Introduction to Bayesian Methods - 5

Edoardo Milotti

Università di Trieste and INFN-Sezione di Trieste

Prior distributions

The choice of prior distribution is an important aspect of Bayesian inference

- prior distributions are one of the main targets of frequentists: how much do posteriors differ when we choose different priors?
- there are two main “objective” methods for the choice of priors (MaxEnt and Jeffreys')
- here we discuss

1. The quest for "objective" priors
2. Review of the Cramer-Rao bound and related concepts
3. Information-theoretic concepts in statistics
4. Jeffreys' method
5. Reference priors
6. The Maximum Entropy Method

Reference priors

In this case we need to consider a sufficient statistic *t*

Recall that a statistic t is sufficient with respect to a statistical model and its associated unknown parameter if "no other statistic that can be calculated from the same sample provides any additional information as to the value of the parameter" (Fisher, 1920)

Given the data **D**, a statistic *t* = T(**D**) is sufficient with respect to the parameter if it contains all the information needed to estimate the parameter.

Examples:

- the sample mean is sufficient for the mean of a normal distribution with known variance. Once the sample mean is known, no further information about the mean can be obtained from the sample itself.
- for an arbitrary distribution the median is not sufficient for the mean: even if the median of the sample is known, knowing the sample itself would provide further information about the population mean.

The idea behind a reference prior is that it must be such that data affect our posterior distribution the most.

We can formalize this by means of the KL divergence by requiring that the KL divergence between prior and posterior be maximal.
To proceed, we utilize a posterior that depends on a sufficient statistic instead of the original data

$$D_{\mathrm{KL}}\left[p(\theta|t)||p(\theta)\right] = \int_{\Theta} p(\theta|t) \ln \frac{p(\theta|t)}{p(\theta)} d\theta$$

then, its expection value over the statistic is

$$\begin{aligned}
\mathbb{E}\left[D_{\mathrm{KL}}\right]_t &= \int_T p(t) \int_{\Theta} p(\theta|t) \ln \frac{p(\theta|t)}{p(\theta)} d\theta \; dt \\
&= \int_T \int_{\Theta} p(\theta|t) p(t) \ln \frac{p(\theta|t) p(t)}{p(\theta) p(t)} d\theta \; dt \\
&= \int_T \int_{\Theta} p(\theta, t) \ln \frac{p(\theta, t)}{p(\theta) p(t)} d\theta \; dt
\end{aligned}$$

Mutual information between the two distributions

A reference prior is a pdf that maximizes the mutual information

$$\int_T \int_\Theta p(\theta, t) \ln \frac{p(\theta, t)}{p(\theta)p(t)} d\theta \; dt$$

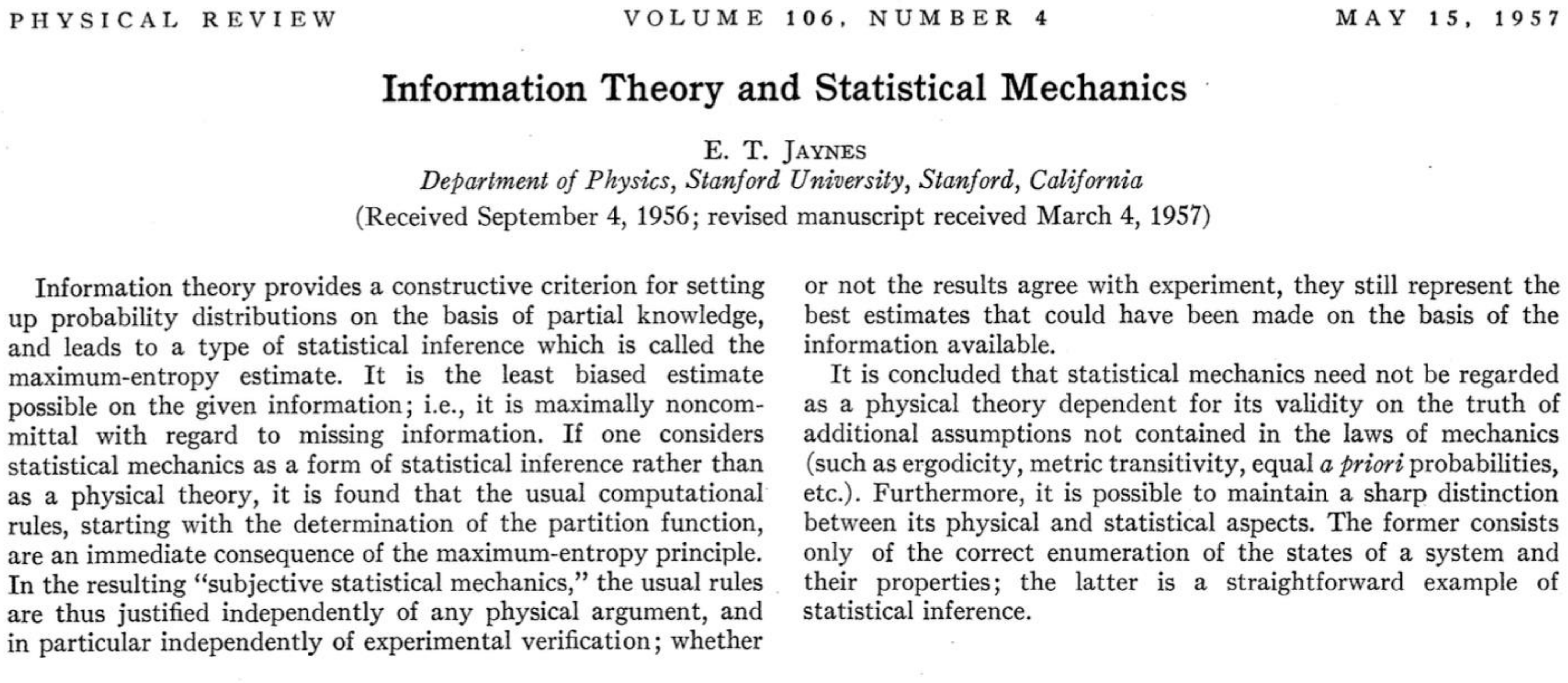
and therefore maximizes the effect of data on the posterior distribution.

