entropy

MDPI

Article

Maximum-Entropy Priors with Derived Parameters in a Specified Distribution

Will Handley [1,2,3,*] **and Marius Millea** [4,5]

1 Astrophysics Group, Cavendish Laboratory, J.J.Thomson Avenue, Cambridge CB3 0HE, UK
2 Kavli Institute for Cosmology, Madingley Road, Cambridge CB3 0HA, UK
3 Gonville & Caius College, Trinity Street, Cambridge CB2 1TA, UK
4 Institut dAstrophysique de Paris (IAP), UMR 7095, CNRS UPMC Universit Paris 6, Sorbonne Universits, 98bis Boulevard Arago, F-75014 Paris, France; mariusmillea@somewhere.com
5 Institut Lagrange de Paris (ILP), Sorbonne Universits, 98bis Boulevard Arago, F-75014 Paris, France
* Correspondence: wh260@cam.ac.uk

Received: 1 February 2019; Accepted: 8 March 2019; Published: 12 March 2019

Abstract: We propose a method for transforming probability distributions so that parameters of interest are forced into a specified distribution. We prove that this approach is the maximum-entropy choice, and provide a motivating example, applicable to neutrino-hierarchy inference.

Keywords: maximum entropy; Bayesian inference; prior; derived distribution; neutrino hierarchy

1. Introduction

In Bayesian analysis, a simple prior on inference parameters can induce a nontrivial prior on critical physical parameters of interest. This arises, for example, when estimating the masses of neutrinos from cosmological observations. Here, three parameters are inferred corresponding to the mass of each of the three neutrino species, (m_1, m_2, m_3). Cosmological observations, however, are mainly sensitive to their sum, $m_1 + m_2 + m_3$. Simple priors, for example, log-uniform priors on individual masses, can induce undesired informative priors on their sum [1].

Another example arises in nonparametric reconstructions. Here, one infers underlying physical function from the data, where the data are a reprocessing of the target function by some physical or instrumental transfer function. Typical approaches involve decomposing the target function into bins, principal component eigenmodes, or generally into any other basis functions. Simple priors on the amplitudes of basis functions can lead to undersized priors on physical quantities derived from the target function. Consideration of these effects is particularly important, for example, when reconstructing the history of cosmic reionization [2].

A natural remedy is to importance-weight the original prior such that the nontrivial distribution on the parameter of interest is transformed to a more desirable one. In this paper, we show that this natural approach is the maximum-entropy prior distribution [3]. Often, the more desirable prior is a uniform distribution, but our proof also holds for any desired target distribution. Our observation provides a powerful justification for the natural solution, as it is the distribution that assumes the least information, and is therefore particularly appropriate for choosing priors [4].

In Section 2, we demonstrate the key ideas with a toy example before providing a rigorous proof in Section 3. We then apply these ideas to a more complicated example, appropriate for constructing priors on neutrino masses, in Section 4.

2. Motivating Example

We begin with a simplified example. Consider a system with two parameters (a, b), with a uniform distribution $q(a, b)$ on the unit square. Analogous to the sum of neutrino masses mentioned earlier, suppose that a derived parameter, $c = a + b$, is of physical interest. Effective distribution $q(a + b)$ is not uniform, but instead symmetric and triangular between $0 < c < 2$, as graphically illustrated in the left-hand side of Figure 1. If one wished to construct a distribution $p(a, b)$ that was uniform in $a + b$, one could do so by dividing out the triangular distribution:

