Introduction to Bayesian Methods - 6

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 (ctd.)
4. Jeffreys' method
5. Reference priors
6. The Maximum Entropy Method

Information theoretic concepts in statistics – additivity of entropy

If symbols are emitted simultaneously and independently by two sources, the joint probability distribution is

$$p(j,k) = p_1(j)p_2(k)$$

and therefore, the joint entropy is

$$\begin{aligned} H &= -\sum_{j,k} p(j,k)\log_2 p(j,k) = -\sum_{j,k} p_1(j)p_2(k)\log_2[p_1(j)p_2(k)] \\ &= -\sum_{j} p_1(j)\log_2 p_1(j) - \sum_{k} p_2(k)\log_2 p_2(k) \\ &= H_1 + H_2 \end{aligned}$$

Information theoretic concepts in statistics – the uniform distribution has maximal entropy

This is an easy result that follows using one Lagrange multiplier to keep probability normalization into account

$$H + \lambda \sum_{k=1}^{N} p_k = -\sum_{k=1}^{N} p_k \log_2 p_k + \lambda \sum_{k=1}^{N} p_k$$

$$= -\frac{1}{\ln 2} \sum_{k=1}^{N} p_k \ln p_k + \lambda \sum_{k=1}^{N} p_k$$

$$\Rightarrow \quad \frac{\partial}{\partial p_j}\left(H + \lambda \sum_{k=1}^{N} p_k\right) = -\frac{1}{\ln 2}(\ln p_j + 1) + \lambda = 0$$

$$\Rightarrow \quad p_j = \exp(\lambda \ln 2 - 1) = 1/N$$

all probabilities have the same value

Information theoretic concepts in statistics – differential entropy

The Shannon entropy cannot be extended to continuous distribution in a straightforward way. Consider a discretized version of the probability distribution:

$$P_k = \int_{k\Delta}^{(k+1)\Delta} p(x)dx = p(x_k^*)\Delta \quad , \text{where } x_k^* \in (k\Delta, (k+1)\Delta)$$

$$H = -\sum_k P_k \ln P_k$$

$$= -\sum_k p(x_k^*) \ln p(x_k^*)\Delta - \sum_k p(x_k^*)\Delta \ \ln \Delta \approx \underbrace{-\int p(x) \ln p(x)dx}_{\text{differential entropy}} - \ln \Delta$$

Information theoretic concepts in statistics – relative entropy

Considering two sets of symbols (same number of symbols), we can consider the relative information carried by each symbol in one set with respect to the corresponding one in the other set

$$(-\log_2 p_k) - (-\log_2 g_k) = -\log_2 \frac{p_k}{g_k}$$

Then, the average difference of the information carried by the p_k's with respect to the reference set (the **relative entropy**) is

$$H_R = -\sum_k p_k \log_2 \frac{p_k}{g_k}$$

This extends without problems to continuous distributions

$$H_R = -\int p(x) \log_2 \frac{p(x)}{g(x)} dx$$

Information theoretic concepts in statistics – the Kullback-Leibler divergence

The Kullback-Leibler (KL) divergence is a simple redefinition of the relative entropy

$$H_R = -\int p(x) \log_2 \frac{p(x)}{g(x)} dx \quad \Rightarrow \quad D_{\mathrm{KL}}(p||g) = \int p(x) \ln \frac{p(x)}{g(x)} dx$$

- Natural logs instead of logs base 2
- Change of sign
- NOT symmetrical with respect to the p,g exchange
- Has several interesting properties

