing so.

Similarly, if we are given the sample $x_1, x_2, \ldots, x_N$ from a unknown probability distribution, $\Pr(x|\vec{\theta})$, which could be continuous, where $\vec{\theta} = (\theta_1, \theta_2, \ldots, \theta_K)$ is a K–dimensional parameter vector, and we wish to determine $\Pr(\vec{\theta}|x_1, x_2, \ldots, x_N, I)$, the prior $\Pr(\vec{\theta}|I)$ must be specified. For the case of sampling from a normal probability distribution with mean μ and standard deviation σ, Jeffreys [4] argued that $\Pr(\sigma|I) \propto 1/\sigma$ and, in [5], suggested the prior $\Pr(\rho|I) \propto 1/\rho$ (now known as Jeffreys' prior) be assigned to any continuous parameter, ρ, known to be positive, on the grounds that the prior over $\Pr(\rho^m|I)$ is then also of this same form. However, no compelling reason was given for basing the prior upon the functions ρ^m, and for excluding other functions of ρ.

Recognising the fragility of existing arguments for the assignment of priors such as $\Pr(P_1|I)$, Jeffreys [6] considered the question of whether there exists a general principle for the assignment of priors that are immune to such objections. By investigating some mathematical expressions that could be plausibly viewed as quantifying the 'distance' between two discrete probability distributions, Jeffreys showed that some of these expressions lead, up to an overall multiplicative constant, to the same metric. Specifically, in the case of a probabilistic model of N possible outcomes, with N–dimensional probability vector $\mathbf{P}(\vec{\theta})$, where $\vec{\theta} = (\theta_1, \theta_2, \ldots, \theta_K)$ with $0 < K < N$, the metric

$$g_{kk'} = \sum_{i=1}^{N} \frac{1}{P_i} \frac{\partial P_i}{\partial \theta_k} \frac{\partial P_i}{\partial \theta_{k'}}, \quad k, k' = 1, 2, \ldots, K. \tag{2}$$

Noting the invariance of $(\det g)^{1/2} d\theta_1 d\theta_2 \ldots d\theta_K$ under non-singular re-parameterisation of $\mathbf{P}$, and appealing to the plausible desideratum that the prior probability $\Pr(\vec{\theta}|I)$ be invariant under such re-parameterisation, Jeffreys proposed the rule (referred to hereafter as *Jeffreys' rule*) that one assign the prior

$$\Pr(\vec{\theta}|I) \propto (\det g)^{1/2}. \tag{3}$$

From Jeffreys' rule, it follows, in contrast to the assumption of Bayes and Laplace, that

$$\Pr(P_1|I) \propto \frac{1}{\sqrt{P_1 P_2}}, \tag{4}$$

and, in the case of an N–outcome probabilistic source,

$$\Pr(P_1, P_2, \ldots, P_N|I) \propto \frac{1}{\sqrt{P_1 P_2 \ldots P_N}}, \tag{5}$$

a prior which will be referred to as Jeffreys' multinomial prior. From Jeffreys' rule, it also follows for a normal distribution that $\Pr(\sigma|I) \propto 1/\sigma$, in agreement with Jeffreys' previous arguments.

Compared with the rules for the assignment of priors that preceded it, Jeffreys' rule has the considerable advantage of being a general mechanical procedure, but it has the disadvantage of being reliant upon the metric in Eq. (2), which is not logically derived

from a set of intuitively plausible postulates. Consequently, since its proposal, numerous other specific priors and assignment rules, such as in [1, 2, 3, 7, 8], have been proposed where an attempt is made to base the derivations on intuitively plausible postulates. The main objective of this paper is to derive Jeffreys' rule from an intuitively plausible information-theoretic invariance principle. Some of the implications for some of the specific priors and rules for the assignment of priors that have been derived by other authors, such as in [1, 2, 3, 7, 8], will also be briefly discussed.

2. DERIVATION OF JEFFREYS' MULTINOMIAL PRIOR

Suppose that an experimenter makes a trial of n interrogations of a probabilistic source with an unknown probability vector $\mathbf{P} = (P_1, P_2, \ldots, P_N)$ and obtains the data string, $D_n = a_1 a_2 \ldots a_N$, of length n, where a_i represents the value of the ith outcome. If the experimenter wishes to estimate $\mathbf{P}$ on the basis of D_n, the only relevant data is the number of instances, m_i of each outcome, i, which can be encoded in the data vector $\mathbf{m} = (m_1, m_2, \ldots, m_N)$, or, equivalently, in the pair $(n, \mathbf{f})$, where $\mathbf{f} = \mathbf{m}/n$ is the frequency vector. Given the data vector, Bayes' theorem can be used to calculate the probability density function, $\Pr(\mathbf{P}|\mathbf{f}, n, I)$, where I represents the knowledge that the experimenter possesses prior to performing the interrogations.

