On the estimation of discrete distributions

J.-P. Laedermann

Institut de Radiophysique CHUV Lausanne

jp.laedermann@laedus.org
ORCID:0000-0001-8922-8914

25.4.05, transl 20.1.26

Contents

1 Introduction

Let us consider a finite set of n situations that may occur according to a distribution $(p_k)_{k=0}^{n-1}$. The probability of each situation is generally determined by repeating the experiment independently and using Laplace's operational definition to obtain the estimators

$$\widehat{p}_k = \frac{\text{Number of occurrences of situation k}}{\text{Number of repetitions of the experiment}}$$

In addition, the central limit theorem provides an estimate of the difference between the true value and the estimate [1].

This empirical method generally works quite well, but poses a few problems. On the one hand, this estimator and its uncertainty depend on the number of repetitions, giving good results only

when this number is sufficiently large, which is a rather vague concept. On the other hand, it is difficult to generalize this definition when the observation of the situation is unclear. However, a situation is often determined using a measuring instrument that produces a noisy result.

The aim of this work is to base the estimation of probabilities associated with events on Bayes' theorem, starting from Jeffreys' non-informative priors.

2 Ingredients

2.1 Elements of Bayesian theory of statistical decision-making

This theory first distinguishes between two spaces: the space of states θ and the space of observations x [3]. Each observation is assumed to be made in the presence of an unknown state. The *measurement model* is the stochastic relation linking the state to the result of the measurement. For a given state, we give a transition probability[1] $\mathrm{p}(x \mid \theta)$, which gives the distribution of observations x given θ. Given an *a priori* distribution $\mathrm{p}(\theta)$ over states and an observation x, Bayes' theorem gives the *a posteriori* distribution over states

$$\mathrm{p}(\theta \mid x) = \frac{\mathrm{p}(x \mid \theta)\mathrm{p}(\theta)}{\int d\theta \mathrm{p}(x \mid \theta)\mathrm{p}(\theta)}$$

When the state space is a differentiable manifold and the measurement model is twice differentiable with respect to θ, we define a prior distribution called *Jeffreys' prior* as follows [2]. First, we form the *Fisher information matrix*

$$I_{ij}(\theta) = -\int dx \mathrm{p}(x \mid \theta) \frac{\partial^2 \ln \mathrm{p}(x \mid \theta)}{\partial\theta_i \partial\theta_j}$$

Jeffreys' prior is then defined as

$$\mathrm{p}_J(\theta)d\theta = \sqrt{\det I(\theta)}d\theta$$

A fundamental property of this distribution is its invariance. Indeed, it can be shown that $\mathrm{p}_J(\theta)d\theta$ does not depend on the choice of coordinate system.

2.2 Hyperspheres

The estimators constructed in this work make use of unit hyperspheres. Let us recall some of their properties [4].

In Euclidean space $\mathbb{R}^n$, the unit hypersphere of dimension $n-1$ is the differentiable manifold

$$\mathbb{S}^{n-1} = \{\boldsymbol{\psi} \in \mathbb{R}^n \mid \sum_k \psi_k^2 = 1\}$$

The use of the letter ψ will be justified below. The usual spherical coordinates are given by

$$\begin{aligned}
\psi_0 &= \sin\theta_{n-2} \sin\theta_{n-3} \ldots \sin\theta_1 \sin\theta_0 \\
\psi_1 &= \sin\theta_{n-2} \sin\theta_{n-3} \ldots \sin\theta_1 \cos\theta_0 \\
\psi_2 &= \sin\theta_{n-2} \sin\theta_{n-3} \ldots \cos\theta_1 \\
&\ldots \\
\psi_{n-1} &= \cos\theta_{n-2}
\end{aligned} \tag{1}$$

[1] The use of the Roman font p will denote a generic probability, the underlying spaces being identifiable by the variables used.

