

# Introduction to Bayesian Methods - 4

*Edoardo Milotti*

Università di Trieste and INFN-Sezione di Trieste

*Edwin T. Jaynes (1922-1998), introduced the method of maximum entropy in statistical mechanics: when we start from the informational entropy (Shannon's entropy) and we use it to introduce Boltzmann's entropy we obtain again the whole of statistical mechanics by maximizing entropy.*

*In a sense, statistical mechanics also arises from a comprehensive "principle of maximum entropy".*

<http://bayes.wustl.edu/etj/etj.html>



## Information Theory and Statistical Mechanics

E. T. JAYNES

*Department of Physics, Stanford University, Stanford, California*

(Received September 4, 1956; revised manuscript received March 4, 1957)

Information theory provides a constructive criterion for setting up probability distributions on the basis of partial knowledge, and leads to a type of statistical inference which is called the maximum-entropy estimate. It is the least biased estimate possible on the given information; i.e., it is maximally noncommittal with regard to missing information. If one considers statistical mechanics as a form of statistical inference rather than as a physical theory, it is found that the usual computational rules, starting with the determination of the partition function, are an immediate consequence of the maximum-entropy principle. In the resulting "subjective statistical mechanics," the usual rules are thus justified independently of any physical argument, and in particular independently of experimental verification; whether

or not the results agree with experiment, they still represent the best estimates that could have been made on the basis of the information available.

It is concluded that statistical mechanics need not be regarded as a physical theory dependent for its validity on the truth of additional assumptions not contained in the laws of mechanics (such as ergodicity, metric transitivity, equal *a priori* probabilities, etc.). Furthermore, it is possible to maintain a sharp distinction between its physical and statistical aspects. The former consists only of the correct enumeration of the states of a system and their properties; the latter is a straightforward example of statistical inference.

## 2. MAXIMUM-ENTROPY ESTIMATES

The quantity  $x$  is capable of assuming the discrete values  $x_i$  ( $i=1,2,\dots,n$ ). We are not given the corresponding probabilities  $p_i$ ; all we know is the expectation value of the function  $f(x)$ :

$$\langle f(x) \rangle = \sum_{i=1}^n p_i f(x_i). \quad (2-1)$$

On the basis of this information, what is the expectation value of the function  $g(x)$ ? At first glance, the problem seems insoluble because the given information is insufficient to determine the probabilities  $p_i$ .<sup>5</sup> Equation (2-1) and the normalization condition

$$\sum p_i = 1 \quad (2-2)$$

would have to be supplemented by  $(n-2)$  more conditions before  $\langle g(x) \rangle$  could be found.

This problem of specification of probabilities in cases where little or no information is available, is as old as the theory of probability. Laplace's "Principle of Insufficient Reason" was an attempt to supply a criterion of choice, in which one said that two events are to be assigned equal probabilities if there is no reason to think otherwise. However, except in cases where there is an evident element of symmetry that clearly renders the events "equally possible," this assumption may appear just as arbitrary as any other that might be made. Furthermore, it has been very fertile in generating paradoxes in the case of continuously variable random quantities,<sup>6</sup> since intuitive notions of "equally possible" are altered by a change of variables.<sup>7</sup> Since the time of Laplace, this way of

<sup>5</sup> Yet this is precisely the problem confronting us in statistical mechanics; on the basis of information which is grossly inadequate to determine any assignment of probabilities to individual quantum states, we are asked to estimate the pressure, specific heat, intensity of magnetization, chemical potentials, etc., of a macroscopic system. Furthermore, statistical mechanics is amazingly successful in providing accurate estimates of these quantities. Evidently there must be other reasons for this success, that go beyond a mere correct statistical treatment of the problem as stated above.

Here we apply the maximum entropy principle (MaxEnt) to solve problems and find prior distributions ...

*The kangaroo problem (Jaynes)*

- *Basic information*: one third of all kangaroos has blue eyes, and one third is left-handed.
- *Question*: which fraction of kangaroos has both blue eyes and is left-handed?
- *Constraints*: the normalization condition must be fulfilled matrixwise + the constraints expressed by the basic information, row by row and column by column.



	left	~left
blue	1/9	2/9
~blue	2/9	4/9

statistical independence

	left	~left
blue	0	1/3
~blue	1/3	1/3

maximum negative correlation

	left	~left
blue	1/3	0
~blue	0	2/3

maximum positive correlation

probabilities  $p_{bl}$   $p_{\bar{b}l}$   $p_{b\bar{l}}$   $p_{\bar{b}\bar{l}}$

entropy (proportional to Shannon's entropy)

$$S = p_{bl} \ln \frac{1}{p_{bl}} + p_{\bar{b}l} \ln \frac{1}{p_{\bar{b}l}} + p_{b\bar{l}} \ln \frac{1}{p_{b\bar{l}}} + p_{\bar{b}\bar{l}} \ln \frac{1}{p_{\bar{b}\bar{l}}}$$

constraints (3 constraints, 4 unknowns)

$$p_{bl} + p_{\bar{b}l} + p_{b\bar{l}} + p_{\bar{b}\bar{l}} = 1$$

$$p_{bl} + p_{b\bar{l}} = 1/3$$

$$p_{\bar{b}l} + p_{\bar{b}\bar{l}} = 1/3$$

# entropy maximization with constraints

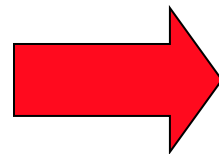
$$S_V = \left( p_{bl} \ln \frac{1}{p_{bl}} + p_{\bar{bl}} \ln \frac{1}{p_{\bar{bl}}} + p_{b\bar{l}} \ln \frac{1}{p_{b\bar{l}}} + p_{\bar{b}\bar{l}} \ln \frac{1}{p_{\bar{b}\bar{l}}} \right) \\ + \lambda_1 (p_{bl} + p_{\bar{bl}} + p_{b\bar{l}} + p_{\bar{b}\bar{l}} - 1) + \lambda_2 (p_{bl} + p_{b\bar{l}} - 1/3) + \lambda_3 (p_{bl} + p_{\bar{bl}} - 1/3)$$

$$\frac{\partial S_V}{\partial p_{bl}} = -\ln p_{bl} - 1 + \lambda_1 + \lambda_2 + \lambda_3 = 0$$

$$\frac{\partial S_V}{\partial p_{\bar{bl}}} = -\ln p_{\bar{bl}} - 1 + \lambda_1 + \lambda_3 = 0$$

$$\frac{\partial S_V}{\partial p_{b\bar{l}}} = -\ln p_{b\bar{l}} - 1 + \lambda_1 + \lambda_2 = 0$$

$$\frac{\partial S_V}{\partial p_{\bar{b}\bar{l}}} = -\ln p_{\bar{b}\bar{l}} - 1 + \lambda_1 = 0$$



$$p_{bl} = \exp(-1 + \lambda_1 + \lambda_2 + \lambda_3)$$

$$p_{\bar{bl}} = \exp(-1 + \lambda_1 + \lambda_3)$$

$$p_{b\bar{l}} = \exp(-1 + \lambda_1 + \lambda_2)$$

$$p_{\bar{b}\bar{l}} = \exp(-1 + \lambda_1)$$

$$\begin{cases} p_{\bar{b}l} = p_{\bar{b}l} \exp(\lambda_3) \\ p_{b\bar{l}} = p_{\bar{b}l} \exp(\lambda_2) \\ p_{bl} = p_{\bar{b}l} \exp(\lambda_2 + \lambda_3) \end{cases} \Rightarrow p_{\bar{b}l} p_{b\bar{l}} = p_{bl} p_{\bar{b}l}$$

$$\begin{cases} p_{bl} + p_{\bar{b}l} + p_{b\bar{l}} + p_{\bar{b}\bar{l}} = 1 \\ p_{bl} + p_{b\bar{l}} = 1/3 \\ p_{bl} + p_{\bar{b}l} = 1/3 \\ p_{\bar{b}l} p_{b\bar{l}} = p_{bl} p_{\bar{b}\bar{l}} \end{cases} \Rightarrow \begin{cases} p_{b\bar{l}} = p_{\bar{b}l} = 1/3 - p_{bl} \\ p_{\bar{b}\bar{l}} = 1/3 + p_{bl} \\ (1/3 - p_{bl})^2 = p_{bl}/3 + p_{bl}^2 \\ 1/9 - 2p_{bl}/3 + p_{bl}^2 = p_{bl}/3 + p_{bl}^2 \end{cases}$$

$$\Rightarrow p_{bl} = \frac{1}{9}; \quad p_{b\bar{l}} = p_{\bar{b}l} = \frac{2}{9}; \quad p_{\bar{b}\bar{l}} = \frac{4}{9}$$

this solution coincides with the least informative distribution given the constraints (statistically independent variables)



What do we learn about Statistical Mechanics using the MaxEnt method?