Let us quantify the change in the experimenter's knowledge about $\mathbf{P}$ using the Shannon-Jaynes entropy functional. If the frequency vector $\mathbf{f}$ is obtained in n interrogations, the gain of information about $\mathbf{P}$,

$$\begin{aligned} \Delta H &= \mathrm{H}[\Pr(\mathbf{P}|I)] - H[\Pr(\mathbf{P}|\mathbf{f}, n, I)] \\ &= \int \cdots \int \Pr(\mathbf{P}|\mathbf{f}, n, I) \ln \frac{\Pr(\mathbf{P}|\mathbf{f}, n, I)}{\Pr(\mathbf{P}|I)} \, dP_1 \ldots dP_{N-1}, \end{aligned} \tag{6}$$

whose value depends upon the the prior probability, $\Pr(\mathbf{P}|I)$.

Consider the case where $N = 2$. Using Bayes' theorem, the posterior probability can be expressed as

$$\Pr(\mathbf{P}|\mathbf{f}, n, I) = \frac{\Pr(\mathbf{f}|\mathbf{P}, n, I) \Pr(\mathbf{P}|I)}{\int \Pr(\mathbf{f}|\mathbf{P}, n, I) \Pr(\mathbf{P}|I) \, dP_1}, \tag{7}$$

where the likelihood,

$$\Pr(\mathbf{f}|\mathbf{P}, n, I) = \frac{n!}{m_1!(n-m_1)!} P_1^{m_1} (1 - P_1)^{n-m_1}. \tag{8}$$

In the limit of large n, the likelihood becomes very sharply peaked around $m_1 = nP_1$ so that the prior probability factors out of the integrand, and the posterior probability can be approximated by a Gaussian function of variance $\sigma^2 = f_1(1 - f_1)/n$.

For the purpose of illustration, suppose the prior probability were chosen to be uniform. The information gain (Eq. 6) would then become

$$\begin{aligned} \Delta H &= -\ln\left(\sigma\sqrt{2\pi e}\right) \\ &= \frac{1}{2}\ln\left(\frac{n}{2\pi e}\right) - \frac{1}{2}\ln\left(f_1(1 - f_1)\right), \end{aligned} \tag{9}$$

whose value is dependent upon f_1. In the limit of large n, f_1 tends to P_1. Hence, with respect to the information gained about the identity of $\mathbf{P}$ in the limit of large n, there would then be a difference between sources with different values of $\mathbf{P}$. But, if we are ignorant as to the origin of the probabilistic source being interrogated, it is reasonable to require that the prior be such that the amount of information that the data D_n provides about $\mathbf{P}$ should not depend upon what $\mathbf{P}$ happens to be. On the ground of this intuitively reasonable notion, we shall select $\Pr(\mathbf{P}|I)$ such, in the limit of large n, the amount of information gained in n detections is independent of $\mathbf{P}$. That is, we shall postulate:

Principle of Information Gain. *In n interrogations of an N-outcome probabilistic source with unknown probability vector $\mathbf{P}$, the amount of Shannon-Jaynes information provided by the data about $\mathbf{P}$ is independent of $\mathbf{P}$ for all $\mathbf{P}$ in the limit as $n \to \infty$.*

In order to implement this principle, we make use of the fact the Shannon-Jaynes entropy is invariant under a change of variables [9]. We shall parameterise the vector $\mathbf{P}$ by the $(N-1)$–dimensional parameter vector $\vec{\lambda} = (\lambda_1, \lambda_2, \ldots, \lambda_{N-1})$, so that $\mathbf{P} = \mathbf{P}(\vec{\lambda})$, and then set the prior probability, $\Pr(\vec{\lambda}|I)$, equal to a constant. This transforms the initial problem of determining the prior probability over $\mathbf{P}$ into one of determining the functions $P_i(\vec{\lambda})$.