and the normalized invariant measure under $SO(n)$ is given by

$$d\mu(\boldsymbol{\psi}) = \frac{1}{A_{n-1}} \sin^{n-2}\theta_{n-2} \sin^{n-3}\theta_{n-3} \ldots \sin\theta_1 d\theta_0 d\theta_1 \ldots d\theta_{n-2} \quad (2)$$

$$A_{n-1} = \frac{2\pi^{n/2}}{\Gamma(n/2)} \quad (3)$$

$$\Gamma(1/2) = \sqrt{\pi} \quad (4)$$

Let $\boldsymbol{m} = (m_0 \ldots m_{n-1})$ be a sequence of positive integers. The even moments relative to this measure are given by

$$H(\boldsymbol{m}) = \int \psi_0^{2m_0} \ldots \psi_{n-1}^{2m_{n-1}} d\mu(\boldsymbol{\psi}) \quad (5)$$

We can show that the function H above is a generalization of the beta function

$$H(\boldsymbol{m}) = \frac{\Gamma(n/2)}{\pi^{n/2}} \frac{\prod_k \Gamma(m_k + 1/2)}{\Gamma(\sum_k (m_k + 1/2))} \quad (6)$$

The normalization factor is chosen so that $H(0) = 1$

3 Probability estimators for known situations

The problem is estimating a distribution p_k. This distribution can be viewed as a point in the $n-1$ dimensional simplex

$$\Delta^{n-1} = \{\boldsymbol{p} \mid 0 \leqslant p_k \leqslant 1, \sum_k p_k = 1\}$$

This simplex will be the state space for a measurement model whose observations are situations. In a canonical way, the probability of obtaining situation k given that the distribution is $\boldsymbol{p}$ is obviously

$$\mathrm{p}(k \mid \boldsymbol{p}) = p_k$$

This canonical measurement model is twice differentiable with respect to $\boldsymbol{p}$. We can therefore calculate the associated Jeffreys prior $\mathrm{p}_J(\boldsymbol{p})$. It was stated above that this distribution is independent of the parameterization of the simplex. Let us choose the new coordinates

$$\psi_k = \sqrt{p_k} \quad (7)$$

Since $\sum_k \psi_k^2 = \sum_k p_k = 1$, we obtain a measure model on a state space that is the positive part of the hypersphere $\mathbb{S}^{n-1}$

$$\mathrm{p}(k \mid \boldsymbol{\psi}) = \psi_k^2$$

Now, it turns out that Jeffreys' prior for this model is the invariant measure μ mentioned above:

$$d\mathrm{p}_J(\boldsymbol{\psi}) = d\mu(\boldsymbol{\psi}) \quad (8)$$

This result allows us to apply Bayes' theorem to a sequence of independent observations. Let

$$\boldsymbol{i} = (i_0 \ldots i_{t-1})$$

be such a sequence of length t, we obtain the posterior on the sphere $\mathbb{S}^{n-1}$ given $\boldsymbol{i}$

$$d\mathrm{p}(\boldsymbol{\psi} \mid \boldsymbol{i}) = \frac{\psi_{i_0}^2 \dots \psi_{i_{t-1}}^2 d\mu(\boldsymbol{\psi})}{\int \psi_{i_0}^2 \dots \psi_{i_{t-1}}^2 d\mu(\boldsymbol{\psi})} \tag{9}$$

Introducing the function $\boldsymbol{m}$ giving the multiplicity of an index k in the sequence $\boldsymbol{i}$:

$$\boldsymbol{m}(\boldsymbol{i})_k = \mathrm{Card}\{s \mid i_s = k\}$$

we obtain (implying the sequence $\boldsymbol{i}$ to simplify the notation)

$$d\mathrm{p}(\boldsymbol{\psi} \mid \boldsymbol{i}) = \frac{\prod_k \psi_k^{2m_k} d\mu(\boldsymbol{\psi})}{H(\boldsymbol{m}(\boldsymbol{i}))} \tag{10}$$