- For one-dimensional parameters, reference priors and Jeffrey's priors are equivalent, while they differ in the multivariate case.

- Since the result is based on the KL divergence, which is transformation-invariant, reference priors are transformation-invariants as well, just as the Jeffrey's priors (and this justifies their equivalence, at least for the univariate case).

- For more information, see, e.g., J. Bernardo, Reference Analysis, Handbook of Statistics, **25** (2005) 17

The principle of Maximum Entropy (MaxEnt)



PHYSICAL REVIEW VOLUME 106, NUMBER 4 MAY 15, 1957

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.

The principle of Maximum Entropy (MaxEnt) – 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

The principle of Maximum Entropy (MaxEnt) – the “Kangaroo problem” (Jaynes) (ctd.)

probabilities

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

entropy (proportional to Shannon’s entropy)

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

constraints

$$\begin{aligned} p_{bl} + p_{\bar{b}l} + p_{b\bar{l}} + p_{\bar{b}\bar{l}} &= 1 \\ p_{bl} + p_{b\bar{l}} &= 1/3 \\ p_{bl} + p_{\bar{b}l} &= 1/3 \end{aligned}$$

Underdetermined system
of linear equations

The principle of Maximum Entropy (MaxEnt) – the "Kangaroo problem" (Jaynes) (ctd.)

Maximization of constrained entropy

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

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

$$\frac{\partial H_C}{\partial p_{b\bar{l}}} = -\ln p_{\bar{b}l} - 1 + \lambda_1 + \lambda_3 = 0$$

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

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

$$\Longrightarrow$$

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

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

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

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

The principle of Maximum Entropy (MaxEnt) – the "Kangaroo problem" (Jaynes) (ctd.)

Solution of the nonlinear system of equations

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

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

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

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

The principle of Maximum Entropy (MaxEnt)

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

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

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

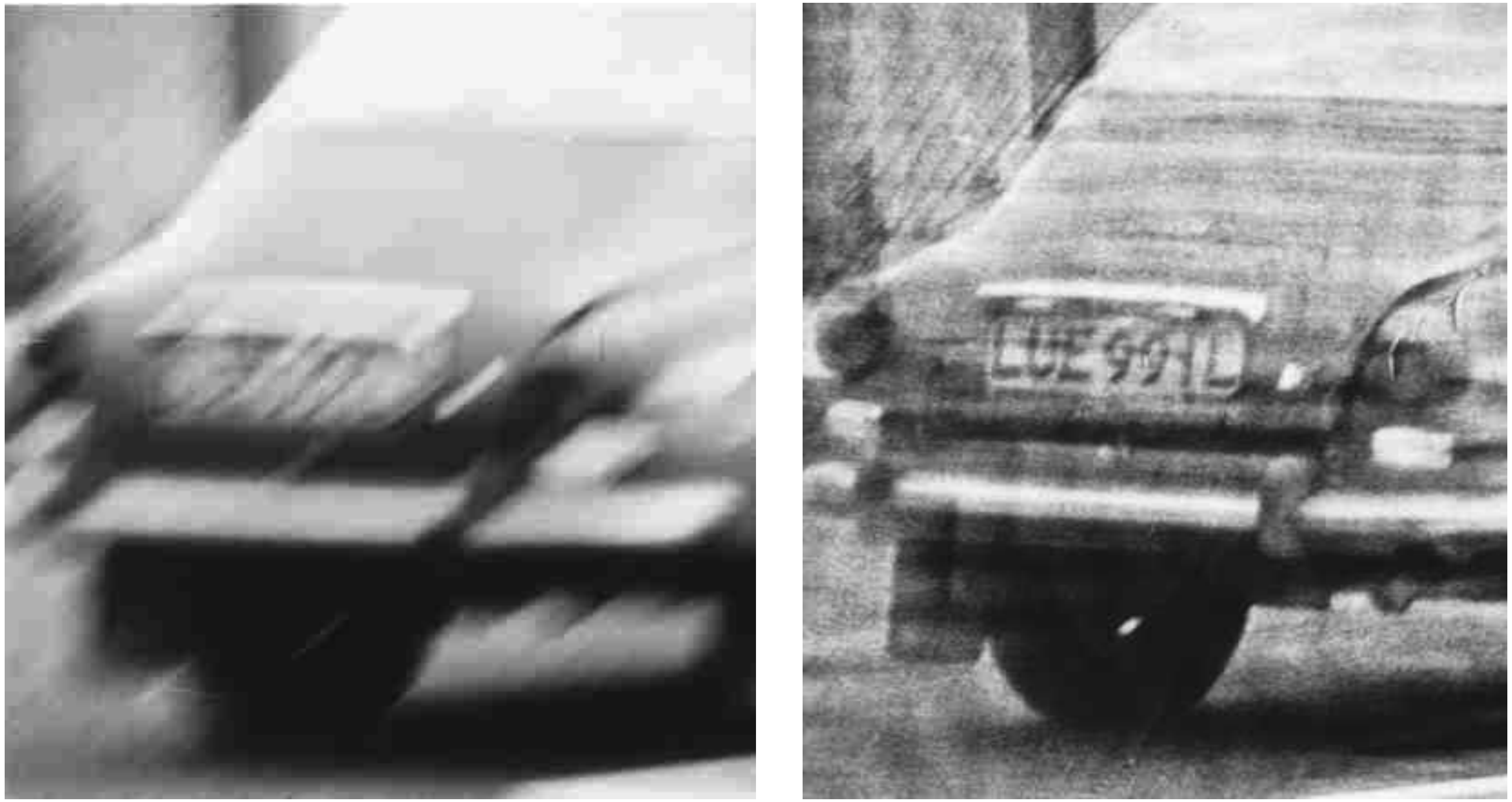
⇒ $$\frac{\partial Q}{\partial p_i} = -(\ln p_i + 1) + (-\lambda + 1) - \mu f(x_i) = 0$$