Information theoretic concepts in statistics – Jensen's inequality

Consider a convex function, then

$$f(a + (b - a)t) \leq f(a) + [f(b) - f(a)]t$$

$$\Rightarrow \quad f[a(1 - t) + bt] \leq f(a)(1 - t) + f(b)t$$

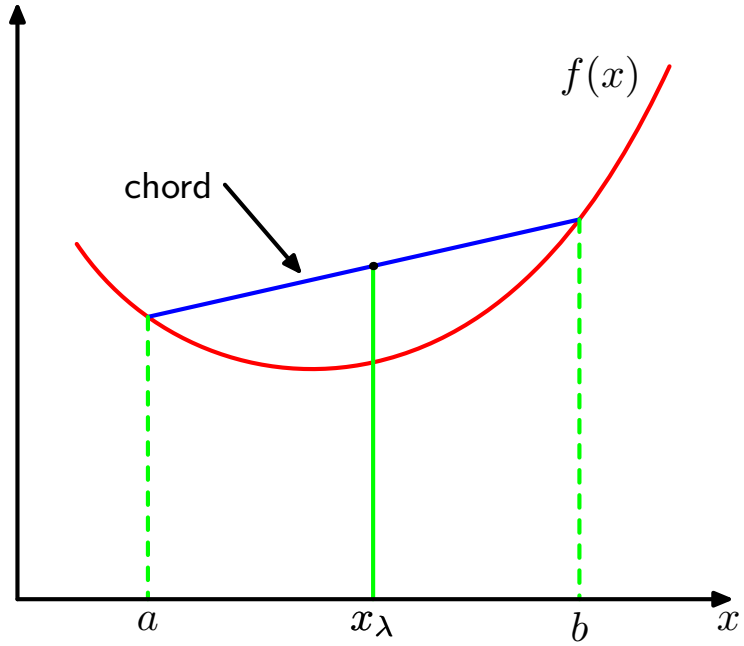
which can be written as

$$f[at_1 + bt_2] \leq f(a)t_1 + f(b)t_2 \quad \text{with } t_1 + t_2 = 1$$

Now, we conjecture the extension

$$f\left(\sum_{k=1}^{n} x_k t_k\right) \leq \sum_{k=1}^{n} f(x_k)t_k \quad \text{with } \sum_{k=1}^{n} t_k = 1$$

and prove the inequality by induction.



A convex function *f*(*x*) is one for which every chord (shown in blue) lies on or above the function (shown in red)

Information theoretic concepts in statistics – Jensen's inequality – 2

If

$$f\left(\sum_{k=1}^{n+1} x_k t_k\right) \leq \sum_{k=1}^{n+1} f(x_k) t_k \qquad \text{with} \sum_{k=1}^{n+1} t_k = 1$$

then isolate the (n+1)-th parameter

$$\sum_{k=1}^{n} t_k = 1 - t_{n+1}, \qquad \text{so that} \qquad \sum_{k=1}^{n} \frac{t_k}{1 - t_{n+1}} = 1$$

and rearrange the l.h.s. of the inequality

$$\begin{aligned} f\left(\sum_{k=1}^{n+1} x_k t_k\right) &= f\left((1 - t_{n+1}) \sum_{k=1}^{n} x_k \frac{t_k}{1 - t_{n+1}} + x_{n+1} t_{n+1}\right) \\ &\leq f\left(\sum_{k=1}^{n} x_k \frac{t_k}{1 - t_{n+1}}\right)(1 - t_{n+1}) + f(x_{n+1}) t_{n+1} \\ &\leq \sum_{k=1}^{n} f(x_k) \frac{t_k}{1 - t_{n+1}} (1 - t_{n+1}) + f(x_{n+1}) t_{n+1} = \sum_{k=1}^{n+1} f(x_k) t_k \end{aligned}$$

Information theoretic concepts in statistics – the Kullback-Leibler divergence - 2

Jensen's inequality can be restated in a simple way if the t's are mapped into probabilities

$$f\left(\sum_{k=1}^{n} x_k p_k\right) \leq \sum_{k=1}^{n} f(x_k) p_k \quad \Rightarrow \quad f\left[\mathbb{E}(x)\right] \leq \mathbb{E}[f(x)]$$

Equivalently

$$f\left(\int x p(x) dx\right) \leq \int f(x) p(x) dx$$

Now we can apply the inequality to the KL divergence (the -log function is convex) and find

$$D_{KL}(p||g) = -\int p(x) \ln \frac{g(x)}{p(x)} dx \geq -\ln\left(\int g(x) dx\right) = 0$$

Information theoretic concepts in statistics – The KL divergence is a quasi-metric (however a local version of the KL divergence is the Fisher information, which is a true metric)

The KL divergence can be used to measure the "distance" between two distributions.

