Probability Theory revisited:
Jaynes' views on Bayesian statistics

Christian P. Robert

Université Paris Dauphine, IuF, and CREST-INSEE
http://www.ceremade.dauphine.fr/~xian

April 10, 2011

Outline

1. Goals of the course
2. Elementary hypothesis testing
3. Elementary parameter estimation
4. Sufficiency, ancillarity and all that
5. The entropy principle
6. Decision theory

First chapter: Goals

1. Goals of the course
    - E.T. Jaynes
    - Against Bourbaki
2. Elementary hypothesis testing
3. Elementary parameter estimation
4. Sufficiency, ancillarity and all that
5. The entropy principle
6. Decision theory

E.T. Jaynes (1922–1998)

Professor of Physics at Washington University in St. Louis. He wrote extensively on statistical mechanics and on foundations of probability and statistical inference, initiating in 1957 the *MaxEnt* interpretation of thermodynamics, as being a particular application of more general Bayesian/information theory techniques.
In 1963, together with Fred Cummings, he modeled the evolution of a two-level atom in an electromagnetic field, in a fully quantized way. This model is known as the Jaynes–Cummings model.

[Wikipedia]

Jaynes and Bayesian statistics

Jaynes strongly promoted the *interpretation of probability theory as an extension of logic*. A particular focus of his work was the construction of logical principles for assigning prior probability distributions; see the principle of maximum entropy, the principle of transformation groups and Laplace's principle of indifference.

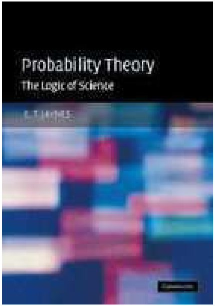
[Wikipedia]

Jaynes and Probability Theory

Although started in 1956, the book *Probability Theory* was published postumously, in 2003, thanks to the editing of Larry Bretthorst.

Jaynes planned two volumes with applications and computer programs (“written in a particularly obscure form of BASIC”).

Main aim of the book is to present *probability theory as extended logic*. Contains Bayesian methods and the Principle of maximum entropy.



Foundations of Probability Theory

"We find ourselves, to our own surprise, in agreement with Kolmogorov" – E.T. Jaynes, p.xxi

"On many technical issues we disagree strongly with de Finetti. It appears to us that his way of treating infinite sets has opened up a Pandora's box of useless and unecessary paradoxes." – E.T. Jaynes, p.xxi

"We sail under the banner of Gauß, Kronecker, and Poincaré rather than Cantor, Hilbert, and Bourbaki." – E.T. Jaynes, p.xxii

Foundations of Probability Theory (2)

> *"There are no really trustworthy standards of rigor in a mathematics that embraced the theory of infinite sets."* – E.T. Jaynes, p.xxvii
> *"Paradoxes are avoided automatically: they cannot arise from correct application of our basic rules, because only finite sets and infinite sets that arise as well-defined and well-behaved limits of finite sets."* – E.T. Jaynes, p.xxii

Refusal to use measure theory as in e.g. Feller (1966, vol. 2), although using limits leads to inconsistencies and dependence on a specific $\sigma-$algebra.

Foundations of Probability Theory (3)

> *"It is ambiguous until we specify exactly what kind of limiting process we propose to use."* – E.T. Jaynes, p.108
> *"Consideration of a continuous variable is only an approximation to the exact discrete theory."* – E.T. Jaynes, p.109

The *everything is finite* assumption gives some intuition but breaks down in complex problems.

Further squibbles

> *"...inappropriate definition of the term 'function'. A delta-function is not a mapping from any set into any other."* – E.T. Jaynes, p.668
> *"The issue has nothing to do with mathematical rigor; it is simply one of notation."* – E.T. Jaynes, p.112

This leads Jaynes to use delta-functions in densities as, e.g. in (4.65), w/o reference to a dominating measure.

$$g(f|X) = \frac{10}{11}(-10^{-6})\delta(f-\frac{1}{6})+\frac{1}{11}(-10^{-6})\delta(f-\frac{1}{3})+10^{-6}\delta(f-\frac{99}{100})$$

confusing functions with measures...

Chapter 4: Elementary hypothesis testing

1. Goals of the course

2. Elementary hypothesis testing
   - Binary choice and Bayesian principles
   - Beyond binary
   - Infinite number of hypotheses

3. Elementary parameter estimation

4. Sufficiency, ancillarity and all that

5. The entropy principle

6. Decision theory

A single hypothesis

"To form a judgement about the likely truth or falsity of any proposition A, the correct procedure is to calculate the probability that A is true." (p.86)

A single hypothesis

> *"To form a judgement about the likely truth or falsity of any proposition A, the correct procedure is to calculate the probability that A is true."* (p.86)

The first part of Chapter 4 is about testing a null hypothesis versus its complement H_1

- Introduction of prior probabilities, conditional on "other information" X
- Immediate call to posterior probabilities and Bayes formula
- Use of the likelihood

Bayes Theorem

Bayes theorem = Inversion of probabilities

If A and E are events such that $P(E) \neq 0$, $P(A|E)$ and $P(E|A)$ are related by

$$P(A|E) = \frac{P(E|A)P(A)}{P(E|A)P(A) + P(E|A^c)P(A^c)} = \frac{P(E|A)P(A)}{P(E)}$$

Bayes Theorem

Bayes theorem = Inversion of probabilities

If A and E are events such that $P(E) \neq 0$, $P(A|E)$ and $P(E|A)$ are related by

$$P(A|E) = \frac{P(E|A)P(A)}{P(E|A)P(A) + P(E|A^c)P(A^c)} = \frac{P(E|A)P(A)}{P(E)}$$

[Thomas Bayes (?)]

Who is Thomas Bayes?

Reverend Thomas Bayes (ca. 1702–1761)

Presbyterian minister in Tunbridge Wells (Kent) from 1731, son of Joshua Bayes, nonconformist minister. Election to the *Royal Society* based on a tract of 1736 where he defended the views and philosophy of Newton.

Who is Thomas Bayes?

Reverend Thomas Bayes (ca. 1702–1761)

Presbyterian minister in Tunbridge Wells (Kent) from 1731, son of Joshua Bayes, nonconformist minister. Election to the *Royal Society* based on a tract of 1736 where he defended the views and philosophy of Newton.
His sole probability paper, *“Essay Towards Solving a Problem in the Doctrine of Chances*”, published posthumously in 1763 by Pierce

Who is Thomas Bayes?

Reverend Thomas Bayes (ca. 1702–1761)

Presbyterian minister in Tunbridge Wells (Kent) from 1731, son of Joshua Bayes, nonconformist minister. Election to the *Royal Society* based on a tract of 1736 where he defended the views and philosophy of Newton.

His sole probability paper, *“Essay Towards Solving a Problem in the Doctrine of Chances*”, published posthumously in 1763 by Pierce and containing the seeds of *Bayes’ Theorem*.

Jaynes' introduction to Bayesian principles

"Almost always the robot will have prior information. (...)

Jaynes' introduction to Bayesian principles

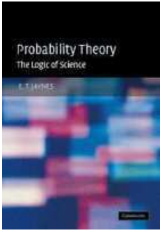
"Almost always the robot will have prior information. (...) Any additional information beyond the immediate data is by definition 'prior information (...)

Jaynes’ introduction to Bayesian principles

“Almost always the robot will have prior information. (...) Any additional information beyond the immediate data is by definition ‘prior information (...) The term ‘a-priori’ *was introduced by Immanuel Kant to denote a proposition whose truth can be known independently of experience, which is what we do* not *mean here. (...)*

Jaynes' introduction to Bayesian principles

"Almost always the robot will have prior information. (...) Any additional information beyond the immediate data is by definition 'prior information (...) The term 'a-priori' *was introduced by Immanuel Kant to denote a proposition whose truth can be known independently of experience, which is what we do* not *mean here. (...) There is no universal rule for assigning prior. At present, four fairly general principles are known—group invariance, maximum entropy, marginalization, and coding theory." (p. 87-88)*



Posterior probability

"The derivation [of Bayes' theorem] does not require any new principle beyond the product rule (...) One man's prior probability is another man's posterior probability (...) There only one kind of probability." (p.88-89)

Posterior probability

> *"The derivation [of Bayes' theorem] does not require any new principle beyond the product rule (...) One man's prior probability is another man's posterior probability (...) There only one kind of probability." (p.88-89)*

The posterior probability of the hypothesis H is given by

$$P(H|DX) = P(H|X)\,\frac{P(D|HX)}{P(D|X)}$$

Posterior probability

> *"The derivation [of Bayes' theorem] does not require any new principle beyond the product rule (...) One man's prior probability is another man's posterior probability (...) There only one kind of probability." (p.88-89)*

The posterior probability of the hypothesis H is given by

$$P(H|DX) = P(H|X)\,\frac{P(D|HX)}{P(D|X)}$$

ⓒ If $P(H|DX)$ is close to one or zero, conclusion about the "truth" of H but if close to $1/2$, need of more evidence.

Likelihood [principle]

"$P(D|HX)$ in its dependence on D for fixed H is called the 'sampling distribution' (...) and in its dependence on H for fixed D is called the 'likelihood' (...) A likelihood $L(H)$ is not itself a probability for H; it is a dimensionless function which, when multiplied by a prior probability and a normalization factor may become a probability. Because of this, constant factors are irrelevant." (p.89)

Odds ratio

Probability ratio

$$\begin{aligned}\frac{P(H|DX)}{P(\bar{H}|DX)} &= \frac{P(H|X)P(D|XH)}{P(\bar{H}|X)P(D|X\bar{H})}\\ &= O(H|DX)\\ &= O(H|X)\,\frac{P(D|HX)}{P(D|\bar{H}X)}\end{aligned}$$

defined as the ‘odds’ or posterior odds.

Evidence

Definition of base 10 logarithm transform as *evidence*:

$$e(H|DX) = 10\log_{10} O(H|DX)$$

measured in *decibels* and additive in the data: if $D = D_1 D_2 \ldots$

$$e(H|DX) = e(H|X) + 10\log_{10}\frac{P(D_1|HX)}{P(D_1|\bar{H}X} + 10\log_{10}\frac{P(D_2|D_1 HX)}{P(D_2|D_1\bar{H}X} + \cdots$$

Jaynes' widgets

> *"11 machines turning out widgets (...) ten of which produce on in six defective [and one machine] one in three defective."(p.93)*

Goal is to find whether a machine is bad (A) or good ($\bar{A}$) from n inspections with n_b defective and $n_g = n - n_b$ non-defective.

Jaynes' widgets

> *"11 machines turning out widgets (...) ten of which produce on in six defective [and one machine] one in three defective."(p.93)*

Goal is to find whether a machine is bad (A) or good ($\bar{A}$) from n inspections with n_b defective and $n_g = n - n_b$ non-defective.

$$e(A|DX) = e(A|X) + 3n_b - n_g = e(A|X) + n(4f_b - 1)$$

with $e(A|X) = -10\text{db}$

Lost w/o a loss

> *"There is nothing in probability theory* per se *which can tell us where to put critical levels [on the evidence] at which we make our decision. This has to be based on (...) decision theory." (p.96)*

Example:

> *"If the evidence* $e(A|HX)$ *is greater than* $+0$ *db, then reject [and] if it goes as low as -13 db, then accept (...) Otherwise continue testing." (p.96)*

Beyond binary

In case of $n > 2$ hypotheses $H_1, \ldots, H_n$, the additivity in the evidence

$$O(H_i|D1, \ldots, d_m X) = O(H_i|X) \frac{P(D_1, \ldots, D_m|H_i X)}{P(D_1, \ldots, D_m|\bar{H}_i X)}$$

Beyond binary

In case of $n > 2$ hypotheses $H_1, \ldots, H_n$, the additivity in the evidence

$$O(H_i|D1, \ldots, d_m X) = O(H_i|X)\frac{P(D_1, \ldots, D_m|H_i X)}{P(D_1, \ldots, D_m|\bar{H}_i X)}$$

only operates for iid data when

$$P(D_1, \ldots, D_m|\bar{H}_i X) = \prod_{j=1}^{m} P(D_j|\bar{H}_i X)$$

Beyond binary

In case of $n > 2$ hypotheses $H_1, \ldots, H_n$, the additivity in the evidence

$$O(H_i|D1, \ldots, d_m X) = O(H_i|X) \frac{P(D_1, \ldots, D_m|H_i X)}{P(D_1, \ldots, D_m|\bar{H}_i X)}$$

only operates for iid data when

$$P(D_1, \ldots, D_m|\bar{H}_i X) = \prod_{j=1}^{m} P(D_j|\bar{H}_i X)$$

which only occurs when *"at most one of the data sets D_j can produce any updating of the probability for H_i"* (p.97)

Local conclusion

“Probability theory does lead us to a useful procedure for multiple hypothesis testing, which (...) makes it clear why the independent additivity cannot, and should not, *hold when* $n > 2$.” (p.98)

Question # 1

Question

What sense does it make to relate $P(D|H_iX)$ with $P(D|\bar{H}_iX)$ when there are *several hypotheses* under comparison?

Question # 1

Question

What sense does it make to relate $P(D|H_iX)$ with $P(D|\bar{H}_iX)$ when there are *several hypotheses* under comparison?

> *"It is always possible to pick two hypotheses and to compare them only against each other (...) Here we are going after the solution of the larger problem directly."* (p.98)

Question # 1

Question

What sense does it make to relate $P(D|H_iX)$ with $P(D|\bar{H}_iX)$ when there are *several hypotheses* under comparison?