$$H = -K \sum_i p_i \ln p_i, \quad \text{with} \quad \sum_i p_i = 1 \quad \text{and} \quad \langle f(x) \rangle = \sum_i f(x_i) p_i$$



$$Q = H + K(-\lambda + 1) \sum_i p_i - K\mu \sum_i f(x_i) p_i$$



$$\frac{\partial Q}{\partial p_i} = -(\ln p_i + 1) + (-\lambda + 1) - \mu f(x_i) = 0$$



$$p_i = \exp(-\lambda - \mu f(x_i))$$



$$\sum_i p_i = e^{-\lambda} \sum_i e^{-\mu f(x_i)} = 1 \quad \text{then, letting} \quad Z(\mu) = \sum_i e^{-\mu f(x_i)} \quad \lambda = \ln Z(\mu)$$



$$\langle f(x) \rangle = -\frac{\partial}{\partial \mu} \ln Z(\mu)$$

The principle of maximum entropy may be regarded as an extension of the principle of insufficient reason (to which it reduces in case no information is given except enumeration of the possibilities  $x_i$ ), with the following essential difference. The maximum-entropy distribution may be asserted for the positive reason that it is uniquely determined as the one which is maximally noncommittal with regard to missing information, instead of the negative one that there was no reason to think otherwise. Thus the concept of entropy supplies the missing criterion of choice which Laplace needed to remove the apparent arbitrariness of the principle of insufficient reason, and in addition it shows precisely how this principle is to be modified in case there are reasons for “thinking otherwise.”

## *Solution of underdetermined systems of equations*

In this problem there are fewer equations than unknowns; the system of equations is underdetermined, and in general there is no unique solution.

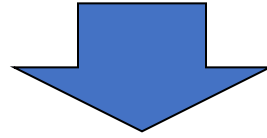
The maximum entropy method helps us find a reasonable solution, the least informative one (least correlation between variables)

*Example:*

$$\begin{aligned} 3x + 5y + 1.1z &= 10 \\ -2.1x + 4.4y - 10z &= 1 \end{aligned} \quad (x, y, z > 0)$$

$$\begin{aligned}
 3x + 5y + 1.1z &= 10 \\
 -2.1x + 4.4y - 10z &= 1
 \end{aligned}
 \quad (x, y, z > 0)$$

this ratio can be taken to be a  
"probability"



$$\begin{aligned}
 S &= - \left( \frac{x}{x+y+z} \ln \frac{x}{x+y+z} + \frac{y}{x+y+z} \ln \frac{y}{x+y+z} + \frac{z}{x+y+z} \ln \frac{z}{x+y+z} \right) \\
 &= - \frac{1}{x+y+z} \left[ x \ln x + y \ln y + z \ln z - (x+y+z) \ln(x+y+z) \right]
 \end{aligned}$$

$$Q = S + \lambda(3x + 5y + 1.1z - 10) + \mu(-2.1x + 4.4y - 10z - 1)$$

$$\begin{aligned}
 \frac{\partial Q}{\partial x} &= - \frac{\ln x - \ln(x+y+z)}{x+y+z} + \frac{x \ln x + y \ln y + z \ln z - (x+y+z) \ln(x+y+z)}{(x+y+z)^2} + 3\lambda - 2.1\mu \\
 &= \frac{(y+z) \ln x + y \ln y + z \ln z}{(x+y+z)^2} + 3\lambda - 2.1\mu = 0
 \end{aligned}$$

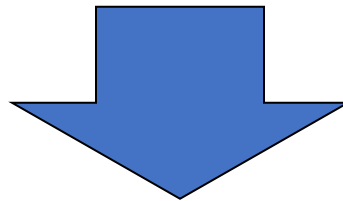
$$\frac{\partial Q}{\partial x} = \frac{(y+z)\ln x + y\ln y + z\ln z}{(x+y+z)^2} + 3\lambda - 2.1\mu = 0$$

$$\frac{\partial Q}{\partial y} = \frac{x\ln x + (x+z)\ln y + z\ln z}{(x+y+z)^2} + 5\lambda + 4.4\mu = 0$$

$$\frac{\partial Q}{\partial z} = \frac{x\ln x + y\ln y + (x+y)\ln z}{(x+y+z)^2} + 1.1\lambda - 10\mu = 0$$

$$10 = 3x + 5y + 1.1z$$

$$1 = -2.1x + 4.4y - 10z$$



$$x = 0.606275; \quad y = 1.53742; \quad z = 0.449148;$$
$$\lambda = 0.0218739; \quad \mu = -0.017793$$

this is an example of an “ill-posed” problem

the Maximum Entropy Method is a kind of **regularization** of the ill-posed problem

# Finding priors with the maximum entropy method

$$S = \sum_k p_k \ln \frac{1}{p_k} = -\sum_k p_k \ln p_k \quad \text{Shannon's entropy}$$

entropy maximization when all information is missing,  
and normalization is the only constraint:


$$\frac{\partial}{\partial p_k} \left[ -\sum_k p_k \ln p_k + \lambda \left( \sum_k p_k - 1 \right) \right] = -(\ln p_k + 1) + \lambda = 0$$

$$p_k = e^{\lambda-1}; \quad \sum_k p_k = \sum_k e^{\lambda-1} = N e^{\lambda-1} = 1 \quad \Rightarrow \quad p_k = 1/N$$

## entropy maximization when the mean is known $\mu$

$$\frac{\partial}{\partial p_k} \left[ -\sum_k p_k \ln p_k + \lambda_0 \left( \sum_k p_k - 1 \right) + \lambda_1 \left( \sum_k x_k p_k - \mu \right) \right]$$
$$= -(\ln p_k + 1) + \lambda_0 + \lambda_1 x_k = 0$$

incomplete  
solution...


$$p_k = e^{\lambda_0 + \lambda_1 x_k - 1};$$

We must satisfy two constraints now ...



$$p_k = e^{\lambda_0 + \lambda_1 x_k - 1}$$

$$\sum_k p_k = \sum_k e^{\lambda_0 + \lambda_1 x_k - 1} = e^{\lambda_0 - 1} \sum_k e^{\lambda_1 x_k} = 1$$

$$\sum_k x_k p_k = \sum_k x_k e^{\lambda_0 + \lambda_1 x_k - 1} = e^{\lambda_0 - 1} \sum_k x_k e^{\lambda_1 x_k} = \mu$$

$$e^{\lambda_0 - 1} = \frac{1}{\sum_k e^{\lambda_1 x_k}}; \quad \frac{\sum_k x_k e^{\lambda_1 x_k}}{\sum_k e^{\lambda_1 x_k}} = \mu$$

no analytic solution, only numerical

## Example : the biased die

(E. T. Jaynes: *Where do we stand on Maximum Entropy?* In *The Maximum Entropy Formalism*; Levine, R. D. and Tribus, M., Eds.; MIT Press, Cambridge, MA, 1978)

mean value of throws for an unbiased die

$$\frac{1}{6}(1 + 2 + 3 + 4 + 5 + 6) = \frac{21}{6} = 3.5$$

mean value for a biased die

$$3.5(1 + \varepsilon)$$

*Problem: for a given mean value of the biased die, what is the probability distribution of each value?*

*The mean value is insufficient information, and we use the maximum entropy method to find the most likely distribution (the least informative one).*

## entropy maximization with the biased die:

$$\frac{\partial}{\partial p_k} \left[ -\sum_{k=1}^6 p_k \ln p_k + \lambda_0 \left( \sum_{k=1}^6 p_k - 1 \right) + \lambda_1 \left( \sum_{k=1}^6 k p_k - \frac{7}{2}(1 + \varepsilon) \right) \right]$$
$$= -(\ln p_k + 1) + \lambda_0 + k\lambda_1 = 0$$

$$p_k = e^{\lambda_0 + \lambda_1 k - 1}$$

$$\sum_{k=1,6} p_k = e^{\lambda_0 - 1} \sum_{k=1,6} e^{\lambda_1 k} = 1$$

$$\sum_{k=1,6} k p_k = e^{\lambda_0 - 1} \sum_{k=1,6} k e^{\lambda_1 k} = \frac{7}{2}(1 + \varepsilon)$$

$$e^{\lambda_0 - 1} = \frac{1}{\sum_{k=1,6} e^{\lambda_1 k}}; \quad \frac{\sum_{k=1,6} k p_k}{\sum_{k=1,6} p_k} = \frac{7}{2}(1 + \varepsilon)$$

we still have to satisfy the constraints ...