Example: the KL divergence

$$D_{\mathrm{KL}}(p||q) = \int_{-\infty}^{+\infty} p(x) \ln \frac{p(x)}{q(x)} dx$$

for the distributions

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

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

$$D_{\mathrm{KL}}(p||q) = \frac{\mu^2}{2\sigma^2}$$

Now consider a family of parametric distributions and evaluate the KL divergence between two close elements of the family

$$D_{\mathrm{KL}}\left[p(x,\theta)||p(x,\theta+\epsilon)\right] = \int_{-\infty}^{+\infty} p(x,\theta)\ln\frac{p(x,\theta)}{p(x,\theta+\epsilon)}dx$$
$$= \mathbb{E}\left[\ln p(x,\theta) - \ln p(x,\theta+\epsilon)\right]$$

Since

$$\ln p(x,\theta+\epsilon) \approx \ln p(x,\theta) + \frac{\partial \ln p(x,\theta)}{\partial\theta}\epsilon + \frac{1}{2}\frac{\partial^2 \ln p(x,\theta)}{\partial\theta^2}\epsilon^2$$

we find, using the first Bartlett identity,

$$D_{\mathrm{KL}}\left[p(x,\theta)||p(x,\theta+\epsilon)\right] = -\mathbb{E}\left(\frac{\partial \ln p(x,\theta)}{\partial\theta}\epsilon + \frac{1}{2}\frac{\partial^2 \ln p(x,\theta)}{\partial\theta^2}\epsilon^2\right)$$
$$= -\frac{1}{2}\mathbb{E}\left[\frac{\partial^2 \ln p(x,\theta)}{\partial\theta^2}\right]\epsilon^2 = \frac{1}{2}I(\theta)\epsilon^2$$

i.e., locally the KL divergence is just the Fisher information

Information theoretic concepts in statistics – The KL divergence can be transformed into a true distance between pdf's

- Jeffreys' distance

$$D_{\mathrm{J}}(p||q) = \frac{1}{2}D_{\mathrm{KL}}(p||q) + \frac{1}{2}D_{\mathrm{KL}}(q||p)$$

- Jensen-Shannon distance

$$D_{\mathrm{JS}}(p||q) = \frac{1}{2}D_{\mathrm{KL}}\left(p||\frac{p+q}{2}\right) + \frac{1}{2}D_{\mathrm{KL}}\left(q||\frac{p+q}{2}\right)$$

Information theoretic concepts in statistics – Using the KL divergence

Suppose that data is being generated from an unknown distribution $p(x)$ that we wish to model. We can try to approximate this distribution using some parametric distribution $q(x|\boldsymbol{\theta})$, governed by a set of adjustable parameters $\boldsymbol{\theta}$, for example, a multivariate Gaussian.

One way to determine $\boldsymbol{\theta}$ is to minimize the Kullback-Leibler divergence between $p(x)$ and $q(x|\boldsymbol{\theta})$ with respect to $\boldsymbol{\theta}$. We cannot do this directly because we don't know $p(x)$. Suppose, however, that we have observed a finite set of training points x_n, for $n = 1, \ldots, N$, drawn from $p(x)$ (an **empirical distribution**). Then the expectation with respect to $p(x)$ can be approximated by a finite sum over these points

$$D_{KL}(p||q) = \mathbb{E}_p\left[\ln \frac{p(x)}{q(x|\boldsymbol{\theta})}\right] \approx \frac{1}{N}\sum_{n=1}^{N}\left[-\ln q(x_n|\boldsymbol{\theta}) + \ln p(x_n)\right]$$

likelihood

parameter-independent term

From this equation we see that we can obtain an approximate distribution by minimizing the KL divergence, i.e., by maximizing the likelihood.