The first step is to determine $\Pr(\vec{\lambda}|\mathbf{f},n,I)$. From Bayes' theorem, the posterior probability,

$$\begin{aligned}\Pr(\vec{\lambda}|\mathbf{f},n,I) &= \frac{\Pr(\mathbf{f}|\vec{\lambda},n,I)\Pr(\vec{\lambda}|n,I)}{\int\cdots\int \Pr(\mathbf{f}|\vec{\lambda},n,I)\Pr(\vec{\lambda}|n,I)\,d\lambda_1\ldots d\lambda_{N-1}} \\ &= \frac{\Pr(\mathbf{f}|\vec{\lambda},n,I)}{\int\cdots\int \Pr(\mathbf{f}|\vec{\lambda},n,I)\,d\lambda_1\ldots d\lambda_{N-1}}.\end{aligned} \tag{10}$$

In the second line, $\Pr(\vec{\lambda}|n,I)$ has been set equal to $\Pr(\vec{\lambda}|I)$. This follows from an application of Bayes' theorem, $\Pr(\vec{\lambda}|n,I)\Pr(n|I) = \Pr(n|\vec{\lambda},I)\Pr(\vec{\lambda}|I)$, and the fact that n is chosen freely by the experimenter and therefore cannot depend upon $\vec{\lambda}$. Hence, the posterior probability is proportional to the likelihood, $\Pr(\mathbf{f}|\vec{\lambda},n,I)$.

When n is large, using Stirling's approximation, $n! = n^n(2\pi n)^{1/2}e^{-n} + O(1/n)$,

$$\begin{aligned}\Pr(\mathbf{f}|\vec{\lambda},n,I) &= \frac{n!}{(nf_1)!\ldots(nf_N)!}\left[P_1(\vec{\lambda})\right]^{nf_1}\ldots\left[P_N(\vec{\lambda})\right]^{nf_N} \\ &= \frac{(2\pi n)^{1/2}}{(2\pi n)^{N/2}}\frac{1}{\sqrt{f_1 f_2\ldots f_N}}\exp\left(-n\sum_i f_i \ln\frac{f_i}{P_i(\vec{\lambda})}\right).\end{aligned} \tag{11}$$

Writing

$$P_i(\vec{\lambda}) = P_i(\vec{\lambda}^{(0)}) + \sum_{l=1}^{N-1}\frac{\partial P_i}{\partial\lambda_l}\bigg|_{\vec{\lambda}^{(0)}}(\lambda_l - \lambda_l^{(0)}) + \ldots, \tag{12}$$

where $\mathbf{f} = \mathrm{P}(\vec{\lambda}^{(0)})$, and retaining only leading order terms in the λ_l,

$$\Pr(\vec{\lambda}|\mathbf{f},n,I) \propto \prod_{l=1}^{N-1}\prod_{l'=1}^{N-1} \exp\left(-\frac{(\lambda_l-\lambda_l^{(0)})(\lambda_{l'}-\lambda_{l'}^{(0)})}{2\sigma_{ll'}^2}\right), \tag{13}$$

where the proportionality constant is a function of the $\lambda_l^{(0)}$ only, and where

$$\frac{1}{\sigma_{ll'}^2} = n\sum_{i=1}^{N}\frac{1}{P_i(\vec{\lambda}^{(0)})}\frac{\partial P_i}{\partial\lambda_l}\bigg|_{\vec{\lambda}^{(0)}}\frac{\partial P_i}{\partial\lambda_{l'}}\bigg|_{\vec{\lambda}^{(0)}}. \tag{14}$$

We shall define the vector $\mathbf{Q} = (Q_1, Q_2, \ldots, Q_N)$ such that $Q_i = \sqrt{P_i}$, so that $\mathbf{Q}$ lies on the positive orthant, S_+^{N-1}, on the unit hypersphere, S^{N-1}, in an N-dimensional Euclidean space with axes $Q_1, Q_2, \ldots, Q_N$. Equation (14) can then be rewritten as

$$\frac{1}{\sigma_{ll'}^2} = 4n\sum_{i=1}^{N}\frac{\partial Q_i}{\partial\lambda_l}\bigg|_{\vec{\lambda}^{(0)}}\frac{\partial Q_i}{\partial\lambda_{l'}}\bigg|_{\vec{\lambda}^{(0)}}. \tag{15}$$