$$e^{\lambda_0 - 1} \sum_{k=1,6} e^{\lambda_1 k} = e^{\lambda_0 - 1} \left( \sum_{k=0,6} e^{\lambda_1 k} - 1 \right) = e^{\lambda_0 - 1} \left( \frac{1 - e^{7\lambda_1}}{1 - e^{\lambda_1}} - 1 \right) = 1$$

$$\begin{aligned} \frac{\sum_{k=1,6} k e^{\lambda_1 k}}{\sum_{k=1,6} e^{\lambda_1 k}} &= \frac{\partial}{\partial \lambda_1} \ln \sum_{k=1,6} e^{\lambda_1 k} = \frac{\partial}{\partial \lambda_1} \ln \left( e^{\lambda_1} \sum_{k=0,5} e^{\lambda_1 k} \right) \\ &= \frac{\partial}{\partial \lambda_1} \left[ \lambda_1 + \ln(1 - e^{6\lambda_1}) - \ln(1 - e^{\lambda_1}) \right] \\ &= 1 - \frac{6e^{6\lambda_1}}{1 - e^{6\lambda_1}} + \frac{e^{\lambda_1}}{1 - e^{\lambda_1}} = \frac{7}{2}(1 + \varepsilon) \end{aligned}$$

The Lagrange multipliers are obtained from nonlinear equations, and we must use numerical methods

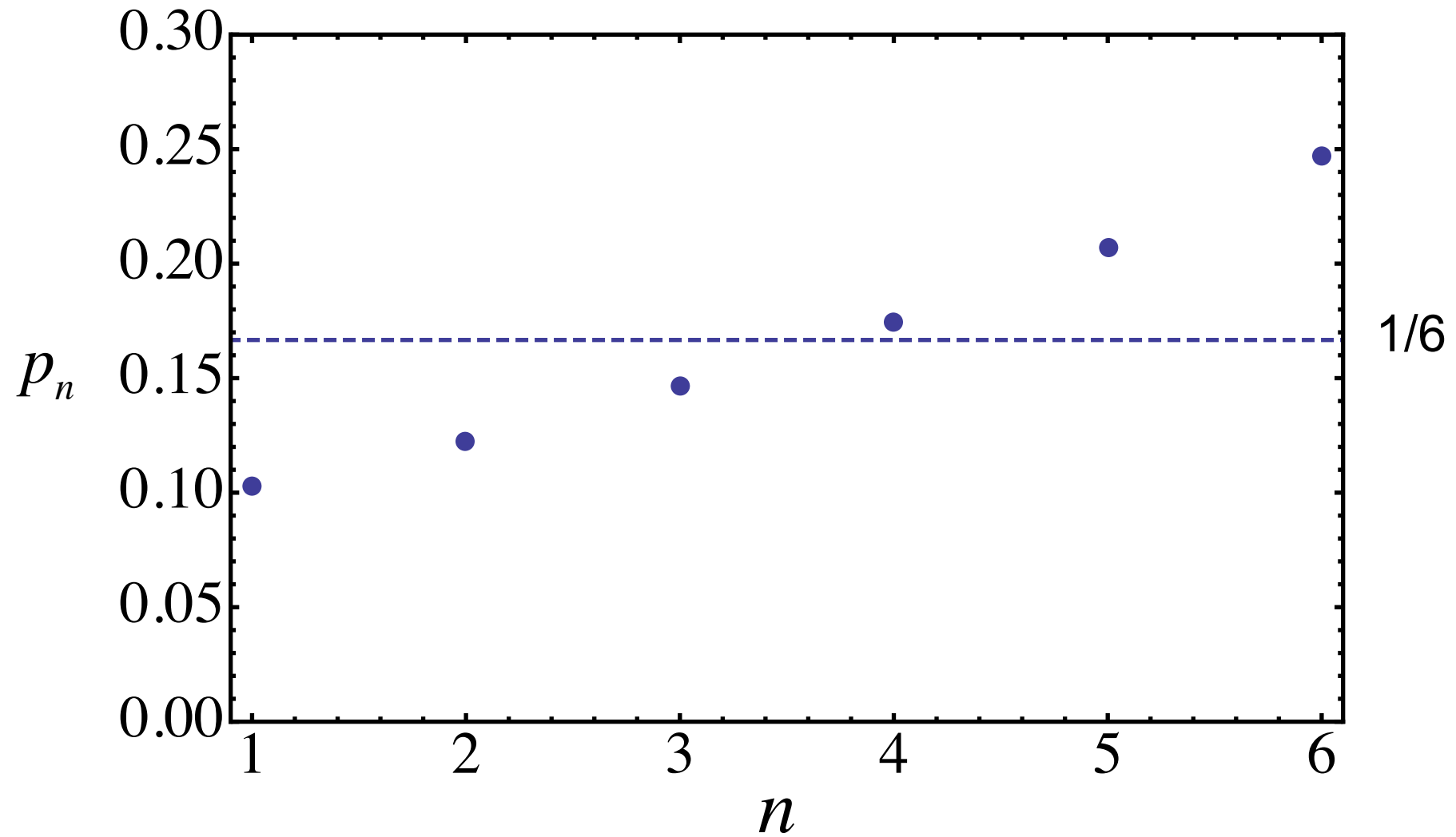
## numerical solution

<b>media</b>	<b><math>p_1</math></b>	<b><math>p_2</math></b>	<b><math>p_3</math></b>	<b><math>p_4</math></b>	<b><math>p_5</math></b>	<b><math>p_6</math></b>
<b>3.0</b>	0.246782	0.20724	0.174034	0.146148	0.122731	0.103065
<b>3.1</b>	0.22929	0.199582	0.173723	0.151214	0.131622	0.114568
<b>3.2</b>	0.212566	0.191659	0.172808	0.155811	0.140487	0.126669
<b>3.3</b>	0.196574	0.183509	0.171313	0.159928	0.149299	0.139377
<b>3.4</b>	0.181282	0.175168	0.16926	0.163551	0.158035	0.152704
<b>3.5</b>	0.166667	0.166667	0.166667	0.166667	0.166666	0.166666
<b>3.6</b>	0.152704	0.158035	0.163551	0.16926	0.175168	0.181282
<b>3.7</b>	0.139377	0.149299	0.159928	0.171313	0.183509	0.196574
<b>3.8</b>	0.126669	0.140487	0.155811	0.172808	0.191659	0.212566
<b>3.9</b>	0.114568	0.131622	0.151214	0.173723	0.199582	0.22929
<b>4.0</b>	0.103065	0.122731	0.146148	0.174034	0.20724	0.246782

with a biased die we obtain skewed distributions.

These are examples of UNINFORMATIVE PRIORS

Example: mean = 4



## Entropy with continuous probability distributions

(we use the relative entropy, i.e., the Kullback-Leibler divergence instead of entropy)

Entropy maximization with additional conditions (partial knowledge of moments of the prior distribution)

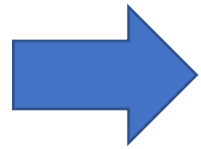
$$\langle x^k \rangle = \int_a^b x^k p(x) dx$$

function (functional) that must be maximized

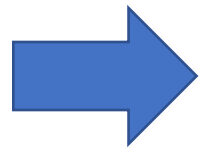
$$Q[p] = - \int_a^b p(x) \ln \frac{p(x)}{m(x)} dx + \sum_k \lambda_k \left\{ \int_a^b x^k p(x) dx - M_k \right\}$$

variation

$$\delta Q = - \int_a^b \delta p \left\{ \ln \frac{p(x)}{m(x)} + 1 - \sum_k \lambda_k x^k \right\} dx = 0$$



$$\ln \frac{p(x)}{m(x)} + 1 - \sum_k \lambda_k x^k = 0$$



$$p(x) = m(x) \exp \left( \sum_k \lambda_k x^k - 1 \right)$$



$$p(x) = m(x) \exp\left(\sum_n \lambda_n x^n - 1\right)$$

$p(x)$  is determined by the choice of  $m(x)$  and by the constraints

The constraints can be the moments themselves:

$$M_k = \int_a^b x^k m(x) \exp\left(\sum_n \lambda_n x^n - 1\right) dx$$

1. no moment is known, normalization is the only constraint, and  $p(x)$  is defined in the interval  $(a,b)$

$$M_0 = \int_a^b m(x) \exp(\lambda_0 - 1) dx = 1$$

we take a reference distribution which is uniform on  $(a,b)$ , i.e.,

$$m(x) = \frac{1}{b-a}$$

$$M_0 = \frac{1}{b-a} \int_a^b \exp(\lambda_0 - 1) dx = \exp(\lambda_0 - 1) = 1$$

$$\Rightarrow \lambda_0 = 1; \quad p(x) = m(x) \exp\left(\sum_{n=0}^0 \lambda_n x^n - 1\right) = \frac{1}{b-a}$$

2. only the first moment – the mean – is known, and  $p(x)$  is defined on  $(a,b)$

$$M_0 = \frac{1}{b-a} \int_a^b \exp(\lambda_0 + \lambda_1 x - 1) dx = 1$$

$$M_1 = \frac{1}{b-a} \int_a^b x \exp(\lambda_0 + \lambda_1 x - 1) dx$$

$$M_0 = 1 = \frac{\exp(\lambda_0 - 1)}{b-a} \int_a^b \exp(\lambda_1 x) dx = \frac{\exp(\lambda_0 - 1)}{b-a} \cdot \frac{\exp(\lambda_1 b) - \exp(\lambda_1 a)}{\lambda_1}$$