Information theoretic concepts in statistics – Mutual information

We can use the KL divergence to measure the degree of statistical dependence between pairs of variates, by measuring the distance between $p(\mathbf{x},\mathbf{y})$ and $p(\mathbf{x})p(\mathbf{y})$

$$I(\mathbf{x},\mathbf{y}) = D_{\mathrm{KL}}[p(\mathbf{x},\mathbf{y})||p(\mathbf{x})p(\mathbf{y})] = \int p(\mathbf{x},\mathbf{y}) \ln \frac{p(\mathbf{x},\mathbf{y})}{p(\mathbf{x})p(\mathbf{y})} d\mathbf{x}d\mathbf{y}$$

This quantity is called the **mutual information**.

We also find,

$$\begin{aligned} I(\mathbf{x},\mathbf{y}) &= -\int p(\mathbf{x},\mathbf{y}) \ln \frac{p(\mathbf{x})p(\mathbf{y})}{p(\mathbf{x},\mathbf{y})} d\mathbf{x}d\mathbf{y} \\ &= -\int p(\mathbf{x}|\mathbf{y})p(\mathbf{y}) \ln \frac{p(\mathbf{x})p(\mathbf{y})}{p(\mathbf{x}|\mathbf{y})p(\mathbf{y})} d\mathbf{x}d\mathbf{y} \\ &= -\int p(\mathbf{x}|\mathbf{y})p(\mathbf{y}) \left[\ln p(\mathbf{x}) - \ln p(\mathbf{x}|\mathbf{y})\right] d\mathbf{x}d\mathbf{y} \\ &= -\int p(\mathbf{x}) \ln p(\mathbf{x}) d\mathbf{x} + \int p(\mathbf{x},\mathbf{y}) \ln p(\mathbf{x}|\mathbf{y}) d\mathbf{x}d\mathbf{y} \\ &= H[\mathbf{x}] - H[\mathbf{x}|\mathbf{y}] \\ &= H[\mathbf{y}] - H[\mathbf{y}|\mathbf{x}] \end{aligned}$$

differential entropy

conditional entropy

we can view the mutual information as the reduction in the uncertainty about **x** by virtue of being told the value of **y** (or vice versa).

An invariant form for the prior probability in estimation problems

By Harold Jeffreys, F.R.S.

(*Received* 23 *November* 1945)

It is shown that a certain differential form depending on the values of the parameters in a law of chance is invariant for all transformations of the parameters when the law is differentiable with regard to all parameters. For laws containing a location and a scale parameter a form with a somewhat restricted type of invariance is found even when the law is not everywhere differentiable with regard to the parameters. This form has the properties required to give a general rule for stating the prior probability in a large class of estimation problems.

Jeffreys' priors – the KL divergence is invariant with respect to generic random variable transformations.

From the definition of KL divergence, and from the transformation formula for pdf's we find

$$\int_{-\infty}^{+\infty} p_y(y) \ln\left(\frac{p_y(y)}{q_y(y)}\right) dy = \int_{-\infty}^{+\infty} p_x(x) \ln\left(\frac{p_x(x)\left|\frac{dx}{dy}\right|}{q_x(x)\left|\frac{dx}{dy}\right|}\right) dx$$

$$= \int_{-\infty}^{+\infty} p_x(x) \ln\left(\frac{p_x(x)}{q_x(x)}\right) dx$$

In this case, our random variables are the parameter estimates, therefore the KL divergence is invariant with respect to parameter (random variable transformations), **therefore the associated Fisher Information from the local expansion of the KL divergence is also invariant with respect to parameter transformations**.

From the equation that relates KL divergence and Fisher Information, we find a corresponding pdf as follows. Equation

$$D_{\mathrm{KL}}\left[p(x|\theta)||p(x|\theta+\epsilon)\right] = -\frac{1}{2}\mathbb{E}\left[\frac{\partial^2 \ln p(x|\theta)}{\partial\theta^2}\right]\epsilon^2 = \frac{1}{2}I(\theta)\epsilon^2$$

means that the KL divergence depends quadratically on small changes of the expansion parameter and that the KL divergence remains constant if the term on the r.h.s. remains constant.