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

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

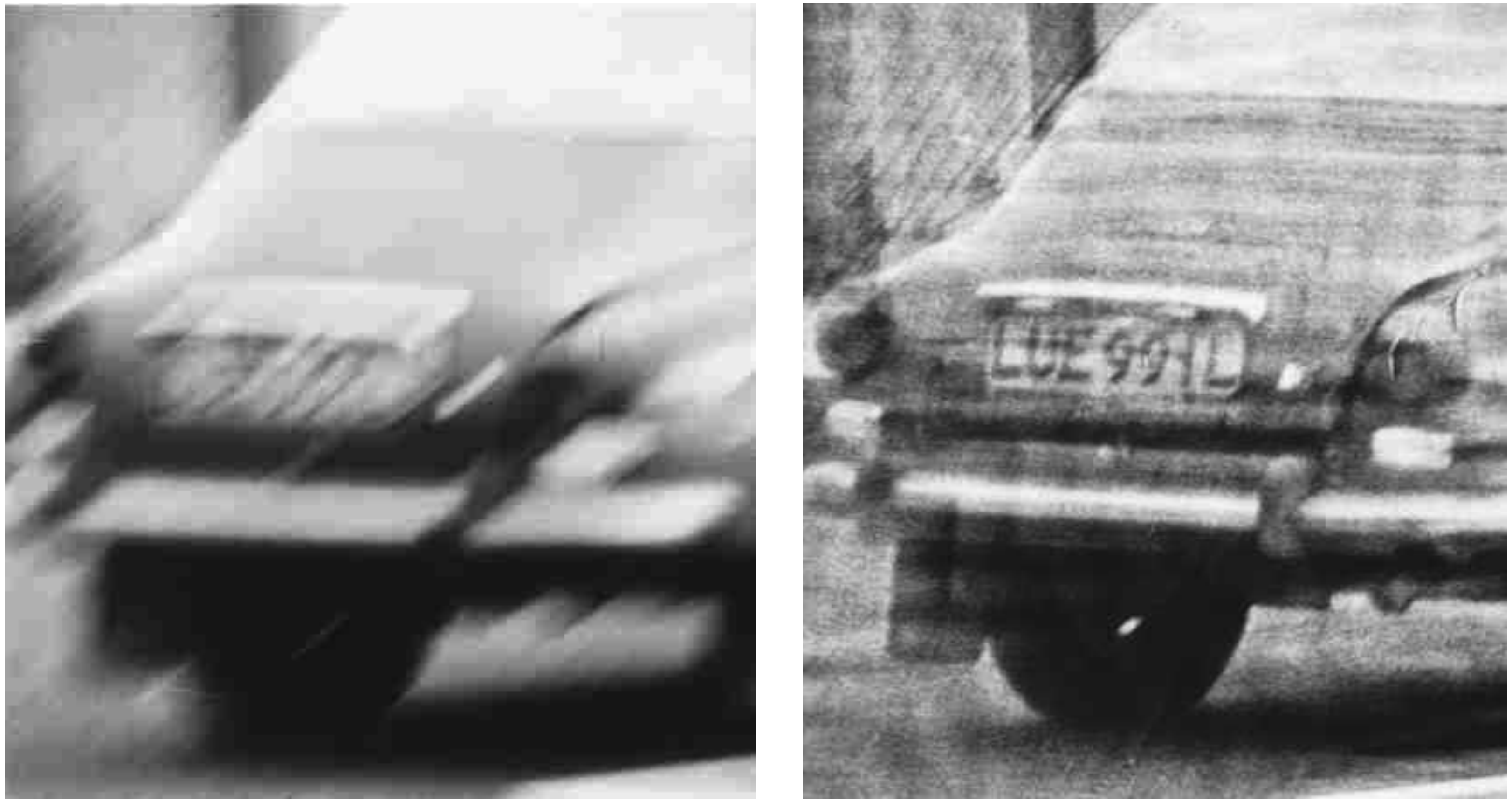
⇒ $$\langle f(x) \rangle = -\frac{\partial}{\partial \mu} \ln Z(\mu)$$

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

The principle of Maximum Entropy (MaxEnt) – Objective priors

$$H = \sum_k p_k \ln \frac{1}{p_k} = -\sum_k p_k \ln p_k$$

Shannon's entropy (in nats)

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

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

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

entropy maximization when the mean μ is known

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

$$\Rightarrow \quad \boxed{p_\ell = e^{\lambda_0 + \lambda_1 x_\ell - 1}}$$

incomplete solution...