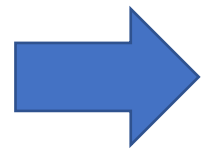
$$M_1 = \frac{\exp(\lambda_0 - 1)}{b-a} \int_a^b x \exp(\lambda_1 x) dx = \frac{\exp(\lambda_0 - 1)}{b-a} \left[ \frac{1}{\lambda_1} (b \exp(\lambda_1 b) - a \exp(\lambda_1 a)) - \frac{1}{\lambda_1^2} (\exp(\lambda_1 b) - \exp(\lambda_1 a)) \right]$$

in general, these equations can only be solved numerically...

special case:

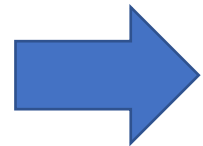
$$a \rightarrow -\frac{L}{2}; \quad b \rightarrow \frac{L}{2}; \quad M_1 = 0$$

$$\frac{\exp(\lambda_0 - 1) \cdot \exp(\lambda_1 L/2) - \exp(-\lambda_1 L/2)}{L \lambda_1} = 1$$



$$\frac{\exp(\lambda_0 - 1)}{L} \left[ \frac{1}{\lambda_1} \left( \frac{L}{2} \exp(\lambda_1 L/2) + \frac{L}{2} \exp(-\lambda_1 L/2) \right) - \frac{1}{\lambda_1^2} (\exp(\lambda_1 L/2) - \exp(-\lambda_1 L/2)) \right] = 0$$

$$\frac{\exp(\lambda_0 - 1) \cdot \exp(\lambda_1 L/2) - \exp(-\lambda_1 L/2)}{L \lambda_1} = 1$$



$$\frac{L}{2} (\exp(\lambda_1 L/2) + \exp(-\lambda_1 L/2)) - \frac{1}{\lambda_1} (\exp(\lambda_1 L/2) - \exp(-\lambda_1 L/2)) = 0$$

$$\exp(\lambda_0 - 1) \frac{\sinh(\lambda_1 L/2)}{\lambda_1 L/2} = 1$$

$$L \cosh(\lambda_1 L/2) - \frac{2}{\lambda_1} \sinh(\lambda_1 L/2) = 0$$

$$\Rightarrow (\lambda_1 L/2) = \tanh(\lambda_1 L/2) \Rightarrow \lambda_1 = 0; \quad \lambda_0 = 1$$

$$p(x) = m(x) \exp\left(\sum_{k=0}^1 \lambda_k x^k - 1\right) = \frac{1}{L}$$

nonzero mean

$$a \rightarrow -\frac{L}{2}; \quad b \rightarrow \frac{L}{2}; \quad M_1 = \varepsilon$$

$$\frac{\exp(\lambda_0 - 1) \cdot \exp(\lambda_1 L/2) - \exp(-\lambda_1 L/2)}{L \lambda_1} = 1$$

$$\frac{\exp(\lambda_0 - 1)}{\lambda_1 L} \left[ \frac{L}{2} (\exp(\lambda_1 L/2) + \exp(-\lambda_1 L/2)) - \frac{1}{\lambda_1} (\exp(\lambda_1 L/2) - \exp(-\lambda_1 L/2)) \right] = \varepsilon$$

$$\frac{\exp(\lambda_0 - 1)}{(\lambda_1 L/2)} \cdot \sinh(\lambda_1 L/2) = 1$$

$$\frac{L}{2} \frac{1}{\tanh(\lambda_1 L/2)} - \frac{1}{\lambda_1} = \varepsilon$$

$$\tanh(\lambda_1 L/2) = \left( \frac{1}{\lambda_1 L/2} + \frac{2\varepsilon}{L} \right)^{-1} \quad \tanh(z) = \left( \frac{1}{z} + \frac{2\varepsilon}{L} \right)^{-1}$$

we find an approximate solution

$$z - \frac{z^3}{3} \approx \left( \frac{1}{z} + \frac{2\varepsilon}{L} \right)^{-1} \Rightarrow \left( z - \frac{z^3}{3} \right) \left( \frac{1}{z} + \frac{2\varepsilon}{L} \right) \approx 1 + \frac{2\varepsilon}{L} z - \frac{z^2}{3} = 1$$

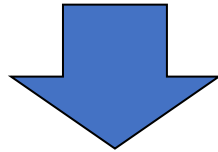
$$\Rightarrow \frac{2\varepsilon}{L} - \frac{z}{3} \approx 0 \Rightarrow z \approx \frac{6\varepsilon}{L}$$

$$\frac{\lambda_1 L}{2} \approx \frac{6\varepsilon}{L} \Rightarrow p(x) \approx \frac{1}{L} \exp(\lambda_1 x) \approx \frac{1}{L} \left( 1 - \frac{12\varepsilon}{L} x \right)$$

another special case  $a = 0; b \rightarrow \infty$

$$M_0 = \frac{1}{b-a} \int_a^b \exp(\lambda_0 + \lambda_1 x - 1) dx = 1$$

$$M_1 = \frac{1}{b-a} \int_a^b x \exp(\lambda_0 + \lambda_1 x - 1) dx$$



$$M_0 = 1 = m_0 \exp(\lambda_0 - 1) \cdot \frac{1}{(-\lambda_1)}$$

$$M_1 = m_0 \exp(\lambda_0 - 1) \left[ \frac{1}{\lambda_1^2} \right] = (-\lambda_1) \left[ \frac{1}{\lambda_1^2} \right] = -\frac{1}{\lambda_1} = \langle x \rangle$$



then

$$m_0 \exp(\lambda_0 - 1) = -\lambda_1 = \frac{1}{\langle x \rangle}$$

and we obtain the exponential distribution

$$\begin{aligned} p(x) &= m(x) \exp\left(\sum_n \lambda_n x^n - 1\right) \\ &= m_0 \exp(\lambda_0 - 1) \exp(\lambda_1 x) = \frac{1}{\langle x \rangle} \exp\left(-\frac{x}{\langle x \rangle}\right) \end{aligned}$$

### 3. both mean and variance are known, and the interval is the whole real axis

$$M_0 = m_0 \int_a^b \exp(\lambda_0 + \lambda_1 x + \lambda_2 x^2 - 1) dx = 1$$

$$M_1 = m_0 \int_a^b x \exp(\lambda_0 + \lambda_1 x + \lambda_2 x^2 - 1) dx$$

$$M_2 = m_0 \int_a^b x^2 \exp(\lambda_0 + \lambda_1 x + \lambda_2 x^2 - 1) dx$$

$$\begin{aligned} \exp(\lambda_0 + \lambda_1 x + \lambda_2 x^2 - 1) &= \exp \left[ \lambda_2 \left( x^2 + 2 \frac{\lambda_1}{\lambda_2} x + \frac{\lambda_1^2}{\lambda_2^2} \right) + \left( \lambda_0 - 1 - \frac{\lambda_1^2}{\lambda_2} \right) \right] \\ &= \exp \left( \lambda_0 - 1 - \frac{\lambda_1^2}{\lambda_2} \right) \exp \left[ \lambda_2 \left( x + \frac{\lambda_1}{\lambda_2} \right)^2 \right] \end{aligned}$$

$$M_0 = m_0 \exp\left(\lambda_0 - 1 - \frac{\lambda_1^2}{\lambda_2}\right) \int_{-\infty}^{+\infty} \exp\left[-\frac{1}{2(-1/2\lambda_2)}\left(x + \frac{\lambda_1}{\lambda_2}\right)^2\right] dx = m_0 \exp\left(\lambda_0 - 1 - \frac{\lambda_1^2}{\lambda_2}\right) \sqrt{-\frac{\pi}{\lambda_2}} = 1$$

$$M_1 = m_0 \exp\left(\lambda_0 - 1 - \frac{\lambda_1^2}{\lambda_2}\right) \int_{-\infty}^{+\infty} x \exp\left[-\frac{1}{2(-1/2\lambda_2)}\left(x + \frac{\lambda_1}{\lambda_2}\right)^2\right] dx = m_0 \exp\left(\lambda_0 - 1 - \frac{\lambda_1^2}{\lambda_2}\right) \sqrt{-\frac{\pi}{\lambda_2}} \left(-\frac{\lambda_1}{\lambda_2}\right) = -\mu$$

$$M_2 = m_0 \exp\left(\lambda_0 - 1 - \frac{\lambda_1^2}{\lambda_2}\right) \int_{-\infty}^{+\infty} x^2 \exp\left[-\frac{1}{2(-1/2\lambda_2)}\left(x + \frac{\lambda_1}{\lambda_2}\right)^2\right] dx = m_0 \exp\left(\lambda_0 - 1 - \frac{\lambda_1^2}{\lambda_2}\right) \sqrt{-\frac{\pi}{\lambda_2}} \left(-\frac{1}{2\lambda_2} + \frac{\lambda_1^2}{\lambda_2^2}\right) = \sigma^2 + \mu^2$$