$$p(a, b) = \frac{q(a, b)}{q(a + b)}. \tag{1}$$

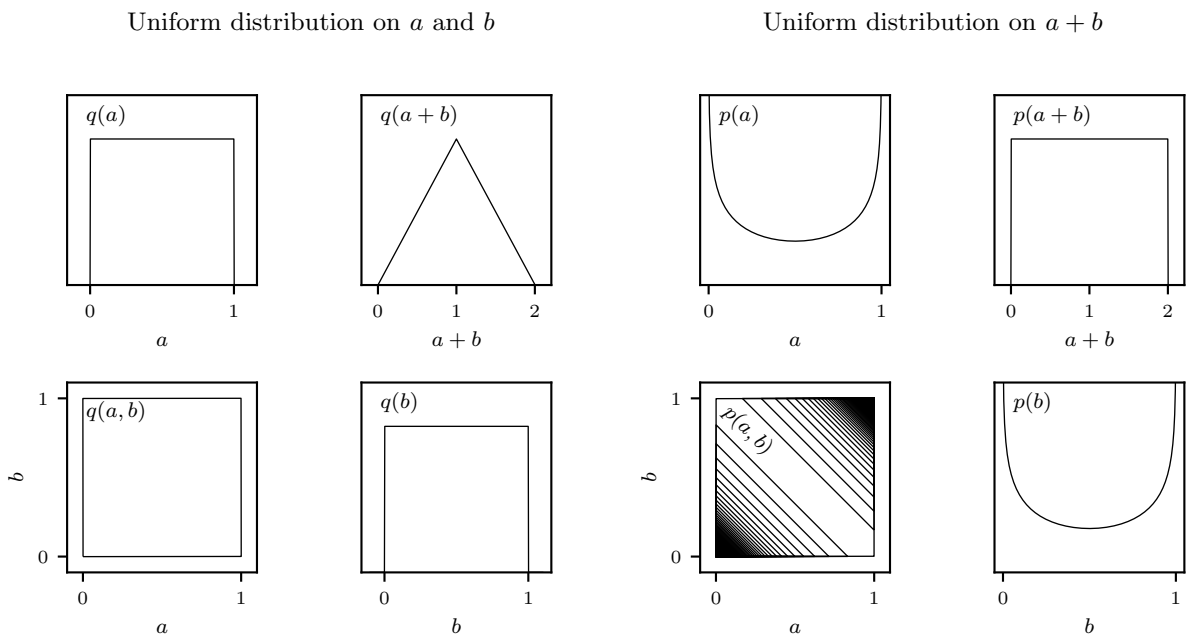

Figure 1. (Left-hand panels) uniform distribution on two parameters $q(a, b)$ inducing triangular distribution on their sum $q(a + b)$. (Right-hand panels) constructing new distribution by dividing out triangular distribution $p(a, b) = q(a, b)/q(a + b)$ renders a uniform distribution induced on the sum $p(a + b)$. Figures were constructed from the analytic forms of the distributions in Python using the Matplotlib package [5].

The resulting transformed distribution is illustrated in the right-hand side of Figure 1. More weight is given to low and higher values of a and b, so that the tails of triangular distribution $q(a + b)$ are counterbalanced. This comes at the price of altering the marginal distributions of a and b, which become $p(a) = -\log[a(1 - a)]/2$ (similarly for b), but which now give a uniform prior, $p(a + b)$. The transformation can be viewed as an importance weighting of the original distribution, and is intuitively the simplest way to force $p(a + b)$ to be uniform.

The aim of this paper is to show that the above intuition is well-founded, as (1) is in fact the maximum-entropy solution. The entropy of a distribution $p(x)$ with respect to an underlying measure $q(x)$ is:

$$H(p|q) = \int \mathrm{d}x\, p(x) \log \left[\frac{q(x)}{p(x)} \right]. \tag{2}$$

The maximum-entropy approach [6,7] finds distribution p that maximises H, subject to user-specified constraints. As it maximises entropy, solution p is generally interpreted as the distribution that assumes the least information given the constraints.

In the next section, we show that (1) is the maximum-entropy solution, subject to the constraint that $p(a+b)$ is uniform. We further generalize to a derived parameter that can be any arbitrary function of the original parameters, for which the desired distribution is in general nonuniform.

In a more usual maximum-entropy setting, user-applied constraints typically take the form of either a domain restriction such as $x \in [-1, 1]$ or $x > 0$, or linear functions of distribution p, such as a specified mean $\mu = \int x p(x)\,\mathrm{d}x$, or variance $\sigma^2 = \int (x-\mu)^2 p(x)\,\mathrm{d}x$. In this work, our constraints contrast with the traditional approach in that, instead of a discrete set of constraints, by demanding that a derived parameter has a distribution in a specified functional form, our constraints form a continuum. In other words, instead of a discrete set of Lagrange multipliers, one must introduce a continuous Lagrange multiplier function.