We must satisfy two constraints now ...

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

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

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

$$\Rightarrow \begin{cases} e^{\lambda_0 - 1} \sum_k e^{\lambda_1 x_k} = 1 \\ e^{\lambda_0 - 1} \sum_k x_k e^{\lambda_1 x_k} = \mu \end{cases}$$

in general, this system does not have an analytical 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:

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

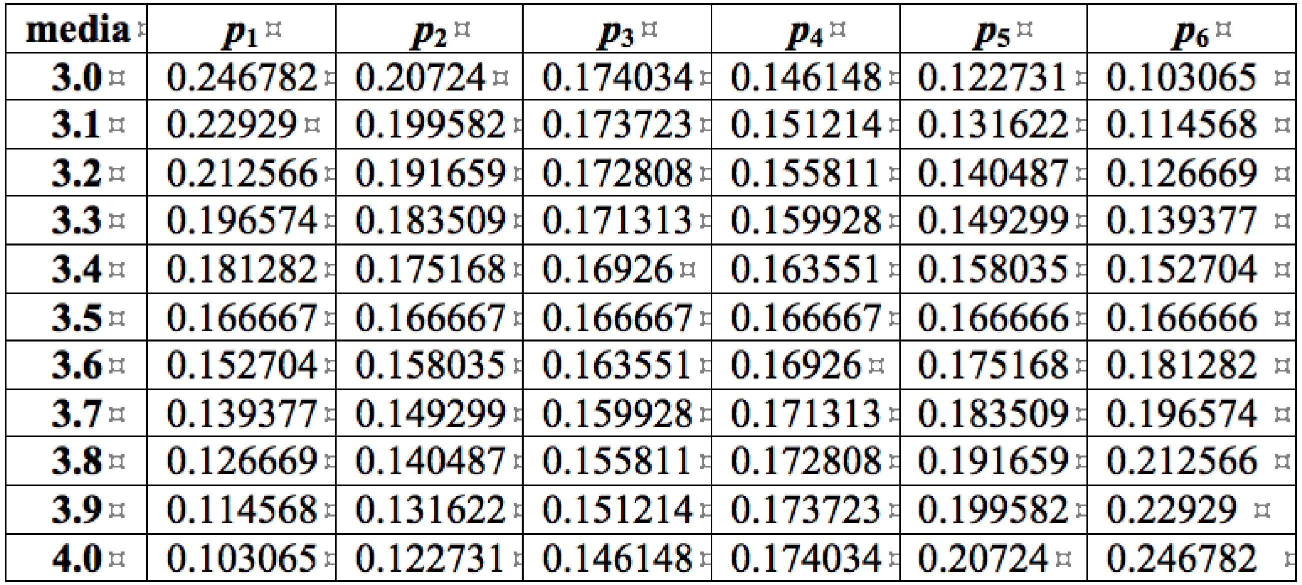
$$\Rightarrow \quad p_\ell = e^{\lambda_0 + \lambda_1 k - 1}$$

$$\boxed{\Rightarrow \begin{cases} e^{\lambda_0 - 1} \sum_{k=1}^{6} e^{\lambda_1 k} = 1 \\ e^{\lambda_0 - 1} \sum_{k=1}^{6} k\, e^{\lambda_1 k} = \frac{7}{2}(1+\varepsilon) \end{cases}}$$

we still have to satisfy the constraints ...

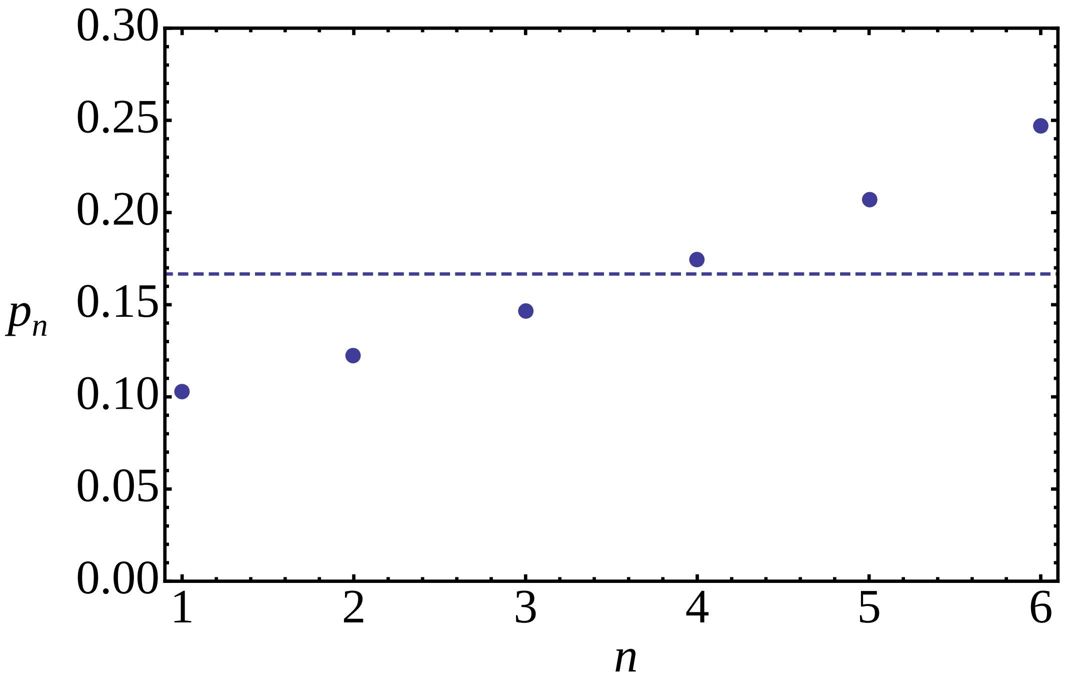
... we have to resort to numerical methods