In the case where $N = 2$,

$$\begin{aligned}\frac{1}{\sigma_{11}^2} &= 4n\left[\left(\frac{dQ_1}{d\lambda_1}\right)^2\bigg|_{\lambda_1^{(0)}} + \left(\frac{dQ_2}{d\lambda_1}\right)^2\bigg|_{\lambda_1^{(0)}}\right] \\ &= 4n\left(\frac{ds}{d\lambda_1}\right)^2\bigg|_{\lambda_1^{(0)}},\end{aligned} \tag{16}$$

where $ds^2 = dQ_1^2 + dQ_2^2$, the Euclidean line element in Q^2-space. The posterior probability, $\Pr(\lambda_1|\mathbf{f},n,I)$, is therefore a Gaussian with standard deviation,

$$\sigma = \frac{1}{2\sqrt{n}}\left|\frac{ds}{d\lambda_1}\right|^{-1}\bigg|_{\lambda_1^{(0)}}, \tag{17}$$

where $\lambda_1^{(0)} = P_1^{-1}(f_1)$, and, since $\Pr(\lambda_1|I)$ is constant,

$$\begin{aligned}\Delta H &= \frac{1}{2}\ln\left(\frac{2n}{\pi e}\right) + \ln\left|\frac{ds}{d\lambda_1}\right|\bigg|_{\lambda_1^{(0)}} - \ln[\Pr(\lambda_1|I)] \\ &= \frac{1}{2}\ln\left(\frac{2n}{\pi e}\right) - \ln[\Pr(s(\lambda_1^{(0)})|I)]\end{aligned} \tag{18}$$

where s is the distance along the circumference of the positive quadrant of the unit circle, and where the relation $\Pr(\lambda_1|I)|d\lambda_1| = \Pr(s|I)|ds|$ has been used in the second line. Independence of ΔH from f_1 can be ensured if and only if $\Pr(s|I)$ at $\lambda_1^{(0)}$ is a constant, independent of $\lambda_1^{(0)}$. Since $\Pr(\lambda_1|I)$ is a constant, it follows from $\Pr(\lambda_1|I)|d\lambda_1| = \Pr(s|I)|ds|$ that $s(\lambda_1) = \alpha\lambda_1 + \beta$, where α and β are arbitrary real constants.

If $\theta \in [0, \pi/2]$ is taken to be the polar angle in Q^2–space, so that $Q_1 = \cos\theta$, then $Q_1 = \cos[s(\lambda_1) + \gamma]$, where γ is an arbitrary real constant. Hence, $Q_1 = \cos(\alpha\lambda_1 + \beta')$ and $P_1 = \cos^2(\alpha\lambda_1 + \beta')$, where β' is an arbitrary real constant. Since $\Pr(\lambda_1|I)$ is constant and since $\Pr(P_1|I)\,|dP_1| = \Pr(\lambda_1|I)\,|d\lambda_1|$, one obtains

$$\Pr(P_1, P_2|I) = \begin{cases} \frac{1}{\pi}\frac{1}{\sqrt{P_1 P_2}} & \text{if } P_1 + P_2 = 1, \\ 0 & \text{otherwise.} \end{cases} \tag{19}$$

The treatment for general N runs parallel to the above. Suppose that the λ_l are chosen such that infinitesimal changes in the λ_l generate orthogonal displacements in Q^N–space. This can be done by using hyperspherical co-ordinates, $(r, \phi_1, \phi_2, \ldots, \phi_{N-1})$, with $r = 1$ and, for $l = 1, \ldots, N$, with ϕ_l being a function of λ_l only. In that case, one finds that

$$\begin{aligned} \Delta H &= -\sum_{l=1}^{N-1} \ln(\sigma_{ll}\sqrt{2\pi e}) \\ &= \frac{N-1}{2}\ln\left(\frac{2n}{\pi e}\right) + \sum_{l=1}^{N-1} \ln\left.\frac{\partial s}{\partial \lambda_l}\right|_{\vec{\lambda}(0)} - \ln[\Pr(\lambda_1, \lambda_2, \ldots, \lambda_{N-1}|I)] \\ &= \frac{N-1}{2}\ln\left(\frac{2n}{\pi e}\right) - \ln[\Pr(s_1, s_2, \ldots, s_{N-1}|I)], \end{aligned} \tag{20}$$

where $ds^2 = dQ_1^2 + dQ_2^2 + \cdots + dQ_N^2$ and where $ds_l = (\partial s/\partial\lambda_l)|_{\vec{\lambda}(0)} d\lambda_l$.