3. Mathematical Proof

Theorem 1. *If one has a D-dimensional distribution on parameters x with probability density function $q(x)$ along with a derived parameter f defined by a function $f = f(x)$, then the maximum-entropy distribution $p(x)$ relative to $q(x)$ satisfying the constraint that f is distributed with probability density function to $r(f)$ is:*

$$p(x) = \frac{q(x) r(f(x))}{P(f(x)|q)}, \tag{3}$$

where $P(f|q)$ is the probability density for the distribution induced by q on $f = f(x)$.

Proof. If we have some function $f(x)$ defining a derived parameter $f = f(x)$, then cumulative density function $C(f|p)$ of $f = f(x)$ induced by p can be expressed as a D-dimensional integral over the region $\Omega(f) = \{x : f(x) < f\}$ with D-dimensional volume element $\mathrm{d}x$:

$$C(f|p) = \int_{f(x)<f} \mathrm{d}x\, p(x). \tag{4}$$

Differentiating (4) with respect to f yields the probability density function of f induced by p, which via the Leibniz integral rule can be expressed as a $(D-1)$-dimensional integral over the boundary surface $\partial\Omega(f) = \{x : f(x) = f\}$, with the induced $(D-1)$-dimensional volume element $\mathrm{d}S(x)$:

$$P(f|p) = \frac{\mathrm{d}}{\mathrm{d}f} C(f|p) \equiv \int_{f(x)=f} \mathrm{d}S(x)\, p(x). \tag{5}$$

We aim to find distribution p that maximises entropy $H(p|q)$ from (2), subject to the constraint that $P(f|p)$ takes a given form with probability density $r(f)$ and cumulative density $c(f)$:

$$C(f|p) = c(f) \quad \Leftrightarrow \quad P(f|p) = r(f) = c'(f) \tag{6}$$

The solution can be obtained via the method of Lagrange multipliers, wherein we maximise the functional F:

$$F(p) = H(p|q) - \lambda \int p(x)\,\mathrm{d}x - \int \mathrm{d}f\, \mu(f) \int_{f(x)<f} \mathrm{d}x\, p(x), \tag{7}$$

subject to normalisation and distribution constraints

$$\int p(x)\,\mathrm{d}x = 1, \tag{8}$$

$$\int_{f(x)<f} p(x)\,\mathrm{d}x = c(f) \quad \Leftrightarrow \quad \int_{f(x)=f} \mathrm{d}S(x)\,p(x) = r(f). \tag{9}$$

Here, we introduced a Lagrange multiplier λ for the normalisation constraint (8), and a continuous set of Lagrange multipliers $\mu(f)$ for the distribution constraints (9).

Functionally differentiating (7) yields:

$$0 = \frac{\delta F}{\delta p(x)} = -1 + \log\frac{p(x)}{q(x)} - \lambda - \int_{f(x)<f} \mathrm{d}f\,\mu(f), \tag{10}$$

$$\Rightarrow p(x) = q(x)e^{1+\lambda+\int_{f(x)<f}\mathrm{d}f\,\mu(f)} = q(x)M(f(x)), \tag{11}$$

where in (10) we have used the fact that:

$$\frac{\delta}{\delta p(x)}\int_{f(x')<f}\mathrm{d}x'\,p(x') = \begin{cases} 1 & : f(x) < f \\ 0 & : \text{otherwise,} \end{cases} \tag{12}$$

and, in (11), defined the new function:

$$\log M(g) = 1 + \lambda + \int_{g<f}\mathrm{d}f\,\mu(f). \tag{13}$$

All that remains to be done is to determine M from Constraints (8) and (9). Taking the right-hand form of distribution Constraint (9), and substituting in $p(x) = q(x)M(f(x))$ from (11), we find:

$$r(f) = \int_{f(x)=f}\mathrm{d}S(x)\,q(x)M(f(x)) = M(f)\int_{f(x)=f}\mathrm{d}S(x)\,q(x) = M(f)P(f|q), \tag{14}$$

where we have used the fact that $M(f(x))$ is constant over the surface $f(x) = f$, and Definition (5) for a constrained probability distribution function. We now have the form of M to substitute into (11), yielding Solution (3). □

Result (3) is precisely what one would expect. The distribution that converts $q(x)$ to one, which instead has $f = f(x)$ distributed according to $r(f)$, is found by first dividing out the distribution on f induced by q, and then modulating by desired distribution $r(f)$.