© The evidence is coherent with a posterior probability accounting for all hypotheses

More of the widgets

Suppose the data is made of $m = 50$ defective widgets in a row. Beside the good (machine) B, and the bad (machine) A, Jaynes introduces the ugly (machine) C where the proportion of defectives is now 99%. His prior evidences are

$$\text{-10 db for } A \tag{1}$$

$$+10 \text{ db for } B \tag{2}$$

$$\text{-60 db for } C \tag{3}$$

More of the widgets

Suppose the data is made of $m = 50$ defective widgets in a row. Beside the good (machine) B, and the bad (machine) A, Jaynes introduces the ugly (machine) C where the proportion of defectives is now 99%. His prior evidences are

$$\text{-10 db for } A \tag{1}$$
$$+10 \text{ db for } B \tag{2}$$
$$\text{-60 db for } C \tag{3}$$

Then

$$P(D|CX) = \left(\frac{99}{100}\right)^m$$

and

$$P(D|\bar{C}X) = \frac{P(D|AX)P(A|X) + P(D|BX)P(B|X)}{P(A|X) + P(B|X)} = \dots$$

More of the widgets

Suppose the data is made of $m = 50$ defective widgets in a row. Beside the good (machine) B, and the bad (machine) A, Jaynes introduces the ugly (machine) C where the proportion of defectives is now 99%. His prior evidences are

$$\text{-10 db for } A \tag{1}$$

$$+10 \text{ db for } B \tag{2}$$

$$\text{-60 db for } C \tag{3}$$

Then

$$e(C|DX) = -60 + 10\log_{10} \frac{\left(\frac{99}{100}\right)^m}{\frac{1}{11}\left(\frac{1}{3}\right)^m + \frac{10}{11}\left(\frac{1}{6}\right)^m}$$

Towards estimation

"Introduction of a continuous range of hypotheses such that

$$H_f \equiv \text{ the machine is putting out a fraction } f \text{ bad}$$

(...) calculate the posterior probabilities for various values of f (...) The extension is so easy." (p.107)

Towards estimation

> *"Introduction of a continuous range of hypotheses such that*
>
> $$H_f \equiv \text{ the machine is putting out a fraction } f \text{ bad}$$
>
> *(...) calculate the posterior probabilities for various values of f (...) The extension is so easy."* (p.107)

Continuous random variables processed as mathematical *"approximation[s] to the discrete set theory"* (p.109)

Rejection of measure theory (Appendix B) leads to loose handling of densities, defined as derivatives of cdf's when the limit is "well-defined" but incorporating delta functions as well (p.108, p.111)

On the nature of priors

When considering a prior on a parameter f, Jaynes stresses that *"what is distributed is not the parameter but the probability (...) f is simply an unknown constant parameter."* (p.108)

On the nature of priors

When considering a prior on a parameter f, Jaynes stresses that *"what is distributed is not the parameter but the probability (...) f is simply an unknown constant parameter."* (p.108)

Question #2

Sounds contradictory (from a mathematical viewpoint): what is the meaning of $P(F < f|X)$ then?

Bayes' theorem redux

"Those who fail to notice this fall into the famous Borel-Kolmogorov paradox, in what a seemingly well-posed problem appears to have many different correct solutions." (p.110)

Bayes' theorem redux

> *"Those who fail to notice this fall into the famous Borel-Kolmogorov paradox, in what a seemingly well-posed problem appears to have many different correct solutions."* (p.110)

Awkward derivation of Bayes' theorem in the continuous case based on the approximations

$$P(F \in (f, f + \mathrm{d}f|X) = g(f|X)\,\mathrm{d}f$$

and

$$P(F \in (f, f + \mathrm{d}f|DX) = g(f|DX)\,\mathrm{d}f$$

Bayes' theorem redux

> *"Those who fail to notice this fall into the famous Borel-Kolmogorov paradox, in what a seemingly well-posed problem appears to have many different correct solutions."* (p.110)

Awkward derivation of Bayes' theorem in the continuous case based on the approximations

$$P(F \in (f, f + \mathrm{d}f|X) = g(f|X)\,\mathrm{d}f$$

and

$$P(F \in (f, f + \mathrm{d}f|DX) = g(f|DX)\,\mathrm{d}f$$

substituting continuity to measurability.

Bayes' widgets

Using a binomial distribution on the proportion of bad widgets (no longer hypotheses A, B and C?),

$$P(D|H_f X) = f^n(1-f)^{N-n} \qquad [\propto]$$

Jayes ends up with Bayes' historical posterior:

$$g(f|DX) = \frac{f^n(1-f)^{N-n} g(f|X)}{\int_0^1 f^n(1-f)^{N-n} g(f|X)\,\mathrm{d}f}$$

Bayes' widgets

Using a binomial distribution on the proportion of bad widgets (no longer hypotheses A, B and C?),

$$P(D|H_f X) = f^n (1-f)^{N-n} \qquad [\propto]$$

Jayes ends up with Bayes' historical posterior:

$$g(f|DX) = \frac{f^n(1-f)^{N-n} g(f|X)}{\int_0^1 f^n (1-f)^{N-n} g(f|X)\,\mathrm{d}f}$$

Jaynes also considers that hypotheses A, B and C can be incorporated via "delta-functions"

Bayes' 1763 paper:

Billiard ball W rolled on a line of length one, with a uniform probability of stopping anywhere: W stops at p.
Second ball O then rolled n times under the same assumptions. X denotes the number of times the ball O stopped on the left of W.

Bayes' 1763 paper:

Billiard ball W rolled on a line of length one, with a uniform probability of stopping anywhere: W stops at p.
Second ball O then rolled n times under the same assumptions. X denotes the number of times the ball O stopped on the left of W.

Bayes' question:

Given X, what inference can we make on p?

Bayes' problem

> *"This result was first found by an amateur* (sic!) *mathematician (...) not Bayes but Laplace who first saw the result in generality and showed how to use it in inference."* (p.112)

Modern translation:

Derive the posterior distribution of p given X, when

$$p \sim \mathcal{U}([0,1]) \text{ and } X|p \sim \mathcal{B}(n,p)$$

Resolution

Since

$$\begin{aligned} P(X = x|p) &= \binom{n}{x} p^x (1-p)^{n-x}, \\ P(a < p < b \text{ and } X = x) &= \int_a^b \binom{n}{x} p^x (1-p)^{n-x} dp \end{aligned}$$

and

$$P(X = x) = \int_0^1 \binom{n}{x} p^x (1-p)^{n-x}\, dp,$$

Resolution (2)

then

$$\begin{aligned} P(a < p < b|X = x) &= \frac{\int_a^b \binom{n}{x} p^x (1-p)^{n-x}\, dp}{\int_0^1 \binom{n}{x} p^x (1-p)^{n-x}\, dp} \\ &= \frac{\int_a^b p^x (1-p)^{n-x}\, dp}{B(x+1, n-x+1)} \, , \end{aligned}$$

Resolution (2)

then

$$\begin{aligned} P(a < p < b | X = x) &= \frac{\int_a^b \binom{n}{x} p^x (1-p)^{n-x}\, dp}{\int_0^1 \binom{n}{x} p^x (1-p)^{n-x}\, dp} \\ &= \frac{\int_a^b p^x (1-p)^{n-x}\, dp}{B(x+1, n-x+1)}\,, \end{aligned}$$

i.e.

$$p|x \sim \mathcal{B}e(x+1, n-x+1)$$

[Beta distribution]

Queer uses (Chapter 5)

> *"There is not the slightest use in rejecting any hypothesis unless we can do it in favor of some definite alternative that better fits the facts."* (p.135)

Disgressions on

- Mrs Stewart's telepathic powers (ESP)
- divergence of opinions
- discovery of Neptune
- horse racing and weather forecasting
- Bayesian jurisprudence

Chapter 6: Elementary parameter estimation

1. Goals of the course

2. Elementary hypothesis testing

3. Elementary parameter estimation
   - Urn parameters
   - Prior selection
   - Bernoulli trials
   - First foray in decision theory
   - Hierarchical models
   - Ride in a taxicab
   - The normal distribution

4. Sufficiency, ancillarity and all that

A simple hypergeometric model

> *"When the hypotheses become very numerous, as in H_t $(1 \leq t \leq n)$, deciding between the hypotheses and estimation the index t are practically the same thing."* (p.149)

Observation of an hypergeometric distribution

$$p(D|NRI) = \binom{N}{n}^{-1} \binom{R}{r} \binom{N-R}{n-r}$$

when both N and R are unknown.

- debate on whether or not the data is informative about N
- early sections are conditional on N

$$p(R|DNI) = p(R|DI) \frac{D|NRI)}{D|NI)}$$

Information on N

> *"Intuition may lead us to expect that the data can only truncate the impossible values [of N]."* (p.151)

We have

$$p(N|DI) \propto \mathbb{I}(N \geq n)\, p(N|I)$$

under the condition

$$\sum_{R=0}^{N} \binom{R}{r} \binom{N-R}{n-r} p(R|NI) = f(n,r) \binom{N}{n}$$

on the prior $p(R|NI)$.

Uniform prior

For instance, if $p(R|NI)$ is uniform,

$$p(R|DNI) \propto \binom{R}{r}\binom{N-R}{n-r}$$

and

$$\sum_{R=0}^{N}\binom{R}{r}\binom{N-R}{n-r}\frac{1}{N+1} = \frac{1}{n+1}\binom{N}{n}$$

so *"the data tells us nothing about* N*, beyond the fact that* $N \geq n$*."* (p.153)

Properties of the conditional posterior

The most probable value of R under $p(R|DNI)$ is

$$R' = (N+1)\frac{r}{n}$$

up to an integer truncation.

Properties of the conditional posterior

The most probable value of R under $p(R|DNI)$ is

$$R' = (N+1)\frac{r}{n}$$

up to an integer truncation.
To find the mean value, Jaynes notices that

$$(R+1)\binom{R}{r} = (r+1)\binom{R+1}{r+1}$$

Hence that

$$\mathbb{E}[R|DNI] + 1 = (r+1)\binom{N+1}{n+1}^{-1}\sum_{R=0}^{N}\binom{R+1}{r+1}\binom{N-R}{n-r}$$

$$= (r+1)\binom{N+1}{n+1}^{-1}\binom{N+2}{n+2} = \frac{(N+2)(r+1)}{(n+2)}$$

Laplace's rule of succession

Predictive distribution for the next draw

$$\begin{aligned} p(R_{n+1}|DNI) &= \sum_{R=0}^{N} p(R_{n+1}|RDNI)\, p(R|DNI) \\ &= \sum_{R=0}^{N} \frac{R-r}{N-n} \binom{N+1}{n+1}^{-1} \sum_{R=0}^{N} \binom{R}{r} \binom{N-R}{n-r} \\ &= \frac{r+1}{n+2} \qquad \text{free of } N \end{aligned}$$

Laplace's rule of succession

Predictive distribution for the next draw

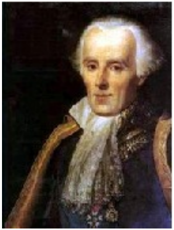
$$
\begin{aligned}
p(R_{n+1}|DNI) &= \sum_{R=0}^{N} p(R_{n+1}|RDNI)\, p(R|DNI) \\
&= \sum_{R=0}^{N} \frac{R-r}{N-n} \binom{N+1}{n+1}^{-1} \sum_{R=0}^{N} \binom{R}{r} \binom{N-R}{n-r} \\
&= \frac{r+1}{n+2} \qquad \text{free of } N
\end{aligned}
$$

© This is Laplace's succession rule

Who's Laplace?

Pierre Simon de Laplace (1749–1827)

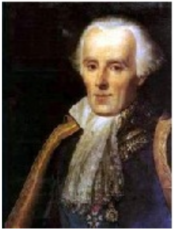
French mathematician and astronomer born in Beaumont en Auge (Normandie) who formalised mathematical astronomy in *Mécanique Céleste*. Survived the French revolution, the Napoleon Empire (as a comte!), and the Bourbon restauration (as a marquis!!).



Who's Laplace?

Pierre Simon de Laplace (1749–1827)

French mathematician and astronomer born in Beaumont en Auge (Normandie) who formalised mathematical astronomy in *Mécanique Céleste*. Survived the French revolution, the Napoleon Empire (as a comte!), and the Bourbon restauration (as a marquis!!).



In *Essai Philosophique sur les Probabilités*, Laplace set out a mathematical system of inductive reasoning based on probability, precursor to Bayesian Statistics.

Jeffreys' criticism on uniform prior

The fundamental trouble is that the prior probabilities $1/(N+1)$ *attached by the theory to the extreme values are utterly so small that they amount to saying, without any evidence at all, that it is practically certain that the population is not homogeneous in respect to the property to be investigated. (...) Now I say that for this reason the uniform assessment must be abandoned for ranges including the extreme values. (Theory of Probability, III, §3.21)*

Jeffreys' criticism on uniform prior

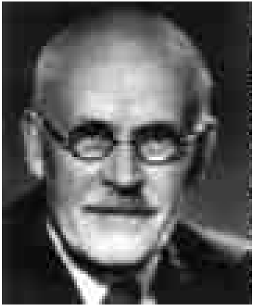
> *The fundamental trouble is that the prior probabilities $1/(N+1)$ attached by the theory to the extreme values are utterly so small that they amount to saying, without any evidence at all, that it is practically certain that the population is not homogeneous in respect to the property to be investigated. (...) Now I say that for this reason the uniform assessment must be abandoned for ranges including the extreme values. (Theory of Probability, III, §3.21)*

Explanation: This is a preparatory step for the introduction of specific priors fitted to point null hypotheses (using Dirac masses).

Who is Harold Jeffreys?

Wikipedia article

Sir Harold Jeffreys (1891–1989)
Mathematician, statistician, geophysicist, and astronomer. He went to St John's College, Cambridge and became a fellow in 1914, where he taught mathematics then geophysics and astronomy. He was knighted in 1953 and received the Gold Medal of the Royal Astronomical Society in 1937.



Jaynes' modification

New prior eliminating boundaries:

$$p(R|NI_1) = \frac{1}{N-1}\,\mathbb{I}(0 < R < N)$$

but no change in $p(R|NI_1)$ if $0 < r < n$.

The noninformative convex posterior

> “ *Find whether there is any prior that would lead to a uniform posterior distribution.*” (p.159)

Jaynes modifies the prior into

$$p(R|NI_{00}) = \frac{A}{R(N-R)}\,\mathbb{I}(0 < R < N)$$

which may be constant for some values of (r, n).