$$M_0 = m_0 \exp\left(\lambda_0 - 1 - \frac{\lambda_1^2}{\lambda_2}\right) \sqrt{-\frac{\pi}{\lambda_2}} = 1$$

$$M_1 = \frac{\lambda_1}{\lambda_2} = \mu$$

$$M_2 = \left(-\frac{1}{2\lambda_2} + \frac{\lambda_1^2}{\lambda_2^2}\right) = \sigma^2 + \mu^2$$

$$\Rightarrow \lambda_1 = -\frac{\mu}{2\sigma^2}; \quad \lambda_2 = -\frac{1}{2\sigma^2}; \quad m_0 \exp\left(\lambda_0 - 1 - \frac{\lambda_1^2}{\lambda_2}\right) = \frac{1}{\sqrt{2\pi\sigma^2}}$$

$$\begin{aligned} p(x) &= m_0 \exp(\lambda_0 + \lambda_1 x + \lambda_2 x^2 - 1) \\ &= m_0 \exp\left(\lambda_0 - 1 - \frac{\lambda_1^2}{\lambda_2}\right) \exp\left[-\frac{1}{2(-1/2\lambda_2)}\left(x + \frac{\lambda_1}{\lambda_2}\right)^2\right] \\ &= \frac{1}{\sqrt{2\sigma^2\pi}} \exp\left[\frac{1}{2\sigma^2}(x - \mu)^2\right] \end{aligned}$$

... in this case where mean and variance are known, the entropic prior is Gaussian

*An alternative form of entropy that incorporates the normalization constraint from the start*

$$Q[p; m] = - \int_x dx p(x) \ln \frac{p(x)}{m(x)} + \lambda \left( \int_x dx p(x) - \int_x dx m(x) \right)$$

$$= \int_x dx \left( -p(x) \ln \frac{p(x)}{m(x)} + \lambda p(x) - \lambda m(x) \right)$$

$$\delta Q = \int_x \delta p dx \left( -\ln \frac{p(x)}{m(x)} - 1 + \lambda \right) = 0$$

$$p(x) = m(x) \exp(\lambda - 1)$$

$$\int_x dx p(x) = \int_x dx m(x) \exp(\lambda - 1) = \exp(\lambda - 1) \int_x dx m(x) = \exp(\lambda - 1) = 1$$

$$\Rightarrow \lambda = 1$$

$$Q[p; m] = \int_x dx \left( -p(x) \ln \frac{p(x)}{m(x)} + p(x) - m(x) \right)$$

Until now we have emphasized the role of the momenta of the distribution, however other information can be incorporated in the same way in the entropic prior.

*A “crystallographic” example (Jaynes, 1968)*

Consider a simple version of a crystallographic problem, where a 1-D crystal has atoms at the positions

$$x_j = jL \quad (L = 1, \dots, n)$$

and such that these positions may be occupied by impurities.

From X-ray experiments it has been determined that impurity atoms prefer sites where

$$\cos(kx_j) > 0$$

furthermore, we take, as an example,

$$\langle \cos(kx_j) \rangle = 0.3$$

which means that we have the constraint

$$\langle \cos(kx_j) \rangle = \sum_{j=1}^n p_j \cos(kx_j) = 0.3$$

where  $p_j$  is the probability that an impurity atom is at site  $j$ .

Then the constrained entropy that must be maximized is

$$Q = -\sum_{j=1}^n p_j \ln p_j + \lambda_0 \left( \sum_{j=1}^n p_j - 1 \right) + \lambda_1 \left( \sum_{j=1}^n p_j \cos(kx_j) - 0.3 \right)$$

from which we find the maximization condition

$$\frac{\partial Q}{\partial p_j} = -(\ln p_j + 1) + \lambda_0 + \lambda_1 \cos(kx_j) = 0$$

i.e.,

$$p_j = \exp \left[ 1 - \lambda_0 - \lambda_1 \cos(kx_j) \right]$$

The rest of the solution proceeds either by approximation or by numerical calculation.



## Example of MaxEnt in action: unconstrained problem in image restoration



J. Skilling, Nature 309 (1984) 748

Car movement introduces linear correlations among pixels. The model of linear corrections does not allow direct inversion to find the corrected image because the number of variables is larger than the number of equations. The MaxEnt methods regularizes the problem and finds a reasonable solution.



J. Skilling, Nature 309 (1984) 748

Reconstruction of missing data  
(from <http://www.maxent.co.uk> )



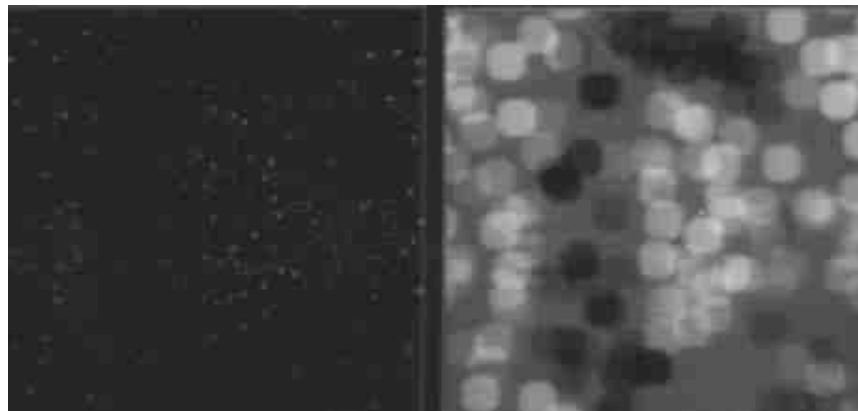
50%

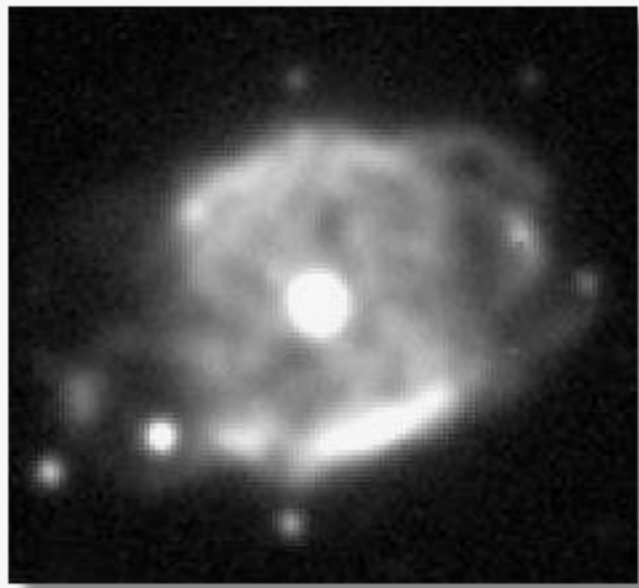


95%

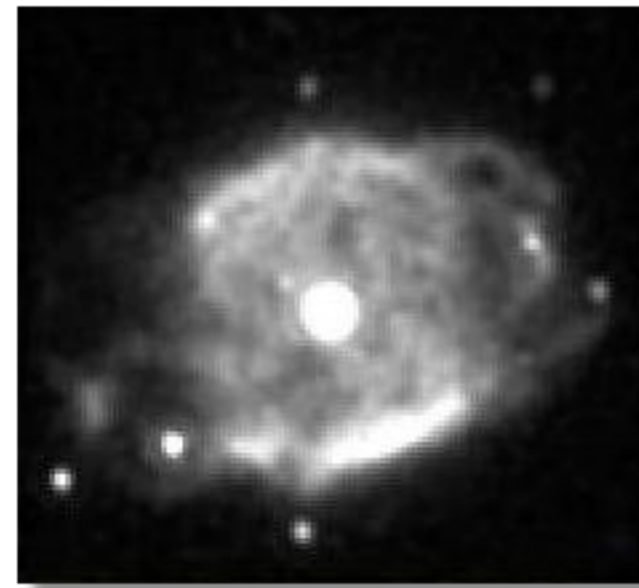


99%

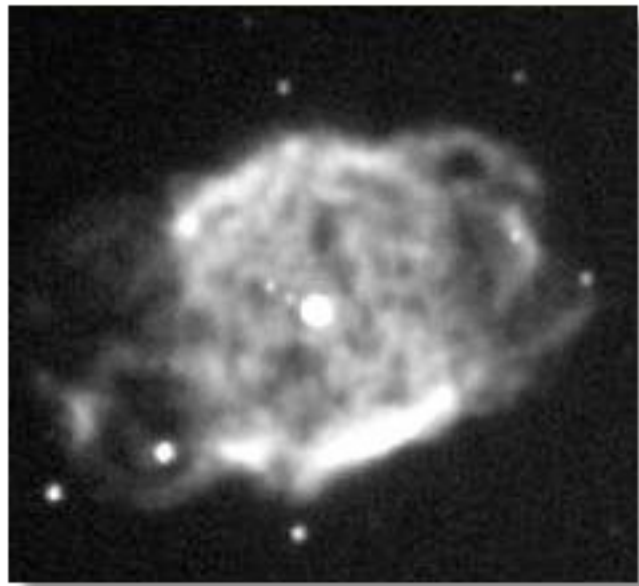




NGC 40



low resolution (MEM enhanced)



low resolution

high resolution

[Home](#)

[About MEDC](#)

[Applications](#)

[Examples](#)

[Products](#)

[Prices](#)

[Documents](#)

[Contact us](#)

[Search MEDC](#)

Quick Search:

Search

### John Skilling: Biographical information

John is Scientific Director of MEDC. He did his Ph.D. (on cosmic rays) in the Department of Physics at Cambridge University, and went on to become a Lecturer in the Department of Applied Mathematics and Theoretical Physics, and a Fellow of St Johns College.