numerical solution



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

Example: mean = 4



with a biased die we obtain skewed distributions.

These are examples of UNINFORMATIVE PRIORS

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

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

equivalent to the minimization of

$$-Q[p;m] = D_{KL}(p||m) - \sum_k \lambda_k \left(\int_a^b x^k p(x) dx - M_k \right)$$

This means that here we minimize the KL divergence with respect to the reference pdf $m(x)$ subject to the constraint(s).

variation

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

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

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

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

$p(x)$ is determined by the choice of $m(x)$ and by the constraints, in this case the moments of the distribution.

The Lagrange multipliers are determined by the equations

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

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

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

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

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

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

$$\Rightarrow \quad \lambda_0 = 1; \quad p(x) = m(x)\exp(\lambda_0 - 1) = \frac{1}{b-a}$$

2. first two moments are known, and $p(x)$ is defined on (a,b), so that

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

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

from which we obtain

$$M_0 = \frac{e^{\lambda_0 - 1}}{(b-a)\lambda_1}\left(e^{\lambda_1 b} - e^{\lambda_1 a}\right) = 1$$

$$M_1 = \frac{e^{\lambda_0 - 1}}{(b-a)\lambda_1^2}\left[(\lambda_1 b - 1)e^{\lambda_1 b} - (\lambda_1 a - 1)e^{\lambda_1 a}\right] = \mu$$

In general, this system can only be solved numerically

special case:

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

$$M_0 = \frac{e^{\lambda_0 - 1}}{\lambda_1 L}\left(e^{\lambda_1 L/2} - e^{-\lambda_1 L/2}\right) = \frac{e^{\lambda_0 - 1}}{\lambda_1 L/2}\sinh(\lambda_1 L/2) = 1$$

$$\begin{aligned} M_1 &= \frac{e^{\lambda_0 - 1}}{\lambda_1^2 L}\left[(\lambda_1 L/2 - 1)e^{\lambda_1 L/2} + (\lambda_1 L/2 + 1)e^{-\lambda_1 L/2}\right] \\ &= \frac{e^{\lambda_0 - 1}}{\lambda_1^2 L}\left[\lambda_1 L \cosh(\lambda_1 L/2) - 2\sinh(\lambda_1 L/2)\right] = 0 \end{aligned}$$

$$M_0 = \frac{e^{\lambda_0 - 1}}{\lambda_1 L/2} \sinh(\lambda_1 L/2) = 1$$

$$M_1 = \frac{e^{\lambda_0 - 1}}{\lambda_1^2 L} \left[\lambda_1 L \cosh(\lambda_1 L/2) - 2 \sinh(\lambda_1 L/2)\right] = 0$$

$$\tanh(\lambda_1 L/2) = \frac{\lambda_1 L}{2} \quad \Rightarrow \quad \lambda_1 = 0 \quad \Rightarrow \quad e^{\lambda_0 - 1} = 1 \quad \Rightarrow \lambda_0 = 1$$

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

another special case: $a = 0; \quad b \to \infty; \quad M_1 = \mu \neq 0$ (improper uniform distribution)

$$\begin{cases} M_0 = \dfrac{1}{b-a}\displaystyle\int_a^b \exp\left(\lambda_0 + \lambda_1 x - 1\right) dx = 1 \\ M_1 = \dfrac{1}{b-a}\displaystyle\int_a^b x \exp\left(\lambda_0 + \lambda_1 x - 1\right) dx = \mu \end{cases} \sim \begin{cases} m_0 \displaystyle\int_0^\infty \exp\left(\lambda_0 + \lambda_1 x - 1\right) dx = m_0 e^{\lambda_0 - 1} \dfrac{1}{(-\lambda_1)} = 1 \\ m_0 \displaystyle\int_0^\infty x \exp\left(\lambda_0 + \lambda_1 x - 1\right) dx = m_0 e^{\lambda_0 - 1} \dfrac{1}{\lambda_1^2} = \mu \end{cases}$$

$$\begin{cases} M_0 = 1 = m_0 e^{\lambda_0 - 1} \dfrac{1}{(-\lambda_1)} = 1 \\ M_1 = \mu = m_0 e^{\lambda_0 - 1} \left(\dfrac{1}{\lambda_1^2}\right) = -\dfrac{1}{\lambda_1} \end{cases} \quad \Rightarrow \quad -\lambda_1 = \frac{1}{\mu}$$

exponential distribution

$$p(x) = m(x) \exp\left(\sum_k \lambda_k x^k - 1\right) = m_0 e^{\lambda_0 - 1} \exp(\lambda_1 x) = m_0 e^{\lambda_0 - 1} \frac{1}{(-1\lambda_1)} (-1\lambda_1) \exp(\lambda_1 x) = \boxed{\frac{1}{\mu} \exp\left(-\frac{x}{\mu}\right)}$$