The noninformative convex posterior

> *" Find whether there is any prior that would lead to a uniform posterior distribution."* (p.159)

Jaynes modifies the prior into

$$p(R|NI_{00}) = \frac{A}{R(N-R)}\,\mathbb{I}(0 < R < N)$$

which may be constant for some values of (r, n).
Not particularly exciting. but preparation for Haldane's prior

> *"As $N \longrightarrow \infty$, the concave prior approaches an improper (non-normalizable) one, which must give absurd answers to some questions."* (p.160)

Towards hierarchical priors

Introduction of an extra (monkey) parameter g governing the prior on R

$$p(R|NI_2) = \binom{N}{R} g^R (1-g)^{N-R}$$

Towards hierarchical priors

Introduction of an extra (monkey) parameter g governing the prior on R

$$p(R|NI_2) = \binom{N}{R} g^R (1-g)^{N-R}$$

Then

$$\begin{aligned} p(R|DNI_2) &\propto \binom{N}{R} g^R (1-g)^{N-R} \binom{R}{r} \binom{N-R}{n-r} \\ &= \binom{N-n}{R-r} g^{R-r} (1-g)^{N-R-n+r} \end{aligned}$$

for $r \leq R \leq N-n+r$

Towards hierarchical priors

Introduction of an extra (monkey) parameter g governing the prior on R

$$p(R|NI_2) = \binom{N}{R} g^R (1-g)^{N-R}$$

Then

$$\begin{aligned} p(R|DNI_2) &\propto \binom{N}{R} g^R (1-g)^{N-R} \binom{R}{r} \binom{N-R}{n-r} \\ &= \binom{N-n}{R-r} g^{R-r} (1-g)^{N-R-n+r} \end{aligned}$$

for $r \leq R \leq N-n+r$

$$\copyright\ R - r \sim \mathcal{B}(N-n, g)$$

Towards hierarchical priors

Introduction of an extra (monkey) parameter g governing the prior on R

$$p(R|NI_2) = \binom{N}{R} g^R (1-g)^{N-R}$$

Then

$$\begin{aligned} p(R|DNI_2) &\propto \binom{N}{R} g^R (1-g)^{N-R} \binom{R}{r} \binom{N-R}{n-r} \\ &= \binom{N-n}{R-r} g^{R-r} (1-g)^{N-R-n+r} \end{aligned}$$

for $r \leq R \leq N-n+r$

Ⓒ The data brings no information about $R-r$

Not such a rare occurence

Example (Cauchy vs. Laplace)

Consider

$$f(x|\theta) = \frac{1}{\pi}\left[1 + (x-\theta)^2\right]^{-1}$$

and

$$\pi(\theta) = \frac{1}{2} e^{-|\theta|}$$

The MAP estimator of θ is then always

$$\delta^*(x) = 0$$

[The Bayesian Choice, Chap. 4]

First occurence of a continuous parameter

> *“We do not see parameter estimation and hypothesis testing as fundamentally different activites.”* (p.163)
> *“The condition of the experiment will tell us whether the order is meaningful or known; and we expect probability theory to tell us whether it is relevant.”* (p.164)

Preparation of the chapter on *sufficiency, ancilarity and all that*.

First occurence of a continuous parameter

> *"We do not see parameter estimation and hypothesis testing as fundamentally different activites."* (p.163)
> *"The condition of the experiment will tell us whether the order is meaningful or known; and we expect probability theory to tell us whether it is relevant."* (p.164)

Preparation of the chapter on *sufficiency, ancilarity and all that*.

Back to Bayes-ics:

$$\theta \sim \mathcal{U}(0,1) \qquad r|\theta \sim \mathcal{B}(n,\theta)$$

$$\theta|r,n \sim \mathcal{B}e(r+1, n-r+1)$$

Haldane's prior

"By passage to the limit $N \longrightarrow \infty$, (...) the concave pre-prior distribution would go into an improper prior for θ.

$$p(\theta|I) \propto \frac{1}{\theta(1-\theta)}$$

for which some sums or integrals would diverge; but that is not the strictly correct method of calculation (...) Under very general conditions this limit is well-behaved, leading to useful results. The limiting improper pre-prior was advocated by Haldane (1932) in the innocent days before the marginalization paradox." (p.165-166)

Jeffreys' Haldane's prior

For a binomial observation, $x \sim \mathcal{B}(n, p)$, and prior $\pi^*(p) \propto [p(1-p)]^{-1}$, the marginal distribution,

$$\begin{aligned} m(x) &= \int_0^1 [p(1-p)]^{-1} \binom{n}{x} p^x (1-p)^{n-x} dp \\ &= B(x, n-x), \end{aligned}$$

is only defined for $x \neq 0, n$.

Jeffreys' Haldane's prior

For a binomial observation, $x \sim \mathcal{B}(n,p)$, and prior $\pi^*(p) \propto [p(1-p)]^{-1}$, the marginal distribution,

$$\begin{aligned} m(x) &= \int_0^1 [p(1-p)]^{-1} \binom{n}{x} p^x (1-p)^{n-x} dp \\ &= B(x, n-x), \end{aligned}$$

is only defined for $x \neq 0, n$.
Missed by Jeffreys:

> *If a sample is of one type with respect to some property there is probability 1 that the population is of that type (Theory of Probability, III, §3.1)*

Noninformative setting

What if all we know is that we know “nothing” ?!

...how can we assign the prior probability when we know nothing about the value of the parameter except the very vague knowledge just indicated? (Theory of Probability, III, §3.1)

Noninformative distributions

> *...provide a formal way of expressing ignorance of the value of the parameter over the range permitted (Theory of Probability, III, §3.1).*

In the absence of prior information, prior distributions solely derived from the sample distribution $f(x|\theta)$

> *It says nothing about the value of the parameter, except the bare fact that it may possibly by its very nature be restricted to lie within certain definite limits (Theory of Probability, III, §3.1)*

▸ Jaynes' ignorance priors

Warning

> Noninformative priors cannot be expected to represent exactly total ignorance about the problem at hand, but should rather be taken as reference or default priors, upon which everyone could fall back when the prior information is missing.

[Kass and Wasserman, 1996]

The stopping rule principle

Another instance of the likelihood principle

> *"It is from the data D that we learn both n and r (...)*
>
> $$p(\theta|nDI) = p(\theta|DI)$$
>
> *(...) some statisticians claim that the stopping rule* does *affect our inference. Apparently, they believe that if a statistic such as r is not known in advance, then parts of the sample space referring to false values of r remain relevant to our inferences, even after the true value of r becomes known from the data."* (p.167)

The survey example

A survey about a simple question gets 6 yes and 9 no.

The survey example

A survey about a simple question gets 6 yes and 9 no.
In a first model, 15 individuals have been selected, out of which 6 replied yes

$$r \sim \mathcal{B}in(15, \theta)$$

The survey example

A survey about a simple question gets 6 yes and 9 no.
In a second model, individuals have been selected until 6 replied yes

$$N \sim \mathcal{N}eg(15, \theta)$$

ⓒ Same likelihood but opposed stopping rules

"Inference must depend on the data that was observed, not on data sets that might have been observed but were not." (p.167)

Compound problems

Study of a multilevel model

$$
\begin{aligned}
n &\sim \mathcal{B}(N, r) \\
c &\sim \mathcal{B}(n, \varphi)
\end{aligned}
$$

Compound problems

Study of a multilevel model

$$n \sim \mathcal{B}(N, r)$$
$$c \sim \mathcal{B}(n, \varphi)$$

As $N \longrightarrow \infty$, $NS \longrightarrow s$, $p(n|Nr) \longrightarrow \exp\{-s\}\frac{s^n}{n!}$

Compound problems

Study of a multilevel model

$$n \sim \mathcal{B}(N, r)$$
$$c \sim \mathcal{B}(n, \varphi)$$

As $N \longrightarrow \infty$, $NS \longrightarrow s$, $p(n|Nr) \longrightarrow \exp\{-s\}\frac{s^n}{n!}$

Then

$$n|\varphi cs) - c \sim \mathcal{P}oi(s(1-\varphi))$$

Loss functions

“Which estimator is best? Laplace gave a criterion that we should make that estimate which minimizes the expected error $|\alpha - \alpha^{\star}|$*.”*

“Laplace’s criterion was generally rejected in favor of the least squares methode of Gauss and Legendre; we seek the estimate that minimizes the expected squares of the error.” (p.172)

The quadratic loss

Choice of a loss function

$$\mathrm{L}(\theta, d) = (\theta - d)^2$$

The quadratic loss

Choice of a loss function

$$\mathrm{L}(\theta, d) = (\theta - d)^2$$

or

$$\mathrm{L}(\theta, d) = ||\theta - d||^2$$

to minimise (a posteriori).

Proper loss

Posterior mean

The Bayes estimator δ^π associated with the prior π and with the quadratic loss is the posterior expectation

$$\delta^\pi(x) = \mathbb{E}^\pi[\theta|x] = \frac{\int_\Theta \theta f(x|\theta)\pi(\theta)\,d\theta}{\int_\Theta f(x|\theta)\pi(\theta)\,d\theta}.$$

Proper loss

Posterior mean

The Bayes estimator δ^π associated with the prior π and with the quadratic loss is the posterior expectation

$$\delta^\pi(x) = \mathbb{E}^\pi[\theta|x] = \frac{\int_\Theta \theta f(x|\theta)\pi(\theta)\,d\theta}{\int_\Theta f(x|\theta)\pi(\theta)\,d\theta}.$$

> *"But thus may not be what we really want (...) the mean value estimate concentrates its attention most strongly in avoiding the very large (but also very improbable errors), at the cost of possibly not doing as well with the far more likely small errors."* (p.173)

Jaynes' criticisms

- Sensitivity to outliers: *"a single very rich man in a poor village..."*
- Influence of fat tails: *"quite sensitive to what happens far away from out in the tails"*
- Lack of consistency under parameter changes: *"the posterior mean estimate of $\lambda(\alpha)$ would not in general satisfy $\lambda^\star = \lambda(\alpha^\star)$"*

The absolute error loss

Alternatives to the quadratic loss:

$$\mathrm{L}(\theta, d) =\mid \theta - d \mid,$$

or

$$\mathrm{L}_{k_1,k_2}(\theta, d) = \begin{cases} k_2(\theta - d) & \text{if } \theta > d, \\ k_1(d - \theta) & \text{otherwise.} \end{cases} \tag{4}$$

The absolute error loss

Alternatives to the quadratic loss:

$$\mathrm{L}(\theta, d) = \mid \theta - d \mid,$$

or

$$\mathrm{L}_{k_1,k_2}(\theta, d) = \begin{cases} k_2(\theta - d) & \text{if } \theta > d, \\ k_1(d - \theta) & \text{otherwise.} \end{cases} \tag{4}$$

L_1 estimator

The Bayes estimator associated with π and (4) is a $(k_2/(k_1 + k_2))$ fractile of $\pi(\theta|x)$.

Jaynes' defence

Robustness of posterior quantiles:

- Insensitivity to outliers: *"One single very rich man in a poor village has no effect..."*
- Lesser influence of fat tails: *"an error twice as large only twice as serious"*
- Consistency under monotonic [and 1D] parameter changes
- Harder to compute but *"today the computational problem is relatively trivial"*

Invariant divergences

Interesting point made by Jeffreys that both

$$\mathrm{L}_m = \int |(dP)^{1/m} - (dP')^{1/m}|^m , \quad \mathrm{L}^e = \int \log \frac{dP'}{dP} d(P' - P)$$

...are invariant for all non-singular transformations of x and of the parameters in the laws (Theory of Probability, III, §3.10)

Intrinsic losses

Noninformative settings w/o natural parameterisation : the estimators should be invariant under reparameterisation

[Ultimate invariance!]

Principle

Corresponding parameterisation-free loss functions:

$$\mathrm{L}(\theta,\delta)=d(f(\cdot|\theta),f(\cdot|\delta)),$$

Examples

1. the *entropy distance* (or *Kullback–Leibler divergence*)

$$\mathrm{L_e}(\theta, \delta) = \mathbb{E}_\theta \left[\log \left(\frac{f(x|\theta)}{f(x|\delta)} \right) \right],$$

Examples

1. the *entropy distance* (or *Kullback–Leibler divergence*)

$$\mathrm{L}_{\mathrm{e}}(\theta,\delta)=\mathbb{E}_\theta\left[\log\left(\frac{f(x|\theta)}{f(x|\delta)}\right)\right],$$

2. the *Hellinger distance*

$$\mathrm{L}_{\mathrm{H}}(\theta,\delta)=\frac{1}{2}\mathbb{E}_\theta\left[\left(\sqrt{\frac{f(x|\delta)}{f(x|\theta)}}-1\right)^2\right].$$

Examples (2)

Example (Normal mean)

Consider $x \sim \mathcal{N}(\theta, 1)$. Then

$$\begin{aligned} \mathrm{L}_{\mathrm{e}}(\theta,\delta) &= \frac{1}{2}\mathbb{E}_\theta[-(x-\theta)^2 + (x-\delta)^2] = \frac{1}{2}(\delta-\theta)^2, \\ \mathrm{L}_{\mathrm{H}}(\theta,\delta) &= 1 - \exp\{-(\delta-\theta)^2/8\}. \end{aligned}$$

When $\pi(\theta|x)$ is $\mathcal{N}(\mu(x), \sigma^2)$, Bayes estimator of θ

$$\delta^\pi(x) = \mu(x)$$

in both cases.

Examples (3)

Example (Normal everything)

Consider $x \sim \mathcal{N}(\lambda, \sigma^2)$ then

$$\mathrm{L}_2((\lambda,\sigma),(\lambda',\sigma')) = 2\sinh^2\zeta + \cosh\zeta\frac{(\lambda-\lambda')^2}{\sigma_0^2}$$

$$\mathrm{L}^e((\lambda,\sigma),(\lambda',\sigma')) = 2\left[1 - \mathrm{sech}^{1/2}\zeta \exp\left\{\frac{-(\lambda-\lambda')^2}{8\sigma_0^2\cosh\zeta}\right\}\right]$$

if $\sigma = \sigma_0 e^{-\zeta/2}$ *and* $\sigma = \sigma_0 e^{+\zeta/2}$ *(III, §3.10, (14) & (15))*

Hierarchy as information

Distinction between iid compound problems ("Mr A")

$$n_i \sim p(n_i|I_A) = \mathbb{I}_{0\leq n_i<N}/N \qquad c_i \sim \mathcal{B}in(n_i, \varphi)$$

and common parameter driving all n_i's ("Mr B")

$$s \sim p(s|I_B) = \mathbb{I}_{0\leq s\leq S_0} \quad n_i \sim \mathcal{P}oi(s) \qquad c_i \sim \mathcal{B}in(n_i, \varphi)$$

Hierarchy as information

Distinction between iid compound problems ("Mr A")

$$p(n_i|\varphi c_i I_A) = \binom{n_i}{c_i} \varphi^{c_i+1}(1-\varphi)^{n_i-c_i}$$

and common parameter driving all n_i's ("Mr B")

$$p(n_i|c_i I_B) = \binom{n_i}{c_i} \varphi^{c_i+1}(1-\varphi)^{n_i-c_i}$$

Making sense of limits?

Jaynes' reasoning starts from

$$p(n_i|I_B) = \int_0^{\infty} p(n_i|s)p(s|\theta I_B)\,\mathrm{d}s = \frac{1}{S_0}\int_0^{s_0} \frac{s^{n_i}\exp\{-s\}}{n_i!} \qquad n_i < S_0$$

and considers the limit of the integral when S_0 goes to $+\infty$:

$$p(n_i|\theta I_B) = \frac{1}{S_0} \times 1$$

without bothering about the remaining S_0...