In the late 1970s, another radio astronomer, [Steve Gull](#), introduced him to the power of the Maximum Entropy Method. John wrote what was to become the first MemSys kernel system, and helped lay the Bayesian foundations for MEM. In 1981 he and Steve founded MEDC to exploit opportunities to apply MEM in other fields.

John resigned his Lectureship in 1990 in order to go fulltime with MSL and MEDC. Thanks to the wonders of modern technology John is able to telecommute from his new home in the West of Ireland, and he makes regular visits to clients both in the UK and further afield.



[Home](#) | [Applications](#) | [Products](#) | [Prices](#) | [Documents](#) | [About MEDC](#) | [Contact Us](#) | [Full search](#)

©MEDC Ltd. Last revised Wed Sep 19 22:19:39 2007

<http://www.maxent.co.uk/>

(the company no longer exists, and the website has disappeared from the web)



Contents lists available at [ScienceDirect](https://www.sciencedirect.com)

SoftwareX

journal homepage: [www.elsevier.com/locate/softx](http://www.elsevier.com/locate/softx)



Original software publication

## PyMaxEnt: A Python software for maximum entropy moment reconstruction

Tony Saad<sup>1,\*</sup>, Giovanna Ruai

*Department of Chemical Engineering, University of Utah Salt Lake City, UT 84102, United States of America*



### ARTICLE INFO

#### Article history:

Received 16 July 2019

Received in revised form 21 October 2019

Accepted 21 October 2019

#### Keywords:

Maximum entropy reconstruction

Inverse moment problem

Particle size distribution

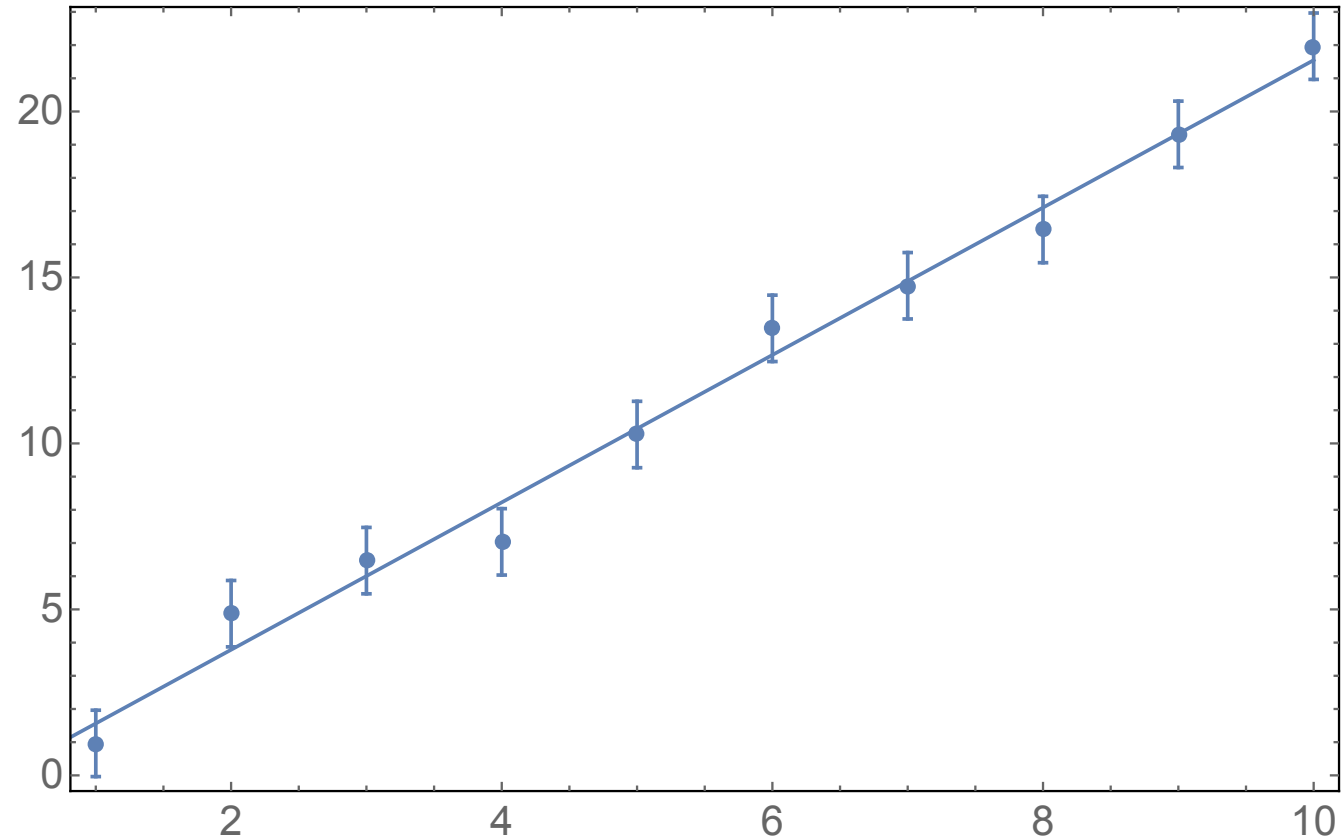
### ABSTRACT

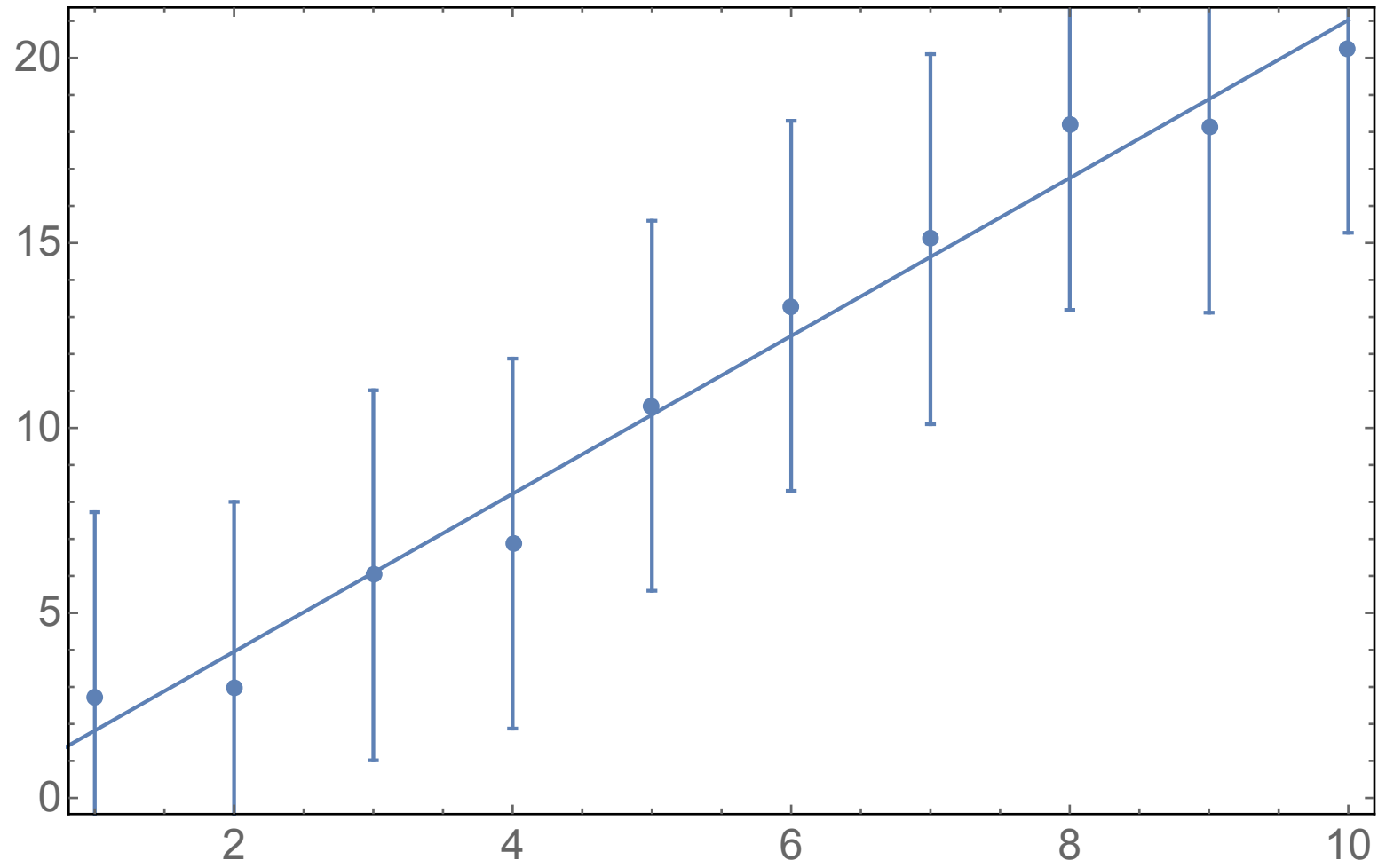
PyMaxEnt is a software that implements the principle of maximum entropy to reconstruct functional distributions given a finite number of known moments. The software supports both continuous and discrete reconstructions, and is very easy to use through a single function call. In this article, we set out to verify and validate the software against several tests ranging from the reconstruction of discrete probability distributions for biased dice all the way to multimodal Gaussian and beta distributions. Written in Python, PyMaxEnt provides a robust and easy-to-use implementation for the community.