Dimensionally, the Fisher information is quadratic with respect to a pdf, therefore we take its square root to define a pdf, i.e.,

$$f(\theta) \sim \sqrt{I(\theta)}$$

This must be normalized to obtain a pdf that is invariant with respect to parameter transformations.

Example: a simple Gaussian Likelihood for *n* datapoints, with known variance

$$L(D|\mu) = \prod_n \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{(x_n-\mu)^2}{2\sigma^2}\right)$$

$$\Rightarrow \quad \ln L(D|\mu) \sim \sum_n \left(-\ln\sigma - \frac{(x_n-\mu)^2}{2\sigma^2}\right)$$ fixed sigma

$$\Rightarrow \quad I(\mu) = \mathbb{E}\left[-\frac{\partial^2 \ln L(D|\mu)}{\partial\mu^2}\right] \sim \text{constant}$$

This points to a uniform prior for μ. In general, this uniform prior is an improper prior.

Example: a simple Gaussian Likelihood for *n* datapoints, with known mean

$$L(D|\mu) = \prod_n \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{(x_n-\mu)^2}{2\sigma^2}\right)$$

$$I(\sigma) = \mathbb{E}\left[-\frac{\partial^2 \ln L(D|\sigma)}{\partial\sigma^2}\right] \sim \frac{1}{\sigma^2}$$

fixed mu

$$\sqrt{I(\sigma)} \sim \frac{1}{\sigma}$$

This power-law pdf is another improper prior.

Example: Poisson distribution

$$L(D|a) = \prod_n \frac{a^{k_n}}{k_n!} e^{-a}$$

$$I(a) = \mathbb{E}\left[-\frac{\partial^2 \ln L(D|a)}{\partial a^2}\right] \sim \frac{1}{a}$$

$$\sqrt{I(a)} \sim \frac{1}{\sqrt{a}}$$

This power-law pdf is yet another improper prior.

Example: binomial distribution

$$L(D|\theta) = \binom{N}{n} \theta^n (1-\theta)^{N-n}$$

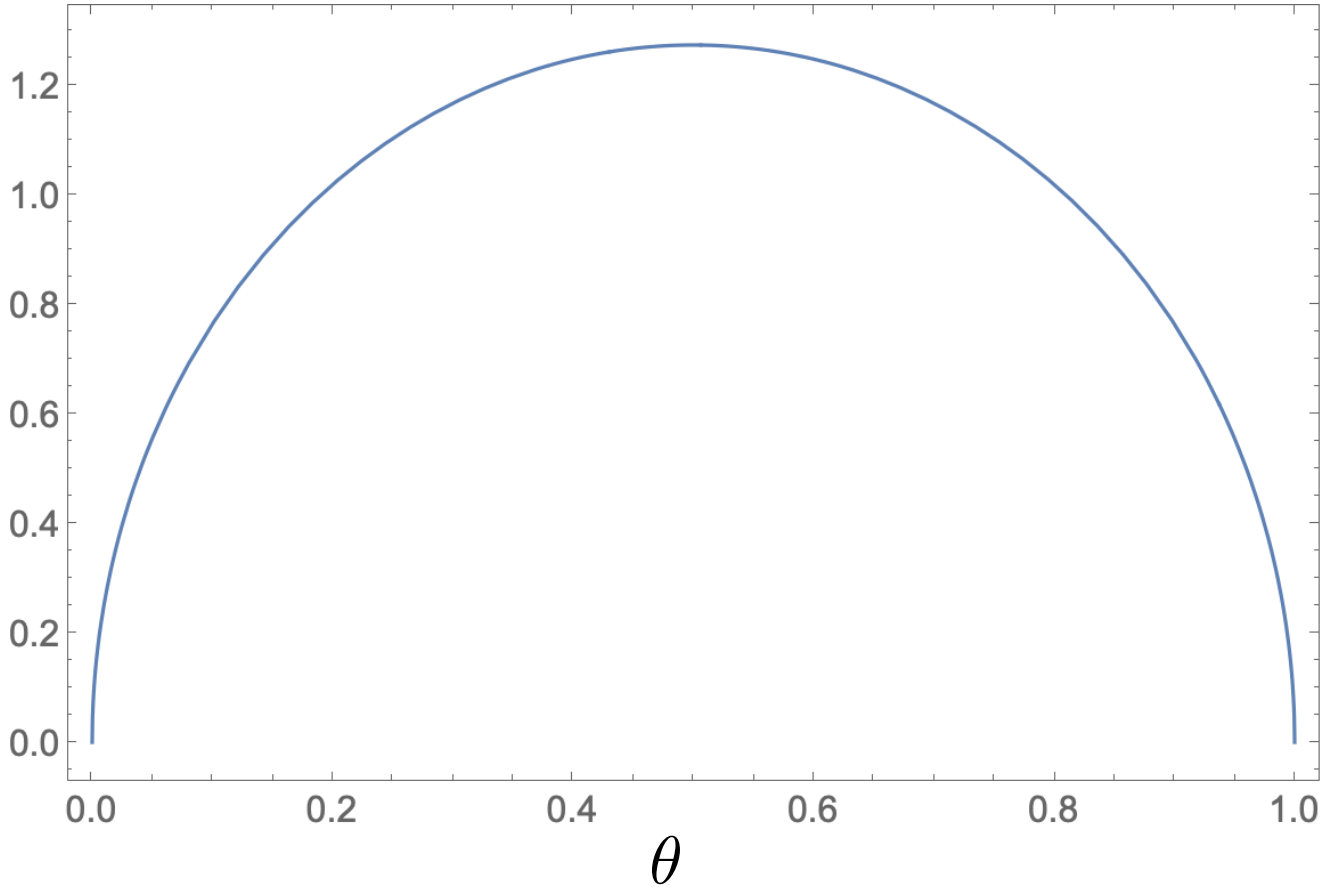
$$\ln L(D|\theta) \sim n \ln \theta + (N-n) \ln(1-\theta)$$