Jeffreys' prior #1

"Harold Jeffreys (...) suggests that the proper way to express "complete ignorance" of a continuous variable known to be positive is to assign uniform probability to its logarithm." (p.181)

$$p(s|I_J) \propto \frac{1}{s} \qquad (0 \leq s < \infty)$$

Jaynes' defence of Jeffreys'prior

"we cannot normalize this, but (...) we can approach this prior as the limit of a sequence of proper priors. If that does not yield a proper posterior distribution (...) the data are too uninformative about either very large s or very small s." (p.181)

Jaynes' defence of Jeffreys'prior

"we cannot normalize this, but (...) we can approach this prior as the limit of a sequence of proper priors. If that does not yield a proper posterior distribution (...) the data are too uninformative about either very large s or very small s." (p.181)

"There is a germ of an important principle here (...) our desideratum of consistency, in the sense that equivalent states of knowledge should be represented by equivalent probability assignements, uniquely determines the Jeffreys rule (...) marginalization theory reinforces this by deriving it Uniquely." (p.182)

[Why "U"?!]

The Jeffreys prior

Based on Fisher information

$$I(\theta) = \mathbb{E}_\theta \left[\frac{\partial \ell}{\partial \theta^\mathsf{T}} \frac{\partial \ell}{\partial \theta} \right]$$

The Jeffreys prior distribution is

$$\pi^*(\theta) \propto |I(\theta)|^{1/2}$$

Note

This general presentation is *not* to be found in **ToP**! And not all priors of Jeffreys' are Jeffreys priors!

Where did Jeffreys hide his prior?!

Starts with second order approximation to both L_2 and L^e:

$$4\mathrm{L}_2(\theta,\theta') \approx (\theta-\theta')^\mathsf{T} I(\theta)(\theta-\theta') \approx \mathrm{L}^e(\theta,\theta')$$

> *This expression is therefore invariant for all non-singular transformations of the parameters. It is not known whether any analogous forms can be derived from [L_m] if $m \neq 2$. (Theory of Probability, III, §3.10)*

Main point

Fisher information equivariant under reparameterisation:

$$\frac{\partial \ell}{\partial \theta^\mathsf{T}}\frac{\partial \ell}{\partial \theta} = \frac{\partial \ell}{\partial \eta^\mathsf{T}}\frac{\partial \ell}{\partial \eta} \times \frac{\partial \eta}{\partial \theta^\mathsf{T}}\frac{\partial \eta}{\partial \theta}$$

The fundamental prior

...if we took the prior probability density for the parameters to be proportional to $||g_{ik}||^{1/2}$ [$= |I(\theta)|^{1/2}$], it could stated for any law that is differentiable with respect to all parameters that the total probability in any region of the α_i would be equal to the total probability in the corresponding region of the α'_i; in other words, it satisfies the rule that equivalent propositions have the same probability (Theory of Probability, III, §3.10)

The fundamental prior

> *...if we took the prior probability density for the parameters to be proportional to $||g_{ik}||^{1/2}$ [$= |I(\theta)|^{1/2}$], it could stated for any law that is differentiable with respect to all parameters that the total probability in any region of the α_i would be equal to the total probability in the corresponding region of the α'_i; in other words, it satisfies the rule that equivalent propositions have the same probability (Theory of Probability, III, §3.10)*

Jeffreys never mentions Fisher information in connection with (g_{ik})

Back to the hierarchy

Jaynes uses Mr B to shows how he learns about n_1 from $c_2, \ldots$ though s:

$$p(n_1|\varphi c_1 c_2 I_B) \propto p(n_1|\varphi c_1 I_B) p(c_2|\varphi n_1 I_B)$$

where

$$\begin{aligned} p(c_2|\varphi n_1 I_B) &= \int_0^\infty p(c_2|\varphi s I_B) p(s|\varphi n_1 I_B)\,\mathrm{d}s \\ &= \int_0^\infty \frac{\exp\{-s\varphi\}\,(s\varphi)^{c_2}}{c_2!} \frac{\exp\{-s\}\,s^{n_1}}{n_1!}\,\mathrm{d}s \\ &= \binom{n_1 + c_2}{c_2} \frac{\varphi^{c_2}}{(1+\varphi)^{n_1+c_2+1}} \end{aligned}$$

Back to the hierarchy

Jaynes uses Mr B to shows how he learns about n_1 from $c_2, \ldots$ though s:

$$p(n_1|\varphi c_1 c_2 I_B) \propto p(n_1|\varphi c_1 I_B) p(c_2|\varphi n_1 I_B)$$

where

$$\begin{aligned} p(c_2|\varphi n_1 I_B) &= \int_0^\infty p(c_2|\varphi s I_B) p(s|\varphi n_1 I_B)\,\mathrm{d}s \\ &= \int_0^\infty \frac{\exp\{-s\varphi\}\,(s\varphi)^{c_2}}{c_2!} \frac{\exp\{-s\}\,s^{n_1}}{n_1!}\,\mathrm{d}s \\ &= \binom{n_1 + c_2}{c_2} \frac{\varphi^{c_2}}{(1+\varphi)^{n_1+c_2+1}} \end{aligned}$$

Hence

$$p(n_1|\varphi c_1 c_2 I_B) = \binom{n_1 + c_2}{c_1 + c_2} \left(\frac{2\varphi}{1+\varphi}\right)^{c_1+c_2+1} \left(\frac{1-\varphi}{1+\varphi}\right)^{n_1-c_1}$$

Impact of prior information

"Mr B's extra information has enabled him to make an appreciable decrease in his probable error (...) Therefore any method of inference which fails to to take prior information into account is capable of misleading us, in a potentially dangerous way." (p.187)

Asymptotic form

As the number of datapoints $c_1, \ldots, c_m$ increases,

$$p(s|\varphi c_1 \cdots c_m I_B) = \frac{(m\varphi)^{c_+ +1}}{c_+!} \exp\{-ms\varphi\}$$

becomes approximately normal

$$p(s|\varphi c_1 \cdots c_m I_B) \approx \exp -\frac{c(s-\hat{s})^2}{2s^2}$$

Asymptotic form

As the number of datapoints $c_1, \ldots, c_m$ increases,

$$p(s|\varphi c_1 \cdots c_m I_B) = \frac{(m\varphi)^{c_+ +1}}{c_+!} \exp\{-ms\varphi\}$$

becomes approximately normal

$$p(s|\varphi c_1 \cdots c_m I_B) \approx \exp - \frac{c(s-\hat{s})^2}{2s^2}$$

Hence in the limit

$$p(n_1|\varphi c_1 \cdots c_m I_B) \longrightarrow \frac{\exp\{-s_0(1-\varphi)\}}{(n_1-c_1)!} \left[s_0(1-\varphi)\right]^{n_1-c_1}$$

(i.e. complete knowledge of the true s)

Asymptotic form

As the number of datapoints $c_1, \ldots, c_m$ increases,

$$p(s|\varphi c_1 \cdots c_m I_B) = \frac{(m\varphi)^{c_+ + 1}}{c_+!} \exp\{-ms\varphi\}$$

becomes approximately normal

$$p(s|\varphi c_1 \cdots c_m I_B) \approx \exp -\frac{c(s-\hat{s})^2}{2s^2}$$

Hence in the limit

$$p(n_1|\varphi c_1 \cdots c_m I_B) \longrightarrow \frac{\exp\{-s_0(1-\varphi)\}}{(n_1-c_1)!} \left[s_0(1-\varphi)\right]^{n_1-c_1}$$

(i.e. complete knowledge of the true s)

The tramcar comparison

> *A man travelling in a foreign country has to change trains at a junction, and goes into the town, of the existence of which he has just heard. The first thing that he sees is a tramcar numbered* $m = 100$*. What can he infer about the number [N] of tramcars in the town? (Theory of Probability, IV, §4.8)*

The tramcar comparison

> *A man travelling in a foreign country has to change trains at a junction, and goes into the town, of the existence of which he has just heard. The first thing that he sees is a tramcar numbered $m = 100$. What can he infer about the number [N] of tramcars in the town? (Theory of Probability, IV, §4.8)*

Famous opposition: Bayes posterior expectation vs. MLE

- Exclusion of flat prior on N
- Choice of the scale prior $\pi(N) \propto 1/N$
- MLE is $\hat{N} = m$

The tramcar (2)

Under $\pi(N) \propto 1/N + \mathrm{O}(n^{-2})$, posterior is

$$\pi(N|m) \propto 1/N^2 + \mathrm{O}(n^{-3})$$

and

$$P(N > n_0|m, H) = \sum_{n_0+1}^{\infty} n^{-2} \Big/ \sum_{m}^{\infty} n^{-2} = \frac{m}{n_0}$$

Therefore posterior median is $2m$

The tramcar (2)

Under $\pi(N) \propto 1/N + \mathrm{O}(n^{-2})$, posterior is

$$\pi(N|m) \propto 1/N^2 + \mathrm{O}(n^{-3})$$

and

$$P(N > n_0|m, H) = \sum_{n_0+1}^{\infty} n^{-2} \Big/ \sum_{m}^{\infty} n^{-2} = \frac{m}{n_0}$$

Therefore posterior median is $2m$

© No mention made of either MLE or unbiasedness

Jaynes' version

"Here we study a continuous version of the [taxicab] *problem, in which more than a taxi may be in view, and then state the exact relationship between the continuous and discrete problems."* (p.191)

Jaynes' version

> *"Here we study a continuous version of the* [taxicab] *problem, in which more than a taxi may be in view, and then state the exact relationship between the continuous and discrete problems."* (p.191)

Case of a uniform (*"rectangular sampling"*) variate

$$p(x_i|\alpha I) = \alpha^{-1}\,\mathbb{I}(0 \leq x_i \leq \alpha)$$

Jaynes' version

> *"Here we study a continuous version of the* [taxicab] *problem, in which more than a taxi may be in view, and then state the exact relationship between the continuous and discrete problems."* (p.191)

Case of a uniform (*"rectangular sampling"*) variate

$$p(x_i|\alpha I) = \alpha^{-1}\,\mathbb{I}(0 \leq x_i \leq \alpha)$$

Under a rectangular prior

$$p(\alpha|I) = (\alpha_1 - \alpha_{00})^{-1}\mathbb{I}(\alpha_{00} \leq \alpha \leq \alpha_1)$$

> *"if any datum is found to exceed the upper prior bounds, the data and the prior information would be logically contradictory."* (p.191)

Small numbers

The corresponding posterior is $(n > 1)$

$$p(\alpha|DI) = \frac{(n-1)\alpha^{-n}}{\alpha_0^{1-n} - \alpha_1^{1-n}} \mathbb{I}(\alpha_0 \leq \alpha \leq \alpha_1)$$

where $\alpha_0 = \max\{\alpha_{00}, x_1, \ldots, x_n\}$

Small numbers

The corresponding posterior is $(n > 1)$

$$p(\alpha|DI) = \frac{(n-1)\alpha^{-n}}{\alpha_0^{1-n} - \alpha_1^{1-n}} \mathbb{I}(\alpha_0 \leq \alpha \leq \alpha_1)$$

where $\alpha_0 = \max\{\alpha_{00}, x_1, \ldots, x_n\}$
The limiting case for $n = 1$

$$p(\alpha|DI) = \frac{\alpha^{-1}}{\log(\alpha_1/\alpha_1)} \mathbb{I}(\alpha_0 \leq \alpha \leq \alpha_1)$$

can also be obtained as a limit when $n \to 1$ *"when n is any complex number"* (p.193)

Small numbers

The corresponding posterior is $(n > 1)$

$$p(\alpha|DI) = \frac{(n-1)\alpha^{-n}}{\alpha_0^{1-n} - \alpha_1^{1-n}} \mathbb{I}(\alpha_0 \leq \alpha \leq \alpha_1)$$

where $\alpha_0 = \max\{\alpha_{00}, x_1, \ldots, x_n\}$
The limiting case for $n = 1$

$$p(\alpha|DI) = \frac{\alpha^{-1}}{\log(\alpha_1/\alpha_1)} \mathbb{I}(\alpha_0 \leq \alpha \leq \alpha_1)$$

can also be obtained as a limit when $n \to 1$ *"when n is any complex number"* (p.193)

typos

a_1 instead of α_1 in (6.169) and a_0 instead of α_0 below.

Away from Jeffreys' prior

Question # 3

How comes the limiting case $\alpha \to \infty$ is left *"as an exercise to the reader"*? While it is an important instance for Jeffreys' prior...

A conclusion from Brittany

“The inhabitants from St. Malo [a small French town on the English channel] are convinced; for a century, in their village, the number of deaths at the time of high tide has been greater than at low tide (...) common sense demands more evidence before considering it even plausible that the tide influences the last hour of the Malouins.”(p.195)

A conclusion from Brittany

"The inhabitants from St. Malo [a small French town on the English channel] are convinced; for a century, in their village, the number of deaths at the time of high tide has been greater than at low tide (...) common sense demands more evidence before considering it even plausible that the tide influences the last hour of the Malouins."(p.195)
"In St Malo, the data does not speak for themselves."(p.196)

[© Nuit Blanche]

A defence of the normal distribution (Chapter 7)

"Bayesian inferences using a Gaussian sampling distribution could be improved upon only by one who had additional information about the actual errors beyong its first two moments."

"For nearly two centuries, the Gaussian distribution has continued to be, in almost all problems, much easier to use and to yield better results (more accurate parameter estimates) than any alternative sampling distribution that anyone had to suggest." (p.210)

In preparation of the MaxEnt principle

- Attempts at justifying the use of the normal distribution, incl. the CLT and Gauß' arithmetic mean
- Possible to fit a Gaussian with only the first two moments (§7.6) and links with sufficiency (§7.11)
- Pseudo-likelihood arguments (§7.6 & §7.10)
- Higher entropy (§7.14)

In preparation of the MaxEnt principle

- Attempts at justifying the use of the normal distribution, incl. the CLT and Gauß' arithmetic mean
- Possible to fit a Gaussian with only the first two moments (§7.6) and links with sufficiency (§7.11)
- Pseudo-likelihood arguments (§7.6 & §7.10)
- Higher entropy (§7.14)

Question # 4

Why is an higher accuracy of parameter estimates relevant when the sampling distribution is incorrect? Is it in the asymptotic variance sense?