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

$$M_0 = m_0 \int_0^\infty \exp\left(\lambda_0 + \lambda_1 x - 1\right) dx = 1$$

$$M_1 = m_0 \int_0^\infty x \exp\left(\lambda_0 + \lambda_1 x - 1\right) dx = \mu$$

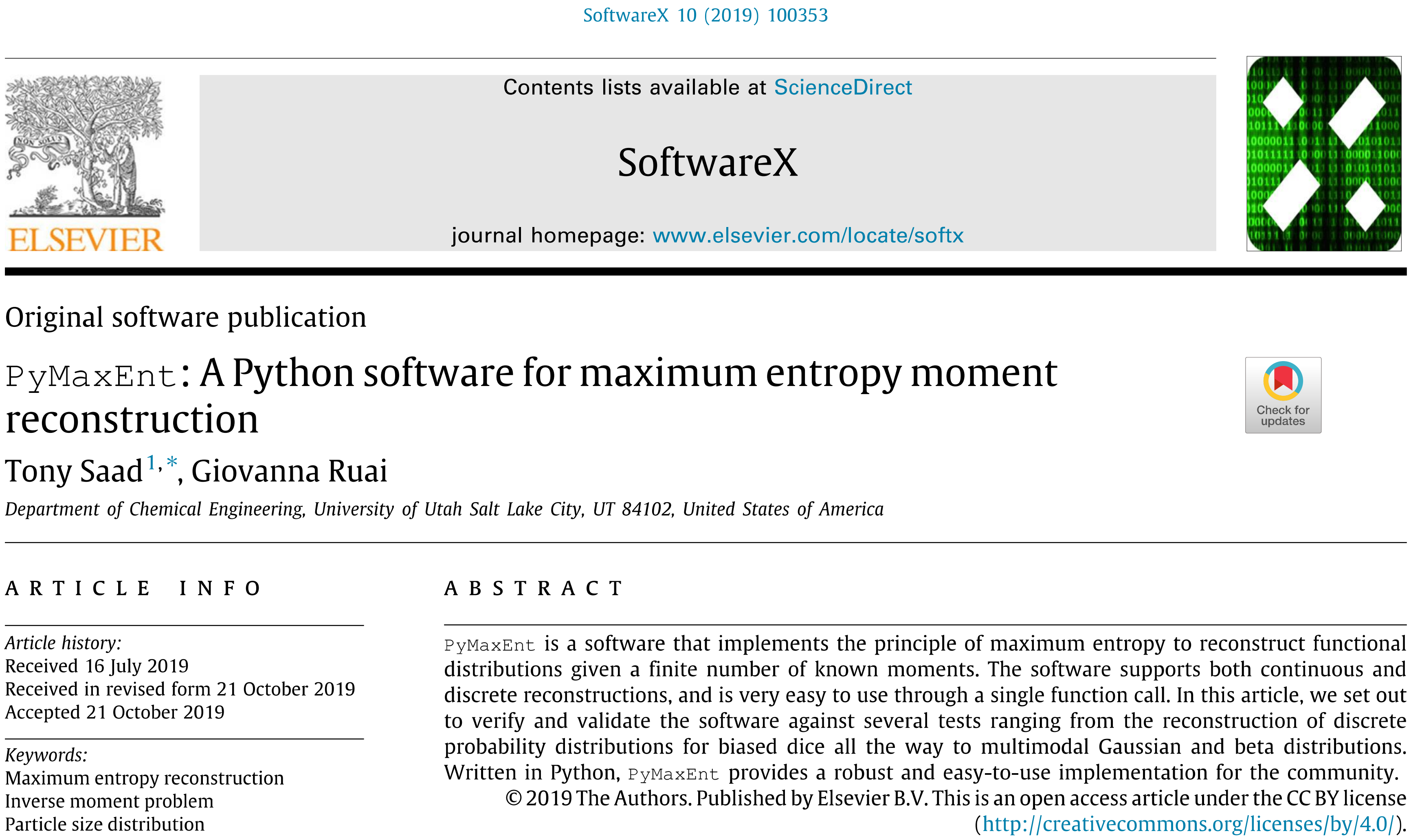
$$M_2 = m_0 \int_0^\infty x^2 \exp\left(\lambda_0 + \lambda_1 x - 1\right) dx = \langle x^2 \rangle$$

starting from these expressions, show that in this case

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

$$p(x) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{(x-\mu)^2}{2\sigma^2}\right)$$

i.e., the entropic prior is a Gaussian pdf



SoftwareX 10 (2019) 100353

Contents lists available at ScienceDirect

SoftwareX

journal homepage: www.elsevier.com/locate/softx

Original software publication

PyMaxEnt: A Python software for maximum entropy moment reconstruction

Tony Saad[1],*, 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.