© 2019 The Authors. Published by Elsevier B.V. This is an open access article under the CC BY license (<http://creativecommons.org/licenses/by/4.0/>).

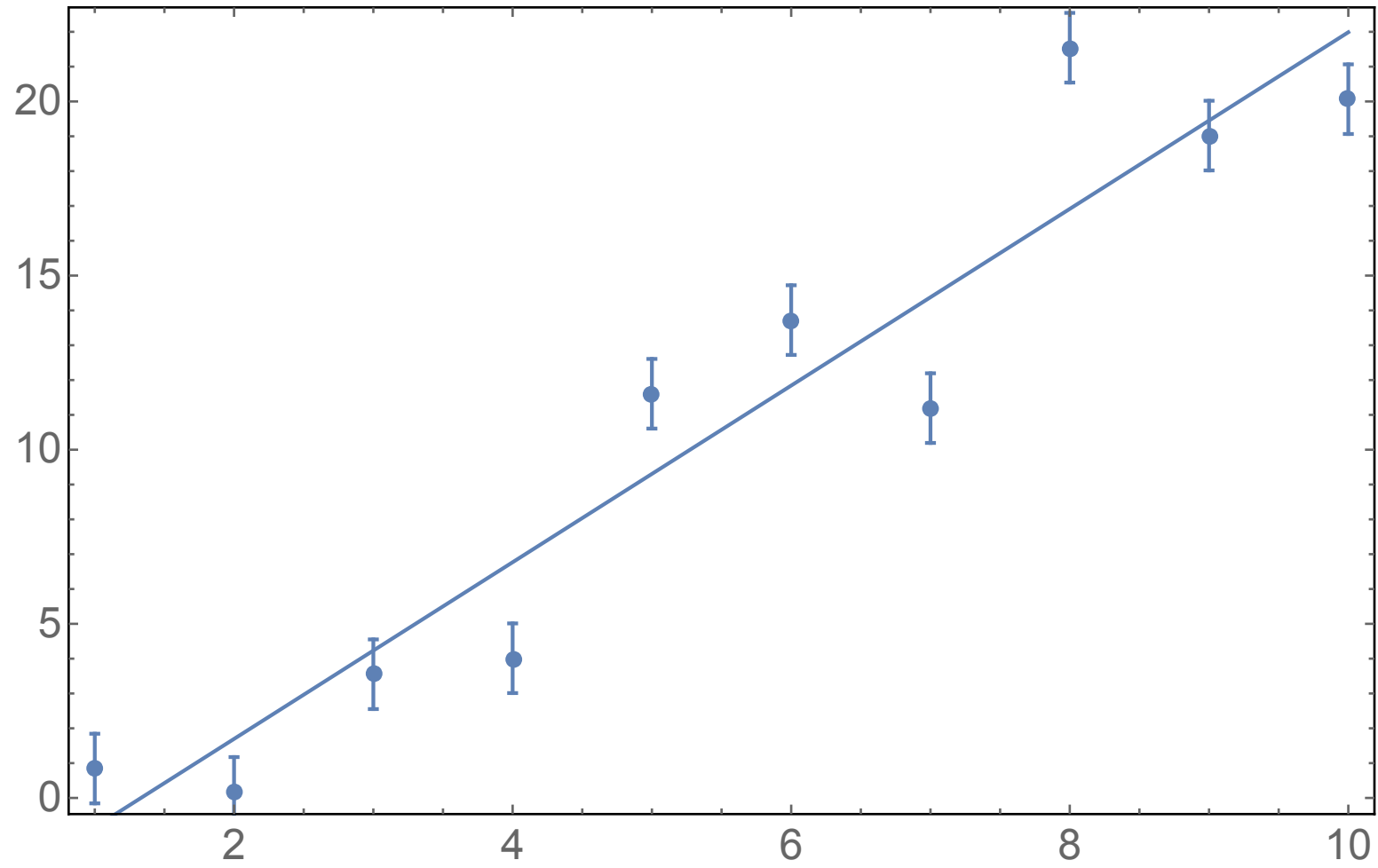
## Example of Bayesian estimate using objective priors: uncalibrated Gaussian measurement uncertainties

Here, we consider the case where we must find the mean value with given measurement uncertainties that are systematically multiplied by an unknown scale factor, under the assumption of Gaussianity.









The likelihood has a Gaussian structure

$$\begin{aligned} P(\mathbf{d} \mid \mu, \boldsymbol{\sigma}, \alpha) &= \prod_{k=1}^N \frac{1}{\sqrt{2\pi\alpha^2\sigma_k^2}} \exp\left[-\frac{(d_k - \mu)^2}{2\alpha^2\sigma_k^2}\right] \\ &= \frac{1}{(2\pi)^{N/2} \alpha^N} \left( \prod_{k=1}^N \frac{1}{\sigma_k} \right) \exp\left[-\frac{1}{2\alpha^2} \sum_{k=1}^N \frac{(d_k - \mu)^2}{\sigma_k^2}\right] \end{aligned}$$

we must rearrange the exponent as usual ...

$$\begin{aligned}\sum_{k=1}^N \frac{(d_k - \mu)^2}{\sigma_k^2} &= \sum_{k=1}^N \frac{d_k^2}{\sigma_k^2} - 2\mu \sum_{k=1}^N \frac{d_k}{\sigma_k^2} + \mu^2 \sum_{k=1}^N \frac{1}{\sigma_k^2} = \frac{ND}{\sigma_M^2} - 2\mu \frac{NM}{\sigma_M^2} + \mu^2 \frac{N}{\sigma_M^2} \\ &= \frac{N}{\sigma_M^2} (D - 2\mu M + \mu^2)\end{aligned}$$

$$\text{dove } \frac{1}{\sigma_M^2} = \frac{1}{N} \sum_{k=1}^N \frac{1}{\sigma_k^2}; \quad M = \sum_{k=1}^N \frac{d_k}{\sigma_k^2} / \sum_{k=1}^N \frac{1}{\sigma_k^2}; \quad D = \sum_{k=1}^N \frac{d_k^2}{\sigma_k^2} / \sum_{k=1}^N \frac{1}{\sigma_k^2}$$

therefore, the likelihood is

$$P(\mathbf{d} \mid \mu, \boldsymbol{\sigma}, \alpha) = \frac{1}{(2\pi)^{N/2} \alpha^N} \left( \prod_{k=1}^N \frac{1}{\sigma_k} \right) \exp \left[ -\frac{N}{2\alpha^2 \sigma_M^2} (D - 2\mu M + \mu^2) \right]$$

Now we estimate the scale factor from Bayes' theorem

$$p(\alpha|\mathbf{d}, \boldsymbol{\sigma}) = \frac{p(\mathbf{d}|\alpha, \boldsymbol{\sigma})}{\int_{\alpha} p(\mathbf{d}|\alpha', \boldsymbol{\sigma})p(\alpha')d\alpha'}p(\alpha)$$

however, we need first to marginalize the likelihood with respect to the mean, which in this case is a *nuisance parameter*

we take a uniform prior for the mean (a Jeffrey's prior)

$$\begin{aligned} P(\mathbf{d}|\boldsymbol{\sigma}, \alpha) &= \int_{\mu} P(\mathbf{d}|\mu, \boldsymbol{\sigma}, \alpha)P(\mu|\boldsymbol{\sigma}, \alpha)d\mu \\ &= \frac{1}{W} \int_{\mu_{\min}}^{\mu_{\max}} P(\mathbf{d}|\mu, \boldsymbol{\sigma}, \alpha)d\mu \\ &\approx \frac{1}{W} \frac{1}{(2\pi)^{N/2} \alpha^N} \left( \prod_{k=1}^N \frac{1}{\sigma_k} \right) \int_{-\infty}^{+\infty} \exp \left[ -\frac{N}{2\alpha^2 \sigma_M^2} (D - 2\mu M + \mu^2) \right] d\mu \end{aligned}$$

$$(W = \mu_{\max} - \mu_{\min})$$

as usual ...