Since the λ_l are independent variables, the prior $\Pr(s_1, s_2, \ldots, s_{N-1}|I)$ must be a constant independent of the λ_l. Therefore, any area element, $dA = \prod_{l=1}^{N-1} ds_l$, on S_+^{N-1} is weighted proportionally to its area independent of its location of the unit hypersphere. Therefore, $\Pr(\mathbf{Q}|I)$ is constant whenever $\mathbf{Q}$ lies on S_+^{N-1}, and zero otherwise. Hence,

$$\Pr(P_1, P_2, \ldots, P_N|I) \propto \begin{cases} \frac{1}{\sqrt{P_1 P_2 \ldots P_N}} & \text{if } \sum P_i = 1, \\ 0 & \text{otherwise,} \end{cases} \tag{21}$$

which is Jeffreys' multinomial prior.

To prove the converse, suppose that Jeffreys' multinomial prior is chosen, in which case $\Pr(\mathbf{Q}|I)$ is constant on S_+^{N-1} and zero otherwise. Then let the Euclidean metric be chosen and $\mathbf{Q}$ be parameterised by $\vec{\lambda}$ such that variations in the λ_l generate orthogonal displacements, ds_l, in Q^N–space such that $ds_l \propto d\lambda_l$. Then, $\Pr(\vec{\lambda}|I)$ is uniform and, from Eq. (20), ΔH is independent of $\mathbf{Q}$. Since the value of ΔH is independent of the choice of parameterisation of $\mathbf{P}$, this conclusion is not dependent upon our choice of metric. Hence, the Principle of Information Gain is satisfied. Furthermore, since Jeffreys' rule implies Jeffreys' multinomial prior, it follows that Jeffreys' rule implies the Principle of Information Gain.

3. DERIVATION OF JEFFREYS' RULE

Using the results derived above, we shall now derive Jeffrey's rule. First, we note that, from the constancy of $\Pr(s_1, s_2, \ldots, s_{N-1}|I)$, it follows that the prior over any subset of

the s_l is also constant and, since the s_l have equal range, is proportional to the number of elements in the subset. For example, the priors $\Pr(s_1|I), \Pr(s_2|I), \ldots, \Pr(s_{N-1}|I)$ are equal to one another.

Second, we note that we are free to reparameterise $\mathbf{P}$ by $\vec{\lambda}$ so that the $(N-1)$ mutually orthogonal directions defined by the increasing λ_l are arbitrarily oriented on the unit hypersphere. Hence, for example, line elements of equal length in *any* direction along the unit hypersphere have equal prior weight.

Consider, now, a discrete probability distribution, given by the N dimensional vector $\mathbf{Q}$. Suppose that the distribution is parameterised by the K parameters $\theta_1, \theta_2, \ldots, \theta_K$, where $0 < K < N$. If the parameter θ_k changes by an amount $d\theta_k$, the corresponding displacement of $\mathbf{Q}$ is

$$ds_k = \left|\frac{\partial \mathbf{Q}}{\partial \theta_k}\right| d\theta_k \hat{\theta}_k, \tag{22}$$

with the unit vector

$$\hat{\theta}_k \equiv \frac{\partial \mathbf{Q}}{\partial \theta_k}\left|\frac{\partial \mathbf{Q}}{\partial \theta_k}\right|^{-1}. \tag{23}$$

The distance, ds, traversed by $\mathbf{Q}$ over the surface of the unit hypersphere due to changes $d\theta_1, d\theta_2, \ldots, d\theta_k$ in the θ_k is thus given by

$$ds^2 = \sum_k \left|\frac{\partial \mathbf{Q}}{\partial \theta_k}\right|^2 d\theta_k^2 + 2 \sum_{k,k',k\neq k'} \left|\frac{\partial \mathbf{Q}}{\partial \theta_k}\right| \left|\frac{\partial \mathbf{Q}}{\partial \theta_{k'}}\right| \hat{\theta}_k \cdot \hat{\theta}_{k'} d\theta_k d\theta_{k'}, \tag{24}$$

which, using Eq. (23), can be written as

$$ds^2 = \sum_{k,k'} M_{kk'} d\theta_k d\theta_{k'}, \tag{25}$$

where M is the K–dimensional orthogonal matrix

$$M_{kk'} = \frac{\partial \mathbf{Q}}{\partial \theta_k} \cdot \frac{\partial \mathbf{Q}}{\partial \theta_{k'}}. \tag{26}$$