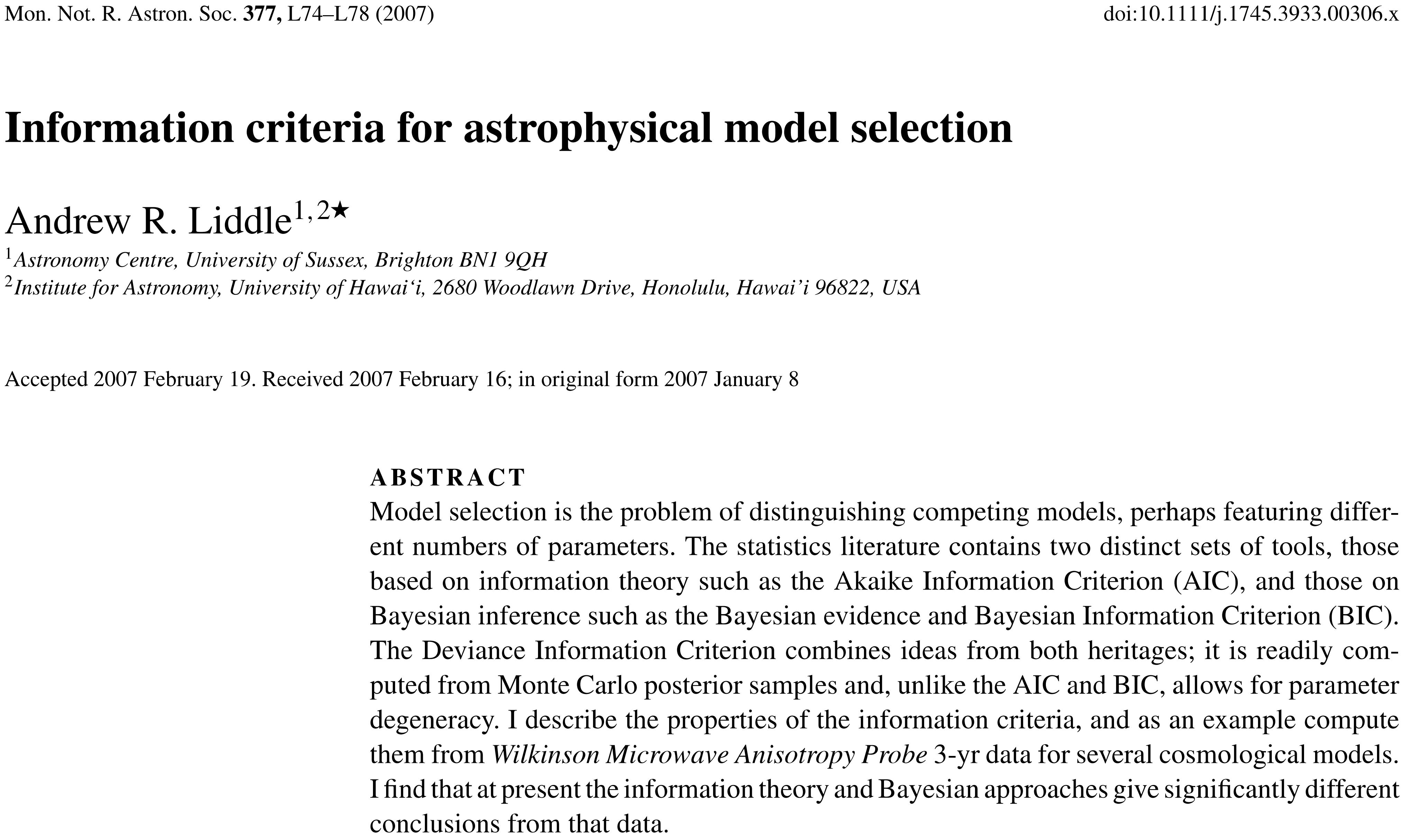
https://www.sciencedirect.com/science/article/pii/S2352711019302456
https://github.com/saadgroup/PyMaxEnt

A short overview of model selection methods

The generic purpose of a model selection statistic is to set up a tension between the predictiveness of a model (for instance indicated by the number of free parameters) and its ability to fit observational data. Oversimplistic models offering a poor fit should of course be thrown out, but so should more complex models that offer poor predictive power.

There are two main types of model selection statistic that have been used in the literature so far. Information criteria look at the best-fitting parameter values and attach a penalty for the number of parameters; they are essentially a technical formulation of "chi-squared per degrees of freedom" arguments. By contrast, the Bayesian evidence applies the same type of likelihood analysis familiar from parameter estimation, but at the level of models rather than parameters. It depends on goodness of fit across the entire model parameter space.