$$\begin{aligned} D - 2\mu M + \mu^2 &= \mu^2 - 2\mu M + M^2 + D - M^2 \\ &= (\mu - M)^2 + D - M^2 \end{aligned}$$

... therefore the marginalized likelihood is:

$$\begin{aligned} P(\mathbf{d} | \boldsymbol{\sigma}, \alpha) &\approx \frac{1}{W} \frac{1}{(2\pi)^{N/2} \alpha^N} \left( \prod_{k=1}^N \frac{1}{\sigma_k} \right) \int_{-\infty}^{+\infty} \exp \left\{ -\frac{N}{2\alpha^2 \sigma_M^2} [(\mu - M)^2 + D - M^2] \right\} d\mu \\ &= \frac{1}{W} \frac{1}{(2\pi)^{N/2} \alpha^N} \left( \prod_{k=1}^N \frac{1}{\sigma_k} \right) \exp \left( -\frac{N(D - M^2)}{2\alpha^2 \sigma_M^2} \right) \sqrt{\frac{2\pi\alpha^2 \sigma_M^2}{N}} \end{aligned}$$

$$\begin{aligned}
p(\alpha|\mathbf{d}, \boldsymbol{\sigma}) &= \frac{p(\mathbf{d}|\alpha, \boldsymbol{\sigma})}{\int_{\alpha} p(\mathbf{d}|\alpha', \boldsymbol{\sigma})p(\alpha')d\alpha'}p(\alpha) \\
&= \frac{\frac{1}{\alpha^{N-1}} \exp\left(-\frac{N(D - M^2)}{2\alpha^2\sigma_M^2}\right)}{\int_{\alpha} \frac{1}{\alpha'^{N-1}} \exp\left(-\frac{N(D - M^2)}{2\alpha'^2\sigma_M^2}\right) p(\alpha')d\alpha'}p(\alpha)
\end{aligned}$$

$P(\alpha) \propto \frac{1}{\alpha}$  for the standard deviation we take again a Jeffreys' prior

$$p(\alpha|\mathbf{d}, \boldsymbol{\sigma}) = \frac{\frac{1}{\alpha^{N-1}} \exp\left(-\frac{N(D - M^2)}{2\alpha^2\sigma_M^2}\right) \frac{1}{\alpha}}{\int_{\alpha} \frac{1}{\alpha'^{N-1}} \exp\left(-\frac{N(D - M^2)}{2\alpha'^2\sigma_M^2}\right) \frac{1}{\alpha'} d\alpha'}; \quad A^2 = \frac{N(D - M^2)}{2\sigma_M^2}$$

$$\Rightarrow p(\alpha|\mathbf{d}, \boldsymbol{\sigma}) \rightarrow \frac{\frac{1}{\alpha^N} \exp\left(-\frac{A^2}{\alpha^2}\right)}{\int_0^{\infty} \frac{1}{\alpha'^N} \exp\left(-\frac{A^2}{\alpha'^2}\right) d\alpha'}$$

evaluation of  $\int_0^\infty \frac{1}{\alpha'^N} \exp\left(-\frac{A^2}{\alpha'^2}\right) d\alpha'$

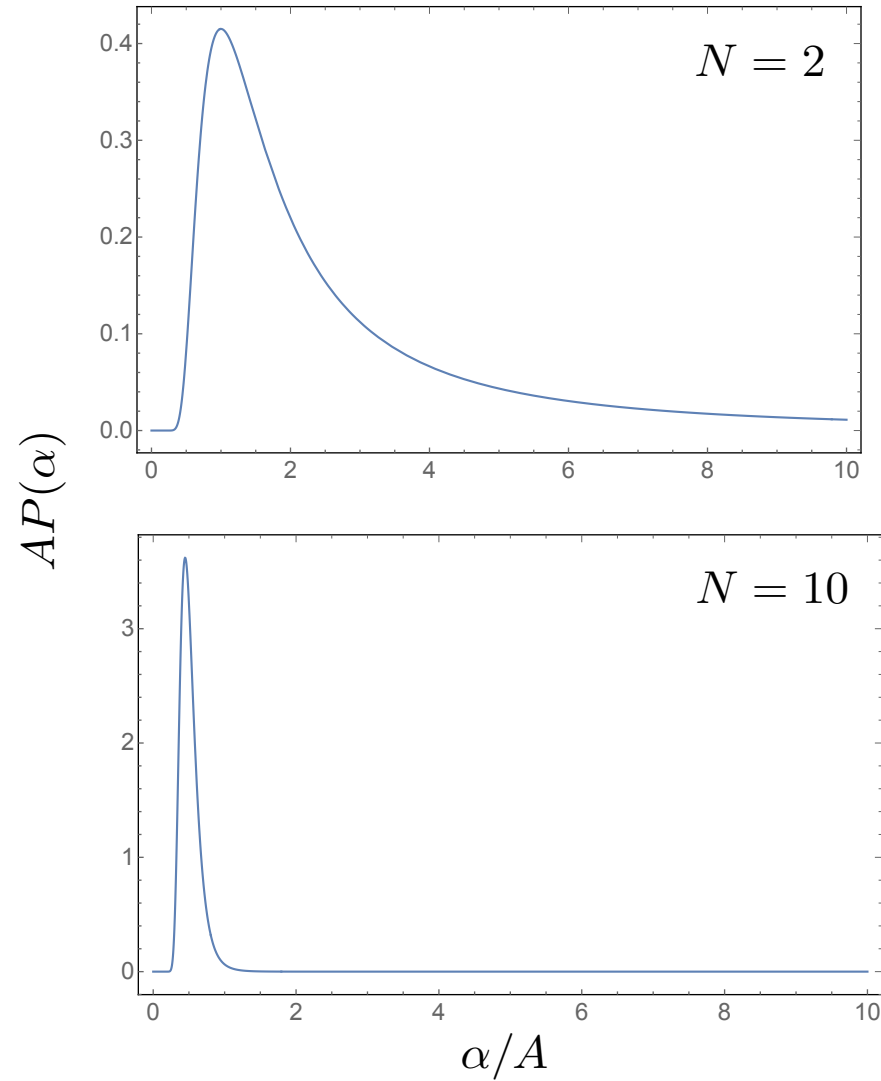
$$\frac{A^2}{\alpha^2} = x; \quad \alpha = \frac{A}{\sqrt{x}}; \quad d\alpha = -\frac{A}{2x^{3/2}} dx$$

$$\int_0^\infty \frac{x^{N/2}}{A^N} \exp(-x) \frac{A}{2x^{3/2}} dx = \frac{1}{2A^{N-1}} \int_0^\infty x^{\frac{N-1}{2}-1} \exp(-x) dx = \frac{1}{2A^{N-1}} \Gamma\left(\frac{N-1}{2}\right)$$

$$p(\alpha|\mathbf{d}, \boldsymbol{\sigma}) \rightarrow \frac{\frac{2A^{N-1}}{\alpha^N} \exp\left(-\frac{A^2}{\alpha^2}\right)}{\Gamma\left(\frac{N-1}{2}\right)}$$



$$P(\alpha|\mathbf{d}, \boldsymbol{\sigma}) = \frac{(2A^{N-1}/\alpha^N) \exp(-A^2/\alpha^2)}{\Gamma[(N-1)/2]}$$



we take the MAP estimate of the scale parameter from the pdf

$$p(\alpha | \mathbf{d}, \boldsymbol{\sigma}) = \frac{2A^{N-1}}{\alpha^N} \exp\left(-\frac{A^2}{\alpha^2}\right) \frac{1}{\Gamma\left(\frac{N-1}{2}\right)}$$

$$\frac{d}{d\alpha} P(\alpha | \mathbf{d}, \boldsymbol{\sigma}) \propto -\frac{N}{\alpha^{N+1}} \exp\left(-\frac{A^2}{\alpha^2}\right) + \frac{2A^2}{\alpha^{N+3}} \exp\left(-\frac{A^2}{\alpha^2}\right) = 0$$

$$\Rightarrow N\alpha^2 = 2A^2 \quad \Rightarrow \alpha_{MAP} = \sqrt{\frac{2}{N}}A = \sqrt{\frac{(D-M^2)}{\sigma_M^2}}$$