An interesting historical aside

"The starting point of Darwin's theory of evolution is precisely the existence of those differences between individual members of a race of species which morphologists for the most part rightly neglect."
W.F. Weldon, Biometrika, **1**, p.1

Chapter 8: Sufficiency, ancillarity and all that

1. Goals of the course

2. Elementary hypothesis testing

3. Elementary parameter estimation

4. Sufficiency, ancillarity and all that
   - Sufficiency
   - Ancillarity
   - Repetitive experiments

5. The entropy principle

6. Decision theory

Ronald Fisher

Ronald Fisher

Meanwhile Ronald Fisher (1890–1962), had rejected the Bayesian approach (1922–1924) and based his work, including maximum likelihood, on frequentist foundations (?).

Fisher sufficiency

> *"Unused parts of the data must be* irrelevant *to the question we are asking (...) then it would not matter if they were unknown"* (p.244)

Fisher sufficiency

> *"Unused parts of the data must be* irrelevant *to the question we are asking (...) then it would not matter if they were unknown"* (p.244)

Fisher sufficiency defined as

$$p(x_1 \dots x_n|\theta) = p(r|\theta)b(x_1, \dots, x_n)$$

(missing comas intentional!)
Criticised by Jaynes: *"Fisher's reasoning that $y_2, \dots, y_n$ can convey no information about θ was only a conjecture (...) which did not use the concepts of prior and posterior probabilities."* (p.245)

First illustrations

- normal model (§8.2.1)
- the "Blackwell–Rao theorem" (§8.2.2) [with a quadratic risk rather than a generic convex loss]

  > *"not compelling to a Bayesian, because the criterion of risk is a purely sampling theory notion that ignores prior information."* (p.248)

- counterexample of the Cauchy distribution

First illustrations

- normal model (§8.2.1)
- the "Blackwell–Rao theorem" (§8.2.2) [with a quadratic risk rather than a generic convex loss]

  > *"not compelling to a Bayesian,*[*] *because the criterion of risk is a purely sampling theory notion that ignores prior information."* (p.248)

- counterexample of the Cauchy distribution

[*] It is clearly compelling for a computational Bayesian!

Generalised sufficiency

“The property R [of factorizing through a function r of the data] *may hold under weaker conditions that depend on which prior we assign. Thus the notion of sufficiency which originated in the Bayesian consideration of Laplace actually has a wider meaning in Bayesian inference.”* (p.249)

Generalised sufficiency

> *"The property R* [of factorizing through a function r of the data] *may hold under weaker conditions that depend on which prior we assign. Thus the notion of sufficiency which originated in the Bayesian consideration of Laplace actually has a wider meaning in Bayesian inference."* (p.249)

Fairly interesting concept, which leads to an integral equation on the prior $f(\theta)$ when $y_2, \ldots, y_n$ is a completion of the sufficient statistic r:

$$\int_\Theta \left\{ g(y|\theta) \frac{\partial g(y|\theta')}{\partial y_i} - g(y|\theta') \frac{\partial g(y|\theta)}{\partial y_i} \right\} f(\theta')\,\mathrm{d}\theta' = 0$$

Prior dependent sufficiency

> *"If there are non-negative solutions* [in f] *they will determine a subclass of priors for which* r *would play the role of a sufficient statistic. [It is a] possibility that, for different priors, different functions* $r(x_1, \ldots, x_n)$ *of the data may take on the role of sufficient statistics."* (p.249)

Quite compelling argument from Jaynes:

> *"As soon as we think of probability distributions as* carriers of information *[it is] trivial and obvious. A piece of information in the data makes a difference in our conclusions only when it tells us something that the prior information does not."* (p.249)

Formal idea?

Question #4

There is no illustration of this generalised sufficiency, is it because nothing except the trivial example of a Dirac point mass makes sense?

Formal idea?

Question #4

There is no illustration of this generalised sufficiency, is it because nothing except the trivial example of a Dirac point mass makes sense?

Question #5

Is sufficiency a sampling property in that it must hold for all datasets and not only for the observed one?

The case of nuisance parameters

If one of the parameters θ_2 is a nuisance parameter, the weaker requirement would be that $p(\theta_1|DI)$ *[with I curiously missing!]* only depends on $r(x_1, \ldots, x_n)$.

The case of nuisance parameters

If one of the parameters θ_2 is a nuisance parameter, the weaker requirement would be that $p(\theta_1|DI)$ *[with I curiously missing!]* only depends on $r(x_1, \ldots, x_n)$.

Partial sufficiency has however been know to suffer from paradoxes, as exposed in Basu (1985).

The Likelihood Principle

Derivation of

> **LP:** the likelihood function $L(D)$ from data D contains all the information about θ that is contained in D.

[Barnard, 1947]

The Likelihood Principle

Derivation of

> **LP:** the likelihood function $L(D)$ from data D contains all the information about θ that is contained in D.

[Barnard, 1947]

from

> **CP:** Recognition of an experiment that might have been performed, but was not, cannot tell us anything about θ.

[Birnbaum, 1962]

The Likelihood Principle

Derivation of

> **LP:** the likelihood function $L(D)$ from data D contains all the information about θ that is contained in D.

[Barnard, 1947]

from

> **CP:** Recognition of an experiment that might have been performed, but was not, cannot tell us anything about θ.

[Birnbaum, 1962]

> *"It is important to note that the likelihood principle refers only to the context of a specific model which is not being questioned."* (p.252)

[Preparation of model choice]

Ancillarity

Obvious concept supplementing sufficiency: a statistic z is ancillary if $p(z|\theta I) = p(z|I)$ and conditioning upon any ancillary statistic provides a more informative distribution.

Ancillarity

Obvious concept supplementing sufficiency: a statistic z is ancillary if $p(z|\theta I) = p(z|I)$ and conditioning upon any ancillary statistic provides a more informative distribution.

Not so obvious for Jaynes:

> *"We do not know Fisher's private reason for imposing this independence [on θ] (...) What Fisher's procedure accomplishes is nothing at all: any method of inference that respects the likelihood principle will lead to just the same inferences about θ."* (p.253)

Ancillarity

Obvious concept supplementing sufficiency: a statistic z is ancillary if $p(z|\theta I) = p(z|I)$ and conditioning upon any ancillary statistic provides a more informative distribution.

Not so obvious for Jaynes:

> *“It is the width of the likelihood function from the one dataset that we actually have that tells us the accuracy of the estimate from that dataset. For a Bayesian the question of ancillarity never comes up at all.”*(p.254)

Generalised ancillarity

Introduction of extra data $Z = (z_1, \ldots, z_m)$ such that

$$p(\theta|ZI) = p(\theta|I)$$

Then

$$p(\theta|DZI) = p(\theta|I)\frac{p(D|\theta ZI)}{p(D|ZI)}$$

and everything is conditional on Z.

Generalised ancillarity

Introduction of extra data $Z = (z_1, \ldots, z_m)$ such that

$$p(\theta|ZI) = p(\theta|I)$$

Then

$$p(\theta|DZI) = p(\theta|I)\frac{p(D|\theta ZI)}{p(D|ZI)}$$

and everything is conditional on Z.

[What's the point?!]

AA=A

Discussion on using the data “twice” leading to

$$p(\theta|EDI) = p(\theta|DI) \text{ if } p(E|\theta DI) = 1$$

[a constant would work as well, i.e. turning D into a sufficient statistic for ED]

Relates to criticisms of data dependent priors, empirical Bayes, Aitkin’s (1991) posterior distribution of the likelihood, &c.

On frequency vs. probability (Chapter 9)

"A frequency is a factual property of the real world that we measure or estimate. The phrase 'estimating a probability' is just as much an incongruity as 'assigning a frequency'. The fundamental, inescapable distinction between probability and frequency lies in this relativity principle: probabilities change when we change our state of knowledge, frequencies do not." (p.292)

First steps towards entropy

Given a state-space $G = \{g_1, \ldots, g_m\}$, definition of a Gibbs-like distribution

$$p(g_i) = \exp\{-\lambda g_i\}/Z(\lambda)$$

with $Z(\lambda)$ the partition function.

Gibbs entropy maximisation

Frequencies f_j that maximise entropy

$$-\sum_j f_j \log f_j$$

under constraint

$$\bar{G} = \sum f_j g_j$$

are given by

$$f_j^\star = \exp\{-\lambda g_i\}/Z(\lambda)$$

with λ given by constraint

[Alternative Lagrangian derivation]

Chapter 11: Discrete prior probabilities: the entropy principle

5 The entropy principle
- Entropy
- Ignorance priors
- Uniform priors
- Transformation groups

Entropy desideratas

Jaynes introduces the entropy by looking for a measure of uncertainty associated with a probability distribution $(p_1, \dots, p_n)$ that is

- numerical, $H(p_1, \dots, p_n)$
- continuous in the p_i's
- naturally increasing in the sense that $H(1/n, \dots, 1/n)$ is increasing with n
- consistent [?]

Entropy desideratas

Jaynes introduces the entropy by looking for a measure of uncertainty associated with a probability distribution $(p_1, \ldots, p_n)$ that is

- numerical, $H(p_1, \ldots, p_n)$
- continuous in the p_i's
- naturally increasing in the sense that $H(1/n, \ldots, 1/n)$ is increasing with n
- consistent [?]

For instance, if moving from $(p_1, q = 1 - p_1)$ to (p_1, p_2, p_3). consistency means

$$H_2(p_1, q) = H_2(p_1, q) + qH_2\left(\frac{p_2}{q}, \frac{p_3}{q}\right)$$

Exercise à la Jaynes

Exercise 11.1

It seems intuitively that the most general condition of consistency would be a functional equation which is satisfied by any monotonic increasing function of H_n. But this is ambiguous unless we say something about how the monotonic functions for different n are to be related; is it possible to invoke the same function for all n? Carry out some new research in this field by investigating this matter; try either to find a possible form of the new functional equations, or to explain why this cannot be done.

Consequences

Solving the above equation in the general case

$$H(p_1,\dots,p_n) = H(w_1,\dots,w_r)+w_1H\left(\frac{p_1}{w_1},\dots,\frac{p_k}{w_k}\right)+w_2H\left(\frac{p_{k+1}}{w_2},\dots,\frac{p_{k+m}}{w_2}\right)+\cdots$$

leads to the functional equation

$$h(\sum n_i) = H\left(\frac{n_1}{\sum n_i},\dots,\frac{n_n}{\sum n_i}\right)+\sum_i \frac{n_i}{\sum n_j}h(n_i)$$

and to the entropy

$$H(p_1,\dots,p_n) = -\sum_i p_i \log(p_i)$$

Properties

- Solution unique up to the choice of a logarithmic basis
- *“in accordance with Gödel’s theorem, one cannot prove that it actually is consistent”* (p.350)
- *“many years of use of the maximum entropy has not revealed any inconsistency”* (p.351)

Maximum entropy solution

Theorem

For a finite state space $\mathfrak{X} = \{x_1, \ldots, x_n\}$, the distribution $\mathbf{p} = (p_1, \ldots, p_n)$ that maximises the entropy $H(p_1, \ldots, p_n)$ under the moment constraints

$$\sum_{i=1}^{n} p_i f_k(x_i) = F_k \qquad 1 \leq k \leq m$$

is given by

$$p_i = \exp\left\{-\lambda_0 - \sum_{j=1}^{m} \lambda_j f_j(x_i)\right\}$$

where the λ_j's are determined by the constraints and

$$\lambda_0 = \log Z(\lambda_1, \ldots, \lambda_m)$$

Maximum entropy solution

Proof]

Introduce Lagrange multipliers

$$\mathfrak{h}(p_1, \ldots, p_n; \lambda_1, \ldots, \lambda_m) = -\sum_i p_i \log(p_i) - (\lambda_0 - 1)\sum_i p_i - \sum_j \lambda_j \left\{ \sum \right.$$

and take derivatives

$$\frac{\partial \mathfrak{h}}{\partial p_i} = -\log p_i - 1 - \lambda_0 + 1 - \sum_j \lambda_j f_j(x_i) = 0$$

Jaynes’s worry

“Our Lagrange multiplier arguments has the nice feature that it gives us the answer instantaneously. It has the bad feature that after we done (sic!) *it, we’re not quite sure it* is *the answer (...) There would always be a little grain of doubt remaining if we do only the variational problem.”*(p.357)

Jaynes's worry

> *"Our Lagrange multiplier arguments has the nice feature that it gives us the answer instantaneously. It has the bad feature that after we done* (sic!) *it, we're not quite sure it* is *the answer (...) There would always be a little grain of doubt remaining if we do only the variational problem."*(p.357)

Question 5

What about the convexity of the entropy function $H(p_1, \ldots, p_n)$?

About the Lagrange multipliers

Due to the constraints,

$$F_k = -\frac{\partial \log Z(\lambda_1, \ldots, \lambda_m)}{\partial \lambda_k}$$

If

$$S(F_1, \ldots, F_m) = \log Z(\lambda_1, \ldots, \lambda_m) + \sum_{k=1}^{m} \lambda_k F_k$$

then

$$\lambda_k = \frac{\partial S(F_1, \ldots, F_k)}{\partial F_k}$$

“in which λ_k is given explicitely” (p.359)

About the Lagrange multipliers

Due to the constraints,

$$F_k = -\frac{\partial \log Z(\lambda_1, \ldots, \lambda_m)}{\partial \lambda_k}$$

If

$$S(F_1, \ldots, F_m) = \log Z(\lambda_1, \ldots, \lambda_m) + \sum_{k=1}^{m} \lambda_k F_k$$

then

$$\lambda_k = \frac{\partial S(F_1, \ldots, F_k)}{\partial F_k}$$

"in which λ_k is given explicitely" (p.359)

Question 6

In which sense is $S(F_1, \ldots, F_k)$ an explicit function of the F_k's?

Objections [in which world?!]

- *“Maximum uncertainty is a negative thing”*
- *“Probabilities obtained by maximum entropy cannot be relevant to physical predictions because they have nothing to do with frequencies”*
- *“The given data* $\{F_1, \ldots, F_n\}$ *are not averages, but definite measured numbers”*
- *“Different people have different information (...) the results are basically arbitrary”* (p.366)

A compelling argument?

Consider N multinomial trials on $\mathfrak{X} = \{x_1, \ldots, x_n\}$ with outcome $(n_1, \ldots, n_n)$.
Under (empirical) constraints like

$$\sum_{i=1}^{n} n_i f_k(x_i) = NF_k \qquad 1 \leq k \leq m$$

that are insufficient to specify the p_i's if $m < n$, what is the best choice for $(p_1, \ldots, p_n)$?