Provided that $r(f)$ is correctly normalised, Expression (3) automatically satisfies normalisation Constraint (8):

$$\begin{aligned}\int \mathrm{d}x\,\frac{q(x)r(f(x))}{P(f(x)|q)} &= \int \mathrm{d}f \int_{f(x)=f}\mathrm{d}S(x)\frac{q(x)r(f(x))}{P(f(x)|q)} \\ &= \int \mathrm{d}f\,\frac{r(f)}{P(f|q)}\int_{f(x)=f}\mathrm{d}S(x)q(x) = \int \mathrm{d}f\,r(f) = 1.\end{aligned} \tag{15}$$

In the above, we first split the volume integral into a set of nested surface integrals, drew out the functions that were constant over the surfaces, applied the definition of induced probability density $P(f|q)$, and then used the normalisation of r. A similar manipulation may be used to confirm that functional Form (3) satisfies distribution Constraint (9).

The proof may be generalised to multiple derived parameters without modification, simply taking $f = f(x)$ to represent a vector relationship, and the cumulative distribution functions to be their multiparameter equivalents.

4. Example: Neutrino Masses

In the past year, there has been interest in the cosmological and particle-physics community regarding the correct prior to put on neutrino masses. Simpson et al. [8] controversially claimed that, with current cosmological parameter constraints ($\sum_\nu m_\nu < 0.13$ eV [9,10]), the normal hierarchy of masses was strongly preferred over an inverted hierarchy, in contrast with the results of Vagnozzi et al. [11]. Later, Schwetz et al. [1] showed that the controversial claim was mostly due to a nontrivial prior that had been put on the neutrino masses. Since then, other choices of prior have been proposed by Caldwell et al. [12], Long et al. [13], Gariazzo et al. [14] and Heavens and Sellentin [15], which reduce the strength of the claim.

Using our methodology, a possible alternative prior to put on the masses can be constructed. Typically, one chooses a broad independent logarithmic prior on each of the masses of the three neutrinos (m_1, m_2, m_3). However, cosmological probes of the neutrino masses typically place a constraint on the sum of the masses $m_1 + m_2 + m_3$. Simple logarithmic priors on the masses place a nontrivial prior on their sum. Using our approach, we can transform the initial distribution into one that has more reasonable distribution on the sum of the masses. Such considerations can be particularly important when determining the strength of cosmological probes.

A concrete example is illustrated in Figure 2. As the original distribution, we take an independent Gaussian prior on the logarithm of the masses. This induces nontrivial distribution on the sum of the masses, approximately log-normal, but with a shifted centre. If one demands that the sum of the masses is instead centred on zero, then the maximum-entropy approach creates a distribution with tails toward low masses in order to compensate for the upward shift in the distribution of the sum of the masses. This tail enters a region of parameter space that would be completely excluded by the original prior; thus, choosing the transformed prior could influence the strength of a given inference on the nature of the neutrino hierarchy. It should be noted that we are not advocating this as the most suitable prior to put on neutrino masses, but merely to show that you may use our procedure to straightforwardly transform a distribution, should one wish to put a flat prior on the sum of the masses. A more physical cosmological example in the context of reionization reconstruction can be found in Millea and Bouchet [2].