The probability of occurrence of a situation k can be estimated by the posterior mean,

$$\widehat{p}_k = \mathbb{E}(p_k) = \frac{\int \psi_k^2 \psi_{i_0}^2 \dots \psi_{i_{t-1}}^2 d\mu(\boldsymbol{\psi})}{H(\boldsymbol{m}(\boldsymbol{i}))} \tag{11}$$

The associated uncertainties will be given by the covariance matrix

$$\widehat{p_k p_l} = \mathbb{E}(p_k p_l) = \frac{\int \psi_k^2 \psi_l^2 \psi_{i_0}^2 \dots \psi_{i_{t-1}}^2 d\mu(\boldsymbol{\psi})}{H(\boldsymbol{m}(\boldsymbol{i}))} \tag{12}$$

Let us define $\boldsymbol{e}_k$ as a sequence of integers that is zero except for the k^{th}, whose value is 1. It is easy to see that

$$\widehat{p}_k = \frac{H(\boldsymbol{m}(\boldsymbol{i}) + \boldsymbol{e}_k)}{H(\boldsymbol{m}(\boldsymbol{i}))} \tag{13}$$

$$\widehat{p_k p_k} = \frac{H(\boldsymbol{m}(\boldsymbol{i}) + \boldsymbol{e}_k + \boldsymbol{e}_l)}{H(\boldsymbol{m}(\boldsymbol{i}))} \tag{14}$$

The fact that $\sum_h m_h = t$ and the properties of the Gamma function [5] immediately give

$$\widehat{p}_k = \frac{m_k + 1/2}{t + n/2} \tag{15}$$

$$\widehat{p_k^2} = \frac{(m_k + 1/2)(m_k + 3/2)}{(t + n/2)(t + n/2 + 1)} \tag{16}$$

$$\widehat{p_k p_l} = \frac{(m_k + 1/2)(m_l + 1/2)}{(t + n/2)(t + n/2 + 1)} \tag{17}$$

4 Self-consistent random sequences

Let $\boldsymbol{i}$ be a sequence of length $t-1$. Equation (15) therefore gives the Bayesian estimates for the p_k if we have observed the sequence $\boldsymbol{i}$, which we will also denote by

$$\widehat{p}_k = \mathrm{p}_J(k \mid i_{t-2}, i_{t-3} \dots i_0)$$

At t, let us randomly draw a new situation i_{t-1} according to these p_k. Let us form a new sequence by concatenation: $(\boldsymbol{i}, i_{t-1})$. Nothing prevents us from repeating the process for this new sequence.

We say that a sequence $\boldsymbol{i}$ is constructed in a **self-consistent** manner if

- i_0 is drawn uniformly from the n situations
- at each step, i_{s-1} is drawn according to the probability $p_J(k \mid i_{s-2} \dots i_0)$

We note that the probability of obtaining a sequence $\boldsymbol{i} = (i_0 \dots i_{t-1})$ in this way is

$$p_{ac}(\boldsymbol{i}) = p_J(i_{t-1} \mid i_{t-2} \dots i_0) p_J(i_{t-2} \mid i_{t-3} \dots i_0) \dots p_J(i_1 \mid i_0) p_J(i_0) \tag{18}$$

Now, it turns out that this probability is given by the function $H \circ \boldsymbol{m}$

$$p_{ac}(\boldsymbol{i}) = H(\boldsymbol{m}(\boldsymbol{i})) \tag{19}$$

Since no additional assumptions are made, we will call the probability p_{ac} **a priori (non-informative) self-consistent** on sequences $\boldsymbol{i}$ of given length t. Sampling according to this prior will be useful later for estimation in uncertain situations.

Note This way of generating a random sequence of numbers provides a possible answer to the question: *Given a sequence, what is its probability of occurrence?* It can be easily seen that the sequences with maximum self-consistent probability are also those with minimum entropy.