A not-so-compelling argument

Jaynes argues that the choice of $(p_1, \ldots, p_n)$ should maximise the number of occurences of $(n_1, \ldots, n_n)$, i.e.

$$W = \frac{N!}{(Np_1)! \cdots (Np_n)!}$$

Then another choice $\mathbf{p}' = (p_1', \ldots, p_n')$ leads to

$$\frac{W}{W'} \longrightarrow \exp\left\{N[H(\mathbf{p}) - H(\mathbf{p}')]\right\}$$

i.e. *"the frequency predicted by maximum entropy can be realized in overwhelmingly many more ways than any other"* (p.368)

Ignorance priors

◂ Noninformative 101

"The natural starting point in translating a number of pieces of prior information is the state of complete ignorance (...) When we advance to complicated problems, a formal theory of how to find ignorance priors becomes more and more necessary"(p.373)

Ignorance priors

◂ Noninformative 101

"The natural starting point in translating a number of pieces of prior information is the state of complete ignorance (...) When we advance to complicated problems, a formal theory of how to find ignorance priors becomes more and more necessary"(p.373)

"Some object to the very attempt on the ground that a state of complete ignorance does not exist."(p.373)

Useless jibe

“There is a large Bayesian community whose members call themselves ‘subjective Bayesians’, who have settled in a position intermediate between ‘orthodox’ statistics and the theory expounded here.”(p.372)

Useless jibe

"There is a large Bayesian community whose members call themselves 'subjective Bayesians', who have settled in a position intermediate between 'orthodox' statistics and the theory expounded here."(p.372)

"Having specified the prior information, we then have the problem of translating that information into a specific prior probability assignment, a formal translation process (...) only dimly perceived in subjective Bayesian theory."(p.373)

Entropy for continuous distributions

Jaynes points out that the quantity

$$H = -\int p(x) \log p(x) \,\mathrm{d}x$$

depends on the parameterisation, hence a difficulty to expand Shannon's information beyond the discrete case.

Entropy for continuous distributions

Jaynes points out that the quantity

$$H = -\int p(x) \log p(x) \,\mathrm{d}x$$

depends on the parameterisation, hence a difficulty to expand Shannon's information beyond the discrete case.
[Necessary but awkward] introduction of a reference measure $m(x)$ in

$$H = -\int p(x) \log \left[\frac{p(x)}{m(x)} \right] \,\mathrm{d}x$$

The unavoidable measure

While Jaynes considers $m(\cdot)$ to be well-enough behaved to allow for Riemann integrals, the choice of the reference measure is paramount to the definition of both the entropy and the maximum entropy priors.

The unavoidable measure

While Jaynes considers $m(\cdot)$ to be well-enough behaved to allow for Riemann integrals, the choice of the reference measure is paramount to the definition of both the entropy and the maximum entropy priors.

> *"If the parameter space is not the result of any obvious limiting process what determines the proper measure $m(x)$? This is the shortcoming from which the maximum entropy has suffered."*(p.376)

Uniform prior on $\mathbb{R}$

If the parameter may have any value in a finite range, or from $-\infty$ to $+\infty$, its prior probability should be taken as uniformly distributed (Theory of Probability, III, §3.1).

Normal illustration

Example (Flat prior)

If $x \sim \mathcal{N}(\theta, 1)$ and $\pi(\theta) = \varpi$, constant, the pseudo marginal distribution is

$$m(x) = \varpi \int_{-\infty}^{+\infty} \frac{1}{\sqrt{2\pi}} \exp\left\{-(x-\theta)^2/2\right\} d\theta = \varpi$$

Normal illustration

Example (Flat prior)

If $x \sim \mathcal{N}(\theta, 1)$ and $\pi(\theta) = \varpi$, constant, the pseudo marginal distribution is

$$m(x) = \varpi \int_{-\infty}^{+\infty} \frac{1}{\sqrt{2\pi}} \exp\left\{-(x-\theta)^2/2\right\} d\theta = \varpi$$

and the posterior distribution of θ is

$$\pi(\theta \mid x) = \frac{1}{\sqrt{2\pi}} \exp\left\{-\frac{(x-\theta)^2}{2}\right\},$$

i.e., corresponds to a $\mathcal{N}(x, 1)$ distribution.

Normal illustration

Example (Flat prior)

If $x \sim \mathcal{N}(\theta, 1)$ and $\pi(\theta) = \varpi$, constant, the pseudo marginal distribution is

$$m(x) = \varpi \int_{-\infty}^{+\infty} \frac{1}{\sqrt{2\pi}} \exp\left\{-(x-\theta)^2/2\right\} d\theta = \varpi$$

and the posterior distribution of θ is

$$\pi(\theta \mid x) = \frac{1}{\sqrt{2\pi}} \exp\left\{-\frac{(x-\theta)^2}{2}\right\},$$

i.e., corresponds to a $\mathcal{N}(x, 1)$ distribution.

[independent of ω]

“Complete ignorance”

Warning – Warning – Warning – Warning – Warning

The mistake is to think of them [non-informative priors] as representing ignorance

[Lindley, 1990]

Over-interpretation

If we take

$$P(d\sigma|H) \propto d\sigma$$

as a statement that σ may have any value between 0 and ∞ (...), we must use ∞ instead of 1 to denote certainty on data H. (..) But (..) the number for the probability that $\sigma < \alpha$ will be finite, and the number for $\sigma > \alpha$ will be infinite. Thus (...) the probability that $\sigma < \alpha$ is 0. This is inconsistent with the statement that we know nothing about σ (Theory of Probability, III, §3.1)

▸ mis-interpretation

Over-interpretation (2)

Example (Flat prior (2))

Consider a $\theta \sim \mathcal{N}(0, \tau^2)$ prior. Then, for any (a, b)

$$\lim_{\tau \to \infty} P^\pi \left(\theta \in [a, b]\right) = 0$$

...we usually have some vague knowledge initially that fixes upper and lower bounds [but] the truncation of the distribution makes a negligible change in the results (Theory of Probability, III, §3.1)

[Not!]

Over-interpretation (3)

Example (Haldane prior)

For a binomial observation, $x \sim \mathcal{B}(n, p)$, and prior $\pi^*(p) \propto [p(1-p)]^{-1}$, the marginal distribution,

$$\begin{aligned} m(x) &= \int_0^1 [p(1-p)]^{-1} \binom{n}{x} p^x (1-p)^{n-x} dp \\ &= B(x, n-x), \end{aligned}$$

is only defined for $x \neq 0, n$.

Missed by Jeffreys:

> *If a sample is of one type with respect to some property there is probability 1 that the population is of that type (Theory of Probability, III, §3.1)*

Uniform difficulties

Lack of reparameterization invariance/coherence

$$\psi = e^\theta \qquad \pi_1(\psi) = \frac{1}{\psi} \neq \pi_2(\psi) = 1$$

There are cases of estimation where a law can be equally well expressed in terms of several different sets of parameters, and it is desirable to have a rule that will lead to the same results whichever set we choose. Otherwise we shall again be in danger of using different rules arbitrarily to suit our taste (Theory of Probability, III, §3.1)

Parameterisation variability

Example (Jeffreys' example, ToP, III, §3.1)

If

$$\pi_V(v) \propto 1\,,$$

then $W = V^n$ is such that

$$\pi_W(w) \propto w^{(n-1)/n}$$

Difficulties (2)

Problems of proper-ness

$$x \sim \mathcal{N}(\theta, \sigma^2), \qquad \pi(\theta, \sigma) = 1$$

$$\begin{array}{rlll} & \pi(\theta, \sigma | x) & \propto & e^{-(x-\theta)^2/2\sigma^2}\sigma^{-1} \\ \Rightarrow & \pi(\sigma | x) & \propto & 1 \qquad (!!!) \end{array}$$

Difficulties (3)

Inappropriate for testing point null hypotheses:

> *The fatal objection to the universal application of the uniform distribution is that it would make any significance test impossible. If a new parameter is being considered, the uniform distribution of prior probability for it would practically always lead to the result that the most probable value is different from zero (Theory of Probability, III,§3.1)*

Difficulties (3)

Inappropriate for testing point null hypotheses:

> *The fatal objection to the universal application of the uniform distribution is that it would make any significance test impossible. If a new parameter is being considered, the uniform distribution of prior probability for it would practically always lead to the result that the most probable value is different from zero (Theory of Probability, III,§3.1)*

but so would any continuous prior!

A strange conclusion

“The way out is in fact very easy”:

> *If v is capable of any value from 0 to ∞, and we take its prior probability distribution as proportional to dv/v, then $\varrho = 1/v$ is also capable of any value from 0 to ∞, and if we take its prior probability as proportional to $d\rho/\rho$ we have two perfectly consistent statements of the same form (Theory of Probability, III, §3.1)*

A strange conclusion

"The way out is in fact very easy":

> *If v is capable of any value from 0 to ∞, and we take its prior probability distribution as proportional to dv/v, then $\varrho = 1/v$ is also capable of any value from 0 to ∞, and if we take its prior probability as proportional to $d\rho/\rho$ we have two perfectly consistent statements of the same form (Theory of Probability, III, §3.1)*

Seems to consider that the objection of ◂ 0 probability result only applies to parameters with $(0, \infty)$ support.

ToP difficulties (§3.1)

Jeffreys tries to justify the prior $\pi(v) \propto 1/v$ as "correct" prior, by usual argument that this corresponds to flat prior on log v, although Jeffreys rejects Haldane's prior which is based on flat prior on the logistic transform $v/(1-v)$

> *...not regard the above as showing that $dx/x(1-x)$ is right for their problem. Other transformations would have the same properties and would be mutually inconsistent if the same rule was taken for all. ...[even though] there is something to be said for the rule (Theory of Probability, III, §3.1)* ▸ Not Jaynes' view

$$P(dx|H) = \frac{1}{\pi} \frac{dx}{\sqrt{x(1-x)}} \,.$$

(continued)

Very shaky from a mathematical point of view:

> *...the ratio of the probabilities that v is less or greater than a is*
>
> $$\int_0^a v^n dv \Big/ \int_a^\infty v^n dv\,.$$
>
> *(...) If $n < -1$, the numerator is infinite and the denominator finite and the rule would say that the probability that v is greater than any finite value is 0. (...) But if $n = -1$ both integrals diverge and the ratio is indeterminate. (...) Thus we attach no value to the probability that v is greater or less than a, which is a statement that we know nothing about v except that it is between 0 and ∞ (Theory of Probability, III, §3.1)*

From ToP to PT

"Jeffreys suggested that we assign a prior $\mathrm{d}\sigma/\sigma$ *to a continuous parameter known to be positive on the grounds that we are saying the same thing whether we use* σ *or* σ^m*."*(p.377)

[Right!]

From ToP to PT

"Jeffreys suggested that we assign a prior $\mathrm{d}\sigma/\sigma$ *to a continuous parameter known to be positive on the grounds that we are saying the same thing whether we use* σ *or* σ^m*."*(p.377)

[Right!]

"We do not want (and obviously cannot have) invariance to more general parameter changes."(p.377)

[Wrong! Witness Jeffreys' priors!]

Invariant priors

Principle: Agree with the natural symmetries of the problem

- Identify invariance structures as group action

$$\begin{array}{lcl} \mathcal{G} & : & x \rightarrow g(x) \sim f(g(x)|\bar{g}(\theta)) \\ \bar{\mathcal{G}} & : & \theta \rightarrow \bar{g}(\theta) \\ \mathcal{G}^* & : & L(d,\theta) = L(g^*(d),\bar{g}(\theta)) \end{array}$$

- Determine an invariant prior

$$\pi(\bar{g}(A)) = \pi(A)$$

Generic solution

Right Haar measure

But...

- Requires invariance to be part of the decision problem
- Missing in most discrete setups (Poisson)
- Invariance must somehow belong to prior setting/information

Location-scale example (§12.4.1)

If

$$p(x|\nu\sigma)\,\mathrm{d}x = \frac{1}{\sigma} h\left(\frac{x-\nu}{\sigma}\right)$$

"a change of scale and shift of location does not change that state of knowledge" (p.379) and the invariant prior under

$$f(\nu,\sigma) = af(\nu+b, a\sigma)$$

is

$$f(\nu,\sigma) \propto \sigma^{-1}$$

[Right-Haar measure and Jeffreys prior]

Location-scale example (§12.4.1)

If

$$p(x|\nu\sigma)\,\mathrm{d}x = \frac{1}{\sigma} h\left(\frac{x-\nu}{\sigma}\right)$$

"a change of scale and shift of location does not change that state of knowledge" (p.379) and the invariant prior under

$$f(\nu,\sigma) = a^2 f(a\nu + b, a\sigma)$$

is

$$f(\nu,\sigma) \propto \sigma^{-2}$$

[Left-Haar measure and not Jeffreys prior]

No group action example (§12.4.2)

Case of the Poisson distribution/process

$$p(n|\lambda t) = \frac{(\lambda t)^n}{n!}\,\exp\{-\lambda t\}$$

which must be invariant under a time scale change, leading to

$$f(\lambda) = qf(q\lambda) \quad \text{i.e.} \quad f(\lambda) \propto \lambda^{-1}$$

Another no-group-action example (§12.4.3)

Case of a binomial distribution

$$p(r|n\theta) = \theta^r(1-\theta)^{n-r}$$

where no group action can take place [except for the trivial $p \to (1-p)$]

Another no-group-action example (§12.4.3)

Case of a binomial distribution

$$p(r|n\theta) = \theta^r(1-\theta)^{n-r}$$

where no group action can take place [except for the trivial $p \rightarrow (1-p)$] Jaynes argues that the prior on θ should not change when provided with a piece of evidence E, modifying θ into

$$\theta' = \frac{\theta p(E|SX)}{\theta p(E|SX) + (1-\theta)p(E|FX)} \triangleq \frac{a\theta}{1-\theta+a\theta}$$

ending up with Haldane's prior

$$f(\theta) \propto 1/\theta(1-\theta)$$

◂ Not Jeffreys' choice

Laplace's succession rule (rev'ed)

Hypergeometric setting

$$p(r|nRN) = \frac{\binom{R}{r}\binom{N-R}{n-r}}{\binom{N}{n}}$$

and *law of natural induction*

$$p(R = N|rnN) = \frac{n+1}{n+2}$$

under uniform prior $p(R|N) = 1/(N+1)$

Laplace's succession rule (rev'ed)

Hypergeometric setting

$$p(r|nRN) = \frac{\binom{R}{r}\binom{N-R}{n-r}}{\binom{N}{n}}$$

and *law of natural induction*

$$p(R = N|rnN) = \frac{n + 1/2}{n + 1}$$

under Jeffreys prior $p(R|N) \approx 1/\pi\sqrt{R(N - R)}$

Laplace's succession rule (rev'ed)

Hypergeometric setting

$$p(r|nRN) = \frac{\binom{R}{r}\binom{N-R}{n-r}}{\binom{N}{n}}$$

and *law of natural induction*

$$p(R=N|rnN) \approx \frac{\sqrt{n}}{\sqrt{n}+\pi^{-1/2}}$$

under reference prior

$$p(R|N) = \begin{cases} 1/2 & \text{if } R=N \\ 1/2\pi\sqrt{R(N-R)} & \text{if } 0 \le R \le N-1 \end{cases}$$

[Jeffreys, 1939; Berger, Bernardo, & Sun, 2009]

Bertrand's random line

"Bertrand's problem was stated originally in terms of drawing a straight line 'at random' intersecting a circle (...) we do no violence to the problem if we suppose we are tossing straws onto the circle (...) What is the probability that the chord thus defined [by a random straw] is longer than the side of the inscribed equilateral triangle?"(p.386)

Bertrand’s undefined randomness

Historically Bertrand (1889) used this illustration to show the relevance of the reference measure (or of the underlying σ-algebra!): depending on whether the reference measure is *“uniform on (a) linear distances between centers of chord and circle; (b) angles of intersections of the chord on the circumference; (c) the center of the chord over the intererior area of the circle, the probabilities are* $1/2$, $1/3$ *and* $1/4$, *respectively.”*

Bertrand's undefined randomness

Historically Bertrand (1889) used this illustration to show the relevance of the reference measure (or of the underlying σ-algebra!): depending on whether the reference measure is *"uniform on (a) linear distances between centers of chord and circle; (b) angles of intersections of the chord on the circumference; (c) the center of the chord over the interior area of the circle, the probabilities are* $1/2$, $1/3$ *and* $1/4$, *respectively."*

Jaynes wonders at *"which answer is correct?"* as if the problem had an answer.