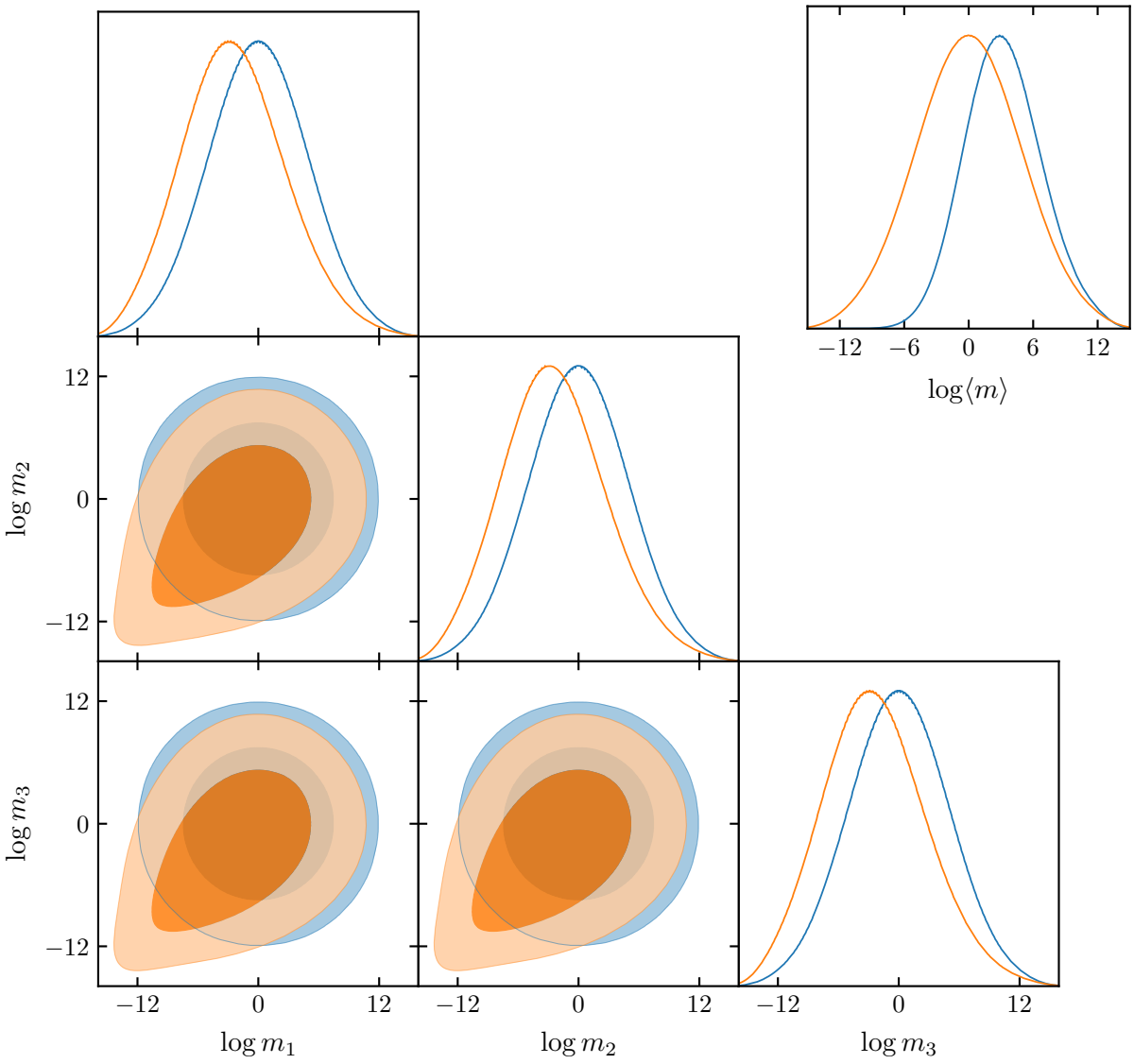

Figure 2. Distribution q (illustrated in blue) is defined as a three-dimensional spherical Gaussian on the logarithm of parameters m_1, m_2, m_3, centred on zero with a standard deviation of five log units on each parameter. Nontrivial distribution is induced on the mean of masses $\langle m \rangle = \frac{1}{3}(m_1 + m_2 + m_3)$, which is approximately log-normal, but with a shifted centre and width. If one demands that the mean of the masses is log-normal centred on zero with width five, as for the original individual masses, then the maximum-entropy approach creates the distribution p, illustrated in orange. Parameters are forced to have a tail toward low values in order to compensate for the upward shift in q-mean distribution.