(Liddle & al., 2006 – Astronomy & Geophysics, Volume 47, Issue 4, pp. 4.30-4.33)



Mon. Not. R. Astron. Soc. **377,** L74–L78 (2007) doi:10.1111/j.1745.3933.00306.x

Information criteria for astrophysical model selection

Andrew R. Liddle[1,2]★

[1]*Astronomy Centre, University of Sussex, Brighton BN1 9QH*

[2]*Institute for Astronomy, University of Hawai‘i, 2680 Woodlawn Drive, Honolulu, Hawai’i 96822, USA*

Accepted 2007 February 19. Received 2007 February 16; in original form 2007 January 8

ABSTRACT

Model selection is the problem of distinguishing competing models, perhaps featuring different numbers of parameters. The statistics literature contains two distinct sets of tools, those based on information theory such as the Akaike Information Criterion (AIC), and those on Bayesian inference such as the Bayesian evidence and Bayesian Information Criterion (BIC). The Deviance Information Criterion combines ideas from both heritages; it is readily computed from Monte Carlo posterior samples and, unlike the AIC and BIC, allows for parameter degeneracy. I describe the properties of the information criteria, and as an example compute them from *Wilkinson Microwave Anisotropy Probe* 3-yr data for several cosmological models. I find that at present the information theory and Bayesian approaches give significantly different conclusions from that data.

Interlude: the Likelihood Ratio Method and Wilks' theorem – 1

- Taylor expansion close to the true value of the parameter(s)

$$\frac{\partial \ln L(D|\theta)}{\partial \theta} \approx -\left.\frac{\partial^2 \ln L(D|\theta)}{\partial \theta^2}\right|_{\theta=\theta_0} (\theta-\theta_0) \approx -E\left[\left.\frac{\partial^2 \ln L(D|\theta)}{\partial \theta^2}\right|_{\theta=\theta_0}\right](\theta-\theta_0)$$

- Integration

$$L(D|\theta) \propto \exp\left\{-\frac{1}{2}E\left[\left.\frac{\partial^2 \ln L(D|\theta)}{\partial \theta^2}\right|_{\theta=\theta_0}\right](\theta-\theta_0)^2\right\}$$

- Extension to more than one parameter (with parameters split into two subsets)

$$L(D|\boldsymbol{\theta}) = L(D|\boldsymbol{\theta}_r, \boldsymbol{\theta}_s) \propto \exp\left[-\frac{1}{2}(\boldsymbol{\theta}-\boldsymbol{\theta}_0)^T I (\boldsymbol{\theta}-\boldsymbol{\theta}_0)\right]$$

where Fisher's information matrix is split into submatrices

$$I = \begin{pmatrix} I_{rr} & \vdots & I_{rs} \\ \cdots & & \cdots \\ I_{sr} & \vdots & I_{ss} \end{pmatrix}$$

Interlude: the Likelihood Ratio Method and Wilks' theorem – 1

- Then, $\boldsymbol{\theta} = \begin{pmatrix} \boldsymbol{\theta}_r \\ \boldsymbol{\theta}_s \end{pmatrix}$ and therefore

$$L(D|\boldsymbol{\theta}_r, \boldsymbol{\theta}_s) \propto \exp\Bigg[-\frac{1}{2}(\boldsymbol{\theta}_r - \boldsymbol{\theta}_{0,r})^T I_{rr}(\boldsymbol{\theta}_r - \boldsymbol{\theta}_{0,r}) \\ -(\boldsymbol{\theta}_r - \boldsymbol{\theta}_{0,r})^T I_{rs}(\boldsymbol{\theta}_s - \boldsymbol{\theta}_{0,s}) - \frac{1}{2}(\boldsymbol{\theta}_s - \boldsymbol{\theta}_{0,s})^T I_{ss}(\boldsymbol{\theta}_s - \boldsymbol{\theta}_{0,s})\Bigg]$$

- When we maximize the likelihood with respect to the whole parameter vector, we find that the estimators for the subvectors are

$$\theta'_r = \hat{\boldsymbol{\theta}}_r; \quad \theta'_s = \hat{\boldsymbol{\theta}}_s$$

and the corresponding maximum likelihood has a fixed value that depends only on data.

Interlude: the Likelihood Ratio Method and Wilks' theorem – 1

- When we maximize the likelihood with respect to the *s* parameters only, we find

$$
\begin{aligned}
L(D|\boldsymbol{\theta}_r, \hat{\boldsymbol{\theta}}_s) &\propto \exp\left[-\frac{1}{2}(\boldsymbol{\theta}_r - \boldsymbol{\theta}_{0,r})^T I_{rr}(\boldsymbol{\theta}_r - \boldsymbol{\theta}_{0,r}) - (\boldsymbol{\theta}_r - \boldsymbol{\theta}_{0,r})^T I_{rs}(\hat{\boldsymbol{\theta}}_s - \boldsymbol{\theta}_{0,s})\right] \\
&\propto \exp\left[-\frac{1}{2}(\boldsymbol{\theta}_r - \boldsymbol{\theta}_{0,r} - b_D)^T I_{rr}(\boldsymbol{\theta}_r - \boldsymbol{\theta}_{0,r} - b_D)\right]
\end{aligned}
$$

- This means that the statistic

$$
\begin{aligned}
\lambda &= -2L(D|\boldsymbol{\theta}_r, \hat{\boldsymbol{\theta}}_s) \\
&\sim (\boldsymbol{\theta}_r - \boldsymbol{\theta}_{0,r} - b_D)^T I_{rr}(\boldsymbol{\theta}_r - \boldsymbol{\theta}_{0,r} - b_D) \\
&\approx (\boldsymbol{\theta}_r - \boldsymbol{\theta}_{0,r})^T I_{rr}(\boldsymbol{\theta}_r - \boldsymbol{\theta}_{0,r})
\end{aligned}
$$

(where the bias vanishes asymptotically) has a chi-square distribution with *r* degrees of freedom for large *n* (Wilks' theorem).