Jaynes' redefined randomness

Jaynes mentions checking by experiment and assessing by frequency, but this seems self defeating, since the “randomness” of the straw draws cannot be defined.

[Just try to write an R code!]

Jaynes' redefined randomness

Jaynes mentions checking by experiment and assessing by frequency, but this seems self defeating, since the "randomness" of the straw draws cannot be defined.

[Just try to write an R code!]

Jaynes introduces invariance under (a) rotation,

$$f(r,\theta) = g(r,\theta-\alpha) = f(r)$$

(b) scale,

$$f(r) = a^2 f(ar) = \frac{qr^{q-2}}{2\pi R^q}$$

and (c) translation tranforms:

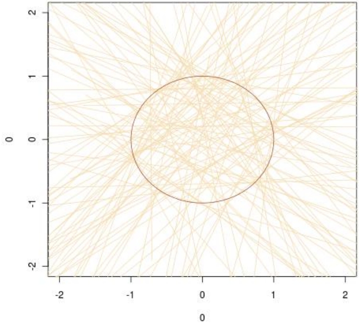
$$f(r) = \frac{1}{2\pi Rr}$$

And the winner is...

R program ran with straw endpoints generated uniformly over a large

$$(-10 \times R, 10 \times R)^2$$

box until the straw covers the unit circle,

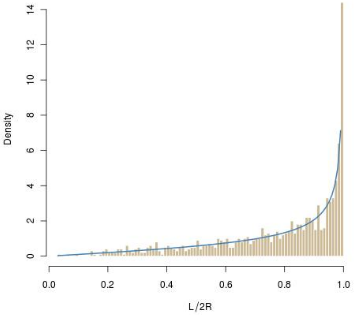


And the winner is...

R program ran with straw endpoints generated uniformly over a large

$$(-10 \times R, 10 \times R)^2$$

box until the straw covers the unit circle, resulting in an aggreement with Borel's and Jaynes' *"universal distribution law"* (p.393)!



A moderate conclusion

" On the one hand, one cannot deny the force of arguments which (...) demonstrate the ambiguity of dangers in the principle of indifference. On the other hand, it is equally undeniable that use of that principle has, over and over, led to correct, nontrivial, and useful predictions (...) Cases to which the principle of indifference has been applied successfully in the past are just the ones in which the actual calculations are seen as an application of indifference between problems, rather than events."(p.395)

Chapters 13-14-20: Decision theory

6 Decision theory
- Background
- Loss functions
- Minimaxity and admissibility
- Model comparison

St Petersburg's paradox

> *"Toss an honest coin till it comes heads for the first time. If it occurs at the nth throw, the player receives 2^n dollars."*(p.399)

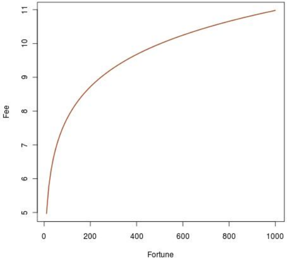
The paradox is that a player should agree to pay an infinite amount since the expected profit is

$$\sum_{k=1}^{\infty} 2^{-k} 2^{k} = +\infty$$

Moral expectation

Bernoulli and Laplace solved the “paradox” with another utility function, $\log(M)$: *“For an initial fortune of m francs, the fair fee* $f(m)$ *is determined by”*

$$\log(m) = \sum_{n=1}^{\infty} 2^{-n} \log(m - f + 2^n)$$



The classical if honest wheatherman

Another example of log utility:

$$\log(n) - \sum_i p_i \log q_i$$

for predicted probabilities q_i and “true” probabilities p_i.

The classical if honest wheatherman

Another example of log utility:

$$\log(n) - \sum_i p_i \log q_i$$

for predicted probabilities q_i and “true” probabilities p_i.
Back to entropy $\log(n) - H(p_1, \ldots, p_n)$ at the maximum

General comments

"In the 1930's and 1940's, a form of decision rules was expounded by J. Neyman and E.S. Pearson. It enjoyed a period of popularity with electrical engineers and economists, but is now obsolete."(p.404)

General comments

> *"In the 1930's and 1940's, a form of decision rules was expounded by J. Neyman and E.S. Pearson. It enjoyed a period of popularity with electrical engineers and economists, but is now obsolete."*(p.404)

Description of Wald's formulation of decision theory via the loss function $L(D_i, \theta_i)$ connected with Raiffa and Schlaifer (1961) and Berger (1985)

Bayesian Decision Theory

Three spaces/factors:

(1) On $\mathcal{X}$, distribution for the observation, $f(x|\theta)$;

Bayesian Decision Theory

Three spaces/factors:

(1) On $\mathcal{X}$, distribution for the observation, $f(x|\theta)$;

(2) On Θ, prior distribution for the parameter, $\pi(\theta)$;

Bayesian Decision Theory

Three spaces/factors:

(1) On $\mathcal{X}$, distribution for the observation, $f(x|\theta)$;

(2) On Θ, prior distribution for the parameter, $\pi(\theta)$;

(3) On $\Theta \times \mathcal{D}$, loss function associated with the decisions, $\mathrm{L}(\theta, \delta)$;

Foundations

> *"This theory is clearly of no use unless by 'making a decision' we mean 'acting as if the decision were correct'."*(p.406)

Theorem (**Existence**)

There exists an axiomatic derivation of the existence of a loss function.

[DeGroot, 1970]

> *"(1) there is a continuous gradation (...) and (2) the consequences of an action will in general depend on what is the true state of nature."*(p.407)

Estimators

Decision procedure δ usually called estimator
(while its *value* $\delta(x)$ called estimate of θ)

Estimators

Decision procedure δ usually called estimator
(while its *value* $\delta(x)$ called estimate of θ)

Fact

Impossible to uniformly minimize (in d) the loss function

$$\mathrm{L}(\theta, d)$$

when θ is unknown

Frequentist Principle

Average loss (or frequentist risk)

$$\begin{aligned} R(\theta,\delta) &= \mathbb{E}_\theta[\mathrm{L}(\theta,\delta(x))] \\ &= \int_{\mathcal{X}} \mathrm{L}(\theta,\delta(x)) f(x|\theta)\, \mathrm{d}x \end{aligned}$$

Frequentist Principle

Average loss (or frequentist risk)

$$\begin{aligned} R(\theta, \delta) &= \mathbb{E}_\theta[\mathrm{L}(\theta, \delta(x))] \\ &= \int_{\mathcal{X}} \mathrm{L}(\theta, \delta(x)) f(x|\theta) \, \mathrm{d}x \end{aligned}$$

Principle

Select the best estimator based on the risk function

Difficulties with frequentist paradigm

> *"Risk and admissibility are evidently sampling theory criteria, not Bayesian, since they invoke only sampling distributions."* (p.408)

(1) Error averaged over the different values of x proportionally to the density $f(x|\theta)$: not so appealing for a client, who wants optimal results for **her** data x, *"what is best for the present specific sample"* (p.411)!

Difficulties with frequentist paradigm

> *"Risk and admissibility are evidently sampling theory criteria, not Bayesian, since they invoke only sampling distributions."* (p.408)

(1) Error averaged over the different values of x proportionally to the density $f(x|\theta)$: not so appealing for a client, who wants optimal results for **her** data x, *"what is best for the present specific sample"* (p.411)!

(2) Assumption of repeatability of experiments (*"belief that $R(\theta, \delta)$ is the limit of the average of actual losses"* (p.411) not always grounded.

Difficulties with frequentist paradigm

> *"Risk and admissibility are evidently sampling theory criteria, not Bayesian, since they invoke only sampling distributions."* (p.408)

(1) Error averaged over the different values of x proportionally to the density $f(x|\theta)$: not so appealing for a client, who wants optimal results for **her** data x, *"what is best for the present specific sample"* (p.411)!

(2) Assumption of repeatability of experiments (*"belief that $R(\theta, \delta)$ is the limit of the average of actual losses"* (p.411) not always grounded.

(3) $R(\theta, \delta)$ is a function of θ: there is no total ordering on the set of procedures (§13.8-§13.9).

Bayesian principle

Principle Integrate over the space Θ to get the posterior expected loss

$$\begin{aligned}\rho(\pi, d|x) &= \mathbb{E}^{\pi}[L(\theta, d)|x] \\ &= \int_{\Theta} \mathrm{L}(\theta, d)\pi(\theta|x)\ \mathrm{d}\theta,\end{aligned}$$

Bayesian principle (2)

Alternative

Integrate over the space Θ and compute *integrated risk*

$$\begin{aligned} r(\pi,\delta) &= \mathbb{E}^{\pi}[R(\theta,\delta)] \\ &= \int_{\Theta}\int_{\mathcal{X}} \mathrm{L}(\theta,\delta(x))\, f(x|\theta)\; \mathrm{d}x\; \pi(\theta)\; \mathrm{d}\theta \end{aligned}$$

which induces a **total** ordering on estimators.

Bayesian principle (2)

Alternative

Integrate over the space Θ and compute *integrated risk*

$$\begin{aligned} r(\pi,\delta) &= \mathbb{E}^\pi[R(\theta,\delta)] \\ &= \int_\Theta \int_{\mathcal{X}} \mathrm{L}(\theta,\delta(x))\, f(x|\theta)\ \mathrm{d}x\ \pi(\theta)\ \mathrm{d}\theta \end{aligned}$$

which induces a **total** ordering on estimators.

Existence of an optimal decision

Bayes estimator

Theorem (Construction of Bayes estimators)

An estimator minimizing

$$r(\pi, \delta)$$

can be obtained by selecting, for every $x \in \mathcal{X}$, the value $\delta(x)$ which minimizes

$$\rho(\pi, \delta | x)$$

since

$$r(\pi, \delta) = \int_{\mathcal{X}} \rho(\pi, \delta(x)|x) m(x) \, \mathrm{d}x.$$

Bayes estimator

Theorem (**Construction of Bayes estimators**)

An estimator minimizing

$$r(\pi, \delta)$$

can be obtained by selecting, for every $x \in \mathcal{X}$, the value $\delta(x)$ which minimizes

$$\rho(\pi, \delta | x)$$

since

$$r(\pi, \delta) = \int_{\mathcal{X}} \rho(\pi, \delta(x) | x) m(x) \, \mathrm{d}x.$$

© Both approaches give the same estimator

Bayes estimator (2)

Definition (Bayes optimal procedure)

A *Bayes estimator* associated with a prior distribution π and a loss function L is

$$\arg\min_{\delta} r(\pi, \delta)$$

The value $r(\pi) = r(\pi, \delta^{\pi})$ is called the *Bayes risk*

The quadratic loss

Historically, first loss function (Legendre, Gauss)

$$\mathrm{L}(\theta, d) = (\theta - d)^2$$

The quadratic loss

Historically, first loss function (Legendre, Gauss)

$$\mathrm{L}(\theta, d) = (\theta - d)^2$$

or

$$\mathrm{L}(\theta, d) = ||\theta - d||^2$$

Proper loss

Posterior mean

The Bayes estimator δ^π associated with the prior π and with the quadratic loss is the posterior expectation

$$\delta^\pi(x) = \mathbb{E}^\pi[\theta|x] = \frac{\int_\Theta \theta f(x|\theta)\pi(\theta)\ \mathrm{d}\theta}{\int_\Theta f(x|\theta)\pi(\theta)\ \mathrm{d}\theta}.$$

The absolute error loss

Alternatives to the quadratic loss:

$$\mathrm{L}(\theta, d) = \mid \theta - d \mid,$$

or

$$\mathrm{L}_{k_1,k_2}(\theta, d) = \begin{cases} k_2(\theta - d) & \text{if } \theta > d, \\ k_1(d - \theta) & \text{otherwise.} \end{cases} \tag{5}$$

The absolute error loss

Alternatives to the quadratic loss:

$$\mathrm{L}(\theta, d) = \mid \theta - d \mid,$$

or

$$\mathrm{L}_{k_1,k_2}(\theta, d) = \begin{cases} k_2(\theta - d) & \text{if } \theta > d, \\ k_1(d - \theta) & \text{otherwise.} \end{cases} \tag{5}$$

L_1 estimator

The Bayes estimator associated with π and (5) is a $(k_2/(k_1 + k_2))$ fractile of $\pi(\theta|x)$.