5. Conclusions

In this paper, we proposed an approach for transforming probability distribution to force a derived parameter into a specified distribution. One importance-weights the original distribution by dividing out the induced distribution on the parameter of interest, and reweights by the desired distribution. We proved that the resulting distribution is the maximum-entropy choice. Finally, we provided some motivating examples.

Author Contributions: conceptualization, M.M.; methodology, W.H. and M.M.; software, W.H. and M.M.; validation, W.M. and M.M.; formal analysis, W.H.; investigation, W.H. and M.M.; writing—original draft preparation, W.H.; writing—review and editing, M.M.; visualization, W.H. and M.M.; project administration, W.H.

Funding: W.H. was supported by a Gonville and Caius College research fellowship. M.M. was supported by the Labex ILP (reference ANR-10-LABX-63).

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

References

1. Schwetz, T.; Freese, K.; Gerbino, M.; Giusarma, E.; Hannestad, S.; Lattanzi, M.; Mena, O.; Vagnozzi, S. Comment on "Strong Evidence for the Normal Neutrino Hierarchy". *arXiv* **2017**, arXiv:1703.04585.
2. Millea, M.; Bouchet, F. Cosmic microwave background constraints in light of priors over reionization histories. *Astron. Astrophys.* **2018**, *617*, A96. [CrossRef]
3. Jeffreys, H. *The Theory of Probability*; Oxford Classic Texts in the Physical Sciences; OUP Oxford: Oxford, UK, 1998.
4. Sivia, D.; Skilling, J. *Data Analysis: A Bayesian Tutorial*; Oxford Science Publications; Oxford University Press: Oxford, UK, 2006.
5. Hunter, J.D. Matplotlib: A 2D graphics environment. *Comput. Sci. Eng.* **2007**, *9*, 90–95. [CrossRef]
6. Jaynes, E.T. Information Theory and Statistical Mechanics. *Phys. Rev.* **1957**, *106*, 620–630. [CrossRef]
7. Jaynes, E.T. Information Theory and Statistical Mechanics. II. *Phys. Rev.* **1957**, *108*, 171–190. [CrossRef]
8. Simpson, F.; Jimenez, R.; Pena-Garay, C.; Verde, L. Strong Bayesian evidence for the normal neutrino hierarchy. *J. Cosmol. Astropart. Phys.* **2017**, *6*, 029. [CrossRef]
9. Cuesta, A.J.; Niro, V.; Verde, L. Neutrino mass limits: Robust information from the power spectrum of galaxy surveys. *Phys. Dark Univ.* **2016**, *13*, 77–86. [CrossRef]
10. Palanque-Delabrouille, N.; Yèche, C.; Baur, J.; Magneville, C.; Rossi, G.; Lesgourgues, J.; Borde, A.; Burtin, E.; LeGoff, J.M.; Rich, J.; et al. Neutrino masses and cosmology with Lyman-alpha forest power spectrum. *J. Cosmol. Astropart. Phys.* **2015**. [CrossRef] *2015*, 011,
11. Vagnozzi, S.; Giusarma, E.; Mena, O.; Freese, K.; Gerbino, M.; Ho, S.; Lattanzi, M. Unveiling ν secrets with cosmological data: Neutrino masses and mass hierarchy. *Phys. Rev. D* **2017**, *96*, 123503. [CrossRef]
12. Caldwell, A.; Merle, A.; Schulz, O.; Totzauer, M. Global Bayesian analysis of neutrino mass data. *Phys. Rev. D* **2017**, *96*, 073001. [CrossRef]
13. Long, A.J.; Raveri, M.; Hu, W.; Dodelson, S. Neutrino mass priors for cosmology from random matrices. *Phys. Rev. D* **2018**, *97*, 043510. [CrossRef]
14. Gariazzo, S.; Archidiacono, M.; de Salas, P.F.; Mena, O.; Ternes, C.A.; Tórtola, M. Neutrino masses and their ordering: Global data, priors and models. *J. Cosmol. Astropart. Phys.* **2018**, *3*, 011. [CrossRef]
15. Heavens, A.F.; Sellentin, E. Objective Bayesian analysis of neutrino masses and hierarchy. *J. Cosmol. Astropart. Phys.* **2018**, *4*, 047. [CrossRef]