Incidentally, for sequences of given length t, we have

$$\sum_{\boldsymbol{i}} H(\boldsymbol{m}(\boldsymbol{i})) = 1$$

From relations (18) and (19), we also have

$$H(\boldsymbol{m}(\boldsymbol{i}) + \boldsymbol{e}_k) = p_J(k \mid \boldsymbol{i}) H(\boldsymbol{m}(\boldsymbol{i})) \tag{20}$$

$$H(\boldsymbol{m}(\boldsymbol{i}) + \boldsymbol{e}_k + \boldsymbol{e}_l) = p_J(k \mid \boldsymbol{i}) p_J(l \mid (\boldsymbol{i}, k)) H(\boldsymbol{m}(\boldsymbol{i})) \tag{21}$$

5 Generalization to fuzzy situations

In reality, situations are often measured by noisy instruments. Instead of knowing the situation at a stage s, the observer only has a measurement result x from a measurement model $\mathrm{p}(x \mid k)$. For a given x, this vector, whose indices are the situations, is often called the likelihood function. For a sequence of steps $s = 0 \dots t-1$, the observations[2] $(x_0 \dots x_{t-1})$ give a **likelihood matrix** whose time is the row index and the situation is the column index

$$\ell^s(k) = \mathrm{p}(x_s \mid k) \tag{22}$$

Let's return to the situation on the hypersphere. Suppose that at time s, the prior probability on $\mathbb{S}^{n-1}$ is $d\eta_s(\boldsymbol{\psi})$. The appearance of a result x_s for the measurement model $\mathrm{p}(x \mid k)$, combined with the canonical model $\mathrm{p}(k \mid \boldsymbol{\psi}) = \psi_k^2$, gives, by Bayes' theorem, an a posteriori that will be taken as a priori in the next step:

$$d\eta_{s+1}(\boldsymbol{\psi}) = \frac{\sum_k \ell^s(k) \psi_k^2 d\eta_s(\boldsymbol{\psi})}{\sum_k \ell^s(k) \int \psi_k^2 d\eta_s(\boldsymbol{\psi})} \tag{23}$$

[2]The measurement model may vary at each step.

We can therefore write

$$d\eta_t(\boldsymbol{\psi}) = \frac{\sum_{\boldsymbol{i}} \ell^0(i_0)\dots\ell^{t-1}(i_{t-1})\psi^2_{i_0}\dots\psi^2_{i_{t-1}}d\mu(\boldsymbol{\psi})}{\sum_{\boldsymbol{i}} \ell^0(i_0)\dots\ell^{t-1}(i_{t-1})\int \psi^2_{i_0}\dots\psi^2_{i_{t-1}}d\mu(\boldsymbol{\psi})} \tag{24}$$

Let

$$\boldsymbol{\ell}(\boldsymbol{i}) = \prod_s \ell^s(i_s)$$

and use the definition of the function H to obtain the estimators at $t-1$

$$\mathbb{E}(p_k) = \frac{\sum_{\boldsymbol{i}} \boldsymbol{\ell}(\boldsymbol{i})H(\boldsymbol{m}(\boldsymbol{i})+\boldsymbol{e}_k)}{\sum_{\boldsymbol{i}} \boldsymbol{\ell}(\boldsymbol{i})H(\boldsymbol{m}(\boldsymbol{i}))} \tag{25}$$

$$\mathbb{E}(p_kp_l) = \frac{\sum_{\boldsymbol{i}} \boldsymbol{\ell}(\boldsymbol{i})H(\boldsymbol{m}(\boldsymbol{i})+\boldsymbol{e}_k+\boldsymbol{e}_l)}{\sum_{\boldsymbol{i}} \boldsymbol{\ell}(\boldsymbol{i})H(\boldsymbol{m}(\boldsymbol{i}))} \tag{26}$$