An urban myth

Jaynes (p.4141) maintains that the MAP estimator

$$\arg\max_{\theta} \ell(\theta|x)\pi(\theta)$$

is associated with the $0-1$ loss

$$\mathrm{L}(\theta, d) = \begin{cases} 0 & \text{if } \theta = d \\ 1 & \text{if } \theta \neq d \end{cases}$$

An urban myth

Jaynes (p.4141) maintains that the MAP estimator

$$\arg\max_\theta \ell(\theta|x)\pi(\theta)$$

is associated with the $0-1$ loss

$$\mathrm{L}(\theta,d) = \begin{cases} 0 & \text{if } \theta = d \\ 1 & \text{if } \theta \neq d \end{cases}$$

Wrong for continuous spaces as the MAP depends on the dominating measure

[Druihlet & Marin, 2007]

Minimaxity

Frequentist insurance against the worst case and (weak) total ordering on $\mathcal{D}^*$

Minimaxity

Frequentist insurance against the worst case and (weak) total ordering on $\mathcal{D}^*$

Definition (Frequentist optimality)

The *minimax risk* associated with a loss L is

$$\bar{R} = \inf_{\delta \in \mathcal{D}^*} \sup_{\theta} R(\theta, \delta) = \inf_{\delta \in \mathcal{D}^*} \sup_{\theta} \mathbb{E}_{\theta}[L(\theta, \delta(x))],$$

Minimaxity

Frequentist insurance against the worst case and (weak) total ordering on $\mathcal{D}^*$

Definition (Frequentist optimality)

The *minimax risk* associated with a loss L is

$$\bar{R} = \inf_{\delta\in\mathcal{D}^*} \sup_\theta R(\theta,\delta) = \inf_{\delta\in\mathcal{D}^*} \sup_\theta \mathbb{E}_\theta[L(\theta,\delta(x))],$$

and a *minimax estimator* is any estimator δ_0 such that

$$\sup_\theta R(\theta,\delta_0) = \bar{R}.$$

Criticisms

- *“Nature is not an intelligent adversary”* (p.407)
- Analysis in terms of the worst case, *“the long-faced pessimist”* (p.407)

Criticisms

- *“Nature is not an intelligent adversary”* (p.407)
- Analysis in terms of the worst case, *“the long-faced pessimist”* (p.407)
- Does not incorporate prior information

Criticisms

- *“Nature is not an intelligent adversary”* (p.407)
- Analysis in terms of the worst case, *“the long-faced pessimist”* (p.407)
- Does not incorporate prior information
- Too conservative

Criticisms

- *“Nature is not an intelligent adversary”* (p.407)
- Analysis in terms of the worst case, *“the long-faced pessimist”* (p.407)
- Does not incorporate prior information
- Too conservative
- Difficult to exhibit/construct

Minimaxity (2)

Existence

If $\mathcal{D} \subset \mathbb{R}^k$ convex and compact, and if $\mathrm{L}(\theta, d)$ continuous and convex as a function of d for every $\theta \in \Theta$, there exists a nonrandomized minimax estimator.

Connection with Bayesian approach

The Bayes risks are always smaller than the minimax risk:

$$\underline{r} = \sup_{\pi} r(\pi) = \sup_{\pi} \inf_{\delta \in \mathcal{D}} r(\pi, \delta) \leq \overline{r} = \inf_{\delta \in \mathcal{D}^*} \sup_{\theta} R(\theta, \delta).$$

Connection with Bayesian approach

The Bayes risks are always smaller than the minimax risk:

$$\underline{r} = \sup_{\pi} r(\pi) = \sup_{\pi} \inf_{\delta \in \mathcal{D}} r(\pi, \delta) \leq \overline{r} = \inf_{\delta \in \mathcal{D}^*} \sup_{\theta} R(\theta, \delta).$$

Definition

The estimation problem *has a value* when $\underline{r} = \overline{r}$, i.e.

$$\sup_{\pi} \inf_{\delta \in \mathcal{D}} r(\pi, \delta) = \inf_{\delta \in \mathcal{D}^*} \sup_{\theta} R(\theta, \delta).$$

$\underline{r}$ is the *maximin risk* and the corresponding π the *favourable prior*

Admissibility

Reduction of the set of acceptable estimators based on “local” properties

Definition (Admissible estimator)

An estimator δ_0 is *inadmissible* if there exists an estimator δ_1 such that, for every θ,

$$R(\theta, \delta_0) \geq R(\theta, \delta_1)$$

and, for at least one θ_0

$$R(\theta_0, \delta_0) > R(\theta_0, \delta_1)$$

Admissibility

Reduction of the set of acceptable estimators based on “local” properties

Definition (Admissible estimator)

An estimator δ_0 is *inadmissible* if there exists an estimator δ_1 such that, for every θ,

$$R(\theta, \delta_0) \geq R(\theta, \delta_1)$$

and, for at least one θ_0

$$R(\theta_0, \delta_0) > R(\theta_0, \delta_1)$$

Otherwise, δ_0 is admissible

Minimaxity & admissibility

If there exists a unique minimax estimator, this estimator is admissible.

The converse is false!

Minimaxity & admissibility

If there exists a unique minimax estimator, this estimator is admissible.

The converse is false!

If δ_0 is admissible with constant risk, δ_0 is the unique minimax estimator.

The converse is false!

The Bayesian perspective

Admissibility strongly related to the Bayes paradigm: Bayes estimators often constitute the class of admissible estimators

The Bayesian perspective

Admissibility strongly related to the Bayes paradigm: Bayes estimators often constitute the class of admissible estimators

- If π is strictly positive on Θ, with

$$r(\pi) = \int_\Theta R(\theta, \delta^\pi)\pi(\theta)\ \mathrm{d}\theta < \infty$$

and $R(\theta, \delta)$, is continuous, then the Bayes estimator δ^π is admissible.

The Bayesian perspective

Admissibility strongly related to the Bayes paradigm: Bayes estimators often constitute the class of admissible estimators

- If π is strictly positive on Θ, with
$$r(\pi) = \int_\Theta R(\theta, \delta^\pi)\pi(\theta)\, \mathrm{d}\theta < \infty$$
and $R(\theta, \delta)$, is continuous, then the Bayes estimator δ^π is admissible.
- If the Bayes estimator associated with a prior π is unique, it is admissible.

Regular ($\neq$generalized) Bayes estimators always admissible

Jaynes' objections

About complete classes:

> *"From Wald's viewpoint it is an highly nontrivial mathematical problem to prove that such a class exists, and to find an algorithm by which any rule in the class can be constructed. From our viewpoint, however, these are unnecessary complications, signifying only an inappropriate definition of the term admissible: an inadmissible decision may be overwhelmingly preferable to an admissible one, because the criterion of admissibility ignores prior information."*(p.408)

Jaynes' objections

About admissibility

> *"An estimation rule that simply ignores the data and always estimate* $\theta^* = 5$ *is admissible if the point* $\theta = 5$ *is in the parameter space (...) This illustrates the folly of inventing noble-sounding names like 'admissibility' and 'unbiased' for principles that are far from noble; and not even fully rational."* (p.409)

Jaynes' objections

Yet again about complete classes but the other way:

> *"Wald's complete class theorem [is that] if the θ_j are discrete then the class of admissible strategies is just the class of Bayes strategies. If the possible θ_j form a continuum, the admissible rules are the proper Bayesian ones; i.e. Bayes rules from proper priors. But few people have ever tried to follow his proof."* (p.415)

[Wrong outside compact cases!]

Jaynes' objections

Against improper priors:

> *"There is a great deal of mathematical nitpicking, also noted by Berger, over the exact situation when one tries to jump into an improper prior in infinite parameter spaces without considering any limit from a proper prior (..) the resulting singular mathematics is only an artifact that corresponds to no singularity in the real problem, where prior information always excludes the region at infinity."*(p.415)

[**Question # 7** How much of the region?!]

Auto-biography

"If a sampling theorist will think his estimation problems through to the end, he will find himself obliged to use the Bayesian mathematical algorithm, even if his ideology still leads him to reject the Bayesian rationale for it."(p.415)

Duality

> *"If one worries about arbitrariness in the prior probabilities then one ought to worry just as much about arbitrariness in the loss functions."*(p.419)

Jaynes remarks on the duality between loss and prior since only the product

$$\mathrm{L}(\theta, d)\pi(\theta)$$

matters

[Rubin, 1983]

Non-statistical quotes about loss

"One person may persuade thousands of others to believe his private myths, as the sordid history of religious, political and military disasters shows."
"All of us have felt the urge to commit robbery, assault, and murder."
"The greatest intellectual gifts sometimes carry with them the inability to perceive simple realities that would be obvious to a moron" (p.422)

Decision theory is not fundamental

"The theory of inference involving priors is more fundamental than that of loss functions (...) Loss functions are less firmly grounded than are prior probabilities (...) In recognising the indefinite and provisional nature of loss functions, we have a more cogent reason for not basing probability theory on decisions."(p.422-425)

Model choice and model comparison

> *"A false premise built into a model that is never questioned cannot be removed by any amount of new data."*(p.601)

Choice of models

Several models available for the same observation

$$\mathfrak{M}_i : x \sim f_i(x|\theta_i), \qquad i \in \mathfrak{I}$$

where $\mathfrak{I}$ can be finite or infinite

Bayesian resolution

B Framework

Probabilises the entire model/parameter space

Bayesian resolution

B Framework

Probabilises the entire model/parameter space

This means:

- allocating probabilities p_i to all models $\mathfrak{M}_i$
- defining priors $\pi_i(\theta_i)$ for each parameter space Θ_i

Formal solutions

Resolution

1. Compute

$$p(\mathfrak{M}_i|x) = \frac{p_i \int_{\Theta_i} f_i(x|\theta_i)\pi_i(\theta_i)\mathrm{d}\theta_i}{\sum_j p_j \int_{\Theta_j} f_j(x|\theta_j)\pi_j(\theta_j)\mathrm{d}\theta_j}$$

Formal solutions

Resolution

1. Compute

$$p(\mathfrak{M}_i|x) = \frac{p_i \int_{\Theta_i} f_i(x|\theta_i)\pi_i(\theta_i)\mathrm{d}\theta_i}{\sum_j p_j \int_{\Theta_j} f_j(x|\theta_j)\pi_j(\theta_j)\mathrm{d}\theta_j}$$

2. Take largest $p(\mathfrak{M}_i|x)$ to determine "best" model, or use averaged predictive

$$\sum_j p(\mathfrak{M}_j|x) \int_{\Theta_j} f_j(x'|\theta_j)\pi_j(\theta_j|x)\mathrm{d}\theta_j$$

Ockham factor

Posterior odds ratio comparing model $\mathfrak{M}_i$ with model $\mathfrak{M}_j$

$$\frac{p(\mathfrak{M}_i|x)}{p(\mathfrak{M}_j|x)} = \frac{p(\mathfrak{M}_i)}{p(\mathfrak{M}_j)} \times \frac{p(x|\mathfrak{M}_i)}{p(x|\mathfrak{M}_j)}$$

Ockham factor

Posterior odds ratio comparing model $\mathfrak{M}_i$ with model $\mathfrak{M}_j$

$$\frac{p(\mathfrak{M}_i|x)}{p(\mathfrak{M}_j|x)} = \frac{p(\mathfrak{M}_i)}{p(\mathfrak{M}_j)} \times \frac{p(x|\mathfrak{M}_i)}{p(x|\mathfrak{M}_j)}$$

Differs from Bayes factor

$$B_{ij}(x) = \frac{p(x|\mathfrak{M}_i)}{p(x|\mathfrak{M}_j)}$$

Who's Occam?

Pluralitas non est ponenda sine neccesitate

William d'Occam (ca. 1290–ca. 1349)

William d'Occam or d'Ockham was a English theologian (and a Franciscan monk) from Oxford who worked on the bases of empirical induction, nominalism and logic and, in particular, posed the above principle later called *Occam's razor*. Also tried for heresy in Avignon and excommunicated by John XXII.

Ockham factor (2)

> *"Some 650 years ago the Franciscan monk William of Ockham perceived the logical error in the mind projection fallacy: Entities are not to be multiplied without necessity."* (p.601)

Jaynes defines the Ockham factor for model $\mathfrak{M}_i$ as

$$\begin{aligned} W_i &= p(x|\mathfrak{M}_i) / \max_\theta L_i(\theta|x) \\ &= \int \frac{L_i(\theta_i|x)}{\max_\theta L_i(\theta|x)} p(\theta_i|\mathfrak{M}_i)\,\mathrm{d}\theta_i \end{aligned}$$

An explanation of Ockham's razor

Special case when $\mathfrak{M}_2$ corresponds to an $(n+1)$-dimensional parameter θ and when $\mathfrak{M}_1$ corresponds to the special case $\theta_{n+1} = 0$
If likelihood concentrated on subsets $\Theta_1' \subset \Theta_1$ and $\Theta_2' \subset \Theta_2$ the subset Θ_2' is getting less prior probability than Θ_1' because $p(\theta|\mathfrak{M}_2)$ spread over a larger space.
Therefore

$$p(x|\mathfrak{M}_2) < p(x|\mathfrak{M}_1)$$

in the case of the smaller model agreeing with the data.

Regression example

Opposition of

$$\mathfrak{M}_1 : \; y_i = \alpha x_i + e_i \quad e_i \sim \mathcal{N}(0, \sigma^2)$$

and

$$\mathfrak{M}_2 : \; y_i = \alpha x_i + \beta x_i^2 + e_i \quad e_i \sim \mathcal{N}(0, \sigma^2)$$

Under $\alpha \sim \mathcal{N}(\alpha_0, \sigma_0^2)$ prior

$$W_1 = \frac{1}{\sqrt{1 + \sigma_0^2 n\bar{x}^2/\sigma^2}} \exp\left\{ -\frac{(\hat{\alpha} - \alpha_0)^2}{1 + \sigma_0^2 n\bar{x}^2/\sigma^2} \right\}$$

[Anticlimactic!]

Regression example

Opposition of

$$\mathfrak{M}_1 : \; y_i = \alpha x_i + e_i \quad e_i \sim \mathcal{N}(0, \sigma^2)$$

and

$$\mathfrak{M}_2 : \; y_i = \alpha x_i + \beta x_i^2 + e_i \quad e_i \sim \mathcal{N}(0, \sigma^2)$$

and under $\beta \sim \mathcal{N}(\beta_0, \sigma_1^2)$ additional prior

$$W_2 = \frac{\sigma_0^{-1} \sigma_1^{-1} \exp\{x\}}{\sqrt{(\sigma_0^{-2} + n\sigma^{-2}\overline{x^2})(\sigma_1^{-2} + n\sigma^{-2}\overline{x^4})}}$$

where x remains undefined in Jaynes (p.613)

[Anticlimactic!]

[Last] comments

"Actual scientific practice does not really obey Ockham's razor, either in its previous 'simplicity' form or in our revised 'plausibility' form (...) In any field, the Establishement is seldom in pursuit of the truth, because they sincerely believe it is composed of those who sincerely believe that they are already in possession of it."(p.613)

[Last] comments

“Actual scientific practice does not really obey Ockham’s razor, either in its previous ‘simplicity’ form or in our revised ‘plausibility’ form (...) In any field, the Establishement is seldom in pursuit of the truth, because they sincerely believe it is composed of those who sincerely believe that they are already in possession of it.”(p.613)

The End