$$\mathbb{E}\left[-\frac{\partial^2 \ln L(D|\theta)}{\partial \theta^2} \right] \sim \frac{N\theta}{\theta^2} + \frac{N - N\theta}{(1-\theta)^2}$$

$$= \frac{N}{\theta} + \frac{N}{1-\theta}$$

$$= \frac{N}{\theta(1-\theta)}$$

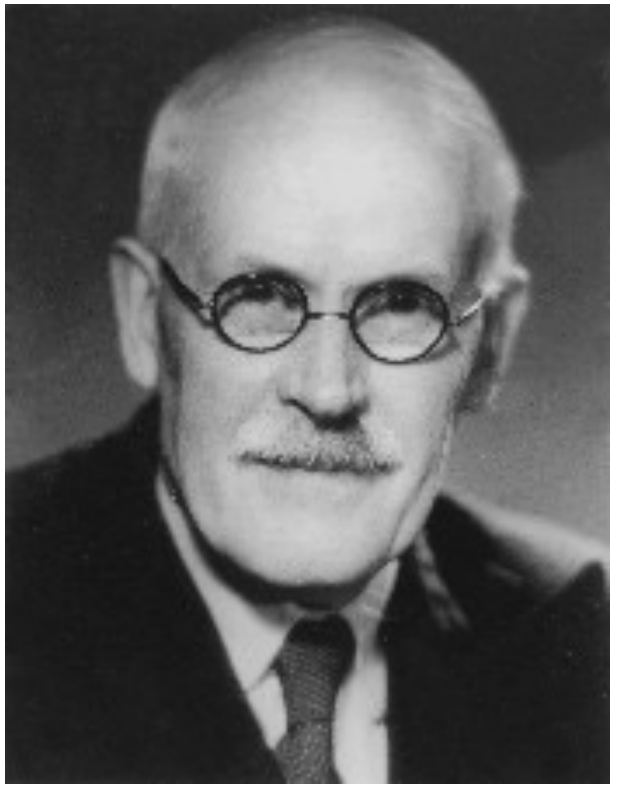
$$I(\theta) \sim \frac{1}{\theta(1-\theta)} \quad \Rightarrow \quad \sqrt{I(\theta)} \sim \frac{\theta^{1/2}(1-\theta)^{1/2}}{B(3/2,3/2)} = \text{Beta}(\theta|3/2,3/2)$$



A lesson learned from Jeffreys' priors

Jeffreys priors are tuned to the Likelihood, but doesn't this sound strange? Shouldn't the prior distribution be related to the prior information alone?