which can also be written as

$$\mathbb{E}(p_k) = \frac{\sum_{\boldsymbol{i}} p_J(k \mid \boldsymbol{i})\boldsymbol{\ell}(\boldsymbol{i})H(\boldsymbol{m}(\boldsymbol{i}))}{\sum_{\boldsymbol{i}} \boldsymbol{\ell}(\boldsymbol{i})H(\boldsymbol{m}(\boldsymbol{i}))} \tag{27}$$

$$\mathbb{E}(p_kp_l) = \frac{\sum_{\boldsymbol{i}} p_J(k \mid \boldsymbol{i})p_J(l \mid (\boldsymbol{i},k))\boldsymbol{\ell}(\boldsymbol{i})H(\boldsymbol{m}(\boldsymbol{i}))}{\sum_{\boldsymbol{i}} \boldsymbol{\ell}(\boldsymbol{i})H(\boldsymbol{m}(\boldsymbol{i}))} \tag{28}$$

The preceding equations suggest a new use for Bayes' theorem. Indeed, given the self-consistent prior p_{ac} on sequences of length t and the product measure model giving the likelihood function $\boldsymbol{\ell}$, we can form the posterior on the sequences

$$p_{ac}(\boldsymbol{i} \mid \boldsymbol{\ell}) = \frac{\boldsymbol{\ell}(\boldsymbol{i})p_{ac}(\boldsymbol{i})}{\sum_{\boldsymbol{j}} \boldsymbol{\ell}(\boldsymbol{j})p_{ac}(\boldsymbol{j})} \tag{29}$$

and we obtain

$$\mathbb{E}(p_k) = \sum_{\boldsymbol{i}} p_J(k \mid \boldsymbol{i})p_{ac}(\boldsymbol{i} \mid \boldsymbol{\ell}) \tag{30}$$

$$\mathbb{E}(p_kp_l) = \sum_{\boldsymbol{i}} p_J(k \mid \boldsymbol{i})p_J(l \mid (\boldsymbol{i},k))p_{ac}(\boldsymbol{i} \mid \boldsymbol{\ell}) \tag{31}$$

Equations (30) and (31) are integrals that can be calculated using Monte Carlo. It is possible to sample the posterior $p_{ac}(\boldsymbol{i} \mid \boldsymbol{\ell})$ as follows:

- draw i_0 according to the weights $\ell^0(k)$
- draw i_1 according to the weights $\ell^1(k)p_J(k \mid i_0)$
- ...
- draw i_{t-1} according to the weights $\ell^{t-1}(k)p_J(k \mid i_{t-2}\dots i_0)$

A sample of r sequences $\boldsymbol{i}_s$ of this type gives the MC estimators

$$\widehat{\mathbb{E}(p_k)} = \frac{1}{r}\sum_s p_J(k \mid \boldsymbol{i}_s) \tag{32}$$

$$\widehat{\mathbb{E}(p_kp_l)} = \frac{1}{r}\sum_s p_J(k \mid \boldsymbol{i}_s)p_J(l \mid (\boldsymbol{i}_s,k)) \tag{33}$$

as well as MC uncertainties

$$u_{\mathbb{E}(p_k)} = \frac{1}{\sqrt{r}}\sqrt{\frac{1}{r}\sum_s p_J(k \mid \boldsymbol{i}_s)^2 - \widehat{\mathbb{E}(p_k)}^2} \tag{34}$$

$$u_{\mathbb{E}(p_k p_l)} = \frac{1}{\sqrt{r}}\sqrt{\frac{1}{r}\sum_s p_J(k \mid \boldsymbol{i}_s)^2 p_J(l \mid (\boldsymbol{i}_s, l))^2 - \widehat{\mathbb{E}(p_k p_l)}^2} \tag{35}$$

Note Simulating this sampling is very easy, for example in C++, and gives excellent results for likelihood matrices with 3 situations and 10,000 observation times with an r of around one million. The advantage of sampling on the posterior is that it directly locates the sequences that contribute significantly to the overall sum. Recall that the sums (30) and (31) contain n^t terms.