The patch with K edges $ds_1, ds_2, \ldots, ds_K$ has area $(\det M)^{1/2} d\theta_1 d\theta_2 \ldots d\theta_K$, and the prior probability of $\mathbf{Q}$ being found in this patch is proportional to its area

$$dA = (\det M)^{1/2} d\theta_1 d\theta_2 \ldots d\theta_K. \tag{27}$$

But this prior probability is equal to the prior probability of the parameter vector $\vec{\theta}$ being found within the K–dimensional volume $d\theta_1 d\theta_2 \ldots d\theta_K$, which is given by

$$\Pr(\vec{\theta}|I) d\theta_1 d\theta_2 \ldots d\theta_K. \tag{28}$$

Hence,

$$\Pr(\vec{\theta}|I) \propto (\det M)^{1/2}. \tag{29}$$

A simple calculation shows that, to within a factor of 4, matrix M is equal to the matrix defined in (2). Hence, Eq. (29) is Jeffreys' rule.

4. DISCUSSION

In the previous two sections, we have shown that the Principle of Information Gain is logically equivalent to Jeffreys' rule. We shall now briefly discuss some of the other specific priors and rules for the assignment of prior probabilities which have been suggested.

A number of alternative multinomial priors have been proposed, and we will consider the proposals due to Jaynes [2] and Skilling [3]. First, Jaynes [2] has derived the prior $\Pr(\mathbf{P}|I) \propto 1/P_1P_2\ldots P_N$. The derivation rests on an invariance principle, the so-called transformation group principle, which is intuitively reasonable, and on the auxiliary assumption that a person who is maximally ignorant is one that is in a state of 'complete confusion', an assumption which is not transparent and whose validity can reasonably be questioned.

Second, the entropic prior, $\Pr(\mathbf{P}|I) \propto \exp(-\alpha\sum P_i \ln P_i)/\sqrt{P_1P_2\ldots P_N}$, where α is a free parameter, has been proposed by Skilling [3] and others authors [7, 8]. Since Jeffreys' multinomial prior is a special case of the entropic prior, it follows that Jeffreys' multinomial prior, and so also the Principle of Information Gain, is consistent with the intuitively plausible axioms, particularly axioms I–IV, formulated by Skilling [3]. However, for $\alpha \neq 0$, the entropic prior violates the Principle of Information Gain, which suggests that the entropic prior cannot be viewed as reflecting a state of ignorance when $\alpha \neq 0$, a conclusion which is supported by the fact that, as $\alpha \to \infty$, the entropic prior tends to a delta function centred on $\mathbf{P} = (1/N, 1/N, \ldots, 1/N)$.

Finally, Hartigan [1] has proposed a general rule for the assignment of priors which differs from Jeffreys' rule. In [1], Hartigan first formulates a set of seven intuitively plausible postulates and shows that they are consistent with, but do not imply, Jeffreys' rule, but then makes additional assumptions that are considerably less obvious which, in conjunction with the set of seven postulates, yield an alternative rule. This alternative rule yields Jaynes' multinomial prior and the prior $\Pr(\sigma|I) \propto 1/\sigma^3$ for the standard deviation of a normal distribution. Since Jeffreys' rule yields the prior $\Pr(\sigma|I) \propto 1/\sigma$, this rule conflicts with Jeffreys' rule both through the multinomial prior and through the prior over σ that it generates.

REFERENCES

1. J. Hartigan, *Ann. Math. Stat.*, **35**, 836–845 (1964).
2. E. Jaynes, *IEEE Transactions on Systems Science and Cybernetics*, **SSC-4**, 227–241 (1968).
3. J. Skilling, "Classical Maximum Entropy," in *Maximum Entropy and Bayesian Methods*, 1989.
4. H. Jeffreys, *Proc. Roy. Soc.*, **138**, 48–55 (1932).
5. H. Jeffreys, *Theory of Probability*, Oxford University Press, 1939.
6. H. Jeffreys, *Proc. Roy. Soc. Lon., Ser. A*, **186**, 453–461 (1946).
7. C. Rodríguez, "Are we cruising a hypothesis space?," in *Maximum Entropy and Bayesian Methods*, 1998.
8. A. Caticha, "Maximum Entropy, fluctuations and priors," in *Maximum Entropy and Bayesian Methods*, 2000.
9. E. Jaynes, "Brandeis Lectures," in [10].
10. R. D. Rosenkrantz, editor, *E.T. Jaynes: Papers on Probability, Statistics, and Statistical Physics*, Kluwer, 1983.