Well ... no, the Likelihood is also constructed using prior information (obviously!). So, in this approach the Likelihood and the priors are both determined using the available prior information.



Harold Jeffreys
(1891-1989)

Additional comments on Jeffreys' priors

- In general, they are NOT conjugate priors, but are limits of conjugate priors
- They work well for single parameter models, but NOT for multivariate models

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 $t = T(\mathbf{D})$ 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.

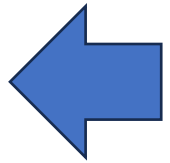
The idea behind a reference prior is that it must be such that data affect our posterior distribution the most.

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

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

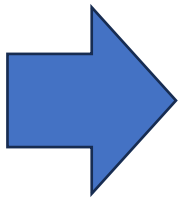
Underdetermined system of linear equations

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

Maximization of constrained entropy

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

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



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

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

Solution of the nonlinear system of equations

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

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

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

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

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

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

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:

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

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

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

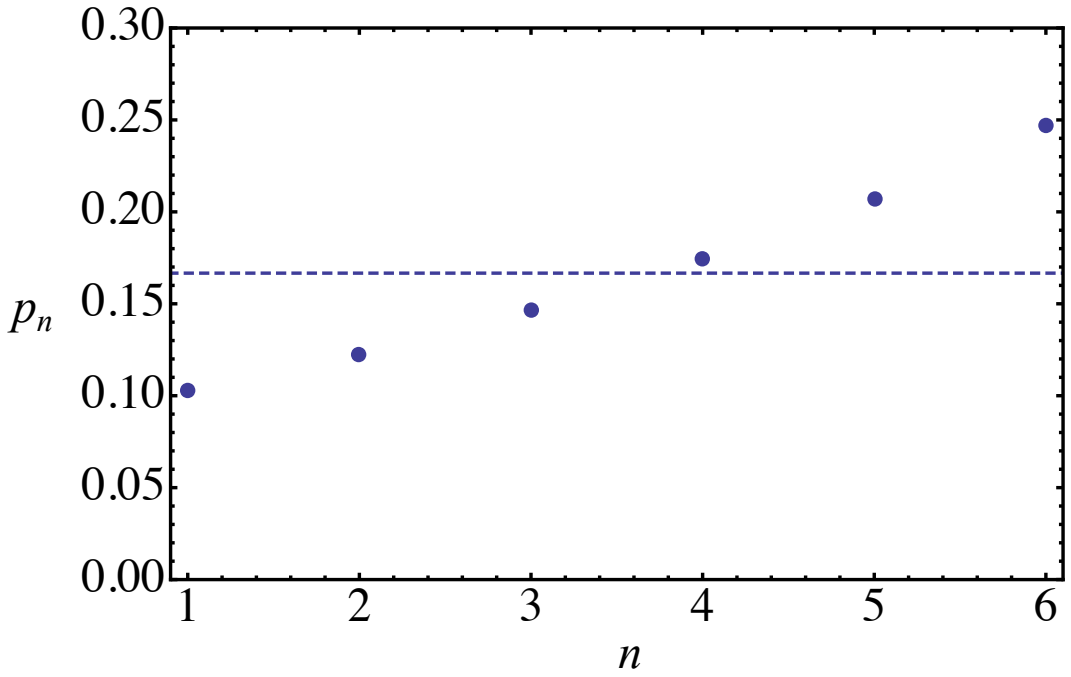
we still have to satisfy the constraints ...

... 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

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

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

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

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

$p(x)$ 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:

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

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

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

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

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

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

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

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

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

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

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