6 Conclusion

Replacing p_k with $\psi_k = \sqrt{p_k}$ makes Jeffreys' prior uniform. In general, probabilistic quantities can be parameterized by the square root, which is their natural expression. This phenomenon is found in quantum mechanics [6], where probability is expressed as the square of the modulus of a wave function, hence the choice of the Greek letter ψ for the points of the hypersphere.

The generalization of beta functions given by the function H bridges the gap between continuous calculus on wave functions and sums over sequences, which are computable by MC.

An application of this work is under development for stationary Markov processes.

7 Summary

7.1 Seen by $\psi \in \mathbb{S}^{n-1}$

$$d\eta^0(\psi) = d\mu(\psi) \text{ invariant} \tag{36}$$

$$d\eta^t(\psi) = d\eta^{t-1}(\psi \mid \ell^{t-1}) = \frac{\sum_k \ell^{t-1}(k)\psi_k^2 d\eta^{t-1}(\psi)}{\sum_k \ell^{t-1}(k)\int \psi_k^2 d\eta^{t-1}(\psi)} \tag{37}$$

$$\mathbb{E}^t(p_k \mid \boldsymbol{\ell}) = \int \psi_k^2 d\eta^t(\psi) \tag{38}$$

$$\mathbb{E}^t(p_k p_l \mid \boldsymbol{\ell}) = \int \psi_k^2 \psi_l^2 d\eta^t(\psi) \tag{39}$$

7.2 Seen by $\boldsymbol{i} \in \{0 \dots n-1\}^{\mathbb{N}}$

$$\begin{aligned}
\boldsymbol{i}^t &= (i_0 \dots i_{t-1}) && (40)\\
p^t_{ac}(\boldsymbol{i}) &= p_J(i_{t-1} \mid \boldsymbol{i}^{t-1}) p^{t-1}_{ac}(\boldsymbol{i}) && (41)\\
\boldsymbol{\ell}^t(\boldsymbol{i}) &= \prod_{s=0}^{t-1} \ell^s(i_s) && (42)\\
p^t_{ac}(\boldsymbol{i} \mid \boldsymbol{\ell}) &= \frac{\boldsymbol{\ell}^t(\boldsymbol{i}) p^t_{ac}(\boldsymbol{i})}{\sum_{\boldsymbol{i}^t} \boldsymbol{\ell}^t(\boldsymbol{i}) p^t_{ac}(\boldsymbol{i})} && (43)\\
\mathbb{E}^t(p_k \mid \boldsymbol{\ell}) &= \sum_{\boldsymbol{i}^t} p_J(k \mid \boldsymbol{i}^t) p^t_{ac}(\boldsymbol{i} \mid \boldsymbol{\ell}) && (44)\\
\mathbb{E}^t(p_k p_l \mid \boldsymbol{\ell}) &= \sum_{\boldsymbol{i}^t} p_J(l \mid \boldsymbol{i}^t k) p_J(k \mid \boldsymbol{i}^t) p^t_{ac}(\boldsymbol{i} \mid \boldsymbol{\ell}) && (45)
\end{aligned}$$

References

[1] Chung K. L. A course in probability theory, chap 7
Academic Press 1968

[2] Bernardo J.M., Smith A.F.M. Bayesian theory p 314, 358
Wiley 1994

[3] Laedermann J.-P. Théorie bayesienne de la décision statistique et mesure de la radioactivité
UNIL, Thèse de doctorat, 2003

[4] Hyperspheric coordinates
https://en.wikipedia.org/wiki/N-sphere Spherical coordinates

[5] Gradshtyn L.S., Ryzbik L.M. Table of integrals, series, and products, 6.4
Alan Jeffrey 2000

[6] Piron C. Quantum Mechanics
Presses polytechniques et universitaires romandes 1998